

## KNOWLEDGE DISCOVERY IN DATABASES

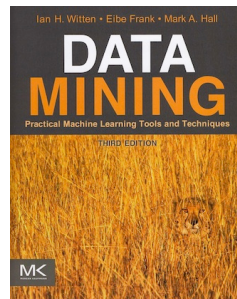
## EVALUACIONES

- Nota por Proyecto (2 Presentaciones + Reporte)
- Nota Trabajo Personal
  - 2-3 Mini-presentaciones
  - Contribución de material (bloggeo)
  - Presentación Herramienta DM
- 2 Controles

## VISITA ILUSTRE

- **Ian Witten** (WEKA, Libro de DM)

- Ira o 2da semana Dic.



## CURSO DM “DATOS”

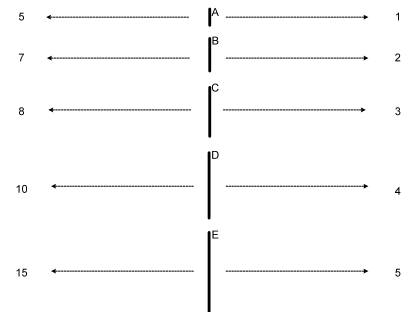
Clase 3  
Bárbara Poblete

**Attributes**

	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

**Objects**

¿QUÉ SON LOS DATOS?



## MIDIENDO LARGO

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

## TIPOS DE ATRIBUTOS

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

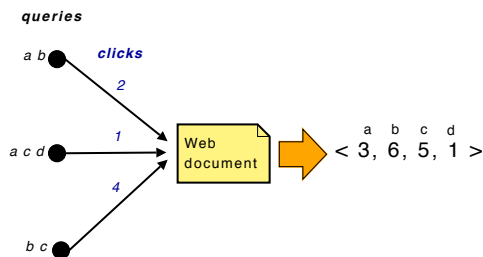
## OPERACIONES O PROPIEDADES

- |                              |                     |
|------------------------------|---------------------|
| 1. Distinción = y $\neq$     | • NOMINAL (1)       |
| 2. Orden $<, \leq, > y \geq$ | • ORDINAL (1,2)     |
| 3. Adición + y -             | • INTERVALO (1,2,3) |
| 4. Multiplicación * y /      | • RAZÓN (1,2,3,4)   |

- Categóricos (Cualitativos), i.e: Nominal, Ordinal
- Numéricos (Cuantitativos), i.e: Intervalo, Razón
- Atributos DISCRETOS
- Atributos CONTINUOS

## CARACTERÍSTICAS GENERALES (SETS DATOS)

- DIMENSIONALIDAD
- DISPERSIÓN (SPARSITY)
- RESOLUCIÓN



## REDUCCIÓN DE DIMENSIONALIDAD (DOCUMENTO WEB)

## DATOS TIPO RECORDS

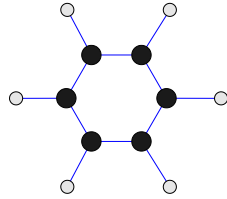
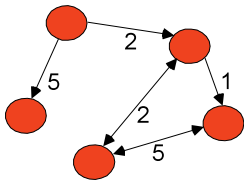
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

	team	coach	ply	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

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

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 GRAFOS



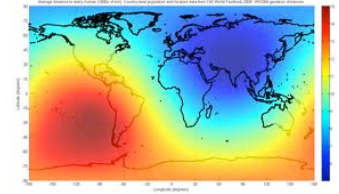
## DATOS ORDENADOS

Items/Events



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

An element of  
the sequence



```

GGTTCGCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCGCGCGTC
GAGAAGGGCCCGCCTGGCGGCG
GGGGGAGGGGGCGCCCGAGC
CCAAACCGAGTCCGACCGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACAGGG
  
```