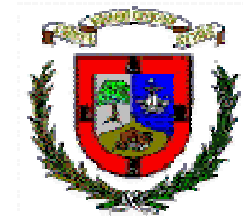


Introducción al Business Intelligence

Marta Zorrilla
Universidad de Cantabria




- ¿Qué es “Business Intelligence”?
- Campos de aplicación
- Evolución de los sistemas de gestión de datos hacia los sistemas de soporte a la decisión
- Data warehouse: justificación, definición, componentes
- Herramientas de análisis y consultas

Situación actual en las organizaciones

- Entorno competitivo y globalizado
 - Optimizar procesos
 - Reducir costes, rentabilidad financiera
 - Anticiparse a la competencia, análisis del mercado
 - Innovar, búsqueda de nuevos productos o estrategias
 - Ganar y fidelizar al “cliente” : Personalizar – simular que cada cliente es único

 - Las empresas maneja cantidades ingentes de información:
 - Fuentes internas (Sistemas corporativos propios, aplicaciones departamentales, etc.)
 - Fuentes externas (INE, INEM, colegios profesionales, encuestas, ... hasta un 20%)

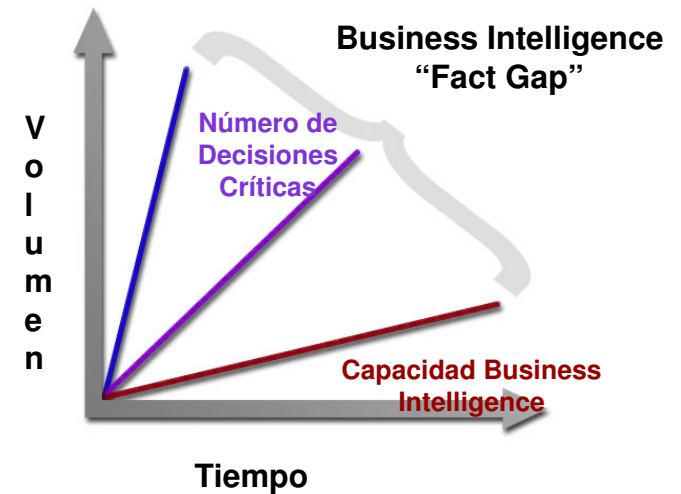
 - Problemas
 - Saturación de información
 - Difícil de acceder
 - No selectiva
-  **Business Intelligence**

¿Qué es Business Intelligence?

- **“Convertir datos en información”**
- Es lograr que los gerentes y directivos de las organizaciones, y por extensión todos los usuarios de la información, tomen las mejores decisiones cada día accediendo de forma directa a la información “clave” de su negocio de manera ágil y sencilla.
- BI suministra el marco para:
 - Definir y medir los indicadores relevantes del negocio, y entender su comportamiento
 - Procesar, resumir, reportar y distribuir la información relevante a tiempo
 - Gestionar y compartir el conocimiento del negocio con la organización
 - Analizar y optimizar los procesos que actúan sobre los indicadores
- Incluye aplicaciones software, tecnología y metodologías para realizar el análisis de datos:
 - Bases de datos
 - Aplicaciones analíticas (OLAP)
 - Reporting y querying
 - Data mining, web mining, text mining, data streaming
 - Técnicas de visualización de datos
 - Herramientas ETL
- Decision Support System (DSS): sinónimo de BI

Business Intelligence “Fact Gap”

Gartner Group (2001) denominó “**Business Intelligence Fact Gap**” a la diferencia que existe entre la información disponible en la empresa y la capacidad de tomar decisiones basándose en dicha información.



"In the absence of BI, a 'fact gap' exists: a condition where users make decisions and assess risk and opportunities based upon anecdotal, incomplete or outdated information. This isn't much better than guessing, leaving most businesses seriously exposed." (Gartner Group 07/01)

A recent research study by the **BusinessWeek Market Advisory Board** (07/2004) surveyed 675 executives throughout North America and Europe and found that **43%** indicated they did **not trust their internal systems**, and an amazing **77%** indicated that they were aware of bad decisions that had been made within their organizations because of a **lack of accurate information**.

- Science
 - astronomy, bioinformatics, drug discovery, ...
- Business
 - CRM (Customer Relationship management), fraud detection, e-commerce, manufacturing, sports/entertainment, telecom, targeted marketing, health care, ...
- Web:
 - search engines, advertising, web and text mining, ...
- Government
 - surveillance, crime detection, profiling tax cheaters, ...

Evolución de las tecnologías de bases de datos

Hito histórico	Pregunta de Negocio	Tecnología que lo posibilita	Suministrador	Característica principal
Data Collection (1960s)	¿Cuáles fueron mis ingresos en los últimos 5 años?	Ordenadores, cintas, discos, DBMS jerárquicos (IMS) y en red	IBM, CDC	Datos históricos
Data Access (1980s)	¿Cuántas unidades vendí el mes pasado en España?	Bases de datos relacionales (RDBMS, SQL, ODBC)	Oracle, Sybase, Informix, IBM, Microsoft	Datos dinámicos a nivel de registro (histórico)
Data Warehousing & Decision Support (1990s)	¿Cuántas unidades vendí el mes pasado en España en relación con Europa?	On-line analytic processing (OLAP), gestores multidimensionales	Cognos, Business Objects, Microstrategy, NCR, SPSS, Comshare, etc.	Datos dinámicos en múltiples niveles o jerarquías (histórico)
Data Mining (2000s)	¿Cuáles serán las ventas del próximo mes en Europa?	algoritmos avanzados (data stream, weblog, bio-data...), RDBMS	SPSS/Clementine, Lockheed, IBM, SGI, SAS, NCR, Oracle, etc.	Datos de prospección (análisis de mercado, de riesgos, ...)

¿Por qué un Data Warehouse?

- Los datos se encuentran en diferentes sistemas de información (uds de medida, convención de nombres y formatos, etc.)
- Estos no están orientadas a la toma de decisiones (KPI), sino a registrar transacciones (BD 3FN).
- La estructura de BD 3FN no es la adecuada para responder de forma ágil a consultas complejas, con cálculo de agregados y para ser analizadas bajo diferentes perspectivas.



Sistema de información específico dirigido por las necesidades de los usuarios de negocio, alimentado desde las fuentes de datos operacionales de la organización y construido y presentado desde una perspectiva sencilla

OLTP

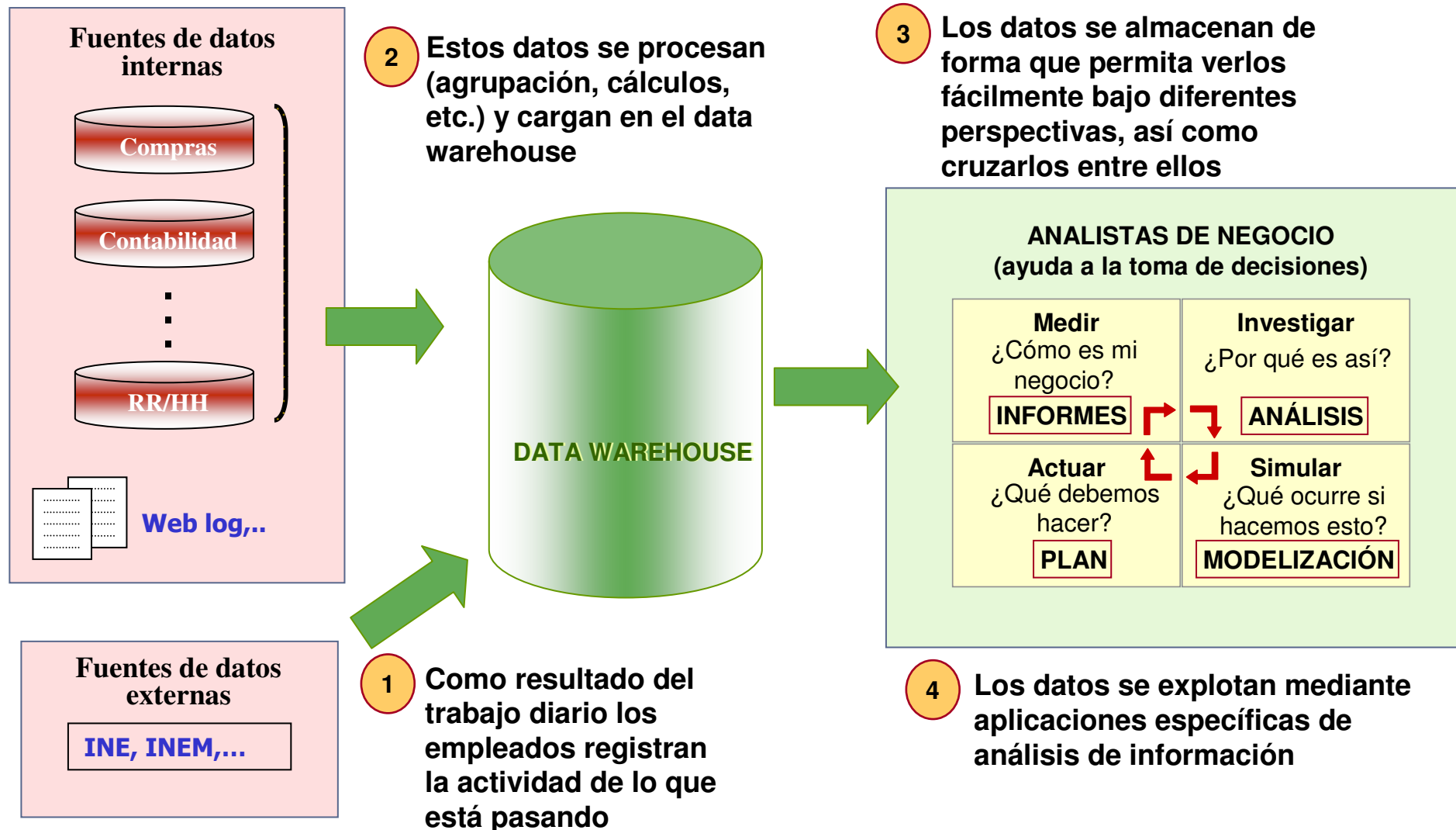
vs

OLAP

- Almacena datos actuales
- Almacena datos de detalle
- Datos dinámicos
- Integridad de datos
- Dedicado al procesamiento de datos (transacción simple)
- Nº de transacciones elevado
- Orientado a los procesos de la organización (aplicación)
- Soporta decisiones diarias
- Sirve a muchos usuarios
- Tamaño BD : 100 Mb-Gb

- Almacena datos históricos
- Almacena datos de detalle y datos agregados a distintos niveles
- Datos estáticos
- Desnormalización, redundancia
- Dedicado al análisis de datos (consultas complejas)
- Nº de transacciones bajo
- Orientado a la información relevante (negocio)
- Soporta decisiones estratégicas
- Sirve a técnicos de dirección
- Tamaño BD : 100 Gb-Tb

¿Cuál es el proceso?



	Informes	Data Mining	Simulación / Optimización
¿A qué responden?	<ul style="list-style-type: none"> ¿Qué está pasando? 	<ul style="list-style-type: none"> ¿Por qué está pasando? 	<ul style="list-style-type: none"> ¿Qué pasaría si....? ¿Cuál es la mejor opción para ... ?
¿Qué hacen?	<ul style="list-style-type: none"> Generan informes y alarmas por perfiles de usuario. <ul style="list-style-type: none"> ✓ Informes estáticos predefinidos ✓ Informes dinámicos configurables por el usuario: simples/complejos ✓ Visualización de resultados (Gráficos, herramientas GIS) 	<ul style="list-style-type: none"> Identifican patrones (tendencias, regularidades, correlaciones) existentes en las BD <ul style="list-style-type: none"> ✓ Modelo descriptivos (indirecto) <ul style="list-style-type: none"> a) Asociación b) Segmentación ✓ Modelos predictivos (directo) <ul style="list-style-type: none"> c) Clasificación d) Estimación 	<ul style="list-style-type: none"> Escenarios futuros y búsqueda de la mejor solución. Diseño de la estrategia óptima <ul style="list-style-type: none"> ✓ Simulación: dinámica de Sistemas (Jay Forrester – M.I.T.) ✓ Optimización: Investigación operativa
¿Cuál es el papel de los usuarios?	<ul style="list-style-type: none"> El usuario introduce una teoría sobre una posible relación en la base de datos, convirtiéndola en una consulta (query) 	<ul style="list-style-type: none"> El usuario no necesita asumir nada, el modelo se encarga de identificar patrones. Los datos conducen 	<ul style="list-style-type: none"> El usuario introduce hipótesis sobre valores futuros y el modelo detecta las mejores soluciones
¿Cómo se obtienen resultados?	<ul style="list-style-type: none"> Razonamiento deductivo 	<ul style="list-style-type: none"> Razonamiento inductivo 	<ul style="list-style-type: none"> Análisis de escenarios + hipótesis
Ejemplo	<ul style="list-style-type: none"> Informes con alarmas en función de la evolución de determinadas medidas 	<ul style="list-style-type: none"> Identificar qué factores (actividad, sector, región, época, etc.) influyen en la evolución de esas medidas 	<ul style="list-style-type: none"> Determinar cómo evolucionaría una determinada medida (por ejemplo ventas) si se realizara una determinada acción (p. ejemplo una campaña publicitaria del tipo 2 por 1)

Informe OLAP

Los informes permiten mostrar la información con diferentes niveles de agrupación.

- Vistas de la misma información según características de la información (dimensiones)
- Navegación multi-dimensional para investigar en los datos

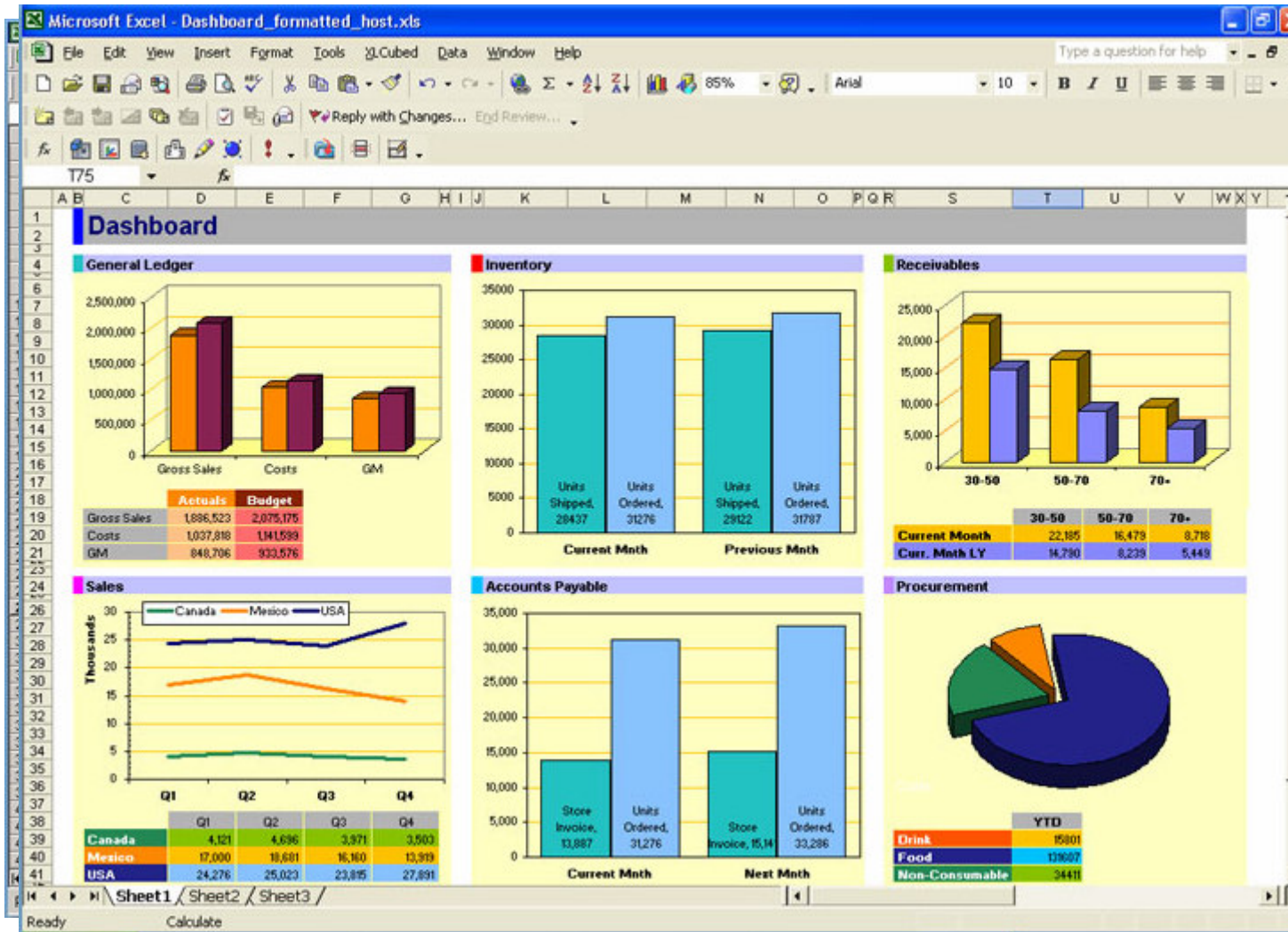
Actividad	Total
Agricultura	1
Comercio	34
Construcción	10
Resto	5
Transporte	10
Total	60

Región	Total
Centro	33
Norte	10
Sur	17
Total	60

Región	Agricultura	Comercio	Construcción	Resto	Transporte	Total
Centro	1	14	3	5	10	33
Norte		6	4			10
Sur		14	3			17
Total	1	34	10	5	10	60

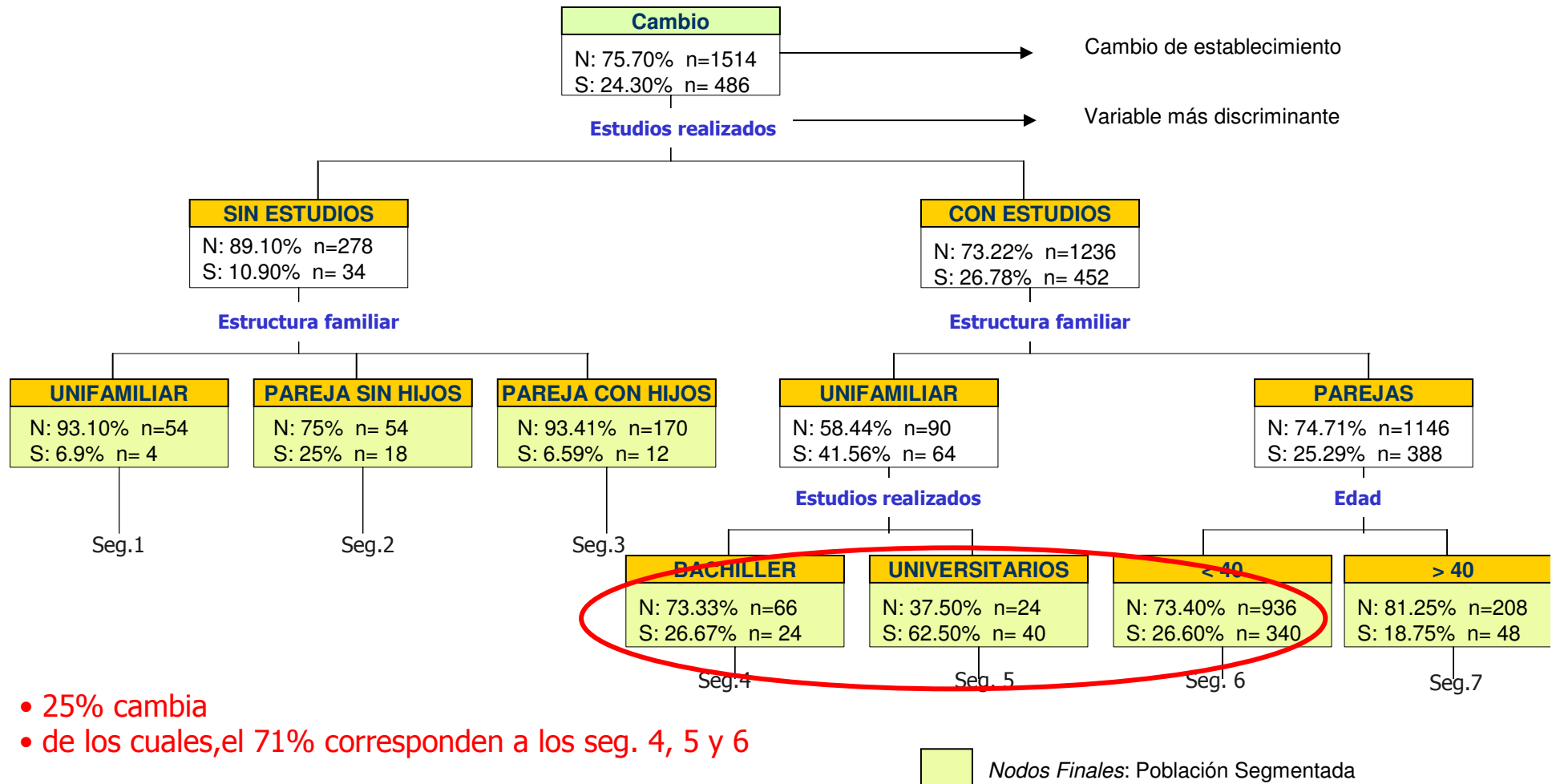
Región	Agricultura			Comercio			Construcción			Transporte					
	Mediana	Total		Grande	Mediana	Pequeña	Total	Grande	Mediana	Pequeña	Total	Grande	Mediana	Pequeña	Total
Centro	1	1		4	4	6	14	1	1	1	3	4	4	2	10
Norte				2	2	2	6	2	1	1	4				
Sur				4	4	6	14	1	1	1	3				
Grand Total	1	1		10	10	14	34	4	3	3	10	4	4	2	10

Cuadros de mando (dashboard, scorecard,..)



Data mining: Caso segmentación

Ejemplo: evaluar qué segmentos de población cambian de establecimiento de compra habitual



- 25% cambia
- de los cuales, el 71% corresponden a los seg. 4, 5 y 6

¿Qué es un Data Warehouse?

- **Ralph Kimball:**

Copia de los datos transaccionales estructurados específicamente para su consulta y análisis. (2002)

Def. extendida: es la plataforma para el business intelligence (DW/BI). (2006)

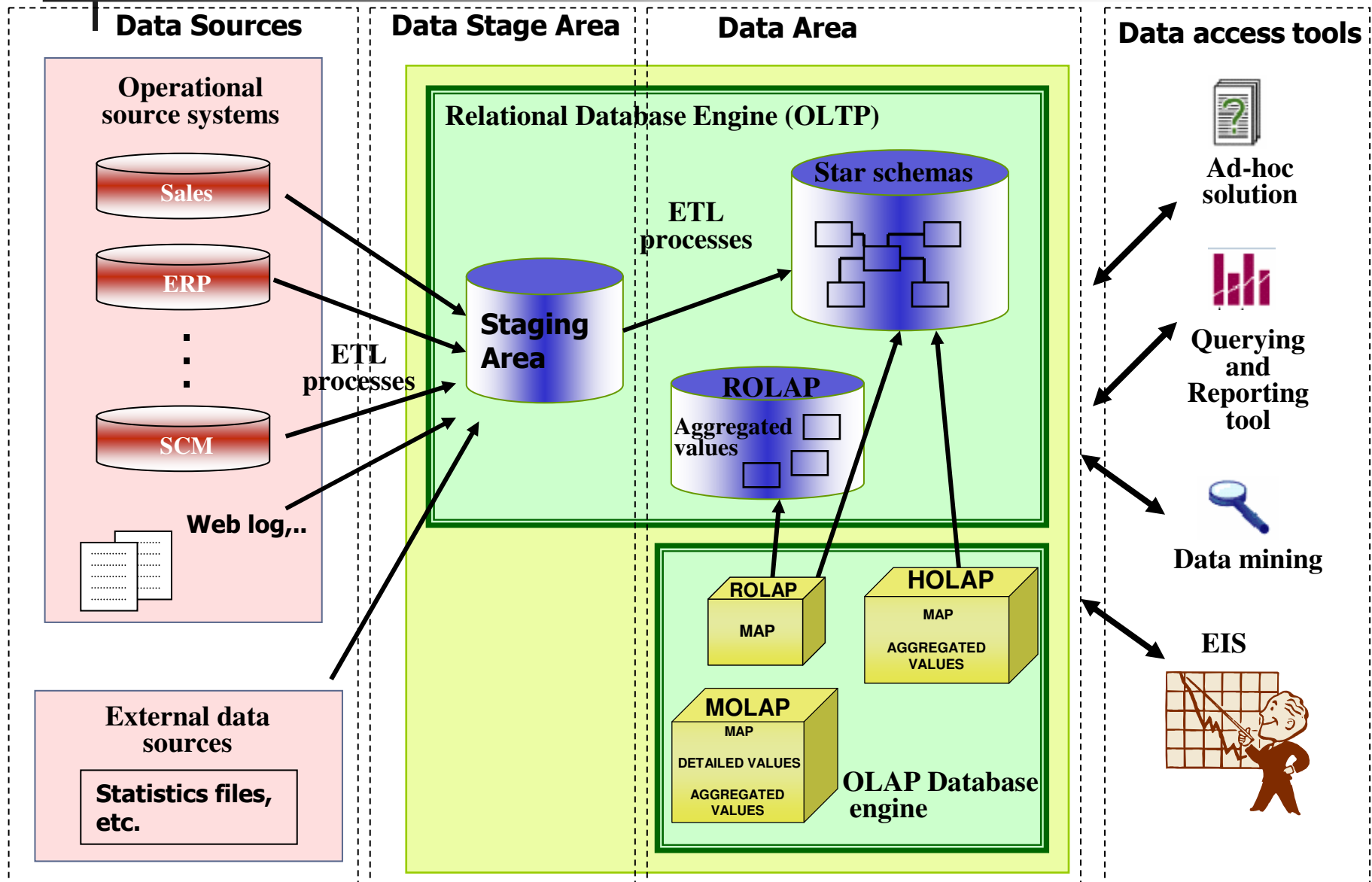
- **Bill Inmom:**

Un Data Warehouse es una colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte del proceso de toma de decisiones de la gerencia.

¿Qué es un Data Warehouse? (y 2)

- Es un sistema de información que:
 - Contiene la información estratégica para la toma de decisiones
 - Se utiliza para analizar datos, detectar tendencias y diseñar estrategias
 - Recoge datos que provienen de diferentes sistemas operacionales (**integración**), consolidados a una determinada fecha (**variante en el tiempo**) y centrados en una determinada materia de negocio (ventas, consumos, uso del sitio Web...).
 - Su estructura se diseña para dar respuesta ágil a las consultas y facilitar la distribución de sus datos, no para soportar procesos de gestión.
 - No se actualizan sus datos, sólo son incrementados (**no volátil**).

Componentes DW/BI



¿Cuál es la diferencia entre EIS y OLAP?

- Un EIS (*Executive Information System*) es un sistema de información empaquetado:
 - Proporciona a los directivos acceso a la información de estado y sus actividades de gestión.
 - Está especializado en analizar el estado diario de la organización (mediante indicadores clave) para informar rápidamente sobre *cambios* a los directivos.
 - La información solicitada suele ser, en gran medida, numérica (*ventas semanales, nivel de stocks, balances parciales, etc.*) y representada de forma gráfica al estilo de las hojas de cálculo.
 - Surgieron en los 80, y son los progenitores del software BI de los 90

- Las herramientas OLAP (*On-Line Analytical Processing*) son más genéricas:
 - Funcionan sobre un sistema de información (relacional o dimensional)
 - Estructura de almacenamiento que permite realizar diferentes agregaciones y combinaciones de datos según distintas perspectivas de observación.

¿Cuál es la diferencia entre “informes avanzados” y OLAP?

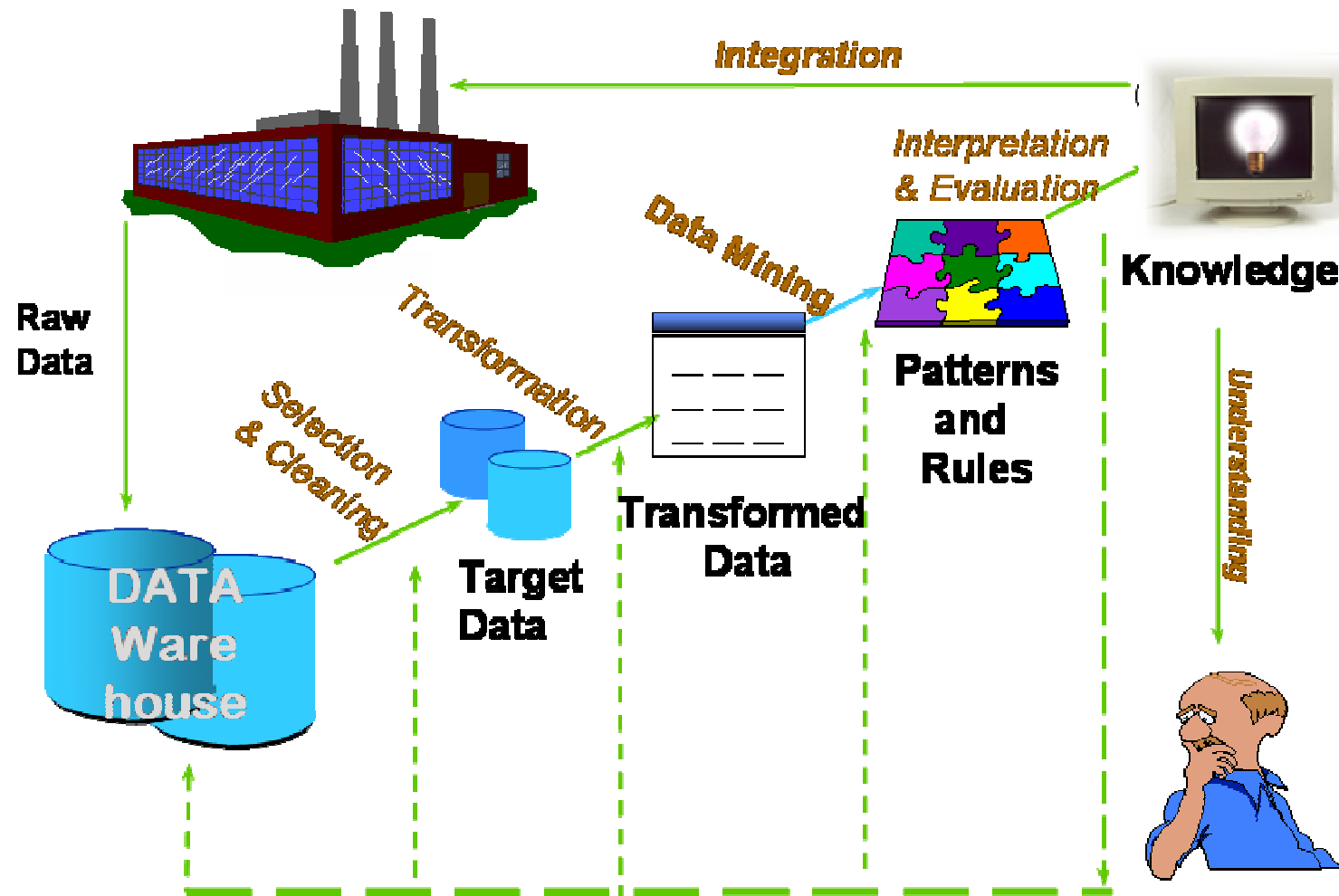
- Los sistemas de informes o consultas avanzadas:
 - Están basados, generalmente, en sistemas *relacionales u objeto-relacionales*,
 - Utilizan los operadores clásicos: concatenación, proyección, selección, agrupamiento, ... (en SQL y extensiones).
 - El resultado se presenta de una manera tabular.
- Las herramientas OLAP
 - Están basadas, generalmente, en sistemas o *interfaces multidimensionales*,
 - Utilizando operadores específicos (además de los clásicos): *drill, roll, pivot, slice & dice, ...*
 - El resultado se presenta generalmente de manera matricial.

¿Cuál es la diferencia entre OLAP y minería de datos?

- Las herramientas OLAP
 - proporcionan facilidades para “manejar” y “transformar” los datos.
 - **producen otros “datos”** (más agregados, combinados).
 - ayudan a analizar los datos porque producen ***diferentes vistas*** de los mismos.
- Las herramientas de Minería de Datos:
 - son muy variadas: permiten “extraer” patrones, modelos, descubrir relaciones, regularidades, tendencias, etc.
 - **producen “reglas” o “patrones” (“conocimiento”)**.

La tecnología OLAP generalmente se asocia a los almacenes de datos, aunque se puede tener DW sin OLAP y viceversa

Knowledge Discovery Process



Piatetsky-Shapiro

Data mining: definición

- Knowledge discovery: the non-trivial process of identifying **valid**, **novel**, potentially **useful**, and ultimately **understandable** patterns in data. (from Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.) (1996). Advances in Knowledge Discovery and Data Mining. Boston, MA: AAAI/MIT Press.)
- “the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and results.” (Berry & Linoff, 1997, 2000)
- Data mining... sometimes refers to the whole process of knowledge discovery and sometimes to the specific machine learning phase.

Qué es (y no) Data Mining?

¿Qué no es DM?

- buscar el producto más vendido
- preguntar a un motor de búsqueda por “estrellas de cine”
- conocer el estado de las cuentas de un cliente

¿Qué es DM?

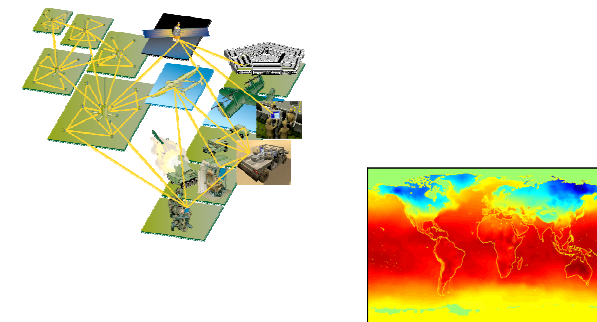
- conocer los productos que se compran juntos
- Agrupar documentos similares retornados por un motor de búsqueda de acuerdo a su contexto
- conocer la probabilidad de que devuelva un crédito

- Data mining es un proceso que trata de buscar relaciones y patrones existentes en grandes bases de datos
- Tareas principales:
 - Clasificación: predecir a qué clase pertenece un ítem
 - Clustering: encontrar clusters en los datos
 - Asociaciones: datos o eventos que ocurren frecuentemente
 - Estimación: predecir un valor continuo
 - Link Analysis: encontrar relaciones
 - Visualización
 - ...

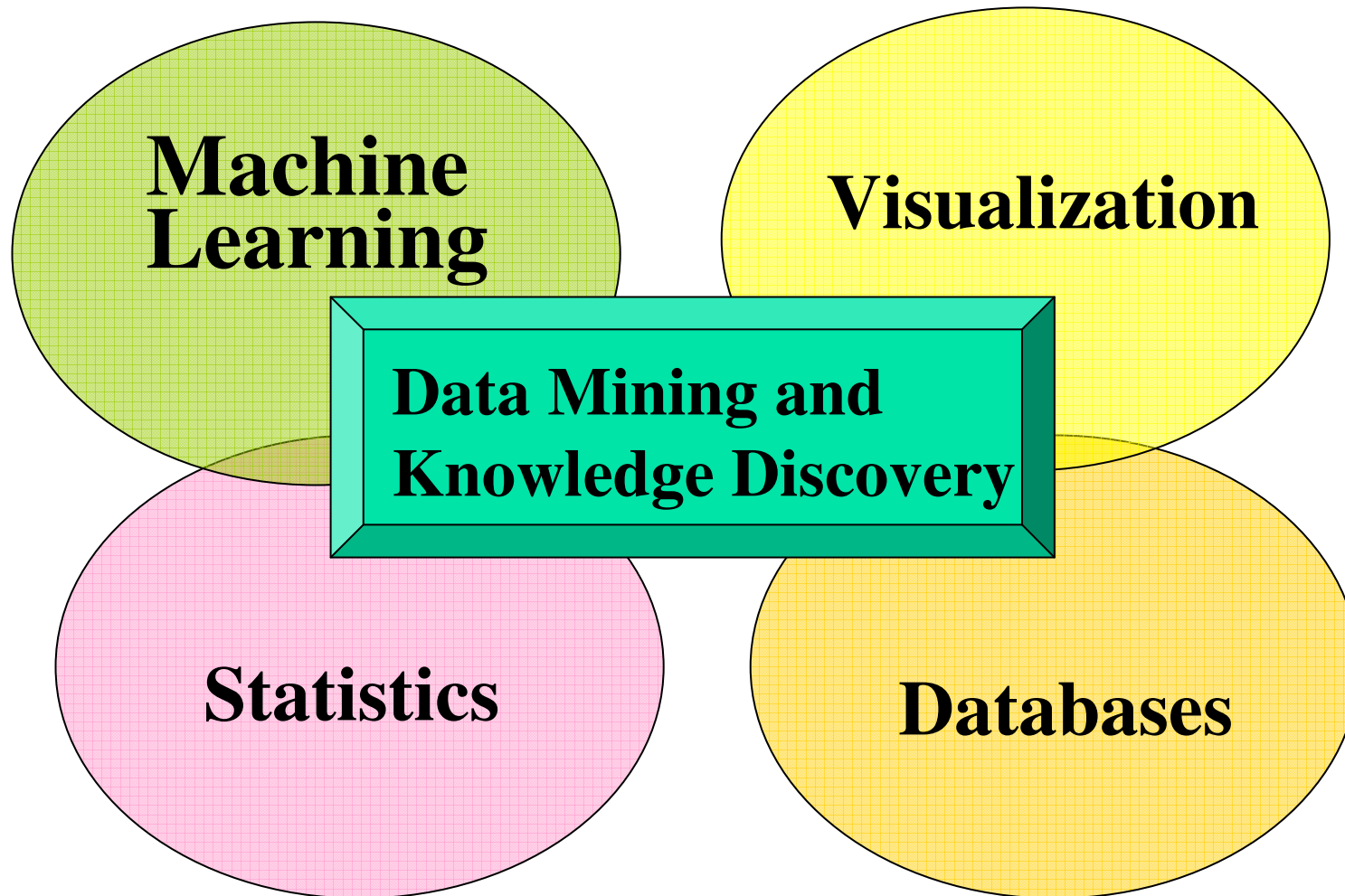
COMERCIAL



CIENTIFICO



Disciplinas relacionadas



- Statistics:
 - more theory-based
 - more focused on testing hypotheses
- Machine learning
 - more heuristic
 - focused on improving performance of a learning agent
 - also looks at real-time learning and robotics – areas not part of data mining
- Data Mining and Knowledge Discovery
 - integrates theory and heuristics
 - focus on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results
- Distinctions are fuzzy

¿Por qué ahora se habla tanto de DM?

- Las técnicas que se verán existían hace años pero la convergencia de los siguientes factores:
 - Cantidad de datos producida
 - Los datos están integrados (data warehouse)
 - La potencia de los ordenadores
 - Fuerte presión de la competencia
 - Software de data mining específico e integración de algoritmos de DM en gestores de BD