

TENEMOS  
MUCHO  
QUE HACER  
JUNTOS

---

Data A

# Machine Learning & Scikit-Learn

Izaskun Mendiola

# Objetivos de la sesión

## Parte 1

- Pequeña introducción a la ciencia de datos.
- Ejemplo clasificador: TITANIC

## Parte 2

1. Ejemplo Clasificador (iris).
  2. Ejemplo Regresor (advertising).
- Tratamiento del dato (limpieza, binarización, estandarización/normalización)
  - Entrenar modelos
  - Sobre entrenamiento
  - Visualización
  - Evaluación de métricas
  - Selección Características
  - Selección Hiper parámetros
  - Evaluación del modelo (matrices de confusión)

# Big Data.



Machine Learning NO es Big Data



Fuente: <http://astrofactoria.webcindario.com/Articulo5.htm>



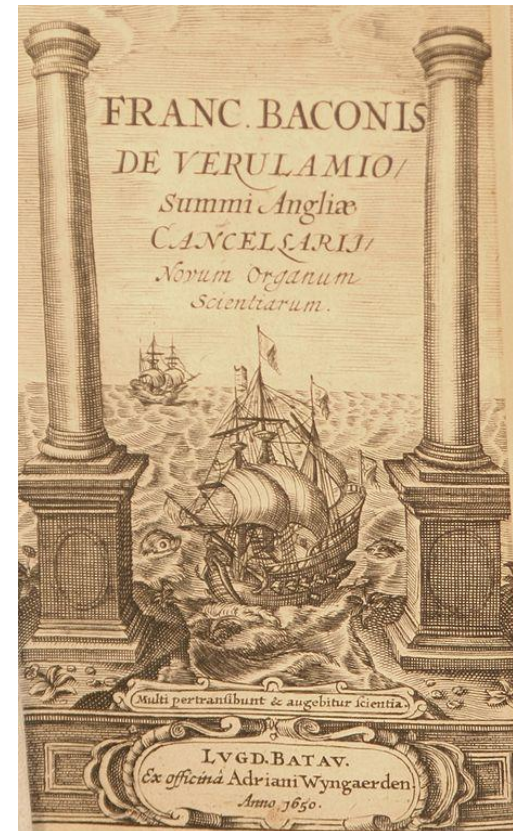
# Empecemos con una Historia de datos....

## Método científico

*“There are and can only be two ways of investigating and discovering truth. The one rushes up from the sense and particulars to axioms of the highest generality and, from these principles and their indubitable truth, goes on to infer and discover middle axioms; and this is the way in current use. The other way draws axioms from the sense and particulars by climbing steadily and by degrees so that it reaches the ones of highest generality last of all; and this is the true but still untrodden way.”*

## **Francis Beacon (1620)**

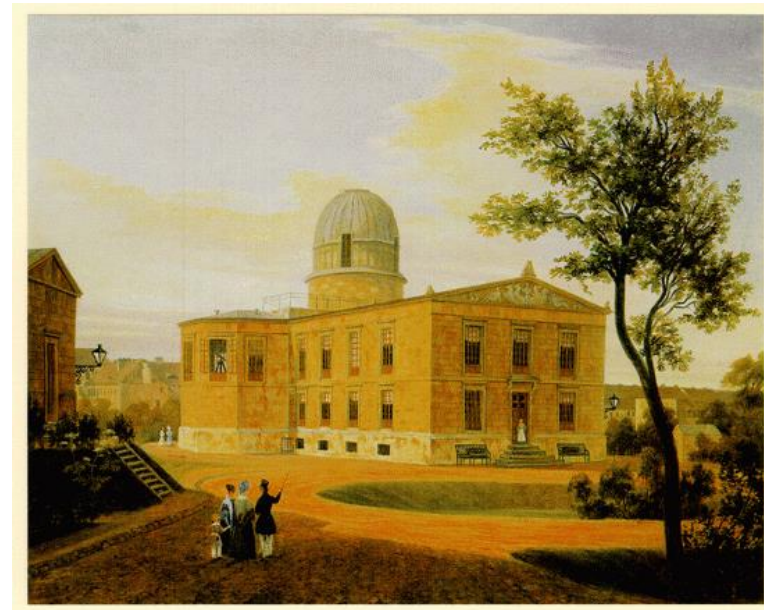
### ***“Su Novum Organum”***



## Determinismo científico

*“We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.”*

— *Pierre Simon Laplace, A Philosophical Essay on Probabilities*



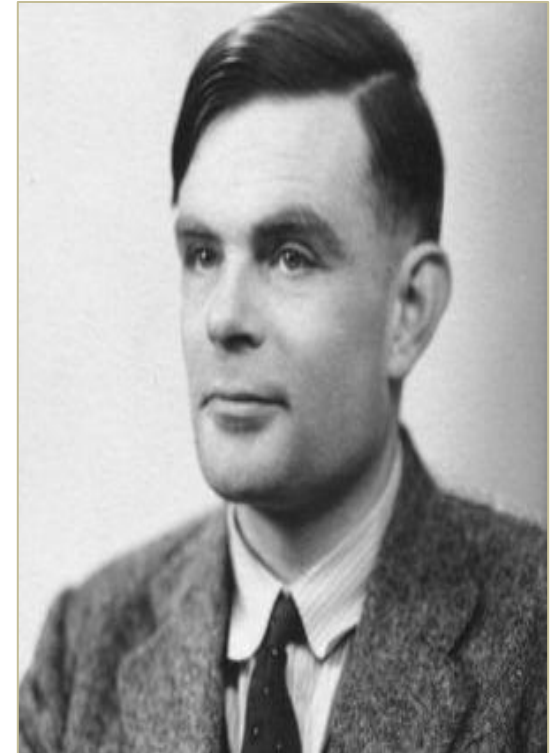
(Sept'1846) Observatorio de Berlín, desde donde Neptuno fue descubierto mediante observaciones.



Karl Pearson (1857 - 1936)



Ronald Fisher (1890 -  
1962)



Alan Turing (1912 - 1954)

# Explosión de los datos

Aunque hace décadas que existen los analistas de datos, también hace décadas que se almacenan datos que no han podido ser procesados hasta hace pocos años:



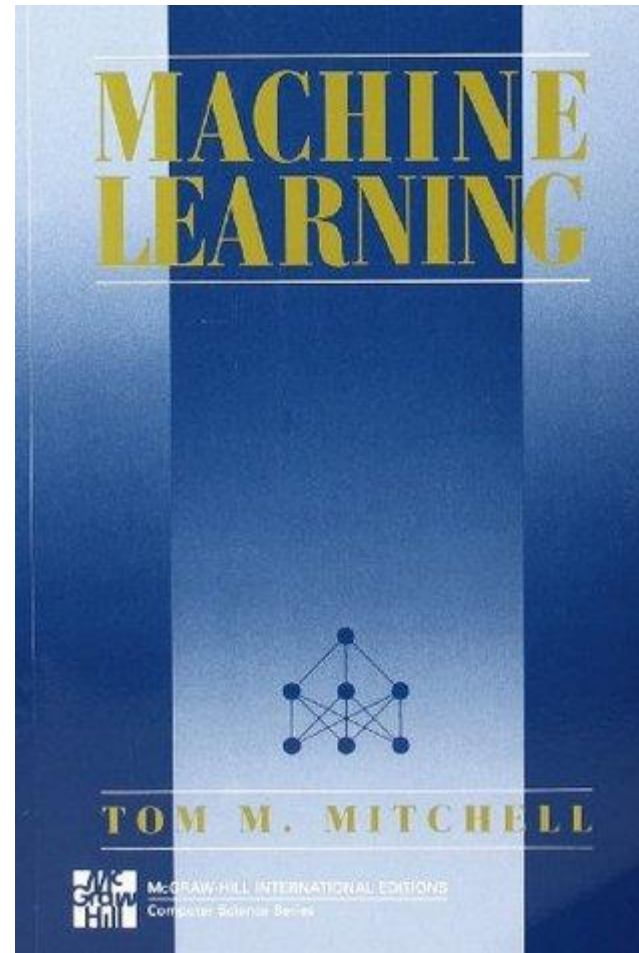
Fuente: Fuente Big Bang Data

- Tecnologías de bases de datos
- Coste del hardware de almacenamiento
- Aumento del ancho de banda
- Aumento de capacidad de procesamiento
- Software científico



# Machine Learning

*“Un programa se dice que aprende de la experiencia  $E$  respecto a alguna tarea  $T$  y alguna medida de eficacia  $P$  si su eficacia en  $T$ , medida por  $P$ , mejora con la Experiencia  $E$ .”*



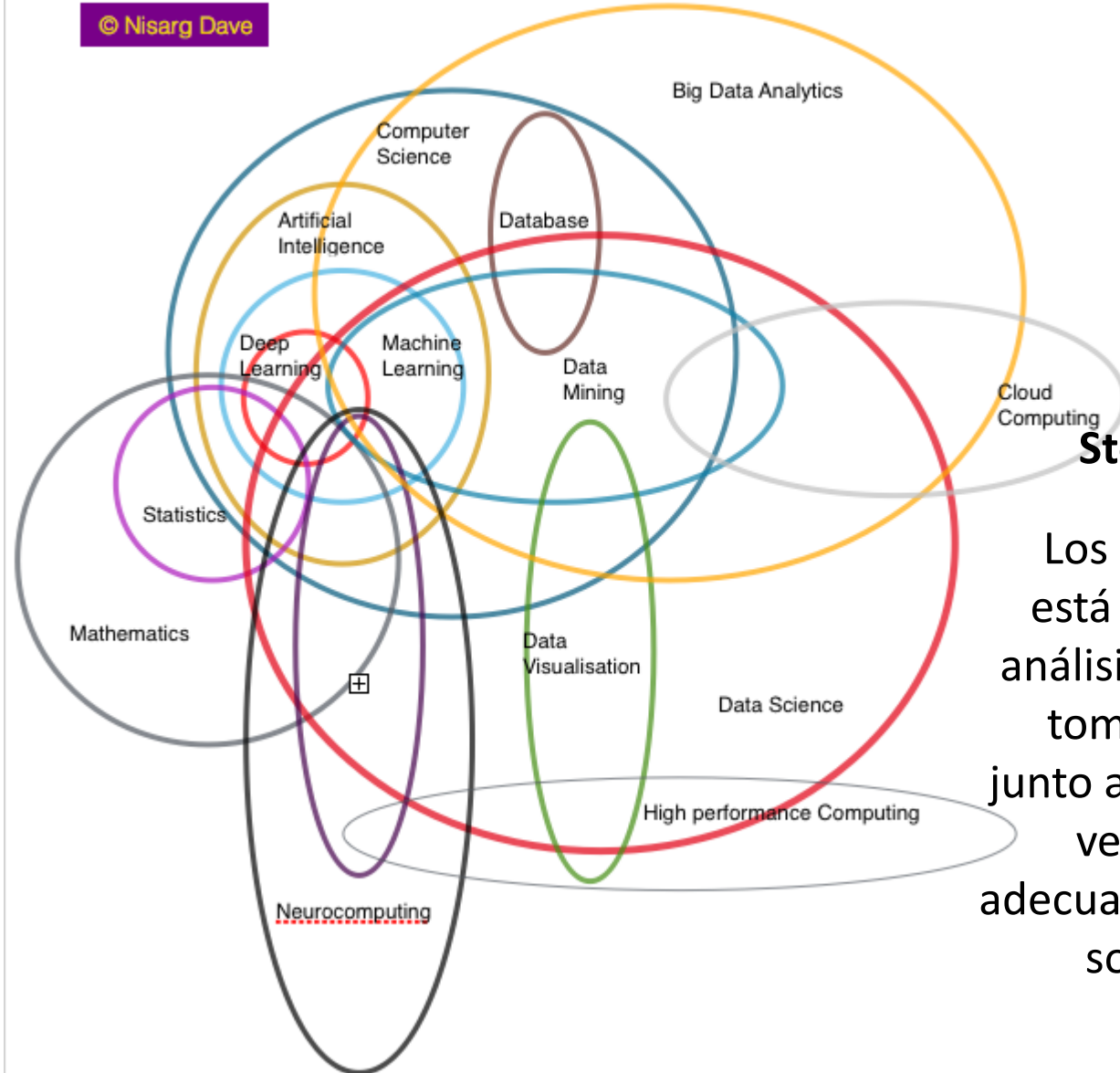
Todo esto nos capacita para pasar  
de la **información** al **conocimiento**

## Ejemplos de éxito

- Traductor de Google
- Reconocimiento de caras de Facebook
- Recomendador compras de Amazon

**This is How I define Data Science &  
Role of Data Scientist !**

© Nisarg Dave



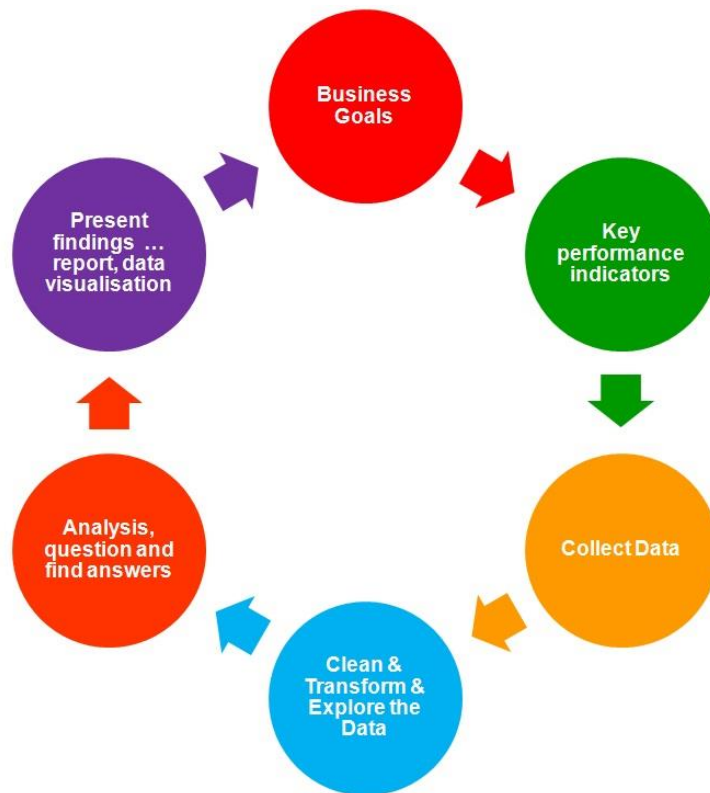
## Stop a la obsesión por los datos

Los datos no lo son todo. También está el conocimiento cualitativo. El análisis de datos puede ayudar en la toma de decisiones, pero siempre junto al conocimiento de expertos. A veces un conocimiento será más adecuado que otros, a veces la mejor solución será la combinación de conocimiento cualitativos y cuantitativos.

Fuente: <http://lod-cloud.net/>



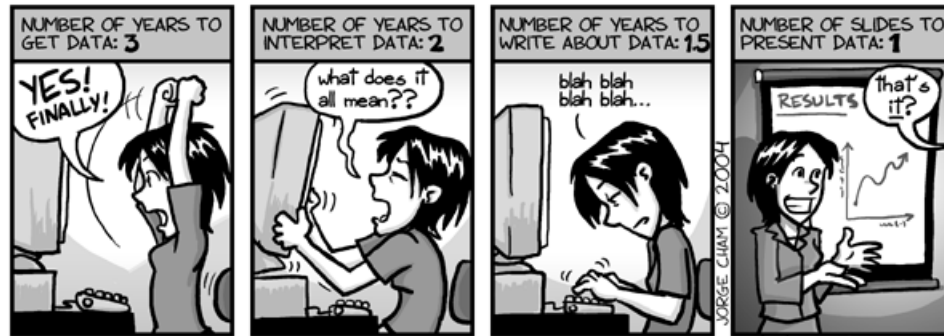
# Etapas en el proceso



- **Integración y recopilación.** Comprensión del dominio de aplicación del problema, identificación de conocimiento a priori.
- **Pre procesamiento.** Selección de datos, limpieza, reducción y transformación.
- **Selección de la técnica** de ML y aplicación de algoritmo concretos de ML
- **Evaluación, interpretación y presentación** de los resultados obtenidos
- **Difusión** y utilización del **nuevo conocimiento**.

# ¿Qué etapa lleva más esfuerzo?

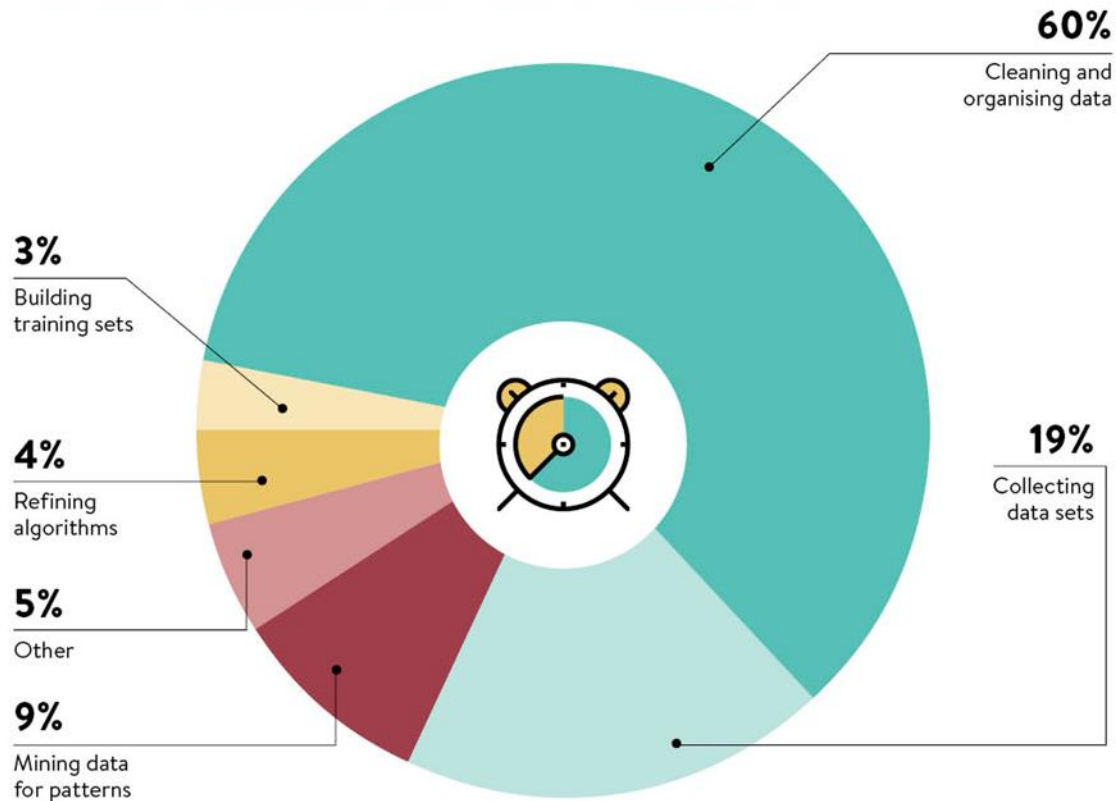
## DATA: BY THE NUMBERS



[www.phdcomics.com](http://www.phdcomics.com)

# ¿Qué etapa lleva más esfuerzo?

## WHAT DATA SCIENTISTS SPEND THE MOST TIME DOING



Source: CrowdFlower 2016

## Análisis de los datos

- Explicar el pasado a través de:
  - Variables numéricas.
  - Variables categóricas.
- Predecir el futuro, mediante modelos:
  - Modelos de clasificación, resultado categoría.
  - Modelos de regresión, resultado numérico.
  - Clustering.
  - Reglas asociativas.
- Prescribir



# Análisis de los datos

## Tipos de aprendizaje

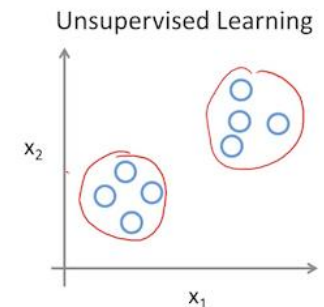
### Aprendizaje supervisado

- Conozco las respuestas.
- Entreno el modelo con las respuestas conocidas.
- Verifico los resultados con las respuestas conocidas.
- Establece una correspondencia entre la entrada y la salida deseada (Clasificación, Regresión).  $[x, f(x)]$



### Aprendizaje no supervisado

- No conozco las respuestas.
- El proceso de modelado se realiza sobre un conjunto ejemplos donde se tiene solo las entradas.  $[x]$



### Aprendizaje por refuerzo

El algoritmo re-aprende constantemente, en función de experiencias, refuerza el aprendizaje si tiene éxito (Feedback)  $[x, f(x)]$

# Librería scikit-learn

## ¿Qué es scikit-learn, ...

- Biblioteca de análisis de datos y Machine Learning.
- Para Python 2 y 3.
- Buen rendimiento.
- Posee implementaciones de algoritmos y modelos de ML.

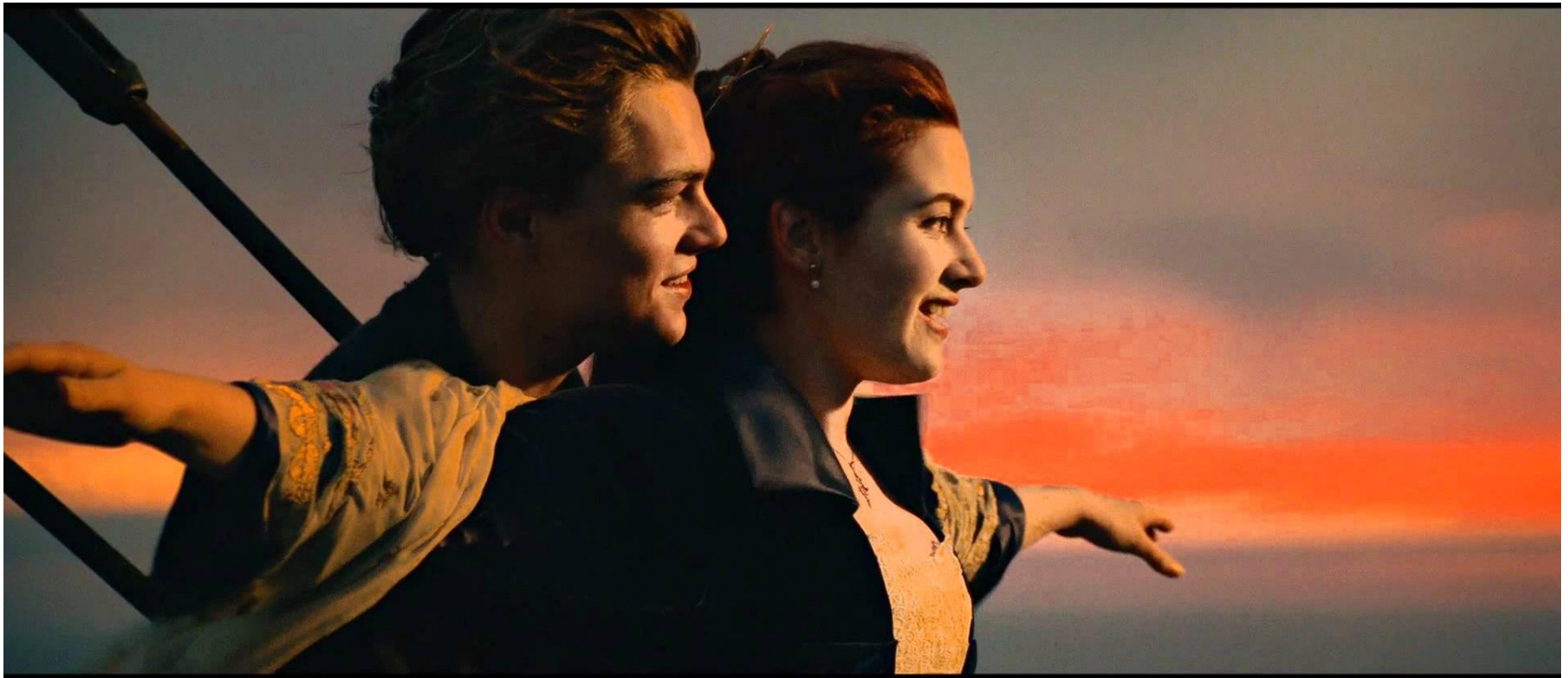
## ... e integración en Python?

Sklearn está perfectamente integrada en Python:

- Numpy, manipulación eficiente de arrays y algebra lineal.
- SciPy: Extensa biblioteca para matemáticas, ciencias e ingeniería.
- Matplotlib: biblioteca de gráficos.
- iPython/Jupyter: Shell de Python (y otros lenguajes) sencillo y versátil.
- Pandas: biblioteca para manipular tablas de forma sencilla

# Ejemplo práctico TITANIC

## ¿Se hubiesen salvado?



Tarea de Clasificación Supervisada

# TITANIC - datos

```

"row.names","pclass","survived","name","age","embarked","home.dest","room","ticket","boat","sex"
"1","1st",1,"Allen, Miss Elisabeth Walton",29.0000,"Southampton","St Louis, MO","B-5","24160 L221","2","female"
"2","1st",0,"Allison, Miss Helen Loraine",2.0000,"Southampton","Montreal, PQ / Chesterville, ON","C26","","","female"
"3","1st",0,"Allison, Mr Hudson Joshua Creighton",30.0000,"Southampton","Montreal, PQ / Chesterville, ON","C26","","(135)","male"
"4","1st",0,"Allison, Mrs Hudson J.C. (Bessie Waldo Daniels)",25.0000,"Southampton","Montreal, PQ / Chesterville, ON","C26","","","female"
"5","1st",1,"Allison, Master Hudson Trevor",0.9167,"Southampton","Montreal, PQ / Chesterville, ON","C22","","11","male"
"6","1st",1,"Anderson, Mr Harry",47.0000,"Southampton","New York, NY","E-12","","3","male"
"7","1st",1,"Andrews, Miss Kornelia Theodosia",63.0000,"Southampton","Hudson, NY","D-7","13502 L77","10","female"
"8","1st",0,"Andrews, Mr Thomas, jr",39.0000,"Southampton","Belfast, NI","A-36","","","male"
"9","1st",1,"Appleton, Mrs Edward Dale (Charlotte Lamson)",58.0000,"Southampton","Bayside, Queens, NY","C-101","","2","female"
"10","1st",0,"Artagaveytia, Mr Ramon",71.0000,"Cherbourg","Montevideo, Uruguay","","","(22)","male"
"11","1st",0,"Astor, Colonel John Jacob",47.0000,"Cherbourg","New York, NY","","17754 L224 10s 6d","(124)","male"
"12","1st",1,"Astor, Mrs John Jacob (Madeleine Talmadge Force)",19.0000,"Cherbourg","New York, NY","","17754 L224 10s 6d","4","female"
"13","1st",1,"Aubert, Mrs Leontine Pauline",NA,"Cherbourg","Paris, France","B-35","17477 L69 6s","9","female"
"14","1st",1,"Barkworth, Mr Algernon H.",NA,"Southampton","Hessle, Yorks","A-23","","B","male"
"15","1st",0,"Baumann, Mr John D.",NA,"Southampton","New York, NY","","","male"
"16","1st",1,"Baxter, Mrs James (Helene DeLaunier Chaput)",50.0000,"Cherbourg","Montreal, PQ","B-58/60","","6","female"
"17","1st",0,"Baxter, Mr Quigg Edmond",24.0000,"Cherbourg","Montreal, PQ","B-58/60","","","male"
"18","1st",0,"Beattie, Mr Thomson",36.0000,"Cherbourg","Winnipeg, MN","C-6","","","male"
"19","1st",1,"Beckwith, Mr Richard Leonard",37.0000,"Southampton","New York, NY","D-35","","5","male"
"20","1st",1,"Beckwith, Mrs Richard Leonard (Sallie Monypeny)",47.0000,"Southampton","New York, NY","D-35","","5","female"
[....]
  
```

1313 pasajero



# Python: Scikit-learn

```
In [1]: %matplotlib inline
import IPython
import sklearn as sk
import numpy as np
import seaborn as sns
import pandas as pd

print 'IPython version:', IPython.__version__
print 'numpy version:', np.__version__
print 'scikit-learn version:', sk.__version__
print 'seaborn version:', sns.__version__
print 'pandas version:', pd.__version__

IPython version: 5.1.0
numpy version: 1.11.1
scikit-learn version: 0.17.1
seaborn version: 0.7.0
pandas version: 0.19.0
```

¡Atención!

El código está en <https://github.com/izmendi/CursoML>

Y para ejecutarlo, copiar URL desde: <https://nbviewer.jupyter.org/>

# 1. Pre procesamiento

```
In [3]: titanic_dataset = pd.read_csv('./data/titanic.txt', sep=',')

# La primera fila tiene los nombres de los atributos
print titanic_dataset.columns.values

['row.names' 'pclass' 'survived' 'name' 'age' 'embarked' 'home.dest' 'room'
 'ticket' 'boat' 'sex']
```

Veamos cómo quedan las tuplas...

```
In [4]: titanic_dataset.iloc[0]
```

```
Out[4]: row.names      1
pclass      1st
survived     1
name      Allen, Miss Elisabeth Walton
age      29
embarked      Southampton
home.dest      St Louis, MO
room      B-5
ticket      24160 L221
boat      2
sex      female
Name: 0, dtype: object
```

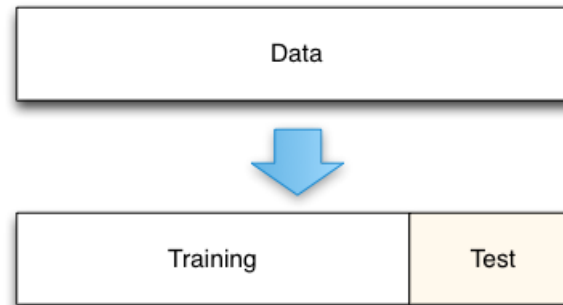
## 2. Ingeniería de atributos

- *¿Qué atributos utilizamos para aprender?*
  - *¿Qué hacemos cuando no hay valores?*
  - *¿Cómo adaptamos nuestros atributos al método que utilizamos para aprender?*
- 
- Estudio de relevancia de variables: age, sex, p\_class
  - Reemplazar valores faltantes:
    - Valores cuantitativos (numéricos) -> media aritmética
    - Valores categóricos -> moda

```
In [14]: print titanic_X.head()
```

	age	sex	primera_clase	segunda_clase	tercera_clase
0	22.0	1	0.0	0.0	1.0
1	38.0	0	1.0	0.0	0.0
2	26.0	0	0.0	0.0	1.0
3	35.0	0	1.0	0.0	0.0
4	35.0	1	0.0	0.0	1.0

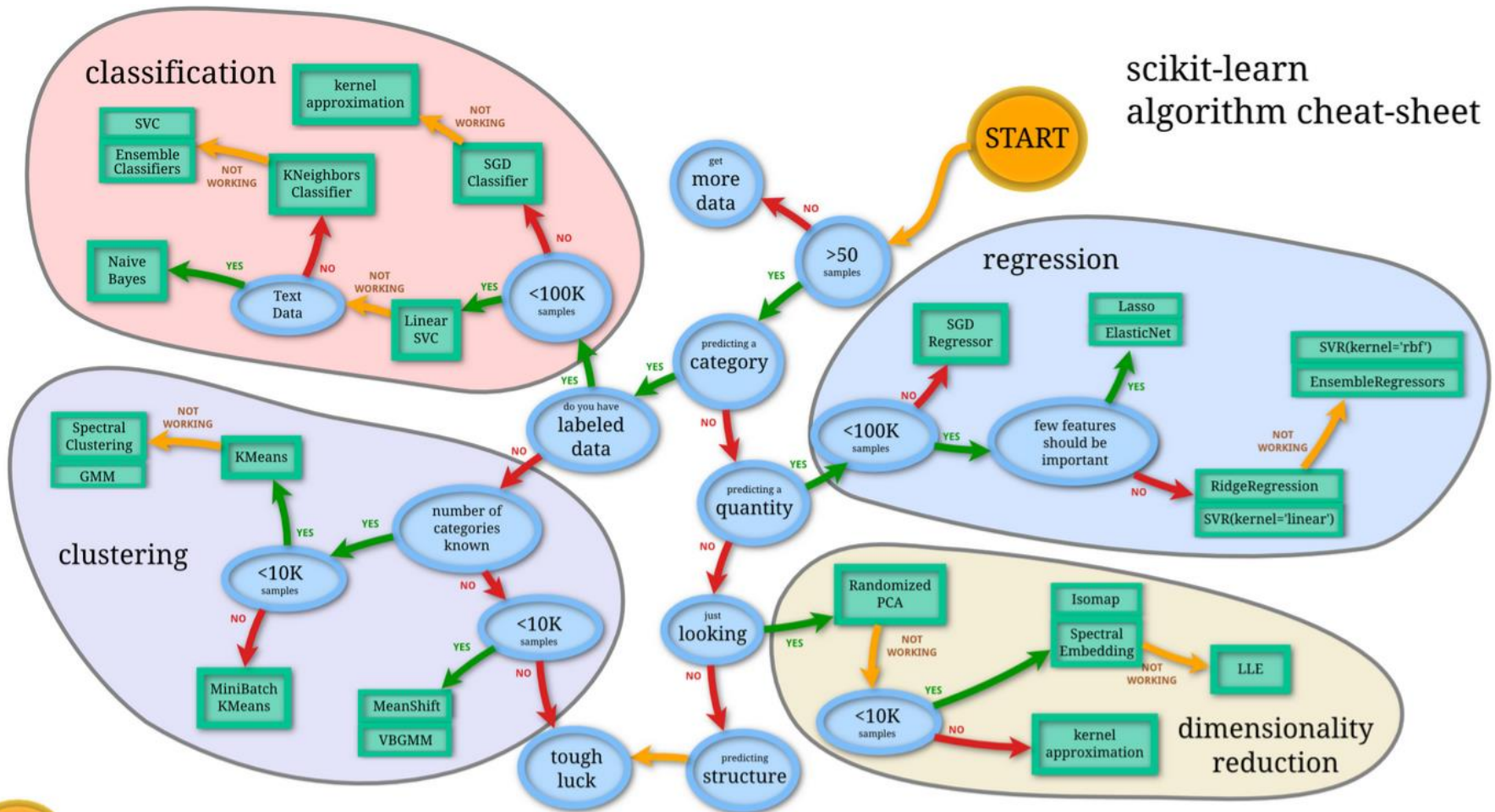
# 3. Separar entrenamiento / testeo (MUY importante)



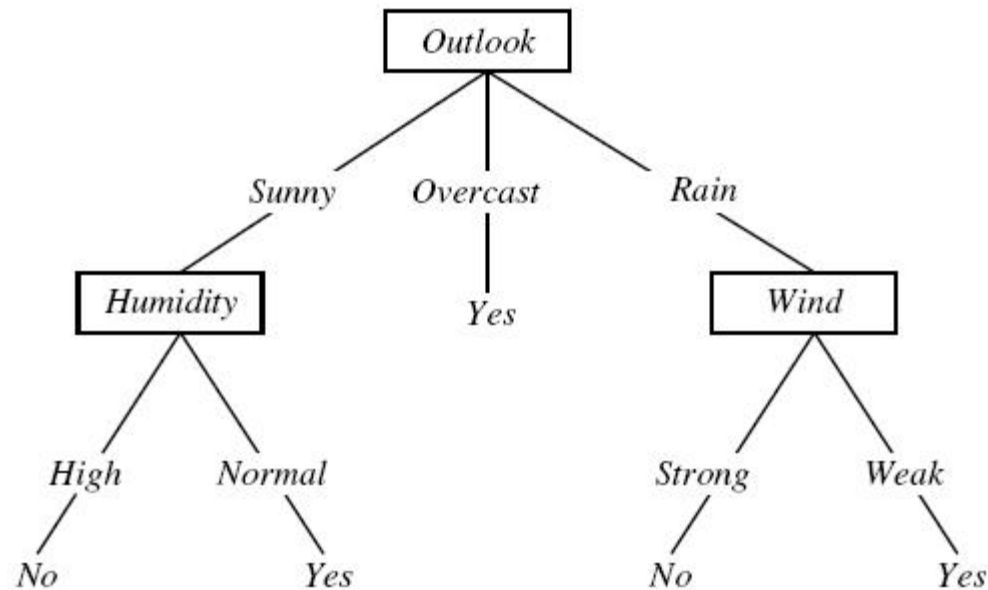
```
In [17]: from sklearn.cross_validation import train_test_split
X_train, X_test, y_train, y_test = train_test_split(titanic_X, titanic_y, test_size=0.25, random_state=33)
```



## 4. Entrenar un modelo



# Árbol de decisión



# ¿Cómo generamos el modelo?





Samuel



Pepe



Pablo



Jorge



Felipe



Clara



Anita



Alfredo



Susana



Ricardo



Paco



Manuel



Germán



David



Bernardo



Alejandro



Tomás



Roberto



Pedro



María



Guillermo



Ernesto



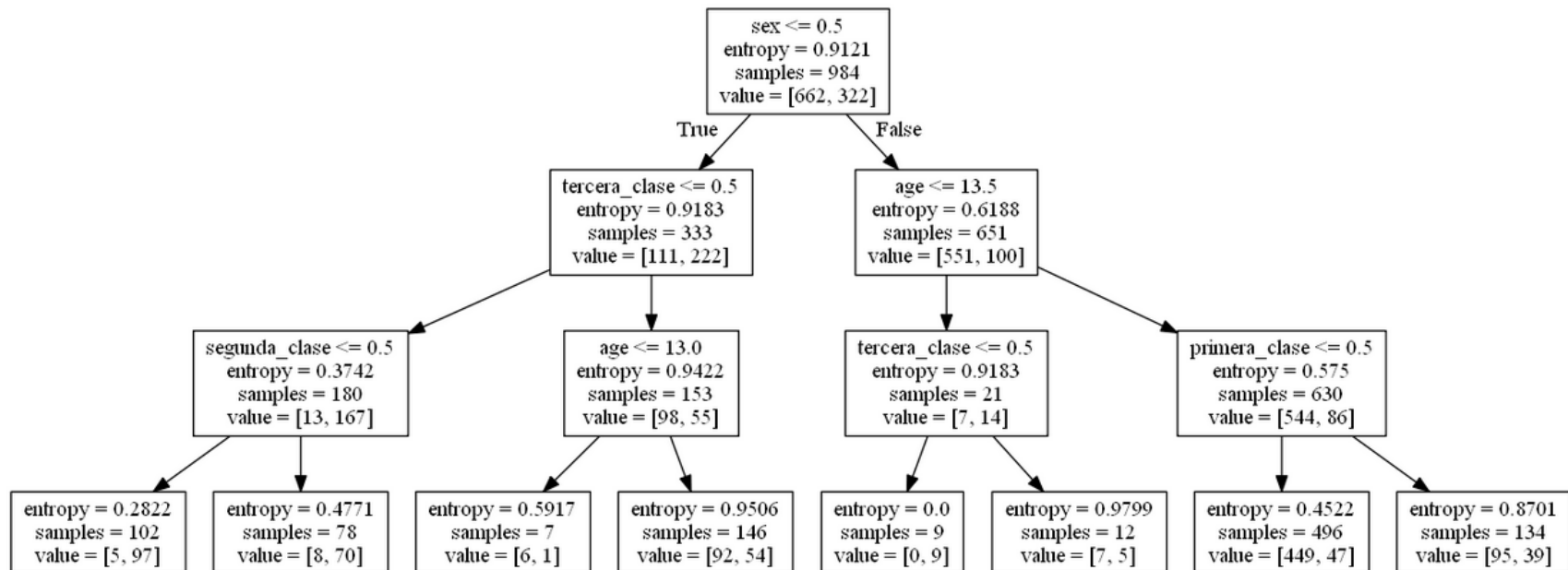
Carlos



Ana



```
In [18]: from sklearn import tree
clf = tree.DecisionTreeClassifier(criterion='entropy', max_depth=3,min_samples_leaf=5)
clf = clf.fit(X_train,y_train)
```



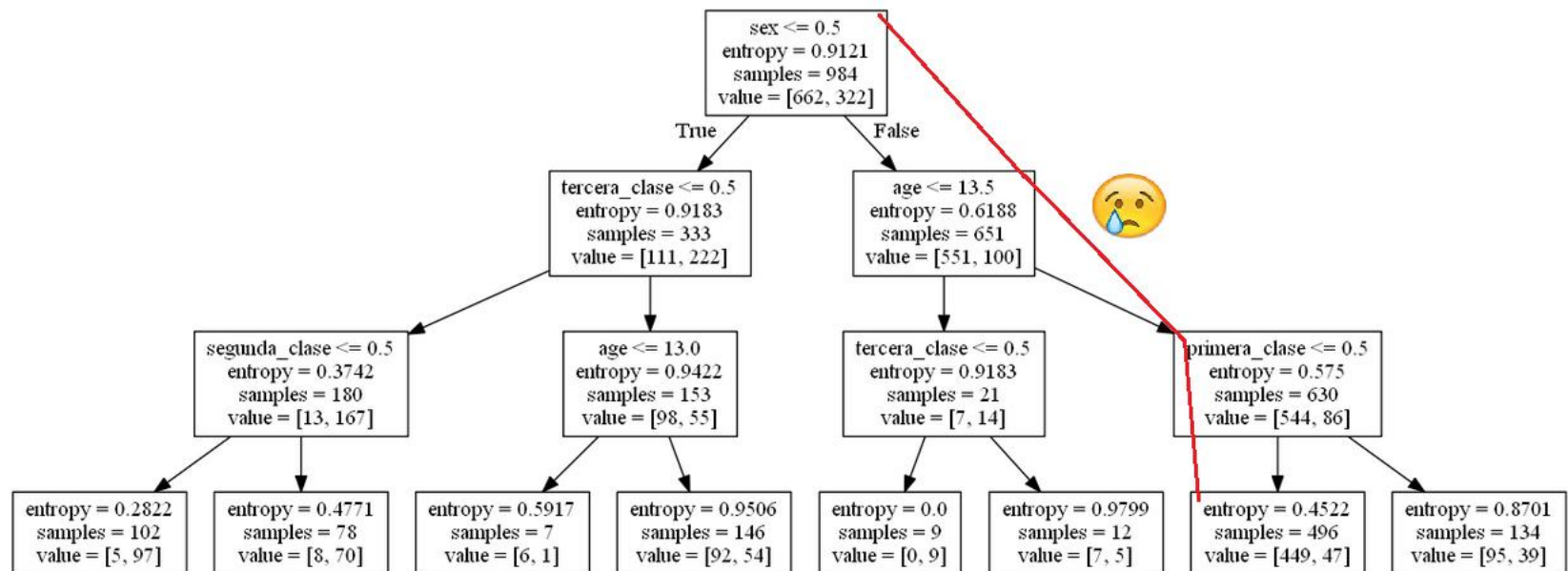


(edad,20), (sexo, 1.0), (primera\_clase, 0.0),  
(segunda\_clase, 0.0), (tercera\_clase, 1.0)

```
print clf.predict([[20.0,1.0,0.0,0.0,1.0]])
```

[0]

Out[42]:





# 5. Evaluar el modelo

## Matriz de confusión en corpus de **entrenamiento**

Valor real \ Valor predicho	No sobrevive (0)	Sobrevive (1)
No sobrevivió (0)	649	13
Sobrevivió (1)	146	176

$$\text{Accuracy} = (TP + TN) / P + N$$

$$\text{Accuracy} = (649 + 176) / 984 = 0,838$$

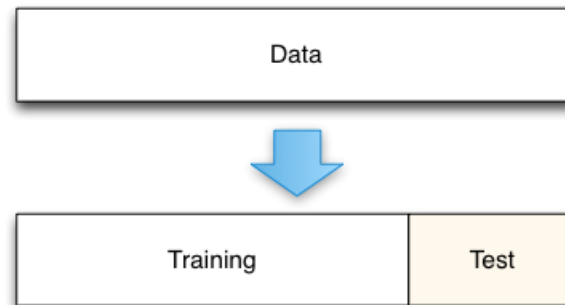
**ESTO NO ES UNA BUENA IDEA!!**

## Matriz de confusión en corpus de **evaluación**

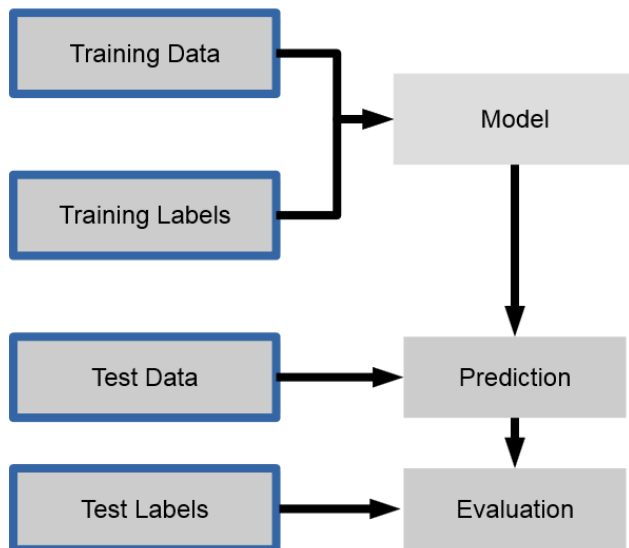
Valor real \ Valor predicho	No sobrevive (0)	Sobrevive (1)
No sobrevivió (0)	193	9
Sobrevivió (1)	59	68

$$\text{Accuracy} = (193 + 68) / 329 = 0,793$$

**ESTA SÍ ES UNA BUENA IDEA!!**



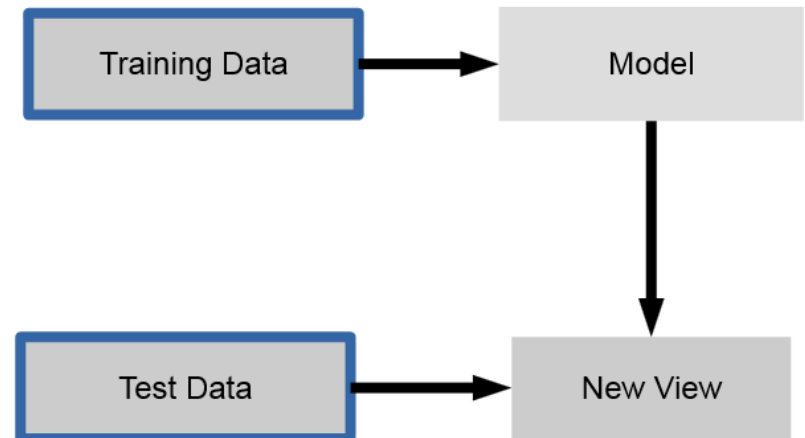
## Aprendizaje Supervisado



Training

Generalization

## Aprendizaje No Supervisado



# Common mistakes

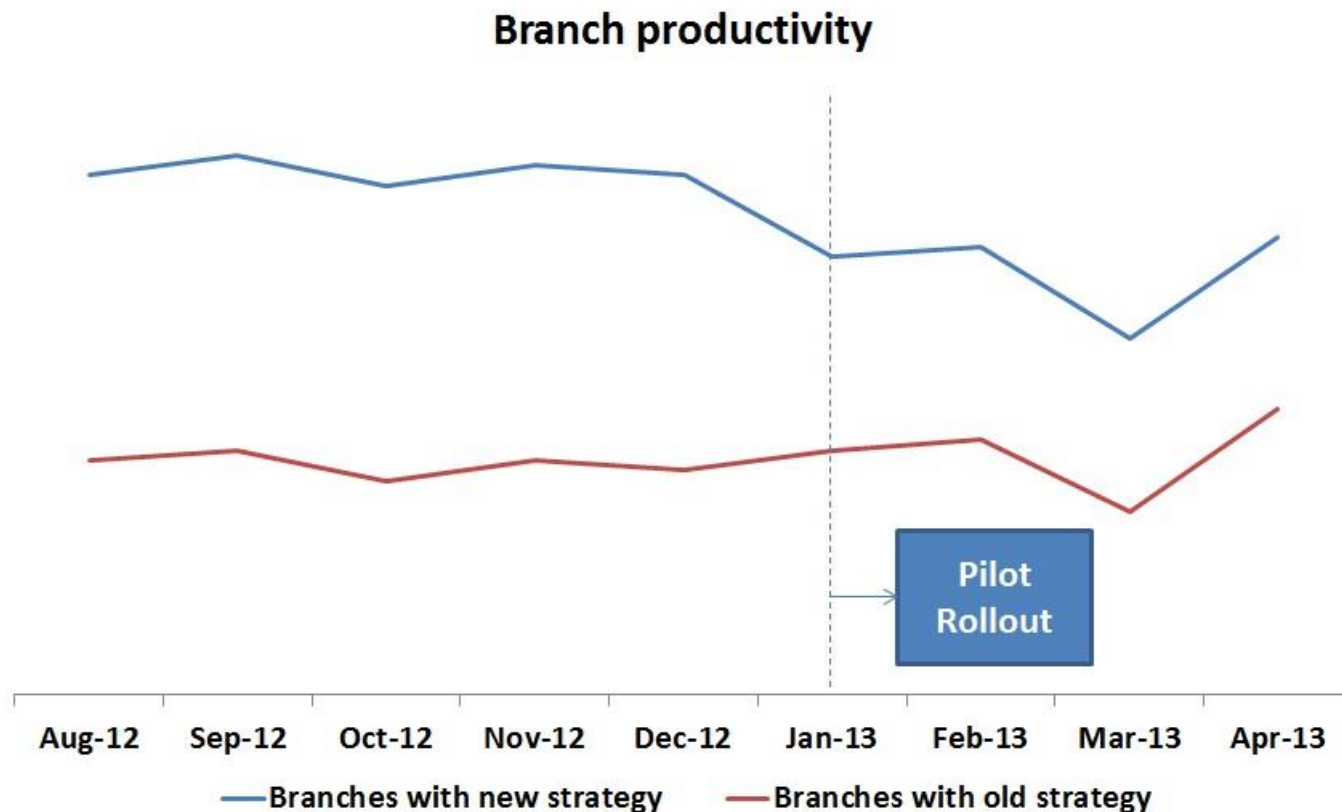
---

---

***“If you torture the data long enough, it will confess.”*** *Ronald Coase, Economist*

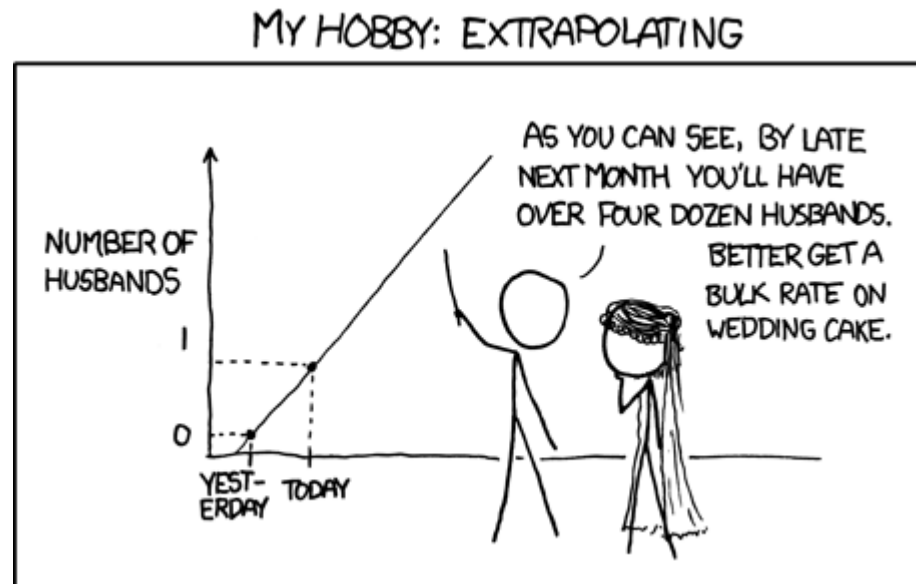
---

---



# Common mistakes

- Drawing inferences on thin data (and extrapolating it)



# Common mistakes

- **Correlation does not mean Causation**

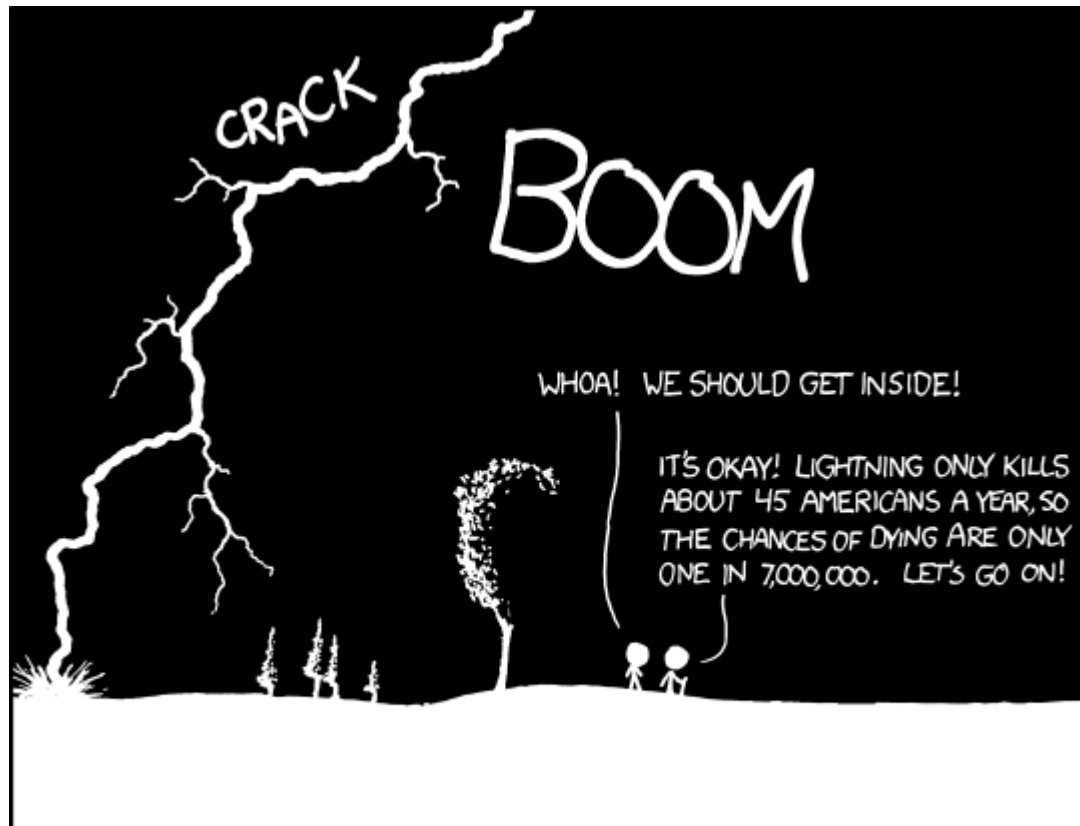


¡Para frenar el calentamiento global: hagámonos piratas!



# Common mistakes

- **Wrong applications of the inferences**



THE ANNUAL DEATH RATE AMONG PEOPLE  
WHO KNOW THAT STATISTIC IS ONE IN SIX.



# Machine Learning es más ....

## COURSERA

- “Introduction to Data Science in Python” by University of Michigan
- 5 course specialization “Applied Data Science with Python” by University of Michigan
- Machine Learning by Stanford University- **Andrew Ng**  
<https://www.coursera.org/learn/machine-learning>
- Machine Learning: Clustering & Retrieval



Introduction  
What is machine  
learning

STANDFORD	<a href="http://online.stanford.edu/courses">http://online.stanford.edu/courses</a>
EdX	<a href="https://www.edx.org/course">https://www.edx.org/course</a>
Udacity	<a href="https://www.udacity.com/courses/all">https://www.udacity.com/courses/all</a>
CLOUDERA	<a href="http://www.cloudera.com">http://www.cloudera.com</a>

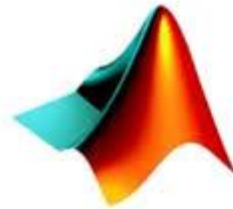
## Vídeos en Youtube

## DATASCHOOL

<https://www.youtube.com/user/dataschool>

Y más ...

# Machine Learning es más ....



MATLAB®



# Python code vs R code



## Python Code



## R Code

### Linear Regression

```
#Import Library
#Import other necessary libraries like pandas,
#numpy...
from sklearn import linear_model
#Load Train and Test datasets
#Identify feature and response variable(s) and
#values must be numeric and numpy arrays
x_train=input_variables_values_training_datasets
y_train=target_variables_values_training_datasets
x_test=input_variables_values_test_datasets
#Create linear regression object
linear = linear_model.LinearRegression()
#Train the model using the training sets and
#check score
linear.fit(x_train, y_train)
linear.score(x_train, y_train)
#Equation coefficient and Intercept
print('Coefficient: \n', linear.coef_)
print('Intercept: \n', linear.intercept_)
#Predict Output
predicted= linear.predict(x_test)
```

```
#Load Train and Test datasets
#Identify feature and response variable(s) and
#values must be numeric and numpy arrays
x_train <- input_variables_values_training_datasets
y_train <- target_variables_values_training_datasets
x_test <- input_variables_values_test_datasets
x <- cbind(x_train,y_train)
#Train the model using the training sets and
#check score
linear <- lm(y_train ~ ., data = x)
summary(linear)
#Predict Output
predicted= predict(linear,x_test)
```

# Pelis ....

- [Moneyball \(Rompiendo las reglas\)](#): Los datos y el beisbol, unidos por la estadística y llevados al triunfo
- [The imitation game \(Descifrando Enigma\)](#): El análisis de cómo se logró destruir al ejército alemán descifrando datos. Realmente la película se centra en la vida del matemático Turing.
- [The Big short \(La gran apuesta\)](#): Paso a paso de la crisis en USA por las hipotecas, precedida por un analista de datos. Es interesante ver el proceso desde que se intuye el fiasco hasta que va ocurriendo.
- [Zero Dark Thirty \(La noche más oscura\)](#): El momento de la operación militar para capturar a Osama Bin Laden, fue clave toda la documentación y datos aportados por una agente de la CIA.

Y más ...



**TO BE  
CONTINUED...**

# Objetivos de la sesión

## Parte 1

- Pequeña introducción a la ciencia de datos.
- Ejemplo clasificador: TITANIC

## Parte 2

1. Ejemplo Clasificador (iris).
  2. Ejemplo Regresor (advertising).
- Tratamiento del dato (limpieza, binarización, estandarización/normalización)
  - Entrenar modelos
  - Sobre entrenamiento
  - Visualización
  - Evaluación de métricas
  - Selección Características
  - Selección Hiper parámetros
  - Evaluación del modelo (matrices de confusión)



Antes de finalizar, un favor:

- Rellena con tu nombre y NIF, y firma la Hoja de Control de Asistencia
  - Rellena el cuestionario de evaluación

Nos sirve para tramitar ayudas a la Fundación Tripartita por formación!

**MUCHAS GRACIAS POR VUESTRA  
PARTICIPACIÓN  
EN ESTA SESIÓN!**

**058585:20200**

Clúster Conocimiento TIC – Equipo Desarrollo SW:

Angulo Redondo, Iñaki [inaki.angulo@tecnalia.com](mailto:inaki.angulo@tecnalia.com); Benguria Elguezabal, Gorka [Gorka.Benguria@tecnalia.com](mailto:Gorka.Benguria@tecnalia.com); Beristain Aizpuru, Joseba [joseba.beristain@tecnalia.com](mailto:joseba.beristain@tecnalia.com); Del Pozo Rojo, Dionisio [dionisio.delpozo@tecnalia.com](mailto:dionisio.delpozo@tecnalia.com); Gil Aguirrebeitia, Guillermo [guillermo.gil@tecnalia.com](mailto:guillermo.gil@tecnalia.com); Martinez Criado, Iñigo [inigo.martinez@tecnalia.com](mailto:inigo.martinez@tecnalia.com); Ozamiz Oyarzabal, Miguel [miguel.ozamiz@tecnalia.com](mailto:miguel.ozamiz@tecnalia.com); Quintano Fernandez, Nuria [Nuria.Quintano@tecnalia.com](mailto:Nuria.Quintano@tecnalia.com); Remazeilles, Anthony [anthony.remazeilles@tecnalia.com](mailto:anthony.remazeilles@tecnalia.com); Ruiz Ruiz, Ana Belen [AnaBelen.Ruiz@tecnalia.com](mailto:AnaBelen.Ruiz@tecnalia.com); Zubizarreta Pikabea, Aitzol [aitzol.zubizarreta@tecnalia.com](mailto:aitzol.zubizarreta@tecnalia.com); Zaitegi Aresti, Koldo [koldo.zaitegi@tecnalia.com](mailto:koldo.zaitegi@tecnalia.com)



Visita nuestro blog:  
<http://blogs.tecnalia.com/inspiring-blog/>



[www.tecnalia.com](http://www.tecnalia.com)