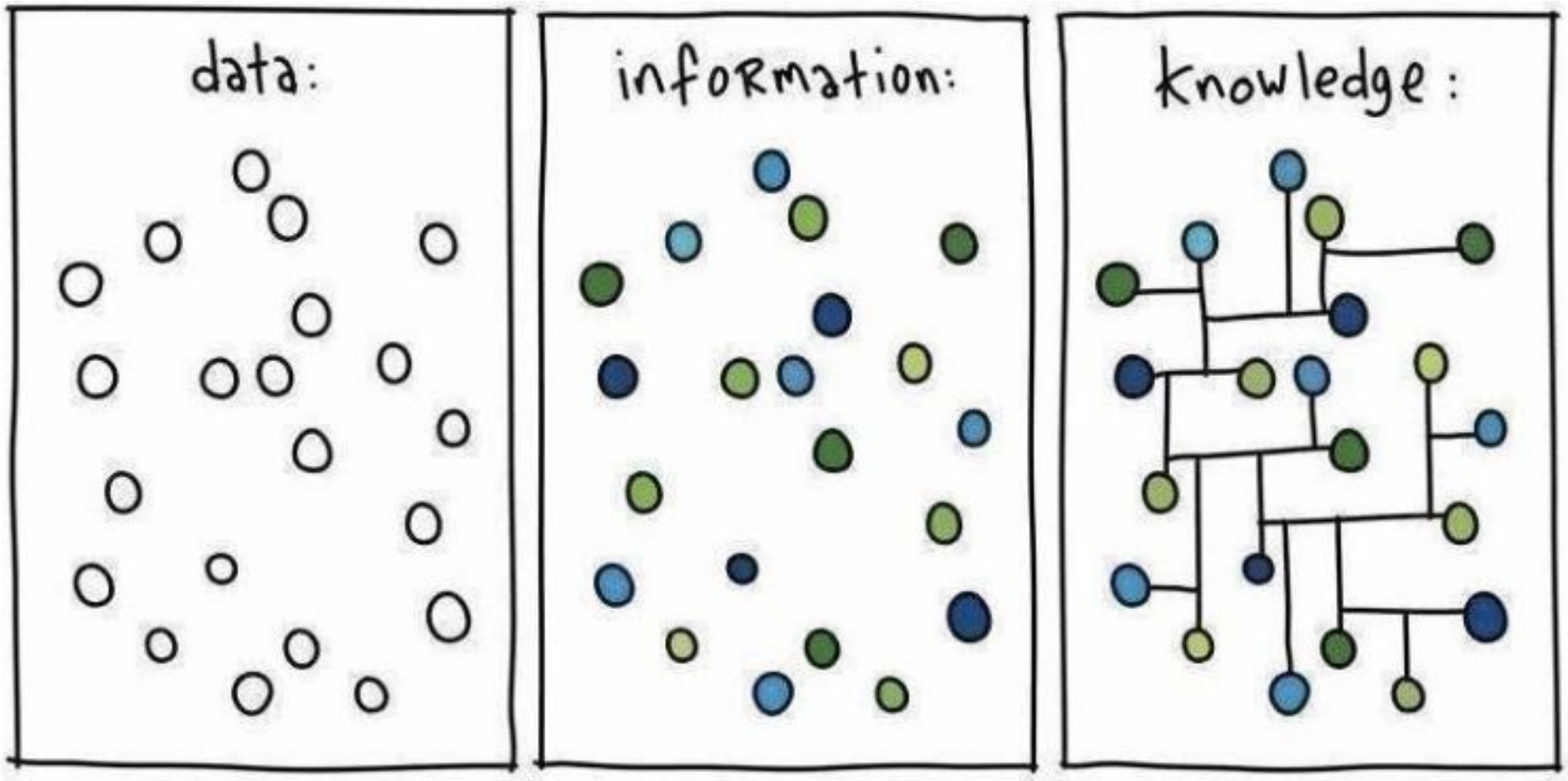
Curso DM
“Datos”

Parte I

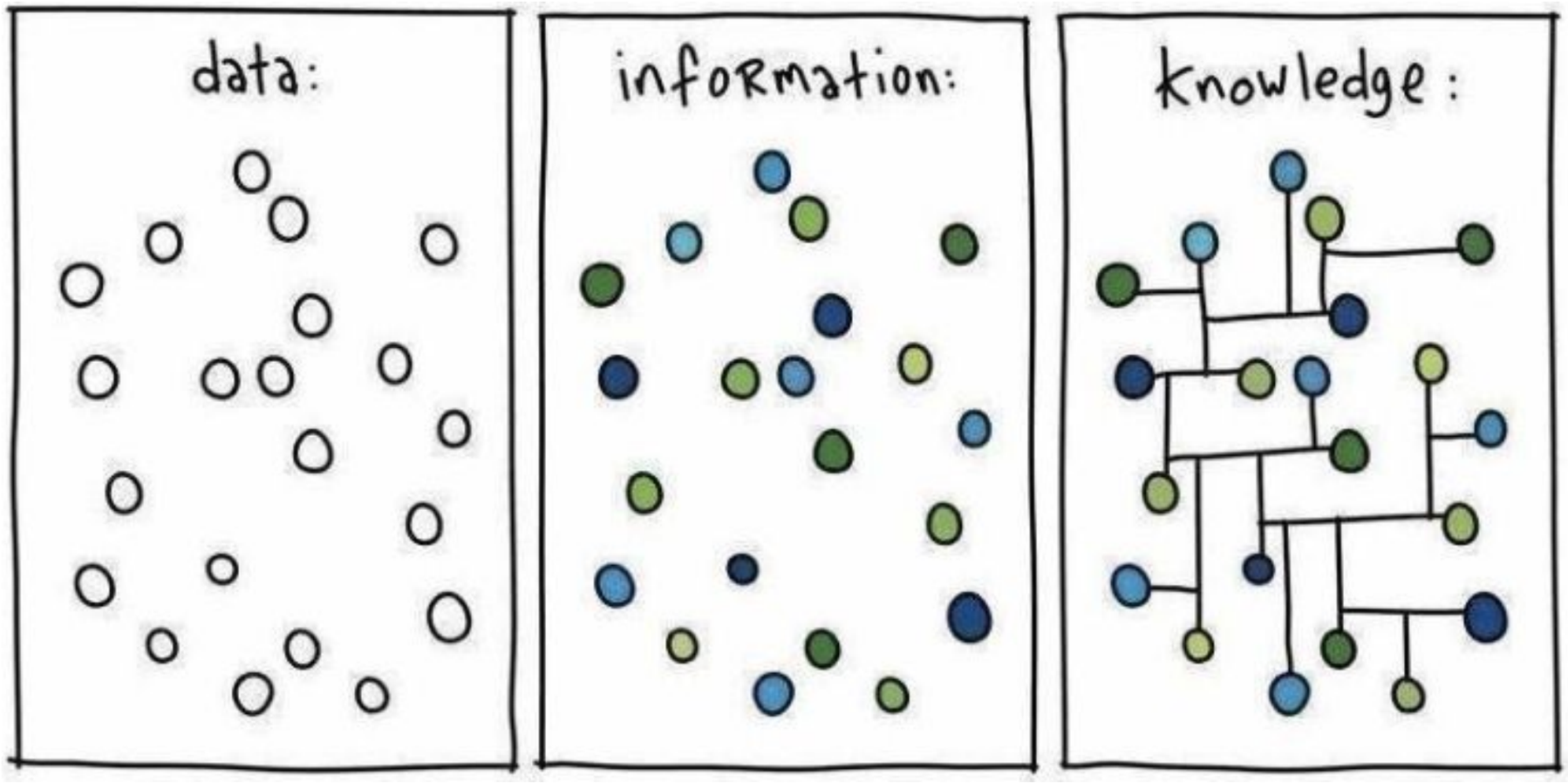
Profesores: Felipe Bravo - Hernán Sarmiento

Basado en las slides de Bárbara Poblete

¿Qué entendemos por dato?



¿Qué entendemos por dato?



Juan Pérez obtuvo un 5.8 en la segunda prueba de historia

Juan Pérez obtuvo mejor nota que el 75% del resto del curso

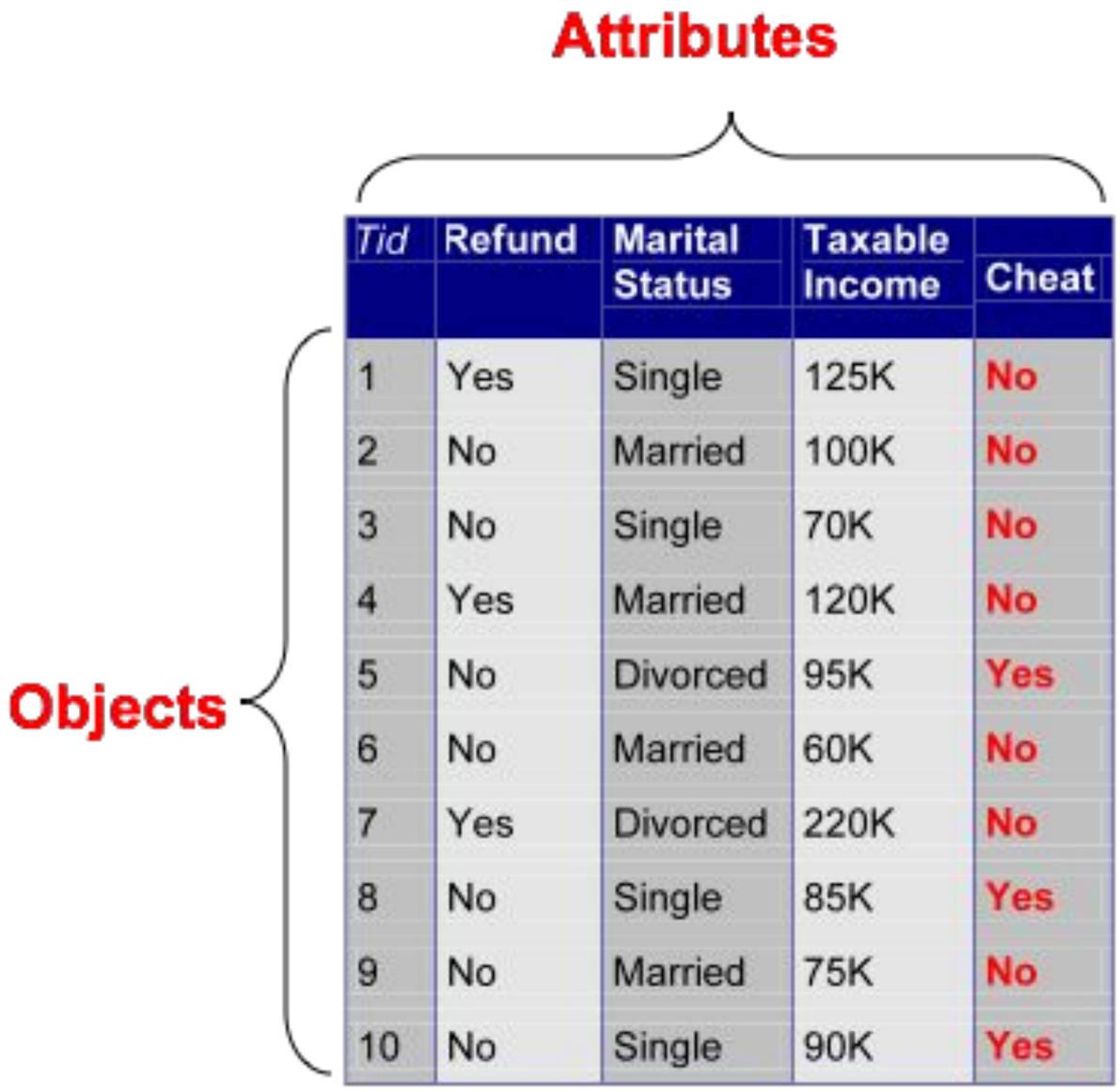
Juan Pérez es un buen alumno

¿Qué entendemos por dato?

- Colección de hechos tales como números, palabras, mediciones, o solo descripción de cosas.

- Utilizado generalmente para **análisis**.

¿Cómo describimos un dato?



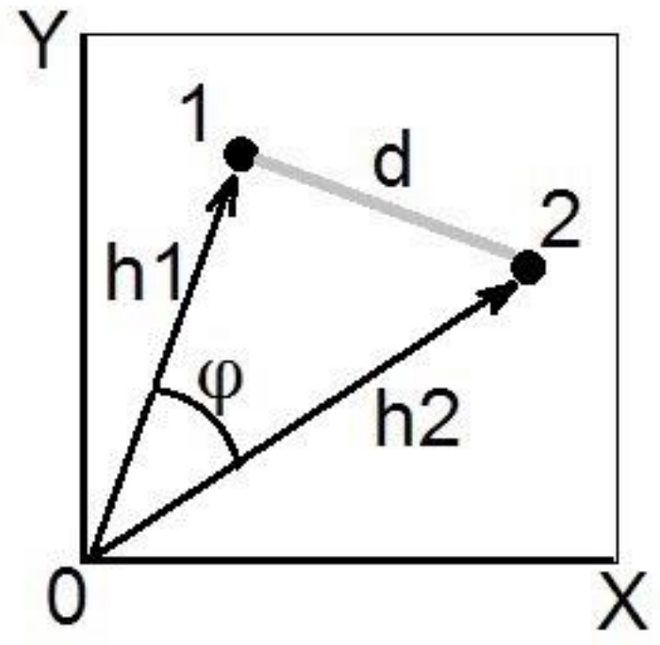
Estructura básica

Dataset, records, atributos

Diferentes enfoques

- Analizar en base a las relaciones entre datos

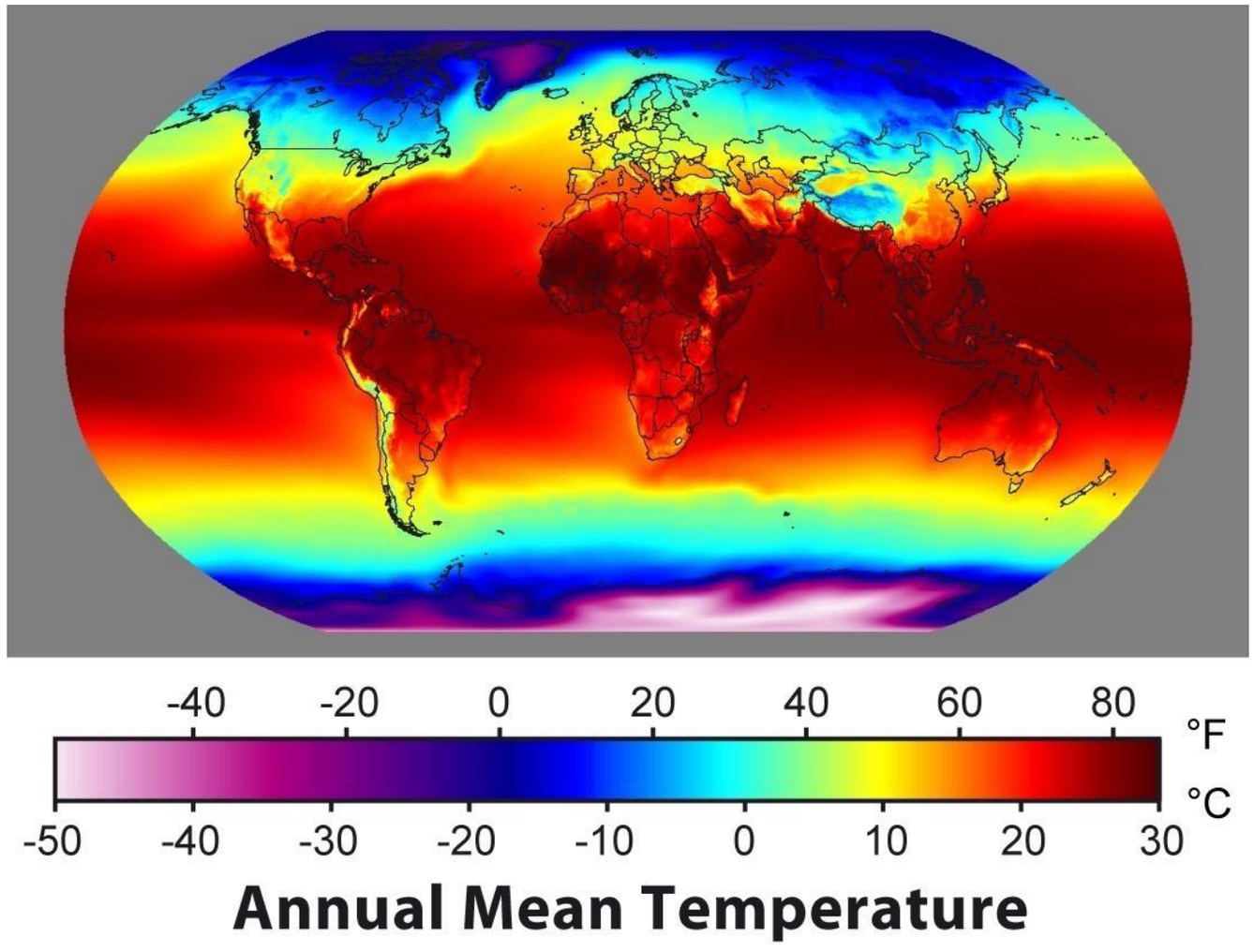
- Extraer relaciones y luego trabajar con estos valores, no con los datos mismos



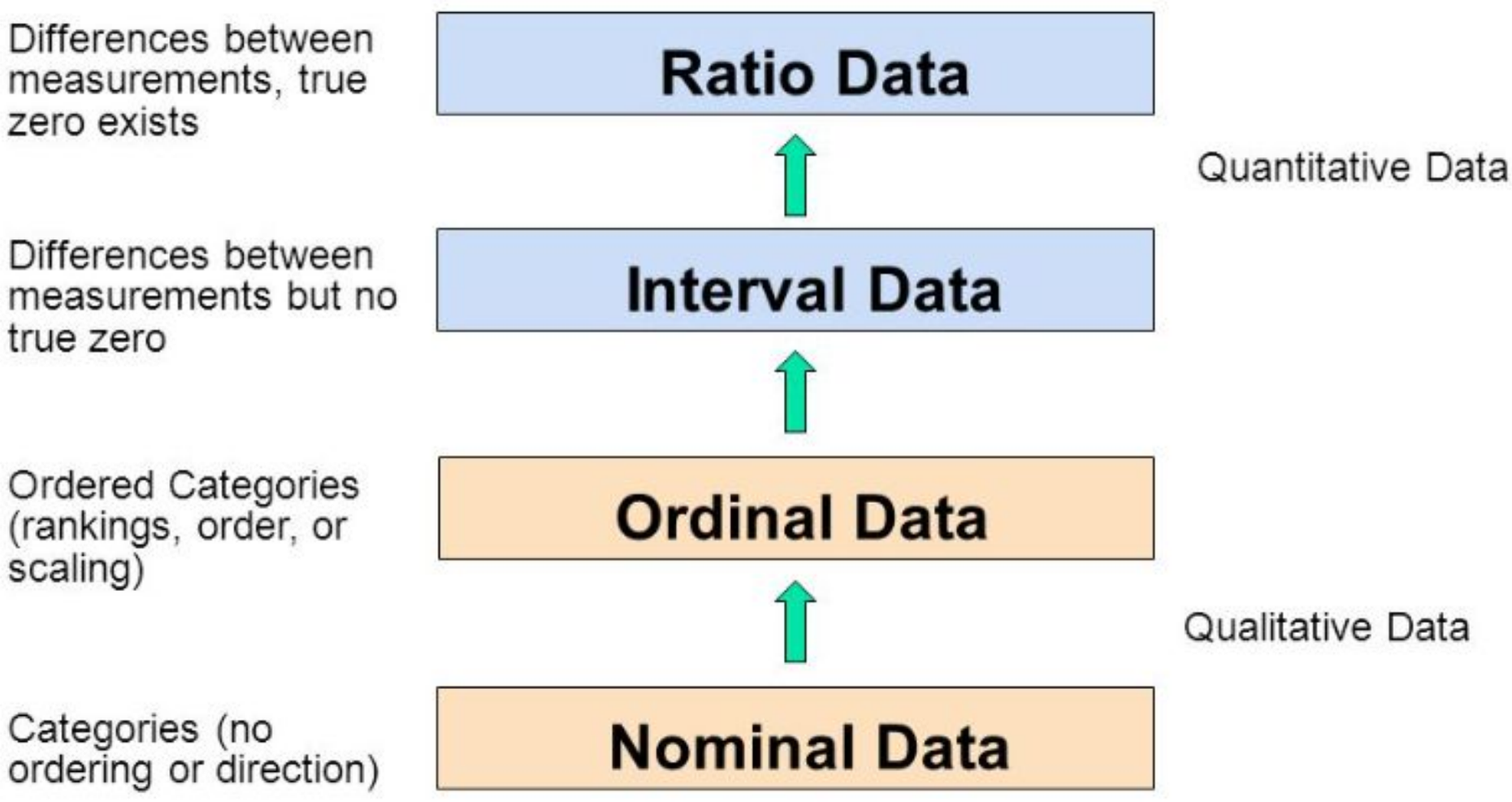
Atributos y métricas

- **Atributo:** propiedad que puede variar de un objeto a otro

- **Escala de medición o métrica:** es necesario definirla de forma exacta, para poder comparar.



TIPOS DE ATRIBUTOS



TIPOS DE ATRIBUTOS

- **NOMINAL** (IDs, color de ojos, categorías de bacterias)
- **ORDINAL** (rankings, notas, altura en {alto, mediano, bajo})
- **INTERVALO** (Fechas, temperaturas ℃ o ℉)
- **RAZÓN** (Temperatura Kelvin, largo, hora)

TIPOS DE ATRIBUTOS

(Operaciones y propiedades)

- **NOMINAL** (IDs, color de ojos, categorías de bacterias)
  Distinción ==, != (solo diferencia de nombre)

- **ORDINAL** (rankings, notas, altura en {alto, mediano, bajo})
  Orden <, <=, >, >= (permite ordenar los datos)

- **INTERVALO** (Fechas, temperaturas ℃ o ℉)
  Adición + -, grado de diferencia entre medidas pero no la razón entre ellos (el cero es arbitrario).

- **RAZÓN** (Temperatura Kelvin, largo, peso)
  Multiplicación * y /

- **CUALITATIVOS (Categóricos)**
  Nominal y Ordinal, aunque sean numéricos deben ser tratados como símbolos

- **CUANTITATIVOS (Numéricos)**
  Intervalo y Razón deben ser tratados como números y tienen propiedades numéricas (continuos o discretos)

Por qué la temperatura en Kelvin es Razón y Celsius es Intervalo?

Operaciones o propiedades

1. **Distinción = y ≠**
2. **Orden <, ≤, > y ≥**
3. **Adición + y -**
4. **Multiplicación * y /**

- NOMINAL (1)
- ORDINAL (1,2)
- INTERVALO (1,2,3)
- RAZÓN (1,2,3,4)

- **Atributos DISCRETOS:**
  - Atributos categóricos.
  - A menudo representados como enteros
  - binarios (finitos)
  - Ej: códigos postales o conjunto de palabras en docs.

- **Atributos CONTINUOS:**
  - Números reales o valores flotantes
  - Ej: temperatura, altura, peso (infinitos)

Formatos de datos

Datos estructurados

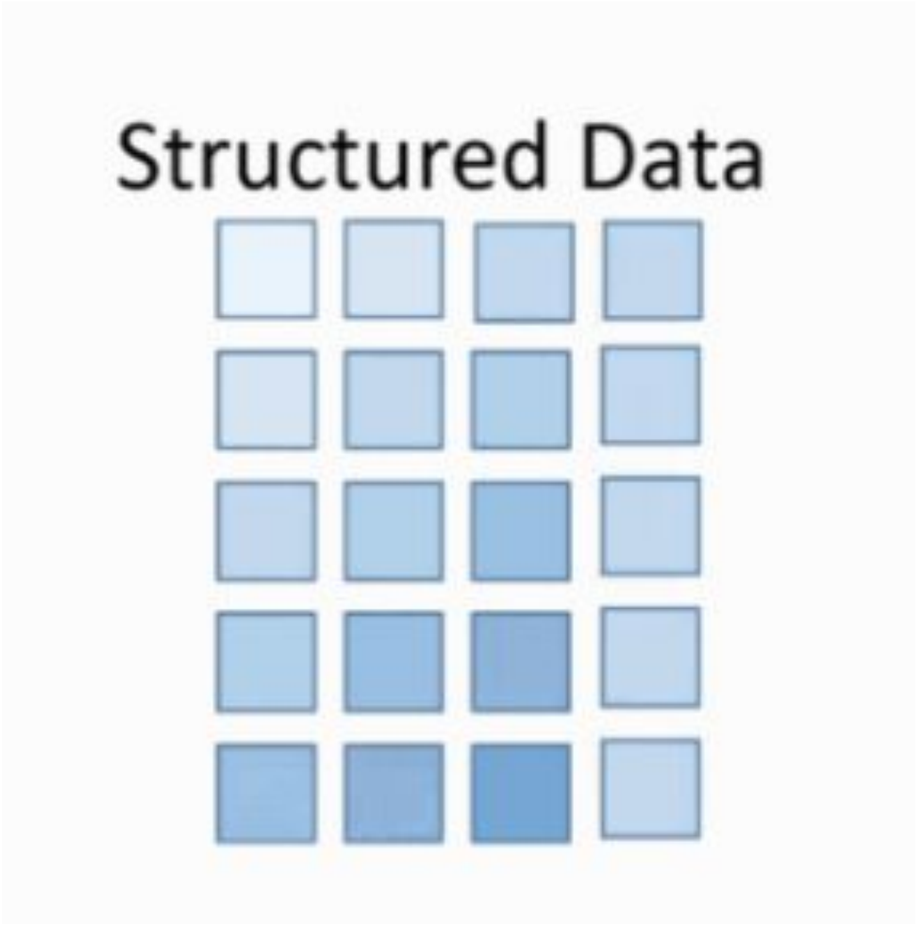
Datos que han sido ***formateados y modelados*** para un “fácil” acceso.

Usualmente se organizan en **tablas.**

Poseen una estructura fija mediante **columnas.**

Un nuevo valor se representa como una **fila.**

Ejemplo: excel, .csv o tablas relacionales (SQL)



Datos estructurados

Nombre mascota	Especie	Raza	Edad	Sexo	Peso
Doris	Canino	Schnauzer	3	Hembra	8
Clotilde	Canino	Mestizo	14	Hembra	7
Tony	Felino	DPC	0.5	Macho	2
Lorenzo	Canino	Mestizo	3	Macho	1.5

Datos estructurados

Nombre mascota	**Especie**	**Raza**	**Edad**	**Sexo**	**Peso**
Doris	Canino	Schnauzer	3	Hembra	8
Clotilde	Canino	Mestizo	14	Hembra	7
Tony	Felino	DPC	0.5	Macho	2
Lorenzo	Canino	Mestizo	3	Macho	1.5

¿Qué **problemas** podríamos tener con **datos** en este formato?

Datos estructurados

Nombre mascota	Especie	Raza	Edad	Sexo	Peso
Doris	Canino	Schnauzer	3	Hembra	8
Clotilde	Canino	Mestizo	14	Hembra	7
Tony	Felino	DPC	0.5	Macho	2
Lorenzo	Canino	Mestizo	3	Macho	1.5

Agregar nuevas columnas:

- Chip
- Celo
- Castración
- Esterilización
- Crías

Datos estructurados

Nombre mascota	Especie	Raza	Edad	Sexo	Peso	Celo	Alergias
Doris	Canino	Schnauzer	3	Hembra	8	NO	SI
Clotilde	Canino	Mestizo	14	Hembra	7	SI	NO
Tony	Felino	Ángora	0.5	Macho	2	??	??
Lorenzo	Canino	Mestizo	3	Macho	1.5	??	SI

Agregar nuevas columnas:

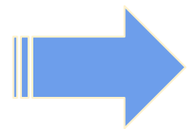
- Chip
- Celo
- Castración
- Esterilización
- Alergías

Datos estructurados

Nombre mascota	Especie	Raza	Edad	Sexo	Peso	Celo	Alergias
Doris	Canino	Schnauzer	3	Hembra	8	NO	SI
Clotilde	Canino	Mestizo	14	Hembra	7	SI	NO
Tony	Felino	Ángora	0.5	Macho	2	??	??
Lorenzo	Canino	Mestizo	3	Macho	1.5	??	SI

Agregar nuevas columnas:

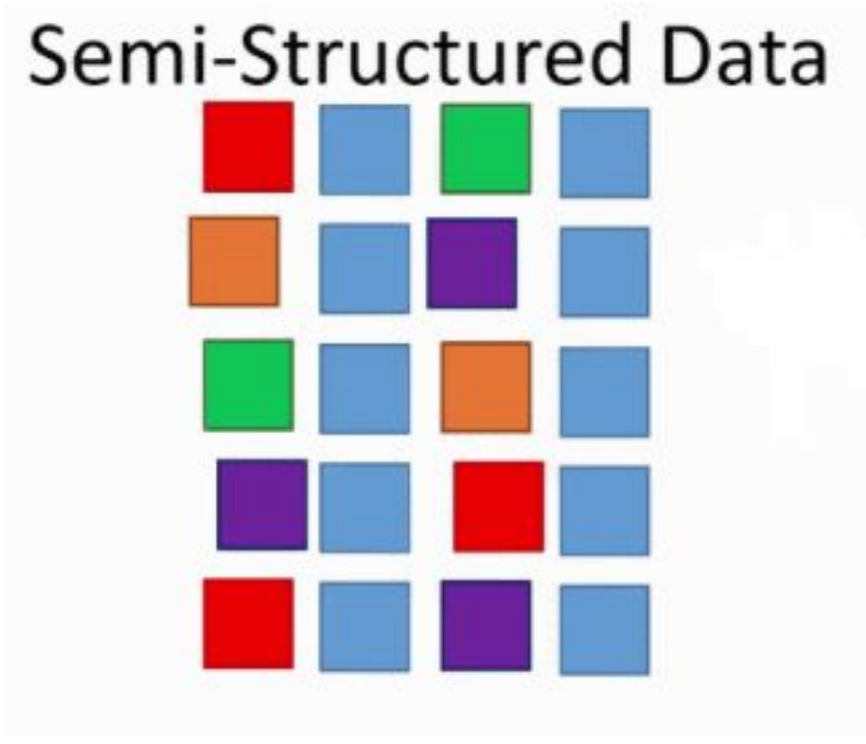
- Chip
- Celo
- Castración
- Esterilización
- Alergias



Muchos valores
VACÍOS

Formatos de datos

Datos semi-estructurados



Datos que sigue una estructura más **dinámica** y **flexible.**

Usualmente se organizan de manera jerárquica.

No necesita modificar la estructura completa al agregar un nuevo atributo.

En lugar de utilizar columnas, se usan **“tags”** para representar atributos

Ejemplos: XML, json, e-mails, algunas web

Datos semi-estructurados

Nombre mascota	Especie	Raza	Edad	Sexo	Peso
Doris	Canino	Schnauzer	3	Hembra	8
Clotilde	Canino	Mestizo	14	Hembra	7

XML

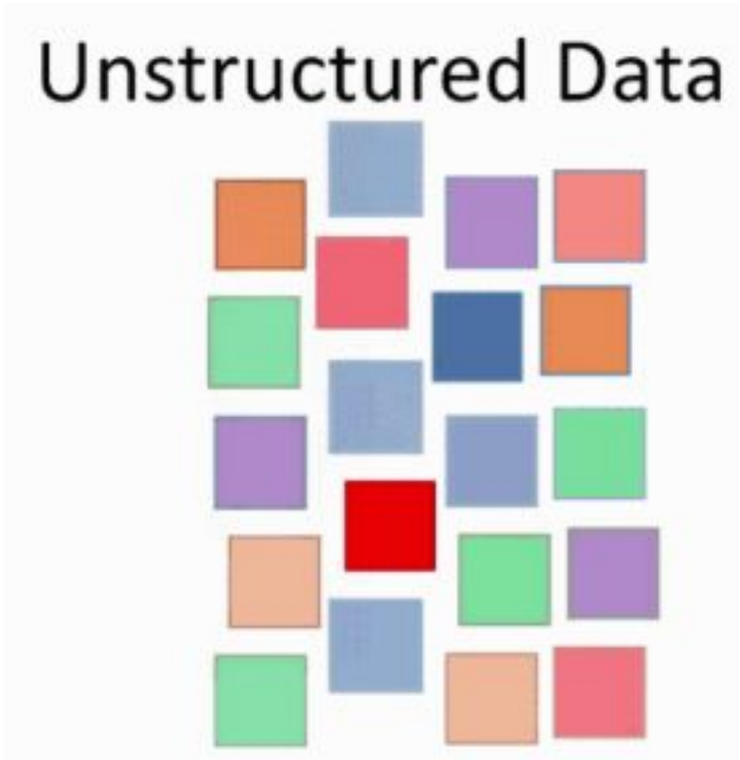
```
<mascota>
  <nombre>Doris</nombre>
  <especie>Canino</especie>
  <raza>Schnauzer</raza>
  <edad>3</edad>
  <color>Sal y pimienta</color>
  <peso>8</peso>
 <celo> SI </celo>
</mascota>
```

json

```
{
  "nombre": "Doris",
  "especie": "Canino",
  "raza": "Schnauzer",
  "edad": "3",
  "color": "Sal y pimienta",
  "peso": "8"
}
```

Formatos de datos

Datos NO estructurados



Datos **no** poseen estructura alguna **previamente** definida.

Es posible encontrar **patrones** de cómo están organizados los datos.

Sin embargo, involucra mucho trabajo de pre-procesamiento de datos.

80% de los datos están en este formato.

Ejemplos: documentos de texto, imágenes, video, audio, etc.

Datos NO estructurados

Nombre mascota	Especie	Raza	Edad	Sexo	Peso
Doris	Canino	Schnauzer	3	Hembra	8
Clotilde	Canino	Mestizo	14	Hembra	7

Se ha ingresado el paciente **Doris**, de especie **canino**. Nació aproximadamente el 2017, por lo que tiene **3 años**. Su manto es color **sal y pimienta** y patas blancas. Este es su primer control, pesando **8kg** aproximadamente.

NO Estructurados **Semi Estructurados** **Estructurados**

(-) formato (+) formato

(+) cantidad (-) cantidad

(+) pre-procesamiento (-) pre-procesamiento

Datos tipo **records**

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

Datos tipo **records**

	season	timeout	lost	win	game	score	ball	player	coach	team
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

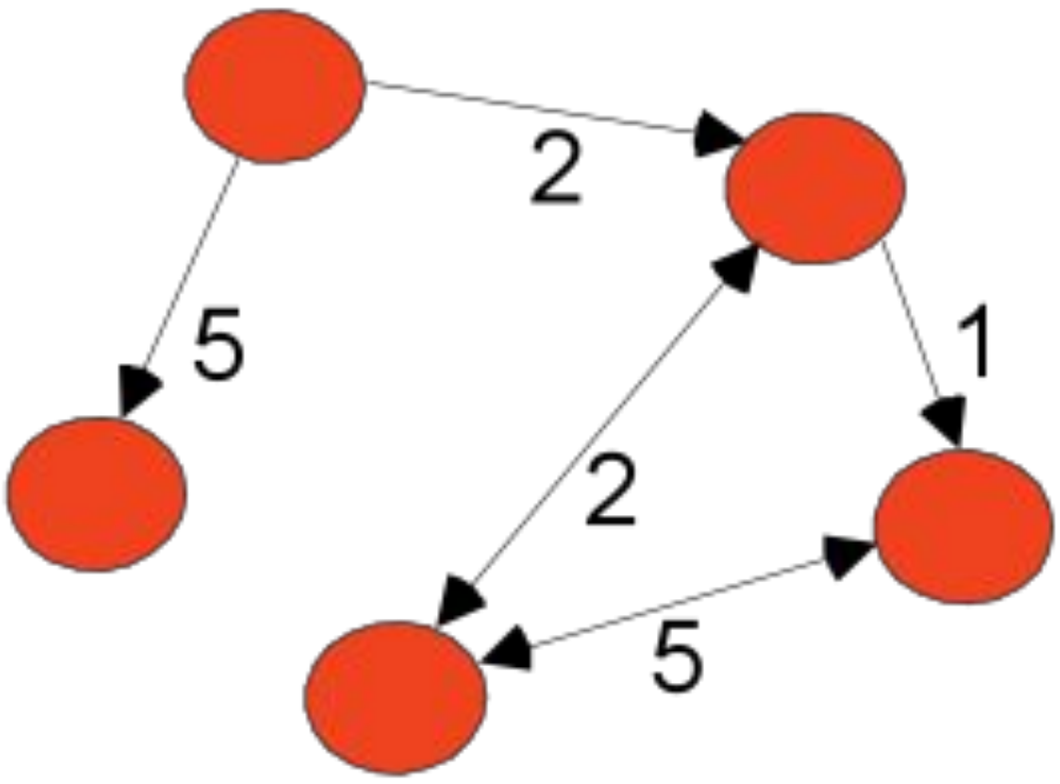
Datos tipo **records**

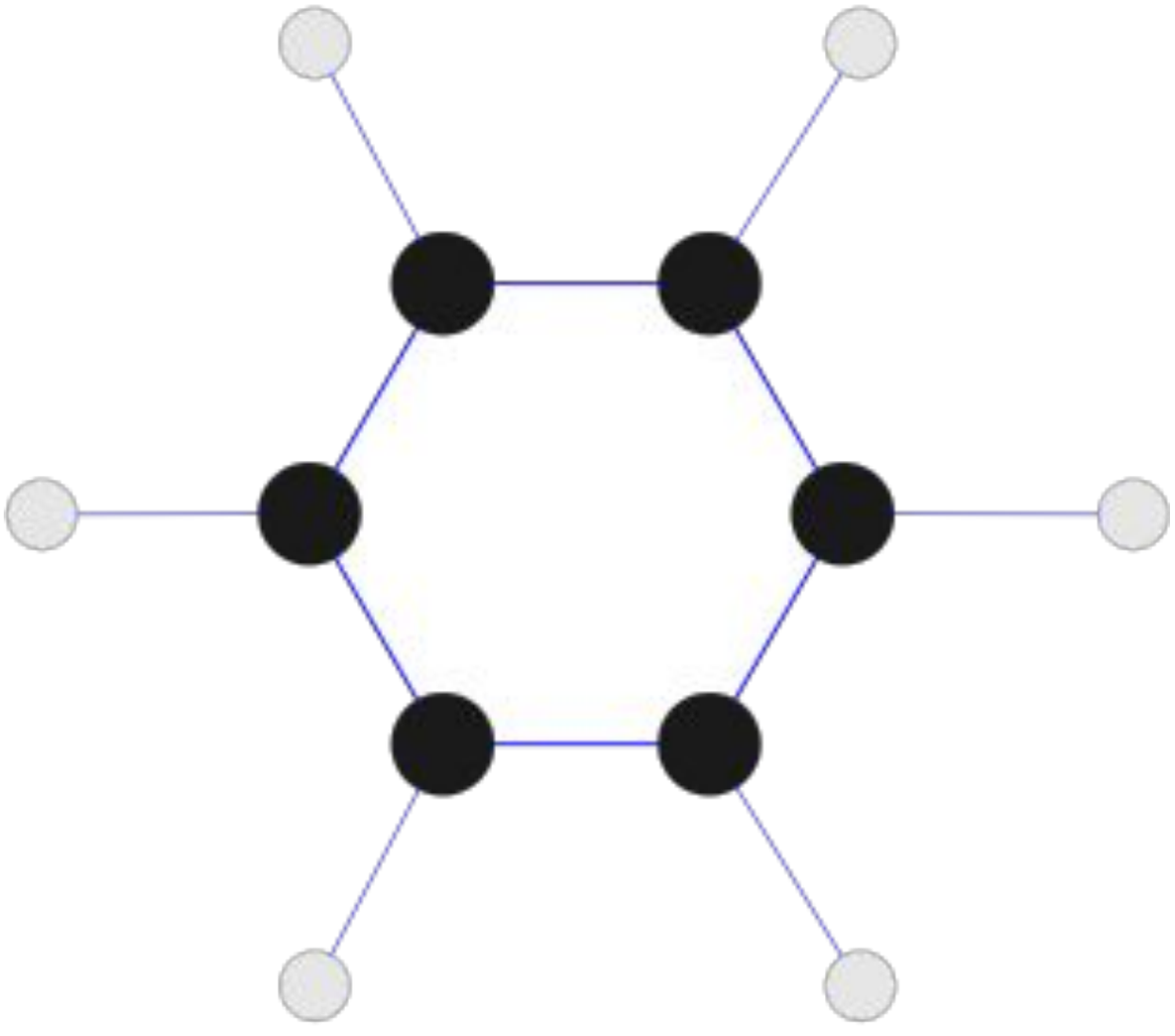
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Datos tipo **records**

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Datos tipo **grafos**





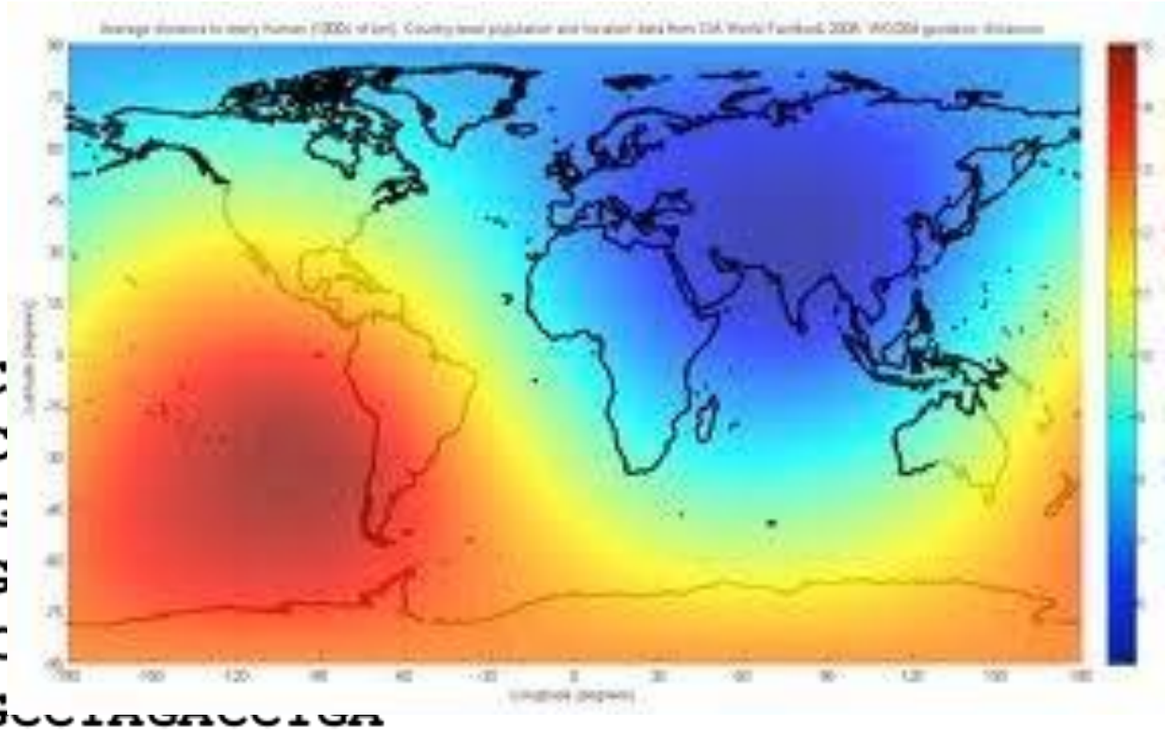
Datos **ordenados**

Items/Events

(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of the sequence

```
GGTTCCGCCTTC
CGCAGGGCCCGC
GAGAAGGGCCCG
GGGGGAGGCGGG
CCAACCGAGTCC
CCCTCTGCTCGG
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```



1) **datos secuenciales** (transacciones con tiempo asociado: libros embarazo -> pañales)
2) **secuencias datos ordenados pero sin tiempo** (ADN, secuencias de genes, etc.)
3) **Datos ordenados en el espacio**
4) **Datos ordenados en el tiempo:** series de tiempo (fluctuaciones de la bolsa)
5) **Autocorrelación temporal:** objetos cercanos en el tiempo son parecidos (mediciones temp en 2 minutos continuos)
6) **Autocorrelación espacial:** obj. cercanos en el espacio son parecidos (i.e., ley del metal)

Características Generales (sets de Datos)

- **Dimensionalidad:** nro. de atributos, maldición de la dimensionalidad (curse of dimensionality) tiene que ver con problemas al trabajar con muchas dimensiones (preprocesamiento: reducción de dimensionalidad)

- **Dispersión:** mayoría de las dimensiones son 0 para los datos, puede tener ventajas, como no necesitar almacenar los valores 0, sólo los 1s. (Ej. Grafo de la Web y sus enlaces).

- **Resolución** (ej. variaciones de presión atmosférica en horas es notoria, pero en meses no se detecta).

Bag of Words Example

Document 1

The quick brown fox jumped over the lazy dog's back.

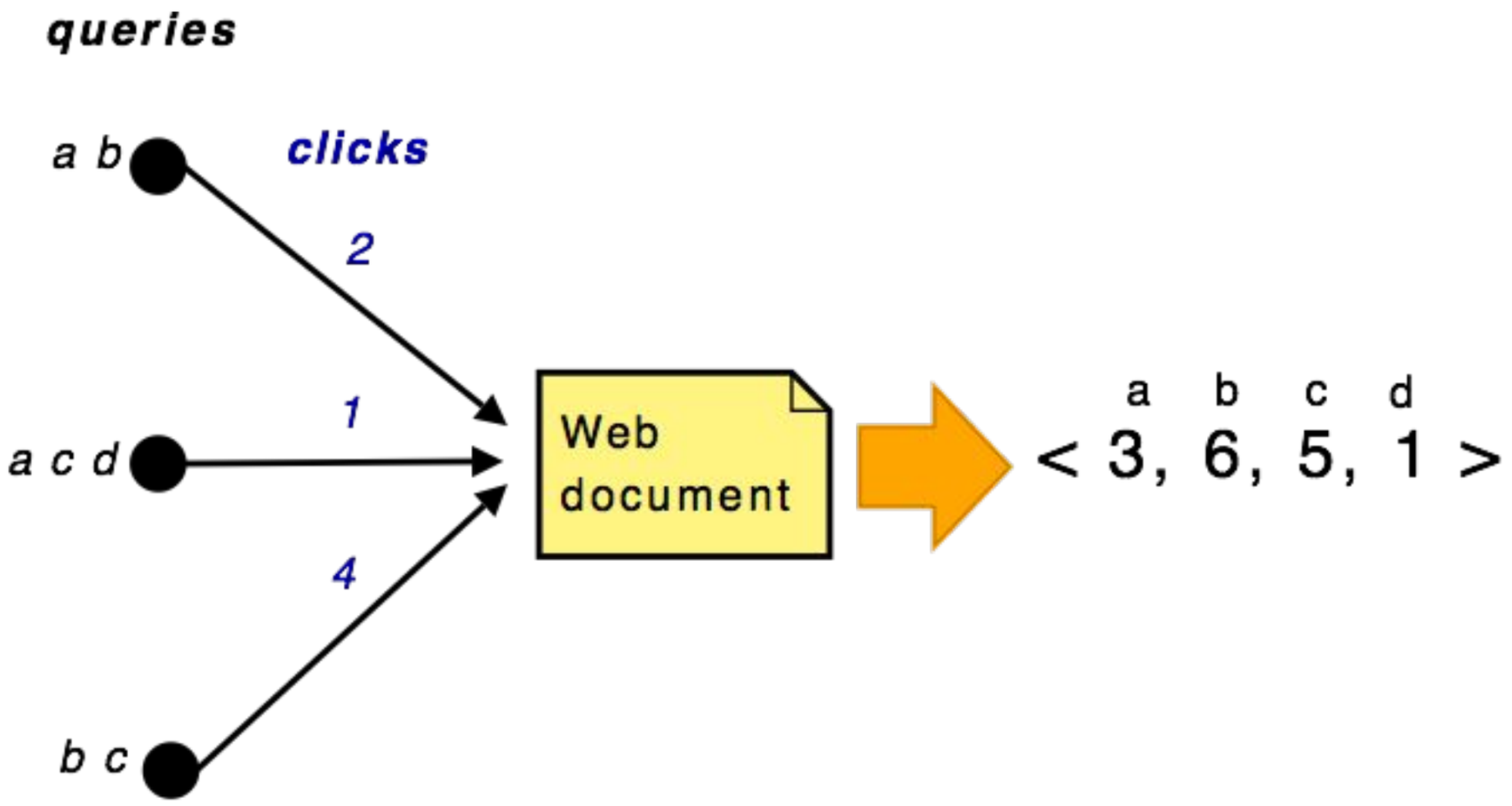
Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

Stopword List

for
is
of
the
to

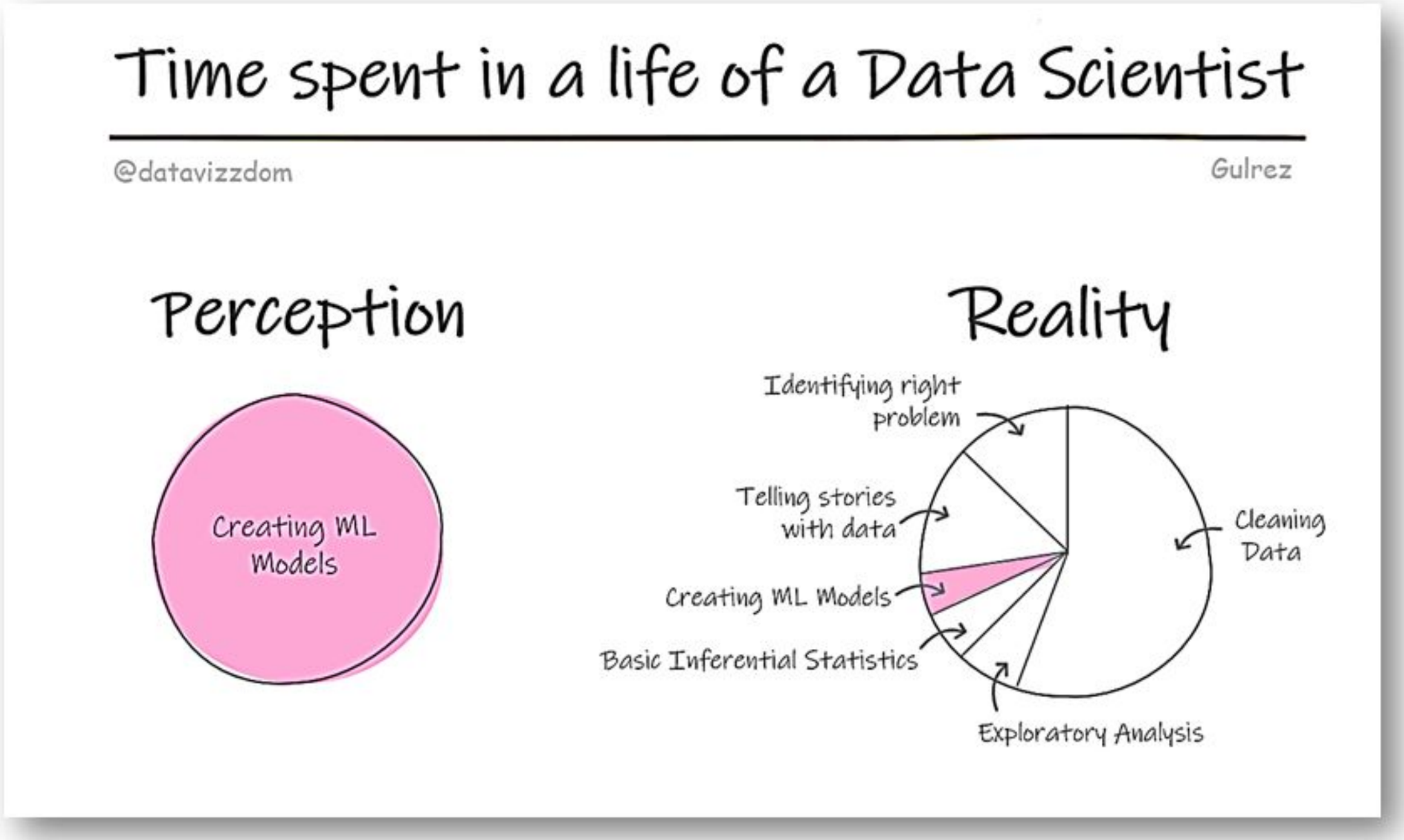


Reducción de dimensionalidad
(DOCUMENTO WEB)

Calidad de los datos

- No poseen la calidad deseada a priori, los algoritmos de DM se enfocan en:
    1. Detección y corrección de problemas de calidad
    2. Usar algoritmos que toleren datos de poca calidad
- i.e., limpieza de datos

Calidad de los datos



¿Por qué se producen errores?

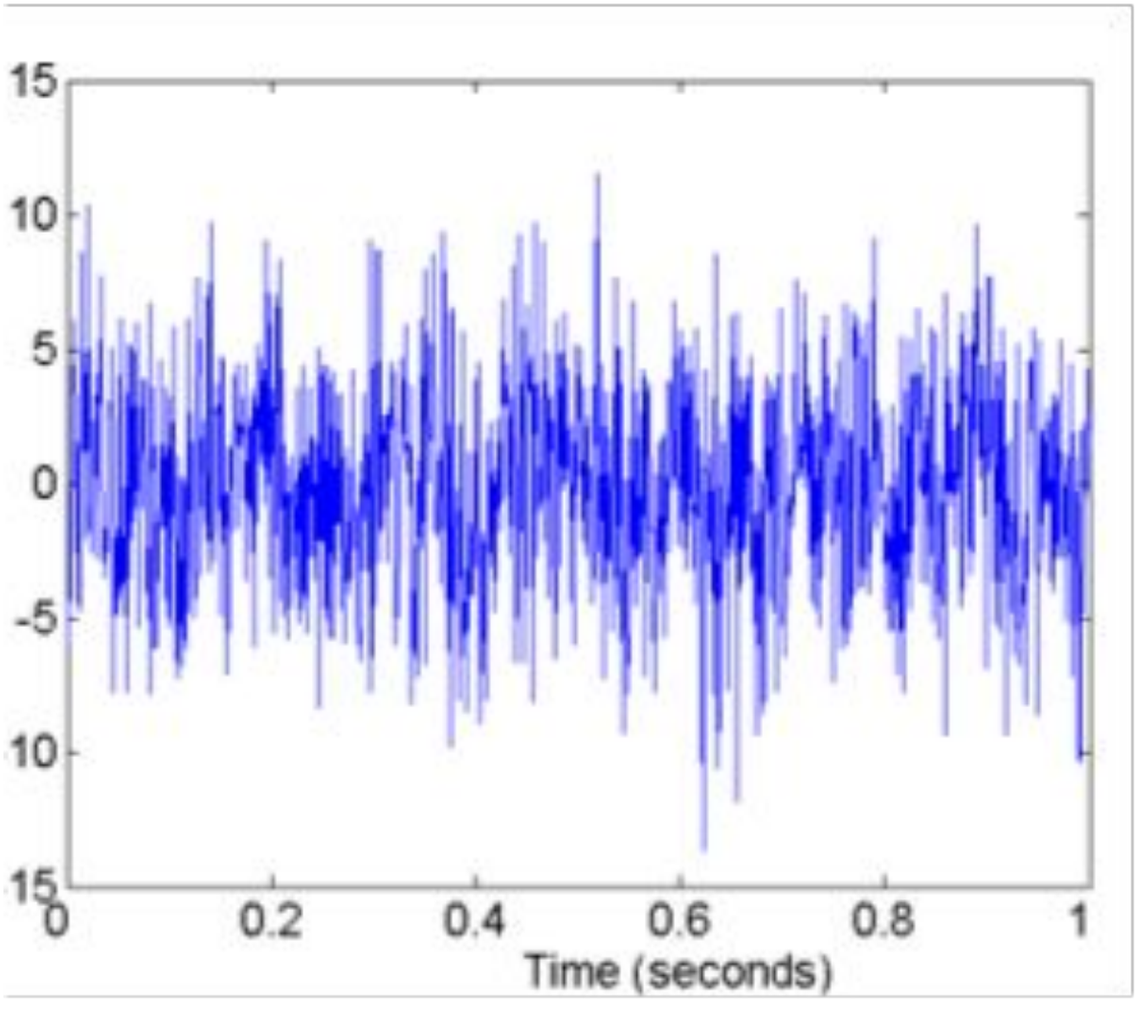
Tipos de errores:

- Ruido y outliers
- Valores faltantes
- Datos duplicados

¿Qué es el ruido?

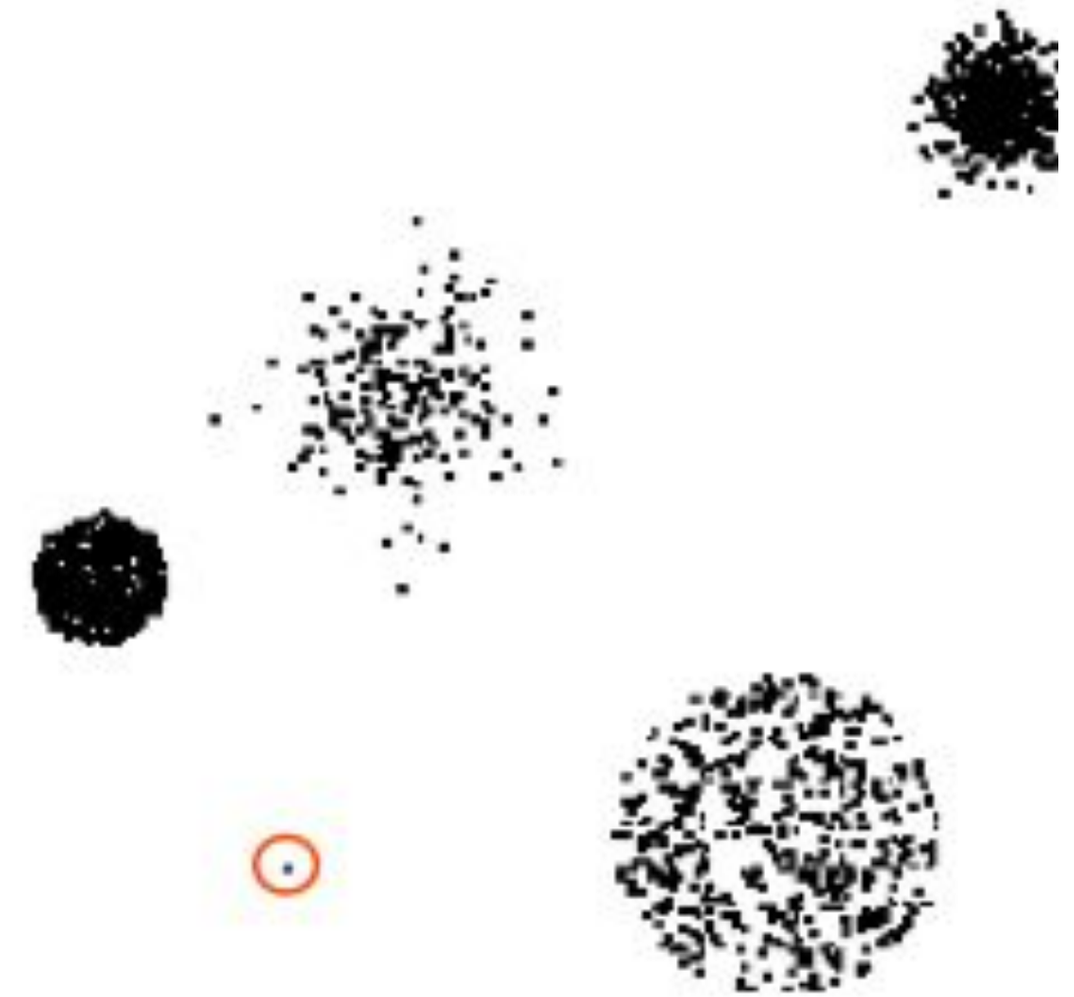
- Componente aleatoria en la medición (distorsión de voz en un teléfono malo)

- Datos espaciales, temporales



Outliers

- Objetos con características considerablemente diferentes a la mayoría



Valores faltantes

- ¿Motivos?
  - Información no recolectada (e.j: no quieren dar edad y/o peso)
  - Atributos no aplicables a todos (e.j: impuesto en niños)

Valores faltantes...

- ¿Cómo los manejo?
- Eliminando el objeto
- Estimando (interpolando) valores
- Ignorar

censo >> 2012

Más moderno, seguro y profesional

15.800.000 = Población efectivamente censada

16.600.000 = Población estimada

incluye

Población efectivamente censada

+

Moradores ausentes

(670.772 Hab.)

censo 2012

Valores inconsistentes

- Datos mal ingresados

Datos duplicados

- El dataset incluye datos duplicados o cuasi-duplicados
- Gran problema al juntar datos de fuentes múltiples
- e.j: RT (casos deseados, no deseados)