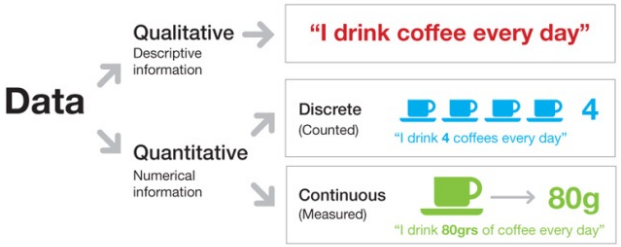
Curso DM
“Datos”

Parte I
Bárbara Poblete

¿qué son los datos?



Attributes

Objects

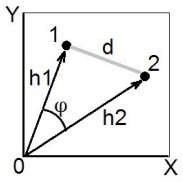
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

¿qué son los datos?

Dataset, records, atributos

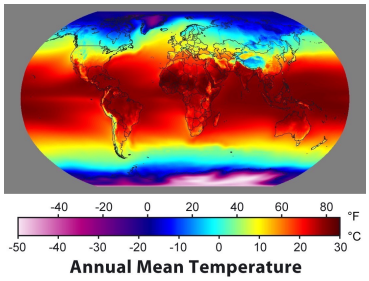
Diferentes enfoques

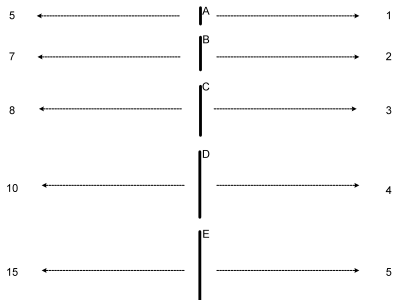
- Analizar en base a las relaciones entre datos
- Extraer relaciones y luego trabajar con estos valores, no con los datos mismos



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.





Midiendo largo

El largo de los segmentos se mide en dos escalas de medición diferentes.
Izquierda: captura orden, Derecha: captura orden y adición

TIPOS DE ATRIBUTOS

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

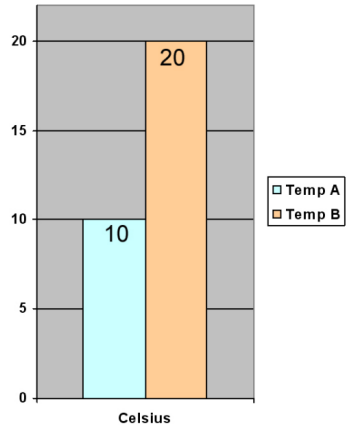
TIPOS DE ATRIBUTOS

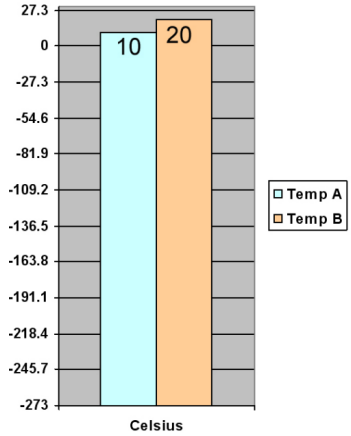
(Operaciones y propiedades)

- **NOMINAL** (IDs, color de ojos, categorías de bacterias)
  (1) **Distinción ==, !=** (solo diferencia de nombre)
- **ORDINAL** (rankings, notas, altura en {alto, mediano, bajo})
  (1) + (2) **Orden <, <=, >, >=** (permite ordenar los datos)
- **INTERVALO** (Fechas, temperaturas °C o °F)
  (1,2)+ (3) **Adición + -**, grado de diferencia entre medidas pero no la razón entre ellos (el cero es arbitrario).
  Ej: No tiene sentido dividir una fecha por otra.
- **RAZÓN** (Temperatura Kelvin, largo, hora)
  (1,2,3)+ (4) **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?





- **Atributos DISCRETOS:** atributos categóricos, binarios (finitos)
- **Atributos CONTINUOS:** temperatura, altura, peso (infinitos)
- Clasificar (discreto, continuo, cualitativo, cuantitativo, nominal, ordinal, intervalo, razón):
  - Rango Militar
  - RUT
  - Distancia al patio

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

	team	coach	play	ball	score	game	win	lost	timeout	season
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

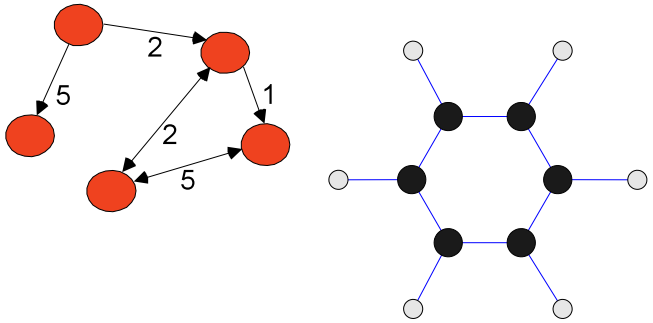
datos tipo records

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

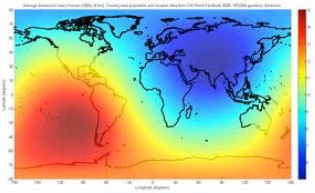


datos ordenados

Items/Events

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

An element of the sequence



GGTTCCGCCTTC
CGCAGGGCCCGC
GAGAAGGGCCCC
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.

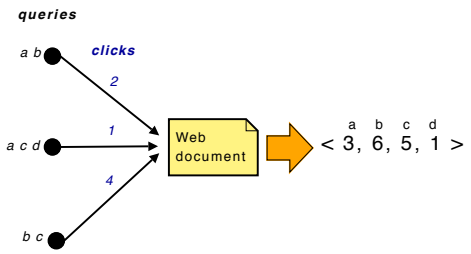
Document 2

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

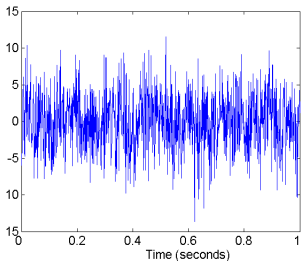
¿Por qué se producen errores?

Tipos de errores:

- Ruido y outliers
- Valores faltantes
- Datos duplicados

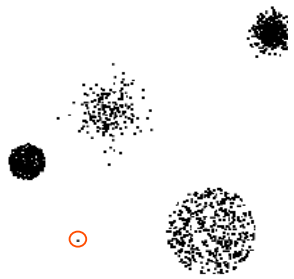
¿Qué es el ruido?

- Componente aleatoria en la medición (distorsión de voz en un teléfono malo)
- Datos espaciales, temporales



Outliers

- Objetos con características considerablemente diferentes a la mayoría

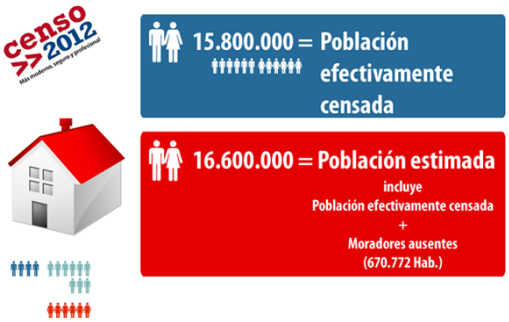


Valores faltantes

- ¿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

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)