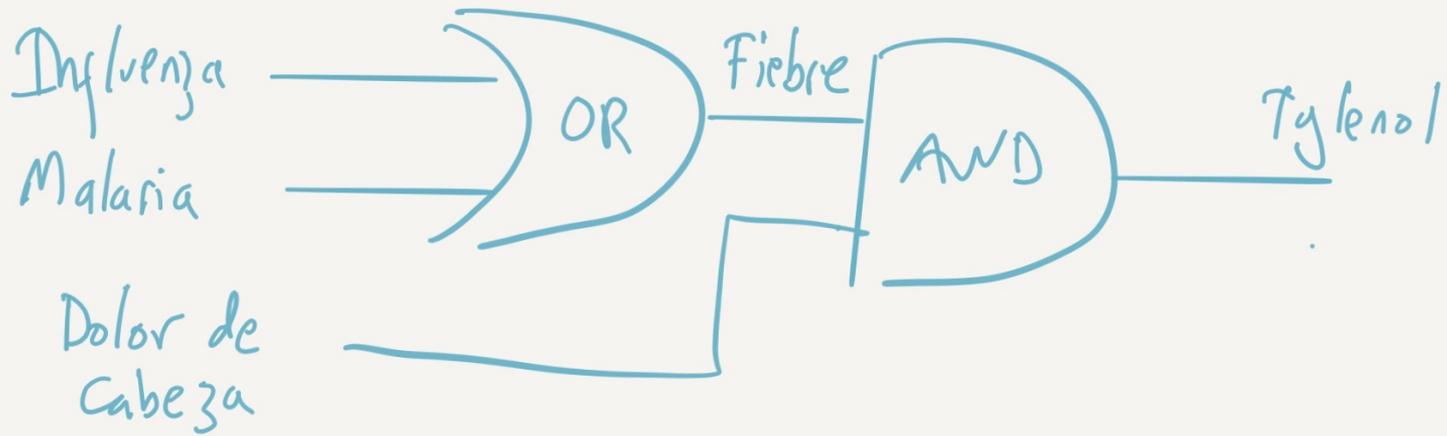


Big Data & Machine Learning

MSc. Ing. Máximo Gurméndez
Universidad de Montevideo



¿Qué es un algoritmo?



¿Qué es Machine Learning?



Vuelca de Tuerta

- Algoritmos creando algoritmos
- La entrada son DATOS la salida es un algoritmo
- ¿Tenemos tantos datos?

¿Qué es Big Data?



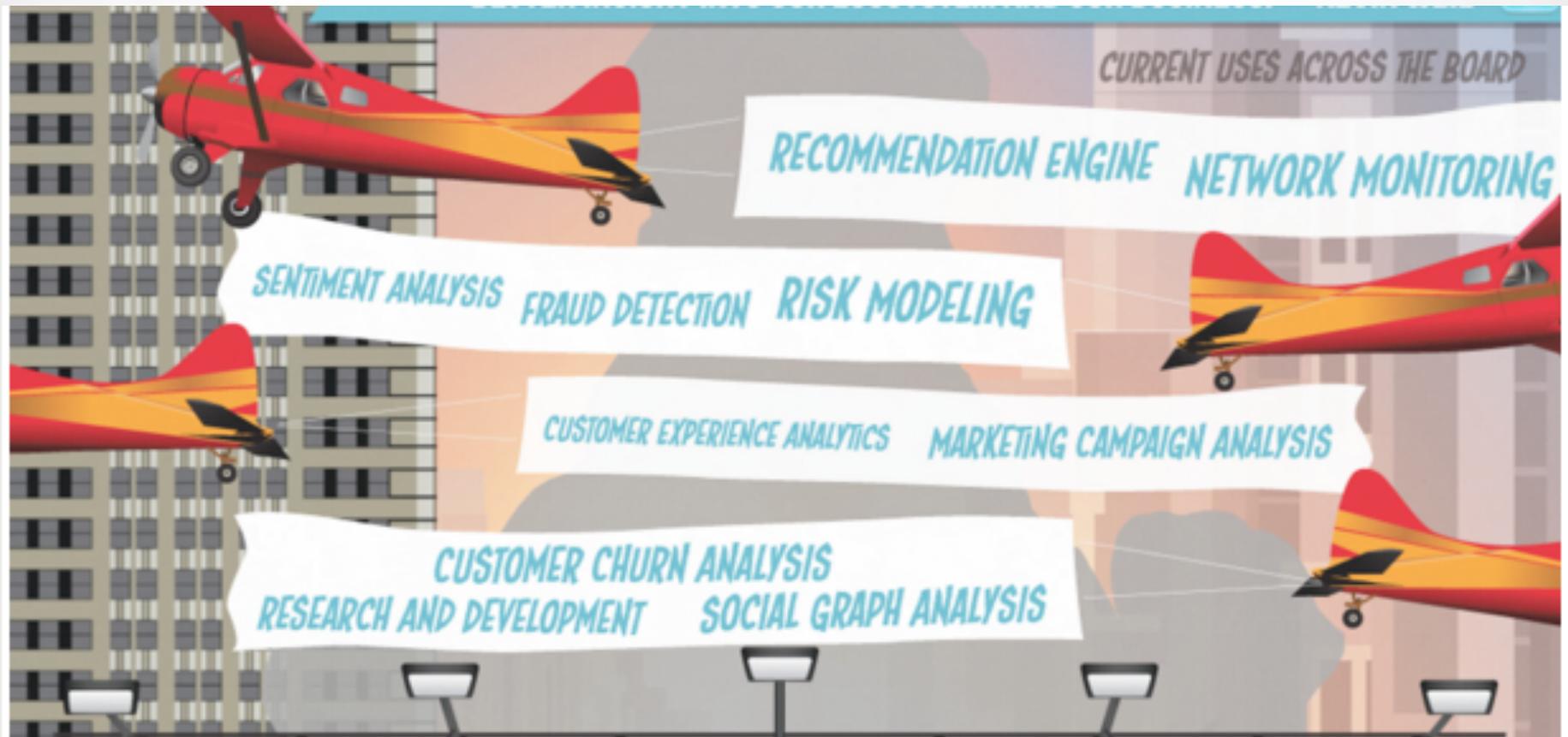
Big Data

- 4 billones de páginas web indexadas
- 300 horas de video subidas a YouTube por minuto
- Walmart maneja 1 millón de transacciones por hora
- 90% de los datos del mundo fueron creados en los últimos 2 años
- Facebook: 30 petabytes de información almacenada, analizada y accedida.
- Twitter: 230 millones de tweets por día
- 300 billones de emails enviados todos los días
-

Big Data

- 2020: 1.7 MB de nueva información producida por segundo por persona en el mundo. 44 Zettabytes
- Google: 40K queries por segundo
- 2015: 1 billón de usuarios usaron facebook en 1 día
- 2015: 1.4 billones de smartphones producidos
- 0.5% de todos los datos no son analizado ni usados.
- 75% de ejecutivos creen que Big Data resultará en > 60% en producción en 2016
- Internet de las cosas

Aplicaciones



Caso: DataXu

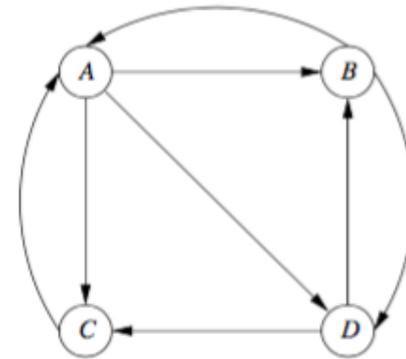


Big Data

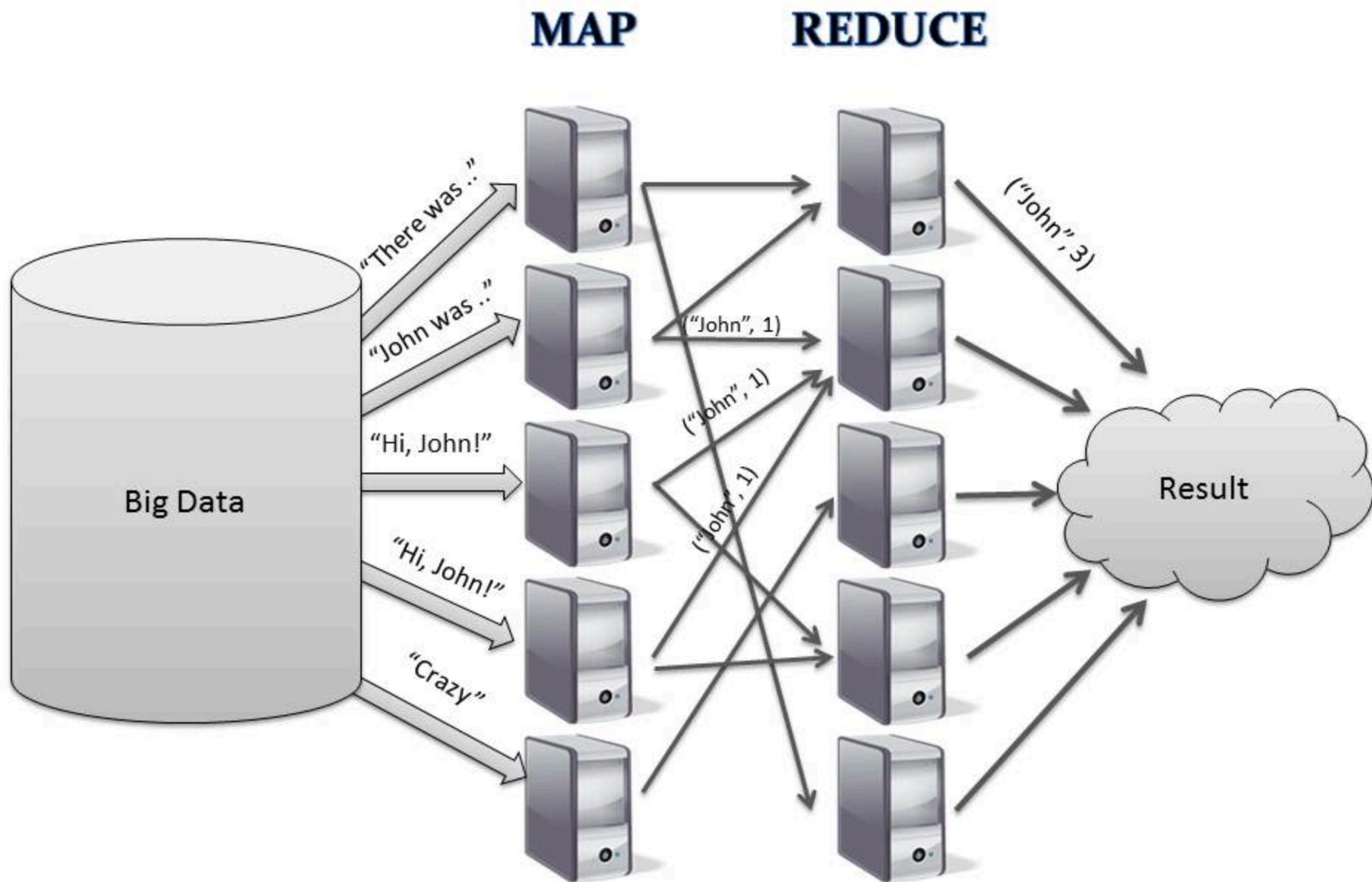
- DataXu:
 - 2 Petabytes procesados cada día
 - 1.8 millones de decisiones de apuestas por segundo
 - 24 X 7 en 5 Continentes
 - Miles de modelos entrenados cada día

Big Data

- Google
 - File System (GFS)
 - Map Reduce
- Hadoop
 - FileSystem (HDFS)
 - Map Reduce, YARN
- Amazon
 - S3
 - Elastic Map Reduce



$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$



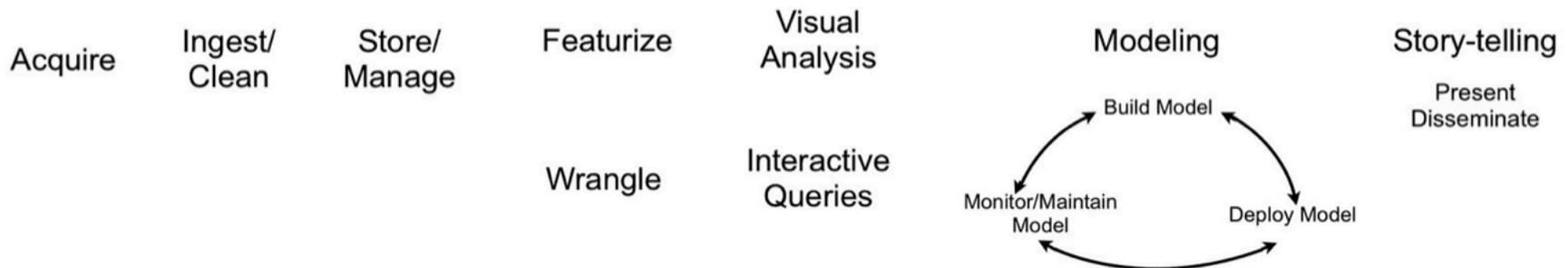
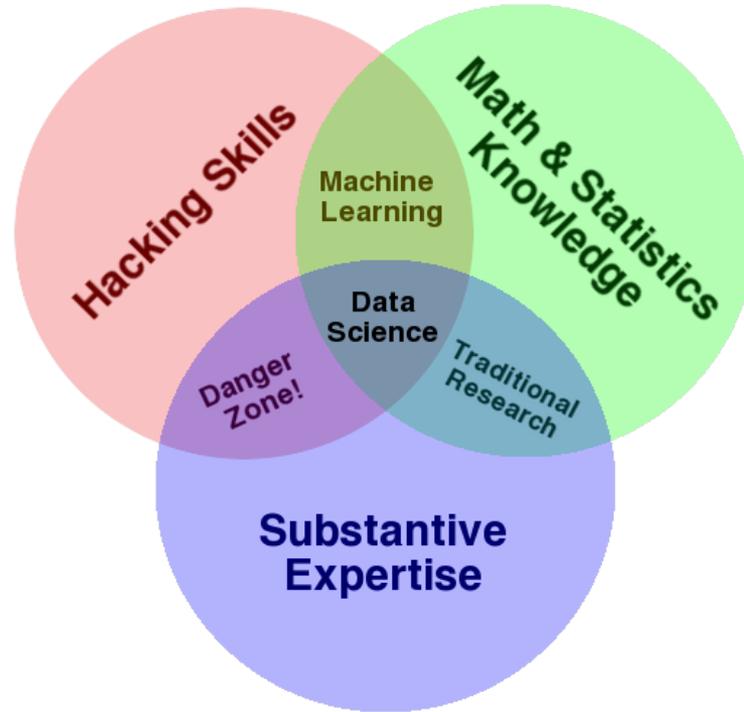
Hadoop Resuelve

- Almacenamiento Distribuido
- Particionar la Información
- Procesamiento distribuido
- Agrupación de datos
- Tolerante a fallas

Apache Spark

- Generalización de MapReduce
- Optimizado para guardar los datos en memoria
- Optimización transparente
- Transformaciones son declarativas
- Compatible con Python, Java y Scala.
- Tiene paquete para ML

¿Qué es Data Science?



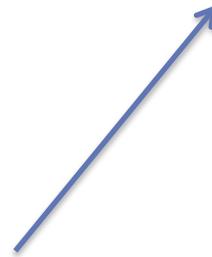
Profesión

- Tim O' Reilly => Data Scientist is the hottest job title in Silicon Valey
- McKinsey: 2018 USA necesitará casi 200 000 más expertos en ML
- ¿Una nueva "Revolución Industrial"?

Ejemplo: Predecir origen de artículos

 República.com.uy

 EL PAIS.com.uy



¿QUÉ DIARIO LO ESCRIBIÓ ?



ARTÍCULO X

Armamos un Data Set



¿Usted decide: ¿sí o no a la baja? ... el clima Montevideo Despejado, 10°

República.com.uy

Martes 16 de septiembre de 2014 | Montevideo - Uruguay

Inicio POLITICA ECONOMIA SOCIEDAD JUSTICIA QUE VOTA TRIBUNA MUNDO OPINION SUPLEMENTOS CHARONA FAMA

DIAS DEL TURISMO NACIONAL

DESCUENTOS EN HOTELES, ALQUILER DE VEHICULOS, RESTORANES Y MUCHOS SERVICIOS TURISTICOS MAS

EVACUADOS EN SIETE DEPARTAMENTOS

Hay 2194 desplazados por las inundaciones en todo el país

El Sistema Nacional de Emergencias (Sinae) informó este martes que asciende a 2194 el número de personas desplazadas en todo el país, de las cuales 303 son evacuadas y 1891 autoevacuadas. Aumentó el número de personas desplazadas en Durazno, se mantiene en San José, Florida y Colonia y se iniciaron evacuaciones en Soriano, Canelones y Treinta y Tres.

➤ Más de 2.000 evacuados en Durazno y la cifra podría aumentar

La República

La enredadera frencampista

Nació hoy un nuevo gobierno: FSI. La idea de Gustavo Carabajal verá gestarse hace tiempo y finalmente toma vida a través de la participación de un grupo de compañeros del diario y de la web.



CRAWLER



Data Set



EL PAIS.com.uy

Montevideo, 17 de septiembre, martes, 16.09.2014 18:30:10

Home Información Mundo Vida actual Opinión Economía Uvación TV Show Divertite SERVICIOS MAS

Elecciones 2014 - Inseguridad ciudadana

Exportadores reclamaron a Vázquez una solución a la pérdida de competitividad

El candidato a la Presidencia por el Frente Amplio, Tabaré Vázquez, se reunió hoy con la Unión de Exportadores. Álvaro Quijón, presidente de esa institución, dijo que se planteó que los Consejos de Salarios deban ir al lado de la productividad.

➤ Mirón: el SA es "soberbio" y lo refleja en "malita" producción de Loreto

FUE EL ÚNICO QUE PUDO FESTEJAR

Cáceres salvó el honor celeste en la Champions League

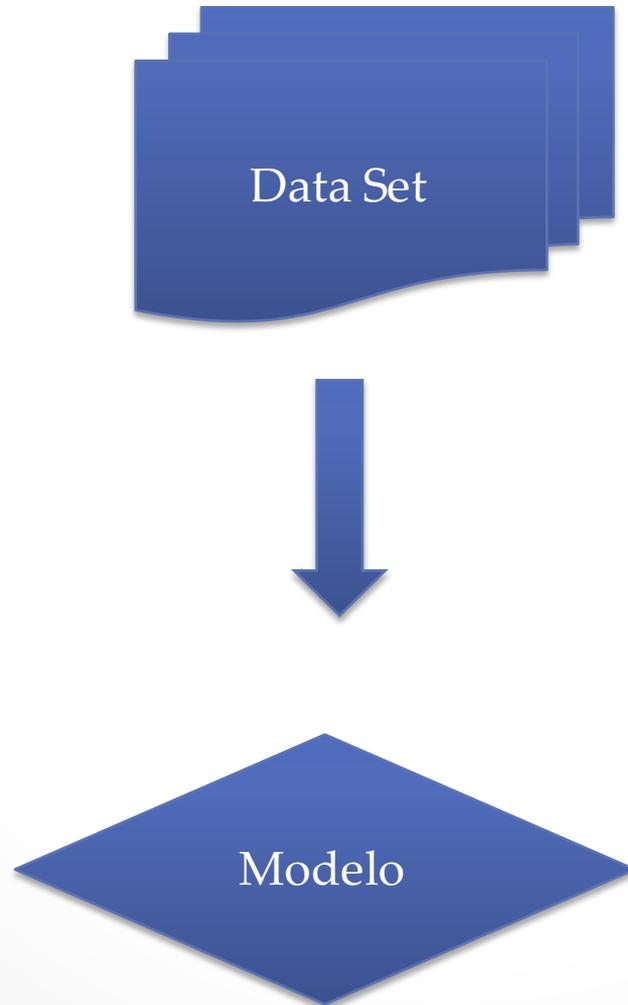
El "Pelado" fue el único uruguayo que pudo festejar en el estadio de la base de grupos del torneo más importante de clubes a nivel europeo.

ARGENTINA

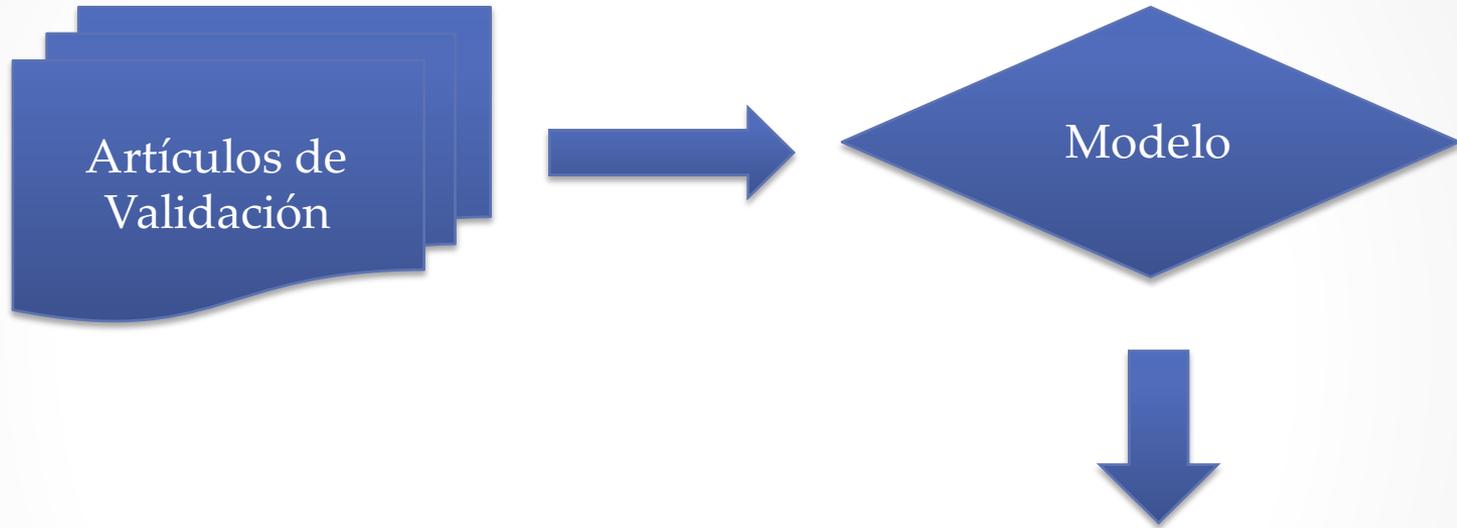
CFK amenaza con echar a embajador de Estados Unidos

La Cancillería argentina transmitió hoy su "profundo malestar" al encargado de negocios de la embajada de Estados Unidos.

Generamos un Modelo



Validamos el Modelo



Origen de Artículo

LR EP LR EP LR EP LR

Predicción:

LR EP EP EP LR LR LR

Precisión 71%

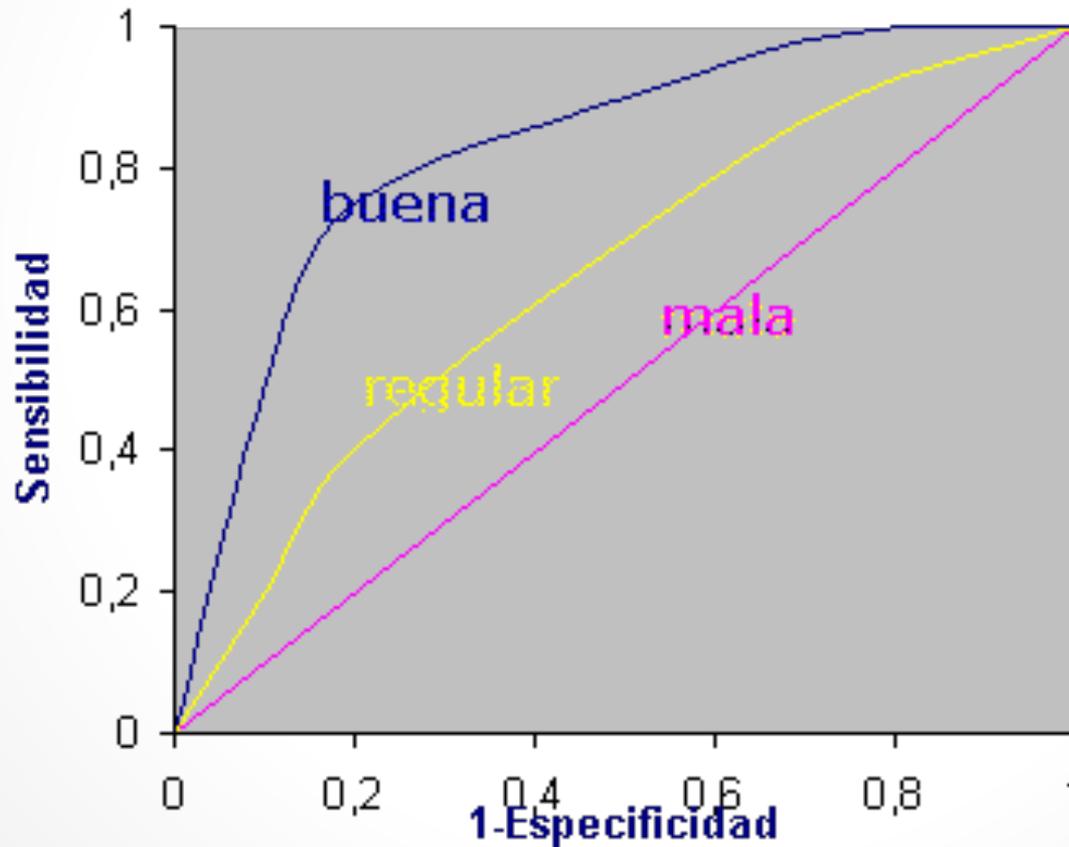
✓ ✓ ✗ ✓ ✓ ✗ ✓

Matriz de Confusión

Diario	Aciertos	Errores	Precisión
El Pais	885	249	78%
La República	2521	201	92%
Precisión			88% (ROC: 0.85)

Curva ROC

Tipos de curvas ROC



Modelo

- Algoritmo:
 - Support Vector Machines (SVM)
- Atributos:
 - n-gramas de las palabras más frecuentes
- Sentiment Analysis:
 - Identificar palabras positivas / negativas
 - Identificar palabras asociadas a los partidos políticos

Proceso

- Selección de los datos
- Pre-procesamiento
- Transformación
- Learning
- Interpretación / Evaluación

¿Cómo funciona?

- Ejemplo:

Predecir el interés que esta charla pueda tener en la audiencia.



Tipo	Edad	Perfil	Interesa_charla
ESTUDIANTE	< 21	TÉCNICO	SI
TRABAJADOR	< 21	EMPRESARIAL	SI
ESTUDIANTE	< 21	EMPRESARIAL	NO
ESTUDIANTE	< 21	ACADÉMICO	SI
ESTUDIANTE	> 21 < 25	EMPRESARIAL	SI
TRABAJADOR	> 21 < 25	ACADÉMICO	NO
ESTUDIANTE	> 21 < 25	EMPRESARIAL	NO
TRABAJADOR	> 21 < 25	EMPRESARIAL	NO
ESTUDIANTE	> 21 < 25	ACADÉMICO	NO
ESTUDIANTE	> 25	TÉCNICO	SI
TRABAJADOR	> 25	TÉCNICO	NO
ESTUDIANTE	> 25	TÉCNICO	SI
ESTUDIANTE	> 25	TÉCNICO	SI
TRABAJADOR	> 25	EMPRESARIAL	NO

Modelo (Árbol de Decisión)

Edad = < 21

| Perfil = TÉCNICO: SI

| Perfil = EMPRESARIAL

| | Tipo = ESTUDIANTE: NO

| | Tipo = TRABAJADOR: SI

| Perfil = ACADÉMICO: SI

Edad = > 21 < 25

| Tipo = ESTUDIANTE

| | Perfil = EMPRESARIAL: SI

| | Perfil = ACADÉMICO: NO

| Tipo = TRABAJADOR: NO

Edad = > 25

| Tipo = ESTUDIANTE: SI

| Tipo = TRABAJADOR: NO

Edad = > 25 : NO

Tipo	Edad	Perfil	Interesa_charla
ESTUDIANTE	< 21	TÉCNICO	SI
TRABAJADOR	< 21	EMPRESARIAL	SI
ESTUDIANTE	< 21	EMPRESARIAL	NO
ESTUDIANTE	< 21	ACADÉMICO	SI
ESTUDIANTE	> 21 < 25	EMPRESARIAL	SI
TRABAJADOR	> 21 < 25	ACADÉMICO	NO
ESTUDIANTE	> 21 < 25	EMPRESARIAL	NO
TRABAJADOR	> 21 < 25	EMPRESARIAL	NO
ESTUDIANTE	> 21 < 25	ACADÉMICO	NO
ESTUDIANTE	> 25	TÉCNICO	SI
TRABAJADOR	> 25	TÉCNICO	NO
ESTUDIANTE	> 25	TÉCNICO	SI
ESTUDIANTE	> 25	TÉCNICO	SI
TRABAJADOR	> 25	EMPRESARIAL	NO

Tipo	Edad	Perfil	Interesa_charla
ESTUDIANTE	< 21	TÉCNICO	SI
ESTUDIANTE	< 21	EMPRESARIAL	NO
ESTUDIANTE	< 21	ACADÉMICO	SI
ESTUDIANTE	> 21 < 25	EMPRESARIAL	SI
ESTUDIANTE	> 21 < 25	EMPRESARIAL	NO
ESTUDIANTE	> 21 < 25	ACADÉMICO	NO
ESTUDIANTE	> 25	TÉCNICO	SI
ESTUDIANTE	> 25	TÉCNICO	SI
ESTUDIANTE	> 25	TÉCNICO	SI
TRABAJADOR	< 21	EMPRESARIAL	SI
TRABAJADOR	> 21 < 25	ACADÉMICO	NO
TRABAJADOR	> 21 < 25	EMPRESARIAL	NO
TRABAJADOR	> 25	TÉCNICO	NO
TRABAJADOR	> 25	EMPRESARIAL	NO

Tipo	Edad	Perfil	Interesa_charla
ESTUDIANTE	< 21	ACADÉMICO	SI
ESTUDIANTE	> 21 < 25	ACADÉMICO	NO
TRABAJADOR	> 21 < 25	ACADÉMICO	NO
ESTUDIANTE	< 21	EMPRESARIAL	NO
ESTUDIANTE	> 21 < 25	EMPRESARIAL	SI
ESTUDIANTE	> 21 < 25	EMPRESARIAL	NO
TRABAJADOR	< 21	EMPRESARIAL	SI
TRABAJADOR	> 21 < 25	EMPRESARIAL	NO
TRABAJADOR	> 25	EMPRESARIAL	NO
ESTUDIANTE	< 21	TÉCNICO	SI
ESTUDIANTE	> 25	TÉCNICO	SI
ESTUDIANTE	> 25	TÉCNICO	SI
ESTUDIANTE	> 25	TÉCNICO	SI
TRABAJADOR	> 25	TÉCNICO	NO

Edad	Perfil	Tipo	Interesa_charla
< 21	ACADÉMICO	ESTUDIANTE	SI
< 21	EMPRESARIAL	ESTUDIANTE	NO
< 21	EMPRESARIAL	TRABAJADOR	SI
< 21	TÉCNICO	ESTUDIANTE	SI
> 21 < 25	ACADÉMICO	ESTUDIANTE	NO
> 21 < 25	ACADÉMICO	TRABAJADOR	NO
> 21 < 25	EMPRESARIAL	ESTUDIANTE	SI
> 21 < 25	EMPRESARIAL	ESTUDIANTE	NO
> 21 < 25	EMPRESARIAL	TRABAJADOR	NO
> 25	TÉCNICO	ESTUDIANTE	SI
> 25	TÉCNICO	ESTUDIANTE	SI
> 25	TÉCNICO	ESTUDIANTE	SI
> 25	TÉCNICO	TRABAJADOR	NO
> 25	EMPRESARIAL	TRABAJADOR	NO

Algoritmo (ID3)

Id3(Ejemplos, Atributo-objetivo, Atributos)

Si todos los ejemplos son positivos devolver un nodo positivo

Si todos los ejemplos son negativos devolver un nodo negativo

Si Atributos está vacío devolver el voto mayoritario del valor del atributo objetivo en
Ejemplos

En otro caso

Sea A Atributo el MEJOR de atributos

Para cada v valor del atributo hacer

Sea Ejemplos(v) el subconjunto de ejemplos cuyo valor de atributo A es v

Si Ejemplos(v) esta vacío devolver un nodo con el voto mayoritario del

Atributo objetivo de Ejemplos

Sino Devolver Id3(Ejemplos(v), Atributo-objetivo, Atributos/{A})

ML @ Negocio

- Google vs Yahoo:
 - Gana quien tenga más datos y mejores algoritmos.
- Etapas:
 - Manual, Automatizado, ML-izado!
- Empresas se tienen que subir al barco

ML @ Ciencias

- Super Método Científico
- ¿Podemos crear teorías desde los datos?
- Experimentación Biológica: Adam analiza información y crea nuevos experimentos

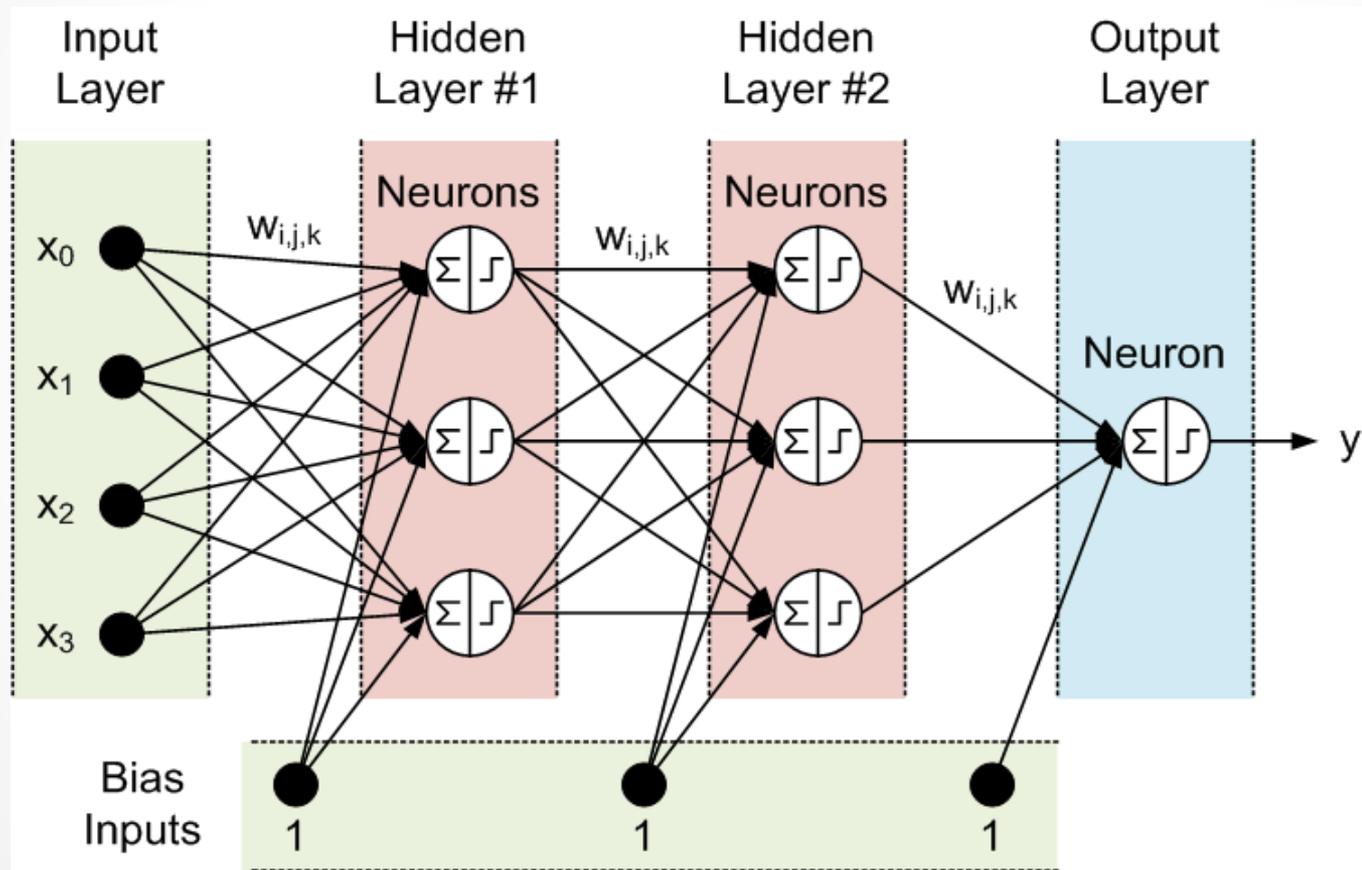
ML @ Política

- Rayid Ghani: Experto ML Chief Scientist de la campaña de Obama.
- Cada noche sistema corría 66000 simulaciones y encausaba a voluntarios en cuanto a quién llamar y qué decir.
- "Un billón de Bill Clintons"

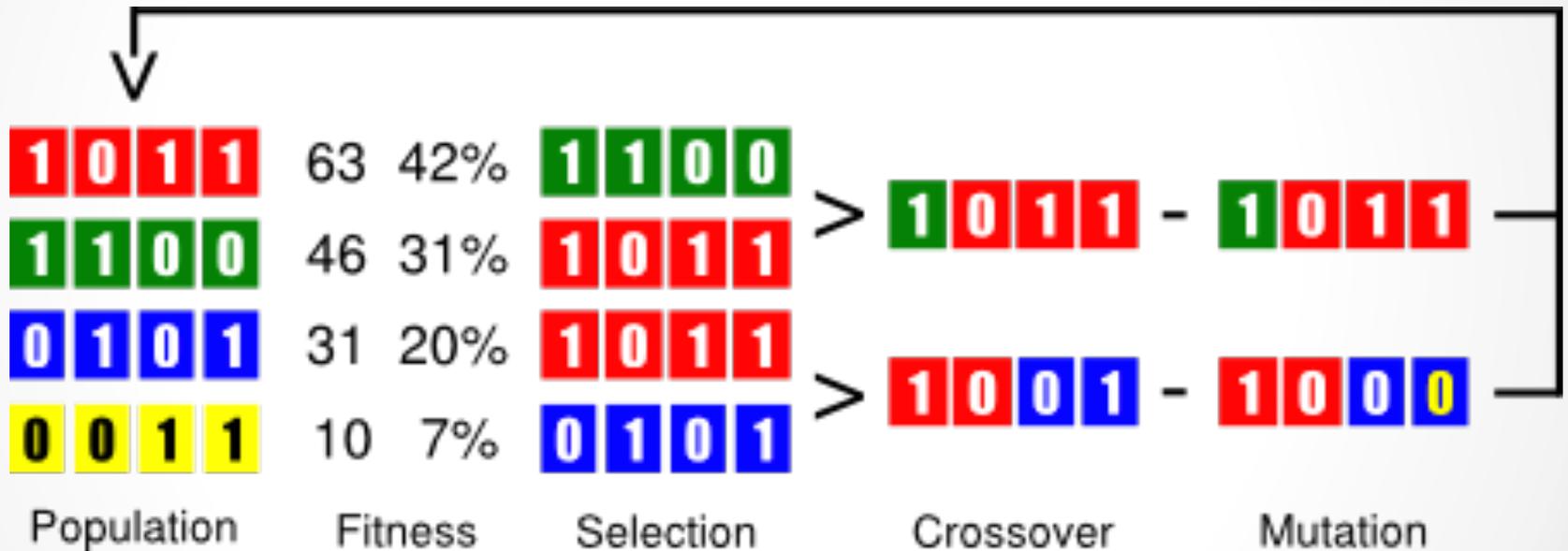
Escuelas de ML

- Symbologists: deducción inversa
- Connectionists: back-propagation
- Evolutionaries: algoritmos genéticos
- Bayesian: Inferencia Bayesiana
- Analogizers: KNN / SVM

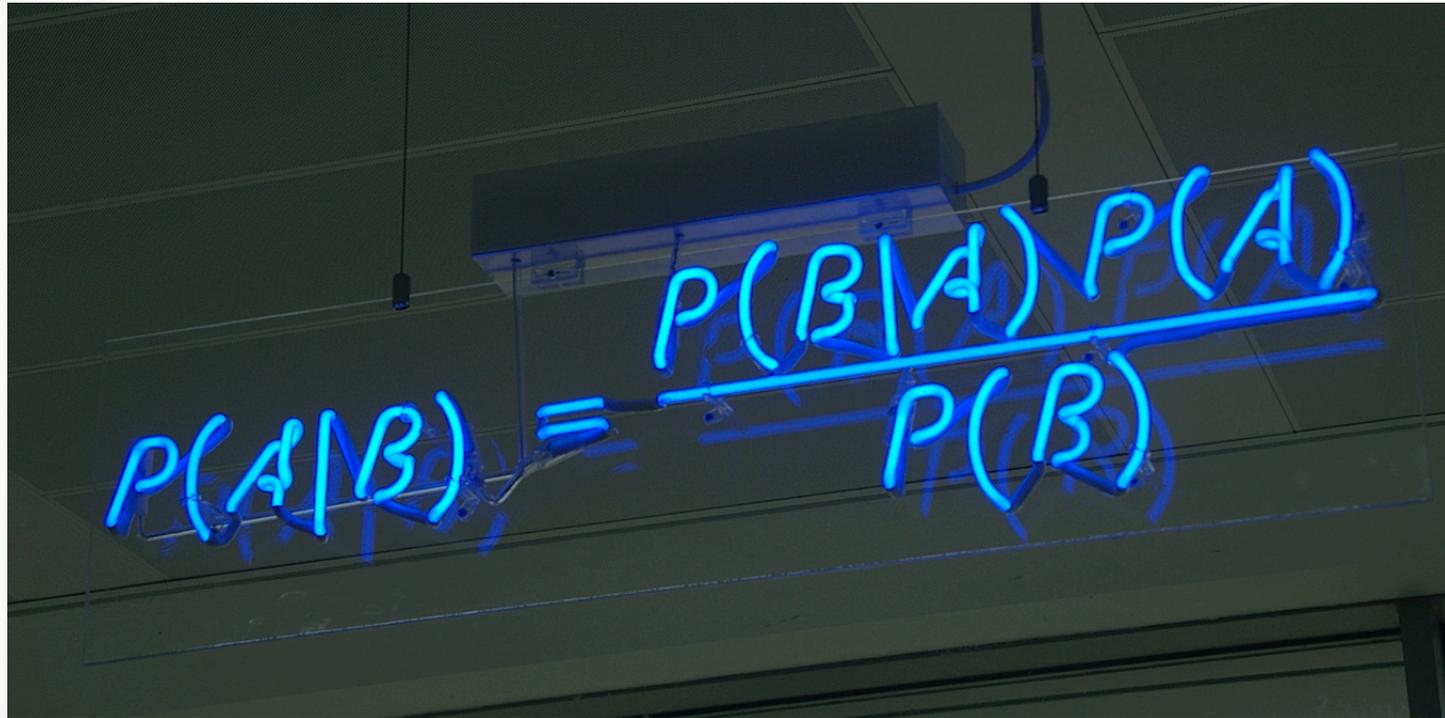
Redes Neuronales



Algoritmos Genéticos



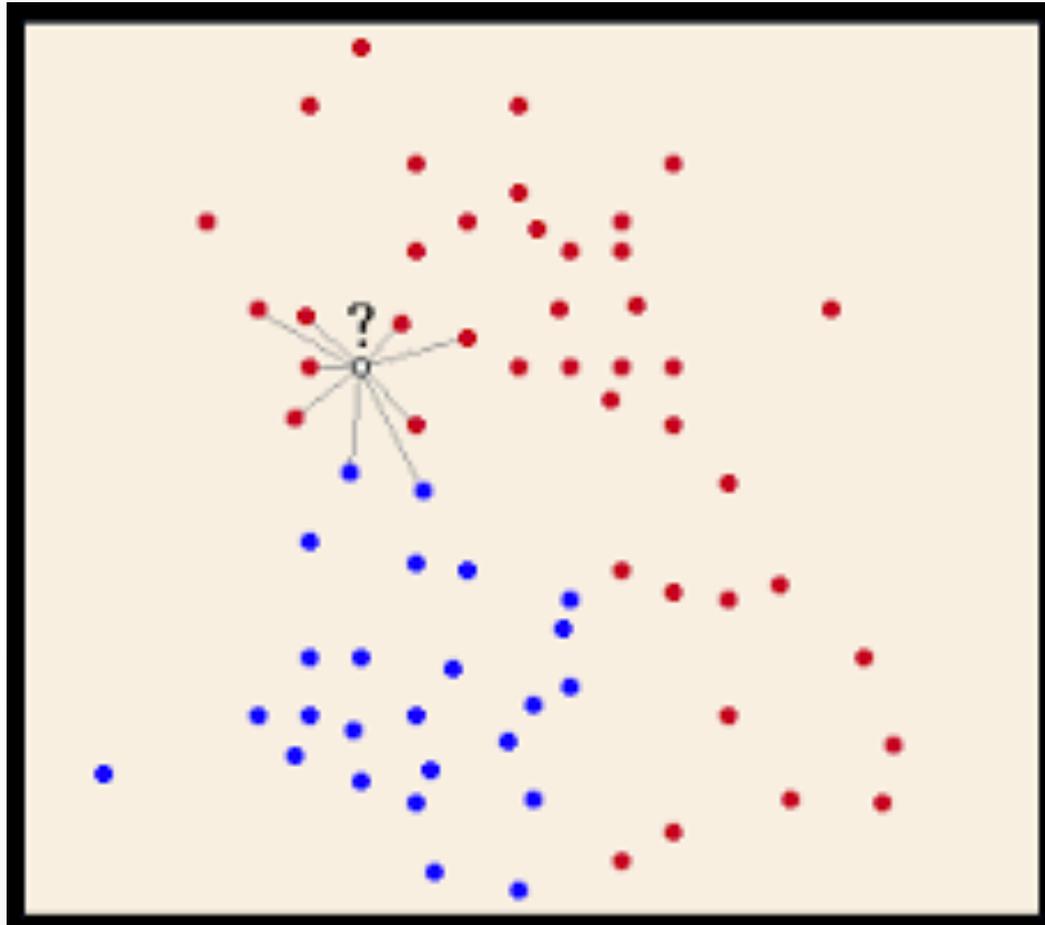
Algoritmos Bayesianos



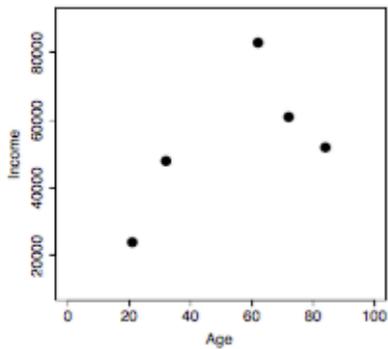
A photograph of a whiteboard with the Bayesian formula $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ written in blue marker. The whiteboard is mounted on a wall, and the lighting is dim, with the blue marker providing the primary illumination for the text.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

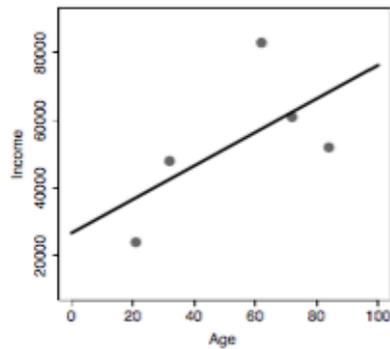
KNN



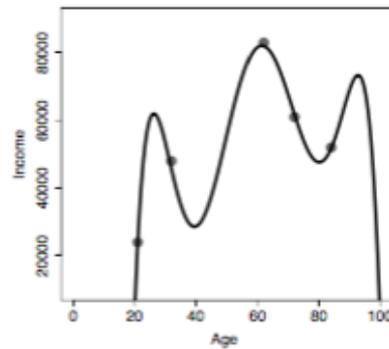
Fitting



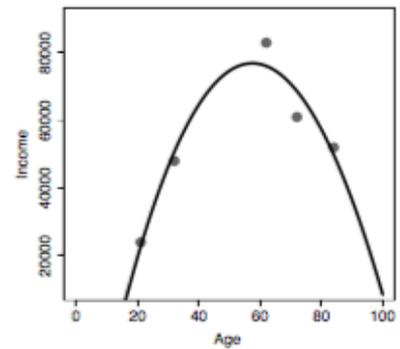
(a) Dataset



(b) Underfitting



(c) Overfitting

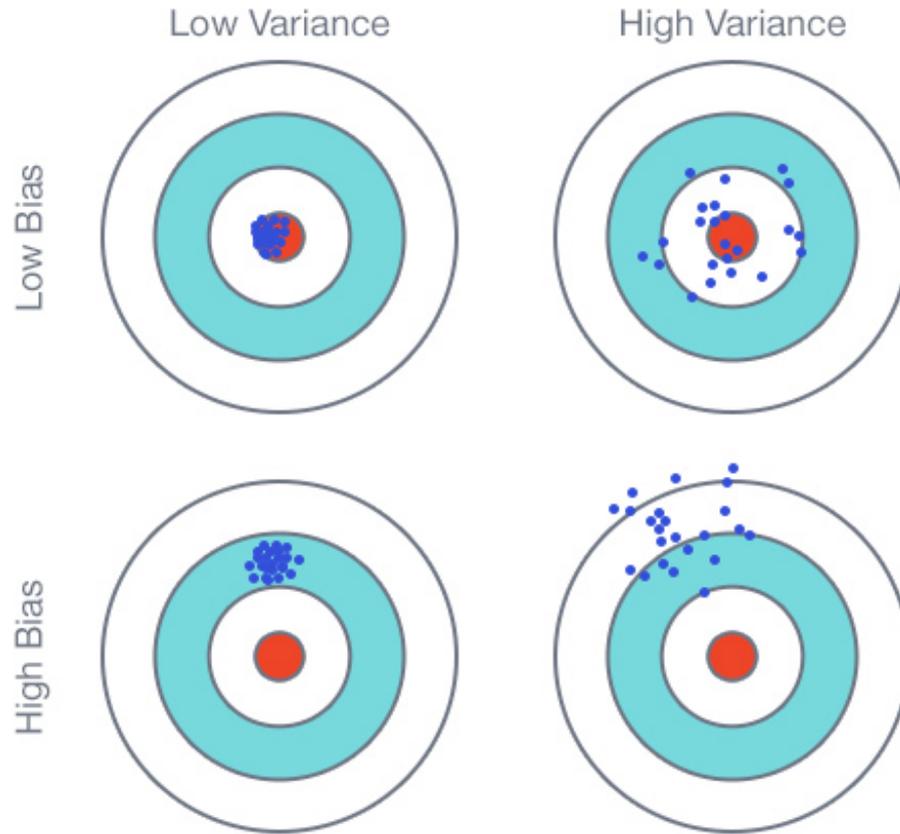


(d) Just right

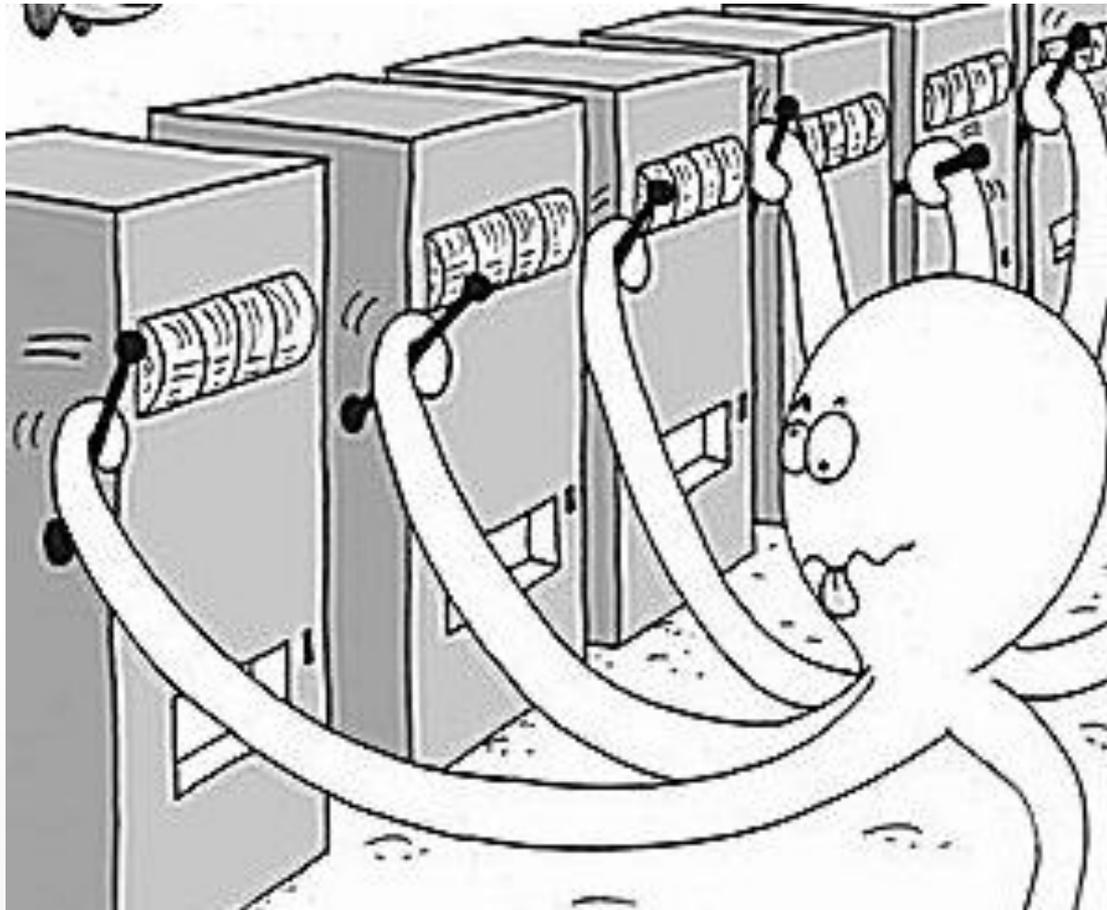
ML – “Ill Posed”

- El problema de Hume
- [Funes el Memorioso](#)
- No Free Lunch Theorem

Bias vs Variance



Exploración vs Explotación



“El gran objetivo de toda ciencia es cubrir el mayor número de hechos empíricos por deducción lógica a partir del menor número de hipótesis o axiomas”

–Albert Einstein

Muchas Gracias

¿Preguntas?

Links

Links

- http://mobile.blogs.wsj.com/cio/2014/07/10/germanys-12th-man-at-the-world-cup-big-data/?mg=blogs-wsj&utm_content=buffer378d4&utm_medium=social&utm_source=linkedin.com&utm_campaign=buffer
- http://en.wikipedia.org/wiki/Big_data
- <https://www.linkedin.com/pulse/article/20131113065157-64875646-the-awesome-ways-big-data-is-used-today-to-change-our-world>
- <https://spark.apache.org/>
- www.dataxu.com
- <https://mahout.apache.org/users/basics/algorithms.html>
- <http://hortonworks.com/hadoop/yarn/>
- <http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html>
- <http://www.ibmbigdatahub.com/gallery/quick-facts-and-stats-big-data>
- <https://yoyoclouds.wordpress.com/tag/hdfs/>
- <http://www.baselinemag.com/analytics-big-data/slideshows/surprising-statistics-about-big-data.html>
- <http://www.forbes.com/sites/tedgreenwald/2012/03/29/big-data-small-decisions-rendered-very-quickly-under-the-hood-at-dataxu/>
- <http://www.worldwidewebsite.com/>
- <http://papers.nips.cc/paper/3150-map-reduce-for-machine-learning-on-multicore.pdf>
- <http://www.cs.ubc.ca/~murphyk/MLbook/>
- http://en.wikipedia.org/wiki/Data_mining#Process

Imágenes

- http://www.pakwheels.com/forums/attachments/guess-humor-hobbies/545611d1210727280-how-many-people-can-u-fit-ur-car-volkswagen-beetle_1938_800x600_wallpaper_1f_bob_pakwheels-com-jpg
- <http://www.enterrasolutions.com/media/images/2013/08/6a00d8341c4ebd53ef0191047c2f5c970c-pi.png>
- <http://www.baselinemag.com/analytics-big-data/slideshows/surprising-statistics-about-big-data.html>
- <http://www.ibmbigdatahub.com/gallery/quick-facts-and-stats-big-data>
- <https://yoyoclouds.wordpress.com/tag/hdfs/>
- http://www.hrc.es/bioest/roc_21.gif
- <http://techblog.baghel.com/index.php?itemid=132>