



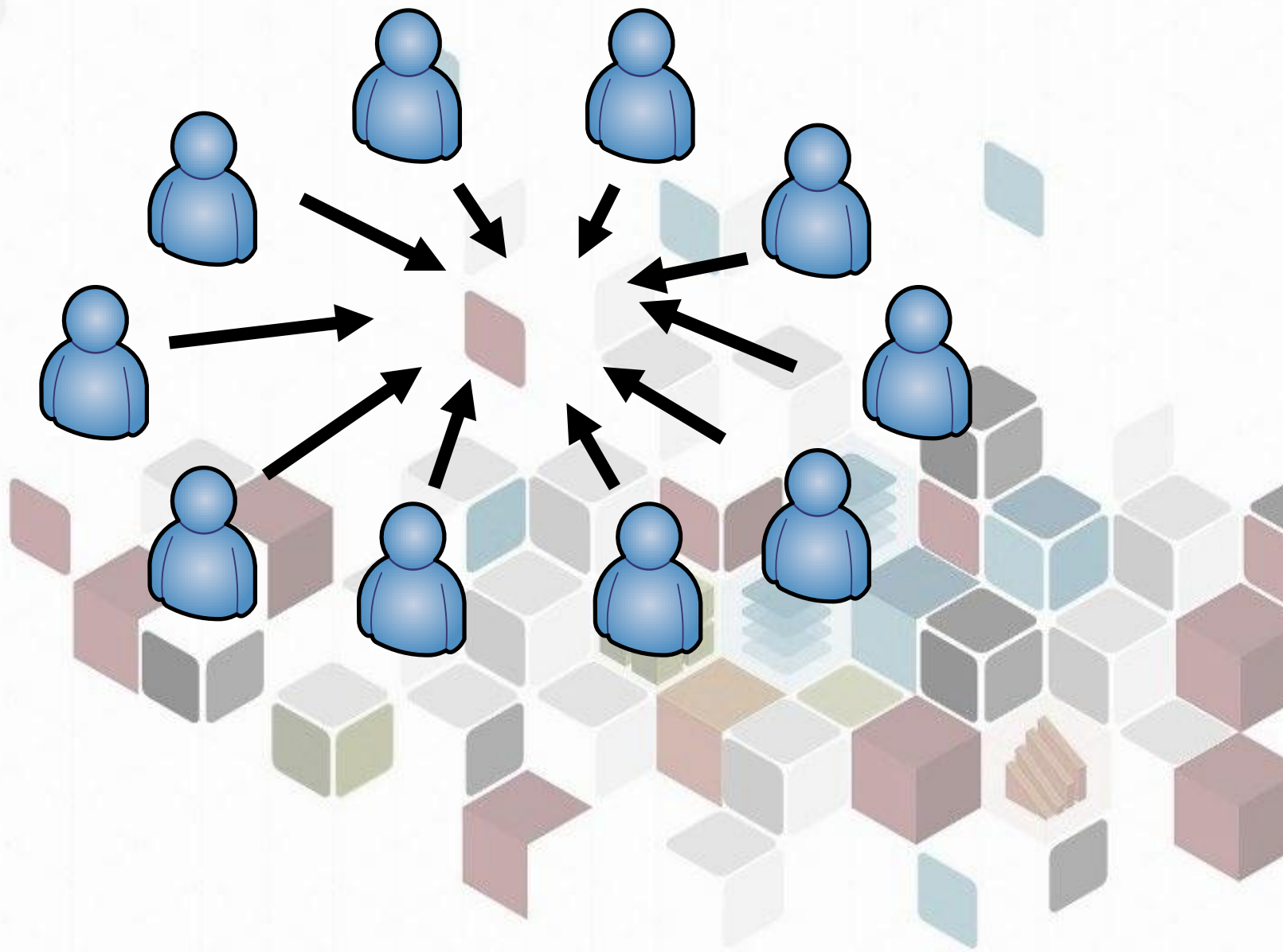
Grandes de Bases de Datos

Pre-procesamiento de datos

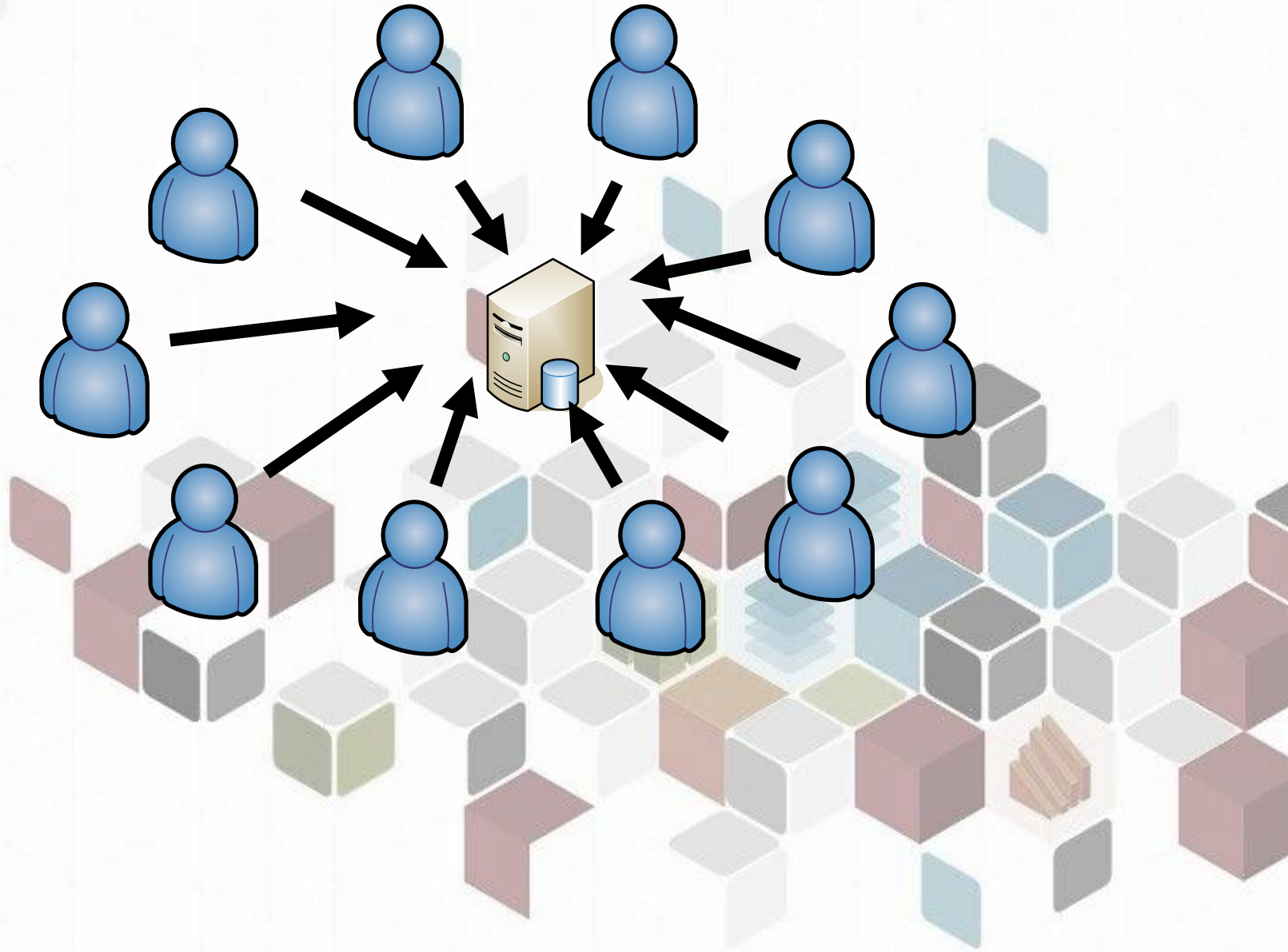
Introducción

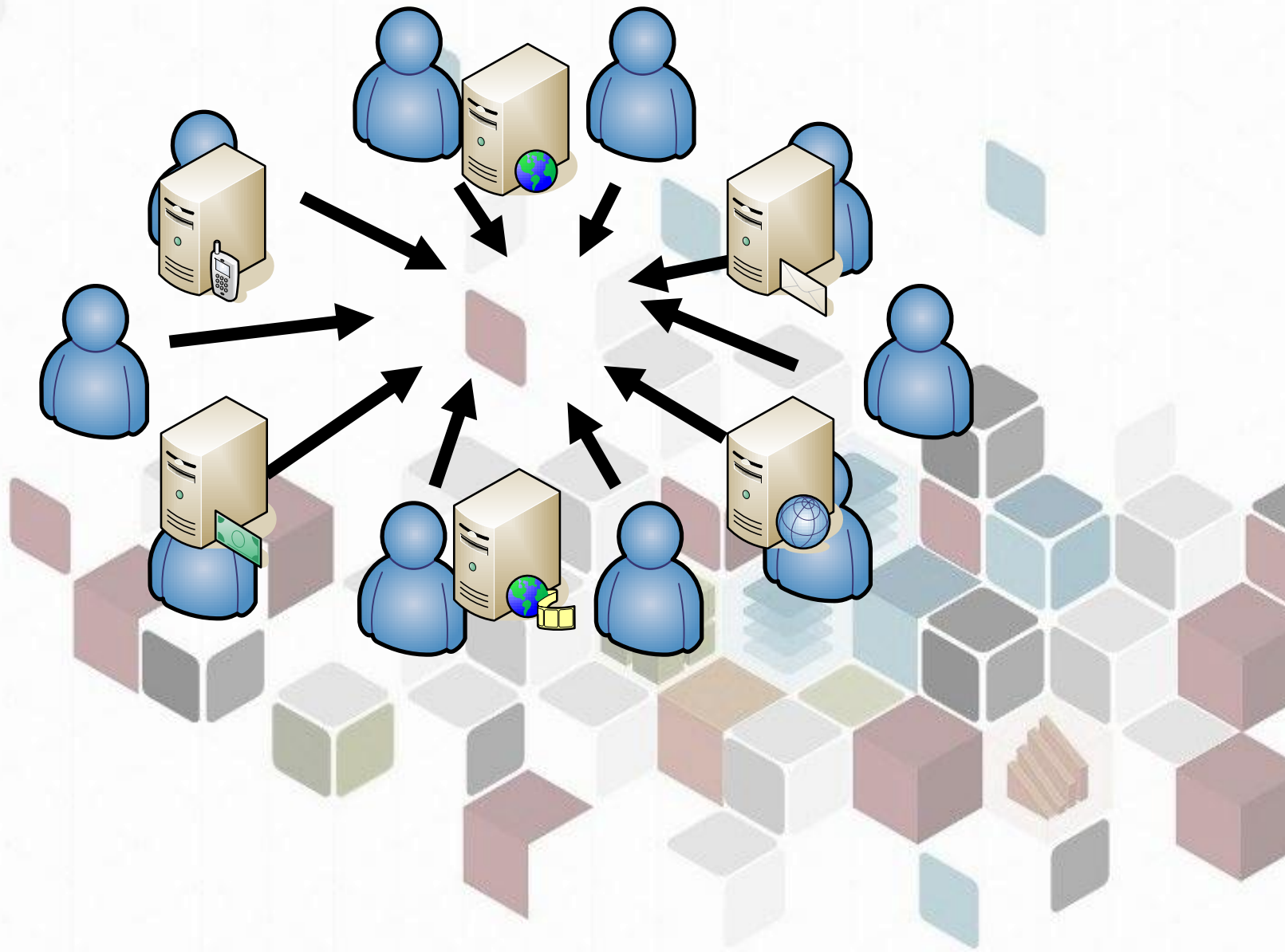


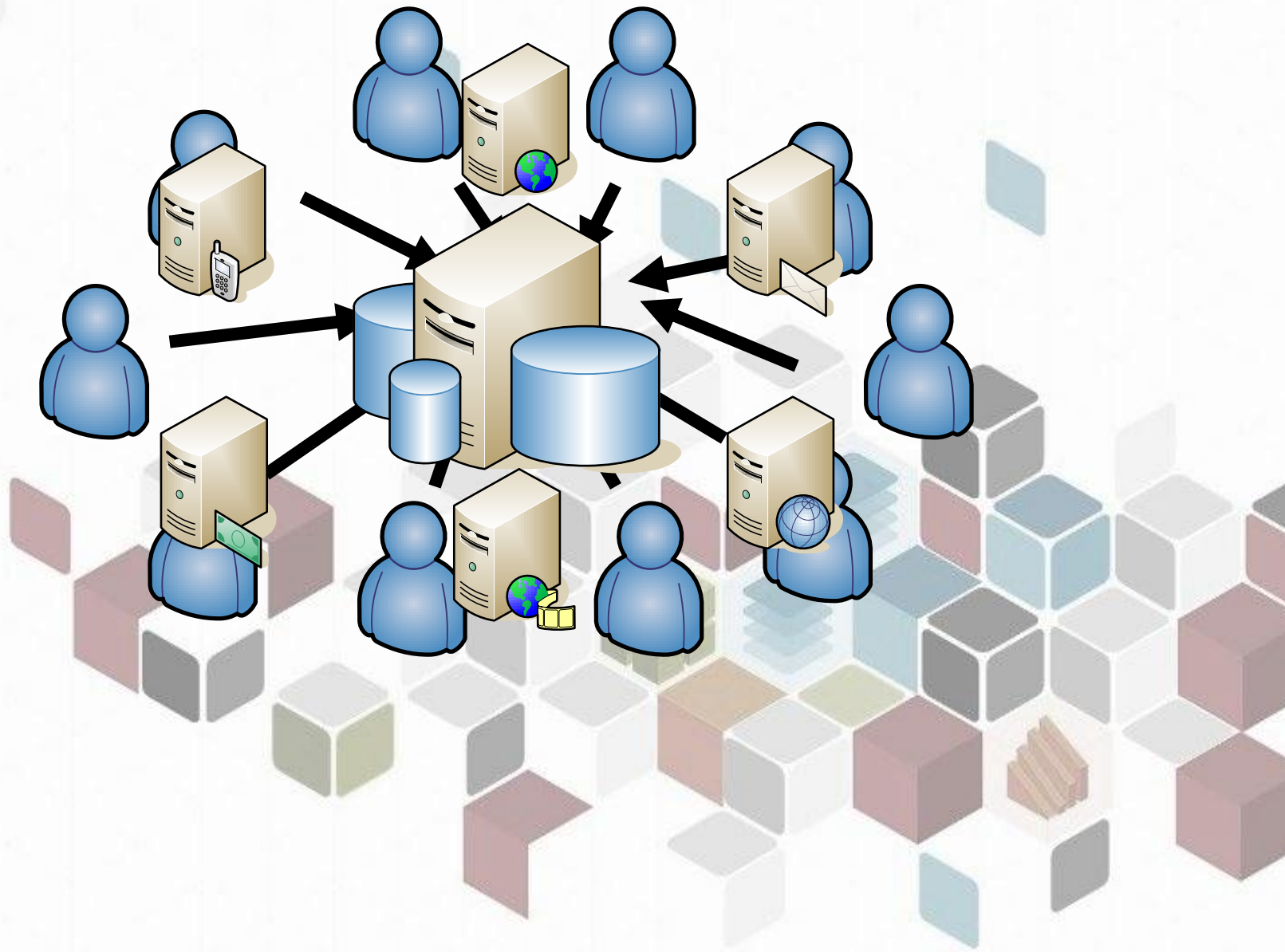
Contexto



Contexto

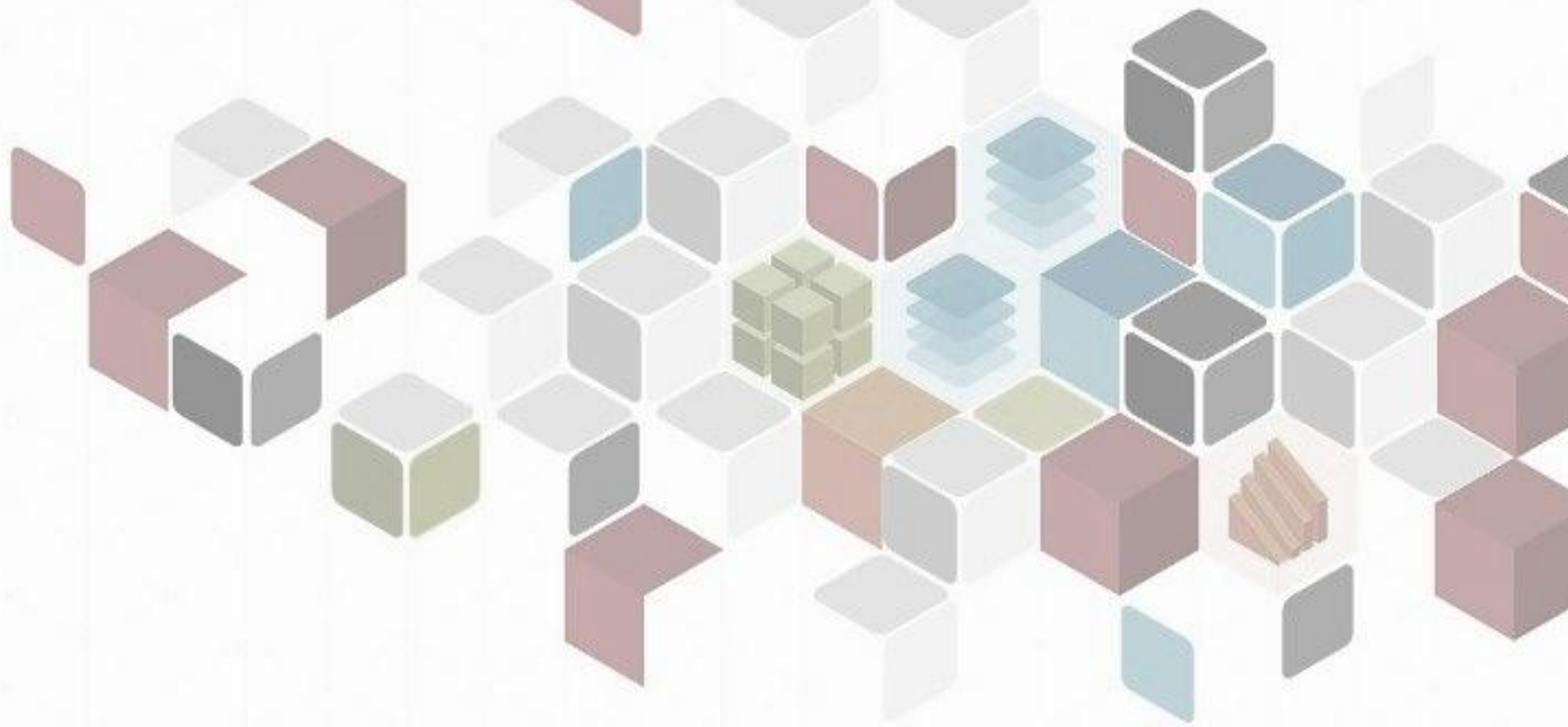






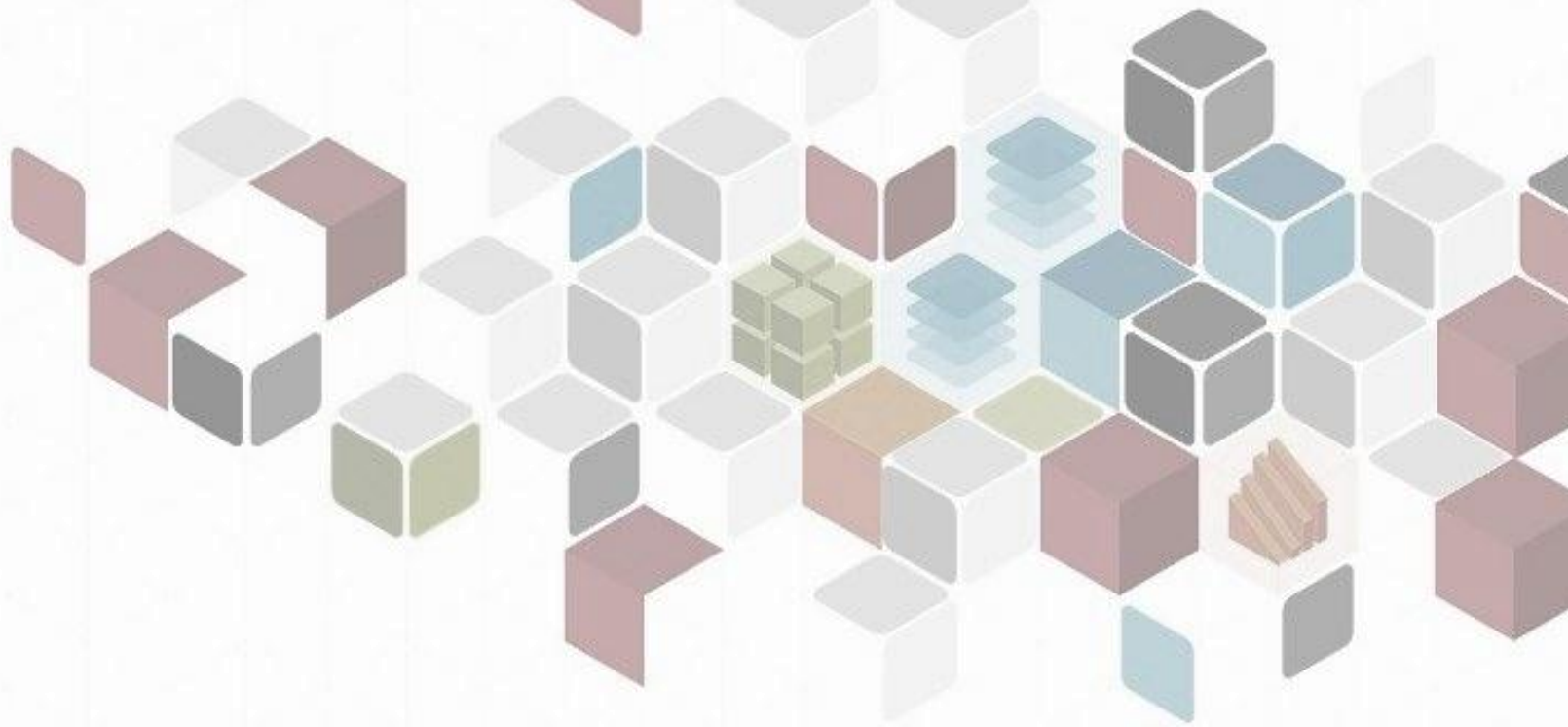
Pre-procesamiento de datos

“Datos con la mayor calidad posible”



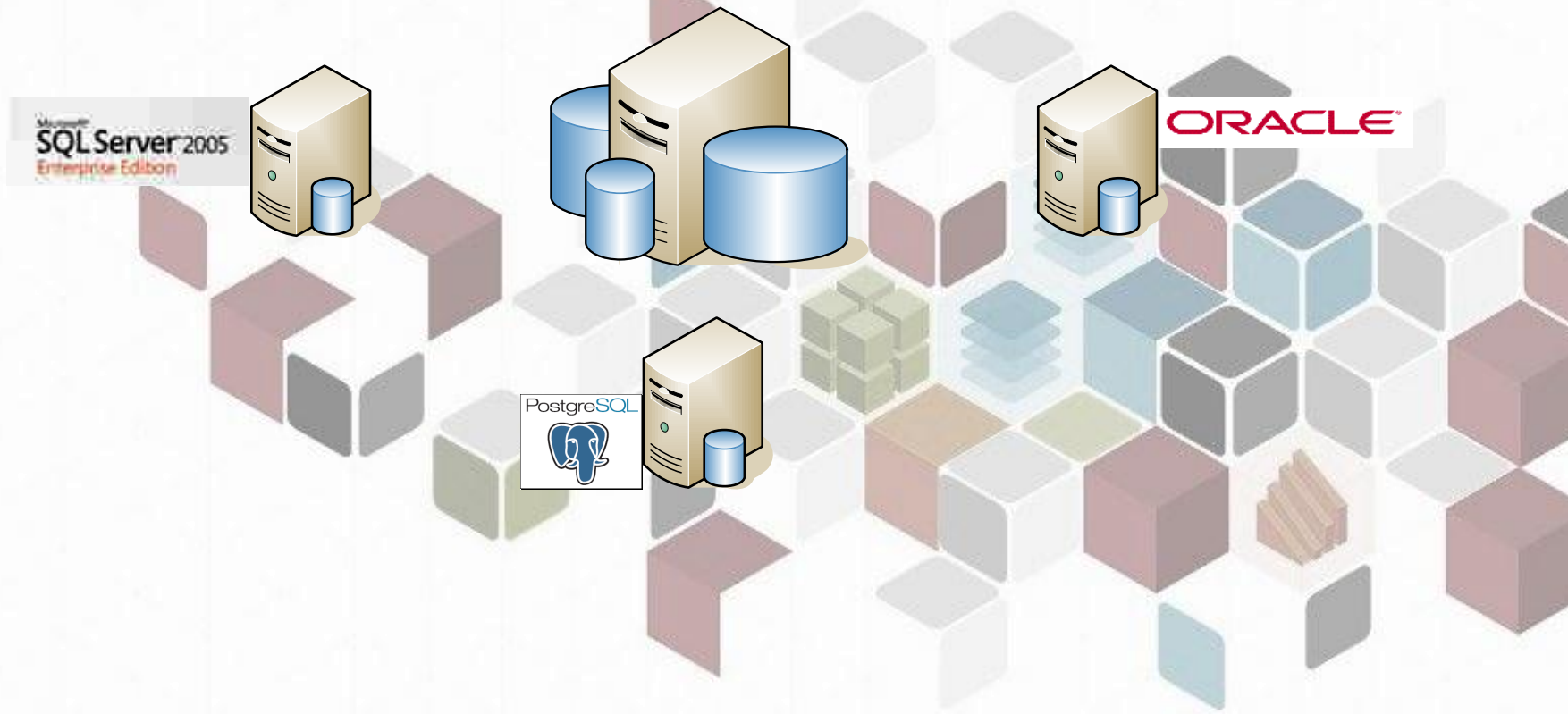
Pre-procesamiento de datos

¿Es posible tener “datos malos”?



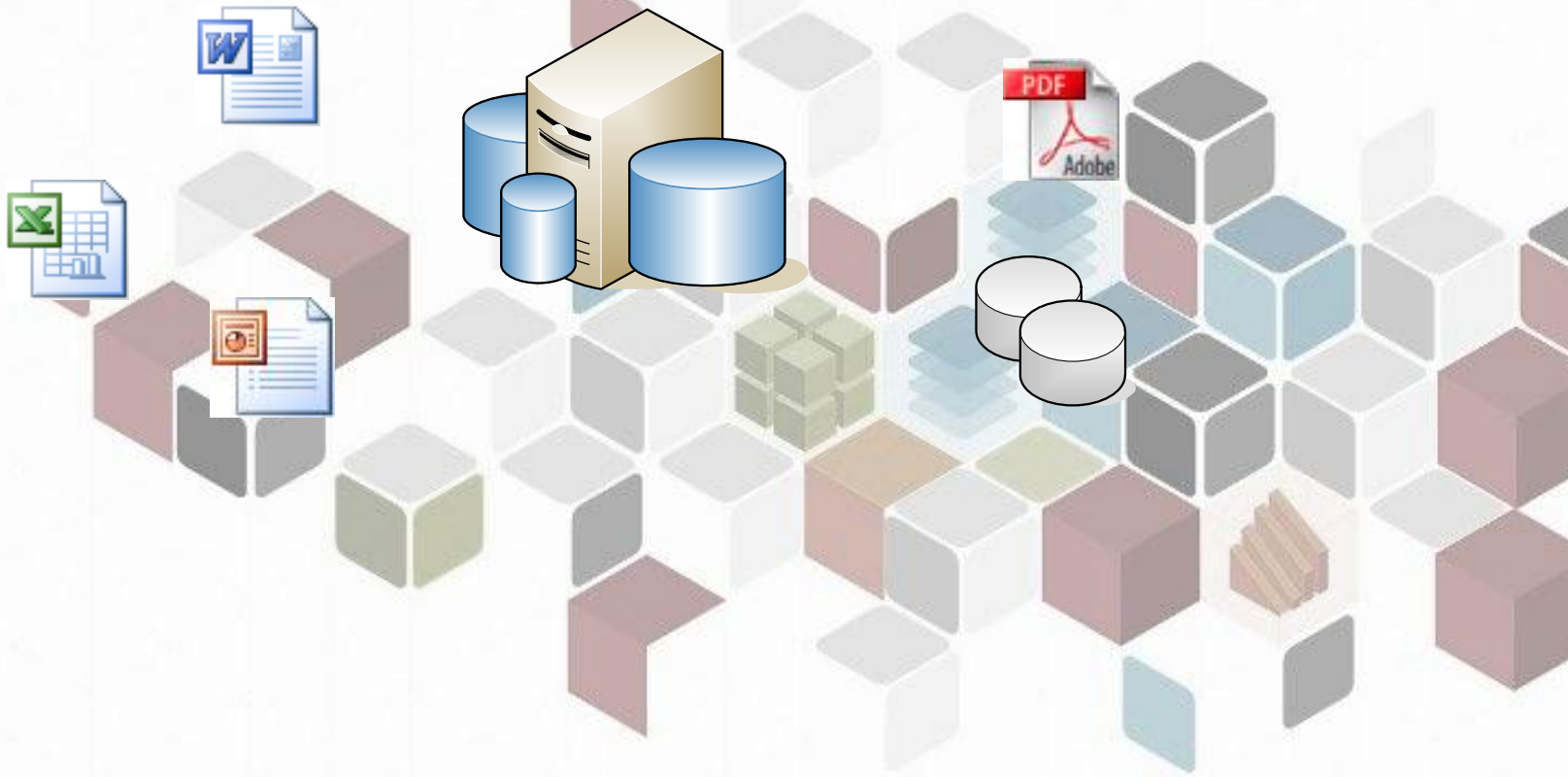
Pre-procesamiento de datos

- Los datos provienen de múltiples fuentes.



Pre-procesamiento de datos

- No siempre se utilizan SMDB para almacenar datos.



Pre-procesamiento de datos

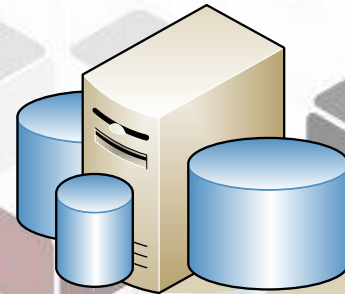
- La transmisión de datos no siempre es 100% confiable.



Pre-procesamiento de datos

- Los datos pueden parecer incompletos.

Nombre	Apellido	Telefono
Rebeca	Uno	55660033
Ricardo	Dos	44552211
Rene	Tres	



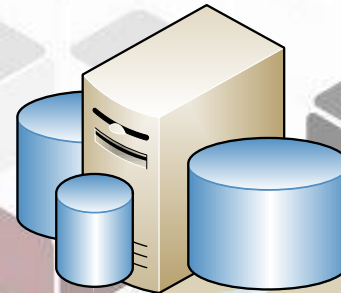
Pre-procesamiento de datos

- El diseño inicial de la BD no consideraba ciertos elementos.

Nombre	Apellido	Telefono
Rebeca	Uno	55660033
Ricardo	Dos	44552211
Rene	Tres	

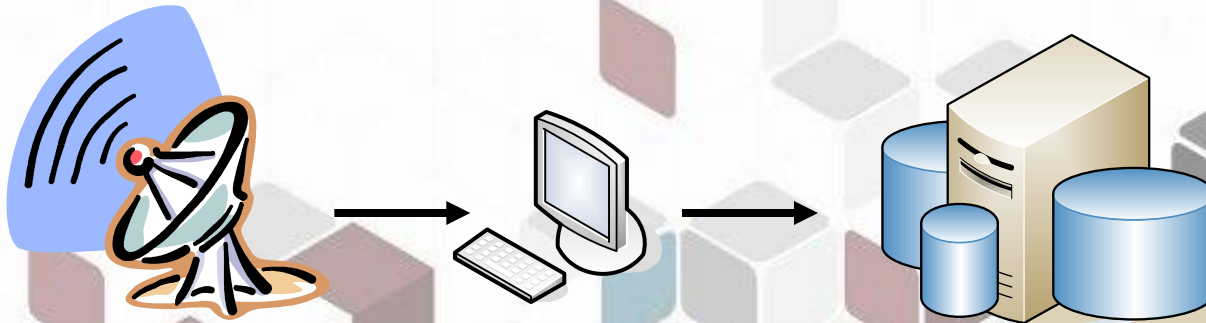


Nombre	Apellido	Telefono	Dependientes	Estado Civil
Rebeca	Uno	55660033	0	
Ricardo	Dos	44552211	0	
Rene	Tres		20	
Hugo	Cuatro	56782345	1	Soltero
Paco	Cinco	55909010	2	Casado
Luis	Seis	90231244	0	Viudo



Pre-procesamiento de datos

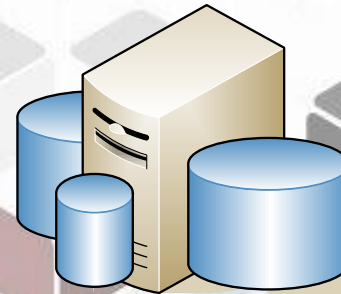
- Los medios por los cuales se obtienen los datos, no son 100% fiables.



Datos malos...

- Incompletos.

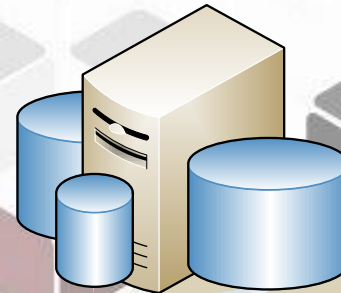
Nombre	Apellido	Telefono	Dependientes	Estado Civil
Rebeca	Uno	55660033	0	
Ricardo	Dos	44552211	0	
Rene	Tres		20	
Hugo	Cuatro	56782345	1	Soltero
Paco	Cinco	55909010	2	Casado
Luis	Seis	90231244	0	Viudo



Datos malos...

- “Ruidosos”

Nombre	Apellido	Telefono	Dependientes	Estado Civil
Rebeca	Uno	55660033	0	
Ricardo	Dos	44552211	0	
Rene	Tres		20	
Hugo	Cuatro	56782345	1	Soltero
Paco	Cinco	55909010	2	Casado
Luis	Seis	90231244	0	Viudo

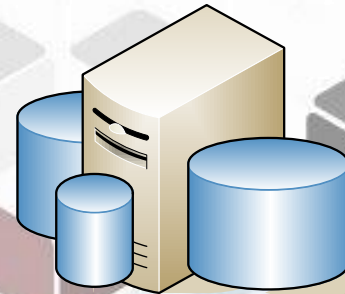


Datos malos...

- Inconsistentes.

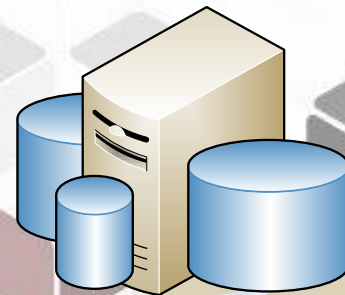
Nombre	Sexo
Rebeca	M
Ricardo	H
Rene	H

Nombre	Sexo
Rebeca	Femenino
Ricardo	Maculino
Rene	Masculino



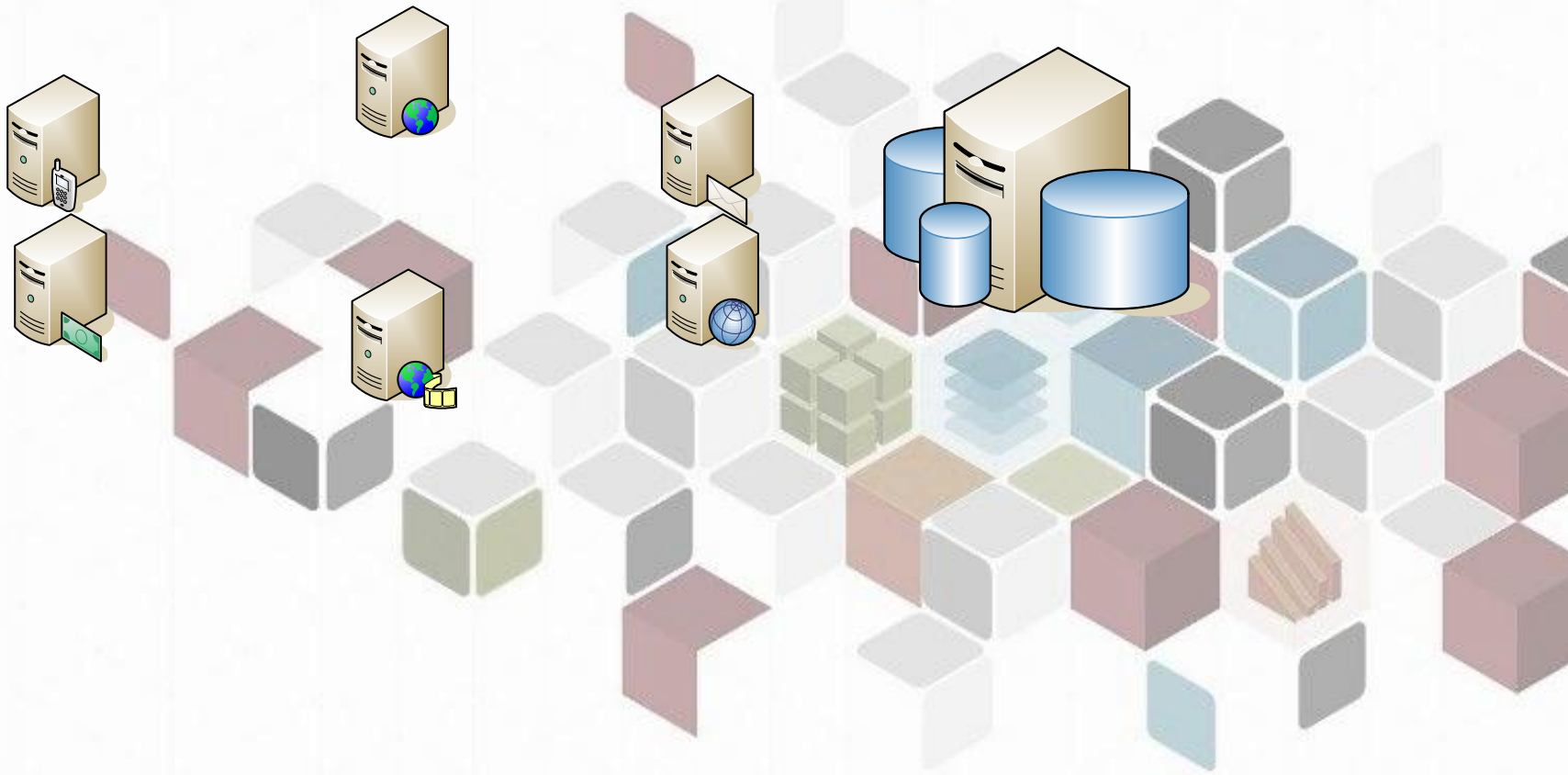
¿Pero por qué ahora?

- La cantidad de datos producida.



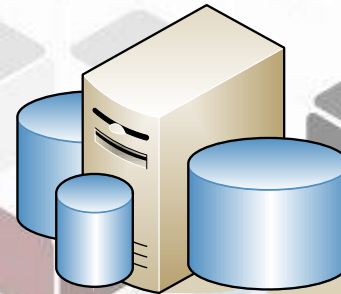
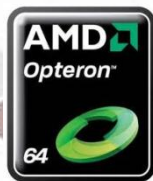
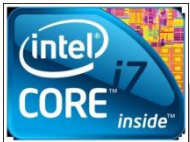
¿Pero por qué ahora?

- Los datos están integrados (DW).



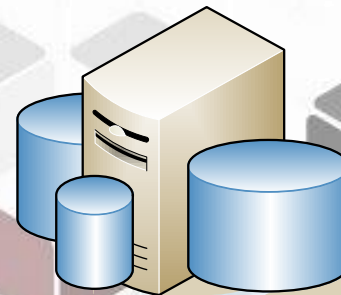
¿Pero por qué ahora?

- La potencia de las computadoras.



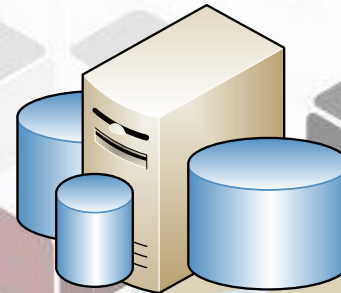
¿Pero por qué ahora?

- Fuerte presión de la competencia.



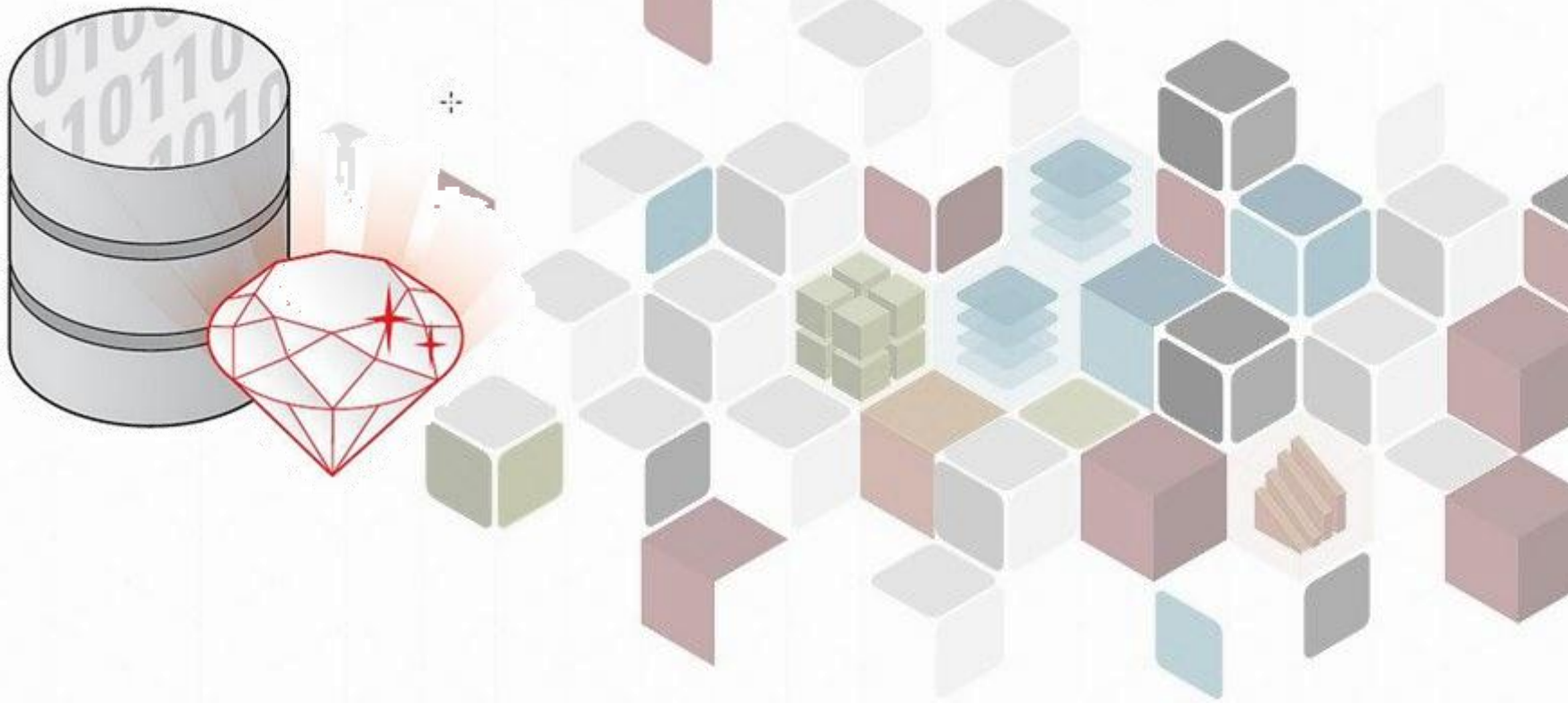
¿Pero por qué ahora?

- Software de minería de datos ha hecho que ahora se vuelva a hablar de él.



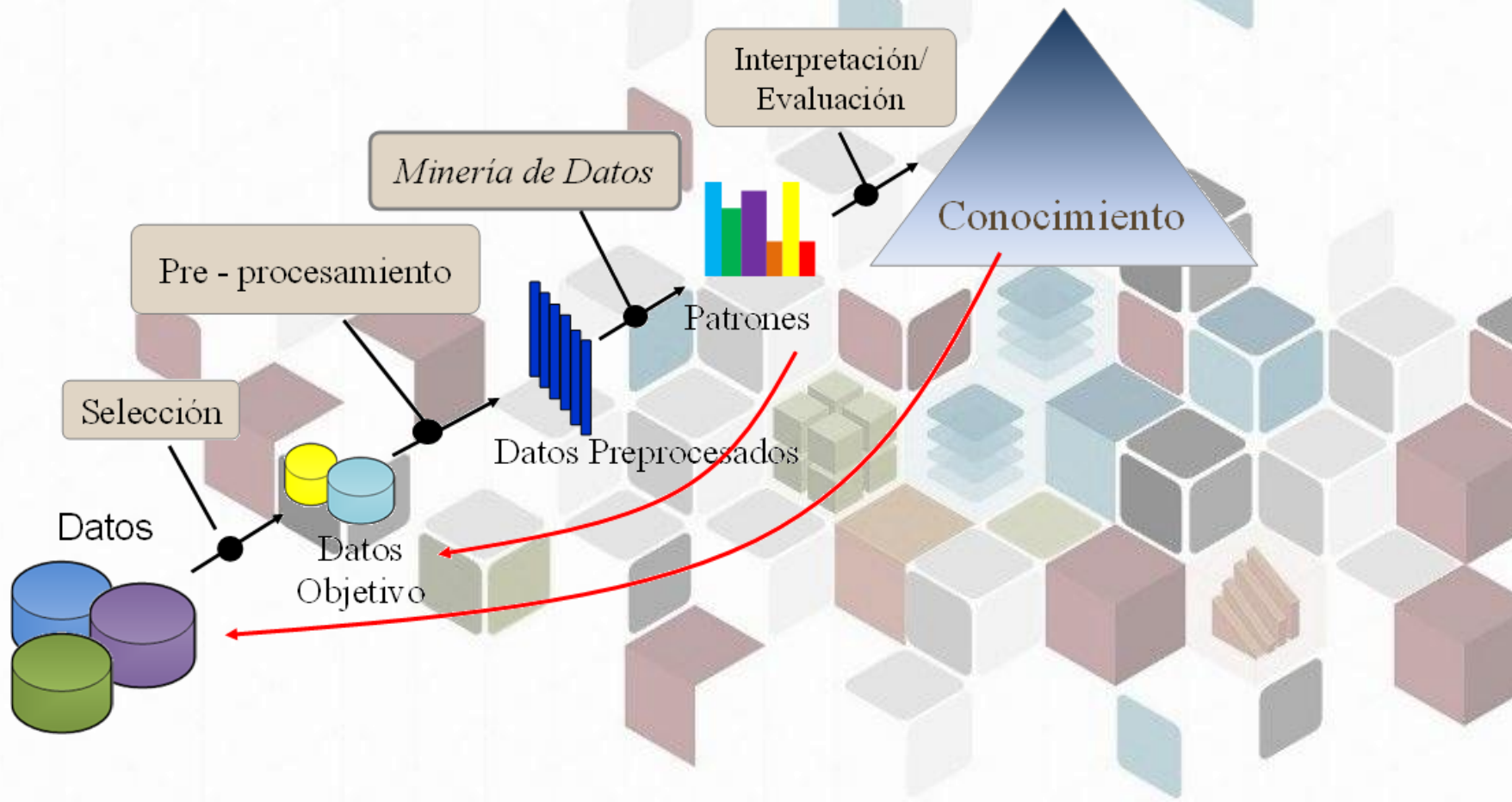
¿Por qué pre-procesar?

- La calidad de los datos es un elemento fundamental.



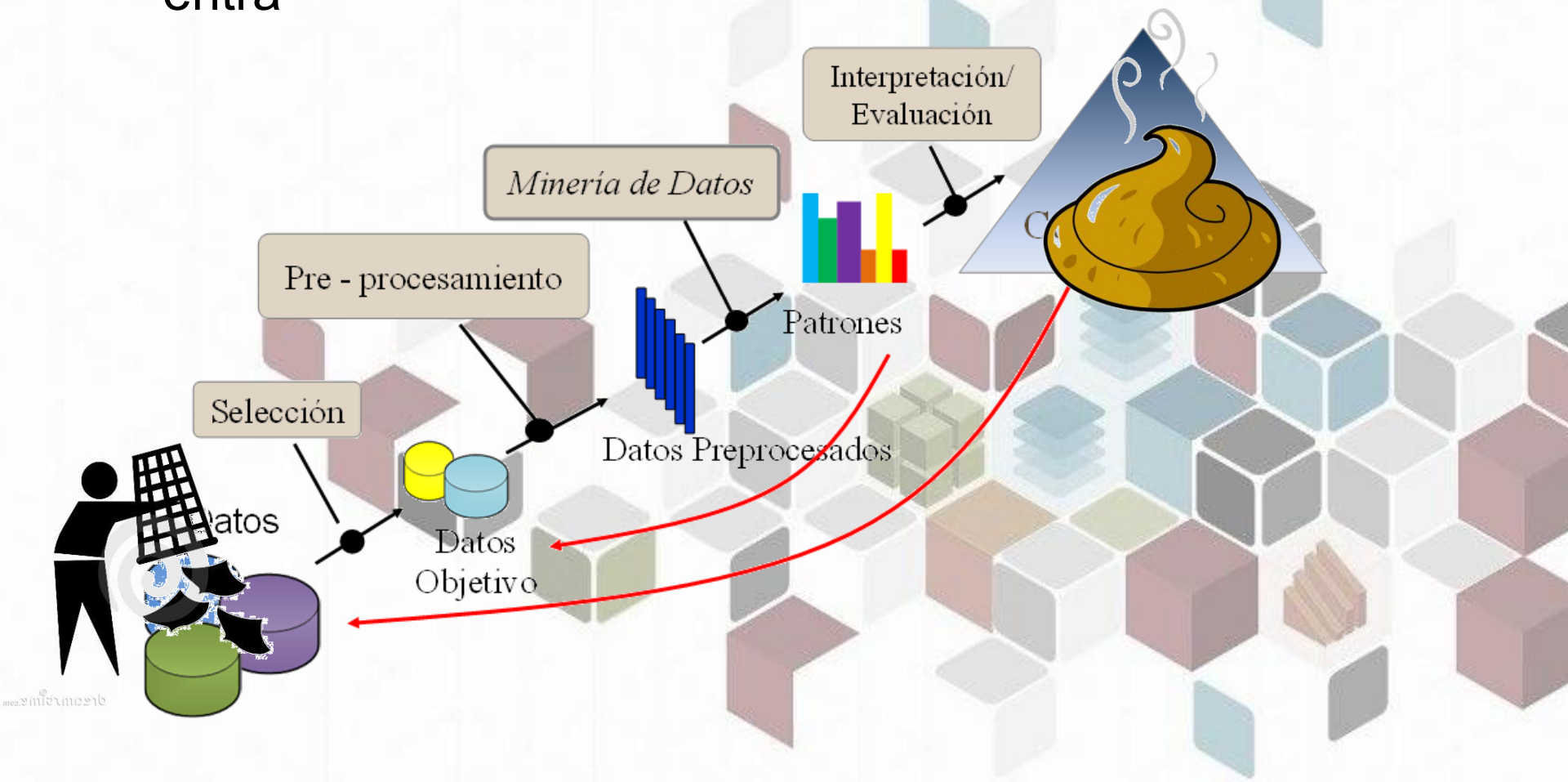
¿Por qué pre-procesar?

- La precisión de lo que “sale” depende de lo que “entra”



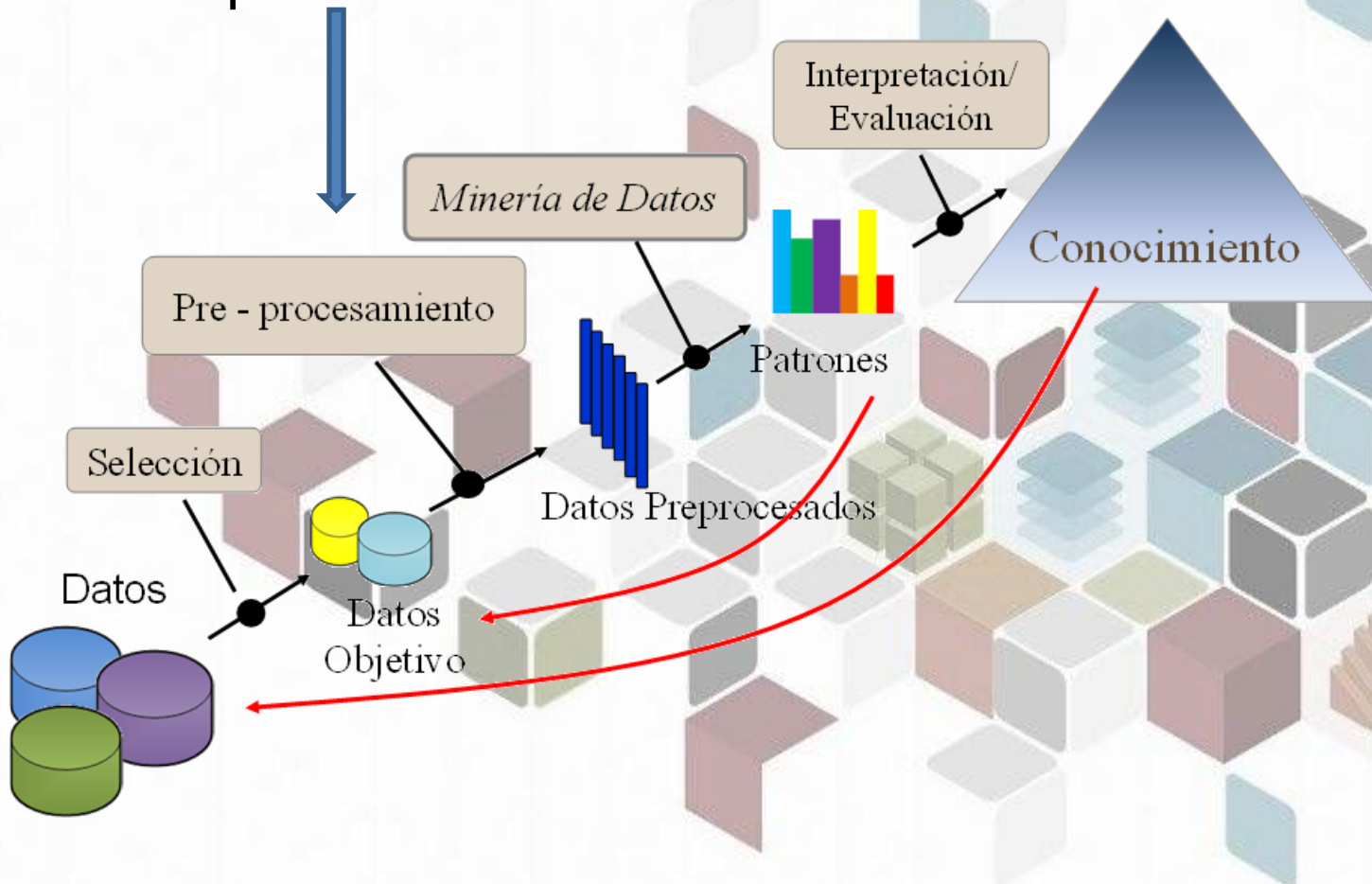
¿Por qué pre-procesar?

- La precisión de lo que “sale” depende de lo que “entra”

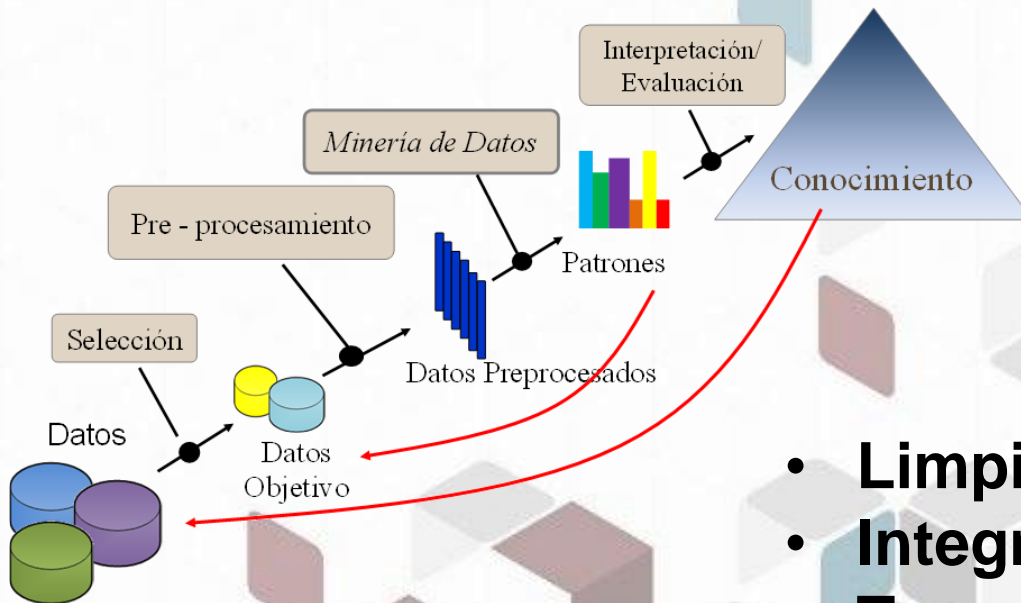


¿Por qué pre-procesar?

- 80% del trabajo realizado en MD, se realiza en esta etapa.



Formas de preprocesamiento



- **Limpieza de datos**
- **Integración de datos**
- **Transformación de datos**
- **Reducción de datos**

Resumen descriptivo de datos

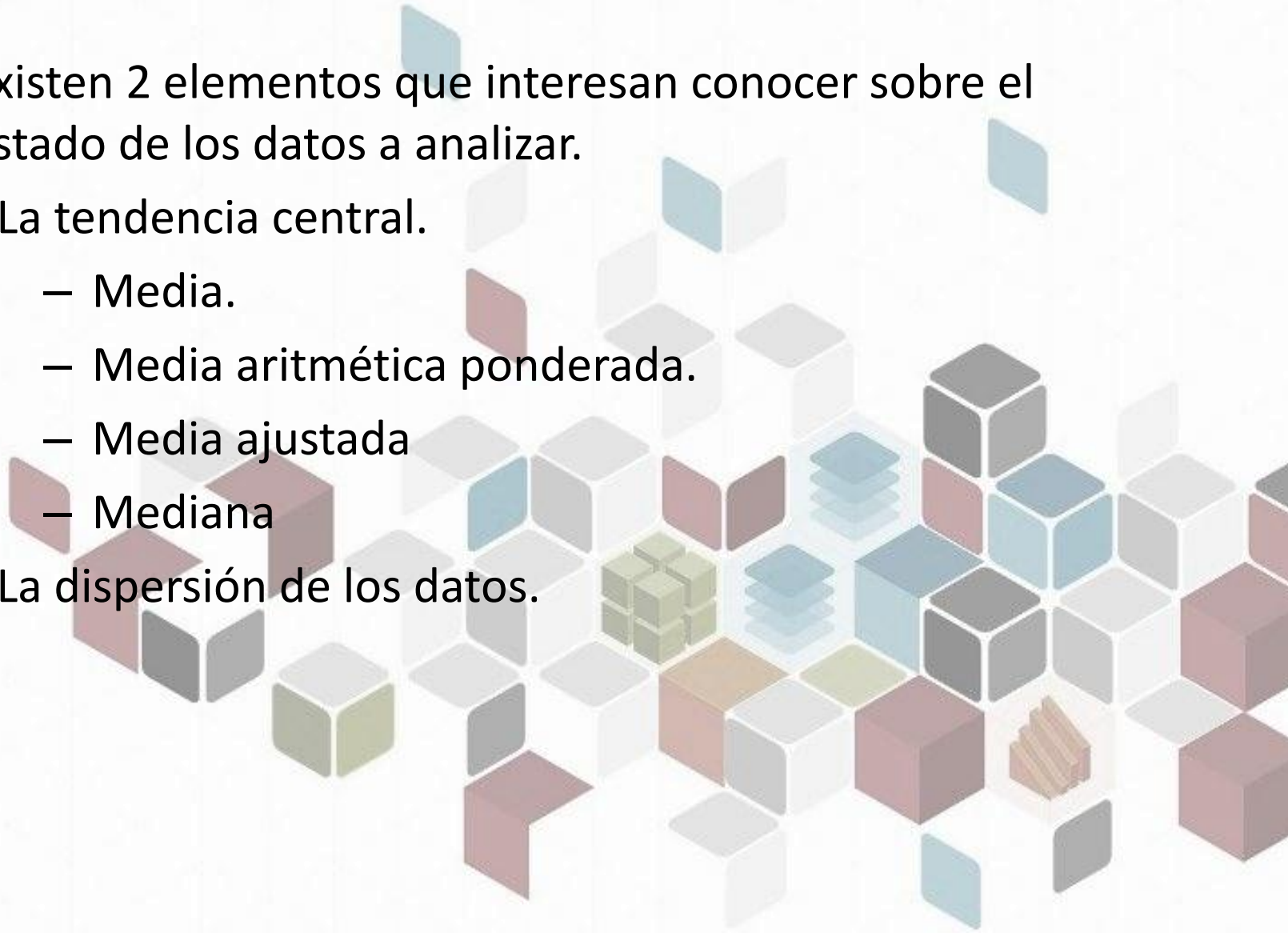
Las métricas se pueden clasificar según sus propiedades en:

- Distributiva
 - Se puede calcular “por partes”
- Algebraica
 - Se puede calcular aplicando funciones algebraicas a métricas distributivas
- Holística
 - Se calculan tomando el total de elementos

Resumen descriptivo de datos

Existen 2 elementos que interesan conocer sobre el estado de los datos a analizar.

- La tendencia central.
 - Media.
 - Media aritmética ponderada.
 - Media ajustada
 - Mediana
- La dispersión de los datos.



Resumen descriptivo de datos

Media:

Sean x_1, x_2, \dots, x_N un conjunto de N valores u observaciones para un atributo, la media es:

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \dots + x_n}{N}$$

¿Su equivalente en SQL es?

Resumen descriptivo de datos

Media aritmética ponderada:

Cada valor x_i en un conjunto puede ser asociado con un peso w_i , para $i = 1, \dots, N$. Los pesos reflejan el significado, importancia o frecuencia de ocurrencia unido a su valor respectivo.

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{\sum_{i=1}^N w_i} = \frac{w_1 x_1 + w_2 x_2 + \dots + w_n x_n}{w_1 + w_2 + \dots + w_n}$$

¿Qué tan representativa es?

Resumen descriptivo de datos

Media ajustada:

Es la medida obtenida al quitar a la media los valores de los extremos (mayor y menor).

$$\bar{x} = \frac{\left(\sum_{i=1}^N x_i \right) - x_l - x_j}{N - 2} = \frac{x_1 + x_2 + \cdots + x_n - x_l - x_j}{N - 2}$$

$$\wedge \begin{matrix} x_l \geq x_i & \forall x_i \in x \\ x_j \leq x_i & \forall x_i \in x \end{matrix}$$

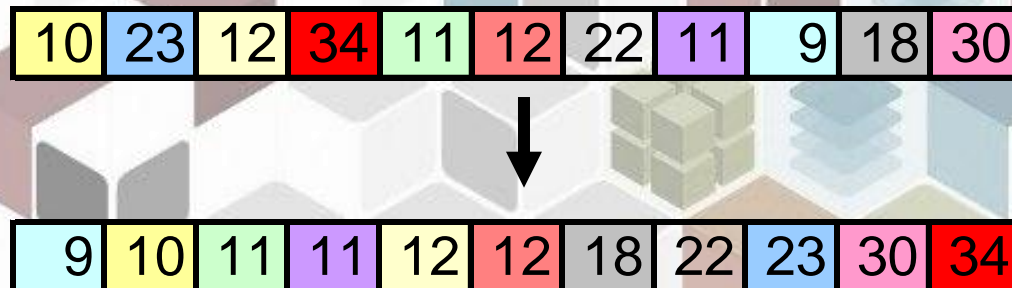
$$\wedge \begin{matrix} x_l \geq x_i & \forall x_i \in x \\ x_j \leq x_i & \forall x_i \in x \end{matrix}$$

¿Cuándo se utiliza? ¿Por qué?

Resumen descriptivo de datos

Mediana:

Sea un conjunto de valores dado de N valores distintos en orden numérico. Si N es impar, entonces la mediana es el valor que está a la mitad del conjunto ordenado; en otro caso, la mediana es el promedio de los dos valores que están en medio del conjunto.



¿Es fácil calcularla? ¿Por qué?

Resumen descriptivo de datos

Sea el intervalo que contiene la frecuencia media el intervalo mediano. Podemos aproximar la mediana del conjunto entero de datos por interpolación usando la formula:

$$mediana = L_1 + \left(\frac{N/2 - (\sum freq)l}{freq_{mediana}} \right) ancho$$

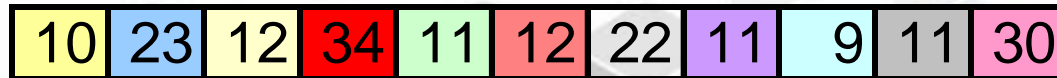
Donde:

- L_1 – Limite inferior del intervalo mediano
- N – Número de valores en el conjunto entero de datos
- $(\sum freq) l$ – Suma de las frecuencias de todos los intervalos que están mas abajo que el intervalo mediano
- **ancho** – es el ancho del intervalo mediano

Resumen descriptivo de datos

Moda:

Es el valor que aparece mas frecuentemente en el conjunto de datos.



Resumen descriptivo de datos

Rango medio:

Promedio del valor más grande y el más pequeño en el conjunto. Esta medida de tendencia central es fácil de calcular utilizando las funciones de agregación de SQL `max()` y `min()`.

10	23	12	34	11	12	22	11	9	11	30
----	----	----	----	----	----	----	----	---	----	----

¿el valor es?

21.50

Resumen descriptivo de datos

Medidas de dispersión de datos.

Rango:

Sean x_1, x_2, \dots, x_N un conjunto de observaciones para algún atributo. El rango del conjunto es la diferencia entre el valor más pequeño y el más grande.



10	23	12	34	11	12	22	11	9	11	30
----	----	----	----	----	----	----	----	---	----	----

¿el valor es?

25

Resumen descriptivo de datos

Rango intercuartil (IQR)

Distancia entre el primer y el tercer cuartil, es una medida que da el rango cubierto por la mitad de los datos.

$$\text{IQR} = Q_3 - Q_1$$

10	23	12	34	11	12	22	11	9	11	30
----	----	----	----	----	----	----	----	---	----	----

¿el valor es?

Resumen descriptivo de datos

¿Qué son los cuartiles?

Ordenamos los datos

- El elemento que denota al 25% es el primer cuartil
- El elemento que denota al 50% es el segundo cuartil
- El elemento que denota al 75% es el tercer cuartil

9	10	11	11	12	12	11	22	23	30	34
---	----	----	----	----	----	----	----	----	----	----

Resumen descriptivo de datos

Rango intercuartil (IQR)

Distancia entre el primer y el tercer cuartil, es una medida que da el rango cubierto por la mitad de los datos.

$$\text{IQR} = Q_3 - Q_1$$

10	23	12	34	11	12	22	11	9	11	30
----	----	----	----	----	----	----	----	---	----	----

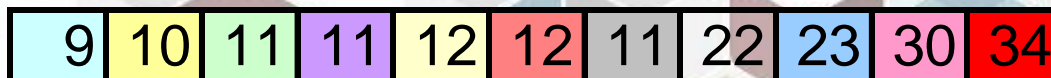
¿el valor es?

Resumen descriptivo de datos

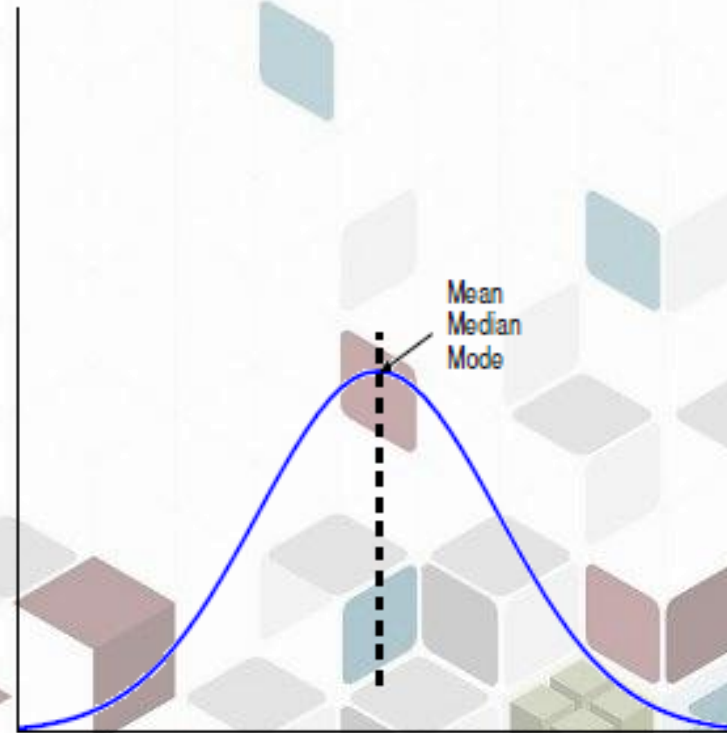
Five-number summary:

Consta de la media, los cuartiles Q1 y Q3, y el valor mas pequeño y el mas grande, escritos en orden quedan:

mínimo, Q1, mediana, Q3, máximo.

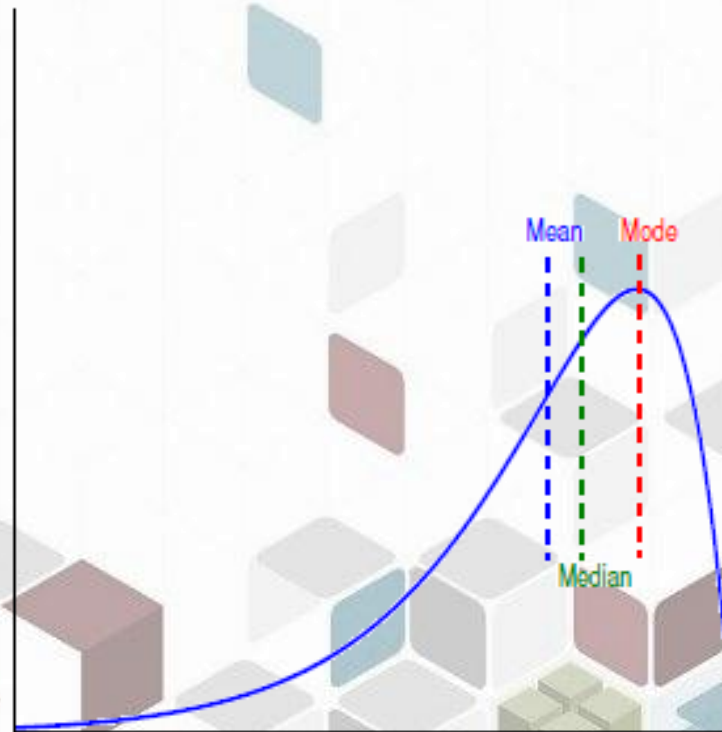


¿Cómo se distribuyen?



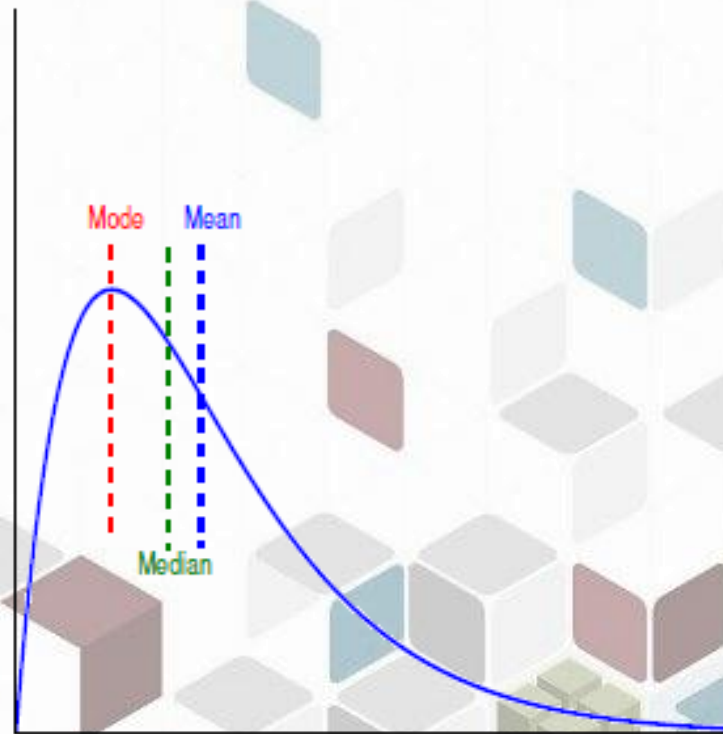
Cuando no hay sesgo

¿Cómo se distribuyen?



Sesgo positivo

¿Cómo se distribuyen?



Sesgo negativo

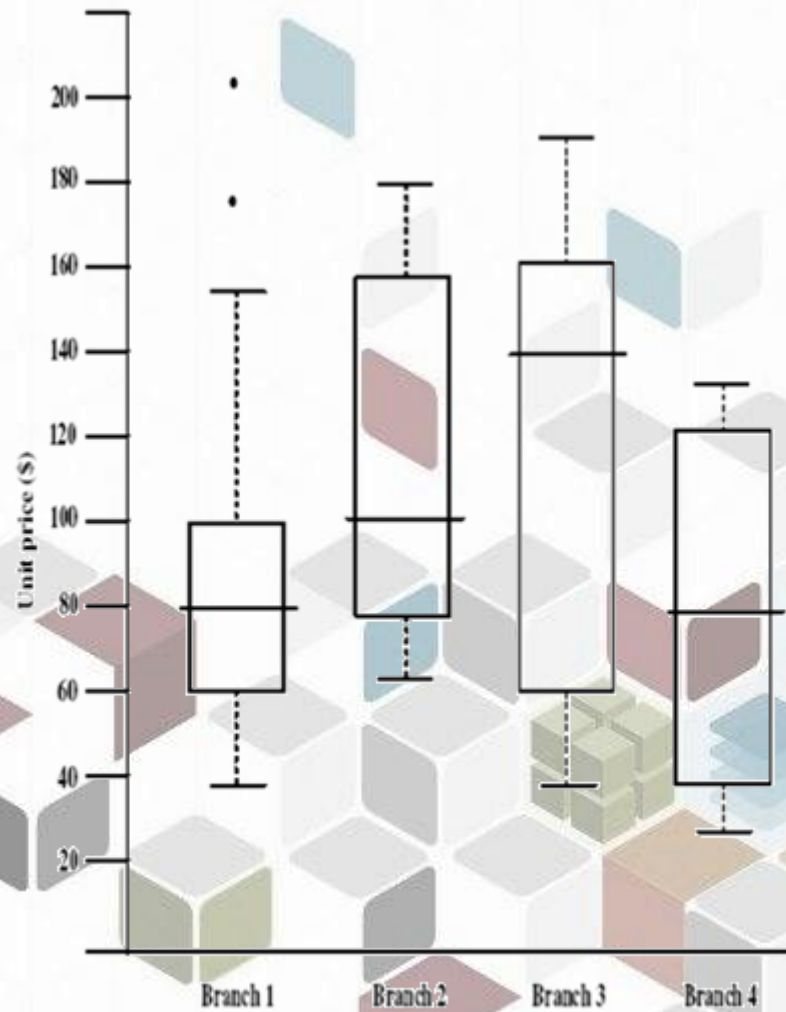
Resumen descriptivo de datos

Boxplots

Representación gráfica del “**five-number summary**” como sigue:

- Los extremos de la caja esta en los cuartiles, así que la longitud de la caja es el rango intercuartil, IQR.
- La mediana esta marcada por una línea dentro de la caja.
- Dos líneas fuera de la caja extendiendo las observaciones mas pequeña y la mas grande.

Resumen descriptivo de datos



Resumen descriptivo de datos

Varianza y desviación estándar:

La varianza de N observaciones, x_1, x_2, \dots, x_N , es:

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 = \frac{1}{N-1} \left[\sum_{i=1}^N x_i^2 - \frac{1}{N} \left(\sum_{i=1}^N x_i \right)^2 \right]$$

Donde \bar{x} es el valor medio de las observaciones. La desviación estándar, s , de las observaciones es el cuadrado de la varianza, s^2 .

La varianza mide, “que tanto varían los datos”, si todos son iguales, es igual a 0

Representaciones gráficas

Información “plana”

Id	Apellido	Meses	Id	Apellido	Meses	Id	Apellido	Meses
1	Washington	94	16	Lincoln	49	30	Coolidge	67
2	Adams	48	17	Johnson	47	31	Hoover	48
3	Jefferson	96	18	Grant	96	32	Roosevelt	146
4	Madison	96	19	Hayes	48	33	Truman	92
5	Monroe	96	20	Garfield	7	34	Eisenhower	96
6	Adams	48	21	Arthur	41	35	Kennedy	34
7	Jackson	96	22	Cleveland	48	36	Johnson	62
8	Van Buren	48	23	Harrison	48	37	Nixon	67
9	Harrison	1	24	Cleveland	48	38	Ford	29
10	Tyler	47	25	McKinley	54	39	Carter	48
11	Polk	48	26	Roosevelt	90	40	Reagan	96
12	Taylor	16	27	Taft	48	41	Bush	48
13	Filmore	32	28	Wilson	96	42	Clinton	96
14	Pierce	48	29	Harding	29	43	Bush	96
15	Buchanan	48						

¿Cómo presentar esta información?

Representaciones gráficas

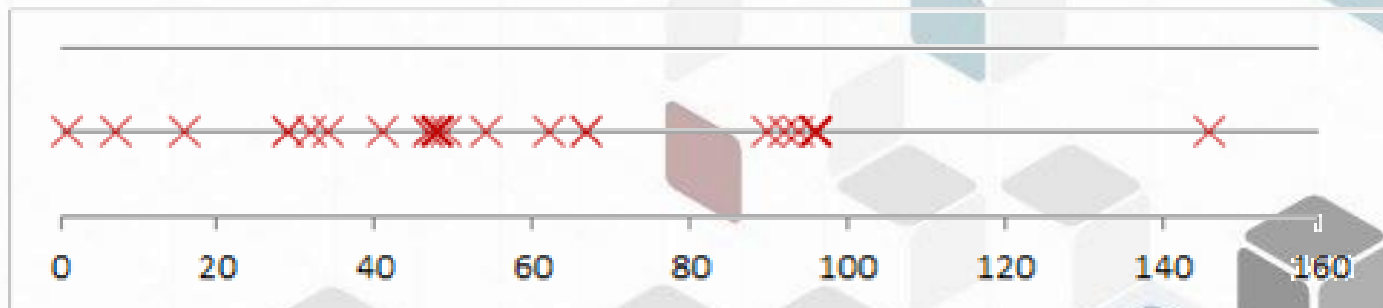
Información “plana”

Id	Apellido	Meses	Id	Apellido	Meses	Id	Apellido	Meses
1	Washington	94	16	Lincoln	49	30	Coolidge	67
2	Adams	48	17	Johnson	47	31	Hoover	48
3	Jefferson	96	18	Grant	96	32	Roosevelt	146
4	Madison	96	19	Hayes	48	33	Truman	92
5	Monroe	96	20	Garfield	7	34	Eisenhower	96
6	Adams	48	21	Arthur	41	35	Kennedy	34
7	Jackson	96	22	Cleveland	48	36	Johnson	62
8	Van Buren	48	23	Harrison	48	37	Nixon	67
9	Harrison	1	24	Cleveland	48	38	Ford	29
10	Tyler	47	25	McKinley	54	39	Carter	48
11	Polk	48	26	Roosevelt	90	40	Reagan	96
12	Taylor	16	27	Taft	48	41	Bush	48
13	Filmore	32	28	Wilson	96	42	Clinton	96
14	Pierce	48	29	Harding	29	43	Bush	96
15	Buchanan	48						

Opción A:
Sobre un eje

Representaciones gráficas

Información “plana”



Opción A:
Sobre un eje

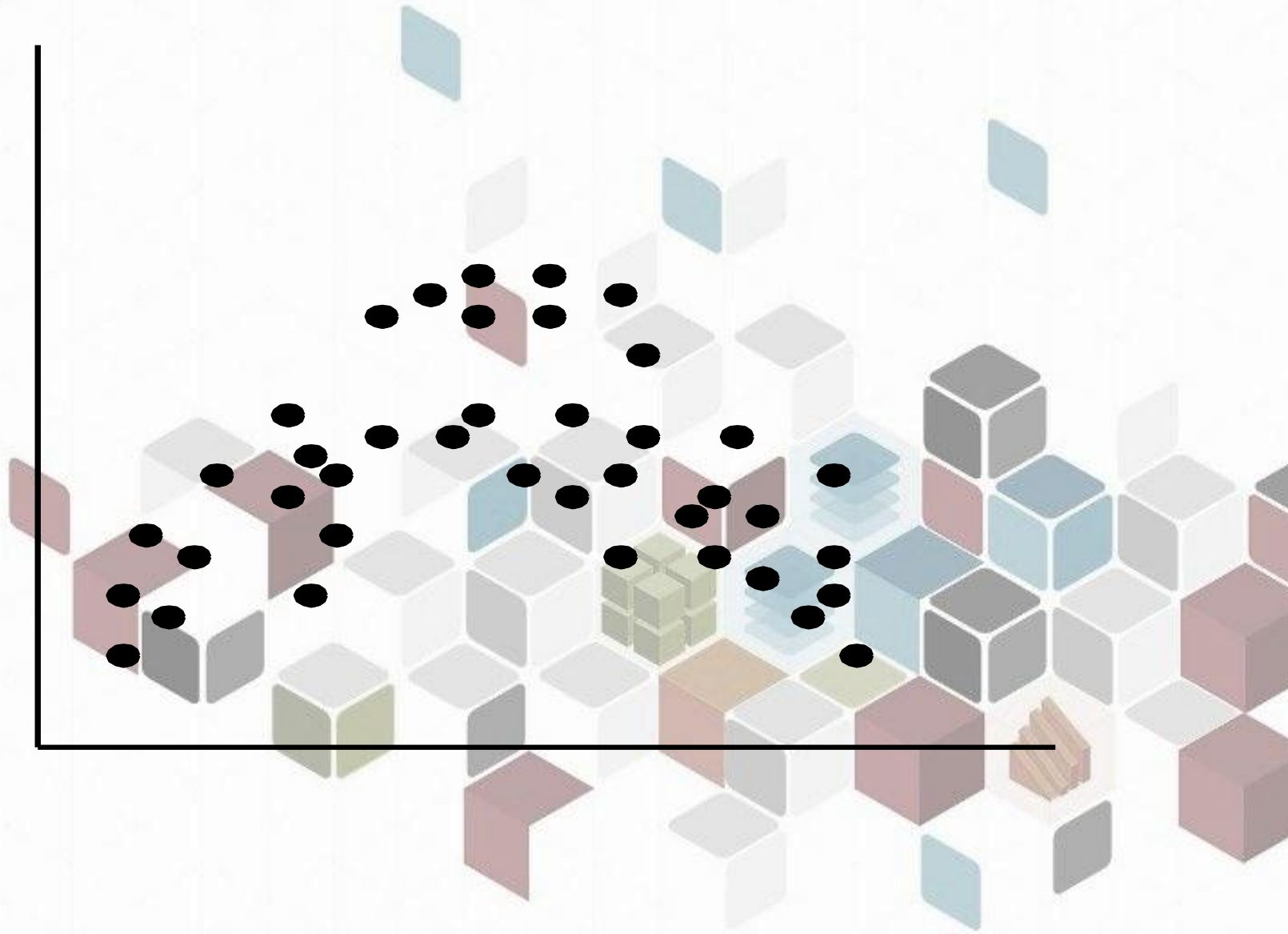
Representaciones gráficas

Scatter plot:

Método efectivo para determinar si hay una relación, patrón o tendencia entre dos atributos numéricos. Para construir un ***scatter plot***, cada par de valores es tratado como un par de coordenadas en un sentido algebraico y graficados como puntos en el plano.

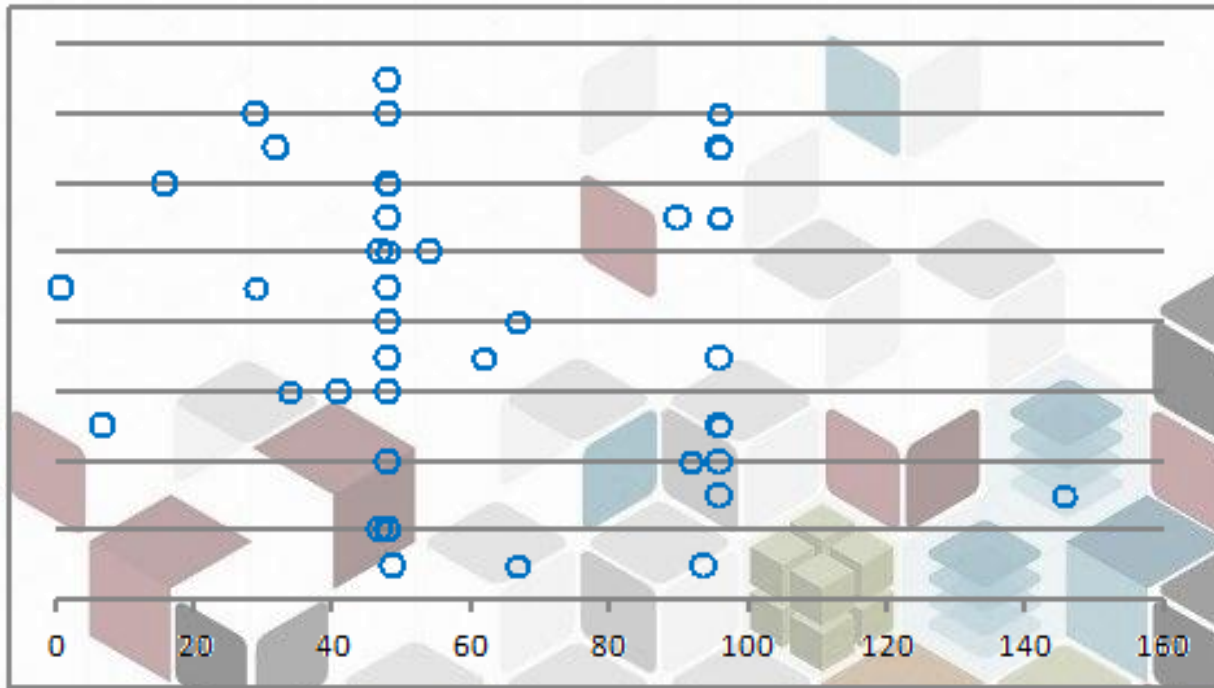


Representaciones gráficas



Representaciones gráficas

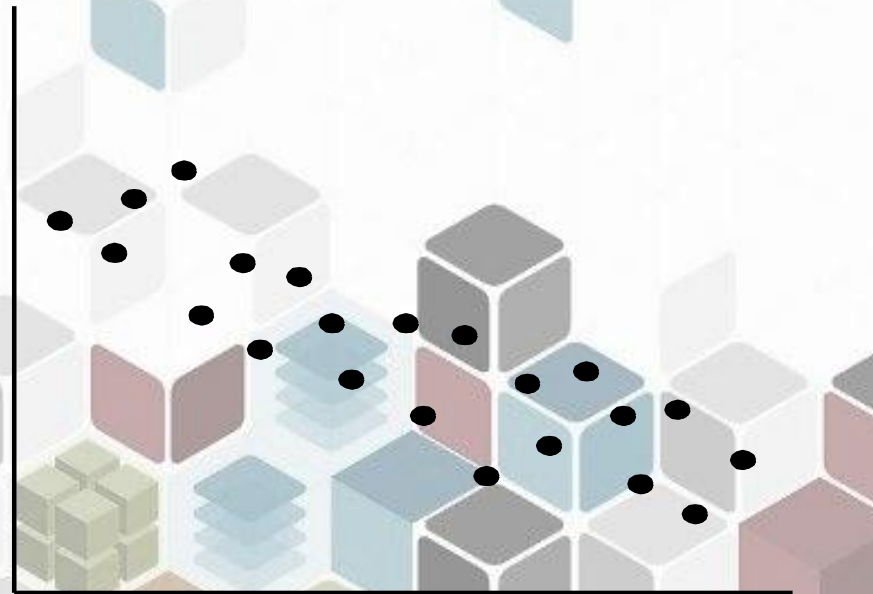
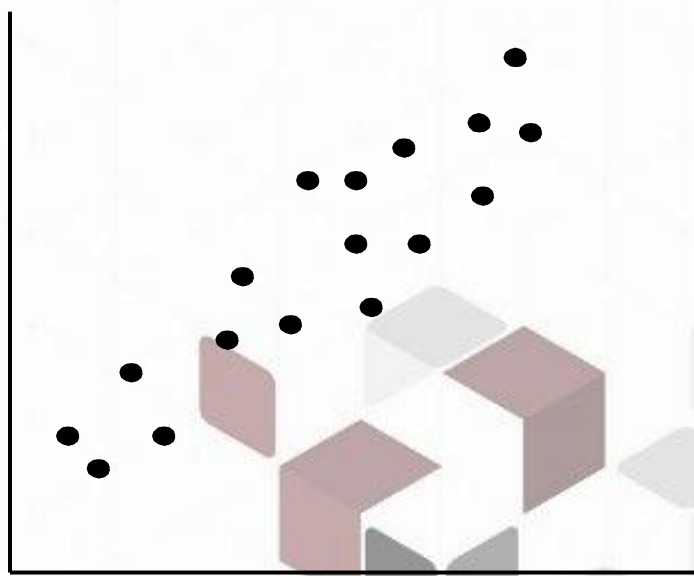
Información “plana”



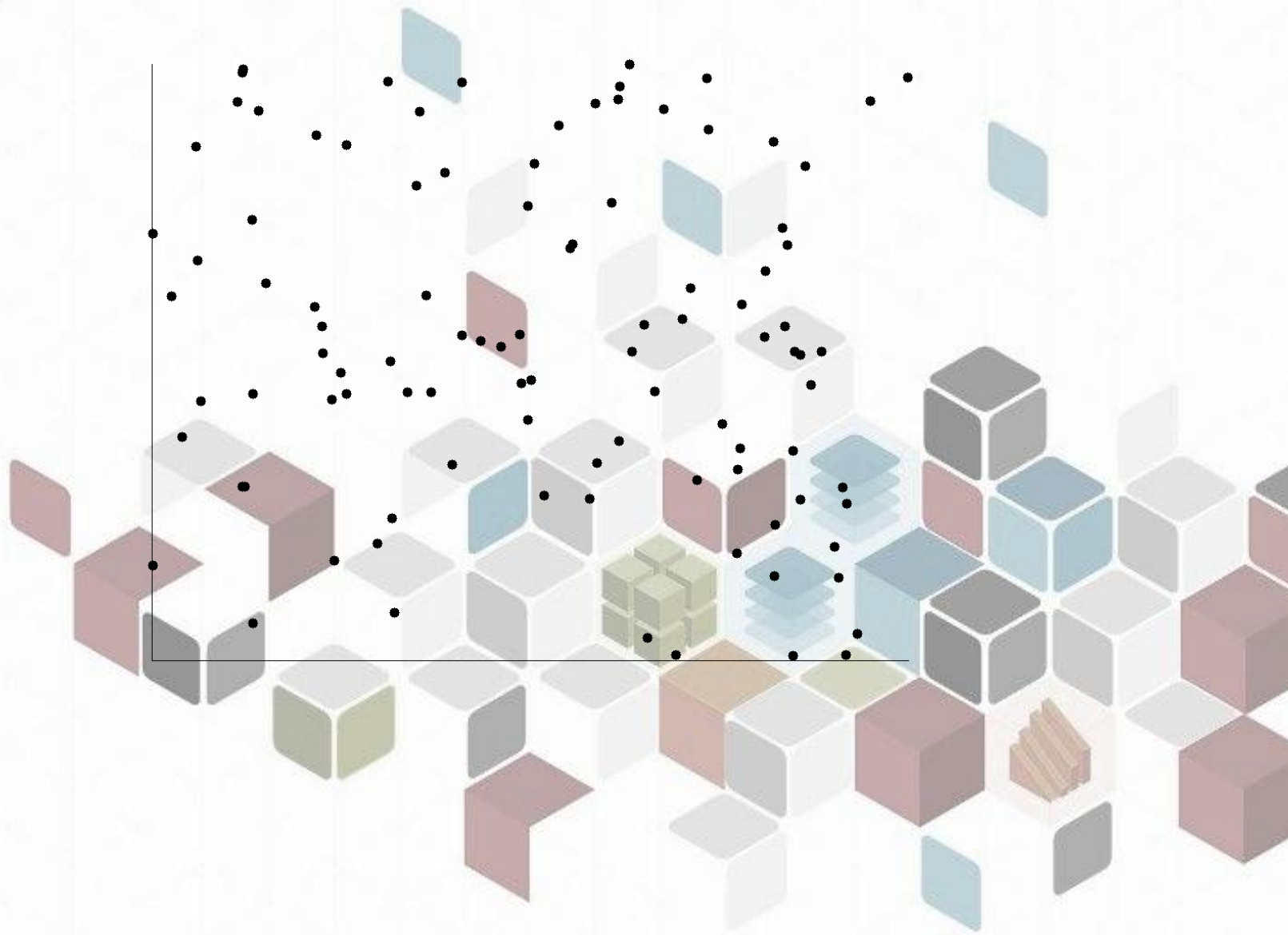
Jitter plot

Opción B:
Scatter plot

Representaciones gráficas



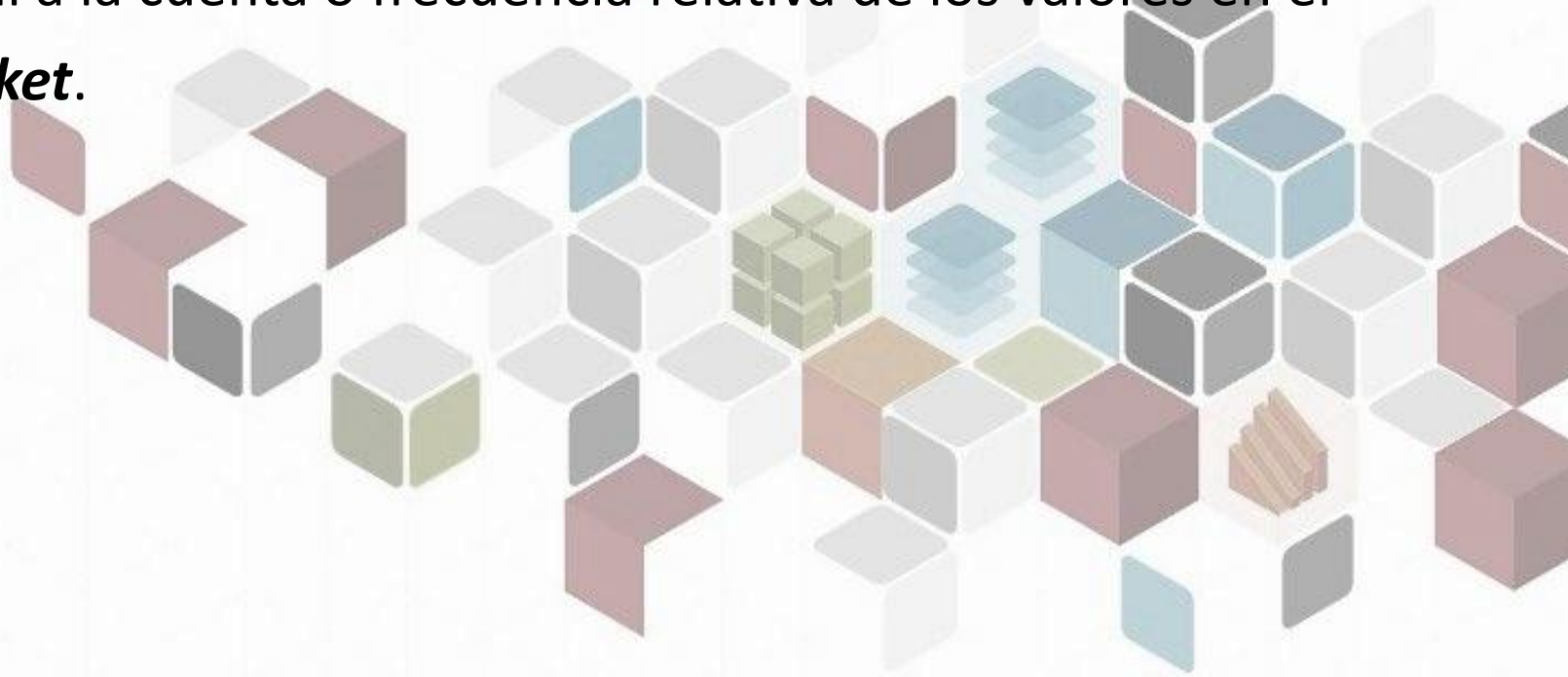
Representaciones gráficas



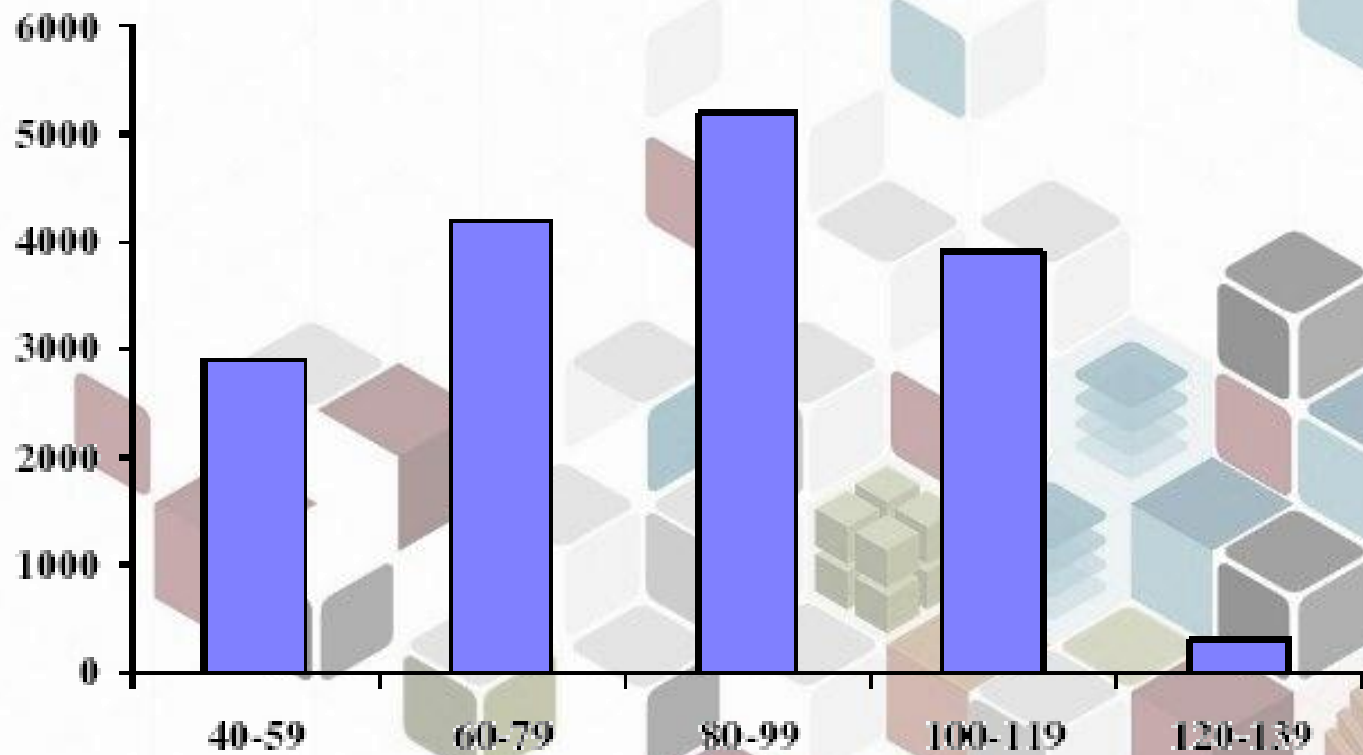
Representaciones gráficas

Histogramas de frecuencia:

Método gráfico para resumir la distribución de un atributo dado. Un histograma para un atributo A divide la distribución de datos de A en subconjuntos ajenos o ***buckets***. Representado por un rectángulo, cuya altura es igual a la cuenta o frecuencia relativa de los valores en el ***bucket***.

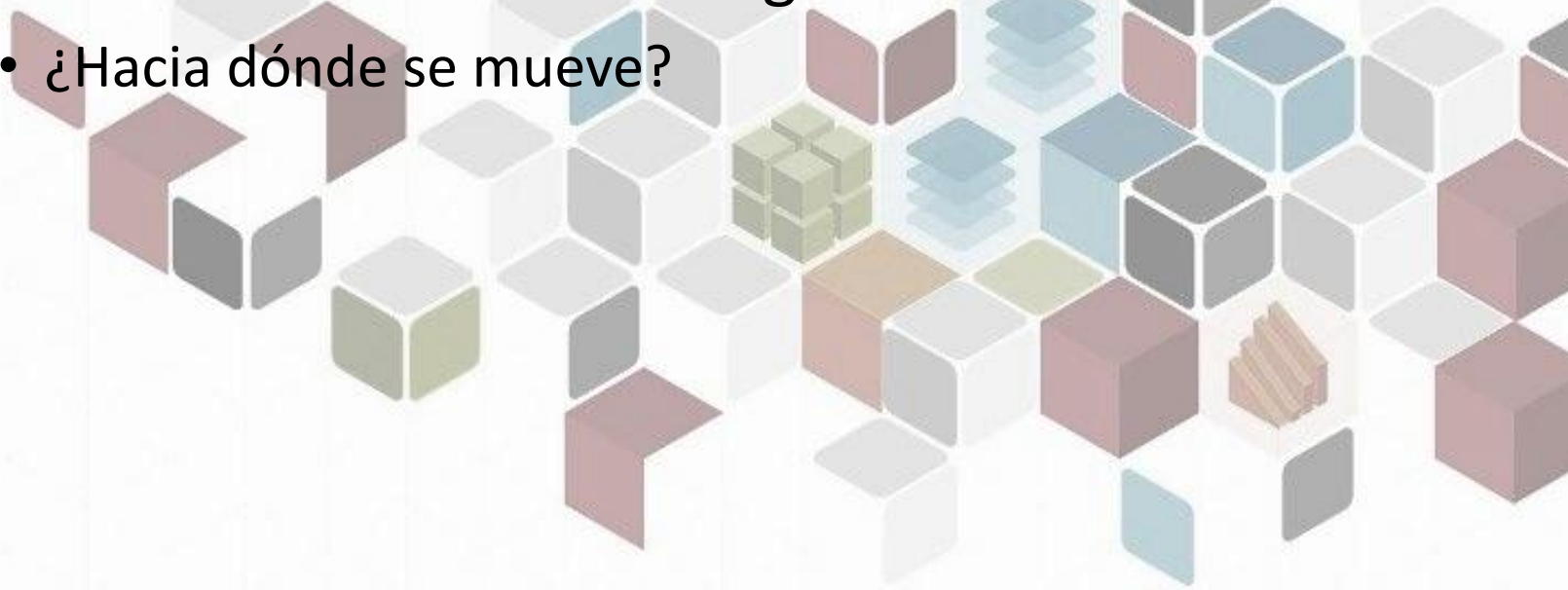


Representaciones gráficas

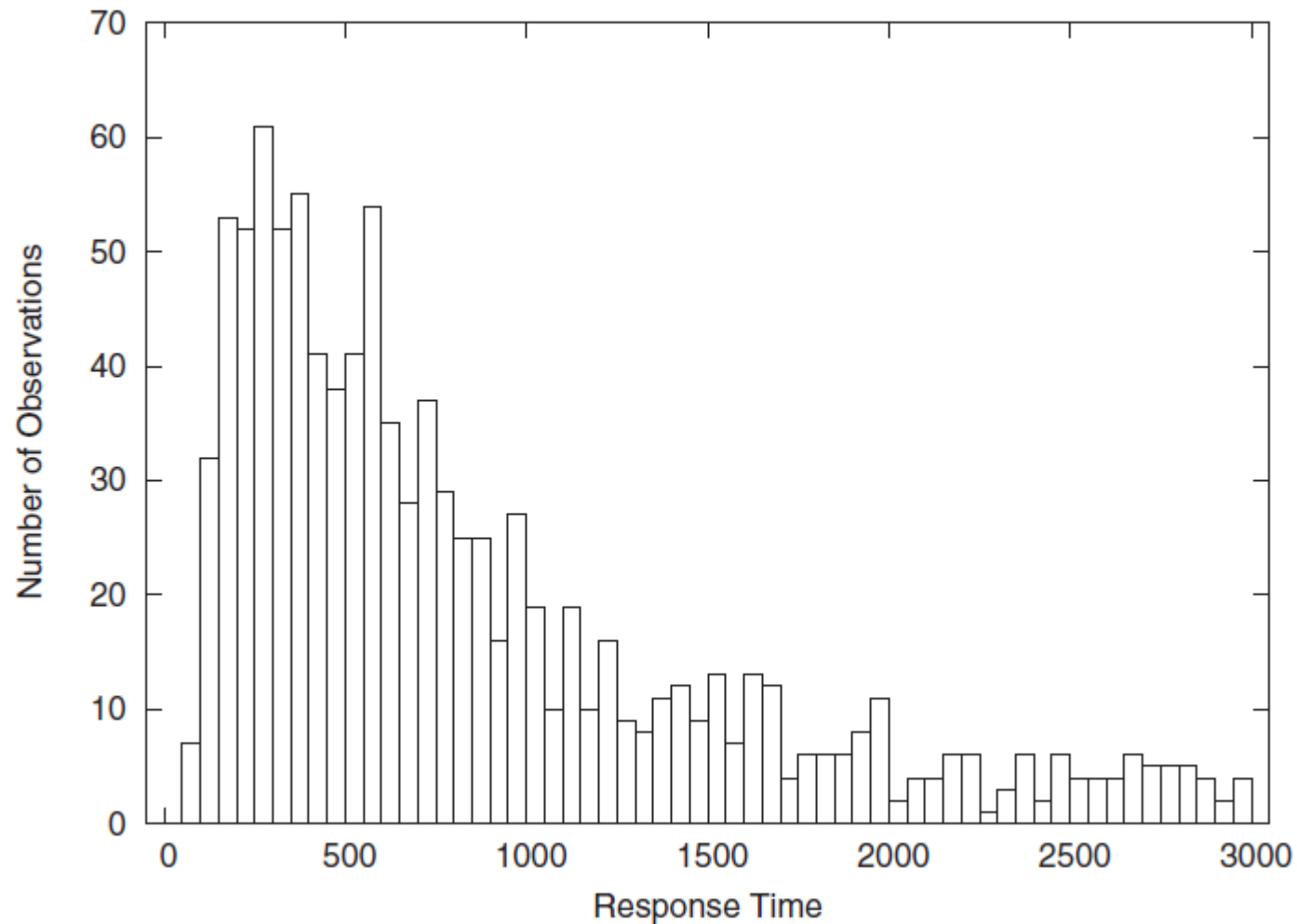


Representaciones gráficas

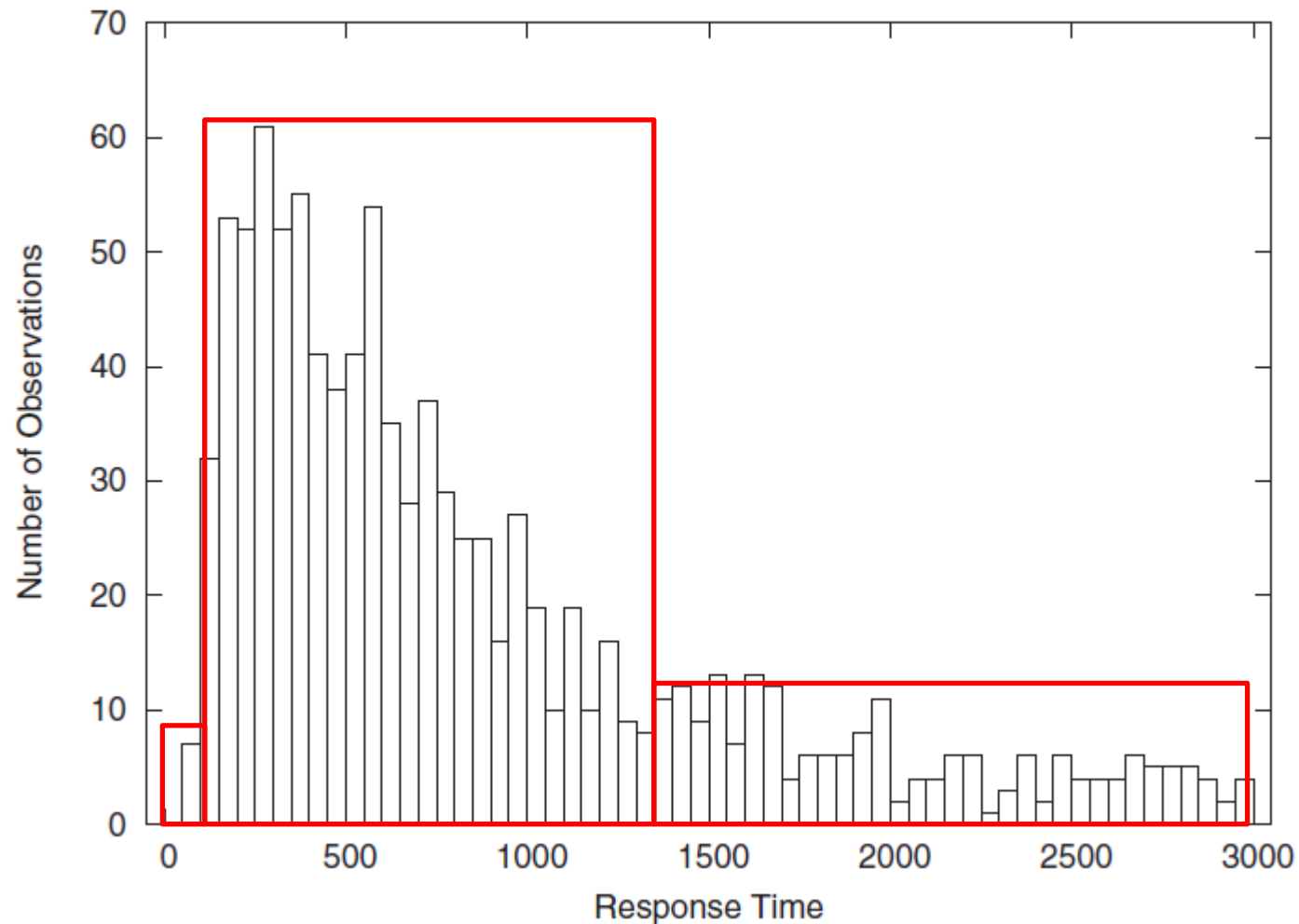
- Existen 2 parámetros
 - Ancho del rectángulo
 - ¿Qué sucede si cambia de tamaño?
 - ¿Cómo saber que tamaño es “bueno”?
 - Posicionamiento del rectángulo
 - ¿Hacia dónde se mueve?



Representaciones gráficas



Representaciones gráficas



Los datos no determinan al histograma

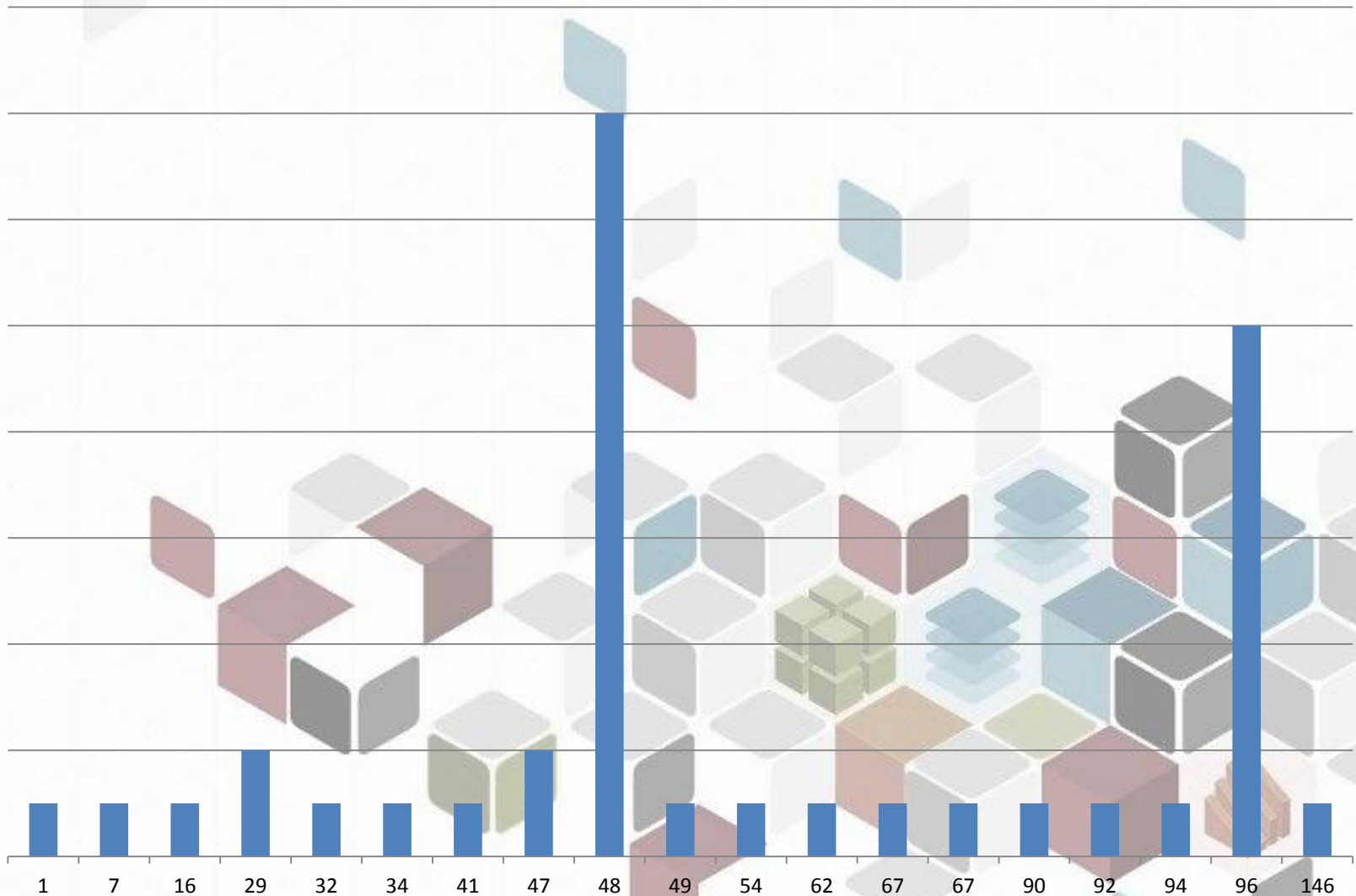
Representaciones gráficas

- Ventajas
 - “Intuitivos”
 - Permiten normalizar datos
 - Fácilmente manipulables
- Desventajas
 - Es viable perder información en su construcción
 - No son únicos
 - No manejan “outliers” de manera eficiente

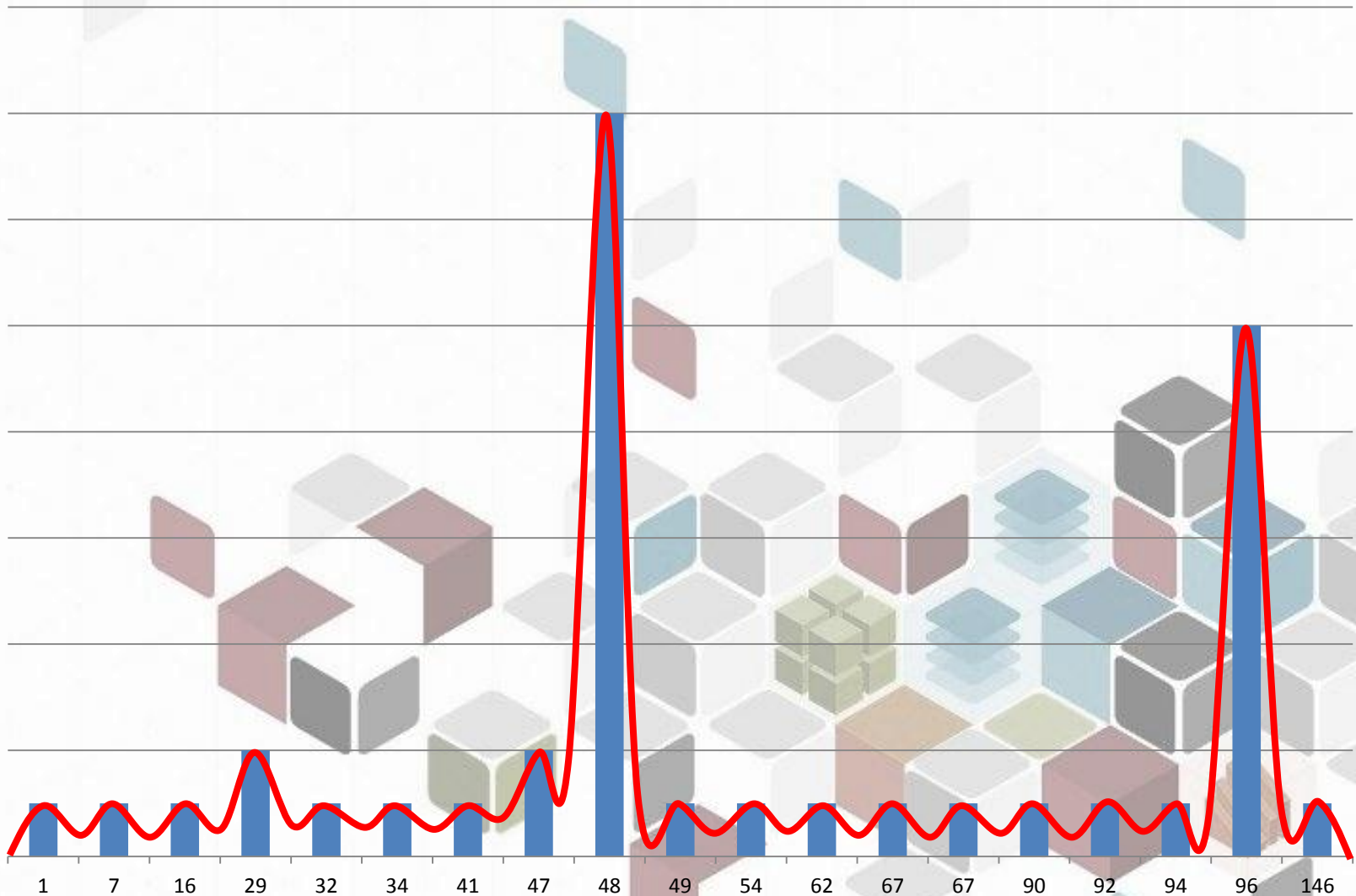
Representaciones gráficas

- KDE
 - No... no es el manejador de ventanas ☺
 - “***Kernel Density Estimates***” – *Estimación de densidad del núcleo*
- Suavizar la rigidez del histograma
- Aplicando una función en cada punto específico y concentrando todas las aplicaciones se logra una curva más suave

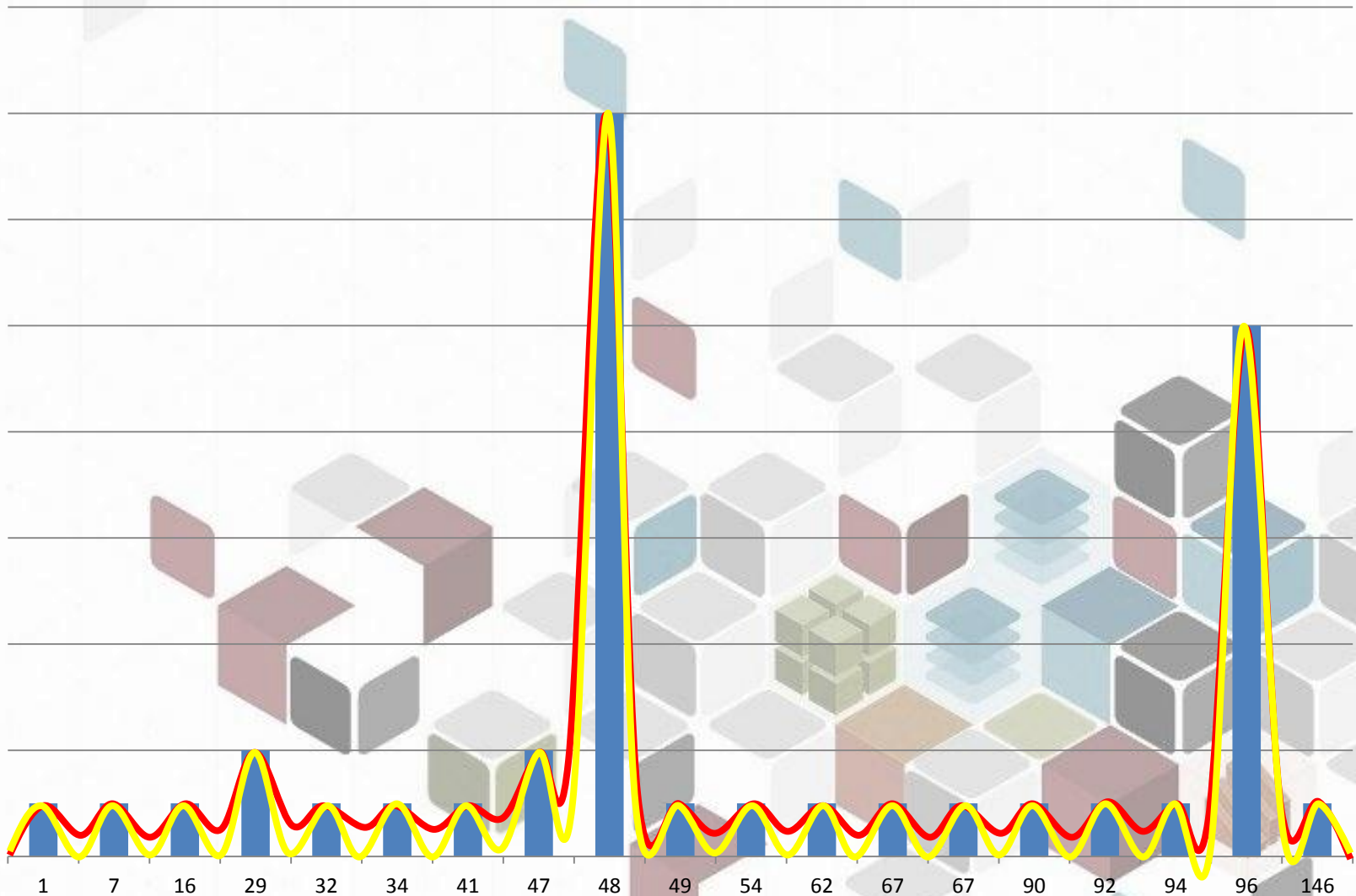
Representaciones gráficas



Representaciones gráficas



Representaciones gráficas

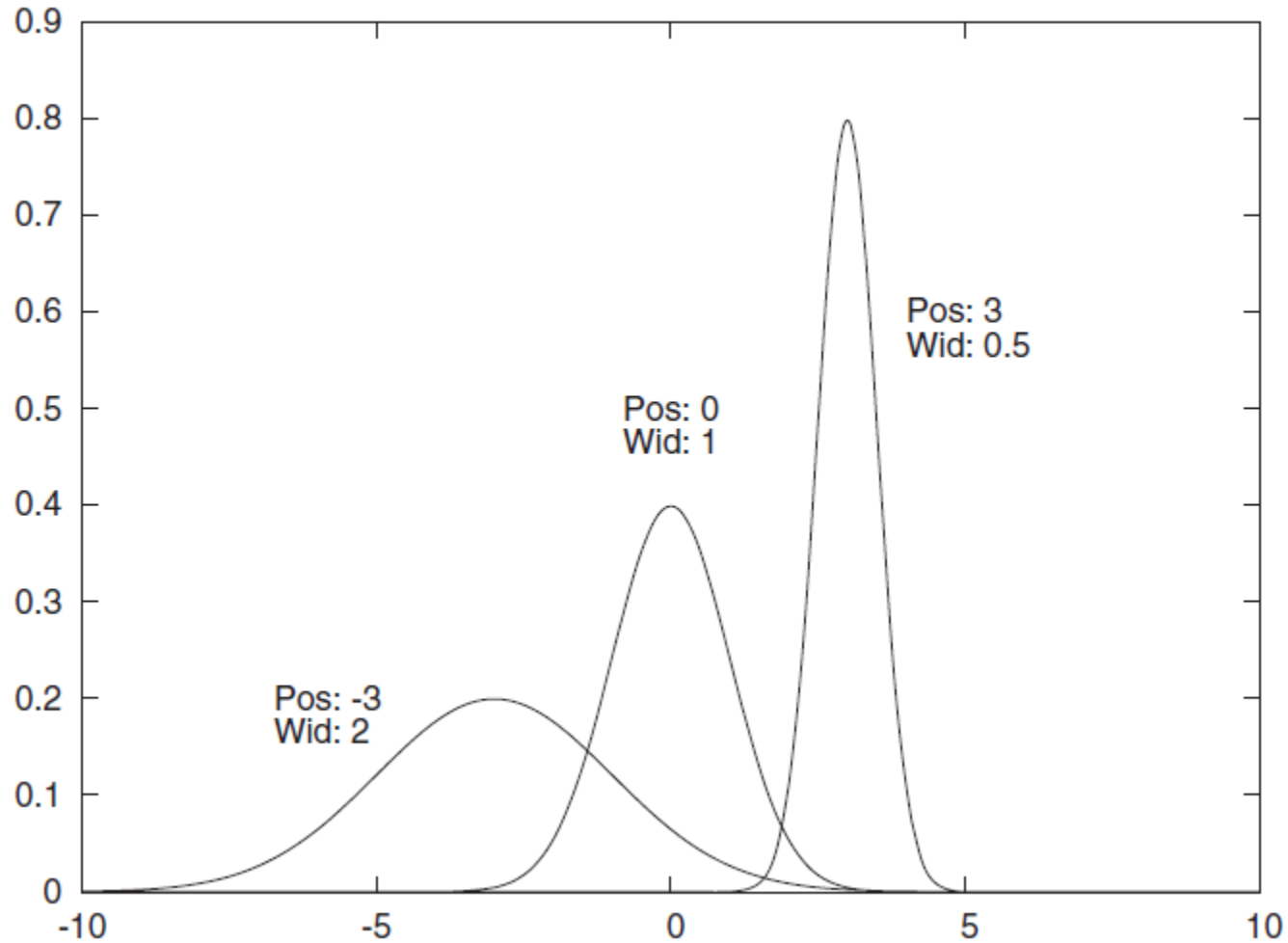


Representaciones gráficas

- E

k

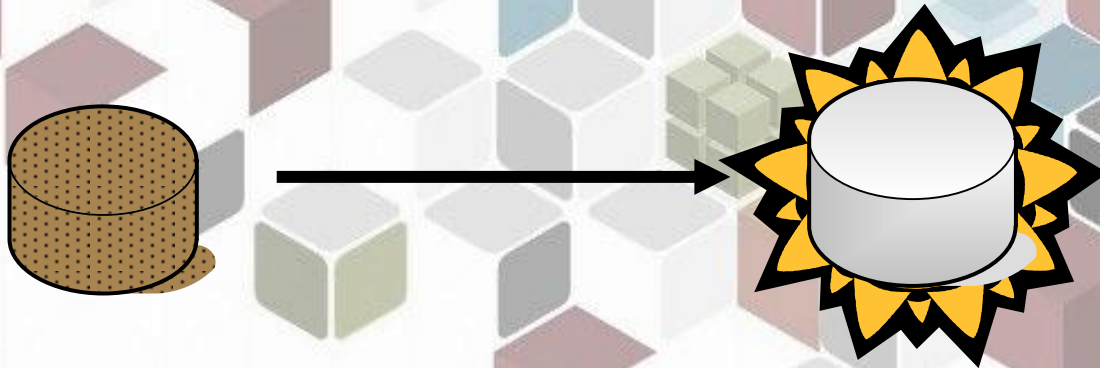
K



Técnicas

- Limpieza de datos.

- “Llenar” datos faltantes.
- “Suavizar” datos “ruidosos”.
- Identificar valores extremos (*outliers*).
- Resolver inconsistencias.



Técnicas

- Integración de datos.

- Mismo concepto, pero diferente nombre.
- Mismo valor expresado de modo distinto.
- Tuplas repetidas en diversas fuentes de datos.



Modificar el esquema.

Técnicas

- Transformación de datos.
 - El rango de algunos atributos difiere mucho.
 - Normalización y agregación
 - Homogeneidad entre atributos
 - Construcción de atributos

Técnicas

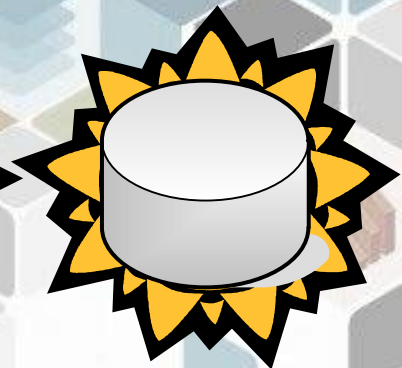
- Reducción de datos.

- Demasiados atributos disminuyen rendimiento
- El análisis se complica proporcionalmente
- Técnicas de muestreo y reducción de dimensiones
- Elección de atributos según ámbito

A	B	C	D
1	10	AAAA	1200
2	20	BBBB	1300
3	12	AABB	1400
4	15	BBAA	1100

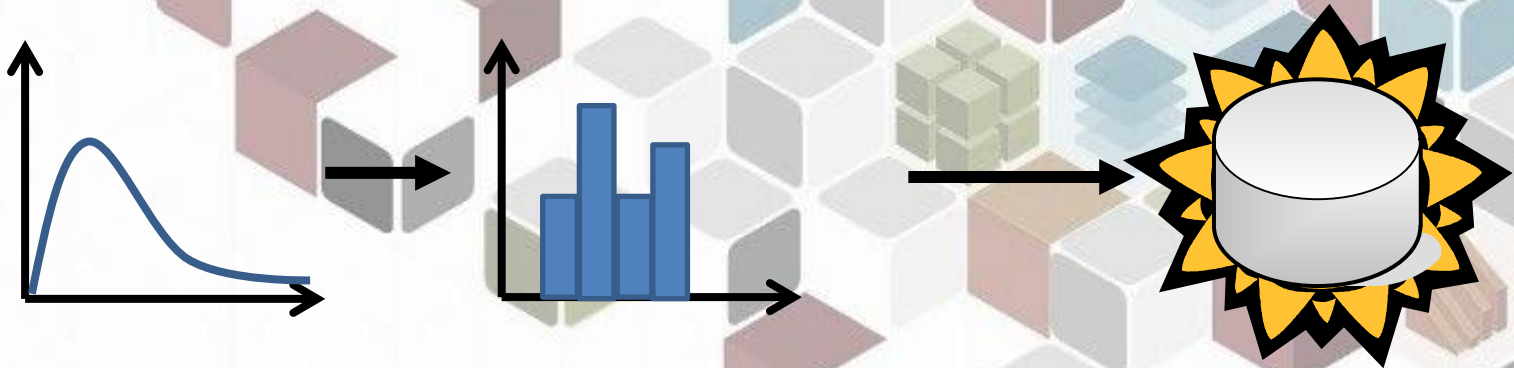


A	D
1	1200
2	1300
3	1400
4	1100

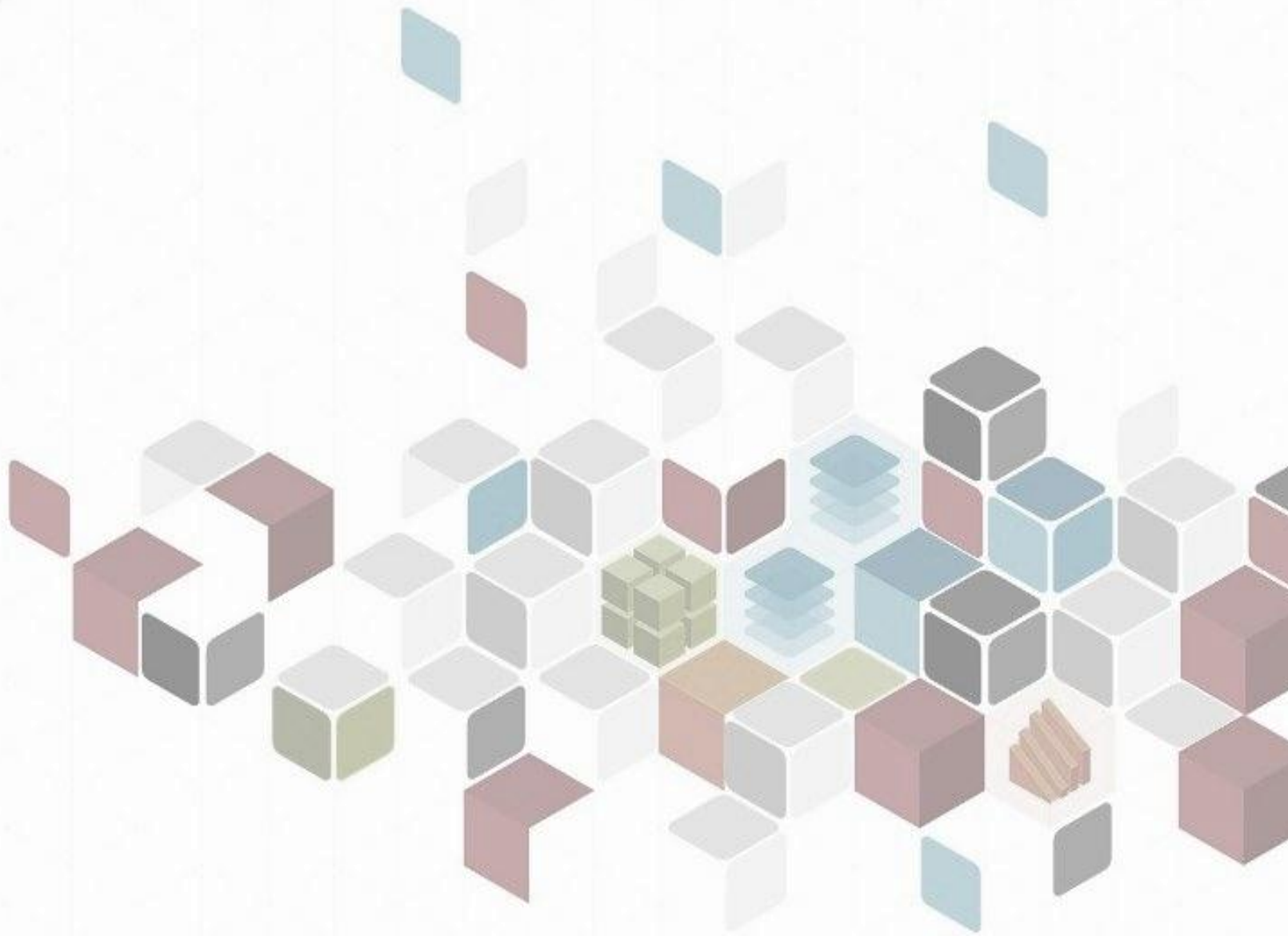


Técnicas

- “Discretización” de datos.
- Parte de reducción de datos, pero importancia para datos numéricos



Técnicas – Limpieza de datos



Técnicas – Limpieza de datos

- “Llenar” datos faltantes.
- Ignorar la tupla.

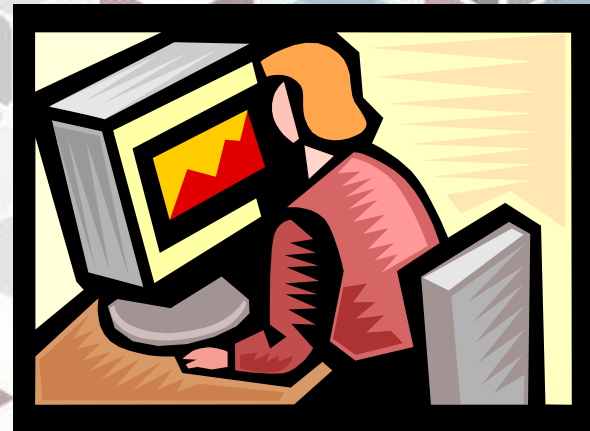
Este método no es muy efectivo, a menos que la tupla contenga muchos atributos con falta de valores. Es especialmente inútil cuando el porcentaje de valores faltantes por atributo varía considerablemente.

Nombre	Apellido	Telefono	Dependientes	Estado Civil
Rebeca	Uno	55660033	0	
Ricardo	Dos	44552211	0	
Rene	Tres		20	
Hugo	Cuatro	56782345	1	Soltero
Paco	Cinco	55909010	2	Casado
Luis	Seis	90231244	0	Viudo

Técnicas – Limpieza de datos

- “Llenar” datos faltantes.
- Rellenar el valor manualmente.

En general este método lleva mucho tiempo y no es factible cuando tenemos un conjunto muy grande de datos con muchos valores faltantes.



Técnicas – Limpieza de datos

- “Llenar” datos faltantes.
- Utilizar una constante global.

Reemplazar todos los valores faltantes de un atributo por la misma constante, por ejemplo “desconocido” ó $-\infty$.

Nombre	Apellido	Telefono	Dependientes	Estado Civil
Rebeca	Uno	55660033	0	Desconocido
Ricardo	Dos	44552211	0	Desconocido
Rene	Tres		20	Desconocido
Hugo	Cuatro	56782345	1	Soltero
Paco	Cinco	55909010	2	Casado
Luis	Seis	90231244	0	Viudo

Técnicas – Limpieza de datos

- “Llenar” datos faltantes.
- Si es numérico, utilizar la media para rellenar.

Utilizar la media para todas las muestras que pertenecen a la misma clase que la tupla dada.

Nombre	Apellido	Telefono	Dependientes	Estado Civil
Rebeca	Uno	55660033	0	
Ricardo	Dos	44552211	0	
Rene	Tres		1	
Hugo	Cuatro	56782345	1	Soltero
Paco	Cinco	55909010	2	Casado
Luis	Seis	90231244	0	Viudo

Técnicas – Limpieza de datos

- “Llenar” datos faltantes.
- Utilizar el valor mas probable a partir de inferencias.

Esto se puede determinar con regresión, herramientas de inferencia utilizando un formalismo Bayesiano o árboles de decisión.

Nombre	Apellido	Telefono	Dependientes	Estado Civil
Rebeca	Uno	55660033	0	Soltero
Ricardo	Dos	44552211	0	Casado
Rene	Tres		20	Soltero
Hugo	Cuatro	56782345	1	Soltero
Paco	Cinco	55909010	2	Casado
Luis	Seis	90231244	0	Viudo

Técnicas – Limpieza de datos

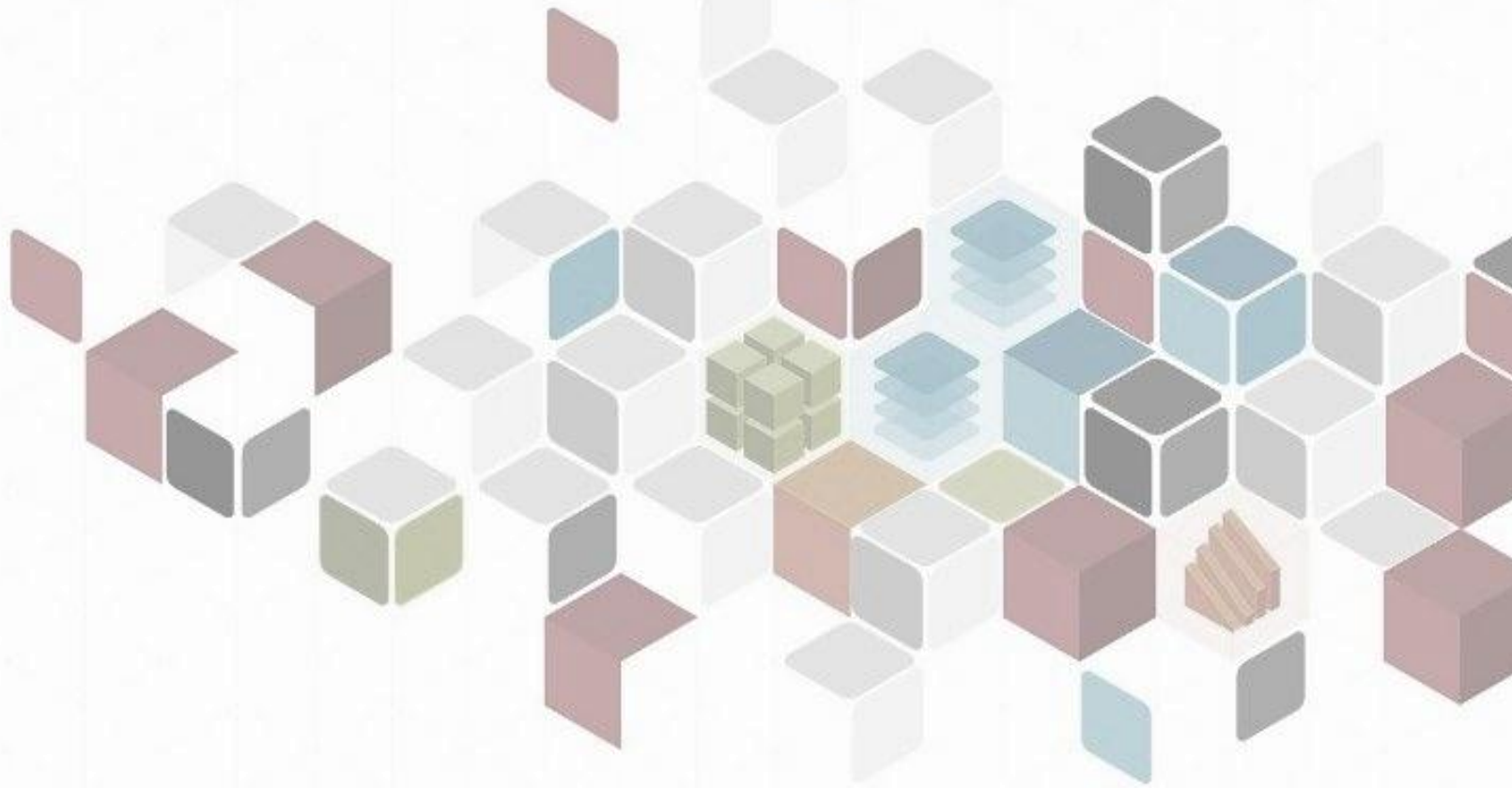
- “Suavizar” datos ruidosos.
- Método “Binning”.
 - Bin por medias
 - Bin por mediana
 - Bin por vecindades

Suavizan un valor de datos ordenados consultando a sus vecinos, esto es, los valores alrededor de él, los valores ordenados se distribuyen en “buckets” o bins (bloques).

Técnicas – Limpieza de datos

Manejo de “cajas” por distribución

4	8	15	21	21	24	25	28	34
---	---	----	----	----	----	----	----	----



Técnicas – Limpieza de datos

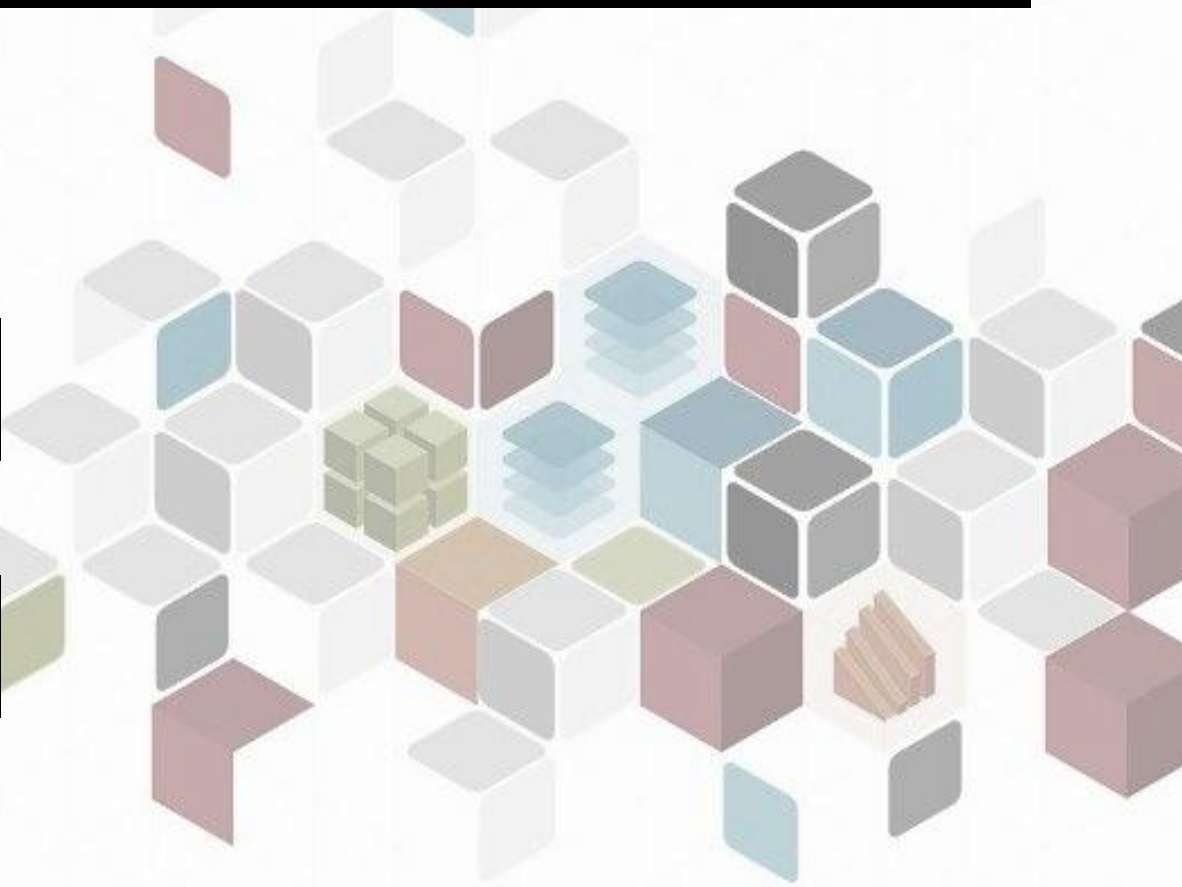
Manejo de “cajas” por distribución

4	8	15	21	21	24	25	28	34
---	---	----	----	----	----	----	----	----

4	8	15
---	---	----

21	21	24
----	----	----

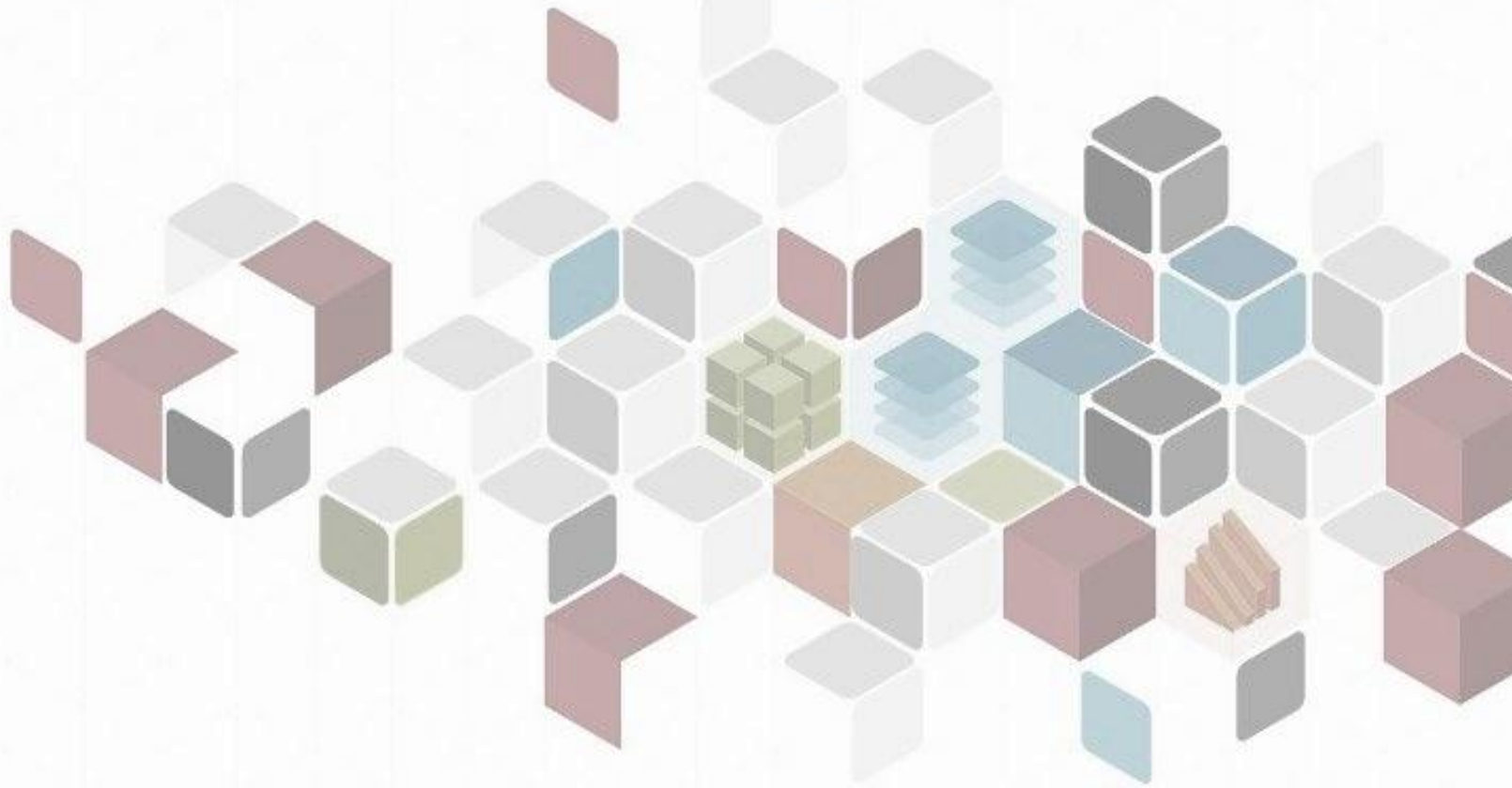
25	28	34
----	----	----



Técnicas – Limpieza de datos

Manejo de “cajas” por medias

4	8	15	21	21	24	25	28	34
---	---	----	----	----	----	----	----	----



Técnicas – Limpieza de datos

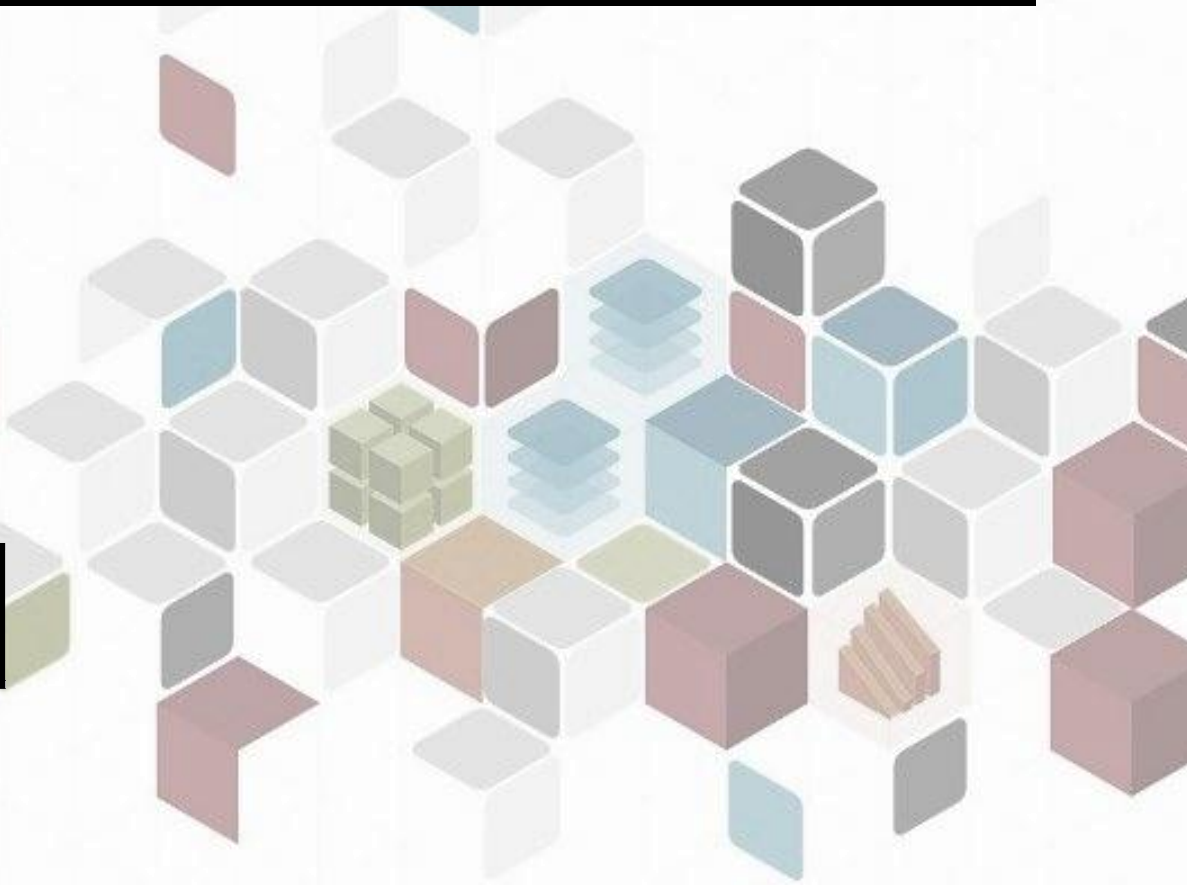
Manejo de “cajas” por medias

4	8	15	21	21	24	25	28	34
---	---	----	----	----	----	----	----	----

9	9	9
---	---	---

22	22	22
----	----	----

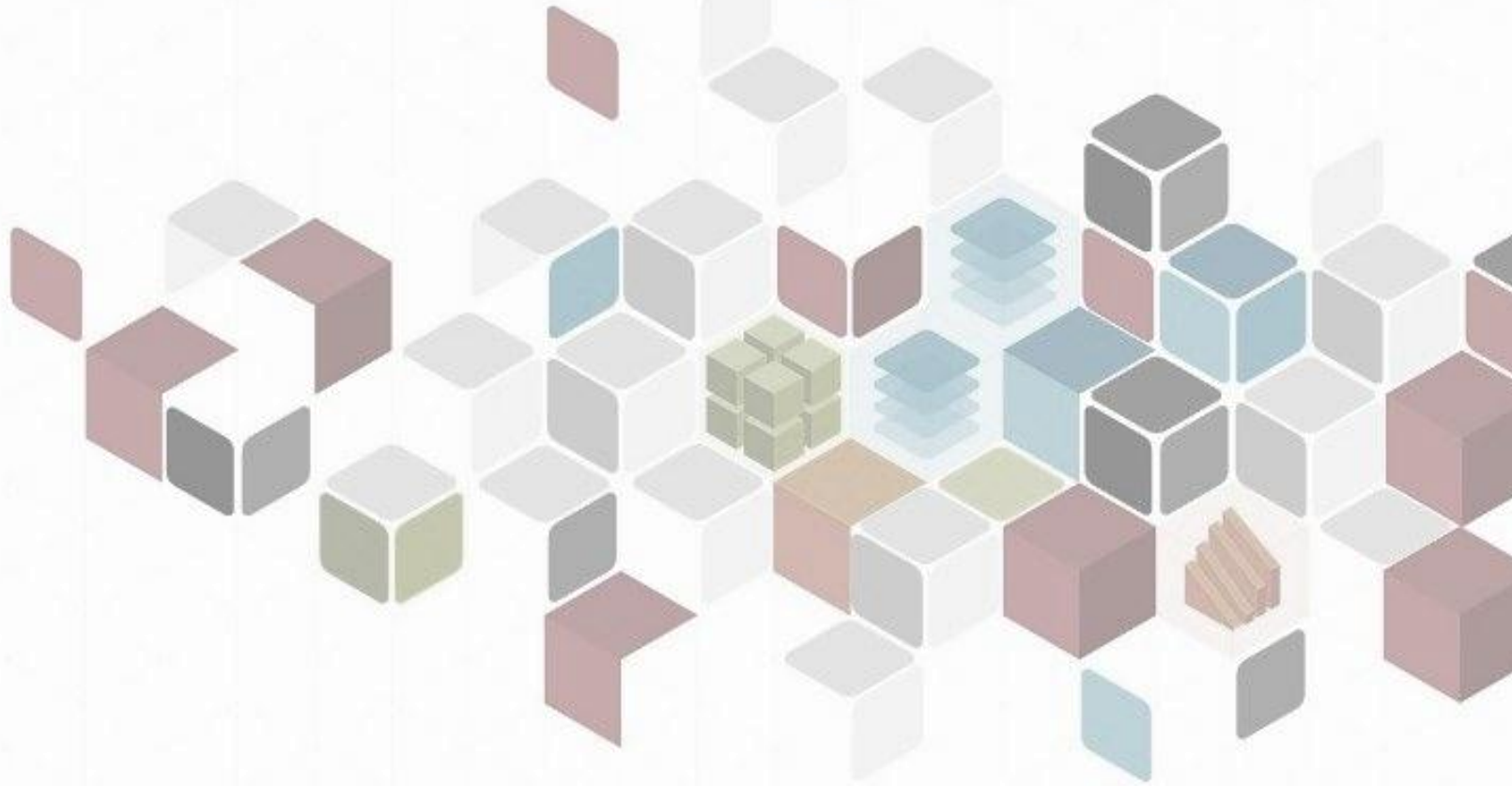
29	29	29
----	----	----



Técnicas – Limpieza de datos

Manejo de “cajas” por limites de caja

4	8	15	21	21	24	25	28	34
---	---	----	----	----	----	----	----	----



Técnicas – Limpieza de datos

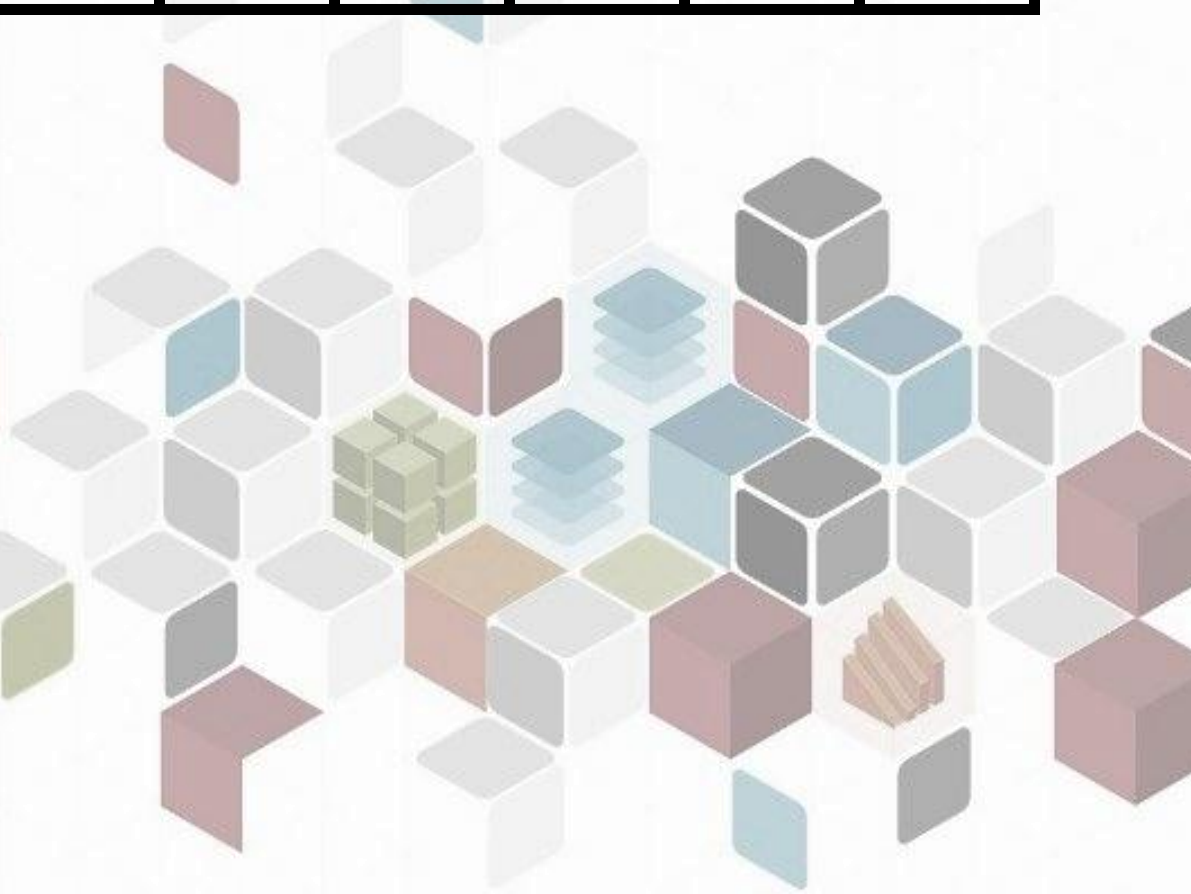
Manejo de “cajas” por limites de caja

4	8	15	21	21	24	25	28	34
---	---	----	----	----	----	----	----	----

4	4	15
---	---	----

21	21	24
----	----	----

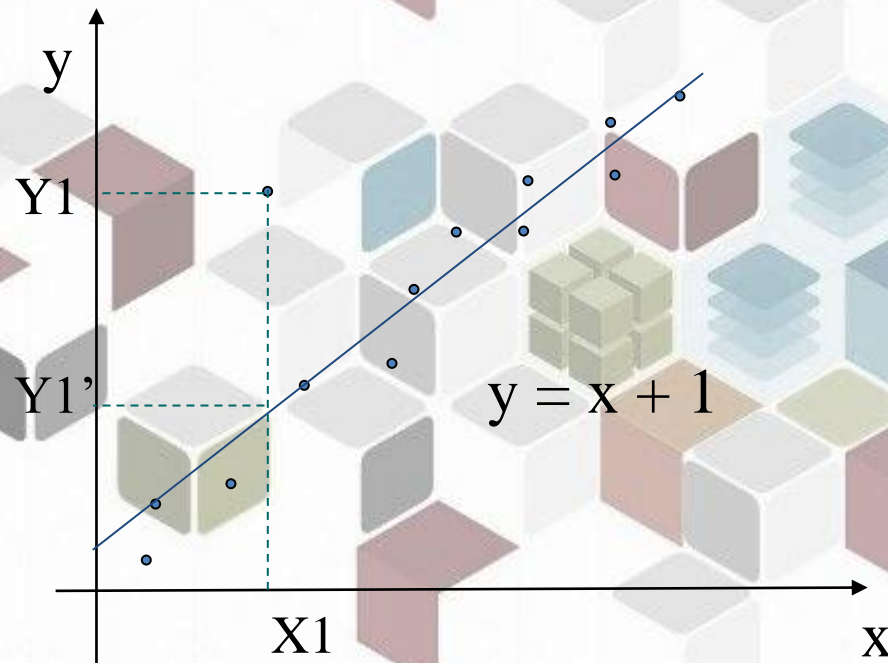
25	25	34
----	----	----



Técnicas – Limpieza de datos

- “Suavizar” datos ruidosos.
- Regresión.

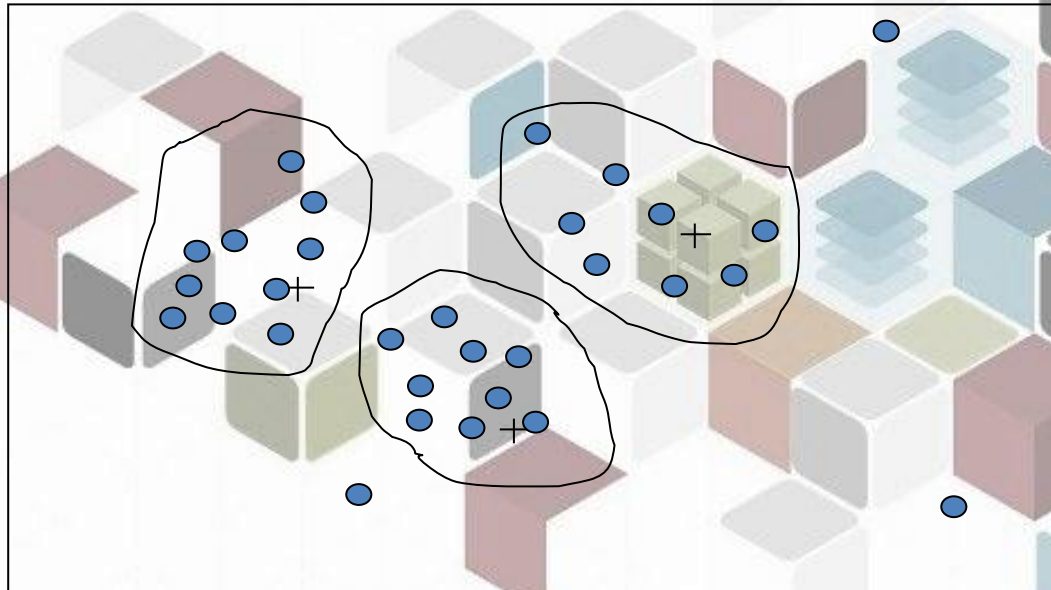
Los datos pueden ser suavizados ajustándolos a una función (regresión).



Técnicas – Limpieza de datos

- “Suavizar” datos ruidosos.
- Agrupamiento.
 - Detectar y remover “outliers”.

Los “outliers” pueden detectarse con agrupamientos, donde valores similares son organizados en grupos o “clusters”.



Técnicas – Limpieza de datos

- “Suavizar” datos ruidosos.
- Inspección automática y supervisada.
 - Valores “sospechosos”
- Regresión.
 - Acoplar datos utilizando funciones de regresión.



Técnicas – Limpieza de datos

- Identificar excepciones o valores extremos

- “Outlier”.

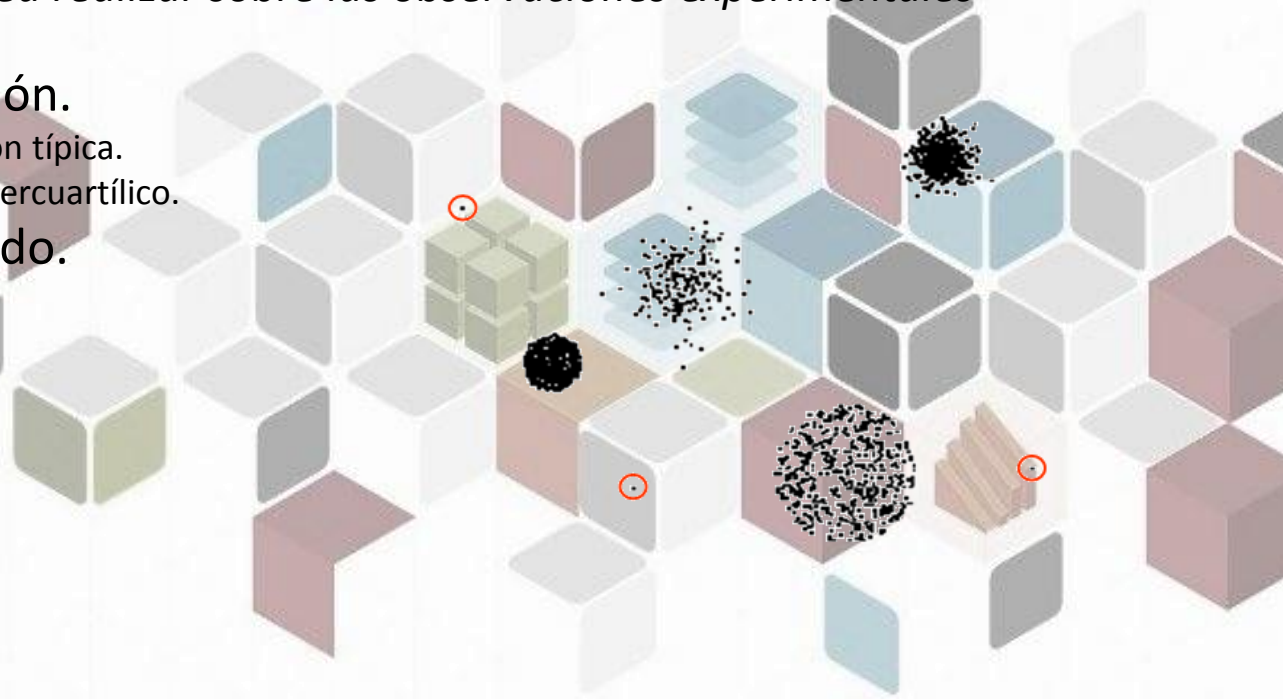
Es aquella observación que siendo atípica y/o errónea, tiene un comportamiento muy diferente con respecto al resto de los datos frente al análisis que se desea realizar sobre las observaciones experimentales

- Métodos de detección.

- Basado en la desviación típica.
- Basado en el rango intercuartílico.

- Métodos de acomodo.

- Recorte
- Reemplazo

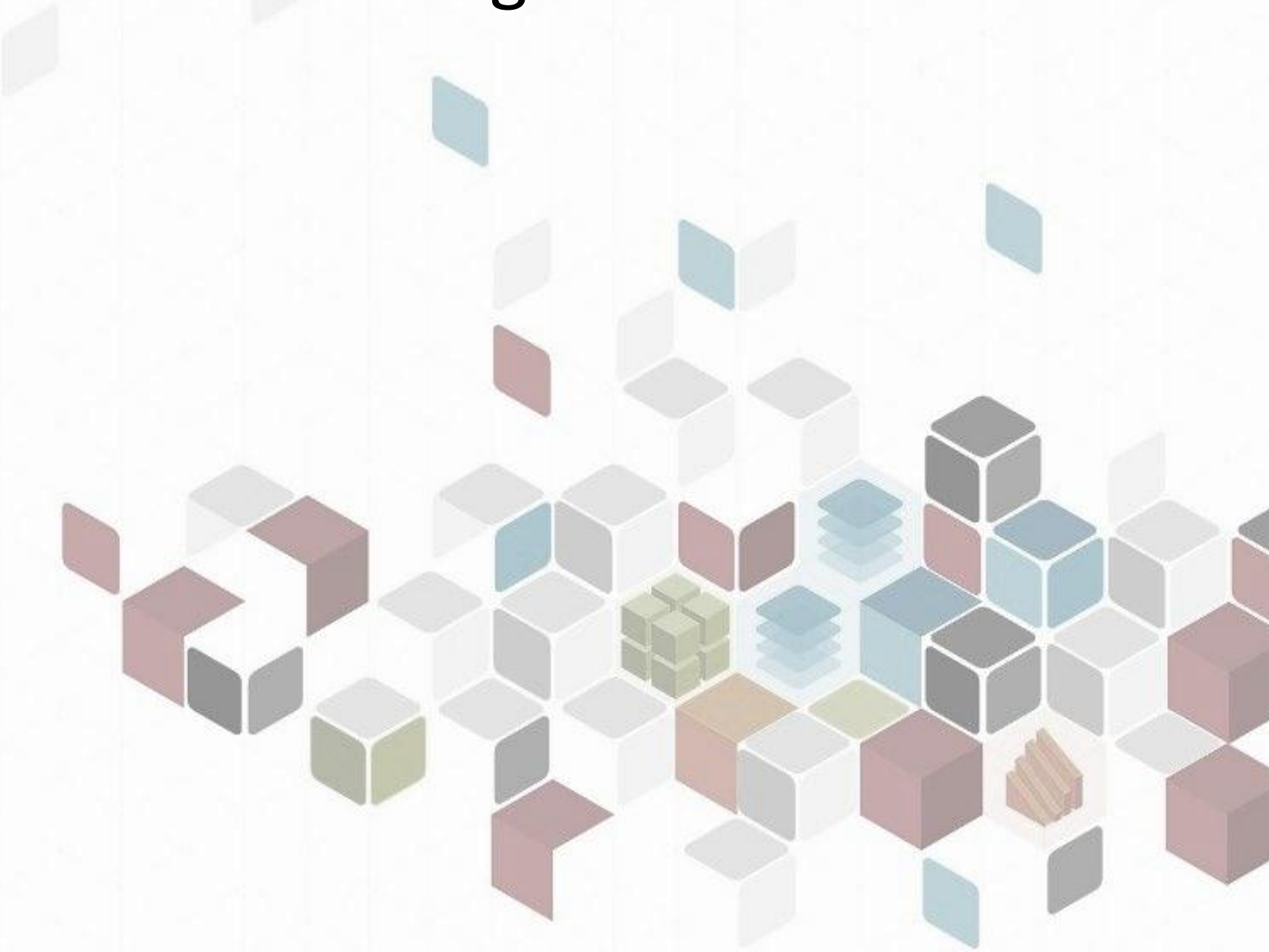


Técnicas – Limpieza de datos

- Resolver inconsistencias
 - Definir estándares.
 - Utilizar restricciones de integridad.
 - Utilizar dependencias funcionales.
 - Definir semántica en atributos
 - Utilización de ETL



Técnicas – Integración de datos



Técnicas – Integración de datos

- Mismo concepto, distinto nombre
- Integración de esquemas.
 - Integración de metadatos.
 - Identificación de entidades.



Oracle

idCliente	sNombre	iEdad	sSexo
1	Hugo	20	M
2	Anni	21	F
3	Paco	22	M
4	Rosa	21	F



PostgreSQL

SecCliente	Nombre	Edad	Genero
C1	Sandy	23	M
C2	Sol	21	M
C3	Rene	27	H
C4	Azul	22	M
C5	Juan	23	H

Técnicas – Integración de datos

- Mismo valor, distinta expresión
- Detección y resolución de conflictos en valores.



Oracle

IdCliente	Nombre	FN	Sexo
1	Hugo	20/07/1995	M
2	Paco	15/05/1998	M
3	Luis	02/12/1993	M
4	Ana	11/04/1990	F
5	Sara	31/08/1992	F



PostgreSQL

SecCliente	Nombre	N	Genero
C1	Sandy	01/31/94	M
C2	Sol	03/21/94	M
C3	Rene	12/22/95	H
C4	Azul	10/15/91	M
C5	Juan	07/08/99	H

Técnicas – Integración de datos

- Datos repetidos, distintas fuentes
- Manejo de datos redundantes.
 - Análisis correlacional.



Oracle

IdCliente	Nombre	FN	Sexo	Salario	Años
1	Hugo	20/07/1995	M	20000	10
2	Paco	15/05/1998	M	30000	12
3	Luis	02/12/1993	M	15000	15
4	Ana	11/04/1990	F	25000	10
5	Sara	31/08/1992	F	20000	11



PostgreSQL

IdCliente	Nombre	FN	Sexo
1	Hugo	20/07/1995	M
2	Paco	15/05/1998	M
3	Luis	02/12/1993	M
4	Ana	11/04/1990	F
5	Sara	31/08/1992	F

SecCliente	Nombre	Salario	Años
C1	Hugo	20000	10
C2	Paco	30000	12
C3	Luis	15000	15
C4	Ana	25000	10
C5	Sara	20000	11

Técnicas – Integración de datos

- Integración de esquemas



UIDCliente	NombreCliente	AP	AM
1	Hugo	Cruz	Nava
2	Luis	Torres	Luna
3	Toño	Castillo	Suárez
4	Ana	Peralta	Liths
5	Sonia	Muñiz	Villa
6	Karla	Zapata	Andre



nIdCliente	sNombre	sApellidoPaterno	sApellidoMaterno
1	Hugo	Cruz	Nava
2	Luis	Torres	Luna
3	Toño	Castillo	Suárez
4	Ana	Peralta	Liths
5	Sonia	Muñiz	Villa
6	Karla	Zapata	Andre

ID	sNombre1	Ap_Pat	Ap_Mat
1	Hugo	Cruz	Nava
2	Luis	Torres	Luna
3	Toño	Castillo	Suárez
4	Ana	Peralta	Liths
5	Sonia	Muñiz	Villa
6	Karla	Zapata	Andre

Técnicas – Integración de datos

- Integración de metadatos de distintas fuentes



nIdCliente	sNombreCompleto
1	Hugo Cruz Nava
2	Luis Torres Luna
3	Toño Castillo Suárez
4	Ana Peralta Liths
5	Sonia Muñiz Villa
6	Karla Zapata Andre

UIDCliente	NombreCliente
1	Cruz Nava Hugo
2	Torres Luna Lucho
3	Castillo Suárez Antonio
4	Peralta Liths Ana
5	Muñiz Villa Sonia
6	Andre Zapata Kar



ID	sNombreCompleto
1	Cruz Nava Hugo
2	Torres Luna Luis
3	Castillo Suárez Toño
4	Peralta Liths Ana
5	Muñiz Villa Sonia
6	Andre Zapata Karla

Técnicas – Integración de datos

- Detectar y resolver conflictos en los valores de los datos



UIDCliente	NombreCliente	Altura
1	Hugo	5.9
2	Luis	5.5
3	Toño	5.10.
4	Ana	6.0.
5	Sonia	5.5
6	Karla	5.7

nIdCliente	sNombre	Estatura
1	Hugo	1.75
2	Luis	1.65
3	Toño	1.78
4	Ana	1.82
5	Sonia	1.65
6	Karla	1.72

ID	sNombre1	Estatura
1	Hugo	175
2	Luis	165
3	Toño	178
4	Ana	182
5	Sonia	165
6	Karla	172

Técnicas – Integración de datos

- La redundancia
 - Análisis de correlación.
- Dados dos atributos, el análisis de correlación puede medir cómo un atributo implica al otro, basándose en los datos disponibles. Para los atributos numéricos podemos evaluar la correlación entre dos atributos A y B , calculando el ***coeficiente de correlación***, esto es:

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum (AB) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

Técnicas – Integración de datos

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum (AB) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

Donde n es el número de tuplas, a_i y b_i son los valores de A y B respectivamente en la tupla i , \bar{A} y \bar{B} es el valor de la media de A y B respectivamente, s_A y s_B son las desviaciones estándar de A y B respectivamente y $(\sum a_i b_i)$ es la suma de AB cross-product (esto es, para cada tupla, el valor de A es multiplicado por el valor de B en esa tupla)

Técnicas – Integración de datos

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum (AB) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

- $r_{A,B} > 0$, A y B correlacionados positivamente ($\nearrow A \leftrightarrow \nearrow B$). Mientras más grande, mayor la correlación.
- $r_{A,B} = 0$: son independientes;
- $r_{A,B} < 0$: existe una correlación negativa

Técnicas – Integración de datos

Para datos categóricos (discretos), una relación de correlación entre dos atributos, A y B , puede descubrirse con una prueba χ^2 .

$$\chi^2 = \sum \frac{(\textit{Observado} - \textit{Esperado})^2}{\textit{Esperado}}$$

Donde la frecuencia observada (cuenta actual) del evento común (A_i, B_j) ; y la frecuencia esperada de (A_i, B_j)

Técnicas – Integración de datos

$$\chi^2 = \sum \frac{(\textit{Observado} - \textit{Esperado})^2}{\textit{Esperado}}$$

Mientras mayor sea el numero χ^2 , entonces es mayor es la probabilidad de que estén relacionados.

Los elementos que contribuyen mas al valor de χ^2 son aquellos cuya cuenta actual es mas diferente que la cuenta esperada

Técnicas – Integración de datos



La correlación **NO** implica
causalidad
de hospitales y # de robos de
autos en una ciudad están
correlacionados
La población

Técnicas – Integración de datos

Cálculo de Chi-Cuadrada

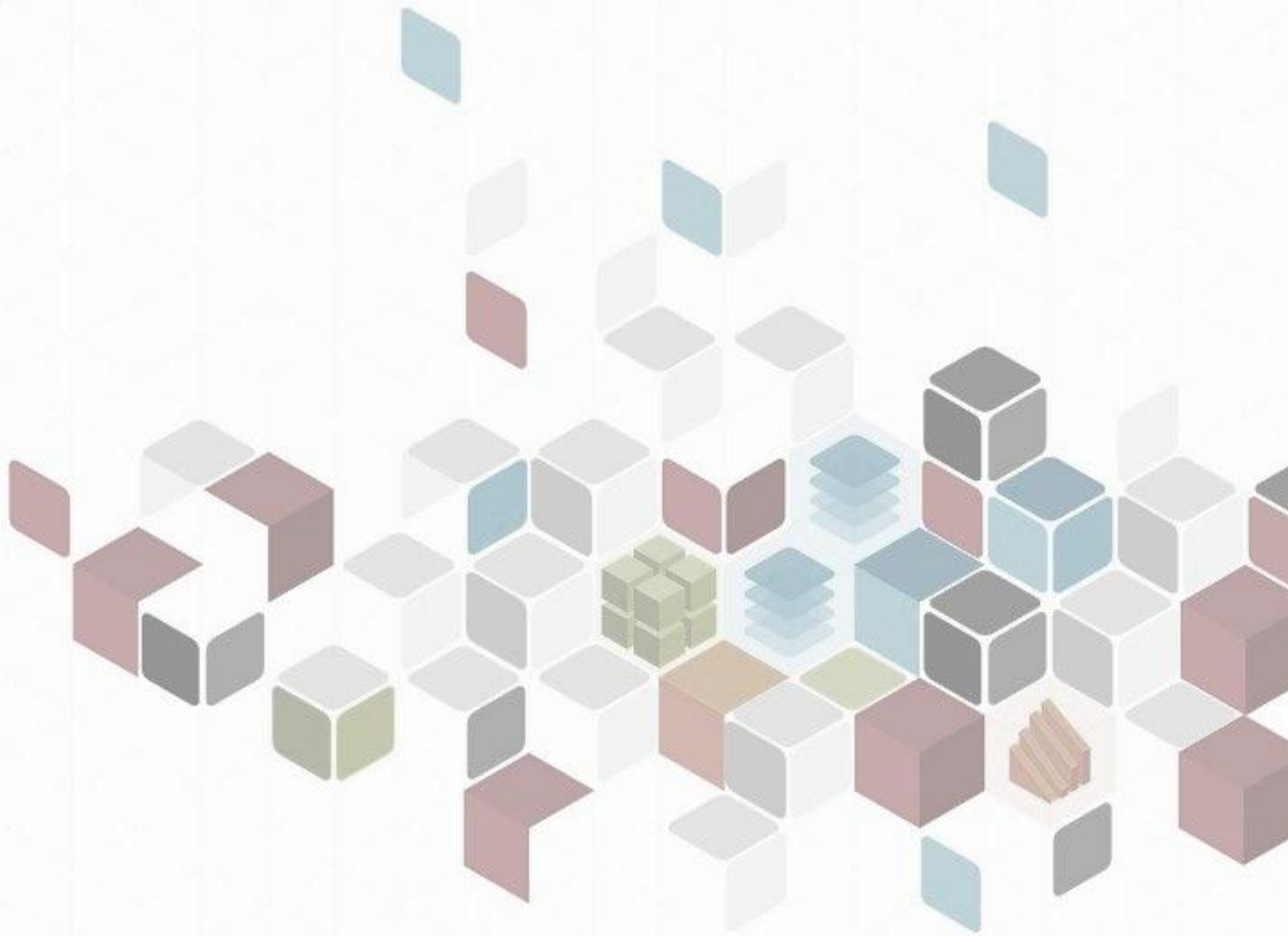
	Jugar ajedrez	No jugar ajedrez	Sum (filas)
Gusta Ciencia Ficción	250(90)	200(360)	450
No gusta Ciencia Ficción	50(210)	1000(840)	1050
Sum(columnas)	300	1200	1500

$$\chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = 507.93$$

El calculo χ^2 (los números en paréntesis son cuentas esperadas calculadas y basadas en la distribución de los datos de las dos categorías)

Se muestra que “Gusta Ciencia Ficción” y “Jugar ajedrez” están correlacionados

Técnicas – Transformación de datos



Técnicas – Transformación de datos

- Rangos que difieren.



Oracle

idAuto	Marca	Modelo	Año	Vmax	Peso
1000	Dodge	Caliber	2009	210	1120
1001	VW	Bora	2010	230	1350
1002	Chevrolet	Astra	2008	210	1210
1003	Chrysler	300	2011	250	1400
1004	Nissan	Murano	2010	215	2200



PostgreSQL

idAuto	Marca	Modelo	Año	Vmax	Peso
2000	Audi	A4	2010	160	2469
2001	BMW	Z3	2006	155	2976
2002	GMC	Accanta	2008	140	2667
2003	Opel	Astra	2009	145	3086
2004	Jeep	Cherokee	2010	140	4850

Técnicas – Transformación de datos

- Generalización.

- Jerarquías de concepto

¡Se pierde detalle, pero se gana significado!

Cereal ⇒ Leche



Kellogg's ⇒ Alpura

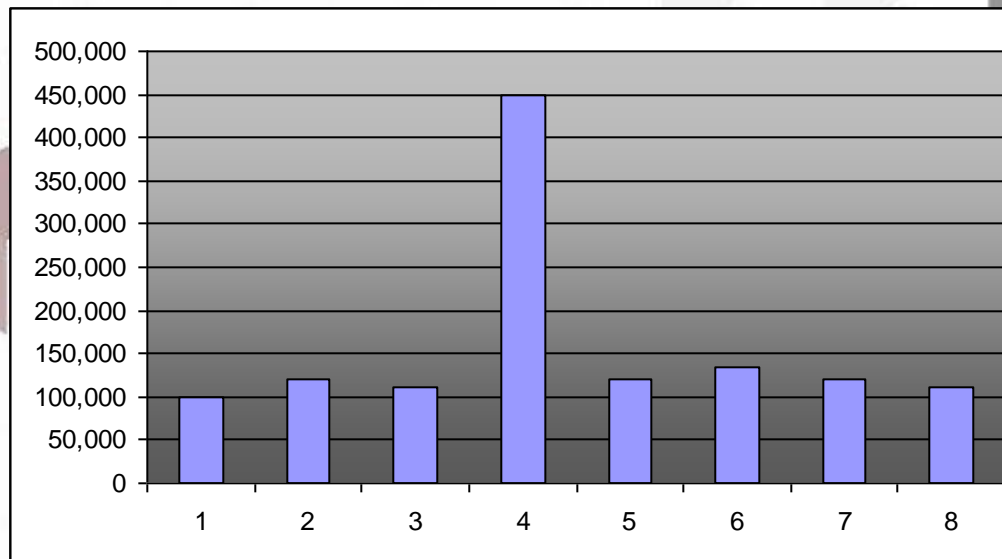


Técnicas – Transformación de datos

- Normalizar.

Ajustar valores para "caer" dentro de un rango pequeño

- Normalización Min-Max (*NMM*)
- Normalización Valor-Z
- Normalización por escala decimal



Técnicas – Transformación de datos

- Normalizar.

Ajustar valores para "caer" dentro de un rango pequeño

- Normalización Min-Max (*NMM*)

Realiza una transformación lineal sobre los datos originales.

Supongamos que \min_A y \max_A son los valores máximos de un atributo A.

NMM mapea un valor, v , de A a v' en el rango $[\text{new_min}_A, \text{new_max}_A]$ calculando:

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

Por ejemplo: Sea el rango de ingresos entre \$12,000 a \$98,000 normalizado o $[0.0, 1.0]$. Entonces \$73,000 se mapea a

$$\frac{73,600 - 12,000}{98,000 - 12,000} (1.0 - 0) + 0 = 0.716$$

Técnicas – Transformación de datos

- Normalizar.

Ajustar valores para "caer" dentro de un rango pequeño

- Normalización Valor-Z

Los valores de un atributo A son normalizados basándose en la media y desviación estándar de A. Un valor, v , de A es normalizado por v' calculando:

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Donde μ y σ_A son la media y desviación estándar respectivamente de un atributo A.

En el ejemplo, sea $\mu = 54,000$, $\sigma = 16,000$. Entonces

$$\frac{73,600 - 54,000}{16,000} = 1.225$$

Técnicas – Transformación de datos

- Normalizar.

Ajustar valores para "caer" dentro de un rango pequeño

- Normalización por escala decimal

Normaliza moviendo el punto decimal de los valores de un atributo A. El número de los puntos decimales movido depende del valor máximo absoluto de A.

Un valor, v , de A es normalizado con v' calculando:

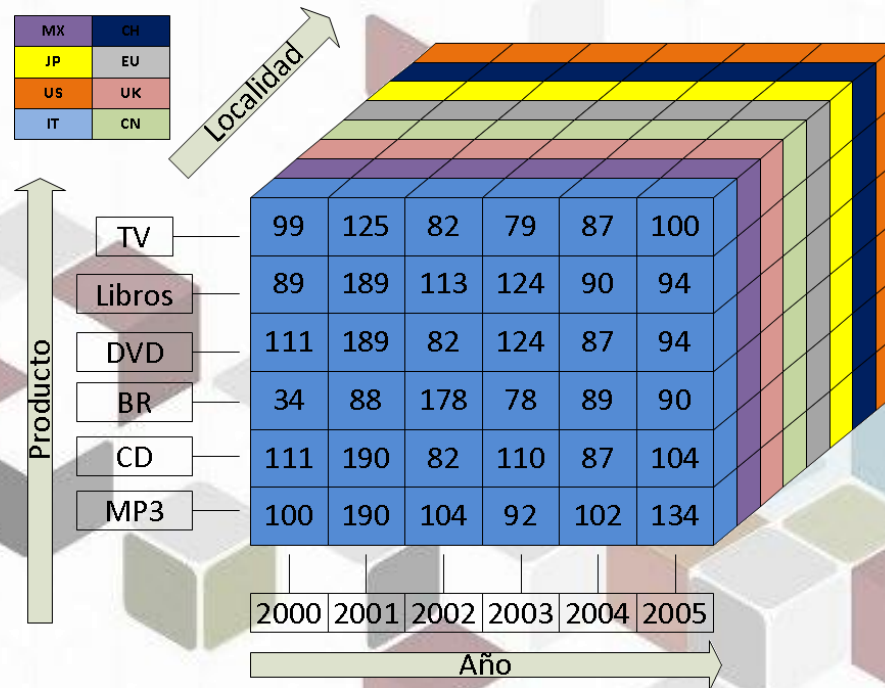
$$v' = \frac{v}{10^j}$$

Donde j es el entero mas pequeño tal que $\text{Max}(|v'|) < 1$.

Técnicas – Transformación de datos

- Agregación.

Resúmenes, construcción de cubos de datos, etc.



Técnicas – Transformación de datos

- Construcción de atributos.

Nuevos atributos a partir de los originales

Fecha de ingreso \Rightarrow Fecha de jubilación



Técnicas – Reducción de datos

¡Premisa!

“El tiempo utilizado en la reducción no debe rebasar el tiempo salvado en el minado de datos”



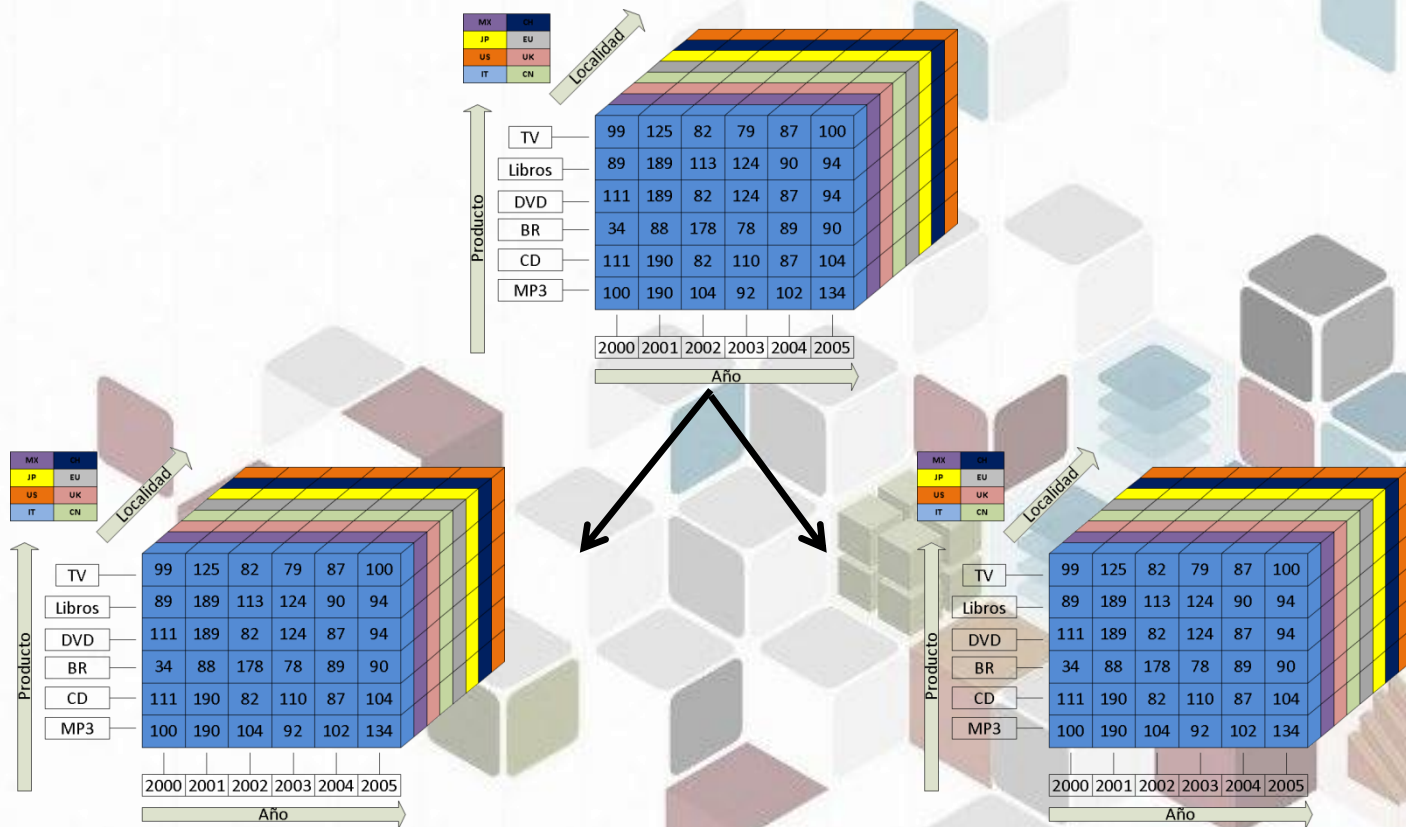
Técnicas – Reducción de datos

¿Por qué?

- BD con información “irrelevante”
- Costo computacional costoso en tiempo,
- Recursos limitados
- Respuesta final “casi igual”

Técnicas – Reducción de datos

- Agregación de cubos de datos



Técnicas – Reducción de datos

- Reducción en atributos

IdCliente	Nombre	FN	Sexo	Salario	Años	Hijos	Casado	Auto	Seguro	Renta	Fing
1	Hugo	20/07/1995	M	20000	10	1	T	T	F	T	1/5/2001
2	Paco	15/05/1998	M	30000	12	0	T	T	F	F	2/2/1999
3	Luis	02/12/1993	M	15000	15	2	F	T	T	T	3/12/1996
4	Ana	11/04/1990	F	25000	10	2	T	F	F	F	15/08/2001
5	Sara	31/08/1992	F	20000	11	1	T	F	T	F	17/11/2000



IdCliente	lombr	FN	Sexo	Salario	Hijos	Casado	Auto	Seguro	Renta
1	Hugo	20/07/1995	M	20000	1	T	T	F	T
2	Paco	15/05/1998	M	30000	0	T	T	F	F
3	Luis	02/12/1993	M	15000	2	F	T	T	T
4	Ana	11/04/1990	F	25000	2	T	F	F	F
5	Sara	31/08/1992	F	20000	1	T	F	T	F

Técnicas – Reducción de datos

- Reducción en atributos
- El uso de **métodos heurísticos** permite hacer una búsqueda en el espacio reducida
- La búsqueda se enfoca a una elección óptima que lleva a una solución
- El resultado es casi siempre estimado
- ¿Estimada?
 - Uso de pruebas de significado de atributos
 - Los atributos son considerados independientes
 - Métrica de ganancia de información

Técnicas – Reducción de datos

- Reducción en atributos

Métodos heurísticos utilizados:

- Stepwise forward selection
- Stepwise backward elimination
- Combinacion de FS y SBE
- Decision Tree Induction



Técnicas – Reducción de datos

- Reducción en atributos

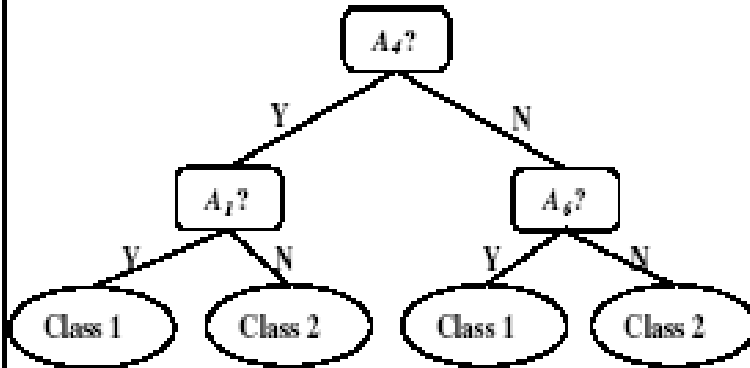
Forward selection	Backward elimination	Decision tree induction
<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p>  <pre> graph TD A4["A4?"] -- Y --> A1["A1?"] A4 -- N --> A6["A6?"] A1 -- Y --> C1_1("Class 1") A1 -- N --> C2_1("Class 2") A6 -- Y --> C1_2("Class 1") A6 -- N --> C2_2("Class 2") </pre> <p>\Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>

Figure 2.15: Greedy (heuristic) methods for attribute subset selection.

Técnicas – Reducción de datos

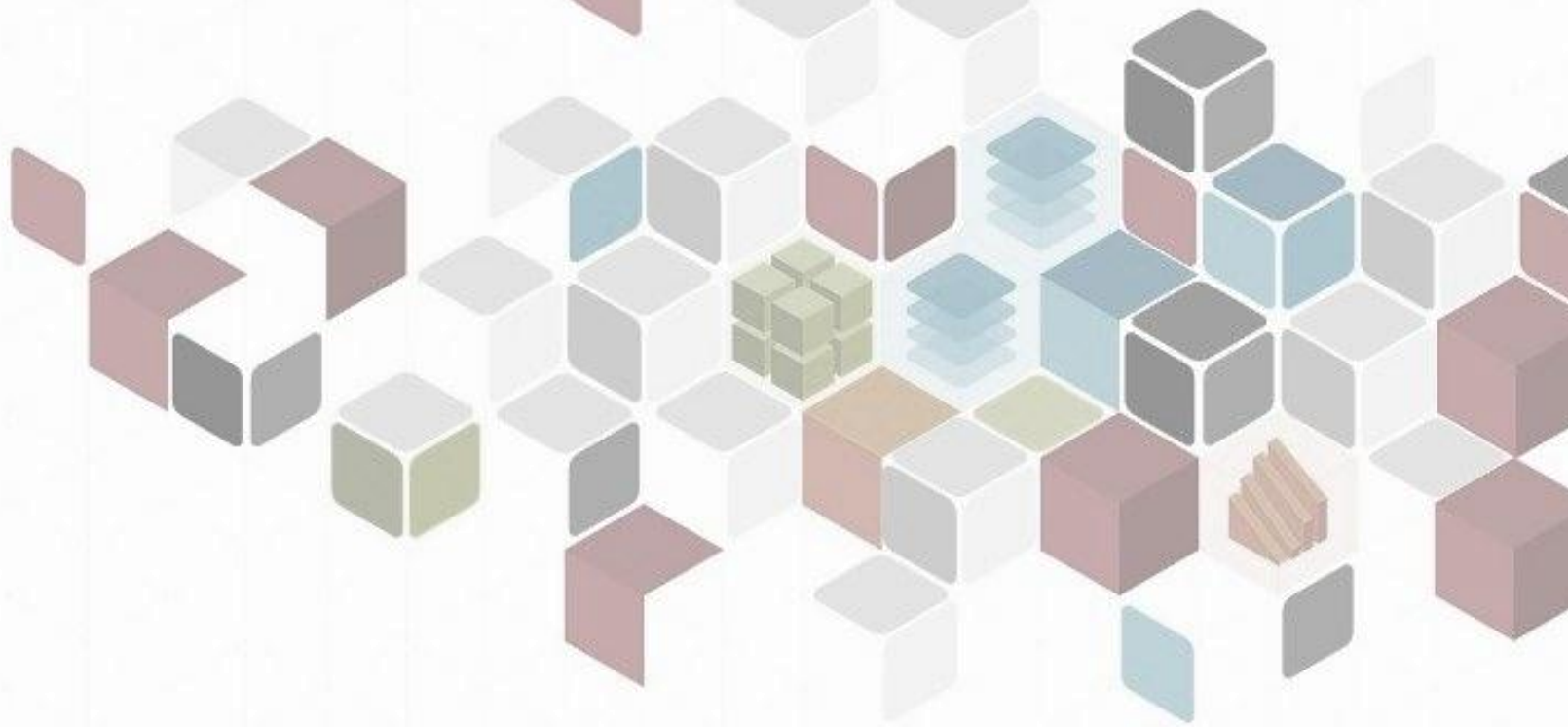
- Reducción de cantidad

IdCliente	Nombre	FN	Sexo	Salario
1	Hugo	20/07/1995	M	20550.00
2	Paco	15/05/1998	M	30250.00
3	Luis	02/12/1993	M	15500.00
4	Ana	11/04/1990	F	15000.00
5	Sara	31/08/1992	F	21250.00
6	Sandy	1/31/1994	F	12000.50
7	Sol	3/21/1994	F	27500.50
8	Rene	12/22/1995	M	18350.50
9	Azul	10/15/1991	F	19450.50
10	Juan	7/8/1999	M	15500.50

IdCliente	Nombre	FN	Sexo	Salario
1	Hugo	20/07/1995	M	Medio
2	Paco	15/05/1998	M	Alto
3	Luis	02/12/1993	M	Bajo
4	Ana	11/04/1990	F	Bajo
5	Sara	31/08/1992	F	Medio
6	Sandy	1/31/1994	F	Bajo
7	Sol	3/21/1994	F	Medio
8	Rene	12/22/1995	M	Medio
9	Azul	10/15/1991	F	Alto
10	Juan	7/8/1999	M	Medio

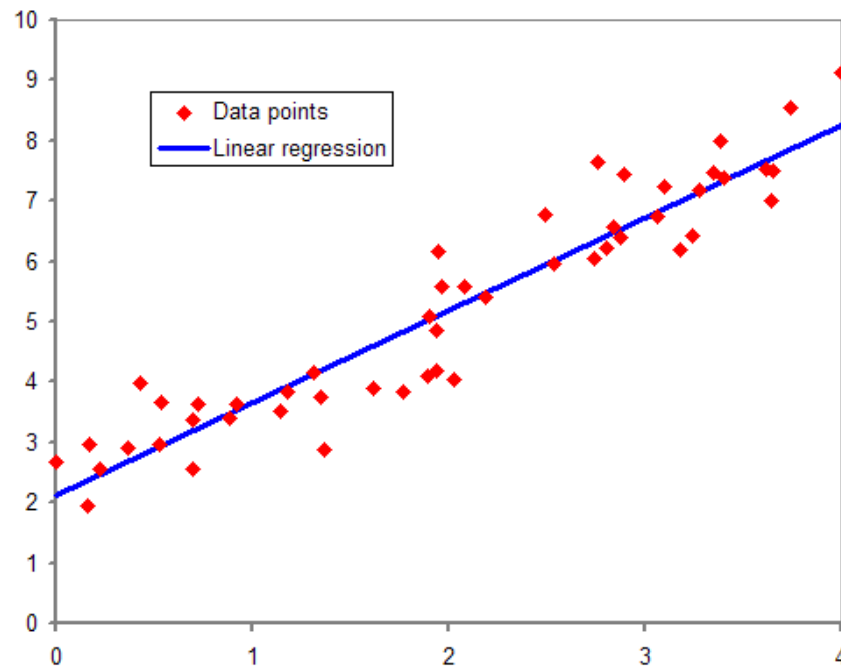
Técnicas – Reducción de datos

- Reducción de cantidad
 - Paramétrica
 - Un modelo es usado para estimar el comportamiento de los datos
 - Se almacenan los parámetros que representan los datos, no los datos en si



Técnicas – Reducción de datos

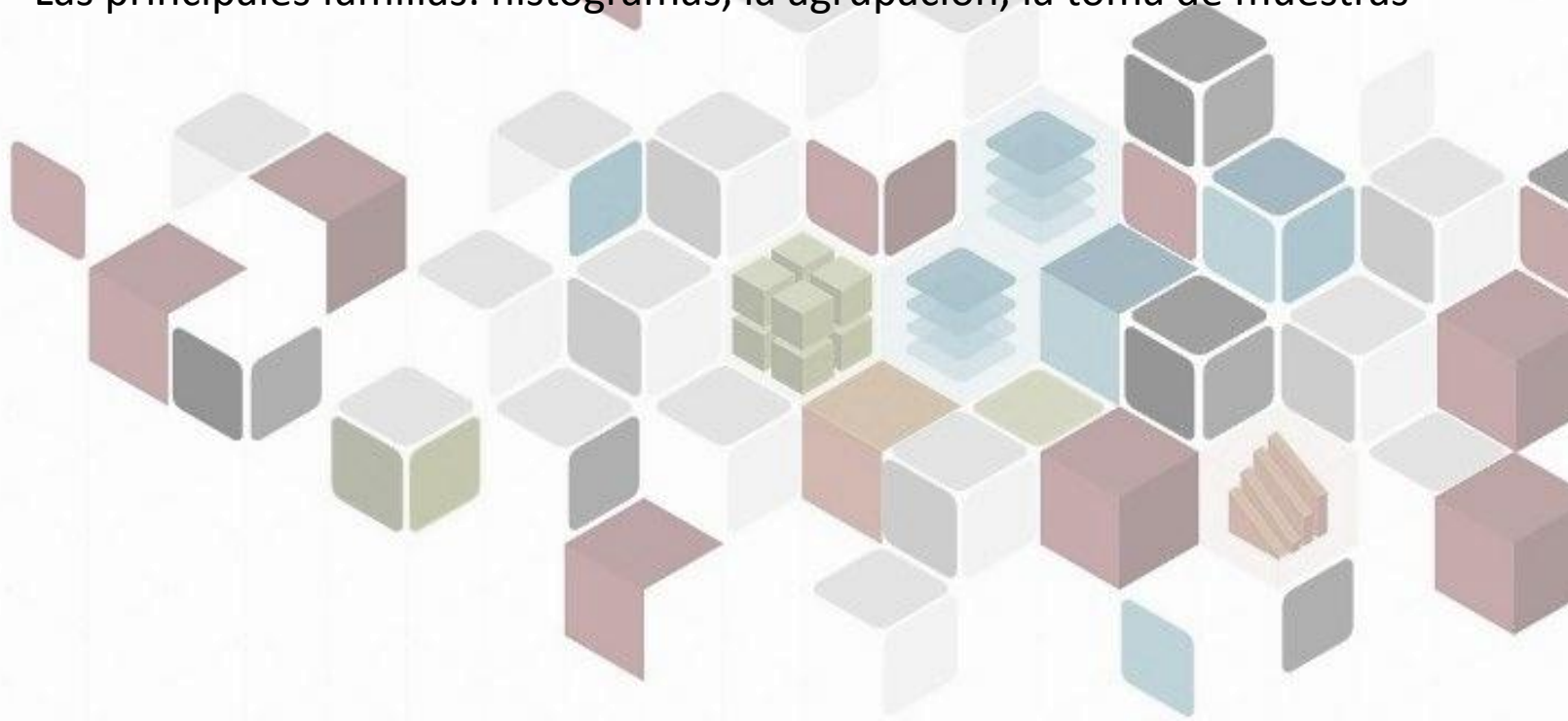
- Reducción de cantidad
 - Paramétrica
 - Modelos de Regresión linear



$$y = wx + b$$

Técnicas – Reducción de datos

- Reducción de cantidad
 - No Paramétrica
 - Almacena formas de representación mas versátiles, ideales para la presentación de información
 - Las principales familias: histogramas, la agrupación, la toma de muestras



Técnicas – Reducción de datos

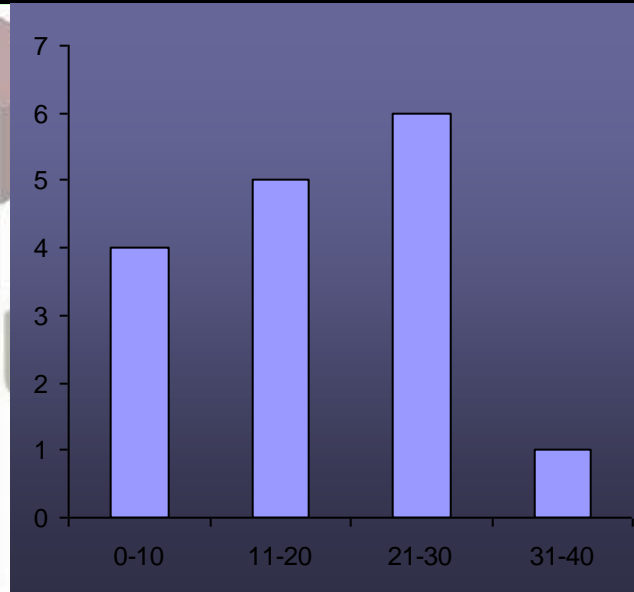
- Reducción de cantidad
 - No Paramétrica
 - Histograma



Técnicas – Reducción de datos

- Reducción de cantidad

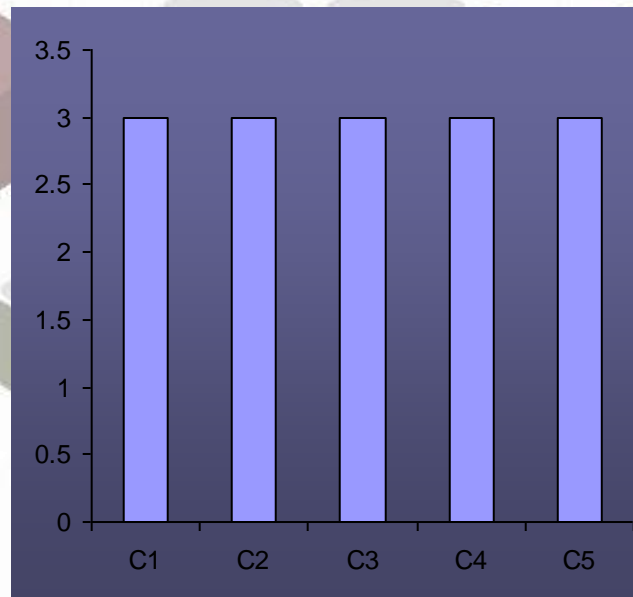
- No Paramétrica
 - Histograma
 - Igual amplitud



Técnicas – Reducción de datos

- Reducción de cantidad

- No Paramétrica
 - Histograma
 - Igual frecuencia



Técnicas – Reducción de datos

- Reducción de cantidad
 - No Paramétrica
 - Clustering (Cúmulos)
 - Los elementos son consideradas como objetos
 - Es mejor si los datos pueden ser organizados en clusters
 - Las representaciones del cluster reemplazan los datos
 - Árboles multidimensionales con índices



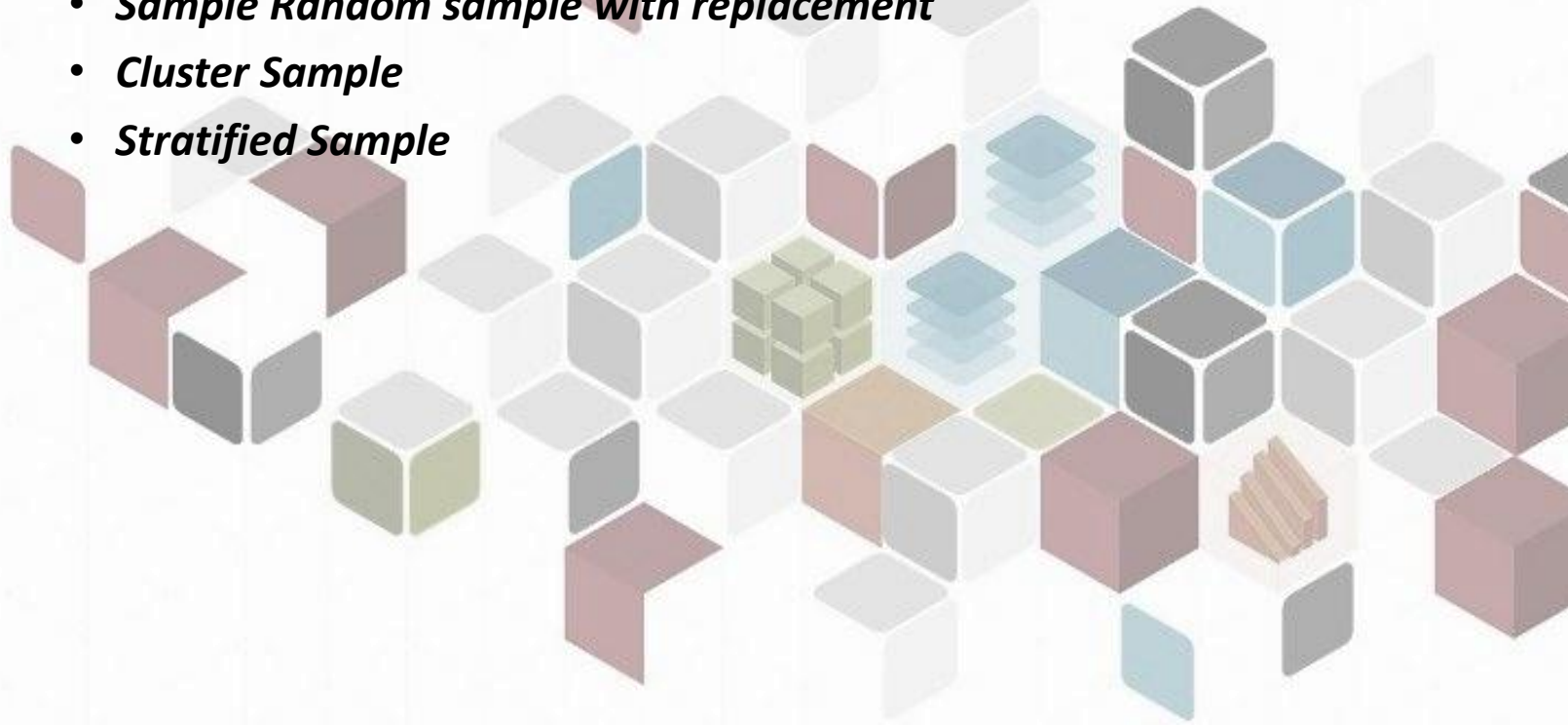
Técnicas – Reducción de datos

- Reducción de cantidad

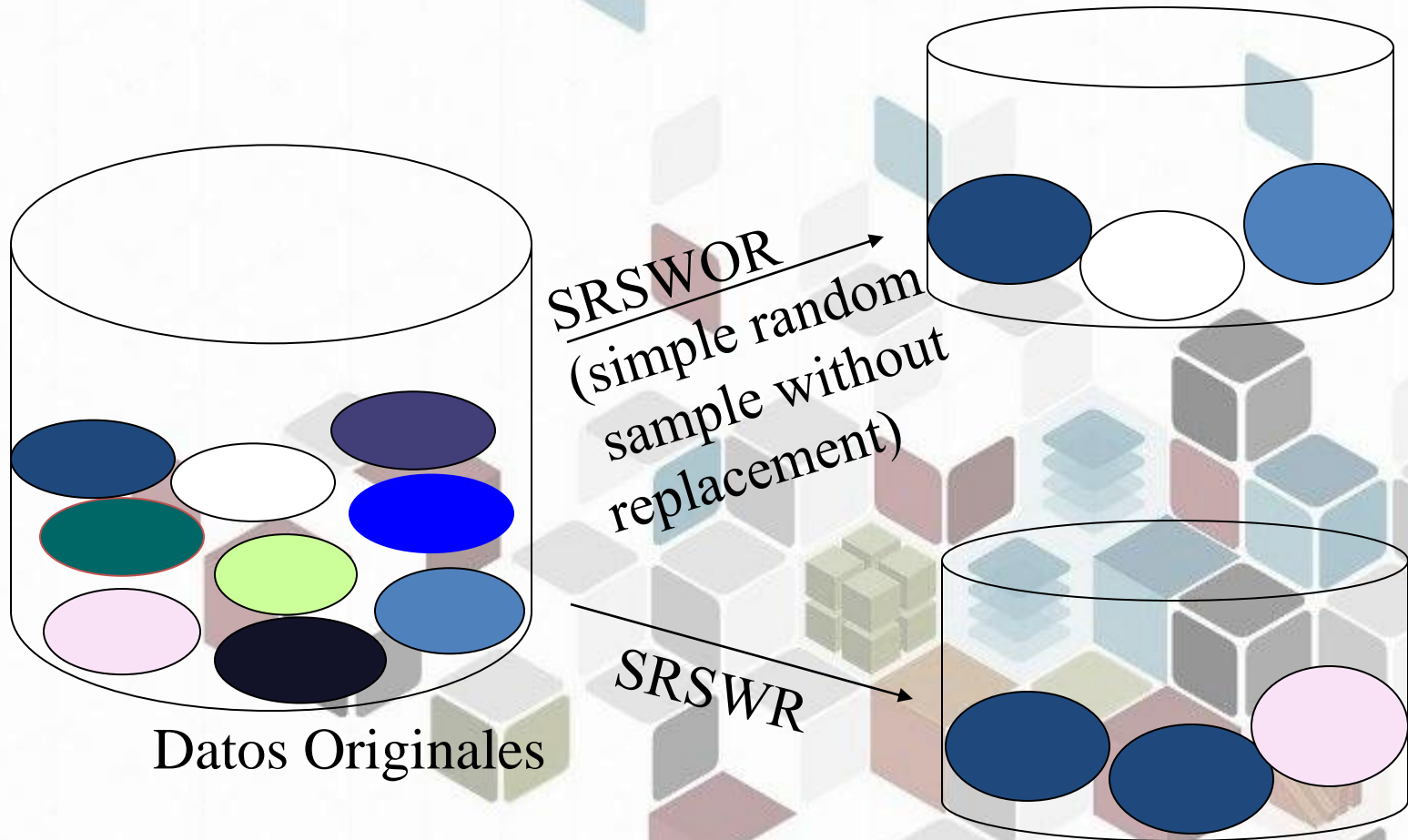
- No Paramétrica

- Muestreo

- *Sample Random sample without replacement*
 - *Sample Random sample with replacement*
 - *Cluster Sample*
 - *Stratified Sample*

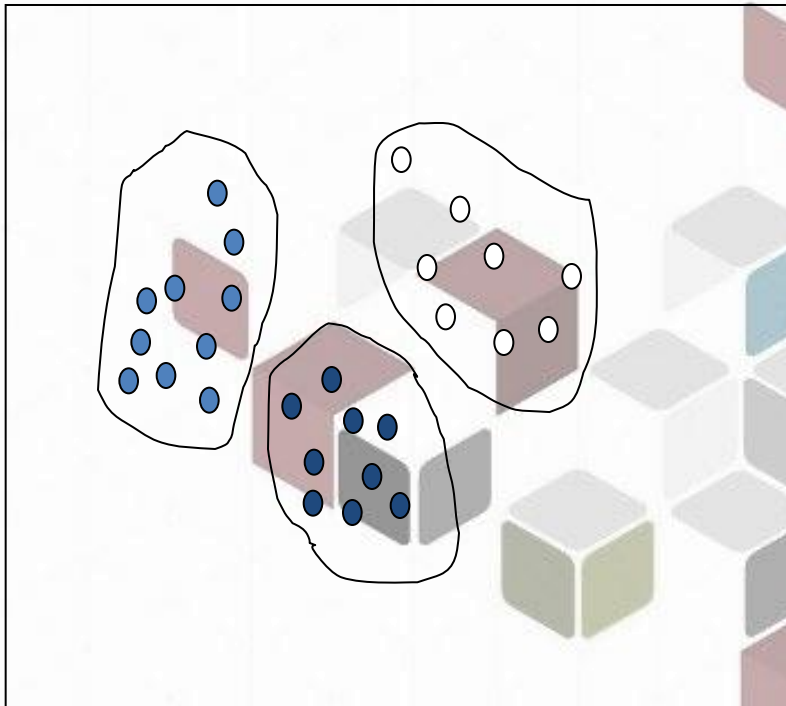


Técnicas – Reducción de datos

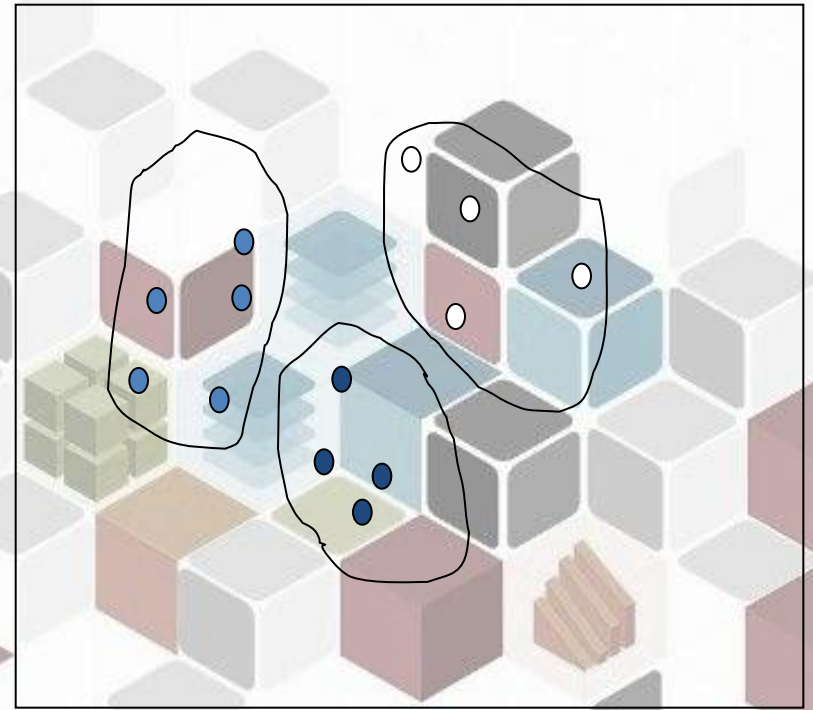


Técnicas – Reducción de datos

Datos Originales

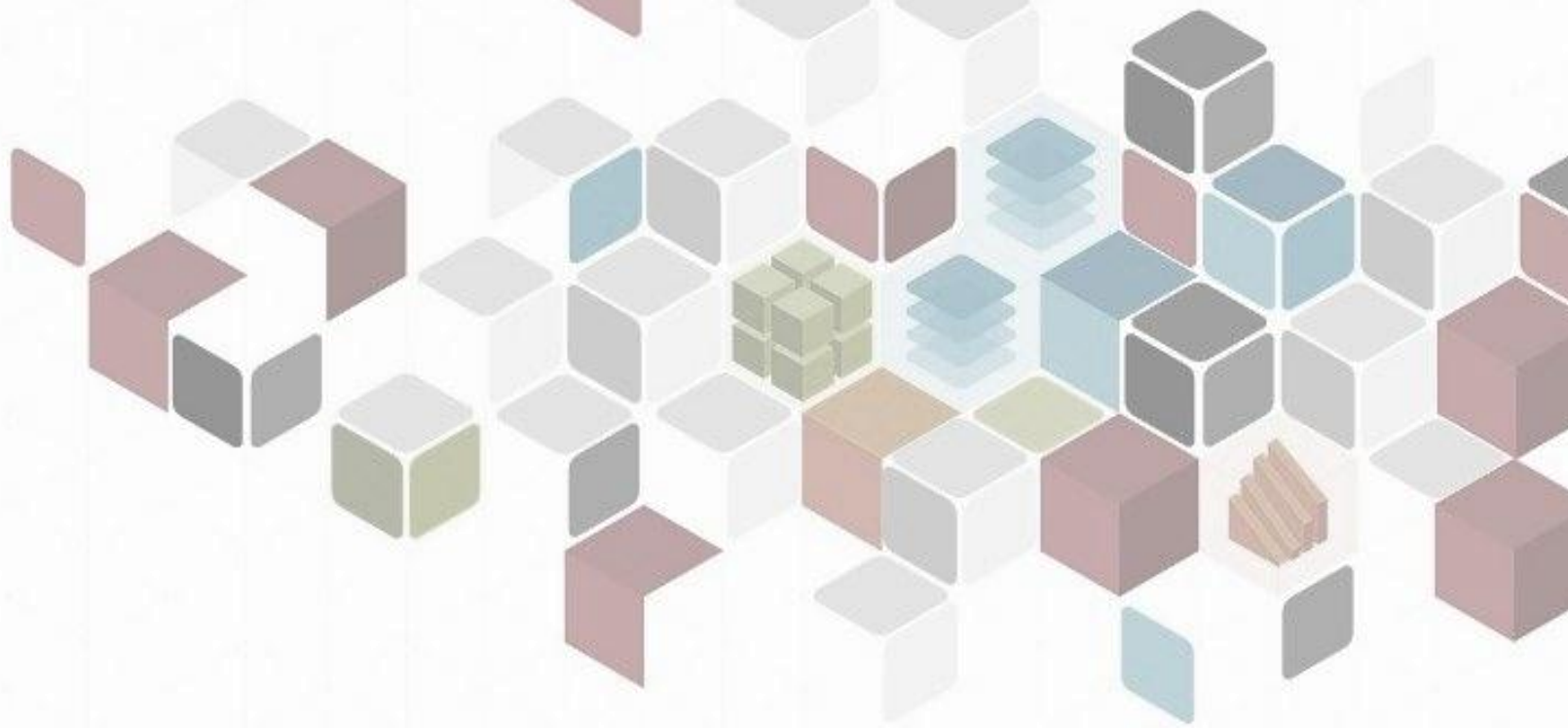


Muestreo Cluster/Estratificado



Técnicas – Reducción de datos

- Muestreo
 - El tiempo del calculo de la muestra es proporcional al tamaño de la muestra
 - La complejidad del muestreo es casi sub-lineal al tamaño de los datos
 - Es de las mas rápidas en responder, pero siempre con un grado de error



Técnicas – Reducción de datos

- Discretización

IdCliente	Nombre	FN	Sexo	Salario
1	Hugo	20/07/1995	M	20550.00
2	Paco	15/05/1998	M	30250.00
3	Luis	02/12/1993	M	15500.00
4	Ana	11/04/1990	F	15000.00
5	Sara	31/08/1992	F	21250.00
6	Sandy	1/31/1994	F	12000.50
7	Sol	3/21/1994	F	27500.50
8	Rene	12/22/1995	M	18350.50
9	Azul	10/15/1991	F	19450.50
10	Juan	7/8/1999	M	15500.50

IdCliente	Nombre	FN	Sexo	Salario
1	Hugo	20/07/1995	M	Medio
2	Paco	15/05/1998	M	Alto
3	Luis	02/12/1993	M	Bajo
4	Ana	11/04/1990	F	Bajo
5	Sara	31/08/1992	F	Medio
6	Sandy	1/31/1994	F	Bajo
7	Sol	3/21/1994	F	Medio
8	Rene	12/22/1995	M	Medio
9	Azul	10/15/1991	F	Alto
10	Juan	7/8/1999	M	Medio

Técnicas – Reducción de datos

- Jerarquías de concepto

- Reducir el número de valores en atributos continuos dividiendo el rango del mismo en intervalos
- Discretización
 - Supervisada
 - Si hace uso de la información de clases
 - Sin Supervisar
 - Proceso
 - Discretización de arriba-abajo
 - Discretización de abajo-arriba