

# PROCESO KDD

*Dra. Rosanna Costaguta (Prof. Responsable)*

**LSI - UNSE - 2019**

A large, multi-pointed pink starburst shape with a thin pink outline, centered in the upper half of the slide. The text 'MINERÍA DE DATOS' is written in white, bold, sans-serif capital letters across its center.

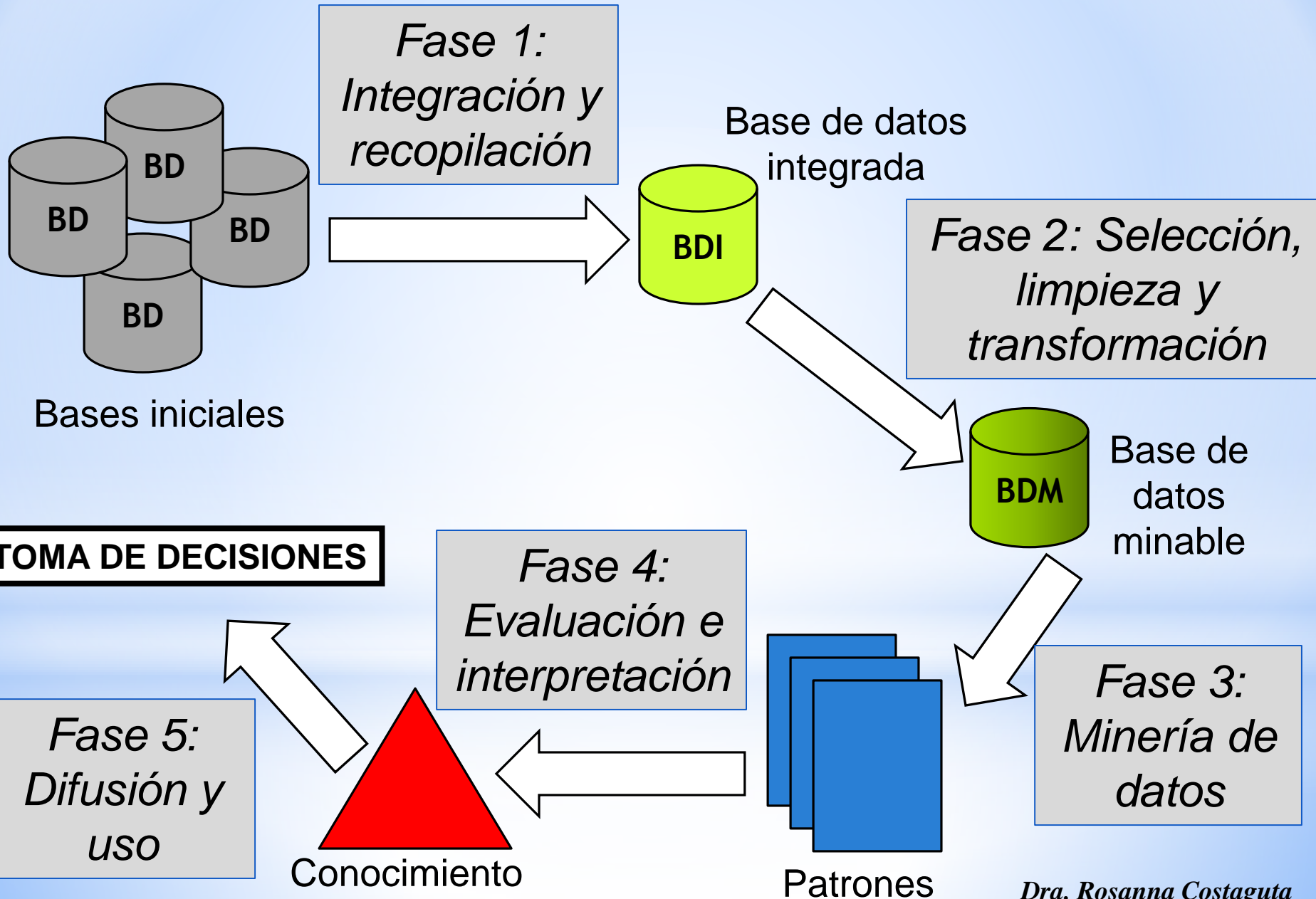
# MINERÍA DE DATOS

***Proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes volúmenes de datos almacenados en diferentes formatos.***

(Witten y Frank, 2000)

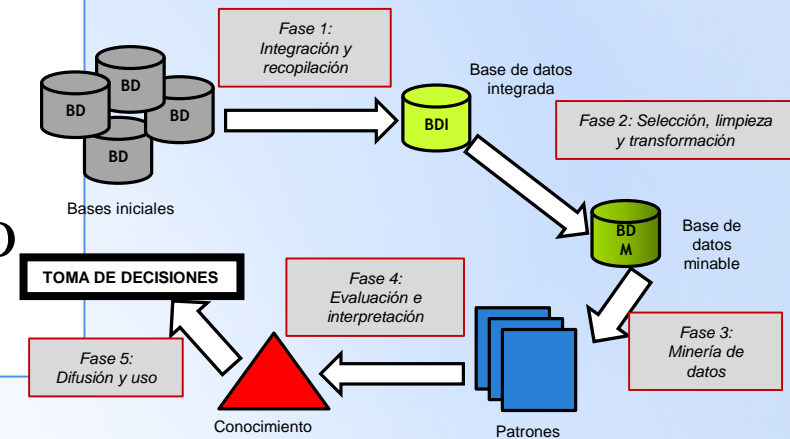
# KDD (Knowledge Discovery from Database)

(Hernández Orallo *et al.*, 2004)



# CARACTERÍSTICAS DE KDD

KDD es un proceso constituido por 5 (cinco) fases, pero además se dice que es un proceso iterativo e interactivo...



- ✓ **Iterativo:** se puede volver atrás para rehacer alguna fase determinada, también es posible comenzar otra vez desde la fase 1
- ✓ **Interactivo:** el usuario (o experto) debe colaborar principalmente en la selección de los datos y en la interpretación de los resultados

# Fase 1: INTEGRACIÓN y RECOPIACIÓN

... primero hay que tener los datos 😊

Algunos interrogantes:

¿de qué fuentes internas/externas se extraerán?

¿cómo se los debería organizar?

¿cómo se los mantendrá actualizados?

¿se requerirá más de una vista minable?

*Algunas veces responder estas preguntas no es simple...*

## EJEMPLO:

Empresa láctea multinacional desea lanzar agresiva campaña publicitaria para incrementar las ventas del producto Z...



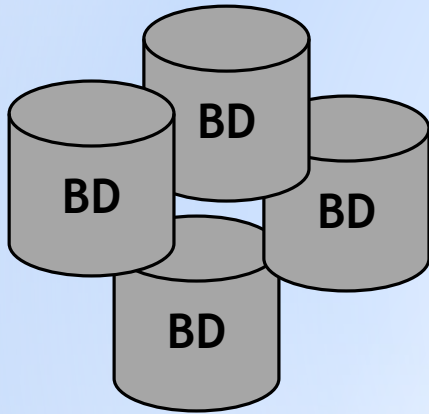
*BD transaccional (datos sobre ventas, productos, proveedores, clientes, recursos humanos, etc.)*

## **SOLUCIÓN!!**

Consultas SQL y sabremos cuáles son los países donde el producto Z tiene pocas ventas



*Datos demográficos de cada país*  
*Distribución etaria en cada país*  
*Preferencias en consumo de lácteos en cada país*  
*Datos económicos de cada país*  
*¿Datos sanitarios de cada país?*  
*¿Datos climáticos de cada país?*  
*¿Datos geográficos de cada país?*



Datos internos  
+externos

## **PREGUNTA**

*¿Siempre debe contarse con un almacén de datos para aplicar KDD?*

## **RESPUESTA**

**No.** Cuando el volumen de datos no es muy grande puede trabajarse directamente con ellos en diferentes formatos (hojas de cálculo, archivos txt, etc.)

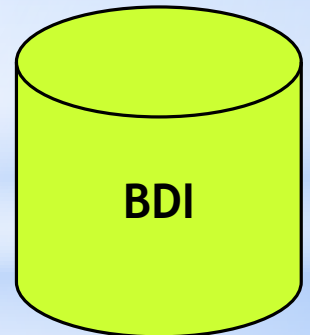
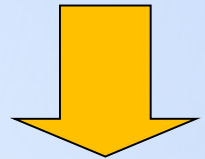
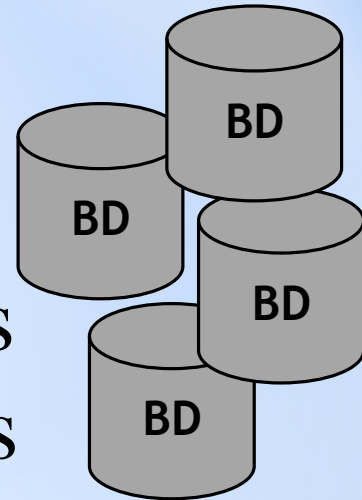


# DIFICULTAD

... nos encontramos con datos que presentan diferentes formatos, distintos grados de integración, diferentes claves primarias, etc.

## RETO

integrar adecuadamente todos los datos



*Base de datos  
integrada*

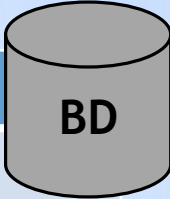
## Fuente 1: *BD Registro Civil*

DNI	Apellido	Nombre	Estado	Hijos	Profesion	...
20076924	Costaguta	Rosanna	Casada	4	Ing.	...
...						



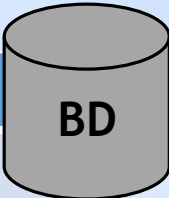
## Fuente 2: *BD Registro del Automotor*

ID-Veh	DNI-Prop	Marca	Modelo	Color	...
NHX542	20076924	Fiat Grand Siena	2014	Rojo	...
...					



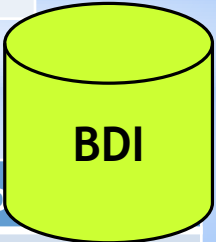
## Fuente 3: *BD Telefonía móvil*

Línea	DNI-Titular	Año	Plan	Aparato	...
3854020XYZ	20076924	2004	Abono - Tipo Q	Samsung	...
...					



## *BD-INTEGRADA*

DNI	Apellido	Nombre	...	Id-Veh	Marca	...	Línea	Año	...
20076924	Costaguta	Rosanna	...	NHX542	Fiat...	...	3854020XYZ	2004	...



*Dra. Rosanna Costaguta*

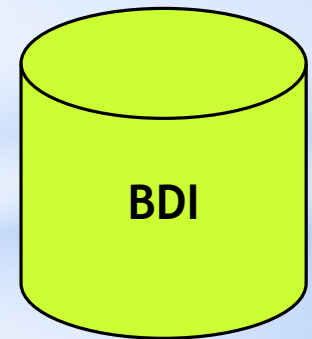
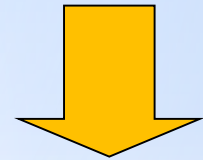
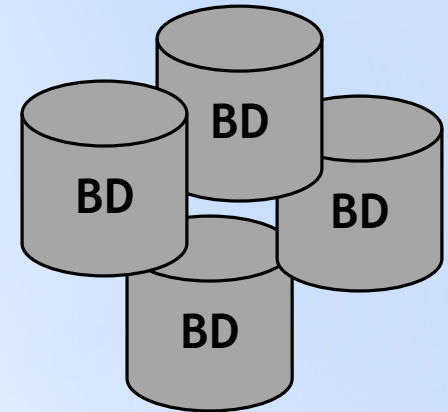
# Fase 1: INTEGRACIÓN y RECOPIACIÓN

*Primer problema:*

Identificar a los objetos...

“Esclarecimiento de la identidad”

- ✓ Dos o más objetos diferentes se unifican
- ✓ Dos o más objetos iguales se dejan separados



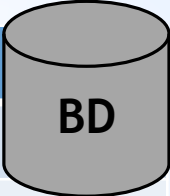
*Base de datos  
integrada*

*Segundo problema:*

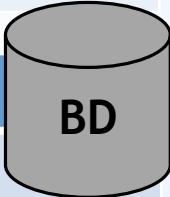
Datos faltantes...

# Dos o más objetos diferentes se unifican...

- ✓ Es el menos frecuente
- ✓ Se usan identificadores externos (claves principales/secundarias en las fuentes de datos)
- ✓ Se suele ser conservador al unificar...

Legajo	Apellido	...		<i>Facultad A</i>
20307/2010	Ponce	...		
20574/2012	Portezuelo	...		

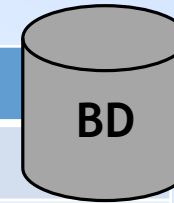
Legajo	Apellido	...		<i>Facultad B</i>
20307/2010	Gerez	...		
20313/2011	Gómez	...		

*Universidad X*

Legajo	Apellido	...	
20307/2010	?????	...	
20313/2011	Gómez	...	
20574/2012	Portezuelo	...	

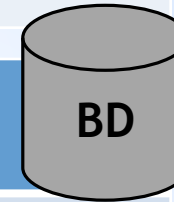
# Dos o más objetos iguales se dejan separados...

ID-Producto	Descripción	Cantidad
Parl20	Parliament x 20	135
...		



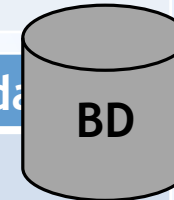
*Kiosko A*

ID-Producto	Descripción	Cant
Parl20	Cig.Parliament X 20	80
...		



*Kiosko B*

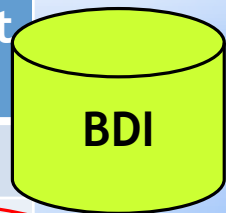
ID-Producto	Descrip-Prod	Cantidad
CigParlGR	Parliament paq x 20	93
...		



*Kiosko C*

*MisKioskos*

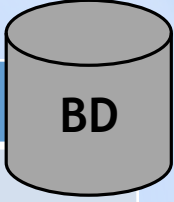
ID-Producto	Descripción	Cantidad	Cant
Parl20	Parliament x 20	135	80
<b>CigParlGR</b>	<b>Parliament paq x20</b>	<b>93</b>	<b>-</b>
...			



# Datos faltantes...

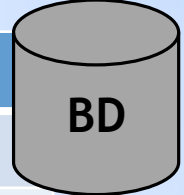
## Fuente 1: *BD Cliente Sucursal 1*

DNI	Apellido	Nombre	Estado	Hijos	Profesión	...
20076924	Costaguta	Rosanna	Casada	4	Dra.	...
...						



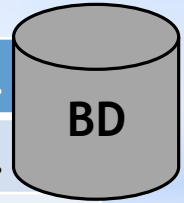
## Fuente 2: *BD Clientes Sucursal 2*

DNI	Apellido	Nombre	Estado	Profesión	...
20076924	Costaguta	Rosanna	Soltera	Ing.	...
...					



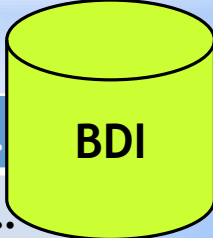
## Fuente 3: *BD Clientes Sucursal 3*

DNI	Apellido	Nombre	Estado	Profesión	...
20076924	Costaguta	Rosanna	Casada	Profesora UNSE	...
...					



## *BD-INTEGRADA*

DNI	Apellido	Nombre	Estado	Hijos	Profesión	...
20076924	Costaguta	Rosanna	-	4	-	...
...						



*Dra. Rosanna Costaguta*



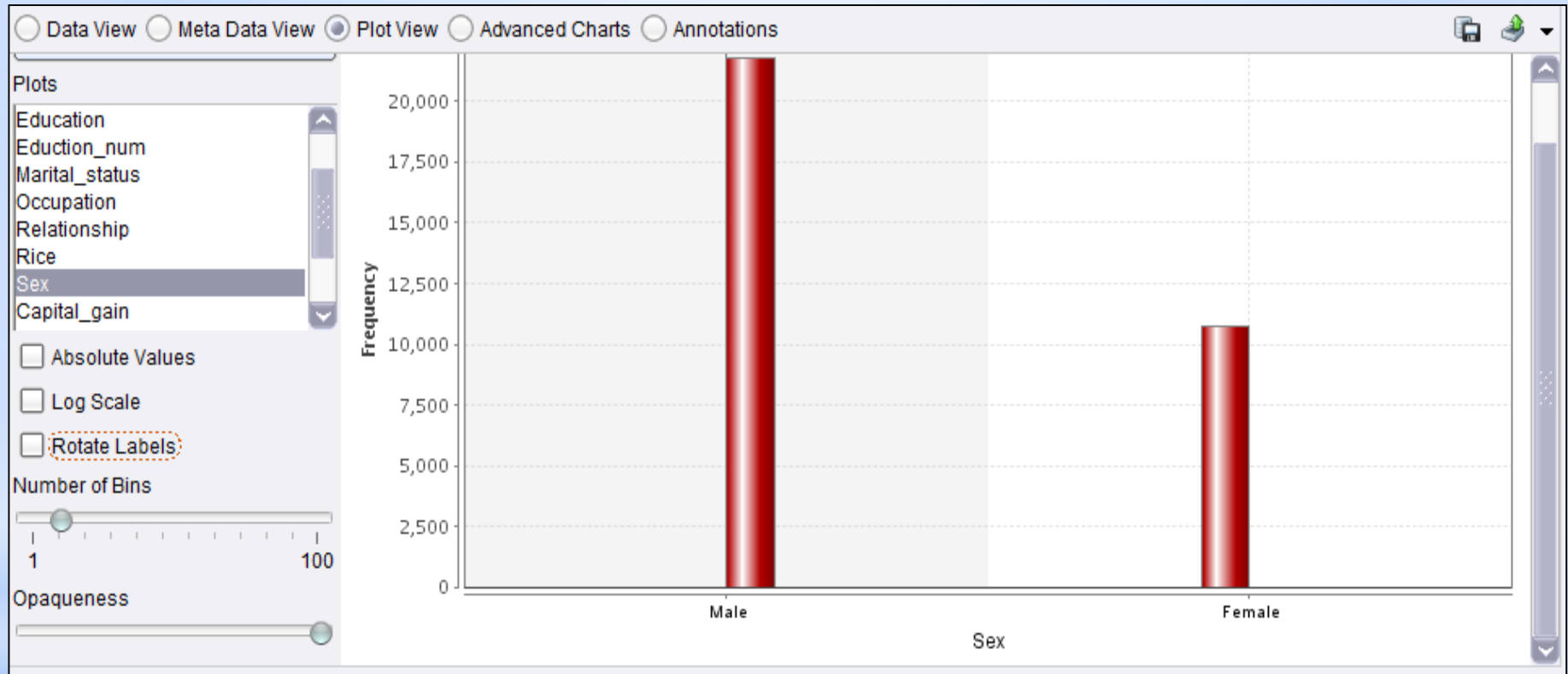
Finalizando la integración conviene realizar un “*reconocimiento*”...

- ✓ Resumen de atributos
  - ✓ Histogramas
- ✓ Gráficos de dispersión

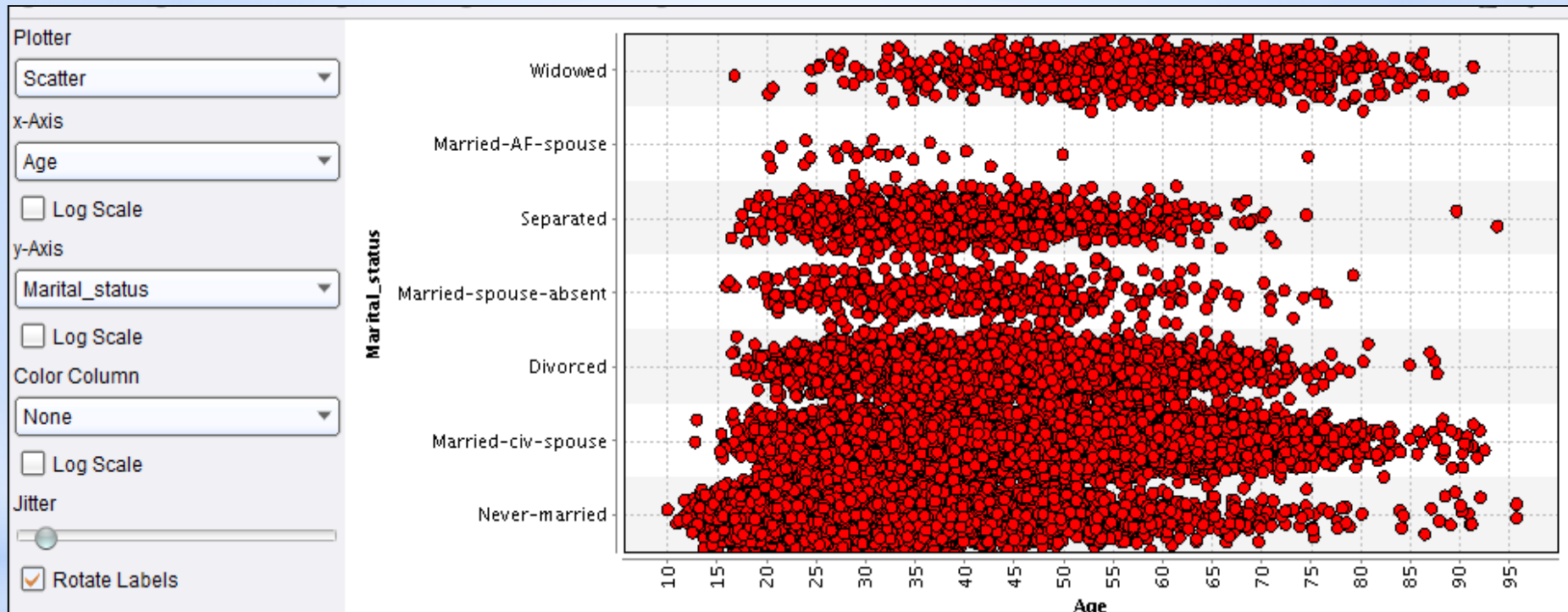
# Resumen de características de los atributos...

ExampleSet (32561 examples, 0 special attributes, 15 regular attributes)					
Role	Name ▲	Type	Statistics	Range	Missings
regular	Age	integer	avg = 38.582 +/- 13.640	[17.000 ; 90.000]	0
regular	Capital_gain	integer	avg = 1077.649 +/- 7385.292	[0.000 ; 99999.000]	0
regular	Capital_loss	integer	avg = 87.304 +/- 402.960	[0.000 ; 4356.000]	0
regular	Class	binominal	mode = <=50K (24720), least = >50K (7841)	<=50K (24720), >50K (7841)	0
regular	Education	polynominal	mode = HS-grad (10501), least = Preschool (51)	Bachelors (5355), HS-grad (10501), 11th (1175), M	0
regular	Eduction_num	integer	avg = 10.081 +/- 2.573	[1.000 ; 16.000]	0
regular	Hours_per_week	integer	avg = 40.437 +/- 12.347	[1.000 ; 99.000]	0
regular	Marital_status	polynominal	mode = Married-civ-spouse (14976), least = Marr	Never-married (10683), Married-civ-spouse (14976)	0
regular	Native_country	polynominal	mode = United-States (29170), least = Holand-N	United-States (29170), Cuba (95), Jamaica (81), In	583
regular	Occupation	polynominal	mode = Prof-specialty (4140), least = Armed-Forc	Adm-clerical (3770), Exec-managerial (4066), Hanc	1843
regular	Relationship	polynominal	mode = Husband (13193), least = Other-relative	Not-in-family (8305), Husband (13193), Wife (1568)	0
regular	Race	polynominal	mode = White (27816), least = Other (271)	White (27816), Black (3124), Asian-Pac-Islander (1	0
regular	Sex	binominal	mode = Male (21790), least = Female (10771)	Male (21790), Female (10771)	0
regular	Work_Class	polynominal	mode = Private (22696), least = Never-worked (7	State-gov (1298), Self-emp-not-inc (2541), Private (	1836

# Histograma...



# Gráfico de dispersión...



## Fase 2:

# SELECCIÓN, LIMPIEZA y TRANSFORMACIÓN

**PRINCIPIO:** La calidad del conocimiento descubierto no sólo depende del algoritmo de minería de datos utilizado sino también de la calidad de los datos minados

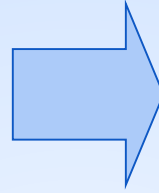
Luego de la recopilación es necesario seleccionar y preparar el subconjunto de datos a *minar*, puesto que seguramente muchos de los datos recolectados resulten irrelevantes o innecesarios para la tarea de minería que se pretende realizar ...

... obtenemos la *vista minable*



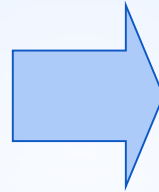
Base de  
datos  
minable

**SELECCIÓN**



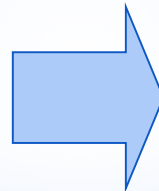
... hay que decidir qué datos se requerirán de todos los recopilados e integrados

**LIMPIEZA**



... puede ser necesario eliminar algunos datos

**TRANSFORMACIÓN**



... puede ser necesario modificar algunos datos, incluso crear nuevos en base a los existentes

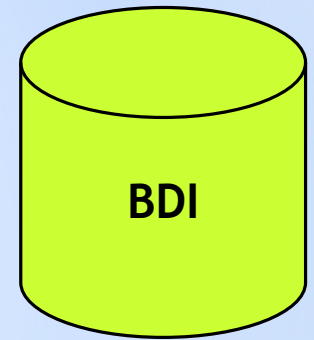
***PREPROCESAMIENTO DE LOS DATOS***



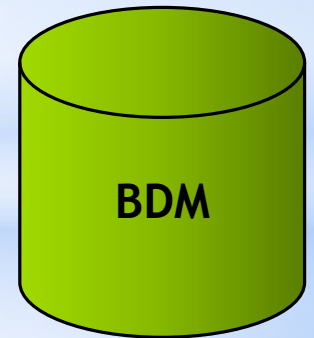
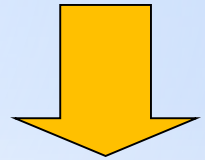
## Fase 2:

# SELECCIÓN, LIMPIEZA y TRANSFORMACIÓN

¿Qué hacemos con los datos faltantes?



*Base de datos  
integrada*

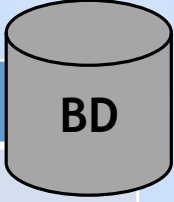


*Base de datos  
minable*

# Datos faltantes...

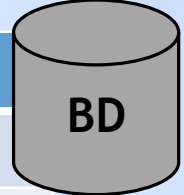
## Fuente 1: *BD Cliente Sucursal 1*

DNI	Apellido	Nombre	Estado	Hijos	Profesión	...
20076924	Costaguta	Rosanna	Casada	4	Dra.	...
...						



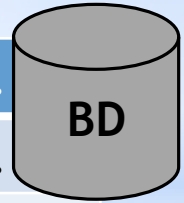
## Fuente 2: *BD Clientes Sucursal 2*

DNI	Apellido	Nombre	Estado	Profesión	...
20076924	Costaguta	Rosanna	Soltera	Ing.	...
...					



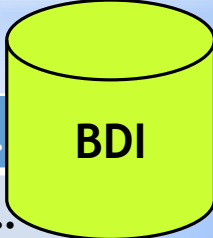
## Fuente 3: *BD Clientes Sucursal 3*

DNI	Apellido	Nombre	Estado	Profesión	...
20076924	Costaguta	Rosanna	Casada	Profesora UNSE	...
...					



## BD-INTEGRADA

DNI	Apellido	Nombre	Estado	Hijos	Profesión	...
20076924	Costaguta	Rosanna	-	4	-	...
...						

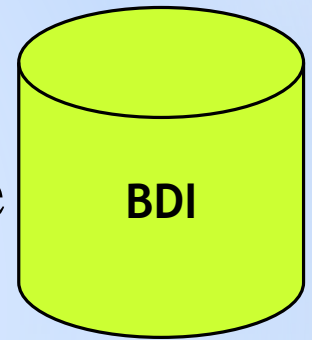


## Fase 2:

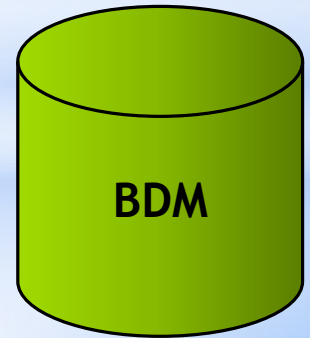
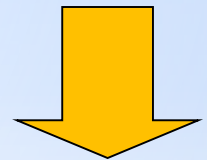
# SELECCIÓN, LIMPIEZA y TRANSFORMACIÓN

¿Qué hacemos con los datos faltantes?

Es fácil detectarlos a través del resumen de atributos.



*Base de datos  
integrada*



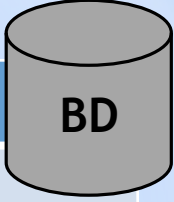
*Base de datos  
minable*

## Estrategias a seguir

- ✓ Ignorarlos
- ✓ Eliminarlos (filtrando la columna)
- ✓ Filtrarlos (quitando la tupla)
- ✓ Reemplazarlos (colocar la *media* en datos numéricos y la *moda* en los nominales, imputarle un *valor estimado*)

## Fuente 1: *BD Cliente Sucursal 1*

DNI	Apellido	Nombre	Estado	Hijos	Profesión	...
20076924	Costaguta	Rosanna	Casada	4	Dra.	...
...						



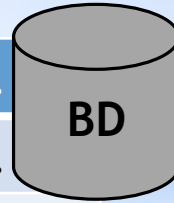
## Fuente 2: *BD Clientes Sucursal 2*

DNI	Apellido	Nombre	Estado	Profesión	...
20076924	Costaguta	Rosanna	Soltera	Ing.	...
...					



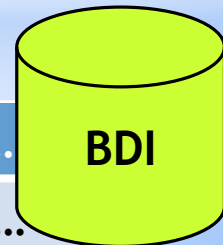
## Fuente 3: *BD Clientes Sucursal 3*

DNI	Apellido	Nombre	Estado	Profesión	...
20076924	Costaguta	Rosanna	Casada	Profesora UNSE	...
...					



## *BD-INTEGRADA*

DNI	Apellido	Nombre	Estado	Hijos	Profesión	..
20076924	Costaguta	Rosanna	Casada	4	Dra.	...
...						



*Dra. Rosanna Costaguta*

# IMPORTANTE

Cuando los datos faltantes se completan o eliminan, se pierde información...  
ya no sabremos cuales “faltaban”.

## *Recomendación:*

- ✓ crear atributo booleano que indique si el atributo anterior era o no faltante
- ✓ crear valor “faltante” para los nominales

## *BD-INTEGRADA*

DNI	Apellido	Nombre	Estado	Hijos	Profesión	Faltante
20076924	Costaguta	Rosanna	Faltante	4	Dra.	SI
...						

**¿Qué son los “nulos camuflados”?**  
Son atributos faltantes no detectados...

*Recomendación:*

✓ Invertir tiempo en reconocerlos

***BD-INTEGRADA***

DNI	Apellido	Nombre	Estado	Hijos	E-mail
20076924	Costaguta	Rosanna	Casada	999	No informado
...					



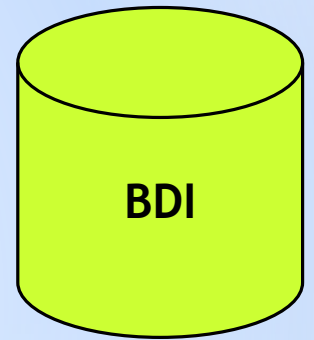
## Fase 2:

# SELECCIÓN, LIMPIEZA y TRANSFORMACIÓN

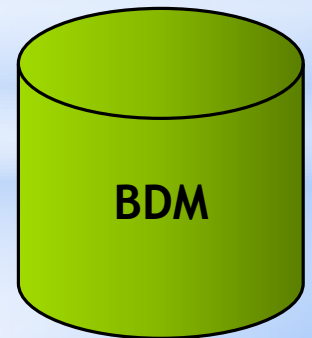
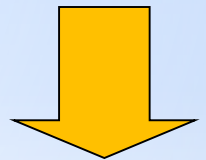
**¿Qué hacemos con los datos erróneos?**

Es fácil detectarlos porque conocemos los valores posibles...

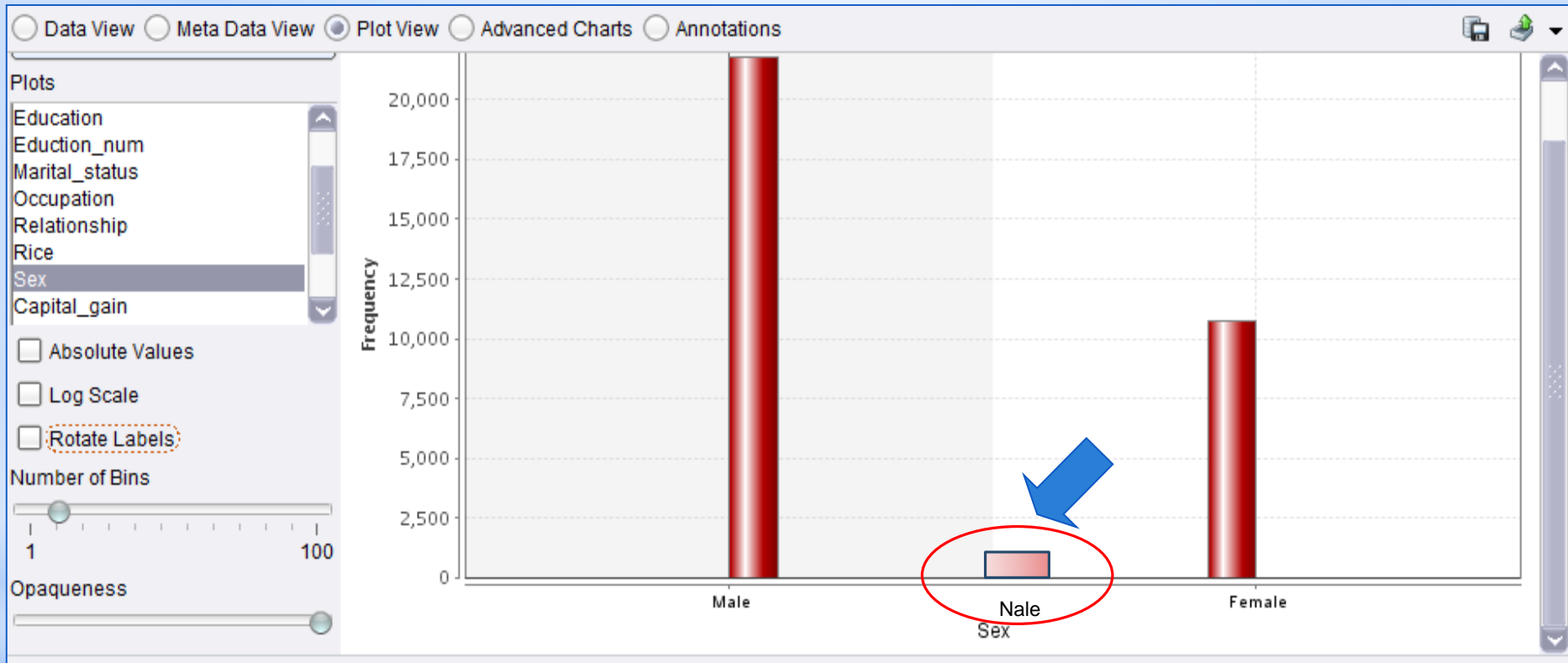
los vemos a través del resumen de atributos y de los histogramas.



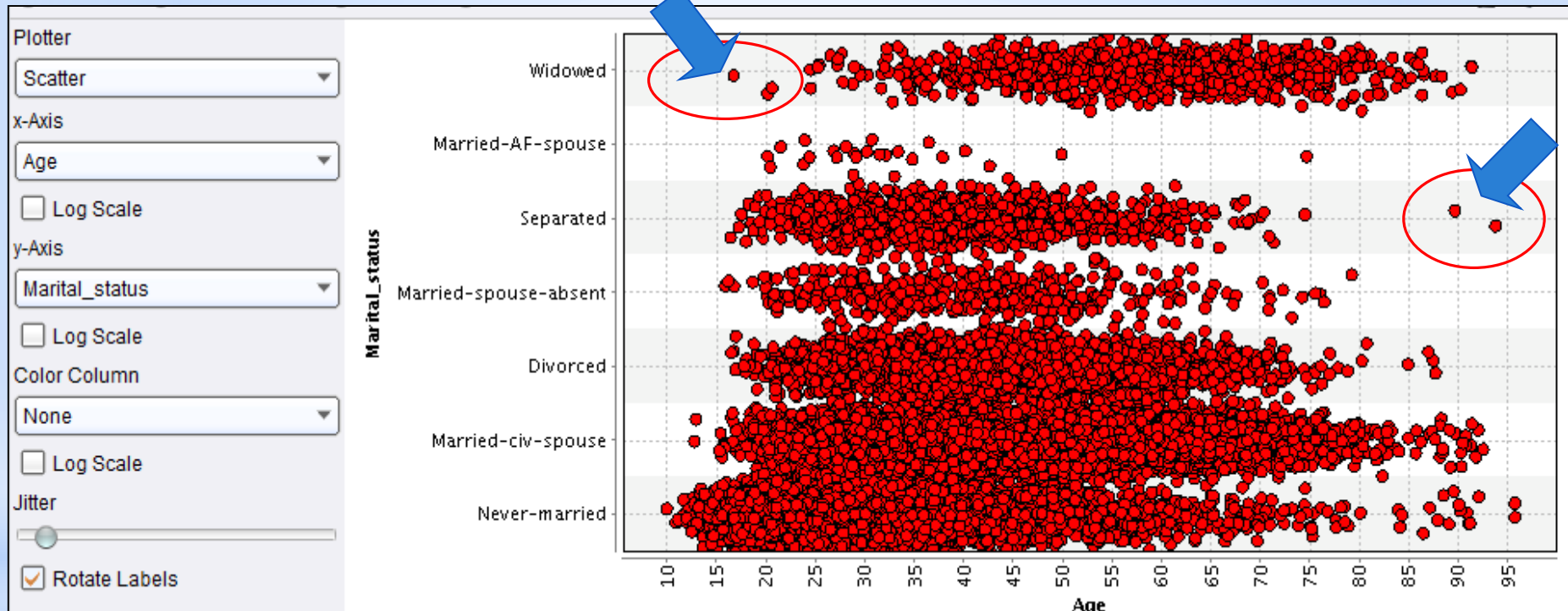
*Base de datos integrada*



# Datos erróneos...



# Datos erróneos...



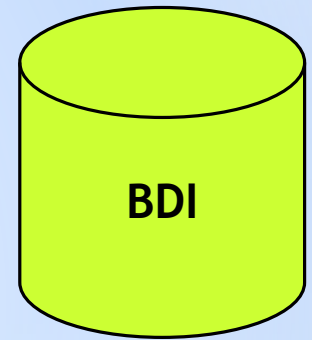
## Fase 2:

# SELECCIÓN, LIMPIEZA y TRANSFORMACIÓN

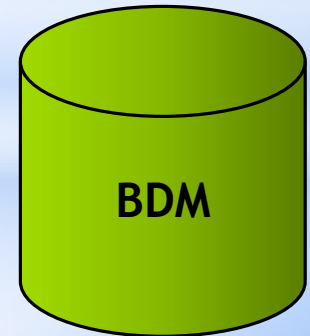
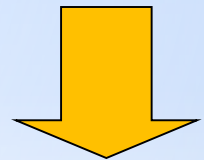
**¿Qué hacemos con los datos erróneos?**

Es fácil detectarlos porque conocemos los valores posibles...

los vemos a través del resumen de atributos y de los histogramas.



*Base de datos integrada*



## Estrategias a seguir

- ✓ Ignorarlos
- ✓ Eliminarlos (filtrando la columna)
- ✓ Filtrarlos (quitando la tupla)
- ✓ Reemplazarlos (imputarle un *valor*)

# IMPORTANTE

Dato erróneo  $\neq$  Dato faltante

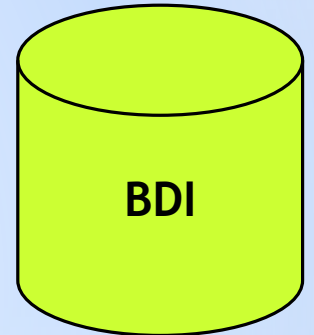
Dato erróneo  $\neq$  Dato anómalo o atípico

## Fase 2:

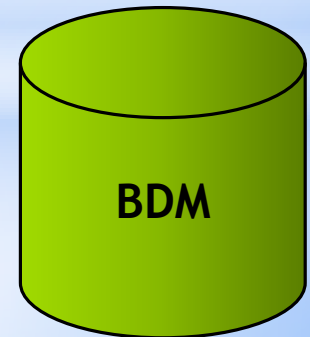
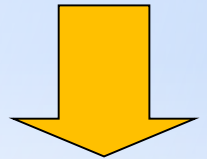
# SELECCIÓN, LIMPIEZA y TRANSFORMACIÓN

## ¿Cómo transformamos datos?

- ✓ Transformamos un atributo en otro
- ✓ Derivamos un nuevo atributo
- ✓ Cambiamos el tipo de dato
- ✓ Cambiamos el rango
- ✓ Reducimos o ampliamos la dimensionalidad



*Base de datos  
integrada*





# ✓ Transformamos un atributo en otro

## a) *Discretización*

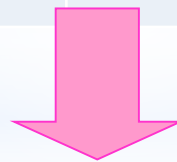
Hijos = 0 → Hijos = Ninguno

Hijos = 1 → Hijos = Chica

Hijos = 2 o 3 → Hijos = Tipo

Hijos = 4 o más → Hijos = Numerosa

DNI	Apellido	Nombre	Estado	Hijos	Profesión	...
20076924	Costaguta	Rosanna	Casada	4	Ing.	...
...						



DNI	Apellido	Nombre	Estado	Hijos	Profesión	...
20076924	Costaguta	Rosanna	Casada	Numerosa	Ing.	...
...						

# ✓ Transformamos un atributo en otro

## b) *Numerización*

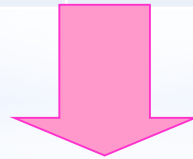
Estado = Soltero/a → Estado = 1

Estado = Casado/a → Estado = 2

Estado = Divorciado/a → Estado = 3

Estado = Viudo/a → Estado = 4

DNI	Apellido	Nombre	Estado	Hijos	Profesión	...
20076924	Costaguta	Rosanna	Casada	4	Ing.	...
...						



DNI	Apellido	Nombre	Estado	Hijos	Profesión	...
20076924	Costaguta	Rosanna	2	4	Ing.	...
...						

## ✓ Creamos nuevo atributo

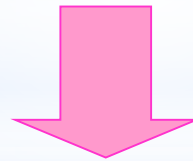
Hijos = 0 → Familia = Matrimonio

Hijos = 1 → Familia = Chica

Hijos = 2 o 3 → Familia = Tipo

Hijos = 4 o más → Familia = Numerosa

DNI	Apellido	Nombre	Estado	Hijos	Profesión	...
20076924	Costaguta	Rosanna	Casada	4	Ing.	...
...						



DNI	Apellido	Nombre	Estado	Hijos	Familia	Profesión	...
20076924	Costaguta	Rosanna	Casada	4	Numerosa	Ing.	...
...							

# ✓ Cambiamos el tipo/el rango

## a) Normalización

$$x' = x - \min / \max - \min$$

(cuando un atributo proviene de fuentes diferentes)

## b) Escalado

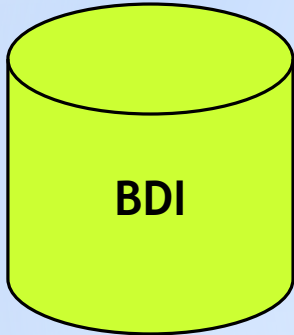
$$x' = 1/\exp(-20(x - 0.5))$$

(cuando existen valores anómalos - *outliers*)

ID	Apellido	Nombre	..	Ingreso mensual	Gasto mensual TC	...
17999881	Carabajal	José	...	7000	2500	Pesos arg.
...						

ID	Apellido	Nombre	..	Ingreso mensual	Gasto mensual TC	...
15764222	Olivas Marti	Gregorio	...	15000	5500	euros
...						

## EJEMPLO:



Tenemos una BDI con los atributos DNI, apellido, nombre, edad, sexo, estado civil y cantidad de hijos.

Queremos aplicar KDD para descubrir alguna vinculación entre la distribución de sexos, edades y estado civil con respecto a los hijos....

DNI	Apellido	Nombre	Edad	Sexo	Estado-Civil	Cant-Hijos
20076924	Costaguta	Rosanna	45	F	Casada	4
...						

*Como preprocesamiento podríamos decidir que los atributos DNI, nombre y apellido resultan innecesarios para la vista minable... también decidimos que no nos interesa el número de hijos sino ciertos rangos predefinidos (1 hijo – familia chica, 2 o 3 hijos – familia tipo, 4 o más hijos - familia numerosa).*

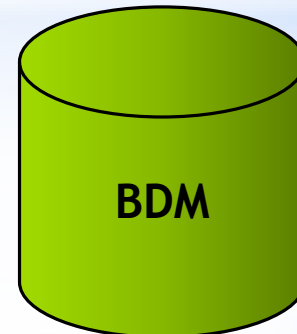
**SELECCIÓN:** edad, sexo, estado civil y cantidad de hijos

**LIMPIEZA:** DNI, nombre y apellido

DNI	Apellido	Nombre	Edad	Sexo	Estado-Civil	Cant-Hijos
20074	Costa	Rosa	45	F	Casada	4
...						

**TRANSFORMACIÓN:** cantidad de hijos → familia

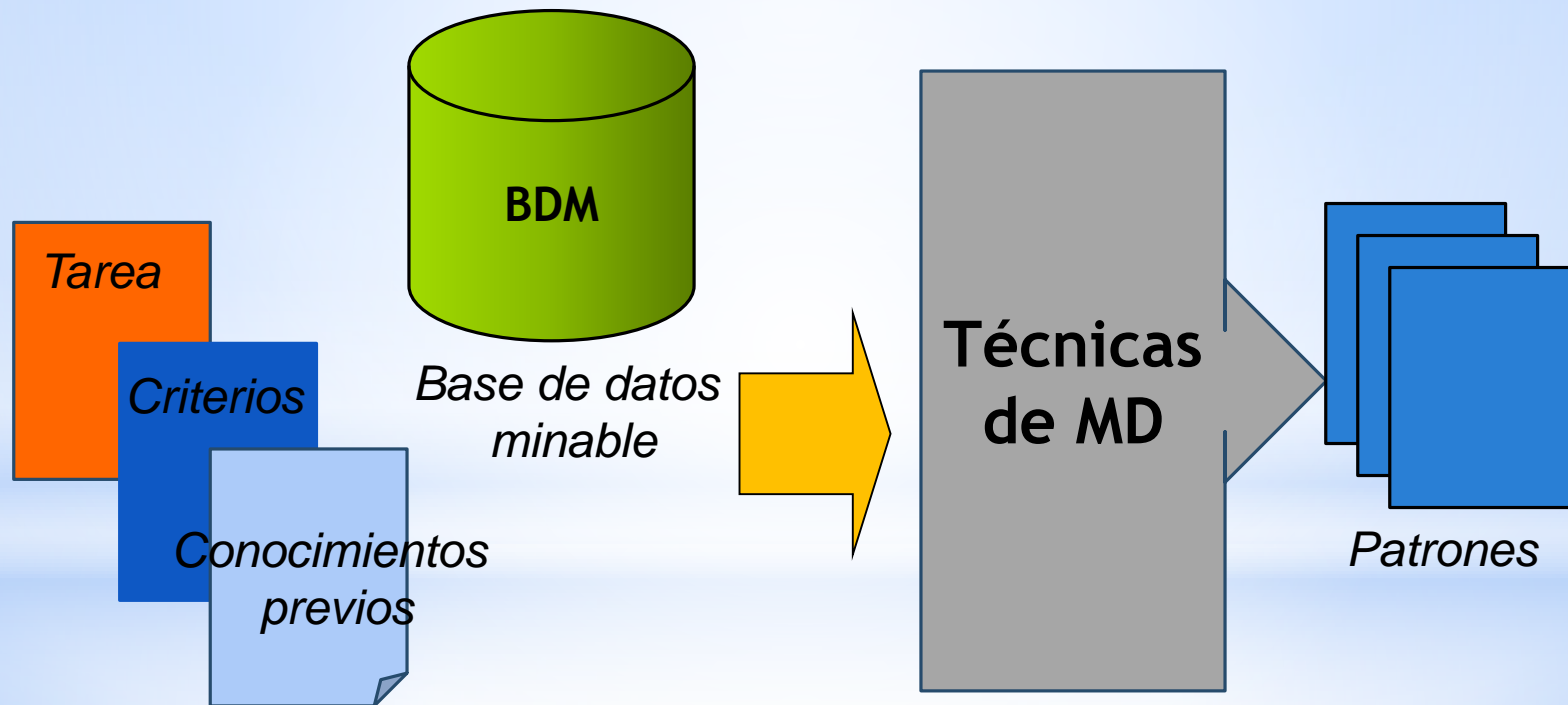
Edad	Sexo	Estado-Civil	Familia
45	F	Casada	Numerosa





# Fase 3: MINERÍA DE DATOS

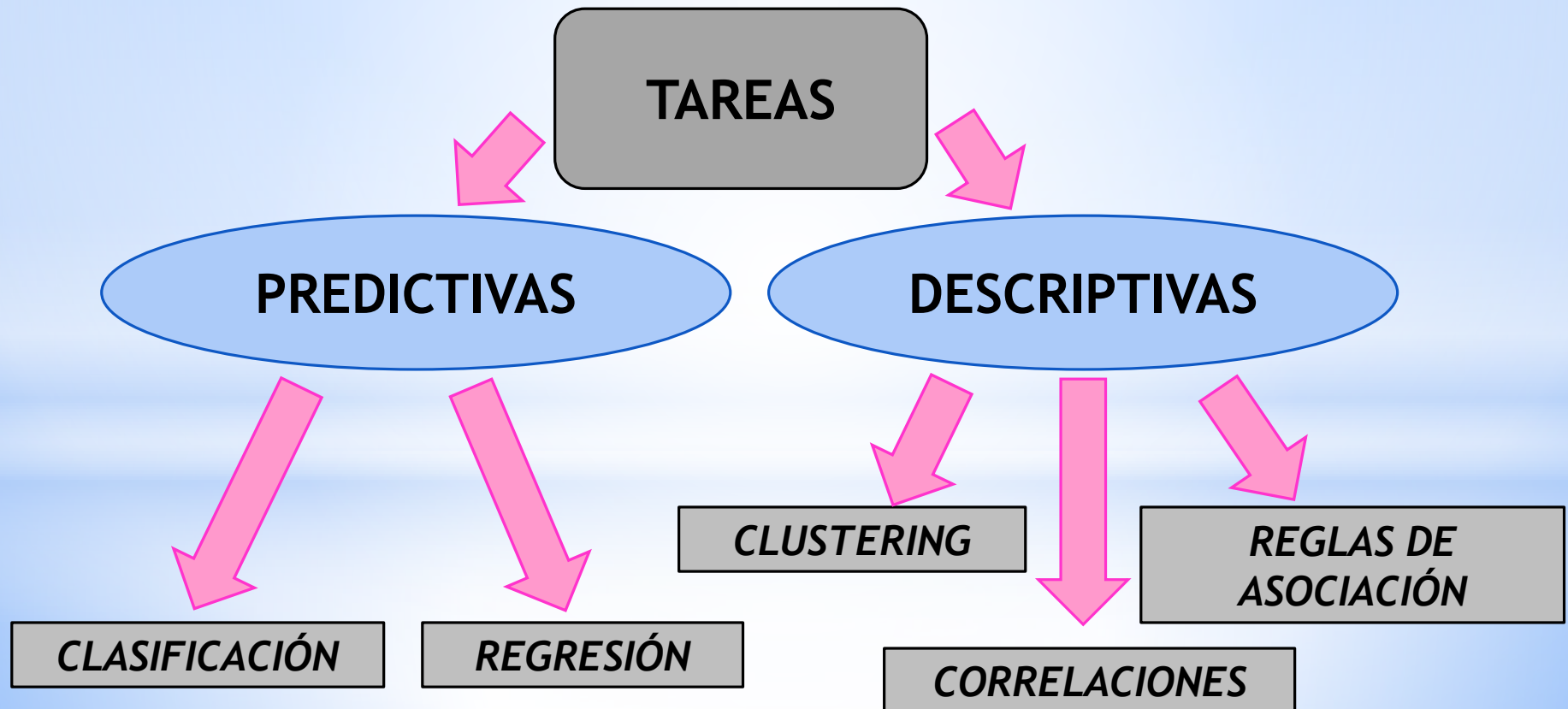
Esta es la fase más característica del proceso de KDD por esto es común que éste sea conocido como minería de datos (*datamining*)



*El proceso no es tan fácil como parece...*

# Fase 3: MINERÍA DE DATOS

En minería de datos cada tarea tiene sus requisitos y la información obtenida con cada una de ellas puede diferir mucho.



# Tareas de MD (Predictivas)

## *CLASIFICACIÓN*

Cada instancia de la BD pertenece a una *clase* indicada en un atributo en particular.

Cada valor posible de ese atributo identifica a las diferentes clases.

El valor de la clase depende del valor de los demás atributos de la tupla.

**EJEMPLO 1:** Una aseguradora posee una BD donde sus clientes están rotulados como *muy riesgosos*, *riesgosos* o *no riesgosos*, en base a los atributos edad, sexo y nivel de ingresos. Cada nuevo cliente es **clasificado** como integrante de una de las tres clases.

**EJEMPLO 2:** Un experto alergista discrimina a sus pacientes como *alérgicos* o *no alérgicos*, en base a ciertos síntomas (estornudos, lagrimeo, irritación de vías respiratorias, y sequedad nasal). Cada nuevo paciente es **clasificado** como perteneciente a una de estas dos clases.

# Tareas de MD (Predictivas)

## *REGRESIÓN*

Mediante una función real se asigna a cada instancia de la BD un *número real*. A diferencia de la clasificación con la regresión a cada instancia se le asigna un valor numérico.

**EJEMPLO 1:** Una empresa constructora posee una BD conteniendo los datos de todas las construcciones realizadas. Utilizando una función de regresión lineal, sobre las instancias contenidas en la BD, es capaz de **predecir** la duración total de construcciones futuras de igual tipo.

**EJEMPLO 2:** Una consultora informática posee una BD conteniendo los datos de todos sus proyectos desarrollados. Utilizando una función de regresión lineal, sobre las instancias contenidas en la BD, es capaz de **predecir** el costo de proyectos informáticos de igual tipo.



# Tareas de MD (Descriptivas)

## *AGRUPAMIENTO*

Consiste en descubrir *grupos* o segmentos a partir de los datos.

El principio del *agrupamiento* o *clustering* es maximizar semejanzas intragrupos y maximizar diferencias intergrupos.

Mediante *clustering* es posible descubrir grupos (*clusters*) en los datos, para luego generar etiquetas y hablar de *clases*.

**EJEMPLO 1:** Una librería que realiza ventas por internet efectúa agrupamientos para reconocer preferencias de compra de sus clientes. Cuando un cliente se interesa por un libro, identifica a que **grupo** pertenece y le sugiere libros adquiridos por otros clientes de ese grupo.

**EJEMPLO 2:** Una agencia de viajes on-line realiza el agrupamiento de sus clientes para descubrir sus preferencias de alojamiento y destino. Cuando un cliente se interesa por paquete turístico en particular, identifica a que **grupo** pertenece y le sugiere destinos visitados por otros clientes de ese grupo.

# Tareas de MD (Descriptivas)

## *CORRELACIÓN*

Examina el *grado de similitud* de los valores asumidos por dos variables numéricas a fin de descubrir si existen comportamientos similares.

Se usa el coeficiente  $r$ ...

- $r = 1$ , variables correlacionadas
- $r = -1$ , variables correlacionadas negativamente
- $r = 0$ , variables no correlacionadas

**EJEMPLO 1:** Una aseguradora desea lanzar a la venta un nuevo seguro. Para definir clientes potenciales podría interesarse en descubrir si el nivel de ingreso de sus asegurados está **correlacionado** con la cantidad de seguros que posee contratados.

**EJEMPLO 2:** La OMS presupone una vinculación entre el uso de cierta vacuna y la proliferación de un determinado virus mutado. Para asegurarse podría intentar descubrir si existe **correlación** entre la cantidad de dosis administradas de la vacuna y la cantidad de nuevos contagios detectados.

# Tareas de MD (Descriptivas)

## *REGLAS de ASOCIACIÓN*

Intenta descubrir relaciones no explícitas entre atributos categóricos de una BD.

La formulación común es “si el atributo  $X$  toma el valor  $a$  entonces el atributo  $Y$  toma el valor  $b$ ”.

*No implica causalidad.*

**EJEMPLO 1:** *Amazon* aplica reglas de asociación para descubrir cuales son los libros que suelen comprarse juntos. Por ejemplo, el 60 % de las veces que alguien compra un libro de Probabilidad también compra uno de Estadística, y esto ocurre en 3 de cada 10 clientes.

**EJEMPLO 2:** *Youtube* aplica reglas de asociación para descubrir cuales son los videos que suelen verse juntos. Por ejemplo, el 80 % de las veces que alguien mira un video de Luis Miguel luego ve otro más, y esto pasa en 7 de cada 10 usuarios.



# Fase 4: EVALUACIÓN y INTERPRETACIÓN

## Evaluación de CLASIFICADORES

Se intenta evaluar la calidad de los patrones encontrados respecto a su precisión predictiva

VP = Verdaderos Positivos (instancias clasificadas correctamente)

FP = Falsos Positivos (instancias clasificadas incorrectamente)

FN = Falsos Negativos (instancias no clasificadas que pertenecían)

$$\textbf{PRECISIÓN} = \text{VP}/(\text{VP}+\text{FP})$$

$$\textbf{RECALL} = \text{VP}/(\text{VP}+\text{FN})$$

*... cuanto más próximo a uno sea el valor de estos indicadores,  
mejor es el resultado de la evaluación...*

# Fase 4: EVALUACIÓN y INTERPRETACIÓN

## Evaluación de CLASIFICADORES

**EJEMPLO 1:** Una aseguradora posee una BD donde sus clientes están clasificados como *riesgoso* o *no riesgoso*, en base a los atributos edad, sexo y nivel de ingresos. Cada nuevo cliente se clasifica como perteneciente a una de estas dos clases.

Sobre 20 nuevas instancias de clientes en la BD se clasificaron correctamente como *no riesgoso* a 17 de ellas, incorrectamente a 2, y como *riesgoso* cuando no lo era a 1.

$$\text{PRECISIÓN} = \text{VP}/(\text{VP}+\text{FP}) =$$

$$17/(17+2) = 17/19 = 0.89$$

$$\text{RECALL} = \text{VP}/(\text{VP}+\text{FN}) =$$

$$17/(17+1) = 17/18 = 0.94$$

# Fase 4: EVALUACIÓN y INTERPRETACIÓN

## Evaluación de Modelos de REGRESIÓN

Se estima la calidad del modelo comparando las predicciones ( $h(x)$ ) con la función objetivo ( $f(x)$ ).

***ERROR CUADRÁTICO MEDIO*** =  $1/n * \text{SUM } (h(x) - f(x))^{**2}$   
considerando n elementos

**EJEMPLO 1:** Una empresa constructora posee una BD conteniendo los datos de todas las construcciones realizadas. Utilizando una función de regresión lineal, sobre las instancias contenidas en la BD, es capaz de predecir la duración total de construcciones futuras de igual tipo.

# Fase 4: EVALUACIÓN y INTERPRETACIÓN

## Evaluación de AGRUPAMIENTOS

Es difícil evaluar ya que no existe clase o valor numérico para contrastar.

**COHESIÓN:** distancia al centroide del grupo desde cada instancia del mismo

**DISTANCIA MEDIA ENTRE GRUPOS:** distancia entre los centroides de los diferentes grupos

**EJEMPLO 1:** Una librería realiza ventas por internet efectúa agrupamientos para reconocer preferencias de compra de sus clientes. Cuando un cliente se interesa por un libro, identifica a que grupo pertenece y le sugiere libros adquiridos por otros clientes de ese grupo.

# Fase 4: EVALUACIÓN y INTERPRETACIÓN

## Evaluación de REGLAS de ASOCIACIÓN

Se busca generar reglas que puedan aplicarse a un mayor número de instancias y que tengan precisión relativamente alta sobre esas instancias

**COBERTURA** = nro. de instancias a las que la regla se aplica correctamente

**CONFIANZA** = proporción de instancias que la regla predice correctamente

# Fase 4: EVALUACIÓN y INTERPRETACIÓN

## Evaluación de REGLAS de ASOCIACIÓN

**EJEMPLO 1:** *Amazon* aplica reglas de asociación para descubrir cuales son los libros que suelen comprarse juntos. Por ejemplo, el 60 % de las veces que alguien compra un libro de Probabilidad también compra uno de Estadística, y esto ocurre en 3 de cada 10 clientes.

**REGLA n:** *IF libro1 = Probabilidad THEN libro2 = Estadística*

**COBERTURA** = nro. de instancias a las que la regla se aplica correctamente  $\leq 60 \%$

**CONFIANZA** = proporción de instancias que la regla predice correctamente =  $30 \%$



**OTRO EJEMPLO:** Un médico neumonólogo posee una BD con información de todos los diagnósticos efectuados. Aplicando Reglas de asociación descubre la siguiente regla...

**IF Fiebre > 36 AND Tos = Si THEN Angina = Si**

Fiebre	Tos	Angina
39	Si	Si
40	Si	No
36	No	No
38	Si	Si
41	Si	Si
38	No	No

$$COBERTURA = 3/6 = 0,50 \%$$

$$CONFIANZA = 3/4 = 0,75 \%$$

**COBERTURA** = nro. de instancias a las que la regla se aplica correctamente

**CONFIANZA** = proporción de instancias que la regla predice correctamente

# Fase 5: DIFUSIÓN, USO y MONITORIZACIÓN

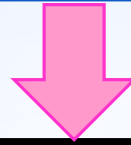
El modelo construido puede incorporarse a una aplicación existente para su ejecución manual o automática...

El nuevo conocimiento debe ser difundido en la organización...

Es necesario monitorear la evolución del modelo. Los patrones pueden cambiar por lo que el modelo debe ser periódicamente reevaluado, reentrenado y hasta quizás reformulado...

# CONSIDERACIONES FINALES

KDD es un proceso *iterativo* porque es conveniente explorar modelos alternativos hasta encontrar aquel que resulte más útil para resolver el problema.



Construido el modelo, y a partir de los resultados obtenidos podríamos decidir cambiar algunos parámetros o utilizar otras técnicas. Esto puede llevarnos a retroceder hasta el *preprocesamiento*...

Cuando se construyen modelos predictivos (clasificación y regresión) se requiere entrenamiento y validación. Se entrena el modelo con una porción de los datos (*training dataset*) y luego se valida con el resto (*test dataset*).