



Knowledge Discovery in Databases

Metodología aplicada al estudio de problemas en minería de datos

Jaime Miranda

**Departamento de Ingeniería Industrial
Universidad de Chile**

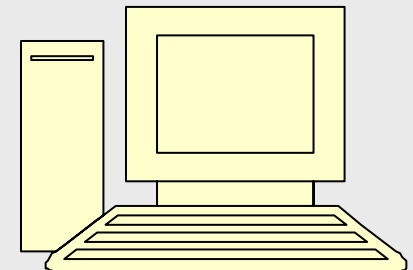
IN47B

Ingeniería de Operaciones

EMPECEMOS CON EN EJEMPLO...

PROBLEMA

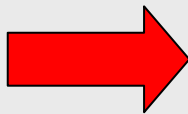
- “Una empresa multinacional de seguros cansada por el enorme y creciente número de clientes que no cancelan sus deudas ,desea poder detectar que clientes son los “mejores” a la hora de incorporarlos a su cartera de clientes con la finalidad de disminuir las pérdidas por estos conceptos”.
- La empresa posee un Datawarehouse con la información de los clientes y de sus transacciones.
- Tamaño de la cartera ~150.000.



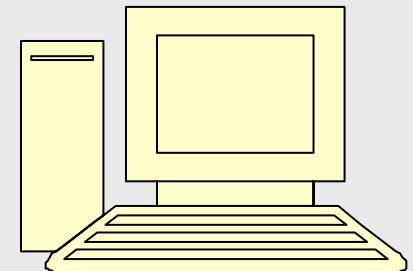
EMPECEMOS CON UN EJEMPLO...

PREGUNTAS

- ¿Qué información tenemos a nuestro alcance?
- ¿Cuál es la calidad de esta información?
- ¿En qué difiere un buen de un mal cliente?
- ¿Cómo es posible caracterizar a un cliente?
- ¿Cuántos clientes en promedio no pagan en un mes?



**Comprender el entorno y
los objetivos del problema**



DESCRIPCIÓN DEL PROBLEMA

EMPRESA: Multinacional de Seguros

DIVISIÓN: Créditos Hipotecarios

GIRO DEL NEGOCIO

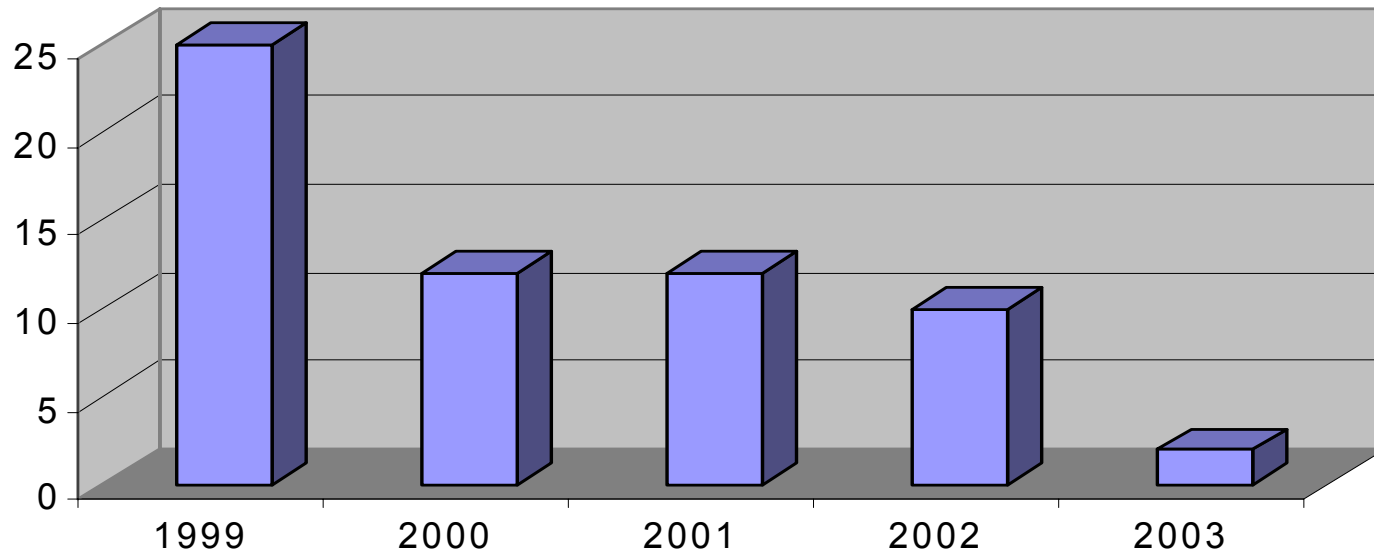
- Financiamiento hipotecario mediante mutuos endosables.
- Mutuo endosable
 - Crédito en \$ para la compra o ampliación de una propiedad urbana, nueva o usada.
 - Es endosable, ya que se puede ceder a algunas instituciones autorizadas por la ley.

CARACTERÍSTICAS

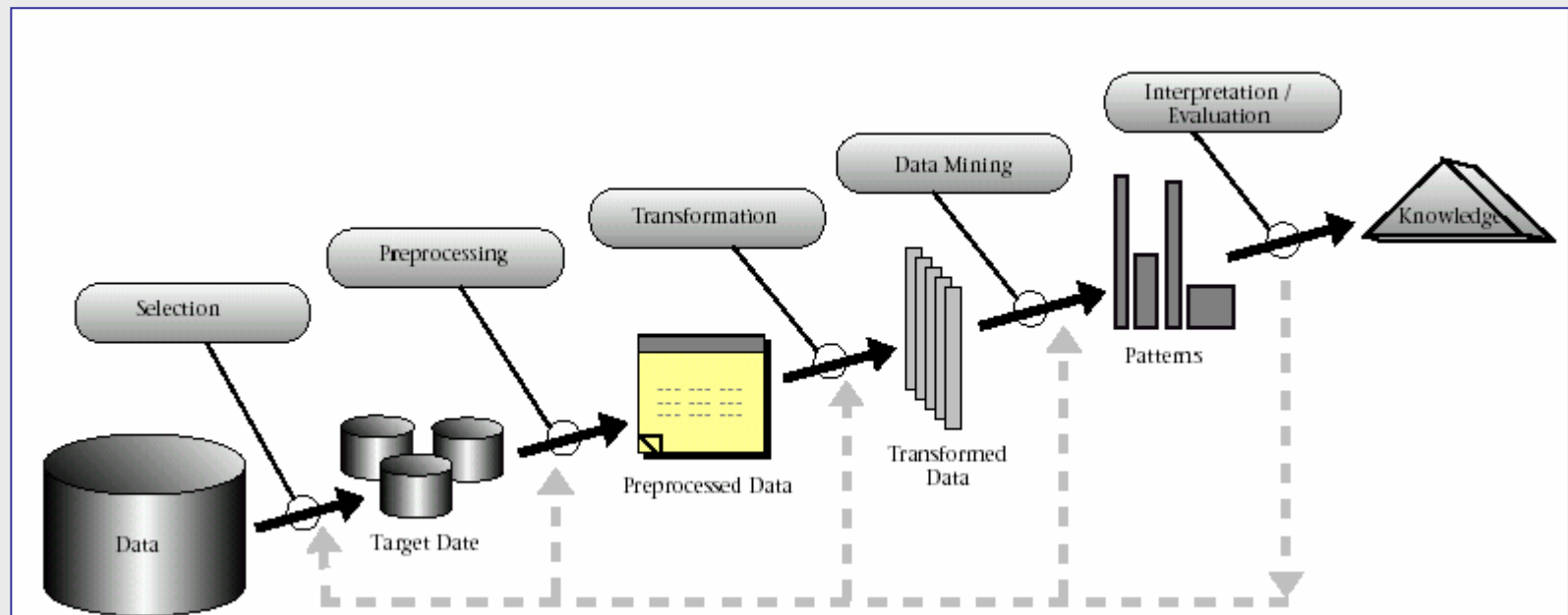
- Cada cierto tiempo algunas instituciones financieras (bancos) andan en búsqueda de liquidez (\$).
- Liquidan parte de sus carteras y transan las deudas hipotecarias de sus clientes.
- La empresa no posee ningún apoyo para la toma de decisión sobre la compra de alguna cartera en especial.
- La empresa no conoce el “comportamiento de pago” pasado de los clientes que potencialmente comprara.
- Al momento de compra todos los clientes parecen “buenos clientes”

SITUACIÓN ACTUAL

Meses de primer atraso



PROCESO KDD



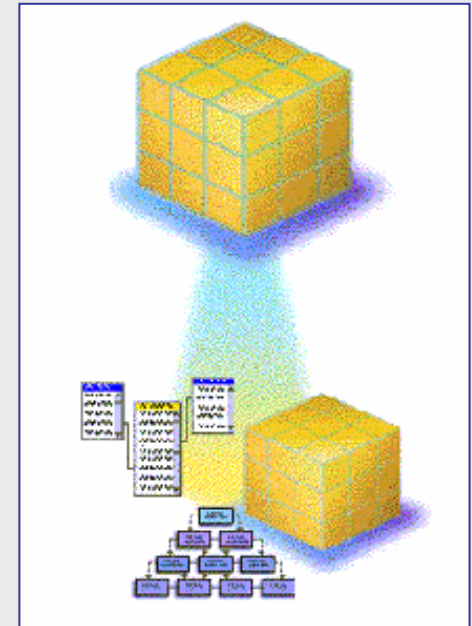
SELECCIÓN DE VARIABLES

FUENTES DE INFORMACIÓN

- Informes
- Data Warehouse – Data Mart.
- Expertos del negocio

VARIABLES IMPORTANTES

- Juicio experto.
- Heurísticas de selección.
- Problema NPC.



CARACTERISTICAS

- Reúne datos esenciales provenientes de bases de datos heterogéneas desde todas las áreas de negocio (Ventas, finanzas, RRHH, etc.)
- Organiza los datos para apoyar decisiones de gestión.
- Maneja elevados volúmenes de información.
- Permite el mejor funcionamiento de los métodos de Data Mining.



DATAWAREHOUSE: Colección de objetos

→ Orientada al sujeto:

- Organizada en torno a los datos más importantes de la empresa.
- Es bueno para realizar filtros y eliminar información poco importante.

→ Unificada:

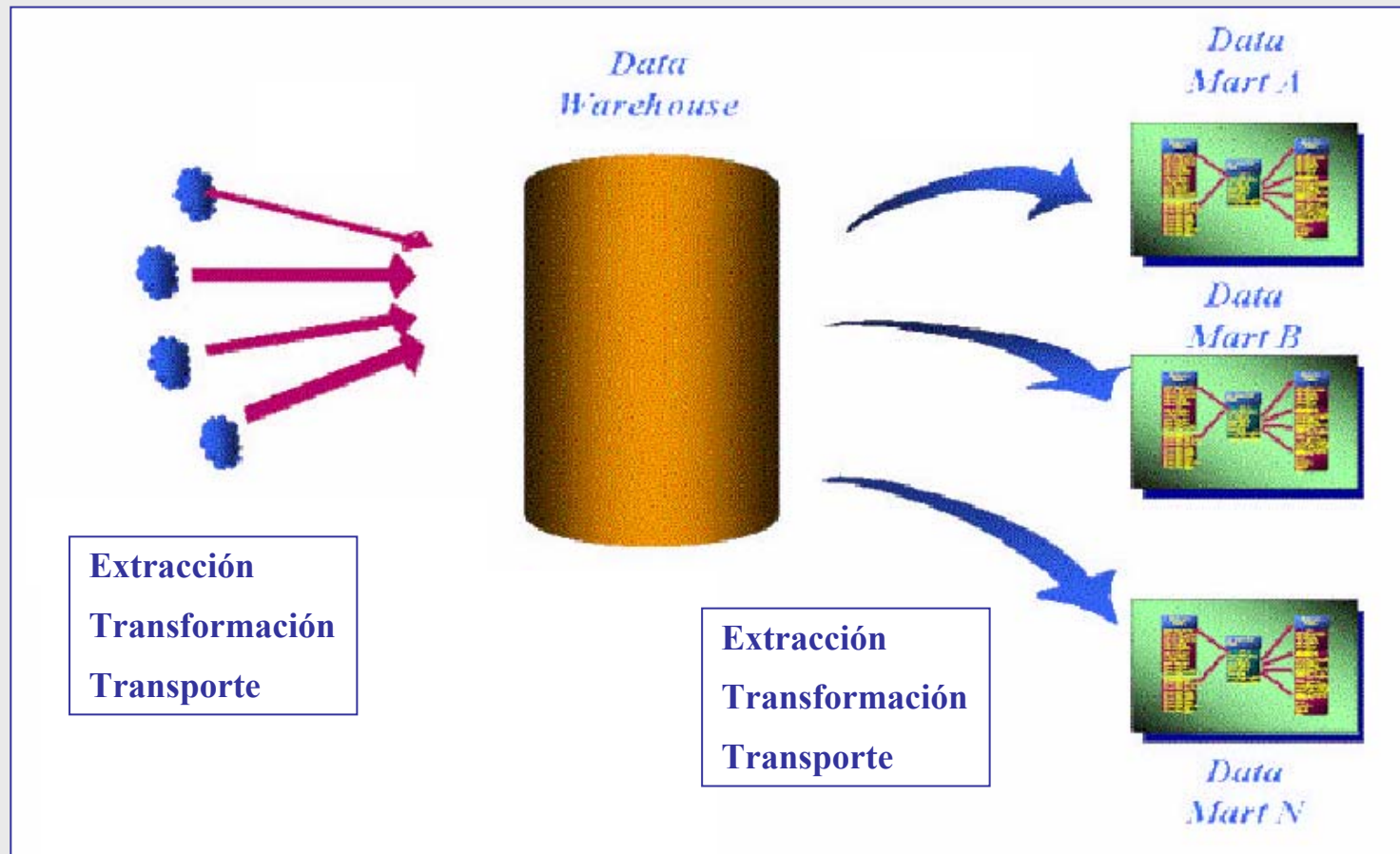
- Basada en unión de información de varias fuentes.
- Asegura la consistencia de la información.

→ Variante en el tiempo

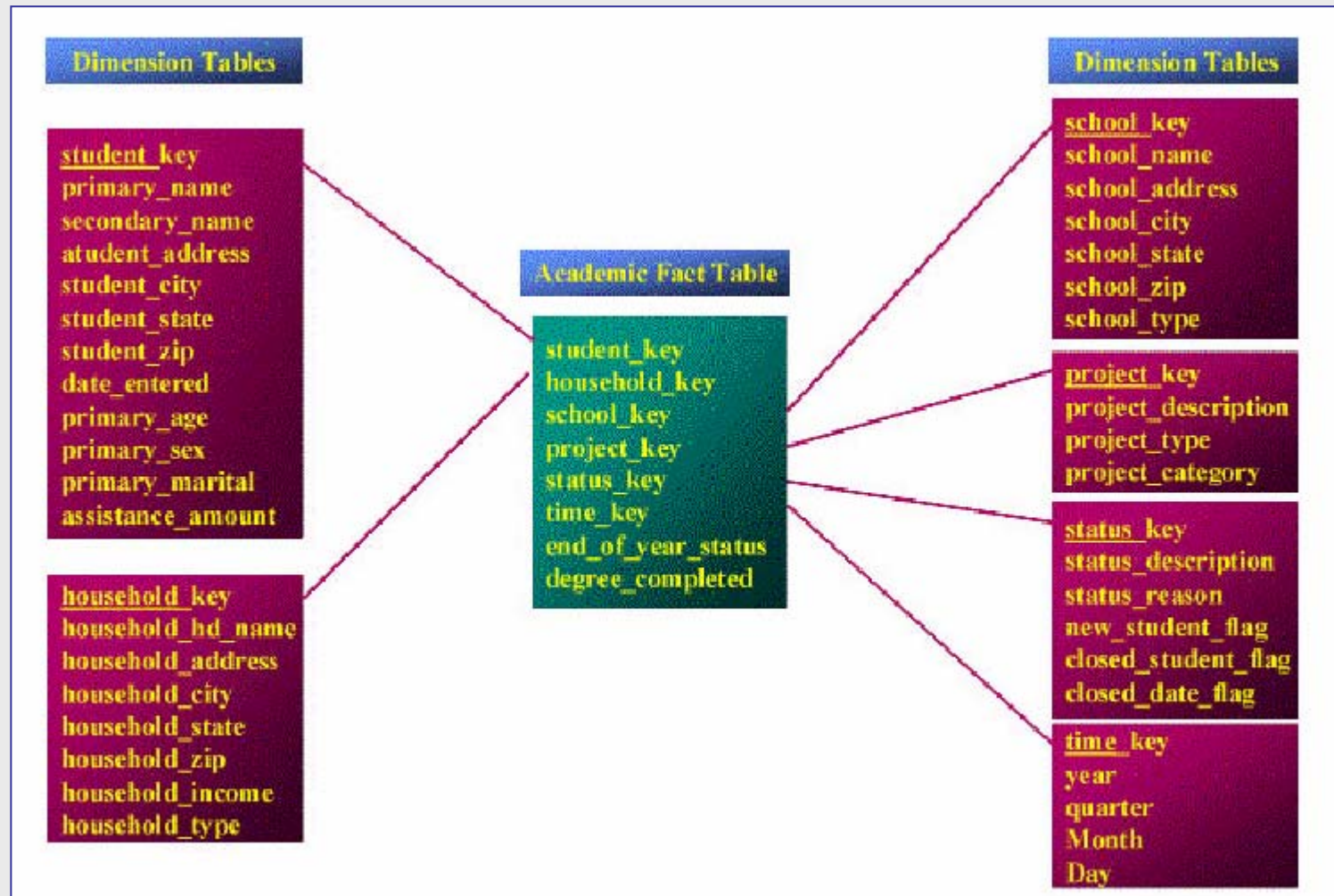
- Guarda información a través del tiempo.
- Posee actualizaciones temporales agregadas: no hay actualizaciones diarias.



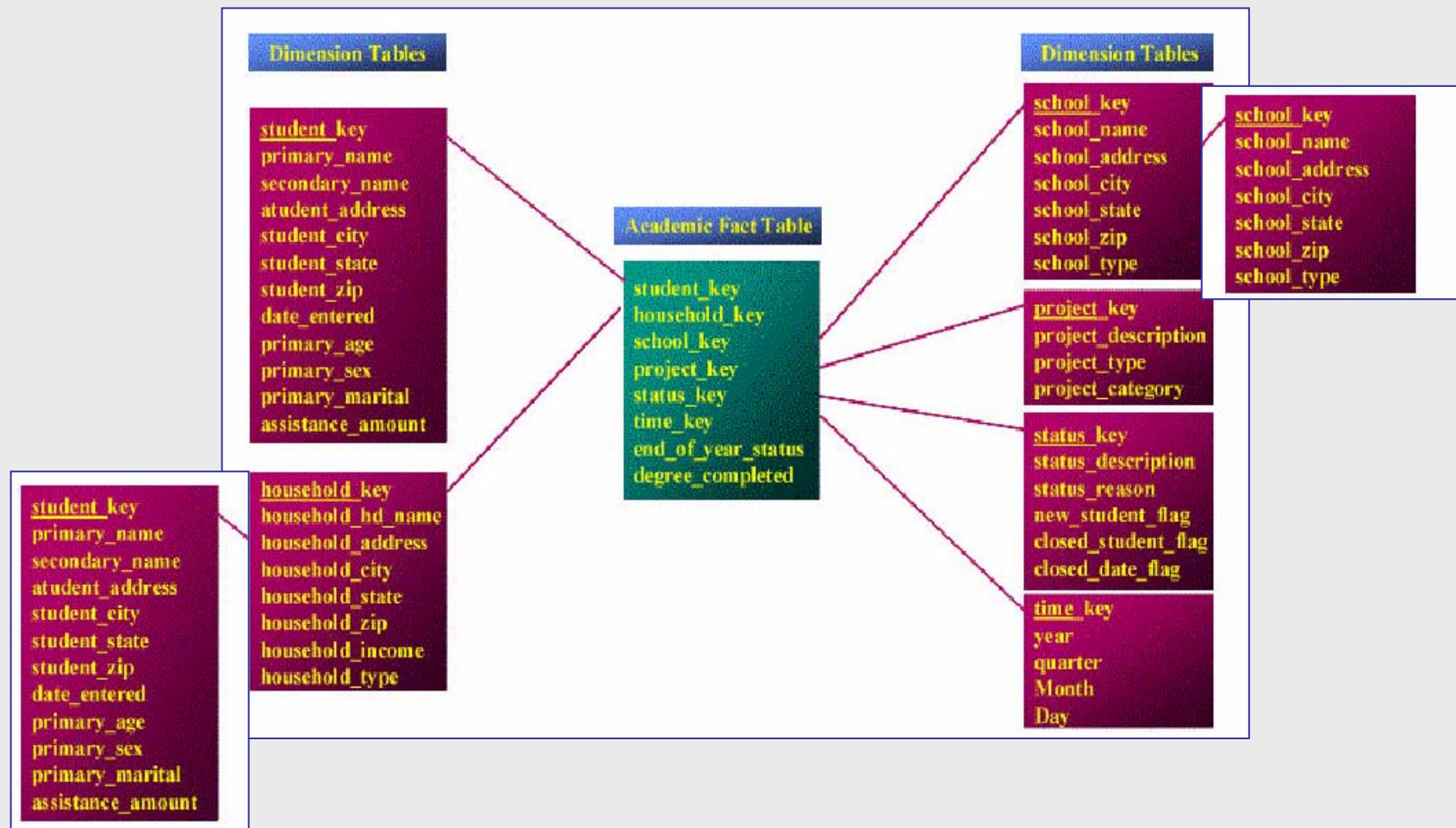
DIAGRAMA GENERAL



ESQUEMA “ESTRELLA”

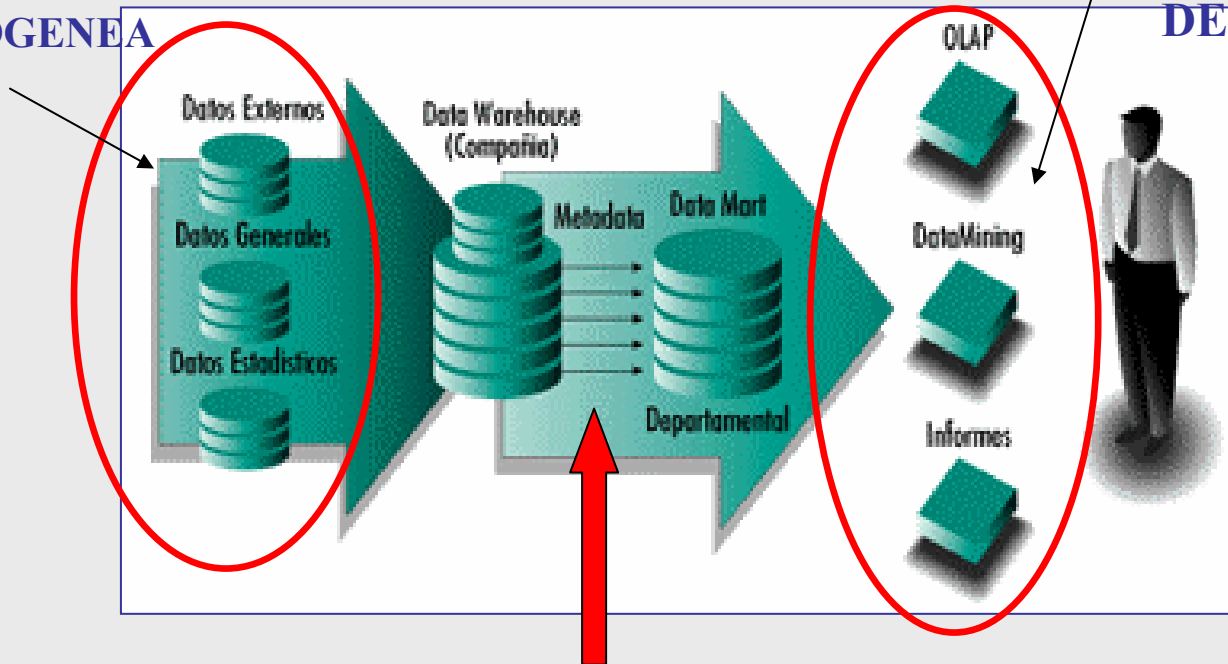


ESQUEMA “COPO DE NIEVE”



ARQUITECTURA MULTICAPAS

INFORMACION
HETEROGENEA



HERRAMIENTAS
DE ANALISIS

METADATOS

GENERACION DE CRITERIOS

Valores perdidos

→ Objetos sin información en una o más variables.

Valores fuera de rango

→ Valores “raros” en una o más variables que definen al objeto.

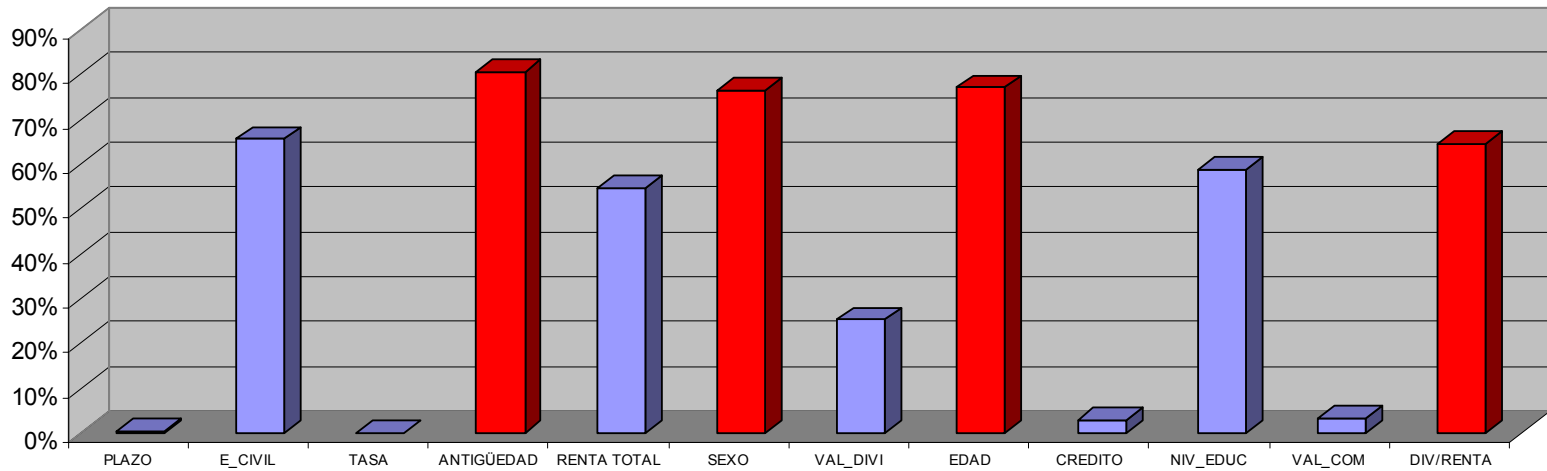
BASE DE TRABAJO MUTUOS

- Posee una cartera de 7.664 clientes.
- Existe un enorme cantidad de valores perdidos la base.
- Más del 80% de la base (6.488), poseía algún valor perdido.



PREPROCESAMIENTO (2)

% de valores perdidos respecto al total



SOLUCIONES BÁSICAS

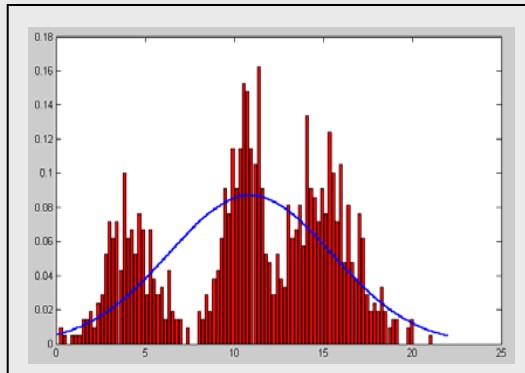
- Eliminación de objetos con algún problema.
- Llenado con valores promedio o modas.

SOLUCIONES AVANZADAS

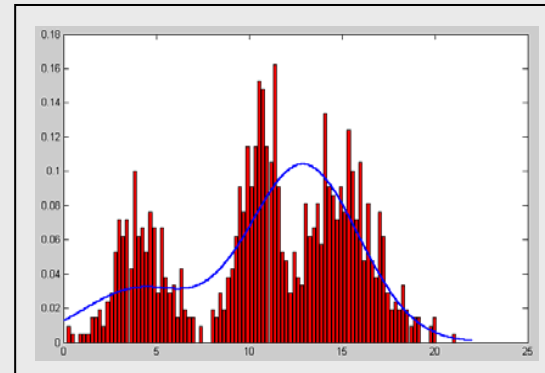
- Modelos predictivos.
- Heurísticas especializadas.



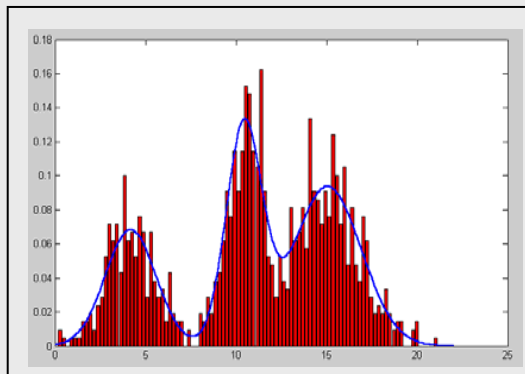
EJEMPLO DE MODELOS MIXTOS: RENTA



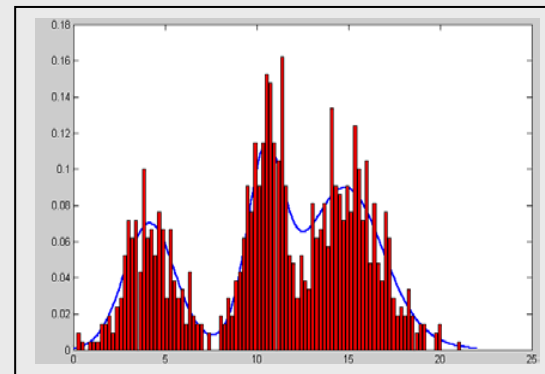
1 componente



2 componentes



3 componentes



4 componentes

MÉTODO: Clamping de atributos

- Heurística especializada
- Generar un ranking de importancia de atributos

$$\textit{Prominencia (salience)}, S(x_i) = 1 - \frac{g_{(x|x_i = \textit{mean_value})}}{g(x)}$$

NORMALIZACIONES

- Escalamiento
 - $[0,1]$ - Rango
- Estandarizaciones.
 - $N(0,1)$

CATEGORIZACIONES

- Variables “string” a números: Nivel educacional
- Generación de rangos en las variables

NUEVAS VARIABLES

- Relaciones nuevas basadas en variables originales
- Variaciones %.

APRENDIZAJE

- “El aprendizaje es una habilidad de la que disponen gran parte de los sistemas naturales para adaptarse al entorno en el que vive”.
- “Adquisición de conocimiento de un proceso por medio del análisis, ejercicio o experiencia”.
- “Un proceso por el cual los parámetros libres del sistema se adaptan a través de un proceso continuo de estimulación a partir del entorno en el que el sistema está inmerso”.



GENERALIDAD

- Modelo representativo, el cual puede predecir el comportamiento a nuevos objetos.

COMPENSIBILIDAD

- Fácil de inspeccionar y entendible.
- Con fáciles mejoras de capacidad de generalizaron



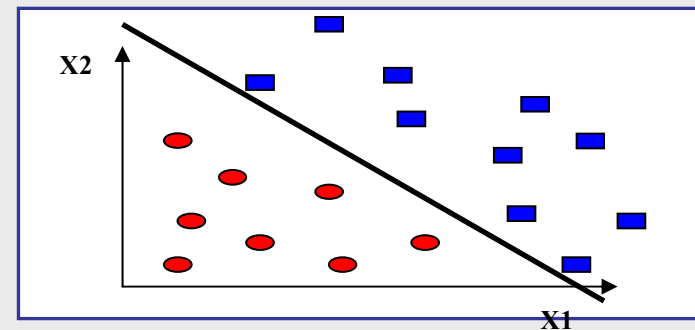
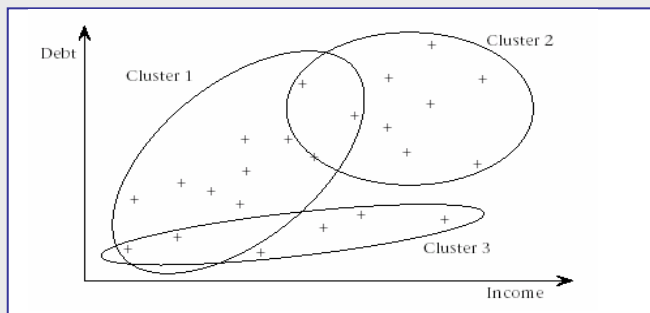
CONCEPTOS BÁSICOS

→ Medida de distancia

→ Prototipo o centro de clase más cercana.

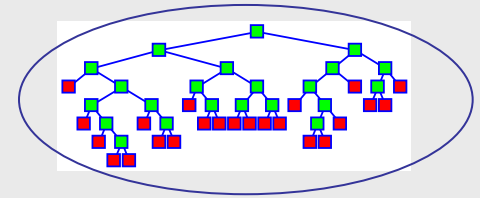
→ Hipersuperficies

→ Clasificación de acuerdo a si los objetos están a uno u otro lado de una hipersuperficie o conjunto de hiperplanos.



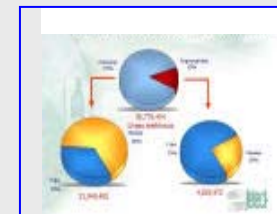
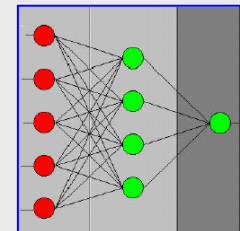
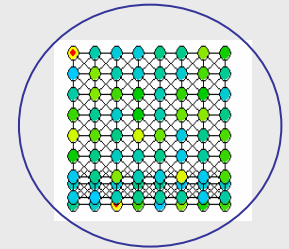
MÉTODOS SUPERVISADOS

- Redes neuronales
- Árboles de decisión.



MÉTODOS NO SUPERVISADOS

- Cluster: Fuzzy C-means
- Mapas de Kohonen (SOM)



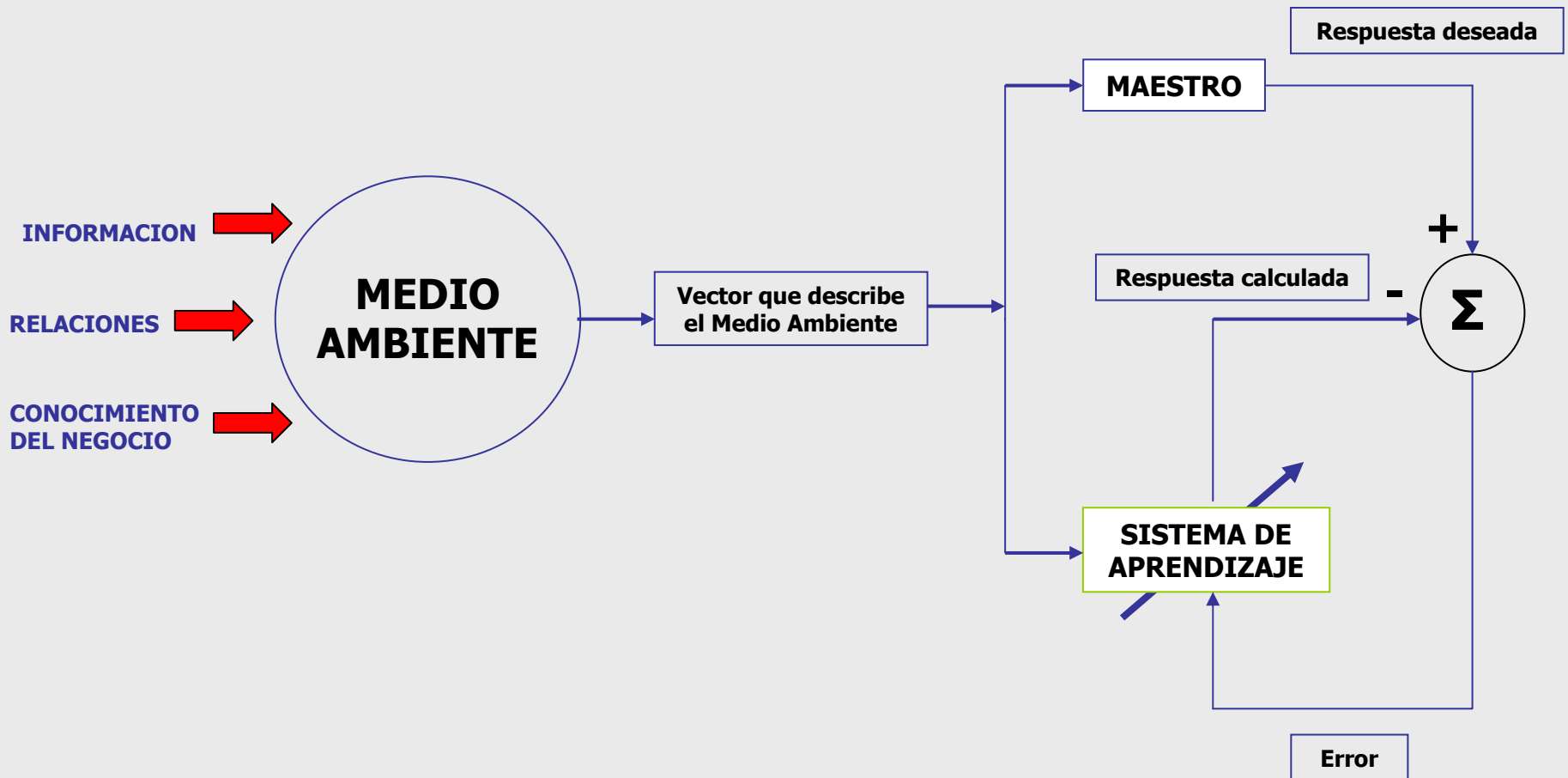
Algunas nociones

- Necesidad de una entrada (inputs) y una salida (outputs).
- Determinar el output a través de “combinaciones” de los inputs.
- Encontrar una función que describa el proceso.
- Uso de la experiencia para describir algún patrón característico.

Conjuntos

- Entrenamiento.
- Test

DIAGRAMA DE APRENDIZAJE SUPERVISADO



Algunas nociones

- Necesidad solo de una entrada (inputs)
- No necesita una salida (outputs) explícita.
- Encontrar una función que describa el proceso.
- Uso de la experiencia para describir algún patrón característico.

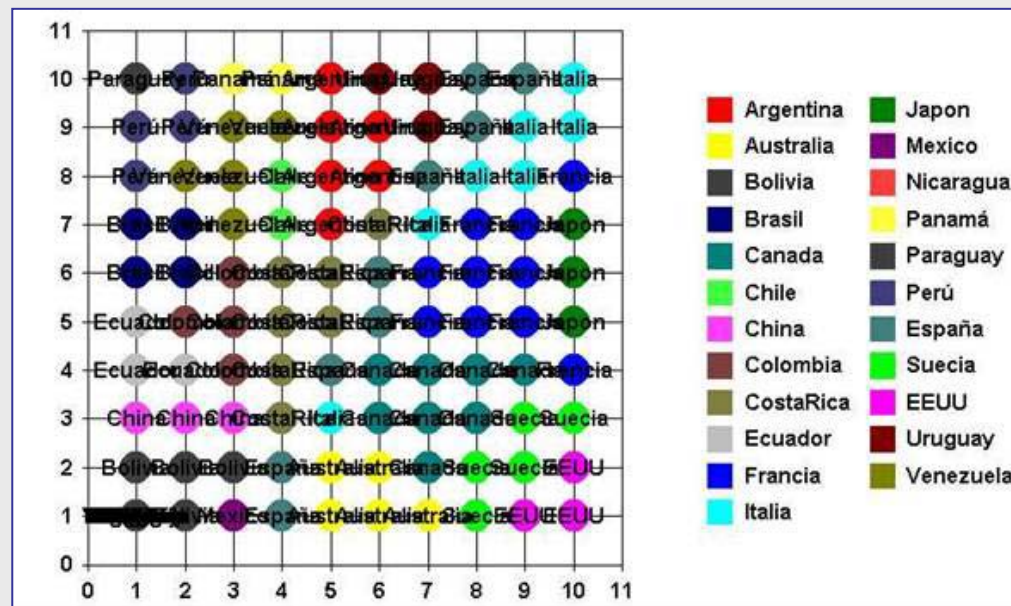
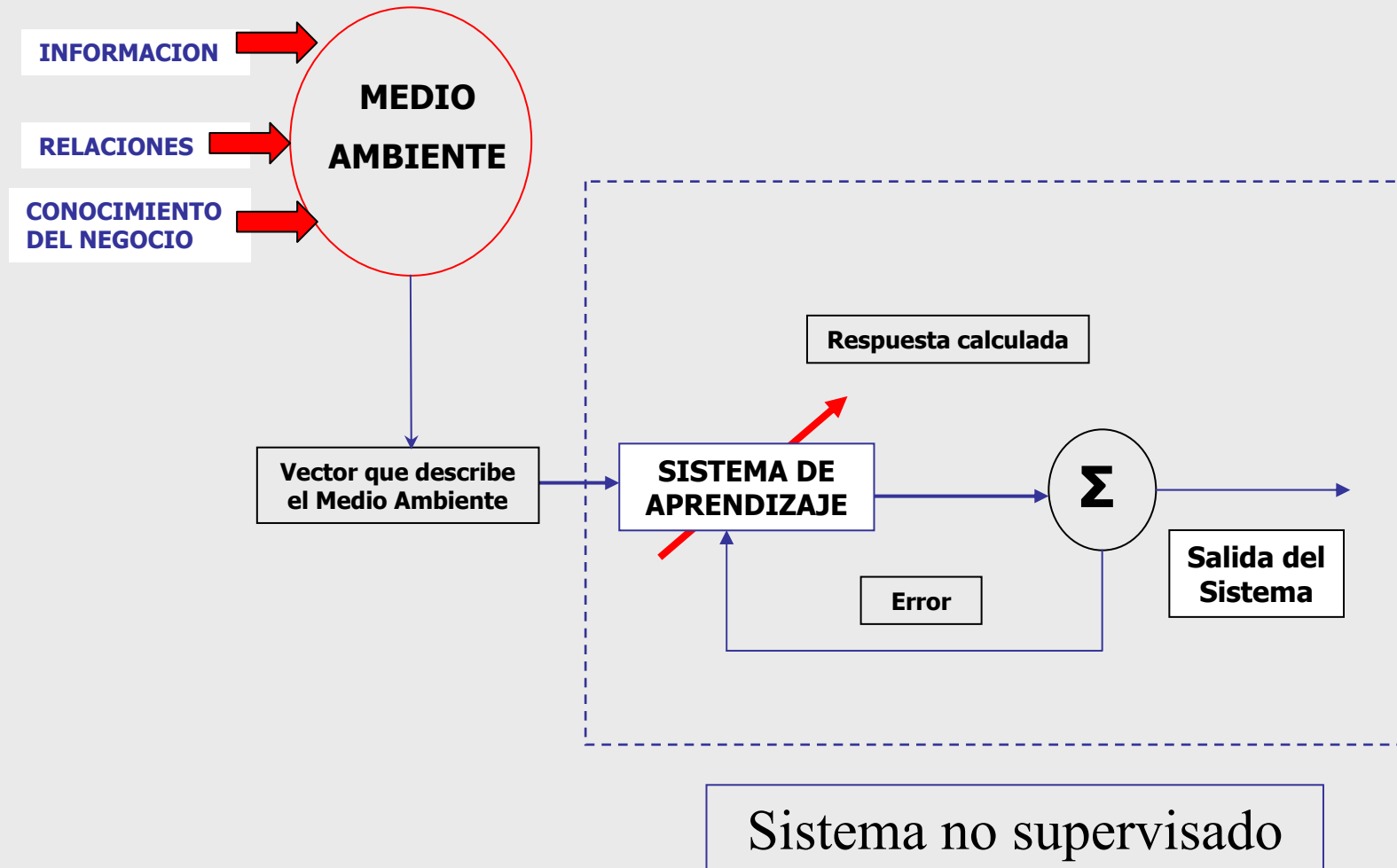


DIAGRAMA DE APRENDIZAJE SUPERVISADO



DESCRIPCIÓN

- Los bancos en busca de liquidez venden parte de su cartera de negocio.
- Se busca construir un modelo que genere un scoring sobre el riesgo crediticio de un cliente que es vendido por una institución financiera.
- Se conoce el comportamiento de pago solo de los clientes que compone la cartera de la empresa.
- Se busca identificar que clientes (“buenos”) comprar a la institución financiera con el fin de minimizar las perdidas por mora.



CASO N°2: FUGA DE CUENTACORRENTISTAS

DESCRIPCIÓN

- Existe un creciente aumento en el número de cierres voluntarios de cuentacorrientes de una institución financiera.
- Se busca un modelo de predicción de fugas, para aumentar eficiencia y eficacia de las políticas comerciales y de retención.
- Se desea disminuir el número de clientes fugados.



CASO N°3: OFERTAS FOCALIZADAS

DESCRIPCIÓN

- Cansados de fallidos y bajos radios de retorno del envío de descuentos y ofertas por correo, una importante empresa de retail desea construir un modelo de predicción de compra (ofertas focalizadas).
- Se desea mandar un descuento para un televisor
- La empresa posee historiales de compras de sus clientes que han comprado alguna vez un televisor de esas características.
- Se desea minimizar los costos de envío del descuento.





Knowledge Discovery in Databases

Metodología aplicada al estudio de problemas en minería de datos

Jaime Miranda

**Departamento de Ingeniería Industrial
Universidad de Chile**

IN47B

Ingeniería de Operaciones