



FUNDAMENTOS DEL BIG DATA

Tipología, captura y preparación de los datos





Big Data

Ciencia de Datos

Machine Learning

Deep Learning

Inteligencia Artificial

Data mining



BIG DATA

- Técnicas para capturar, almacenar, homogeneizar, transferir, consultar, visualizar y analizar datos a gran escala y de manera sistemática
- Aumento exponencial de los datos y de las fuentes que los generan

Volumen

- Muchos tipos de datos para administrar y proteger

Variedad

THE INTERNET IN 2023 EVERY MINUTE





BIG DATA

Retos con información masiva:

- **Calidad de los datos**
 - Privacidad y seguridad de la información
- 

¿QUÉ ES LA CIENCIA DE DATOS?

- Extracción de información relevante de conjuntos de datos: KDD (Knowledge Discovery in Databases)
- Métodos, procesos y sistemas que involucran tratamiento de datos para extraer conocimiento

¿QUÉ ES LA CIENCIA DE DATOS?

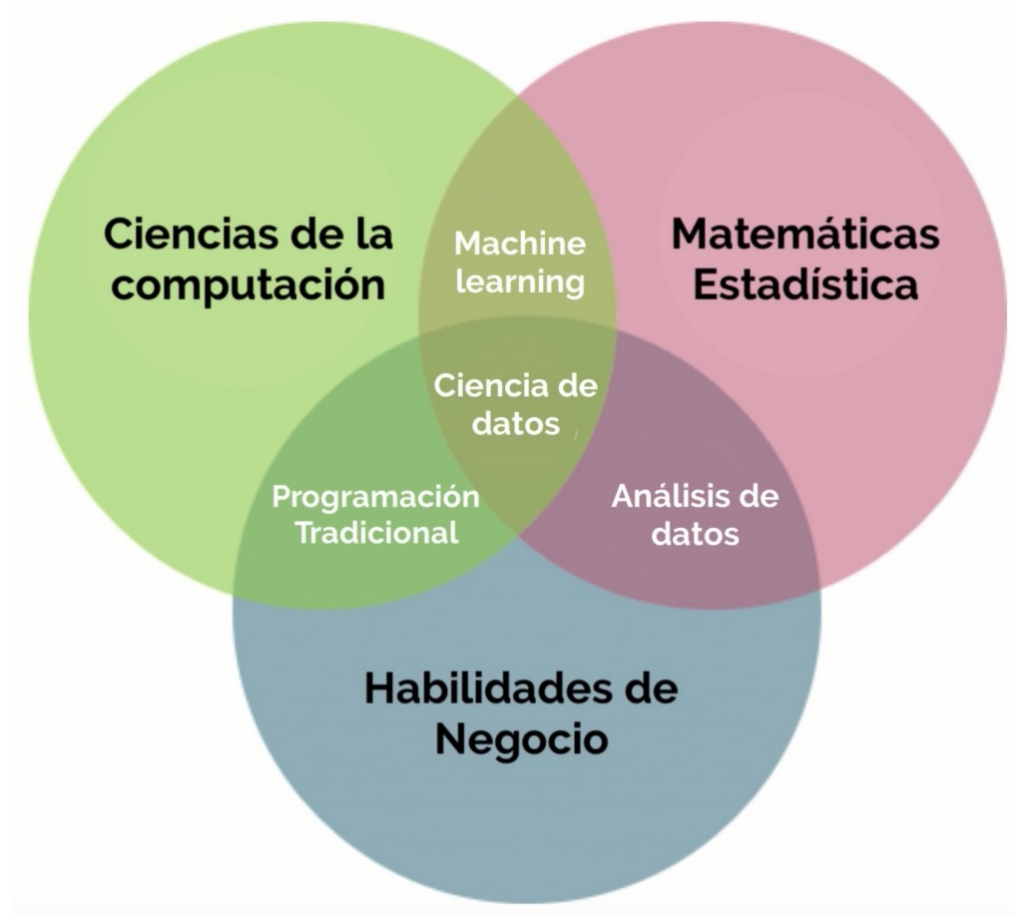
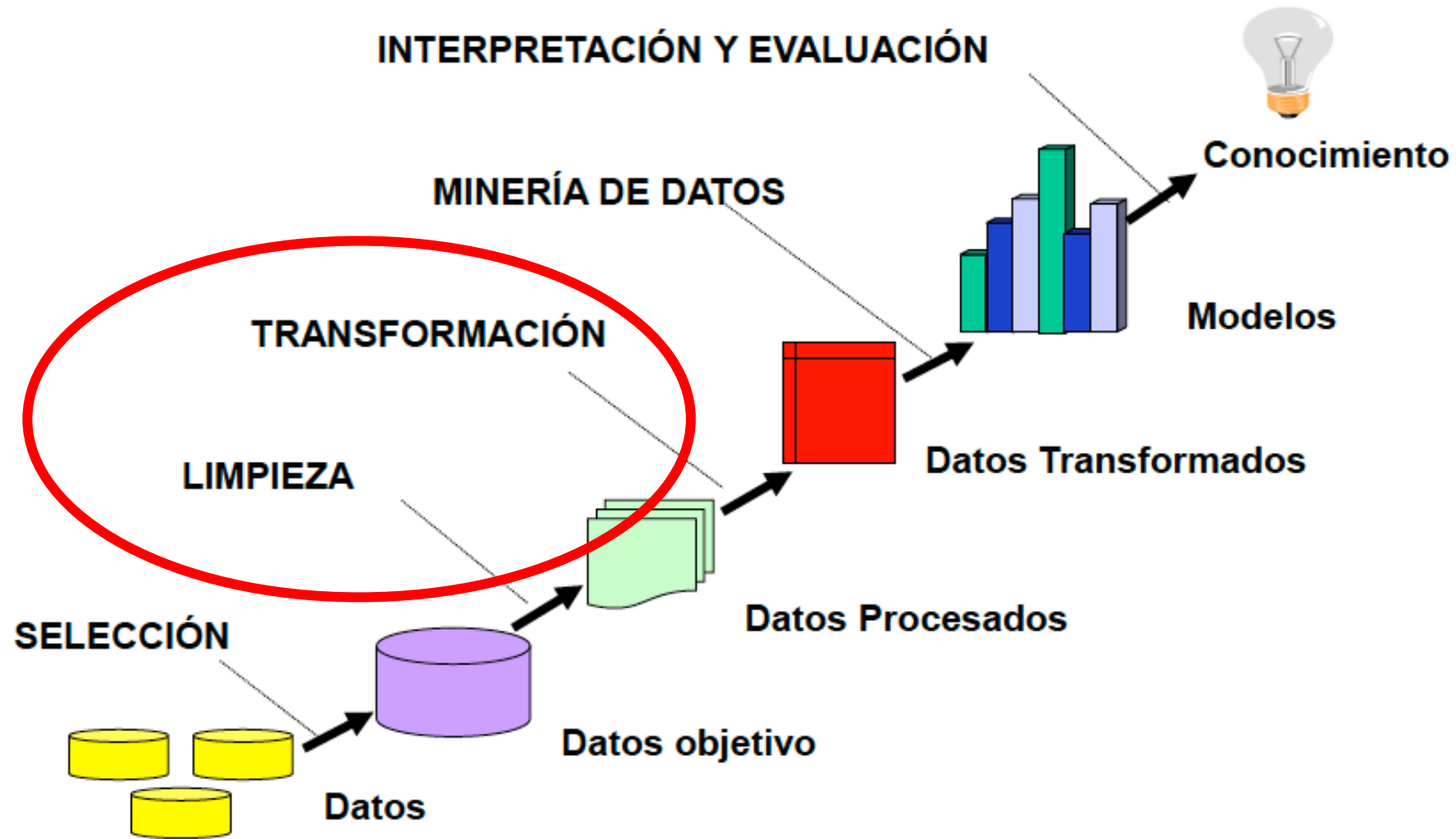


Diagrama de Venn de la Ciencia de datos de Drew Conway

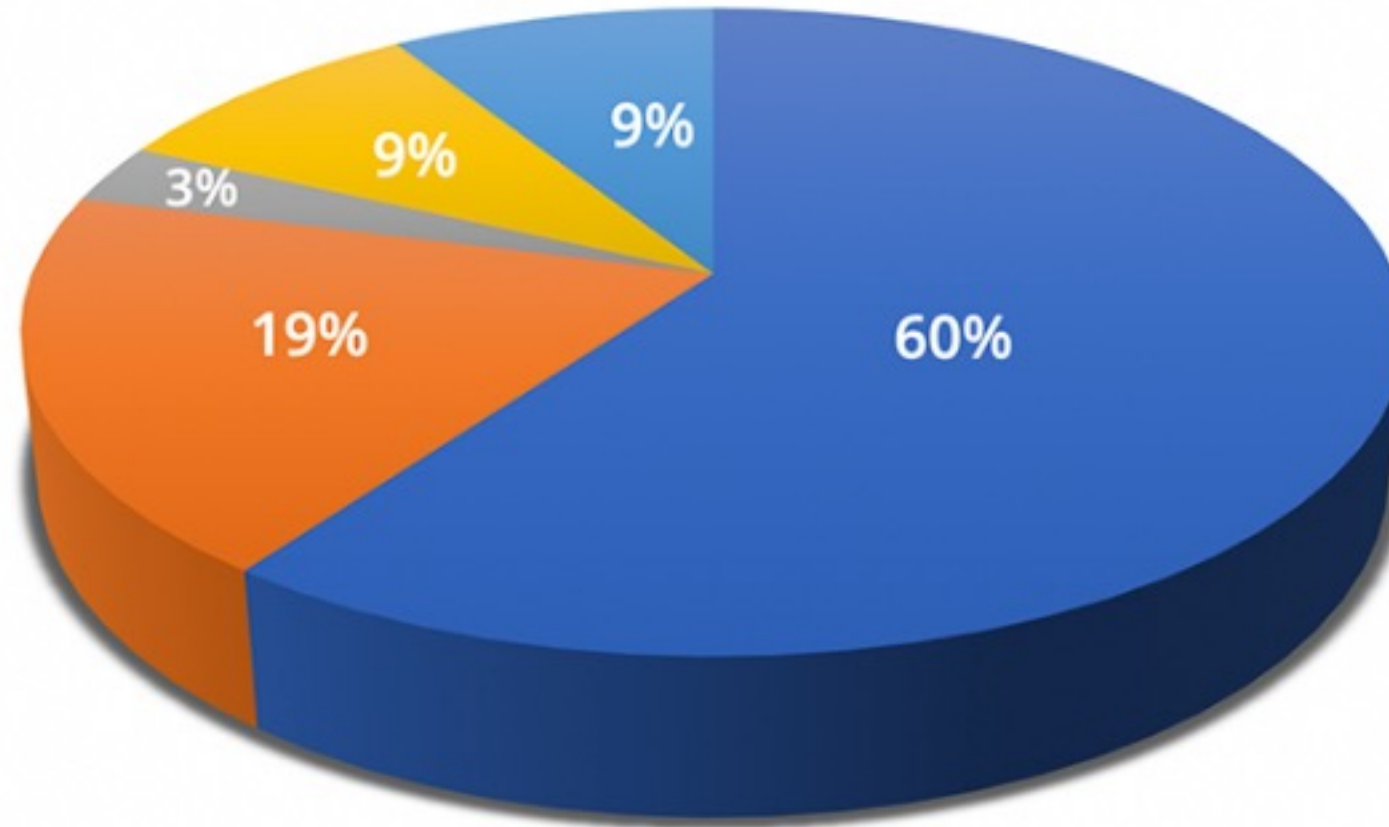
¿QUÉ ES LA CIENCIA DE DATOS?



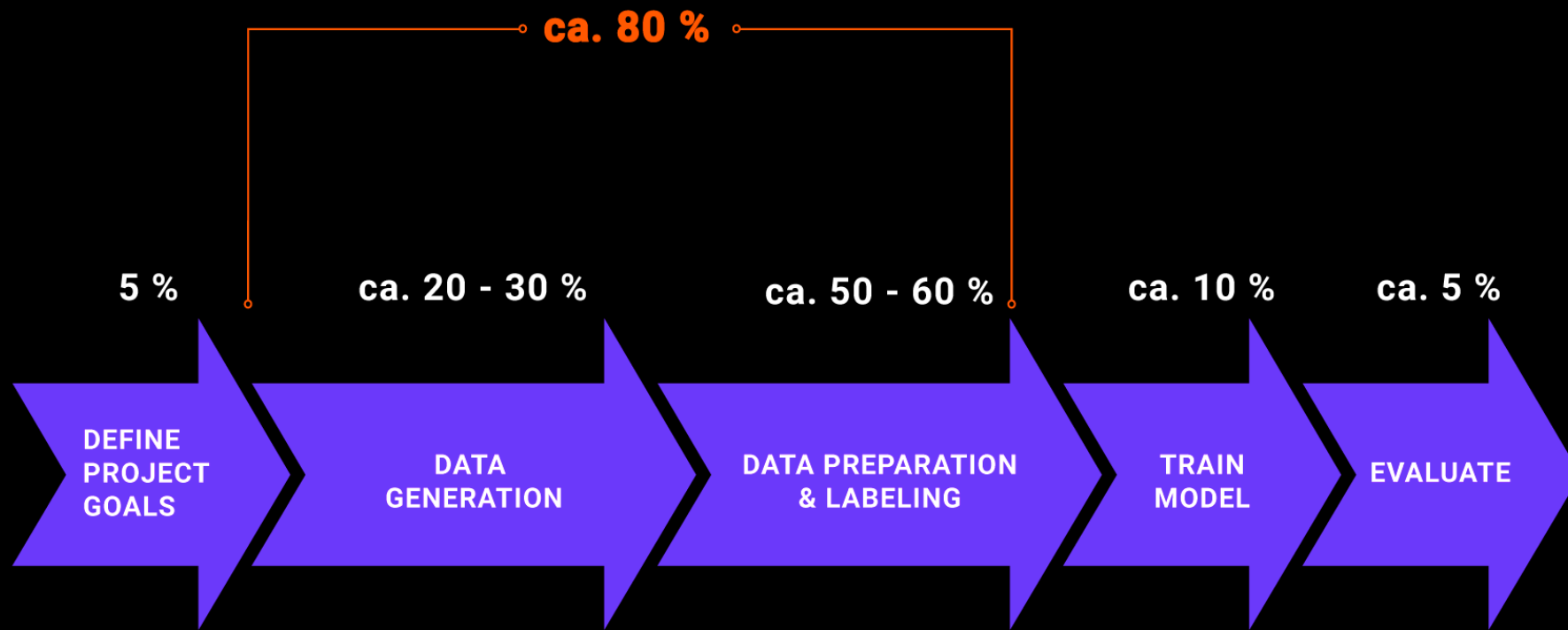
¿QUÉ ES LA CIENCIA DE DATOS?

- Limpiar y optimizar los datos es uno de los mayores desafíos de la ciencia de datos: **Dilema 80/20 de la ciencia de datos**
- Se dedica aproximadamente el 80% del tiempo a generar, preparar y etiquetar datos y solo el 20% a construir y entrenar modelos

¿A qué dedica el tiempo un científico de datos?



■ Preprocesamiento de datos ■ Obtener datos ■ Construir modelos ■ Explorar datos ■ Otros



DATA GENERATION

- Acquire data (search, make or buy)
- Data generation
- Data Augmentation

DATA PREPARATION

- Store / load data
- Organize data
- Correct, normalize ..
- Label & annotate data


TRAIN & EVALUATE MODEL

- Choose model
- Train model
- Evaluate model
- Deploy model



¿QUÉ ES LA CIENCIA DE DATOS?


Para sacar beneficio a la cantidad de información disponible es necesario comprender cuáles son las **categorías de datos** y las **fuentes de origen** de los mismos





TIPOS DE DATOS

En base a su estructura:

- Estructurados
 - Semiestructurados
 - No estructurados
- 

DATOS ESTRUCTURADOS

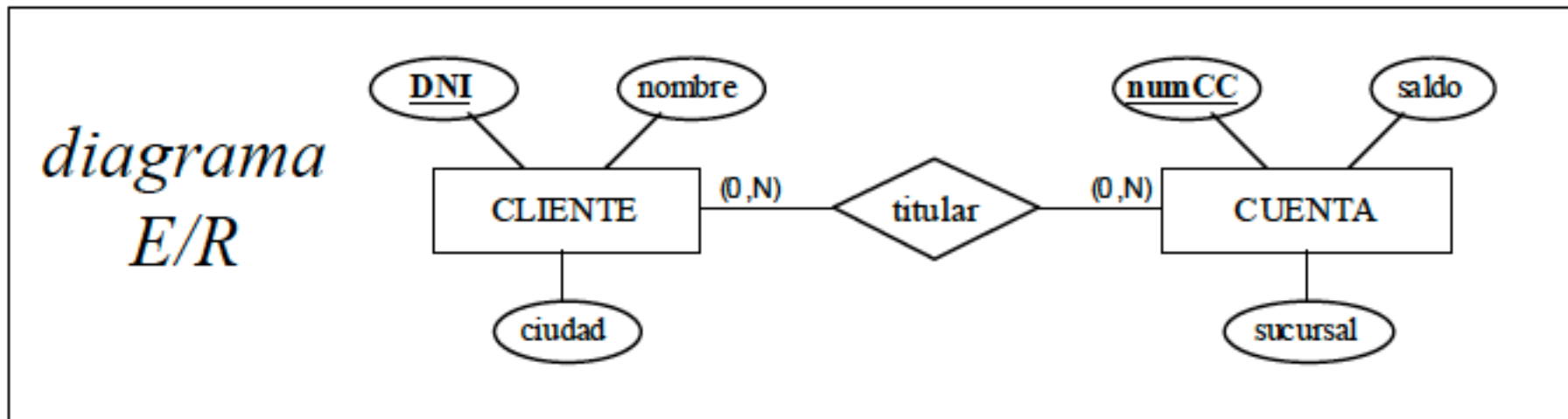
- Información que ha sido formateada y transformada en un modelo de datos bien definido
- Están altamente organizados de tal manera que se pueden buscar fácilmente
- Proviene de sistemas transaccionales, bases de datos relacionales y aplicaciones administrativas (por ejemplo, sistemas de ERP)

DATOS ESTRUCTURADOS

- Son los datos típicos de la mayoría de bases de datos relacionales
- Tienen un esquema determinado que define las tablas en las que se almacenan, qué tipo de campos tienen y cómo se relacionan entre ellas
- Se gestionan mediante un tipo de lenguaje de programación estructurado conocido como SQL (Structured Query Language)


DATOS ESTRUCTURADOS

- Modelo Entidad-Relación, BBDD Relacional



DATOS ESTRUCTURADOS

```
create table libros(  
    titulo varchar(20),  
    autor varchar(30),  
    editorial varchar(15),  
    precio float,  
    cantidad integer  
);  
  
insert into libros (titulo,autor,editorial,precio,cantidad)  
values ('El aleph','Borges','Emece',45.50,100);  
insert into libros (titulo,autor,editorial,precio,cantidad)  
values ('Alicia en el pais de las maravillas','Lewis Carroll','Planeta',25,200);  
insert into libros (titulo,autor,editorial,precio,cantidad)  
values ('Matematica estas ahi','Paenza','Planeta',15.8,200);  
  
select * from libros;  
  
select titulo,autor,editorial from libros;  
  
select titulo,precio from libros;  
  
select editorial,cantidad from libros;
```



```
Query 1 x  
1 drop table if exists libros;  
2  
3 create table libros(  
4     codigo integer unsigned auto_increment,  
5     titulo varchar(20) not null,  
6     autor varchar(30),  
7     editorial varchar(15),  
8     precio float unsigned,  
9     cantidad integer unsigned,  
10    primary key (codigo)  
11 );  
12
```

DATOS NO ESTRUCTURADOS

- Datos en bruto y no organizados
- Sin estructura interna identificable
- Se presentan en muchos formatos con diversos grados de complejidad
- Fuentes heterogéneas y generados por humanos o máquinas



DATOS NO ESTRUCTURADOS

- ¿Qué relevancia tiene esta variedad de datos? ¿Qué relación guarda con el Big Data?
- Alrededor del 80 % de la información relevante para un negocio se origina en forma no estructurada
- El desafío para las organizaciones radica en comprender y extraer valor de los datos no estructurados
- Ventaja competitiva

DATOS SEMIESTRUCTURADOS

- Se encuentran a medio camino entre los estructurados y los no estructurados
- Tienen un cierto nivel de estructura, jerarquía y organización, aunque carecen de un esquema fijo
- Se organizan mediante etiquetas semánticas que permiten agruparlos y crear jerarquías: **Metadatos**
- Se refieren a cualquier información que utilice un esquema de **autodescripción**

DATOS SEMIESTRUCTURADOS

- Dos de los formatos más comunes de datos semiestructurados son:
 - XML (eXtensible Markup Language)
 - JSON (JavaScript Object Notation)

DATOS SEMIESTRUCTURADOS

XML (EXTENSIBLE MARKUP LANGUAGE)

- Desarrollado por W3C (World Wide Web Consortium)
- Basado en SGML (Standard Generalized Markup Language)
- Utilizado para el almacenamiento e intercambio de datos entre distintas plataformas
- Define etiquetas personalizadas para la descripción y organización de datos

DATOS SEMIESTRUCTURADOS

XML (EXTENSIBLE MARKUP LANGUAGE)

- Los documentos XML están formados por texto plano (sin formato) y contienen marcas o etiquetas
- Sintaxis:

`<etiqueta>valor</etiqueta>`

- Declaración XML

`<?xml version="1.0" encoding="UTF-8"?>`

DATOS SEMIESTRUCTURADOS

XML (EXTENSIBLE MARKUP LANGUAGE)

<persona>

<nombre>Elsa</nombre>

<mujer/>

<fecha-de-nacimiento>

<día>18</día>

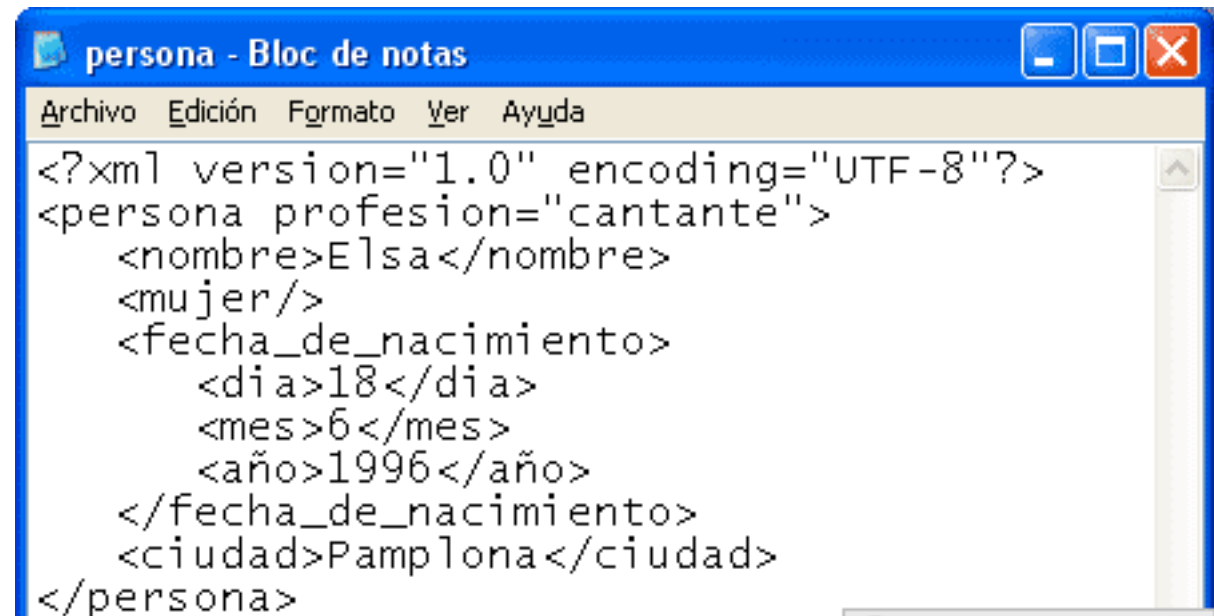
<mes>6</mes>

<año>1996</año>

</fecha-de-nacimiento>

<ciudad>Pamplona</ciudad>

</persona>



```
persona - Bloc de notas
Archivo  Edición  Formato  Ver  Ayuda
<?xml version="1.0" encoding="UTF-8"?>
<persona profesion="cantante">
  <nombre>Elsa</nombre>
  <mujer/>
  <fecha_de_nacimiento>
    <día>18</día>
    <mes>6</mes>
    <año>1996</año>
  </fecha_de_nacimiento>
  <ciudad>Pamplona</ciudad>
</persona>
```


DATOS SEMIESTRUCTURADOS

XML (EXTENSIBLE MARKUP LANGUAGE)

- Un atributo proporciona información extra del elemento que lo contiene

```
<producto codigo="G45">
```

```
  <nombre color="negro" precio="12.56">Gorro de lana</nombre>
```

```
</producto>
```

DATOS SEMIESTRUCTURADOS

XML (EXTENSIBLE MARKUP LANGUAGE)

```
<?xml version="1.0" encoding="UTF-8"?>
<biblioteca>
  <libro>
    <titulo>La vida está en otra parte</titulo>
    <autor>Milan Kundera</autor>
    <fechaPublicacion año="1973"/>
  </libro>
  <libro>
    <titulo>Pantaleón y las visitadoras</titulo>
    <autor fechaNacimiento="28/03/1936">Mario Vargas Llosa</autor>
    <fechaPublicacion año="1973"/>
  </libro>
  <libro>
    <titulo>Conversación en la catedral</titulo>
    <autor fechaNacimiento="28/03/1936">Mario Vargas Llosa</autor>
    <fechaPublicacion año="1969"/>
  </libro>
</biblioteca>
```

DATOS SEMIESTRUCTURADOS

JSON (JAVASCRIPT OBJECT NOTATION)

- Es un formato ligero de intercambio de datos
- Consisten en pares de atributos y valores
- Surgió como alternativa más simple y ligera al XML
- Aunque en sus orígenes estuvo ligado a JavaScript, se ha convertido en un estándar independiente de datos
- Rápida aceptación por la rapidez en la lectura y su menor tamaño

DATOS SEMIESTRUCTURADOS

JSON (JAVASCRIPT OBJECT NOTATION)

```
{
  "departamento":8,
  "nombredepto":"Ventas",
  "director": "Juan Rodríguez",
  "empleados":[
    {
      "nombre":"Pedro",
      "apellido":"Fernández"
    },{
      "nombre":"Jacinto",
      "apellido":"Benavente"
    }
  ]
}
```

```
{
  "localizaciones": [
    {
      "latitude": 40.416875,
      "longitude": -3.703308,
      "city": "Madrid",
      "description": "Puerta del Sol"
    },
    {
      "latitude": 40.417438,
      "longitude": -3.693363,
      "city": "Madrid",
      "description": "Paseo del Prado"
    },
    {
      "latitude": 40.407015,
      "longitude": -3.691163,
      "city": "Madrid",
      "description": "Estación de Atocha"
    }
  ]
}
```

DATOS SEMIESTRUCTURADOS

JSON (JAVASCRIPT OBJECT NOTATION)

XML

```
<empinfo>
  <employees>
    <employee>
      <name>James Kirk</name>
      <age>40</age>
    </employee>
    <employee>
      <name>Jean-Luc Picard</name>
      <age>45</age>
    </employee>
    <employee>
      <name>Wesley Crusher</name>
      <age>27</age>
    </employee>
  </employees>
</empinfo>
```

JSON

```
{ "empinfo" :
  {
    "employees" : [
      {
        "name" : "James Kirk",
        "age" : 40,
      },
      {
        "name" : "Jean-Luc Picard",
        "age" : 45,
      },
      {
        "name" : "Wesley Crusher",
        "age" : 27,
      }
    ]
  }
}
```

DATOS SEMIESTRUCTURADOS

JSON (JAVASCRIPT OBJECT NOTATION)

XML	JSON
<pre><Servers> <Server> <name>Server1</name> <owner>John</owner> <created>123456</created> <status>active</status> </Server> </Servers></pre>	<pre>{ Servers: [{ name: Server1, owner: John, created: 123456, status: active }] }</pre>

DATOS SEMIESTRUCTURADOS

JSON (JAVASCRIPT OBJECT NOTATION)

Archivo colores1.json	Archivo colores2.json	Archivo colores3.json
<pre>{ "arrayColores":[{ "nombreColor":"rojo", "valorHexadec":"#f00" }, { "nombreColor":"verde", "valorHexadec":"#0f0" }, { "nombreColor":"azul", "valorHexadec":"#00f" }, { "nombreColor":"cyan", "valorHexadec":"#0ff" }, { "nombreColor":"magenta", "valorHexadec":"#f0f" }, { "nombreColor":"amarillo", "valorHexadec":"#ff0" }, { "nombreColor":"negro", "valorHexadec":"#000" }]</pre>	<pre>{ "arrayColores":[{ "rojo":"#f00", "verde":"#0f0", "azul":"#00f", "cyan":"#0ff", "magenta":"#f0f", "amarillo":"#ff0", "negro":"#000" }]</pre>	<pre>{ "rojo":"#f00", "verde":"#0f0", "azul":"#00f", "cyan":"#0ff", "magenta":"#f0f", "amarillo":"#ff0", "negro":"#000" }</pre>

DATOS SEMIESTRUCTURADOS

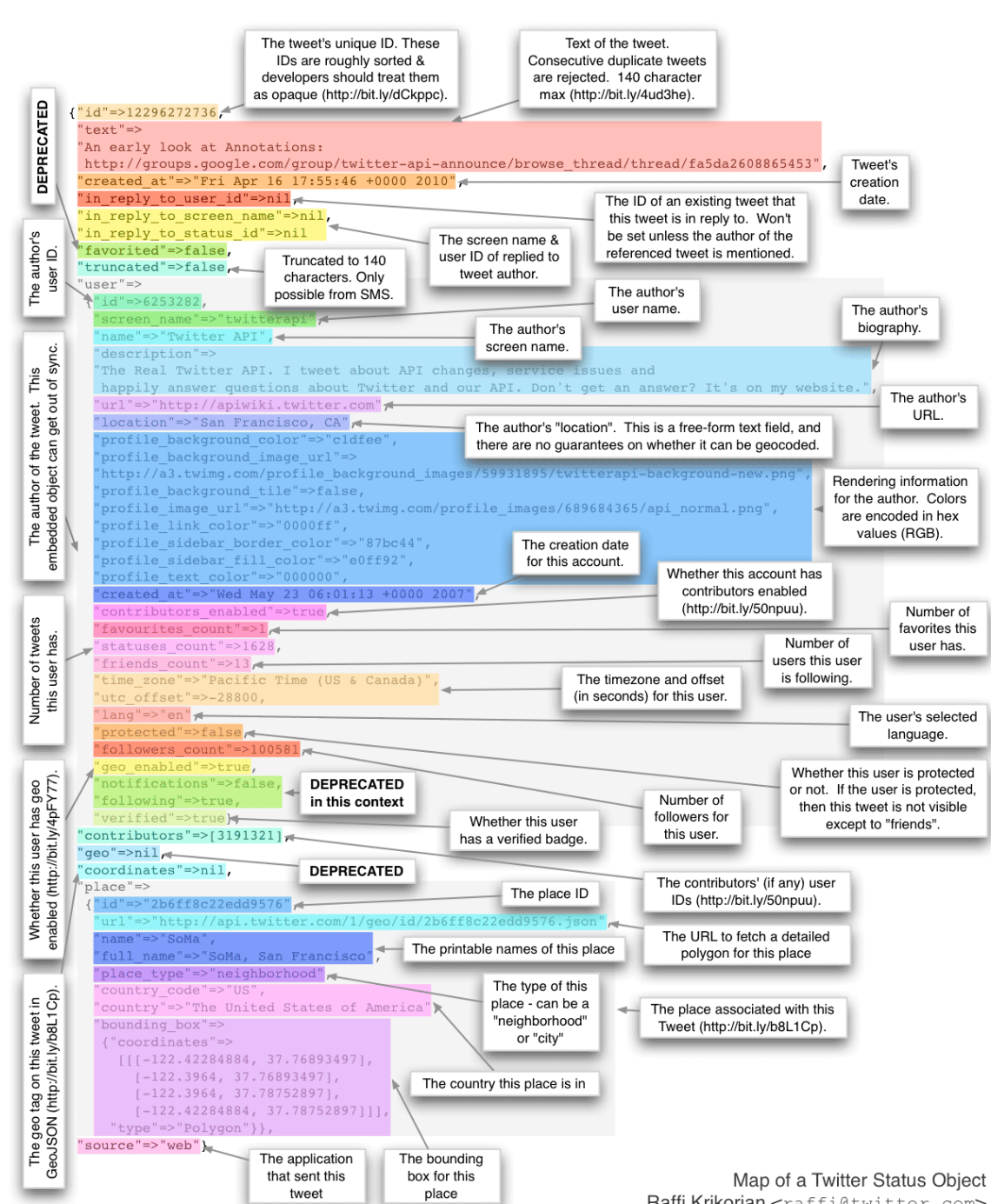
METADATOS

- Los metadatos son los "datos que describen datos"
- Están asociados a la mayoría de la información que se produce en el mundo digital
- Tratados masivamente y con técnicas de Big Data son una fuente extraordinaria - e involuntaria - de información personal
- Es común codificar metadatos usando XML o JSON

DATOS SEMIESTRUCTURADOS METADATOS

Anatomía de un tweet:

You are your Metadata: Identification and Obfuscation of Social Media Users using Metadata Information (2018) University College de Londres e Instituto Alan Turing

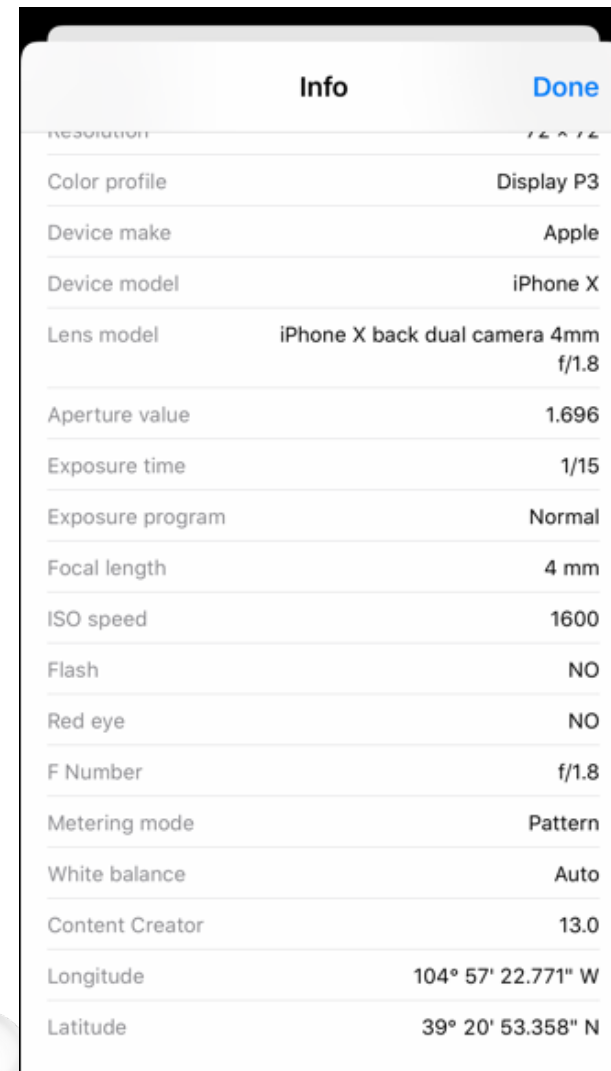


DATOS SEMIESTRUCTURADOS

METADATOS

EXIF (Exchangeable image file format)

Metadatos incrustados dentro del propio
archivo de imagen

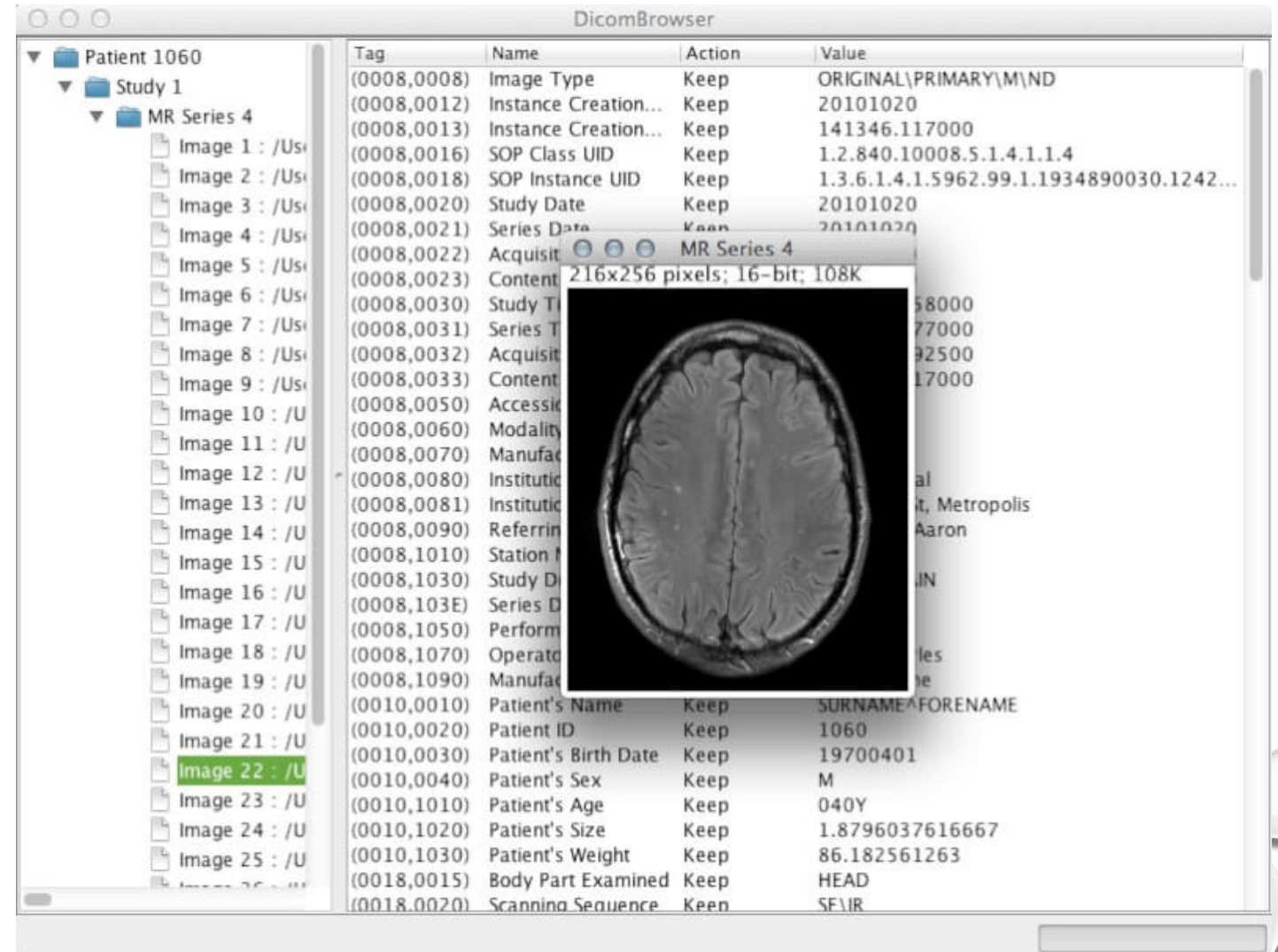


Info		Done
Resolution	14.0 x 14.0	
Color profile	Display P3	
Device make	Apple	
Device model	iPhone X	
Lens model	iPhone X back dual camera 4mm f/1.8	
Aperture value	1.696	
Exposure time	1/15	
Exposure program	Normal	
Focal length	4 mm	
ISO speed	1600	
Flash	NO	
Red eye	NO	
F Number	f/1.8	
Metering mode	Pattern	
White balance	Auto	
Content Creator	13.0	
Longitude	104° 57' 22.771" W	
Latitude	39° 20' 53.358" N	

DATOS SEMIESTRUCTURADOS METADATOS

DICOM (Digital Imaging and
Communication On Medicine)

Estándar de transmisión y
almacenamiento de imágenes
médicas



DATOS SEMIESTRUCTURADOS

METADATOS

- En algunas ocasiones, los datos no estructurados se pueden clasificar dentro de los semiestructurados porque tienen uno o más atributos de clasificación

Ejemplo: Correo electrónico

TIPOS DE DATOS ALMACENAMIENTO

- Datos estructurados: Bases de Datos **SQL**
- Datos semiestructurados y no estructurados: Bases de Datos **NoSQL**
(Not Only SQL)
- Existe una amplia variedad de Bases de Datos NoSQL

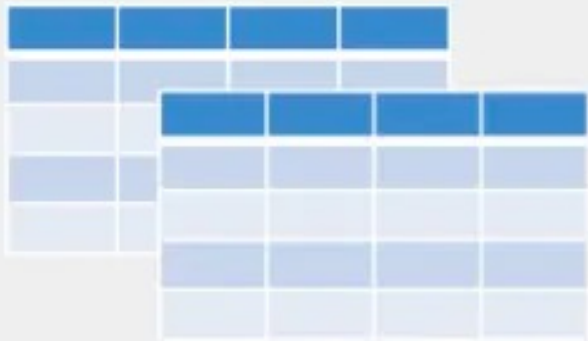
¿Flexibilidad?

¿Facilidad de análisis?

BBDD NoSQL

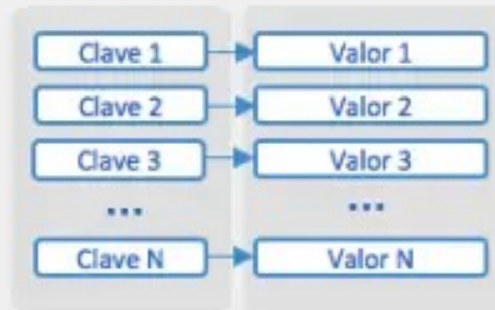
Base de datos SQL

Relacional

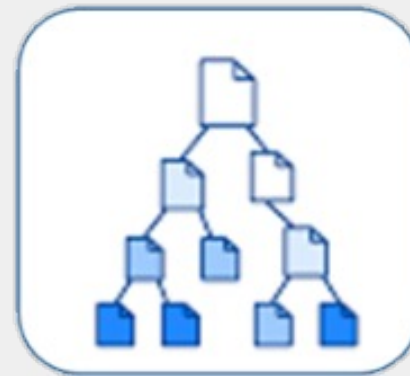


Base de datos NoSQL

Clave-Valor



Documental



Grafos



BBDD NoSQL

MODELO

CARACTERÍSTICAS

BB Clave-Valor

La más sencilla de las bases de datos NoSQL, los datos se representan como una colección de pares clave-valor. Los valores no requieren un esquema fijo. No existe el concepto de relaciones, pensadas para almacenar información básica y que pueda ser consultada de forma muy rápida.

BBDD de documentos

Los datos se almacenan de forma jerárquica en documentos basados en JSON, XML y BSON. Cada documento puede tener la misma estructura o una estructura diferente.

BBDD de grafos

Los datos se almacenan en una estructura de grafos

Clave - Key

70248160-N

32123844-Z

...

233456879-A

Valor - Value

Juan Antonio Pérez natural de Villa del Mar, estudio en las mejores universidades del EEUU y ha sido padre recientemente.

Manuel Rodríguez es fontanero de Olmedo y no esta casado pero le encantaría encontrar pareja.

...

Paloma García es una importante escritora de literatura infantil muy querida en su ciudad natal.

BBDD NoSQL

student_id	age	score
1	12	77
2	12	68
3	11	75



```
[  
  {  
    "student_id":1,  
    "age":12,  
    "score":77  
  },  
  {  
    "student_id":2,  
    "age":12,  
    "score":68  
  },  
  {  
    "student_id":3,  
    "age":11,  
    "score":75  
  }  
]
```

CALIDAD DE LOS DATOS

- La mala calidad de los datos es el **principal riesgo** al que se enfrenta la ciencia de datos
- Uno de los problemas "ocultos" más graves y persistentes. Implica:
 - Toma de decisiones estratégicas no acertadas
 - Deterioro en la imagen corporativa de la compañía
 - Ineficiencia en la toma de decisiones
 - Mala gestión de los clientes

CALIDAD DE LOS DATOS

- En Big Data resulta todavía más complicado lidiar con la calidad ya que los datos se suelen originar fuera del proyecto y tienen una vida independiente más allá de éste
- El investigador no tiene un control total sobre los datos
- Garbage in, garbage out (GIGO)
- El éxito depende principalmente de los datos de entrada



CALIDAD DE LOS DATOS

Big Data vs Smart Data






CALIDAD DE LOS DATOS

Big Data vs Smart Data

Los datos en sí mismos no generan ventajas competitivas, hay que extraer su valor a partir de su procesamiento y análisis



CALIDAD DE LOS DATOS

Big Data vs Smart Data

- Calidad de los datos: punto de inflexión entre ambos
- Smart Data son datos de calidad, listos para ser utilizados en la extracción de conocimiento y toma de decisiones
- **Preprocesamiento de datos:** fundamental para convertir datos almacenados en datos de calidad

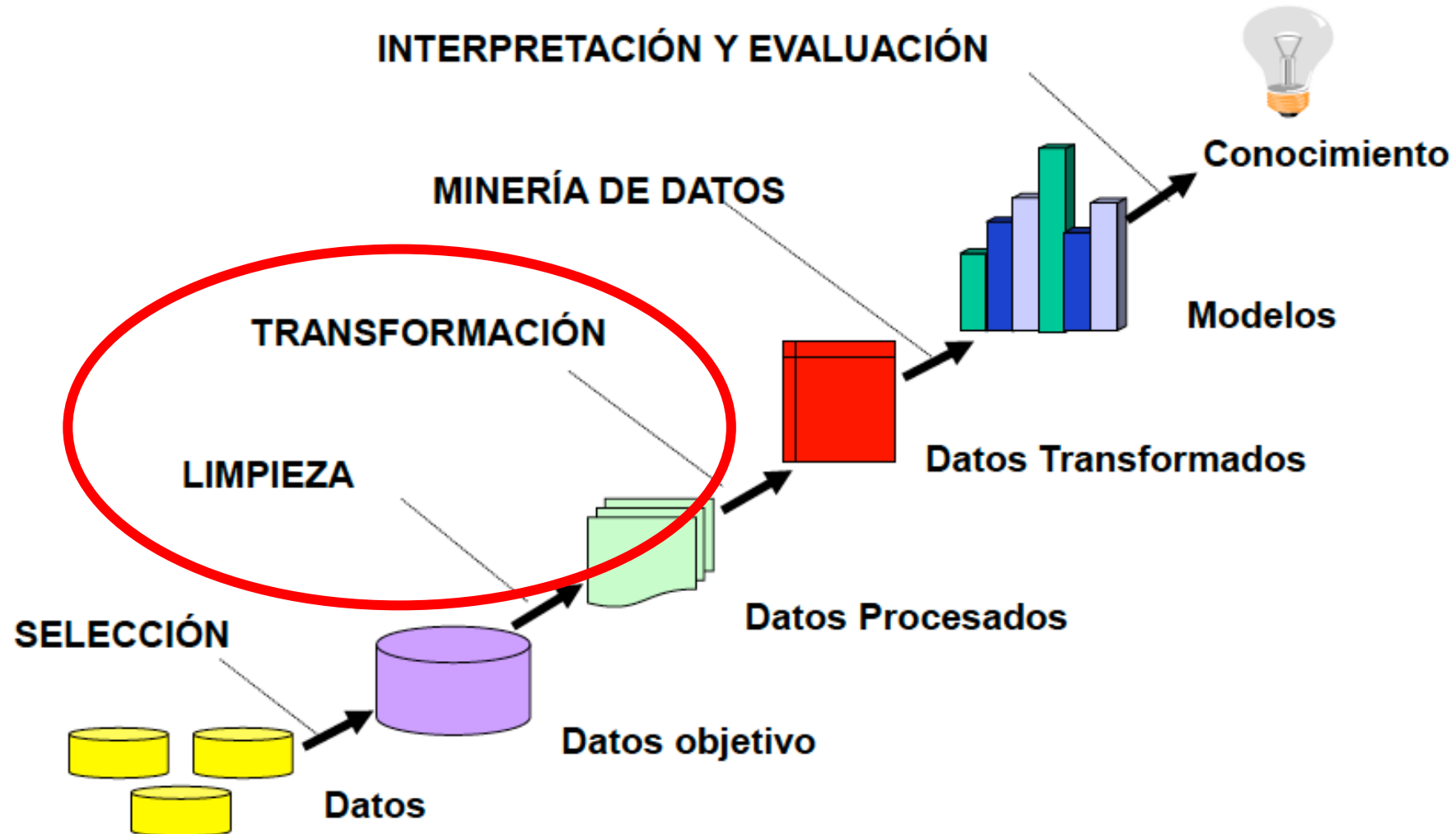
PREPROCESAMIENTO DE LOS DATOS

¿Por qué es importante preprocesar datos?

Los datos reales pueden ser impuros, lo que provoca la extracción de patrones o reglas poco útiles. Causas:

- Datos incompletos: falta de valores de atributos, ...
- Datos con ruido
- Datos inconsistentes (incluyendo discrepancias)

PREPROCESAMIENTO DE LOS DATOS



PREPROCESAMIENTO DE LOS DATOS

- Obtener un conjunto de datos final que sea de calidad y útil para la fase de extracción de conocimiento
- Pasos:
 - **Limpieza de los datos**
 - **Transformación de los datos**
 - Reducción de la dimensionalidad

LIMPIEZA DE DATOS

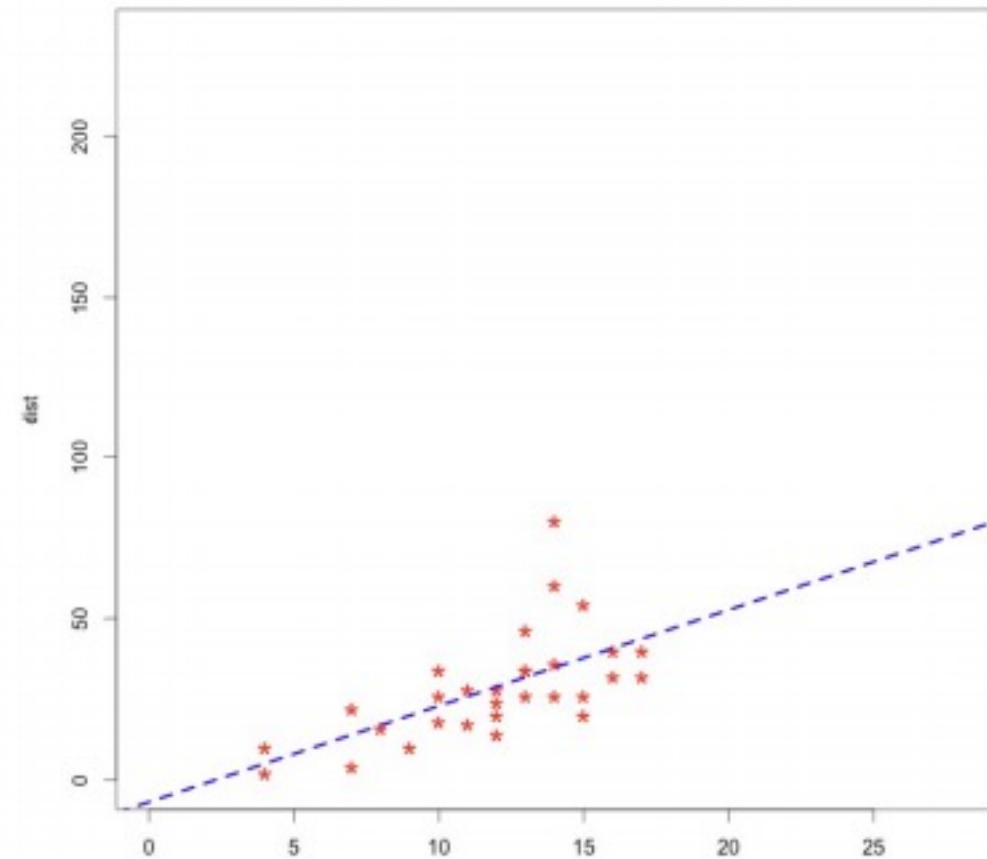
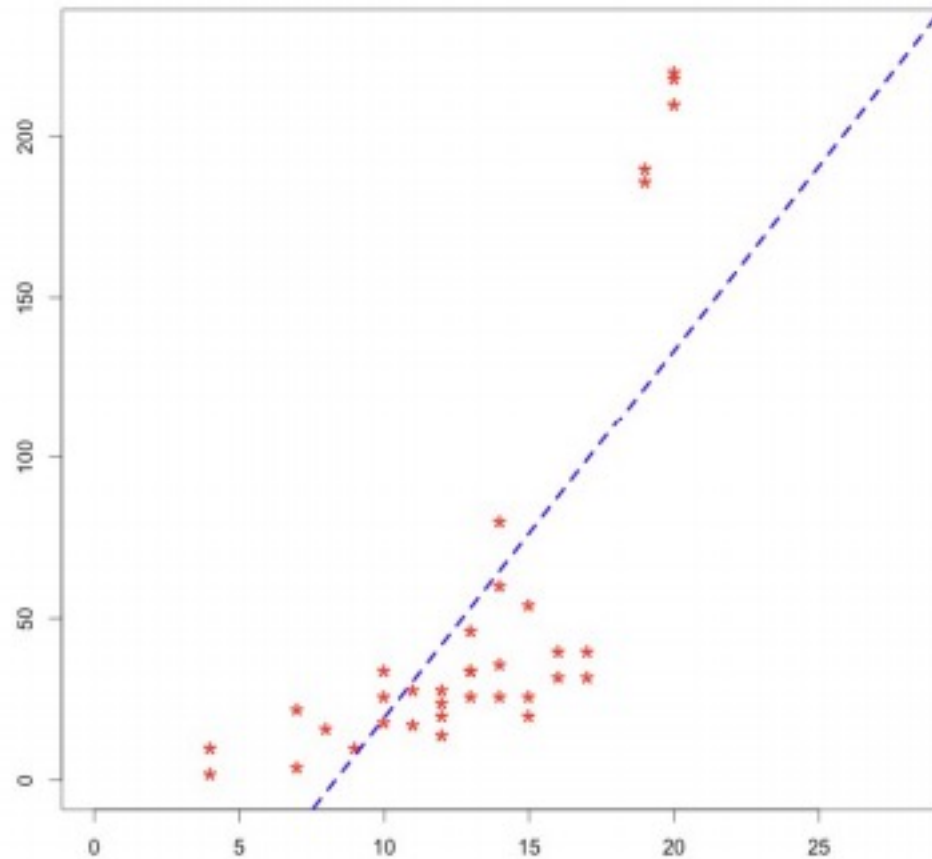
1. Valores ausentes o nulos:

- **Interpolación:** datos temporales, interpolar teniendo en cuenta los datos próximos
- **Rellenar con un valor fijo:** media, moda o valor 0.
- **Rellenar utilizando regresión:** predecir el valor perdido utilizando el resto de las variables del conjunto de datos.
- **Considerar el vacío como una categoría:** variables categóricas
- **Eliminar el registro completo**

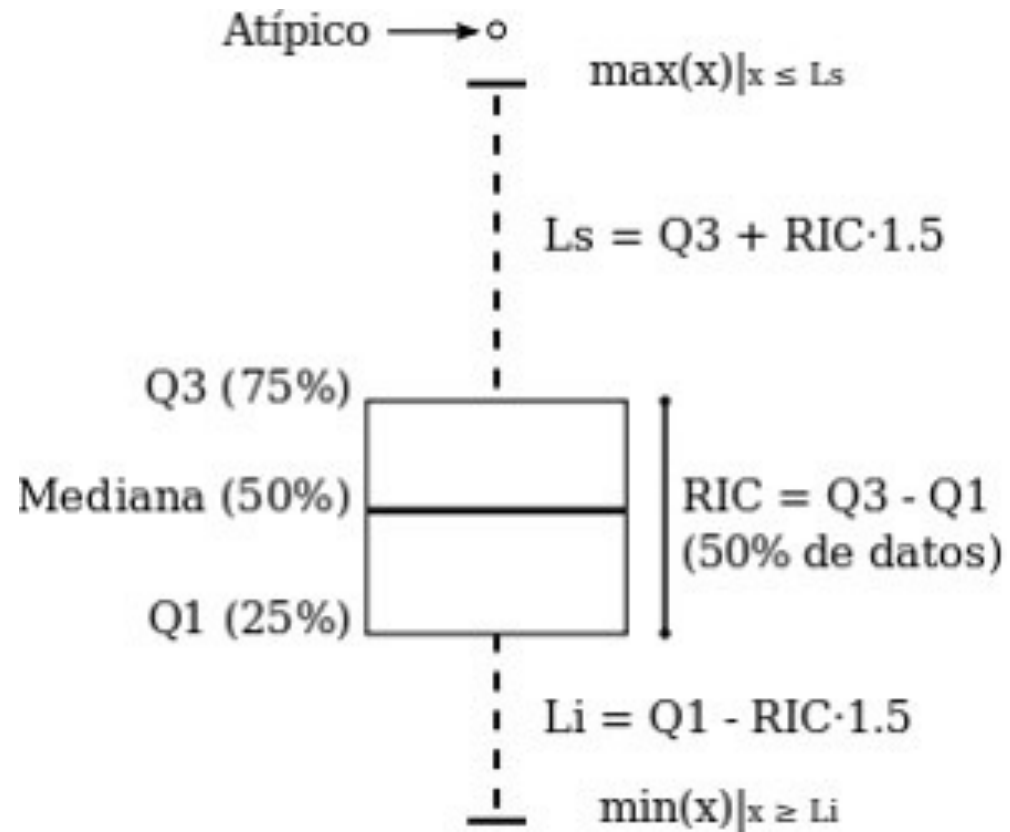
LIMPIEZA DE DATOS

2. **Inconsistencia de datos**: Errores en el formato o tipo de datos
3. **Valores duplicados**: Eliminar para evitar sesgos
4. **Datos anómalos o atípicos (Outliers)**: Distorsión de los datos

LIMPIEZA DE DATOS



LIMPIEZA DE DATOS



LIMPIEZA DE DATOS

Alrededor del 90% de los potenciales errores de calidad de datos se resuelven con:

1. Análisis de nulos y atípicos
2. Estadísticos básicos
3. Análisis longitudinal
4. Coherencia entre variables

TRANSFORMACIÓN DE DATOS

1. **Normalizar**: Escalar las variables a un intervalo (para distancias)

- Mínima-máxima

$$\frac{x - \min(x)}{[\max(x) - \min(x)]}$$

- Puntuación z

$$\frac{x - \text{mean}(x)}{\text{stdev}(x)}$$

2. **Discretizar**: Convertir valores continuos en intervalos. Reducir el tamaño de datos y mejorar la precisión

- Misma amplitud
- Misma frecuencia

3. **Variables sintéticas o derivadas**

4. **Formatear los datos para facilitar el análisis**

TRANSFORMACIÓN DE DATOS

Día	Encargados		Día	Encargado 1	Encargado 2	Encargado 3		Día	Encargado
Lunes	Jose, Carlos, Antonio	➔	Lunes	Jose	Carlos	Antonio	➔	Lunes	Jose
Martes	David, Jose, Pablo		Martes	David	Jose	Pablo		Lunes	Carlos
Miércoles	Carlos, Javier, Pablo		Miércoles	Carlos	Javier	Pablo		Lunes	Antonio
								Martes	David
								Martes	Jose
								Martes	Pablo
								Miércoles	Carlos
								Miércoles	Javier
								Miércoles	Pablo

Comunidad Autónoma	Población		Comunidad Autónoma	2016	2015	2014
Andalucía			Andalucía	8.388.107	8.399.043	8.402.305
	2016	8.388.107	Aragón	1.308.563	1.317.847	1.325.385
	2015	8.399.043	Asturias, Principado de	1.042.608	1.051.229	1.061.756
	2014	8.402.305	Balears, Illes	1.107.220	1.104.479	1.103.442
Aragón			Canarias	2.101.924	2.100.306	2.104.815
	2016	1.308.563				
	2015	1.317.847				
	2014	1.325.385				
Asturias, Principado de						
	2016	1.042.608				
	2015	1.051.229				
	2014	1.061.756				
Balears, Illes						
	2016	1.107.220				
	2015	1.104.479				
	2014	1.103.442				
Canarias						
	2016	2.101.924				
	2015	2.100.306				
	2014	2.104.815				

Comunidad Autónoma	Año	Población
Andalucía	2016	8.388.107
Andalucía	2015	8.399.043
Andalucía	2014	8.402.305
Aragón	2016	1.308.563
Aragón	2015	1.317.847
Aragón	2014	1.325.385
Asturias, Principado de	2016	1.042.608
Asturias, Principado de	2015	1.051.229
Asturias, Principado de	2014	1.061.756
Balears, Illes	2016	1.107.220
Balears, Illes	2015	1.104.479
Balears, Illes	2014	1.103.442
Canarias	2016	2.101.924
Canarias	2015	2.100.306
Canarias	2014	2.104.815