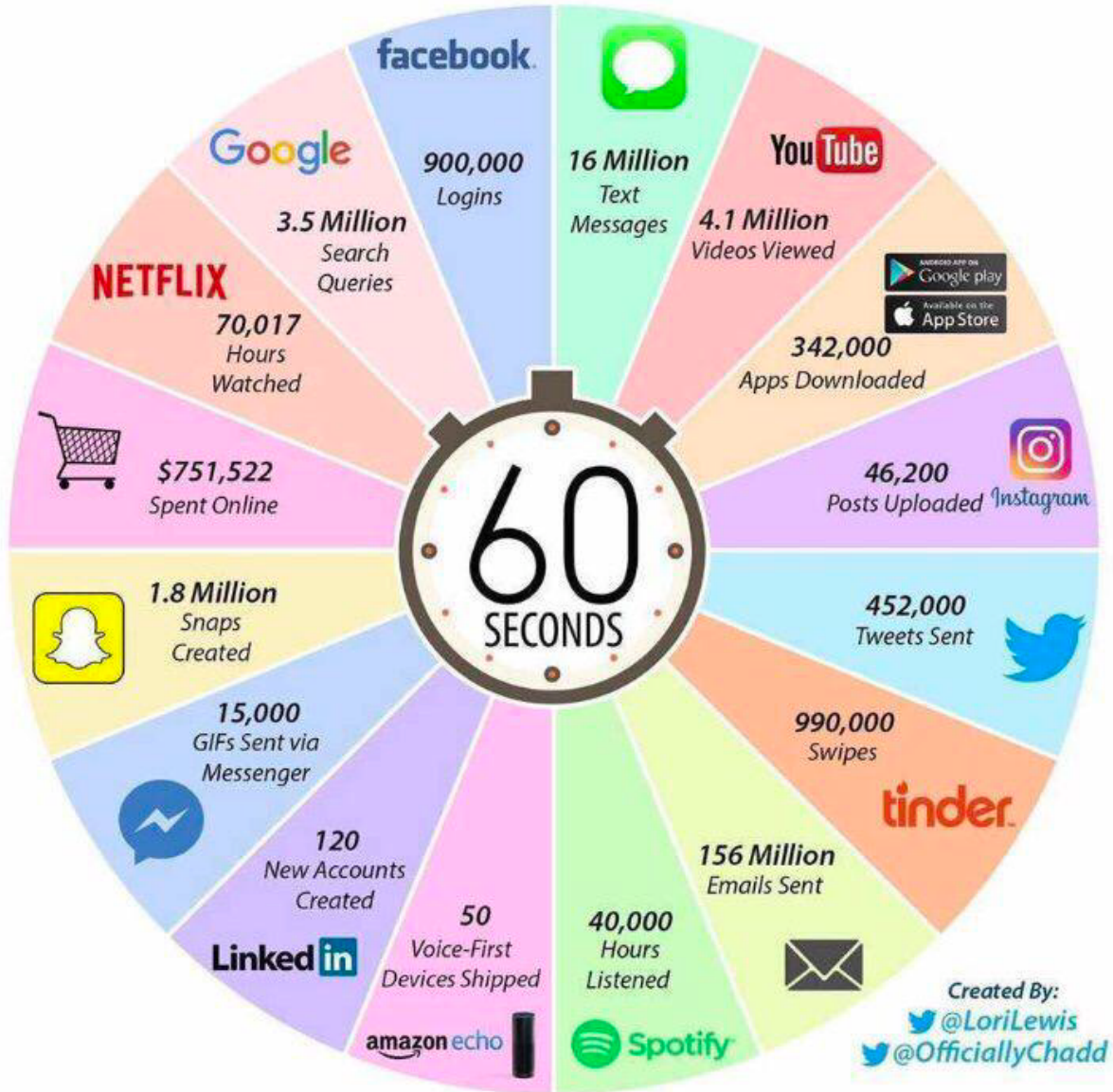
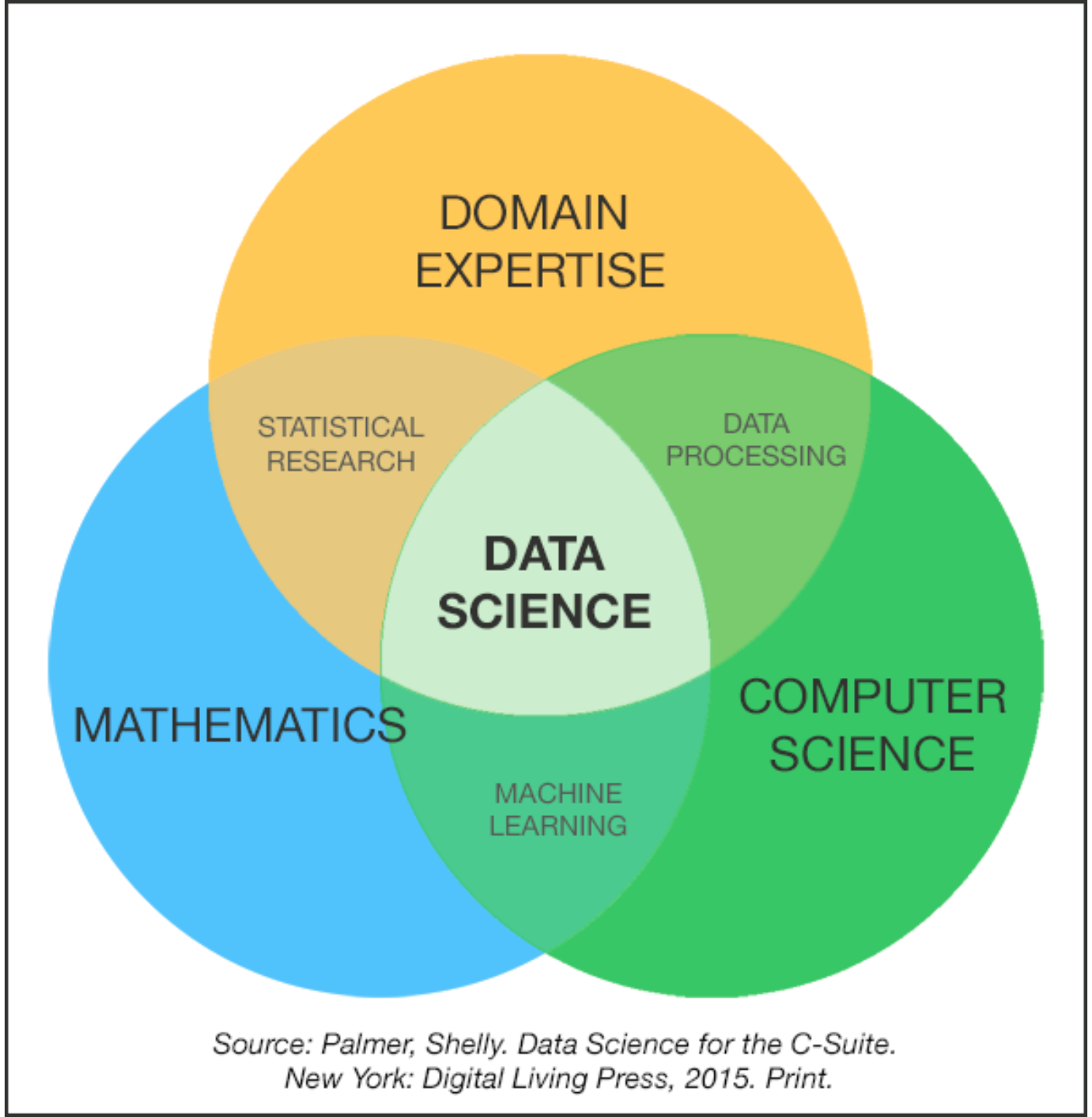
Machine Learning

Agenda

- Conceptos de Big Data
- Machine Learning
- Lenguaje R
- Práctica de entrenamiento de modelos predictivos en R

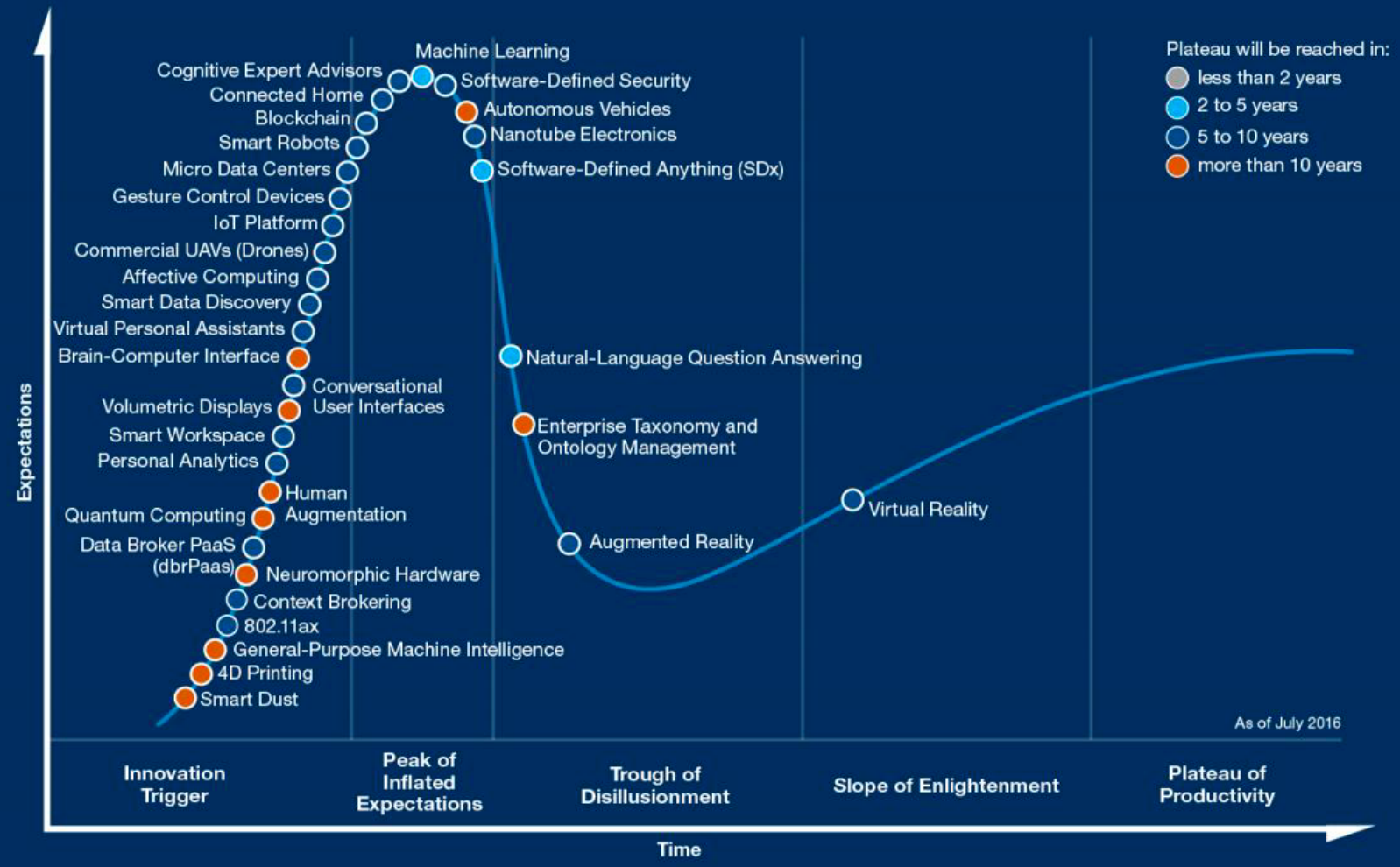
2017 This Is What Happens In An Internet Minute





Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

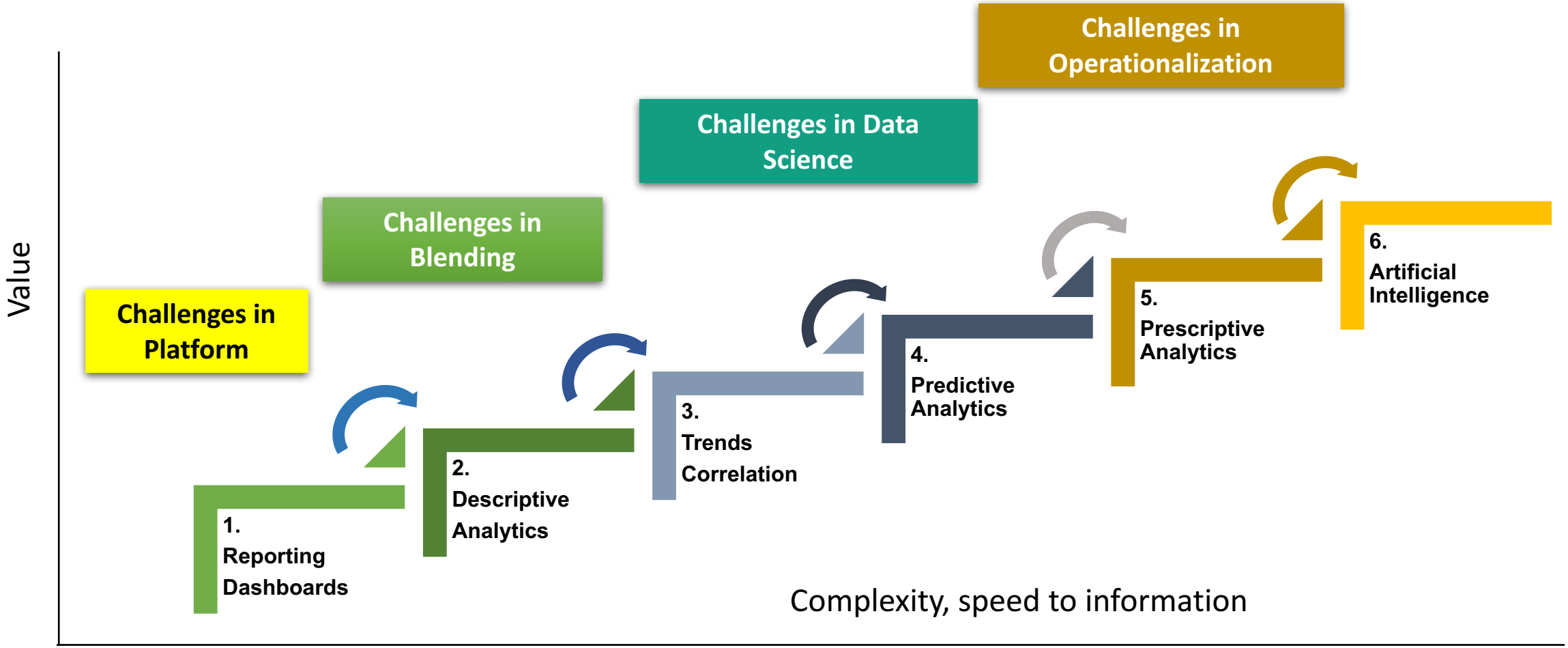
Gartner Hype Cycle for Emerging Technologies, 2016



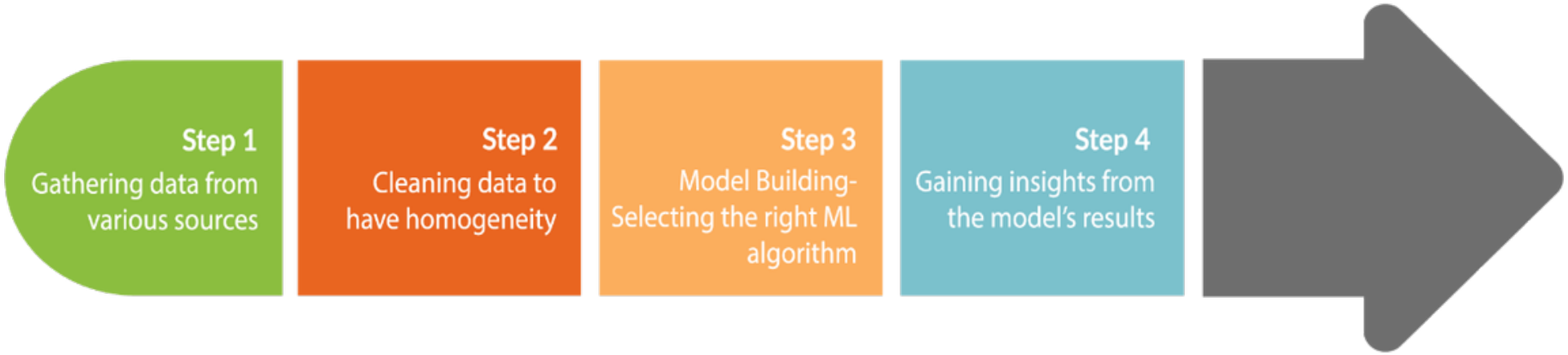
Source: Gartner

Gartner

Data Understanding Analytics Maturity

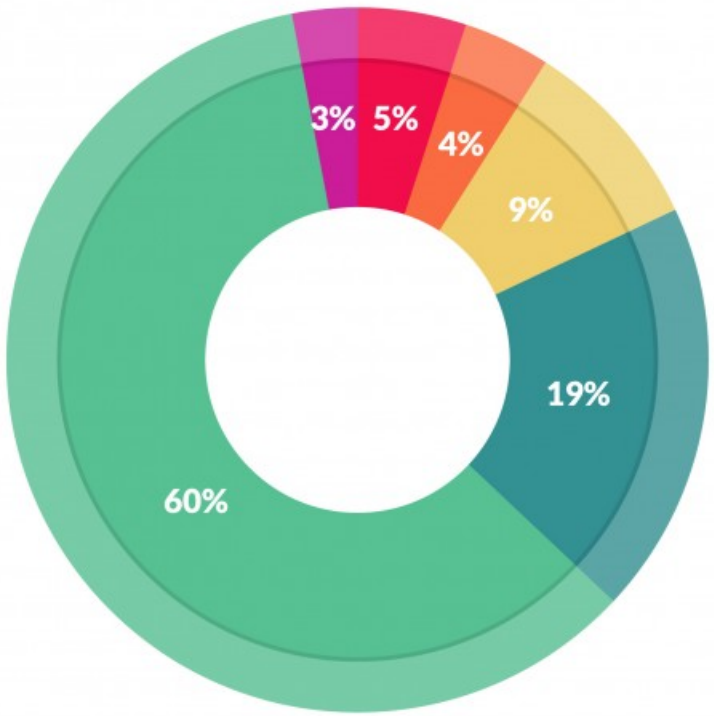


Proceso de Machine Learning



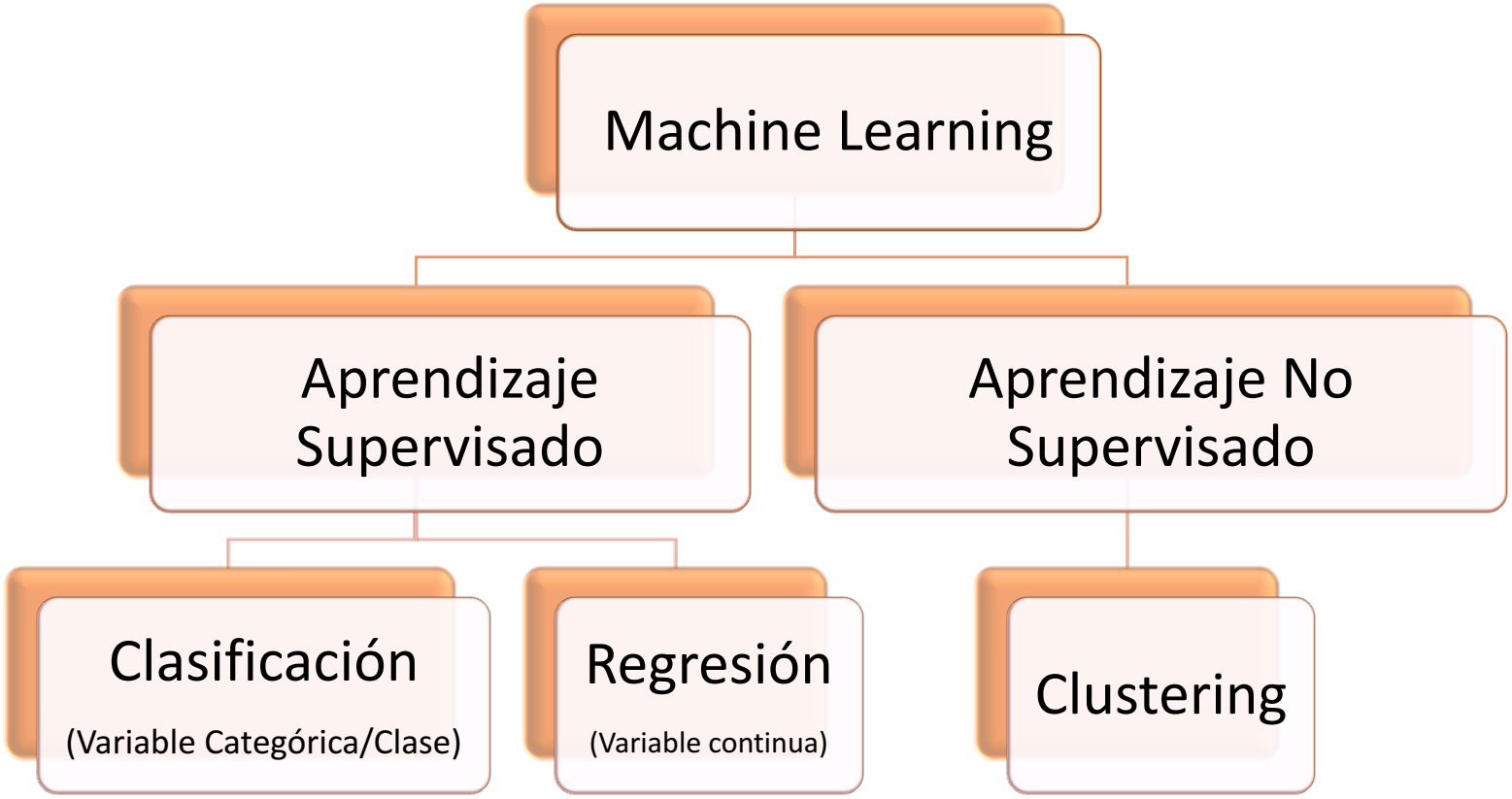
- Gathering data
- Preparing that data
- Choosing a model
- Training
- Evaluation
- Hyper parameter tuning
- Prediction

Data Cleaning



What data scientists spend the most time doing

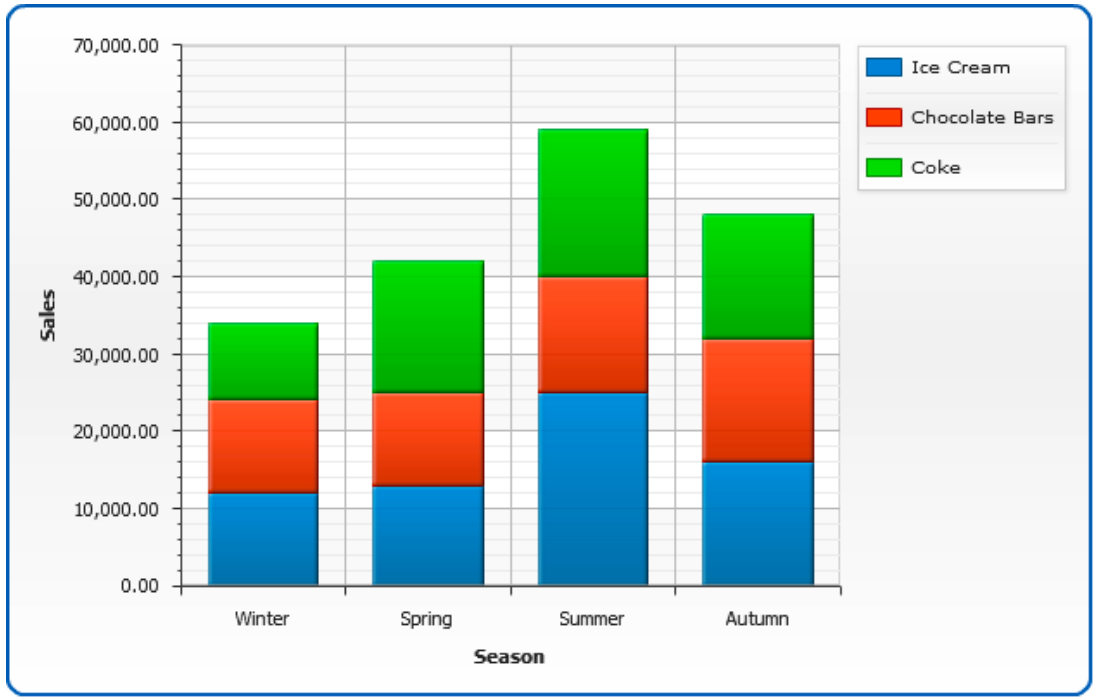
- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*


Machine Learning
Aprendizaje Supervisado
Aprendizaje No Supervisado
Clasificación
(Variable Categórica/Clase)
Regresión
(Variable continua)
Clustering
Regresión Logística
Redes Neuronales
RandomForest
NaiveBayes
XGBOOST
SVMs
Regresión lineal
Redes Neuronales
XGBOOST
RandomForest
SVMs
K-Means
Dendograma

Variables

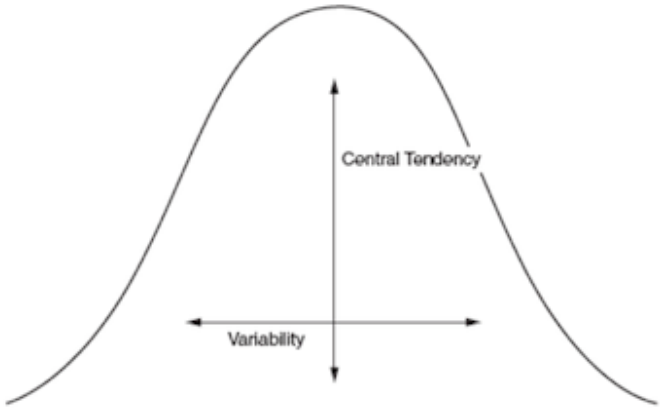
Variables Categóricas / Cualitativas

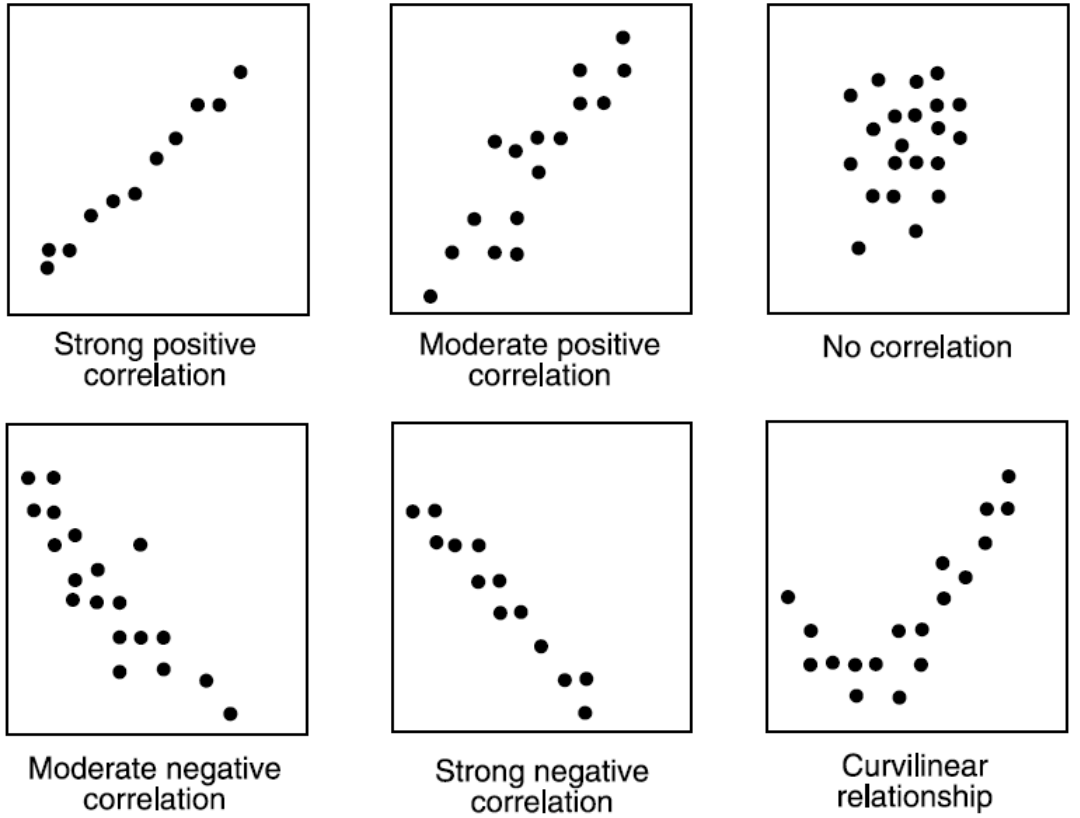
- Binarias: Solo dos opciones. Ej.: SI/NO
- Ordinales: Pueden ser ordenadas lógicamente o “rankeadas”. Ej: Chico, Mediano, Grande.
- Nominales: Ej.: Sexo, Profesión. Frecuencias para medir distribución
- Bi-variado o Multivariado: BarPlot



Variables Numéricas (Cuantitativas)

- Discretas
- Continuas: Estadística Descriptiva, tendencia central, dispersión, Boxplot, Histogramas
- Bi-variado o Multivariado: ScatterPlot, Correlación





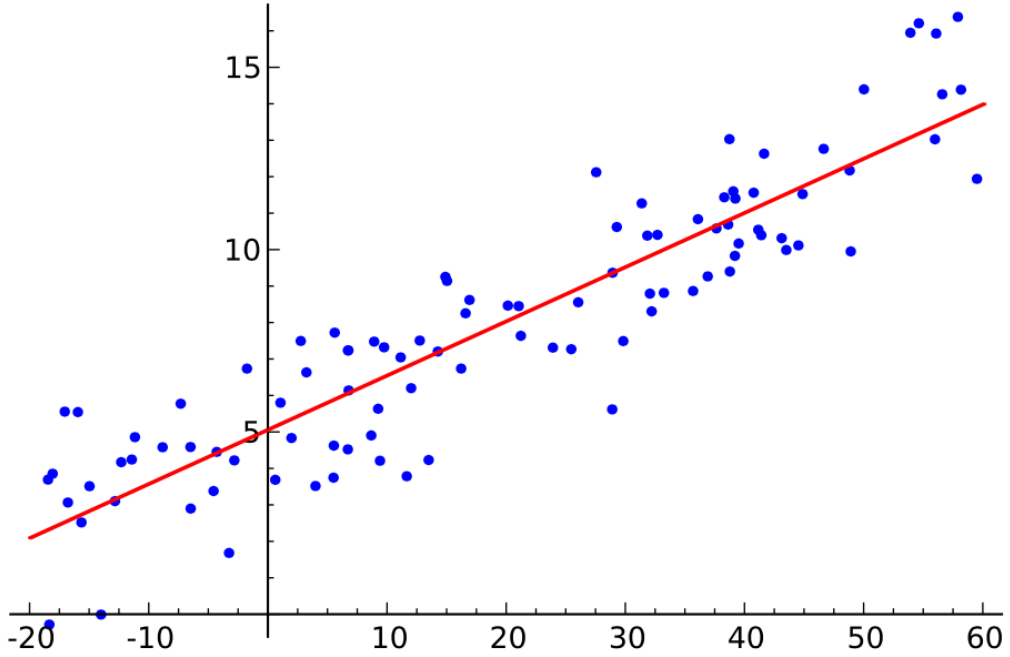
Variables de un dataset

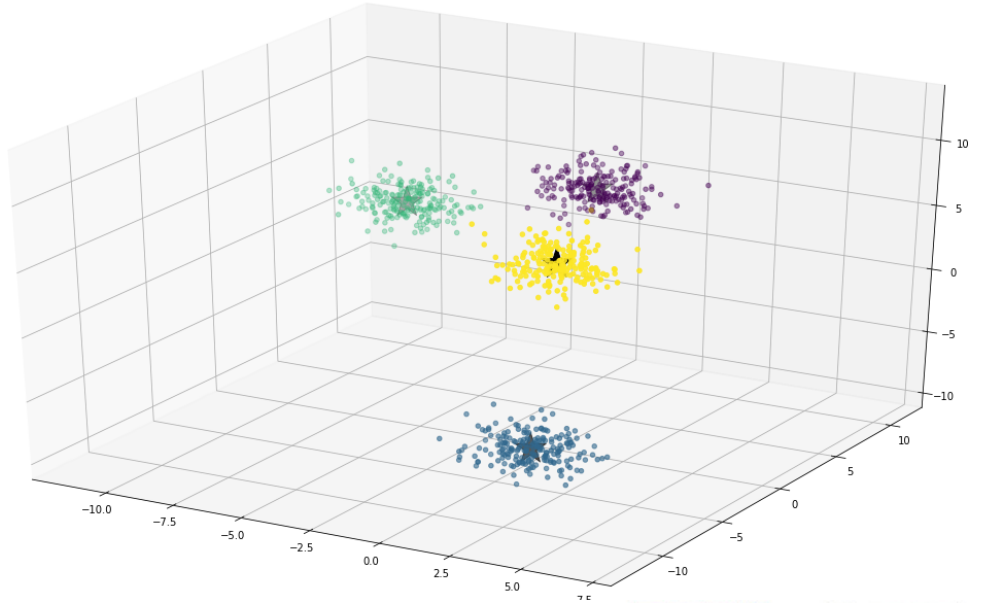
Variable Dependiente

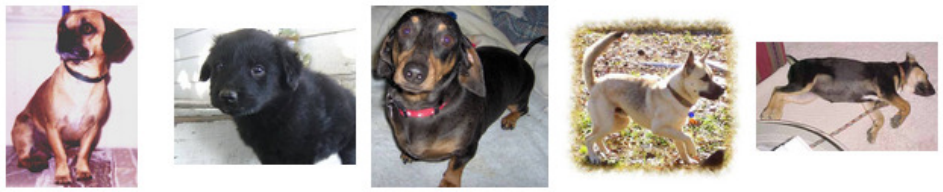
age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	hours-per-week	native-country	income
39	State-gov	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	40	United-States	<=50K
50	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	13	United-States	<=50K
38	Private	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	40	United-States	<=50K
53	Private	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	40	United-States	<=50K
28	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	40	Cuba	<=50K
37	Private	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	40	United-States	<=50K
49	Private	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	16	Jamaica	<=50K
52	Self-emp-not-inc	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	45	United-States	>50K
31	Private	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	50	United-States	>50K
42	Private	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	40	United-States	>50K
37	Private	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	80	United-States	>50K
30	State-gov	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	40	India	>50K
23	Private	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	30	United-States	<=50K
32	Private	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	50	United-States	<=50K
40	Private	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	40	?	>50K
34	Private	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	45	Mexico	<=50K
25	Self-emp-not-inc	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	35	United-States	<=50K
32	Private	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	40	United-States	<=50K
38	Private	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	50	United-States	<=50K
43	Self-emp-not-inc	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	45	United-States	>50K
40	Private	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	60	United-States	>50K
54	Private	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	20	United-States	<=50K

Variables Predictoras

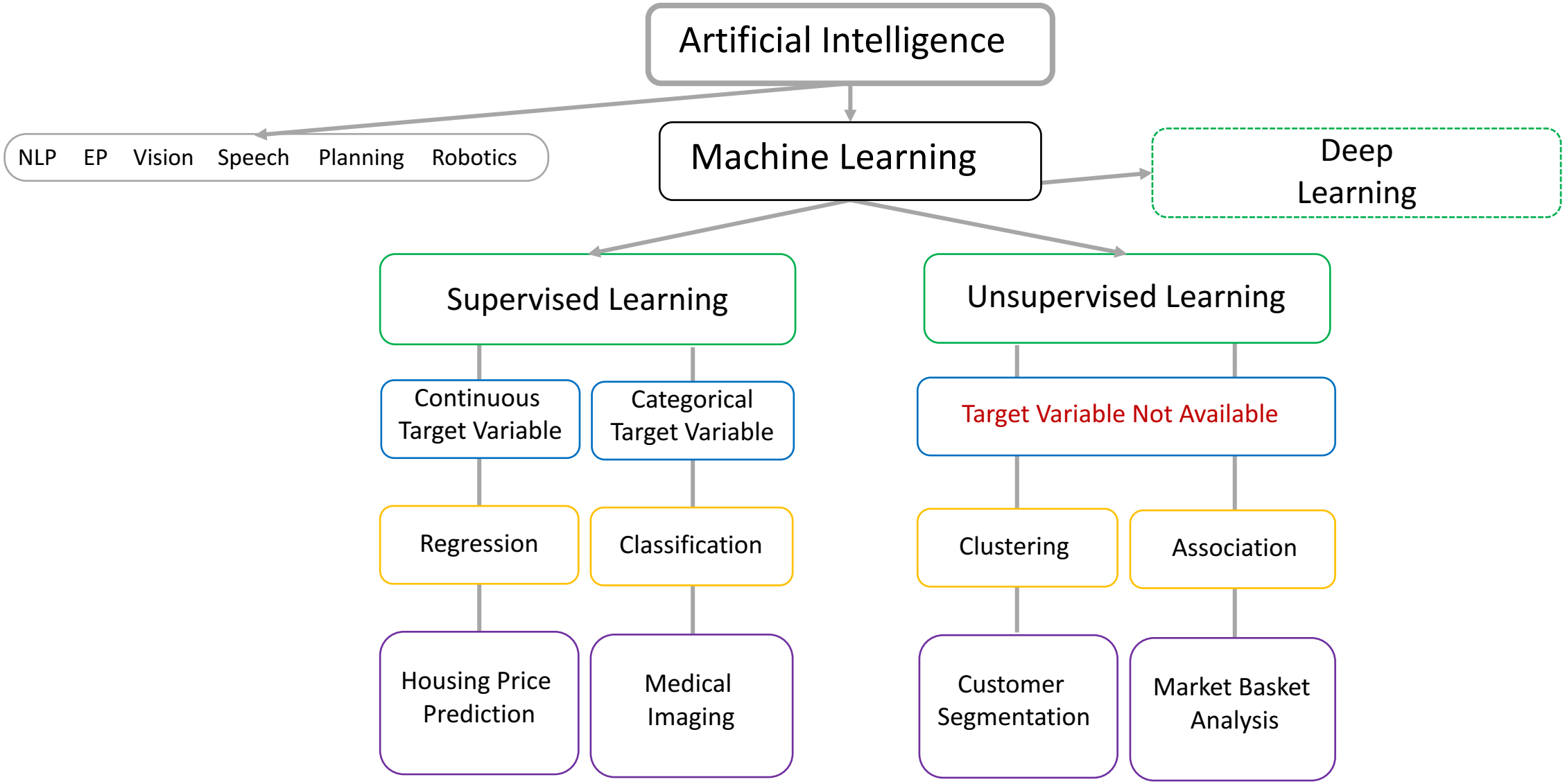
Ejemplos de aplicación de algoritmos


15
10
5
-20
-10
10
20
30
40
50
60


10
5
0
−5
−10
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0

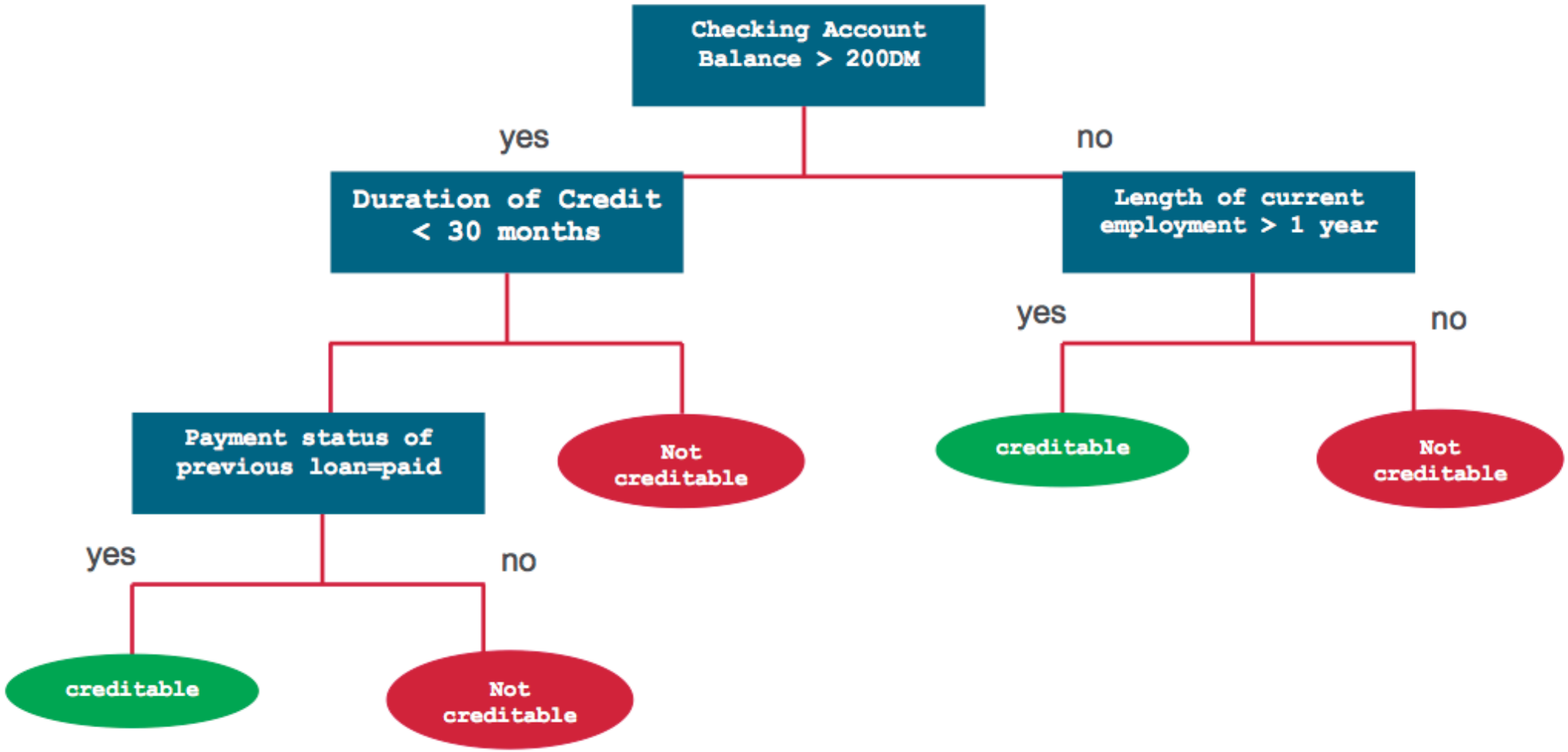


Algunos Casos de Uso

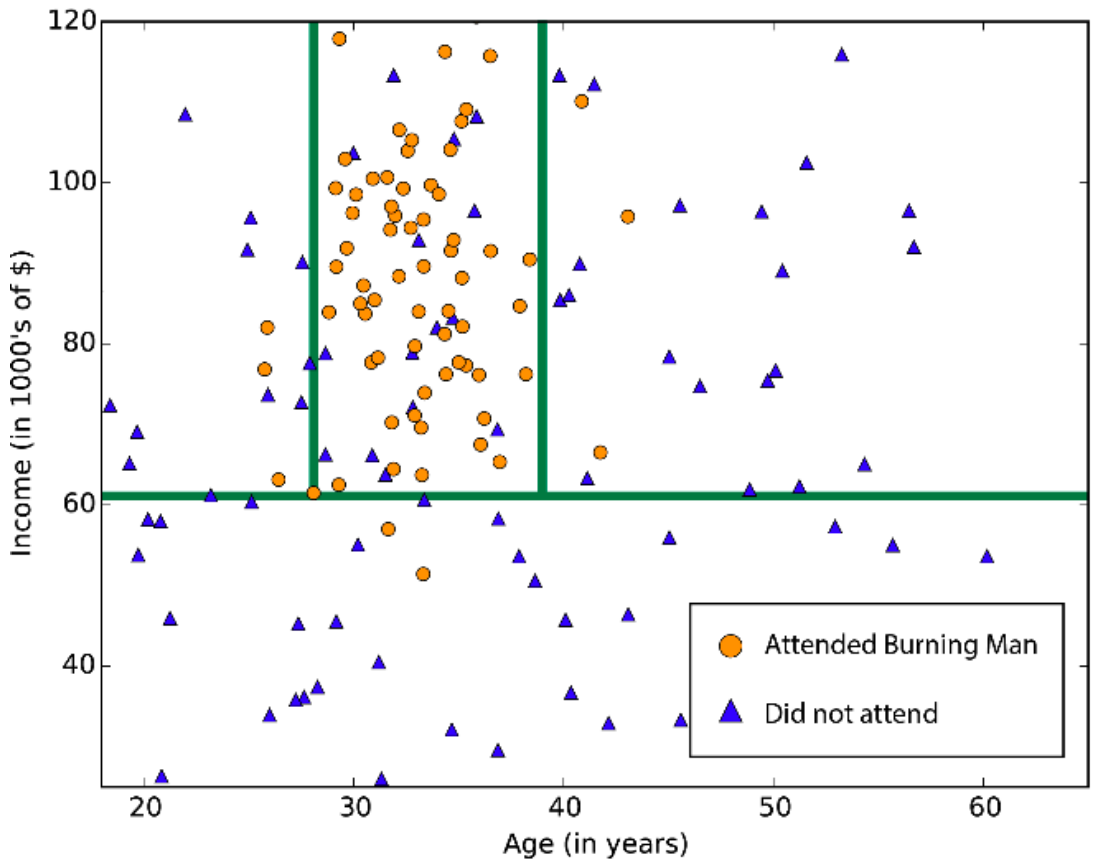


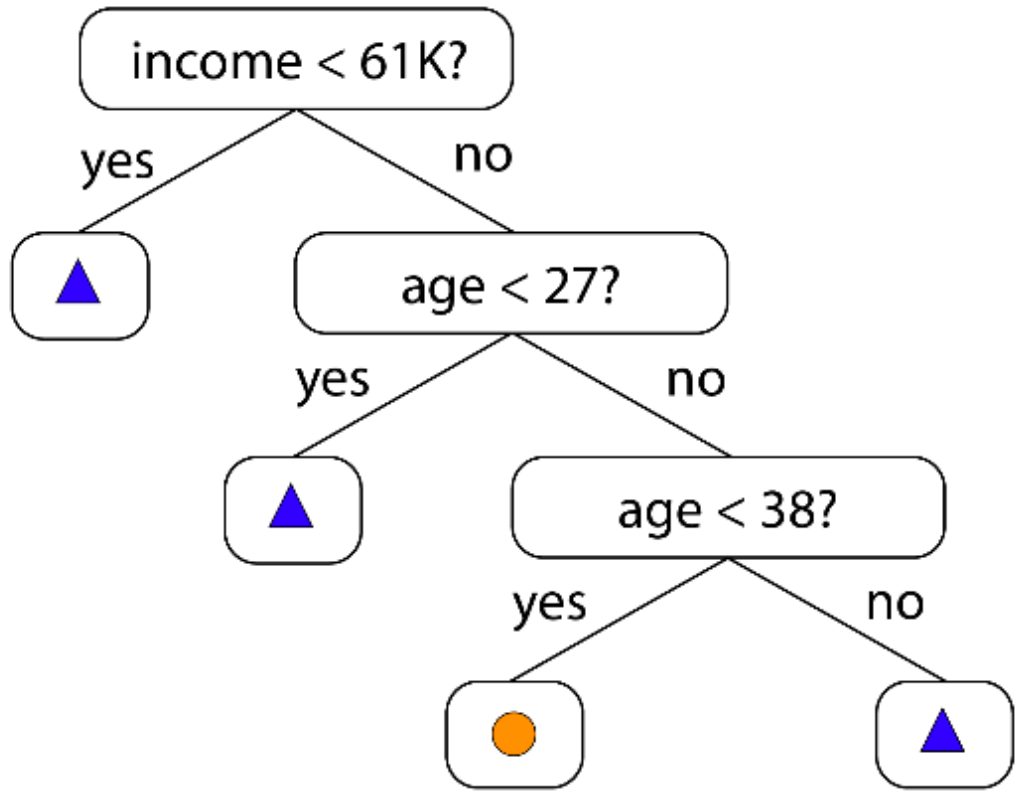
Arboles y Random Forests

Decision Trees

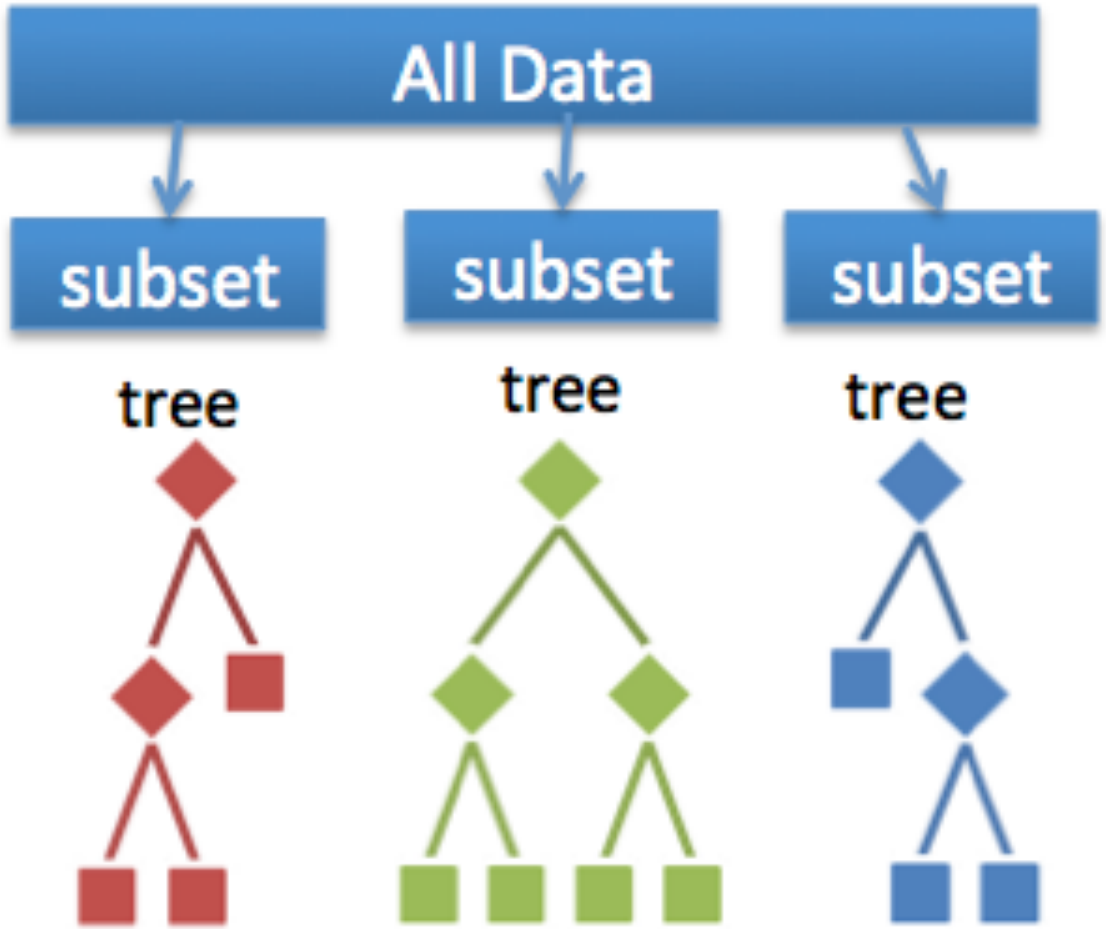


Arboles de decisión



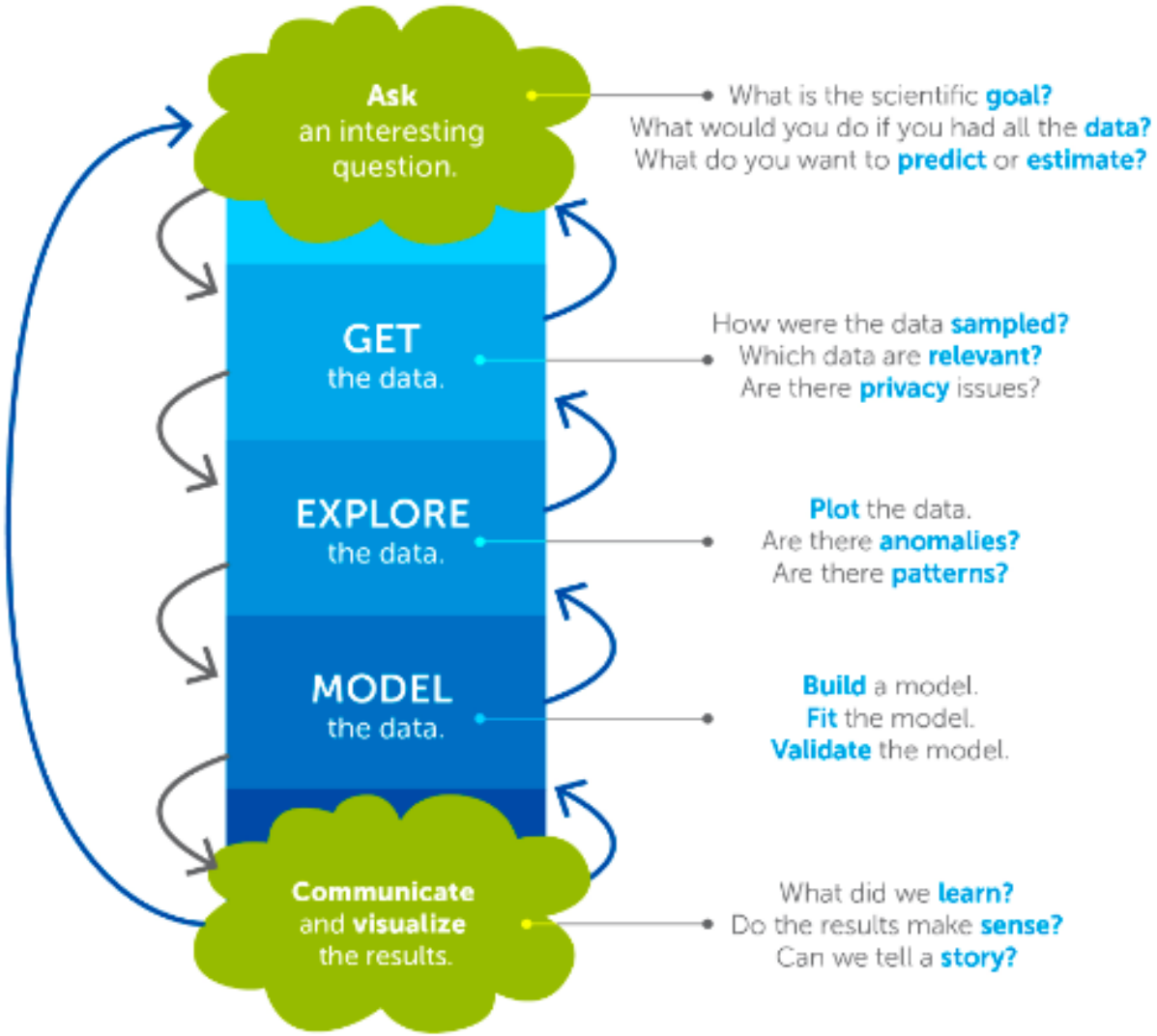


Random Forests

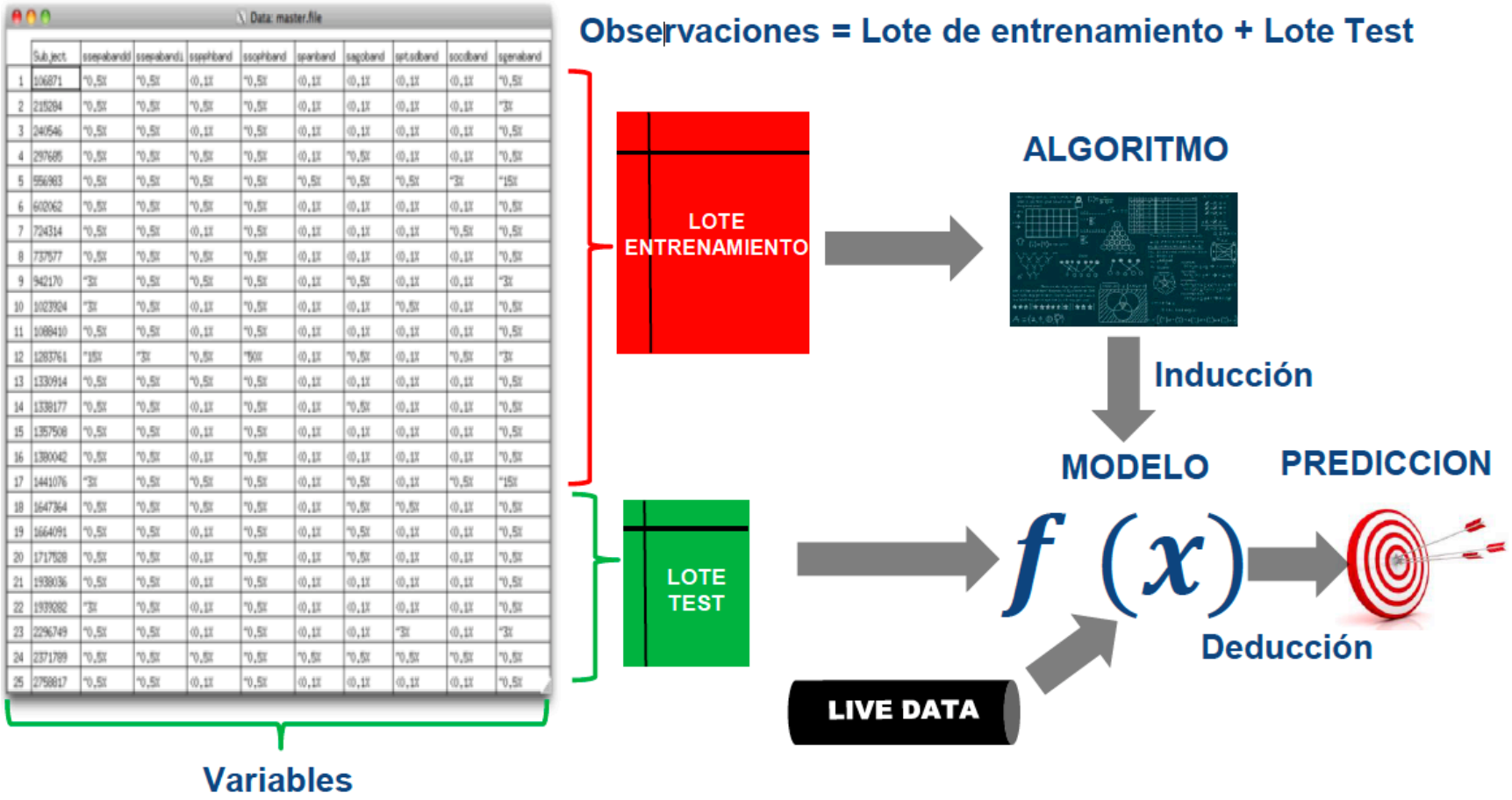


Modelado

Como es el proceso del data scientist?

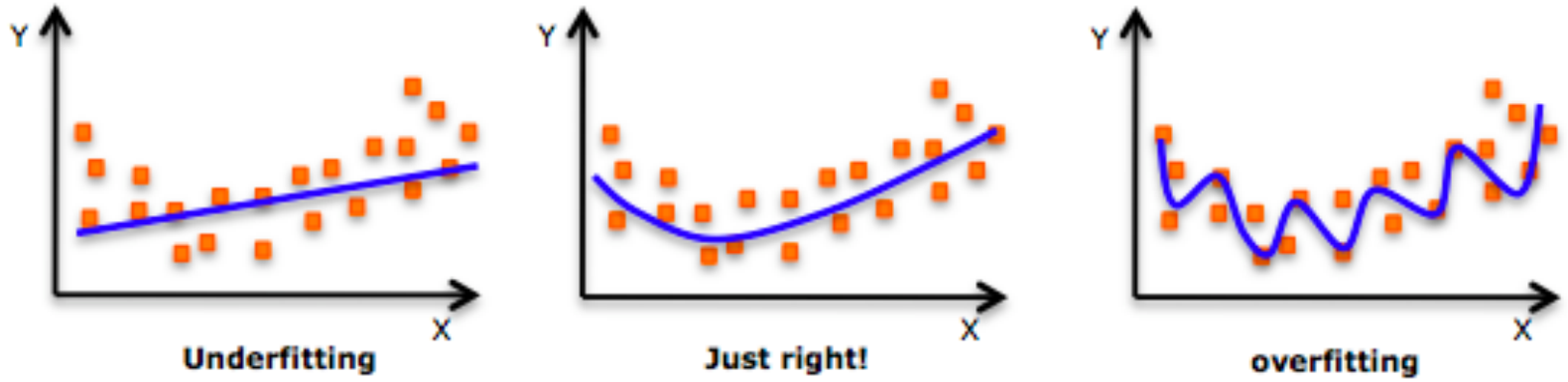


Entrenamiento y Armado del Modelo

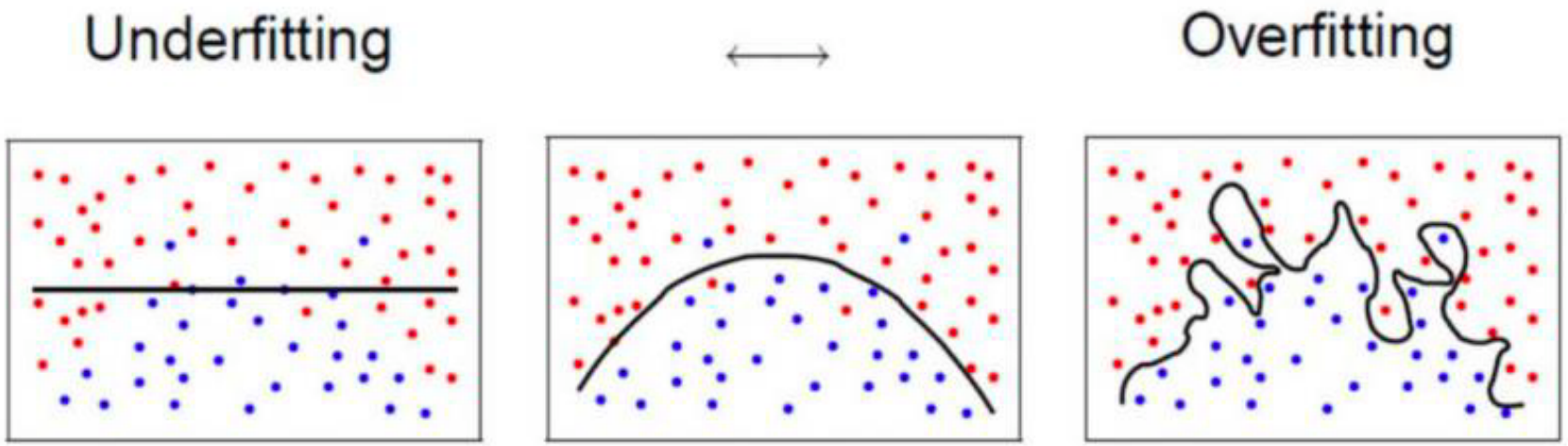


Overfitting y Underfitting

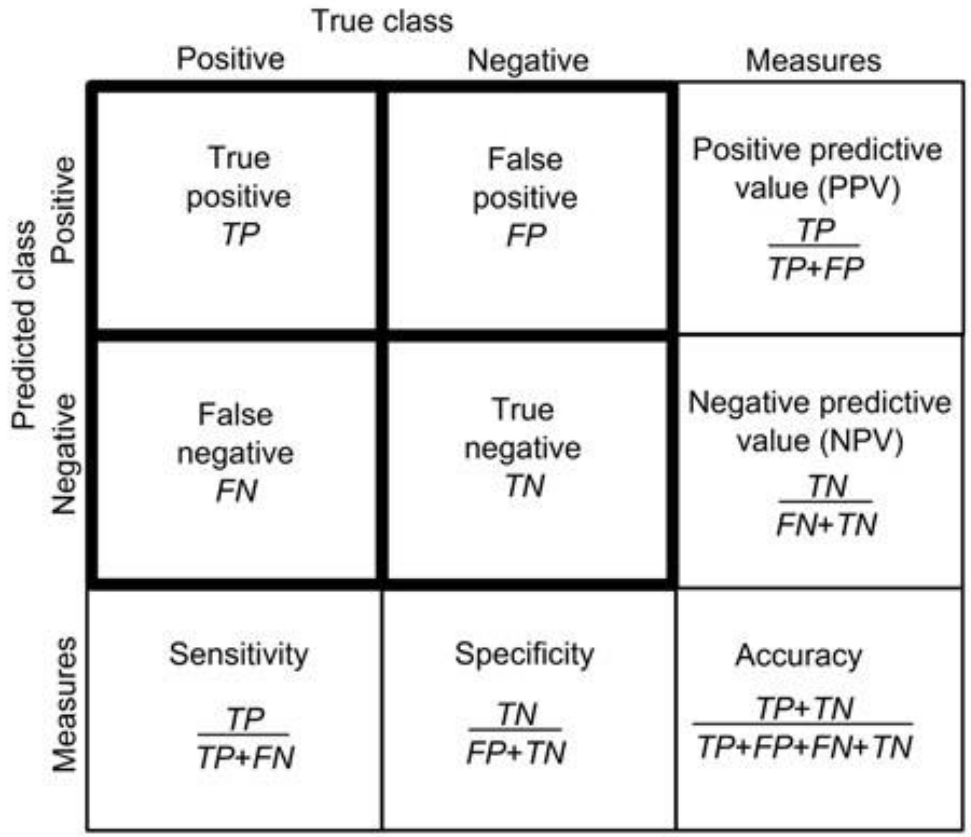
Regresión



Clasificación



Comparando Modelos: Matriz de Confusión



Problemas del Accuracy

A) Clases Desbalanceadas

- Considerar un problema de dos clases con:
  - 9.990 observaciones de clase 0
  - 10 Observaciones de clase 1

Si el modelo predice todo como clase 0 su exactitud es 9.990/10.000 = 99,9 % →La exactitud es engañosa porque el modelo no detecta ningún ejemplo de clase 1 (accuracyparadox)

B) No distingue los FP de los FN →¿Son ambos igual de importantes?

Lenguaje / Plataforma R

Lenguaje R

- Qué es el Lenguaje R?
  - Sistema Estatístico, Matemático y Computacional que auxilia profesionales de diversas áreas del conocimiento. Busca respuestas para grandes volumes de datos a través de Gráficos, Análisis y clasificación de datos

- Por qué utilizar R?
  - R posee muchos componentes, paquetes y facilidades para un buen análisis estatístico de datos
  - Es software libre

CRAN

- Comprehensive R Archive Network

- Website responsable de almacenar y tener disponible los códigos, paquetes y manuales.

- R: http://www.r-project.org
- Rstudio: http://www.rstudio.com

Ayuda en R

- Ayuda en HTML
  - Paquetes
  - Puede darse uma palabra clave y se obtiene información
  - help(keyword)

Operaciones básicas en R

Operadores : “+”, “-”, “*”, “/”, “^”.

Función	Descripción
`sqrt()`	raiz quadrada
`abs()`	valor absoluto
`exp()`	Exponencial
`log10()`	logaritmo en base 10
`log()`	Logaritmo en base e
`sin() cos() tan()`	Funciones trigonométricas
`asin() acos()` `atan()`	Funciones trigonométricas inversas
`mean()`	Média aritmética
`length()`	Número de elementos
`max(), min()`	Mayor y menor valor

Características de las funciones

Función (argumento(s) obligatorio(s), argumento(s) opcional(es))

Ejemplo: `log(2187, base=3)`

o simplemente `log(2187, 3)`

Algunos comandos importantes

Visualización de objetos: `ls()`

Remoción de objetos: `rm()`

Para eliminar todo lo que hay en memoria:

```
rm(list=ls())
```

Vectores

- Tipo de objeto más simples
  - Numérico
  - Caracter “ ”
- R deja reescribir el objeto sin advertir!!!
- varios valores
  - `objeto <- c(conjunto de valores separados por comas)`
  - `objeto<-c(1,2,5,0)`

Generando Sequencias de números

objeto <- seq(from=*inicio*, to=*fin*)

Se puede omitir la descripción de parámetros

objeto <- seq(*inicio*, *fin*)

O:

objeto <- *inicio*:*fin*

Seleccionando una posición en un vector:

vector[*valor o vector de posiciones*]

Ejemplo: objeto[42]

Operaciones Vectoriales

- Ejemplo:

```
peso <- c(62, 70, 52, 98, 90, 70)
altura <- c(1.70, 1.82, 1.75, 1.94, 1.84,
            1.61)
```

Calculando el índice de masa corporal (IMC)...

```
imc <- peso/altura^2
imc
[1] 21.45329 21.13271 16.97959 26.03890
    26.58318 27.00513
```

Vectores Lógicos

```
x <- c(1:6)
  y <-(x <= 4)
    [1]  TRUE  TRUE  TRUE  TRUE FALSE FALSE
```

- Otros valores posibles para una variable:
  - NA (not avaliable): valor no disponble
    - Función is.na(x)
  - NaN (not a number): indeterminaciones matemáticas.
    - Ejemplo: 0/0

Matrices

```
objeto <- matrix(conjunto de valores
     (vector), ncol = número de colunas)
```

Obteniendo valores de una matriz:

```
matriz[fila, columna]  o
matriz[intervalo, columna]
   o
matriz[fila,]
```

Data frames

"Tabla de datos" donde:

- Columnas: variables
- Líneas: registros

Se accede a los datos igual a una matriz

`names(`*`data frame`*`): variables`

Acceso a datos: *`data frame`*`[,5]` (5ª variable)

o *`dataframe$nombre_de_la_variable`*

Sustitución de valores:

`dataframe$variable``[`*`condición`*`] = `*`valor`*

Manipulación de Paquetes

- Instalación
  - Install packages from CRAN
  - Install packages from .zip files
- Atualización
  - Update packages from CRAN
- Lectura
  - Load package
  - library(nombre_del_paquete)

Manejo de Archivos

Entrada

```
read.table(“Test.txt”)
read.csv(“Test.csv”)
source(“Test.r”)
Save(“nombre.Rdata”)y load para cargarlo
```

Salida

```
write.table(iris,file="Test.csv",sep=",",col.name=NA)
```

Programando en R

buscarmaximo

```
buscarmaximo <- function(vector) {
  mayor = -1
  for (i in 1:length(vector)) {
    if (vector[i] > mayor)
      mayor = vector[i]
  }
  return(mayor)
}
```

Código en R

```
vet = c(3,6,8,9,1,2)
 buscarmaximo(vet)
 [1] 9
```

Algunas operaciones útiles

Funciones cbind rbind y t. Otra forma de obtener matrices.

```
col1<-c(1,2,3)
col2<-c(0,1,1)
A<-cbind(col1,col2)
 A
col1 col2
[1,] 1 0
[2,] 2 1
[3,] 3 1
```

“Pegamos” los vectores en columnas

Algunas operaciones útiles

Funciones cbind rbind y t. Otra forma de obtener matrices.

```
B<-rbind(col1,col2)
B
[,1] [,2] [,3]
col1 1 2 3
col2 0 1 1
```

“Pegamos” los vectores en filas

Función apply

La función apply permite realizar un cálculo por fila ó por columna.

```
> A
[,1] [,2] [,3] [,4] [,5]
[1,] 1 5 9 13 17
[2,] 2 6 10 14 18
[3,] 3 7 11 15 19
[4,] 4 8 12 16 20
> apply(A,2,sum) #Sumamos las columnas de la matriz A
[1] 10 26 42 58 74
```

Función apply

```
> apply(A,1,sum) #Sumamos las filas de la matriz A
[1] 45 50 55 60
```

Si A es una matriz de n x p entonces

```
> apply(A,2,sum)
```

dará un vector de longitud p (cantidad de columnas) con las sumas de las columnas. El argumento 2 en "apply(A,2,sum)" indica que el cálculo debe realizarse en la segunda dimensión.

Función apply

```
> apply(A,1,sum)
```

dará un vector de longitud n (cantidad de filas) con las sumas de las filas.

Con la función apply() se pueden calcular varianzas, medianas, sumas, productos

```
> apply(A,2,mean) # medias por columnas
> apply(A,1,var) # varianzas por filas
> sqrt(apply(A,2,var)) # desvíos estándar por columna
```

Gráficos en R

- R posee algunos gráficos utilizados para análisis de datos:
    - Scatterplot
    - Histograma
    - Boxplot

Scatterplot

- Parámetros principales:

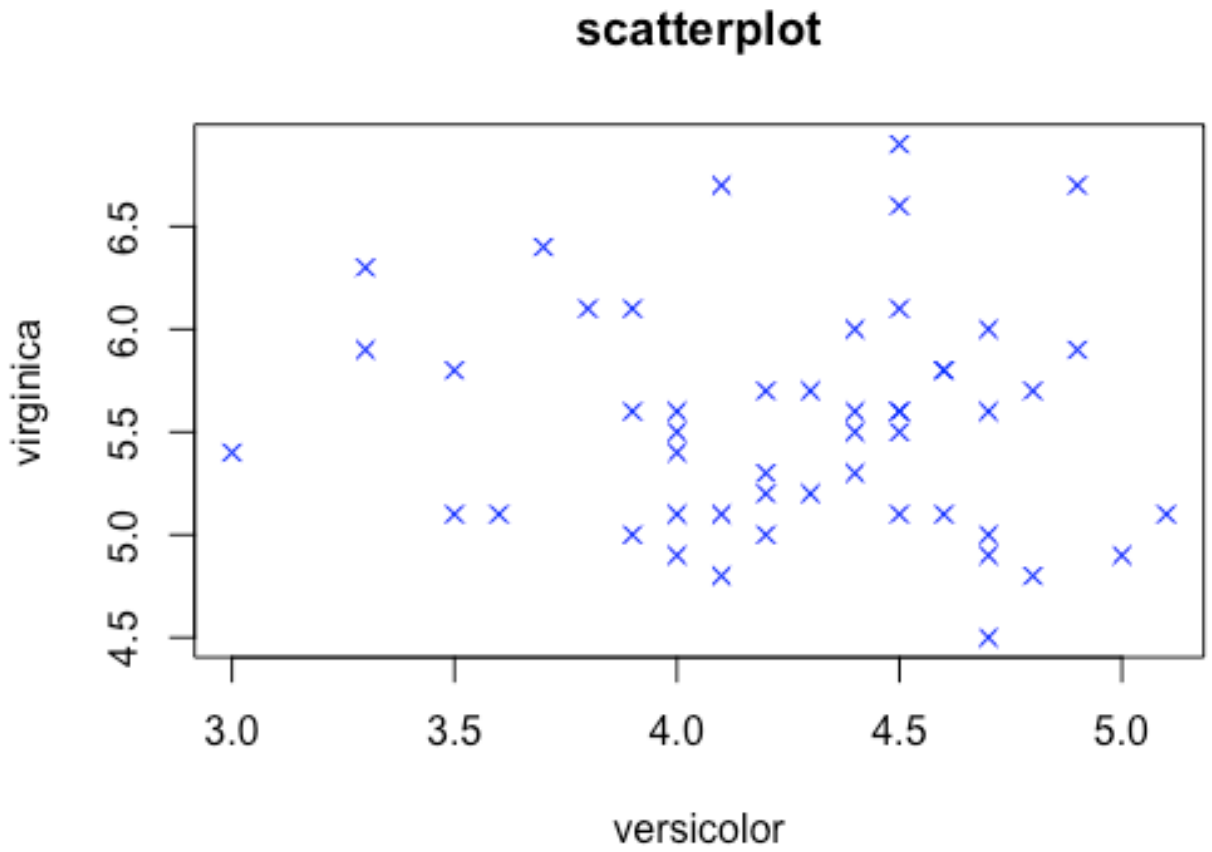
- col: selecciona el color
- pch: selecciona el caracter a ser dibujado
- cex: tamaño del caracter
- main: título principal
- xlab: label para el eje x
- ylab: label para el eje y

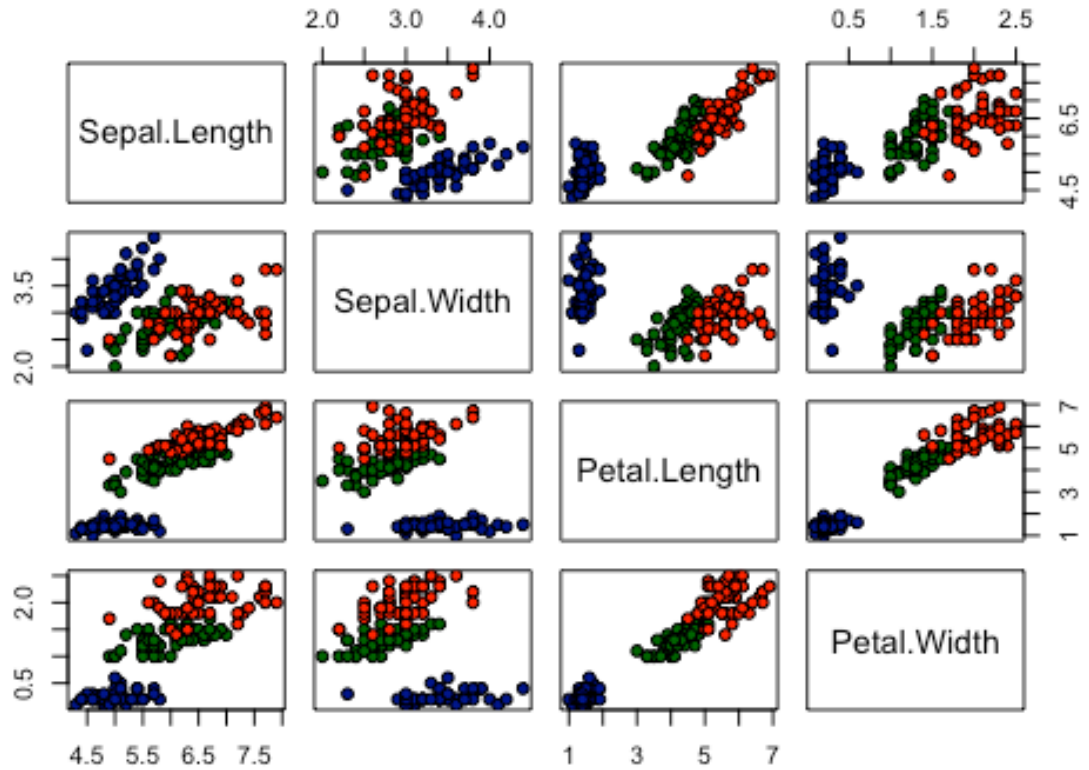
type	
“p”	Puntos
“l”	Líneas
“o”	Ambos
“n”	Nada

- Ejemplo:

```
plot(iris$Sepal.Length, iris$Sepal.Width, col="blue", pch=2, cex=1,
main="Aprendiendo scatterplot", xlab="vectorX", ylab="vectorY",type='p')
```

Scatterplot





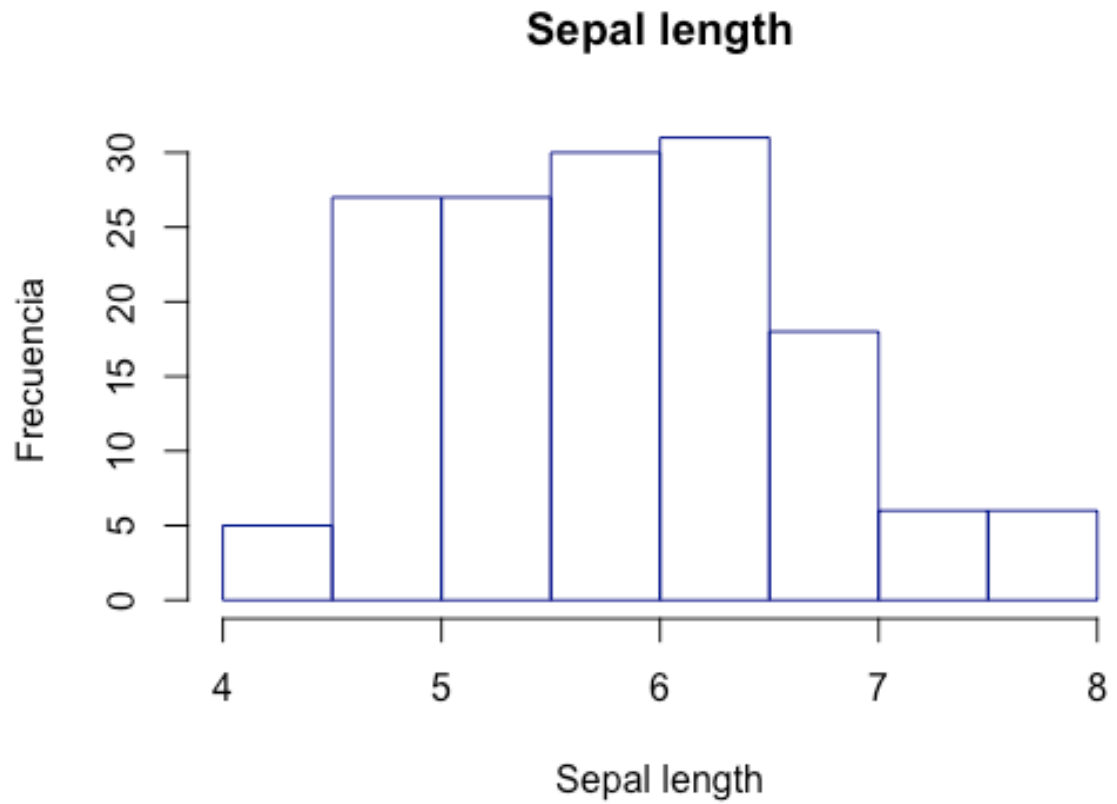
Histograma

- Dada una muestra X, el histograma es la frecuencia de cada una de las muestras dentro de un intervalo.

```
hist(x, nclass, breaks, plot=TRUE, angle, density, col, inside) # todos los parámetros posibles, hay que darle valores.
```

```
hist(iris$Sepal.Length, xlab = "Sepal length", ylab = "Frecuencia", main = "Sepal length", border = "darkblue", breaks = 10)
```

Histograma

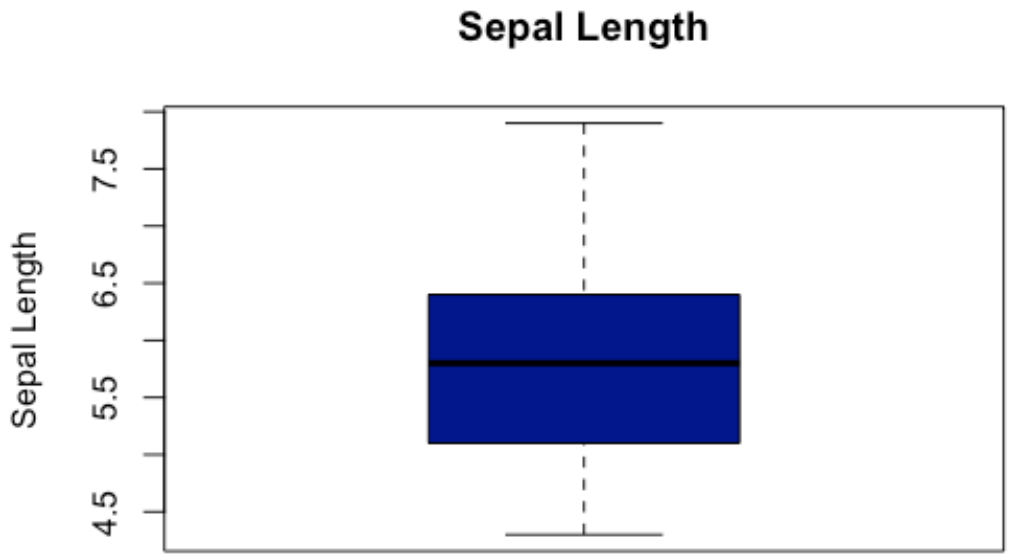


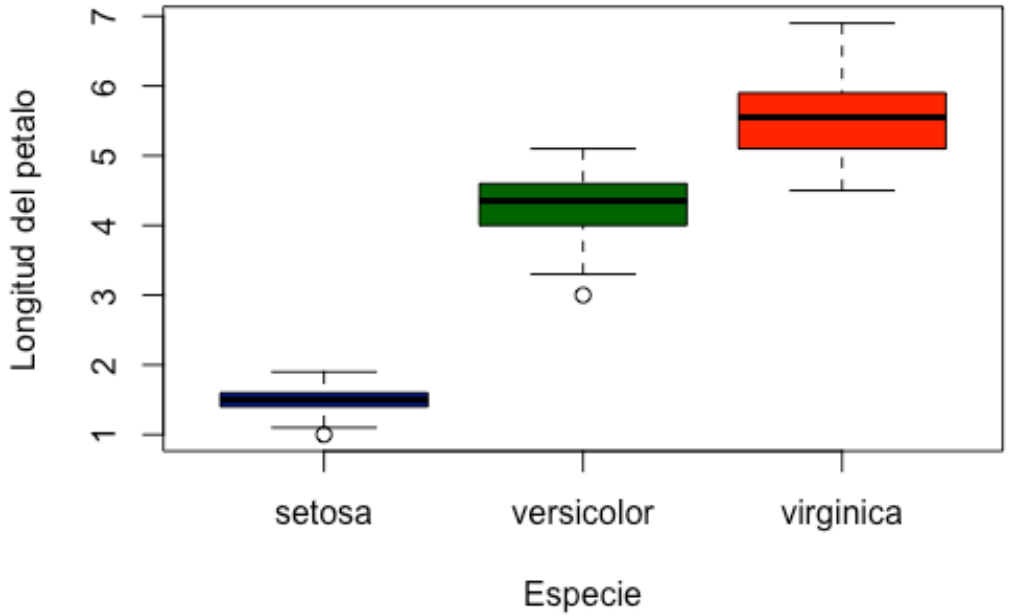
Boxplot

- Dibuja las columnas de una matriz contra las columnas de otra.
- Permite observar de una forma clara la distribución de los datos.

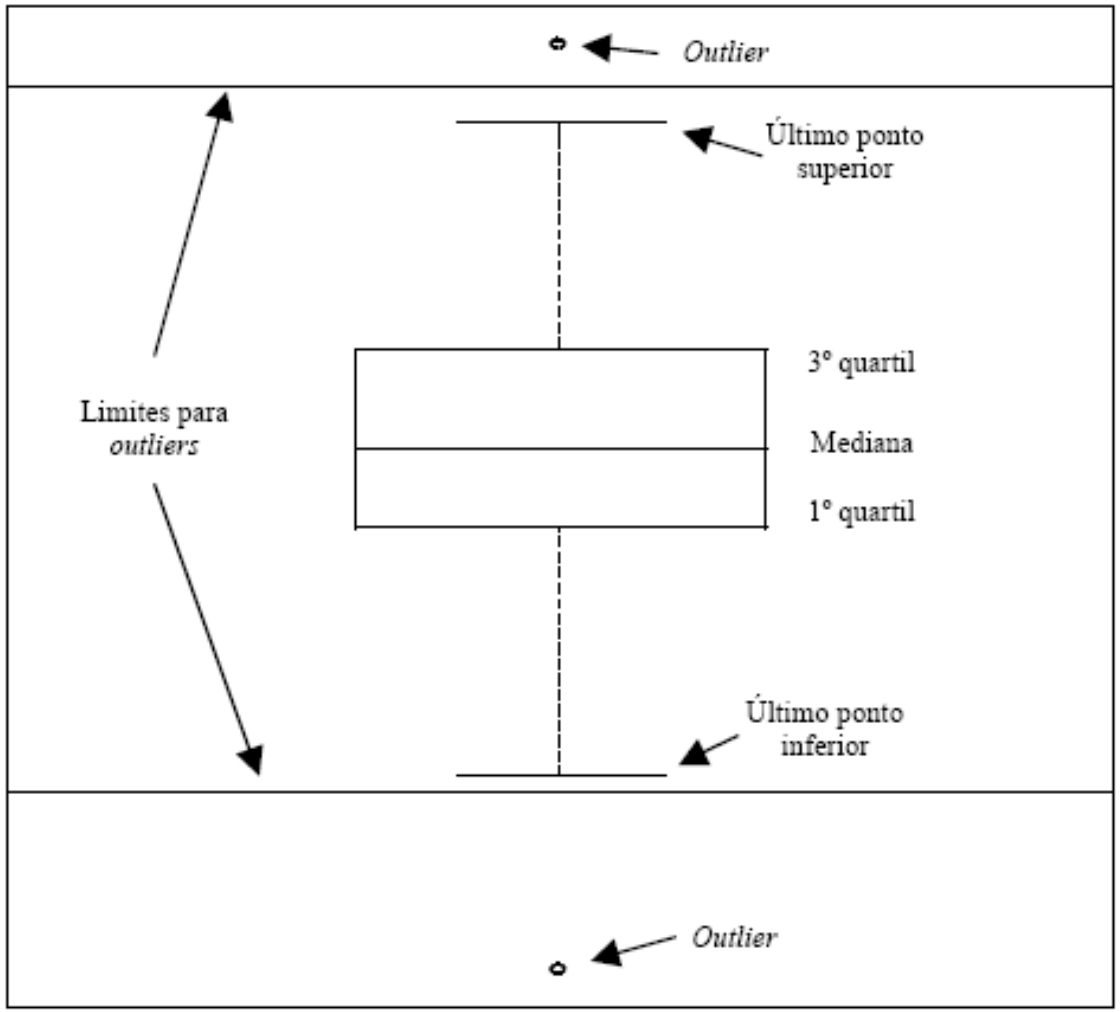
```
boxplot(iris$Sepal.Length, ylab = "Sepal Length",
main = "Sepal Length", col = "darkblue")
```

Boxplot


Sepal Length
Sepal Length
4.5
5.5
6.5
7.5


Longitud del petalo
1
2
3
4
5
6
7
setosa
versicolor
virginica
Especie

Boxplot




```
boxplot(variable de análisis ~ variable de referencia,    main="Título")
```

Dataset Iris

Conjunto de Datos IRIS

- Histórico
  - Datos recolectados en la Península de Gaspe (Quebec) por Edgar Anderson em 1935
  - Datos utilizados por R. A. Fisher em 1936
- El dataset se compone de 150 observaciones de flores de la planta iris.
- Existen tres tipos de clases de flores iris: virginica, setosa y versicolor.

Conjunto de Datos lRIS

- Hay 50 observaciones de cada una.
- Las variables o atributos que se miden de cada flor son:
- 1 El tipo de flor como variable categorica.
- 2 El largo y el ancho del pétalo en cm como variables numéricas.
- 3 El largo y el ancho del sépalo en cm como variables numéricas.

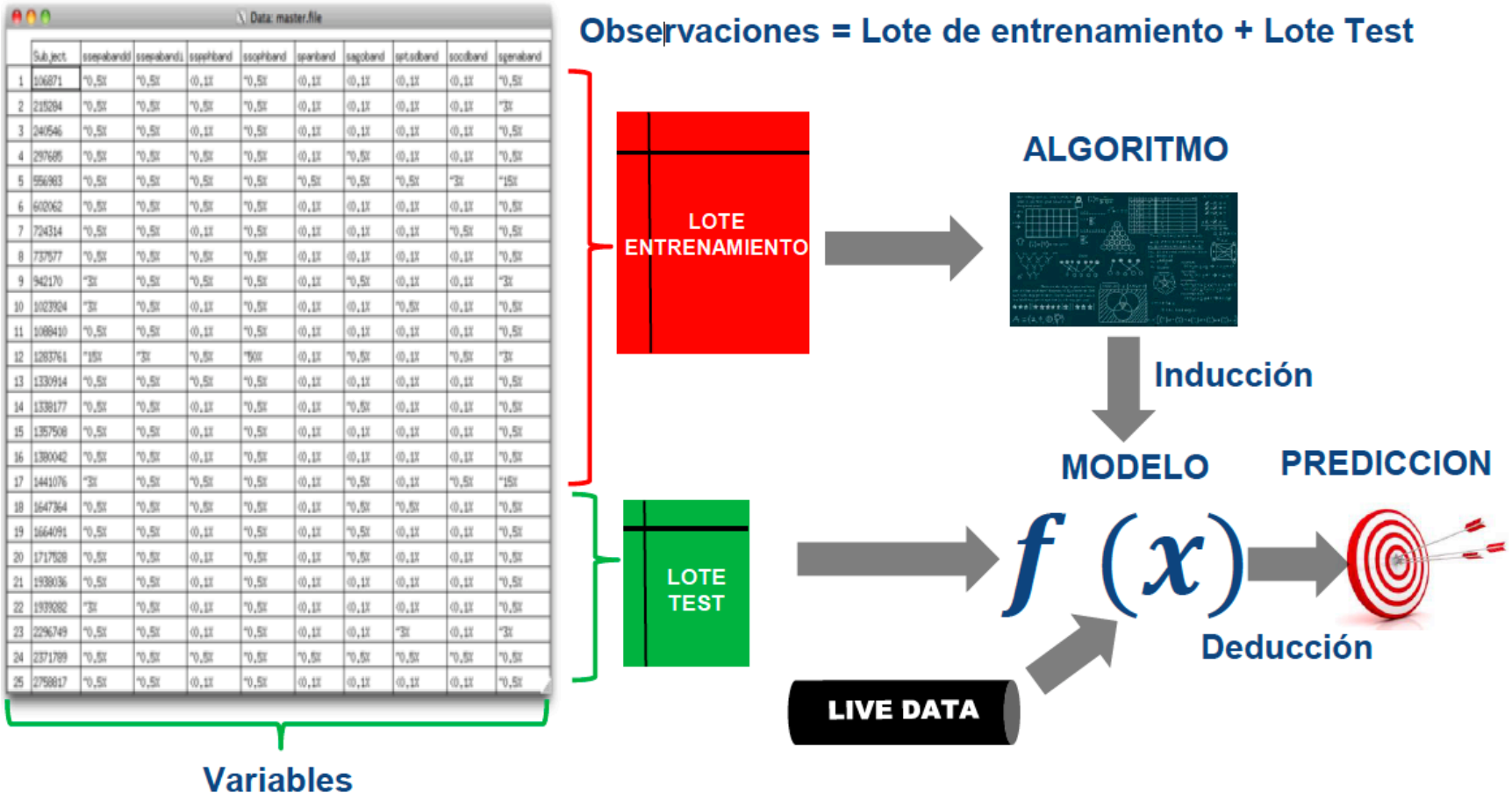
Las flores

Tabla del Conjunto IRIS

	Comp. Sépala	Larg. Sépala	Comp. Pétala	Larg. Pétala	Espécies
1	5.1	3.5	1.4	0.2	Setosa
...	...	...	...	...	...
50	5.0	3.3	1.4	0.2	Setosa
51	7.0	3.2	4.7	1.4	Versicolor
...	...	...	...	...	...
100	5.7	2.8	4.1	1.3	Versicolor
101	6.3	3.3	6.0	2.5	Virginica
...	...	...	...	...	...
150	5.9	3.0	5.1	1.8	Virginica

Práctica de Entrenamiento de modelo predictivo en R

¿Como vamos a Trabajar en la Práctica?



¿Como vamos a Trabajar en la Práctica?

Entrenamiento

```
nombre_modelo<-modelo(fórmula, dataset_entrenamiento, parámetros)
```

Fórmula con variables predictoras seleccionadas

```
variable_dependiente ~ Var1+Var3
```

Fórmula con todas variables predictoras

```
variable_dependiente ~ .
```

Test

```
nombre_predicción <- predict(nombre_contenedor, dataset_test, parámetros)
```

El dataset de test es igual al de entrenamiento pero sin la variable dependiente