Clase 2
Tipos de Datos

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

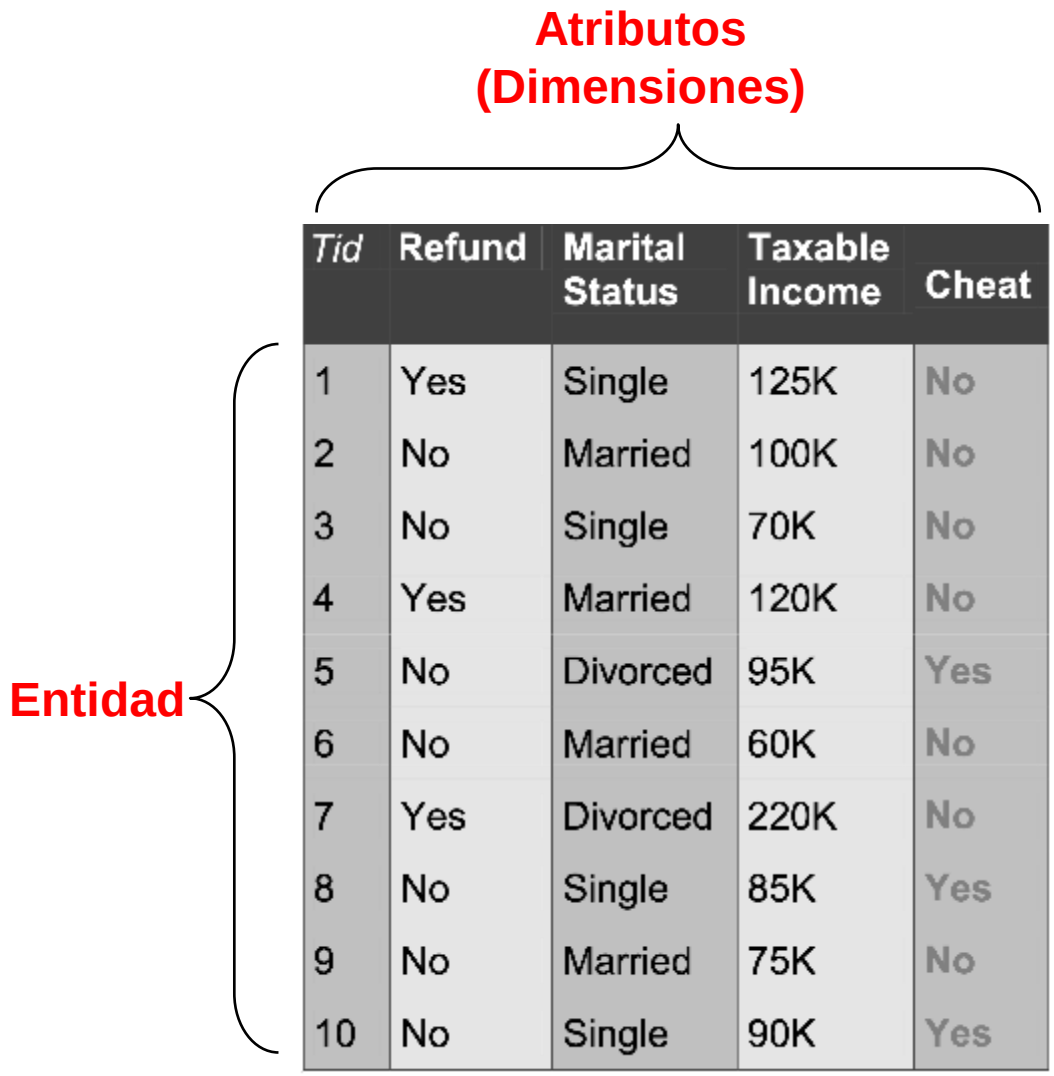
@raimun2

- **Horario ayudantía**
- **Proyecto 1**

¿Qué son los datos?

- Colección de entidades y sus atributos
- Atributo: propiedad o característica de una entidad (p. Ej., Color de ojos, temperatura)
- Entidad: colección de atributos

  Aka: registro, punto, caso, muestra, objeto o instancia
- Los valores de los atributos son números o signos asignados al atributo



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

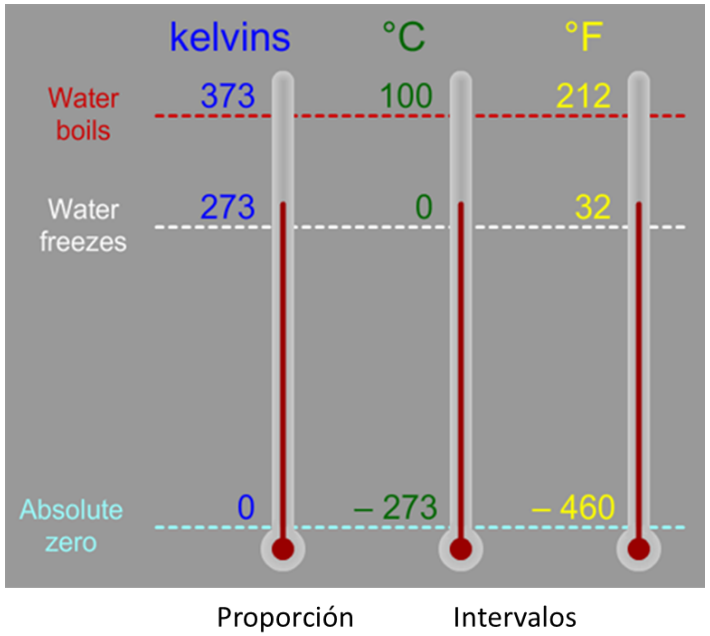
Ejemplo: Datos de personas (csv)

edad, clase de trabajo, peso final, educación, número de educación, estado civil, ocupación, relación, raza, sexo, ganancia de capital, pérdida de capital, horas por semana, país de origen, salario anual
39, gobierno estatal, 77516, solteros, 13, nunca casados, administrativo, no en la familia, blanco, hombre, 2174, 0, 40, Estados Unidos, <= 50K
50, Self-emp-not-inc, 83311, Solteros, 13, Cónyuge-civil-casado, Ejecutivo-gerencial, Esposo, Blanco, Hombre, 0, 0, 13, Estados Unidos, <= 50K
40, Privado, 121772, Assoc-voc, 11, Cónyuge casado, Reparación artesanal, Esposo, Asian-Pac-Islander, Hombre, 0, 0, 40,?,> 50K
52, Self-emp-not-inc, 209642, HS-grad, 9, Casado-civ-cónyuge, Ejecutivo-gerencial, Esposo, Blanco, Hombre, 0, 0, 45, Estados Unidos,> 50K

Tipo de medidas

- **Nominal:** Valores categóricos, sin ningún orden
- **Ordinal:** Valores ordenados, sin distancia significativa entre puntos.
- **Intervalo:** Valores ordenados, con distancia significativa entre puntos.
- **Proporción:** Valores ordenados, con una distancia significativa entre puntos y una definición clara de cero.
- **Discreto:** Tiene solo un conjunto de valores finito o numerablemente infinito
- **Continuo :** Tiene números reales como valores de atributo.

Ordinal



Nominal

Tipos de datos

Tipos de datos: Datos Tabulares (Tidy Data)

- Colección de registros, cada uno de los cuales consta de un conjunto fijo de atributos.
- Características
  - Cada variable tiene que ser en una columna.
  - Cada observación tiene que ser en una fila diferente.
  - Si tienes múltiples tablas, debe existir una columna en cada tabla que permita enlazarlas.

Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

Tipos de datos: Datos de documentos (Minería de texto)

Cada documento se representa como un vector de término, donde cada atributo registra el número de veces que el término aparece en el documento.

	Docs																			
Terms	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
data	1	1	0	0	2	0	0	0	0	0	1	2	1	1	1	0	1	0	0	0
examples	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
introduction	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
mining	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0
network	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1
package	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Tipos de datos: Datos transaccionales
(Reglas de asociación)

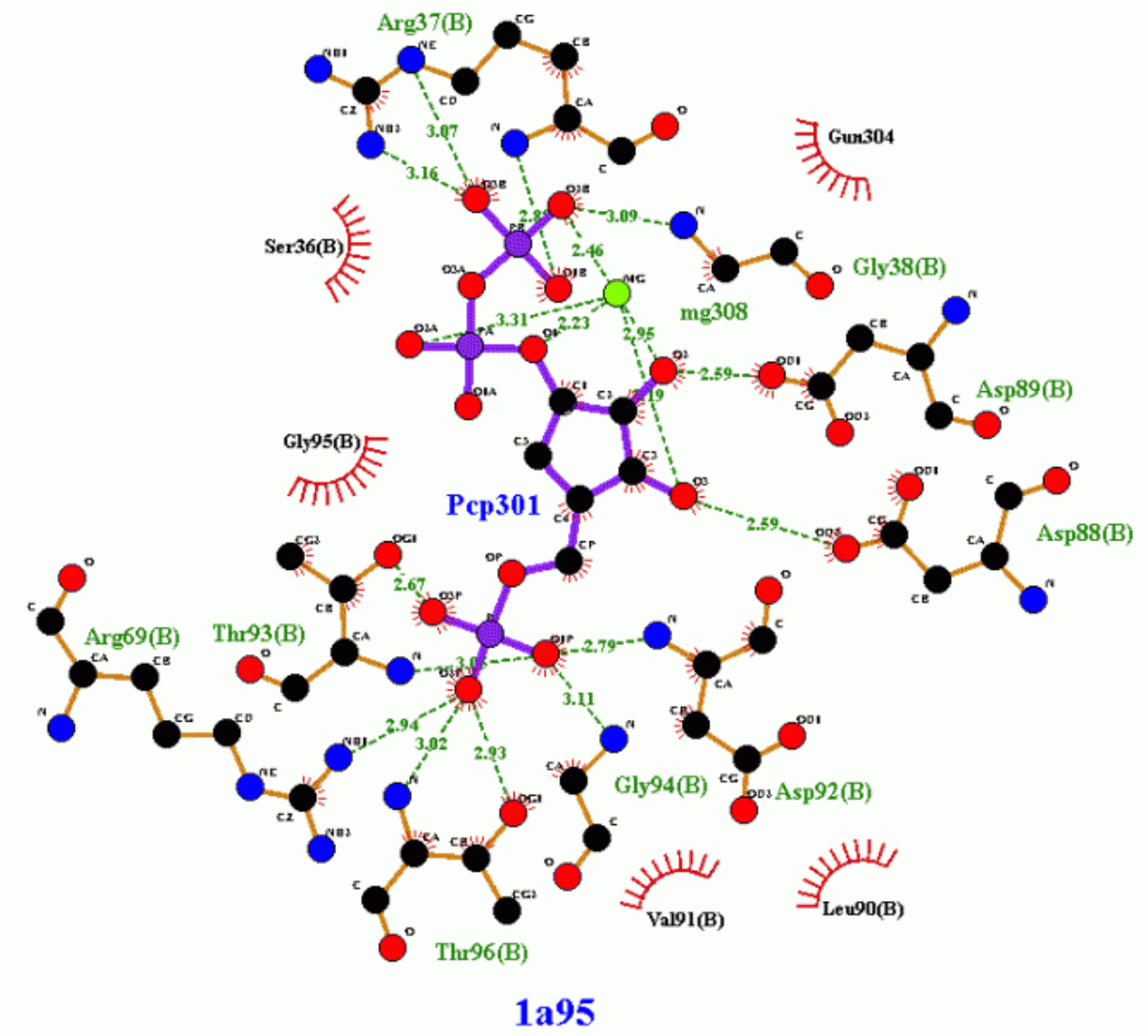
Cada registro corresponde a una transacción que involucra un conjunto de elementos

Table 6.22. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	{a,d,e}
1	0024	{a,b,c,e}
2	0012	{a,b,d,e}
2	0031	{a,c,d,e}
3	0015	{b,c,e}
3	0022	{b,d,e}
4	0029	{c,d}
4	0040	{a,b,c}
5	0033	{a,d,e}
5	0038	{a,b,e}

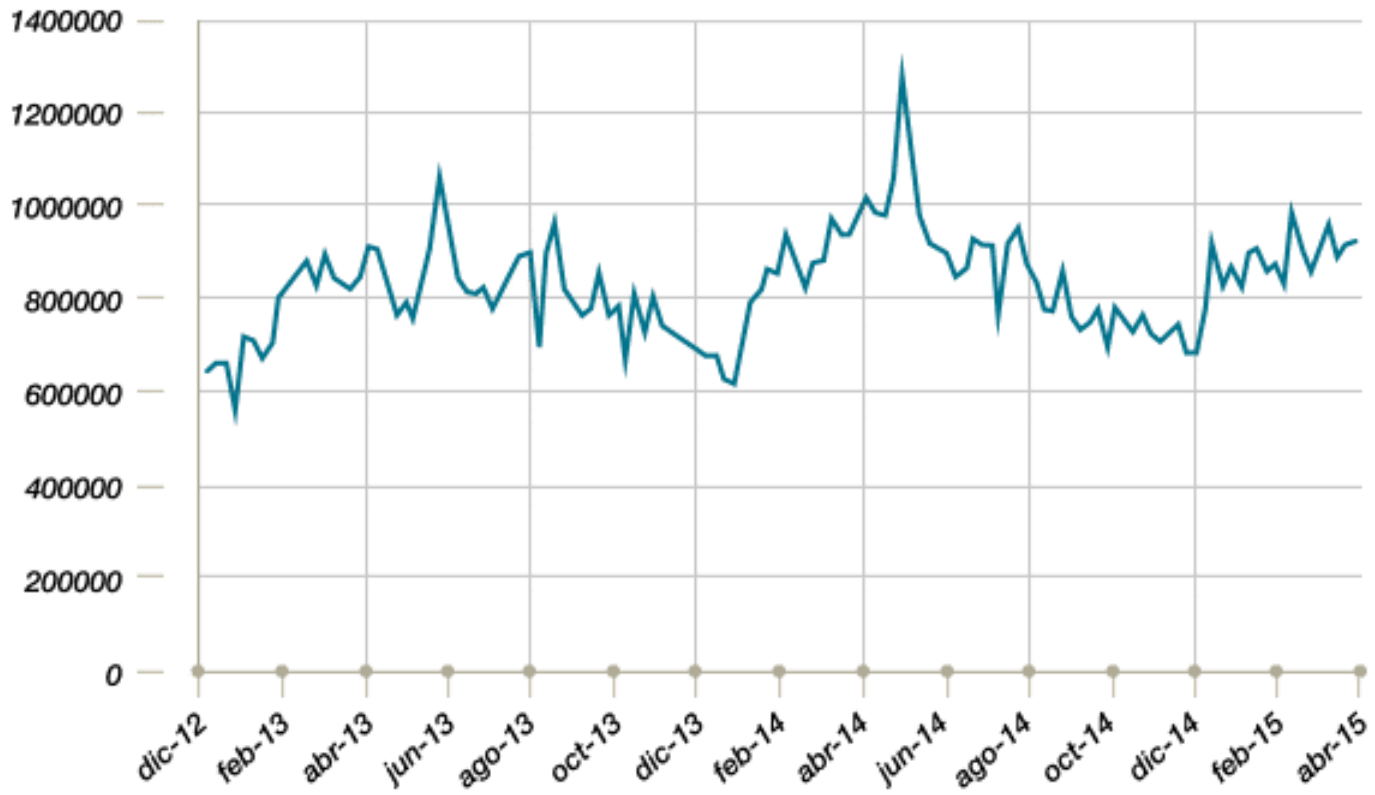
Tipos de datos: Datos de redes (Grafos)

Los nodos corresponden a entidades, los bordes corresponden a relaciones



Tipos de datos: Series de tiempo
(Análisis de series temporales)

Datos generados a través de un proceso continuo en el tiempo.



Tipos de variables

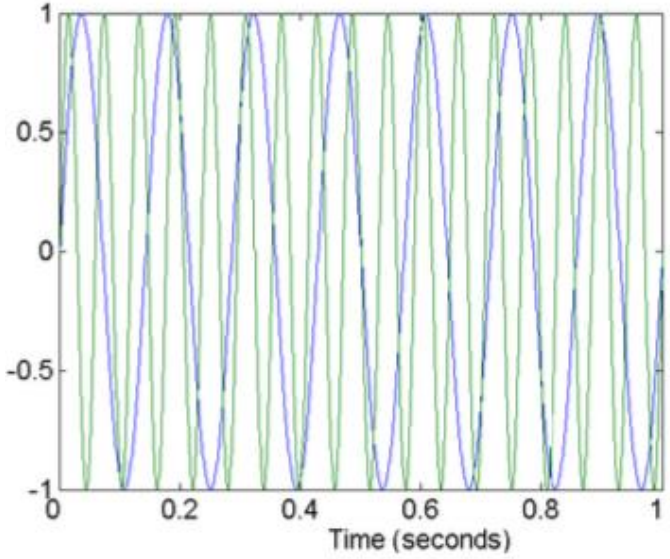
- Character
- Integer
- Double
- Date
- Logical
- Factor
- Matrix
- Vector
- List
- Dataframe
- Tibble

Calidad de los datos

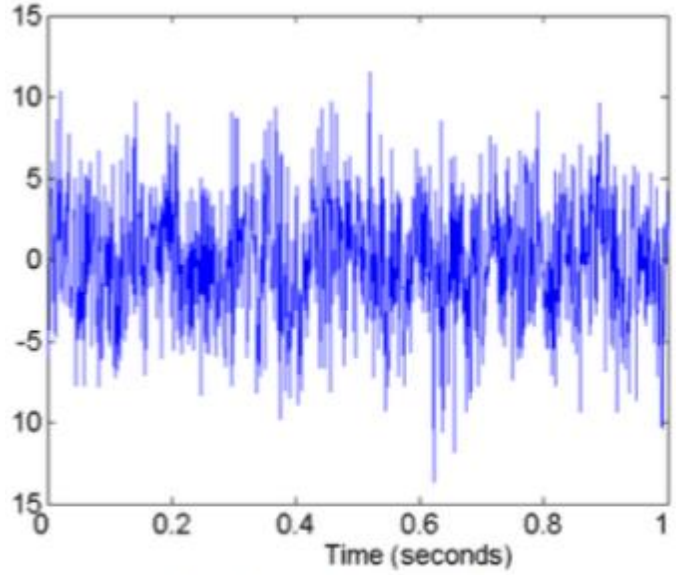
Muchas veces los datos recopilados presentan algunos problemas importantes como, ruido, valores atípicos, valores faltantes o datos duplicados

Calidad de los datos: Ruido

- El ruido se refiere al error de medición en los valores de los datos.
- Podría ser un error aleatorio o un error sistemático
- Graficar o usar técnicas como la descomposición de Fourier puede ayudar a detectar estos casos.



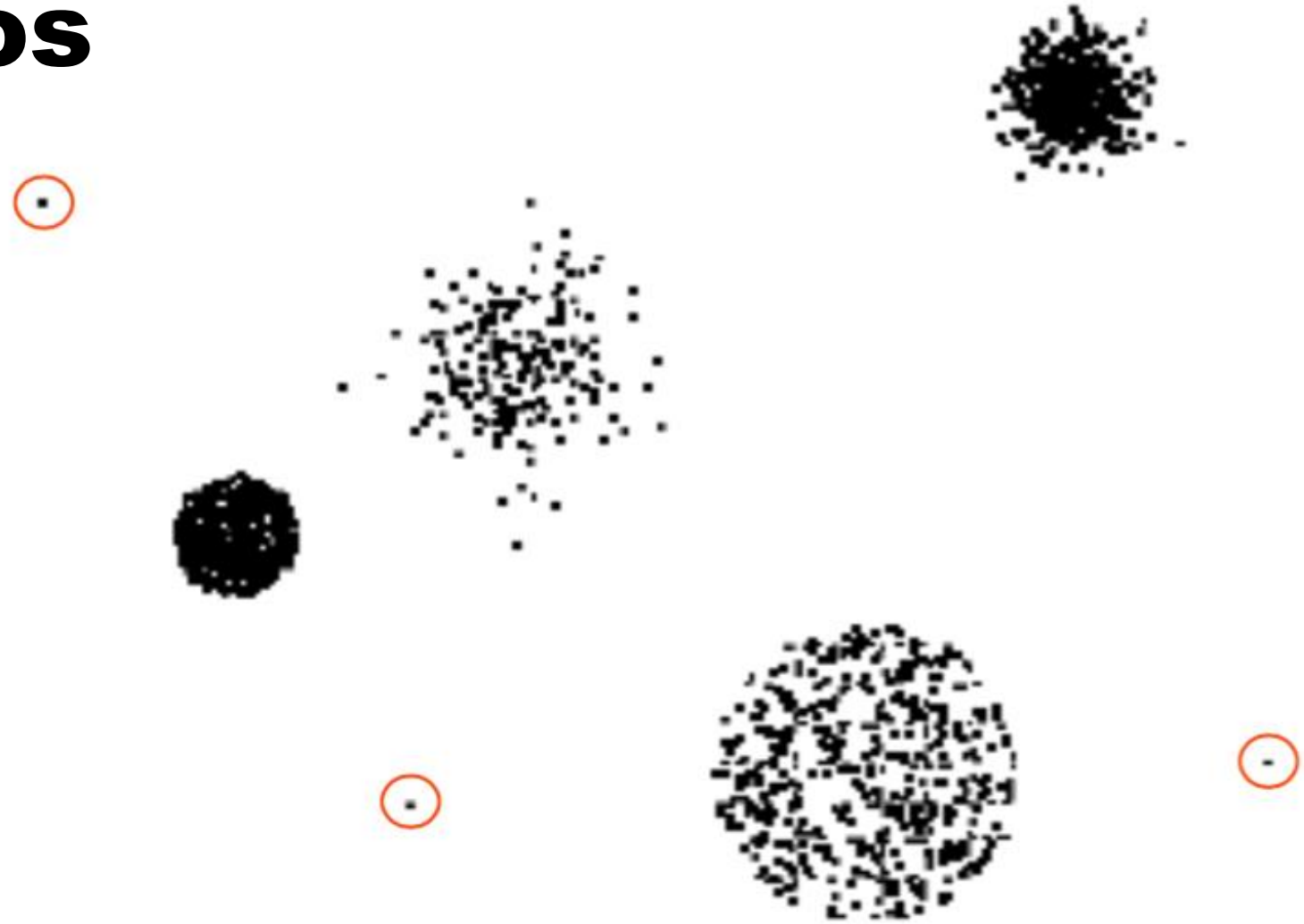
Two Sine Waves



Two Sine Waves + Noise

Calidad de los datos: Valores atípicos

- Valores atípicos son datos con características diferentes al resto del conjunto de datos.
- Pudo indicar casos genuinamente atípicos, o casos erróneos.
- Visualizar la data permite identificar estos casos.



Calidad de los datos: Valores faltantes

- Razones de los valores perdidos
  - No se recopila información (por ejemplo, las personas se niegan a dar su edad)
  - Es posible que los atributos no se apliquen a todos los casos (por ejemplo, el ingreso anual no se aplica a los niños)
- Maneras de manejar los valores perdidos
  - Eliminar entidades con valores perdidos
  - Estimar atributos con valores perdidos
  - Ignore los valores perdidos durante el análisis
  - Imputar valores perdidos

U/A	C		
	Temperature	Headache	Nausea
1	High	?	no
2	very-high	yes	yes
3	?	no	yes
4	High	yes	?
5	?	no	no
6	normal	yes	yes
7	Very-high	no	yes

Calidad de los datos: Valores duplicados

- El conjunto de datos puede incluir entidades de datos duplicadas o casi duplicadas entre sí.
- Problema surge al fusionar datos de fuentes heterogéneas

Pre procesamiento de datos

Proceso que prepara la data para el análisis, resolviendo asuntos sobre la calidad de los datos, y también seleccionando las piezas de información necesarias para realizar el análisis posterior.

- Encontrar y tratar con entidades duplicadas
- Encontrar y corregir errores de medición
- Lidiar con los valores perdidos
- Normalización / Estandarización
- Muestreo
- Selección de variables
- Agregación
- Discretización
- Reducción de dimensionalidad

Pre procesamiento de datos: Normalización / Estandarización

- Muchos algoritmos de ML calculan distancias entre puntos
- Las variables tienen diferente escala, lo que lleva a conclusiones erróneas.
- Cuando los datos están estandarizados, las variables numéricas deben tener una escala similar entre ellas.

Escalador

$$y_i^j = \frac{x_i^j - \min(x_1^j, \ldots, x_n^j)}{\max(x_1^j, \ldots, x_n^j) - \min(x_1^j, \ldots, x_n^j)}$$

Los valores varían entre 0 y 1.

No es resistente a valores atípicos.

Estandarización

$$z_i^j = \frac{x_i^j - \bar{x}_i}{s_j}$$

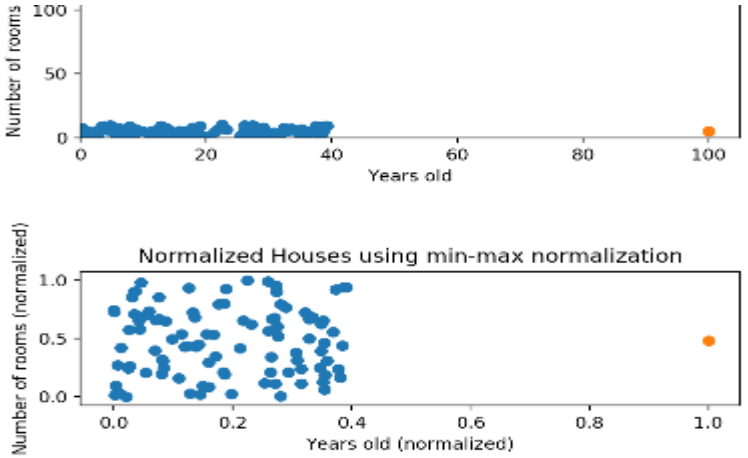
Más robusto para oultiers.

Asume distribución normal

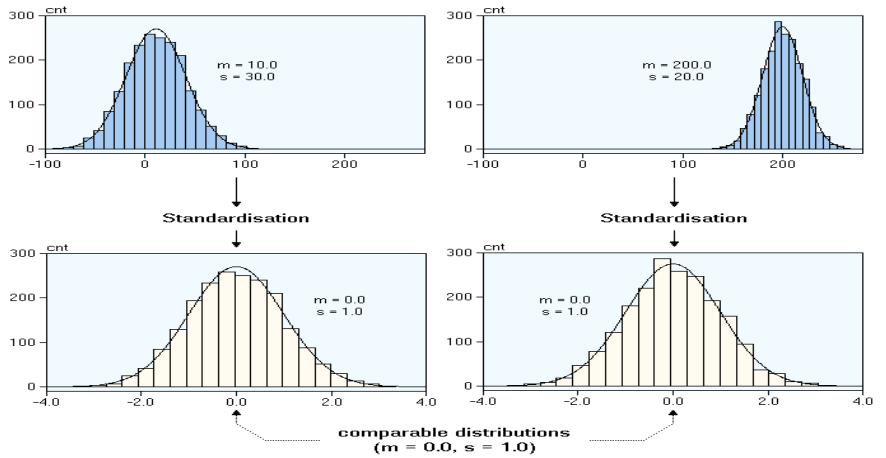
Pre procesamiento de datos: Normalización / Estandarización

- Muchos algoritmos de ML calculan distancias entre puntos
- Las variables tienen diferente escala, lo que lleva a conclusiones erróneas.
- Cuando los datos están estandarizados, las variables numéricas deben tener una escala similar entre ellas.

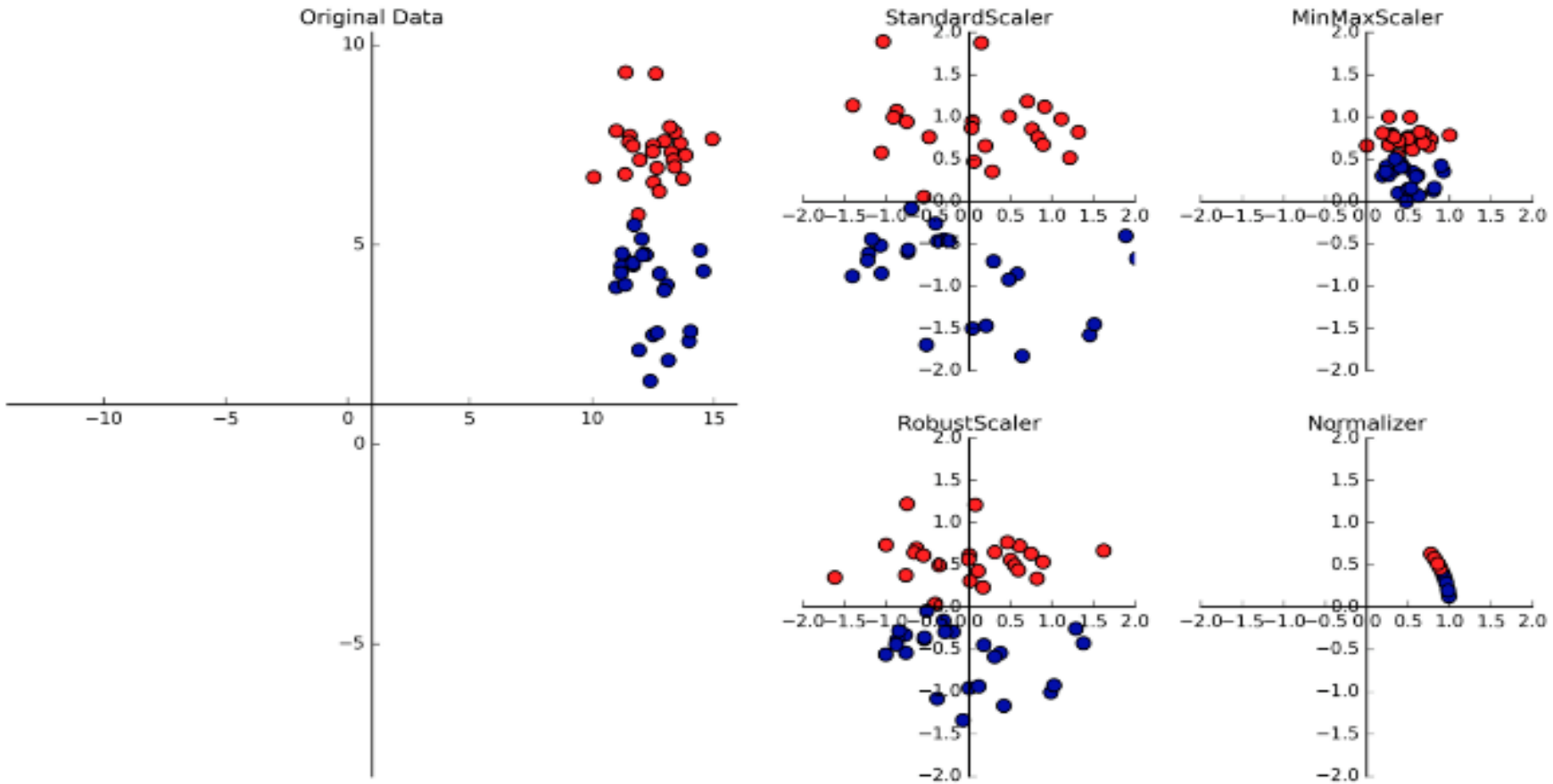
Escalador



Estandarización

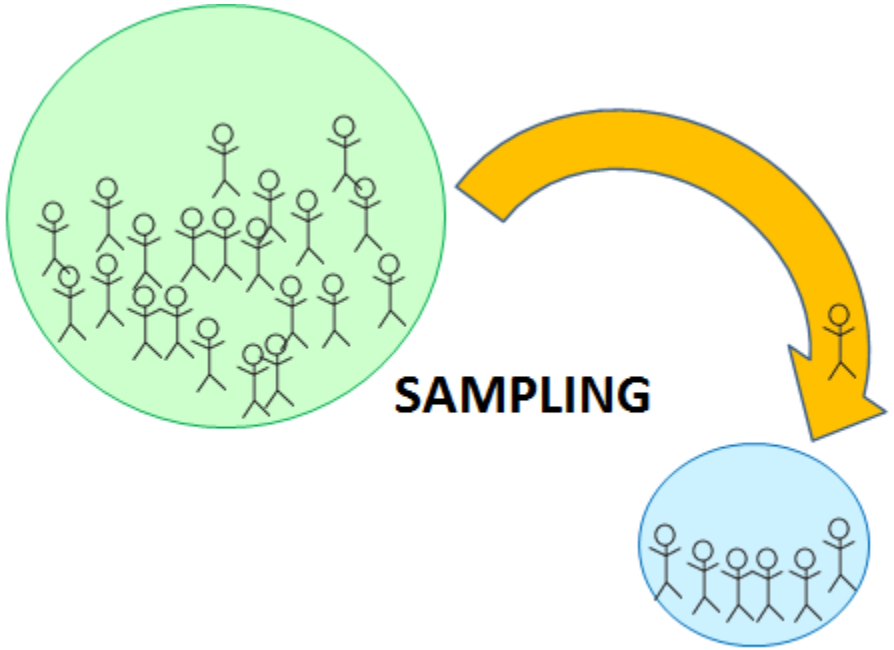


Pre procesamiento de datos: Normalización / Estandarización



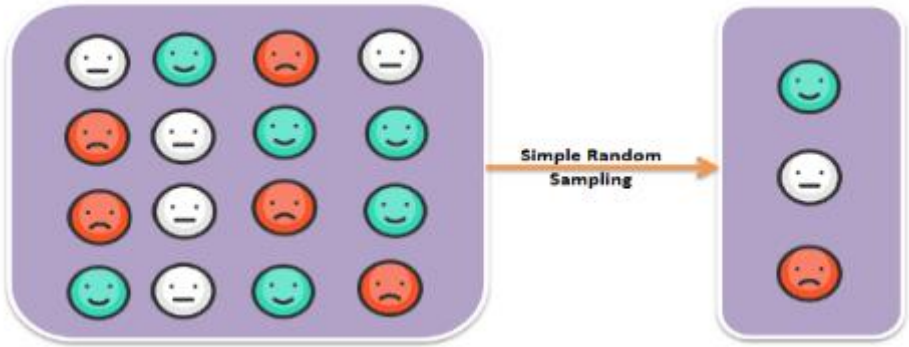
Pre procesamiento de datos: Muestreo

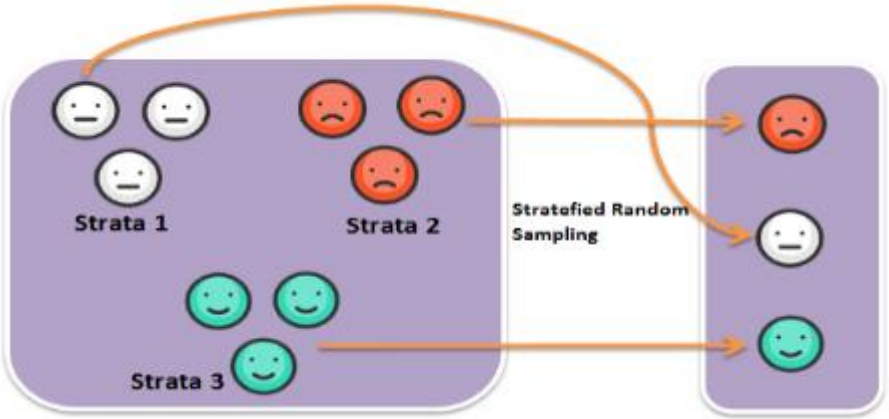
- El muestreo es la principal técnica empleada para la selección de datos.
- Se utiliza porque procesar / obtener el conjunto completo de datos de interés es demasiado caro o requiere mucho tiempo (BIG DATA).
- La muestra debe ser representativa (tiene aproximadamente la misma propiedad (de interés) que el conjunto original de datos).



Pre procesamiento de datos: Tipos de muestreo:

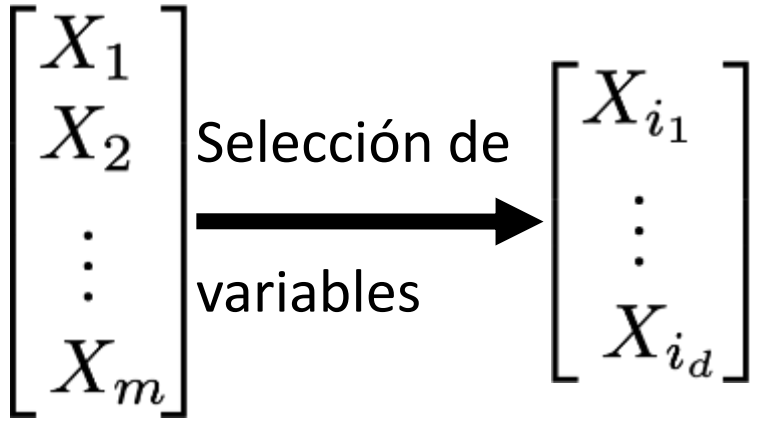
- Muestreo aleatorio simple: Existe la misma probabilidad de seleccionar cualquier elemento en particular.
- Muestreo aleatorio estratificado: divida los datos en varias particiones de interés; luego extrae muestras aleatorias de cada partición.
- Muestreo sin reemplazo: a medida que se selecciona cada elemento, se elimina de la población.
- Muestreo con reemplazo: los objetos no se eliminan de la población a medida que se seleccionan para la muestra.





Pre procesamiento de datos: Selección de variables

- Dado un conjunto de **m** variables, un subconjunto de **d** las variables están seleccionadas (**d** < **m**).
- Variables redundantes: duplican gran parte o toda la información contenida en uno o más atributos.
- Variables irrelevantes: no contienen información que sea útil para la tarea de minería de datos en cuestión



Pre procesamiento de datos: Agregación

- Proceso donde la información se combina disminuyendo su complejidad, mediante la reducción del número de variables y / o registros.
- Disminuye la variabilidad de la data
- **Agregación de registros**: los sensores generan datos cada milisegundo. Agregue los datos para obtener la información cada segundo.
- **Agregación variable**: dos variables que muestran la entrada y salida de un sensor. Agregue la variable para obtener la diferencia entre estas variables.

user	day	transaction
A	1	10
A	1	-100
B	1	5
A	1	5
B	1	-500
B	2	-100
A	2	60
A	2	65
B	2	54
A	2	-30

→

user	day	transaction
A	1	-85
B	1	-495
A	2	95
B	2	-46

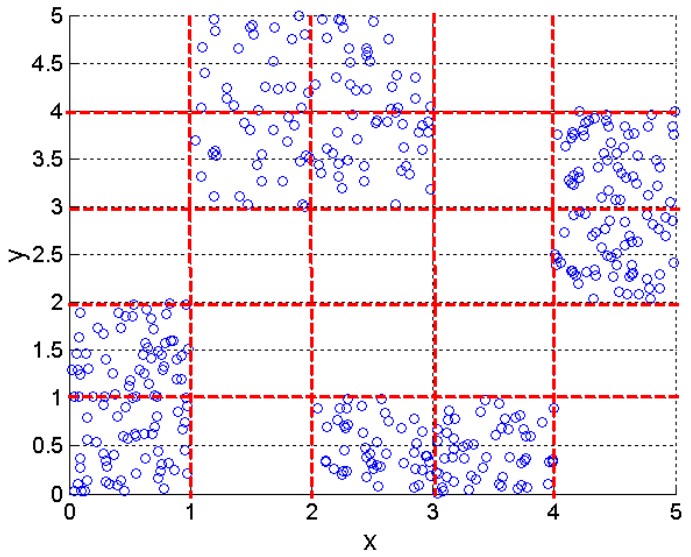
id	t1	t2	t3	t4	t5
1	21	16	32	20	23
2	35	17	31	30	25
3	39	50	30	30	19
4	25	34	35	27	15
5	39	23	33	20	15
6	30	18	31	26	22
7	21	24	31	29	18
8	26	20	34	28	17
9	26	19	30	22	25
10	27	48	33	29	15

→

id	prom t	DesvEst
1	16,25	8,43
2	18,10	7,92
3	20,15	7,44
4	18,20	6,76
5	19,80	11,63
6	21,05	11,08
7	17,90	7,57
8	20,65	7,64
9	19,50	11,68
10	18,80	9,44

Pre procesamiento de datos: Discretización

- La discretización es el proceso de poner valores en depósitos para que haya un número limitado de estados posibles.
- **Discretización**: cambia una variable continua por una categórica.
- Dada la diversidad de los datos (binarios, discretos, ordinales, etc.), cambiar los datos podría mejorar el rendimiento de los modelos.
- **Binarización**: cambia una variable categórica a múltiples variables binarias, funciona para modelos específicos.



Person	Marital status
xxx	Single
yyy	Married
zzz	Divorcee

Categorical to binary

Person	Single	Married	Divorcee
xxx	1	0	0
yyy	0	1	0
zzz	0	0	1

Clase 2
Tipos de Datos

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2