Universidad de Santiago de Chile
Facultad de Ingeniería
Depto. de Ingeniería Informática

Análisis de Datos
Capítulo I
"Introducción"

Profesor: Dr. Max Chacón.

Objetivos:

- Conocer las diferentes de definiciones del análisis de datos y sus relaciones con la obtención de conocimiento en Bases de Datos.
- Definir el problema desde el punto de vista de aprendizaje no lineal.
- Comparar con modelos lineales de regresión.
- Identificar las etapas del proceso de adquisición de conocimiento en Bases de Datos.
- Examinar las hipótesis de los modelos basados en aprendizaje.
- Identificar las diferencias entre Bases de datos operacionales y analíticas.

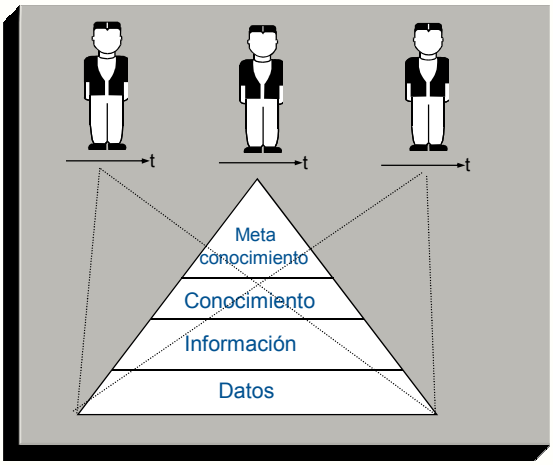
1.1 Definiciones

Identificación

- **Análisis de datos**
- **Análisis inteligente de datos**
- **Aprendizaje automático**
- **Algoritmos de aprendizaje**
- **Aprendizaje basado en ejemplos**
- **Aprendizaje de maquinas**
- **Minería de datos**
- **Inteligencia computacional**
- **Maquinas inteligentes**
- **Inteligencia Artificial.**

- ***Datos*:** Hechos o medidas que describen características de objetos, eventos o personas, es la materia prima de la cual se obtendrá la información.
- ***Información*:** Datos procesados y presentados en forma adecuada, de interés para un observador en un tiempo determinado.
- ***Conocimiento*:** Información procesada para emitir juicios que llevan a conclusiones.
- ***Meta Conocimiento*:** Reglas que permiten obtener conocimiento.

Estos conceptos se pueden representar en una estructura piramidal que representa una reducción en cantidad, como se muestra.



Análisis de Datos y Minería de Datos (MD).

Conjunto de técnicas que permiten extraer información y conocimiento a partir de Bases de Datos.

Analogía: Similar al proceso de extracción de minerales se requiere remover grandes cantidades de datos (materia prima) para obtener información o conocimiento (mineral puro).

Muchas veces estos términos son mal usados queriendo indicar la ***Obtención de Conocimiento de Bases de Datos (Knowledge Discovery in Database, KDD*).**

El termino *MD* aparece en 1989, es atribuido a Frawlay, Restetsky, Shapiro y Mathus.

"*Proceso no trivial de identificación válido, novedoso, potencialmente útil, y esencialmente entendible de obtención de patrones de los datos*".

El terminó patrones está usado en sentido amplio y considera:

- *Relaciones, Correlaciones, Tendencias, Descripción de eventos raros*, etc.
- En la primera conferencia internacional de *KDD* Canadá, 1995 el termino *KDD* es empleado para describir el proceso de extracción de conocimiento de los datos.
- *KDD* es *"La extracción no-trivial de conocimiento implícito en los datos que resulte ser previamente desconocido y potencialmente útil".*
- El conocimiento debe ser nuevo, no obvio y debe estar disponible para el uso.

- Fayyad y col. (1996) define *KDD* como:

  "*La utilización de las Bases de Datos a lo largo de un proceso de selección, pre-procesamineto, sub-muestreo y transformación; aplicando los métodos de minería de datos (algoritmos) para enumerar patrones y evaluar los productos de la minería, como un proceso de identificación de subconjuntos de patrones enumerables, denominado conocimiento*".

- *KDD* no es una técnica nueva, es un campo multidisplinario de investigación que cubre diversas áreas del conocimiento como: Bases de datos (operacionales y analíticas), Redes de Computadores, Estadística, Reconocimiento de patrones, Redes neuronales, Sistemas expertos, Aprendizaje automático de máquinas, Computación evolutiva, y otras.

- *Minería de datos*: Es usado para descubrir exclusivamente la etapa de obtención (descubrimiento) del conocimiento del proceso de *KDD*.

- Una forma de ver los objetivos de la minería de datos es clasificarla en niveles de generalidad de la información que se requiere.

- Esta clasificación se relaciona con la pregunta para la obtención de conocimiento.

Modelo del sistema de información operacional.

- Un *Sistema de Información* (*SI*) es una representación de datos generados de la medida de algún fenómeno físico como imagen, voz, texto, proceso industrial, etc.
- Un *SI* está compuesto de 4-tupla como: $SI=<U, Q, V, f>$

Donde:

- U universo cerrado: un conjunto finito, no vacío, de n *objetos*, $\{x_1, x_2, ..., x_n\}$
- Q: un conjunto finito, no vacío, de p *atributos* $\{q1, q2, ..., q_p\}$
- $V= \bigcup_{q \in Q} V_q$, donde V_q es un dominio (valor) de los atributos q.
- $f: UxQ \rightarrow V$ es una función de decisión *llamada función de información*, tal que $f(x,q) \in Vq$ para cualquier $q \in Q$, $x \in U$.
- Un par (q,v) para $q \in Q$, $v \in Vq$ es llamado *descriptor* en un sistema de información *SI*.

El *SI* puede ser representado por una tabla finita de datos, donde las columnas están indicadas por los atributos, las filas por los *objetos* y la entrada por la columna *q* y la fila *xi* resulta la instancia de la función de información $f(x_i,q)$.

Ej: Descripción de automóviles.

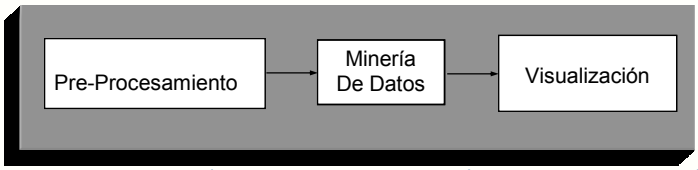
Objeto	Atributos		
U	Potencia	Caja Vel.	Tipo
x_1	140 HP	4	Sedan
x_2	120 HP	5	Hashback
x_3	100 HP	Autom.	SW
x_4	120 HP	Autom.	Sedan
x_5	100 HP	5	Hashback
x_6	140 HP	5	SW

- U={x1, x2, x3, x4, x5, x6}
- Q={Potencia, Caja Velocidades, Tipo}
- Vpotencia= {100, 120, 140}, Vc.v.{4, 5, Automático}, Vtipo{Sedan, Hashback, SW}
- f(x3,Tipo)=SW.

*En general este sistema se **denomina Sistema de Información Operacional** o **Base de Datos Operacional**, pues está destinado a la realización de consultas (tipo SQL) que tienen que ver con la operación normal del proceso que apoya, desde el punto de vista informático.*

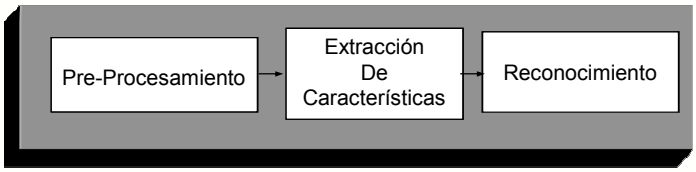
1.2. Estructura del Proceso de Obtención de Conocimiento.

La estructura general del proceso de obtención de conocimiento se puede resumir en una etapa de pre-procesamiento, una de minería de datos y una etapa de visualización o generación de informes.

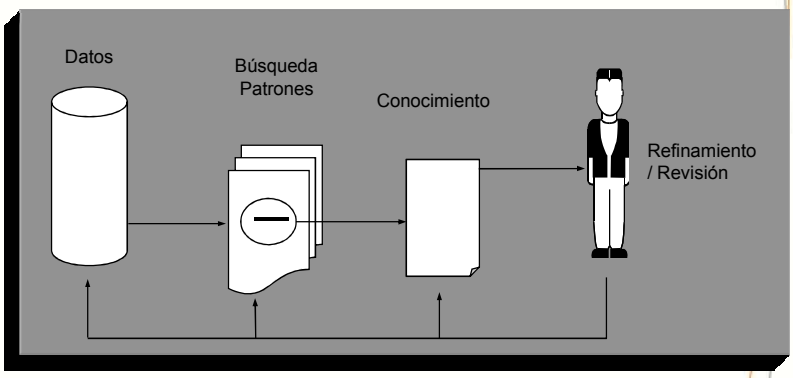


La etapa de pre-procesamiento se puede dividir en varias sub etapas, como son: selección de datos, limpieza de datos (o filtrado), enriquecimiento y codificación.

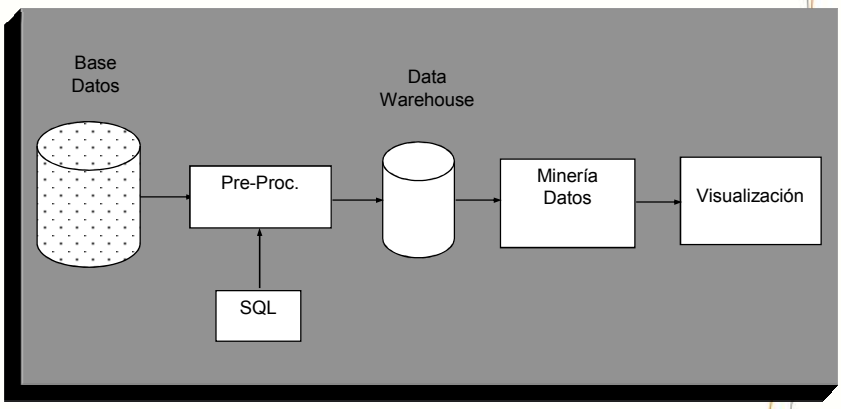
- El proceso parece secuencial con desarrollo lineal, pero en la práctica, en cualquier etapa se detiene y vuelve atrás.
- Esta estructura general no es fija para cada problema y varias de estas etapas o fases no existen necesariamente para cada aplicación o se deben incorporar algunas nuevas variantes.
- En un contexto amplio se puede incorporar el reconocimiento de patrones cuya estructura se muestra en la figura.



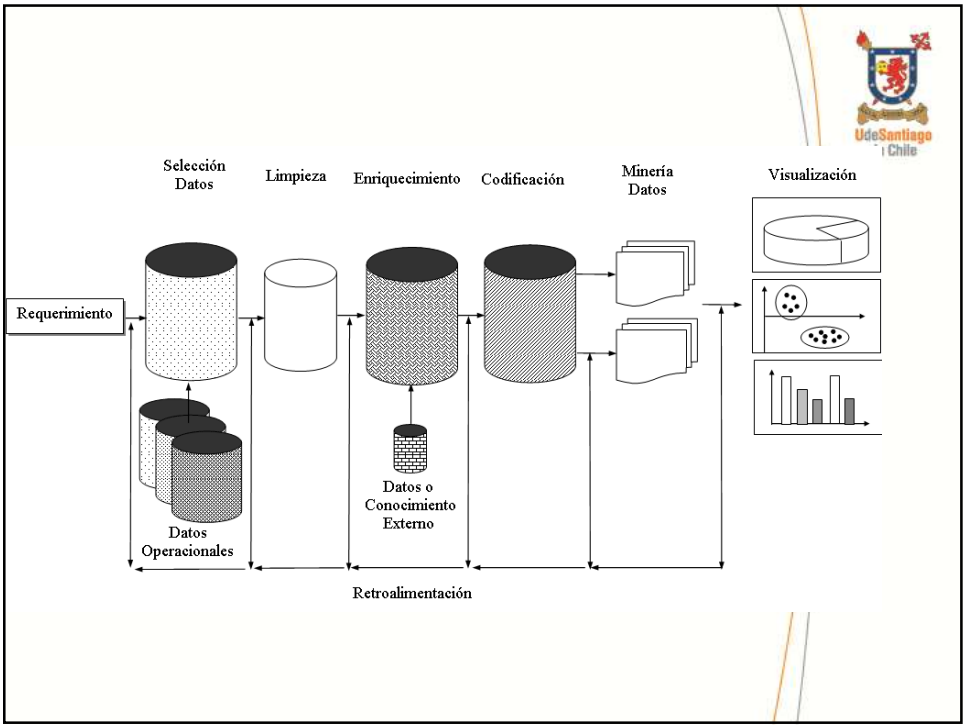
La estructura del proceso de obtención de conocimiento está íntimamente relacionada con las bases de datos, de la cual se extraen patrones con los cuales se producirán piezas de conocimiento.



Una forma general de mostrar el proceso es construir, después de un pre-procesamiento, un Data Warehouse (DW) para realizar posteriormente el proceso de minería de datos.



- El pre-procesamiento, en general, está formado por diferentes sub-procesos, muchos de los cuales no constituyen minería de datos.
- En general el pre-procesamiento se puede considerar otra disciplina diferente a la *MD*, que incluso utiliza alguna de las mismas técnicas que incluye la MD.



Para ejemplificar las diferentes fases de la etapa de pre-procesamiento se supondrá que se cuenta con una base de datos operacional de una librería que realiza ventas por Internet y vende diferentes tipos de libros como: naturaleza, arquitectura, computación, arte, educación, medicina, música, ficción, etc.

Los objetivos de la minería de datos pueden ser múltiples, por nombrar algunos:

- Requerimientos del departamento de marketing para el diseño de catálogos.
- Perfil de un lector de libros de computación.
- Determinar si existe una relación en el interés del lector de libros de computación y de ficción.

- **Selección de datos.**

Se realiza generalmente de una base de datos operacional. Para facilitar el proceso, los datos son copiados en otra base de datos denominada generalmente base de datos analítica.

El principal objetivo es seleccionar datos que contengan la información o el conocimiento que se desea obtener.

Para realizar este proceso se requiere conocimiento, experticia en el área de trabajo, además de algunos dominios de muestreo estadístico.

- **Limpieza.**

Existen varios tipos de limpieza o filtrado de datos que pueden ser aplicados en esta etapa, pero es común que alguna polución o ruido en los datos sólo se detecte en la etapa de codificación o de minería.

Algunos de los problemas más comunes que pueden ser detectados en esta etapa son:

- Duplicación de registros
- Fechas fuera de rango
- Falta de campos en registros
- Registros diferentes con campos iguales.

Ej: Librería que realiza ventas por INTERNET

Cliente Nº	Nombre	Dirección	Fecha Compra	Tipo
54011	Johnson	12 road stret 20, USA	25/12/98	Arte
54012	Pacheco	359 Maipu, AR	*01/01/01*	Musica
54013	Stabros	129 Liking, GR	29/06/95	Edu.
54014	Martinez	2 Plaza, SP	05/11/97	Ficción
54015	Matuz	25 Av. Terra, SP	*11/11/11*	Arte
54016	Müller	134 Lutero, GR	15/12/98	Edu.

- En muchos casos no basta con la detección visual o intuitiva, muchas veces se requieren métodos automáticos de detección con cierto grado de inteligencia.
- El filtrado o limpieza de los datos es en general una disciplina diferente de la minería de datos, pero tienen muchas cosas en común.
- Los algoritmos de reconocimiento de patrones, que pueden ser usados en minería de datos también son aplicados a la limpieza de los datos.

- **Enriquecimiento.**

Se agrega información a los registros que "enriquece" la información inicial, esta información puede ser nuevos datos o conocimiento que transforme los datos originales.

Ej:

- Agregar la ciudad a la dirección.
- Agregar la distancia de la ciudad de los centros de distribución.
- Se puede conectar con la próxima etapa y usar conocimiento extra para codificar la información (semántica →cuantitativa).

Cliente Nº	Nombre	Edad	Dirección	País/ciudad	Fecha Compra	Tipo
54011	Johnson	54	12 road street 20	USA/NY	25/12/98	Arte
54012	Pacheco	23	13 Lasalle	AR/Bu.Ai.	*01/01/01*	Musica
54013	Stabros	43	129 Liking	GR/Colo.	29/06/95	Edu.
54014	Martinez	33	2 Plaza	SP/Bilbao	05/11/97	Ficción
54015	Matuz	27	25 Av. Terra	SP/Madrid	*11/11/11*	Arte y Edu.
54016	Müller	19	134 Lutero	GR/Berl.	15/12/98	Edu.

• Codificación.

En general las etapas anteriores pueden ser realizadas usando sentencias SQL (excepto limpieza).

En esta etapa se debe decidir lo que sucede con los registros que falta información o con los registros que contienen información inconsistente. En general estos registros son eliminados, puesto que en MD se cuenta con suficiente información para tener consistencia estadística. Pero se debe tener cuidado puesto que estos casos pueden ser una fuente potencial de fraude o la falta de información pueden entregar patrones de interés para su análisis.

Cuando los registros son escasos es posible aplicar algunas técnicas para completar los faltantes.

Cliente Nº	Edad	Región	Cantidad	Arte	Música	Fic.	Edu.
54011	54	1	10	1	0	0	0
54012	23	10	7	0	1	0	0
54013	43	20	3	0	0	0	1
54014	33	21	5	0	0	1	0
54015	27	21	4	1	0	0	1
54016	19	27	2	0	0	0	1

Cuando una variable es de tipo cualitativo (de cardinalidad *n*) es común utilizar una representación con *n* variables binarias "flattening".

1.3. Hipótesis del aprendizaje automático.

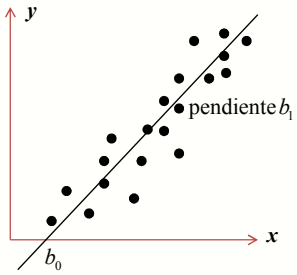
"Se dice que un programa computacional aprende de la experiencia ε una tarea τ, con una medida de eficiencia ρ. Si el desempeño en τ, medido con ρ, mejora con la experiencia ε.

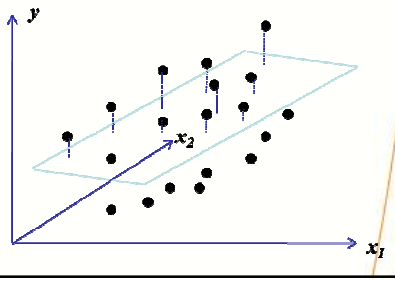
En general esas tares serán:

- identificación de grupos
- clasificación
- determinación de funciones desconocidas (predicción).

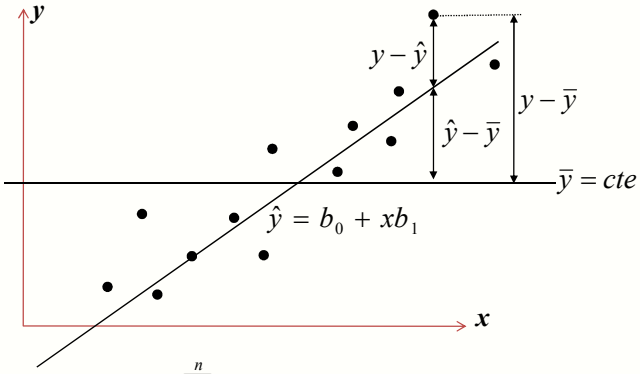
1.3.1. Problema sesgo v/s varianza

Modelo lineal. $$\hat{y} = \boldsymbol{E}[\mu_i] = \beta_0 + \sum_{i=1}^{i=p} \boldsymbol{x}_i \beta_i = \vec{\beta}^T \vec{\boldsymbol{x}}$$





Errores de la regresión



$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1)\,\mathrm{var}(y) = (n-1)\hat{\sigma}_y$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

SCT = SCR + SCE

Modelo no lineal

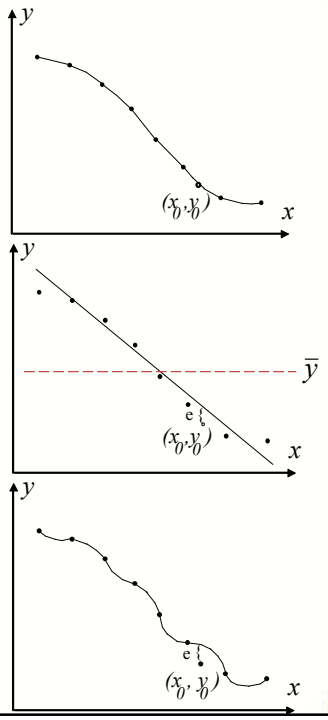
- Función no lineal a estimar, se el punto (x_0, y_0) desconocido.

- Si realizamos un ajuste lineal, y se requiere averiguar el punto desconocido, se genera error.

$$sesgo^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{y} - f(x_i))^2$$

- Si ajustamos un polinomio a todos los datos, también hay error

$$\text{var}\,ianza(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - \bar{y})^2$$

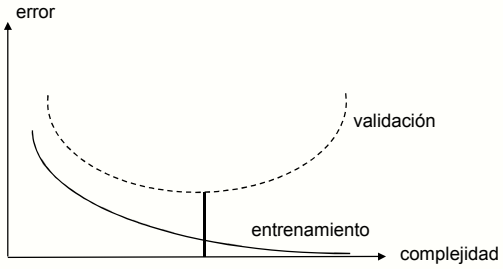


- Cuando el modelo tiene un gran sesgo, el modelo esta lejos de la solución y existe un sub-ajuste a los datos.
- Cuando el modelo se ajusta a todos los datos incluso el ruido, el modelo esta sobre-ajustado.

Se requiere un balance ente sesgo y varianza.

1.3.2. Procedimiento de ajuste de complejidad

Se utiliza un conjunto de datos disjunto (conjunto de validación)



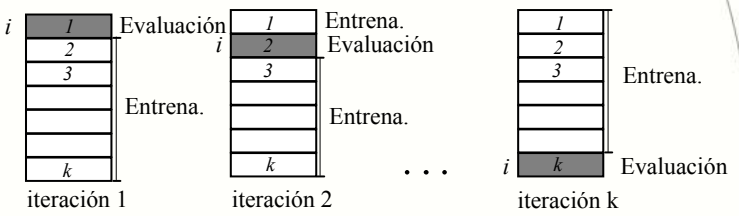
Para evaluar el modelo final se requiere un tercer conjunto llamado de prueba ("publication set")

Validación Cruzada

Usar sólo un conjunto de datos disjuntos, no garantiza que el error se mantendrá con otro conjunto de datos y no es posible definir con una cierta probabilidad que el error se mantenga dentro de un intervalo.

Para solucionar este problema se recurre al concepto de la validación-cruzada que consiste en entrenar y evaluar varias veces el mismo conjunto de datos.

Suponiendo que se cuenta con un conjunto de *n* datos tanto para entrenamiento como para prueba, se separa en *k* conjuntos diferentes seleccionados aleatoriamente, obteniéndose conjuntos de tamaño *n/k*, se obtiene un modelo con *k-1* grupos (aprendizaje), *n-n/k* casos y se evalúa con el grupo que no se entrenó, *k* casos.



Esta operación se realiza *k* veces y se calculan los errores d_i cometidos en cada uno de los grupos de prueba (o evaluación).

Con estos valores se calcula la media de los errores en la totalidad de grupos como:

$$\bar{\delta} = \frac{1}{\boldsymbol{k}} \sum_{i=1}^{k} \delta_i$$

dado que no se conoce la desviación estándar de este parámetro se debe estimar por:

$$\hat{\sigma}_{\bar{\delta}} = \sqrt{\frac{\sum_{i=1}^{k} (\delta_i - \bar{\delta})^2}{\boldsymbol{k}(\boldsymbol{k}-1)}}$$

En este caso es posible estimar un intervalo de confianza para estos errores medios con un grado o nivel de confianza *Con=1-a*.

$$\bar{\delta} \pm \boldsymbol{t}_{(\boldsymbol{k}-1),\alpha/2}\ \hat{\sigma}_{\bar{\delta}}$$

Una condición para ajustar la aproximación de la normal es que el tamaño de los grupos *n/k >30*.

Es fácil notar que, a medida que el número de grupos crece, el tamaño del intervalo de confianza disminuye y la distribución ***t*** tienda la normal.

Proceso de regularización

Existen otros métodos para evitar usar un conjunto de datos de validación.

Se crea una nueva función de error:

$$\varepsilon = (error\ de\ los\ datos) + \lambda(complejidad\ del\ \mathrm{mod}elo)$$

Minimizando ε se busca aumentar el ajuste de los datos y simultáneamente castigar los modelos mas complejos.

Cuando se escoge un λ grande, se restringe la elección a modelos simples.

La navaja de Occam y Descriptor del largo mínimo

Occanismo:

Si se tienen dos o más hipótesis, lo más razonable es aceptar la más simple; o sea la que presenta menos supuestos no probados.

Principio parsimonia: si hay dos o mas explicaciones en igualdad de condiciones, no hay que tener en cuanta una explicación complicada si existe una más simple.

"No significa que la que la explicación más simple sea la más correcta, sino que existen más probabilidades que sea cierta y que es preferible elegirla hasta que haya razones bien fundamentadas para adoptar una alternativa más compleja"

Longitud de Descripción Mínima

Este trabajo se basa en la teoría de la complejidad estocástica, basada en las teorías de Kolmogorov.

- Alfonso X "el sabio": *Si Dios Nuestro Señor me hubiera consultado antes de crear el mundo, le hubiera recomendado que hiciera algo más sencillo.*

Este metodología intenta dar formalidad al principio de parsimonia.

La mínima longitud de descripción de un vector $\vec{x} = [x_1, x_2, ..., x_n]$ usando *p* parámetros $\vec{\theta} = [\theta_1, \theta_2, ..., \theta_n]$

$$MDL(p) = -\log(P(\vec{x} / \hat{\theta})\pi(\hat{\theta})) + \frac{p}{2}\log(n) + O(p)$$

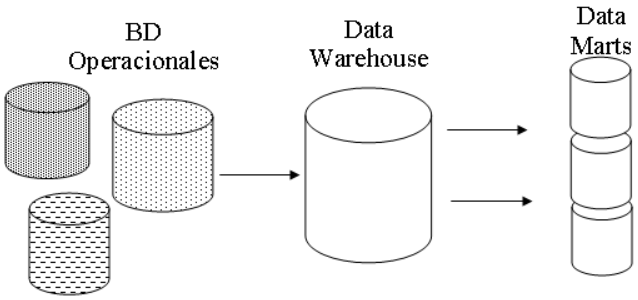
Donde $\pi(\theta)$ es la distribución de probabilidades en función de los parámetros y *O(p)* la complejidad del modelo.

1.4. Bases de datos Analíticas (Data Warehouse)

Las Bases de Datos operacionales no están diseñadas para realizar análisis sobre su contenido. Las sentencias SQL no facilitan las consultas tendientes a obtención de conocimiento, la mejor aproximación se logra con consultas de tipo estadísticas descriptivas.

Para realizar el trabajo analítico se requieren bases especialmente diseñadas para la toma de decisiones estratégicas, las cuales usan como fuentes de información las BD operacionales.

En general, los datos operacionales sufren una reducción de dimensionalidad mediante un muestreo.



- El diseño del DW requiere de especialistas en el conocimiento contenido, BD, redes de computadores y hardware.

- En general se utilizan diseños espaciales en la BD, puesto que muy a menudo se requiere un acceso de alta velocidad al módulo de datos.

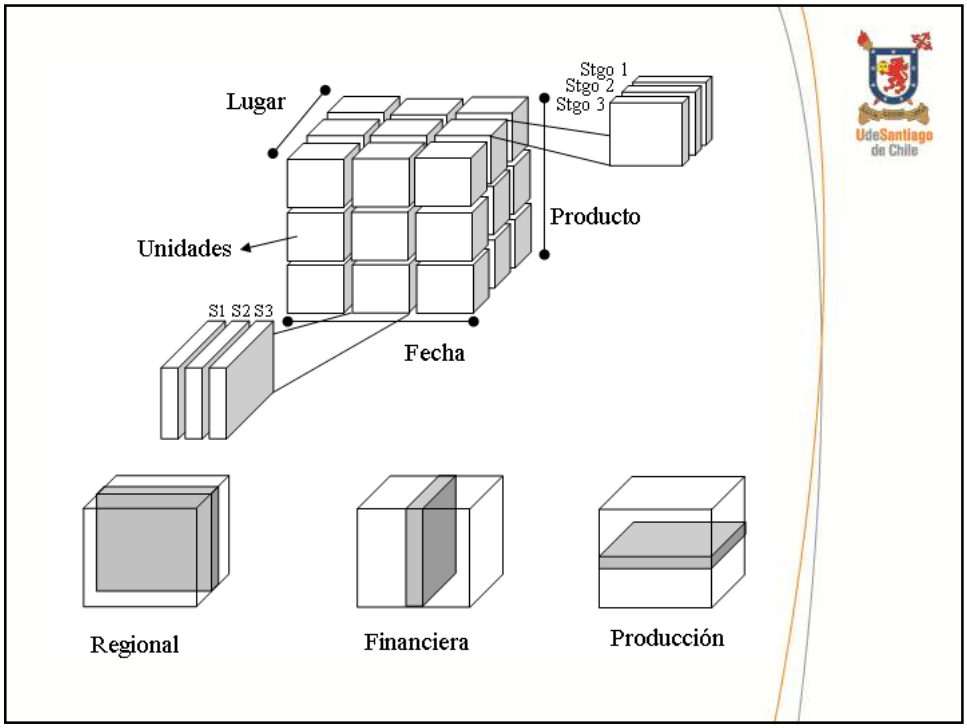
Para el análisis es posible extraer parte de los datos del DW y procesarlos en servidores locales donde existan herramientas especiales de minería de datos para satisfacer los requerimientos del usuario.

Para la toma de decisiones en línea, muchas veces se requiere trabajar con un conjunto elevado de tablas, aumentando la carga del sistema. Para esto el DW requiere máquinas de alta velocidad y una variedad de procesos optimizados.

Una de las estructuras mas usadas en los DW son los arreglos multidimensionales (hipercubo).

Ej: Análisis de una librería con tiendas a nivel nacional.

Producto	Lugar	Fecha Compra	Unidades
CD	Santiago1	Mes 1	1500
Libro	Linares	Mes 1	150
Revista	Temuco1	Mes 1	506
CD	Santiago2	Mes 2	1020
CD	Santiago3	Mes 3	1567


Lugar
Stgo 1
Stgo 2
Stgo 3
Producto
Unidades
S1 S2 S3
Fecha
Regional
Financiera
Producción