CURSOS DE VERANO 2014

APROXIMACIÓN PRÁCTICA A LA CIENCIA DE DATOS Y BIG DATA: HERRAMIENTAS KNIME, R, HADOOP Y MAHOUT

Introducción a la Ciencia de Datos, Minería de Datos y Big Data

Francisco Herrera

CURSOS VERANO UNIA 2014

CAMPUS ANTONIO MACHADO DE BAEZA

Del 25 al 28 de agosto

CURSO / 3476

APROXIMACIÓN PRÁCTICA A LA CIENCIA DE DATOS Y BIG DATA: HERRAMIENTAS KMINE, R, HADOOP Y MAHOUT

Introducción a la Ciencia de Datos, Minería de Datos y Big data

Francisco Herrera

Dpto. Ciencias de la Computación e I.A.

Universidad de Granada

herrera@decsai.ugr.es

Grupo de investigación SCI²S

http://sci2s.ugr.es

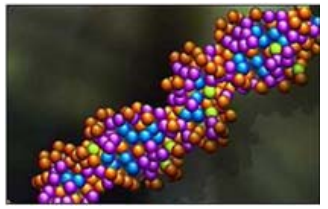
DECSAI

Universidad de Granada

Ciencia de Datos, Minería de Datos, Big Data

Nuestro mundo gira en torno a los datos

- Ciencia
  - Bases de datos de astronomía, genómica, datos medio-ambientales, datos de transporte, …



- Ciencias Sociales y Humanidades
  - Libros escaneados, documentos históricos, datos sociales, …

- Negocio y Comercio
  - Ventas de corporaciones, transacciones de mercados, censos, tráfico de aerolíneas, …

- Entretenimiento y Ocio
  - Imágenes en internet, películas, ficheros MP3, …
- Medicina
  - Datos de pacientes, datos de escaner, radiografías …

- Industria, Energía, …
  - Sensores, …

Motivación

El problema de la explosión de información:

- existencia de herramientas para la recolección de información
- madurez de la tecnología de bases de datos
- bajo precio del hardware

➔ Disponemos de cantidades gigantescas de datos almacenados en bases de datos, *data warehouses* y otros tipos de almacenes de información

Somos ricos en datos pero pobres en conocimiento

El progreso y la innovación ya no se ven obstaculizados por la capacidad de recopilar datos, sino por la capacidad de gestionar, analizar, sintetizar, visualizar, y descubrir el conocimiento de los datos recopilados de manera oportuna y en una forma escalable

Ciencia de Datos, Minería de Datos, Big Data

ELMUNDO.es
Líder mundial en español | Miércoles 04/09/2013. Actualizado 16:27h.

Alex ' Sandy' Pentland, director del programa de emprendedores del 'Media Lab' del Massachusetts Institute of Technology (MIT)

INTERNET | Campus Party Europa 2013

'Es la década de los datos y de ahí vendrá la revolución'

Alex 'Sandy' Pentland, durante su ponencia. | M. Sáinz

Considerado por 'Forbes' como uno de los siete científicos de datos más influyentes del mundo

Ben Chams - Fotolia

http://www.elmundo.es/elmundo/2013/09/03/navegante/1378243782.html

Objetivos de esta sesión:

- Introducir los conceptos de ciencia de datos, minería de datos y big data
- Conocer las etapas del proceso de minería de datos
- Introducir las técnicas clásicas de minería de datos, casos de estudio, lenguajes de programación utilizados, ...

Objetivo del curso: Ciencia de datos es un campo muy amplio de conocimiento y de tecnologías asociadas. En el curso trataremos de introducir brevemente las amplias áreas de estudio en minería de datos y big data, y formar a nivel de iniciación práctica en cuatro herramientas y lenguajes muy utilizados: KNIME, R, Hadoop y Mahout.

Página web con el material del curso

http://sci2s.ugr.es/docencia/index.php

http://sci2s.ugr.es/docencia/asignatura.php?id_asignatura=16

Ciencia de Datos y 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
- Minería de Datos: Casos de uso
- Herramientas y Lenguajes en Ciencia de Datos. Repositorio Kaggle
- Comentarios Finales

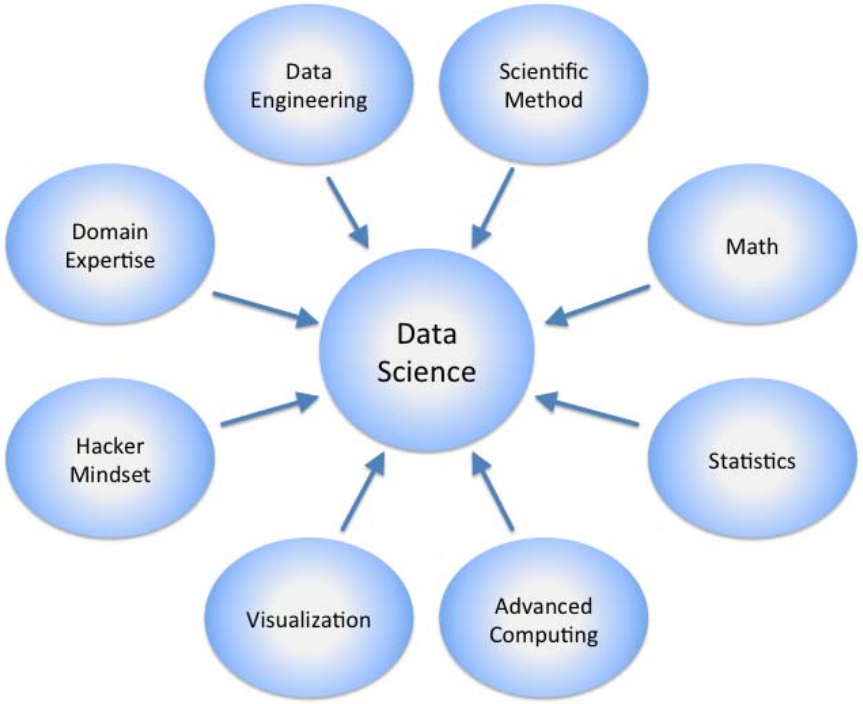
Ciencia de Datos y Minería de Datos

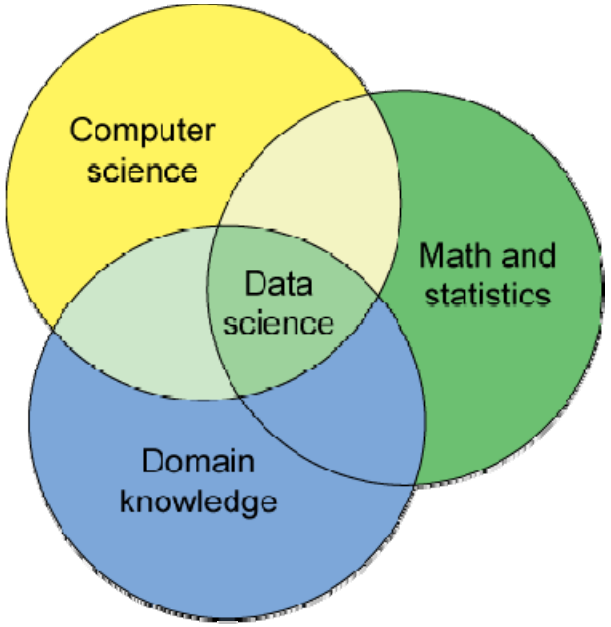
- **¿Qué es la Ciencia de Datos?**
- Minería de Datos
- Proceso de Minería de Datos
- Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación y Otros
- Minería de Datos: Casos de uso
- Herramientas y Lenguajes en Ciencia de Datos. Repositorio Kaggle
- Comentarios Finales

Ciencia de Datos

Data Science

Ciencia de Datos es el ámbito de conocimiento que engloba **las habilidades asociados al procesamiento de datos, incluyendo Big Data**





Ciencia de Datos

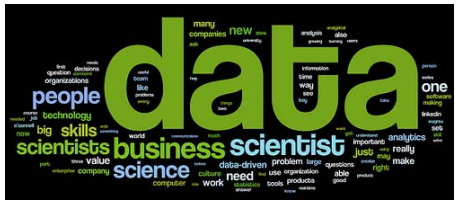
Data Science o la **Ciencia de Datos** incorpora diferentes elementos y se basa en las técnicas y teorías de muchos campos, incluyendo las matemáticas, estadística, ingeniería de datos, reconocimiento de patrones y aprendizaje, computación avanzada, visualización, modelado de la incertidumbre, almacenamiento de datos y la informática de alto rendimiento con el objetivo de extraer el significado de datos y la creación de productos de datos.

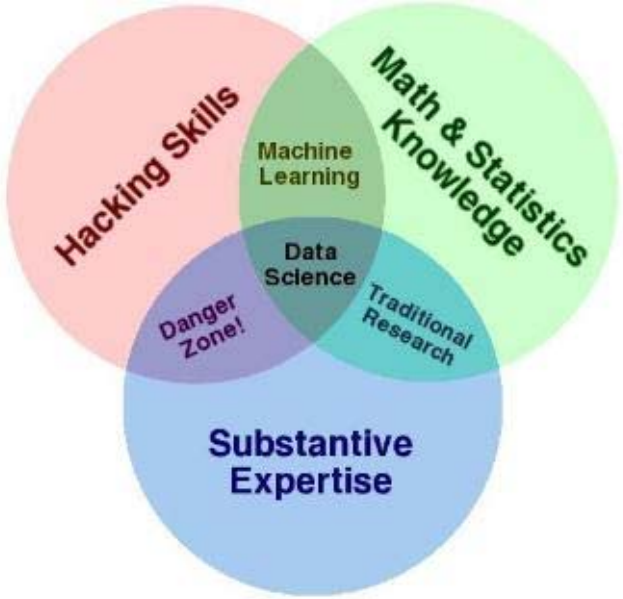
Es un término relativamente nuevo que se utiliza a menudo de manera intercambiable con inteligencia o analítica de negocio ó analítica de datos. La ciencia de datos busca utilizar todos los datos disponibles y relevantes para **"extraer conocimiento"** que pueda ser fácilmente comprendido por los expertos en el área de aplicación. Un experto de la ciencia de datos se denomina un científico de datos.

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, conocimientos de programación (y sus múltiples lenguajes), ciencias de la computación y analítica.





Data Science
Machine Learning
Traditional Research
Danger Zone =
Traditional software

Ciencia de Datos

José Antonio Guerrero: uno de los mejores científicos de datos del mundo (Plataforma Kaggle)

¿Qué es un científico de datos?

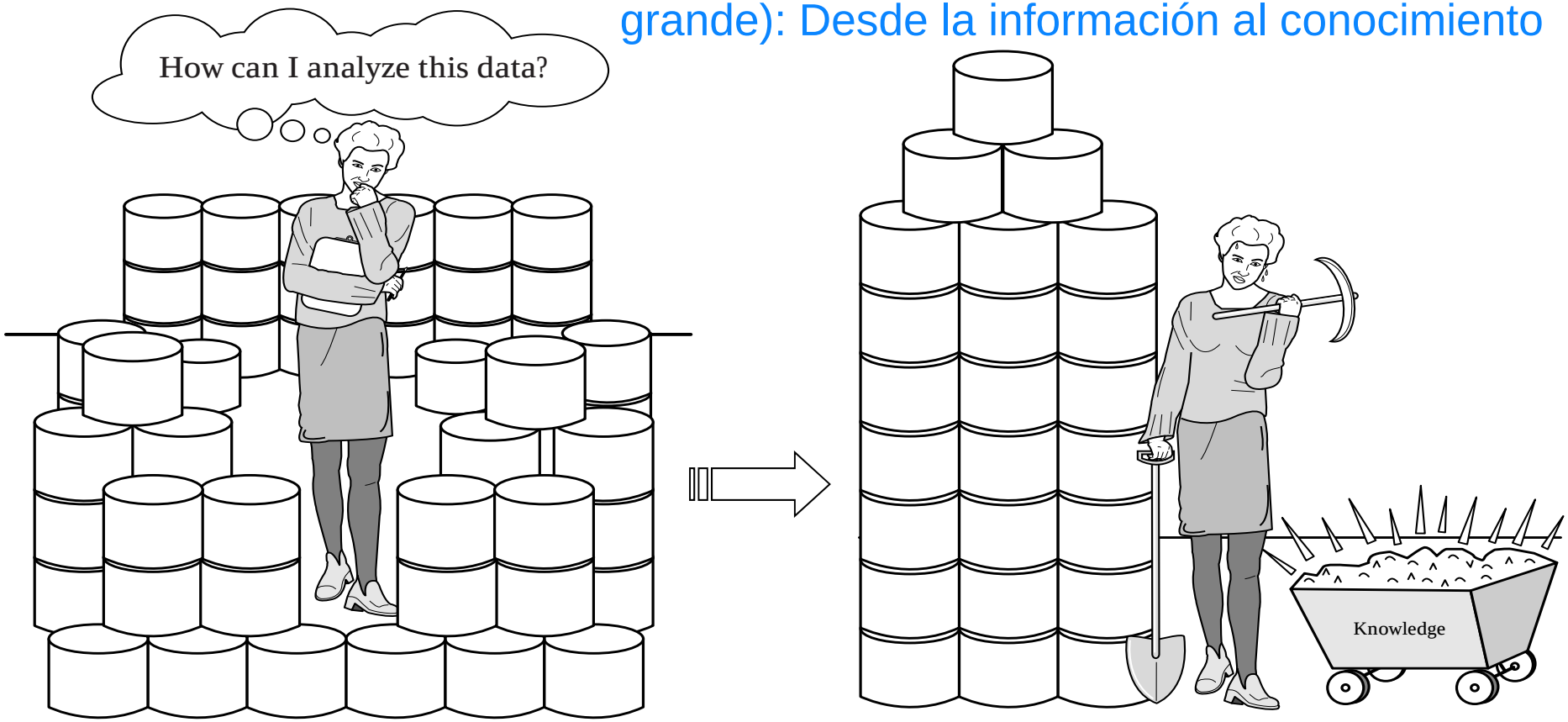
"Es una persona con fundamentos en matemáticas, estadística y métodos de optimización, con conocimientos en lenguajes de programación y que además tiene una experiencia práctica en el análisis de datos reales y la elaboración de modelos predictivos.
De las tres características quizás la más difícil es la tercera; no en vano la modelización de los datos se ha definido en ocasiones como un arte. Aquí no hay reglas de oro, y cada conjunto de datos es un lienzo en blanco."

Leer más: http://www.elconfidencial.com/tecnologia/2013-12-19/un-matematico-andaluz-desconocido-es-el-mejor-cientifico-de-datos-del-mundo_67675/

Ciencia de Datos

Minería de Datos

Descubrimiento de patrones interesantes en una base de datos (usualmente grande): Desde la información al conocimiento



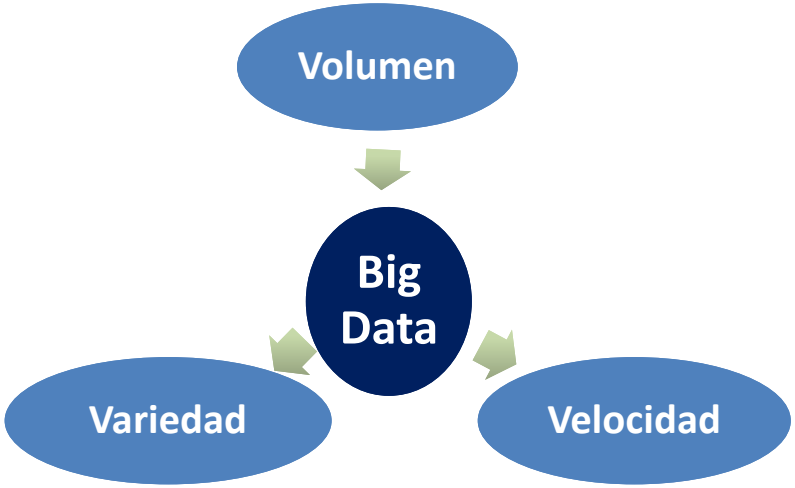
We have rich data,
but poor information

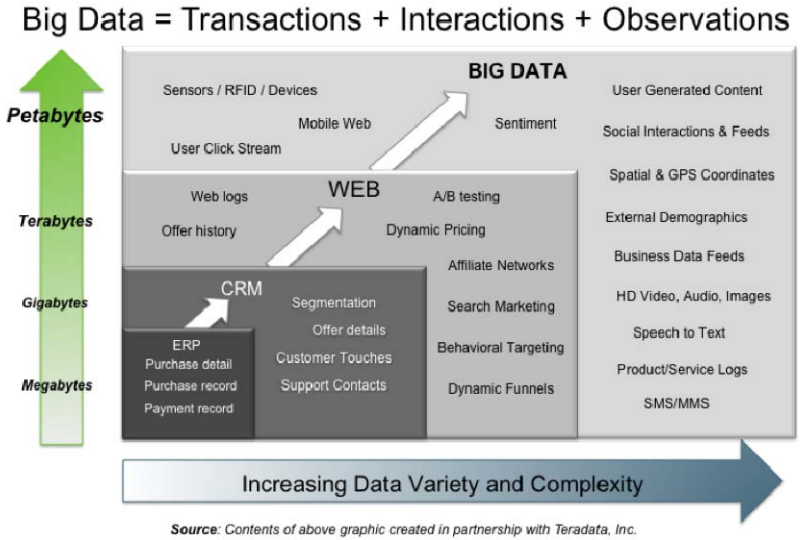
Data mining-searching for knowledge
(interesting patterns) in your data.

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 ...



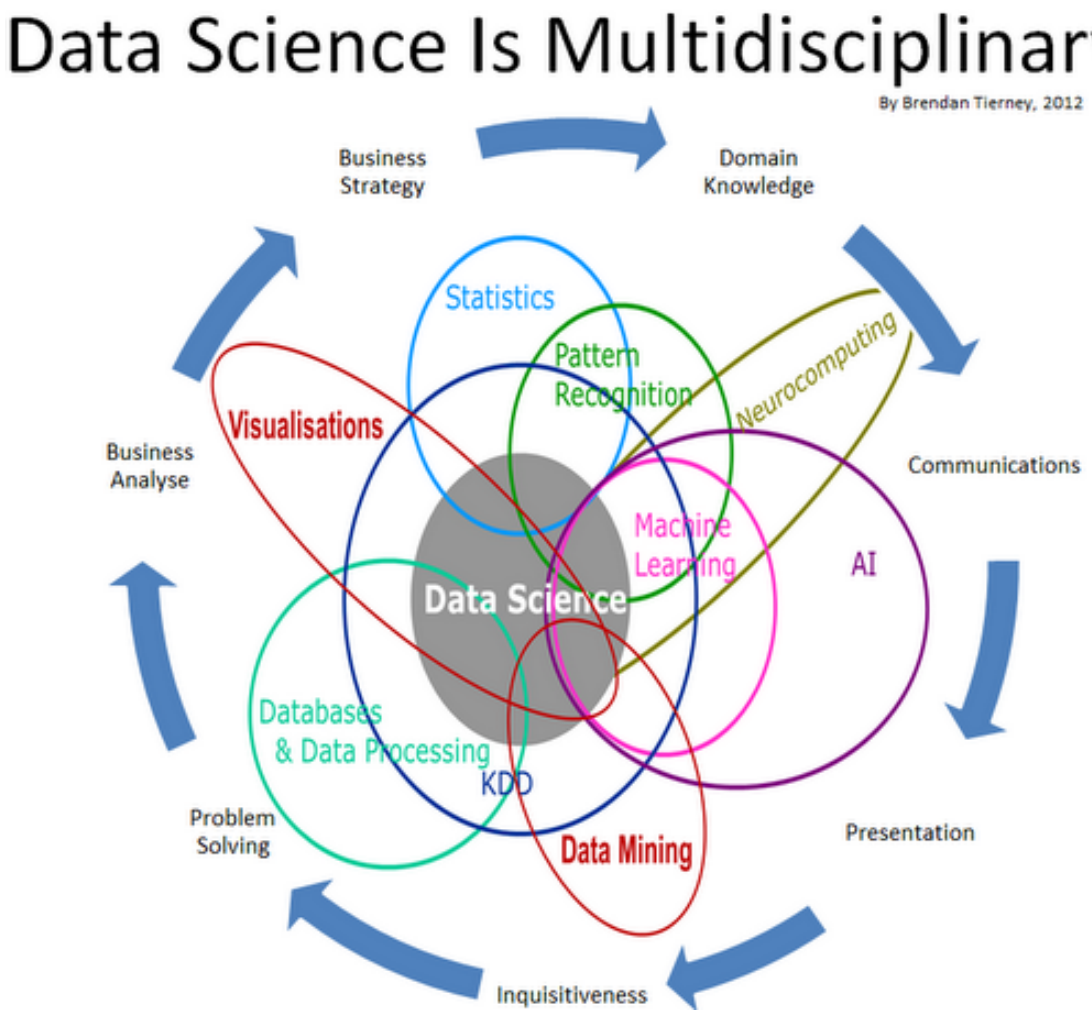


Ciencia de Datos

Big Data

Big data es cualquier característica sobre los datos que represente un reto para las funcionalidades de un sistema.

Ciencia de Datos



Big Data

- ¿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
- Minería de Datos: Casos de uso
- Herramientas y Lenguajes en Ciencia de Datos. Repositorio 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,
- ...

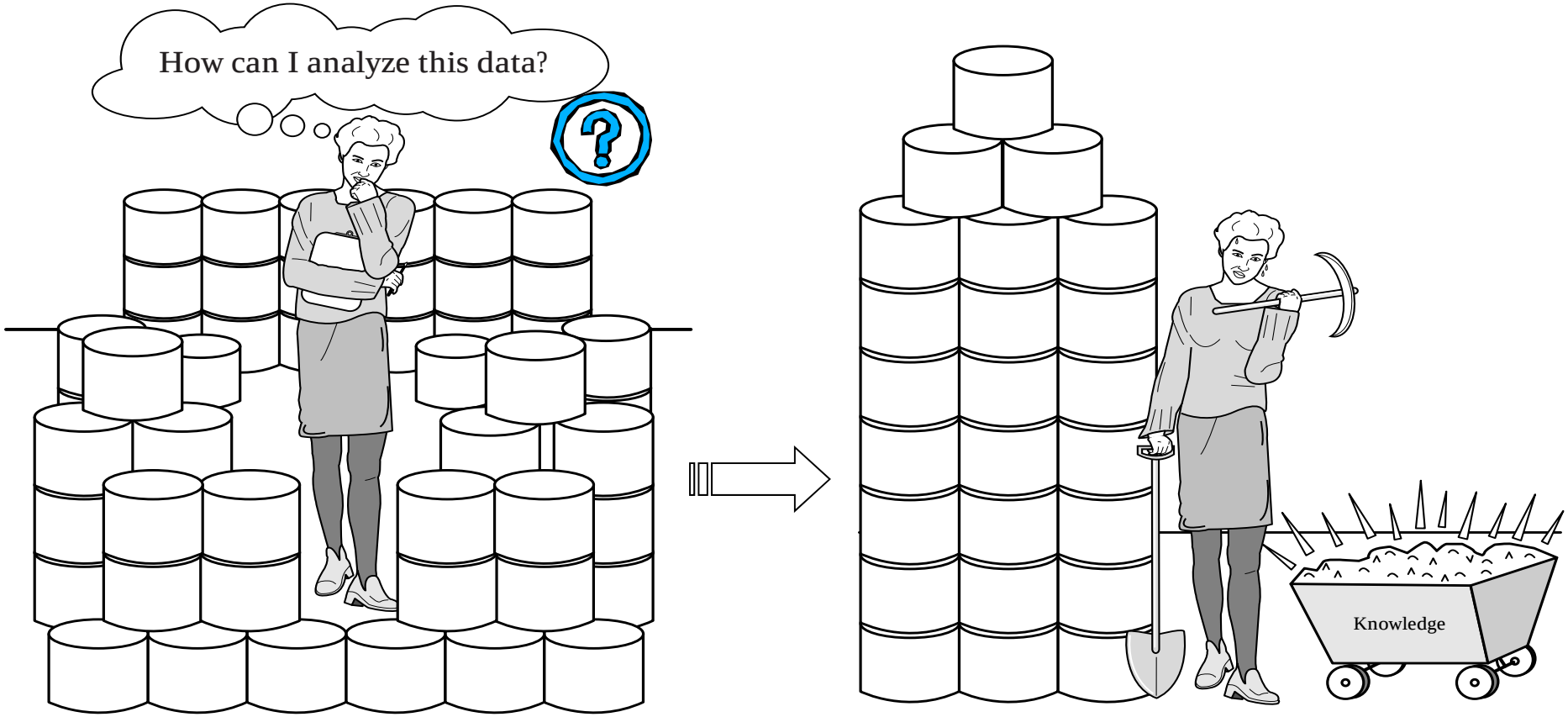
¿Qué es la Minería de Datos?

- Muchas de las técnicas utilizadas en MD ya se conocían previamente, ¿a qué se debe?

- En los 90's convergen los siguientes factores:

1. Los datos se están produciendo
2. Los datos se están almacenando
3. La potencia computacional necesaria es abordable
4. Existe una gran presión competitiva a nivel empresarial
5. Las herramientas software de MD están disponibles

¿Qué es la Minería de Datos?



¿Qué es la Minería de Datos?

¿Para qué se utiliza el 'conocimiento' obtenido?

- hacer predicciones sobre nuevos datos
- explicar los datos existentes
- resumir una base de datos masiva para facilitar la toma de decisiones
- visualizar datos altamente dimensionales, extrayendo estructura local simplificada, ...

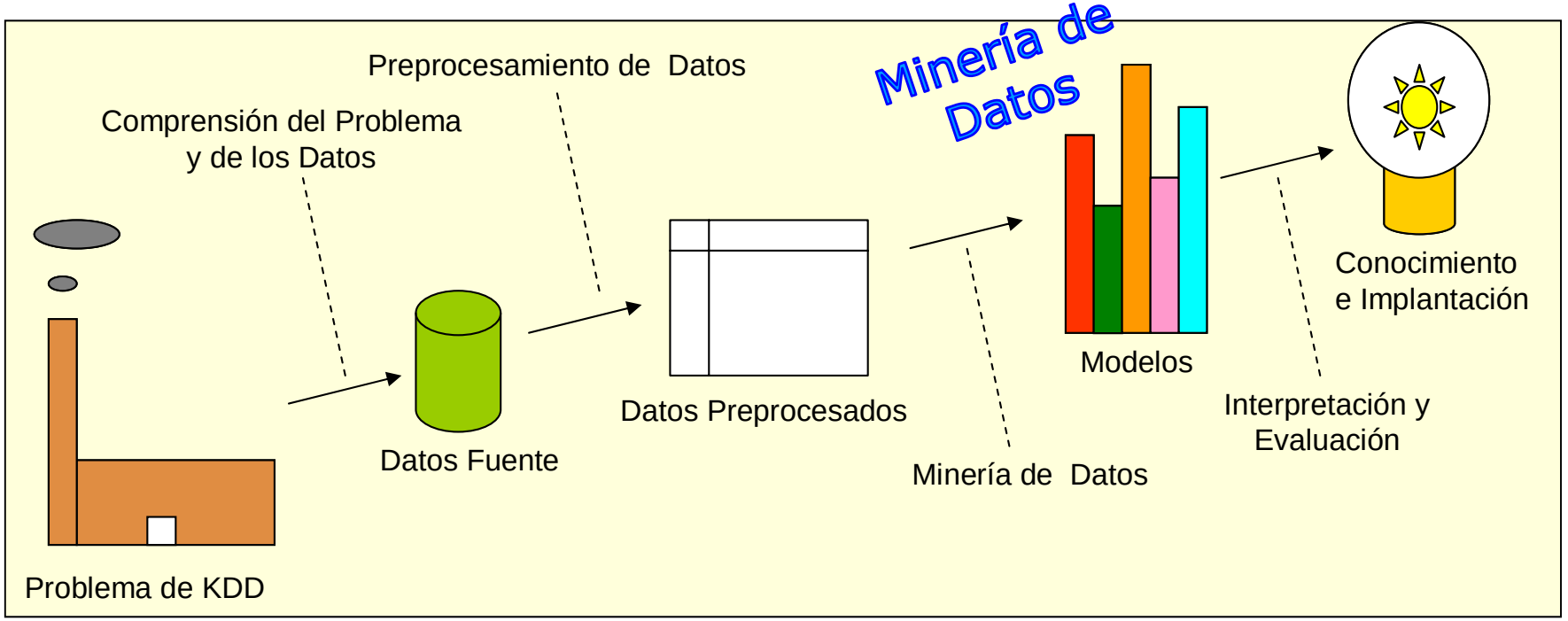
Nuevas necesidades de análisis datos

¿Qué es la Minería de Datos?

- 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

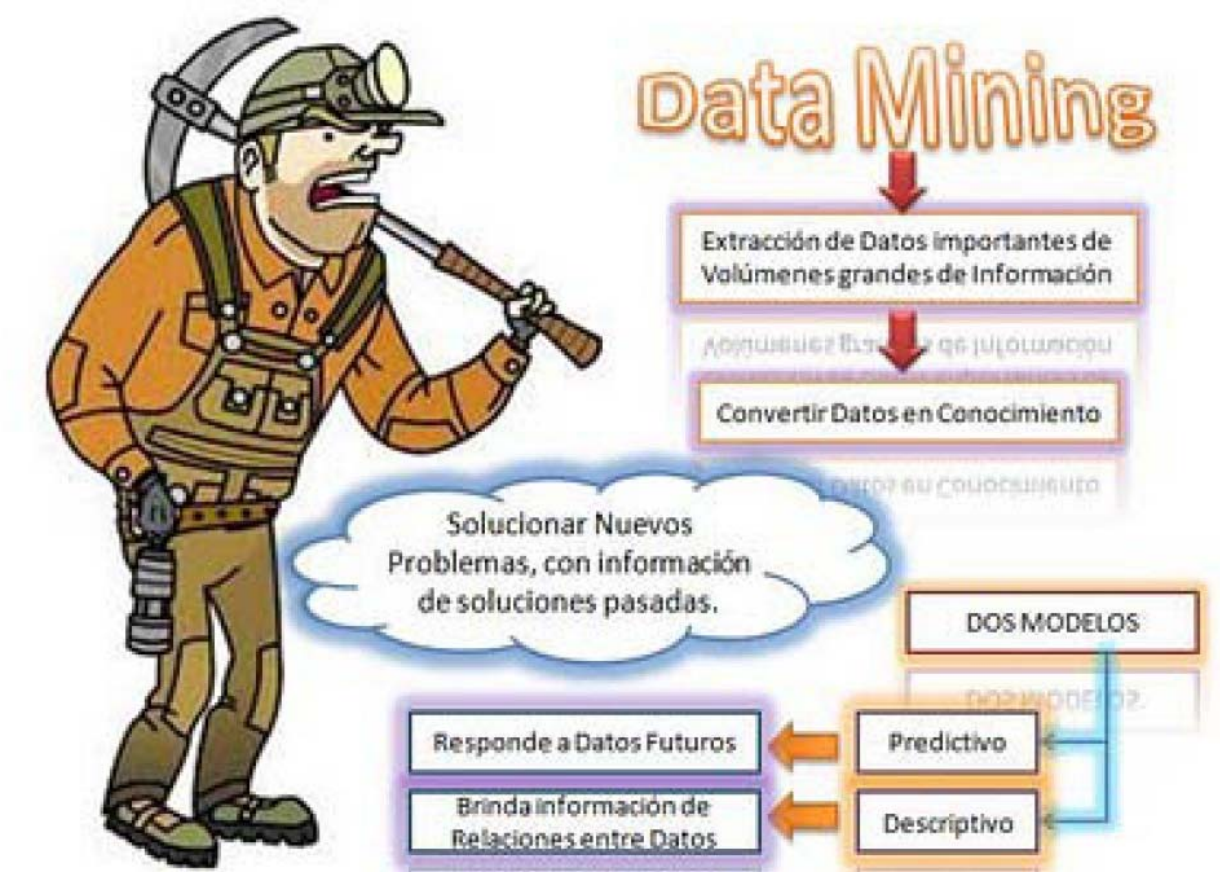
¿Qué es la Minería de Datos?

Etapas en un proceso de KDD



Informalmente se asocia Minería de Datos con KDD

¿Qué es la Minería de Datos?



Minería de Datos. Tipos de datos

¿A qué tipos de datos puede aplicarse las técnicas de Minería de Datos?

En principio, a cualquier tipo

- Bases de datos relacionales
- Bases de datos espaciales
- Bases de datos temporales
- Bases de datos documentales (Text mining)
- Bases de datos multimedia
- World Wide Web (Web mining)
  - El almacén de información más grande y diverso de los existentes
  - Existe gran cantidad de datos de los que extraer información útil
- **…. Grandes volúmenes de datos: Big Data, Social Big Data**

Minería de Datos. Áreas de aplicación

- *Aplicaciones empresariales / industriales*

Toma de decisiones en banca, seguros, finanzas, marketing, control de calidad, retención de clientes, predicción, políticas de acción (sanidad, etc.), ...

- *Aplicaciones en investigación científica*

Medicina, astronomía, geografía, genética, bioquímica, meteorología, etc.

- *Aplicaciones en Internet/Redes Sociales*

Minería de textos y de datos en la web

Minería de Datos. Áreas de aplicación

Análisis y gestión de mercados (I)

- ***Fuentes*:** transacciones con tarjetas de crédito, tarjetas de descuento, quejas de cliente, estilos de vida publicados, comentarios en redes sociales…

- ***Identificación de objetivos para marketing***: encontrar grupos (*clusters)* que identifiquen un modelo de cliente con características comunes (intereses, nivel de ingresos, hábitos de gasto, …)

- ***Determinar patrones de compra en el tiempo:*** Unificación de cuentas bancarias, compra de determinados productos simultáneamente,…

Minería de Datos. Áreas de aplicación

Análisis y gestión de mercados (II)

- ***Análisis de cestas de mercado*:** asociaciones / co-relaciones entre ventas de producto, predicción basada en asociación de informaciones,...

- ***Perfiles de cliente*:** Identificar qué tipo de clientes compra qué productos (*clustering* y/o clasificación), usar predicción para encontrar factores que atraigan nuevos clientes, retención de clientes,...

- ***Generar información resumida*:** informes multidimensionales, información estadística (tendencia central y variación), ...

Minería de Datos. Áreas de aplicación

Análisis de riesgo en banca y seguros

- **Banca**
  - Detectar patrones de uso fraudulento en tarjetas
  - Estudio de concesión de créditos y/o tarjetas
  - Determinación del gasto en tarjeta por grupos
  - Identificar reglas de comportamiento del mercado de valores a partir de históricos
- **Seguros**
  - Predicción de clientes propensos a suscribir nuevas pólizas
  - Identificar grupos/patrones de riesgo
  - Identificar tendencias de comportamiento fraudulento
- **Ambos:** Identificación de clientes leales, identificación de fuga de clientes

Minería de Datos. Áreas de aplicación

Minería de datos en industria

- **Control de calidad**
  - Detección precisa de productos defectuosos
  - Localización precoz de defectos
  - Identificación de causas de fallos

- **Procesos industriales**
  - Automatizar el control del proceso
  - Optimización del rendimiento de forma adaptativa
  - Implementar programas de mantenimiento predictivo

Minería de Datos. Áreas de aplicación

Medicina / diagnóstico

- Identificación de terapias para diferentes enfermedades
- Estudio de factores de riesgo en distintas patologías
- Segmentación de pacientes en grupos afines
- Gestión hospitalaria y planificación temporal de salas, urgencias,...
- Recomendación priorizada de fármacos para una misma patología
- Estudios en genética (ADN,...)
- Selección de embriones en reproducción artificial

Minería de Datos. Áreas de aplicación

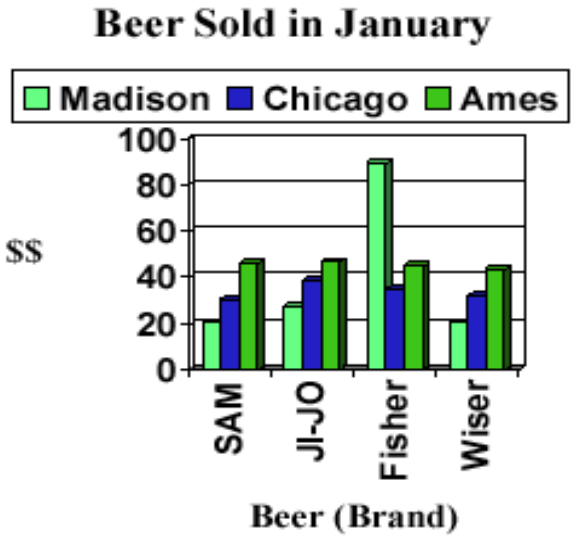
Web mining / minería de datos web

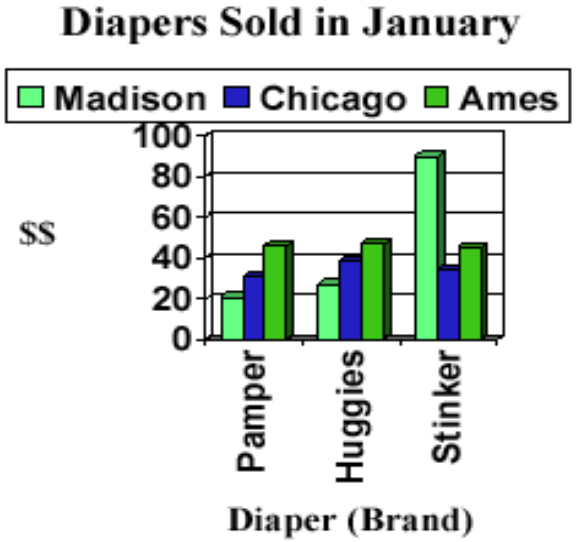
- La mayoría de las herramientas actuales analizan los ficheros .log y generan estadísticas, pero ningún conocimiento acerca de las características del cliente ni de su comportamiento
- Minería de datos web en un sitio de e-comercio, generaría análisis del comportamiento y perfiles del visitante
- Lo que interesa es responder preguntas del tipo: ¿quién compra qué producto y en qué porcentaje?
- Hay que capturar información en el servidor desde los .log, cookies, formularios, y completar con información geográfica, etc.,…
- En función de esto y de su actividad, generar perfiles de cliente y estudiar posibilidades de venta cruzada (*cross-selling*)
- Recuperación de información (*information retrieval*)

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

Ciencia de Datos y Minería de Datos

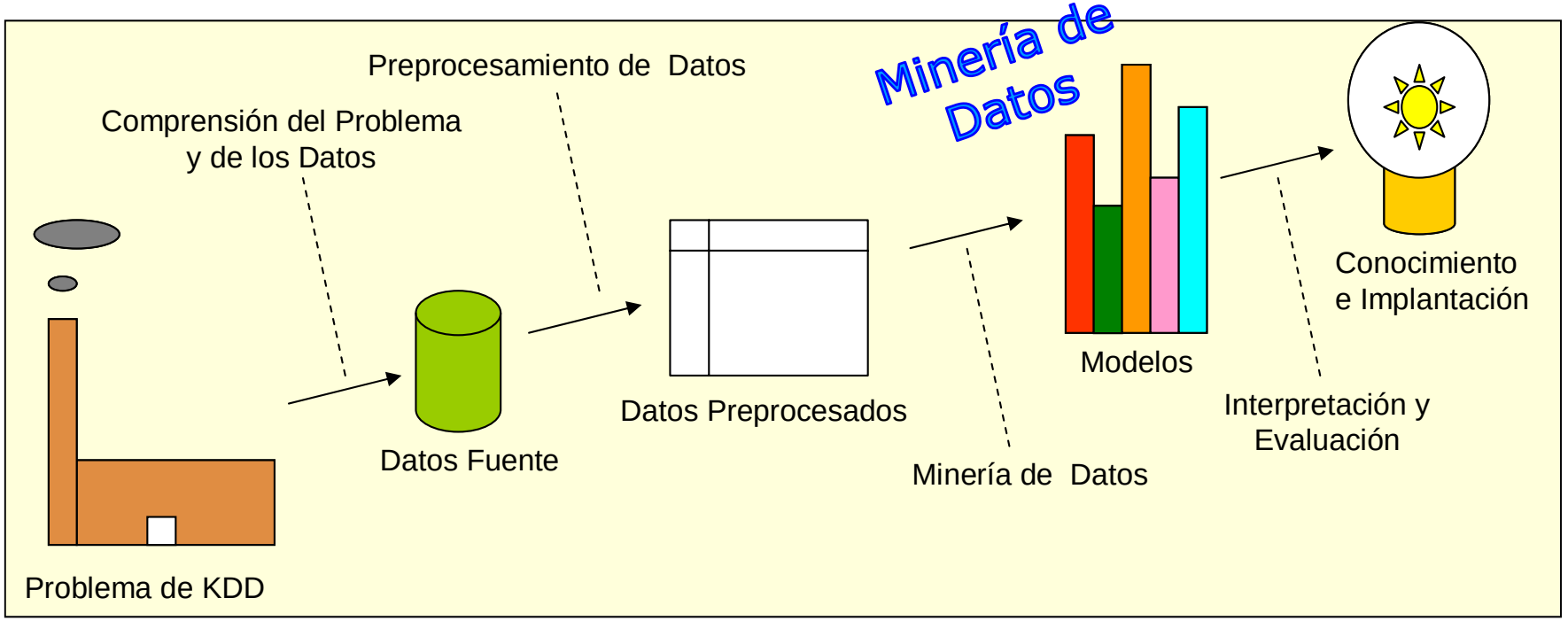
- ¿Qué es la Ciencia de Datos?
- Minería de Datos
- **Proceso de Minería de Datos**
- Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación y Otros
- Minería de Datos: Casos de uso
- Herramientas y Lenguajes en Ciencia de Datos. Repositorio Kaggle
- Comentarios Finales

Etapas en el proceso de KDD

1. **Integración y recopilación**: Comprensión del dominio de aplicación del problema, identificación de conocimiento a priori y creación del Datawarehouse
2. **Preprocesamiento**: Selección de datos, limpieza, reducción y transformación
3. **Selección de la técnica de MD** y aplicación de algoritmos concretos de MD
4. **Evaluación**, interpretación y presentación de los resultados obtenidos
5. **Difusión y utilización del nuevo conocimiento**

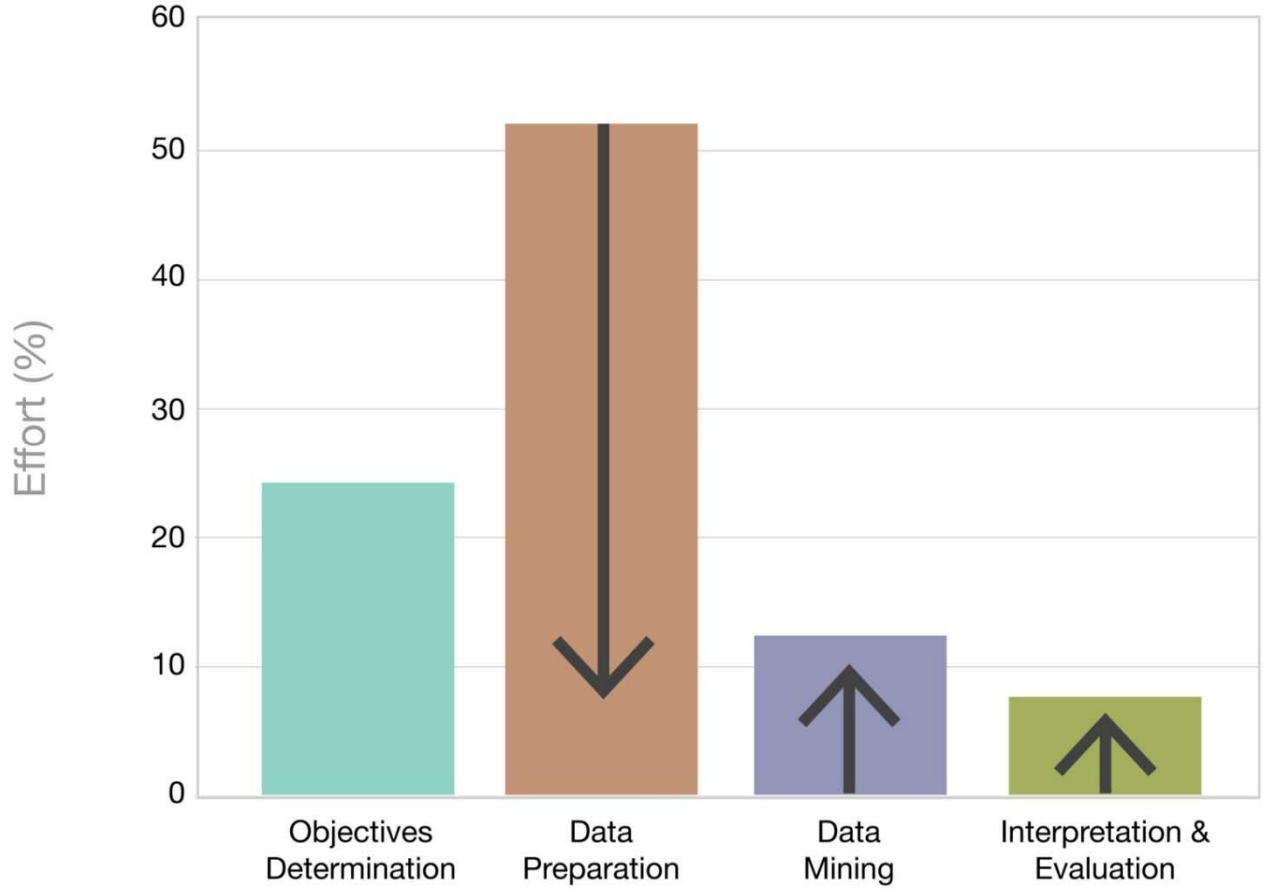
Etapas en el proceso de KDD

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

Etapas en el proceso de KDD

Selección, limpieza, reducción y transformación

- La calidad del conocimiento descubierto no depende sólo del algoritmo de DM sino de la calidad de los datos minados
- Objetivo general de esta fase: seleccionar el conjunto de datos adecuado para el resto del proceso de KDD
- Las tareas de esta etapa se agrupan en:
  - Limpieza de datos (*data cleaning*)
  - Transformación de los datos
  - Reducción de la dimensionalidad

Etapas en el proceso de KDD

Minería de datos

- Objetivo: Producir nuevo conocimiento que pueda utilizar el usuario
- ¿Cómo? Construyendo un modelo, basado en los datos recopilados, que sea una descripción de los patrones y relaciones entre los datos con los que se puedan hacer predicciones, entender mejor los datos o explicar situaciones pasadas
- Decisiones a tomar:
  - ¿Qué tipo de conocimiento buscamos?
    - Predictivo, Descriptivo
  - ¿Qué técnica es la más adecuada?
    - Clasificación, Regresión, clustering, Asociaciones, ...
  - ¿Qué tipo de modelo?
    - P.e. Clasificación: reglas, árboles de decisión, SVM, etc.
  - ¿Es necesaria la incertidumbre en el modelo resultante? Certeza, probabilidad, lógica difusa,...
  - ¿Qué algoritmo es el más adecuado?

Etapas en el proceso de KDD

Evaluación, interpretación y presentación de resultados

- La fase de MD puede producir varias hipótesis de modelos
- Es necesario establecer qué modelos son los más válidos
- Criterios: los patrones descubiertos deben ser
  - precisos,
  - comprensibles, e
  - interesantes (útiles, novedosos)
- Técnicas de evaluación: Al menos se divide el conjunto de datos en dos (entrenamiento y test)
  - Entrenamiento: Para extraer el conocimiento
  - Test: Para probar la validez del conocimiento extraído
  - Alternativas:
    - Validación simple
    - n-validación cruzada
    - *Bootstrapping*,…
- Medidas de evaluación de modelos: Dependen de la tarea:
  - Clasificación: precisión predictiva (%acierto)
  - Regresión: Error cuadrático medio
  - Agrupamiento: Medidas de cohesión y separación entre grupos
  - Reglas de asociación: cobertura, confianza…
- La interpretación de los mejores modelos (visualización, simplicidad, posibilidad de integración, ventajas colaterales,…) ayuda a la selección del modelo(s) final(es)

Etapas en el proceso de KDD

Difusión y utilización del nuevo conocimiento

Una vez construido y validado el modelo puede utilizarse:

- para recomendar acciones
- para aplicar el modelo a diferentes conjuntos de datos

En cualquier caso, es necesario:

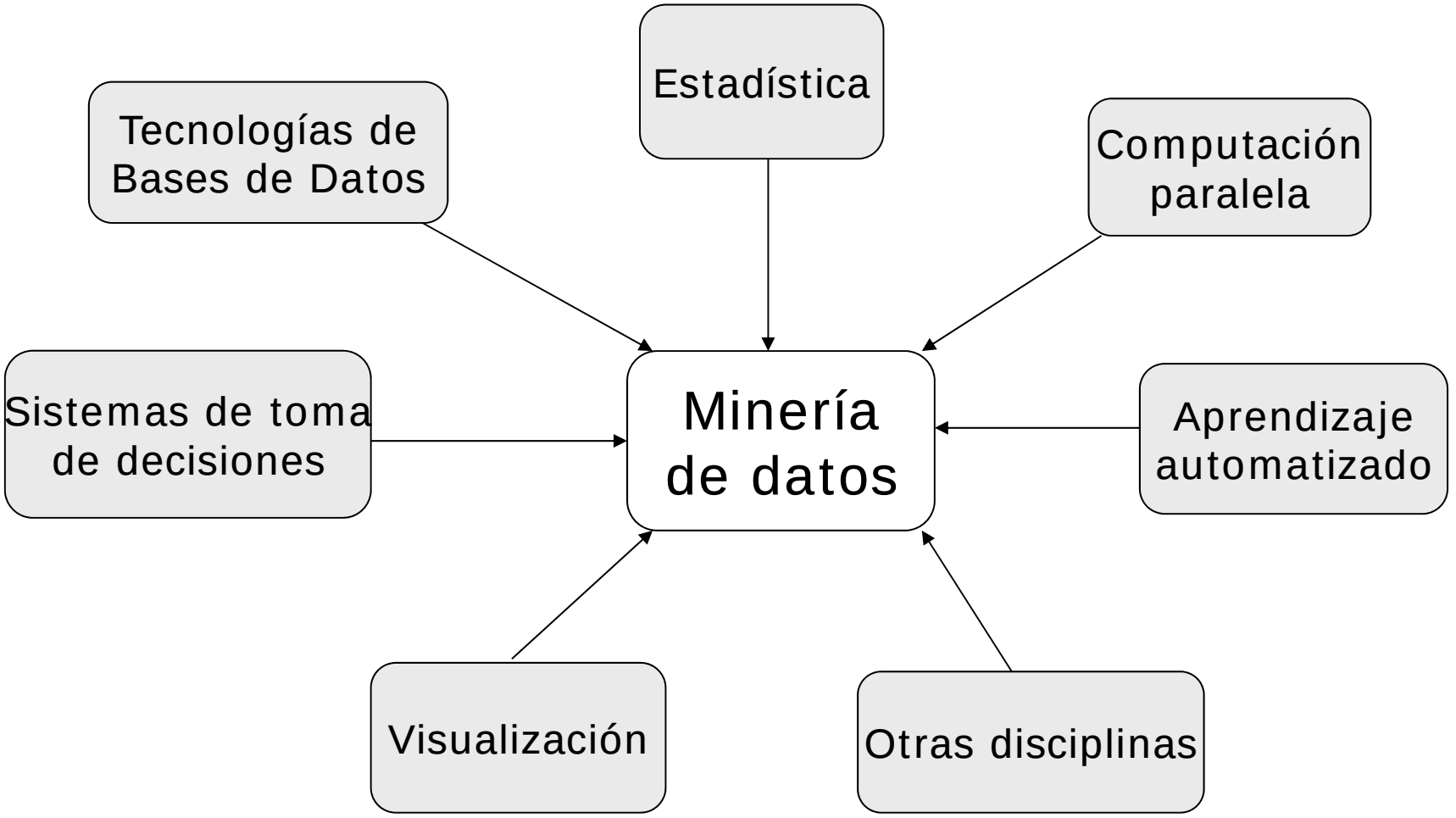
- Difusión: Elaboración de informes para su distribución
- Utilización del nuevo conocimiento de forma independiente
- Incorporación a sistemas ya existentes

  → comprobar con el conocimiento ya utilizado para evitar inconsistencias y posibles conflictos

La monitorización del sistema en acción dará lugar a nuevos casos que realimentarán el ciclo del KDD

Las conclusiones iniciales pueden variar, invalidando el modelo adquirido

Relación con otras disciplinas



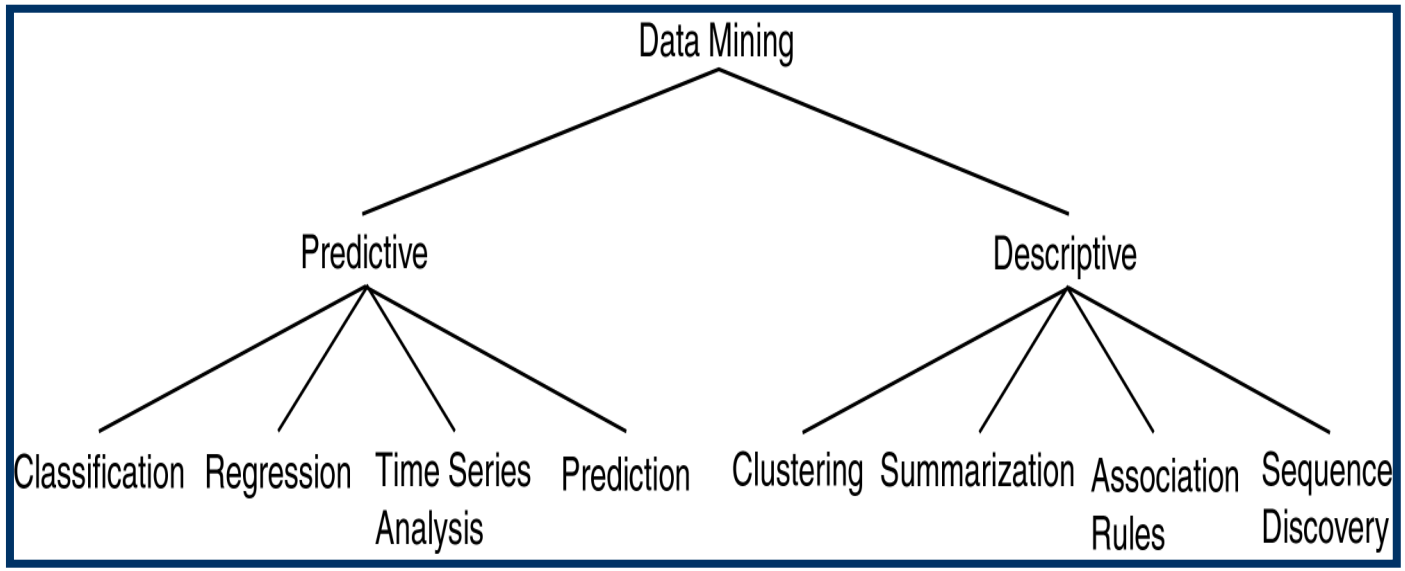
Disciplinas del científico de datos

Ciencia de Datos y Minería de Datos

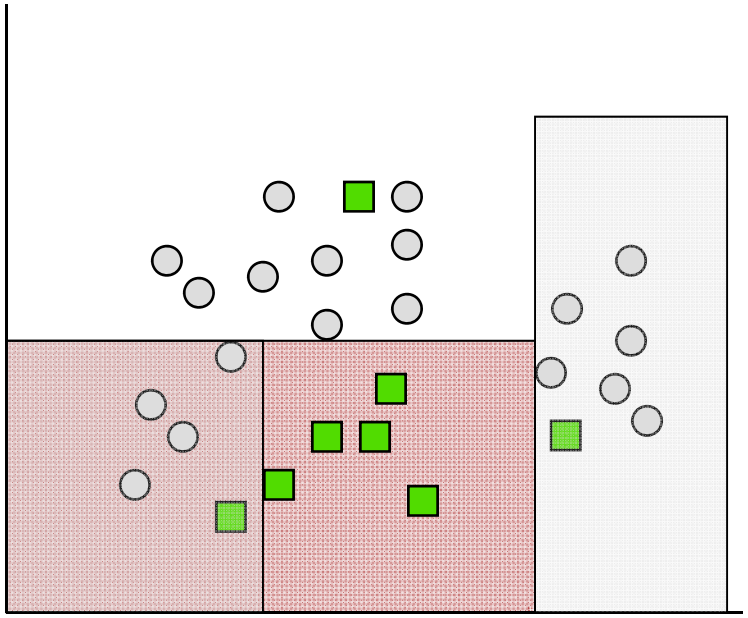
- ¿Qué es la Ciencia de Datos?
- Minería de Datos
- Proceso de Minería de Datos
- **Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación y Otros**
- Minería de Datos: Casos de uso
- Herramientas y Lenguajes en Ciencia de Datos. Repositorio Kaggle
- Comentarios Finales

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



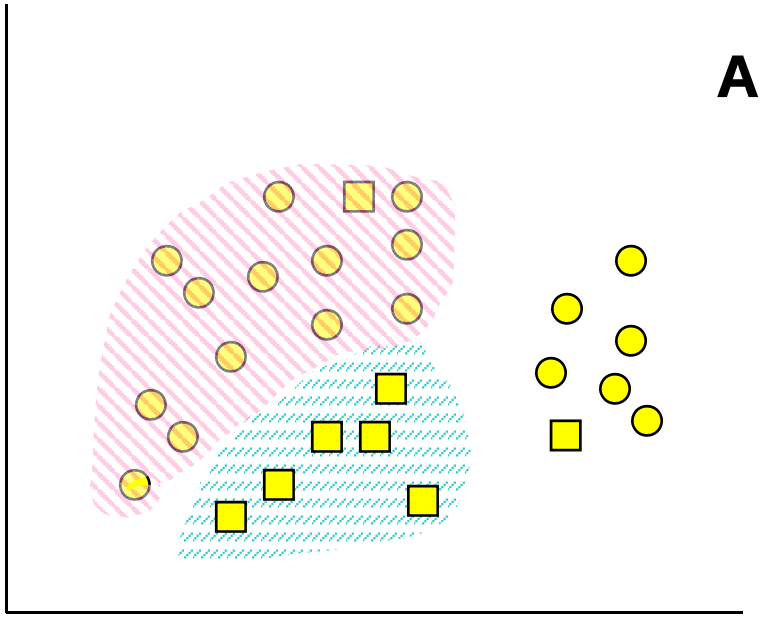
Aprendizaje Supervisado vs No Supervisado



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)

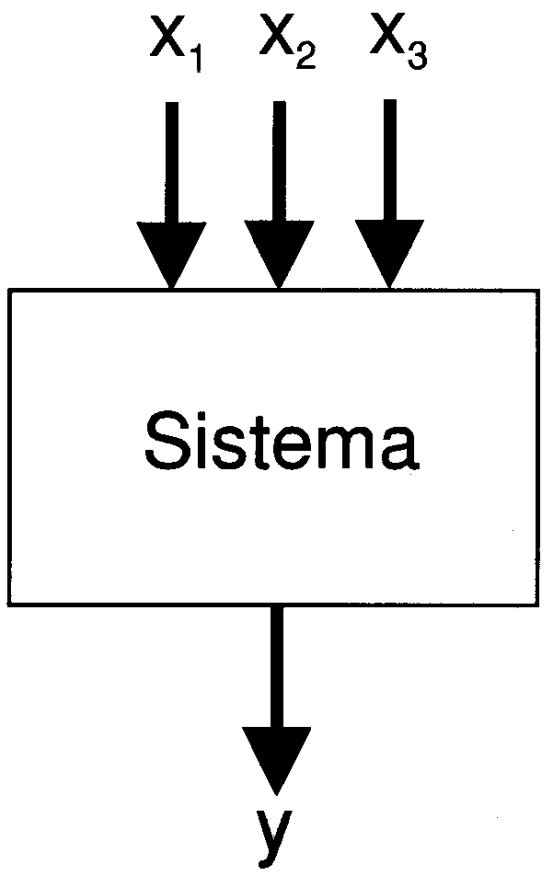
Técnicas de Minería de Datos

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

Regresión

- **Modelado o Predicción**

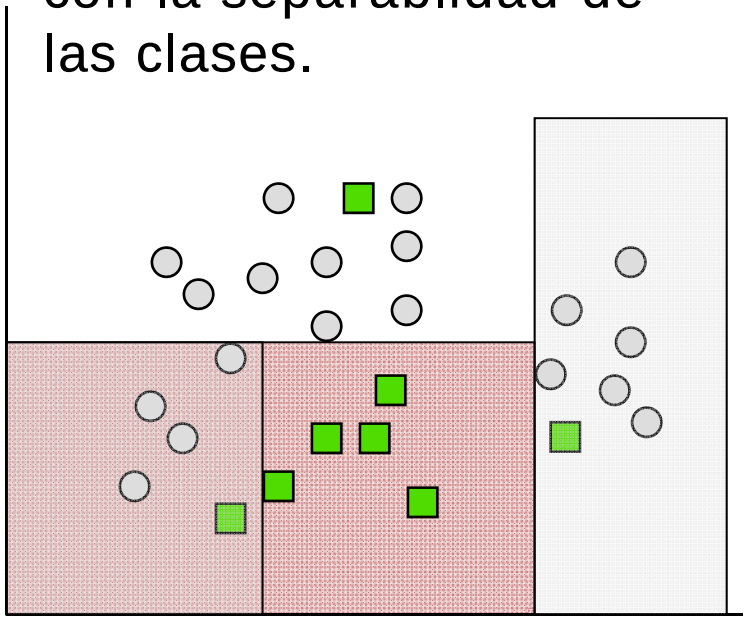
El problema fundamental de la predicción está en modelar la relación entre las variables de estado para obtener el valor de la variable de control.

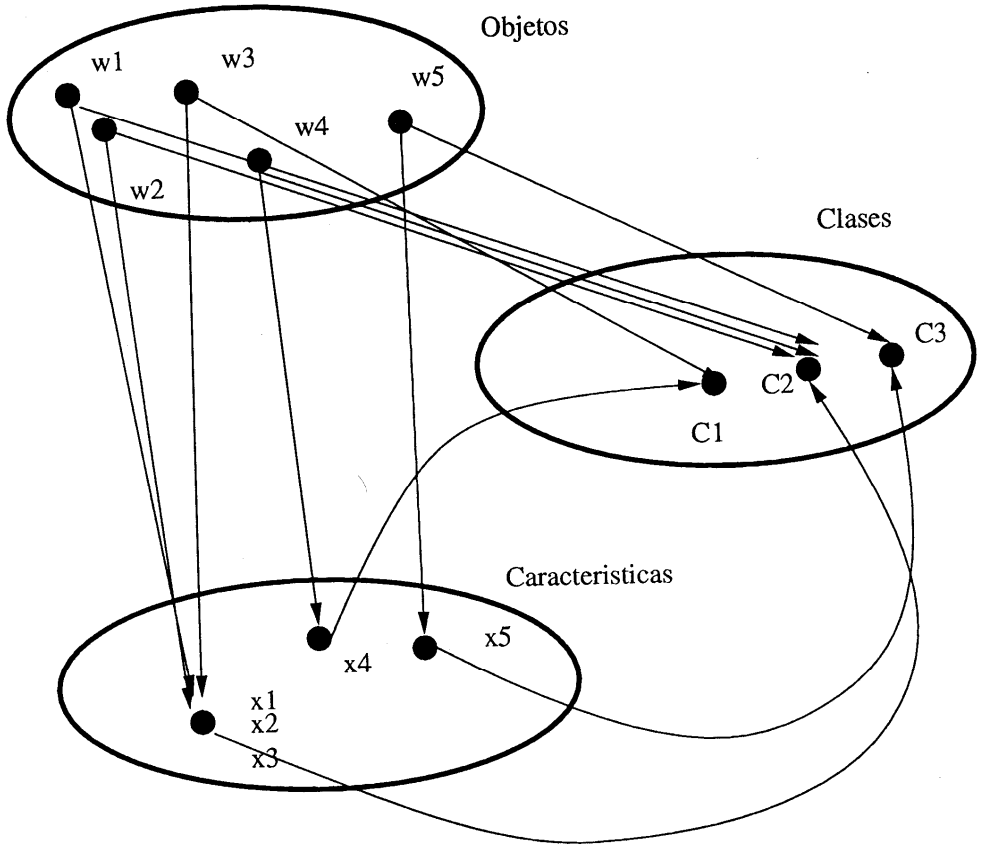


Clasificación

- **Clasificación**

El problema fundamenta
de la clasificación está
directamente relacionad
con la separablidad 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

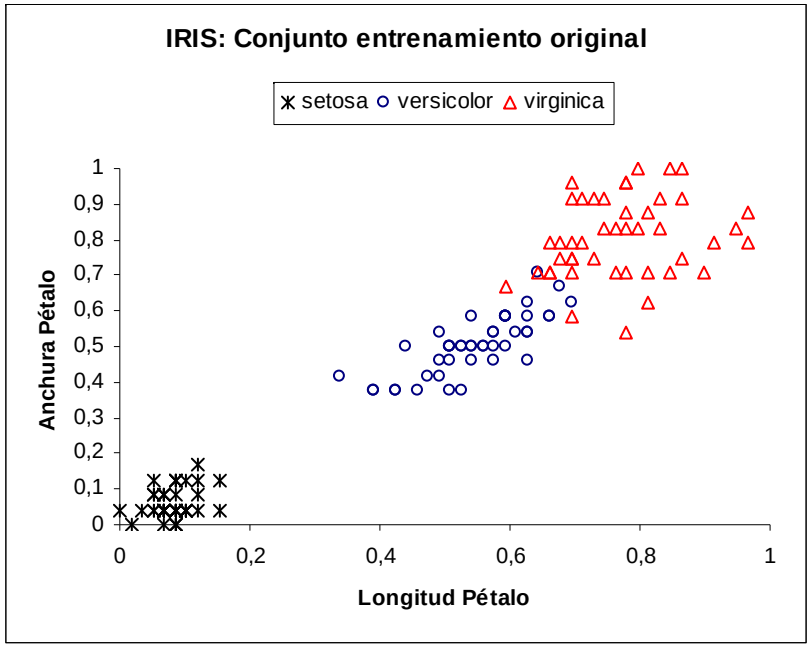
setosa

versicolor

virginica

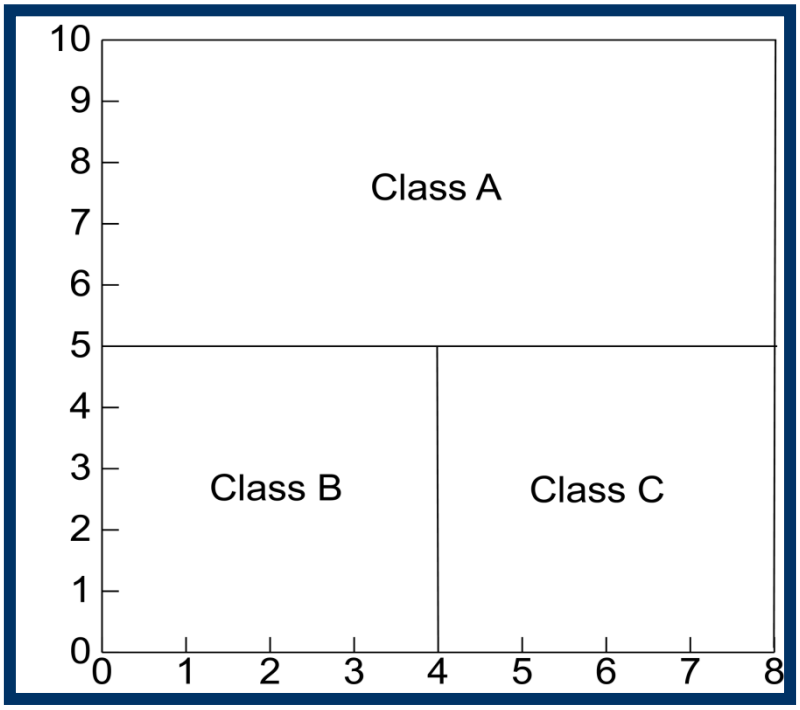
Clasificación. Ejemplo

Ejemplos de conjuntos seleccionados sobre *Iris:*



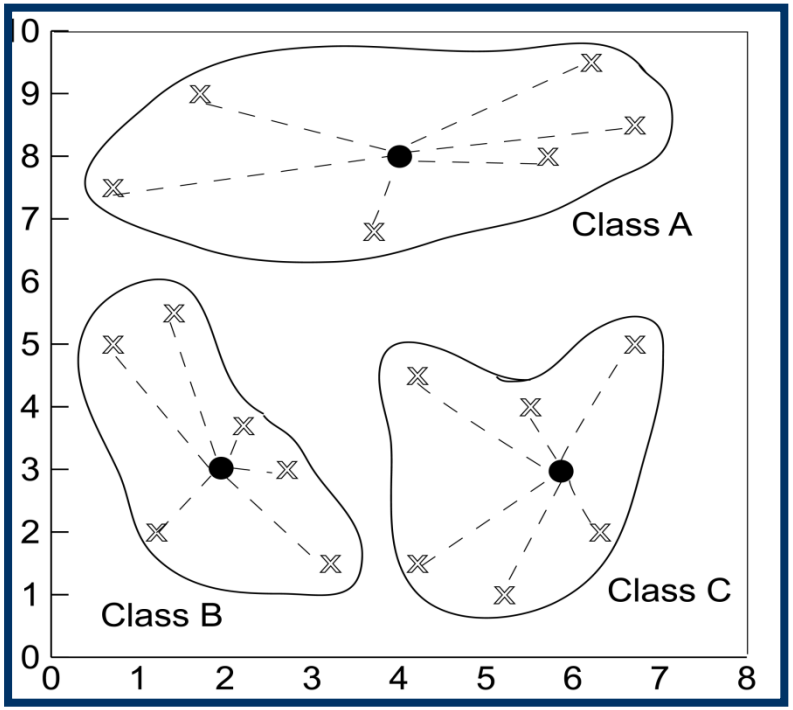
Clasificación. Ejemplo

Clases Definidas



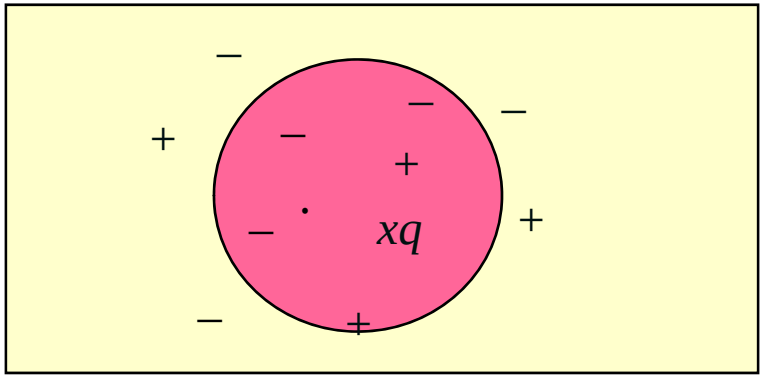
Basado en Particiones

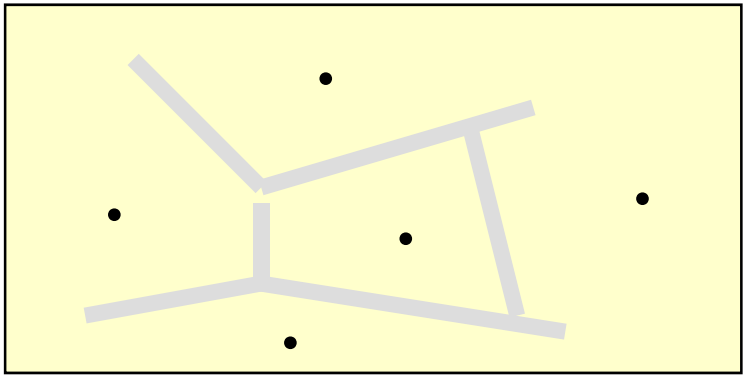
Basado en Distancias



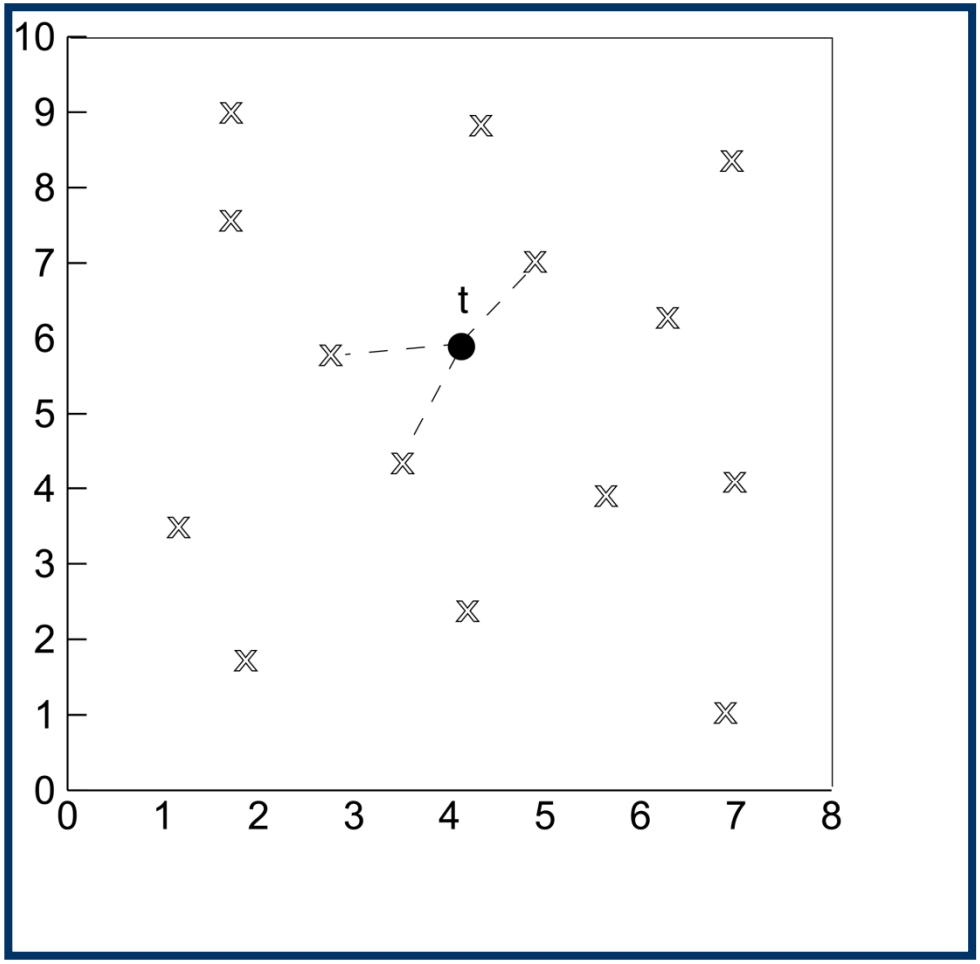
Ejemplo de Clasificador: k-NN

- *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

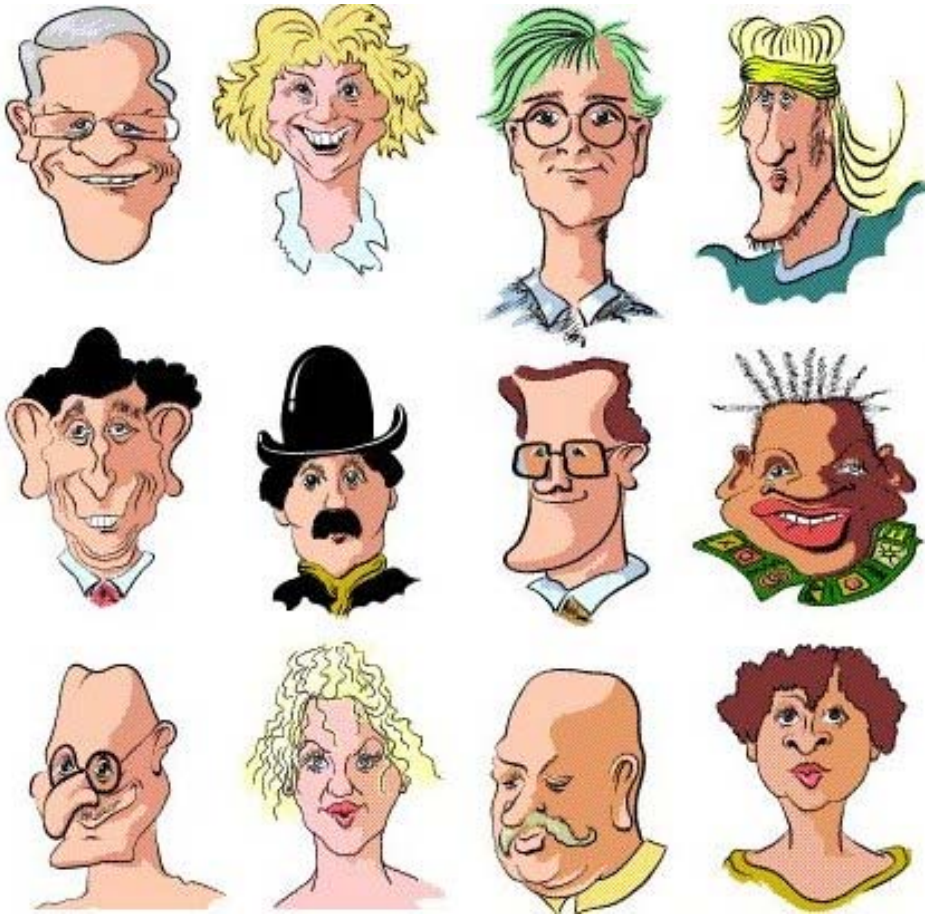


k = 3

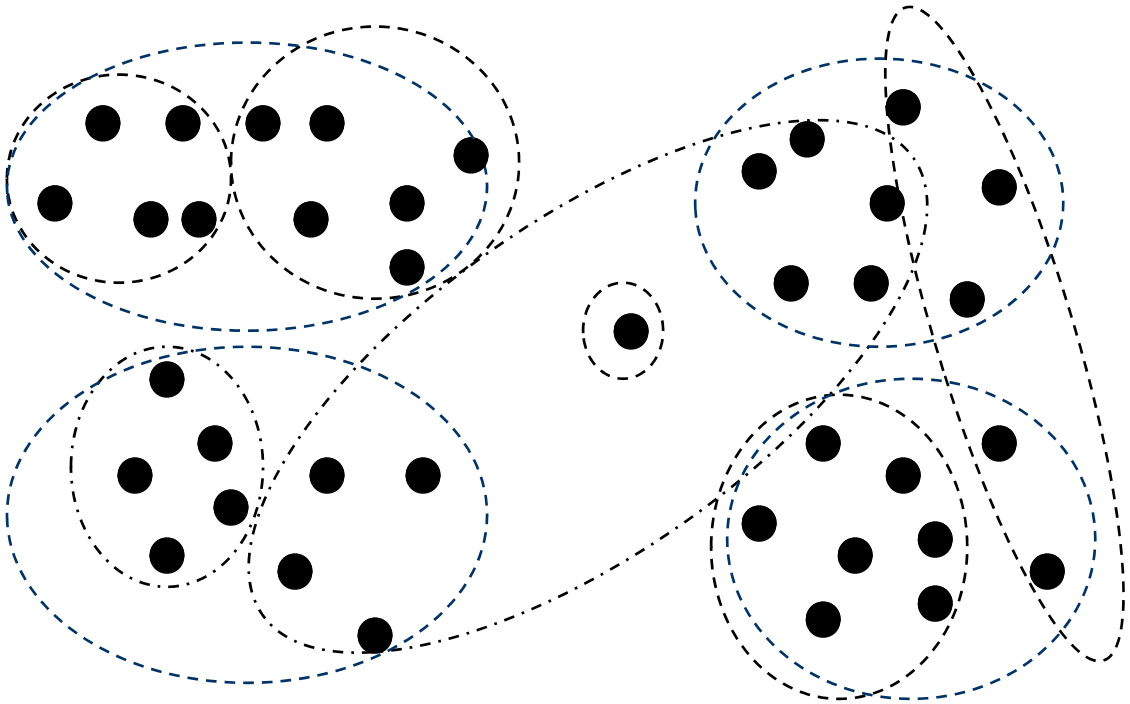
Agrupamiento

Hay problemas en los que deseamos agrupar las instancias creando clusters de similares características

Ej. Segmentación de clientes de una empresa



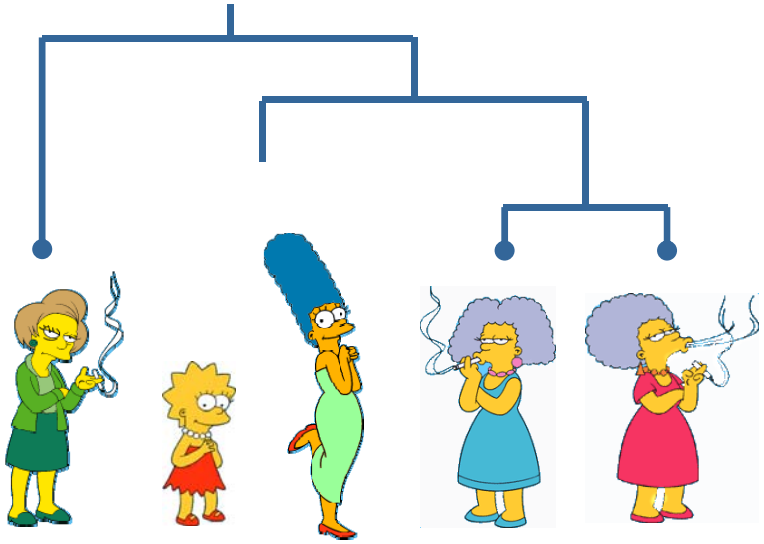
Agrupamiento. Niveles



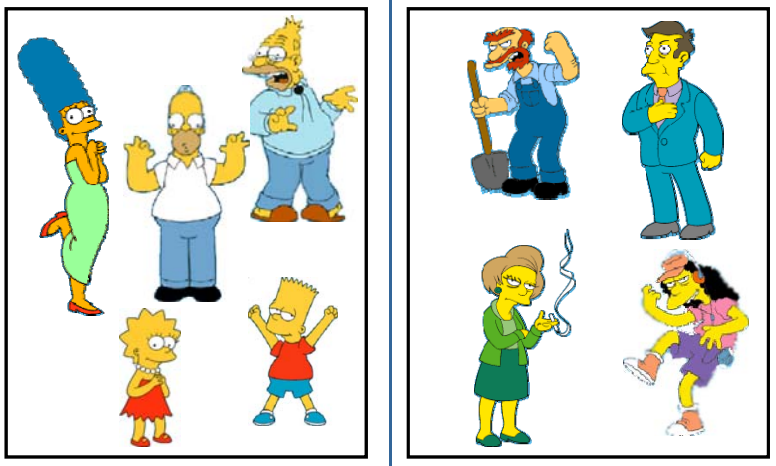
La decisión del número de clusters
es uno de los retos en agrupamiento

Agrupamiento. 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, ...

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

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

¿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?

Descubrimiento de Asociaciones. Ejemplo

Ejemplo: Asociación Cervezas y Pañales

- Los clientes que compran cerveza también compran patatas

  ¡Para eso no es necesario el uso de técnicas de Minería de Datos!

- Los viernes por la tarde, con frecuencia, quienes compran pañales, compran también cerveza.

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

Descubrimiento de asociaciones

Market Basket Analysis

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

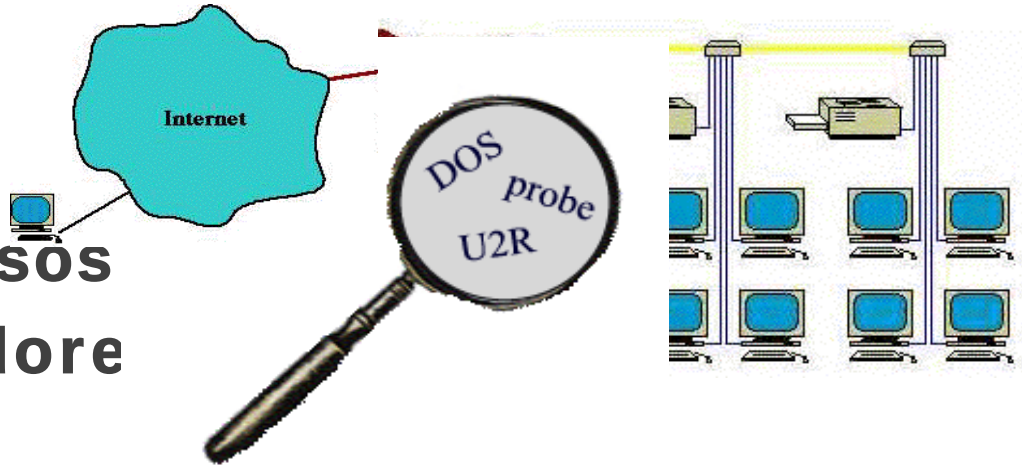
{Diaper, Milk} --> {Beer}

Detección de Desviaciones/Anomalías

Detección de desviaciones significativas de datos normales

- Aplicaciones
  - **Detección de fraude en tarjetas de crédito**

  - **Detección de intrusos en redes de ordenadore**



Ciencia de Datos y 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
- **Minería de Datos: Casos de uso**
- Herramientas y Lenguajes en Ciencia de Datos. Repositorio Kaggle
- Comentarios Finales

Minería de Datos. Casos de estudio

- Procesamiento de préstamos
- Estudio de imágenes
- Planificación de recursos
- Diagnóstico de fallos
- Marketing y ventas
- Bioinformática
- Minería web

Minería de Datos. Casos de estudio

Procesamiento de préstamos (clasificación)

- Entrada: cuestionario de datos personales y financieros
- Problema: ¿se le concede el préstamo?
- Muchas solicitudes
  → estudiadas por ordenador (estadísticos)
- 90% se procesan directamente, pero el 10% están en la duda
  → estudiar por un experto en préstamos
- De los préstamos concedidos en esta franja de duda, ¡el 50% no devuelven el dinero!
- La solución NO es denegar todos los préstamos de esta franja

Minería de Datos. Casos de estudio

Procesamiento de préstamos (clasificación)

- Datos: 1000 ejemplos de casos en la franja completa
- 20 atributos: edad, antigüedad en la dirección actual, tarjetas de crédito, salario, posesiones, historial en el banco,...
- Enfoque: reglas. Las reglas aprendidas clasifican correctamente 2/3 de los casos en un conjunto de prueba (*test*) distinto
- Ventaja adicional: el conocimiento extraído (reglas) sirve al agente para explicar su decisión

Minería de Datos. Casos de estudio

Estudio de imágenes (clasificación)

- Entrada: imágenes de satélite de aguas costeras
- Problema: detección de mareas negras
- Una marea negra suele aparecer en la imagen como una región oscura de tamaño y forma cambiante
- Complejidad: situaciones parecidas pueden ser provocadas por vientos y tormentas
- El estudio de las imágenes es un proceso costoso tanto en tiempo como en dinero (personal muy especializado)

Minería de Datos. Casos de estudio

Estudio de imágenes (clasificación)

- Dado el gran mercado, una empresa decide abordar el problema mediante un producto software
- Problema: trabajar con la imagen directamente es inviable
- Preprocesamiento: de los pixeles a docenas de atributos (extracción de características)
- Atributos: tamaño de la mancha, geometría, intensidad,...
- Problemas encontrados en el desarrollo:
  - Escasez de ejemplos positivos → datos no balanceados
  - Complicado de generalizar, muy dependiente de la zona

Minería de Datos. Casos de estudio

Planificación de recursos (regresión/series temporales)

- Las compañías eléctricas necesitan predicciones de demanda futura
- La predicción con exactitud de un intervalo de carga para cada hora → ahorrar mucho dinero
- Problema: se dispone de un modelo estático de predicción que asume condiciones climáticas normales, el objetivo es ajustar la predicción en función del clima
- Modelo estático: demanda usual en el año, fechas vacacionales, ...

Minería de Datos. Casos de estudio

Planificación de recursos (regresión/series temporales)

- Predicción basada en estudio de días “más similares”
- Datos: La predicción estática, archivos históricos, datos climáticos
- Se genera una base de datos para los 15 años anteriores con atributos como temperatura, humedad, velocidad del viento, nubosidad y la diferencia entre la predicción de consumo estática y el consumo real
- Se añade la diferencia media de los tres días más similares a la predicción del modelo estático
- Se usa regresión lineal como modelo de predicción

Minería de Datos. Casos de estudio

Diagnóstico de fallos (clasificación o detección de anomalías)

- El diagnóstico es el dominio por excelencia de los sistemas expertos
- Conjuntos de reglas elicitados a partir del experto son viables en problemas pequeños, pero no en problemas medianos/grandes
- Problema: realizar diagnóstico de fallos y mantenimiento predictivo en dispositivos electromecánicos como motores y generadores, en una planta química de unos 1000 dispositivos
- Datos: se miden vibraciones en determinados puntos y se realiza un análisis de Fourier
- Objetivo: determinar fallos y realizar mantenimiento predictivo
- Actualmente: se usa un conjunto de reglas diseñadas por el experto

Minería de Datos. Casos de estudio

Diagnóstico de fallos (clasificación)

- Datos: provenientes de diagnósticos realizados por el experto, 600 casos
- Después de depurar se descartan 300
- Se incrementa la dimensión del problema (atributos) con conceptos intermedios (razonamiento causal)
- El conjunto de reglas resultante muestra una gran exactitud, pero no le gusta al experto, porque no está en línea con su forma de actuar
- Tras añadir conocimiento de *background*, el conjunto de reglas es más complicado pero le gusta al experto porque está en consonancia con su mecánica

✓ Las reglas se usan no porque sean buenas, si no porque le gustan al experto

Minería de Datos. Casos de estudio

Marketing y ventas (asociaciones)

- Empresa de supermercados con más de 1000 tiendas
- Vende aproximadamente 20.000 artículos distintos
- Los datos de las ventas se almacenan (lector de código de barras + Pc)
- Todas las transacciones + datos adicionales de cada tienda se almacenan y actualizan diariamente en una sede central
- Dispone de una tarjeta de cliente frecuente

Se generan informes diarios, semanales y mensuales, mostrando para cada artículo y cada marca: ventas, inventario, ofertas, precios, ...

Minería de Datos. Casos de estudio

Marketing y ventas (asociaciones)

¿Dónde se deberían colocar los detergentes para maximizar las ventas?

¿Se compra limpia cristales si se compra simultáneamente zumo de naranja y refrescos?

¿Cómo afecta la demografía del entorno a lo que compran los clientes?

Minería de Datos. Casos de estudio

Compras a través de internet (asociaciones)

amazon.com

- Una persona compra un libro (producto) en Amazon.com
- Tarea: Recomendar otros libros (productos) que esa persona pueda comprar
- Amazon hace *clustering* basándose en las compras de libros: clientes que compran “Advances in Knowledge Discovery and Data Mining”, también compran “Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations”
- El programa de recomendación es bastante éxitoso

Minería de Datos. Casos de estudio

Genomic Microarrays (Clasificación)

Dado un conjunto de datos de microarrays para un número de ejemplos (pacientes), podemos

- ¿Diagnosticar de forma precisa la enfermedad?
- ¿Predecir resultados para un tratamiento dado?
- ¿Recomendar el mejor tratamiento?

Minería de Datos. Casos de estudio

Descubrimiento de secuencias en páginas web

- Objetivo:

  Determinar patrones secuenciales en los datos

- Estos patrones son asociaciones en los datos pero con una relación en el tiempo

- Ejemplo: Descubrimiento de secuencias en el análisis de un web log para determinar como acceden los usuarios a determinadas páginas

Minería de Datos

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

Ciencia de Datos y 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
- Minería de Datos: Casos de uso
- **Herramientas y Lenguajes en Ciencia de Datos. Repositorio Kaggle**
- Comentarios Finales

Herramientas, Lenguajes, Kaggle

Una web sobre el software libre para Ciencia de Datos ...

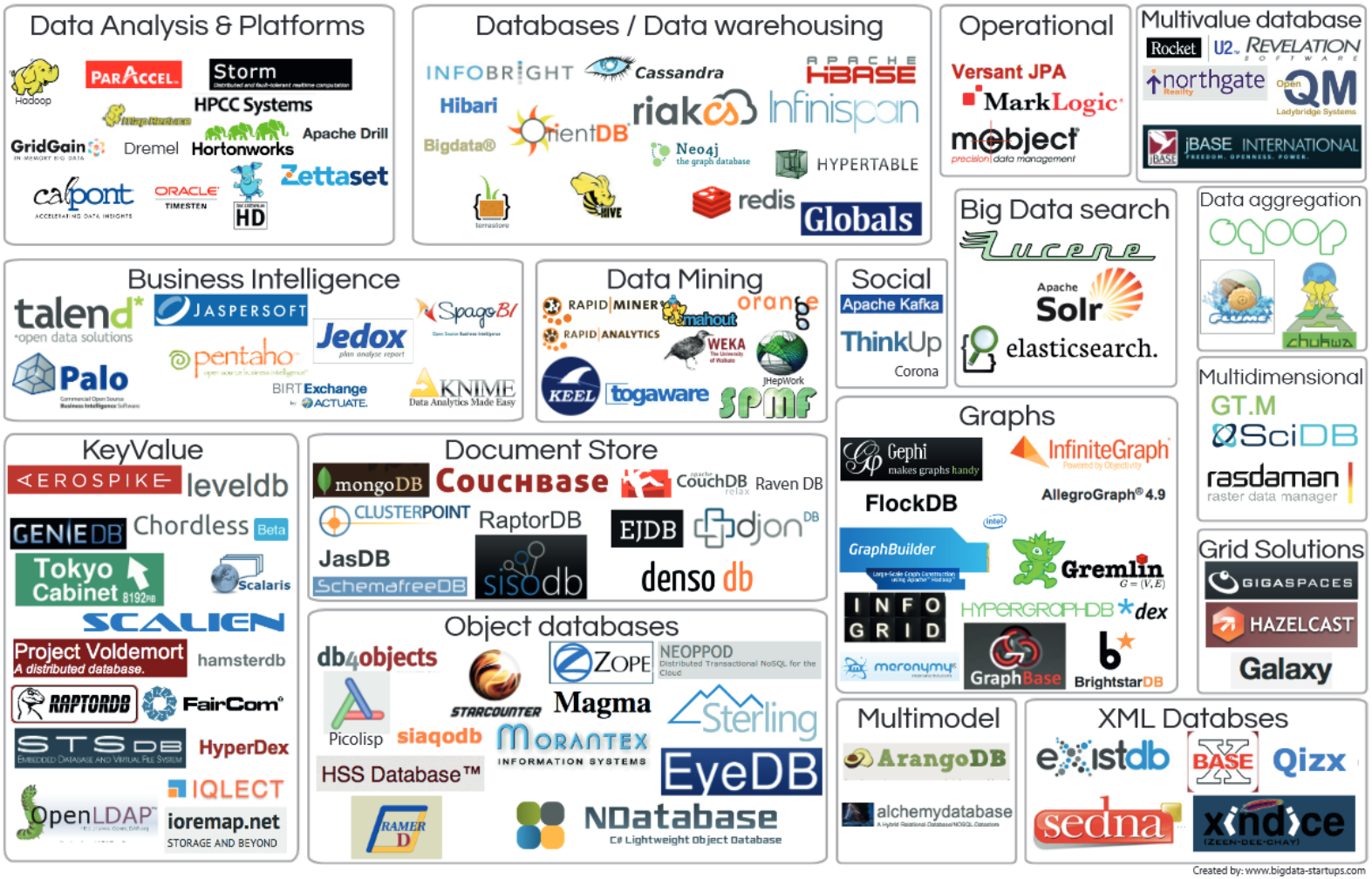
Software (open source tools)

http://www.bigdata-startups.com/open-source-tools/

Herramientas, Lenguajes, Kaggle

Una web sobre el software libre para Ciencia de Datos ...

http://www.bigdata-startups.com/open-source-tools/



Herramientas, Lenguajes, Kaggle

Generation	1ª Generación	2ª Generación
Ejemplos	KNIME, SAS, R, Weka, SPSS, KEEL	Mahout, Pentaho, Cascading
Escalabilidad	Vertical	Horizontal (over Hadoop)
Algoritmos disponibles	Huge collection of algorithms	Small subset: sequential logistic regression, linear SVMs, Stochastic Gradient Descendent, k-means clustering, Random forest, etc.
Algoritmos No disponibles	Practically nothing	Vast no.: Kernel SVMs, Multivariate Logistic Regression, Conjugate Gradient Descendent, ALS, etc.
Tolerancia a Fallos	Single point of failure	Most tools are FT, as they are built on top of Hadoop

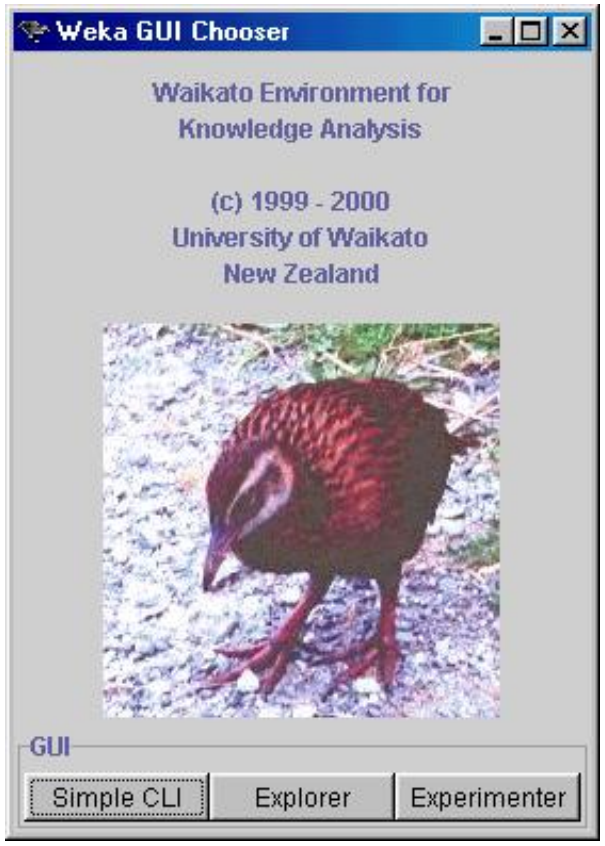
Herramientas, Lenguajes, Kaggle

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

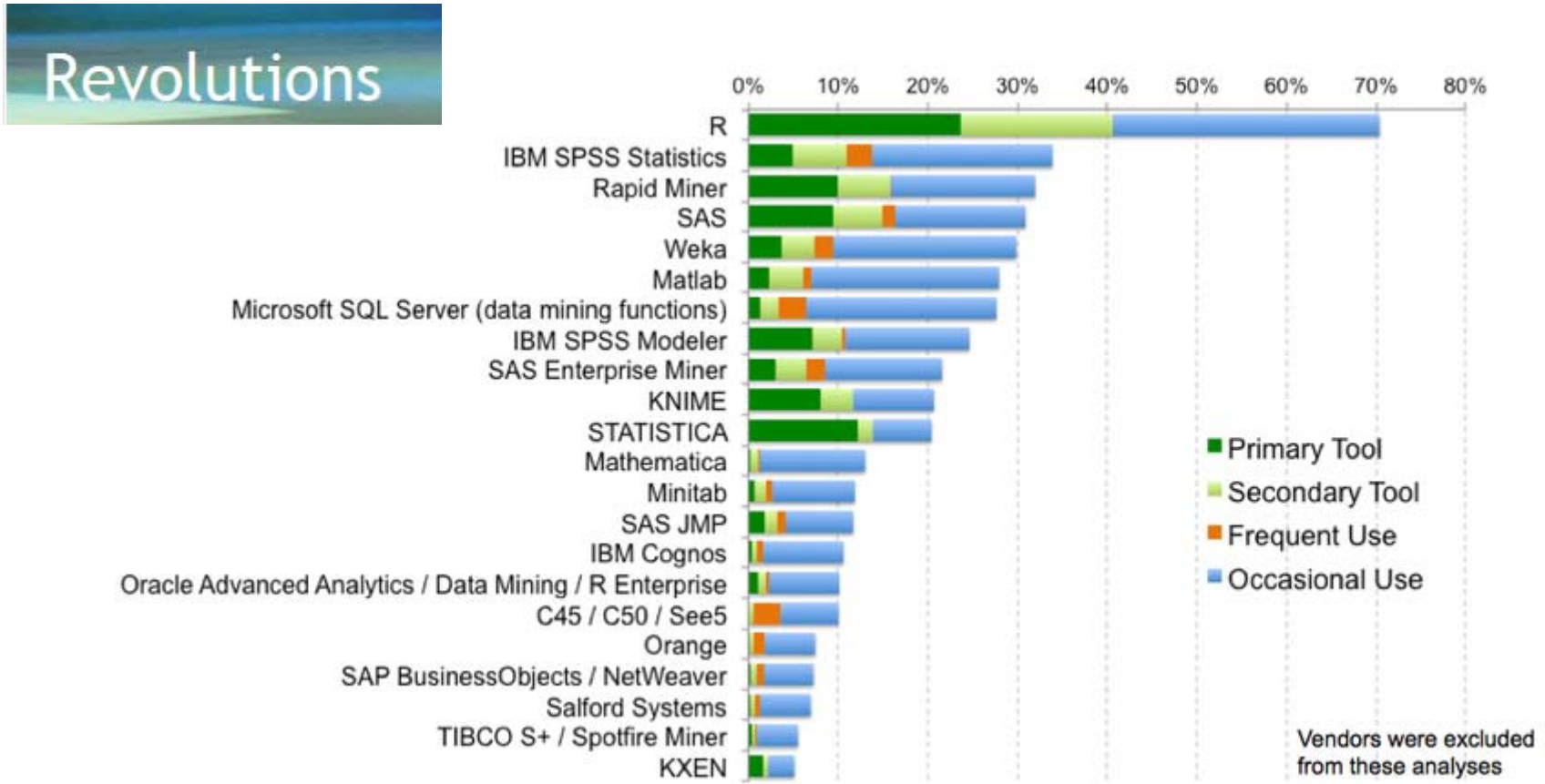
KEEL

- University of Granada
- Machine learning software in Java implementation

http://www.keel.es/

Herramientas, Lenguajes, Kaggle

Sobre herramientas de minería de datos

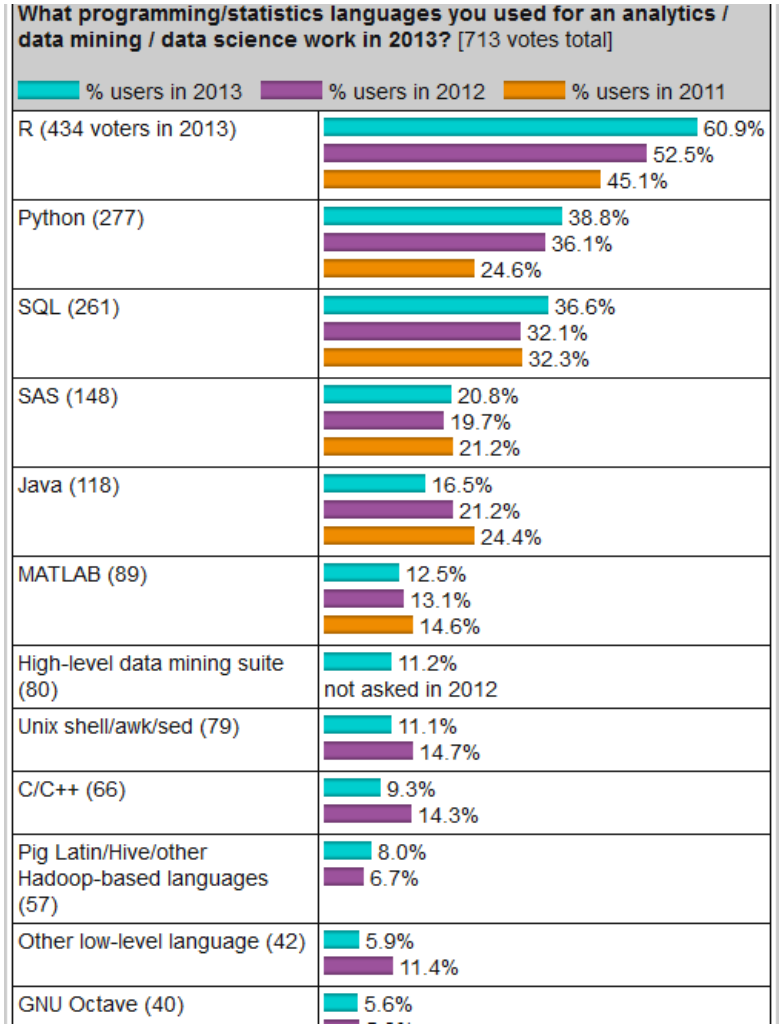


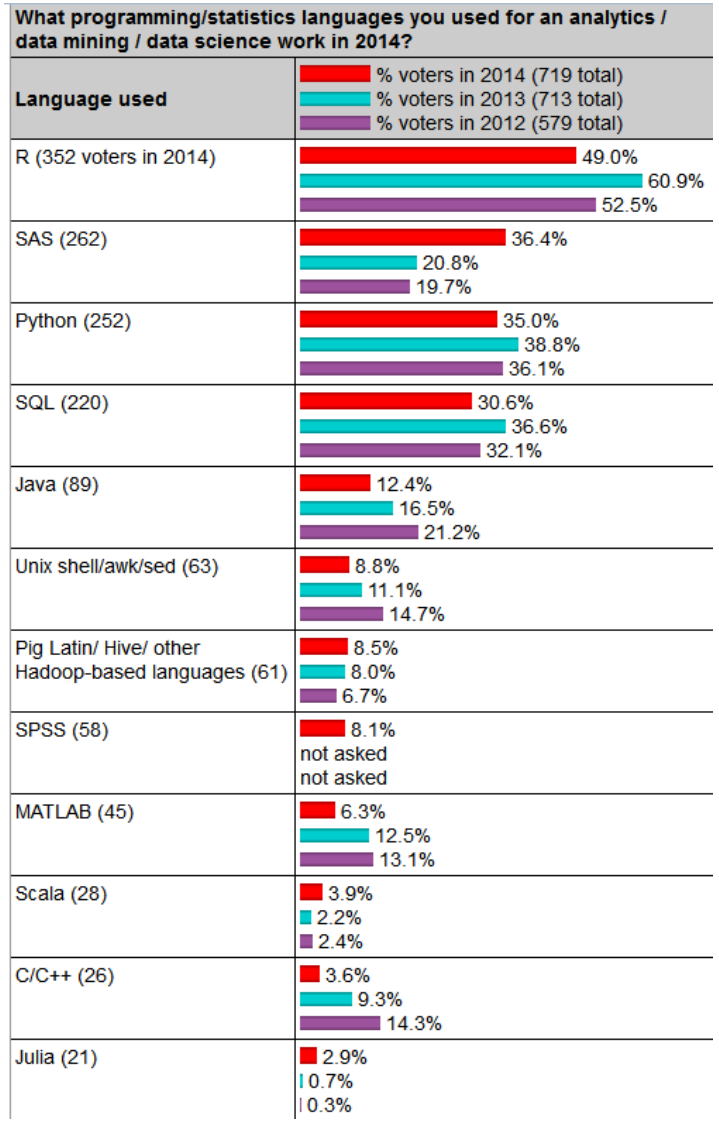
http://blog.revolutionanalytics.com/2013/10/r-usage-skyrocketing-rexer-poll.html

Herramientas, Lenguajes, Kaggle

Sobre los lenguajes de programación (R, Phyton, …).

Lenguajes a usar para Data Science





Herramientas, Lenguajes, Kaggle

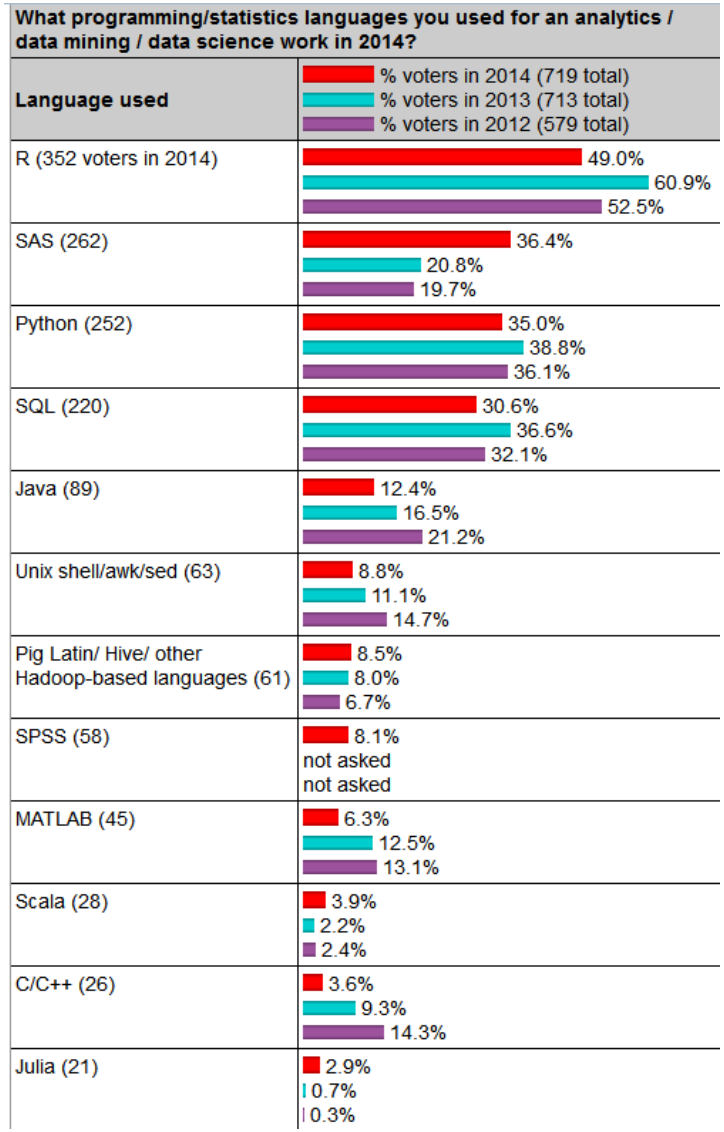
Sobre los lenguajes de programación (R, Phyton, …).

Consolidation among top 4 languages: R, SAS, Python, and SQL, and decline in usage of less popular languages for data mining: Java, Unix shell, MATLAB, C/C++, Perl, Octave, Ruby, Lisp, F.

Languages with the highest growth in 2014 were
Julia, 316% growth, from 0.7% share in 2013 to 2.9% in 2014
SAS, 76% growth, from 20.8% in 2013 to 36.4% in 2014
Scala, 74% growth, from 2.2% in 2013 to 3.9% in 2014

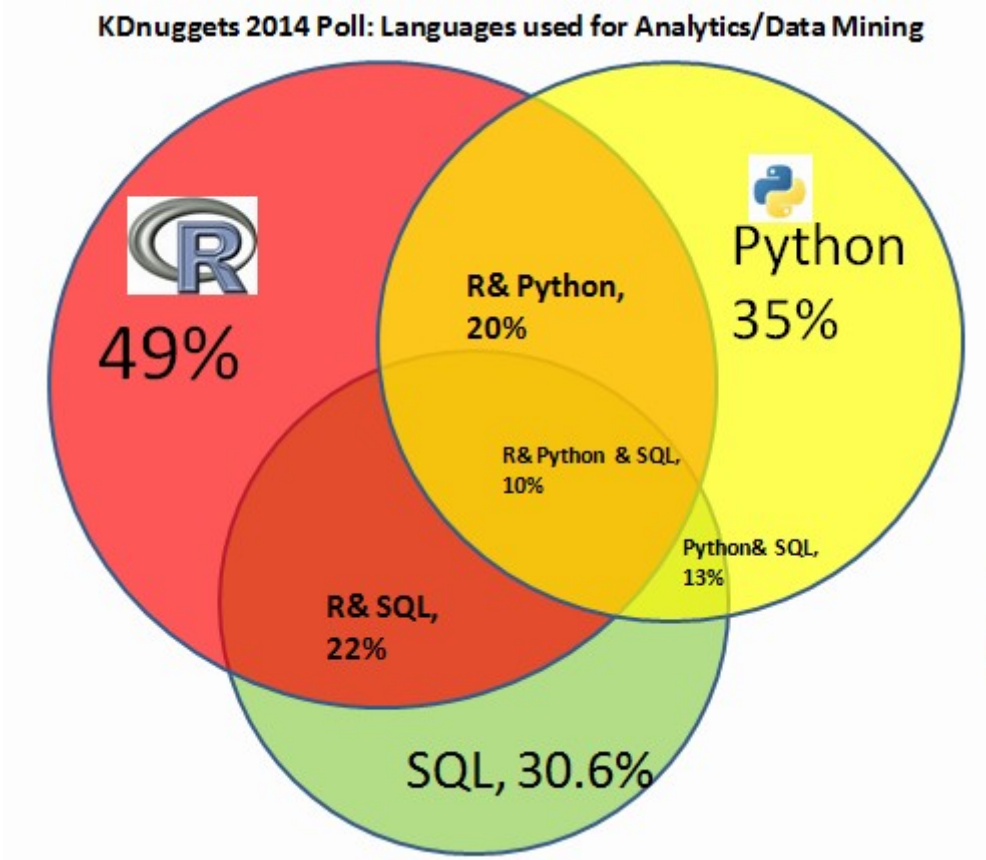
By Gregory Piatetsky, Aug 18, 2014.

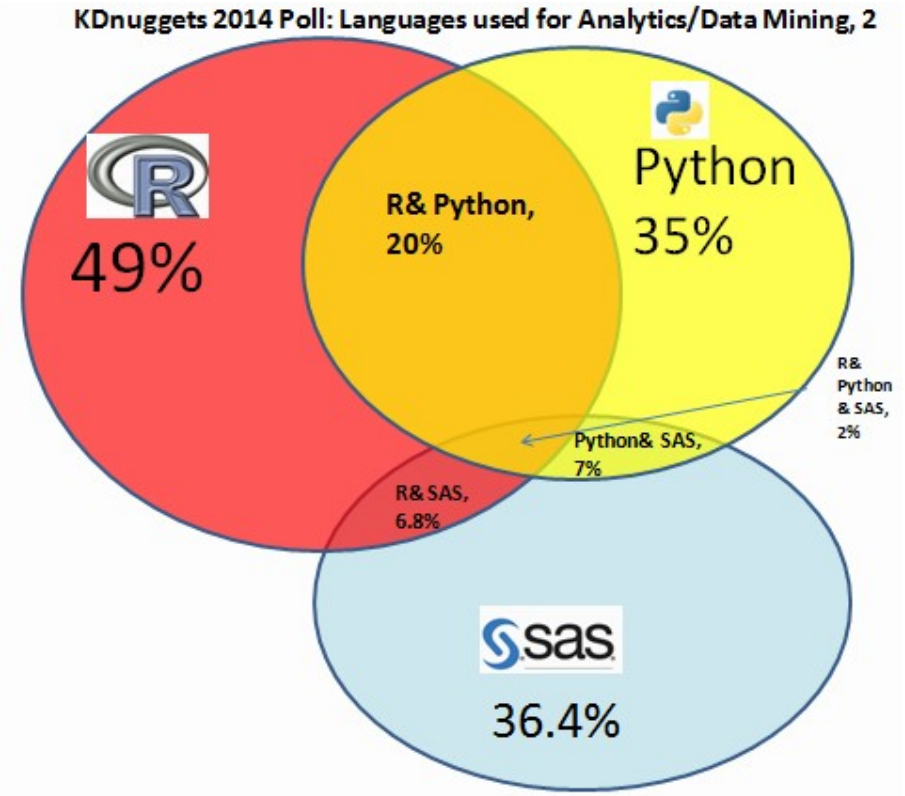
http://www.kdnuggets.com/polls/2014/languages-analytics-data-mining-data-science.html



Herramientas, Lenguajes, Kaggle

Sobre los lenguajes de programación (R, Phyton, …).





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

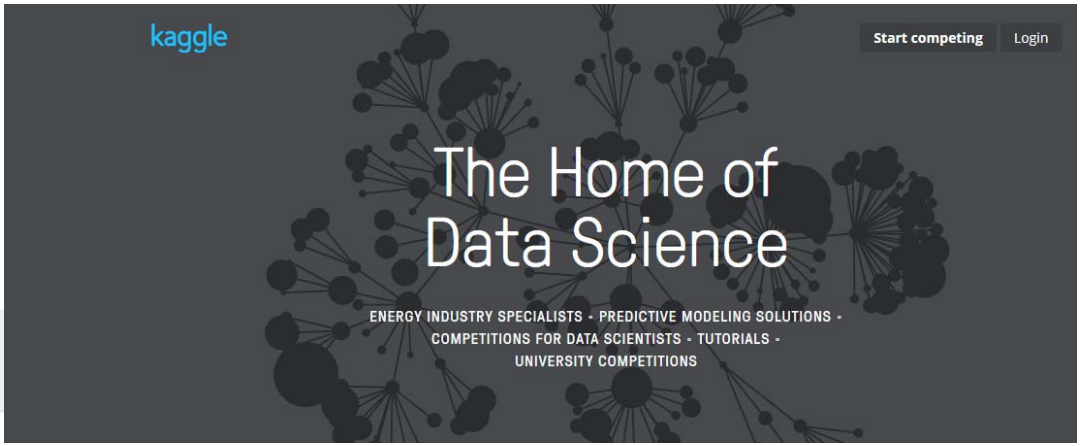
Herramientas, Lenguajes, Kaggle

… 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 oppurtunity to apply machine learning to real problems.

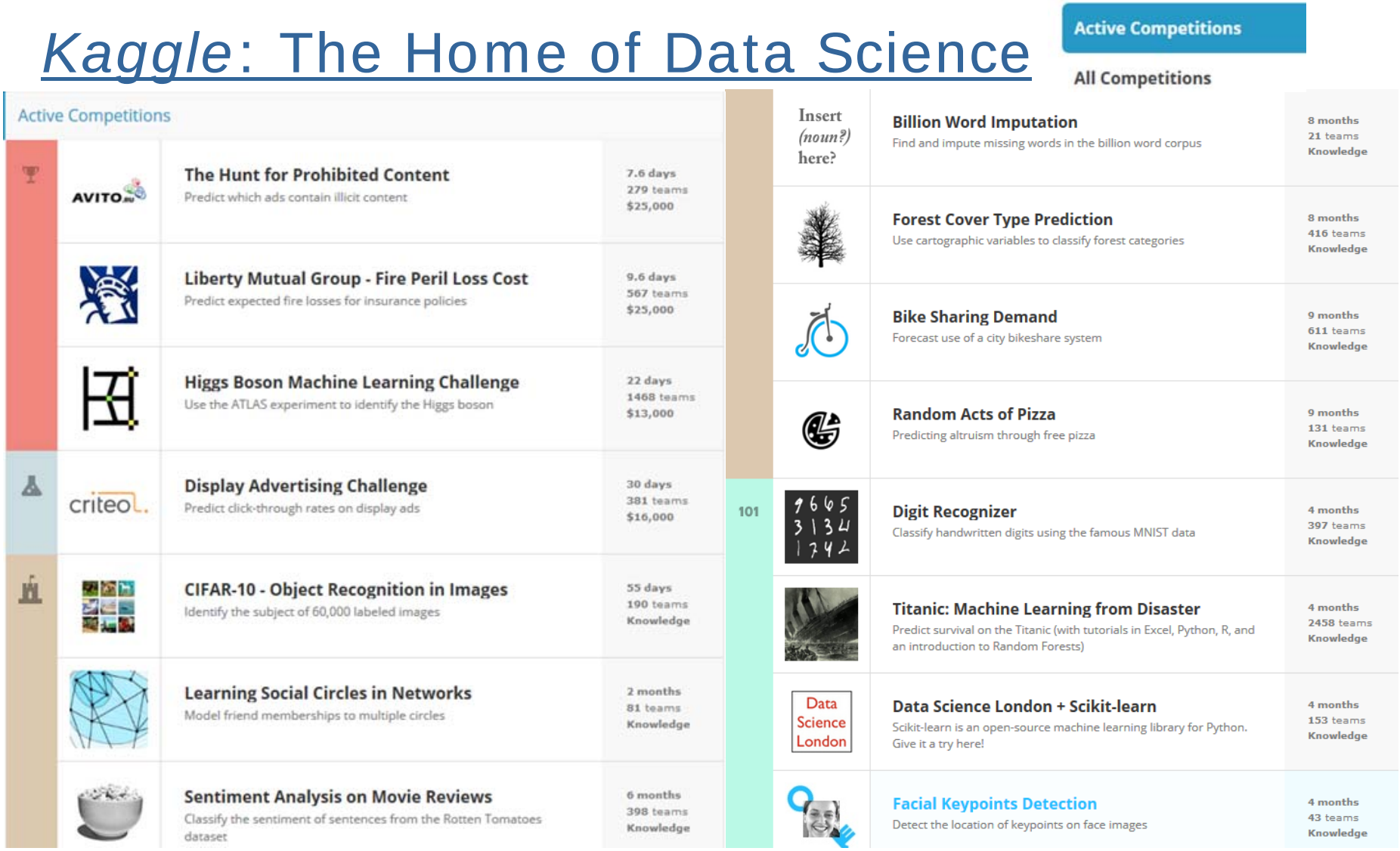
Learn about hosting

Berkeley · 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



Herramientas, Lenguajes, Kaggle

… y un buen enlace para comenzar a practicar, KAGGLE

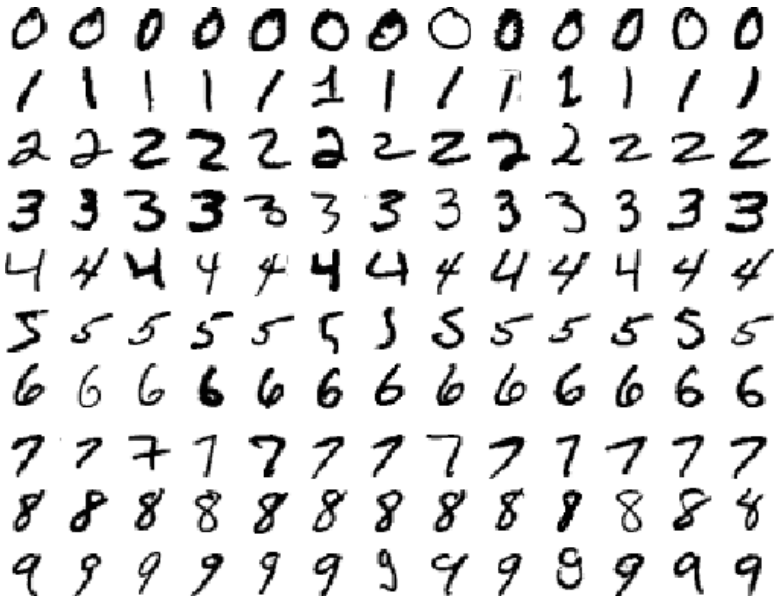
Kaggle: The Home of Data Science

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.

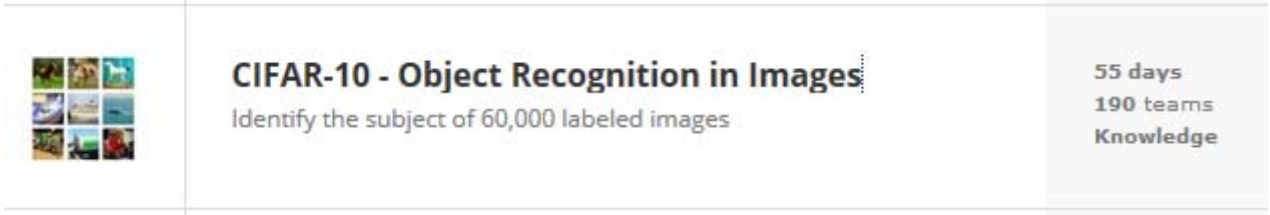
See someone using multi

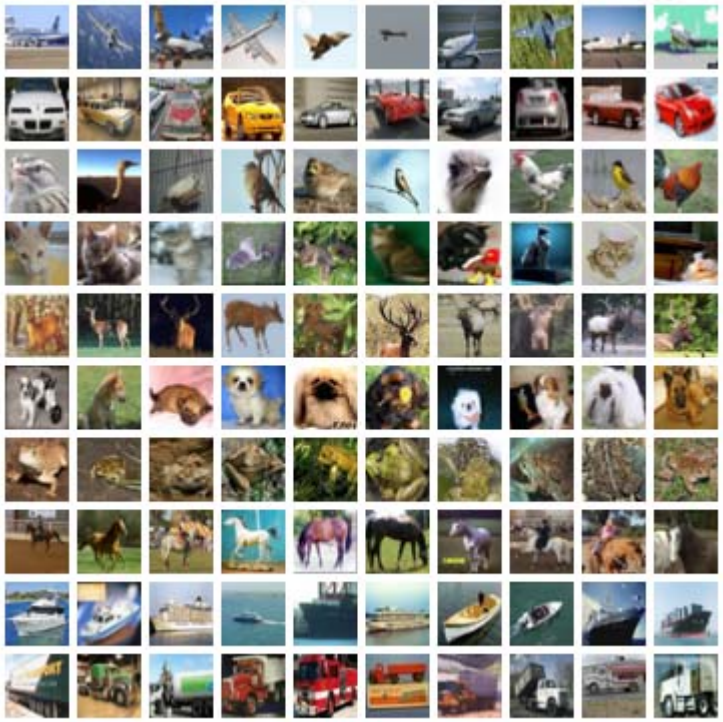
#	Δ1w	Team Name	Score	Entries	Last Submission UTC (Best – Last Submission)
1	–	teppei	1.00000	1	Tue, 08 Jul 2014 15:44:55
2	–	never	1.00000	1	Wed, 16 Jul 2014 13:32:08
3	–	Yonghong	1.00000	2	Tue, 22 Jul 2014 13:41:07
4	–	Shicai Yang	1.00000	3	Wed, 23 Jul 2014 01:52:53
5	–	Aviad_Abigail	0.99757	4	Mon, 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

Herramientas, Lenguajes, Kaggle

… y un buen enlace para comenzar a practicar, KAGGLE

Kaggle: The Home of Data Science





60,000 32x32 color images containing one of 10 object classes, with 6000 images per class.

Dashboard

Leaderboard - CIFAR-10 - Object Recognition in Images

This leaderboard is calculated on all of the test data.

See someone using multiple

#	Δ1w	Team Name	Score	Entries	Last Submission UTC (Best – Last Submission)
1	–	Raillou	0.92610	7	Wed, 16 Apr 2014 14:45:45
2	–	DeepCNet	0.91820	1	Mon, 20 Jan 2014 12:12:53
3	↑3	git.io/nagadomi-cifar10	0.91730	7	Thu, 21 Aug 2014 21:40:00
4	↓1	Dmitriy Anisimov	0.90670	2	Thu, 17 Apr 2014 09:18:53
5	↓1	fastml.com/dropconnect	0.90660	4	Sun, 22 Dec 2013 15:08:58
6	↓1	Yan	0.90530	13	Fri, 29 Nov 2013 05:33:29

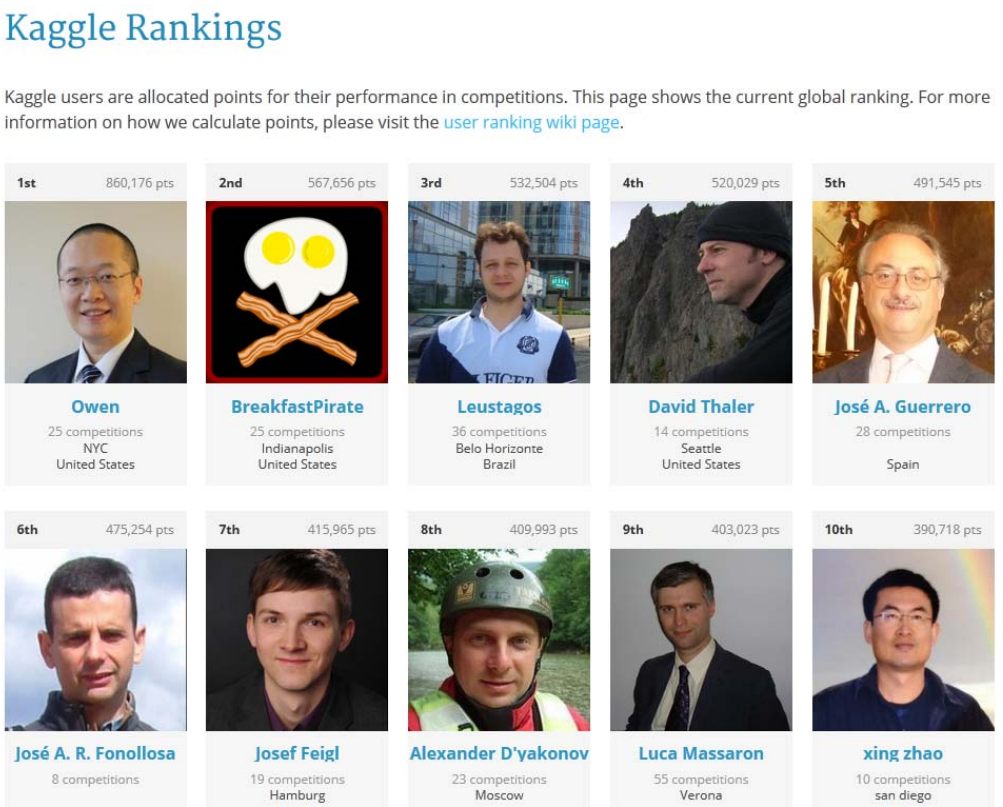
Herramientas, Lenguajes, Kaggle

… y un buen enlace para comenzar a practicar, KAGGEL

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.

Comunidad Kaggle



Ciencia de Datos y Minería de Datos

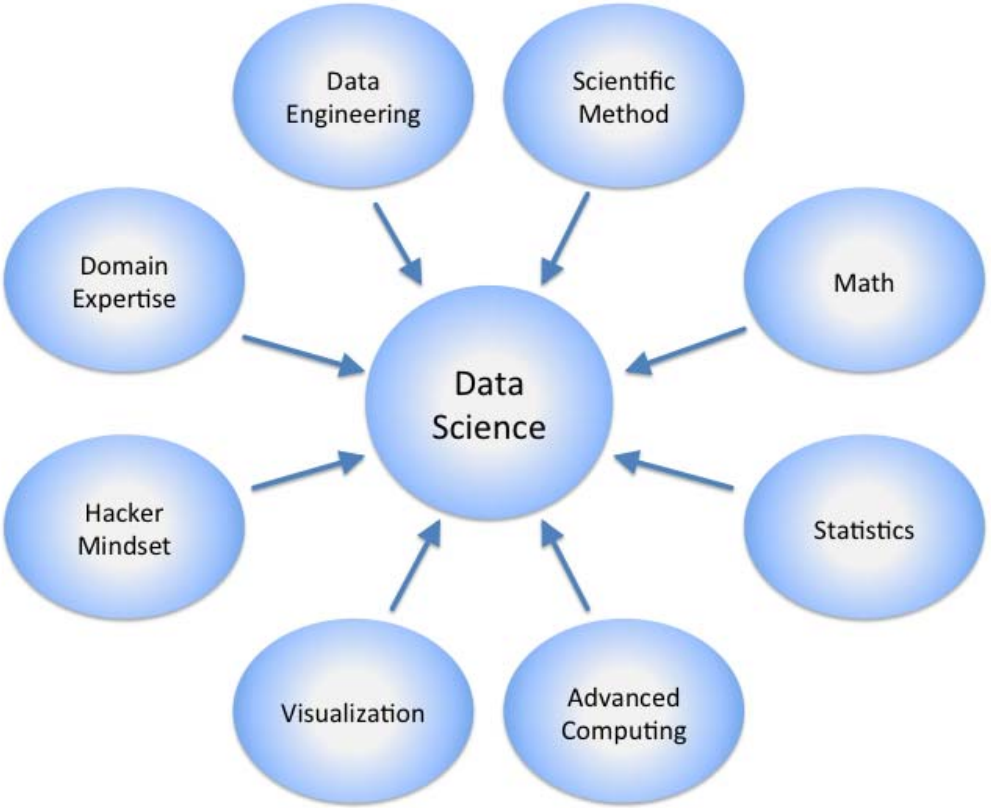
- ¿Qué es la Ciencia de Datos?
- Minería de Datos
- Proceso de Minería de Datos
- Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación y Otros
- Minería de Datos: Casos de uso
- Herramientas y Lenguajes en Ciencia de Datos. Repositorio Kaggle
- **Comentarios Finales**

Comentarios Finales

- **Ciencia de datos:** Es el ámbito de conocimiento que engloba **las habilidades asociados al procesamiento 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, …

Comentarios Finales

(Data Science, Business Analytics, Data Analytics)
Análisis de Datos en un contexto amplio



Business Analytics

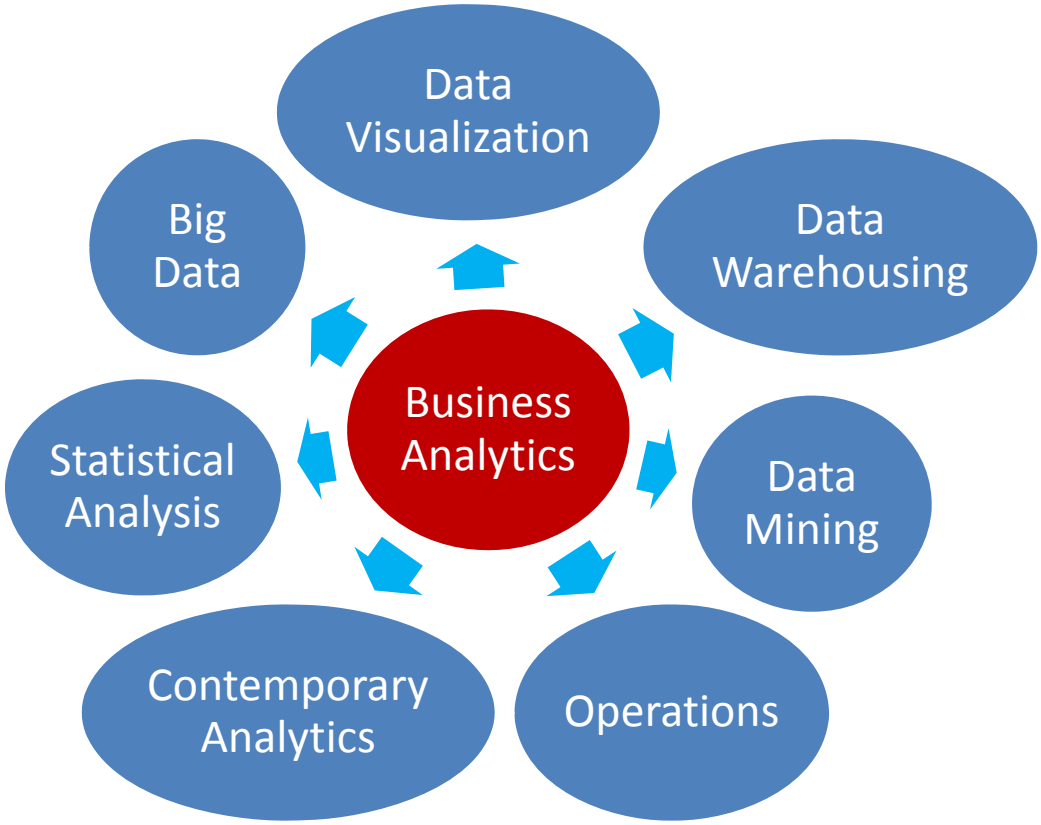
Data Science

=

Business Analytics

=

Data Analytics



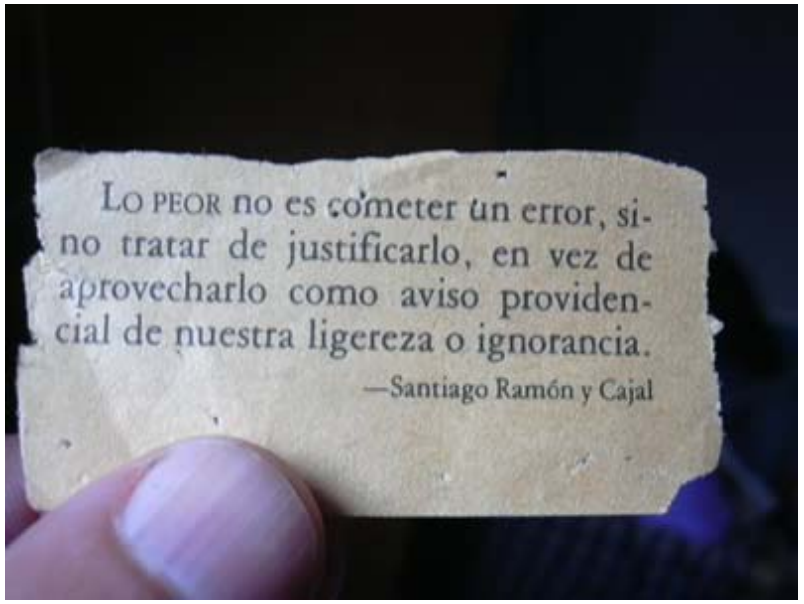
Data mining: Data Preprocessing, Supervised learning, unsupervised learning, forecasting
Contemporary Analytics: text mining, network analytics, social analytics, customer analytics, web analytics, risk analytics, information retrieval and recommendations
Statistical Analysis: Estimation and inference; and regression models
Operations: Simulation and optimization

Comentarios Finales

Hay que obtener conocimiento útil

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

Surge como profesión el "Científico de Datos"

Científico de Datos

Oportunidad profesional: En 2015, Gartner predice que 4,4 millones de empleos serán creados en torno a big data. (Gartner, 2013)

Fuente: http://www.gartner.com/technology/topics/big-data.jsp

Comentarios Finales

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

Oportunidades en Big Data

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*.


EL PAÍS

PORTADA INTERNACIONAL

ECONOMÍA

ECONOMÍA EMPRESAS MERCADOS BOLSA FINANZAS PERSONALES VIVIENDA

ESTÁ PASANDO Multa a la banca Revuelo en Hacienda Eléctricas y renovables

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

Comentarios Finales

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

Oportunidades en Big Data (en España)

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 Toledo.

“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.

CAMPUS ANTONIO MACHADO DE BAEZA

CURSOS VERANO UNIA 2014

Del 25 al 28 de agosto

CURSO / 3476

APROXIMACIÓN PRÁCTICA A LA CIENCIA DE DATOS Y BIG DATA: HERRAMIENTAS KMINE, R, HADOOP Y MAHOUT

Contenido

Lunes 25: Minería de Datos. Herramienta KNIME

Martes 26: Minería de Datos, Visualización y Datos Temporales en el Lenguaje R

Miércoles 27: Big Data. Plataforma Hadoop y Librería Mahout.

Jueves 25: Mahout. Kaggle (Comunidad, repositorio, competiciones …)

Comentarios Finales

Para terminar, un video de la UMUC sobre Big Data y Data Analytics