



INTELIGENCIA DE NEGOCIO

2018 - 2019

- Tema 1. Introducción a la Inteligencia de Negocio
- Tema 2. Minería de Datos. Ciencia de Datos
- Tema 3. Modelos de Predicción: Clasificación, regresión y series temporales
- Tema 4. Preparación de Datos
- Tema 5. Modelos de Agrupamiento o Segmentación
- Tema 6. Modelos de Asociación
- Tema 7. Modelos Avanzados de Minería de Datos.
- Tema 8. Big Data



Ben Chams - Fotolia

Objetivos:

- Introducir los conceptos de Ciencia de Datos, Minería de Datos, Big Data
- Conocer las etapas del proceso de minería de datos
- Conocer los problemas clásicos de minería de datos

[illegible]

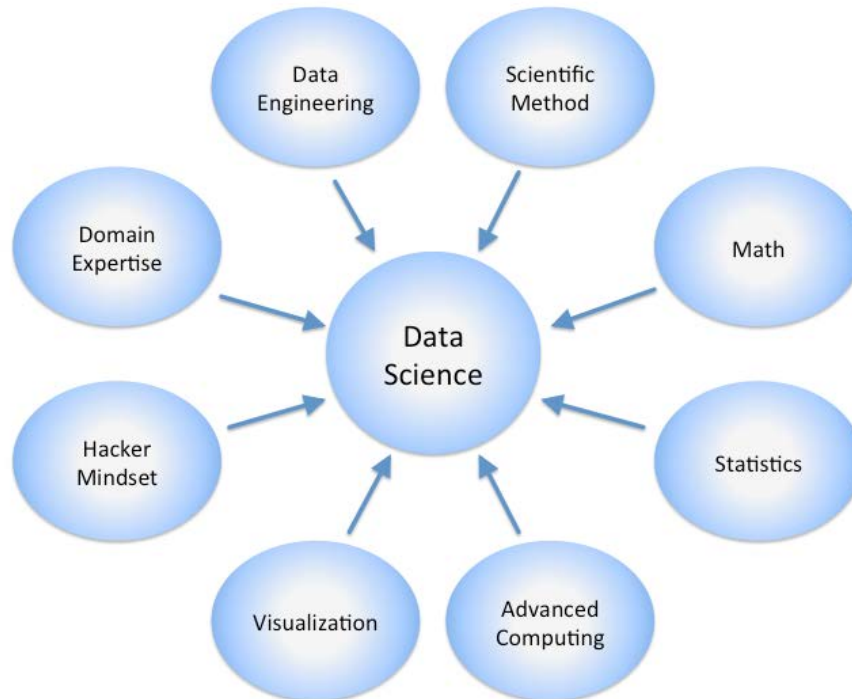
- ❑ ¿Qué es la Ciencia de Datos?
- ❑ Minería de Datos
- ❑ Proceso de Minería de Datos
- ❑ Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación y Otros
- ❑ El Poder de los Datos
- ❑ Herramientas y Lenguajes en Ciencia de Datos. Repositorio de Kaggle
- ❑ Comentarios Finales

- ❑ **¿Qué es la Ciencia de Datos?**
- ❑ Minería de Datos
- ❑ Proceso de Minería de Datos
- ❑ Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación y Otros
- ❑ El Poder de los Datos
- ❑ Herramientas y Lenguajes en Ciencia de Datos. Repositorio de Kaggle
- ❑ Comentarios Finales

Ciencia de Datos

Data Science

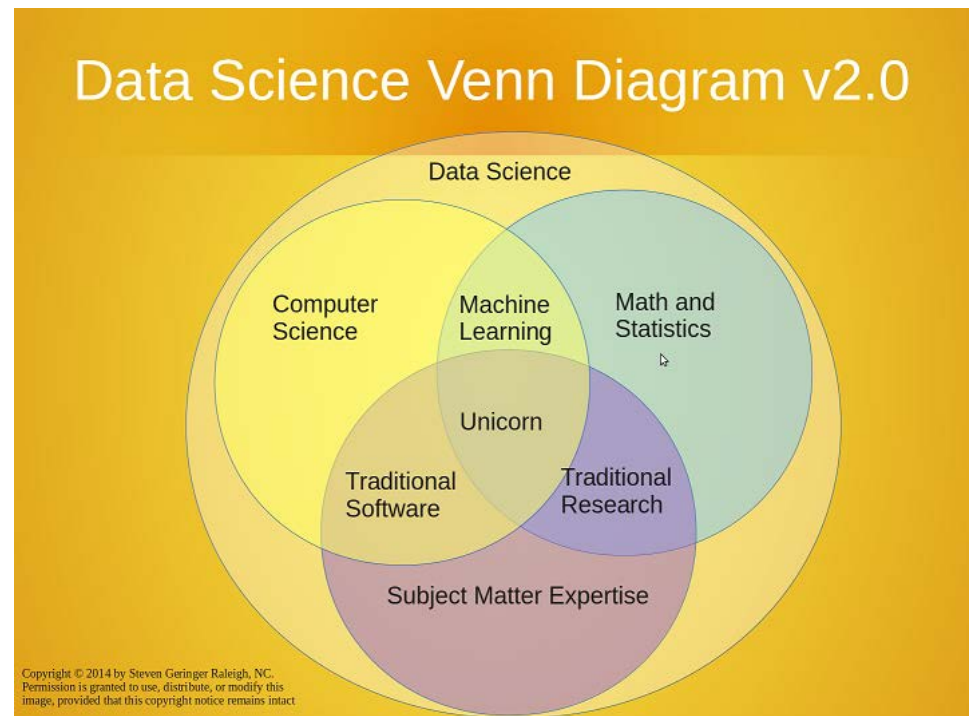
Ciencia de Datos es el ámbito de conocimiento que engloba las habilidades asociadas a la extracción de conocimiento de datos, incluyendo Big Data



Ciencia de Datos

¿Qué es un Científico de Datos?

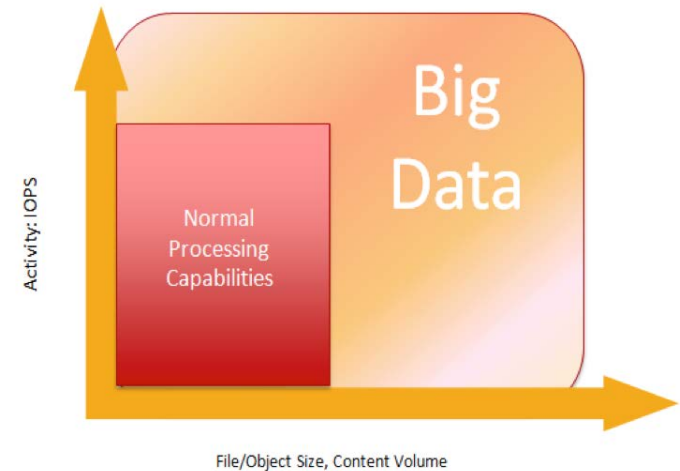
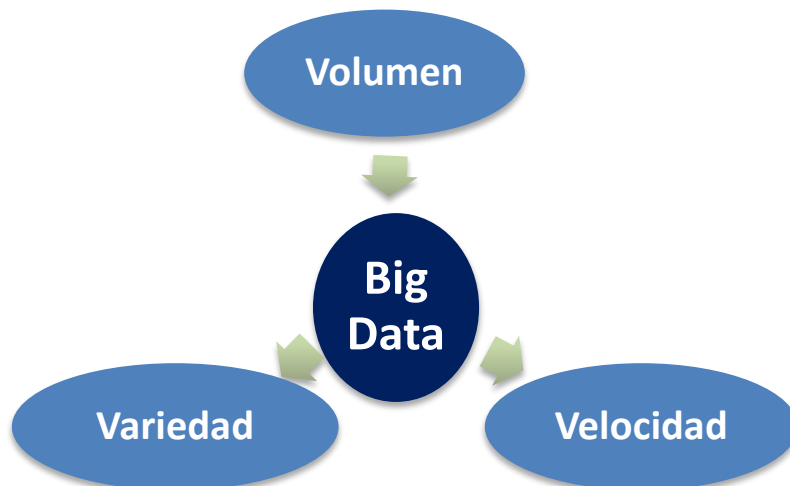
Un científico de datos es un profesional que debe dominar las ciencias matemáticas y la estadística, acabadados conocimientos de programación (y sus múltiples lenguajes), ciencias de la computación y analítica.



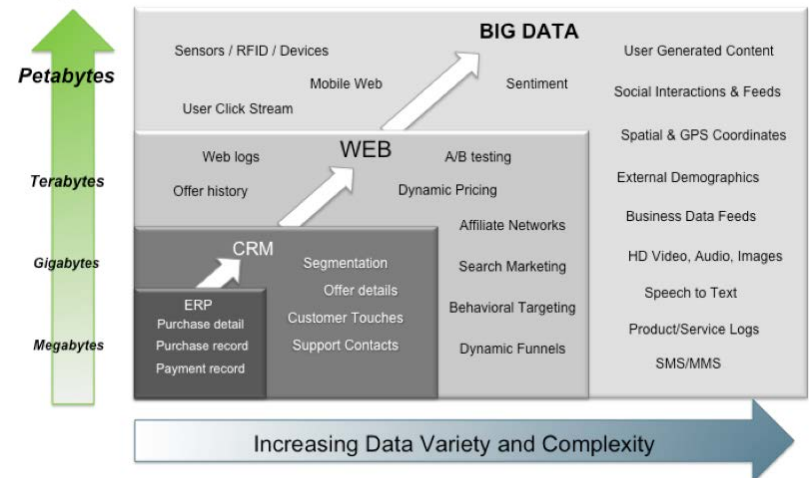
Ciencia de Datos

Big Data

“**Big Data**” son datos cuyo volumen, diversidad y complejidad **requieren nueva arquitectura, técnicas, algoritmos y análisis** para gestionar y extraer valor y conocimiento oculto en ellos ...



Big Data = Transactions + Interactions + Observations



Source: Contents of above graphic created in partnership with Teradata, Inc.

[illegible]

- ❑ ¿Qué es la Ciencia de Datos?
- ❑ **Minería de Datos**
- ❑ Proceso de Minería de Datos
- ❑ Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación y Otros
- ❑ El Poder de los Datos
- ❑ Herramientas y Lenguajes en Ciencia de Datos. Repositorio de Kaggle
- ❑ Comentarios Finales

¿Qué es la Minería de Datos?



La Minería de datos (MD) es el proceso de extracción de patrones de información (implícitos, no triviales, desconocidos y potencialmente útiles) a partir de grandes cantidades de datos



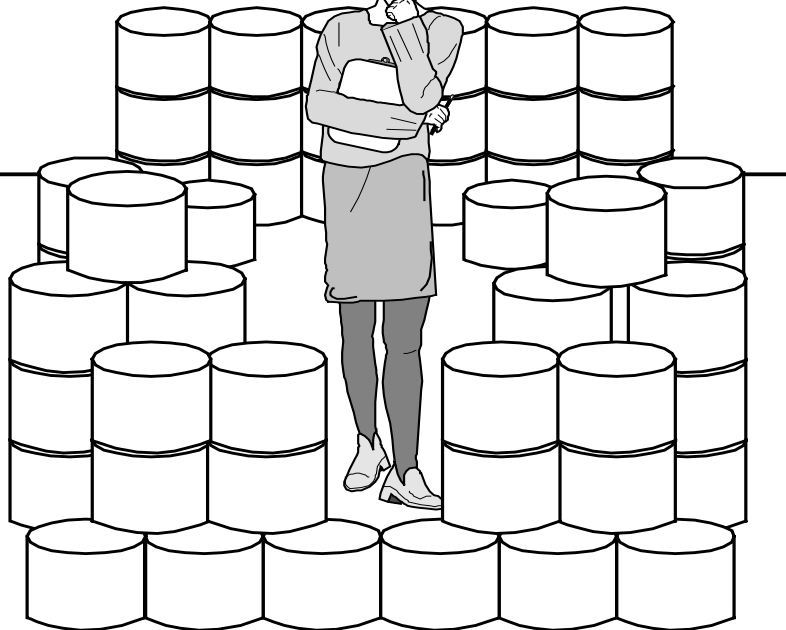
También se conoce como:

- Descubrimiento de conocimiento en bases de datos (KDD),
- extracción del conocimiento,
- análisis inteligente de datos / patrones,
- ...

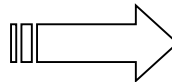
¿Qué es la Minería de Datos?



How can I analyze this data?



**We have rich data,
but poor information**

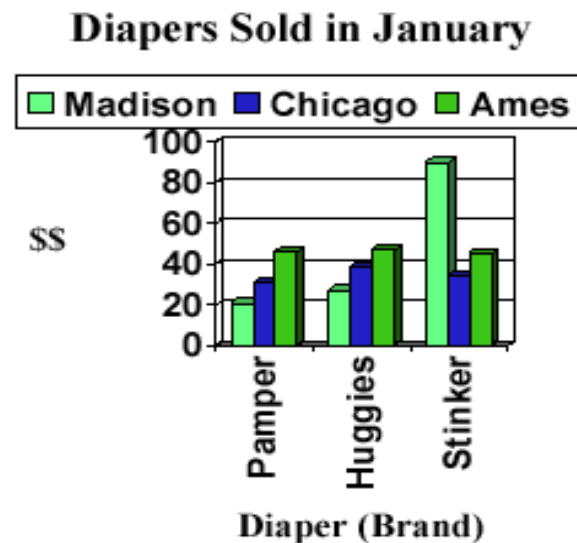
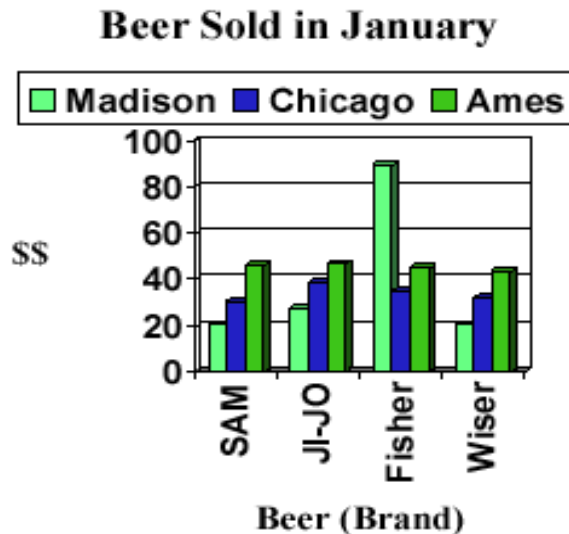


**Data mining-searching for knowledge
(interesting patterns) in your data.**

Minería de Datos. Caso de estudio

Marketing y ventas (asociaciones)

- Si se realiza sólo toma de decisión en función de los informes (datos), por ejemplo para dos productos, cerveza y pañales



¿Qué información aporta?

Minería de Datos. Caso de estudio

Marketing y ventas (asociaciones)

- Objetivo: determinar grupos de items que tienden a ocurrir juntos en transacciones (=tickets de compra pagados con o sin tarjeta)
- Se utilizan técnicas de asociación, que pueden descubrir información como:
 - Los clientes que compran cerveza también compran patatas **¡Para eso no es necesario el uso de técnicas de DM!**
 - Los viernes por la tarde, con frecuencia, quienes compran pañales, compran también cerveza.

- ✓ ¿Qué significa?
- ✓ ¿A qué se debe?
- ✓ Acciones a realizar



Minería de Datos. Caso de estudio

Marketing y ventas (asociaciones)

Explicación más probable

- Se acerca el fin de semana
- Hay un bebé en casa
- No quedan pañales
- El padre/madre compra pañales al salir del trabajo
- ¡No pueden salir!
- Comprar cervezas para el fin de semana (y un partido/película PPV)

- Se acerca el fin de semana
- Hay un bebé en casa luego nada de ir fuera
- Hay que comprar pañales
- Quedarse en casa → ver partido / película
- Comprar cervezas para el partido / película

Pañales → Cerveza



Minería de Datos. Caso de estudio

Marketing y ventas (asociaciones)

Acciones a realizar:

- Planificar disposiciones alternativas en el almacén
- Limitar descuentos especiales a sólo uno de los dos productos que tienden a comprarse juntos
- Poner los aperitivos que más margen dejan entre los pañales y las cervezas
- Poner productos de bebé en oferta cerca de las cervezas
- Ofrecer cupones descuento para el producto “complementario”, cuando uno de los productos se venda por separado...



La profleración de “tarjetas de lealtad” se debe al interés por identificar el historial de ventas individual del cliente...

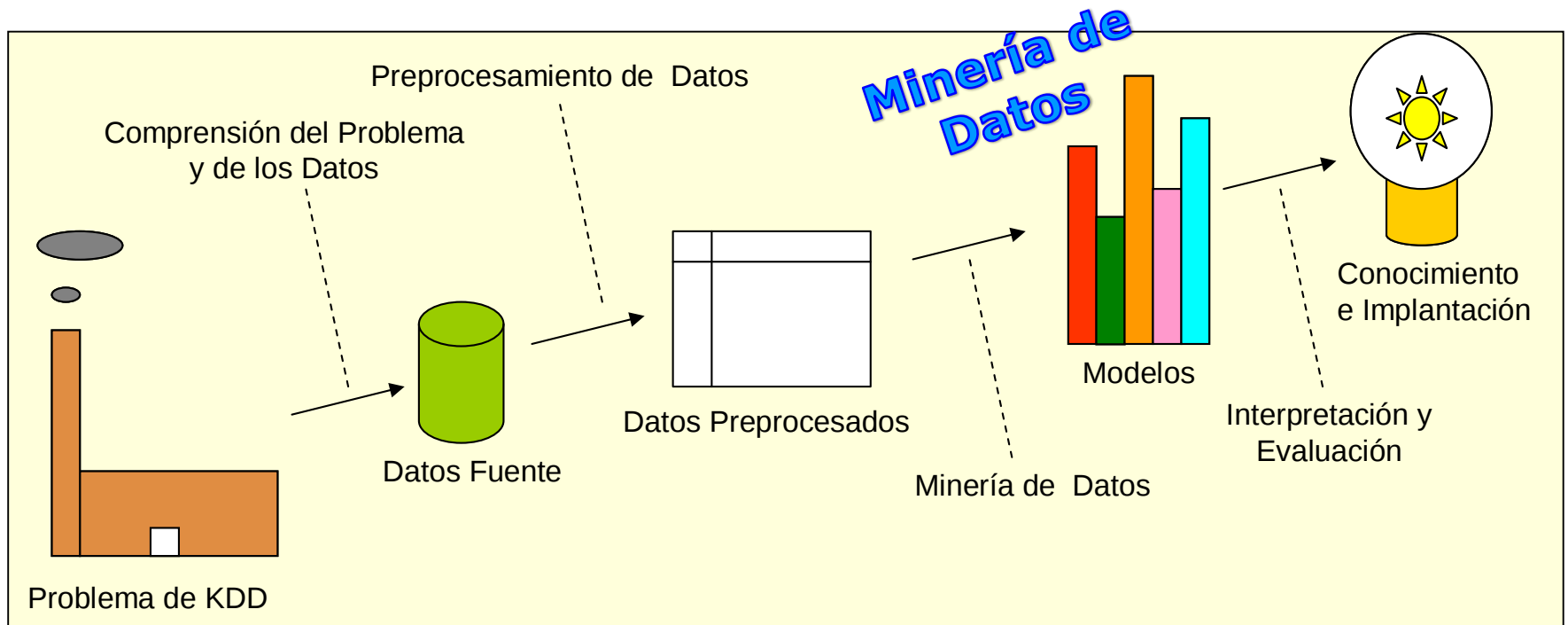
- ❑ ¿Qué es la Ciencia de Datos?
- ❑ Minería de Datos
- ❑ **Proceso de Minería de Datos**
- ❑ Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación y Otros
- ❑ El Poder de los Datos
- ❑ Herramientas y Lenguajes en Ciencia de Datos. Repositorio de Kaggle
- ❑ Comentarios Finales

Etapas en el proceso de KDD

- KDD = *Knowledge Discovery from Databases*
- El KDD es el proceso completo de extracción de conocimiento a partir de bases de datos
- El término se acuñó en 1989 para enfatizar que el conocimiento es el producto final de un proceso de descubrimiento guiado por los datos
- La Minería de Datos es sólo una etapa en el proceso de KDD
- Informalmente se asocia Minería de Datos con KDD

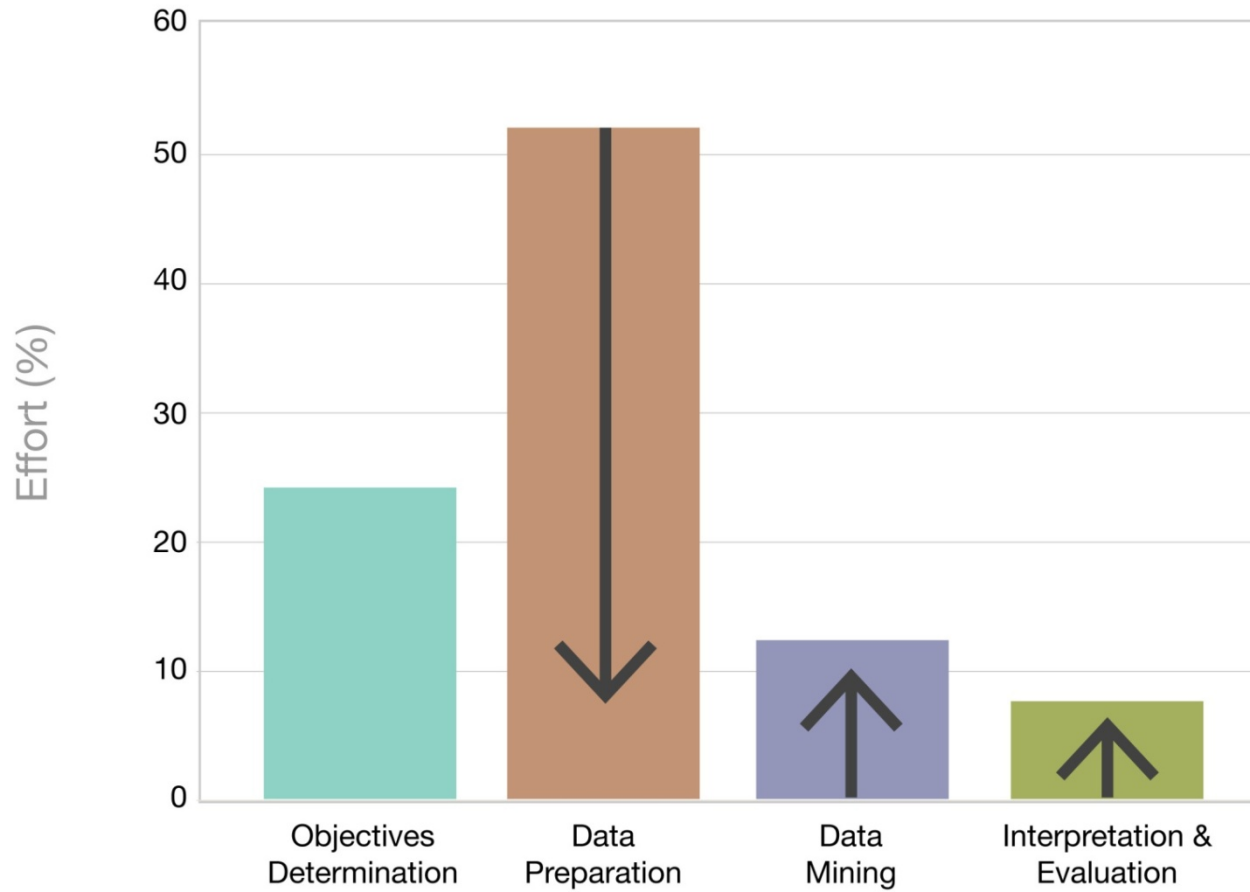
Etapas en el proceso de KDD

Etapas en un proceso de KDD



Informalmente se asocia Minería de Datos con KDD

Etapas en el proceso de KDD



Tiempos estimados en el análisis de un problema mediante técnicas de minería de datos

- ❑ ¿Qué es la Ciencia de Datos?
- ❑ Minería de Datos
- ❑ Proceso de Minería de Datos
- ❑ **Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación y Otros**
- ❑ El Poder de los Datos
- ❑ Herramientas y Lenguajes en Ciencia de Datos.
Repositorio de Kaggle
- ❑ Comentarios Finales

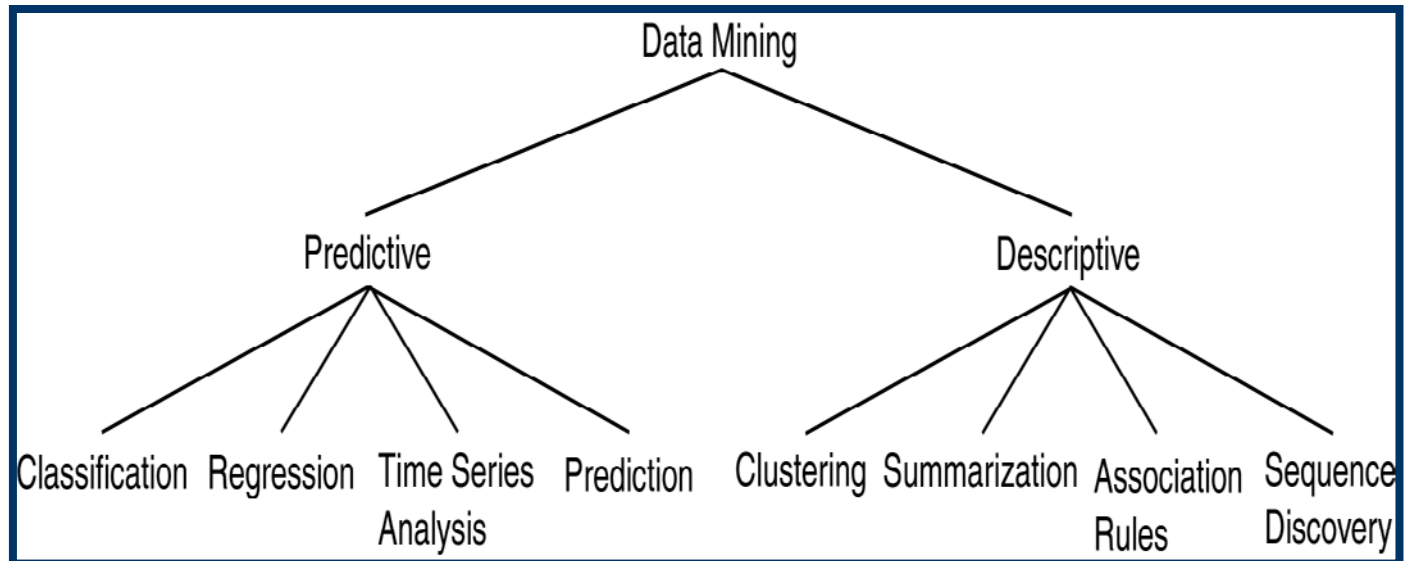
Técnicas de Minería de Datos

■ Métodos predictivos

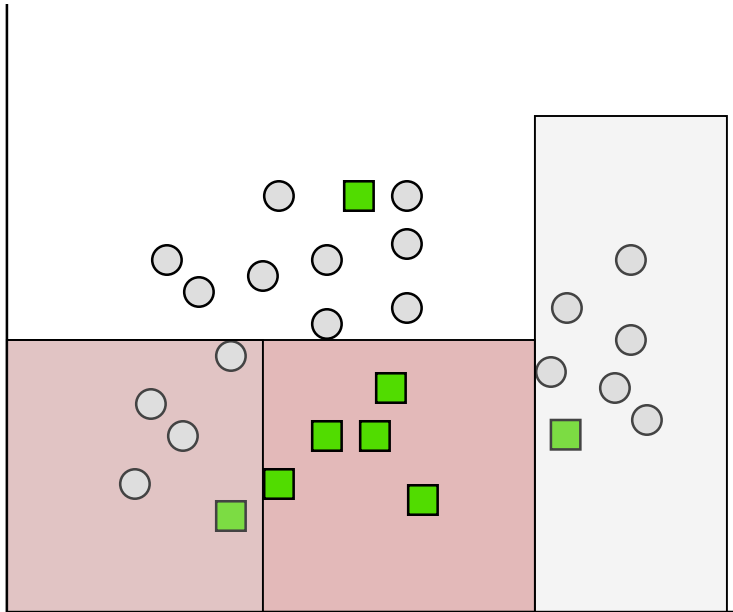
- Se utilizan algunas variables para predecir valores desconocidos de otras variables

■ Métodos descriptivos

- Encuentran patrones interpretables que describen los datos

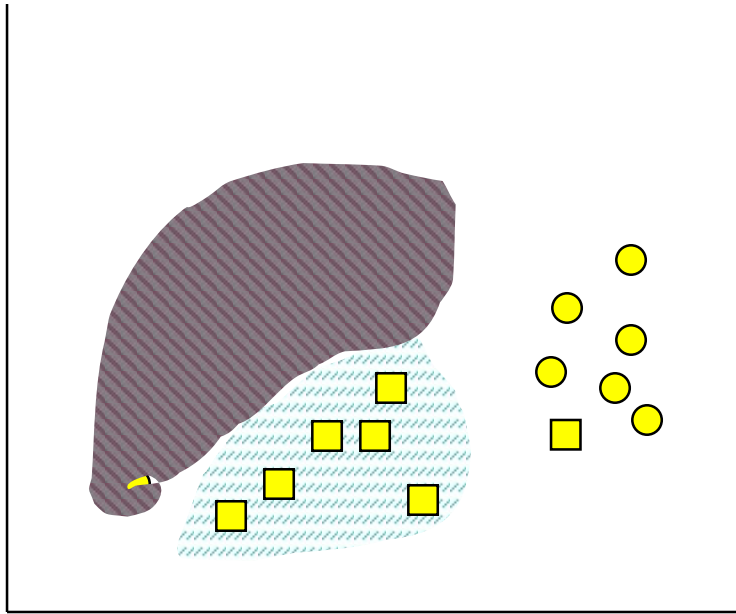


Aprendizaje Supervisado vs No Supervisado



Aprendizaje supervisado:
Aprende, a partir de un conjunto de instancias pre-etiquetadas un metodo para predecir (Ejemplo, clasificación: la clase a que pertenece una nueva instancia)

Aprendizaje Supervisado vs No Supervisado



Aprendizaje no supervisado:

No hay conocimiento a priori sobre el problema, no hay instancias etiquetadas, no hay supervisión sobre el procedimiento.

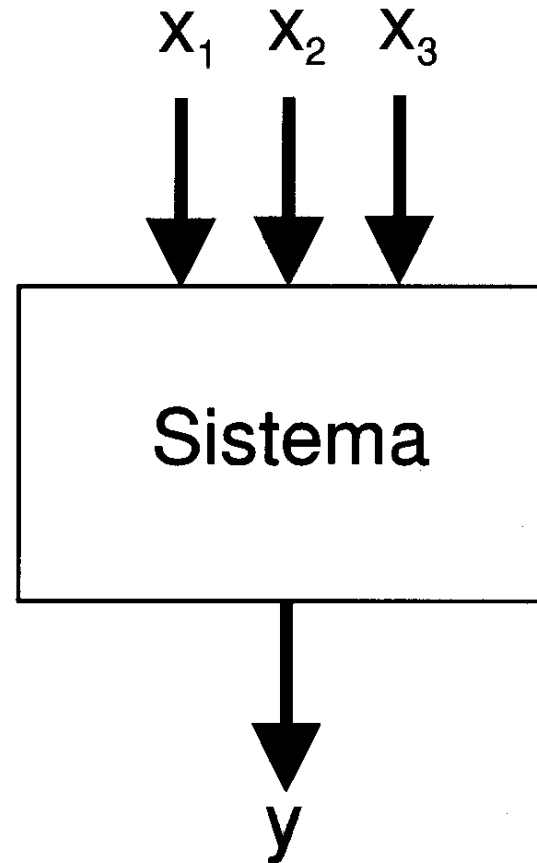
(Ejemplo, clustering: Encuentra un agrupamiento de instancias “natural” dado un conjunto de instancias no etiquetadas)

Técnicas de Minería de Datos

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation/Anomaly Detection [Predictive]
- Time Series [Predictive]
- Summarization [Descriptive]

Regresión

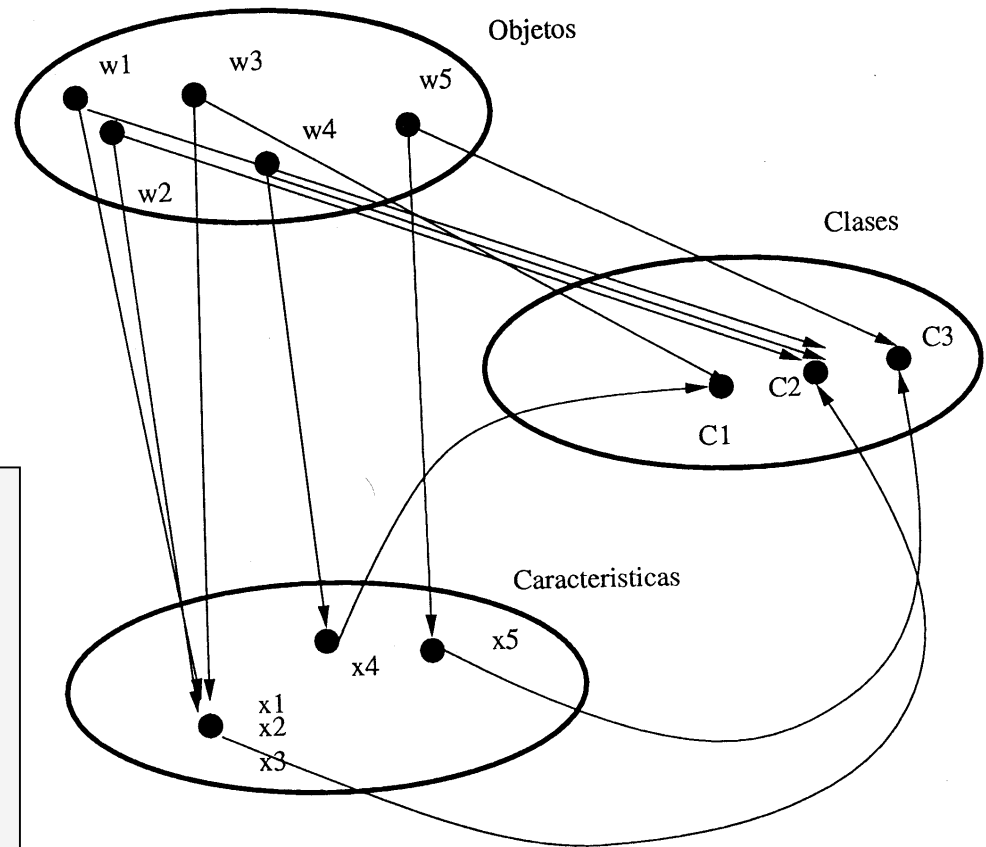
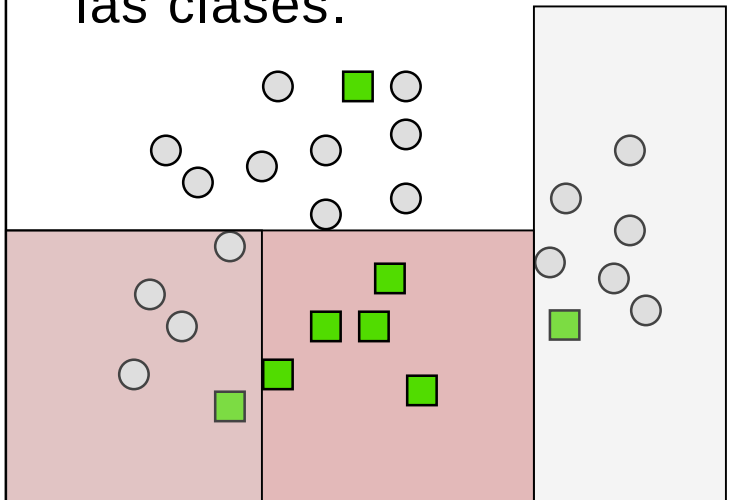
El problema fundamental de la predicción está en modelar la relación entre las variables de estado para obtener el valor de la variable de control.



Clasificación

■ Clasificación

El problema fundamental de la clasificación está directamente relacionado con la separabilidad de las clases.



Clasificación. Ejemplo

- **Ejemplo:** Diseño de un Clasificador para *Iris*
 - Problema simple muy conocido: *clasificación de lirios*.
 - Tres clases de lirios: *setosa*, *versicolor* y *virginica*.
 - Cuatro atributos: *longitud* y *anchura* de *pétalo* y *sépalo*, respectivamente.
 - 150 ejemplos, 50 de cada clase.
 - Disponible en <http://www.ics.uci.edu/~mlearn/MLRepository.html>



setosa



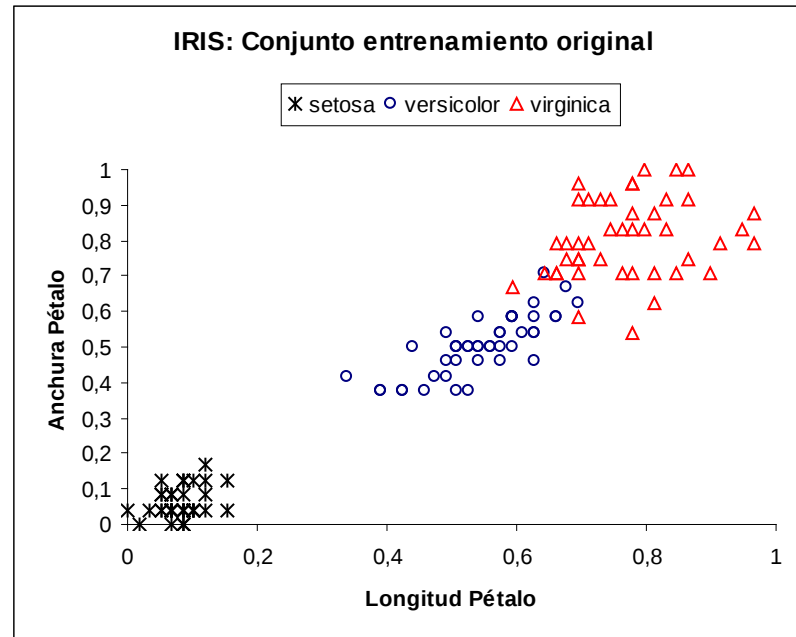
versicolor



virginica

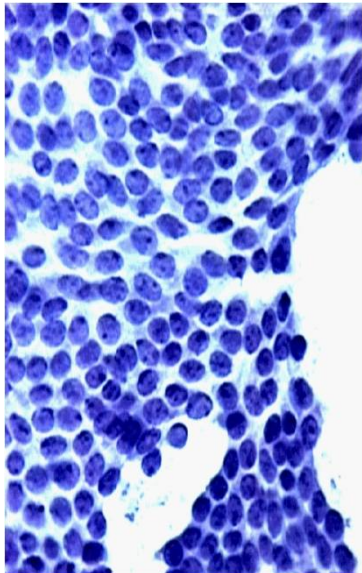
Clasificación. Ejemplo

Ejemplos de conjuntos seleccionados sobre *Iris*:



Clasificación. Ejemplo

Wisconsin Breast Cancer: Predict malignant / benign



Attribute name	Description
RADIUS	<i>Mean of distances from center to points on the perimeter</i>
TEXTURE	<i>Standard deviation of grayscale values</i>
PERIMETER	<i>Perimeter of the mass</i>
AREA	<i>Area of the mass</i>
SMOOTHNESS	<i>Local variation in radius lengths</i>
COMPACTNESS	<i>Computed as: $\text{perimeter}^2 / \text{area} - 1.0$</i>
CONCAVITY	<i>Severity of concave portions of the contour</i>
CONCAVE POINTS	<i>Number of concave portions of the contour</i>
SYMMETRY	<i>A measure of the nuclei's symmetry</i>
FRACTAL DIMENSION	<i>'Coastline approximation' - 1.0</i>
DIAGNOSIS (Target)	<i>Diagnosis of cell sample: malignant or benign</i>

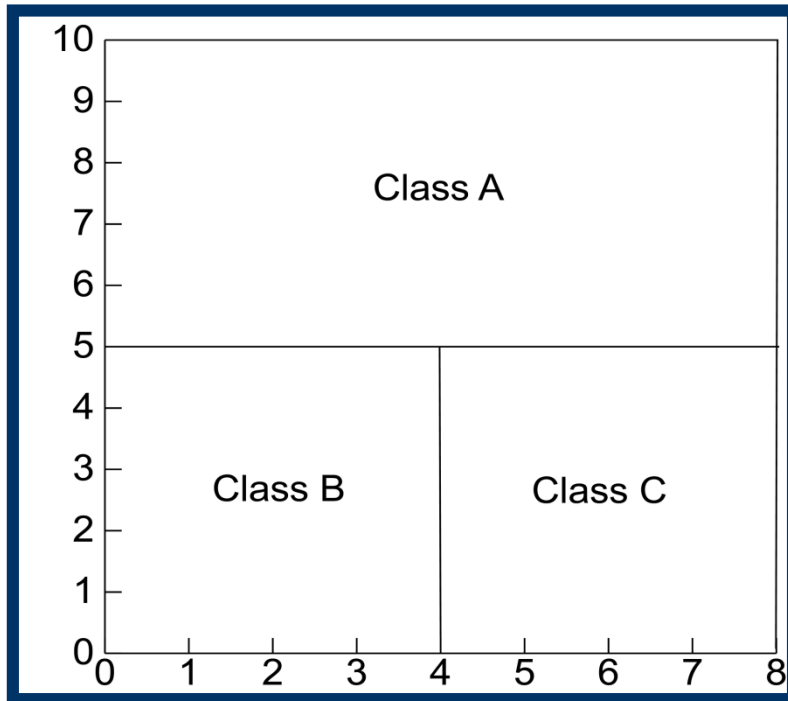
Clasificación. Ejemplo

Handwritting recognition.
Assign a digit from 0 to 9.



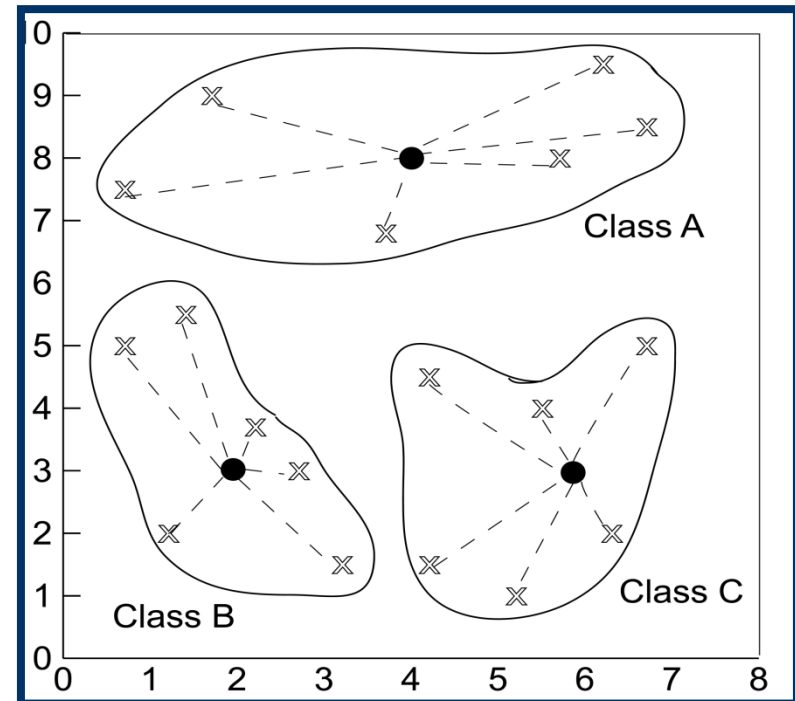
Clasificación. Ejemplo

Clases Definidas



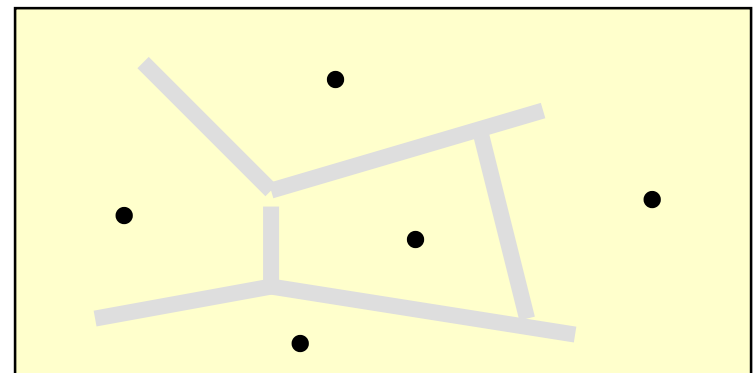
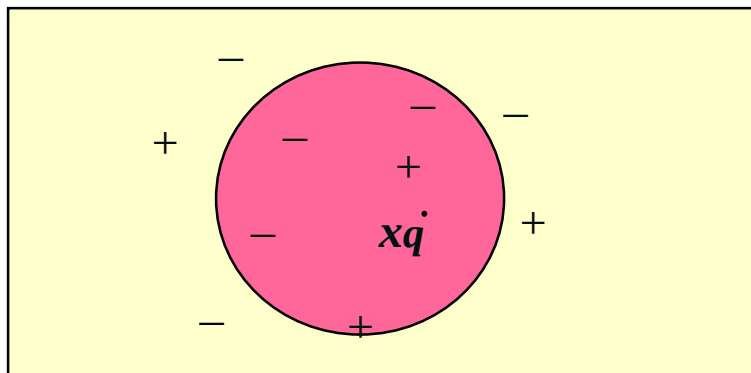
Basado en Particiones

Basado en Distancias

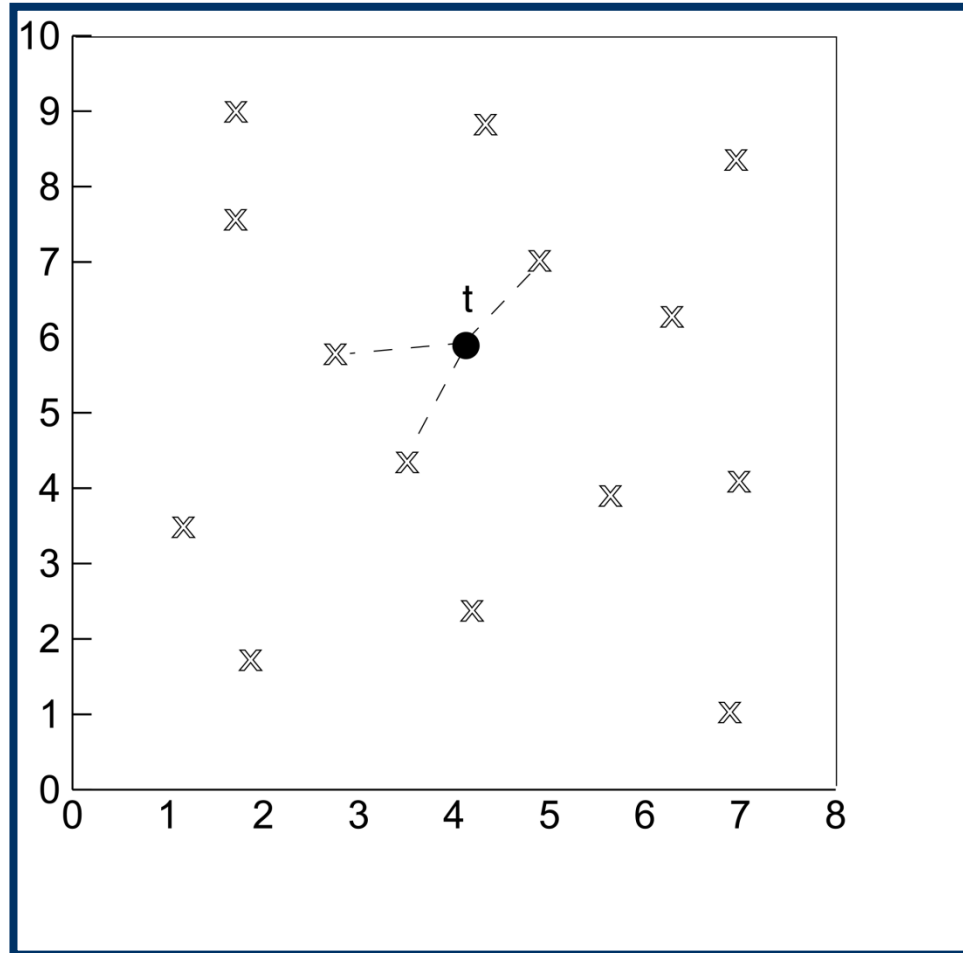


Ejemplo de Clasificador: k-NN

- k -NN devuelve la clase más repetida de entre todos los k ejemplos de entrenamiento cercanos a xq .
- Diagrama de Voronoi: superficie de decisión inducida por 1-NN para un conjunto dado de ejemplos de entrenamiento.



Ejemplo de Clasificador: k-NN



$k = 3$

Clasificación

Se pueden construir distintos tipos de clasificadores:

Modelos Interpretables:

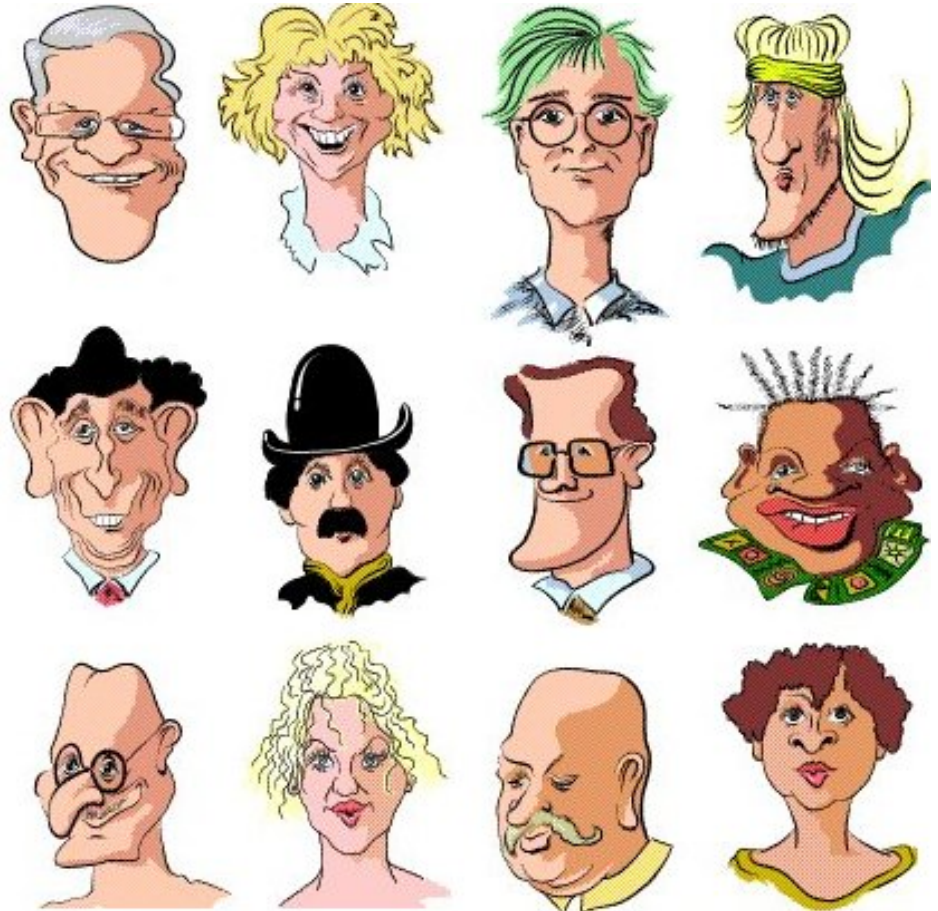
- Árboles de decisión
- Reglas (p.ej. listas de decisión)

Modelos no interpretables:

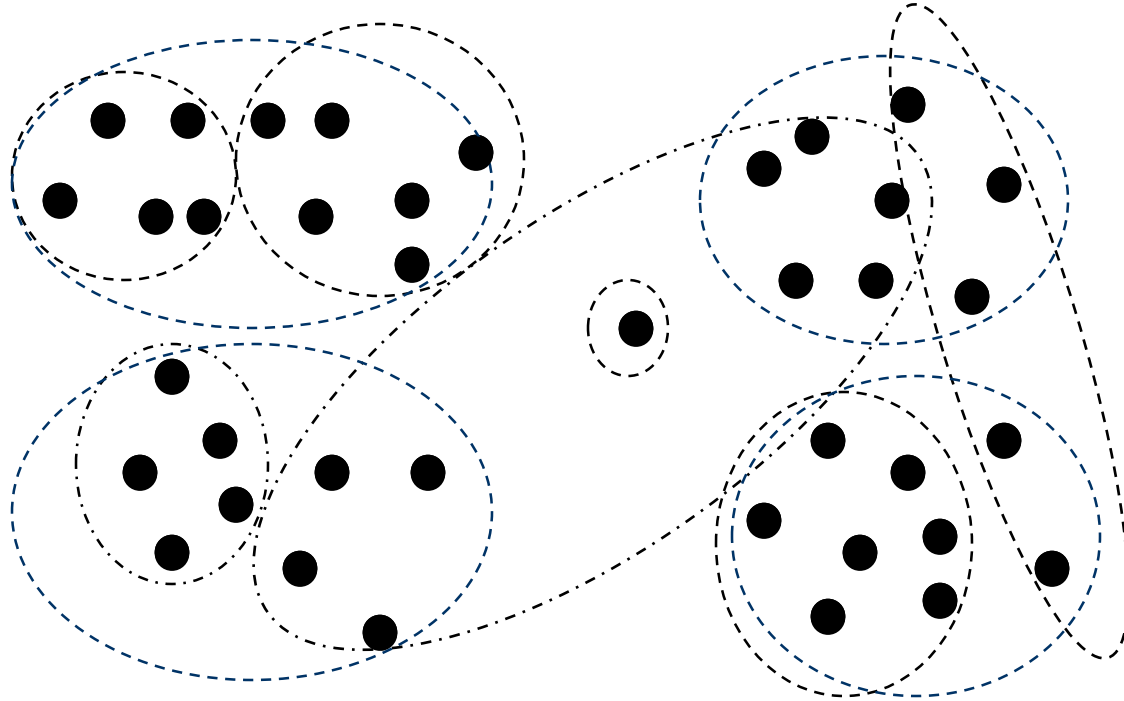
- Clasificadores basados en casos (k-NN)
- Redes neuronales
- Redes bayesianas
- SVMs (Support Vector Machines)
- ...

Agrupamiento

Hay problemas en
los que deseamos
agrupar las
instancias
creando clusters
de similares
características
Ej. Segmentación
de clientes de una
empresa



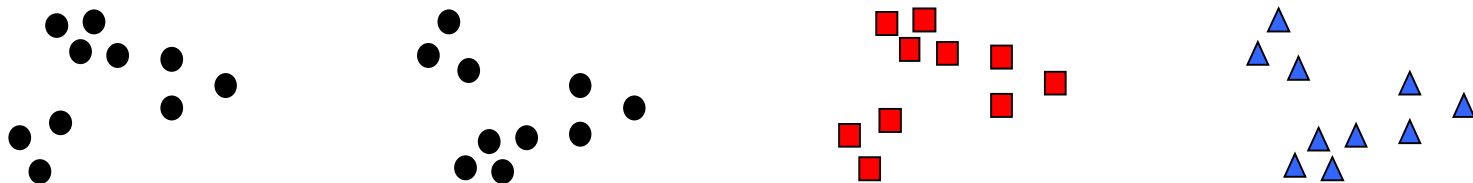
Agrupamiento. Niveles



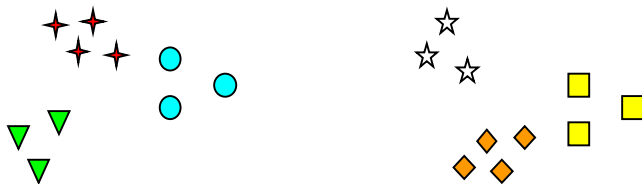
**La decisión del número de clusters
es uno de los retos en agrupamiento**

Agrupamiento. Niveles

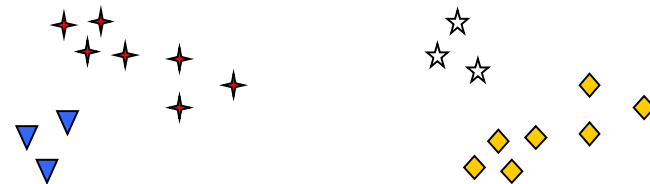
$k = 2$



$k = 6$



$k = 4$

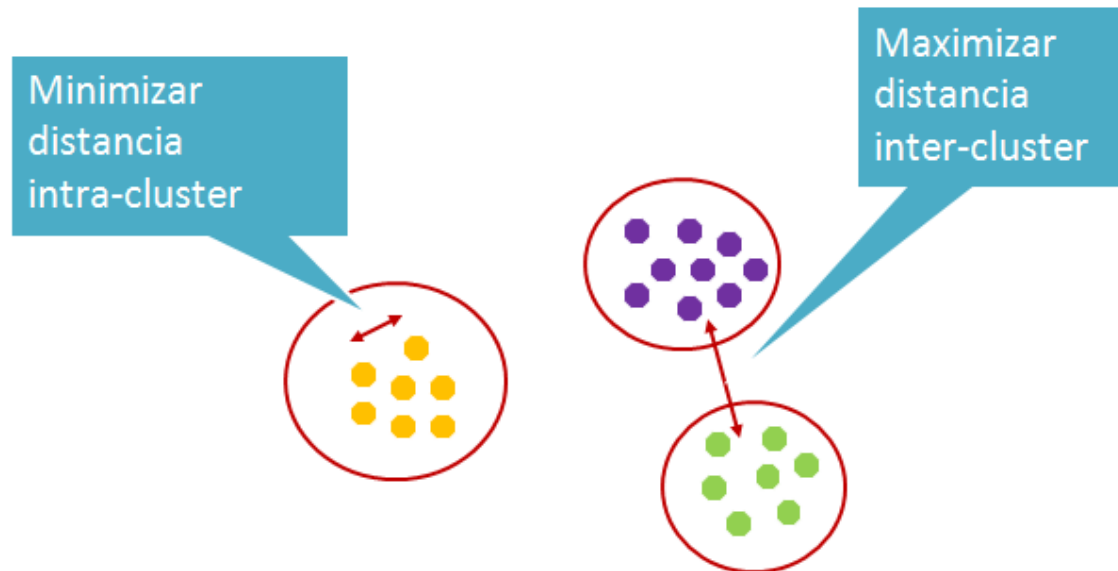


La decisión del número de clusters
es uno de los retos en agrupamiento

Agrupamiento. Modelos

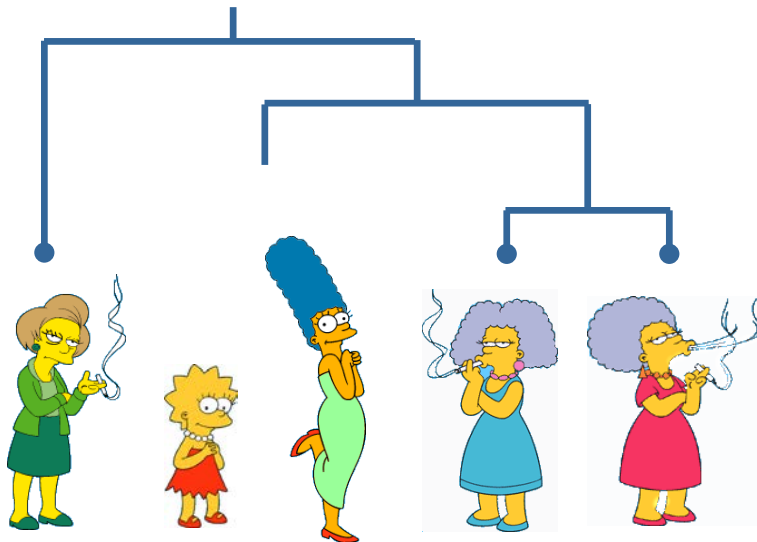
Objetivo

Encontrar agrupamientos de tal forma que los objetos de un grupo sean similares entre sí y diferentes de los objetos de otros grupos [*clusters*].

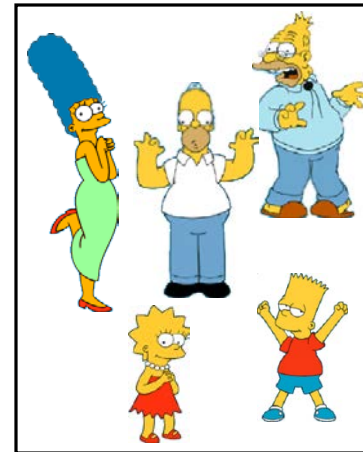


Agrupamiento. Modelos

Modelos Jerárquicos



Modelos Particionales



Ejemplos de Agrupamiento

- **Marketing:** descubrimiento de distintos grupos de clientes en la BD. Usar este conocimiento en la política publicitaria, ofertas, ...
- **Uso de la tierra:** Identificación de áreas de uso similar a partir de BD con observaciones de la tierra (cultivos, ...)
- **Seguros:** Identificar grupos de asegurados con características parecidas (siniestros, posesiones,). Ofertarles productos que otros clientes de ese grupo ya poseen y ellos no
- **Planificación urbana:** Identificar grupos de viviendas de acuerdo a su tipo, valor o situación geográfica
- **WWW:** Clasificación de documentos, analizar ficheros .log para descubrir patrones de acceso similares, ...

Descubrimiento de Asociaciones

- Descubrimiento de reglas de asociación:
 - Búsqueda de patrones frecuentes, asociaciones, correlaciones, o estructuras causales entre conjuntos de artículos u objetos (datos) a partir de bases de datos transaccionales, relacionales y otros conjuntos de datos
 - Búsqueda de secuencias o patrones temporales
 - Aplicaciones:
 - análisis de cestas de la compra (*Market Basket analysis*)
 - diseño de catálogos,...
 - ¿Qué hay en la cesta? Libros de Jazz
 - ¿Qué podría haber en la cesta? El último CD de Jazz
 - ¿Cómo motivar al cliente a comprar los artículos que es probable que le gusten?

Descubrimiento de asociaciones

Market Basket Analysis

Compra: zumo de naranja, plátanos, detergente para vajillas, limpia cristales, gaseosa, ...

¿Cómo afecta la demografía de la vecindad a la compra de los clientes?

¿Es típico comprar gaseosa y plátanos?
¿Es importante la marca de la gaseosa?

¿Aumenta la compra del limpia cristales cuando se compran a la vez detergente para vajillas y zumo de naranja?

¿Dónde deberían colocarse los detergentes para maximizar sus ventas?



A word cloud visualization showing the relationship between 'data' and 'science'. The words are arranged in a circular pattern around a central point. 'data' is the largest word at the top left, and 'science' is the largest word at the bottom right. Other prominent words include 'research', 'university', 'students', 'faculty', 'projects', 'career', 'space', 'education', 'bridge', 'evolution', 'graduate', 'activity', 'follows', 'institute', 'program', 'work', 'tended', 'new', 'growth', 'complexity', 'social', 'economics', 'methodology', 'environment', 'reproducibility', 'tools', 'teaching', 'learning', 'innovation', 'incubator', 'discovery', 'collaboration', 'interdisciplinary', 'multidisciplinary', 'transdisciplinary', 'convergence', 'integration', 'synergy', 'partnership', 'alliance', 'consortium', 'network', 'community', 'ecosystem', 'landscape', 'arena', 'stage', 'platform', 'arena', 'stage', 'platform', 'arena', 'stage', 'platform'.

- ❑ ¿Qué es la Ciencia de Datos?
- ❑ Minería de Datos
- ❑ Proceso de Minería de Datos
- ❑ Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación y Otros
- ❑ **El poder de los datos**
- ❑ Herramientas y Lenguajes en Ciencia de Datos. Repositorio de Kaggle
- ❑ Comentarios Finales

El poder de los datos

Análisis de asociaciones

Pañales y cerveza. Ficción y leyenda para ilustrar el análisis de transacciones



Si compro cerveza, entonces compro pañales

60%

Si compro pañales, entonces compro cerveza

100% ✓

El poder de los datos

Análisis de asociaciones en transacciones de tarjetas de crédito



El poder de los datos

Análisis de asociaciones: Salud



El poder de los datos

Análisis de transacciones



El poder de los datos

Análisis de transacciones: Un chivo expiatorio



Target (cadena de grandes almacenes) que utiliza el análisis de transacciones y asociaciones.



Unos días después el director llamó al padre para disculparse.

Respuesta conciliadora del padre:

“He estado hablando con mi hija –dijo el padre– Resulta que en mi casa han tenido lugar ciertas actividades de las que yo no estaba del todo informado. Mi hija sale de cuentas en agosto. Soy yo el que les debe una disculpa”.



El poder de los datos

Análisis de transacciones

Amazon: Sistema de recomendación



Los datos incrementaron tremendamente las ventas
Ahora más de 1/3 de las ventas son gracias a las recomendaciones



Críticos y editores literarios
La voz de Amazon (1995)



Dilema: ¿Lo que los clics decían o lo que opinaban los críticos?



Greg Linde (1997) propuso un sistema de recomendaciones, filtrado colaborativo
“artículo a artículo”



Banca: Identificación de personas con las compras de tarjetas de crédito

http://elpais.com/elpais/2015/01/29/ciencia/1422520042_066660.h

PRIVACIDAD EN INTERNET »

Cuatro compras con la tarjeta bastan para identificar a cualquier persona

- Los patrones de uso de las tarjetas permiten descubrir la identidad del 90% de una muestra de 1,1 millones de personas anónimas, según demuestra un estudio del MIT



Unique in the shopping mall: On the reidentifiability of credit card metadata

Yves-Alexandre de Montjoye,^{1*} Laura Radaelli,² Vivek Kumar Singh,^{1,3} Alex "Sandy" Pentland¹

Large-scale data sets of human behavior have the potential to fundamentally transform the way we fight diseases, design cities, or perform research. Metadata, however, contain sensitive information. Understanding the privacy of these data sets is key to their broad use and, ultimately, their impact. We study 3 months of credit card records for 1.1 million people and show that four spatiotemporal points are enough to uniquely reidentify 90% of individuals. We show that knowing the price of a transaction increases the risk of reidentification by 22%, on average. Finally, we show that even data sets that provide coarse information at any or all of the dimensions provide little anonymity and that women are more reidentifiable than men in credit card metadata.

<http://www.sciencemag.org/content/347/6221/536>

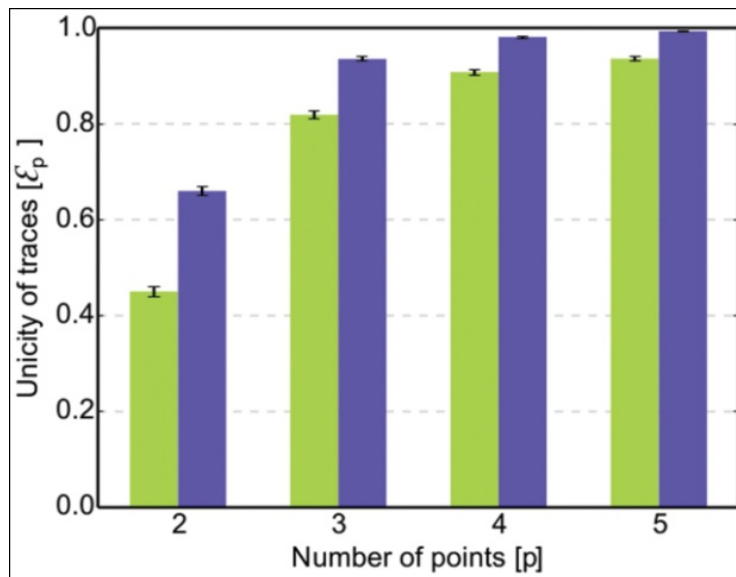


Banca: Identificación de personas con las compras de tarjetas de crédito

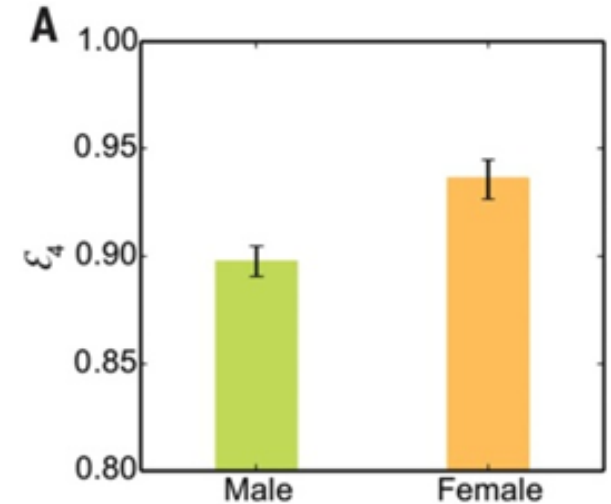
PRIVACIDAD EN INTERNET »

Cuatro compras con la tarjeta bastan para identificar a cualquier persona

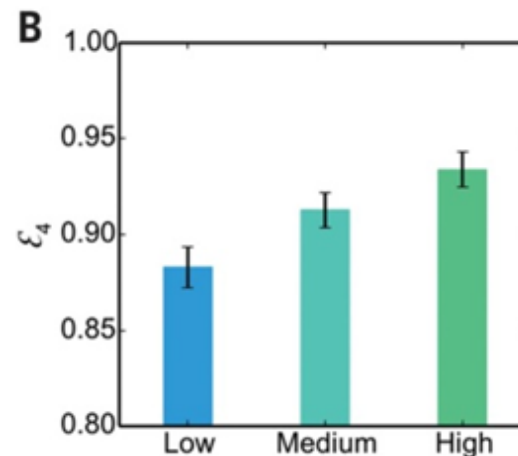
- Los patrones de uso de las tarjetas permiten descubrir la identidad del 90% de una muestra de 1,1 millones de personas anónimas, según demuestra un estudio del MIT



Identificación por el número de compras



Identificación por el género



Identificación por el poder adquisitivo

Analizando Twitter para medir la Salud Pública

You Are What You Tweet

Un sistema de filtrado de datos de Twitter puede inferir aspectos de salud analizando 144M de tuits (2011-2013)



Se obtienen 13 grupos coherentes de mensajes correlacionados

- Gripe estacional ($r = 0.689$) y alergias ($r = 0.810$)
- Ejercicio y obesidad relacionados con datos geográficos, ..

Discovering Health Topics in Social Media Using Topic Models

Michael J. Paul, Mark Dredze, Johns Hopkins University, Plos One 9(8)

e103408, 2014

doi:10.1371/journal.pone.0103408

TFM-UGR (Master Ciencia de Datos): Sentiment Analysis For Touristic Attractions: A Case Study On The Alhambra

Ana Valdivia(Sept. 2016)

On The Alhambra

Turismo – Web 2.0



"Amazing Experience!"

★★★★★ Reviewed yesterday via mobile

A must visit in Granada! An amazing journey through history makes you truly appreciate this gem. We really enjoyed our visit. Plenty of walking required.

Helpful?



1

Thank surf_golf

Report



Sentiment Analysis For Touristic Attractions: A Case Study On The Alhambra

TFM – Ana Valdivia(Sept. 2016)



TripAdvisorAlhambra
data set

8140 instances

10 features

id
username
location
userop
quote
review
quote+review
rating
date
page

SentimentValue



positive

neutral

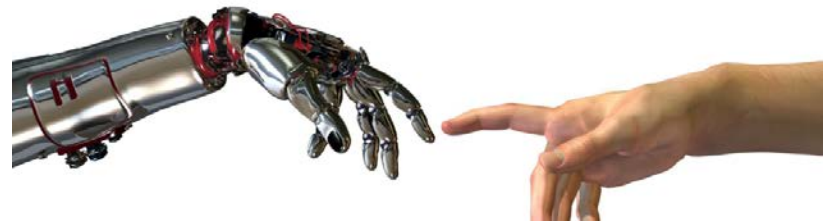
negative

2 class labels

Class labels

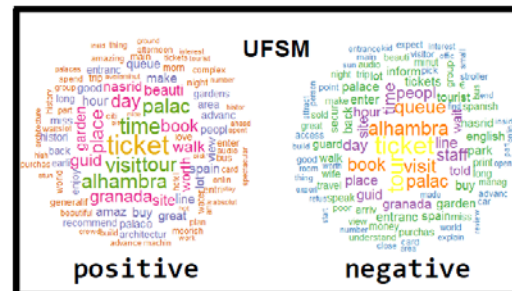
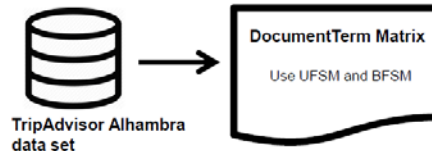
SentimentCoreNLP

SentimentValue



Sentiment Analysis For Touristic Attractions: A Case Study On The Alhambra

TFM – Ana Valdivia(Sept. 2016)

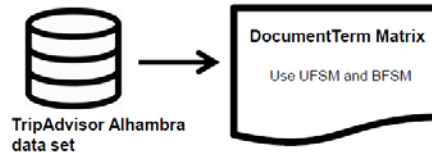


SentimentValue	SentimentCoreNLP			Total
	positive	neutral	negative	
positive	4,049	1,071	2,508	7,628
neutral	51	32	260	343
negative	5	6	158	169
Total	4,105	1,109	2,926	8,140

Table 4.2: Correlation between SentimentValue and SentimentCoreNLP

53.08 % of coincidence

TFM – Ana Valdivia(Sept. 2016)



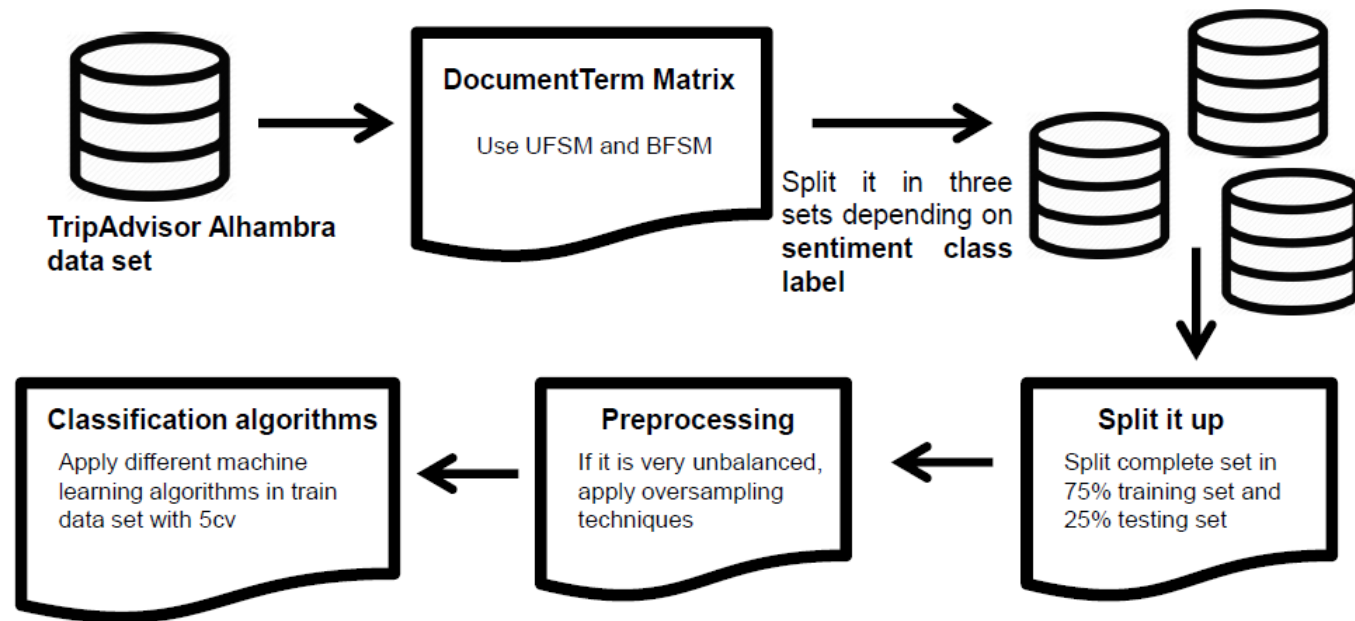
	SentimentCoreNLP			
SentimentValue	positive	neutral	negative	Total
positive	4,049	1,071	2,508	7,628
neutral	51	32	260	343
negative	5	6	158	169
Total	4,105	1,109	2,926	8,140

Table 4.2: Correlation between `SentimentValue` and `SentimentCoreNLP`

93.49 % of coincidence

Sentiment Analysis For Touristic Attractions: A Case Study On The Alhambra

TFM – Ana Valdivia(Sept. 2016)



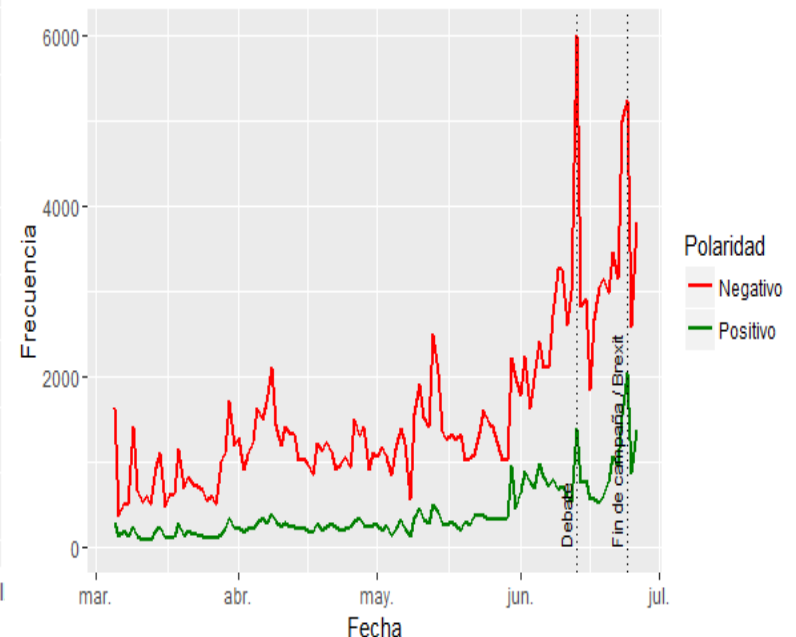
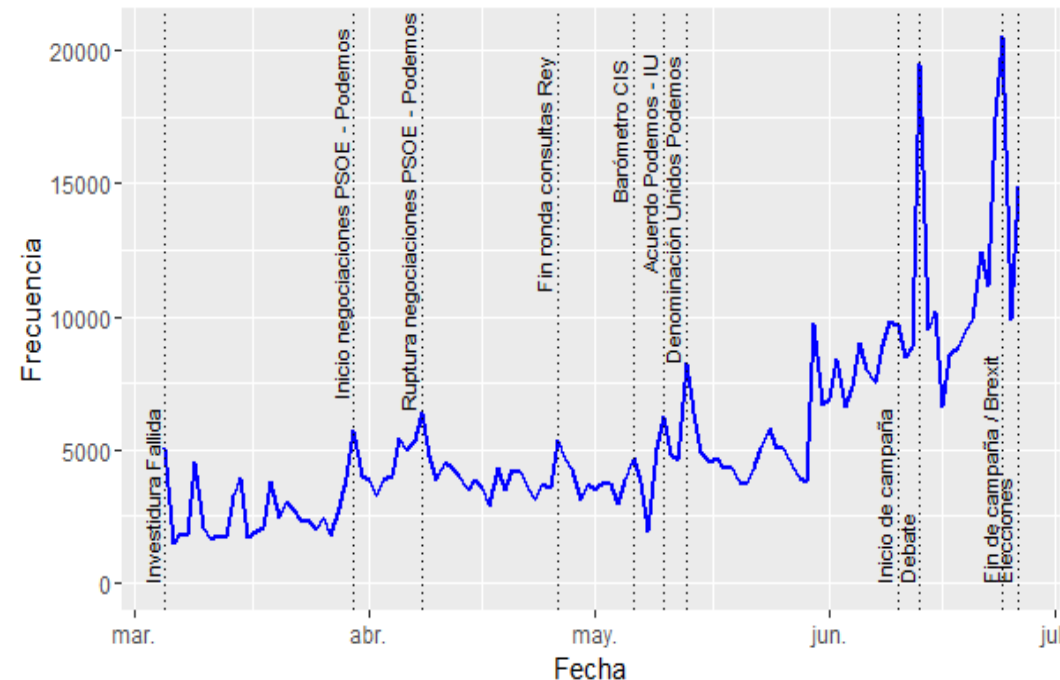
DISCOVERING INTERESTING PATTERNS IN NEGATIVE OPINIONS



TFM-UGR (Master Ciencia de Datos): Uso de técnicas de análisis de polaridad y fiabilidad de *tweets* en campaña electoral

Gabriel Quesada (Sept. 2016)

- Descarga de *tweets* a través de la *REST API* de *Twitter*.
- Los datos se obtienen a partir los *#hashtags* de los cuatro principales partidos y sus líderes.
- Período de recogida de datos: del 5 de marzo (tras la investidura fallida de Pedro Sánchez) hasta el 26 de junio (Elecciones Generales).
- Número total de *tweets* obtenidos: 1.593.624 (0,5 GB).

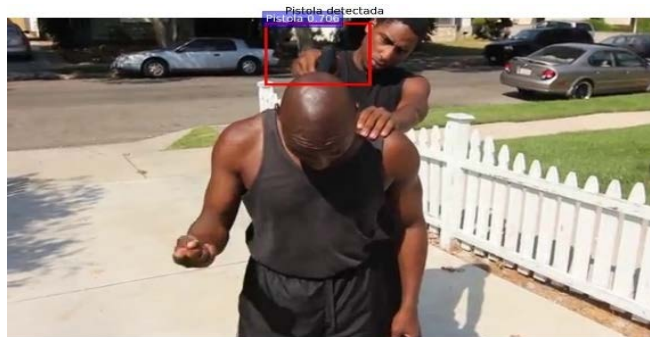
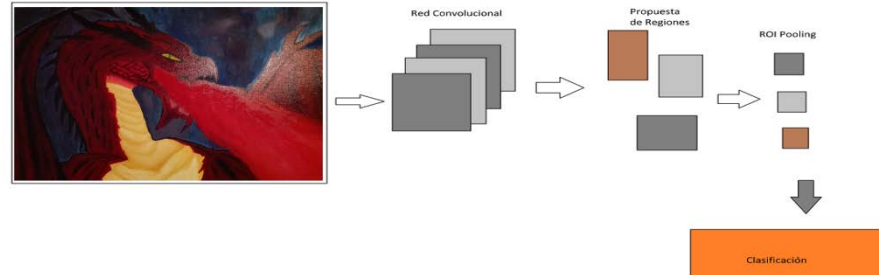


TFM-UGR (Master Ciencia de Datos): Identificación de Armas en Imágenes

Roberto Olmos (Sept. 2016)

DEEP LEARNING

Librería Caffe: Uso de
algoritmo Faster R-CNN



Impacto en la sociedad

Salud

Redes sociales como fuente de datos

Industria, comercio, banca, ...

Ocio y cultural (Ej. Recomendaciones)

Política, Bien social (Social good)





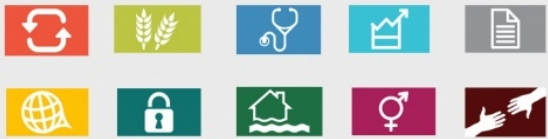
UNITED NATIONS GLOBAL PULSE

Harnessing big data for development and humanitarian action

GLOBAL PULSE PROJECT SERIES

The Global Pulse Project Series is a collection of case studies of 20 data innovation projects covering global issues ranging from public health to climate change, food security to employment.

[Read More /](#)



Global Pulse Project Series

Gran Impacto en la Sociedad y presencia en los medios de comunicación

EL MUNDO

Líder mundial en español | Miércoles 04/09/2013. Act

INTERNET | Campus Party Europa 2013

'Es la década d ahí vendrá la re



Alex 'Sandy' Pentland, durante su ponenci

<http://www.elmundo.es/elm>

EL PAÍS

ECONOMÍA

ECONOMÍA EMPRESAS MERCADOS BO

El maná de los datos

- La conversión de datos en información útil pe millones de dólares en 2015. La herramienta

SUSANA BLÁZQUEZ | Madrid | 29 SEP 2013 - 01

Archivado en: Citigroup Cap Gemini Sogeti IBM Telefónica Aplicaciones informáticas Tecn



EL PAÍS

ECONOMÍA

ECONOMÍA EMPRESAS MERCADOS BOLSA MIS AHORROS VIVIENDA TECNOLOGÍA OF

EMPRENDEDORES »

El Big Data echa una mano al campo

- Una empresa española recoge miles de datos para predecir las cosechas

MARÍA FERNÁNDEZ | 30 NOV 2014 - 00:00 CET

Archivado en: Bases datos Emprendedores Aplicaciones informáticas Empresas Programas informáticos Economía Informática Industria



Impacto Económico

La demanda de profesionales formados en Ciencia de Datos y *Big Data* es enorme.

Se estima que la conversión de datos en información útil generó un mercado de 132.000 millones de dólares en 2015 y que se crearán más de 4.4 millones de empleos.

España necesitaba para 2015 más de 60.000 profesionales con formación en Ciencia de Datos y *Big Data*.

http://www.revistacloudcomputing.com/2013/10/espana-necesitara-60-000-profesionales-de-big-data-hasta-2015/?goback=.gde_4377072_member_5811011886832984067#!

España necesitará 60.000 profesionales de Big Data hasta 2015

22 octubre, 2013 Eventos 18



España necesitará 60.0

Toledo.

EL PAÍS

PORTADA

INTERNACIONAL

POL

ECONOMÍA

ECONOMÍA EMPRESAS MERCADOS BOLSA FINANZAS PERSONALES VIVIENDA TECNOLOGÍA

▶ ESTÁ PASANDO

Multa a la banca

Revuelo en Hacienda

Eléctricas y renovables

Paro

El maná de los datos

- La conversión de datos en información útil para las empresas generará un mercado de 132.000 millones de dólares en 2015. La herramienta 'big data' sacará del mercado a quien no la use

SUSANA BLÁZQUEZ Madrid | 29 SEP 2013 - 01:00 CET

10

Archivado en: Citigroup Cap Gemini Sogeti SAP Oracle ING Bank BBVA Mapfre Bases datos IBM Telefónica Aplicaciones informáticas Tecnología Empresas Programas informáticos Economía



http://economia.elpais.com/economia/2013/09/27/actualidad/1380283725_938376.html

- ❑ ¿Qué es la Ciencia de Datos?
- ❑ Minería de Datos
- ❑ Proceso de Minería de Datos
- ❑ Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación y Otros
- ❑ El Poder de los Datos
- ❑ **Herramientas y Lenguajes en Ciencia de Datos. Repositorio de Kaggle**
- ❑ Comentarios Finales

Herramientas, Lenguajes, Kaggle

KNIME (o Konstanz Information Miner) es una plataforma de minería de datos que permite el desarrollo de modelos en un entorno visual. KNIME está desarrollado sobre la plataforma Eclipse y programado, esencialmente, en java.

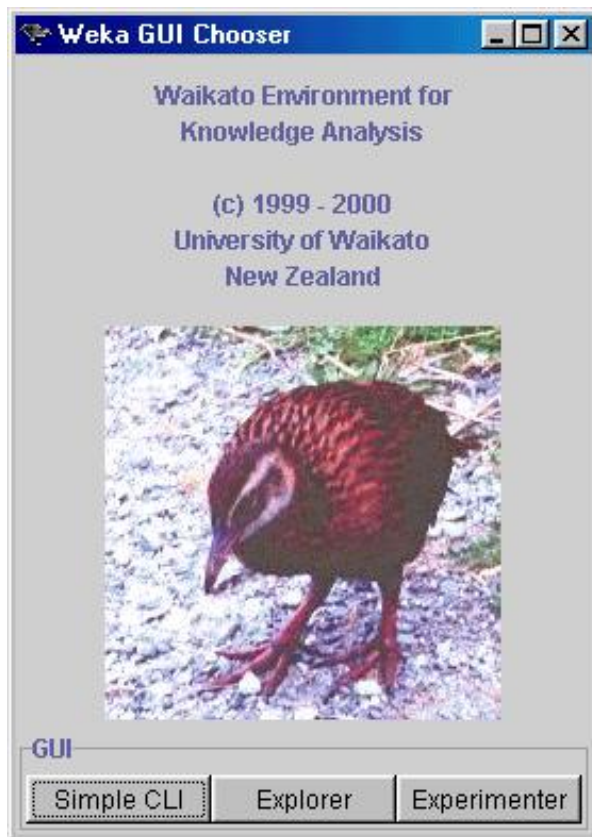
Fue desarrollado originalmente en el departamento de bioinformática y minería de datos de la Universidad de Constanza, Alemania, bajo la supervisión del profesor Michael Berthold. En la actualidad, la empresa KNIME.com GmbH, radicada en Zúrich, Suiza, continúa su desarrollo además de prestar servicios de formación y consultoría.



<https://www.knime.org/>

Herramientas, Lenguajes, Kaggle

Weka



- The University of Waikato, New Zealand
- Machine learning software in Java implementation

<http://www.cs.waikato.ac.nz/ml/weka/>

Herramientas, Lenguajes, Kaggle

KEEL



- University of Granada
- Machine learning software in Java implementation

<http://www.keel.es/>

Herramientas, Lenguajes, Kaggle

Sobre los lenguajes de programación (R, Phyton, ...). El
website CRAN

cran.r-project.org /

The Comprehensive R Archive Network



CRAN
[Mirrors](#)
[What's new?](#)

Contributed Packages

Available Packages

Currently, the CRAN package repository features 5799 available packages.

[Table of available packages, sorted by date of publication](#)

[Table of available packages, sorted by name](#)

<http://cran.r-project.org/web/views/MachineLearning.html>

Herramientas, Lenguajes, Kaggle

Sobre herramientas de minería de datos



<http://scikit-learn.org/stable/>

Classification

Identifying to which set of categories a new observation belong to.

Applications: Spam detection, Image recognition.

Algorithms: *SVM, nearest neighbors, random forest, ...* — Examples

Regression

Predicting a continuous value for a new example.

Applications: Drug response, Stock prices.

Algorithms: *SVR, ridge regression, Lasso, ...* — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: *k-Means, spectral clustering, mean-shift, ...* — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency

Algorithms: *PCA, feature selection, non-negative matrix factorization.* — Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning

Modules: *grid search, cross validation, metrics.* — Examples

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.

Modules: *preprocessing, feature extraction.* — Examples

Herramientas, Lenguajes, Kaggle



<http://www.kdnuggets.com/>

[Subscribe to KDnuggets News](#) | | [Contact](#)

[SOFTWARE](#) | [NEWS](#) | [Top stories](#) | [Opinions](#) | [Tutorials](#) | [JOBS](#) | [Companies](#) | [Courses](#) | [Datasets](#) | [EDUC](#)

**Machine Learning, Data Science,
Data Mining, Big Data, Analytics**

[Software](#) (Suites, Text, Visualization)
[Jobs - Industry](#) | [Academic](#)
[Meetings, Conferences](#)
[Companies](#) (Consulting, Products)
[Courses in Big Data, Data Science](#)
[Datasets](#) (APIs/Markets, Gov)
[Data Mining Course](#) | [Gregory Piatetsky](#)
[Education](#) (online, USA, Europe, cert)
[FAQ](#) | [Polls](#) | [Publications \(Books\)](#)
[Solutions](#) (Fraud, Data Cleaning)
[Webcasts](#) | [Websites](#) (Blogs, Cartoons,
Podcasts)

Most Recent

- [Benchmarking Big Data SQL Platforms in the Cloud](#)
- [Tensorflow Tutorial: Part 1 - Introduction](#)
- [A Solution to Missing Data: Imputation Using R](#)
- [Edelweiss: Data Scientist](#)
- [How To Lie With Numbers](#)
- [Toptal: Data Scientist](#)

[Past poll results: Python overtakes R, becomes the leader in Data Science, Machine Learning platforms](#)

Latest
[News](#) | [Software](#) | [Tutorials](#)

- [Benchmarking Big Data SQL Platforms in the Cloud](#)
- [Tensorflow Tutorial: Part 1 - Introduction](#)
- [A Solution to Missing Data: Imputation Using R](#)
- [Top tweets, Sep 13-19: Top Books on NLP; What Else Can ...](#)
- [How to Get Ready for AI and Machine Learning](#)

KDnuggets News

Herramientas, Lenguajes, Kaggle



<http://www.kdnuggets.com/2017/09/python-vs-r-data-science-machine-learning.html>

Data Science Jobs in USA on indeed.com

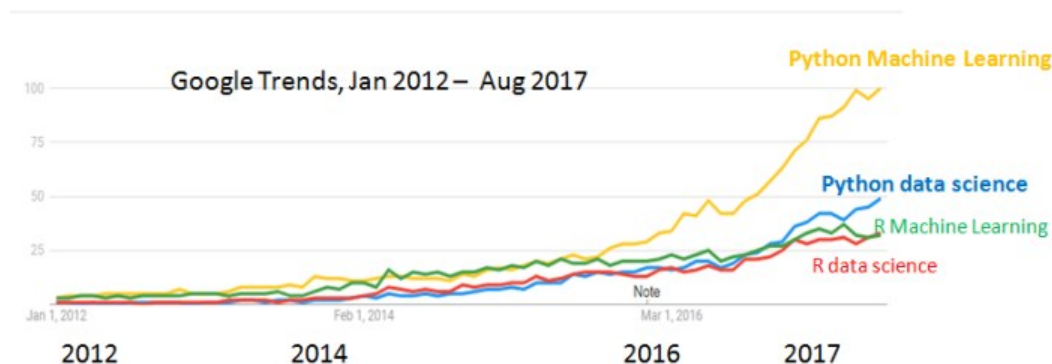
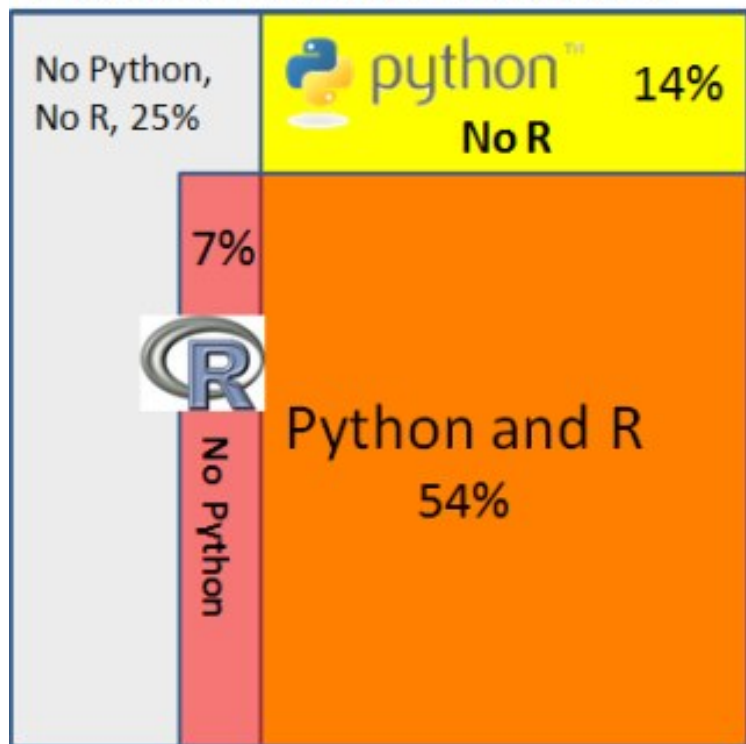


Fig. 2: Google Trends, Jan 2012 - Aug 2017, "Python Machine Learning", "R Machine Learning", "Python data science", and "R data science".

Fig. 3: Snapshot of indeed.com Data Scientist job ads in USA that also include Python and/or R, Sep 2017

Herramientas, Lenguajes, Kaggle

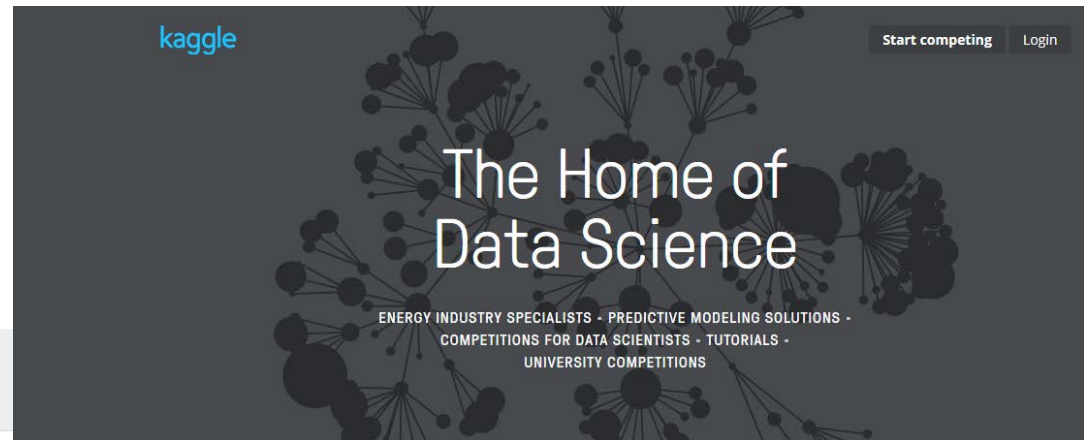
... y un buen enlace para comenzar a practicar, KAGGLE

Kaggle: The Home of Data Science

<http://www.kaggle.com/>

Es un portal web que
ofrece
competiciones,
tutoriales, actividades
académicas ...

kaggle
in Class



Academic Machine Learning Competitions

Theory, meet practice.

Kaggle hosts free projects for hundreds of universities around the globe. Engage students with an opportunity to apply machine learning to real problems.

[Learn about hosting](#)

Berkeley
UNIVERSITY OF CALIFORNIA

UNIVERSITY
OF CALIFORNIA

UCL

COLUMBIA

Cornell

ERASMUS
UNIVERSITY

HARVARD

THE UNIVERSITY OF
MELBOURNE

MICHIGAN

UNIVERSITY OF
OXFORD

Stanford
University

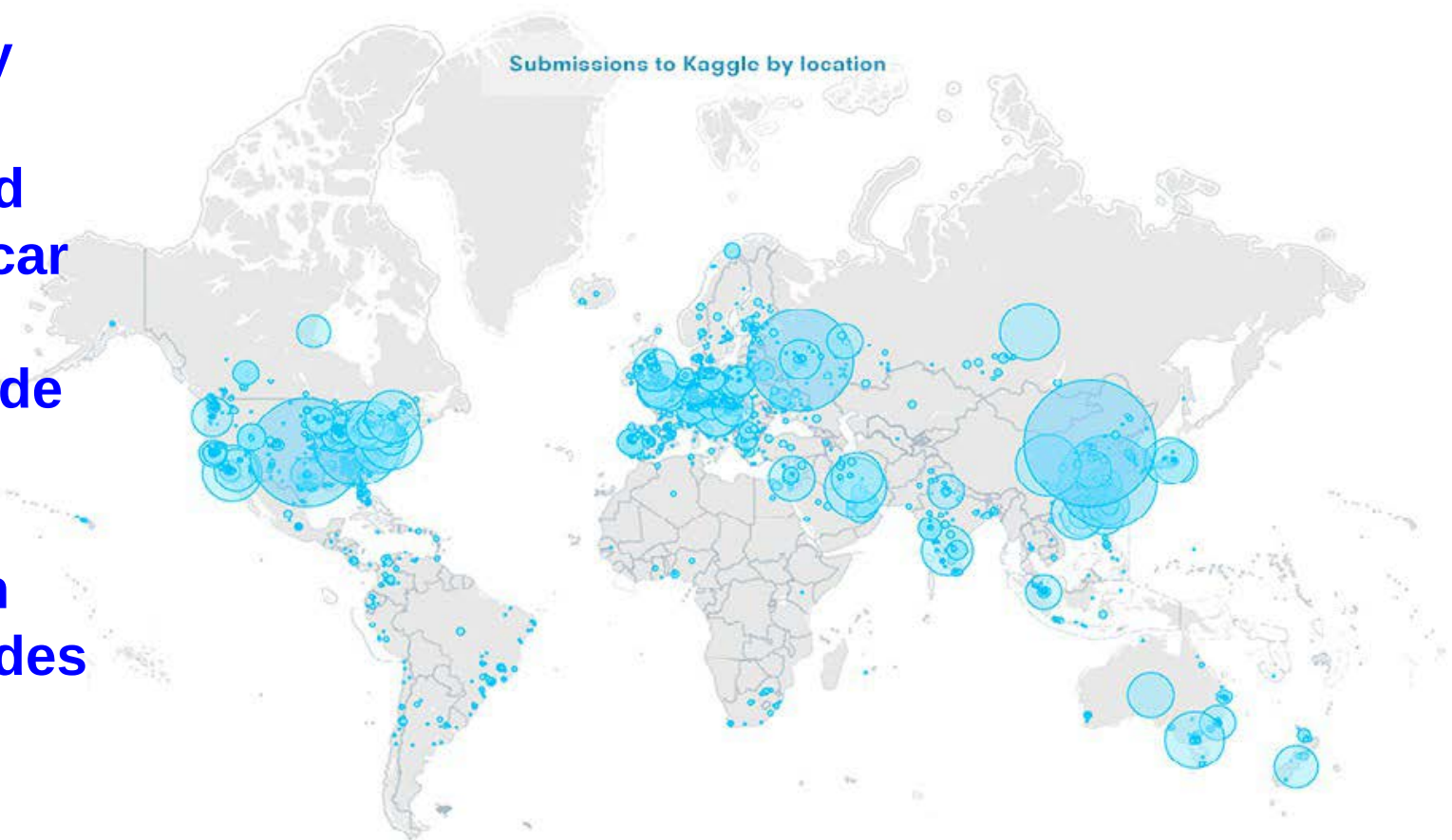
UNIVERSITY
OF TORONTO

Herramientas, Lenguajes, Kaggle

... y un buen enlace para comenzar a practicar, KAGGLE

Kaggle: The Home of Data Science

Es una muy buena oportunidad para practicar en la resolución de problemas reales y la adquisición de habilidades en Data Science.



Herramientas, Lenguajes, Kaggle

... y un buen enlace para comenzar a practicar, KAGGLE

Kaggle: The Home of Data Science (2016)

Active Competitions

All Competitions



Bosch Production Line Performance

Reduce manufacturing failures

2 months to go - **Featured**

557 teams
342 kernels
\$30,000



Melbourne University AES/MathWorks/NIH Seizure Prediction

Predict seizures in long-term human intracranial EEG recordings

2 months to go - **Research**

360 teams
258 kernels
\$20,000

Herramientas, Lenguajes, Kaggle

... y un buen enlace para comenzar a practicar, KAGGLE

Kaggle: The Home of Data Science (2017)

Active

All

Entered

Sort by Prize

18 active competitions

All Categories

Search competitions



Passenger Screening Algorithm Challenge

Improve the accuracy of the Department of Homeland Security's threat recognition algorithms

Featured · 3 months to go · terrorism, image, object detection

\$1,500,000

265 teams



Zillow Prize: Zillow's Home Value Prediction (Zestimate)

Can you improve the algorithm that changed the world of real estate?

Featured · 4 months to go · housing, real estate

\$1,200,000

3,076 teams



Cdiscount's Image Classification Challenge

Categorize e-commerce photos

Featured · 3 months to go · multiclass classification

\$35,000

102 teams



Carvana Image Masking Challenge

Automatically identify the boundaries of the car in an image

Featured · 6 days to go · object segmentation, automobiles, image

\$25,000

699 teams



Web Traffic Time Series Forecasting

Forecast future traffic to Wikipedia pages

Research · 2 months to go · forecasting, internet, tabular, time series

\$25,000

1,095 teams

74

Herramientas, Lenguajes, Kaggle

... y un buen enlace para comenzar a practicar, KAGGLE

Kaggle: The Home of Data Science (2017)

Active Competitions

All Competitions

IMAGENET

Identify and label everyday objects in images

Research · 12 years to go · image, object detection

IMAGENET

ImageNet Object Localization Challenge

Identify the objects in images

Research · 12 years to go · object detection, image

IMAGENET

ImageNet Object Detection from Video Challenge

Identify and label ordinary objects in videos

Research · 12 years to go · image, object detection



Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Getting Started · 3 years to go · binary classification, tabular

8,238 teams



House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Getting Started · 2 years to go · tabular, regression

1,720 teams



Digit Recognizer

Learn computer vision fundamentals with the famous MNIST data

Getting Started · 2 years to go · image, tabular, object identification, multiclass classification

1,679 teams

Herramientas, Lenguajes, Kaggle

... y un buen enlace para comenzar a practicar, KAGGLE

Kaggle: The Home of Data Science

101



Digit Recognizer

Classify handwritten digits using the famous MNIST data

4 months
397 teams
Knowledge



MNIST data

Dashboard ▾

Leaderboard - Digit Recognizer

This leaderboard is calculated on approximately 25% of the test data. The final results will be based on the other 75%, so the final standings may be different.

See someone using multi

#	Δ1w	Team Name	Score 🏆	Entries	Last Submission UTC (Best - Last Submission)
1	—	tepei	1.00000	1	Tue, 08 Jul 2014 15:44:55
2	—	never	1.00000	1	Wed, 16 Jul 2014 13:32:08
3	—	Yonghong	1.00000	2	Tue, 22 Jul 2014 13:41:07
4	—	Shicai Yang	1.00000	3	Wed, 23 Jul 2014 01:52:53
5	—	Aviad_Abigail	0.99757	4	Mon, 30 Jun 2014 11:17:11 (-25.4h)
6	—	Jonathan	0.99671	2	Thu, 26 Jun 2014 04:49:48
7	—	trinh	0.99657	1	Wed, 23 Jul 2014 12:13:59
8	—	Yuxin Wu	0.99643	2	Sat, 09 Aug 2014 18:30:25
9	—	田中 浩一	0.99620	1	Fri, 18 Jul 2014 05:15:13

Herramientas, Lenguajes, Kaggle

... y un buen enlace para com

Kaggle Rankings

Competitions

Kernels

Discussion

Learn more about rankings >

95

Grandmasters

885

Masters

2,630

Experts

45,894

Contributors

15,815

Novices

Rank	Tier	User	Medals	Points
1		 Gilberto Titericz Junior joined 5 years ago	 34  25  19	194,882
2		 Stanislav Semenov joined 4 years ago	 28  9  0	190,356
3		 Μαριος Μιχαηλιδης KazAnova joined 4 years ago	 26  23  21	168,976
4		 Faron joined 3 years ago	 14  4  3	132,862
5		 Eureka joined 4 years ago	 16  13  3	131,759
6		 raddar joined 2 years ago	 9  6  3	119,285
7		 idle_speculation joined 4 years ago	 7  8  6	116,367
8		 weiwei joined a year ago	 5  3  1	108,836
9		 bestfitting joined a year ago	 5  3  0	107,497
10		 Silogram joined 5 years ago	 10  24  9	97,850

A KAGGLE competition with a Multiclass and imbalanced problem

Classify products into the correct category

The Otto Group is one of the world's biggest e-commerce companies, with subsidiaries in more than 20 countries, including Crate & Barrel (USA), Otto.de (Germany) and 3 Suisses (France). We are selling millions of products worldwide every day, with several thousand products being added to our product line.

A consistent analysis of the performance of our products is crucial. However, due to our diverse global infrastructure, many identical products get classified differently. Therefore, the quality of our product analysis depends heavily on the ability to accurately cluster similar products. The better the classification, the more insights we can generate about our product range.

A KAGGLE competition with a Multiclass and imbalanced problem



Competición IN-UGR



UGR15: Preprocessing for Protein Structure Pre...

Deal with a dataset on Protein Structure Prediction by preprocessing methods (feature/instance selection, noise filters, missing values)

21 days
25 teams
Knowledge



Knowledge • 25 teams

UGR15: Preprocessing for Protein Structure Prediction

Thu 12 Nov 2015

Wed 6 Jan 2016 (21 days to go)

Dashboard

Home



Data



Make a submission



Information



Description

Evaluation

Rules

Forum



Leaderboard



Competition Details » [Get the Data](#) » [Make a submission](#)



This competition is private-entry. You can view but not participate.

Deal with a dataset on Protein Structure Prediction by preprocessing methods (feature/instance selection, noise filters, missing values)

This leaderboard is calculated on approximately 50% of the test data.
The final results will be based on the other 50%, so the final standings may be different.

See someone using multiple accounts
[Let us know](#)

#	Δ1w	Team Name	Score ?	Entries	Last Submission UTC (Best - Last Submission)
1	↑4	WhiteShadow	0.70407	113	Tue, 15 Dec 2015 17:44:40
2	↓1	Nicolas García	0.70313	28	Tue, 01 Dec 2015 20:58:45 (-4d)
3	↓1	Antonio José Navarro Céspedes	0.70267	31	Sat, 12 Dec 2015 07:21:46 (-6.9h)
4	↑2	javiDiaz	0.70020	48	Sun, 13 Dec 2015 16:10:16 (-3.2d)
5	↓2	moulay	0.70007	99	Wed, 16 Dec 2015 05:51:07 (-4.4d)
6	↓2	Silviam	0.69887	35	Sun, 13 Dec 2015 22:32:37 (-19.2d)
7	—	José Manuel Rodríguez-Rabadán	0.69220	38	Sun, 29 Nov 2015 14:43:28 (-7.8d)
8	—	Barranco	0.69193	21	Tue, 15 Dec 2015 21:54:19 (-15.4d)
9	—	Medina	0.69153	25	Sat, 12 Dec 2015 08:36:45 (-15h)
10	—	Antonio Espinosa Navarro	0.68853	47	Tue, 01 Dec 2015 21:33:01 (-8.5d)
11	—	Daniel Vinuesa	0.65927	12	Wed, 25 Nov 2015 17:01:51 (-4.2h)
12	—	nael	0.64340	10	Thu, 03 Dec 2015 22:15:02 (-5.3d)
13	↑1	Miguel Ángel Cano Mesa	0.64007	16	Wed, 16 Dec 2015 03:12:51 (-31.9h)
14	↓1	Serafín Díaz	0.63493	17	Tue, 15 Dec 2015 16:33:27 (-9d)



<https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings>

¿En qué país van a reservar los nuevos usuarios?



Airbnb Recruiting: New User Bookings

Data Scientist at **Airbnb**
San Francisco

57 days
434 teams
507 scripts
Jobs



Jobs • 434 teams

Airbnb Recruiting: New User Bookings

Wed 25 Nov 2015

Thu 11 Feb 2016 (57 days to go)

Dashboard

Home

Data

Make a submission

Information

Competition Details » [Get the Data](#) » [Make a submission](#)

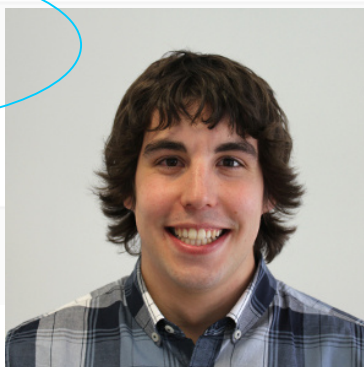
Where will a new guest book their first travel experience?



Información de la entrega

■ Posición 66 (Dic. 15) usando Gradient Boosting

62	new	 Khaled Fayed	0.86642	1	Fri, 11 Dec 2015 10:23:07
63	↓18	 SecondPlan	0.86642	28	Tue, 15 Dec 2015 09:11:25 (-3.8d)
64	↑7	 kprab1990	0.86642	5	Fri, 11 Dec 2015 16:39:26
65	↑104	 A.Antonov	0.86642	7	Fri, 11 Dec 2015 18:08:18
66	↓19	 David Gasquez	0.86642	25	Tue, 15 Dec 2015 23:43:53 (-4.1d)
67	new	Bluefool			45:42 (-2.3d)
68	↑201	 Johnny Strings			0:58



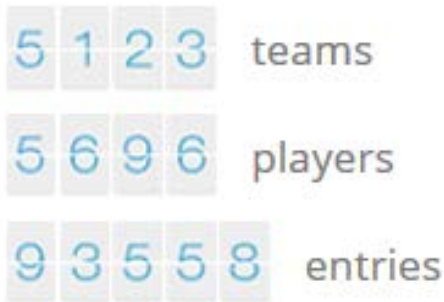
Save time managing your social media
Schedule, publish and analyze all your posts in one place

<https://buffer.com/>

Herramientas, Lenguajes, Kaggle

... y un buen enlace para comenzar a practicar, KAGGEL

Competición de Kaggle: Santander Customer Satisfaction



**Banco
Santander**

- **Objetivo:** crear un modelo que prediga qué clientes no están satisfechos
 - Premio de 60.000€ entre las 3 mejores soluciones
- Competición activa desde el 2 de marzo hasta el 2 de mayo, 2016
 - Conjunto de datos: 76020 instancias y 371 variables
 - Datos públicos 50 %

Herramientas, Lenguajes, Kaggle

... y un buen enlace para comenzar a practicar, KAGGLE

Antonio Navarro (TFG, 2016): Tenía un modelo... ¡que lo habría dejado en la posición 17 si lo hubiese seleccionado!

13	1266	Ouranos	0.828086	229	Mon, 02 May 2016 23:48:03 (-0.5h)
14	1318	g a	0.828011	37	Mon, 02 May 2016 18:00:11 (-0.6h)
15	11298	Kide	0.827992	194	Sun, 01 May 2016 15:46:40 (-38.9d)
16	11144	PradeepHegde	0.827986	19	Mon, 02 May 2016 23:01:45
-		Antonio José Navarro Céspedes	0.827968	-	Tue, 03 May 2016 17:00:38 Post-Deadline
Post-Deadline Entry If you would have submitted this entry during the competition, you would have been around here on the leaderboard.					
17	11712	scndll	0.827952	25	Mon, 02 May 2016 16:21:43 (-22.3d)
18	11895	Niels de Brabander	0.827939	25	Mon, 25 Apr 2016 06:18:55 (-8.8d)

- ❑ ¿Qué es la Ciencia de Datos?
- ❑ Minería de Datos
- ❑ Proceso de Minería de Datos
- ❑ Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación y Otros
- ❑ El Poder de los Datos
- ❑ Herramientas y Lenguajes en Ciencia de Datos. Repositorio de Kaggle
- ❑ **Comentarios Finales**

Comentarios Finales



La Minería de Datos
es una forma de
aprender del pasado
para tomar mejores
decisiones en el
futuro

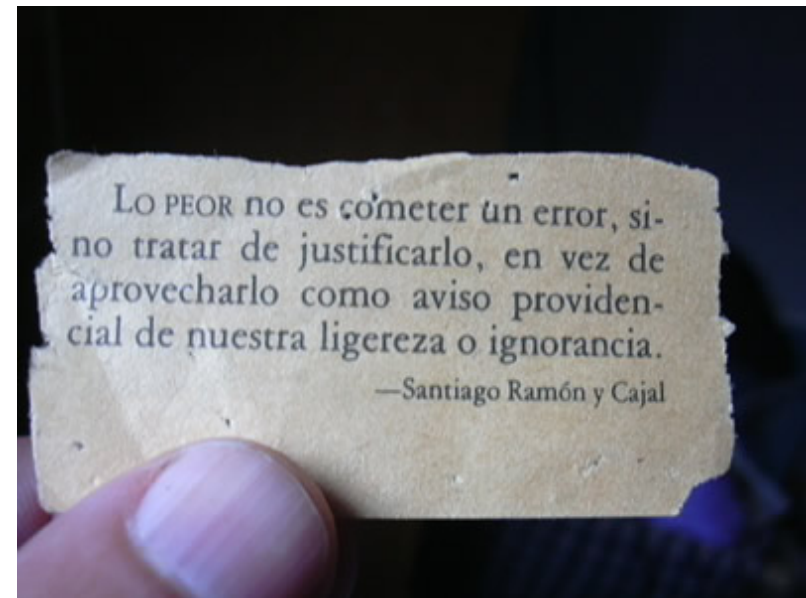
Business Intelligence



Comentarios Finales

Hay que evitar los errores comunes

- Aprender de cosas que no son ciertas
 - Patrones que no representan ninguna regla subyacente
 - Datos que no reflejan lo relevante
 - Datos con un nivel de detalle erróneo
- Aprender cosas ciertas, pero inútiles
 - Aprender información ya conocida
 - Aprender cosas que no se pueden utilizar



**Hay que obtener
conocimiento útil**

Comentarios Finales

- **Ciencia de datos:** Es el ámbito de conocimiento que engloba las habilidades asociadas al procesamiento de datos, extracción de conocimiento de datos, incluyendo Big Data.
- **Minería de datos:** descubrimiento de patrones interesantes en una base de datos (usualmente grande).
- **Un proceso de KDD incluye:** limpieza de datos, integración, reducción de datos, transformación, minería de datos, evaluación, y presentación del conocimiento.
- La minería de datos puede utilizarse sobre una gran variedad de fuentes de información (numérica, textos, ...).
- **Funcionalidades en Minería de Datos:** caracterización, asociación, regresión, characterization, agrupamiento, detección outlier, tendencias, minería de textos, big data, ...

Comentarios Finales

Una demanda creciente de profesionales en “Big Data” y “Ciencia de Datos”

Oportunidades en Big Data y Ciencia de Datos

La demanda de profesionales formados en Ciencia de Datos y *Big Data* es enorme.

Se estima que la conversión de datos en información útil generará un mercado de 132.000 millones de dólares en 2015 y que se crearán más de 4.4 millones de empleos.

España necesitará para 2015 más de 60.000 profesionales con formación en Ciencia de Datos y *Big Data*.



http://economia.elpais.com/economia/2013/09/27/actualidad/1380283725_938376.html

Comentarios Finales

Una demanda creciente de profesionales en “Big Data” y “Ciencia de Datos”

Oportunidades en Big Data y Ciencia de Datos

http://www.revistacloudcomputing.com/2013/10/espana-necesitara-60-000-profesionales-de-big-data-hasta-2015/?goback=.gde_4377072_member_5811011886832984067#!

España necesitará 60.000 profesionales de Big Data hasta 2015

📅 22 octubre, 2013 📍 Eventos 💬 18



España necesitará 60.000 profesionales de Big Data hasta 2015

“España va a necesitar alrededor de sesenta mil profesionales del Big Data de aquí a 2015”, así lo ha asegurado Francisco Javier Antón, Subdirector General de Tecnologías del Ministerio de Educación, Cultura y Deportes en una mesa redonda sobre beneficio y aplicación de Big Data en pymes, moderada por Daniel Tapias de [Sigma Technologies](#), celebrada durante el [4º Congreso Nacional de CENTAC](#) de

“Existe una demanda mundial para formar a 4,4 millones de profesionales de la gestión Big Data desde ingenieros, gestores y científicos de datos”, comenta Antón. Sin embargo, “las empresas todavía no ven en el Big Data un modelo de negocio”, lamenta. “Solo se extrae un 1 % de los datos disponibles en la red”, añade. “Hace falta formación y concienciación.”

Bibliografía

- Libros buenos e interesantes:

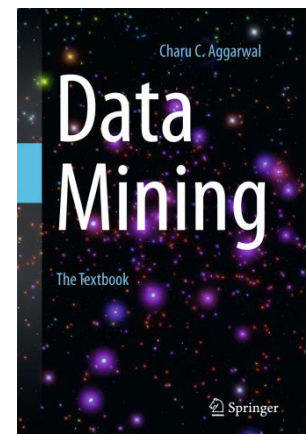
- <http://www.datasciencecentral.com/profiles/blogs/15-books-every-data-scientist-should-read>

- Libros gratis:

- <http://www.wzchen.com/data-science-books/>

- Libro reciente sobre Data Mining:

C.C. Aggarwal, Data Mining, Springer, 2015





INTELIGENCIA DE NEGOCIO

2018 - 2019

- Tema 1. Introducción a la Inteligencia de Negocio
- Tema 2. Minería de Datos. Ciencia de Datos
- Tema 3. Modelos de Predicción: Clasificación, regresión y series temporales
- Tema 4. Preparación de Datos
- Tema 5. Modelos de Agrupamiento o Segmentación
- Tema 6. Modelos de Asociación
- Tema 7. Modelos Avanzados de Minería de Datos.
- Tema 8. Big Data