Ingeniería del Conocimiento CIF-8458

Introducción al KDD (Knowledge Discovery in Databases)

Carlos Valle Vidal

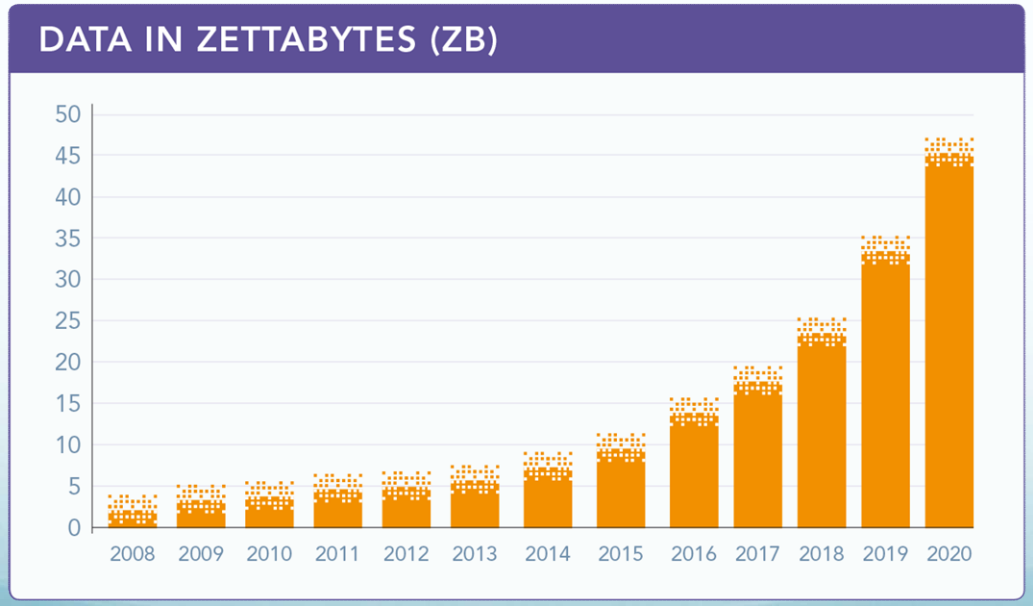
Segundo Semestre 2023

Abundancia de datos

- Estamos en la era de la información, caracterizada por la generación masiva de datos.
- Estos datos pueden ser generados por personas o cosas (sistemas de información, dispositivos, etc.)

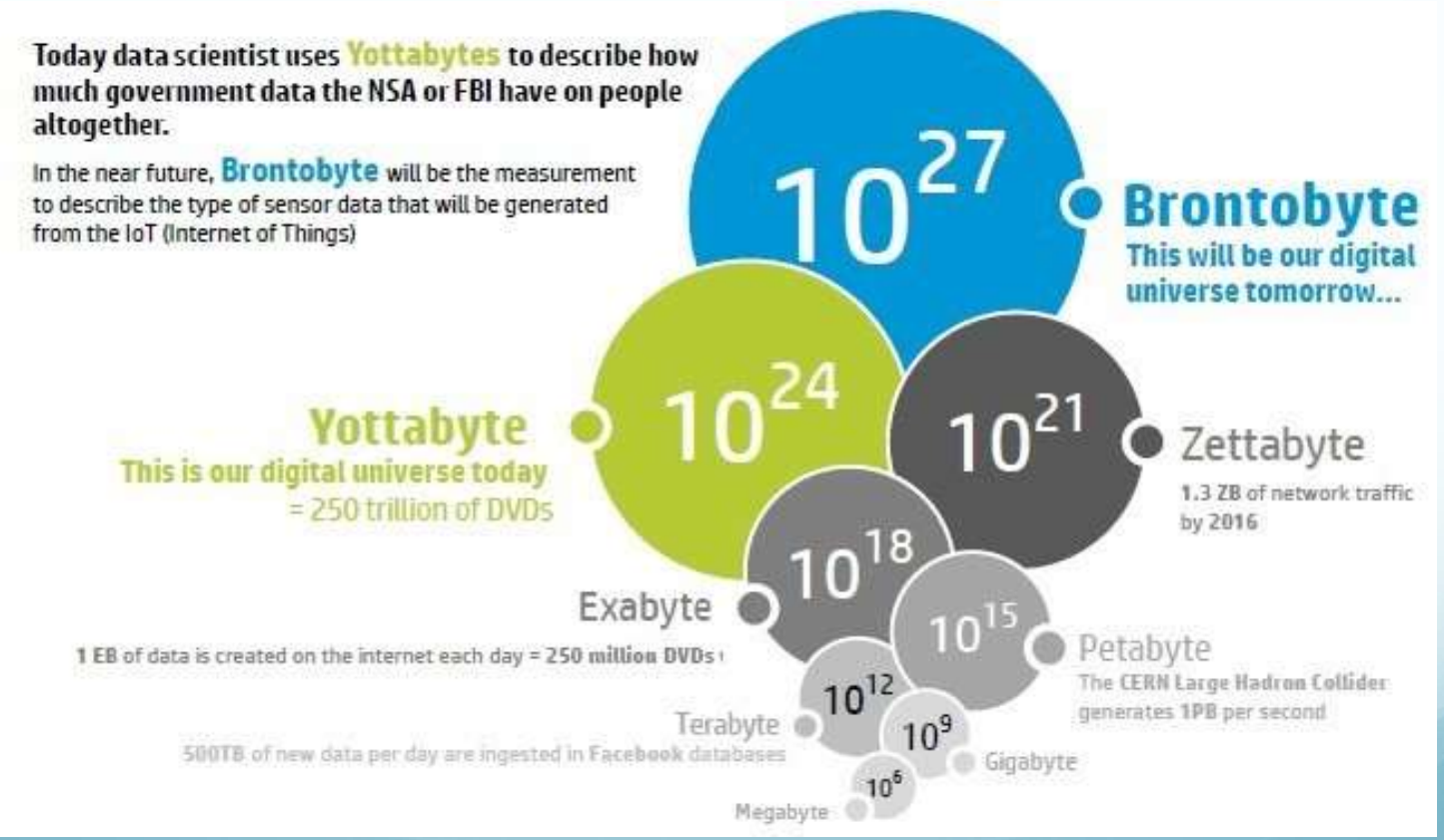
Ley de Moore de procesamiento y almacenamiento

- "La cantidad de información se está doblando cada 2 años"
- La capacidad de información que podemos procesar suele ser menor a la que podemos almacenar.



¿Zettabytes?

- 5 ZB es el equivalente informativo de 4.500 pilas de libros impresos de la Tierra al Sol.



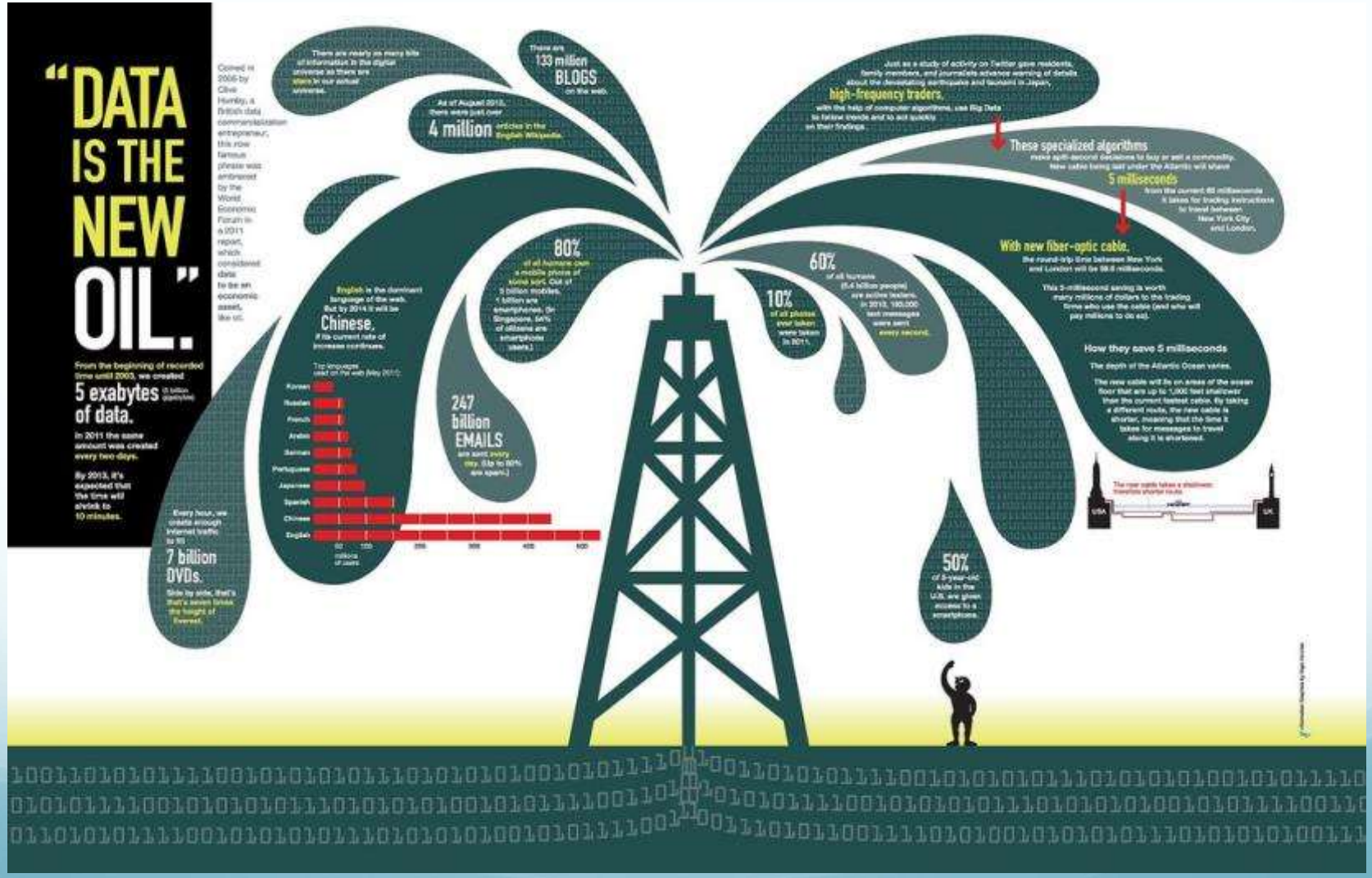
Cambio de paradigma

- El progreso y la innovación ya no se ven obstaculizados por la capacidad de recopilar datos, sino que por la capacidad de gestionar, analizar, resumir, visualizar y descubrir conocimiento de los datos recopilados de manera oportuna y de manera escalable.

Meta

- La tarea de la **industria de la información** es ser capaz de descubrir **conocimiento** que permita satisfacer las necesidades de los clientes (personas u organizaciones).

¿Son los datos el nuevo petróleo?



¿Son los datos el nuevo petróleo?

Here's Why Data Is Not The New Oil
Bernard Marr:
Forbes, 5 de Marzo 2018.

https://www.forbes.com/sites/bernardmarr/2018/03/05/heres-why-data-is-not-the-new-oil/#27f81203aa96

¿Cómo las compañías se enteran de nuestro secretos?

- Las tarjeta del retail, más que beneficiar a estas empresas con el pago a crédito, ellas se enteran de nuestras preferencias, de cuando y donde compramos, etc.

¿Cómo las compañías se enteran de nuestro secretos?

- La empresa TARGET descubrió cuándo una madre tendrá un bebé antes de que comience a comprar pañales.
- Andrew Pole analizó los datos y algunos patrones interesantes emergieron.
- Por ejemplo, las madres en su segundo trimestre de embarazo compran lociones sin esencias. Además durante las primeras 20 semanas compran suplementos como calcio, magnesio y zinc. Además cuando compran jabones y grandes bolsas de algodones, además de desinfectantes y toallas de mano, entonces están cerca del día del parto."
- Además Andrew Pole identificó 25 productos que juntos son buenos predictores de embarazo.
- Una familia recibió supones para ropas de bebés y el padre de una adolescente recibió la información de que su hija estaba embarazada.

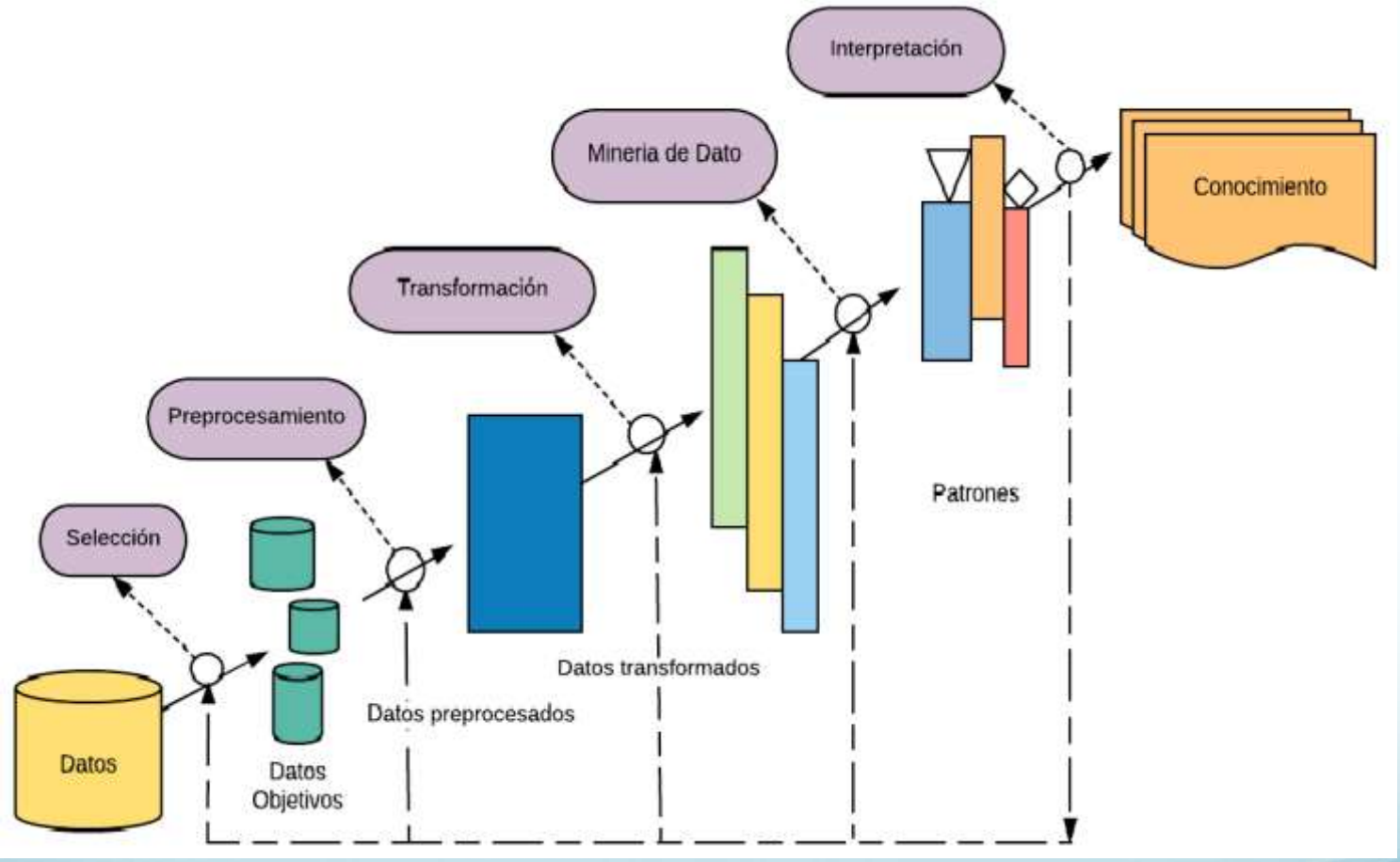
Estadísticas para conformar equipos deportivos

- El equipo de baseball Oakland Athletics (conocido como los A's) de bajo presupuesto para contratar, tuvieron exitosas campañas entre 2000 y 2003.
- Su Gerente General de ese entonces Billy Beane y Paul DePodesta analizaban las estadísticas del baseball para contratar jugadores subvalorados.
- Más allá de los sesgos y preferencias de los entendidos del baseball, lo que causó gran controversia.

Knowledge Discovery in Databases (KDD)

- Cuando hablamos de grandes cantidades de datos, el Descubrimiento de Conocimiento en Bases de Datos o KDD se refiere al proceso **de identificar patrones válidos, novedosos, potencialmente útiles y principalmente entendibles**.
- Consiste en integrar varias tecnologías para la gestión de los datos. Por ejemplo, manejo de Bases de Datos, data ware-housing, máquinas de aprendizaje, visualización, computación paralela, etc.

Proceso del Descubrimiento de Conocimiento en Bases de Datos (KDD)



Etapas del Proceso KDD

- Comprensión del dominio a estudiar y establecimiento de objetivos.
- Creación de un conjunto de datos (dataset) de interés.
- Limpieza y transformación de datos.
- Minería de datos.
- Evaluación e interpretación.
- Difusión de Resultados.

Comprensión del dominio a estudiar y establecimiento de objetivos.

- Como en cualquier tipo de investigación, es fundamental tener muy claros los límites y objetivos de lo que pretendemos. Es muy fácil perder el rumbo en el océano infinito de datos a nuestra disposición.
- En este paso es cuando reconocemos las fuentes de información más importantes y quienes tienen control sobre ellas. También es relevante incluir toda la metadata relacionada, dimensionar la cantidad de datos, y formatos.
- Toda la información más importante que se encuentre solamente en medios físicos debe ser digitalizada, previo a iniciar las actividades de KDD.

Creación de un conjunto de datos (dataset) de interés.

- En esta etapa debemos seleccionar e integrar los datos de interés provenientes de fuentes múltiples y heterogéneas.
- Es importante homogeneizar los formatos para que los datos sean más fáciles de procesar y analizar.

Limpieza y transformación de datos

- Se eliminan o corrigen los datos incorrectos y se decide la estrategia a seguir con los datos incompletos.
- Se proyectan los datos para considerar únicamente aquellas variables o atributos que van a ser relevantes (selección y extracción de características).

Minería de Datos

- La minería de datos se define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos". (Witten and Frank 2000.)
- En esta etapa se decide cuál es la tarea a realizar.
- Se elige el método a utilizar (algoritmo).
- Utilizamos el método y aplicamos medidas de interés.

Evaluación e interpretación de datos

- Es importante que comprendamos la diferencia entre dos términos clave:
- Patrones: son estructuras locales que pueden identificarse dentro de un conjunto mayor de datos.
- Modelos: son estructuras globales que relacionan las variables del problema. Nos permiten predecir el valor de alguna otra variable.
- https://science20.wordpress.com/2013/03/04/the-models-vs-patterns-problem/

Evaluación e interpretación de datos (2)

- Los resultados deben presentarse en un formato entendible. Por esta razón las técnicas de visualización son importantes para que los resultados sean útiles, dado que los modelos matemáticos o descripciones en formato de texto pueden ser difíciles de interpretar para los usuarios finales.
- Desde este punto del proceso es posible regresar a cualquiera de los pasos anteriores.

Evaluación e interpretación de datos (3)

- Un resultado es interesante si es:
  - **Válido**: Los patrones deben seguir siendo precisos para datos nuevos, y no sólo para aquellos que han sido usados en su obtención.
  - **Novedoso:** Que aporte algo desconocido tanto para el sistema y preferiblemente para el usuario
  - **Potencialmente útil:** La información debe conducir a acciones que reporten algún tipo de beneficio para el usuario.
  - **Comprensible:** La extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión y validación y uso en la toma de decisiones.
- El interés de los resultados se puede evaluar objetivamente (criterios estadísticos)
- Subjetivamente (perspectiva del usuario)

Difusión de resultados

- Se hace uso del nuevo conocimiento y se hace partícipe de él a todos los posibles usuarios.

KDD es multidisciplinario

- Ciencias de la computación
- Matemáticas
- Estadística