



Minería de Datos

Motivación e introducción

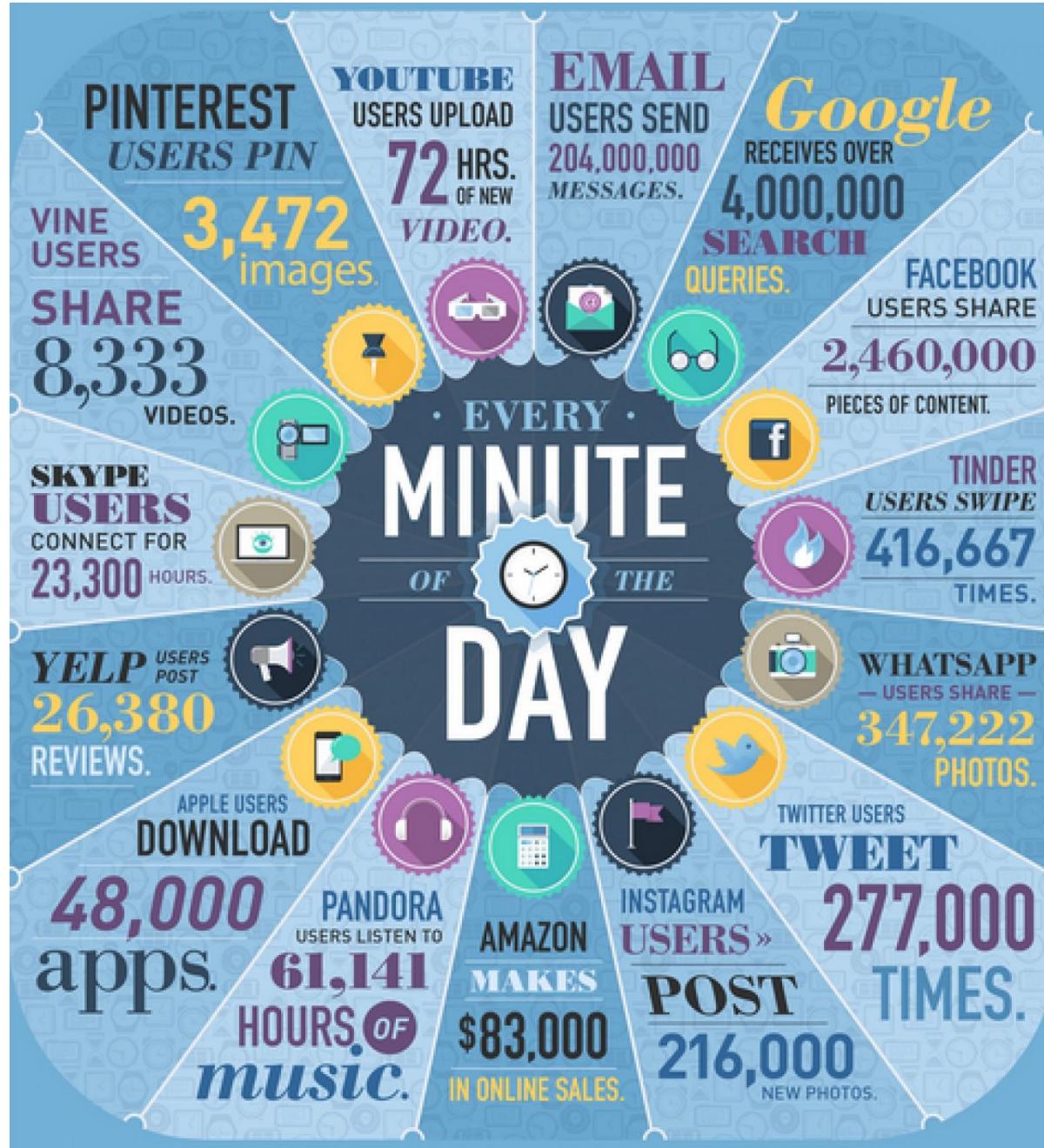


Algunos números

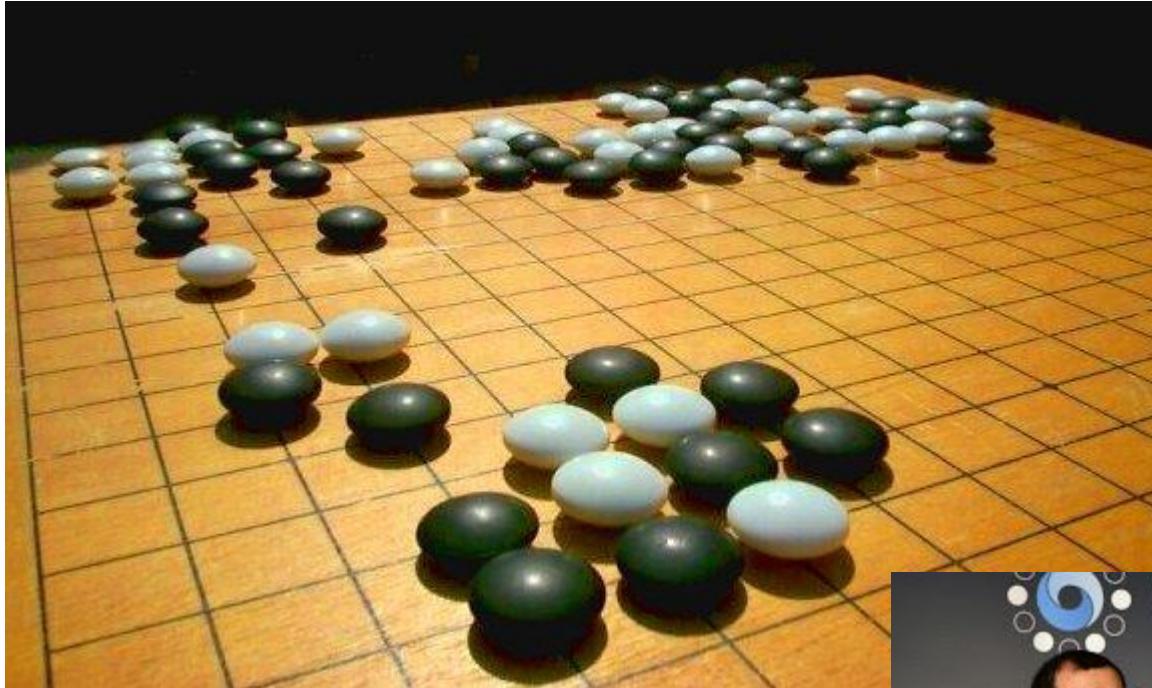
- ▶ “Big Data, for better or worse: 90% of world's data generated over last two years” – SINTEF

Cada minuto, alrededor de:

- ▶ Facebook: 2.5 millones de publicaciones.
- ▶ Twitter: twitts 300,000.
- ▶ Instagram: 220,000 nuevas fotografías.
- ▶ YouTube: 72 horas de video.
- ▶ Apple: 50,000 apps descargadas de la tienda de aplicaciones.
- ▶ Email: 200 millones de mensajes.
- ▶ Amazon: \$80,000 generados por ventas en línea.



Teaching Deep Convolutional Neural Networks to Play Go



mining
data

useful predictive process

hidden extraction

Discovery information

business databases

Twiki Generally new

info sets

allows Introduction

society Doug large

Emerging accurate

super AKA Dictionary

great ODM

needles funny different advanced haystacks science

insights important guides techniques automatic

Discovering step interdisciplinary

help basis April links talks

cool field Maybe

site

provide objectives meet

summarizing analyzing

objectives meet previously MIT

impact focus

young

Launched

Library Microsoft Server

powerful dea

Learn introduced tracor

Overview forecasting historical

ability technology unknown language

Alexander video



- Data
 - Mining
 - Información
 - Conocimiento
 - Descubrimiento
 - Escondido
 - Extracción
 - Útil
 - Proceso
 - Bases de datos
 - Grande

Definición

- **Data Mining:** *Es el proceso de descubrimiento y extracción de información útil que se encuentra escondida en grandes conjuntos de datos con la finalidad de generar conocimiento.*

Introducción

- ★ ¿Qué es Minería de Datos?
 - Extracción de información o de patrones (no trivial, implícita, previamente desconocida y potencialmente útil) de grandes bases de datos.



Introducción

★ ¿Qué es Minería de Datos?

- Es analizar datos para encontrar patrones ocultos usando medios automatizados.



Introducción



★ ¿Qué es Minería de Datos?

- La Minería de Datos es un proceso no elemental de búsqueda de relaciones, correlaciones, dependencias, asociaciones, modelos, estructuras, tendencias, clases (clústeres), segmentos, los que se obtienen de grandes juegos de datos, los cuales generalmente están almacenados en bases de datos (relacionales o no).
- Esta búsqueda se lleva a cabo utilizando métodos matemáticos, estadísticos o algorítmicos.

Introducción



★ ¿Qué es Minería de Datos?

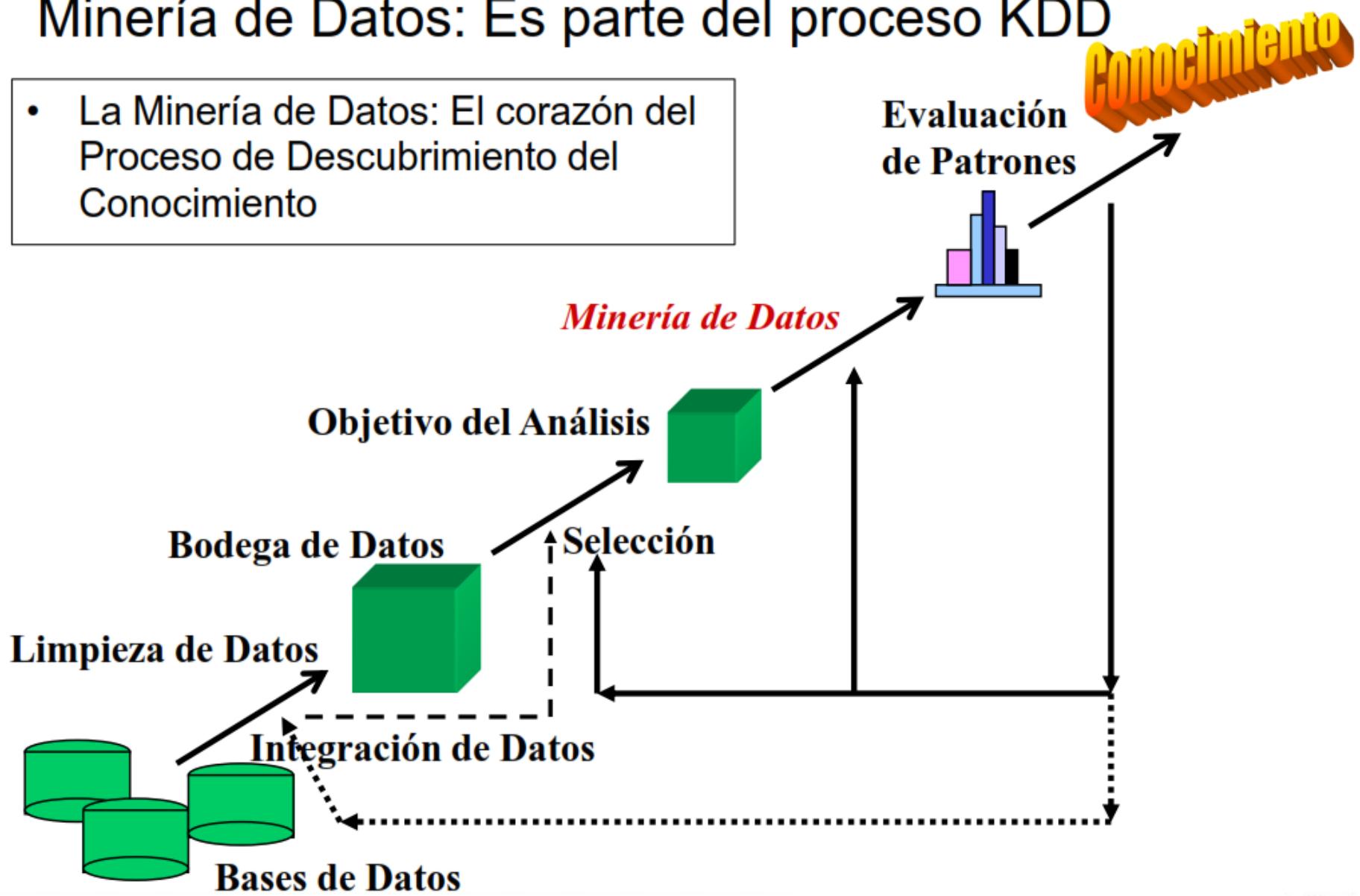
- Se considera la Minería de Datos como un el proceso, lo más automatizado posible, que va de los datos elementales disponibles en una Bodega de Datos a la decisión.
- El objetivo principal de la Minería de Datos es crear un proceso automatizado que toma como punto de partida los datos y cuya meta es la ayuda a la toma de decisiones.

Introducción

- ★ Minería de Datos versus KDD (*Knowledge Discovery in Databases*)
 - Usualmente ambos términos son intercambiables.
- ★ KDD (Knowledge Discovery in Databases): Es el proceso de encontrar información y/o patrones útiles en los datos.
- ★ Minería de Datos: es el uso de algoritmos para extraer información y/o patrones como parte del proceso KDD.

Minería de Datos: Es parte del proceso KDD

- La Minería de Datos: El corazón del Proceso de Descubrimiento del Conocimiento



Introducción

★ Minería de Datos versus Estadística

- ✓ La estadística generalmente analiza muestras de datos para luego hacer inferencia a toda la población, mientras que la minería de datos pretende buscar información útil usando toda la base datos.
- ✓ La estadística en la mayoría de los casos supone que los datos se comportan de acuerdo a ciertas distribuciones de probabilidad (normal, binomial, geométrica, Poisson, etc), mientras que la minería de datos usa técnicas mucho más exploratorias que vienen de la IA, o del “Analyse des Données”.

Introducción

★ Minería de Datos versus Análisis de Datos

- Con el advenimiento de las computadoras, aproximadamente en 1960, un nuevo concepto surgió del “matrimonio” entre la informática y la estadística: *El Análisis de Datos* (conocido en como: Analyse des Données - Exploratory Data Analysis).
- Esta nueva manera de analizar los datos con un objetivo decisional usa mucho más la informática y los métodos analíticos (el análisis de factorial, la clasificación automática, la discriminación, etc.) que los métodos estadísticos clásicos, las pruebas de hipótesis, que parten de supuestos matemáticos muy difíciles de verificar en la práctica. (Ej. no se supone que los datos siguen cierta distribución de probabilidad – los datos se muestran por si mismos).
- A diferencia de la minería de datos, el análisis de datos usualmente no es automatizado, ni trata con volúmenes de datos tan grandes.

Introducción



Minería de Datos versus Bodegas de Datos

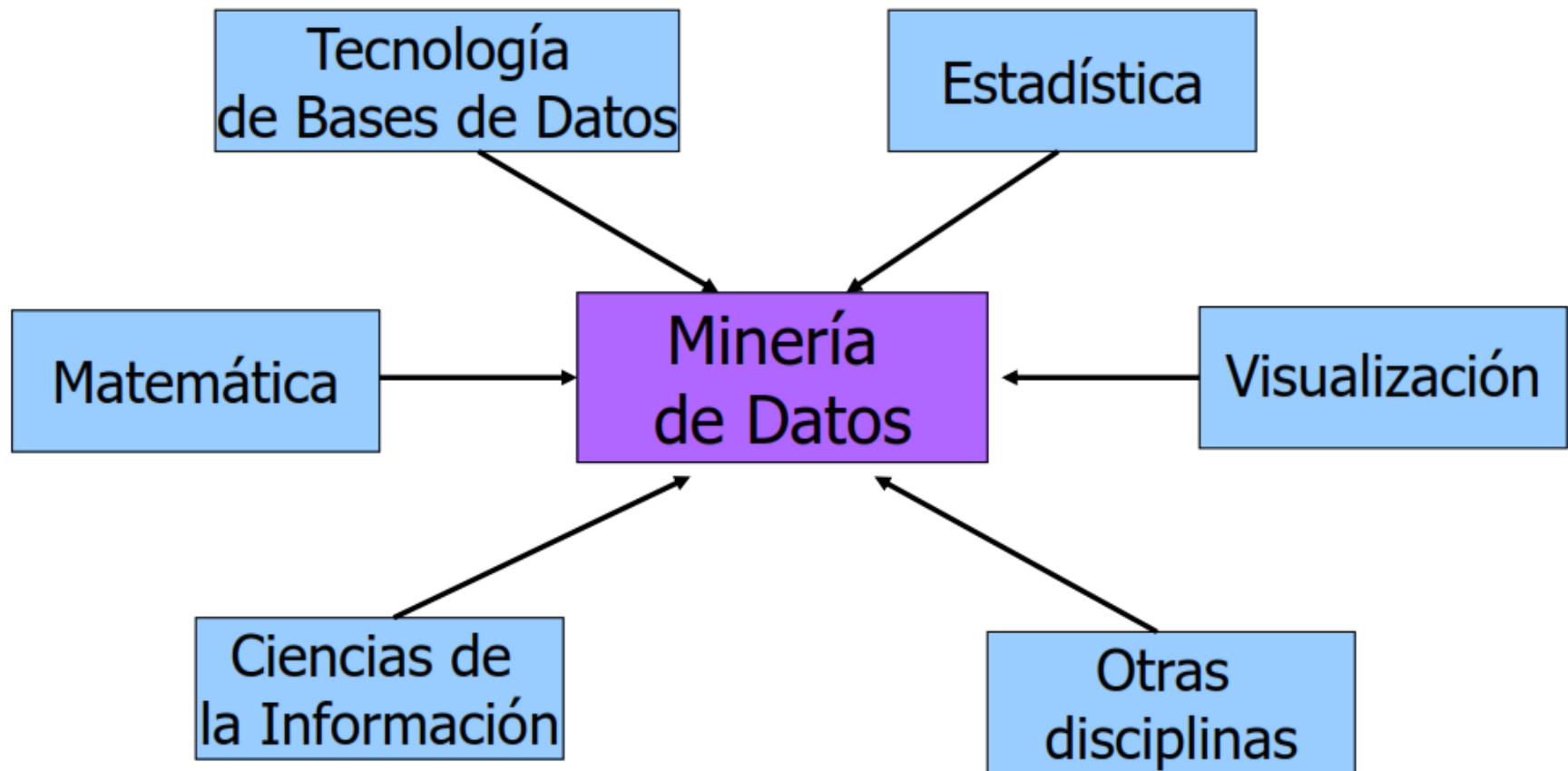
- Una *Bodega de Datos* es un almacén de datos de una compañía que contiene algunos datos operacionales, datos agregados (sumarizaciones), datos del históricos, datos evolutivos y posiblemente aquellos datos externos a la compañía pero que tienen una posible relación con las actividades de esta.
- Estos datos se depositan en una o más bases de datos relacionales y son accesibles a todas las aplicaciones orientadas a la toma de decisiones.
- Evidentemente bodegas de datos y minería de datos son cosas muy diferentes. Una bodega de datos es usualmente apenas el punto de partida de la minería de datos. Podría decirse que ambos, las bodegas de datos y la minería de datos son parte del proceso KDD.

Introducción

● Minería de Datos versus Machine Learning

- “*Machine Learning*”: es un área de la Inteligencia Artificial (IA) que trata sobre como escribir programas puedan aprender.
- En “Data Mining” es usualmente usado para predicción y clasificación.
- Se divide en dos: aprendizaje supervisado (learns by example) y aprendizaje no supervisado.

La Minería de Datos: Confluencia de Múltiples Disciplinas



Aplicaciones de la Minería de Datos

- *Retención de Clientes* ¿Cuáles clientes se van ir para la competencia?
- *Patrones de Compra* ¿Cuándo un cliente compra un producto cuál otro le podría interesar?
- *Detección de Fraude* ¿Cuáles transacciones son fraudulentas?
- *Manejo del Riesgo* ¿A qué clientes les doy un préstamo?
- *Segmentación de clientes* ¿Quiénes son mis clientes?
- *Predicción de Ventas* ¿Cuánto voy a vender el próximos mes?

¿Porqué usar Minería de Datos?

➤ **Desde un punto de vista comercial**

- Muchos datos están siendo generados y almacenados, datos de la Web, comercio electrónico.
- Las compras
- Bancos / tarjeta de crédito
- Millones de transacciones



➤ **Las computadoras se han vuelto más baratas y más potentes**

➤ **La presión de la competencia es fuerte:**

- Proporcionar mejores y más servicios personalizados

Tareas de la Minería de Datos

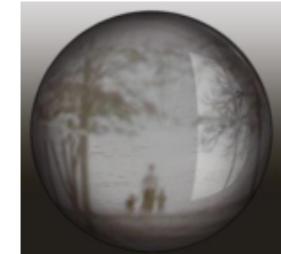
- **Descriptivas:**

- Buscar patrones humano-interpretables que describen los datos



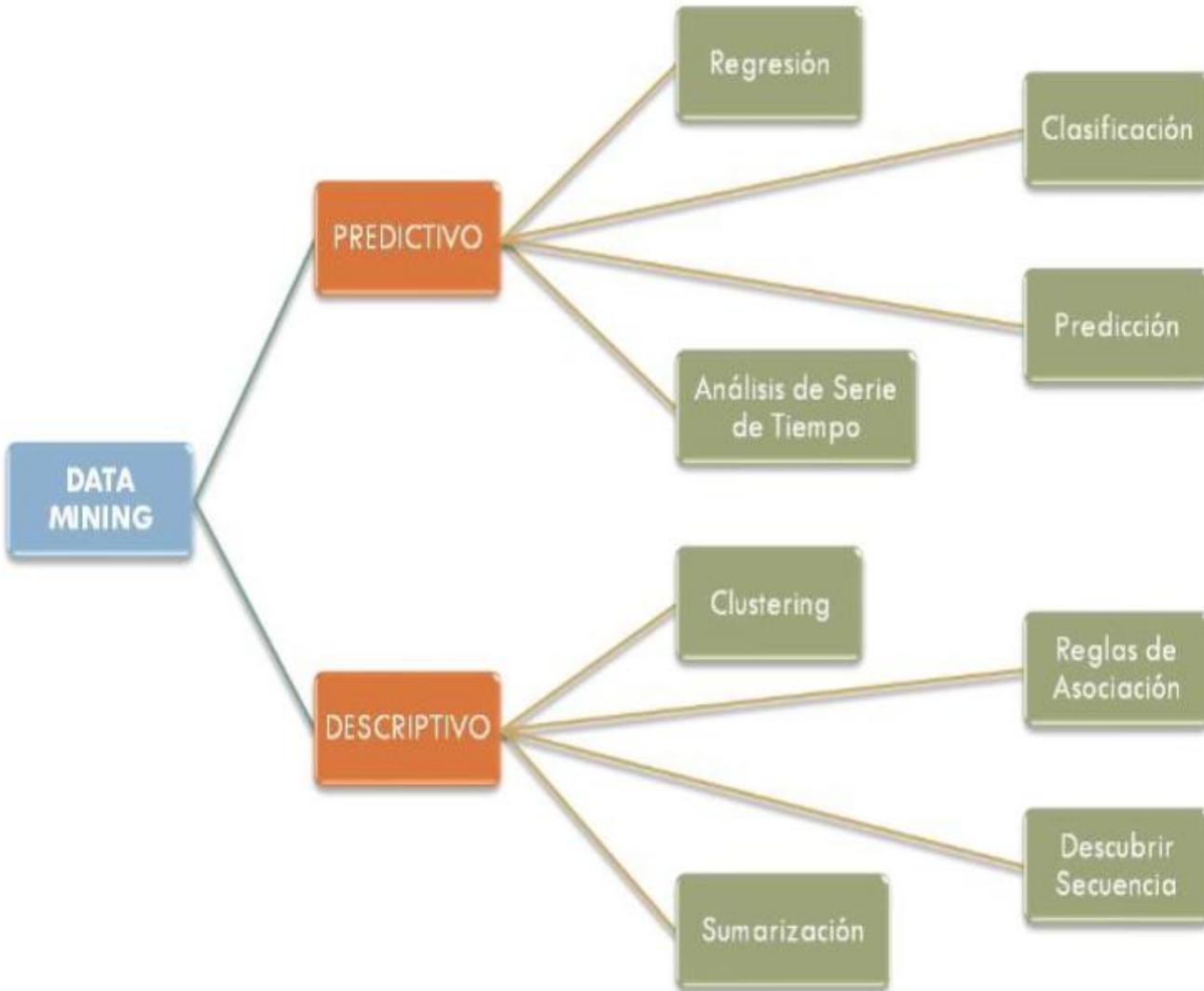
- **Predictivas:**

- Utiliza algunas de las variables para predecir los valores futuros desconocidos de la misma variable o bien de otras variables



Tareas de la Minería de Datos

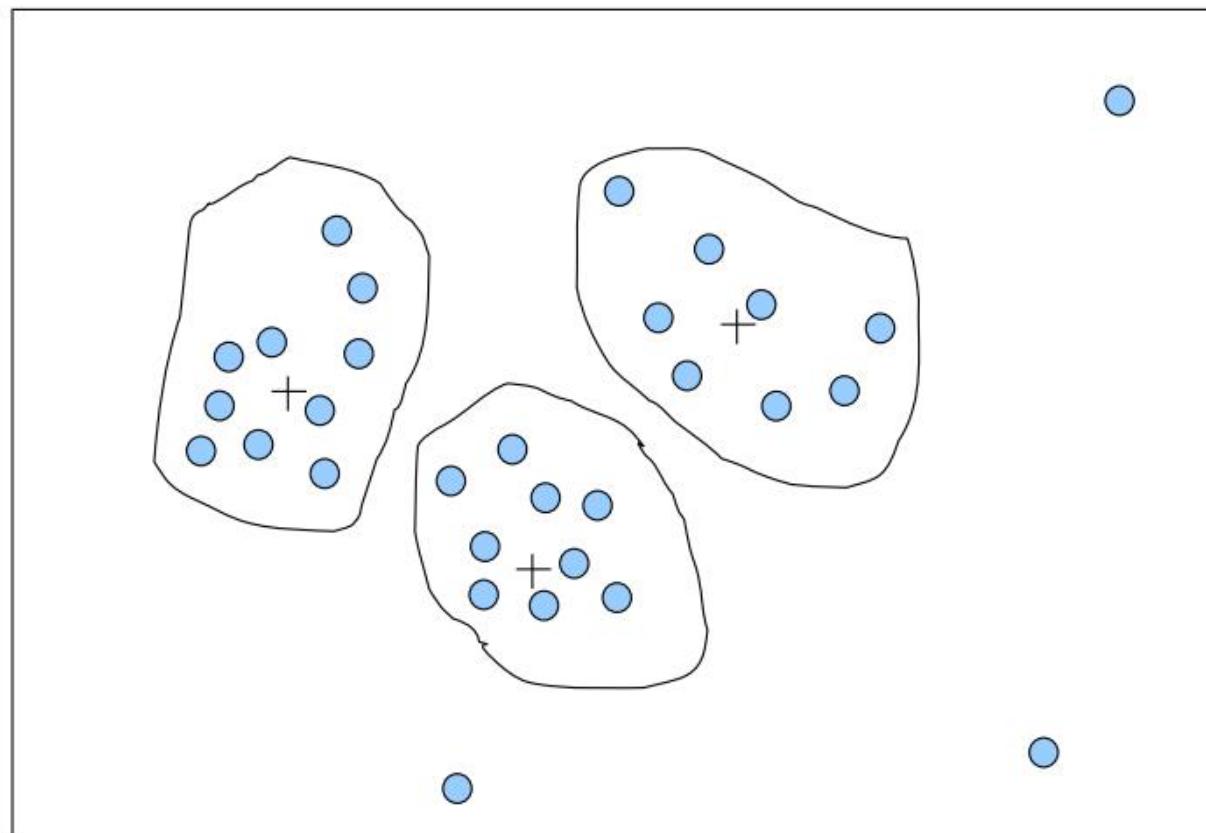
- **Descriptivas:**
 - OLAP (visualización).
 - “Clustering”.
 - Métodos Factoriales como ACP, AFC.
- **Predictivas:**
 - Series de Tiempo.
 - Análisis Discriminante.
 - Regresión.
 - Árboles de Decisión.



Tareas de la Minería de Datos

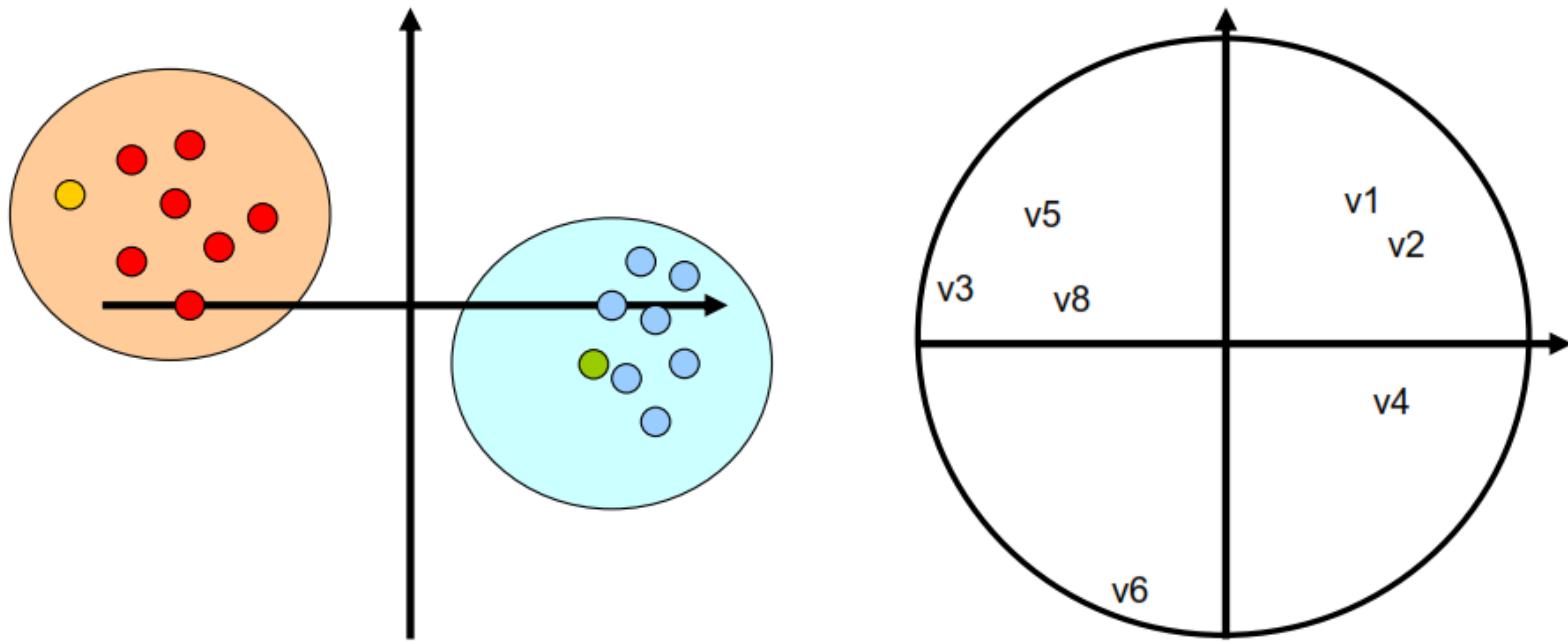
- “***Clustering***”: (clasificación no supervisada, aprendizaje no supervizado): Es similar a la clasificación, excepto que los grupos no son predefinidos. El objetivo es particionar o segmentar un conjunto de datos o individuos en grupos que pueden ser disjuntos o no. Los grupos se forman basados en la similaridad de los datos o individuos en ciertas variables. Como los grupos no son dados a priori el experto debe dar una interpretación de los grupos que se forman.
- **Métodos:**
 - Clasificación Jerárquica (grupos disjuntos).
 - Nubes Dinámicas (grupos disjuntos).
 - Clasificación Piramidal (grupos NO disjuntos).

Clustering o Búsqueda de Conglomerados



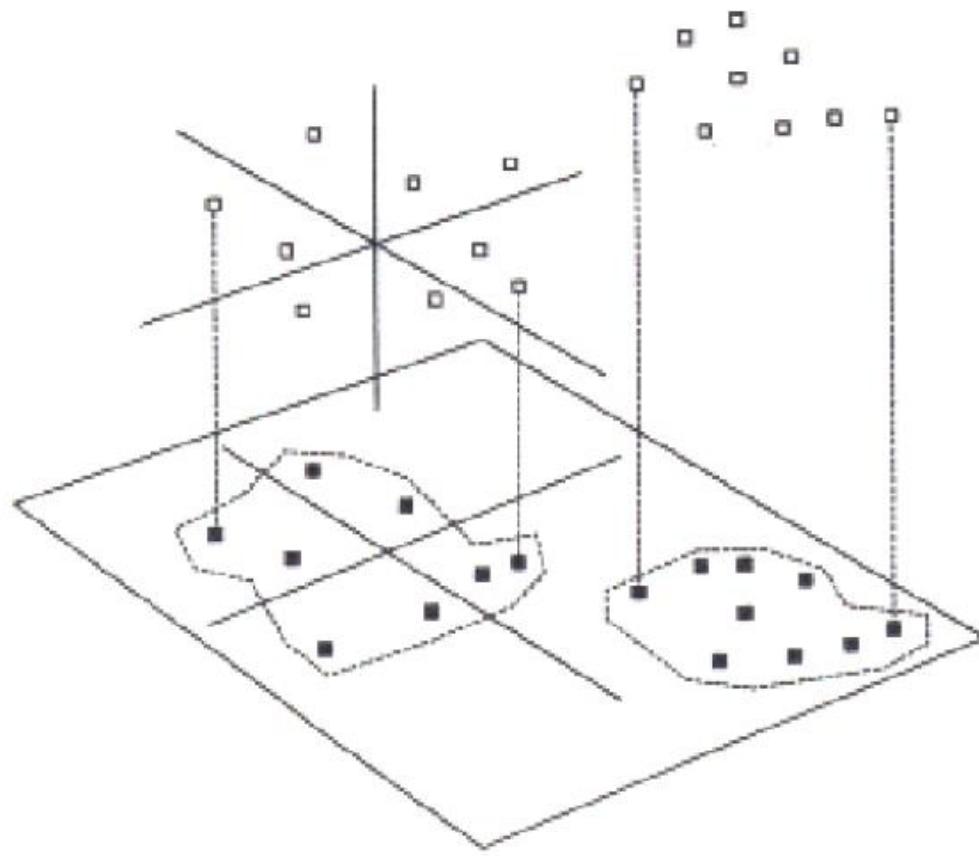
Tareas de la Minería de Datos

- **Clasificación** (discriminación): Mapea o asocia datos a grupos predefinidos (aprendizaje supervisado).
 - Encuentra modelos (funciones) que describen y distinguen clases o conceptos para futuras predicciones.
 - Ejemplos: Credit scoring.
 - Métodos: Análisis discriminante, decision-tree, classification rule, neural network



Tareas de la Minería de Datos

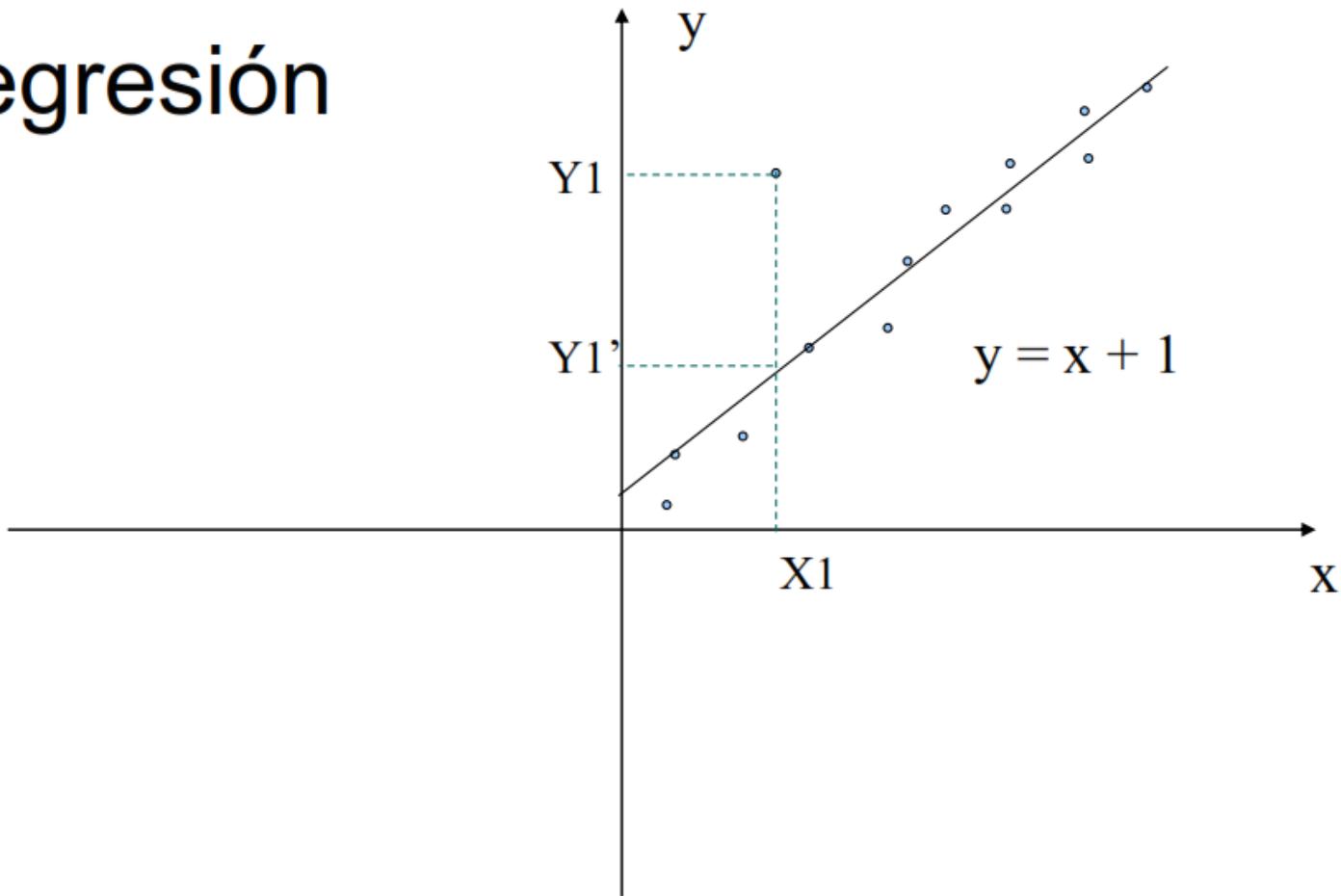
- **Descubrimiento de Factores (Análisis Factorial):**
 - El análisis factorial es un nombre genérico que se da a una clase de métodos multivariantes cuyo propósito principal es encontrar la estructura subyacente en una tabla de datos (factores ocultos).
 - Generalmente hablando, aborda el problema de cómo analizar la estructura de las interrelaciones (correlaciones) entre un gran número de variables con la definición de una serie de dimensiones subyacentes comunes, conocidas como factores.
- **Métodos:**
 - Análisis en Componentes Principales (ACP).
 - Análisis Factorial de Correspondencias simples y múltiples (AFC).
 - Análisis Canónico (AC).
 - Análisis Discriminante (AD).



Tareas de la Minería de Datos

- **Regresión**: Se usa una regresión para predecir los valores ausentes de una variable basándose en su relación con otras variables del conjunto de datos.
- Hay regresión lineal, no lineal, logística, logarítmica, univariada, multivariada, entre otras.

Regresión

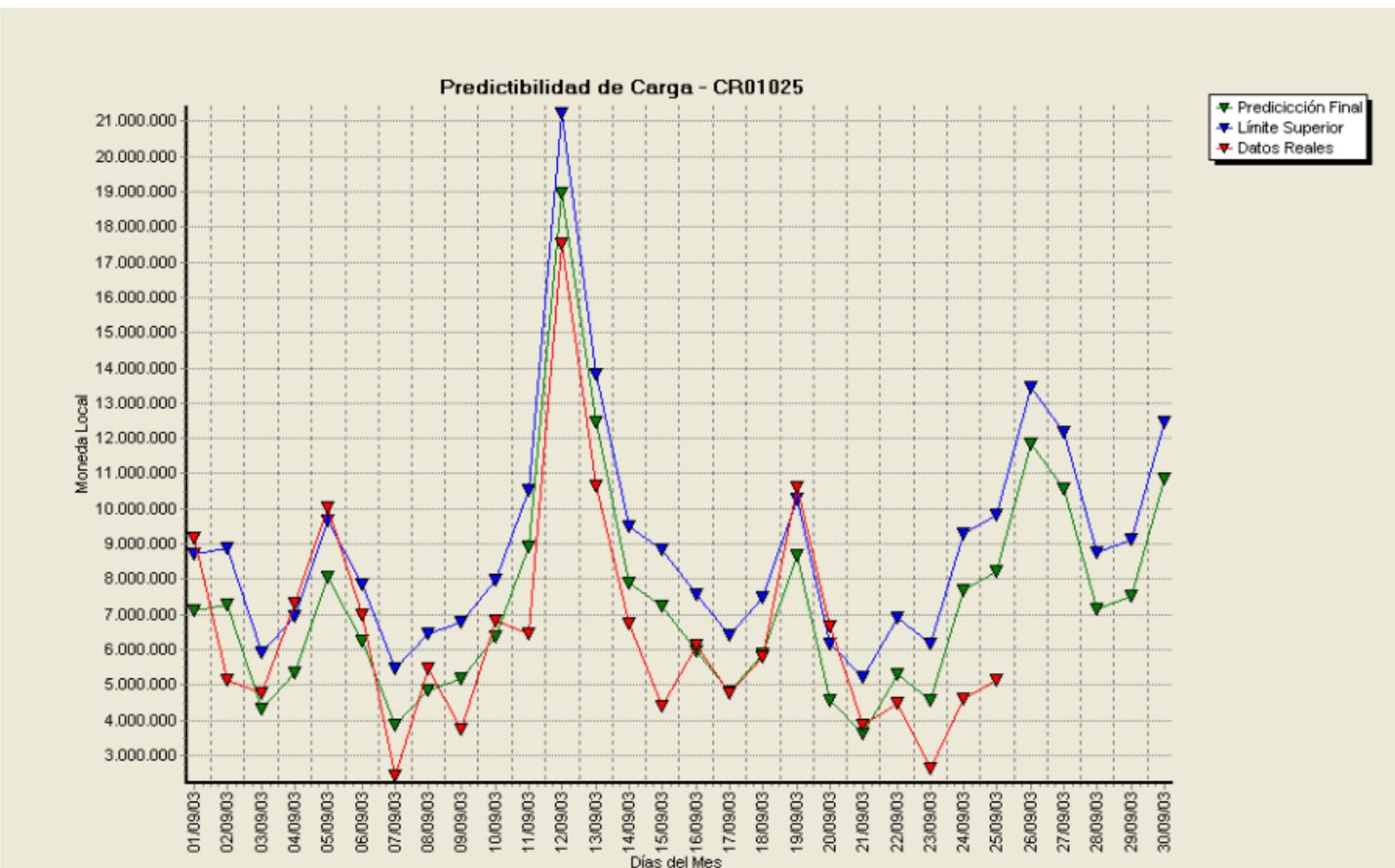


Tareas de la Minería de Datos

- **Descubrimiento de secuencias:**
 - “Secuential analysis” es usado para descubrir secuencias de patrones en los datos, estos patrones son similares a los encontrados con reglas de asociación pero tales relaciones son basadas en el tiempo.
- **Métodos:**
 - Redes neuronales.
 - Series de tiempo.

Tareas de la Minería de Datos

- **Series de Tiempo:** Una serie de tiempo corresponde a un conjunto de observaciones hechas respecto a una variable en momentos equidistantes en el tiempo, pasos:
 1. X_t : Serie de tiempo.
 2. Corregir errores sistemáticos.
 3. Transformaciones matemáticas.
 4. $X_t = \text{Tendencia} + \text{Estacionalidad} + \text{Ciclos} + E_t$.
 5. Para E_t ([Si no es un ruido blanco](#))
 1. Elegir el modelo (Box-Jenkins).
 1. ARMA(p,q) (AutoRegressive Moving Average)
 2. ARIMA(p,d,q) (AutoRegressive-Integrated Moving Average)
 2. Estimar parámetros.
 6. Pronósticos.



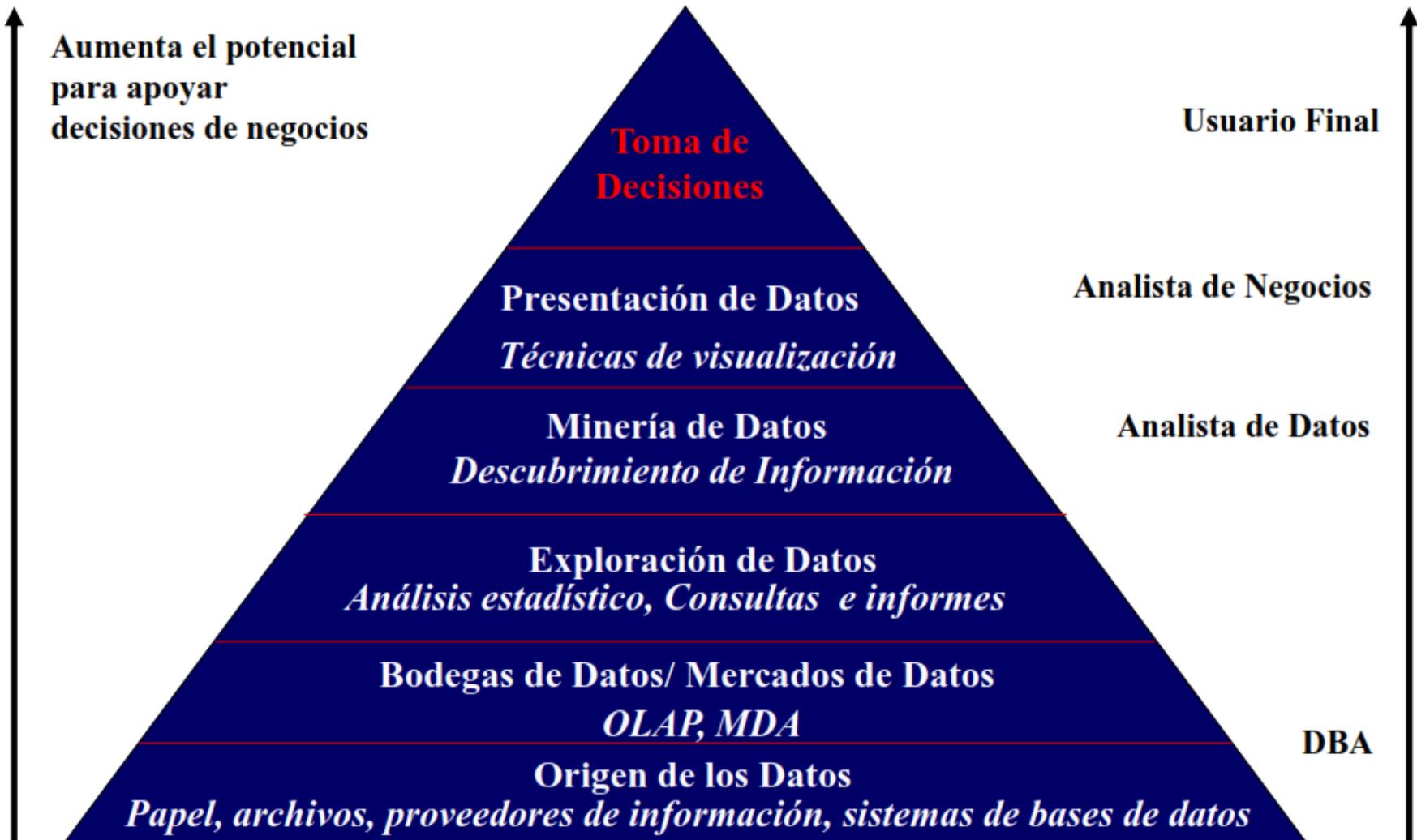
Tareas de la Minería de Datos

- **Sumarización:**
 - Los métodos de sumarización asignan los datos a conjuntos (individuos de segundo orden) que tienen asociadas descripciones.
 - Estos métodos permiten extraer o derivar datos representativos de una base de datos.
 - Permite el análisis de conceptos.
- **Métodos:**
 - Análisis de datos simbólicos.
 - Lógica difusa.
 - Interval Analysis.

Minería de Datos: ¿En qué tipo de datos?

- Bases de datos relacionales
- Bodegas de datos
- Bases de datos transaccionales
- Bases de datos orientadas a objetos y simbólicas
- Bases de datos espaciales Sistemas de Información Geográfica - GIS
- Series cronológicas de datos y los datos temporales
- Bases de datos de texto
- Bases de datos multimedia
- www (web mining)

Minería de Datos y “Business Intelligence”



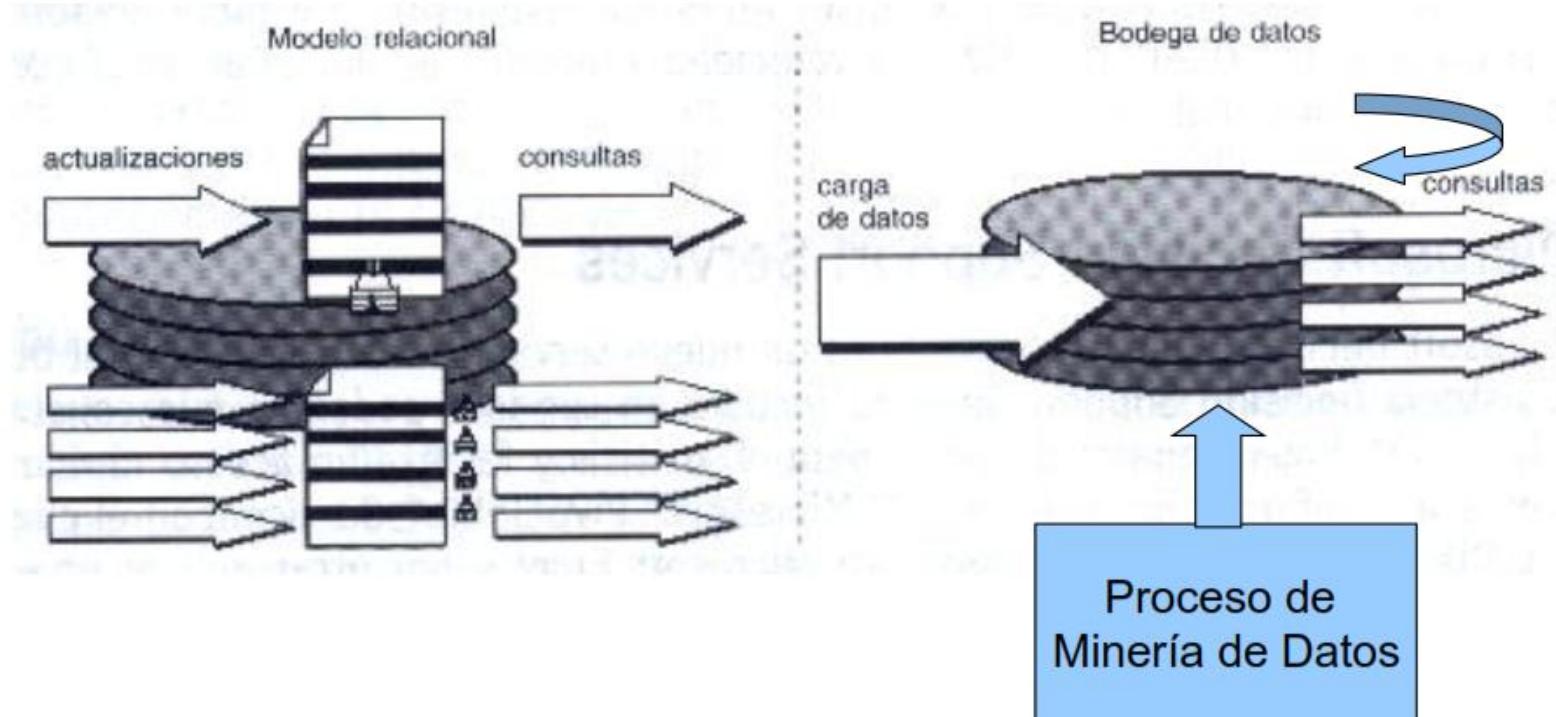
¿Qué es una Bodega de Datos? (Data Warehouse)

- Una bodega de datos es una base de datos orientada a consultas, como resultado de un análisis extenso y de la transformación de datos de la empresa.
- La bodega de datos se usa como punto de partida de un sistema de toma de decisiones.

¿Qué es una Bodega de Datos? (Data Warehouse)

- Una bodega de datos tiene datos consolidados y consistentes, orientados hacia un tema, históricos y solamente de lectura.
- Una bodega de datos podría ser el resumen un conjunto de bases de datos de una empresa.

¿Qué es una Bodega de Datos? (Data Warehouse)



	Bases de datos operativas	Almacenes de Datos
Contenido de datos	Valores actuales	Datos históricos, derivados, agregados (actual + histórico)
Estabilidad de datos	Dinámica, cambia continuamente	Estática hasta que se actualice (Estable)
Estructura de datos	Optimizada para las transacciones (orientada a la aplicación)	Optimizada para consultas muy complejas (orientada al sujeto)
Frecuencia de acceso	Alta (muchas transacciones cortas)	De Mediana a baja
Tipo de acceso	Lecturas, modificaciones, borrados, registro por registro	Lectura, agregación
Concurrencia	Alta	Baja
Número de usuarios concurrentes	Muchos	Pocos
Frecuencia de modificación de datos	Alta	Ninguna
Número de registros por acceso	Pocos	Millones
Tipo de usuario	Empleados de oficina, operadores, etc.	Gerencia, directores, ejecutivos
Redundancia	Poca	Mucha
Análisis	Ninguno	Multidimensional

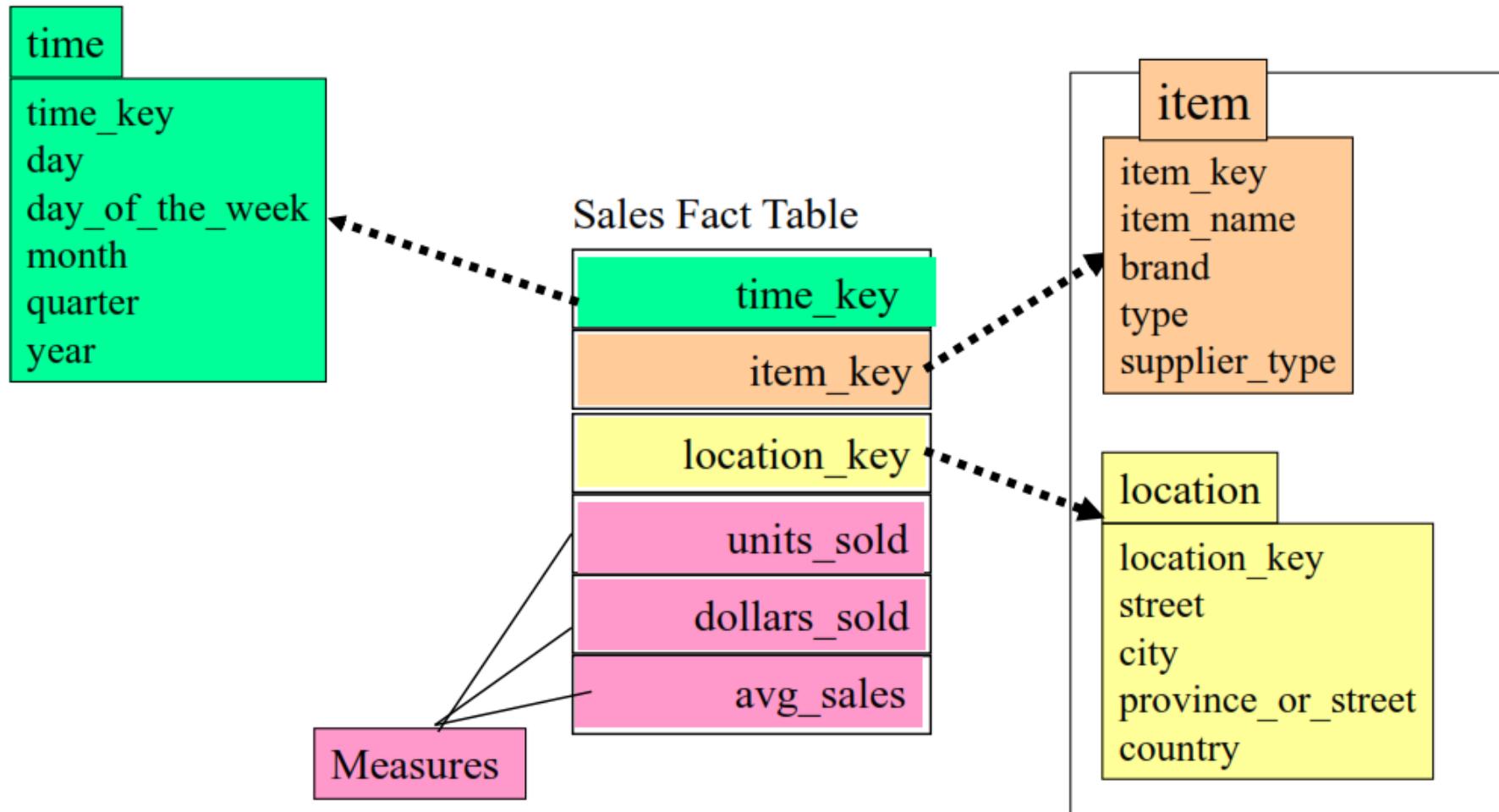
¿Qué es un Mercado de Datos? (Data Mart)

- Un Mercado de Datos (*Data Mart*) tiene las mismas características que una *bodega de datos*, pero a un nivel más refinado, pues contiene información más detallada perteneciente a un solo departamento de la empresa.

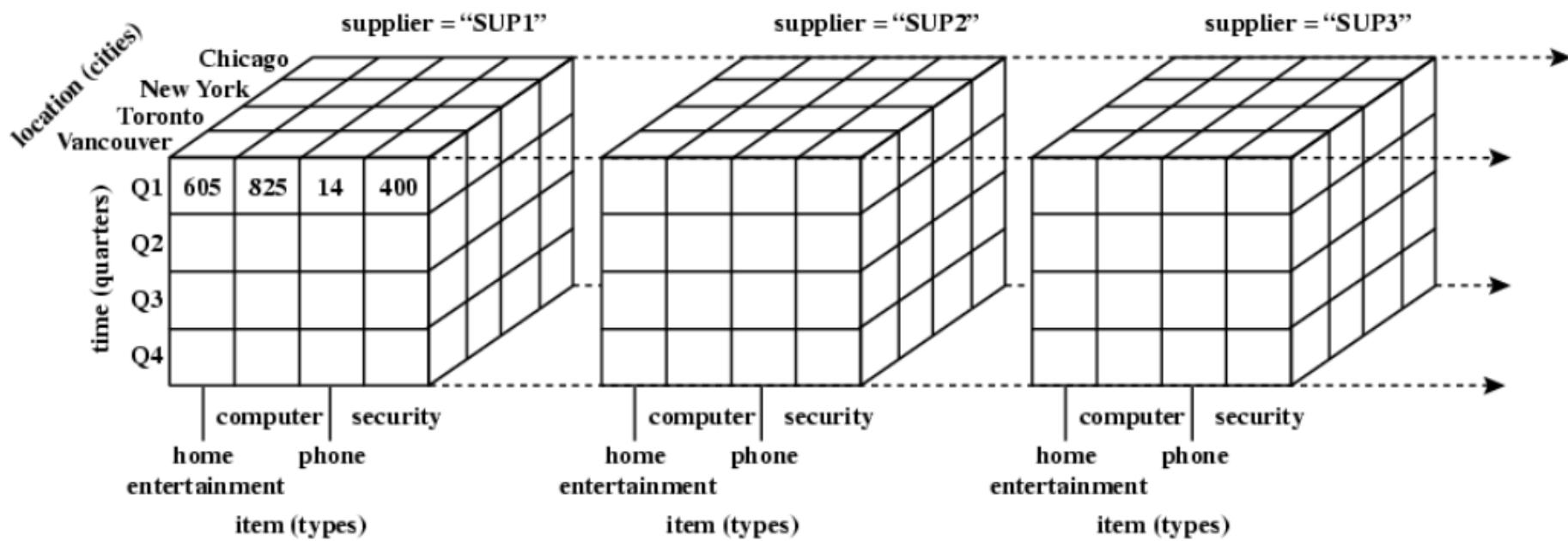
¿Qué es OLAP?

- OLAP (*Online Analytical Processing*)
- OLAP es una tecnología que procesa información de una bodega de datos en estructuras multidimensionales que proporcionan una respuesta rápida a consultas complejas.
- El objetivo de OLAP es resumir y organizar grandes cantidades de datos para ser analizados y evaluados rápidamente.

Ejemplo de un Modelo Estrella



		location (cities)							
		Chicago	854	882	89	623			
		New York	1087	968	38	872			
		Toronto	818	746	43	591			
		Vancouver							
time (quarters)		Q1	605	825	14	400	682	925	698
		Q2	680	952	31	512	728	1002	789
		Q3	812	1023	30	501	784	984	870
		Q4	927	1038	38	580			
			computer		security				
			home		phone				
			entertainment						
				item (types)					



Ciclo de un proyecto de minería de datos

1. Aprender sobre el negocio
2. Recolectar los datos. Usualmente las compañías tienes muchas bases de datos que deben ser centralizadas.
3. Limpieza y transformación de datos (mucho esfuerzo).
4. Definir la meta del proyecto y así encontrar el modelo adecuado.
5. Escoger los algoritmos que permitan optimizar el modelo.
6. Generar reportes.
7. Generar predicciones y/o “Scoring”.
8. Aplicación de los resultados en el negocio.
9. Actualización de los modelos (calibración constante de los modelos).

Estándares en Minería de Datos

- En Minería de Datos estamos como en Base de Datos hace 20 años, es decir, se están haciendo esfuerzos por definir estándares.
- XML for Analysis: es otro estándar de la industria y está a cargo del “XML / A Council”. Así surge el lenguaje de consultas “query language Data Mining eXtensions” (DMX) que permite consultas basadas en XML a los servidores de Minería de Datos.
- SQL MM: (SQL/ Multimedia for Data Mining) fue propuesto por IBM.
- Java Data Mining API. Es un paquete JAVA para minería de datos propuesto por ORACLE. El objetivo es permitir a las aplicaciones JAVA con motores de minería de datos.
- PMML, Crisp-DM, CMW (extensión de UML) y otros.

CRISP-DM

**Metodología para el Desarrollo
de Proyectos en Minería de
Datos**

CRISP-DM

**CRoss-Industry Standard Process
for Data Mining**

¿Por qué debería ser un proceso estándar?

El proceso de minería de datos debe ser confiable y repetible para personas con escasos conocimientos de minería de datos.

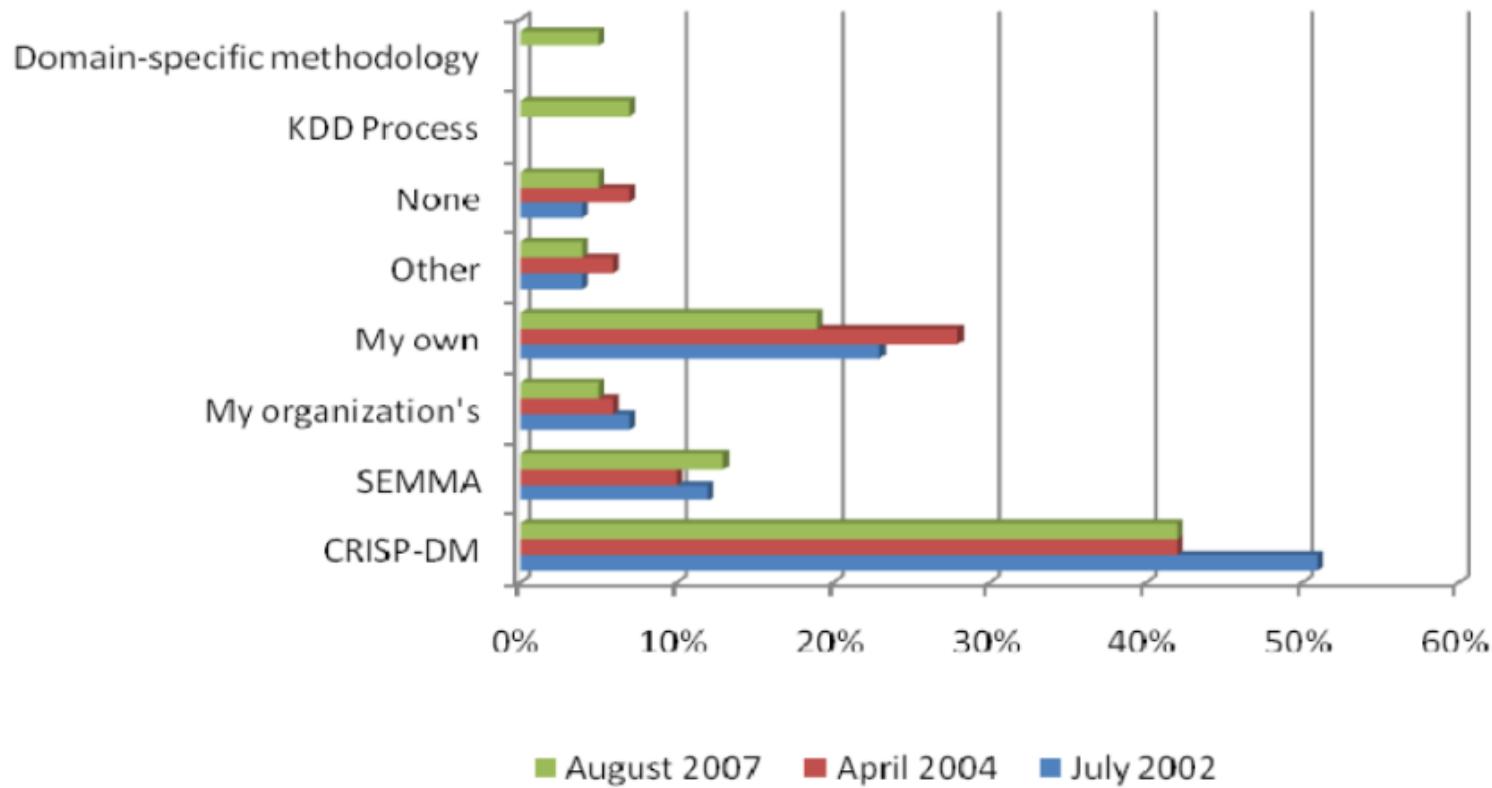


CRISP-DM

- No tiene propietario
- Aplicación / Industria neutral
- Se centra en cuestiones de negocios
- Así como en el análisis técnico y de métodos

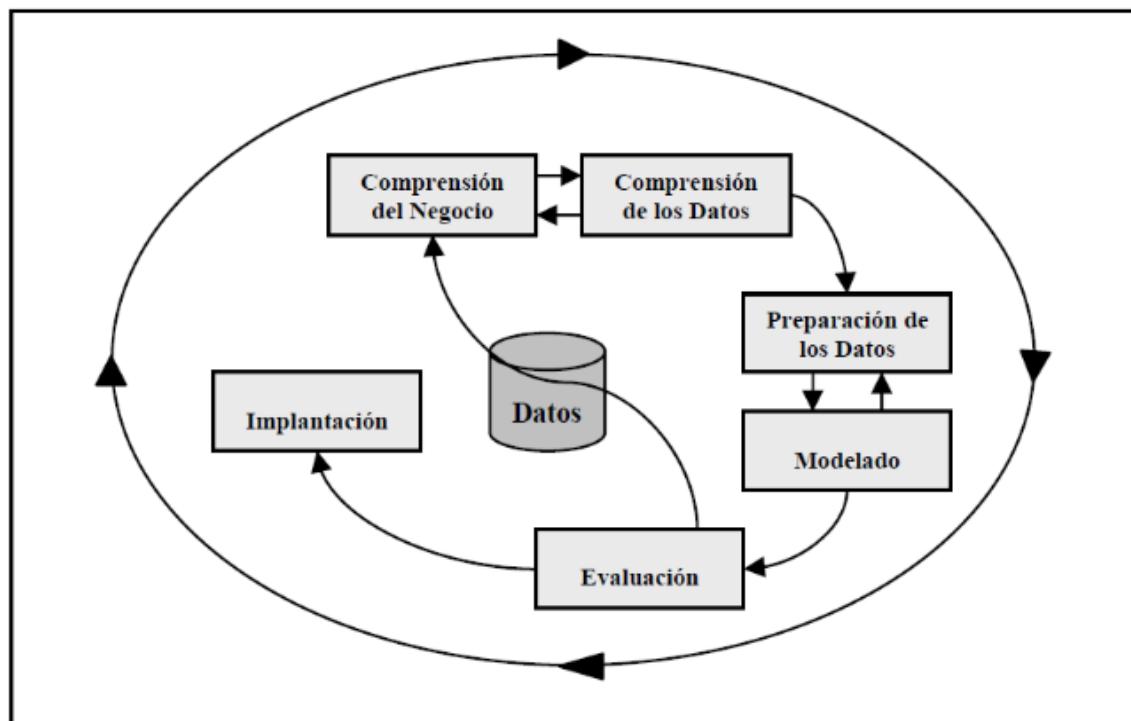


Metodologías utilizadas en Minería de Datos



Fases de CRISP-DM

- CRISP-DM, está dividida en una serie de seis fases:



Fases de CRISP-DM

- La sucesión de fases no es necesariamente rígida.
- Cada fase es estructurada en varias tareas generales de segundo nivel.
- Las tareas generales se proyectan a tareas específicas, donde finalmente se describen las acciones que deben ser desarrolladas para situaciones específicas.

¿Qué NO es Minería de Datos?



¿Qué NO es Minería de Datos?

- En general la Minería de Datos NO se basa en modelos **Determinísticos**.
- Un modelo **Determinístico** es un modelo matemático donde las mismas entradas producirán invariablemente las mismas salidas, no contemplándose la existencia del azar ni el principio de incertidumbre.

¿Qué NO es Minería de Datos?

- En general la Minería de Datos se basa en modelos **Probabilísticos**.
- Un modelo **Probabilístico** es un modelo matemático que nos ayuda a predecir la conducta de futuras repeticiones de un experimento aleatorio mediante la estimación de una probabilidad de ocurrencia de dicho evento concreto.