IN643 – Introducción a la Minería de Datos
Otoño 2010

Aplicaciones I

Universidad de Chile

Departamento de Ingeniería Industrial

Profesor: Richard Weber (rweber@dii.uchile.cl)

Prof. Auxiliar: Gastón L'Huillier (glhuilli@dcc.uchile.cl)

Contenido

- Análisis de comportamiento de clientes en un banco
- Predicción de demanda

Identifying web usage behavior of bank customers

Sandro Araya[1)], Mariano Silva[2)], Richard Weber[3)]

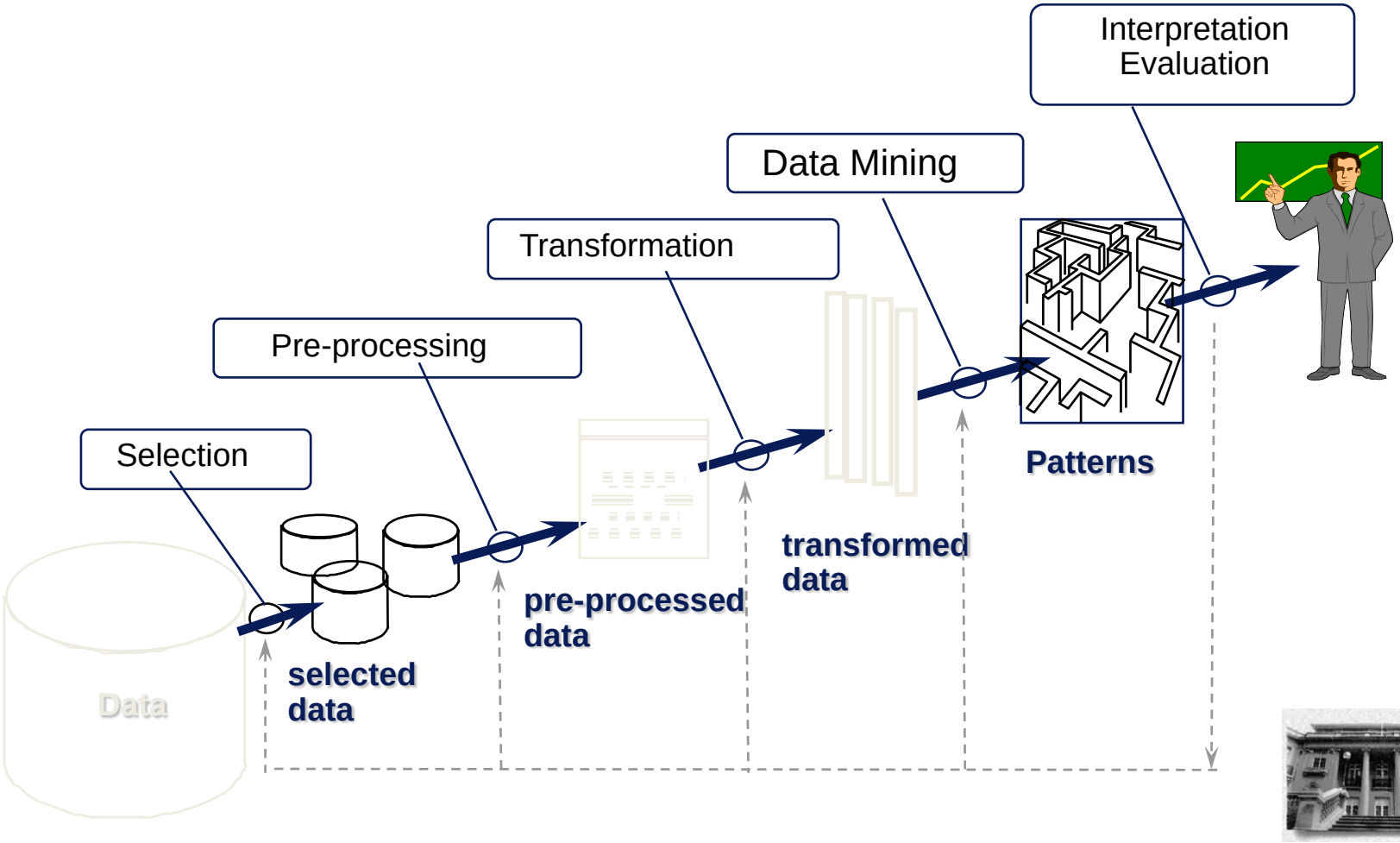
1) BCI Bank, Santiago, Chile
2) webmining.cl, Santiago, Chile

3) Department of Industrial Engineering, Universidad de Chile, Santiago, Chile

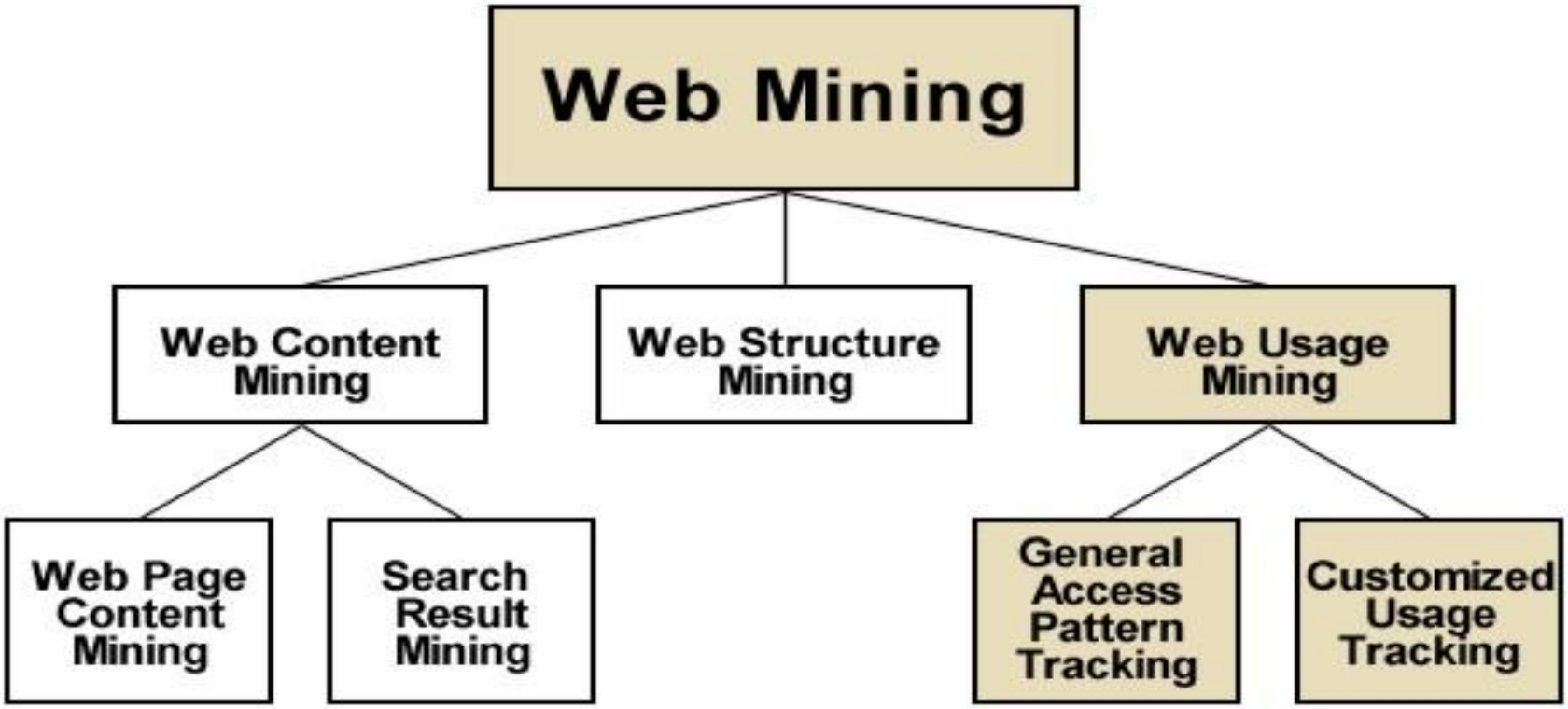
BCI - Banco de Crédito e Inversiones (www.bci.cl)

- Founded in 1937
- Started Virtual Bank in 1996
- 10,000+ Internet transactions daily

Process of knowledge discovery in databases (KDD)

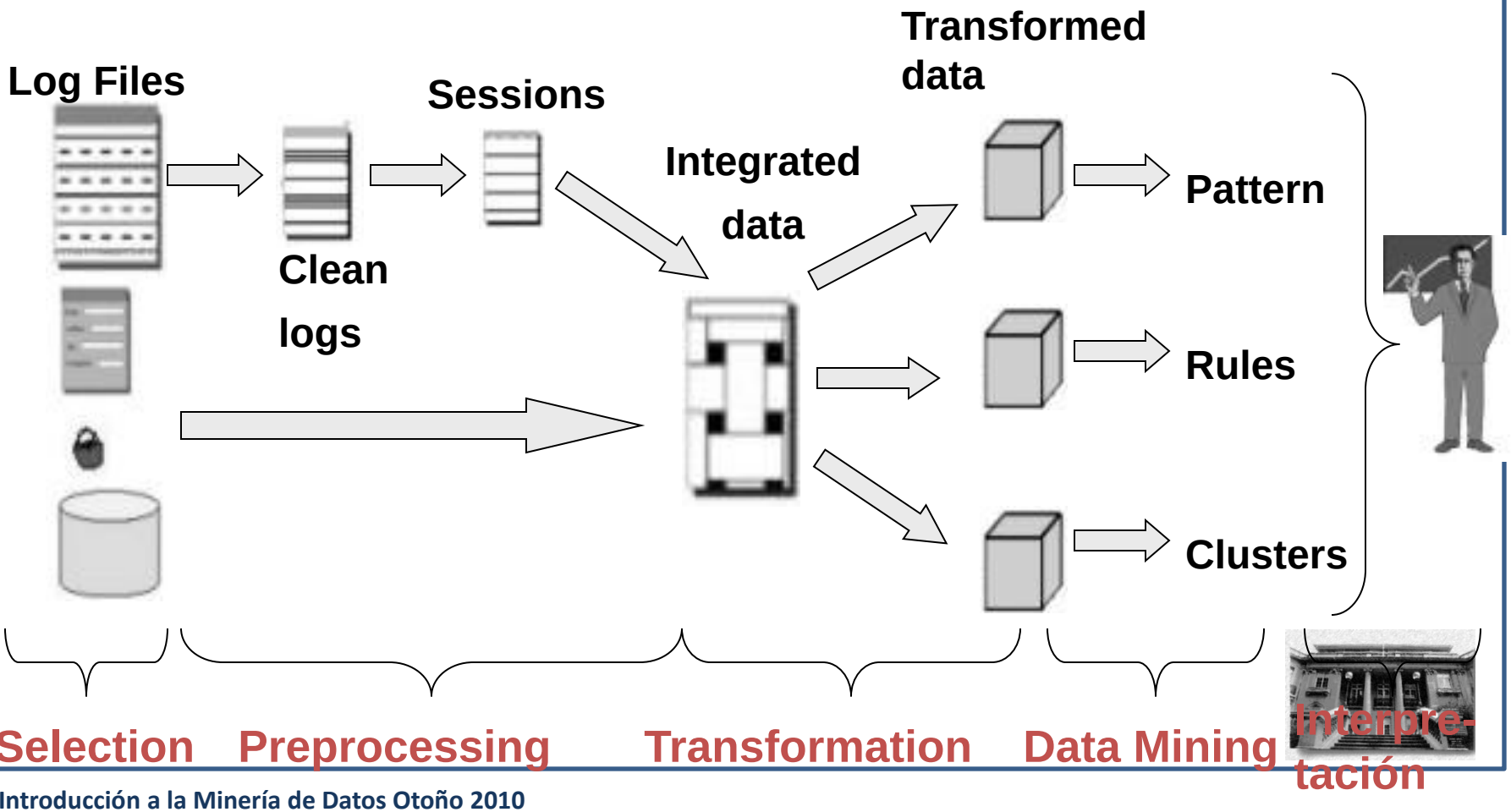


Application areas of Web Mining



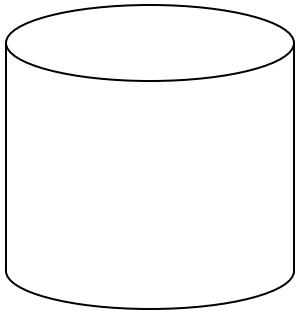
Methodology of Web Mining

Combination of KDD process and Web Traffic analysis



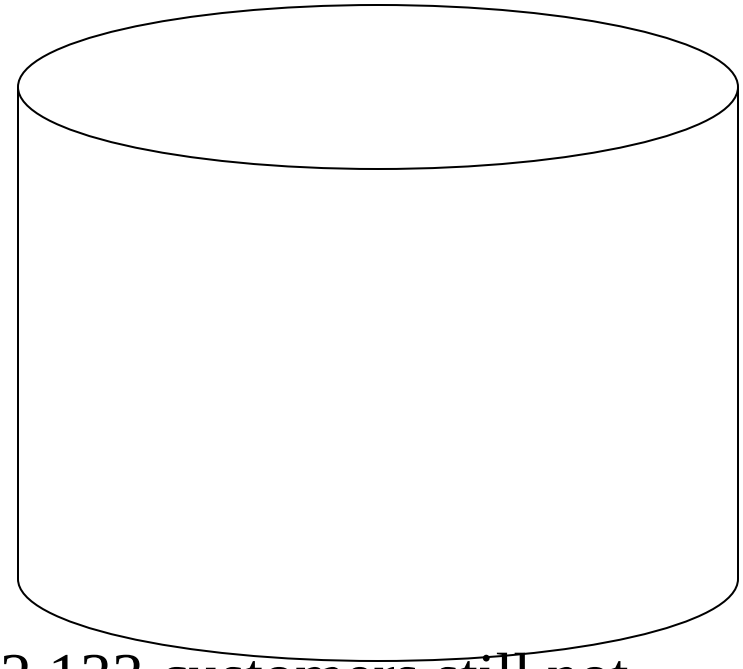
Current situation

Registed Visitors of Virtual Bank



41,563 navigating customers

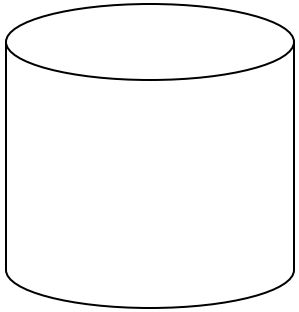
(Traditional) Bank Customers



142,133 customers still not visitors of the web site

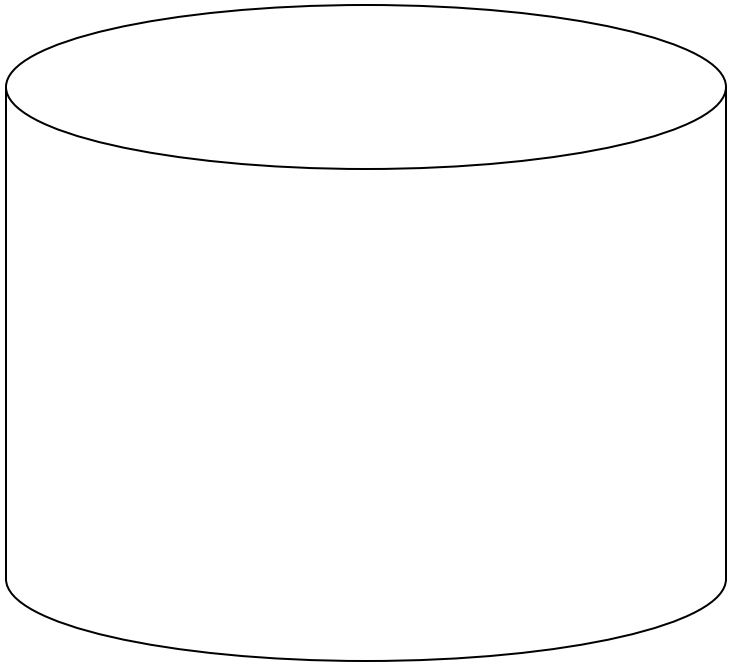
Questions

Virtual bank



- How do my navigating customers behave?
- Are there segments of "typical visitors"?
- Is it possible to identify "heavy users"?

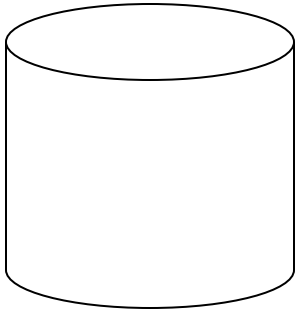
Traditional bank



- Are there customers that look like "heavy users"?
- How can I convert these "twins of heavy users" to users of my web site?

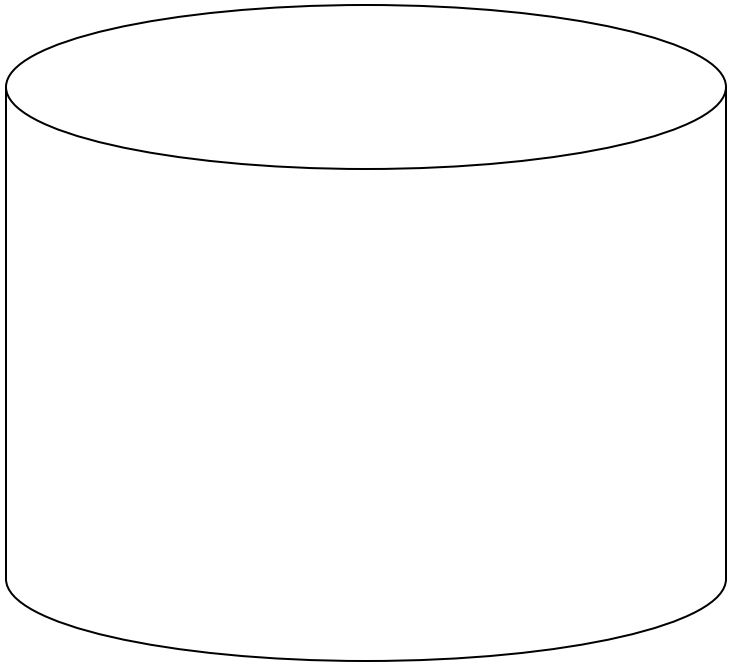
Two step approach

Virtual bank



- Clustering of navigating customers
- Determine profile of "heavy users"
- => Fuzzy Clustering

Traditional bank

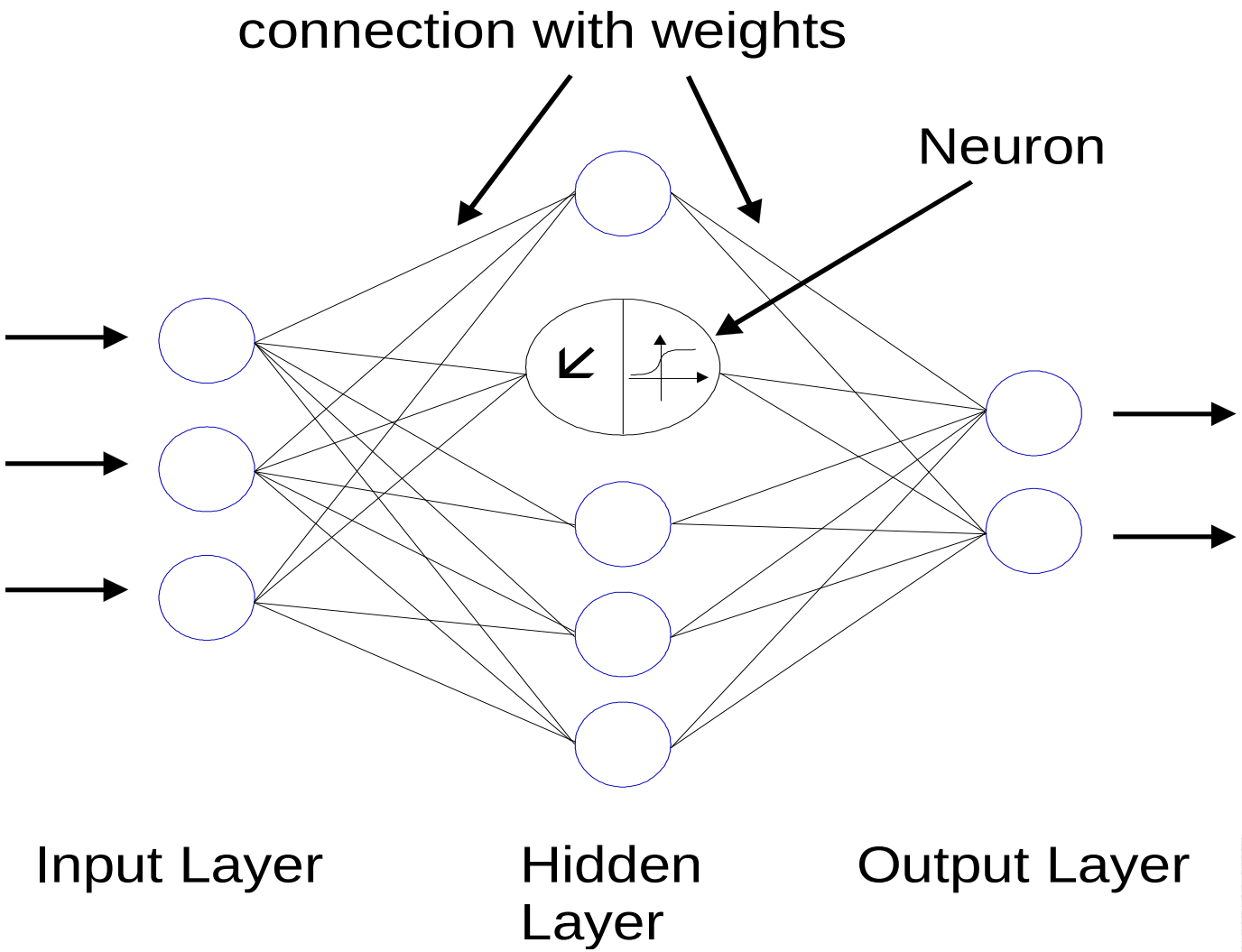


- Search for (traditional) customers that have a profile similar to that of "heavy users"
- Marketing campaign directed to these "twins of heavy users"
- => Neural Network

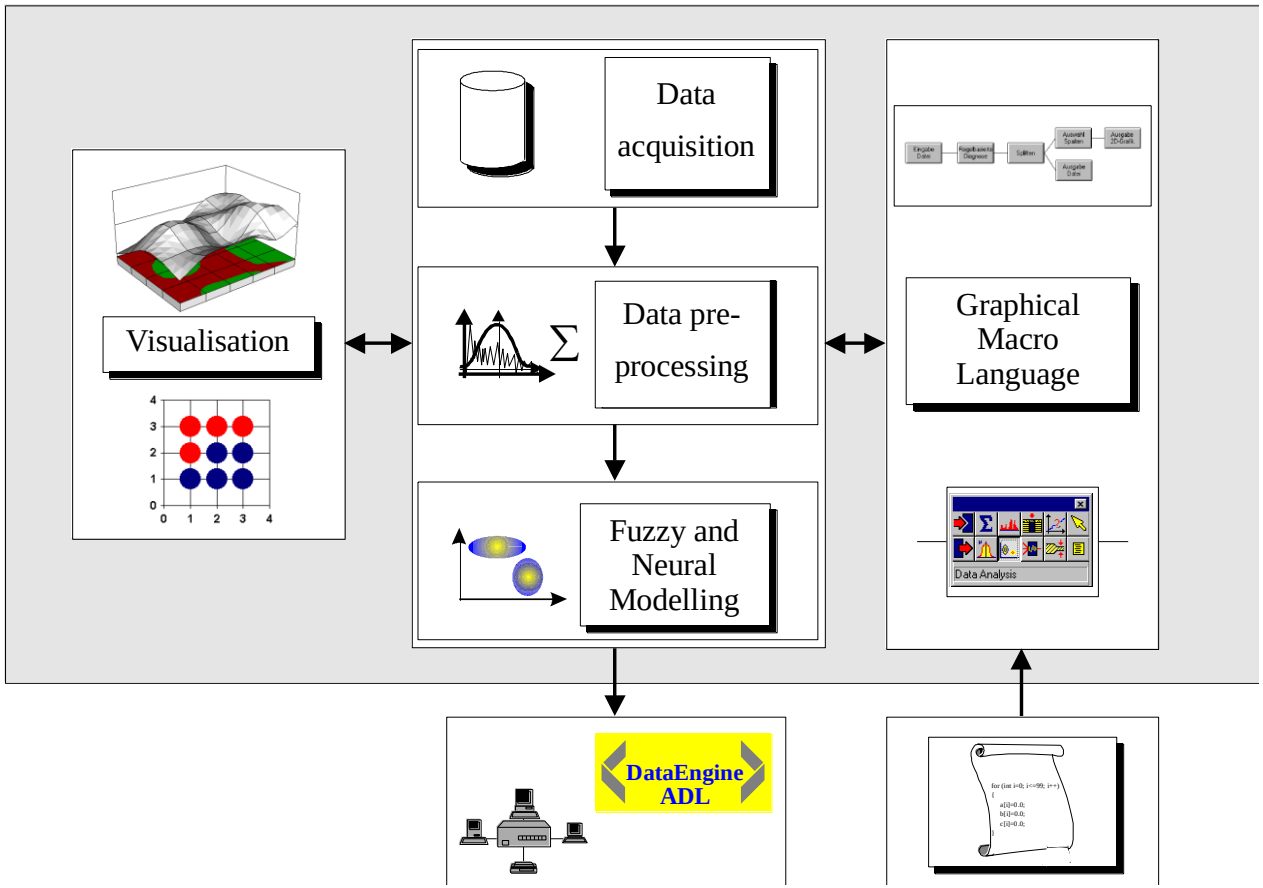
Results of Segmentation

Class	Age (years)	Trx Web	N° of Cases	% Cases
Class L1	38	25	9130	22.0%
Class L2	29	26	4277	10.3%
Class M1	58	31	4599	11.1%
Class M2	47	32	11829	28.5%
Class H	34	141	11728	28.2%
TOTAL			41563	100.0%

Neural networks (Multilayer Perceptron)



DataEngine



www.dataengine.de

Identification of twins with Neural networks

Architecture of the Multilayer Perceptron:

Number of input neurons: 6,
corresponding to the attributes: age, gender, civil status, education, income, and antiquity.

Number of neurons in the hidden layer: 12 (transfer function: sigmoid)

Number of output neurons: 5,
corresponding to the 5 classes of customers: H, L1, L2, M1 and M2.

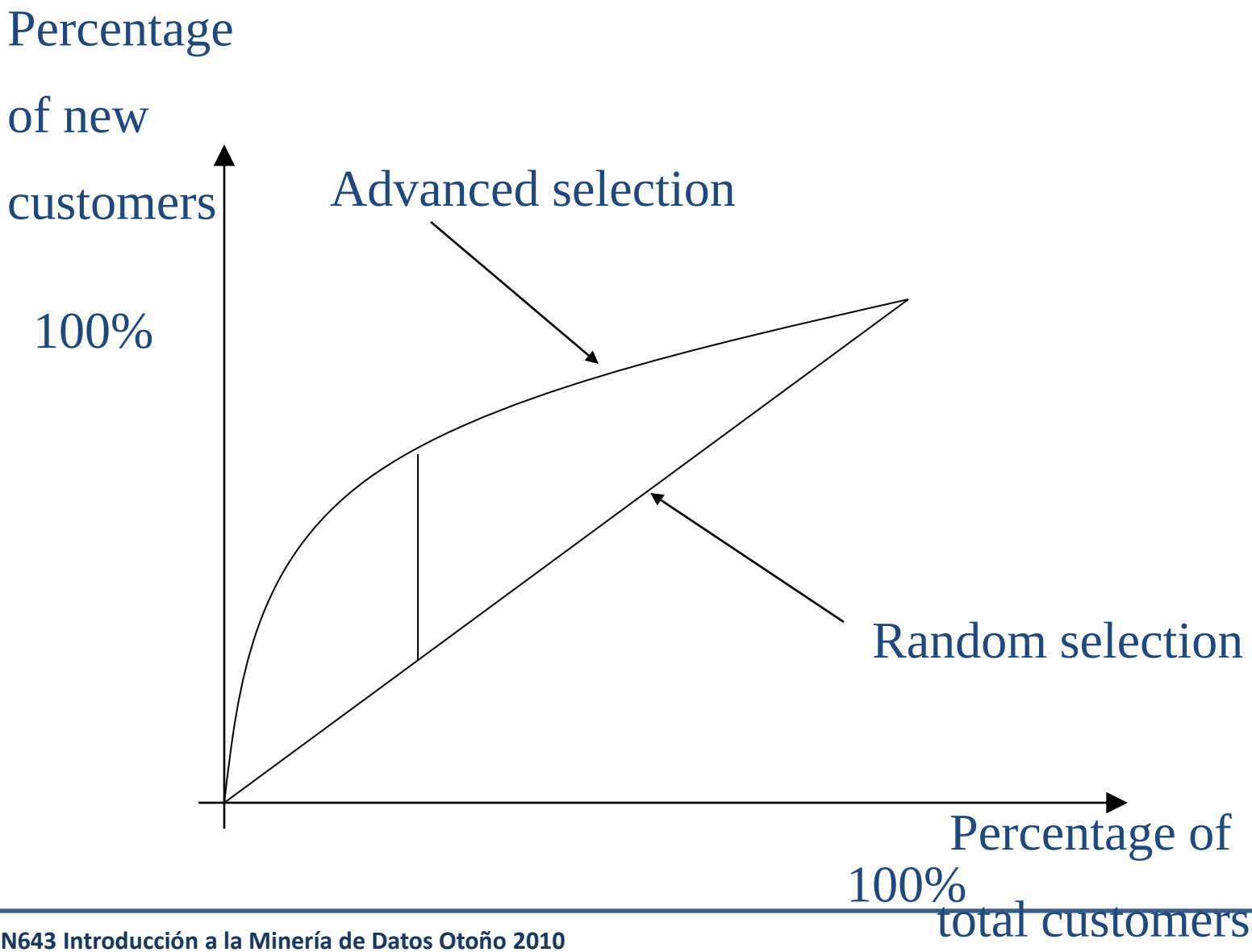
Neural Network Results

Class	Selected Cases	% Cases
L1	32,602	22.9%
L2	25,216	17.7%
M1	35,805	25.2%
M2	18,608	13.1%
H	29,902	21.0%
TOTAL	142,133	100.0%

Marketing Campaign

	Received mailing	Did not receive mailing	Total
Customers from class H	11,567	18,335	29,902
Other customers	15,806	96,425	112,231
Total	27,373	114,760	142,133

Gains Chart


Percentage
of new
customers
100%
Advanced selection
Random selection
Percentage of
100%
total customers

Marketing Campaign

Week	New visitors from class H that received the mailing	New visitors from class H that did not receive the mailing	New visitors from class H (total)
13	737	256	993
14	153	264	417
15	114	212	326
16	101	204	305

Results

New visitors from class H
that received the mailing

Week	New Visitors
13	737
14	153
15	114
16	101
TOTAL	**1,105**

$$\text{Response rate Twins} = \frac{1.105}{11.0567} = 10\%$$

Results

New visitors from class H
that did not receive the mailing

Week	New Visitors
13	256
14	264
15	212
16	204
TOTAL	936

Connection rate of twins of heavy users without mailing $\frac{936}{18.300} = 5\%$

Conclusion

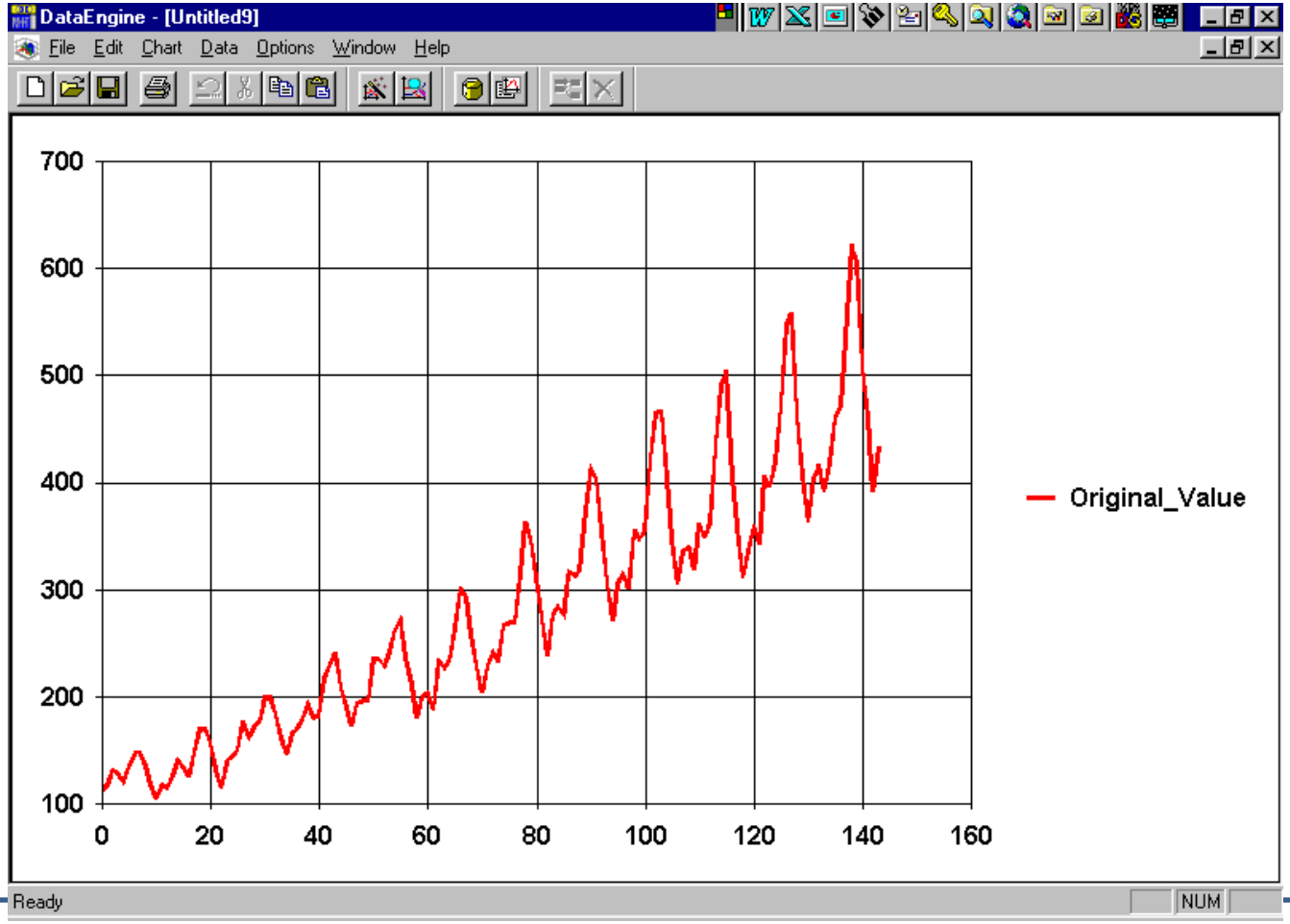
”Natural connecting rate” ~ 1.050 new customers /week

~ 2% of web site users

Response rate after mailing to “twins of heavy users” = 10%

Natural connecting rate of “twins of heavy users”
(i.e. without receiving mailing) = 5%

Predicción de una Serie de Tiempo



Serie de tiempo:

Número de pasajeros de una línea aérea

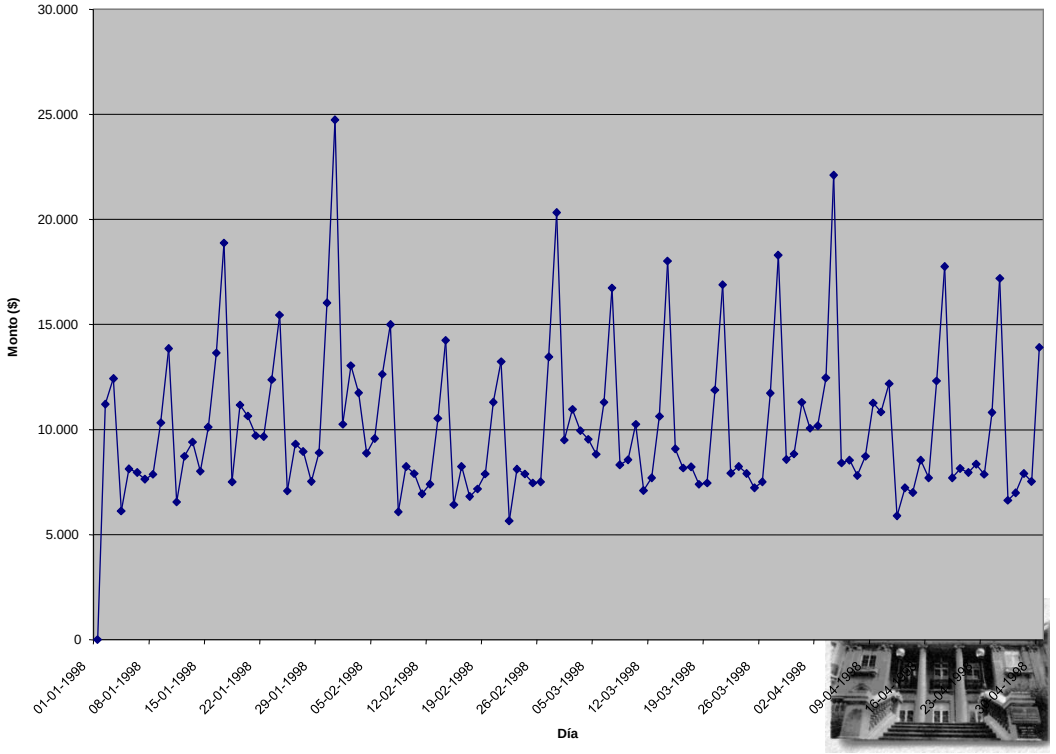
Motivación del Problema

Necesidad de pronóstico de ventas en el corto plazo

Motivación del Problema

- Ventas ... De qué dependen?

- Ventas pasadas
- Precios
- Campañas Publicitarias
- Estacionalidad
- Festivos
- Clima
- Venta de productos similares



Motivación del Problema

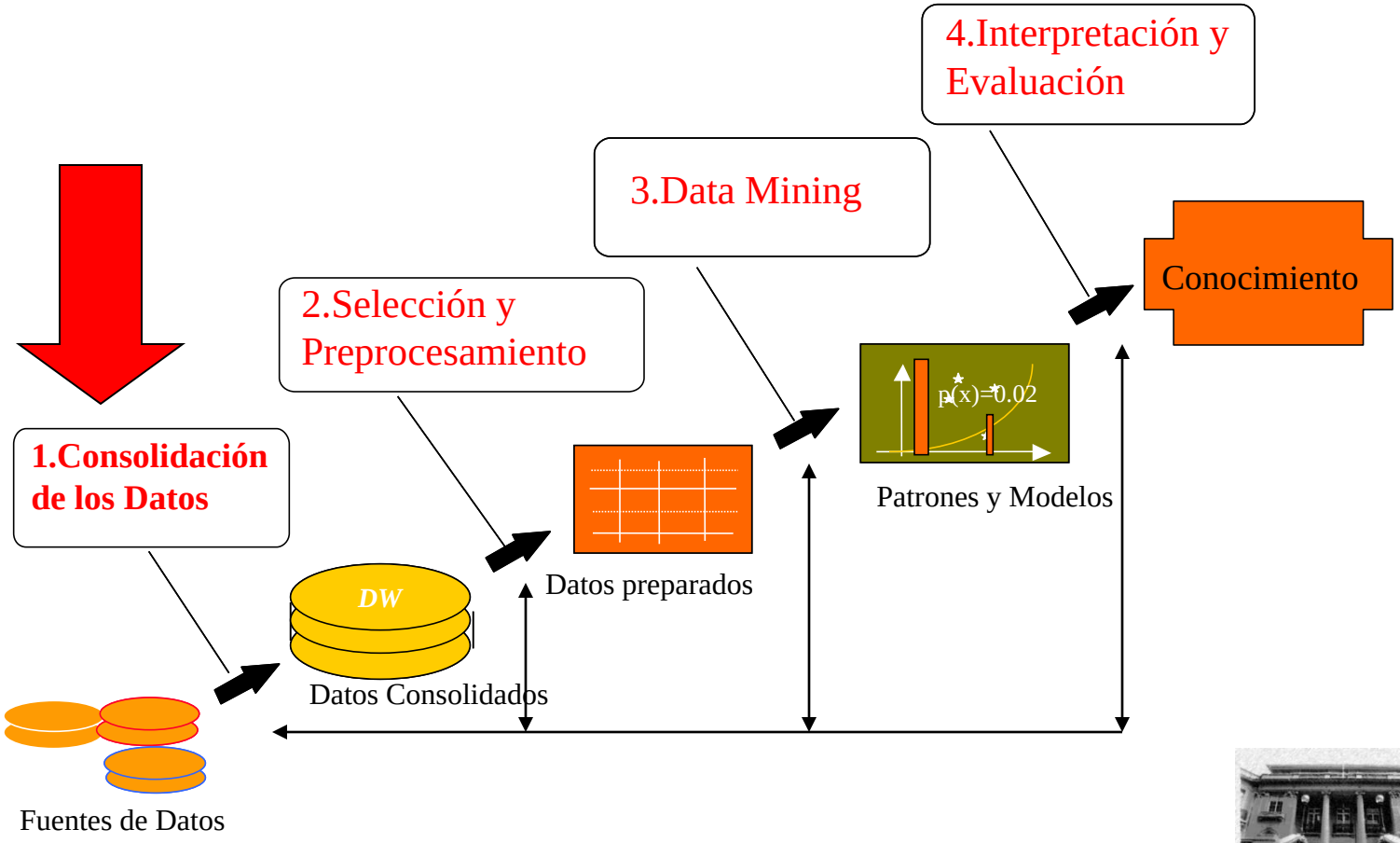
- ¿Cómo administrar el inventario?
  - **Muy poco → Quiebres de Venta. Clientes insatisfechos**
  - **Mucho → Costos de capital**
- Desarrollar mejores técnicas de pronóstico y de acuerdo a esto gestionar nuestro inventario

Alcances del Proyecto

Se acotará el ámbito de estudio a:

- Local La Pintana: Supermercado Tradicional con 4.500 m^2
- Un subconjunto de productos: 50 PLU's más vendidos en el local (representan el 23,18% de las ventas)
- Con datos desde el 12/09/2000 a 31/07/2001

Knowledge Discovery in Databases: KDD



1.Consolidación de los Datos

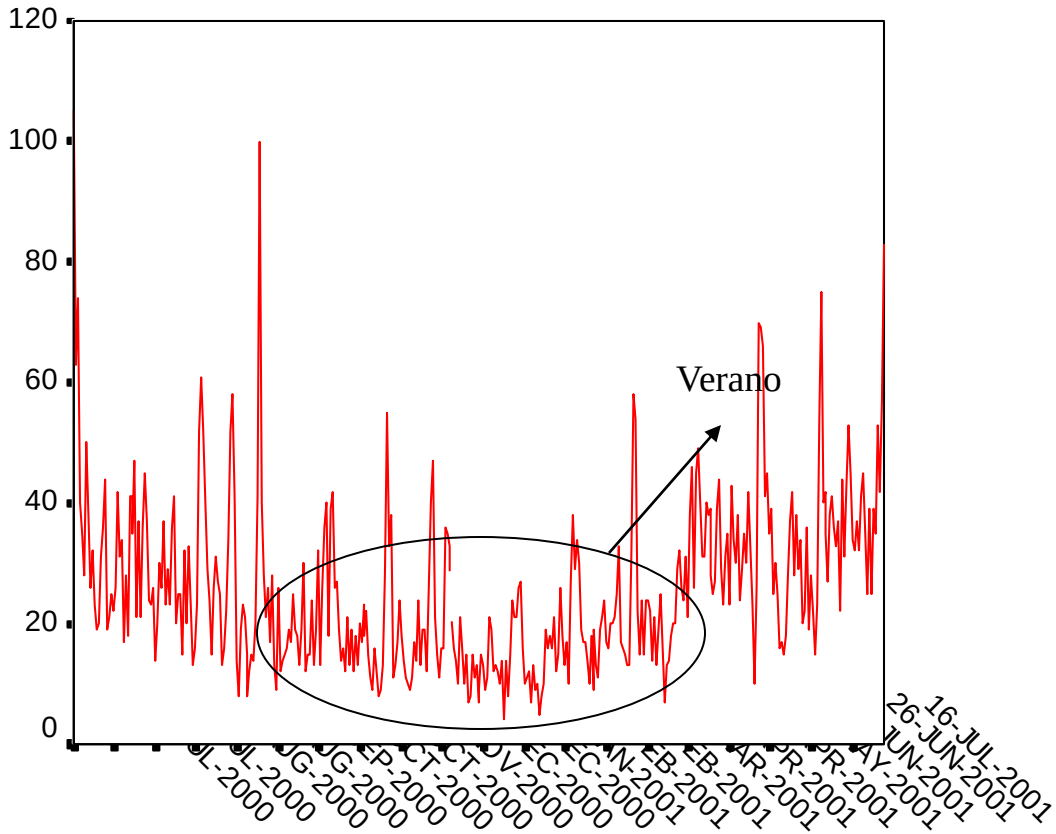
- Datos de diferentes fuentes:

- ORION ➔ Unidades Vendidas en local La Pintana desde 01/07/00 al 31/07/01 para los 50 PLU's más vendidos

- AC Nielsen ==> Precios semanales de los productos en el local de estudio y la competencia del micromercado (Santa Isabel, Ekono y Lider)

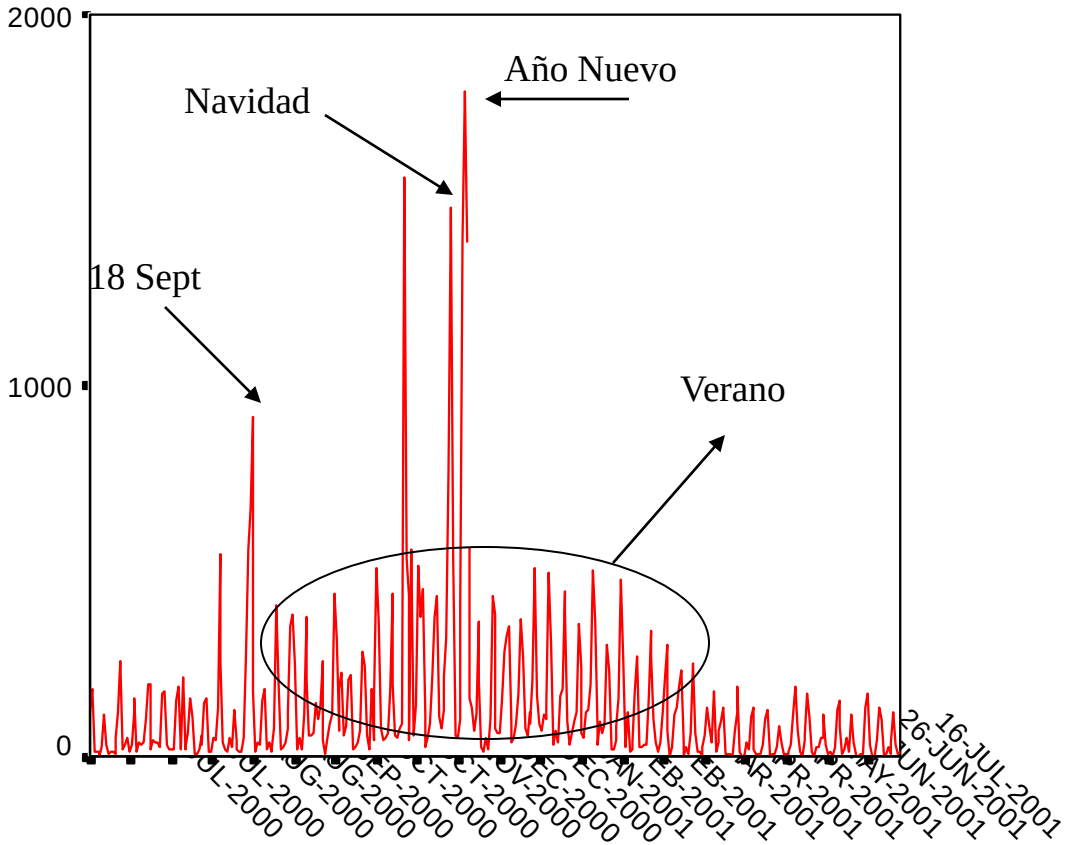
1.Consolidación de los Datos:

Café 170 grs.



1.Consolidación de los Datos:

Cerveza 1 Lt.

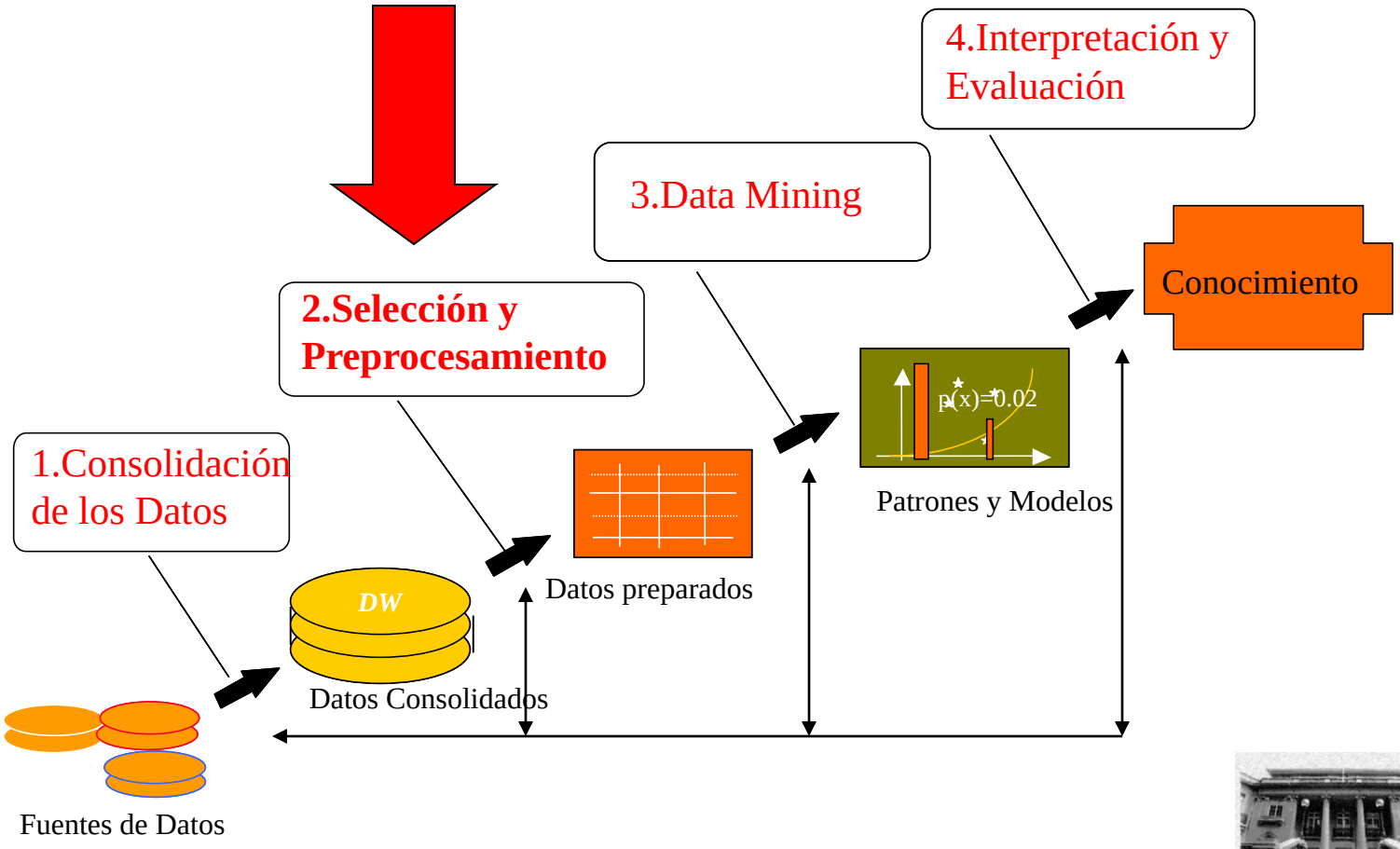


1.Consolidación de los Datos

Características del día. Variables binarias (0,1)

- pago :Días de pago de fin de mes.
- quincena :Días de pago de quincena
- prefest :Días anteriores a feriados
- feriado :Días festivos
- patrias :Días de fiestas patrias
- santa :Días de semana santa
- vacation :Días de vacaciones (Enero y Febrero)
- verano :Días de meses estivales (desde 01/10 al 31/03)
- a_nuevo :1 de Enero. Único día del año donde los supermercados no venden.

Knowledge Discovery in Databases: KDD



2.Selección y Preprocesamiento

- "En la vida real los datos no están como quisiéramos"

- De los 50 PLU's originales hay 3 correspondientes a promociones
- De los 47 PLU's restantes 9 presentan ausencia de datos de más del 25% en la serie de tiempo

LIMPIEZA DE DATOS!!!

2.Preprocesamiento

- Las ventas se escalaron entre 0 y 1
- En base a los precios se crean las siguientes variables:

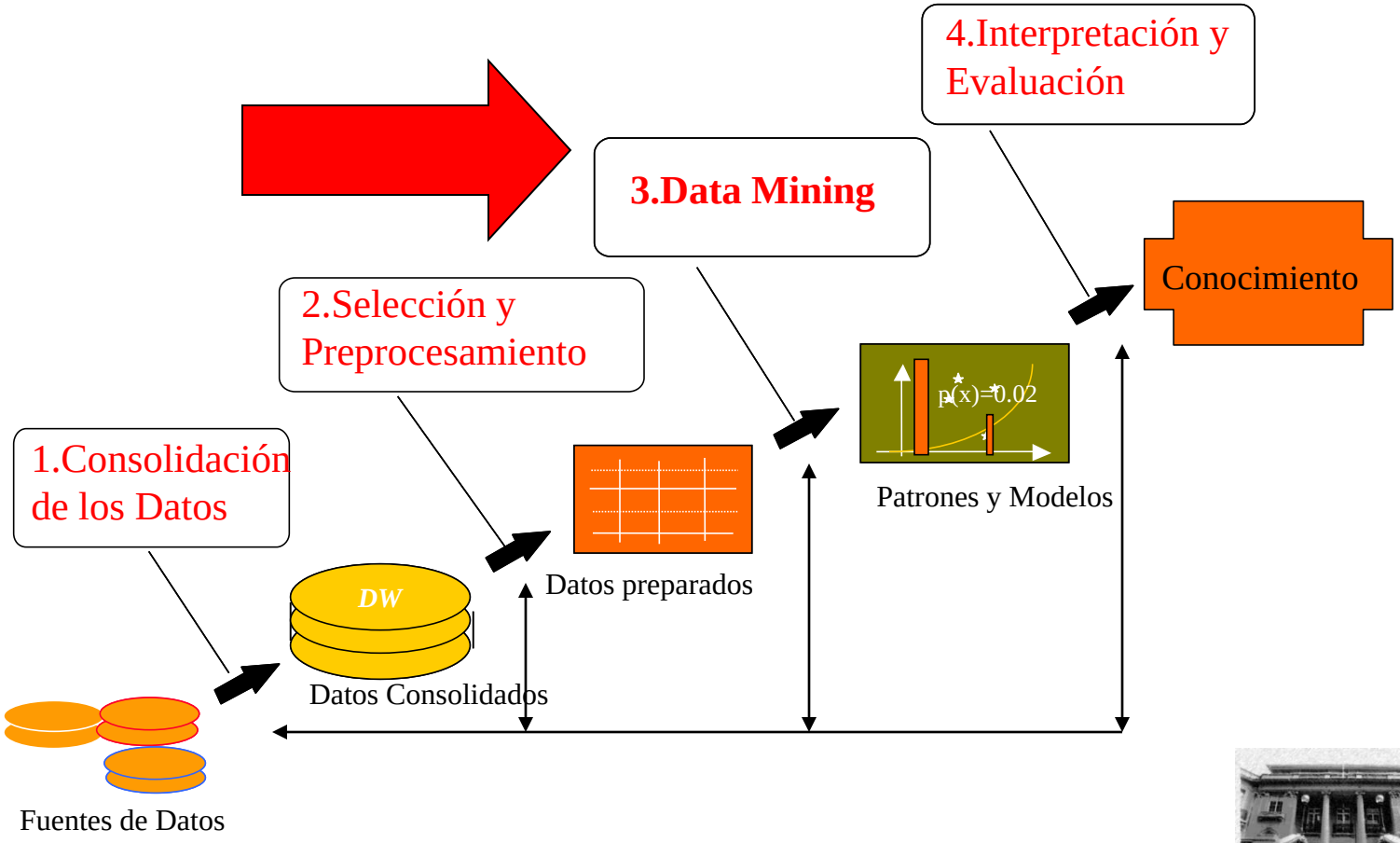
$$PA(N^{o}PLU) = precioPLU\_Economax$$

$$PB(N^{o}\,PLU) = \frac{precioPLU\_Economax}{MAX(precioPLU\_micromercado)}$$

$$PC(N^{o}\,PLU) = \frac{precioPLU\_Economax}{MIN(precioPLU\_micromercado)}$$

Estas variables también se reescalan entre 0 y 1

Knowledge Discovery in Databases: KDD



3.Data Mining: Enfoques de Solución

- Modelos Ingenuos (enfoque actual)
- Modelos Box Jenkins
  - SARIMAX (p,d,q) (sp,sd,sq) Y
- Redes Neuronales
  - Perceptrón Multicapas (MLP)

Análisis de Series de Tiempo

- Box, Jenkins (1976)
- MA(q) (FIR)

$$Xt = \sum_{n=1}^{q} b_n * e_{t-n} = b_1 e_{t-1} + \ldots + b_p e_{t-q}$$

- AR(p) (IIR)

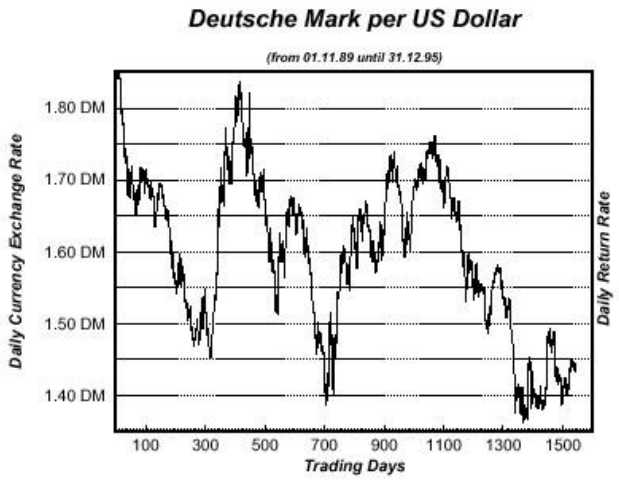
$$Xt = \sum_{i=1}^{p} a_i * x_{t-i} + e_t$$

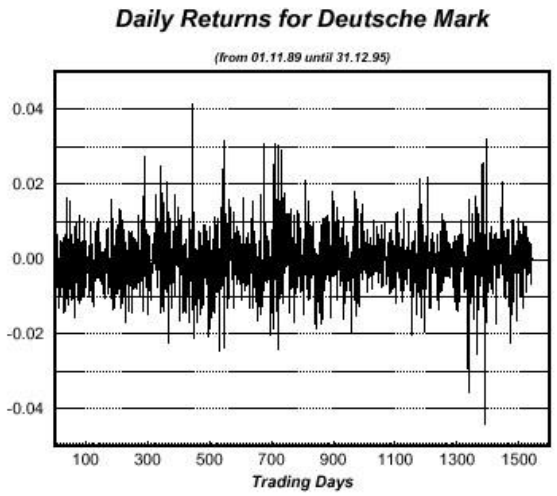
- ARMA (p,q)

$$Xt = \sum_{i=1}^{p} a_i * x_{t-i} + \sum_{n=1}^{q} b_n * e_{t-n} + e_t$$

Modelos Box Jenkins

- Requisitos de ARMA
  - **Al menos 50 observaciones**
  - **La serie debe ser estacionaria**



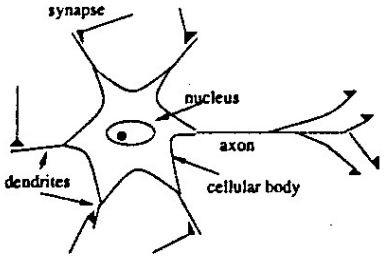


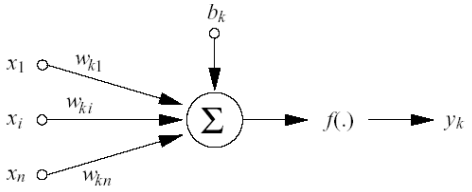
Modelos Box Jenkins

- Para convertir una serie no estacionaria en otra estacionaria se puede:
  - Aplicar transformaciones logarítmicas
  - Diferenciar la serie (X_t - X_{t-1})
- ARIMA(p,d,q) donde d es Nº de términos diferenciados
- Seasonal ARIMA: SARIMA (p,d,q) (sp,sd,sq)
- SARIMAX con X variables externas (regresores)

Redes Neuronales

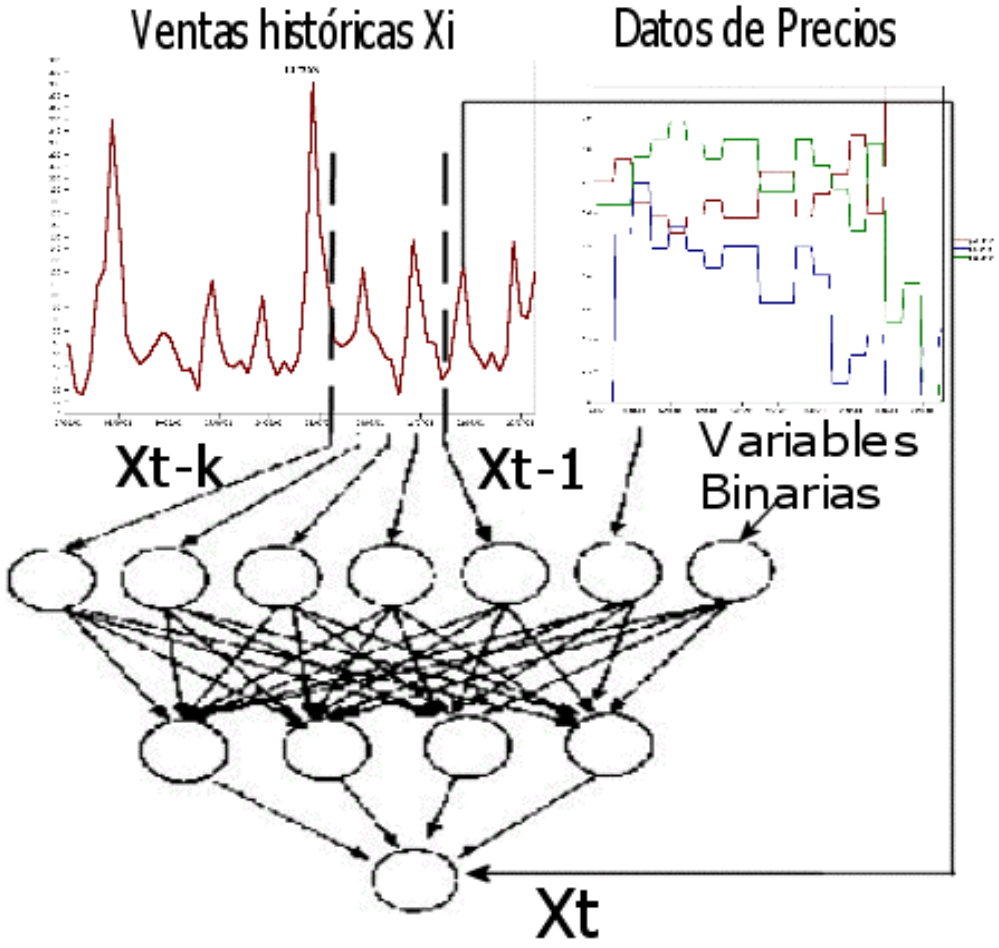
- Modelos de conectividad
- Resuelven problemas de:
  - Clasificación de patrones
  - Aproximación de funciones
  - Clustering
  - Optimización
  - Memoria asociativa
  - Predicción o pronóstico





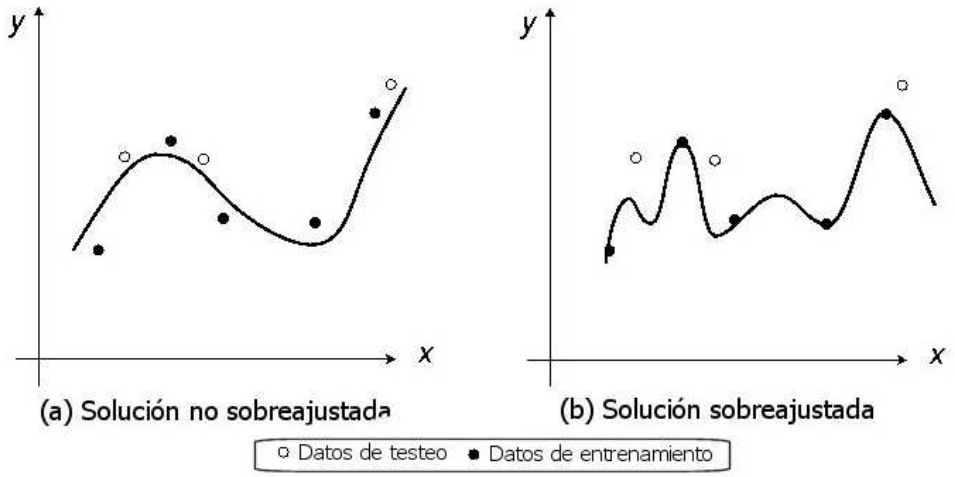
$$y_k = f(\sum_{i=0}^{n} w_{ik} x_i)$$

MLP para forecasting



Overfitting o Sobreajuste

- Sobreajuste de la red a los datos del problema y no al problema en sí



ARIMA v/s MLP

Modelo Estadístico (ARIMA)	Redes Neuronales (MLP)
Modelo lineal: asume un comportamiento de la serie a priori	Modelo no lineal: más grados de libertad para el modelo
La modelación requiere que la serie sea estacionaria	No impone requisitos estadísticos a la serie de tiempo a analizar
Requieren de conocimientos en Estadística e interacción con el usuario en la modelación	Requieren menor interacción con el usuario
El modelo entrega conocimiento e información en sus parámetros	Difícil lectura del modelo (caja negra)
Bajo peligro de sobreajustar el modelo	Fácil de sobreajustar el modelo a los datos

Desempeño del pronóstico: medidas de error

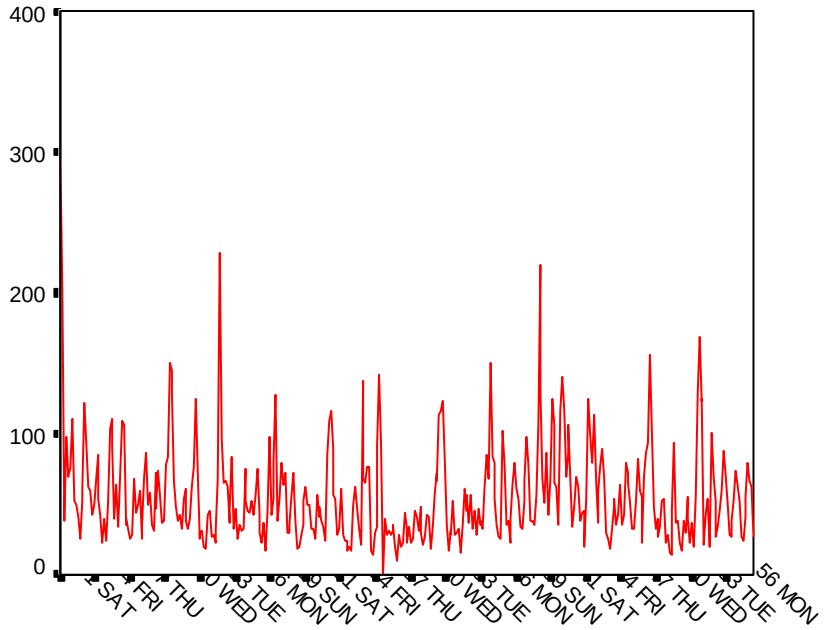
- Error Porcentual (Error porcentual absoluto medio)

$$\frac{1}{N}\sum_{k}\left|\frac{(y(k)-\hat{y}(k))}{y(k)}\right|$$

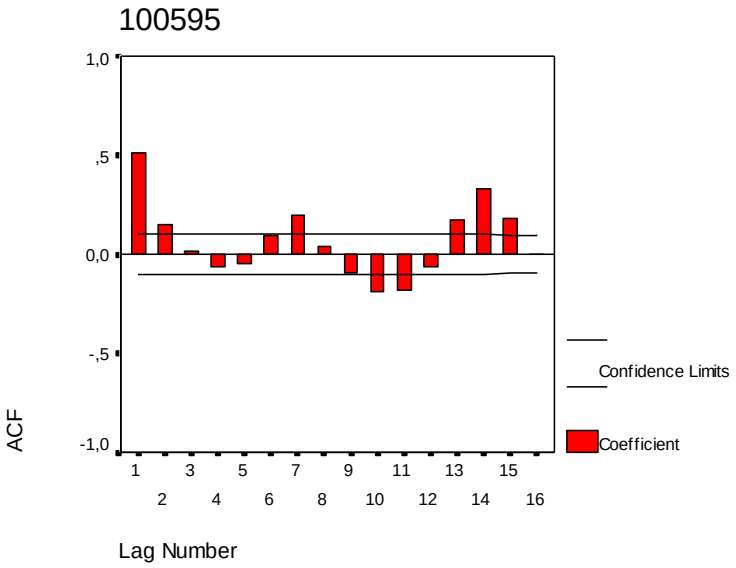
- Error Normalizado (Error cuadrático medio normalizado)

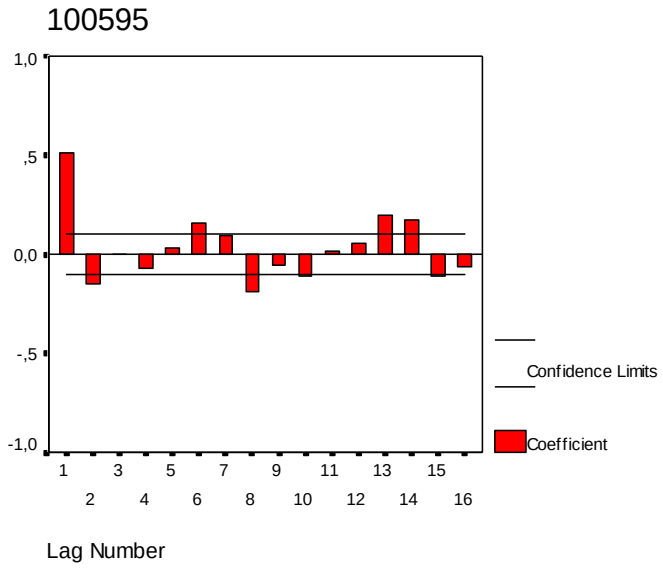
$$\frac{\sum_{k}(y(k)-\hat{y}(k))^2}{\sum_{k}(y(k)-\bar{y}(k))^2}=\frac{1}{\sigma^2 N}\sum_{k}(y(k)-\hat{y}(k))^2$$

Aplicación a PLU 100595 (Aceite Vegetal 1 Lt.)



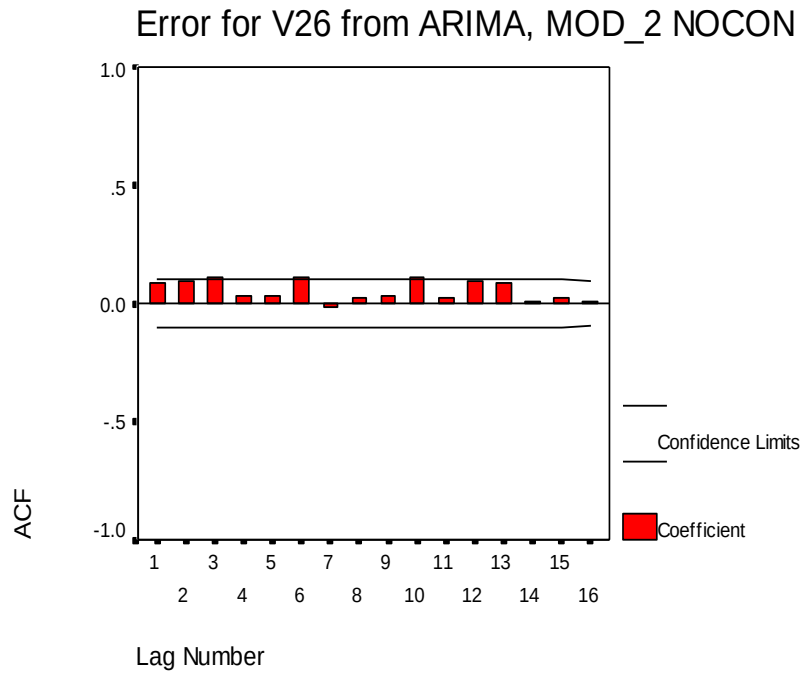
Aplicación de Box Jenkins

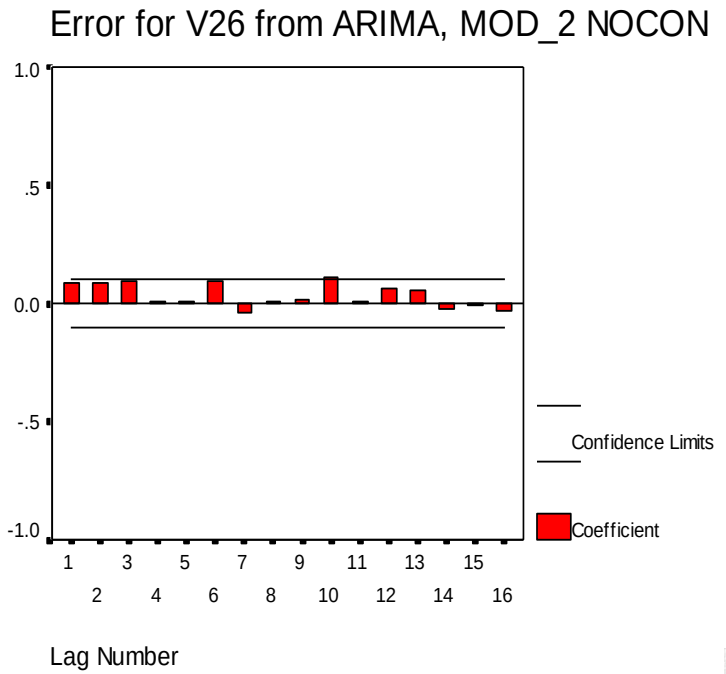




SARIMA (1,0,0) (2,0,0)

Aplicación de Box Jenkins





Modelos Tradicionales y MLP

100595	Conjunto de Entrenamiento		Conjunto de Testeo	
	Error Porcentual	Error Normalizado	Error Porcentual	Error Normalizado
ARIMA	36.21%	0.3301	40.49%	0.6090
Ingenuo	44.28%	0.6972	56.83%	1.2481
Ingenuo Estacional	64.67%	1.2212	45.75%	1.0217
Media Incondicional	59.98%	0.7759	48.54%	0.9689

100595	Conjunto de Entrenamiento		Conjunto de Testeo	
	Error Porcentual	Error Normalizado	Error Porcentual	Error Normalizado
MLPtw21	32.93%	0.4633	31.85%	0.4973
MLPtw14	31.15%	0.3115	34.64%	0.5703
MLPtw7	30.00%	0.3092	35.44%	0.5490
MLPtw6	32.45%	0.3761	33.53%	0.5112
MLPtw5	30.26%	0.3526	35.61%	0.5540
MLPtw3	29.61%	0.3002	34.36%	0.5281
MLPtw1	30.00%	0.3405	35.31%	0.5340
MLPtw0	34.12%	0.4760	31.80%	0.6244

En Resumen...

Se realizaron pruebas con otros cinco productos, y se obtuvo que:

- ARIMA mejora los pronósticos obtenidos por métodos ingenuos
- Generalmente se obtienen mejores resultados con Redes Neuronales (RN) que con ARIMA
- ARIMA entrega un modelo comprensible y buenos resultados, pero con costos no despreciables (requerimientos estadísticos, y de conocimientos del usuario)
- RN obtienen los mejores resultados de forma más automática, pero con modelo tipo "black box"

Sistema de Pronóstico de Ventas

Unidades Vendidas

Precios

Promociones

Parámetros:

Tiempo de predicción

Producto, local

Chequeo de datos y corrección

Datos limpios y verificados

Modelo Predictivo

Pronóstico

Reentrenamiento con nuevos datos

Chequeo de Predicción

Sistema de Pronóstico de Ventas

Pronóstico para el período deseado

Sistema de Reposición de Inventarios

Sistema de Reposición

Proveedor

ORION

Información de Proveedores:

Tiempos de entrega

Confiabilidad

Orden de Compra via EDI

Ajustes del tomador de decisiones

Cuánto y/o Cuándo pedir

Sistema de Reposición de Inventarios

Sistema de Pronóstico de Ventas

Pronóstico para el período deseado

Sistema de Reposición Periódica

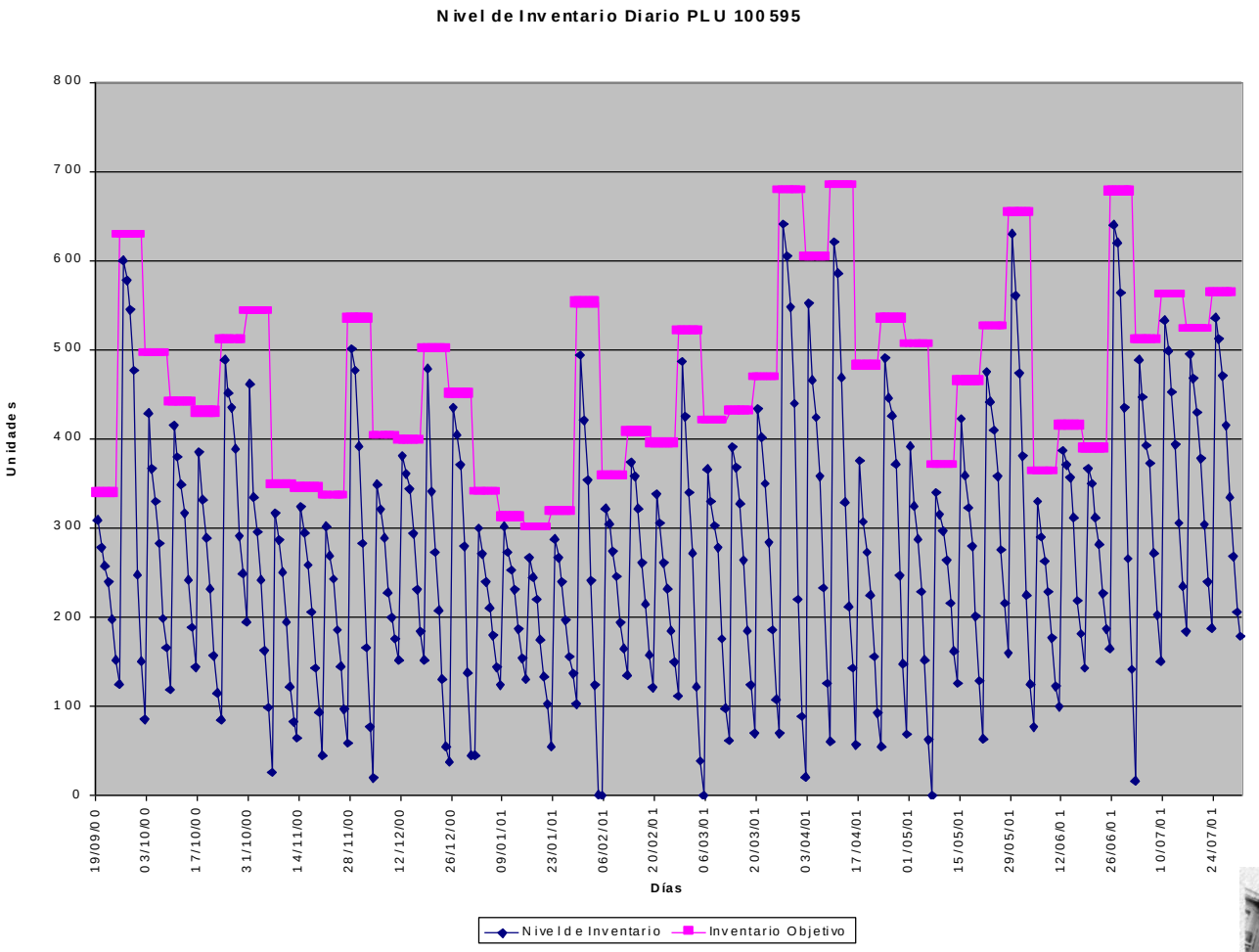
- Reposición cada P días, con tiempo de entrega de L días.

INVENTARIO OBJETIVO T

$$T = m' + z\sigma$$

- Con:
- m': demanda promedio durante P+L días (del sistema de pronóstico)
- Z σ: stock de seguridad (nivel de servicio*desviación ventas)

Reposición de Inventarios



Quiebres de venta: 1% con 5 días de alcance en inventario