

Introducción a la Minería de Datos

Verónica Guarín Escudero
Escuela de Estadística

Correo: jvguarine@unal.edu.co

Contenido del Curso

► Introducción a la Minería de Datos:

- ¿qué es la minería de datos? - Ejemplos de aplicaciones y disciplinas - Tipos de Datos - Tipos de Modelos.
 - Proceso de extracción de conocimiento.
 - Fuentes de Información, recuperación de información, Web scrapping.
 - Pre procesamiento de información, reglas de asociación y correlación.
 - Repaso general de clasificación y predicción.
- Relación con otras

► Introducción a la minería de texto:

- Introducción ¿qué es la minería de texto?
- Manipulación de datos. N-gramas, la ley de zipf, importancia de las palabras td-idf.
- Análisis de sentimientos.
- Modelaje de tópicos.

Contenido del Curso

- ▶ Evaluaciones:
 - Asistencia - Tareas y actividades (25%)
 - Trabajo 1: Fuentes de Información - Depuración - Web scrapping. (25%)
 - Trabajo 2: Regresión - Clasificación - Reglas de asociación. (25%)
 - Trabajo 3: Minería de Texto. (25%)
- ▶ Software Estadístico R (pero abiertos a usar otros como python).

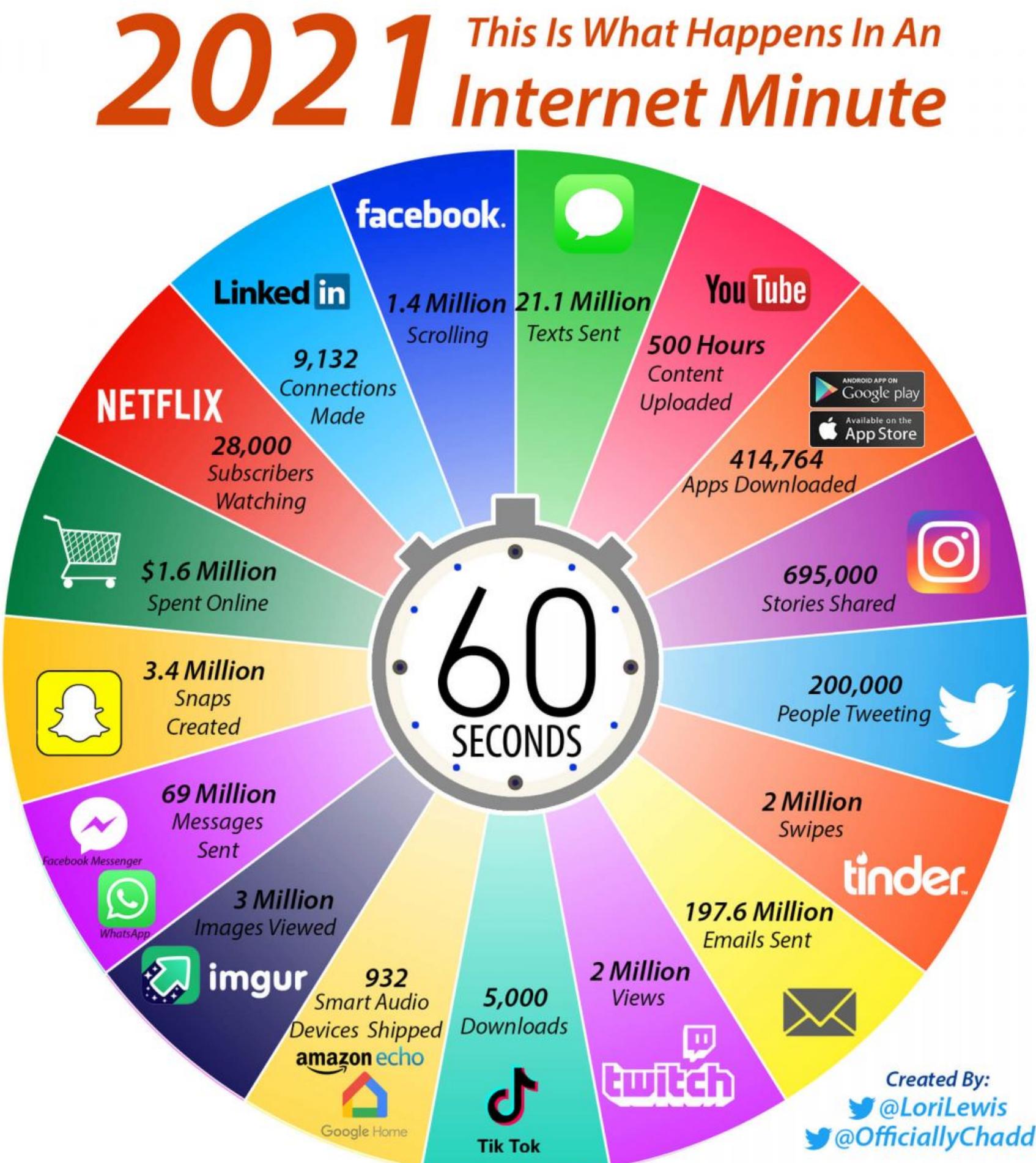
★ **Horario de Asesoría.** Miércoles de 6-7:30 pm vía google meet.

Introducción

- ▶ ¿Qué es Minería de Datos?

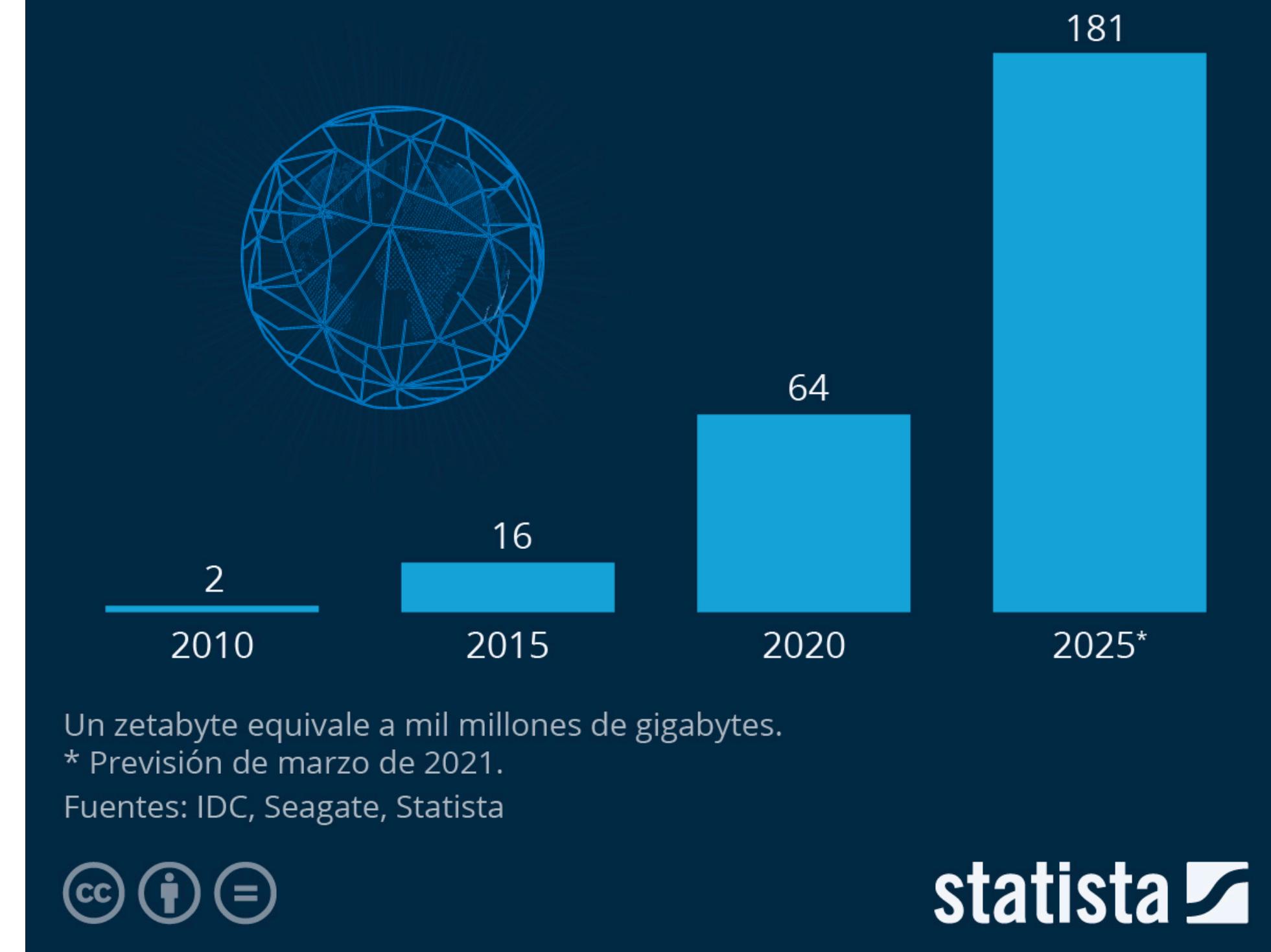


Introducción



El Big Bang del Big Data

Volumen estimado de datos digitales creados o replicados en todo el mundo, en zetabytes



Introducción

► ¿Qué es Minería de Datos?

Hay muchas definiciones...

En [Witten & Frank 2000] se define la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos.

One of many definitions:

"Data mining is the science of extracting useful knowledge from huge data repositories."

Tomado de: P. Tan, M. Steinbach, V. Kumar. Introduction to Data Mining. Addison-Wesley. 2006.

Introducción

► ¿Qué es Minería de Datos? - ¿Por qué?

Desde un punto de vista comercial...

- Los negocios y empresas recogen y almacenan gran cantidad de información.
 - Negocios de ventas de productos.
 - Bancos y entidades crediticias.
 - Redes sociales.
 - Empresas telefónicas.
- Computadores son más baratos y más poderosos.
- Competencia por proporcionar mejores servicios.
 - Sistemas de recomendación.
 - Segmentar clientes.

Introducción

► ¿Qué es Minería de Datos? - Ejemplos

El primer ejemplo pertenece al ámbito de la banca. Un banco por Internet desea obtener reglas para predecir qué personas de las que solicitan un crédito no lo devuelven.

IDC	D-crédito (años)	C-crédito (euros)	Salario (euros)	Casa propia	Cuentas morosas	...	Devuelve- crédito
101	15	60.000	2.200	sí	2	...	no
102	2	30.000	3.500	sí	0	...	sí
103	9	9.000	1.700	sí	1	...	no
104	15	18.000	1.900	no	0	...	sí
105	10	24.000	2.100	no	0	...	no
...

A partir de éstos, las técnicas de minería de datos podrían sintetizar algunas reglas, como por ejemplo:

SI Cuentas-Morosas > 0 **ENTONCES** Devuelve-crédito = no

SI Cuentas-Morosas = 0 **Y** [(Salario > 2.500) **O** (D-crédito > 10)] **ENTONCES**
Devuelve-crédito = sí

Introducción

► ¿Qué es Minería de Datos? - Ejemplos

Un supermercado quiere obtener información sobre el comportamiento de compra de sus clientes. Piensa que de esta forma puede mejorar el servicio que les ofrece: reubicación de los productos que se suelen comprar juntos, localizar el emplazamiento idóneo para nuevos productos, etc. Para ello dispone de la información de los productos que se adquieren en cada una de las compras o cestas.

Idcesta	Huevos	Aceite	Pañales	Vino	Leche	Mantequilla	Salmón	Lechugas	...
1	sí	no	no	sí	no	sí	sí	sí	...
2	no	sí	no	no	sí	no	no	sí	...
3	no	no	sí	no	sí	no	no	no	...
4	no	sí	sí	no	sí	no	no	no	...
5	sí	sí	no	no	no	sí	no	sí	...
6	sí	no	no	sí	sí	sí	sí	no	...
7	no	no	no	no	no	no	no	no	...
8	sí	sí	sí	sí	sí	sí	sí	no	...
...

Analizando estos datos el supermercado podría encontrar, por ejemplo, que el 100 por cien de las veces que se compran pañales también se compra leche, que el 50 por ciento de las veces que se compran huevos también se compra aceite o que el 33 por ciento de las veces que se compra vino y salmón entonces se compran lechugas. También se puede analizar cuáles de estas asociaciones son frecuentes, porque una asociación muy estrecha entre dos productos puede ser poco frecuente y, por tanto, poco útil.

Introducción

► ¿Qué es Minería de Datos? - Ejemplos

Una gran cadena de tiendas de electrodomésticos desea optimizar el funcionamiento de su almacén manteniendo un stock de cada producto suficiente para poder servir rápidamente el material adquirido por sus clientes. Para ello, la empresa dispone de las ventas efectuadas cada mes del último año de cada producto

Producto	mes-12	...	mes-4	mes-3	mes-2	mes-1
televisor plano 30' Phlipis	20	...	52	14	139	74
vídeo-dvd-recorder Miesens	11	...	43	32	26	59
discman mp3 LJ	50	...	61	14	5	28
frigorífico no frost Jazzussi	3	...	21	27	1	49
microondas con grill Sanson	14	...	27	2	25	12
...

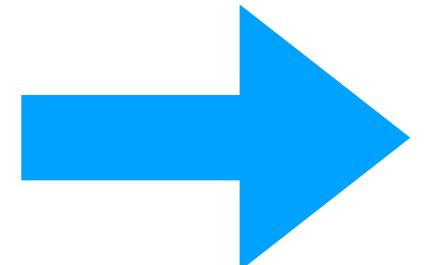
Esta información permite a la empresa generar un modelo para predecir cuáles van a ser las ventas de cada producto en el siguiente mes en función de las ventas realizadas en los meses anteriores, y efectuar así los pedidos necesarios a sus proveedores para disponer del stock necesario para hacer frente a esas ventas.

Introducción

► ¿Qué es Minería de Datos? - Ejemplos

El departamento de recursos humanos de una gran empresa desea categorizar a sus empleados en distintos grupos con el objetivo de entender mejor su comportamiento y tratarlos de manera adecuada. Para ello dispone en sus bases de datos de información sobre los mismos (sueldo, estado civil, si tiene coche, número de hijos, si su casa es propia o de alquiler, si está sindicado, número de bajas al año, antigüedad y sexo).

Id	Sueldo	Casado	Coche	Hijos	Alq/prop	Sindicado	Bajas/año	Antigüedad	Sexo
1	1.000	Sí	No	0	Alquiler	No	7	15	H
2	2.000	No	Sí	1	Alquiler	Sí	3	3	M
3	1.500	Sí	Sí	2	Prop	Sí	5	10	H
4	3.000	Sí	Sí	1	Alquiler	No	15	7	M
5	1.000	Sí	Sí	0	Prop	Sí	1	6	H
6	4.000	No	Sí	0	Alquiler	Sí	3	16	M
7	2.500	No	No	0	Alquiler	Sí	0	8	H
8	2.000	No	Sí	0	Prop	Sí	2	6	M
9	2.000	Sí	Sí	3	Prop	No	7	5	H
10	3.000	Sí	Sí	2	Prop	No	1	20	H
11	5.000	No	No	0	Alquiler	No	2	12	M
12	800	Sí	Sí	2	Prop	No	3	1	H
13	2.000	No	No	0	Alquiler	No	27	5	M
14	1.000	No	Sí	0	Alquiler	Sí	0	7	H
15	8 00	No	Sí	0	Alquiler	No	3	2	H
...



Grupo 1:
Sueldo: 1.535,2€
Casado: No -> 0,777
Sí -> 0,223
Coche: No -> 0,82
Sí -> 0,18
Hijos: 0,05
Alq/Prop: Alquiler -> 0,99
Propia -> 0,01
Sindic.: No -> 0,8
Sí -> 0,2
Bajas/Año: 8,3
Antigüedad: 8,7
Sexo: H -> 0,61
M -> 0,39

Grupo 2:
Sueldo: 1.428,7€
Casado: No -> 0,98
Sí -> 0,02
Coche: No -> 0,01
Sí -> 0,99
Hijos: 0,3
Alq/Prop: Alquiler -> 0,75
Propia -> 0,25
Sindic.: Sí -> 1,0
Sí -> 0,33
Bajas/Año: 2,3
Antigüedad: 8
Sexo: H -> 0,25
M -> 0,75

Grupo 3:
Sueldo: 1.233,8€
Casado: Sí -> 1,0
Coche: No -> 0,05
Sí -> 0,95
Hijos: 2,3
Alq/Prop: Alquiler -> 0,17
Propia -> 0,83
Sindic.: No -> 0,67
Sí -> 0,33
Bajas/Año: 5,1
Antigüedad: 8,1
Sexo: H -> 0,83
M -> 0,17

Estos grupos podrían ser interpretados por el departamento de recursos humanos de la siguiente manera:

Grupo 1: sin hijos y con vivienda de alquiler. Poco sindicados. Muchas bajas.

Grupo 2: sin hijos y con coche. Muy sindicados. Pocas bajas. Normalmente son mujeres y viven en casas de alquiler.

Grupo 3: con hijos, casados y con coche. Mayoritariamente hombres propietarios de su vivienda. Poco sindicados. su vivienda. Poco sindicados.

Introducción

► ¿Qué es Minería de Datos?

Para extraer conocimiento, los datos necesitan...

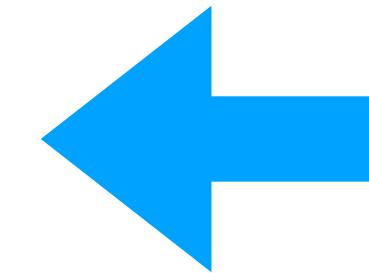
- Ser almacenados
- Administrados
- Analizados

Introducción

► ¿Qué es Minería de Datos?

Para extraer conocimiento, los datos necesitan...

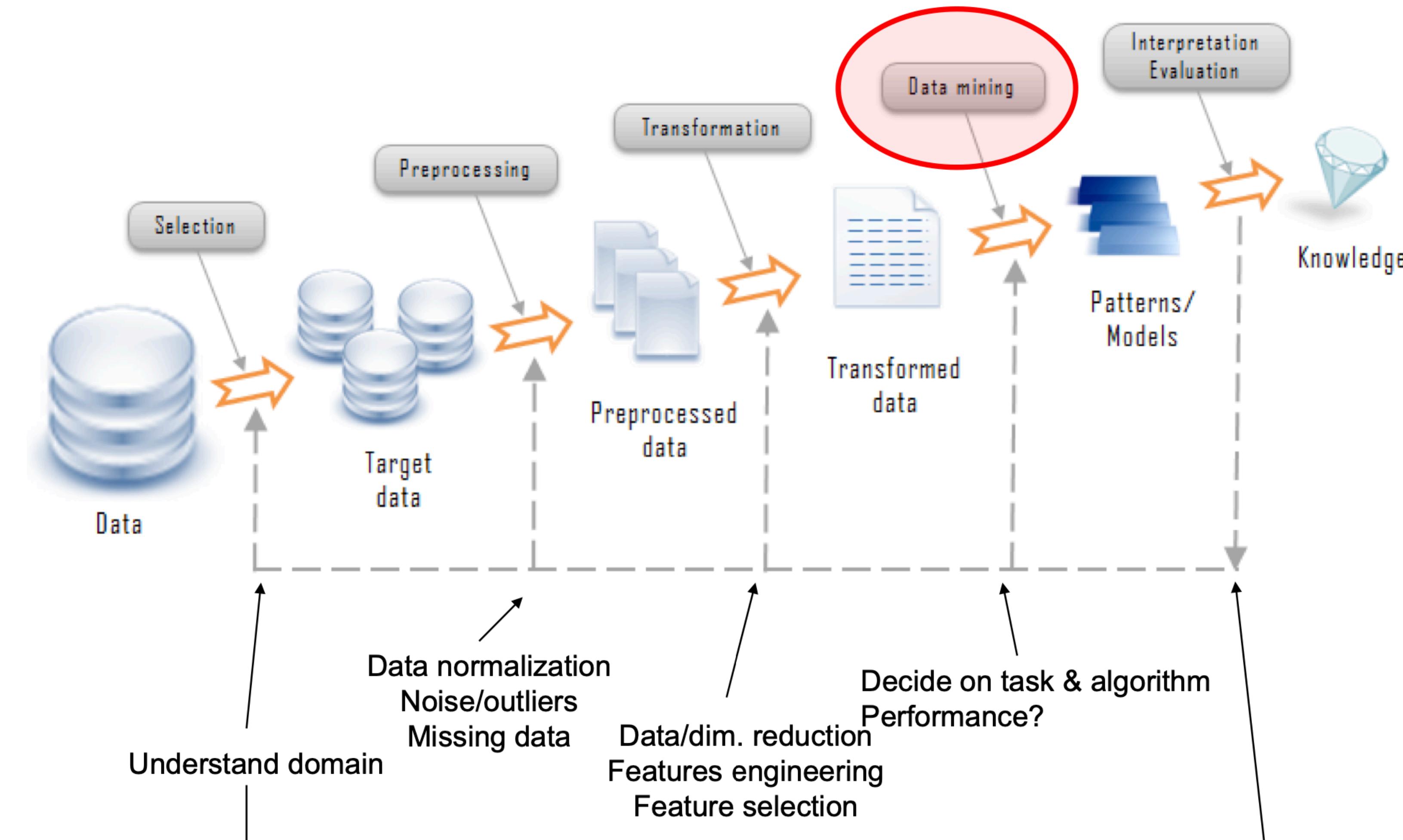
- Ser almacenados
- Administrados
- Analizados



Data mining ≈ big data ≈ Analítica Predictiva ≈ Data Science

Introducción

► Knowledge Discovery in Database (KDD) Process



Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery: an overview.

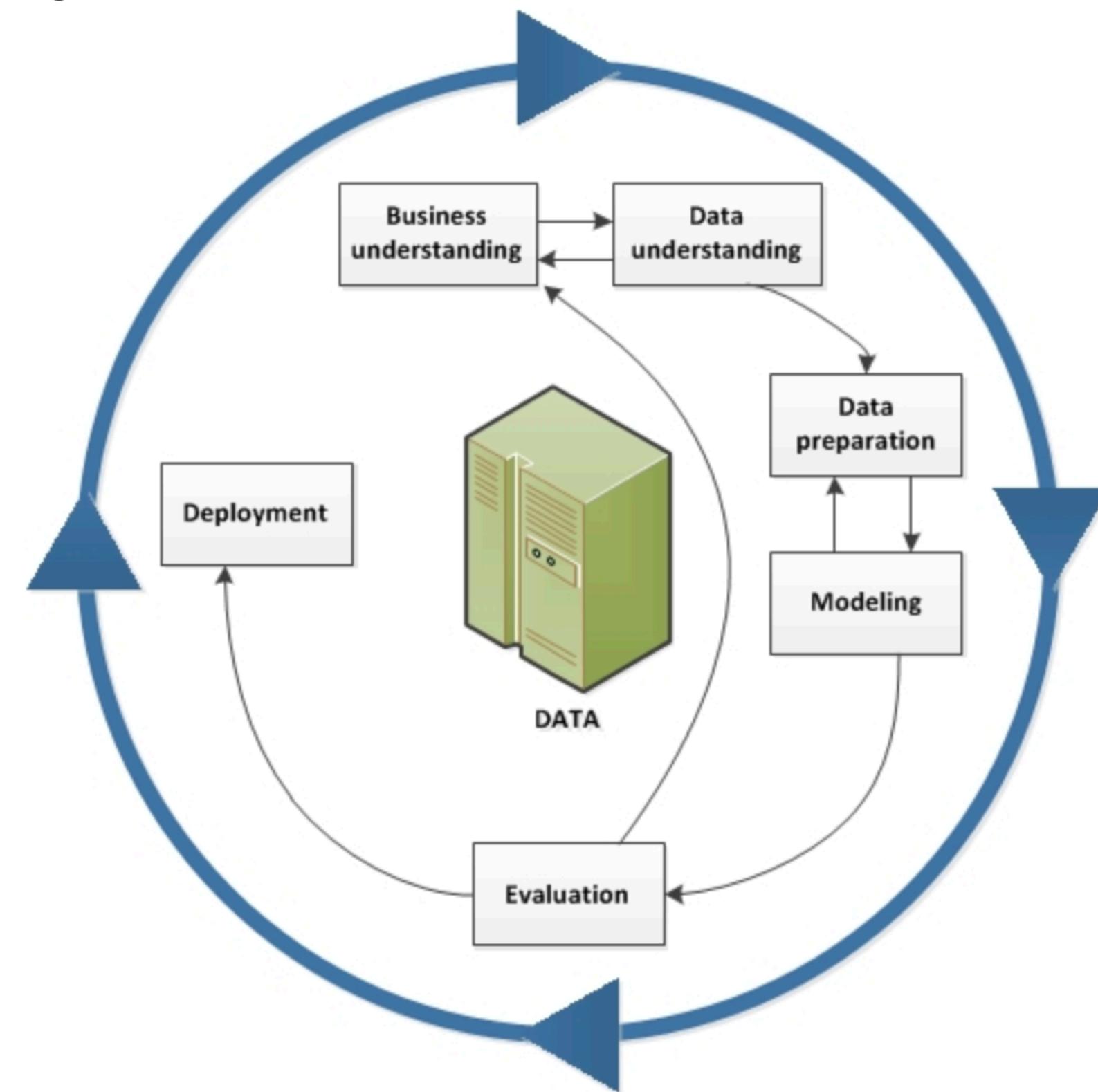
Introducción

► CRISP-DM

CRISP-DM, que son las siglas de Cross-Industry Standard Process for Data Mining, es un método probado para orientar sus trabajos de minería de datos.

Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.

Como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de minería de datos.



Introducción

► Tareas en el modelo CRISP-DM

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives <i>Background</i> <i>Business Objectives</i> <i>Business Success Criteria</i>	Collect Initial Data <i>Initial Data Collection Report</i>	Select Data <i>Rationale for Inclusion/Exclusion</i>	Select Modeling Techniques <i>Modeling Technique</i> <i>Modeling Assumptions</i>	Evaluate Results <i>Assessment of Data Mining Results w.r.t. Business Success Criteria</i> <i>Approved Models</i>	Plan Deployment <i>Deployment Plan</i>
Assess Situation <i>Inventory of Resources Requirements, Assumptions, and Constraints</i> <i>Risks and Contingencies</i> <i>Terminology</i> <i>Costs and Benefits</i>	Describe Data <i>Data Description Report</i>	Clean Data <i>Data Cleaning Report</i>	Generate Test Design <i>Test Design</i>	Review Process <i>Review of Process</i>	Plan Monitoring and Maintenance <i>Monitoring and Maintenance Plan</i>
Determine Data Mining Goals <i>Data Mining Goals</i> <i>Data Mining Success Criteria</i>	Explore Data <i>Data Exploration Report</i>	Construct Data <i>Derived Attributes</i> <i>Generated Records</i>	Build Model <i>Parameter Settings</i> <i>Models</i> <i>Model Descriptions</i>	Determine Next Steps <i>List of Possible Actions</i> <i>Decision</i>	Produce Final Report <i>Final Report</i> <i>Final Presentation</i>
Produce Project Plan <i>Project Plan</i> <i>Initial Assessment of Tools and Techniques</i>	Verify Data Quality <i>Data Quality Report</i>	Integrate Data <i>Merged Data</i>	Format Data <i>Reformatted Data</i>	Assess Model <i>Model Assessment</i> <i>Revised Parameter Settings</i>	Review Project <i>Experience Documentation</i>

Figure 3: Generic tasks (**bold**) and outputs (*italic*) of the CRISP-DM reference model

Introducción

► Relación con otras disciplinas

Bases de datos. De donde provienen los datos. Técnicas de indexación y acceso a datos.

Recuperación de la información. Obtener información a partir de datos textuales. Ejemplo: clasificación de documentos en función de palabras clave.

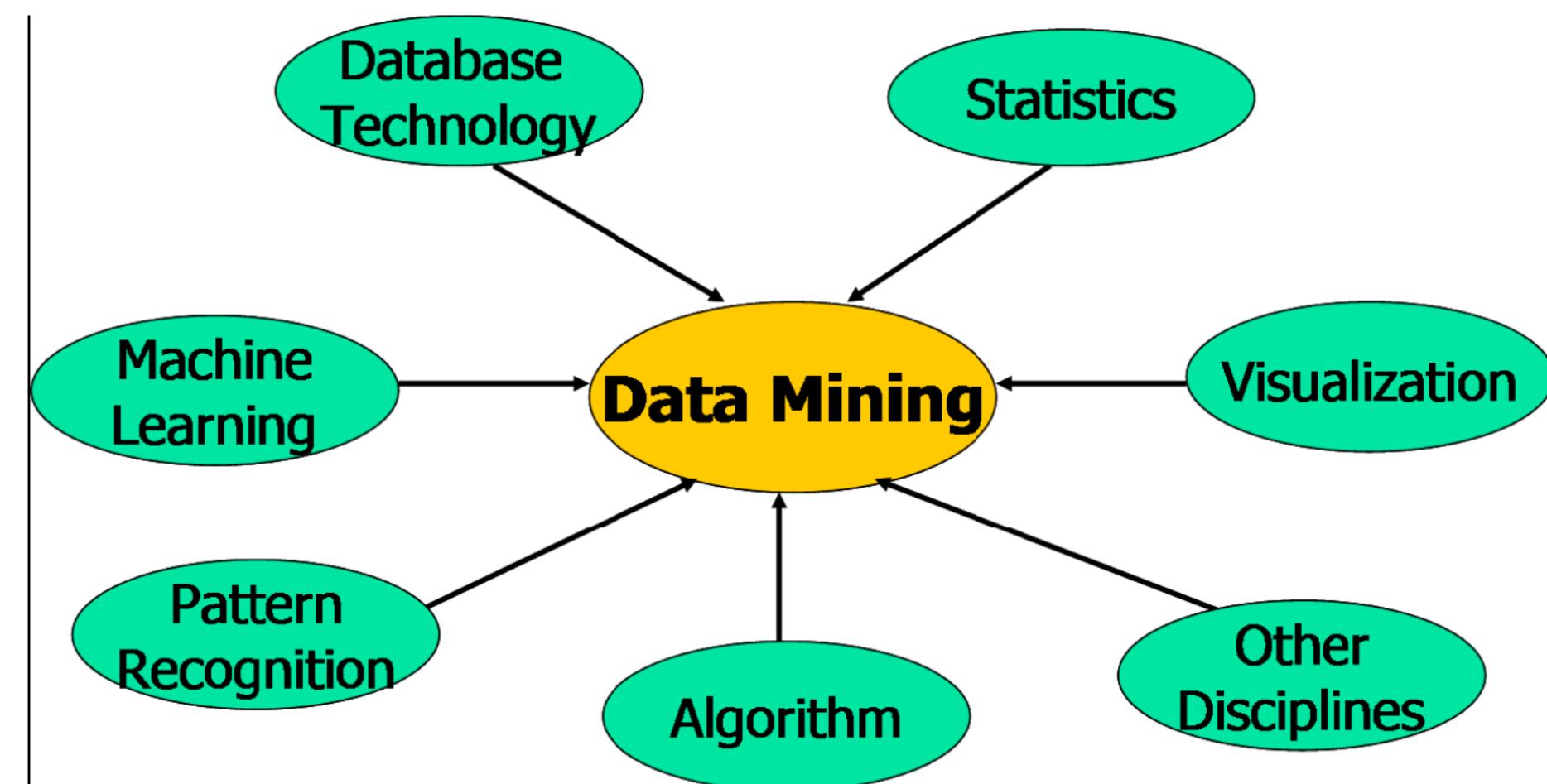
Estadística. Fuente de conceptos, algoritmos, técnicas. Comprobar hipótesis frente a encontrar hipótesis.

Aprendizaje automático. Área de la IA, algoritmos capaces de aprender.

Visualización de datos. Describir, intuir o entender patrones. Difíciles de comprender a partir de fórmulas matemáticas o descripciones textuales.

Computación paralela y distribuida. Elevado coste computacional de las tareas más complejas.

Otras. Dependientes del tipo de datos. Procesamiento de lenguaje natural, análisis de imágenes, procesamiento de señales.



Introducción

► Tareas en Minería de Datos:

Métodos Descriptivos

Encontrar patrones, correlaciones, grupos, anomalías que resuman las relaciones implícitas en los datos y que sean interpretables por el analista.

Métodos Predictivos

El objetivo de esta tarea es predecir el valor particular de un atributo basado en los valores de otros atributos.

Introducción

► Tareas en Minería de Datos:

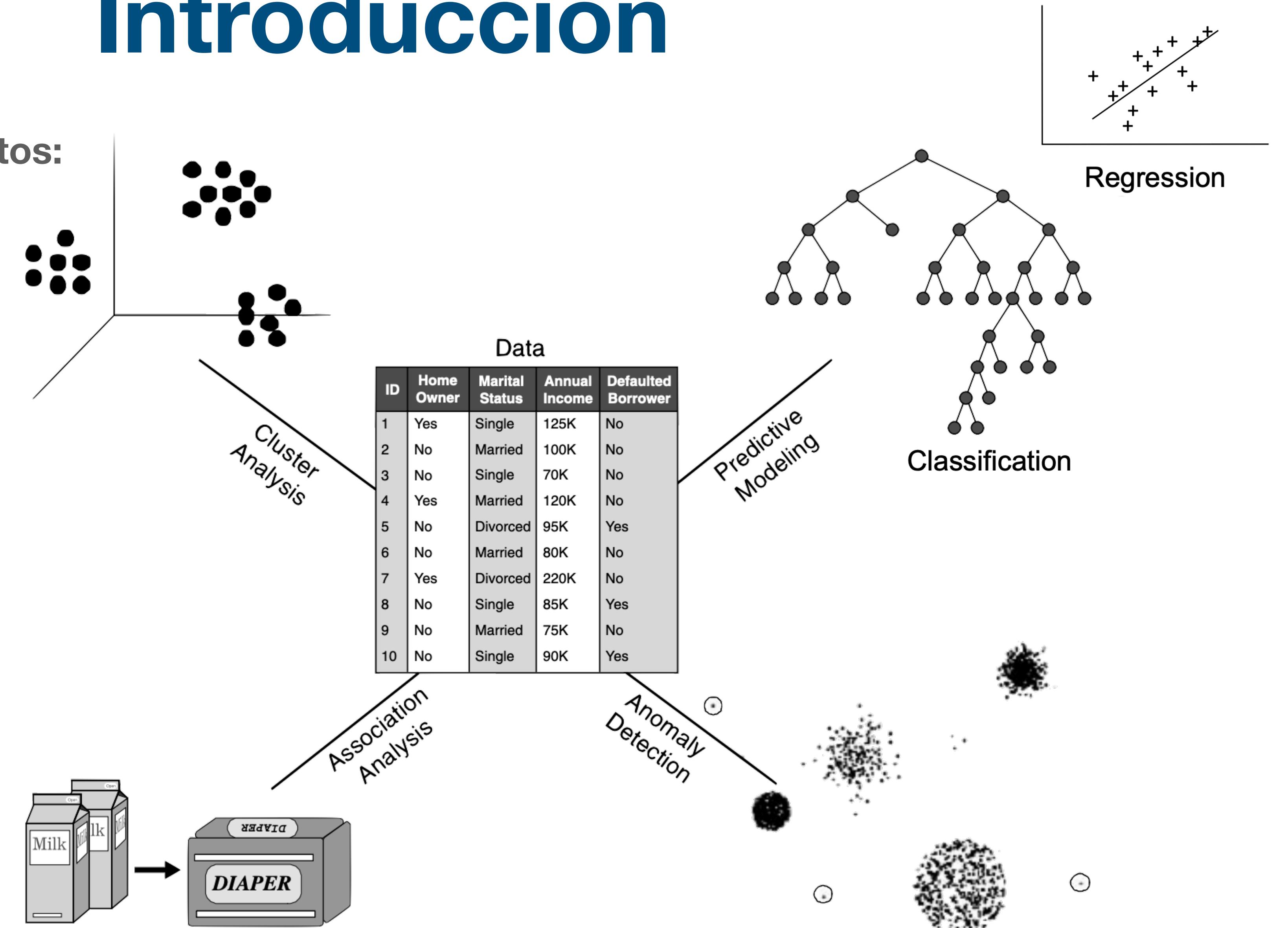


Figure 1.3. Four of the core data mining tasks.

Introducción

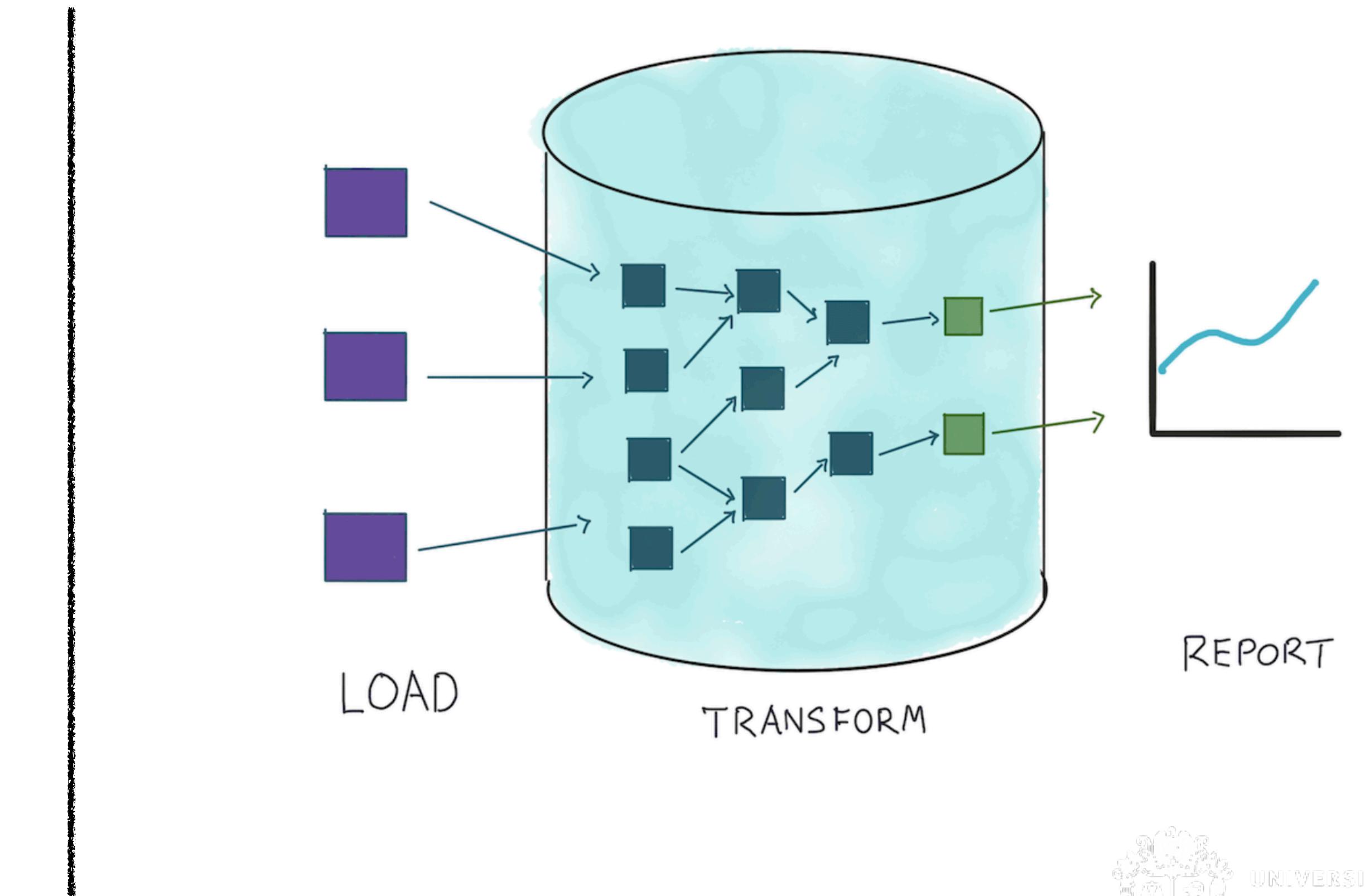
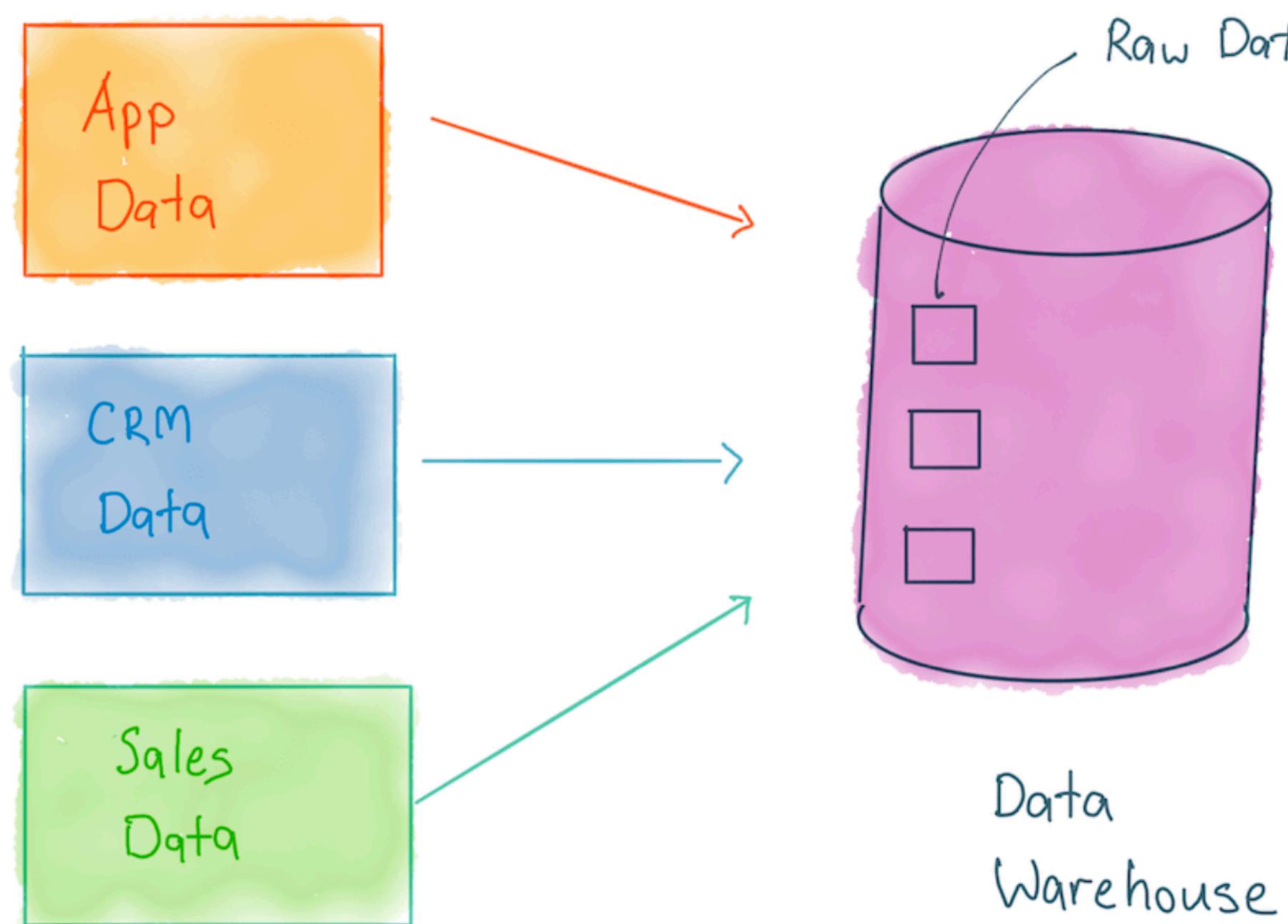
► Tareas en Minería de Datos:

Llegados a este punto surge una pregunta obligada, ¿a qué tipo de datos puede aplicarse la minería de datos? En principio, ésta puede aplicarse a cualquier tipo de información, siendo las técnicas de minería diferentes para cada una de ellas. En esta sección damos una breve introducción a algunos de estos tipos. En concreto, vamos a diferenciar entre datos estructurados provenientes de bases de datos relacionales, otros tipos de datos estructurados en bases de datos (espaciales, temporales) e información no estructurada (multimedia).

Introducción

► Bases de datos:

Un Data Warehouse es un almacén electrónico donde generalmente una empresa u organización mantiene una gran cantidad de información. Los datos de un data warehouse deben almacenarse de forma segura, fiable, fácil de recuperar y fácil de administrar.



Introducción

► Bases de Datos Relacionales:

Una base de datos relacional es una colección de relaciones (tablas). Cada tabla consta de un conjunto de atributos (columnas o campos) y puede contener un gran número de tuplas (registros o filas). Cada tupla representa un objeto, el cual se describe a través de los valores de sus atributos y se caracteriza por poseer una clave única o primaria que lo identifica. Por ejemplo, la siguiente figura ilustra una base de datos con dos relaciones: empleado y departamento. La relación empleado tiene seis atributos: el identificador o clave primaria (IdE), el nombre del empleado (Enombre), su sueldo (Sueldo), su edad (Edad), su sexo (Sexo) y el departamento en el que trabaja (IdD), y la relación departamento tiene tres atributos: su identificador o clave primaria (IdD), el nombre (Dnombre) y su director (Director). Una relación puede además tener claves ajenas, es decir, atributos que hagan referencia a otra relación, como por ejemplo el sexto atributo de la relación empleado, IdD, que hace referencia (por valor) al IdD de departamento.

Introducción

► Bases de Datos Relacionales:

<i>empleado</i>					
<i>IdE</i>	<i>Enombre</i>	<i>Sueldo</i>	<i>Edad</i>	<i>Sexo</i>	<i>IdD</i>
1	Juan	2.100	45	H	Ge
2	Elena	2.400	40	M	Ma
3	María	?	53	M	Ge
4	Pedro	1.000	20	H	Ge
5	Lucía	3.500	35	M	Ma

<i>departamento</i>		
<i>IdD</i>	<i>Dnombre</i>	<i>Director</i>
Ge	Gestión	Rubio
Ve	Ventas	Burriel
Ma	Márketing	Torrubia



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Gracias