UNIVERSIDAD NACIONAL DE LA MATANZA

INTELIGENCIA DE NEGOCIOS

Introducción a la Minería de Datos y la Explotación de Información

Profesor: Mg. Diego Basso

Curso 2017

Ejemplo de un Problema Motivador

#	EDAD	PADECIMIENTO	ASTIGMATISMO	LAGRIMEO	TIPO DE LENTE
1	joven	hipermétrope	si	reducido	ninguno
2	pre-presbiópico	miope	no	reducido	ninguno
3	joven	hipermétrope	no	normal	suave
4	presbiópico	miope	no	reducido	ninguno
5	joven	miope	si	reducido	ninguno
6	pre-presbiópico	hipermétrope	si	reducido	ninguno
7	joven	hipermétrope	si	normal	duro
8	pre-presbiópico	hipermétrope	no	reducido	ninguno
9	presbiópico	hipermétrope	si	normal	ninguno
10	presbiópico	hipermétrope	si	reducido	ninguno
11	joven	miope	si	normal	duro
12	joven	miope	no	reducido	ninguno
13	pre-presbiópico	hipermétrope	no	normal	suave
14	presbiópico	hipermétrope	no	normal	suave
15	presbiópico	miope	si	reducido	ninguno
16	joven	hipermétrope	no	reducido	ninguno
17	presbiópico	miope	no	normal	ninguno
18	pre-presbiópico	miope	si	reducido	ninguno

- ¿Qué tipo de lentes necesita bajo estas condiciones?

joven	miope	no	normal	?

Otro Problema

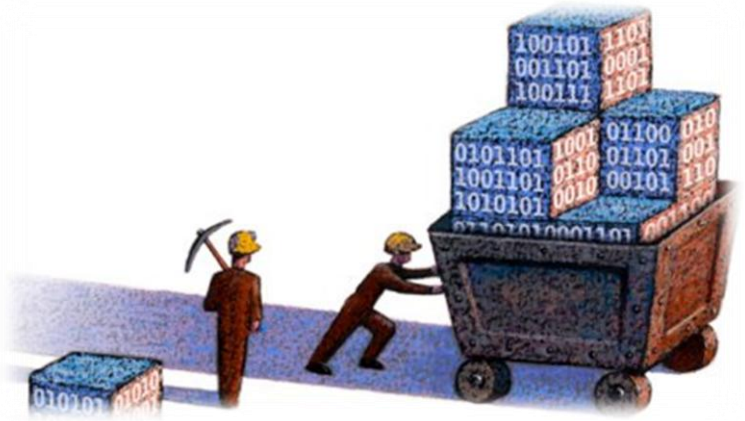
- Un banco trata de evitar que sus clientes se vayan a otros bancos. Por eso quieren detectar los *signos tempranos de deserción*, es decir las actitudes que toma un cliente antes de irse del banco.
- Esto le permitirá identificar qué clientes están haciendo eso mismo, e intentar retenerlos mientras son clientes .
- El banco tiene información histórica sobre movimientos de los clientes (entre ellos la baja), y quiere usar esa información para detectar cuáles son estos signos, para luego identificar qué clientes presentan los mismos "comportamientos".

La Solución con OLAP

- Se formulan hipótesis:
  - ✓ *Los clientes que no renovaron plazos fijos tienen tendencia a irse.*
  - ✓ *Los clientes que disminuyeron sus operaciones de cajero automático tienen tendencia a irse.*
  - ✓ *Los clientes que cerraron cuentas tienen tendencia a irse.*
- Utilizando las variables adecuadas del DW se analiza qué porcentaje de clientes en esas condiciones se fueron del banco.
  - Esto confirma o rechaza las hipótesis.
- Puede ocurrir que no hayan respuestas satisfactorias.

Una Nueva Técnica

- Esta técnica, que permite extraer el conocimiento necesario de los datos históricos para resolver el problema, se denomina **DATA MINING** (o Minería de Datos).



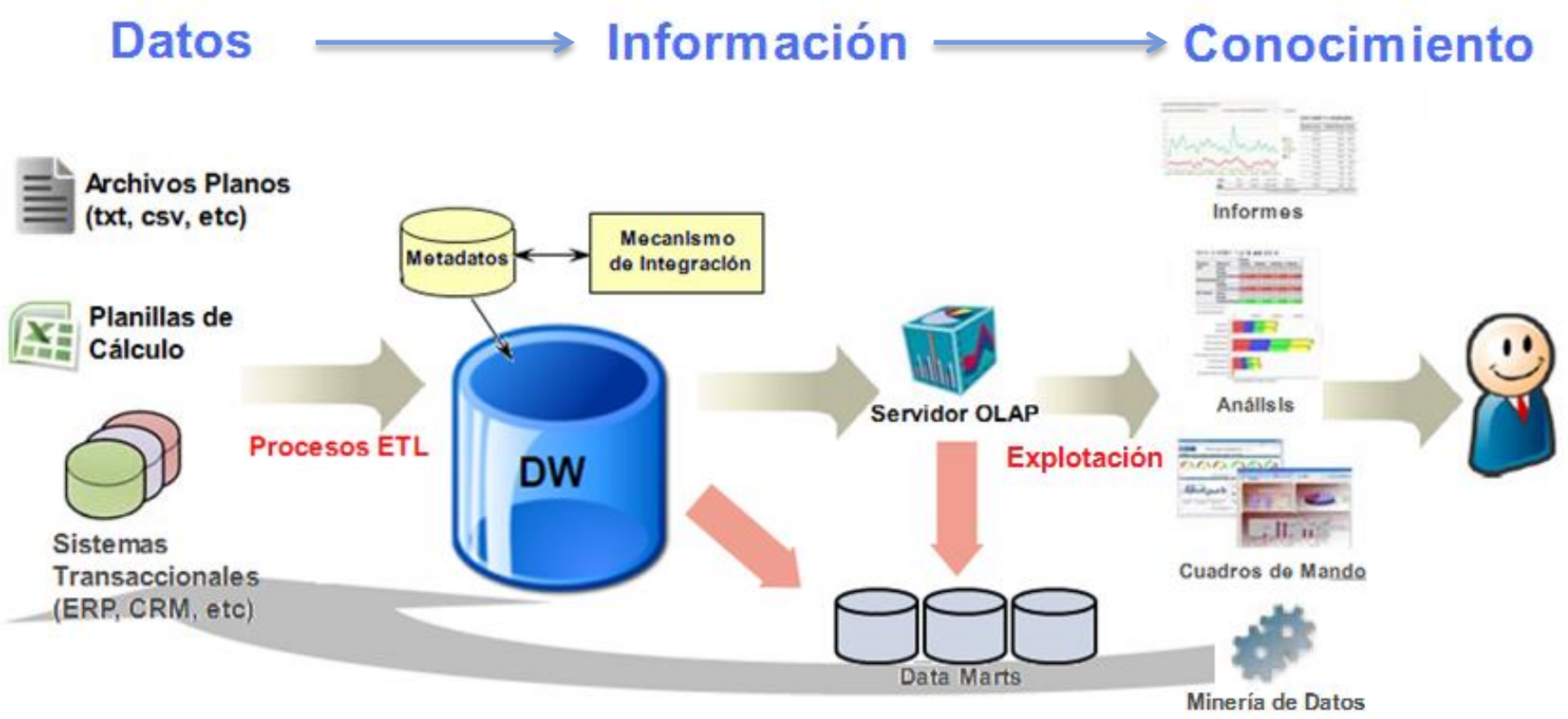
Motivación de la Minería de Datos

- Necesidad de analizar grandes volúmenes de datos para obtener información desconocida que sea útil para tomar decisiones.
  - Volumen y variedad de información informatizada que **desborda la capacidad humana**.
  - Uso de técnicas que imiten la cualidad humana del **aprendizaje**, es decir, con capacidad de extraer nuevo conocimiento a partir de experiencias (ejemplos).
  - Las **decisiones** se basan en la información de **experiencias** pasadas extraídas de fuentes muy diversas.
  - Se cuenta con información histórica que es útil para predecir.

Minería de Datos

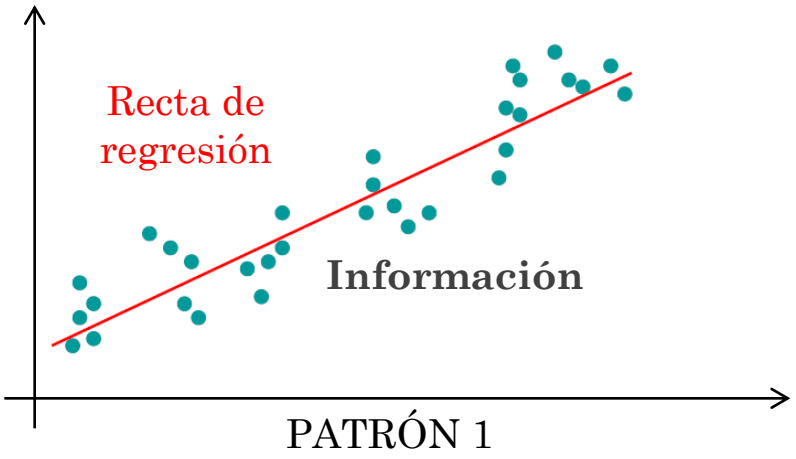
- Proceso automático que permite extraer y descubrir patrones de **conocimiento interesantes**, **no triviales**, **previamente desconocidos** y **potencialmente útiles** de los datos y descubrir relaciones entre variables.
- Sirve de ayuda en el proceso de toma de decisiones, formando parte del conjunto de tecnologías aplicables a la Inteligencia de Negocio.
- Fase del proceso de "*Descubrimiento de Conocimiento a partir de Bases de Datos*" (KDD, del inglés Knowledge Discovery from Databases), aunque los términos suelen ser usados como sinónimos.

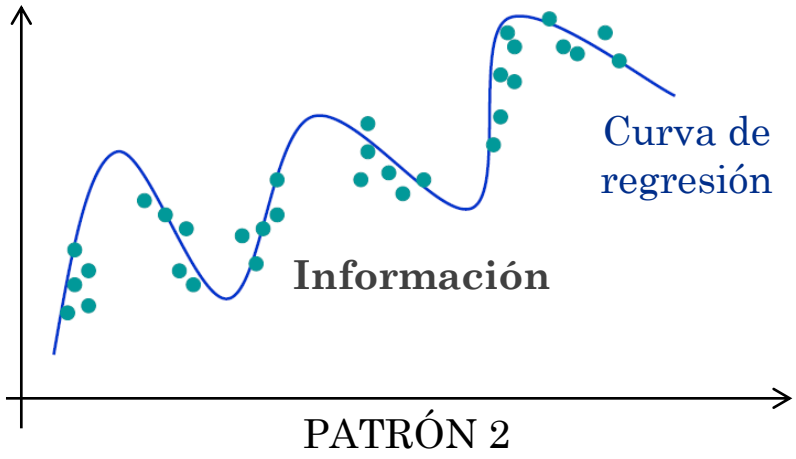
Arquitectura BI

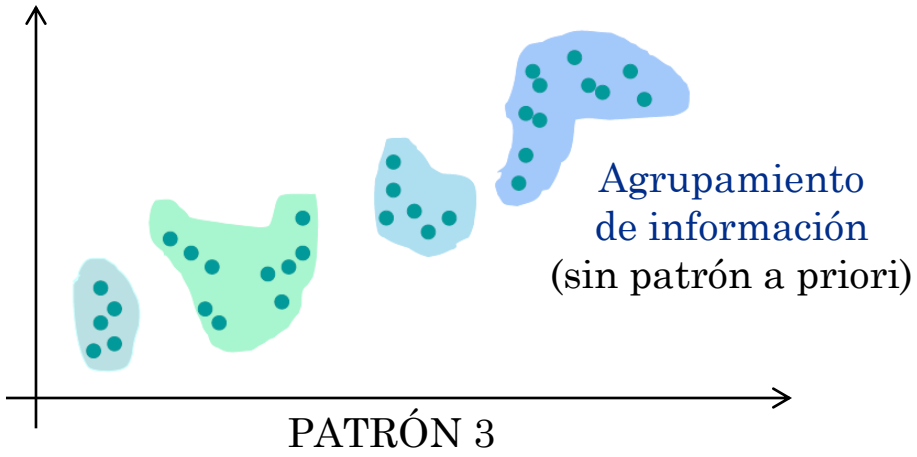


Patrón de Conocimiento

- Es una unidad o pieza de conocimiento que nos resume una información.







OLAP vs Minería de Datos

Herramientas OLAP	Minería de Datos
Facilidad para manejar y transformar datos.	Extrae patrones a partir de los datos, se construyen modelos, descubre relaciones entre atributos, tendencias, etc.
Producen **información** (datos agregados y combinados, medidas derivadas)	Produce patrones de **conocimiento** a partir de reglas.
Permite al usuario analizar los datos desde diferentes vistas.	**Analiza** los datos y ayuda al usuario a tomar decisiones a partir del conocimiento descubierto.

OLAP vs Minería de Datos

- El análisis OLAP puede responder a preguntas como:
  - ¿Han subido las ventas en el mes de Abril?
  - Las ventas del producto X bajan cuando se promociona el producto Y?
  - ¿Venden más las sucursales del Gran Buenos Aires o del Interior?
- La minería de datos puede responder a preguntas como:
  - ¿Qué factores influyen en la venta del producto X?
  - ¿Cuál será el producto más vendido si se abre una sucursal en Córdoba?
  - ¿Cuándo un cliente compra el producto Y, qué otro/s producto/s suele comprar mayormente?

Requerimientos

- ¿Qué se necesita para hacer minería de datos?
  - Herramientas de SW
  - Datos, digitalizados y de buena calidad
  - RRHH especialistas: técnico, analítico y de negocios

Áreas de Aplicación

- Comercio / Marketing
  - Identificar patrones de compra de los clientes.
  - Buscar asociaciones entre clientes y características demográficas.
  - Predecir respuesta a campañas de mailing.
- Análisis de canasta de compra

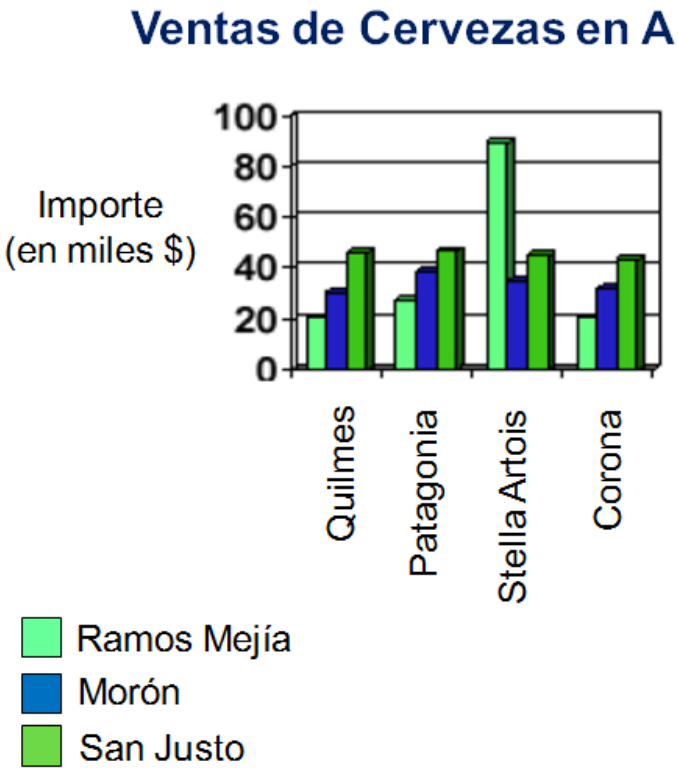
Áreas de Aplicación

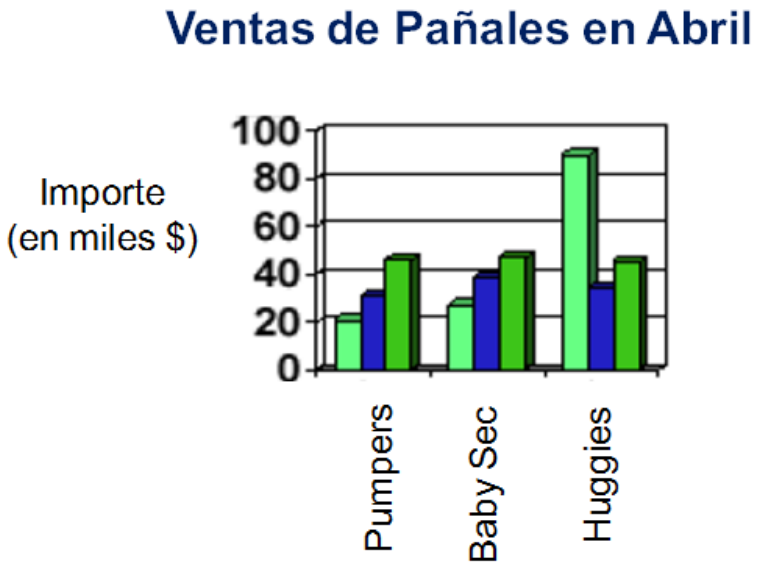
- Bancos
  - Detectar patrones de uso fraudulento de tarjetas de crédito.
  - Identificar clientes leales.
  - Predecir clientes con probabilidad de darse de baja.
  - Determinar gasto en tarjetas de crédito por grupos.
  - Encontrar correlaciones entre indicadores financieros.

Áreas de Aplicación

- Salud Privada
  - Identificar patrones de comportamiento de clientes con alto riesgo.
  - Análisis de procedimientos médicos.
- Medicina
  - Segmentación de pacientes para una atención más inteligente según su grupo.
  - Estudio de factores (genéticos, neurológicos, alimenticios, etc.) de riesgo/salud en distintas patologías.
  - Identificación de terapias médicas satisfactorias para diferentes enfermedades.

Caso de Estudio: Marketing-Ventas





- Si se realiza sólo la toma de decisión en función de los informes (datos) de ventas de cervezas y pañales.

¿Qué información aporta?

Caso de Estudio: Marketing-Ventas

- *Objetivo*: determinar grupos de ítems que tienden a ocurrir juntos en una misma transacción de compra.
- Utilizando minería de datos se puede descubrir información como:
  - Los clientes que compran cervezas también compran papas fritas y leche.
  - Los viernes por la tarde, con frecuencia, quienes compran pañales también compran cerveza.
- ¿Qué significa esto? ¿A qué se debe?
- ¿Qué acciones debemos realizar?

Caso de Estudio: Marketing-Ventas

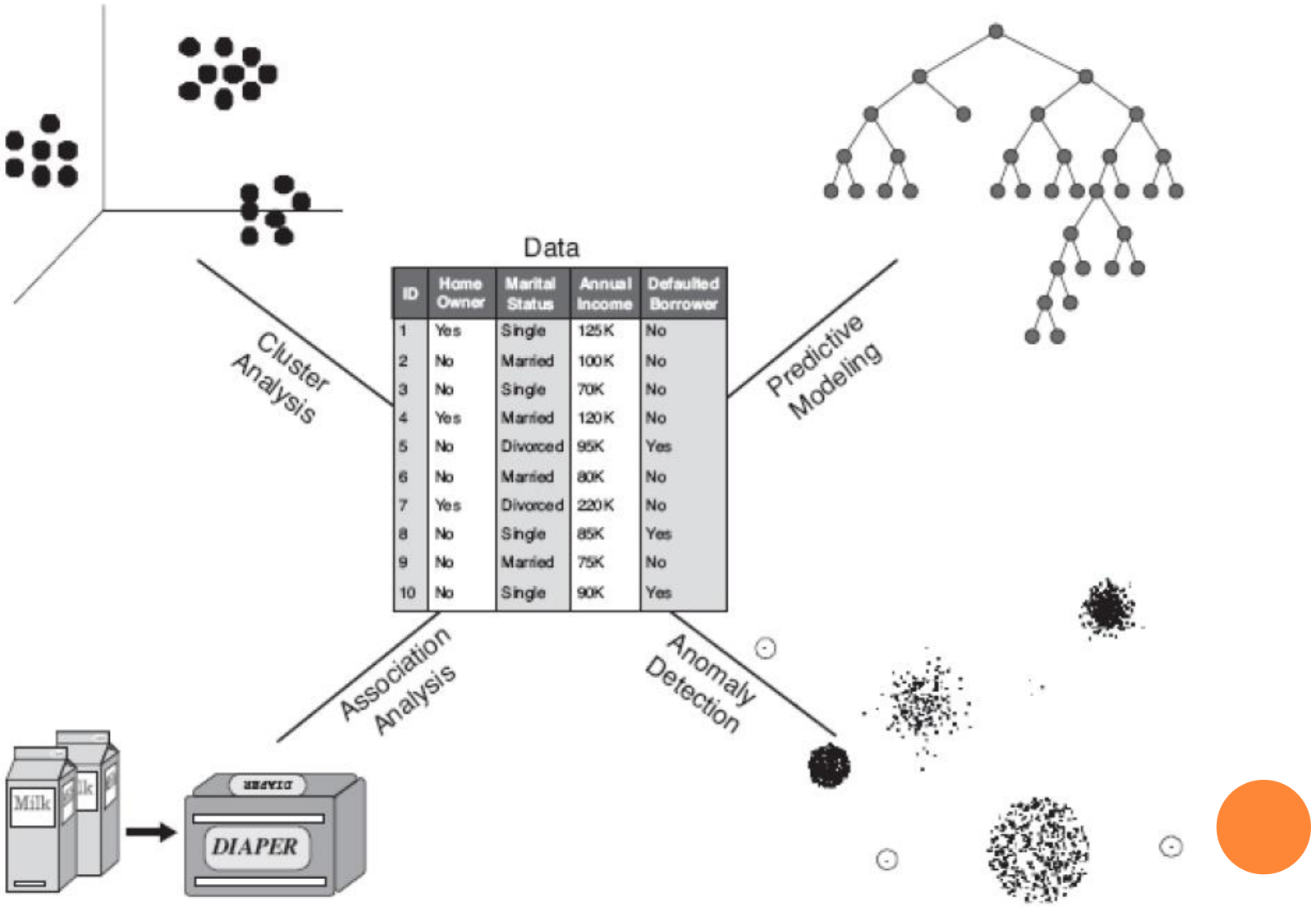
- Algunas explicaciones probables:
  - Se acerca el fin de semana
  - Hay un bebé en casa
  - Los padres no pueden salir!
  - No quedan pañales
  - Se compra cerveza para ver un partido/película

- Aparecen asociaciones: Pañales → Cerveza
  Pañales → Cerveza, Leche

Caso de Estudio: Marketing-Ventas

- Acciones a realizar:
  - Planificar la disposición de los productos en las góndolas:
    - Las leches al lado de los alimentos lácteos para bebés y niños
    - Las cervezas frente a la góndola de snacks.
  - Poner los aperitivos que más margen dejan entre los pañales y las cervezas.
  - Poner ofertas de pañales.
  - Poner productos de bebés en oferta y cerca de las cervezas.
  - Ofrecer cupones de descuento para el producto "complementario", cuando uno de los productos se venda por separado.

Tareas de Minería de Datos



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Técnicas de Minería de Datos

- Las técnicas de minería de datos son herramientas que facilitan el descubrimiento de conocimiento.

Técnicas de Minería de Datos

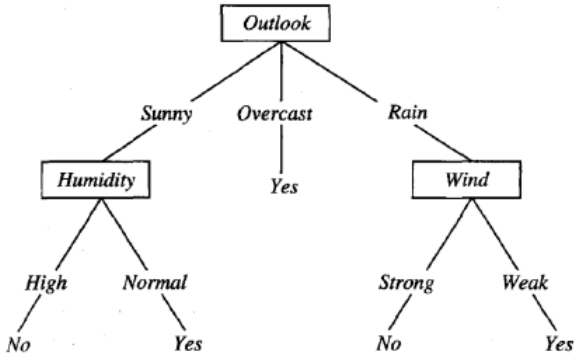
Métodos Predictivos – Supervisados

- Usan algunas variables para predecir valores desconocidos de otras variables.
- Ejemplos
  - ¿Esta transacción es fraudulenta?
  - ¿Qué tipo de seguro es más probable que contrate el cliente Carlos Gómez?

Técnicas de Minería de Datos

- Tareas de Clasificación
  - Predicen un valor discreto
    - SI / NO
    - Alto / Mediano / Bajo

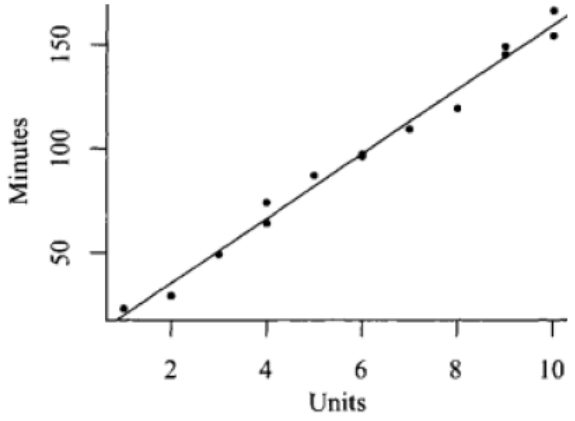
Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes



- Tareas de Regresión
  - Predicen un valor continuo
    - Importes
    - Cantidades

Row	Minutes	Units
1	23	1
2	29	2
3	49	3
4	64	4
5	74	4
6	87	5
7	96	6

$$\text{Minutes} = 4.162 + 15.509 \cdot \text{Units}$$



Técnicas de Minería de Datos

Métodos Descriptivos – No Supervisados

- Encuentran patrones interpretables para las personas que describen los datos.
- Proporcionan información sobre las relaciones entre los datos y sus características.
- Ejemplos
  - Los clientes que compran pañales suelen comprar cerveza.
  - El tabaco y el alcohol son los factores que más inciden en la enfermedad Y.
  - Los clientes sin televisión y con bicicleta tienen características muy diferenciadas del resto.

Técnicas de Minería de Datos

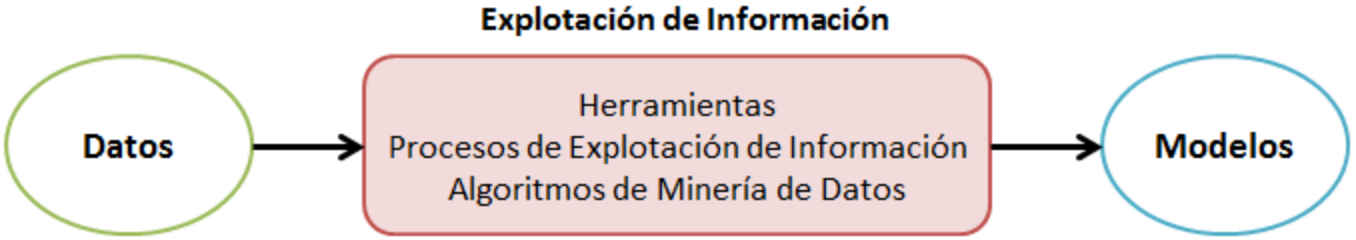
- Tareas de Asociación
  - Descubren por medio de reglas de asociación hechos que ocurren en común dentro de un determinado conjunto de datos.
  - Utilizado en análisis de canasta (market basket analysis).
    - {cebollas, vegetales} ⇒ {carne}
    - {cerveza} ⇒ {leche, pañales}
- Tareas de Segmentación (Clustering)
  - Agrupamiento jerárquico o no jerárquico de datos de acuerdo a un determinado criterio.
    - Jerárquico: Puede ser aglomerativo o divisivo.
    - No Jerárquico: N° Grupos determinados de antemano.

Explotación de Información

- La Explotación de Información es la sub-disciplina informática que aporta a la Inteligencia de Negocios las herramientas (procesos y tecnologías) para la transformación de información en conocimiento.
- Utiliza la Minería de Datos.
- Aborda la solución a problemas de predicción, clasificación y segmentación.
- La minería de datos y la explotación de información no son conceptos equivalentes.

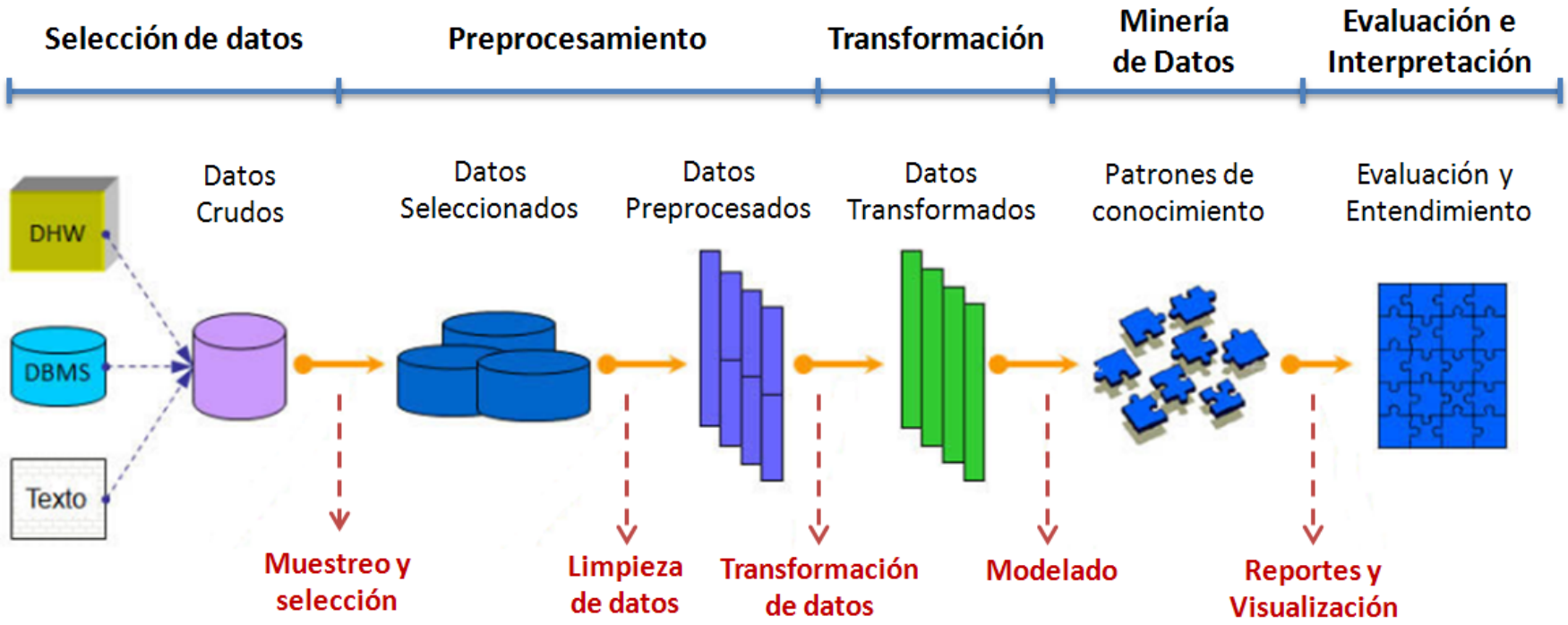
Explotación de Información

- La minería de datos está relacionada a la programación, a los algoritmos para resolver un problema de inteligencia de negocios.
- La explotación de información está relacionada a tareas de la Ingeniería de Software, a la aplicación de técnicas y procesos ingenieriles para construir la solución de un problema de inteligencia de negocios.
- La minería de datos describe la tecnología que da soporte a la explotación de la información.



Proceso Descubrimiento de Conocimiento

- También conocido como KDD, del inglés Knowledge Discovery in Databases.



Proceso Descubrimiento de Conocimiento

- **Selección de datos**: Datos sobre los que se trabajará.
- **Preprocesamiento**: Preparación y limpieza de los datos. Estrategias para manejar datos faltantes o nulos, datos inconsistentes o que están fuera de rango.
- **Transformación**: Tratamiento preliminar de los datos, transformación, agregación, normalización y generación de nuevas variables a partir de los datos existentes.

Proceso Descubrimiento de Conocimiento

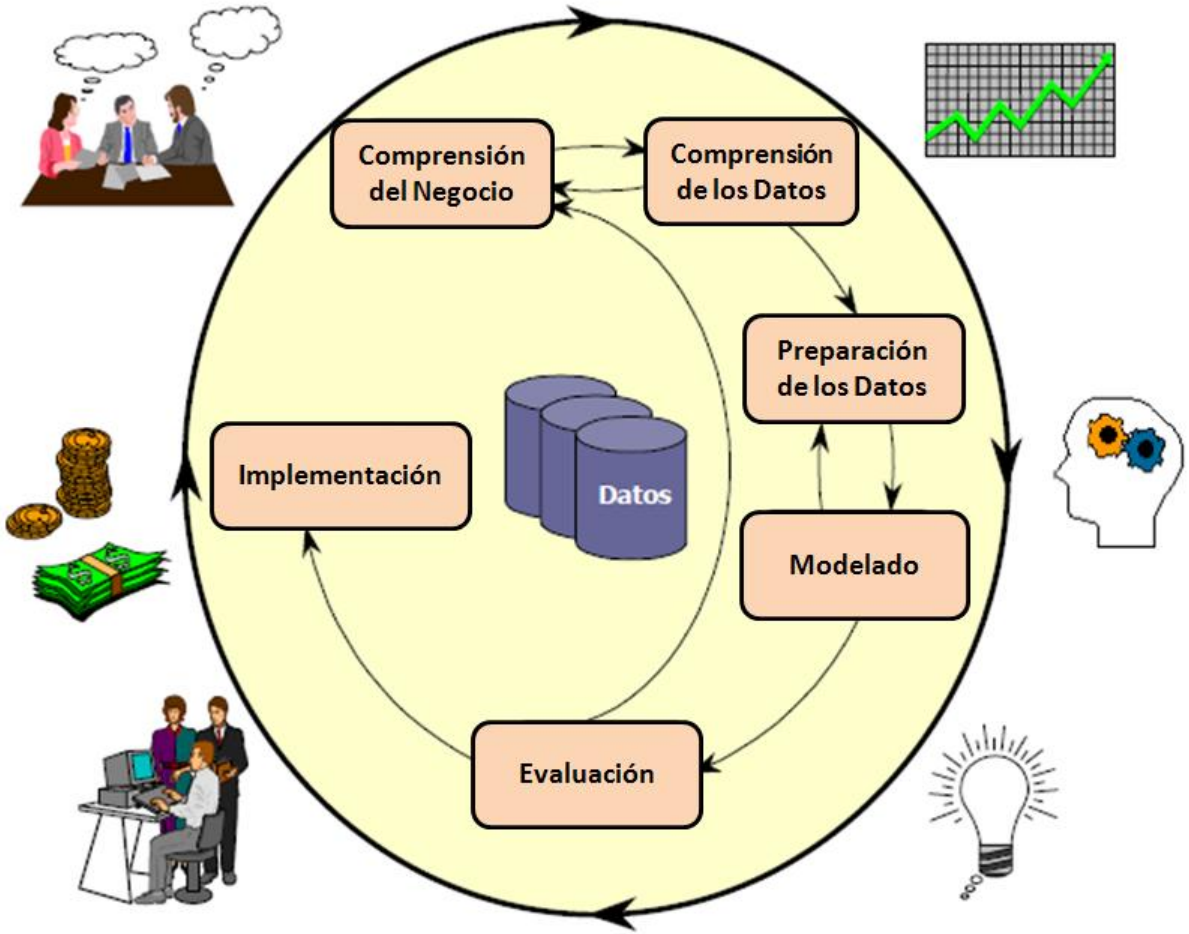
- **Minería de Datos**: Construcción de modelos con técnicas de minería de datos y procesos de explotación de información para extracción de patrones de conocimiento.
  - Técnicas Predictivas
  - Técnicas Descriptivas
- **Evaluación e interpretación**: Evaluación del modelo construido, del conocimiento obtenido y validación si los resultados son satisfactorios en el dominio del problema.

Metodologías de Explotación de Información

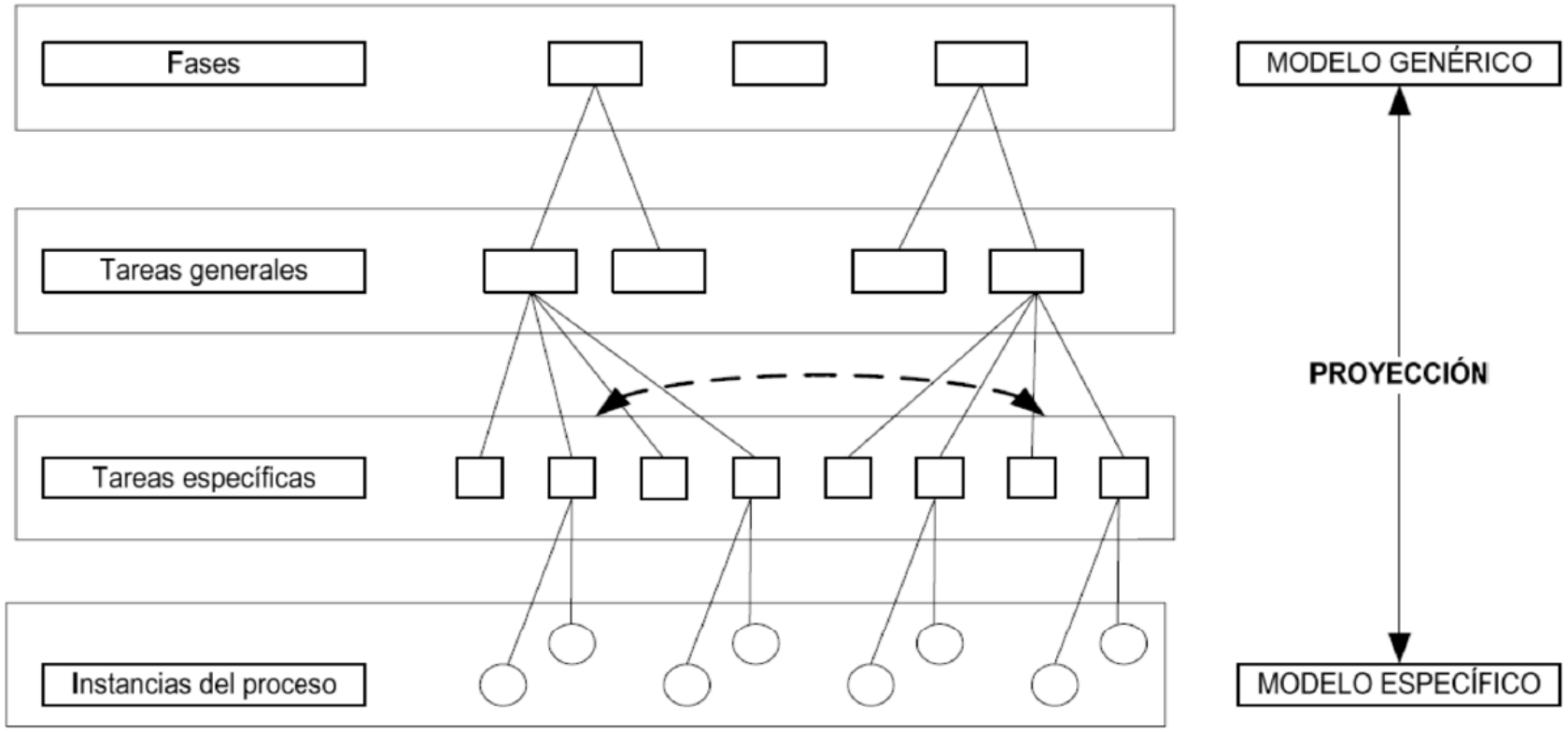
- Conjunto de actividades organizadas que tienen como objetivo la realización de un proyecto de explotación de información.
- Para cada actividad se define, las entradas, las salidas y la forma en la que debe llevarse a cabo.
- Metodologías probadas por la comunidad científica:
  - CRISP-DM
  - SEMMA
  - P^3TQ

Metodología CRISP-DM

- **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining.



METODOLOGÍA CRISP-DM



Esquema de los cuatro Niveles de Abstracción de CRISP-DM

Metodología CRISP-DM

Comprensión del Negocio

- Se determinan los objetivos y requerimientos del proyecto desde una perspectiva del negocio, definiendo el problema de minería y el plan de trabajo.
  - Objetivos de negocio y criterios de éxito
    - Detectar fraude con tarjetas de crédito
    - Captar nuevos clientes bancarios
    - Detectar signos tempranos de algún padecimiento clínico
    - Etc.
  - Análisis del problema
  - Objetivos de minería de datos

METODOLOGÍA CRISP-DM

Comprensión de los Datos

- Se recolectan los datos que se utilizarán y se analizan las características de los mismos. Surgen las primeras hipótesis acerca de la información que podría estar oculta.
- Atributos Nominales
  - Llamados Categóricos o Discretos
  - Número finito de valores, no tienen orden.
    - Ejemplo: género, color de ojos, sucursales, booleanos, etc.
- Atributos Ordinales
  - Llamados Numéricos o Continuos
  - Número finito de valores (enteros o reales), tienen orden
    - Ejemplo: puntuación, rangos, altura, importes, temperaturas, fechas, etc.

Metodología CRISP-DM

Preparación de los Datos

- Comprenden actividades de tratamiento de los datos o conjunto de datos final sobre el cual se aplicarán procesos de explotación de información y minería de datos.
  - Selección, Limpieza y Transformación
- Análisis de la calidad de los datos
  - ¿Qué tipos de problemas de calidad podemos encontrar?
    - Valores anómalos (ruido, outlier)
    - Valores faltantes o nulos
    - Datos Duplicados
  - ¿Cómo podemos detectarlos en los datos?
  - ¿Qué podemos hacer al respecto?

Metodología CRISP-DM

- Preprocesamiento de los datos
  - Transformaciones de los datos necesarias en función del análisis previo, con el objetivo de prepararlos para aplicarles explotación de información según el problema de negocio.
    - Agregación
    - Seleccionar conjunto de atributos
    - Creación de atributos
    - Discretización
    - Transformación de atributos

Metodología CRISP-DM

- **Modelado**: se aplican procesos de explotación de información y algoritmos de minería sobre el conjunto de datos para obtener información oculta y patrones de conocimiento.
- **Evaluación**: se analizan los patrones obtenidos en función de los objetivos organizacionales. Se determina si se ha omitido algún objetivo importante del negocio y si el nuevo conocimiento será implementado.
- **Implementación**: se comunica e implementa el nuevo conocimiento, el cual debe ser representado de forma entendible para el usuario.

Caso de Estudio – Créditos Personales

- Un banco dispone de una muestra de 144 clientes históricos a los que se les otorgó un crédito personal.
- Las muestras contienen los siguientes atributos:
  - Nivel de ingresos
  - Composición familiar
  - Tipo de vivienda
  - Servicios que posee
  - Antecedente de otros créditos
  - Resultado del otorgamiento de crédito
- El banco quiere lanzar una línea de créditos y necesita analizar la información, en base a las siguientes necesidades:
  - Identificar criterios de otorgamiento de créditos
  - Identificar y caracterizar grupos de clientes en orden a estudiar líneas de crédito diferenciales por grupo.
  - Identificar los factores de incidencia en cada grupo de clientes con ingresos superiores a $ 15.000.

Caso de Estudio – Créditos Personales

- Comprensión de los datos

Atributo	Valor	Descripción
Ingreso	1	Entre $ 8.000 y $ 15.000
	2	Más de $ 15.000
Composición familiar	1	Soltero
	2	Casado sin hijos
	3	Casado con un hijo
	4	Casado con dos hijos
Vivienda	1	Alquila
	2	Propia
Servicios	1	Básicos
	2	Básicos y TV por cable
	3	Básicos, TV por cable y celular
Otros créditos	1	Un crédito
	2	Dos créditos
	3	Tres créditos
Otorga Crédito	Sí	Préstamo otorgado
	No	Préstamo rechazado

Caso de Estudio – Créditos Personales

- Comprensión de los datos

View dataset 1 [All] (144 examples, 6 attributes)

	Ingreso	Composición_Familiar	Vivienda	Servicios	Otros_Creditos	Otorga_Credito
1	1	1	1	1	1	si
2	1	1	1	1	2	si
3	1	1	1	1	3	no
4	1	1	1	2	1	si
5	1	1	1	2	2	si
6	1	1	1	2	3	no
7	1	1	1	3	1	si
8	1	1	1	3	2	si
9	1	1	1	3	3	no
10	1	1	2	1	1	
11	1	1	2	1	2	
12	1	1	2	1	3	
13	1	1	2	1	1	
14	1	1	2	2	2	
15	1	1	2	2	3	
16	1	1	2	3	1	
17	1	1	2	3	2	

Dataset description

6 attribute(s)
144 example(s)

Attribute	Category	Informations
Ingreso	Continue	-
Composición_Familiar	Continue	-
Vivienda	Continue	-
Servicios	Continue	-
Otros_Creditos	Continue	-
Otorga_Credito	Discrete	2 values

Herramientas para Minería de Datos

- **Licenciadas**
  - SAS (Analytics, Enterprise Miner)
  - SPSS (IBM SPSS Statistics, IBM SPSS Modeler – ex Clementine)
- **Libres**
  - WEKA (http://www.cs.waikato.ac.nz/ml/weka/)
  - Tanagra (https://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html)
  - R (https://www.r-project.org/)
  - Rapid Miner (https://rapidminer.com/)
  - Otros