

Los Datos

- Definición: ¿ Qué son los datos ?
- Fuentes de Datos
- Datos Estructurados versus No Estructurados
- Ejemplos Datos Estructurados y No Estructurados
- Ejemplo de Uso de los Datos

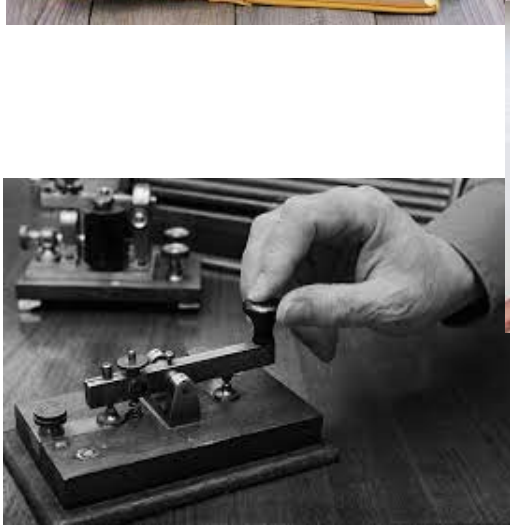
¿ Qué son los Datos ?

Es **Información** que ha sido transformada de forma tal que sea eficiente su procesamiento.

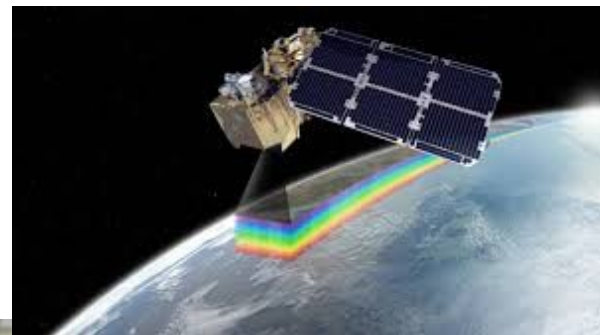
Es Información susceptible de ser almacenada y procesada por una **computadora**.

Es una representación de hechos o ideas capaz de ser comunicada y **manipulada** mediante algún proceso.

Estos **No** son Datos



Estos **Sí** son Datos



¿ Cuando se Produjo el Salto ?

Terreno Analógico



Terreno Digital



1948

Tiempo

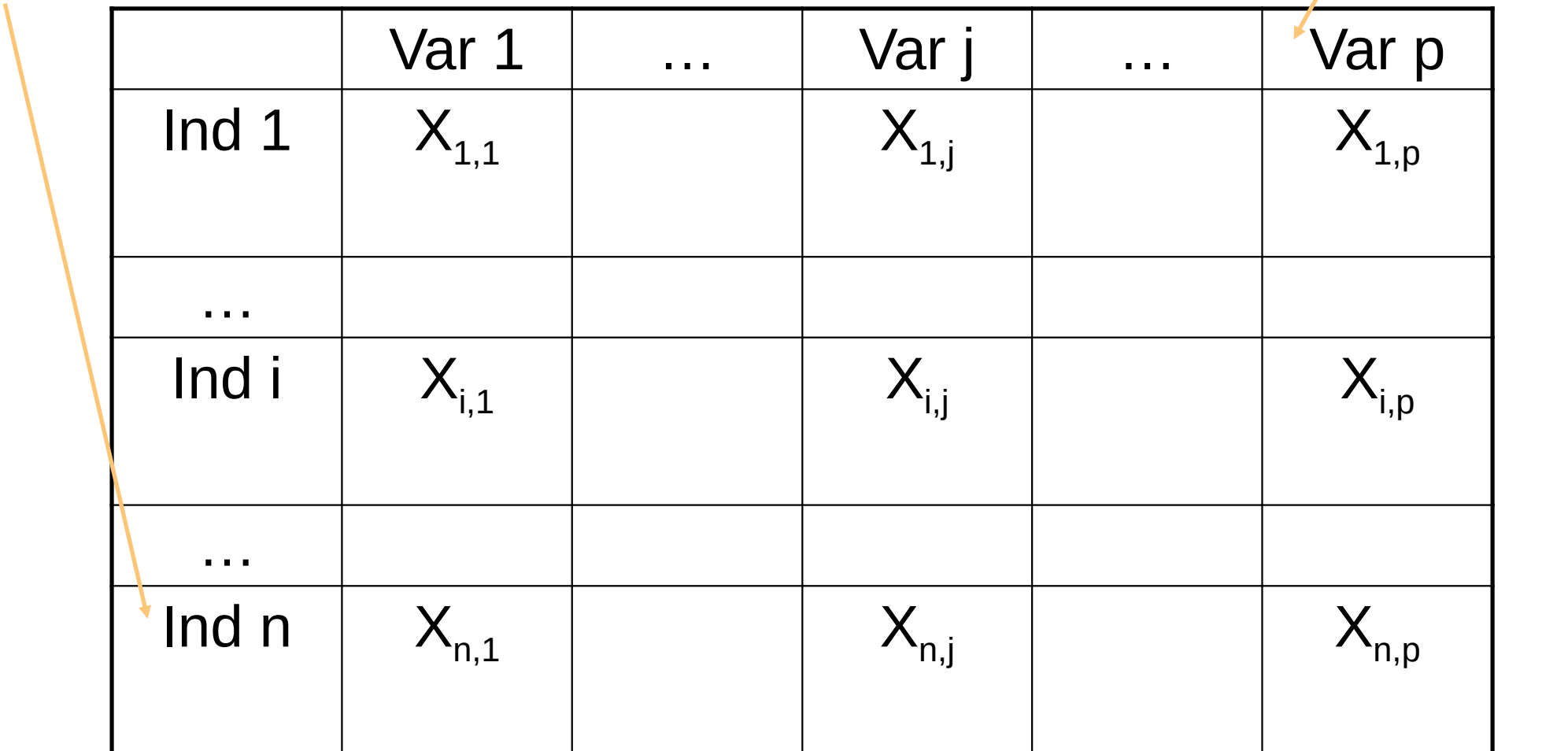
Fuentes de Datos

- API
- Web Scrapping
- Archivos Planos (i.e. csv)
- Bases de Datos Relacionales
- FTP
- Páginas Web
- Sensores
- Teléfono Movil
- Juegos en Red
- Experimentos
- Streaming

Datos Estructurados

n Observaciones

p Atributos



	Var 1	...	Var j	...	Var p
Ind 1	$X_{1,1}$		$X_{1,j}$		$X_{1,p}$
...					
Ind i	$X_{i,1}$		$X_{i,j}$		$X_{i,p}$
...					
Ind n	$X_{n,1}$		$X_{n,j}$		$X_{n,p}$

Formato Ancho versus Largo

wide

id	x	y	z
1	a	c	e
2	b	d	f

long

id	key	val
1	x	a
2	x	b
1	y	c
2	y	d
1	z	e
2	z	f

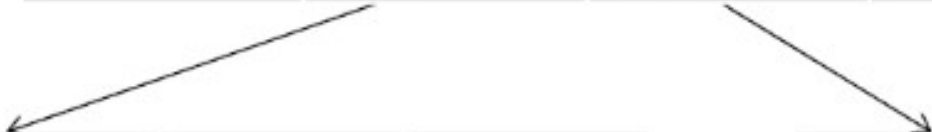
Fuente: **Statistical Modelling**

Léo Belzile (HEC Montréal)

Ejemplos de Datos Estructurados

Purchase table

Transaction ID	Customer ID	Product ID	Purchase date
1112	24221	8977	03-22-2010
1113	24222	8978	03-22-2010
1114	24223	8979	03-22-2010



Customer ID	Customer	Address
24221	Bob	123 East street
24222	Alice	223 Main street
24223	Martha	465 North street

Customer table

Product ID	Name	Price
8977	Banana	.79
8978	TV	400
8979	Watch	50

Product table

Ejemplos de Datos Estructurados

id	ad_type	start_date	end_date	created_on	lat	lon	l1
<chr>	<chr>	<date>	<date>	<date>	<dbl>	<dbl>	<chr>
oyj+f764ALCYodIqBvWAww==	Propiedad	2019-04-14	2019-07-10	2019-04-14	-34.65225	-58.38556	Argentina
HdjpKrQdwYfH9YU1DKjltg==	Propiedad	2019-04-14	2019-04-15	2019-04-14	-34.62825	-58.40652	Argentina
YwWE3rTb2+gmsBwjUHmAPQ==	Propiedad	2019-04-14	2019-06-30	2019-04-14	-34.59280	-58.42093	Argentina
6AxnSWOhblU8TUCqb+paBg==	Propiedad	2019-04-14	9999-12-31	2019-04-14	-34.56563	-58.46513	Argentina
U4fk+co3Rd8JDMot0pQl6Q==	Propiedad	2019-04-14	2019-05-21	2019-04-14	-34.62218	-58.52272	Argentina
AfdcsqUSelai1ofCAq2B0Q==	Propiedad	2019-04-14	9999-12-31	2019-04-14	-34.6321	-58.48888	Argentina

6 rows | 1-8 of 24 columns

surface_covered	price	currency	price_period
<dbl>	<dbl>	<chr>	<chr>
180	320000	USD	Mensual
240	500000	USD	Mensual
157	350000	USD	Mensual
NA	470000	USD	NA
110	155000	USD	NA
69	199900	USD	NA

l2	l3	l4	l5	l6	rooms	bedrooms	bathrooms	surface_total
<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Capital Federal	Barracas	NA	NA	NA	NA	NA	NA	300
Capital Federal	Boedo	NA	NA	NA	6	NA	2	178
Capital Federal	Palermo	NA	NA	NA	NA	NA	2	240
Capital Federal	Belgrano	NA	NA	NA	3	NA	4	157
Capital Federal	Versalles	NA	NA	NA	NA	NA	1	140
Capital Federal	Velez Sarsfield	NA	NA	NA	3	NA	2	95

7 rows | 9-17 of 24 columns

Fuentes de Datos No Estructurados

- Páginas Web
- Redes Sociales
- Historias Clínicas
- Encuestas Complejas
- Documentos de Texto
- Material Multimedia

Ejemplos de Formatos de Datos No Estructurados: Listas

```
> x
$a
[1] 2.5
$b
[1] TRUE
$c
[1] 1 2 3
> typeof(x)
[1] "list"
> length(x)
[1] 3
```

Un número

Elemento lógico

Un Vector

La Clase

Cantidad de Elementos

Ejemplos de Formatos de Datos No Estructurados:

XML

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<breakfast_menu>
  <food>
    <name>Belgian Waffles</name>
    <price>$5.95</price>
    <description>
      two of our famous Belgian Waffles with
      plenty of real maple syrup
    </description>
    <calories>650</calories>
  </food>
```

```
<food>
  <name>Strawberry Belgian Waffles</name>
  <price>$7.95</price>
  <description>
    light Belgian waffles covered
    with strawberries and whipped cream
  </description>
  <calories>900</calories>
  <sides>
    <side>
      <name>Sausage</name>
      <price>$2.00</price>
    </side>
    <side>
      <name>Bacon</name>
      <price>$2.50</price>
    </side>
  </sides>
</food>
```

Ejemplo de Uso de los Datos

¿ Qué barrio es más barato, Belgrano o Caballito ?

I3	surface_covered	price
<chr>	<dbl>	<dbl>
Belgrano	68	255000
Belgrano	61	175000
Caballito	62	125000
Caballito	69	289000
Belgrano	63	314200
Belgrano	62	177000
Belgrano	65	265000
Belgrano	68	320000
Belgrano	63	270000
Belgrano	66	295000

Proceso de
manipulación

Datos

Resultado

```
datos1c %>% filter(l3=="Belgrano"|l3=="Caballito",property_type=="Departamento",surface_covered>60,  
surface_covered<70) %>% group_by(l3) %>% summarise(Promedio=mean(price),Cantidad=length(l3))
```

I3	Promedio	Cantidad
<chr>	<dbl>	<int>
Belgrano	250460.0	386
Caballito	197130.2	287