Tema 1
Métodos de Captura de Información

Métodos de captura de información

Origen y calidad de los datos

Evaluación de calidad:

- **Completitud**: Grado en el que los valores se encuentran en un conjunto de datos.
- **Credibilidad**: Nivel de fiabilidad del organismo que proporciona el conjunto de datos.
- **Consistencia**: Grado en el que los datos carecen de contradicciones.
- **Interpretabilidad**: Grado en el que los datos deben ser interpretados por una persona.
- Precisión: Nivel de exactitud del valor.

Niveles de abstracción:

- **Datos**: Conjunto de hechos discretos y objetivos sobre un evento.
- **Información**: Datos con significado.
- **Conocimiento**: Combinación de información contextualizada, experiencias, valores e intuición.

Fuentes de datos:

- Captura manual: Encuestas y observaciones.
- Análisis de documentos estructurados: estructurados (HTML) y sin formato (lenguaje natural)
- Salida de aplicaciones: Logs o bases de datos.
- Sensores: Dispositivos de medición.
- Datos de acceso público: Gubernamentales y servicios web públicos.

Organización de los datos

Ficheros planos:

- **CSV** (comma separated values): RFC 4180. Define un registro por línea y separar los campos por comas.
- **JSON**: RFC 7159 y ECMA-404. Describe objetos encapsulados por llaves y listas por corchetes.
- **XML**: Descrito en el estándar XML 1.0 de W3C. Permite almacenar información de forma legible utilizando etiquetas.

Bases de datos:

- Conjunto de datos persistentes, utilizados por sistemas de aplicación.
- En el modelo Entidad-Relación (E/R) una entidad es cualquier objeto repensado en la BBDD y un vínculo representa relaciones entre ellos.
- La unidad básica de almacenamiento es el campo, agrupados en registros y estos en ficheros almacenados.

Bases de datos relacionales y SQL

- Los ficheros almacenados se representan en forma de tablas (relaciones), con columnas (campos) y filas (registros).
- El estándar SQL define un lenguaje para la consulta y modificación de los datos.
- El comando SELECT permite consultar información de tablas.
- Los comandos INSERT, UPDATE y DELETE permiten la inserción, edición y eliminación de registros respectivamente.

Casos de estudio

- Procesamiento de sitio web sobre cursos online
- Procesamiento de *logs* de servidor web
- API de acceso a transacciones bancarias
- Almacenamiento de información sobre productos en un fichero CSV
- Representación de información geolocalizada en formato JSON
- Almacenamiento de información sobre clientes de una base de datos relacional

Persistencia

- El método más básico para almacenar datos es mediante el uso del sistema de ficheros del sistema operativo.
- Los ficheros pueden tener un formato plano, donde toda la información es legible para una persona; o un formato binario, donde la información puede escribirse y leerse de forma directa por una aplicación, pero no puede ser analizada directamente de forma manual.

- Se habla de dos aproximaciones: la utilización de ficheros planos, que son comúnmente utilizados para el almacenamiento y compartición de datos.
- Las bases de datos dan un paso más allá y proporcionan una consistencia en la información y el hecho de poder **consultar** y **modificar** de **manera eficiente** un conjunto de datos en específico.

Almacenamiento de datos

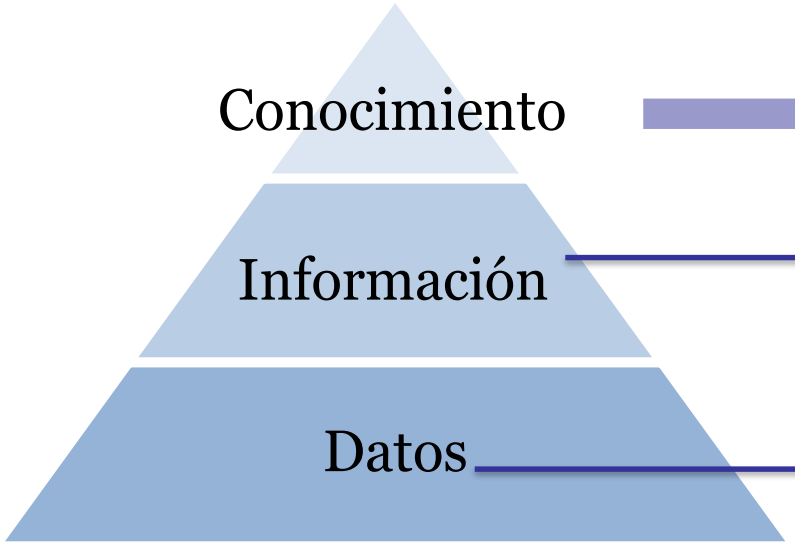
- Sistema de ficheros planos
  - CSV, XML, JSON, TXT…
- Sistema de ficheros binarios
  - BSON, CLASS…

```
CrimeId;OriginalCrimeTypeName;OffenseDate;CallTime;CallDateTime
160903280;Assault / Battery;2016-03-30T00:00:00;18:42;2016-03-3
160912272;Homeless Complaint;2016-03-31T00:00:00;15:31;2016-03-
160912590;Susp Info;2016-03-31T00:00:00;16:49;2016-03-31T16:49:
160912801;Report;2016-03-31T00:00:00;17:38;2016-03-31T17:38:00;
160912811;594;2016-03-31T00:00:00;17:42;2016-03-31T17:42:00;REP
160913003;Ref'd;2016-03-31T00:00:00;18:29;2016-03-31T18:29:00;G
160913050;Homeless Complaint;2016-03-31T00:00:00;18:43;2016-03-
160913056;Homeless Complaint;2016-03-31T00:00:00;18:47;2016-03-
160913078;Agg Assault / Adw Dv;2016-03-31T00:00:00;18:52;2016-0
160913103;Encampment;2016-03-31T00:00:00;18:57;2016-03-31T18:57
160913118;Burglary;2016-03-31T00:00:00;18:59;2016-03-31T18:59:0
160913148;Suspicious Person;2016-03-31T00:00:00;19:08;2016-03-3
160913167;Ip;2016-03-31T00:00:00;19:13;2016-03-31T19:13:00;HAN;
```

```
- <employees>
  - <person id="1392">
      <name>John Smith</name>
      <dob>1974-07-25</dob>
      <start-date>2004-08-01</start-date>
      <salary currency="USD">35000</salary>
    </person>
  - <person id="1395">
      <name>Clara Tennison</name>
      <dob>1968-03-15</dob>
      <start-date>2003-05-16</start-date>
      <salary currency="USD">27000</salary>
    </person>
  </employees>
```

```
[
  {
    "id_terraza": 7,
    "id_local": 280067128,
    "id_distrito_local": 20,
    "desc_distrito_local": "SAN BLAS-CANILLEJAS",
    "desc_barrio_local": "ROSAS",
    "clase_vial_edificio": "CALLE",
    "num_edificio": 177,
    "Cod_Postal": 28022,
    "coordenada_x_local": "448900,55",
    "coordenada_y_local": "4474755,41",
    "desc_situacion_local": "Abierto",
    "Nombre": "SR",
    "id_periodo_terraza": 2,
    "desc_periodo_terraza": "Estacional",
    "id_situacion_terraza": 1,
    "desc_situacion_terraza": "Abierta",
    "Superficie_ES": "7,2"
  },
```

Jerarquía del conocimiento



Datos de ejemplo					
Marca	**Modelo**	**Precio**	**Motor**	**Consumo**	**Fecha**
Audi	La Ferrari	15.000 €	963 CV diésel	alto	2014
Lamborghini	Murciélago	300.000 €	580 CV 7500rpm		2011
BMW	M6	158.000€	560CV 4935rpm	13,6/7,6/9,9	2012
Porsche	911 turbo	127.000 €	420CV 7500rpm	11 litros	2017

Transformación del dato

- Contextualización: conocer el propósito del dato obtenido.
- Categorización: conocer la unidad de medida y los componentes del dato.
- Cálculo: realizar una operación matemática sobre el dato.
- Corrección: eliminar errores del dato.
- Agregación: resumir o minimizar un dato de forma más concisa.

Transformación de la información

- Comparación: relación entre información obtenida en distintas experiencias.
- Repercusión: implicación de la información en decisiones y acciones.
- Conexión: relación entre distintos tipos de información.
- Conversación: opinión de otras personas sobre la información.

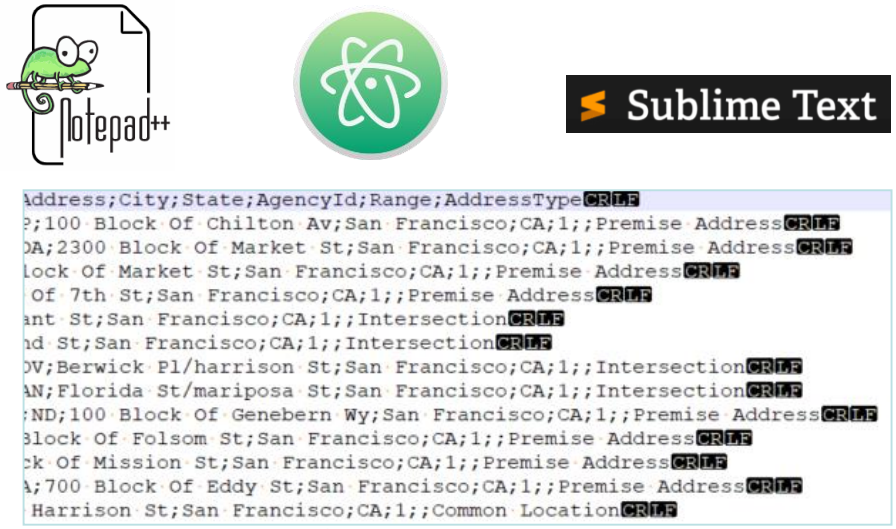
Calidad del dato

- Completitud: % de datos disponibles respecto a la población total que representan dichos datos.
- Credibilidad: fiabilidad que se le brinda al organismo que proporciona el conjunto de datos.
- Precisión: porcentaje de datos correctos respecto al total disponible.
- Consistencia: nivel con el que los datos son coherentes entre ellos.
- Interpretabilidad: grado en el que los datos pueden ser entendidos.

Ficheros planos: CSV

- Intercambio de información entre aplicaciones.
- La primera línea contiene los **nombres de los campos*** *opcional.
- Cada registro se delimita por **cambio de línea** (CR y LF).
- Los valores de cada registro están **separados por comas**.
- Los valores pueden estar entre **comillas dobles**.
- Si un valor contiene comillas, estas deben **escaparse**

  "Negocio ""grande"" "



Elegid un editor y quedaos con él.

Ficheros planos: JSON (JavaScript Object Notation)

- Intercambio de información entre aplicaciones.
- **Objeto** o registro definido por conjuntos pares nombre/valor.
- Un **array** o lista ordenada de valores.
- Delimitado por { } y los pares nombre/valor separados por comas.
- Un array se delimita por [ ] y los valores se separan con comas.
- **Tipo de valor**: “cadena”, número, booleano, nulo (NULL), otro objeto, un array.

```
[{
    “Nombre”: “Juan”,
    “Edad”: 45,
    “Cargo”: “Director”
}, {
    “Nombre”: “Antonio”,
    “Edad”: 35,
    “Cargo”: “Gestor de proyectos”
}]
```

Ficheros planos: XML (eXtended Markup Language)

- Intercambio de información entre aplicaciones.
- Inicia con la línea: <?xml version="1.0">.
- Tiene solamente un elemento raíz.
- Elemento apertura <etiqueta> y elemento de cierre </etiqueta>.
- Los elementos pueden tener atributos <etiqueta id="miEle">.
- Contenido del elemento: texto, uno o más elementos, combinación de ambos.

```
- <employees>
  - <person id="1392">
      <name>John Smith</name>
      <dob>1974-07-25</dob>
      <start-date>2004-08-01</start-date>
      <salary currency="USD">35000</salary>
    </person>
  - <person id="1395">
      <name>Clara Tennison</name>
      <dob>1968-03-15</dob>
      <start-date>2003-05-16</start-date>
      <salary currency="USD">27000</salary>
    </person>
  </employees>
```

Tratar ficheros desde Python

Crear un fichero JSON desde Python

In [9]:
```
import json

data = {}

data['people'] = []

data['people'].append({
    'name': 'Maria',
    'email': 'maria@unir.net',
    'country': 'Ecuador'
})

data['people'].append({
    'name': 'Laura',
    'email': 'laura@unir.net',
    'country': 'Colombia'
})

data['people'].append({
    'name': 'Ana',
    'email': 'ana@unir.net',
    'country': 'Panamá'
})

with open('data.txt', 'w') as outfile:
    json.dump(data, outfile)
```

In [10]:
```
import json

with open('data.txt') as json_file:
    data = json.load(json_file)
    for p in data['people']:
        print('Name: ' + p['name'])
        print('Email: ' + p['email'])
        print('Country: ' + p['country'])
        print('')
```

ANACONDA®

python

Bases de datos

- Conjunto de datos persistente utilizado por un sistema de software.
- Componentes:
  - Datos
  - Hardware
  - Software: DBMS Database Management System
  - Usuarios: programadores, usuarios finales, administradores de BD.
- Bases de datos relacionales SQL.
- Bases de datos No relacionales NoSQL.
- **CRUD**: Create, Read, Update and Delete

Muchas gracias por tu atención