Analytics

Data Science mediante casos prácticos

FABIO GÓMEZ-ESTERN
UNIVERSIDAD LOYOLA ANDALUCÍA

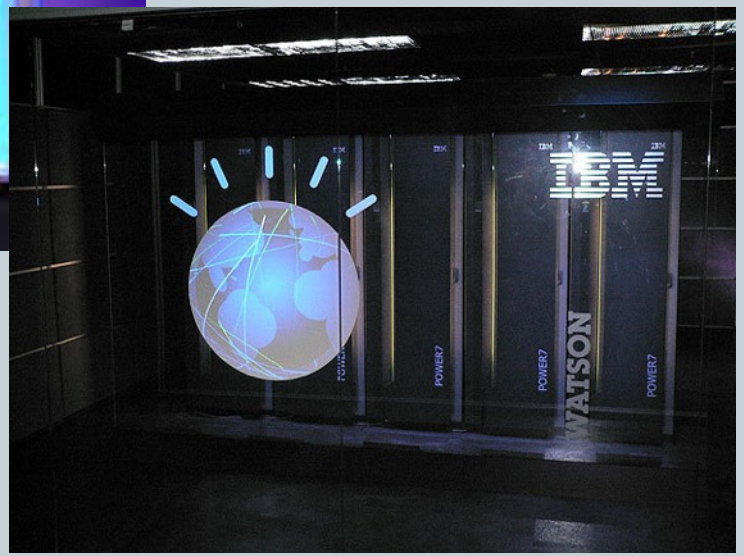
Organización de la presentación

- Orígenes del Big Data
- Conceptos matemáticos
- Big Data en entorno R
- Caso práctico
- Big Data vs. intuición
- Conclusiones

La historia de la informática corporativa

- La carrera de IBM
  - Deep Blue: Ajedrez 1960-1990. Máquina: 1996
  - Blue Gene (2005). 36 Teraflops
  - Watson: Jeopardy! 2006-2011. 11,38 Gigaflops
- Big Data y Business Intelligence
  - Google Flu Trends (2009)
  - MoneyBall (2011, Oackland Athletics 2002)

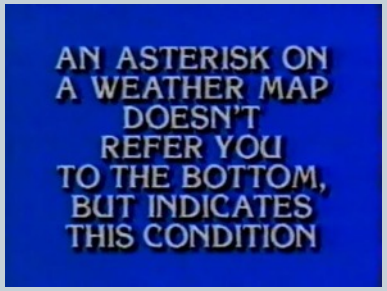
IBM Watson gana Jeopardy!

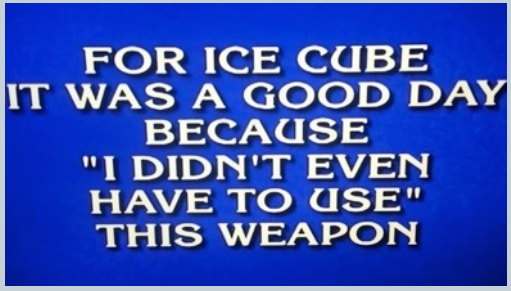


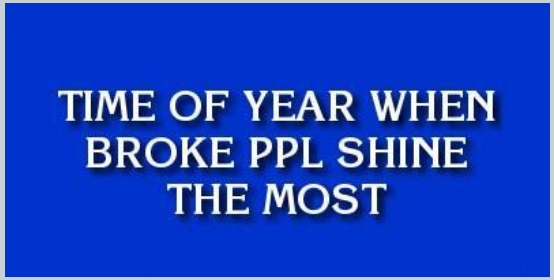
IBM Watson

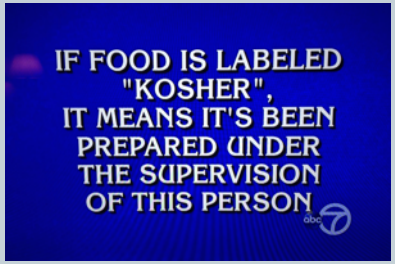
- Jeopardy!
  - Concurso de preguntas y respuestas
  - Las preguntas no son explícitas
  - Las respuestas no forman parte de ninguuna definición. Requiere la asociación de una serie de ideas
- La máquina ha de realizar una búsqueda profunda en miles de artículos
  - La máquina no está conectada a la red

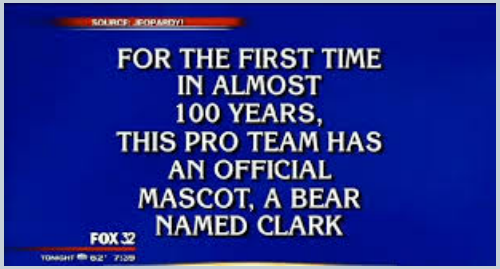
Preguntas Jeopardy!

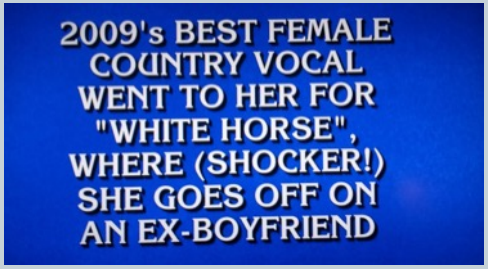




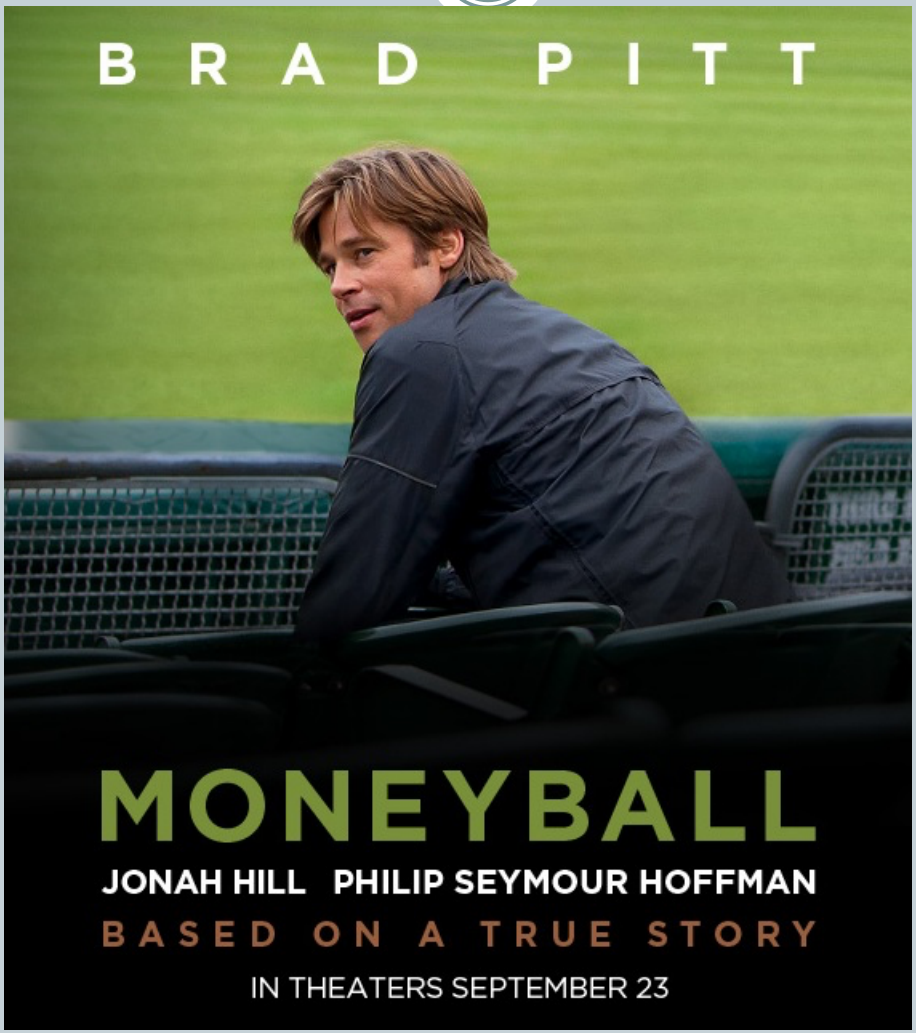








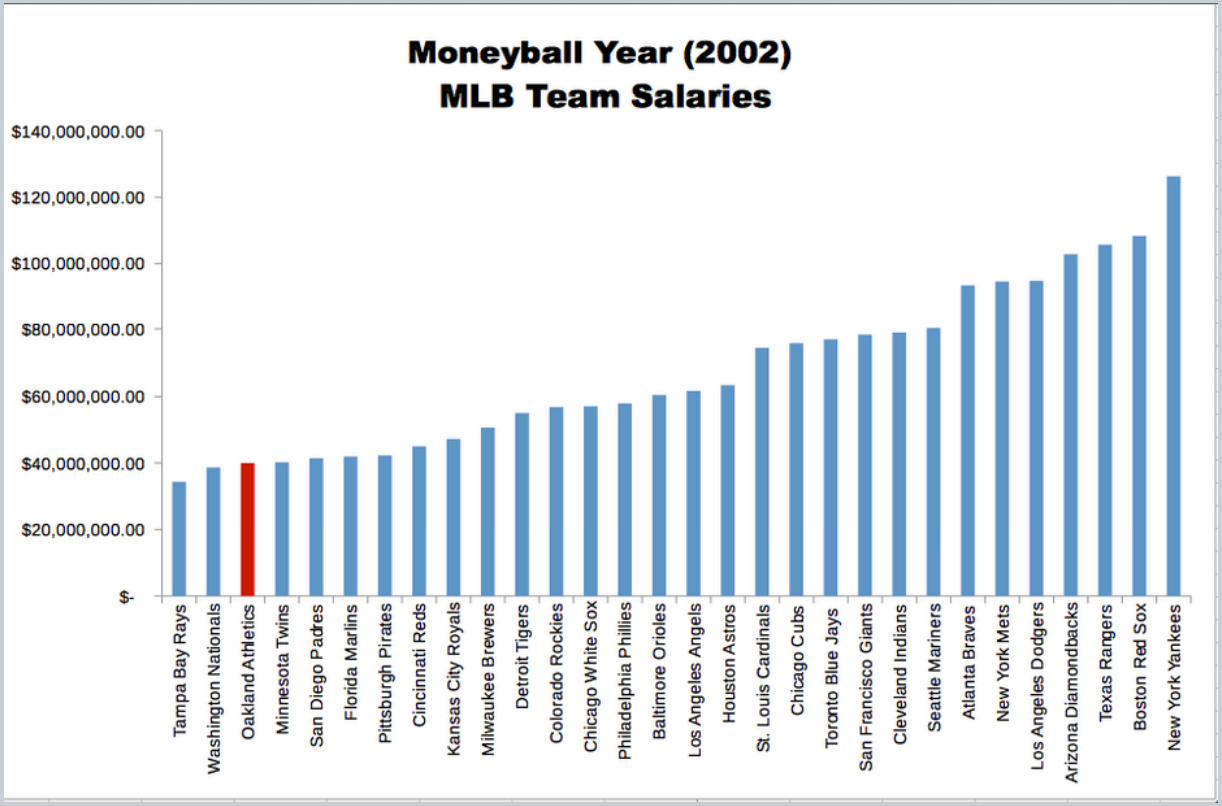
Moneyball


BRAD PITT
MONEYBALL
JONAH HILL PHILIP SEYMOUR HOFFMAN
BASED ON A TRUE STORY
IN THEATERS SEPTEMBER 23

Moneyball - sabermetrics

- Una historia real sobre un entrenador de beisbol
- Billy Beane, entrenador del Oakland Atletics en 2002
  - Un equipo de poca categoría
  - Contrató a Peter Brand, un economista recién licenciado de Harvard
  - Este hace un análisis de regresión múltiple
    - Consisten en encontrar los factores numéricos que son los responsables del éxito en un equipo
    - Intuitivamente, todos contratan al mejor bateador, y su precio sube
    - Pero hay otros factores decisivos, que escapan al ojo del mercado, lo que permite comprar a precios más bajos.

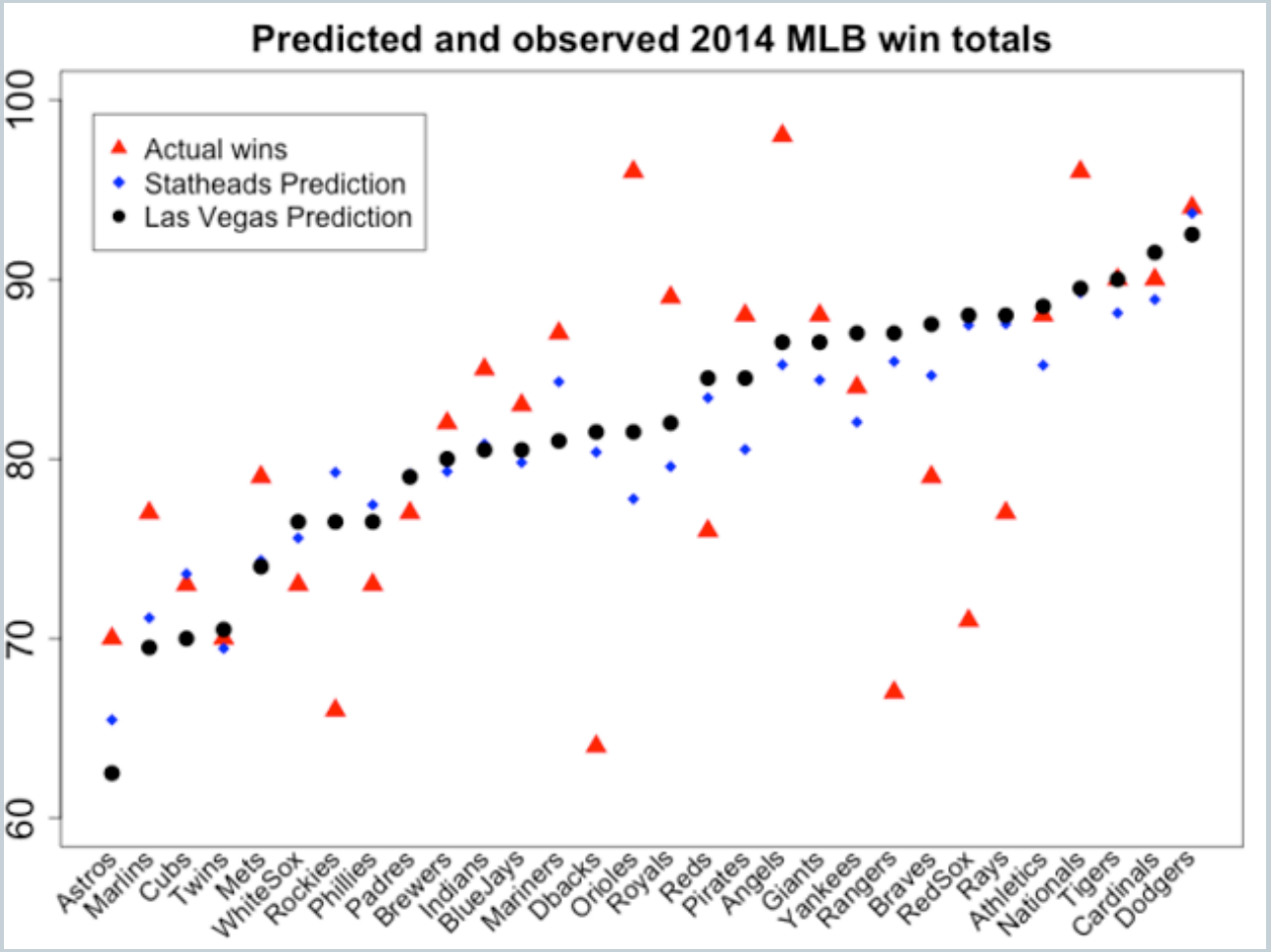
Salarios pagados por los equipos



Moneyball

- El resultado: Oakland Athletics pgana 20 partidos seguidos, el récord de la liga americana.
- Al final, los grandes equipos acaban aprendiendo la técnica, y vuelven a recuperar su dominio.
- Sabermetrics
  - Batting measurements
  - Pitching measurements
  - Fielding measurements

Predicción de victorias



Google Flu Trends

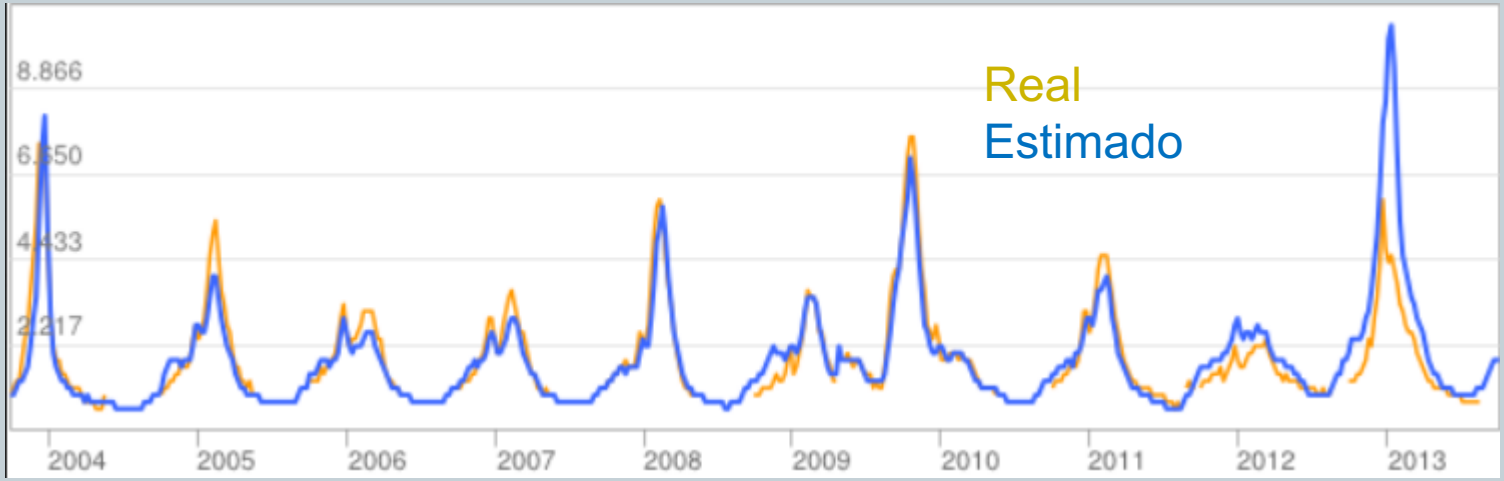
- En 2014, Google se propuso hacer predicciones para ayudar al servicio nacional de salud (NHS).
- Se trataba de detectar el virus de la gripe antes de que llegaran los informes médicos (2 semanas)

Google Flu Trends

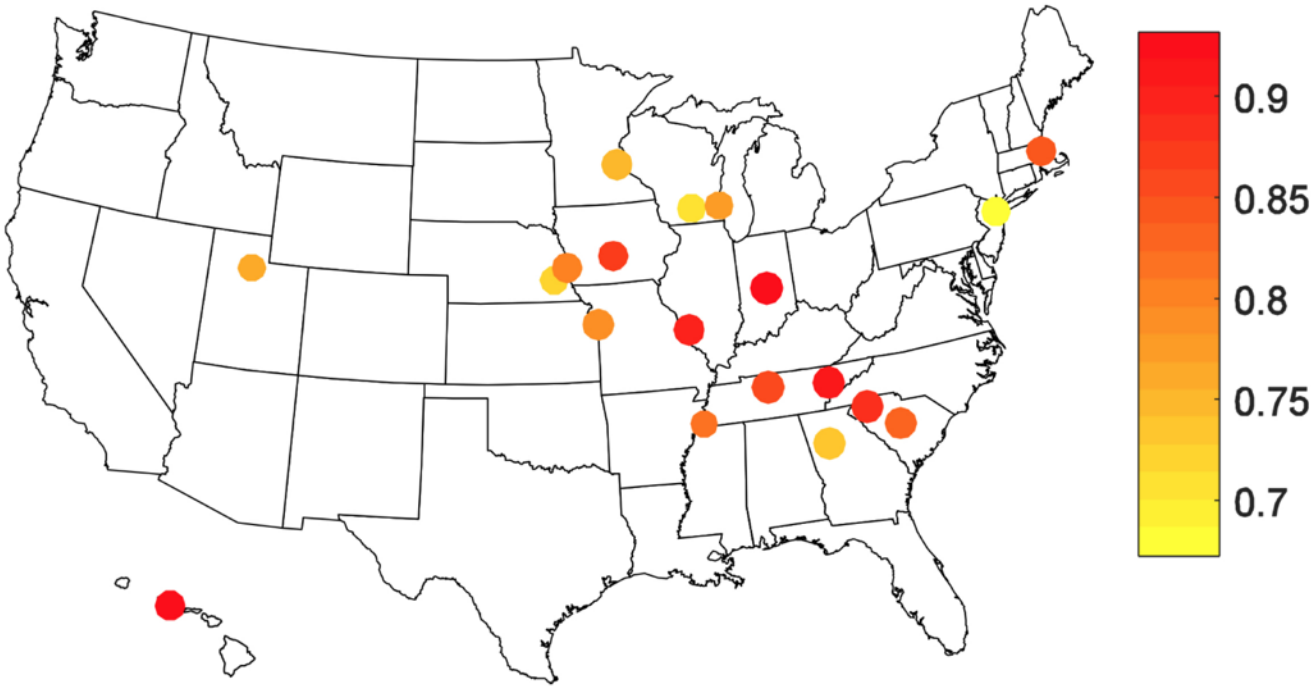
- Se propuso usar las búsquedas de Google para determinar si había algún brote de virus
  - Supuestamente, en la zona donde se produce el brote, hay una incidencia especial de cierto tipo de búsquedas
  - ¿Sabemos a priori qué tipo de búsquedas son especialmente frecuentes cuando hay gripe?
  - No es necesario! El sistema será capaz de encontrarlas

Google Flu Trends

- Es un caso de Big Data: cientos de millones de búsquedas procesadas en 5 años
- Estimación inmediata de la incidencia de gripe en EEUU (frente a 2 semanas del S.Nac. De Salud)
- Artículo *Nature* $$logit(P) = \beta_0 + \beta_1 \times logit(Q) + \varepsilon$$



Brotes de gripe H1N1



Análisis de Google Flu Trends

We designed an **automated method of selecting ILI-related search queries**, requiring **no prior knowledge about influenza**.

We measured how effectively our model would fit the CDC ILI data in each region if we used only a single query as the explanatory variable Q.

Each of the **50 million candidate queries** in our database was separately tested in this manner, to identify the search queries which could most accurately model the CDC ILI visit percentage in each region. [...]

Análisis de Google Flu Trends

The automated query selection process produced a **list of the highest scoring search queries**, sorted by mean Z-transformed correlation across the nine regions.

To decide which queries would be included in the ILI-related query fraction Q, we considered different sets of N top scoring queries.

We measured the performance of these models based on the sum of the queries in each set, and picked N such that we obtained the best fit against out-of-sample ILI data across the nine regions (Figure 1).

Seleccionador de búsquedas

Search Query Topic	Top 45 Queries N	Weighted
Influenza Complication	11	18.15
Cold/Flu Remedy	8	5.05
General Influenza Symptoms	5	2.60
Term for Influenza	4	3.74
Specific Influenza Symptom	4	2.54
Symptoms of an Influenza Complication	4	2.21
Antibiotic Medication	3	6.23
General Influenza Remedies	2	0.18
Symptoms of a Related Disease	2	1.66
Antiviral Medication	1	0.39
Related Disease	1	6.66
Unrelated to Influenza	0	0.00
	45	**49.40**

Otros usos de Bikg Data

- Administración pública
  - Persecución del crimen (minority report)
- Publicidad
- Seguros
- Sistemas de tarifas variables
  - Transportes
  - Hoteles
- Estimación de las probabilidades de éxito de un negocio
- Salud
- Educación…

Pero ¡hay riesgos!

- Conocimiento en manos de empresas y gobierno
- Huida de los datos al extranjero
- Control total vía smartphone
- Indestructibilidad de los datos

Nosotros también podemos hacer Big Data

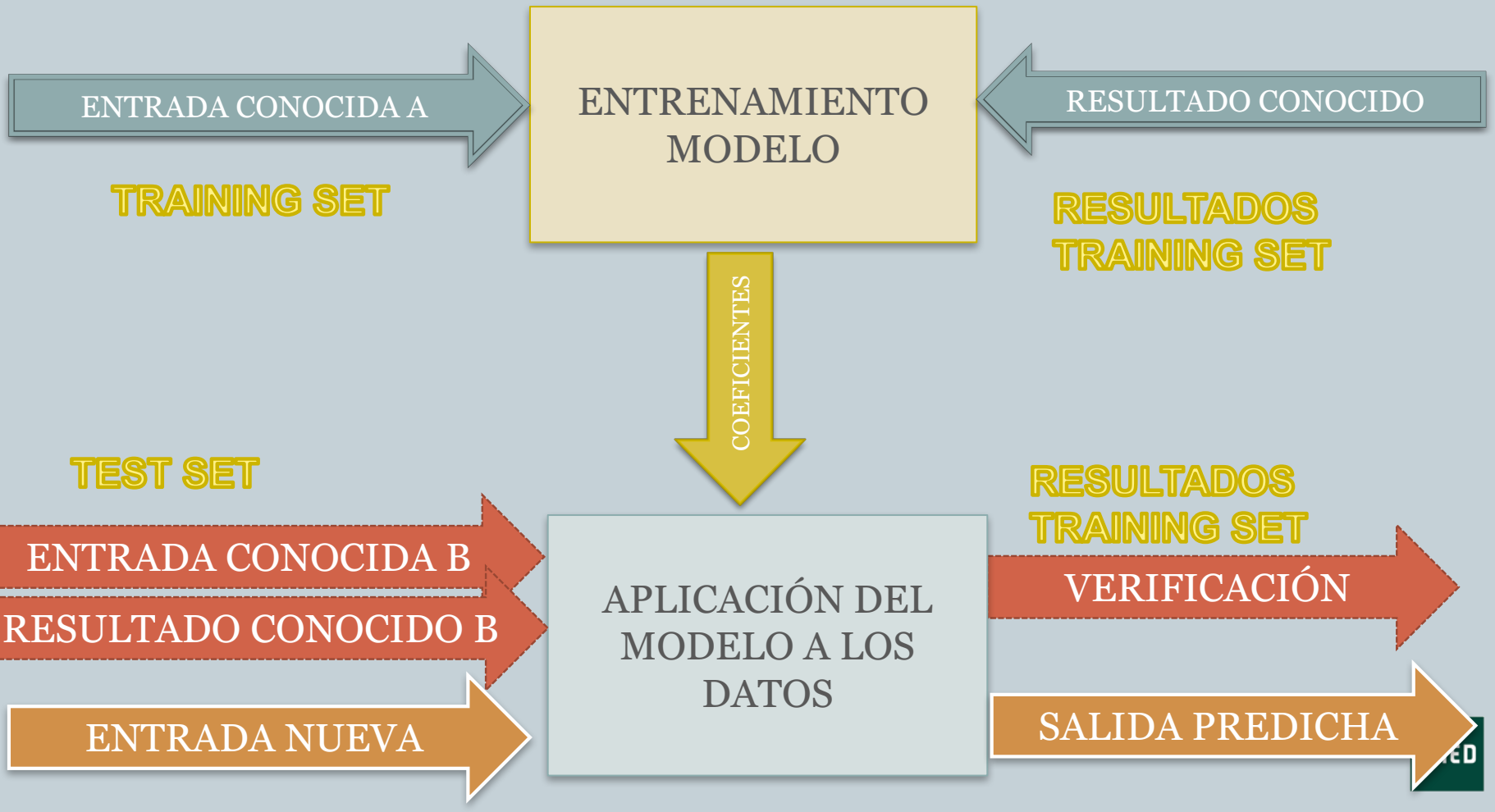
- Curso MIT online: The Analytics Edge
- Lenguaje de programación preferido: R
  - Muy similar a MATLAB, pero con mayores capacidades de proceso de listas y registros
- Muchas organizaciones hacen pública la información que recopila su actividad: OMS, Administración USA…
  - A partir de dicha información es posible hacer análisis y extraer información útil
- También hay empresas que acumulan enormes bases de datos
  - Amazon
  - Spotify
  - Facebook
- Usaremos una base de datos de la industria discográfica

Actividad: predicción de éxitos musicales

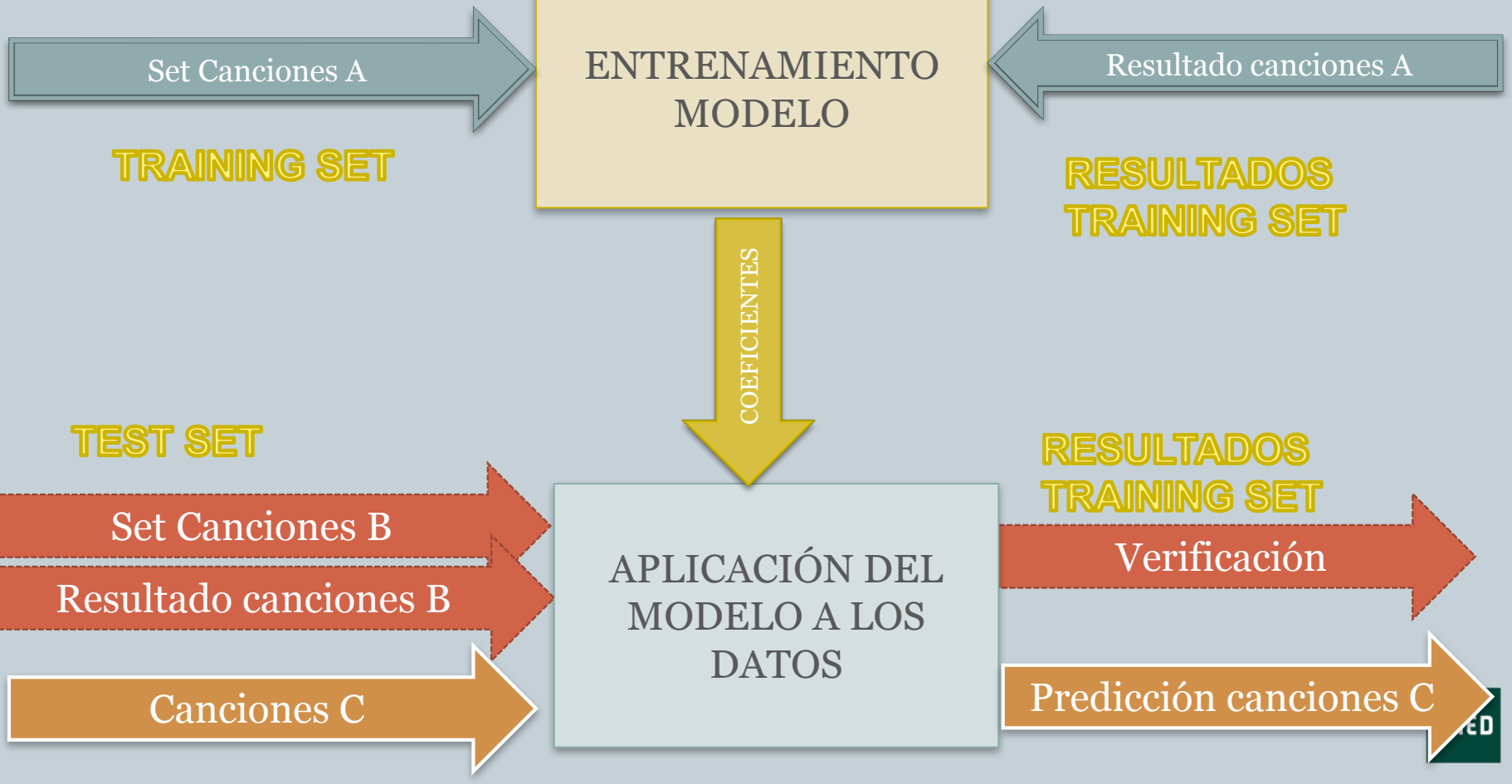
Base de datos de música

- La industria musical: 15.000 M$ anuales
- 3 grandes compañías tienen el 82% del mercado
  - Ofrecen a los artistas recursos para la venta (marketing, distribución…)
  - La inversión por artista es alta
- Sería deseable un predictor de éxito ¿es posible?

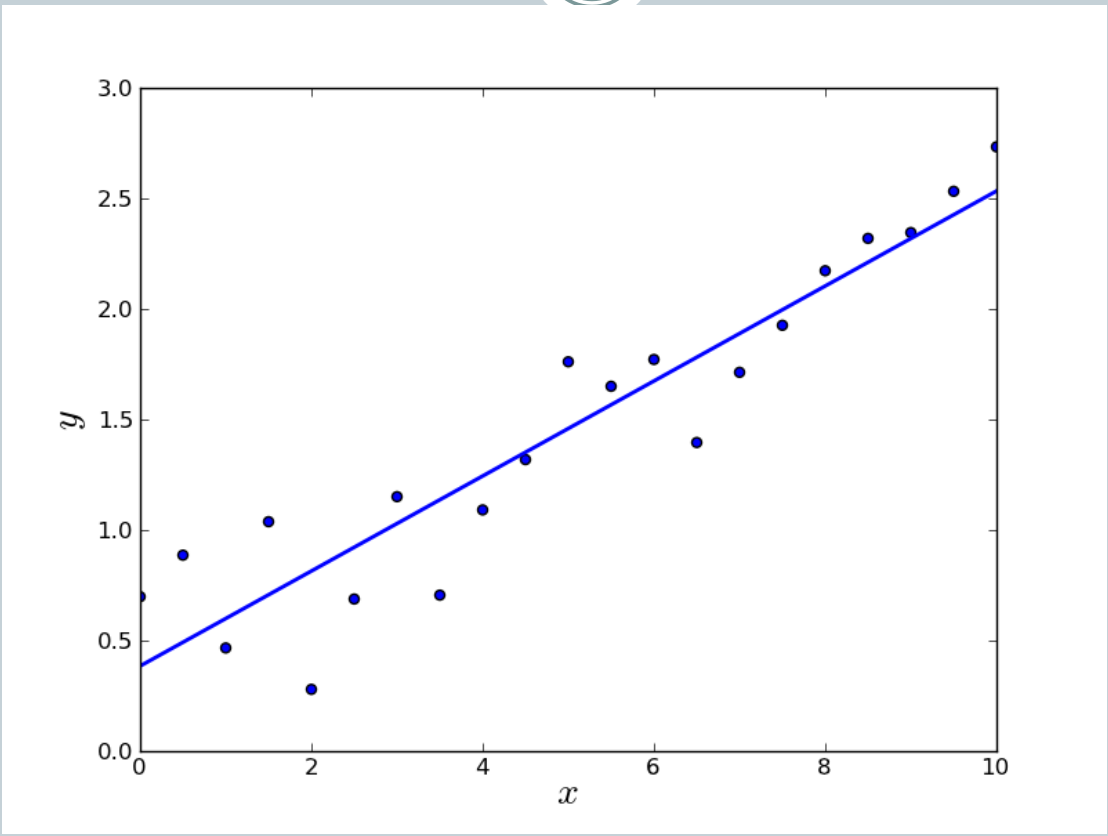
¿Qué es un predictor?



En el caso de los éxitos musicales

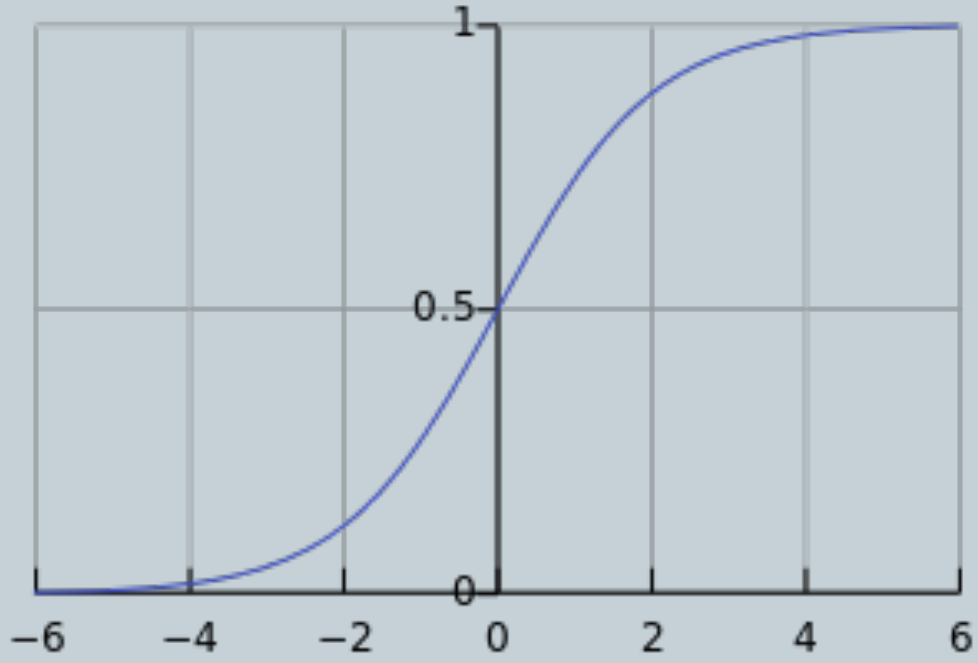


Regresión lineal. Predice un valor y en función de x conocido



$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

Regresión logística: predice una probabilidad de un suceso binario



$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}}$$

La base de datos

Campo	Descripción	Factor?
songtitle	título de la canción	NO
artistname	nombre del artista	
songID y artistID	identificador de la canción	
timesignature	medida del compás y fidelidad del mismo	SI
timesignature_confidence	fidelidad medida del compás	SI
loudness	amplitud media en decibelios	SI
Tempo	golpes de percusión por minuto	SI
tempo_confidence	fidelidad en la medida de tempo	SI
key y key_confidence	estimación de la clave de la canción (sol; fa#, . . ., si…), y al fidelidad de la estimación	SI
energy	estimación de la energía de la canción, usando una mezcla de variables	SI
pitch	variable continua que indica el tono de la canción (agudo, grave)	SI
timbre_0_min, timbre_0_max, timbre_1_min, timbre_1_max, . . . , timbre_11_min, y timbre_11_max	variables que indicsan los valores máximo y mínimo de todos los segmentos para cada uno de los 12 valores en el vector de timbres (24 variables continuas)	SI
Top10	Variable binaria que indica si llegó a la lista de los 10 más vendido del Billboard Hot 100 Chart (1 si llegó al top 10, 0 si no)	NO

Objetivos

- A partir de la base de datos, entrenar un modelo de regresión logística
- El modelo servirá para predecir si una canción será un éxito o no
- Además, compararemos la intuición humana con la capacidad de predicción de la máquina.

Muy breve introducción a R

- Entorno de trabajo por comandos,
  - usa variables, vectores y listas
  - Software libre, Universidad de Atlanta
- Principalmente orientado a estadística
- Es la herramienta más popular en Data Science
- Su mayor arma: la descarga de paquetes de un repositorio
  - Permite el acceso a miles de algoritmos de análisis, resultados de investigaciones en área

Análisis de datos

- Resolveremos el problema en el entorno R
- Con unas pocas líneas construiremos un predictor
- Fases:
  - Carga de datos csv
  - Separación de datos (entrenamiento y test)
  - Entrenamiento
  - Creación del modelo de regresión logística
  - Uso del modelo para predecir el éxito
  - Comprobación de los datos de test

Algunos comandos de R

- Línea de comandos
- Similitudes con MATLAB
  - Matlab ha incorporado numerosas herramientas basadas en R
- Ejemplos de comandos sencillos

```
# VARIABLE NAMES x <- 5
x # 1 way to print out contents of a variable var1 <- 7/2
print(var1) # another way to print contents
valid.variable.name <- 18.6 # you can even have long variable names
valid.variable.name
# if you are of that kind of wierdness
```


```
# VARIABLE NAMES x <- 5
x # 1 way to print out contents of a variable var1 <- 7/2
print(var1) # another way to print contents
valid.variable.name <- 18.6 # you can even have long variable names
valid.variable.name
# if you are of that kind of wierdness
```

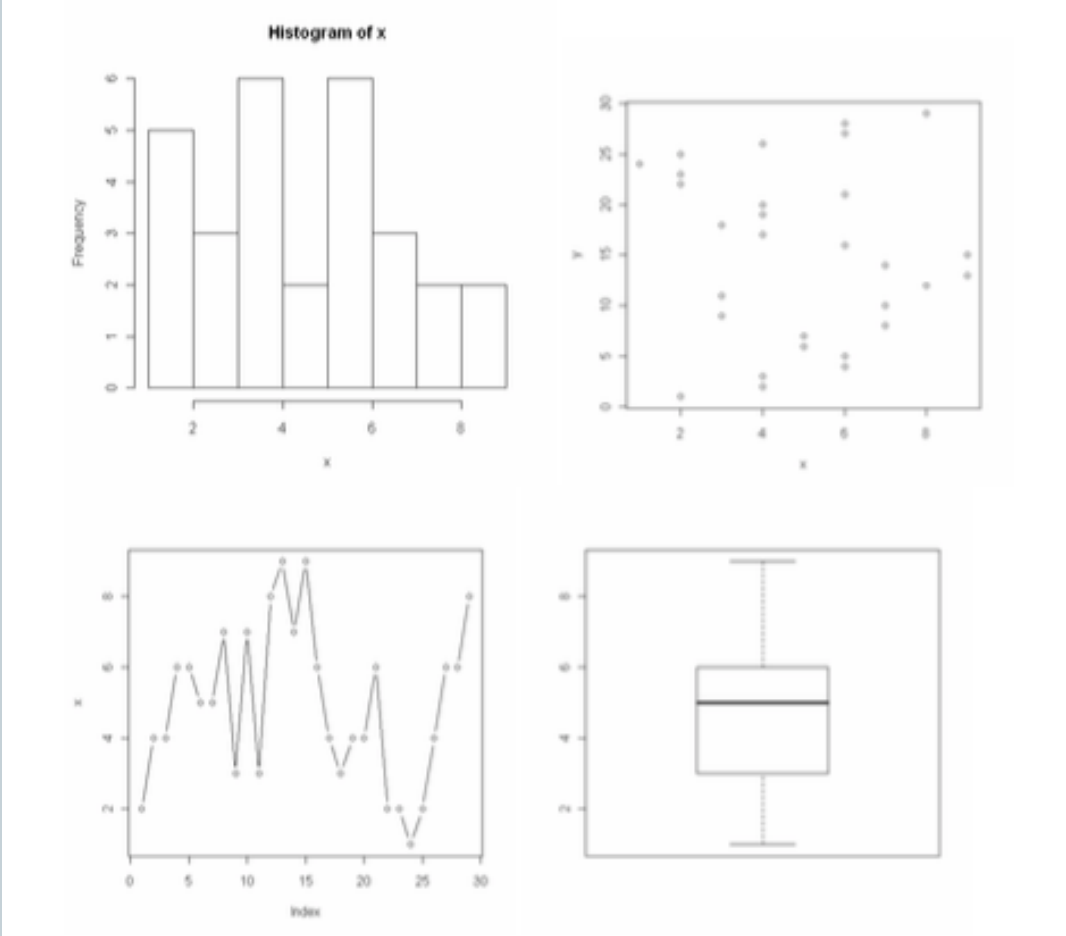
Algunos comandos: vectores y matrices

```
# VECTORS
x <- c(2,3,5,1,4,4)
sum(x)
mean(x)
sd(x)
median(x)
sqrt(x)
x^2
seq(1,10) l's seq(1,10, 2) seq(1:10) seq(1,10,by=2) y<- c(1:7)
y
z<- 1:7
w<-c(1:12,0,-6)
```

Gráficas

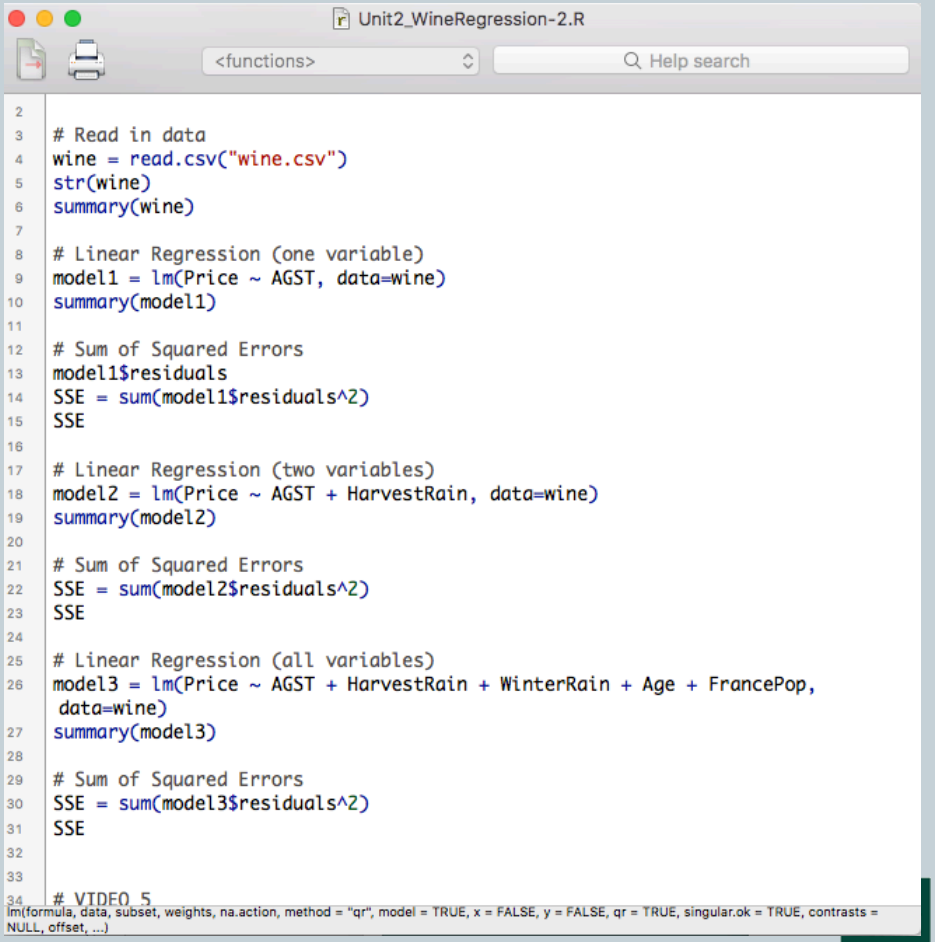
```
# INTRO TO GRAPHING
# input a vector x with data
x <- c(2,4,4,6,6,5,5,7,3,7,3,8,9,7,9,6,4,3,4,4,6,2,2,1,2,4,6,6,8)
y <- c(1:29)
hist(x)
plot(x,y)
plot(x, type="b")
y <- rbinom(20, 12, .4)
hist(y)
y <- rbinom(200, 12, .4)
hist(y)
boxplot(x)
boxplot(x,y)
```

Resultado



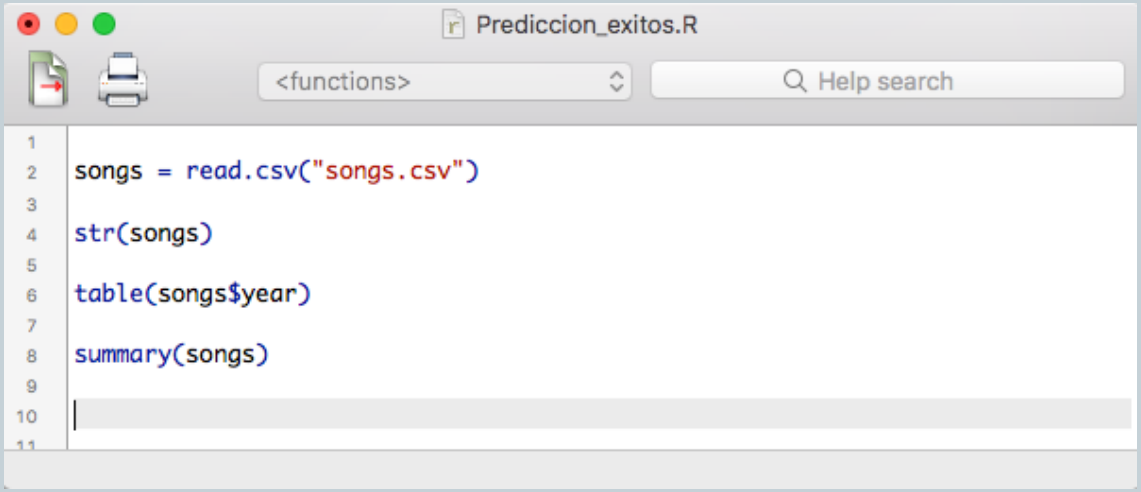
Scripts

- Son secuencias de comandos en R, almacenados en un fichero
- Extensión .R
- Se ejecutan invocando el nombre en línea de comandos



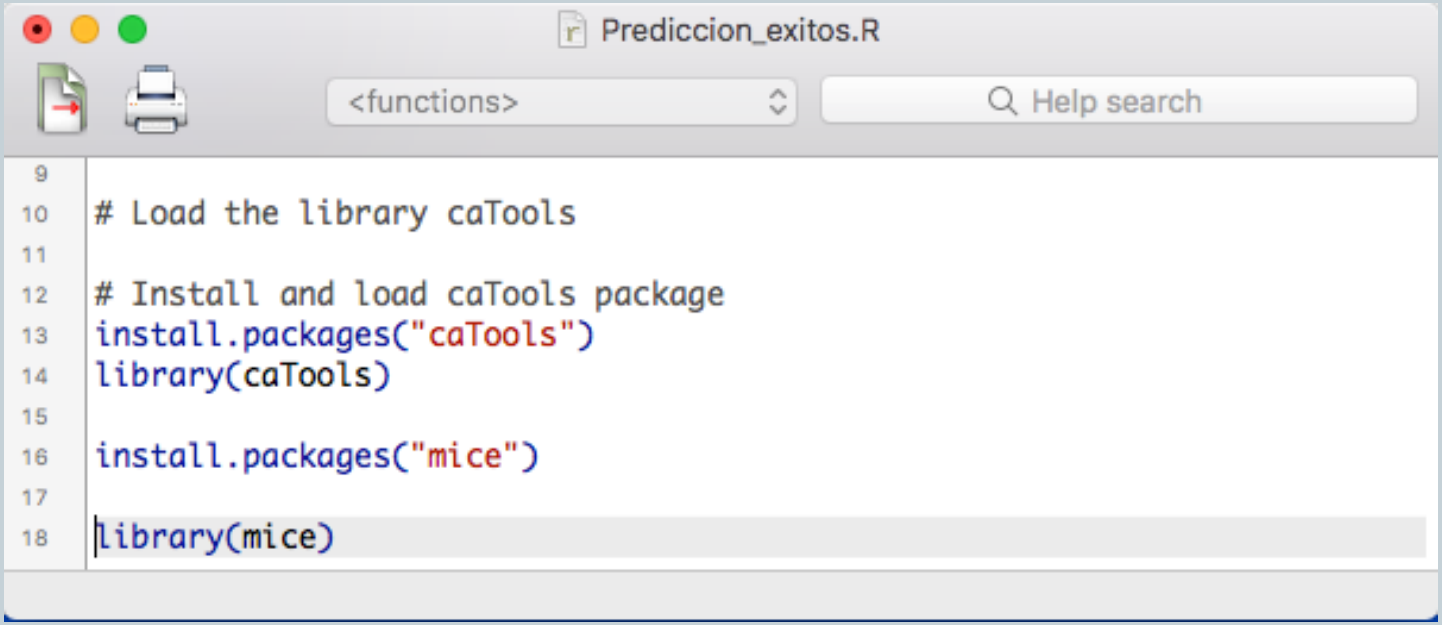
Trabajo con datos reales: BDde música

- El movimiento OpenData
- Descarga de una base de datos real
- Carga de ficheros CSV
- Análisis de las variables



Carga de paquetes

- Alta en un repositorio
- Descarga de paquetes
- Instalación de paquetes



Regresión

- Definición
- Cómputo
- Utilidad

Unit2_WineRegression (1).R

<functions> Help search

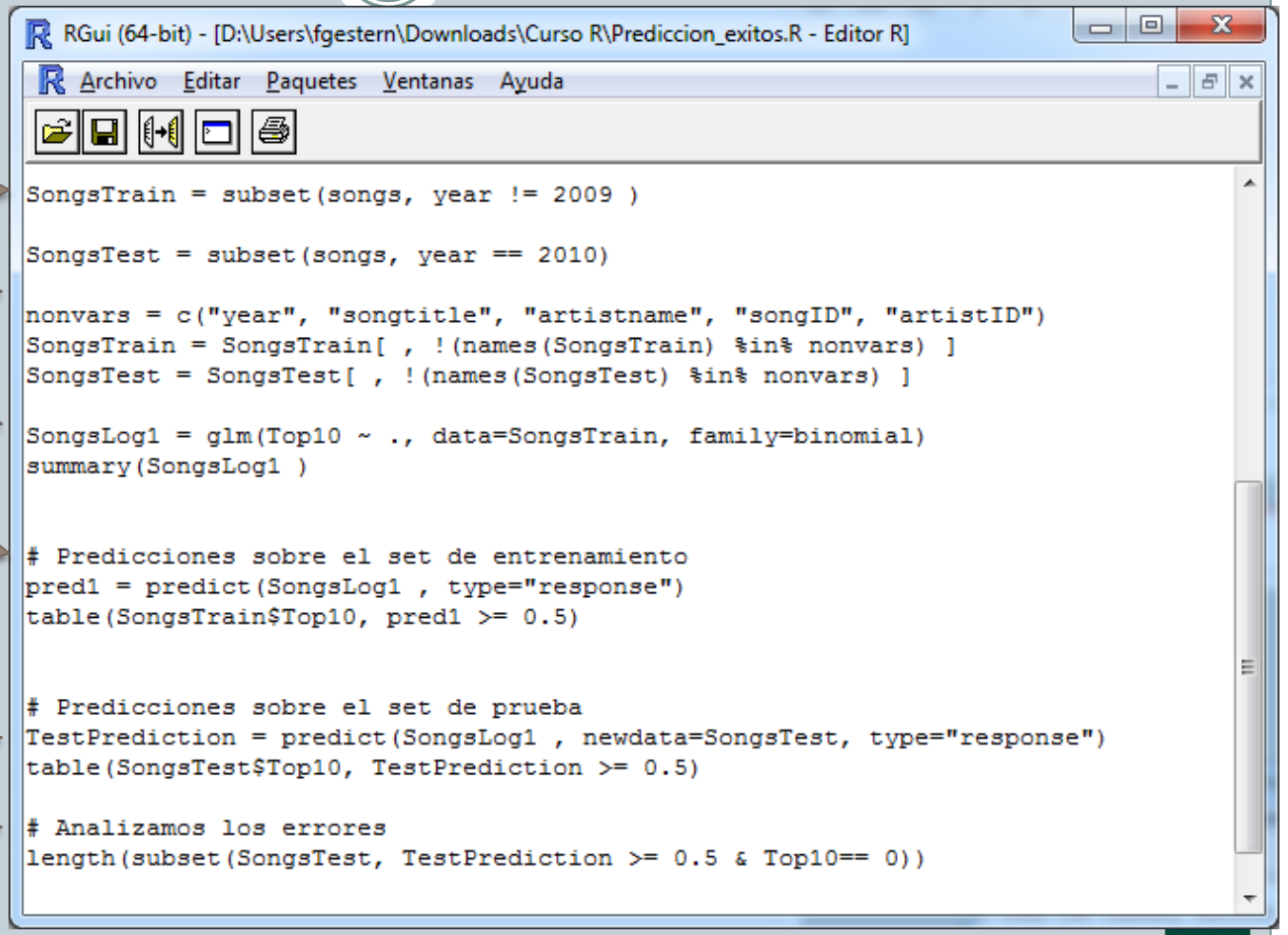
```r
# LECTURA

# Read in data
wine = read.csv("wine.csv")
str(wine)
summary(wine)

# REGRESIÓN LINEAL (UNA VARIABLE)
model1 = lm(Price ~ AGST, data=wine)
summary(model1)

# SUMA DE LOS ERRORES CUADRÁTICOS
SSE = sum(model1$residuals^2)
SSE

# REGRESIÓN BOIVARIABLE
model2 = lm(Price ~ AGST + HarvestRain, data=wine)
summary(model2)

# SUMA DE LOS ERRORES CUADRÁTICOS
SSE = sum(model2$residuals^2)
SSE

# REGRESIÓN MULTIVARIABLE
model3 = lm(Price ~ AGST + HarvestRain + WinterRain + Age + FrancePop, data=wine)
summary(model3)

# SUMA DE LOS ERRORES CUADRÁTICOS
SSE = sum(model3$residuals^2)
SSE

# CORRLACIONES
cor(wine$WinterRain, wine$Price)
cor(wine$Age, wine$FrancePop)
cor(wine)

# ELIMINAMOS CAMPOS CORRELADOS
model5 = lm(Price ~ AGST + HarvestRain + WinterRain, data=wine)
summary(model5)
```

Vuelta al ejemplo de los éxitos musicales

- Pasos
  1. Cargar la base de datos
  2. Preparar/limpiar los datos
  3. Separar el set de entrenamiento del set de verificación
  4. Generar un modelo de regresión logística (glm)
  5. Obtener los coeficientes del modelo
  6. Tratar de predecir el éxito en los datos de verificación
  7. Calcular la tabla de cofusión

Predicción canciones

- Set de aprendizaje
- Set de verificación
- Regresión logística
- Construcción del modelo
- Aplicación del modelo a los datos de prueba
- Validación



Otros algoritmos y aplicaciones

- Predicción de éxitos de vinos
- Recomendaciones en clustering en Netflix
- Evaluación de sentimiento en red

Experiencia

- Trataremos de poner nuestra intuición en competición con la máquina.
  - No modificaremos los pesos, pero sí su orden de importancia
  - Lo haremos con una votación - mqlicker
  - Cada usuario debe puntuar cada uno de estos parámetros por orden de importancia
    - 1: poco importante, 5: muy importante
  - Nos conectamos a http://respond.cc
  - Cada usuario proporciona los valores de 1 a 5 para los parámetros elegidos
  - Comparamos los resultados con el modelo

Conclusiones

- Hemos conocido
  - La historia reciente en torno al Big Data
  - Los conceptos matemáticos básicos
  - La herramienta fundamental: R
  - La construcción de un predictor sobre una base de datos
  - Otros casos prácticos
  - La intuición frente a la máquina

Despedida

Muchas gracias por su atención

fgestern@uloyola.es