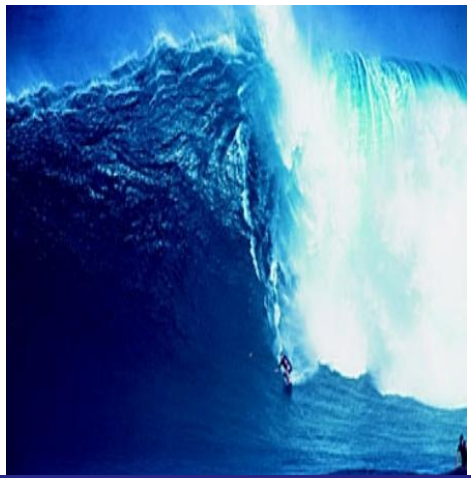
Minería de datos

Dra. Amparo López Gaona

Fac. Ciencias, UNAM
Abril 2018

Introducción

- El problema:
  - Los sistemas de información producen gran cantidad y variedad de datos almacenados en BD digitales. Su crecimiento es exponencial.
  - Esta información puede ser extremadamente valiosa para la organización que la genera, podría por ejemplo:
    - Optimizar procesos.
    - Maximizar la satisfacción del cliente.
    - Crear campañas dirigidas a las preferencias de los clientes.
    - Detectar transacciones fraudulentas o sospechosas.
    - etc.
  - Pero, ... simplemente es demasiada información que debe ser analizada.



... Introducción

- La solución:

- Ejemplo:

... Introducción

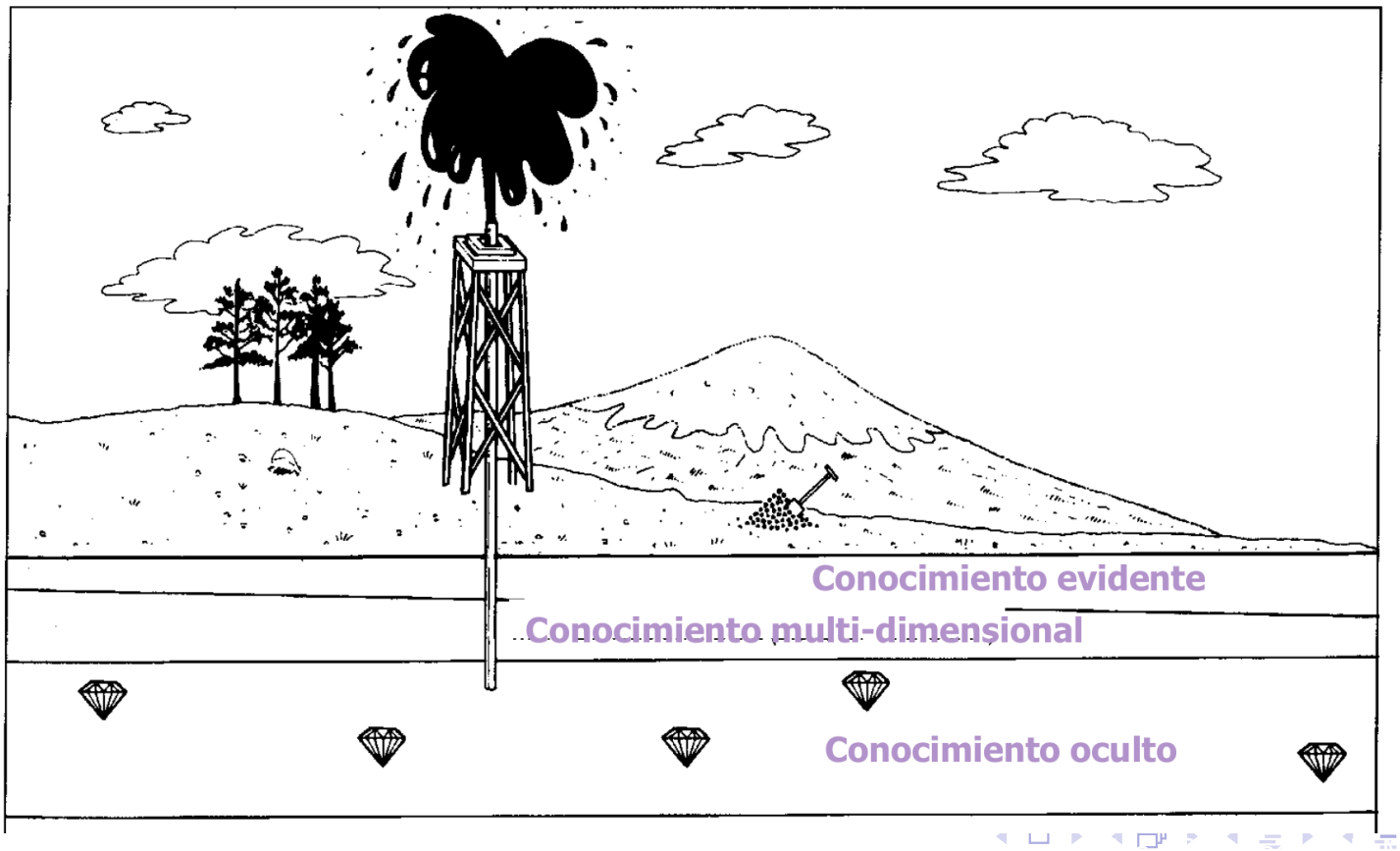
- Minería de datos es la extracción o "minado" de conocimiento de grandes volúmenes de datos.
- DM es un proceso que intenta descubrir patrones interesantes en grandes volúmenes de datos.

... ¿Qué es la minería de datos?

¿Ejemplos?:

Tipos de conocimiento

- Para decidir la técnica más apropiada para cada situación se requiere distinguir el tipo de conocimiento que se quiere extraer.
- De acuerdo al nivel de abstracción, el conocimiento obtenido de los datos puede clasificarse.



... Tipos de conocimiento

- Conocimiento evidente:
  - La información se recupera con facilidad usando consultas a la BD.
  - Ejemplos:
  - Técnica: OLTP.
- Conocimiento mulitidimensional.
  - Los datos se consideran con estructura mulitidimensional, en lugar de considerar cada transacción individual se organizan de acuerdo a dimensiones como tiempo, zona geográfica, etc.
  - Es un reacomodo de una BD relacional con lo cual se pueden detectar algunas regularidades difícilmente observadas en una tabla normal.
  - Ejemplos:
  - Técnica: OLAP

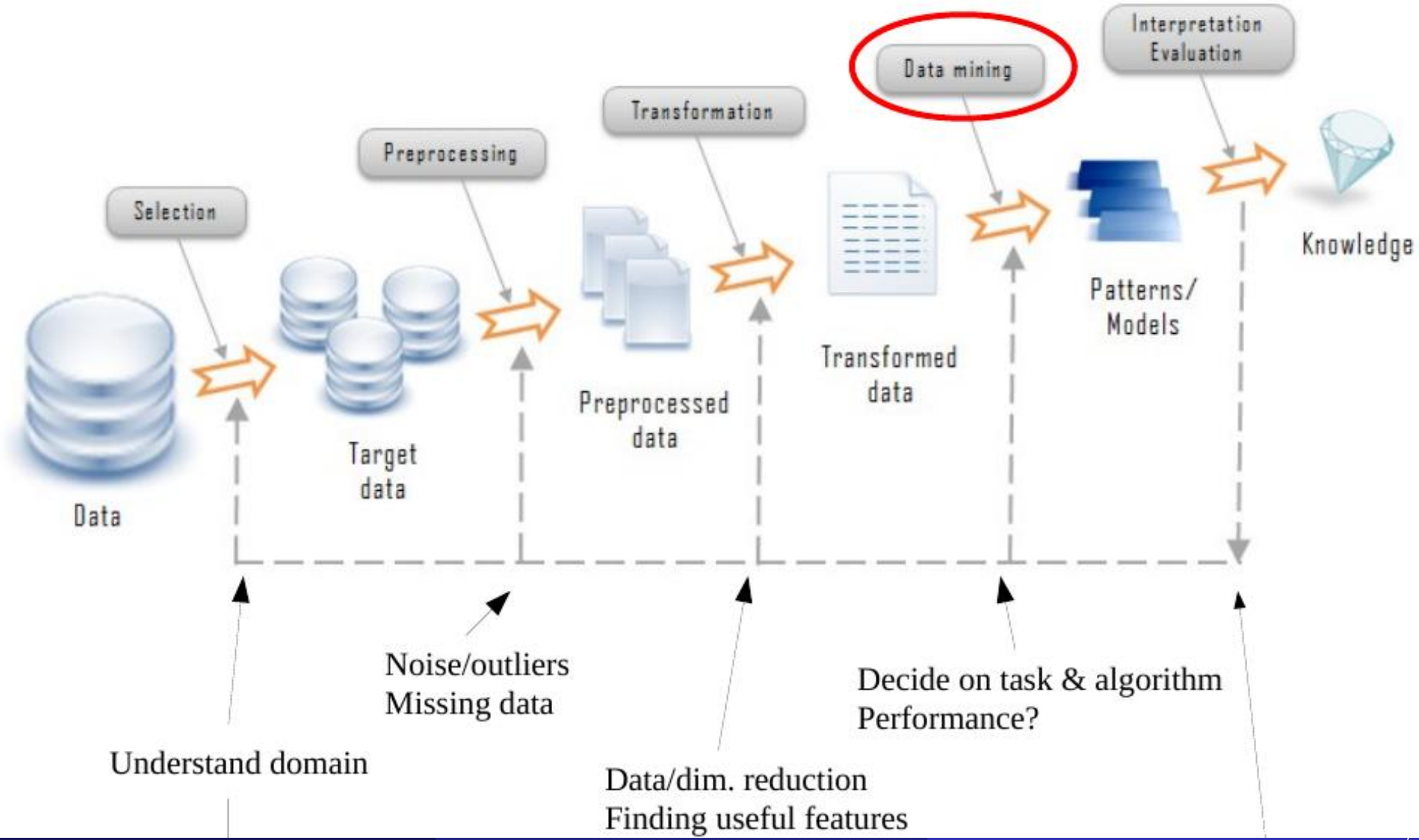
... Tipos de conocimiento

- Conocimiento oculto.
  - Información no evidente, desconocida a priori y potencialmente útil.
  - Información de gran valor, conocimiento novedoso que permite una nueva visión del problema.
  - Ejemplos: ¿Qué tipos de clientes se tienen? ¿Perfil de cada clase de clientes?
  - Técnica: minería de datos.

¿Qué es la minería de datos?

- La minería de datos también es llamada KDD (*knowledge discovery in database*).
  - "Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams." Han
- DM es la extracción de patrones interesantes de diversas fuentes de datos.
  - Fuentes de datos: bases de datos, textos, almacenes de datos, web, imágenes, etc.
  - Patrones: potencialmente útiles, interesantes, novedosos, inesperados, entendibles.

P. de descubrimiento de conocimiento en BDs



Procesamiento en BD vs. DM

- Bases de datos
  - Consultas:
    - Bien definidas.
    - SQL.
  - Datos:
    - Datos operacionales.
  - Salida
    - Precisa.
    - Subconjunto de la BD.
- Minería de datos
  - Consultas:
    - Pobremente definidas.
    - No precisa un lenguaje de consulta.
  - Datos:
    - No necesariamente datos operacionales.
  - Salida:
    - Difusa.
    - No es un subconjunto de una BD.

Ejemplos de consultas

- EN bases de datos:
  - Encontrar todas las solicitudes de crédito de clientes con apellido López.
  - Identificar los clientes que han comprado más de $10,000 el último mes.
  - Encontrar los clientes que han comprado leche.
- En minería de datos:
  - Encontrar todas las solicitudes de crédito de clientes que tienen poco riesgo de ser morosos. (clasificación)
  - Identificar los clientes que tienen hábitos de compra similares. (agrupación)
  - Encontrar los artículos que frecuentemente se compran junto con leche. (reglas de asociación).

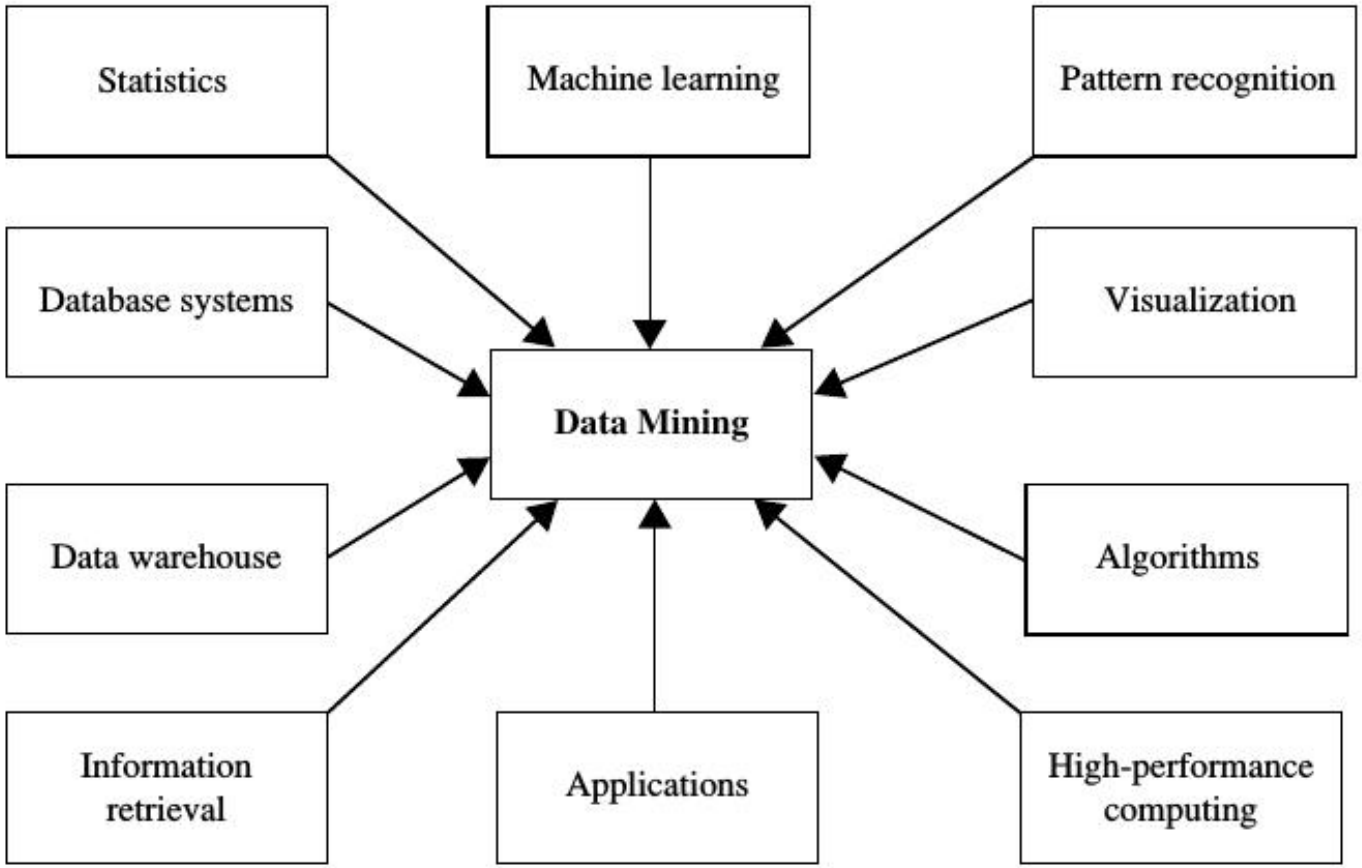
¿Porqués

- ¿Por qué es importante la DM?
  - El uso masivo de las computadoras en las organizaciones produce grandes cantidades de datos.
  - ¿Cómo se puede hacer uso de ellos?
  - El conocimiento descubierto de ellos puede ayudar a tomar ventaja competitiva.
- ¿Por qué es necesaria la DM?
  - Permite utilizar datos valiosos.
  - Hay una brecha entre los datos almacenados y el conocimiento; y la transición no ocurre automáticamente.
  - Se puede obtener información interesante que con sólo consultas a la BD. Ejems.
    - Encontrar las personas que probablemente compren mis productos.
    - Quiénes problablemente responderán a cierta promoción.

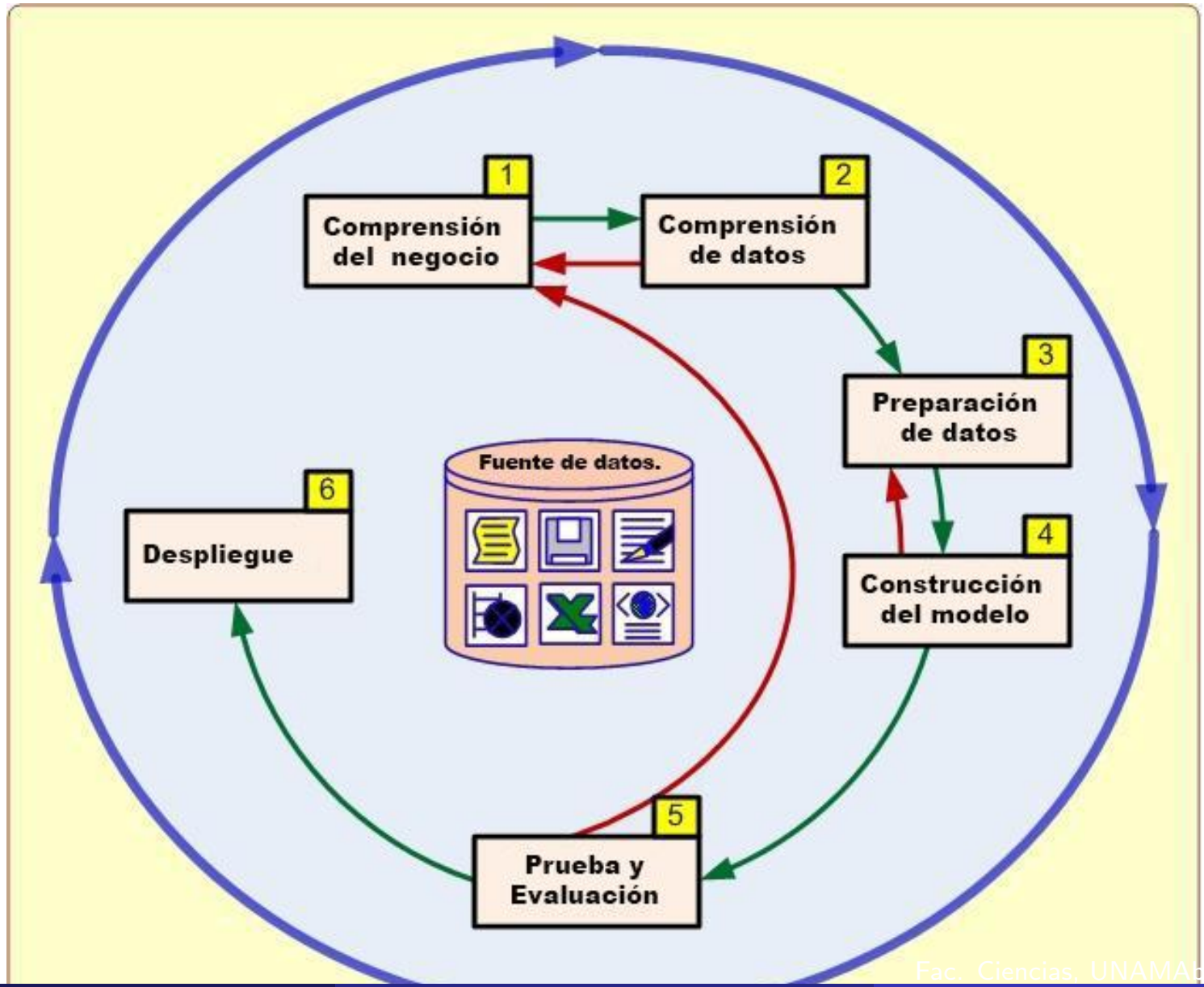
¿Porqués

- ¿Por qué ahora?
  - Hay abundancia de datos.
  - Las computadoras son asequibles.
  - La presión competitiva es fuerte.
  - Las herramientas para DM son accesibles.

La minería de datos es multi disciplinaria

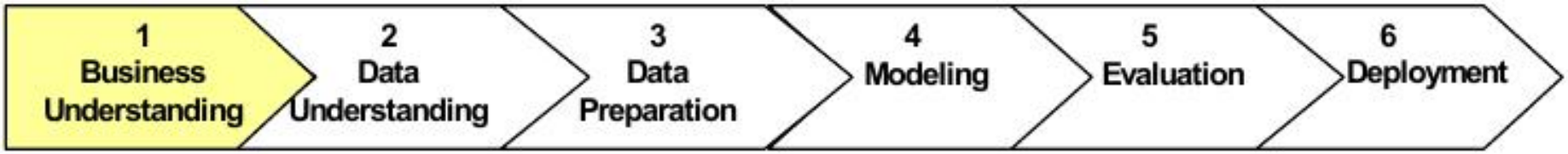


CRISP-DM (*Cross Industry Standard Process for DM*)



Proceso de minería de datos

Fase 1: Conocer la organización. Definición del problema.



Conocer tanto objetivos como requerimientos del proyecto, desde la perspectiva de la organización, y traducirlos a la definición del problema de minería de datos.

... Proceso de MD (Definición del problema)

- Es la etapa más importante, y de mayor reto. Sin comprender cuál es el problema que se quiere resolver y cómo se usarán los resultados, las expectativas pueden ser vagas y poco realistas.

```
Alicia: Would you tell me, please, which way I ought to go from
here?
Gato: That depends a good deal on where you want to get to.
Alicia: I don't much care where.
Gato: Then it doesn't matter which way you go.
Alicia: ... So long as I get SOMEWHERE,"
Gato: Oh, you're sure to do that, if you only walk long enough.
```

- Personal de la organización y de computación deben trabajar juntos para definir el problema.

... Proceso de MD (Definición del problema)

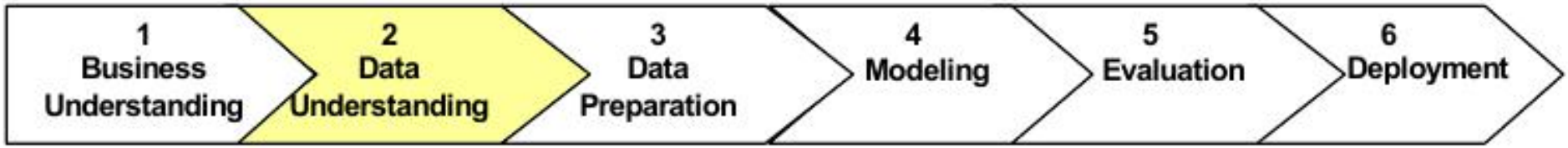
- Determinar los objetivos de la organización, desde la perspectiva del cliente.
  - "Reducir el costo de fraude en reclamaciones de seguro".
- Evaluar la situación.
  - Recursos, restricciones, suposiciones y otros factores a considerar al determinar el propósito del análisis de datos.
- Determinar la meta, medible y cuantificable, en términos de la MD:
  - "Determinar los factores que aparecen juntos en una forma de reclamación, combinados con la demografía del solicitante para identificar reclamaciones fraudulentas; predecirlas y ordenarlas por valor monetario".
  - Meta medible: "Ser mejores al momento de determinar si una reclamación es fraudulenta o no" ¿? NO.
  - Reducir la cantidad de fraudes al menos en un 10 %, usando un modelo que prediga solicitudes con alta probabilidad de ser fraudulentas.

Una meta concreta lleva a establecer condiciones de terminación y criterios de aceptación concretos.

- Producir un plan de trabajo.

... Proceso de minería de datos (Datos)

Fase 2: Conocer los datos.



Esta etapa empieza con la recolección inicial de los datos. Se procede con actividades para:

- Familiarizarse con los datos. (Descripción de ellos)
- Identificar problemas de calidad de datos.
- Descubrir primeras impresiones/ideas de los datos o bien,
- Detectar subconjuntos interesantes para formar hipótesis de información oculta.
- Generalmente es la estapa que consume más tiempo.

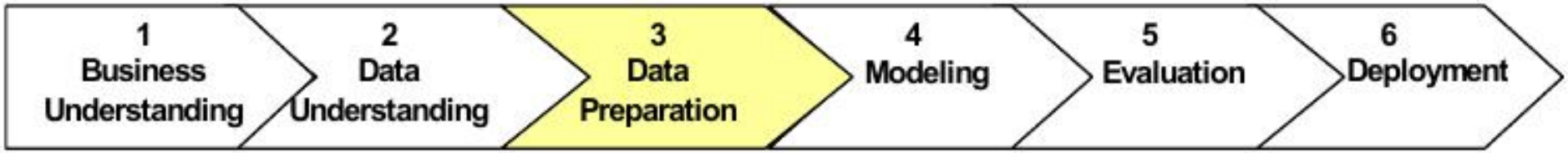
... Proceso de minería de datos (Datos)

- Colección de datos.
  - Adquirir los datos especificados en los recursos del proyecto.
  - Frecuentemente, es necesario integrar varias bases de datos.

... Proceso de minería de datos (Datos)

- Describir los datos:
  - Examinar las propiedades de los datos adquiridos y reportarlas.
- Explorar los datos.
  - tipo de datos, distribución de datos, relación entre ellos, calidad, etc.
- Verificar la calidad de los datos.
  - ¿Los datos están completos?, ¿son correctos? ¿Hay valores perdidos? ¿Qué tan común es esto?

Fase 3: Preparación de los datos.

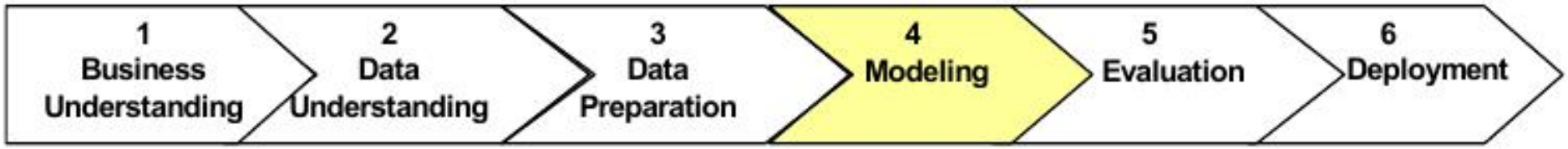


- Abarca las actividades necesarias para construir el dataset final a partir de los datos iniciales.
- Estas tareas incluyen selección de datos, registros, tablas, así como transformación y limpieza de datos.
- Estas tareas pueden ser realizadas en varias ocasiones y sin un orden determinado.

... Proceso de MD (Preparación)

- Selección: Antes de empezar con el proceso de MD, se debe seleccionar un subconjunto de los datos para trabajar con él.
  - Criterios de selección: importancia para el objetivo de MD, calidad, restricciones técnicas tales como límites en los volúmenes.
- Limpieza:
  - Algunos datos pueden no ser relevantes para la metas del sistema.
  - En ocasiones puede haber demasiados valores perdidos o algunos datos poco o nada confiables.
  - Puede desearse eliminar estos datos para que no "contaminen" el conocimiento que se va descubrir.
- Transformación/Construcción de datos:
  - Debido a que tenemos datos integrados de diferentes fuentes, es frecuente tener que cambiar los formatos para homogeneizar los datos.
  - En ocasiones se requiere aplicar varias transformaciones a los datos sobre los cuales se desea aplicar las técnicas de MD.
  - Construcción de datos atributos derivados.

Fase 4: Generación del modelo de datos.



En esta fase, se seleccionan y aplican varias técnicas de modelado y se calibran sus parámetros para obtener valores óptimos.

- Es importante distinguir entre tareas y técnicas.
- Tareas de MD: Son "qué deseamos lograr". Hay dos tipos principales:
  - Predictivas: Realizan inferencia sobre los datos con la intensión de predecir el valor de un atributo particular (variable fuente, objetivo o dependiente) tomando como base el valor de otros atributos (variables independientes).
  - Descriptivas: tratan de describir las propiedades generales de los datos en la BD. El objetivo es determinar patrones que resuman las relaciones subyacentes en los datos.

... Proceso de minería de datos

- Técnicas: son las herramientas disponibles para lograr las tareas.
  - Reglas de asociación, redes Bayesianas, árboles de decisión, clustering, redes neuronales, algoritmos genéticos, etc.
- Con frecuencia, una tarea puede atacarse con diferentes técnicas.

... Proceso de minería de datos (Modelo)

Tareas predictivas. Tratan de predecir uno o más valores para cada instancia:

- Clasificación: Determinar a qué clase pertenece una nueva instancia en la BD. Las posibles clases se conocen de antemano. La meta es determinar/definir una función que asocie una clase a cada instancia.
- Estimar la probabilidad de clasificación: se desea aprender una función que indique la probabilidad de que una instancia dada pertenezca a cada clase.
- Análisis de patrones secuenciales. Detección de secuencias de datos en el tiempo.
- Regresión. Se trata de encontrar una función que asigne a cada instancia un valor numérico.

... Proceso de minería de datos (Modelo)

Tareas descriptivas: En lugar de hacer una predicción, tratan de describir los datos.

- Agrupamiento o clustering: busca crear grupos con los datos. No se sabe cuántos hay, se requiere descubrirlos.
- Reglas de asociación: Intentan identificar relaciones entre atributos categóricos.
- Correlaciones: intentan identificar dependencias (co-relaciones) entre diferentes atributos.
- Detección de valores anómalos o instancias anómalas. Tratan de encontrar valores de atributos, o instancias, los cuales son anómalos.

... Proceso de minería de datos (Modelo)

Los pasos en esta etapa son:

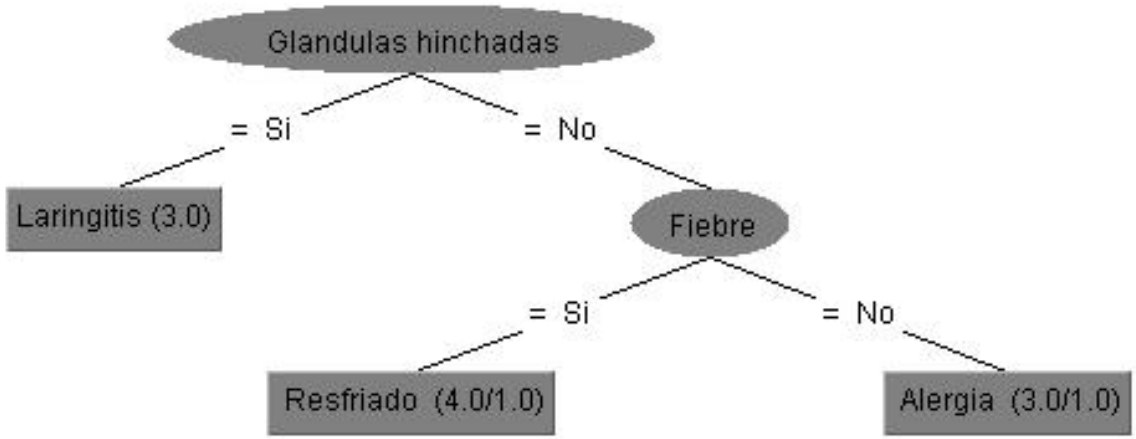
- Seleccionar la técnica de modelado.
- Generar un diseño de prueba.
- Construir un modelo.
- Probar el modelo.

... Proceso de minería de datos (Modelo)

Otra clasificación de las tareas de DM: supervisadas y no-supervisadas.

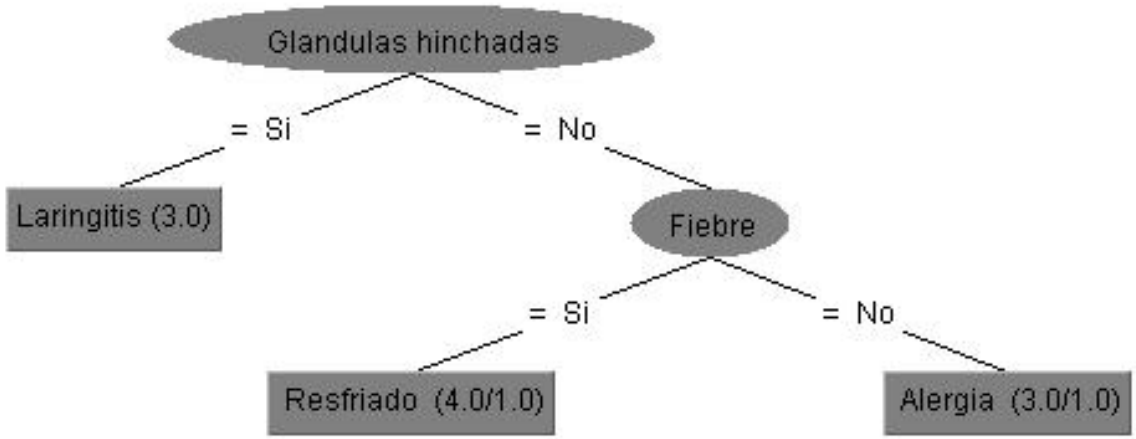
- Propósitos del aprendizaje supervisado:
  - Construir un modelo de clasificación a partir de datos clasificados previamente.
  - Usar el modelo construido para clasificar nuevas instancias de los datos.
- Ejemplo:

Pac. ID	Dolor de garganta	Fiebre	Glándulas hinchadas	Congestión	Dolor de cabeza	Diagnóstico
1	Sí	Sí	Sí	Sí	Sí	Faringitis
2	No	No	No	Sí	Sí	Alergia
3	Sí	Sí	No	Sí	No	Resfriado
4	Sí	No	Sí	No	No	Faringitis
5	No	Sí	No	Sí	No	Resfriado
6	No	No	No	Sí	No	Alergia
7	No	No	Sí	No	No	Faringitis
8	Sí	No	No	Sí	Sí	Alergia
9	No	Sí	No	Sí	Sí	Resfriado
10	Sí	Sí	No	Sí	Sí	Resfriado



El árbol generaliza los datos de la tabla específicamente en:

- Si un paciente tiene inflamación de glándulas, el diagnóstico es laringitis.
- Si un paciente no tiene inflamación de glándulas pero sí tiene fiebre el diagnóstico es resfriado.
- Si un paciente no tiene inflamación de glándulas ni fiebre el diagnóstico es alergia.



Pac. ID	Dolor de garganta	Fiebre	Glándulas hinchadas	Conges-tión	Dolor de cabeza	Diagnóstico
11	Sí	Sí	No	No	Sí	?
12	No	No	Sí	Sí	Sí	?
13	No	No	No	No	Sí	?

Aprendizaje no-supervisado

- Para construir el modelo no se cuenta con clases predefinidas.
- Los datos se agrupan basados en un esquema de similitud.
- De acuerdo a ciertas técnicas de evaluación, se decide el significado de los grupos.
- Ejemplo: Se tiene la siguiente BD de inversionistas:

ID	Tipo cuenta	Cuenta Princ.	Método Trans.	Trans/ Mes	Sexo	Edad	Hobbie	Ingreso Anual
1005	Conjunta	Si	Online	12.5	F	30-39	Tenis	40-59K
1013	Custodia	Si	Broker	0.5	F	50-59	Bucear	80-99 K
1245	Conjunta	Si	Online	3.6	M	20-29	Golf	20-39K
2110	Individ.	No	Broker	22.3	M	30-39	Pescar	40-59K
1001	Individ.	No	Online	5.0	M	40-49	Golf	60-79K

... Aprendizaje no-supervisado

Posibles preguntas:

- ¿Puedo desarrollar un perfil de los inversionistas Online? En ese caso, ¿qué características los distinguen de los que usan un corredor?
- ¿Puedo determinar si un cliente que no ha abierto una cuenta principal, lo hará en un futuro?
- ¿Puedo construir un modelo capaz de predecir la cantidad promedio de transacciones mensuales para un nuevo inversionista?
- ¿Qué características diferencian a los inversionistas de las inversionistas?

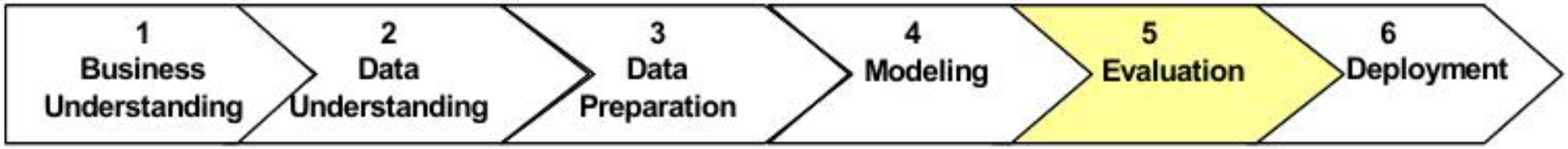
- ¿Qué atributos comparten grupos de clientes de Inversiones ACME?
- ¿Qué diferencias en atributos permiten segmentar los clientes de la BD?

... Aprendizaje no-supervisado

- Grupo 1
  - Cuenta Principal = no, edad = 20-29 e ingresos = 40-59K.
  - Precisión 80 %
  - Cobertura 25 %
- Grupo 2
  - Tipo de Cuenta = custodia, hobbie = Bucear e ingresos = 80-90K.
  - Precisión 95 %
  - Cobertura 15 %
- Grupo 3
  - Tipo de cuenta = conjunta, mov./mensuales > 5 y transacciones online
  - Precisión 82 %
  - Cobertura 55 %

... Proceso de minería de datos (Evaluación)

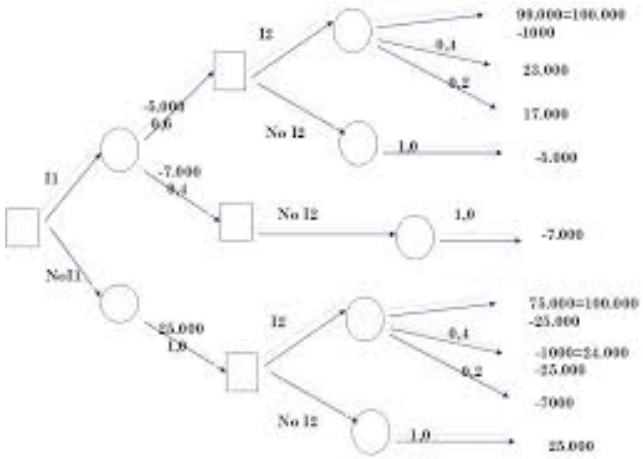
Fase 5: Evaluación.



- Este paso evalúa el grado en el cual el modelo cumple con los objetivos del negocio y busca determinar si hay alguna razón del negocio por la cual el modelo sea deficiente.
- Compara los resultados con el criterio de evaluación definido al principio.

... Proceso de minería de datos (Evaluación)

- Una vez que se ha descubierto conocimiento, es necesario interpretar qué significa y evaluarlo:
  - ¿El conocimiento descubierto es nuevo o ya se sabía?
  - ¿El conocimiento descubierto es nuevo y además útil?
  - ¿Qué tan confiable es el conocimiento descubierto? ¿Qué garantiza que se pueda aplicar a otros datos?
  - Si se han construido diferentes modelos, usando técnicas diferentes, o la misma técnica con diferentes parámetros. ¿pueden combinarse para mejorar resultados?



Evaluación

Preguntas generales:

- ¿Los beneficios recibidos del proyecto de MD compensan los costos del proceso de MD?
- ¿Cómo interpretar los resultados de MD?
- ¿Podemos utilizar los resultados del proceso de MD con confianza?

... Proceso de minería de datos (Evaluación)

Si el proceso es supervisado, suele usarse una matriz de confusión para evaluarlo.

	C_1	C_2	C_3
C_1	C_{11}	C_{12}	C_{13}
C_2	C_{21}	C_{22}	C_{23}
C_3	C_{31}	C_{32}	C_{33}

- C_{ij} i es la clase real y, j la clase calculada.
- En el renglón C_i se tienen los datos que pertenecen a la clase C_i.
- En la columna C_i están los datos que se clasificaron en la clase C_i.
- Elementos en la diagonal, son los elementos clasificados correctamente.

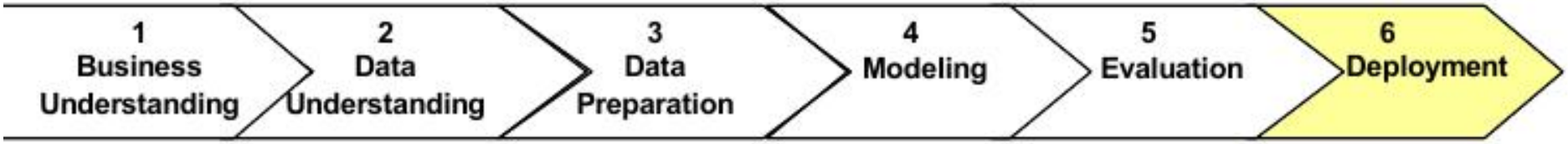
Ejemplo:

	BadLoan	GoodLoan
BadLoan	41	259
GoodLoan	13	687

Precisión del modelo = (diagonal/total de datos) = 72 %

... Proceso de minería de datos (Explotación)

Fase 6: Explotación de nuevo conocimiento.



- La creación del modelo, generalmente, no es el fin del proyecto.
- El conocimiento adquirido necesita organizarse y presentarse en una forma que el cliente pueda usar.
- Es necesario incorporar el nuevo conocimiento en el proceso de la organización y usarlo en producción.
- Diferentes audiencias requieren diferente tipo de información.
  - Para los altos ejecutivos relacionados con préstamos bancarios, lo más importante es saber cómo la aplicación del modelo reducirá la cantidad de dinero que el banco pierde con malos créditos.
  - Los ejecutivos que otorgarán el préstamo necesitan saber cómo utilizar/interpretar el modelo.

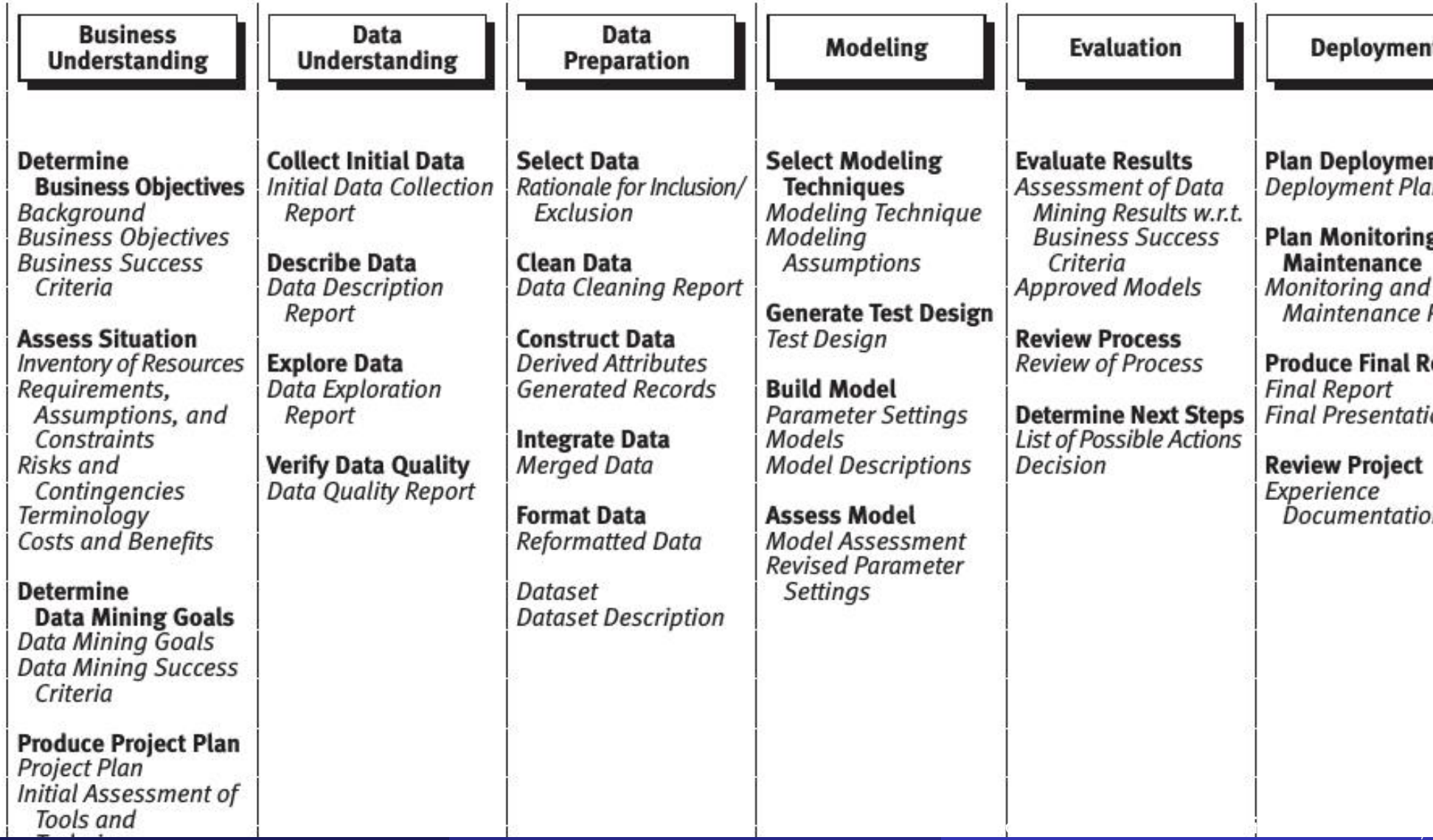
... Proceso de minería de datos (Explotación)

¡¡ Finalmente el modelo se pone en operación!!

... Proceso de minería de datos

- Estas fases no son lineales, se puede tener ciclos de retroalimentación entre ellas.
- Ejemplos:
  - En la fase de modelado se puede descubrir que se han perdido datos relevantes, y se debe regresar a la fase de colección.
  - En la fase de validación se puede encontrar que nuestro modelo no tiene el comportamiento adecuado con datos diferentes de los usados y se debe regresar a la etapa de modelado.

... Proceso de minería de datos (Resumen)

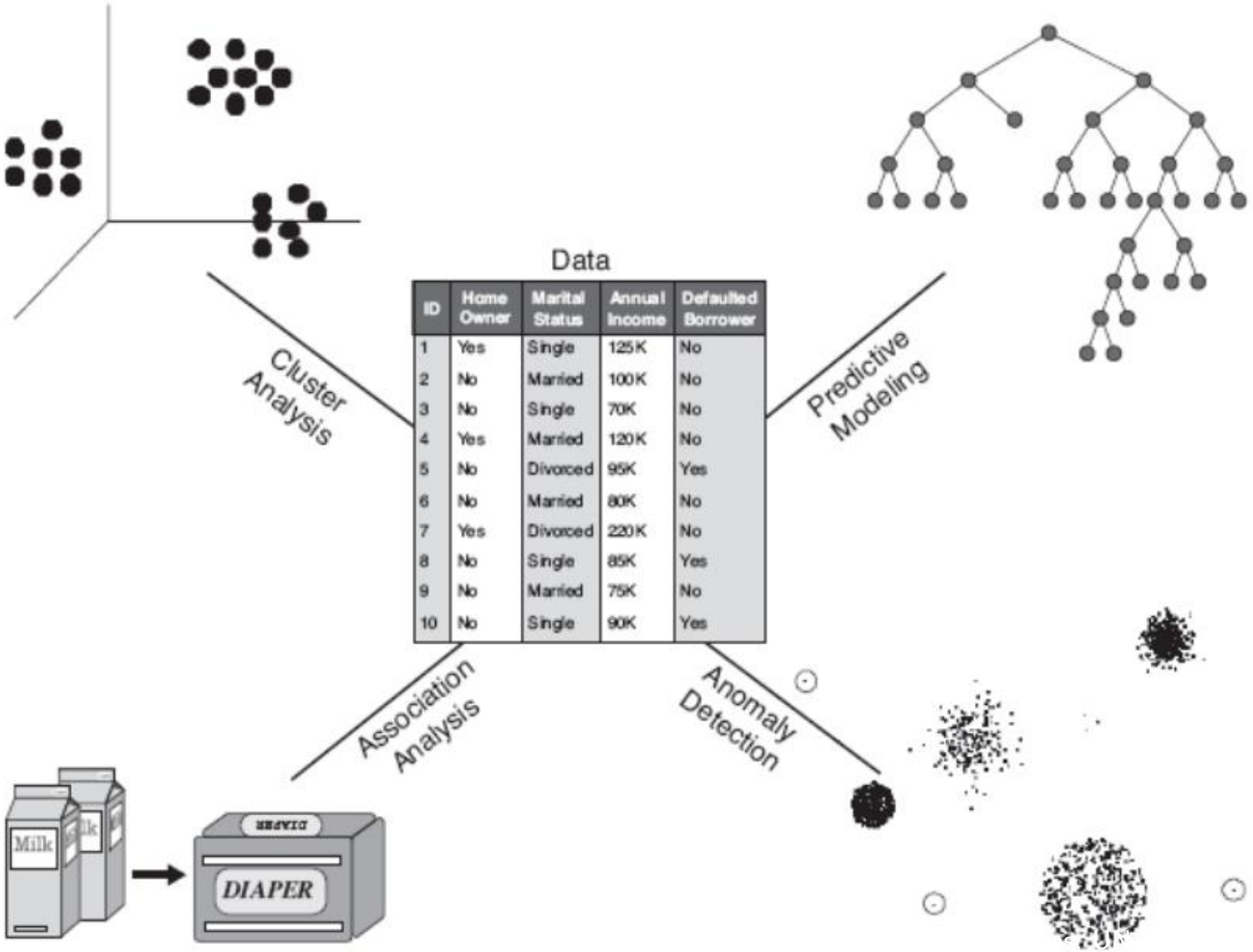


Aplicaciones

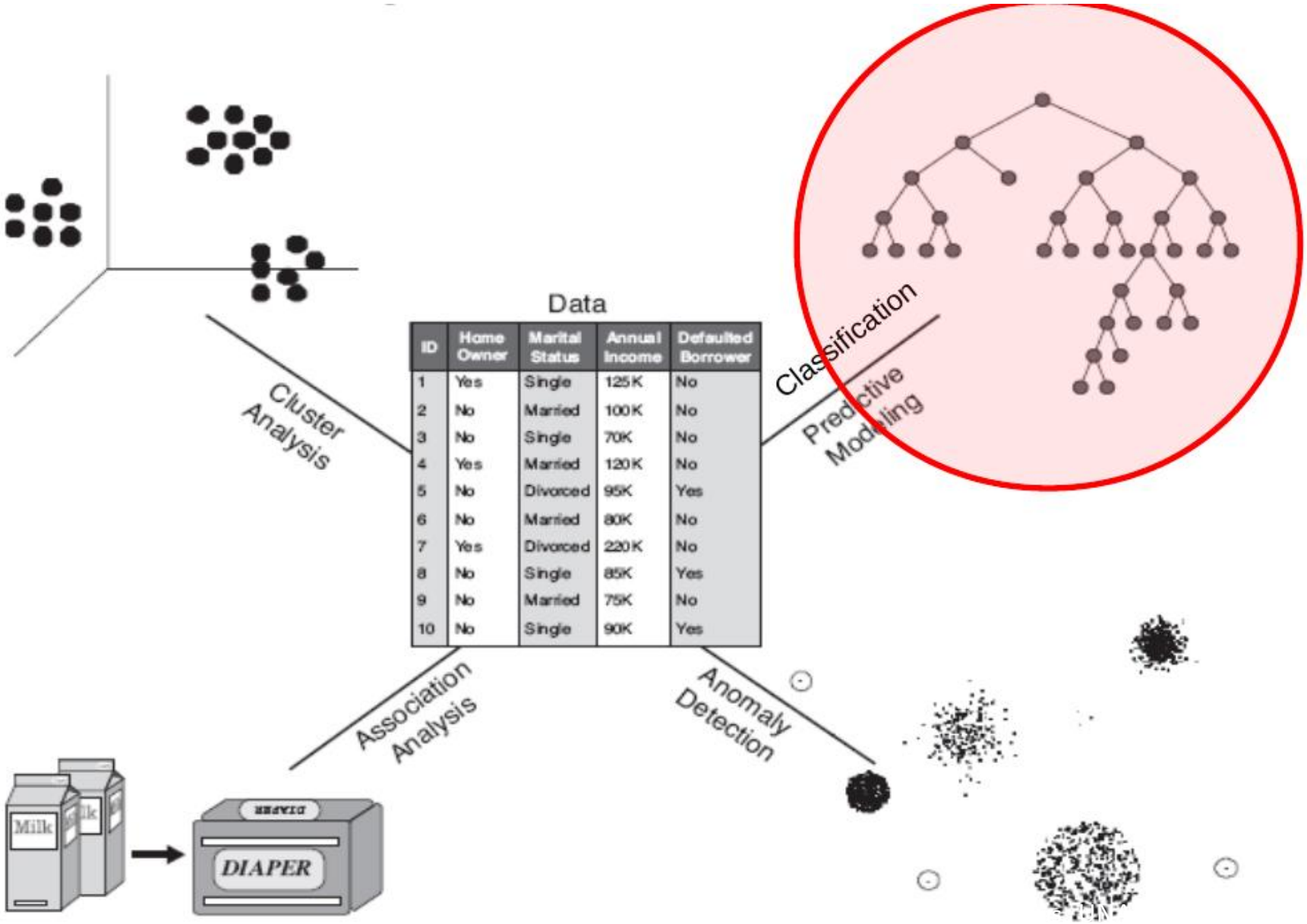
Las aplicaciones de DM pueden agruparse como sigue:

1. Relaciones de mercadotecnia.
2. Perfilado de clientes.
3. Segmentación de clientes.
4. Detección de fraudes.
5. Diseño de sitios web y promociones.

Técnicas de minería de datos



... Técnicas de minería de datos



Clasificación: Definición

- Dado un conjunto de registros (conjunto de entrenamiento)
  - Cada registro contiene un conjunto de atributos, uno de los cuales es su clase.
- Encontrar un modelo para la clase como una función de los valores de los otros atributos.
- Meta: Asignar una clase a registros que no se han clasificado, de la manera más precisa posible.
  - Un conjunto de prueba se usa para determinar la precisión del modelo. Usualmente, el conjunto de datos dado, se divide en conjuntos de entrenamiento y de prueba.

... Clasificación: Ejemplo

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

... Clasificación: Aplicación

Marketing directo.

- **Objetivo:** Reducir el costo de envío, por correo, de propaganda, enfocándose en un conjunto de consumidores que probablemente comprarán un nuevo producto.
- **Enfoque:**
  - Usar los datos para productos similares introducidos antes.
  - Sabemos cuáles clientes decidieron comprarlo y cuáles no.
  - Recabar información relacionada con la demografía, estilo de vida e interacción con la compañía, por parte de tales clientes:
    - Tipo de negocios, dónde están, cuánto ganan, dónde compran, etc.
  - Usar esta información como atributos de entrada para obtener un modelo de clasificación.

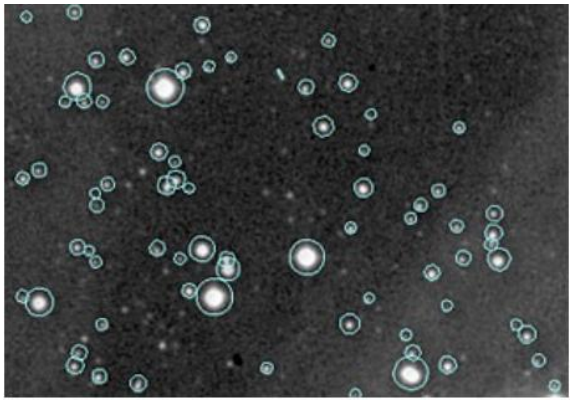
... Clasificación: Aplicación

Retención de clientes (Churn):

- **Objetivo:** Predecir si es probable que ciertos clientes se vayan con la competencia.
- **Enfoque:**
  - Usar un registro detallado de transacciones con cada uno de los clientes anteriores y actuales, encontrar sus atributos:
    - Frecuencia de llamadas del cliente, duración de las llamadas, a dónde llama, horario de llamadas, estatus financiero, estado civil, etc.
  - Etiquetar los clientes como leales o desleales.
  - Encontrar un modelo de lealtad.

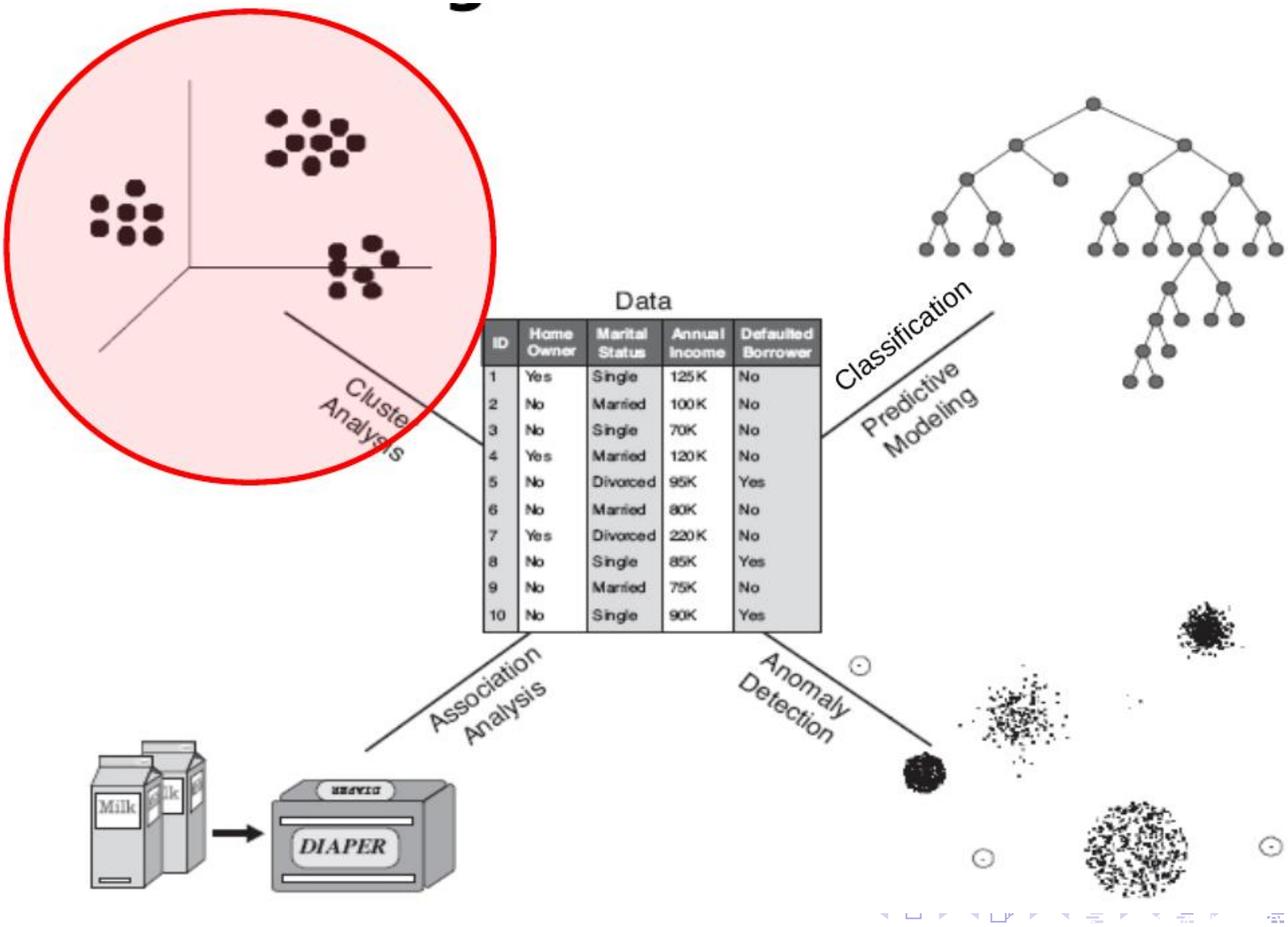
... Clasificación: Aplicación

Catálogo de estudios del cielo.



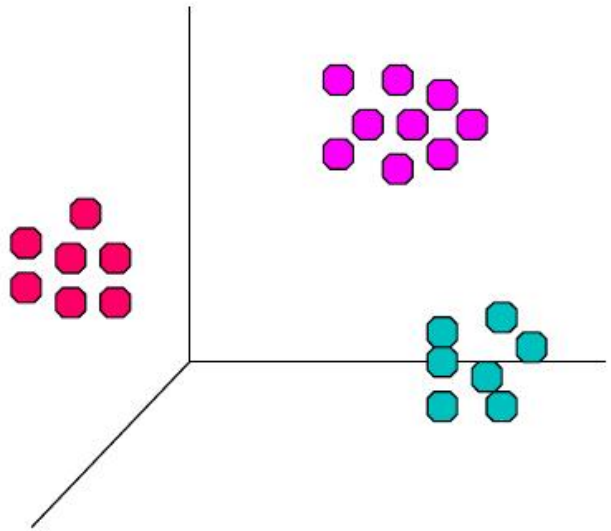
- **Objetivo:** Predecir la clase (estrella o galaxia) de los objetos en el cielo, especialmente los poco visibles, basados en las imágenes de telescopio. Ejemplo observatorio del monte Palomar en USA.
  - 3,000 imágenes con 23,040 $\times$ 23,040 píxeles por imagen.
- **Enfoque:**
  - Segmentar la imagen.
  - Medir los atributos de la imagen (características)
    - Alrededor de 40 por imagen.
  - Modelar la clase basado en estas características.
  - Historia de éxito: Se encontraron 16 nuevos cuerpos celestes algunos muy difíciles de encontrar por su lejanía.

... Técnicas de minería de datos



Agrupamiento: Definición

- Dado un conjunto de datos, cada uno con un conjunto de atributos, y una medida de similitud entre ellos, encontrar grupos (clusters) tales que:
  - Los elementos en el grupo son muy similares entre sí, y
  - no son tan similares a elementos en otros grupos.

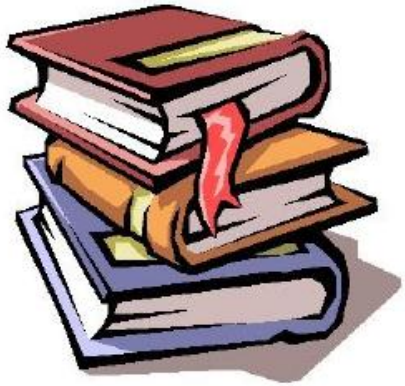


... Agrupamiento: Aplicación

Segmentación de mercado.

- **Objetivo:** Dividir un mercado en distintos subconjuntos de clientes donde cualquier subconjunto puede ser seleccionado como un mercado objetivo dentro de una mezcla de distintos marketings.
- **Enfoque:**
  - Colectar diferentes atributos de clientes basados en su información geográfica y estilo de vida.
  - Encontrar grupos de clientes similares.
  - Medir la calidad de los grupos observando patrones de compra de los clientes en el mismo grupo vs. los de diferentes grupos.

... Agrupamiento: Aplicación



Agrupar documentos.

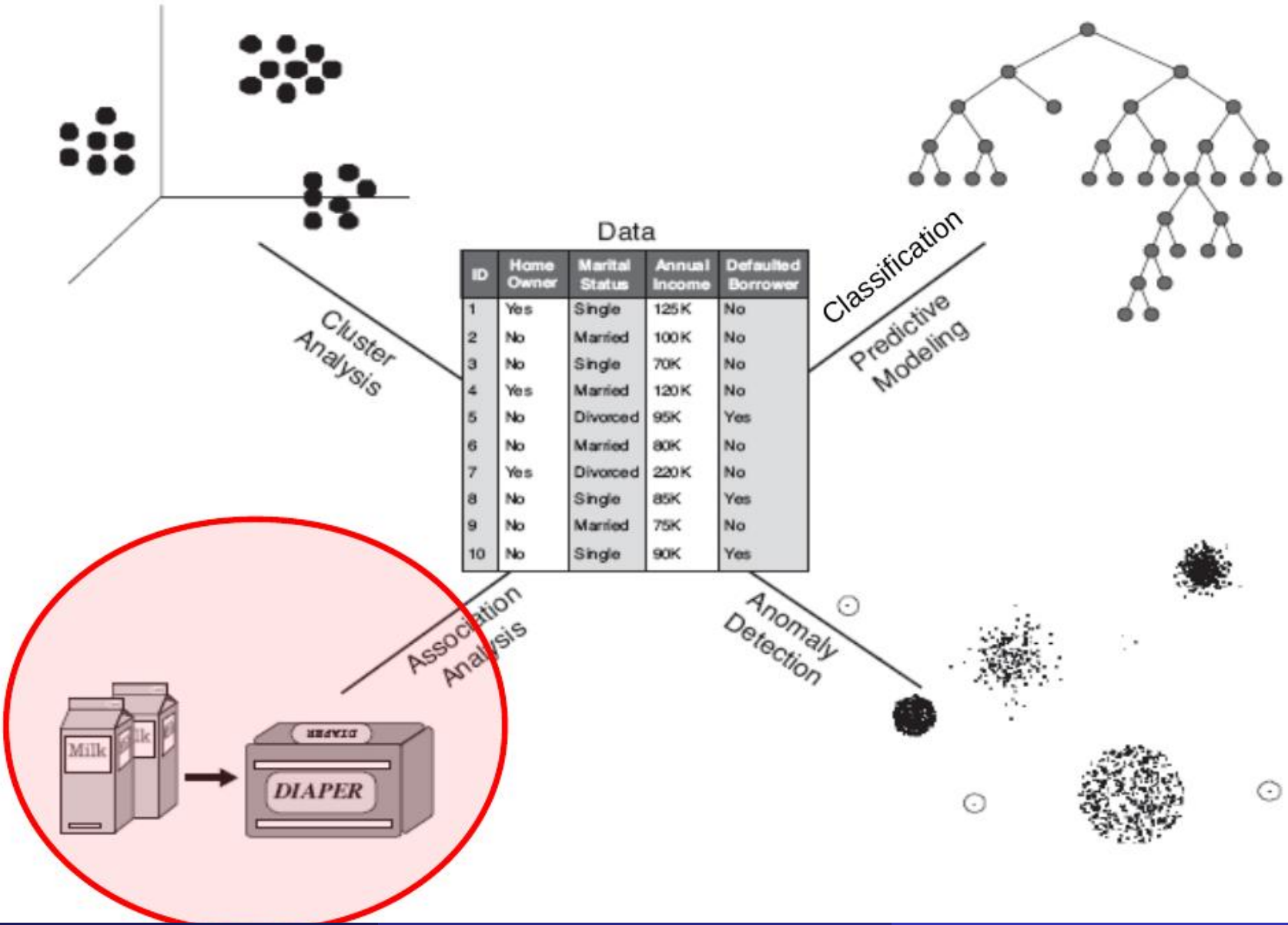
- **Objetivo:** Encontrar grupos de documentos que son similares entre sí basados en la aparición de términos importantes en ellos.
- **Enfoque:**
  - Identificar la frecuencia de aparición de términos en cada documento.
  - Formar una medida de similitud basada en las frecuencias de los diferentes términos.
  - Usar la medida para agrupar.
- **Beneficio:** Se pueden utilizar técnicas de recuperación de información para agrupar nuevos documentos relacionados o buscar términos en documentos agrupados.

... Agrupamiento: Aplicación

- Puntos de agrupación: 3204 artículos del Los Ángeles Times.
- Medida de similitud: ¿Cuántas palabras tienen en común estos documentos (después de algún filtrado).

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Técnicas de minería de datos



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Reglas de asociación: Definición

- Dado un conjunto de registros, cada uno de los cuales contiene elementos de una colección dada:
  - Generar reglas de dependencia que predecirán la ocurrencia de un elemento basado en la ocurrencia de otros.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Discovered Rules:

{Milk} → {Coke}

{Diaper, Milk} → {Beer}

Reglas de asociación: Aplicación

Administración de anaqueles.

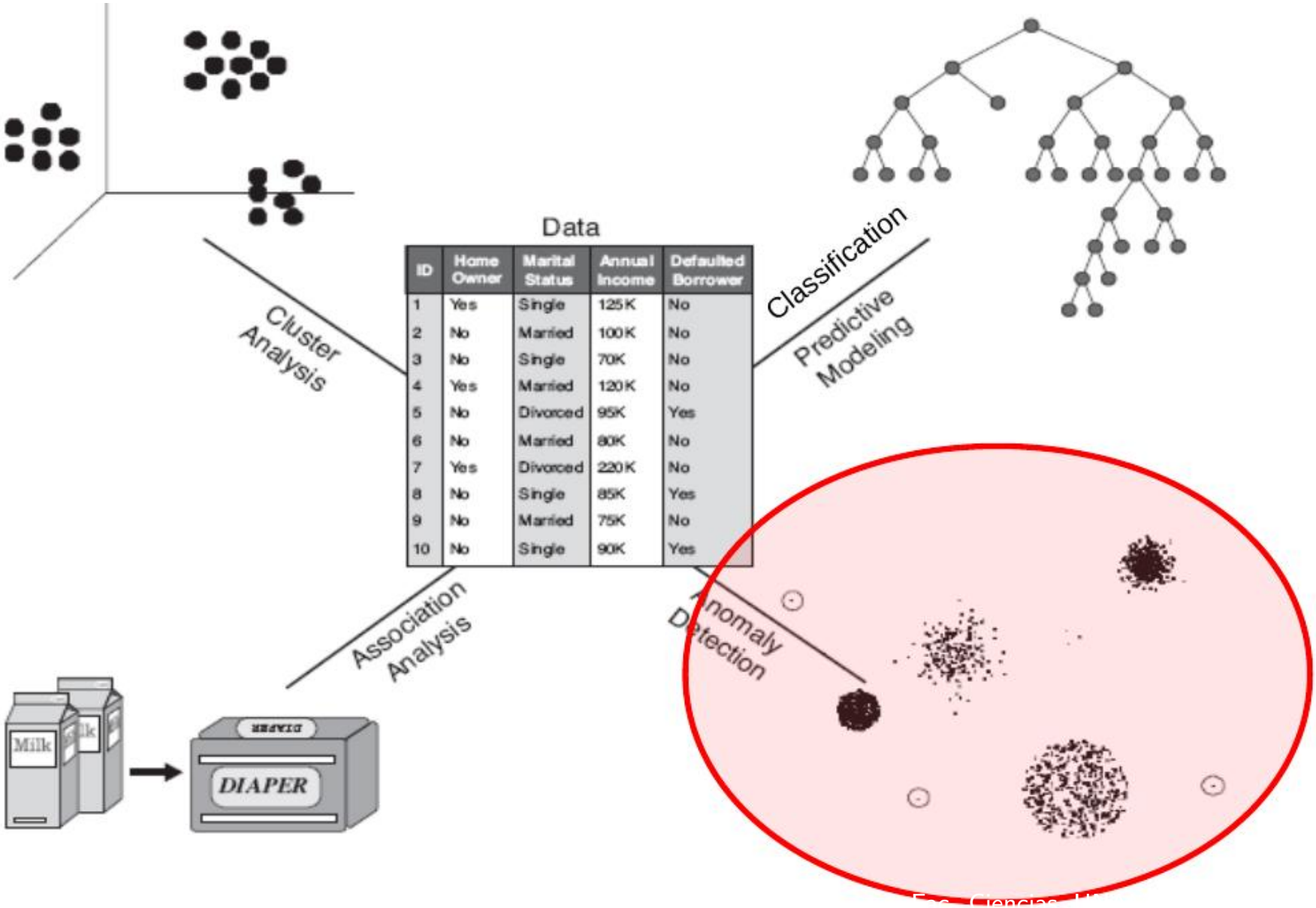
- **Objetivo:** Identificar artículos que se compran juntos por una cantidad considerable de clientes.
- **Enfoque:** Procesar los datos generados en los puntos de venta al escanear los códigos de barras para encontrar dependencias entre los artículos.
- Una regla clásica:
  - Si un cliente compra pañales y leche, entonces es muy probable que compre cerveza.
  - Por lo tanto, no es sorprendente encuentrar los paquetes de cerveza cerca de los pañales.

Reglas de asociación: Aplicación

Administración de inventarios:

- **Objetivo:** Una compañía de reparación de electrodomésticos desea anticiparse a la naturaleza de las reparaciones sobre sus productos y mantener los vehículos de servicio equipados con las partes adecuadas para reducir la cantidad de visitas a los domicilios de los consumidores.
- **Enfoque:** Procesar los datos sobre herramientas y partes utilizadas en reparaciones anteriores en diferentes localidades de consumidores y descubrir patrones de co-ocurrencia.

Técnicas de minería de datos



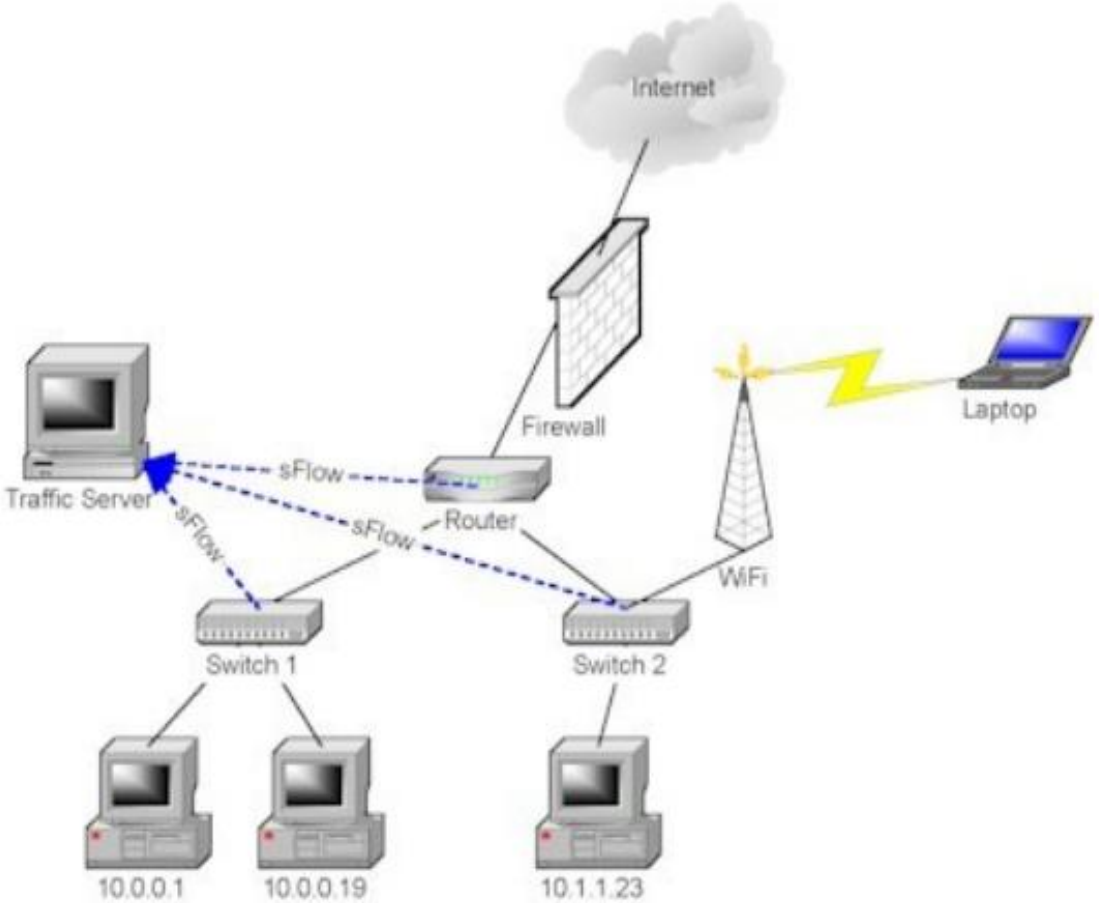
Detección de anomalías/atípicos

- Detectar desviaciones significativas del comportamiento normal.
- Detección de fraudes

  - **Objetivo:** Predecir casos fraudulentos en transacciones con tarjetas de crédito.
  - **Enfoque:**
    - Uso de operaciones con tarjetas de crédito y la información acerca del titular de la tarjeta.
    - Dónde y cuándo ha comprado, qué ha comprado, con qué frecuencia paga a tiempo.

... Detección de atípicos (Aplicación)

- Detección de intrusiones en redes.



Patrones

La forma en que se presentan los patrones encontrados depende de la tarea aplicada.

- Reglas del tipo: `Si antecedente entonces consecuente`.
- Árboles de decisión.
- Listas de datos, etc.

Patrones Interesantes

- ¿Son interesantes todos los patrones o reglas encontrados?
  Respuesta: sólo una pequeña fracción de ellos lo es.
- Esto lleva a las preguntas
  1. ¿Qué hace interesante a un patrón?
  2. ¿Puede un SMD generar todos los patrones interesantes?
  3. ¿Puede un SMD generar sólo patrones interesantes?
- Un patrón es interesante si:
  1. Es fácilmente entendible por humanos.
  2. Es válido sobre datos nuevos o de prueba con algún grado de certeza,
  3. Es potencialmente útil, y
  4. Es nuevo.
  5. Valida una hipótesis que el usuario desea confirmar.

  En una palabra un patrón interesante representa conocimiento.

... Patrones Interesantes

- La segunda pregunta se refiere a la **completez** del algoritmo. Es irreal e ineficiente para SMD generar todos los patrones posibles, en lugar de ello el usuario proporciona restricciones y medidas de interés que deberían ser usadas en la búsqueda.
- La tercera pregunta es un problema de **optimización**. Sólo se deberían generar patrones interesantes, así los usuarios no tendrían que buscar a través de patrones generados para identificar la veracidad de ellos. Esto es un reto en MD.
- Las medidas de interesantes de patrones son esenciales para descubrir patrones de valor para el usuario. Estas medidas pueden usarse después de la MD para ordenar los patrones descubiertos de acuerdo a los intereses y filtrar los no interesantes.