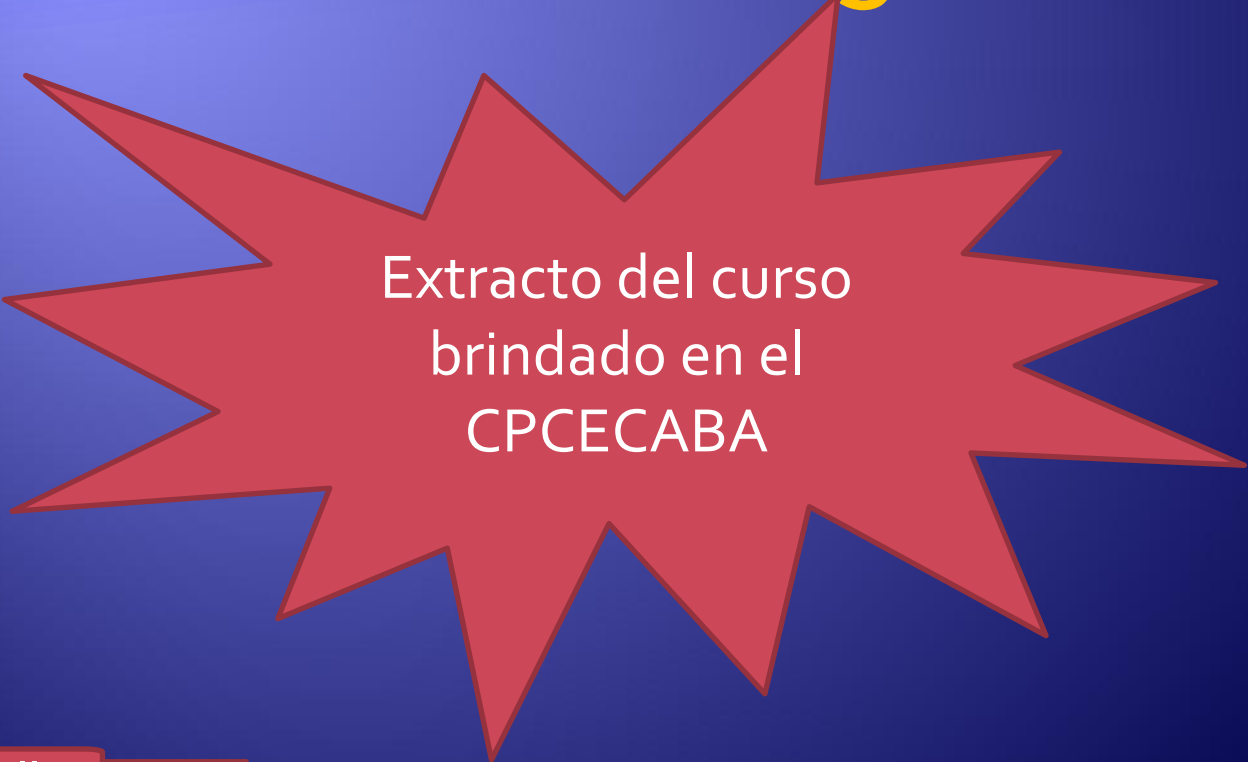


Business Intelligence

Data Mining

Big Data

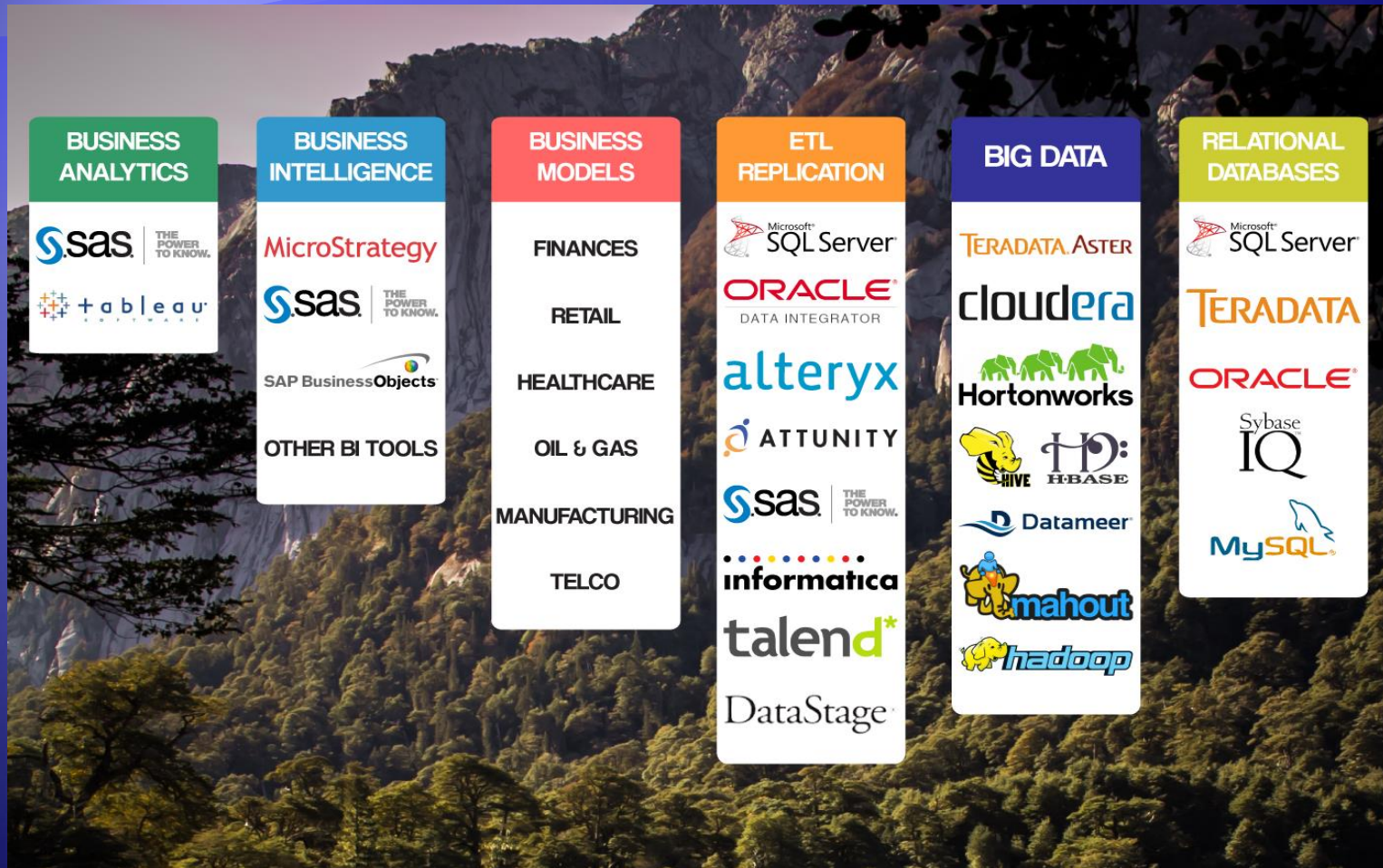
Machine Learning



Extracto del curso
brindado en el
CPCECABA



Tecnologías para la toma de decisiones



¿Que es Datamining ?



Conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Que es CONCRETAMENTE Data Mining

Un *algoritmo* en minería de datos (o aprendizaje automático) es un conjunto de TECNICAS heurísticas y cálculos que permiten crear un modelo a partir de datos.

El ser humano hace “Mining” en su vida... ¡pero no lo sabe!

¿Qué es Data Mining?

- ◆ 1. Extracción no trivial de información a partir de los datos.
 - ◆ La información debe ser útil y deber estar implícita y ser desconocida previamente.
- ◆ 2. Exploración y análisis, por medios automáticos o semiautomáticos, de grandes cantidades de datos de manera de poder descubrir patrones con significado que incrementen el conocimiento del dominio.

¿Porqué hacer minería de datos?

- ♦ Se recolectan y almacenan gran cantidad de datos
 - ♦ Internet, celulares Comercios Bancos, tarjetas de crédito, CRM, Redes Sociales...
- ♦ Las computadoras son más económicas y potentes
- ♦ La presión competitiva de los negocios
 - ♦ Proveer mejores servicios y más focalizados en las necesidades de cada cliente

Ejes Principales de la minería de datos



¿Cual estadística?

Descriptiva = obtiene, organiza, presenta y describe un conjunto de datos con el propósito de facilitar el uso: MINIMOS / MAXIMOS / PROMEDIOS / CUARTILES / MEDIAS / DISTRIBUCIÓN

Inferencial = Probabilidades / comprende los métodos y procedimientos que por medio de la inducción determina propiedades de una población. Obtener conclusiones útiles para hacer deducciones

¿Como Funciona?

Para crear un modelo, el algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias.

El algoritmo usa los resultados de este análisis en un gran número de iteraciones para determinar los parámetros óptimos para crear el modelo de minería de datos

A continuación, estos parámetros se aplican en todo el conjunto de datos para extraer patrones procesables y estadísticas detalladas.

¿Como Funciona?

El modelo de minería de datos que crea un algoritmo a partir de los datos puede tomar diversas formas, incluyendo:

- Un conjunto de clústeres que describe cómo se relacionan los casos de un conjunto de datos.
- Un árbol de decisión que predice un resultado y que describe cómo afectan a este los distintos criterios.
- Un modelo matemático que predice las ventas.
- Un conjunto de reglas que describen cómo se agrupan los productos en una transacción, y las probabilidades de que dichos productos se adquieran juntos. *** Cervezas y Pañales *** 1,67 +/- 20cm

¿Como Elegir el algoritmo correcto?

La elección del mejor algoritmo para una tarea analítica específica puede ser un desafío.

Aunque puede usar diferentes algoritmos para realizar la misma tarea, cada uno de ellos genera un resultado diferente, y algunos pueden generar más de un tipo de resultado

Algoritmos mas comunes

- **Algoritmos de clasificación**, que predicen una o más variables discretas, basándose en los demás atributos del conjunto de datos.
- **Algoritmos de regresión**, que predicen una o más variables numéricas continuas, como pérdidas o ganancias, basándose en otros atributos del conjunto de datos.
- **Algoritmos de segmentación**, que dividen los datos en grupos, o clústeres, de elementos que tienen propiedades similares.

Algoritmos más comunes

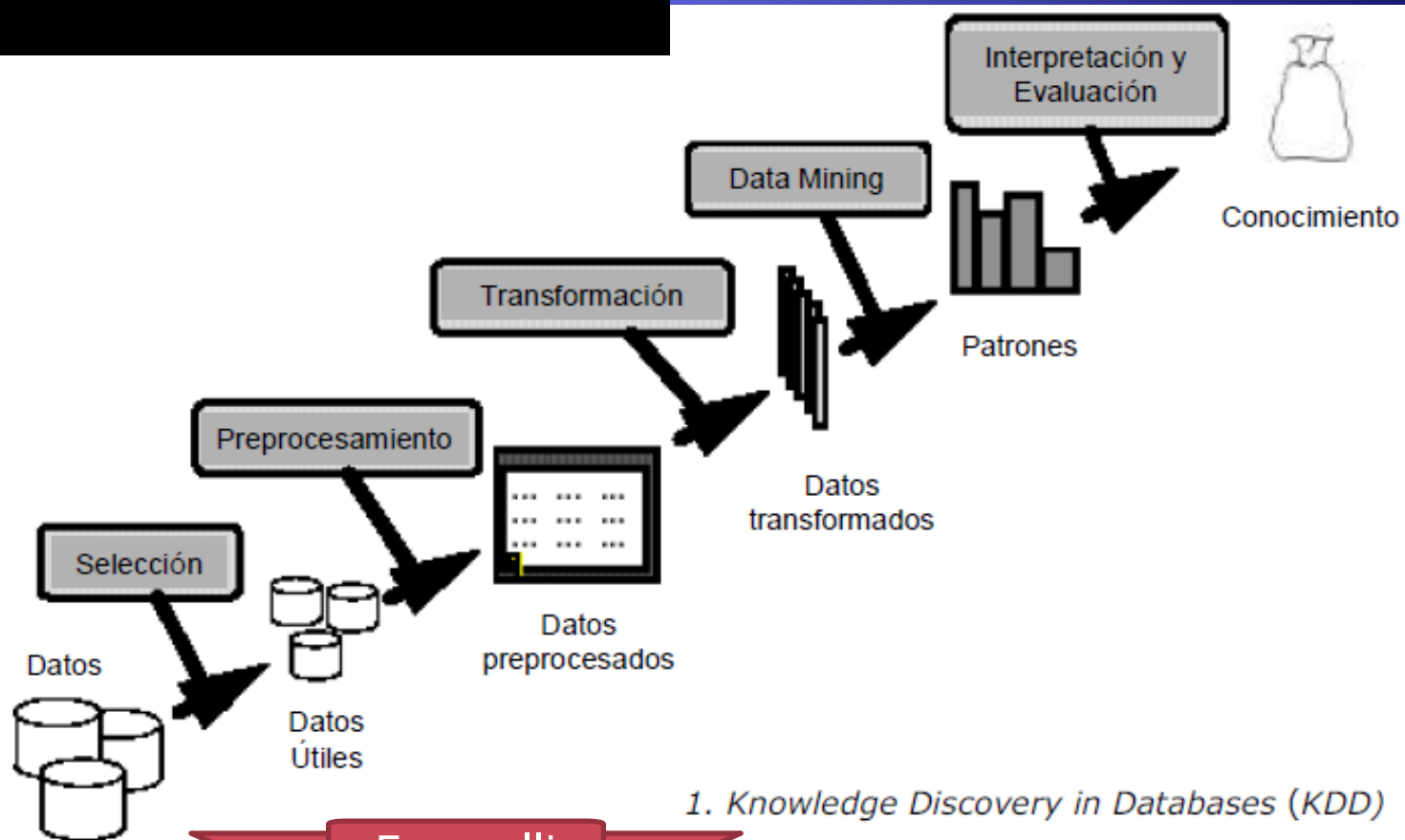
Algoritmos de asociación, que buscan correlaciones entre diferentes atributos de un conjunto de datos. La aplicación más común de esta clase de algoritmo es la creación de reglas de asociación, que pueden usarse en un análisis de la cesta de compra.

Algoritmos de análisis de secuencias, resumen las secuencias frecuentes o episodios en los datos, como una serie de clics en un sitio web o una serie de eventos de registro que preceden al mantenimiento del equipo.

Algoritmos combinados

Los analistas experimentados usarán a veces un algoritmo para determinar las entradas más eficaces (es decir, variables) y luego aplicarán un algoritmo diferente para predecir un resultado concreto basado en esos datos

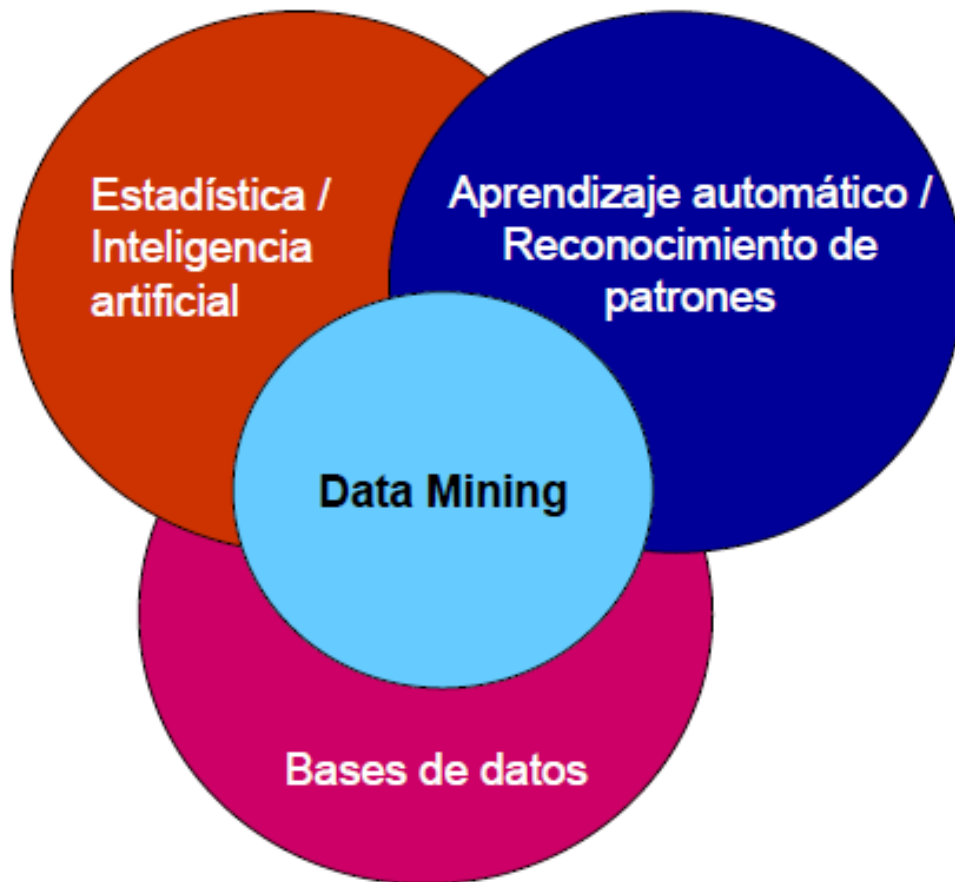
¿Cuáles son las etapas del proceso de descubrimiento de conocimiento a partir de los datos?



1. Knowledge Discovery in Databases (KDD)

¿Qué es y qué no es Data Mining?

- ◆ “No” es Data Mining:
 - ◆ Buscar un número telefónico en una guía telefónica
 - ◆ Consultar en un buscador en la Web por información sobre “Amazon”
- ◆ “Es” Data Mining:
 - ◆ Determinar los 3 apellidos más comunes en cierta ciudad (Pérez, Martinez, González)
 - ◆ Agrupar documentos similares devueltos por un buscador, de acuerdo a su contexto (Amazon, Google, Redes Sociales)



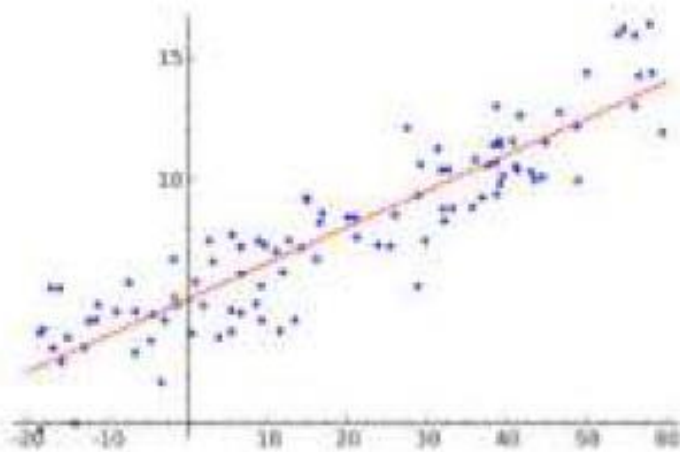
Las técnicas tradicionales pueden ser inadecuadas debido a:

- Gran volumen de datos
- Alta dimensionalidad de los datos
- La naturaleza heterogénea y distribuida de los datos

Tareas Data Mining

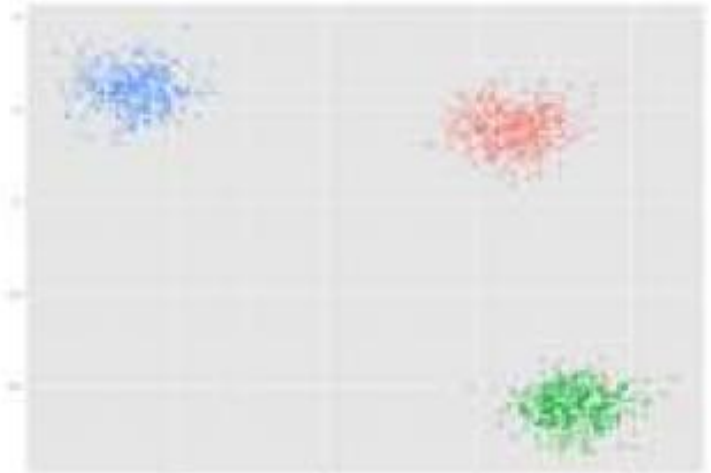
- **Método predictivo**

- Utilizan algunas variables para predecir valores futuros o desconocidos de otras variables



- **Método descriptivo**

- Encuentran patrones que describen los datos, que son interpretables por los humanos.



- **Método predictivo**

- Clasificación
- Regresión
- Detección de desvíos

- **Método descriptivo**

- Clustering
- Reglas de asociación
- Patrones secuenciales

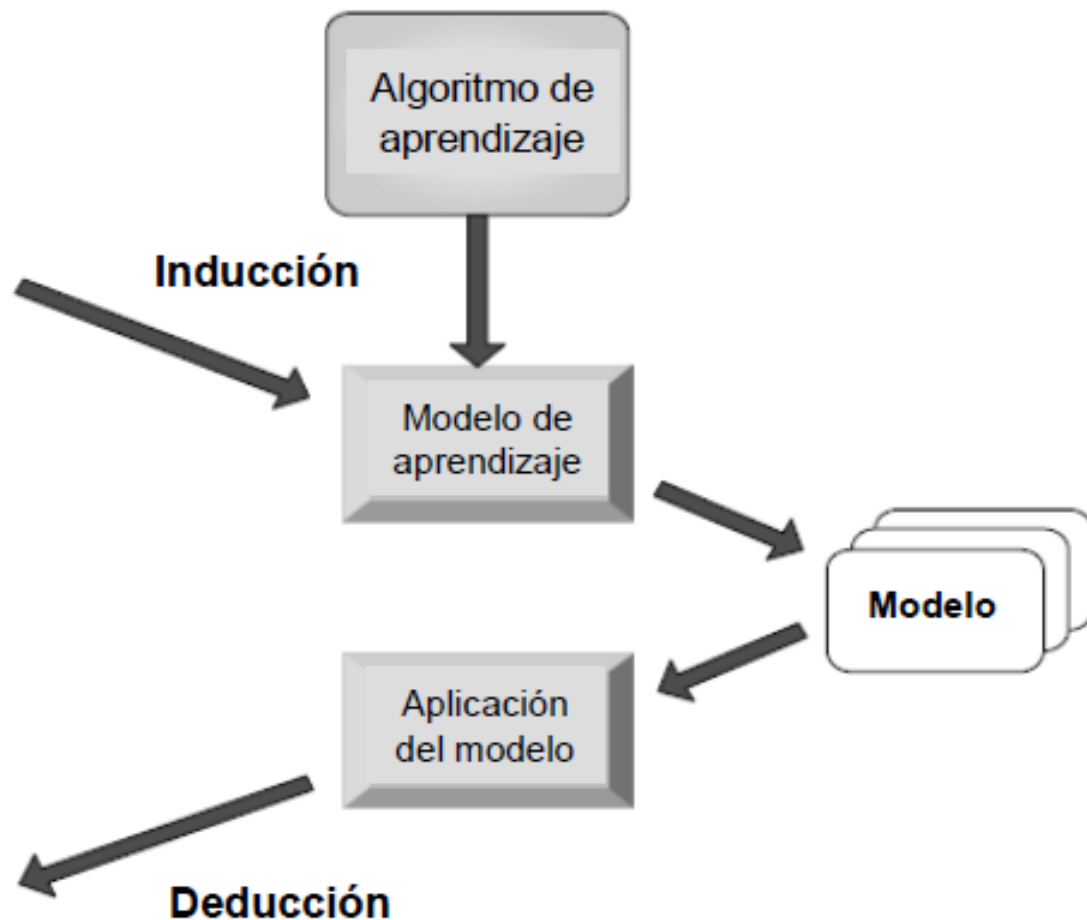
Clasificación: ejemplo

Nº	Atrib1	Atrib2	Atrib3	Clase
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Datos de entrenamiento

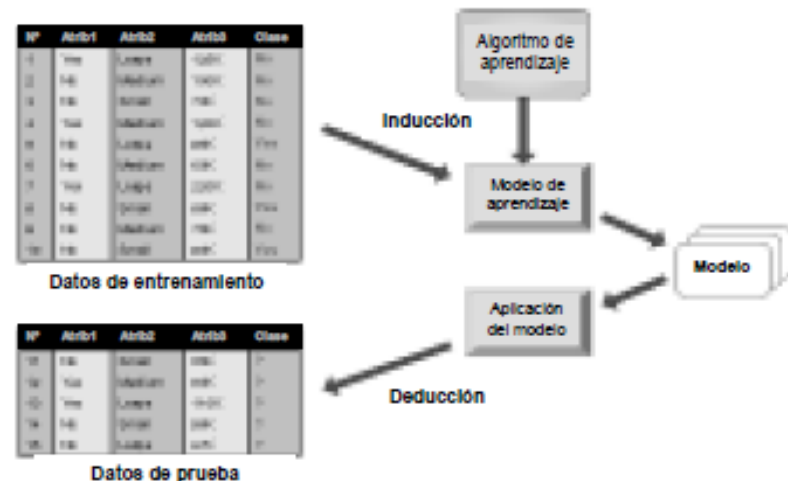
Nº	Atrib1	Atrib2	Atrib3	Clase
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Datos de prueba



Clasificación: definición

Dado un conjunto de registros o instancias (conjunto de entrenamiento o *training set*), en el cual cada instancia se define por un conjunto de atributos, y donde uno de los atributos es la clase.

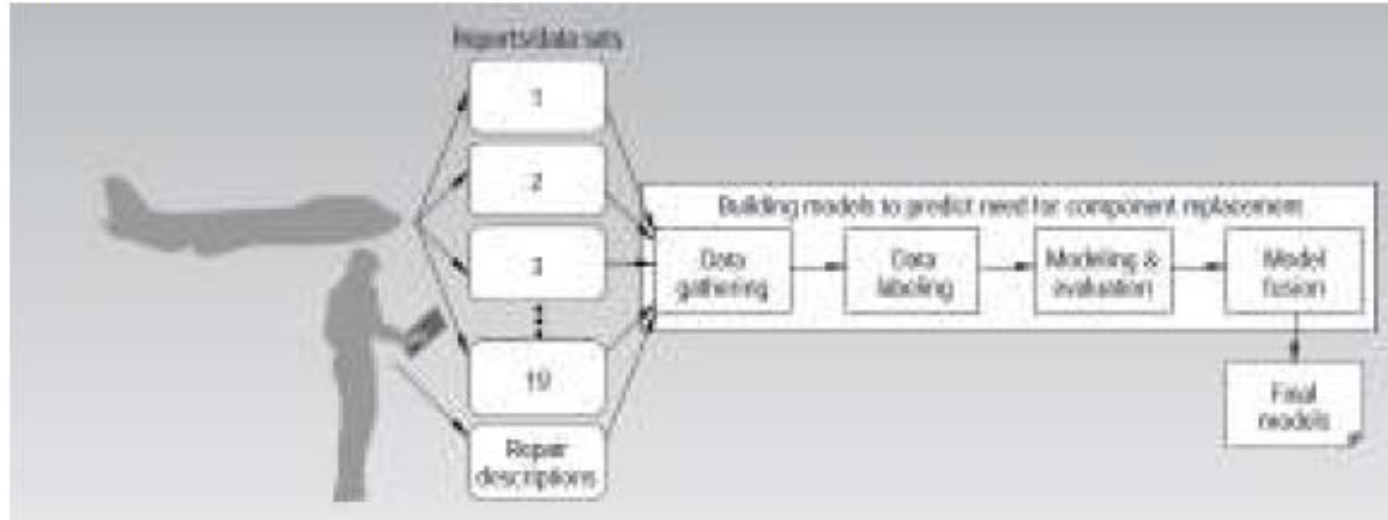


La clasificación permite identificar la clase a la cual pertenece una nueva instancia (no vista previamente), generando para ello un modelo predictivo.

Un conjunto de prueba o *test set* se usa para determinar la precisión del modelo. Usualmente, el conjunto de datos dados (*data set*) se divide en *training set* y *test set*, utilizando el *training set* para construir el modelo (inducción) y el *test set* para validarlo (deducción).

Clasificación: aplicaciones

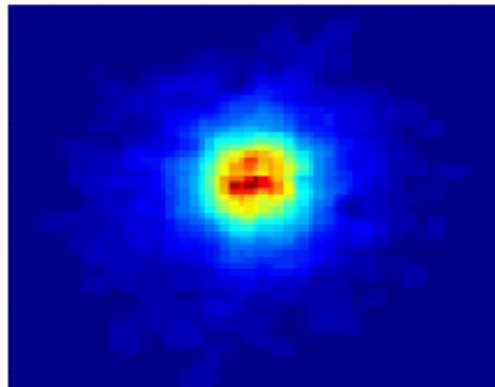
1. **Detección de fraudes:** predecir casos fraudulentos en transacciones de tarjetas de crédito.
2. **Pérdida de clientes** (*customer attrition* o *churn*): predecir si un cliente en particular es probable que se pierda.
3. **Reemplazo de componentes:** predecir el reemplazo o no del componente de un avión(1).



Clasificación: aplicaciones

4. **Catalogación astronómica:** predecir la clase (estrella o galaxia) de objetos astronómicos, especialmente aquellos débilmente visibles, basado en imágenes telescópicas.

Temprana

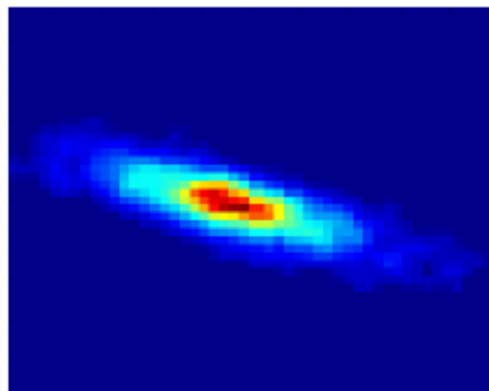


From: <http://aps.umn.edu>

Clase:

- Estado de formación

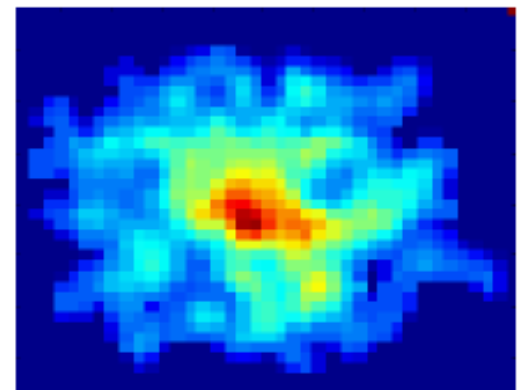
Intermedia



Atributos:

- Parámetros de la imagen
- Característica de las ondas de luz recibidas, etc.

Tardía

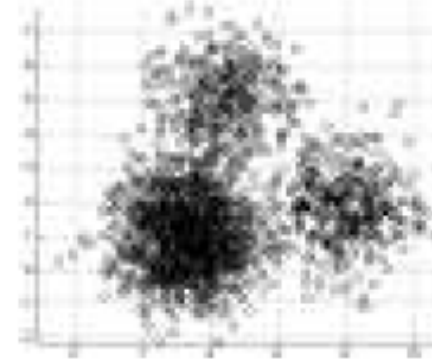


Tamaño del data set:

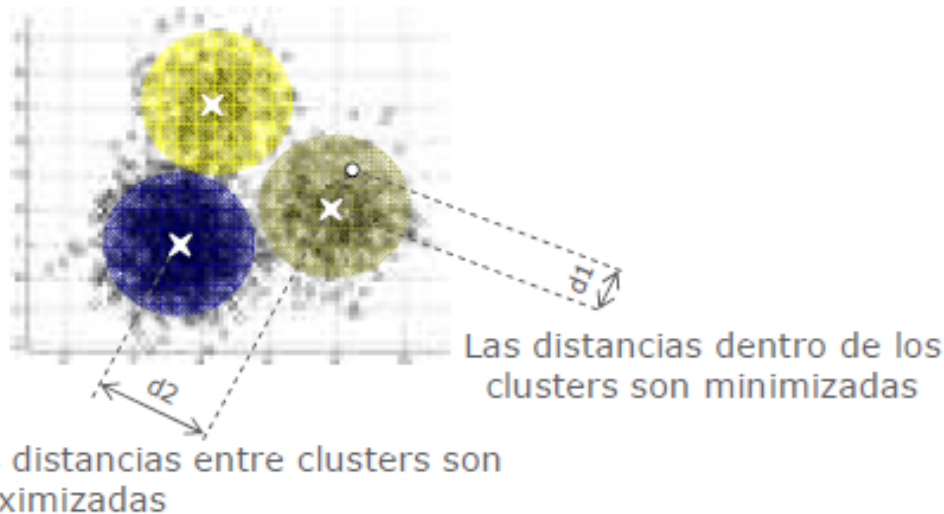
- 72 millones de estrellas, 20 millones de galaxias
- Objetos catalogados: 9 GB
- Base de datos de imágenes: 150 GB

Clustering: definición

Dado un conjunto de objetos de dato (puntos), cada uno con un conjunto de atributos y una medida de similitud entre ellos, hallar agrupamientos (*clusters*) tal que:



- ✓ Los objetos en un *cluster* son más similares entre sí.
- ✓ Los objetos en *clusters* separados son menos similares entre sí.



Medidas de Similitud:

- Distancia euclídea, si los atributos son continuos.
- Otras medidas específicas para ciertos problemas.

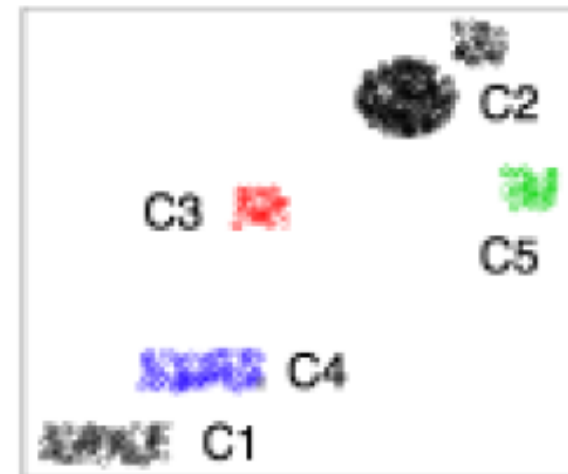
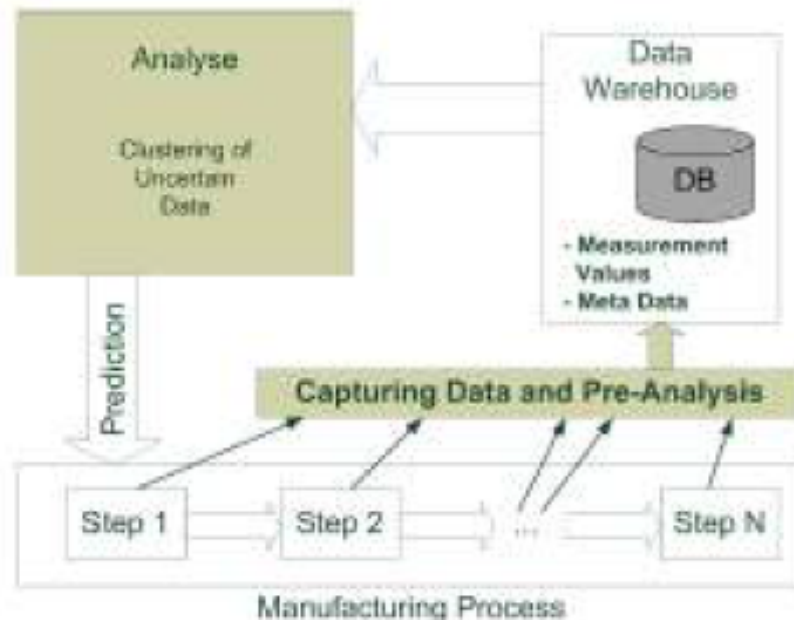
Clustering: aplicaciones

1. **Segmentación de mercado:** subdividir un mercado en distintos subconjuntos de clientes, donde cualquier subconjunto puede ser seleccionado como un mercado *target* a ser alcanzado con un *mix* distinto.
2. **Agrupamiento de documentos similares:** Hallar grupos de documentos similares entre sí basado en los términos importantes que aparezcan en ellos.

<i>Categoría (Cluster)</i>	<i>Total artículos</i>	<i>Correctamente ubicados</i>
<i>Financiera</i>	555	364
<i>Exterior</i>	341	260
<i>Nacional</i>	273	36
<i>Local</i>	943	746
<i>Deportes</i>	738	573
<i>Entretenimiento</i>	354	278

Clustering: aplicaciones

3. **Calibración de máquina:** en la producción de microchips es necesario re-calibrar los equipos para cada pedido específico de un cliente. El clustering de clientes con características de calibración similares permite reducir sustancialmente el tiempo de preparación de máquina (2)



Clustering: aplicaciones

4. **Acciones de la bolsa:** agrupación de acciones con igual comportamiento. Ejemplo tomado de S&P 500 Stock Data.

	<i>Discovered Clusters</i>	<i>Industry Group</i>
1	Applied-Matl-DOWN, Bay-Network-Down, 3-COM-DOWN, Cabletron-Sys-DOWN, CISCO-DOWN, HP-DOWN, DSC-Comm-DOWN, INTEL-DOWN, LSI-Logic-DOWN, Micron-Tech-DOWN, Texas-Inst-Down, Tellabs-Inc-Down, Natl-Semiconduct-DOWN, Oracle-DOWN, SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN, Autodesk-DOWN, DEC-DOWN, ADV-Micro-Device-DOWN, Andrew-Corp-DOWN, Computer-Assoc-DOWN, Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN, Microsoft-DOWN, Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN, Fed-Home-Loan-DOWN, MBNA-Corp-DOWN, Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP, Dresser-Inds-UP, Halliburton-HLD-UP, Louisiana-Land-UP, Phillips-Petro-UP, Unocal-UP, Schlumberger-UP	Oil-UP

Aprendizaje Supervisado vs. Aprendizaje No-supervisado

Aprendizaje Supervisado (e.g. Clasificación)

- Los datos de entrenamiento (instancias) están acompañadas de etiquetas que indican la clase de dichas instancias.
- Los nuevos datos son clasificados a partir de esos datos de entrenamiento.

Aprendizaje No-supervisado (e.g. Clustering)

- La clase de los datos de entrenamiento es desconocido.
- Dado un grupo de instancias el objetivo es establecer la existencia de clases o clusters en los datos.

Reglas de asociación: definición

Dado un conjunto de registros, cada uno de los cuales contiene un cierto número de ítems, tomados de una cierta colección:

El descubrimiento de reglas de asociación genera reglas de dependencia que predecirán la ocurrencia de un ítem basado en las ocurrencias de otros ítems.



Nº	Items
1	Pan, Gaseosa, Leche
2	Cerveza, Pan
3	Cerveza, Gaseosa, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Gaseosa, Pañales, Leche

Reglas de asociación descubiertas:

{Leche} → {Gaseosa}

{Pañales, Leche} → {Cerveza}

Reglas de asociación: aplicaciones

1. **Marketing:** dada una regla de asociación descubierta se pueden tomar acciones de promoción de ventas específicas.

Regla de asociación descubierta:
{Leche} → {Gaseosa}

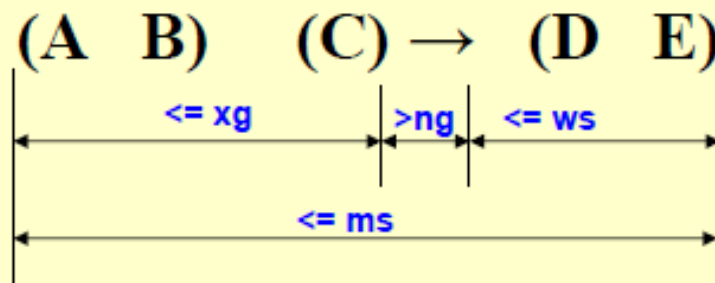
2. **Gestión de Inventarios:** por ejemplo una compañía que repara electrodomésticos desea anticipar la naturaleza de las reparaciones sobre sus productos y mantener los vehículos de servicio equipados con las partes correctas, para reducir el número de visitas a los hogares de sus clientes.

Patrones secuenciales: definición

Dado un conjunto de objetos, con cada objeto asociado con su propia secuencia de eventos, hallar las reglas que predicen fuertes **dependencias secuenciales** entre diferentes eventos.

$$(A \ B) \ (C) \rightarrow (D \ E)$$

Las reglas se forman descubriendo patrones, en primer término. Las ocurrencias de eventos en los patrones están gobernadas por restricciones de tiempo.



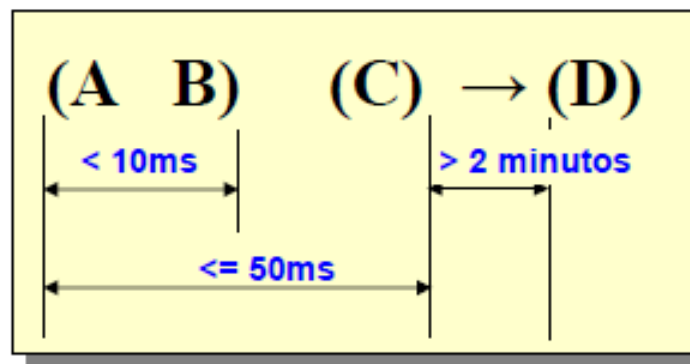
Patrones secuenciales: aplicaciones

1. **Logs de alarmas de telecomunicaciones:** encontrar el patrón secuencial de parámetros que generan una alarma, de manera de poder tomar acciones preventivas.

(A B) (C) → (D)

¿Corte de Energía?

Donde A={Problemas en el inversor}, B={Excesiva corriente de línea},
C={Alarma del rectificador}, y D={Alarma de incendio}



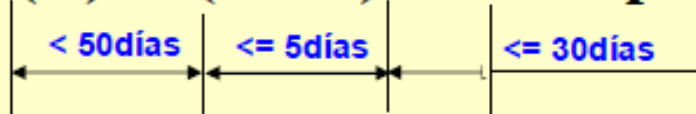
Patrones secuenciales: aplicaciones

2. **Transacciones en punto de venta:** encontrar el patrón secuencial de las compras de clientes en un negocio de deportes, de manera de poder tomar acciones de telemarketing para incentivar las ventas.

(A) (B C) → Compra de Remera

Donde $A = \{\text{Zapatillas}\}$, $B = \{\text{Raqueta de tenis}\}$, y $C = \{\text{Pelotas de tenis}\}$

(A) (B C) → Compra de Remera



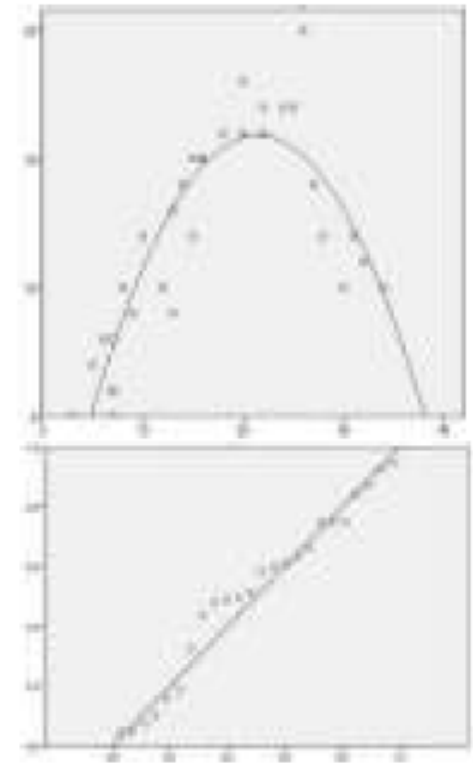
Regresión

Predecir el valor de una variable continua dada, en base a los valores de otras variables, asumiendo un modelo de dependencia lineal o no-lineal.

✓ Vastamente estudiada en el campo de la Estadística.

Aplicaciones:

- Predicción de los montos de venta de nuevos productos en base a los gastos de publicidad.
- Predicción de velocidades del viento como una función de la temperatura, humedad, presión atmosférica, etc.
- Predicción de series de tiempo de índices de acciones en la Bolsa de Comercio.



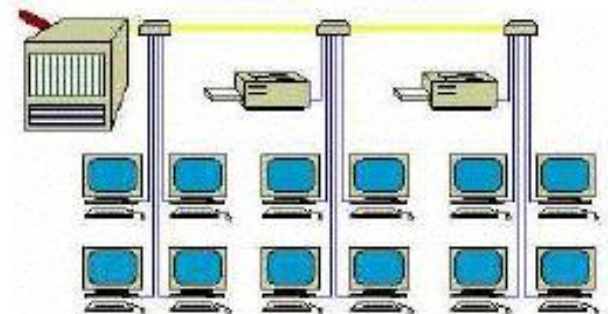
Detección de desvíos

Detectar anomalías o desvíos significativos del comportamiento normal.



Aplicaciones:

- Detección de fraudes en tarjetas de crédito.
- Detección de intrusos en redes informáticas.



¿Fijamos Conceptos?

1. ¿Qué es Data Mining?
2. Describir los pasos del proceso de descubrimiento de conocimiento KDD.
3. Discutir si cada una de las siguientes actividades es o no una tarea de data mining:
4. Suponer que usted tiene un empleo como consultor en data mining para una compañía de un buscador de Internet. Describir cómo puede data mining ayudar a la compañía por medio de ejemplos específicos de cómo se pueden aplicar técnicas como Clustering, Clasificación, Descubrimiento de Reglas de Asociación y Detección de Anomalías.

1. ¿Qué es Data Mining?

- ♦ El Data Mining es un proceso que nos permite descubrir o incrementar el conocimiento que se posee de una cierta área a partir de la aplicación de una serie de técnicas, haciendo uso de las tecnologías desarrolladas a partir de las bases de datos, estadística y aprendizaje automático.
- ♦ Para algunos esta definición debe completarse con la adición de que tal cometido se logra a partir del análisis de grandes volúmenes de datos almacenados en sistemas informáticos (Big Data).

2.Describir los pasos del proceso de descubrimiento de conocimiento KDD.

- ♦ Las etapas del proceso de descubrimiento de conocimiento o Data Mining a partir de los datos de una o varias base de datos son:
 - ♦ Selección,
 - ♦ Pre-procesamiento,
 - ♦ Transformación,
 - ♦ Data Mining propiamente dicho, e
 - ♦ Interpretación y Evaluación.

3. Discutir si cada una de las siguientes actividades es o no una tarea de data mining:

- a. Dividir los clientes de una compañía de acuerdo a su género.
- b. Dividir los clientes de una compañía de acuerdo a su rentabilidad.
- c. Calcular el total de ventas de una compañía
- d. Clasificar una base de datos de estudiantes basado en los números de identificación.
- e. Predecir los resultados de arrojar un par de dados.
- f. Predecir el precio futuro de las acciones de una compañía usando registros históricos.
- g. Monitorear los latidos del corazón de un paciente para buscar anomalías.
- h. Monitorear ondas sísmicas para detectar actividades de terremotos.
- i. Extraer las frecuencias de una onda de sonido.

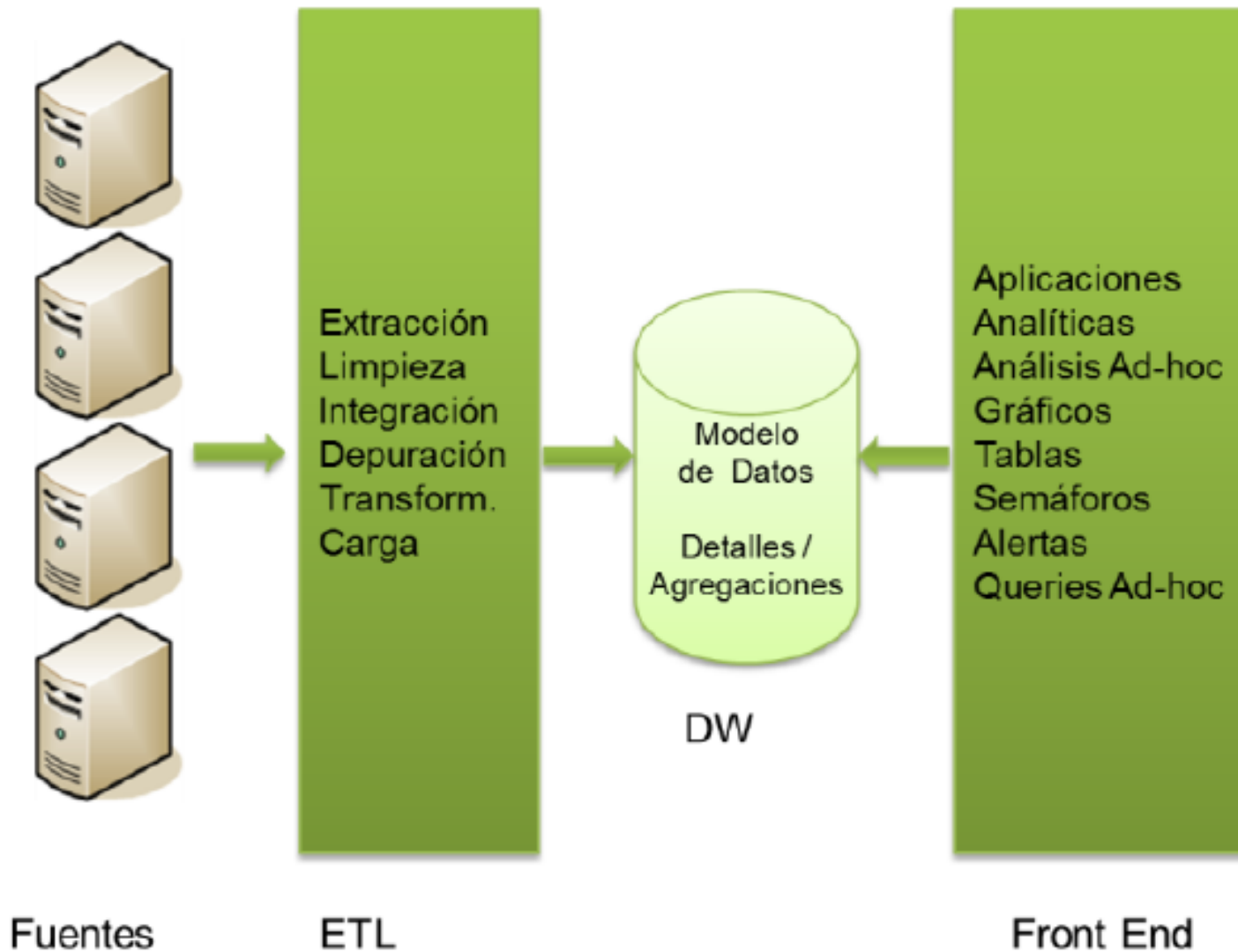
4. Suponer que usted tiene un empleo como consultor en data mining para una compañía de un buscador de Internet. Describir cómo puede data mining ayudar a la compañía por medio de ejemplos específicos de cómo se pueden aplicar técnicas como Clustering, Clasificación, Descubrimiento de Reglas de Asociación y Detección de Anomalías.

- ♦ Clustering: determinar tipos de grupos de usuarios que acceden a sitios de compra de productos por Internet a partir de algunas características particulares básicas (edad, género, país, etc.): jóvenes de países del tercer mundo, personas adultas solteras, etc.
 - ♦ De esta manera se pueden mejorar las estrategias de marketing para diferentes productos a ser ofertados a través de Internet.
- ♦ Clasificación: determinar la clase de producto de interés de cada usuario del buscador en base a sus datos básicos.
 - ♦ De esta manera se pueden recomendar opciones de compra a los diferentes usuarios del buscador. Netflix? Spotify? Youtube? Tinder? Campañas Políticas...?

Continua...

- ◆ Reglas de asociación: en base a los registros históricos de las últimas visitas a páginas web determinar las posibles recomendaciones a realizar en base a las búsquedas realizadas en la actualidad.
 - ◆ Otras personas que vieron XXX también consultaron ZZZ
- ◆ Detección de anomalías: detectar hackers en base a la secuencia de páginas visitadas.
 - ◆ Han accedido a tu dispositivo desde una IP no registrada previamente...

DW – Un Esquema Genérico y Simple





Las tres V del Big Data

Volumen

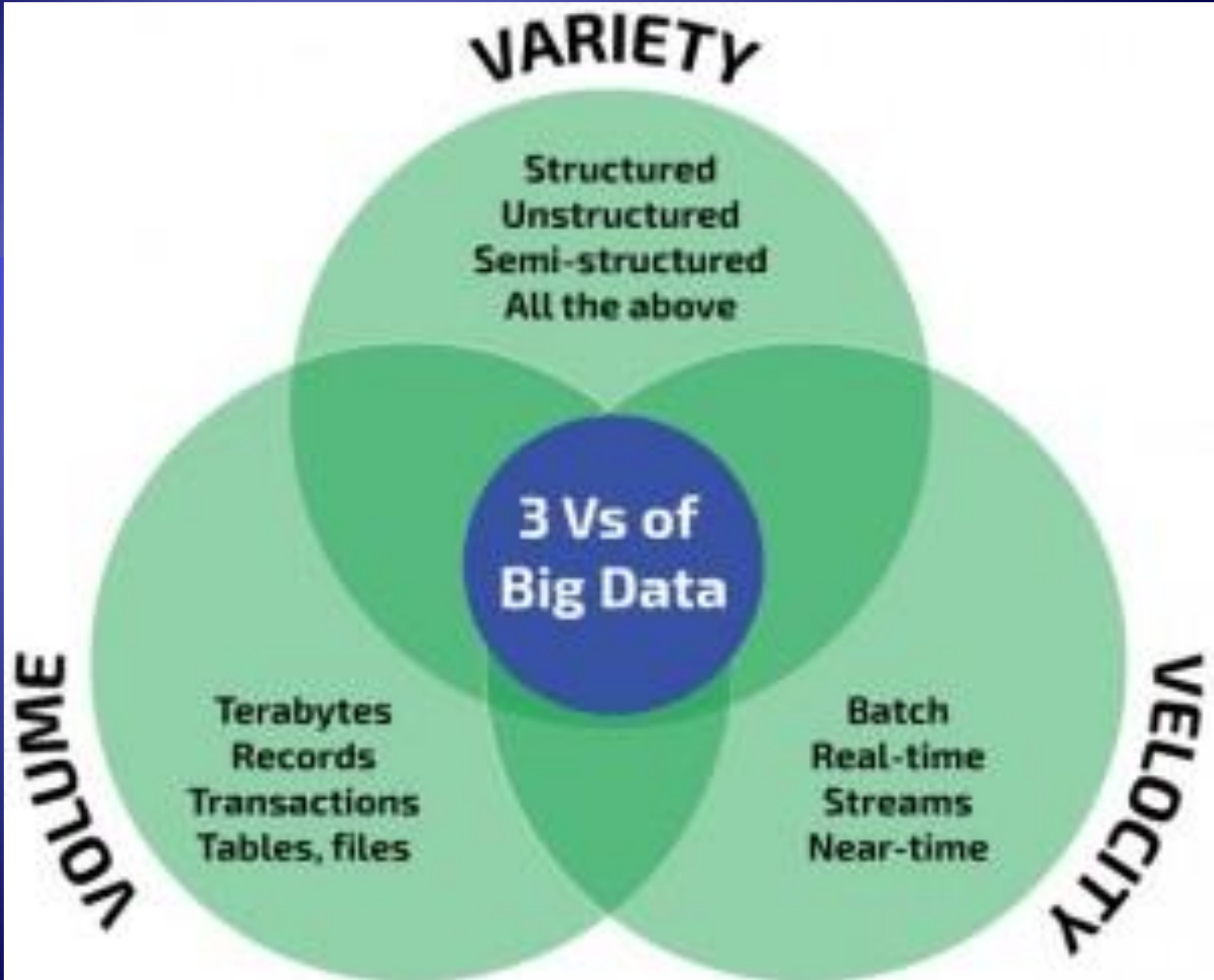
- Gigantesca cantidad de datos
- PETABYTES

Velocidad

- Problemas a considerar:
 - Vel en que la info está disponible para analisis
 - Velocidad en que podemos actuar luego de procesar

Variedad

- Diferentes tipos de Data:
 - Estructurada (20 %)
 - Semi – Estructurada (10%)
 - No Estructurada (70%)



Unidades de almacenamiento (tradicionales)

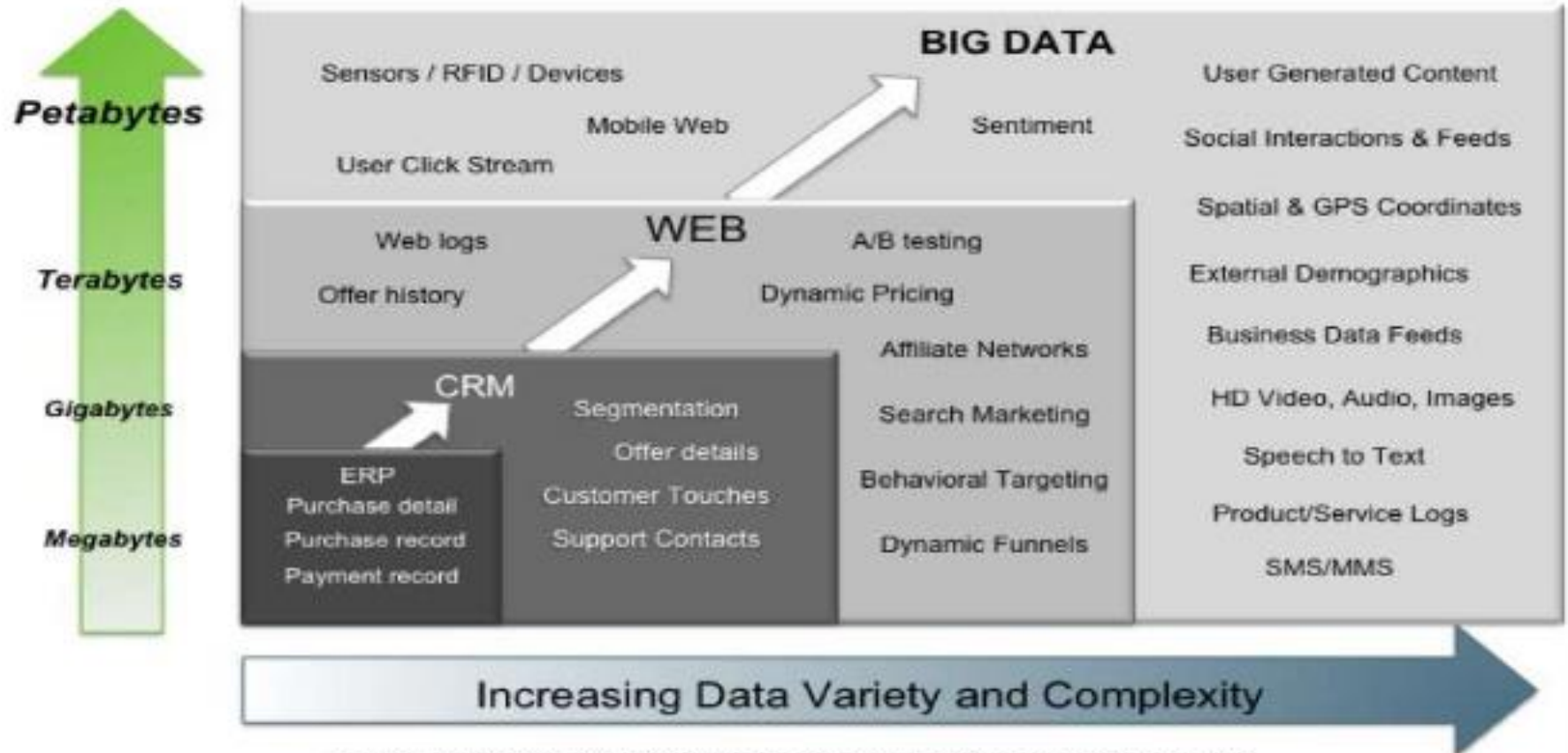
Nombre	Símbolo	Binario	Número de bytes	Equivale
<u>kilobyte</u>	KB	2^{10}	1.024	=
<u>megabyte</u>	MB	2^{20}	1.048.576	1.024KB
<u>gigabyte</u>	GB	2^{30}	1.073.741.824	1.024MB
<u>terabyte</u>	TB	2^{40}	1.099.511.627.776	1.024GB
<u>petabyte</u>	PB	2^{50}	1.125.899.906.842.624	1.024TB
<u>exabyte</u>	EB	2^{60}	1.152.921.504.606.846.976	1.024PB
<u>zettabyte</u>	ZB	2^{70}	1.180.591.620.717.411.303.424	1.024EB
<u>yottabyte</u>	YB	2^{80}	1.208.925.819.614.629.174.706.176	1.024ZB

2021 *This Is What Happens In An Internet Minute*



Evolución Procesamiento de datos

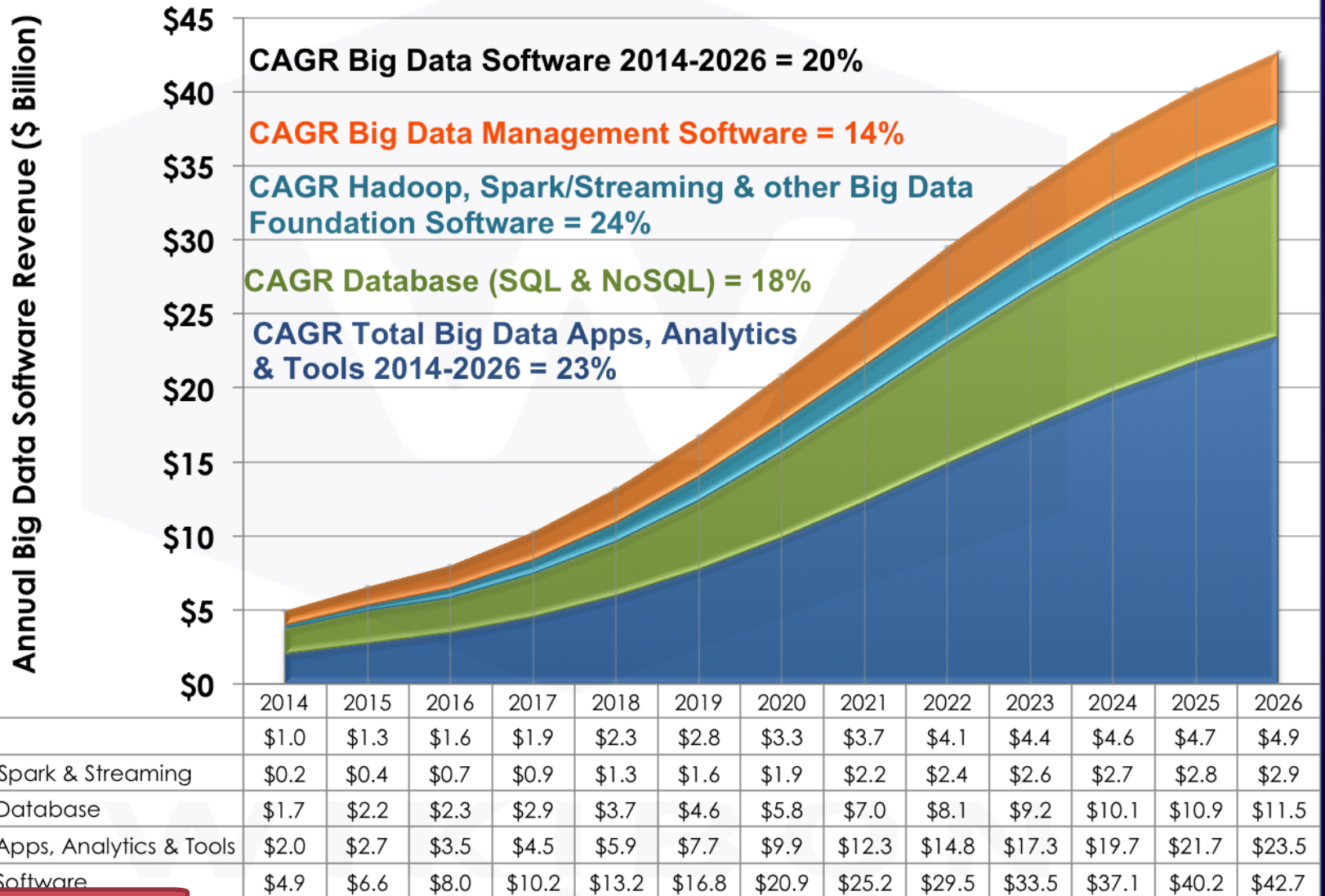
Big Data = Transactions + Interactions + Observations



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

Tamaño del Mercado

Wikibon Big Data Software Breakout Projection 2014-2026 (\$B)



Big Data Analytcs

- Examinar grandes cantidades de data en tiempos razonables
- Información adecuada y oportuna
- Identificación de patrones escondidos o correlaciones desconocidas en datos estructurados y no estructurados
- Ventajas competitivas por patrones o mapas de calor en segmentos de Mercado
- Mejores decisiones de Negocio: Estrategicas y operacionales
- Eficiencia en Marketing, satisfacción del cliente -> Aumento del Revenue (por poco stock, mejor logistica, menor cantidad empleados para procesos operativos).

Machine Learning

- ◆ Convertir “Datos” en “Conocimiento”
- ◆ Test de Turing (Lenguaje Natural)
- ◆ Inteligencia Artificial
- ◆ Inducción de Conocimiento
- ◆ Sistemas: Ayudante – Colega – Autómata
- ◆ Aprendizaje:
 - ◆ Supervisado / No supervisado / semi supervisado
 - ◆ Por refuerzo / transducción / Multi tarea (reaprendizaje)



Preguntas y Respuestas