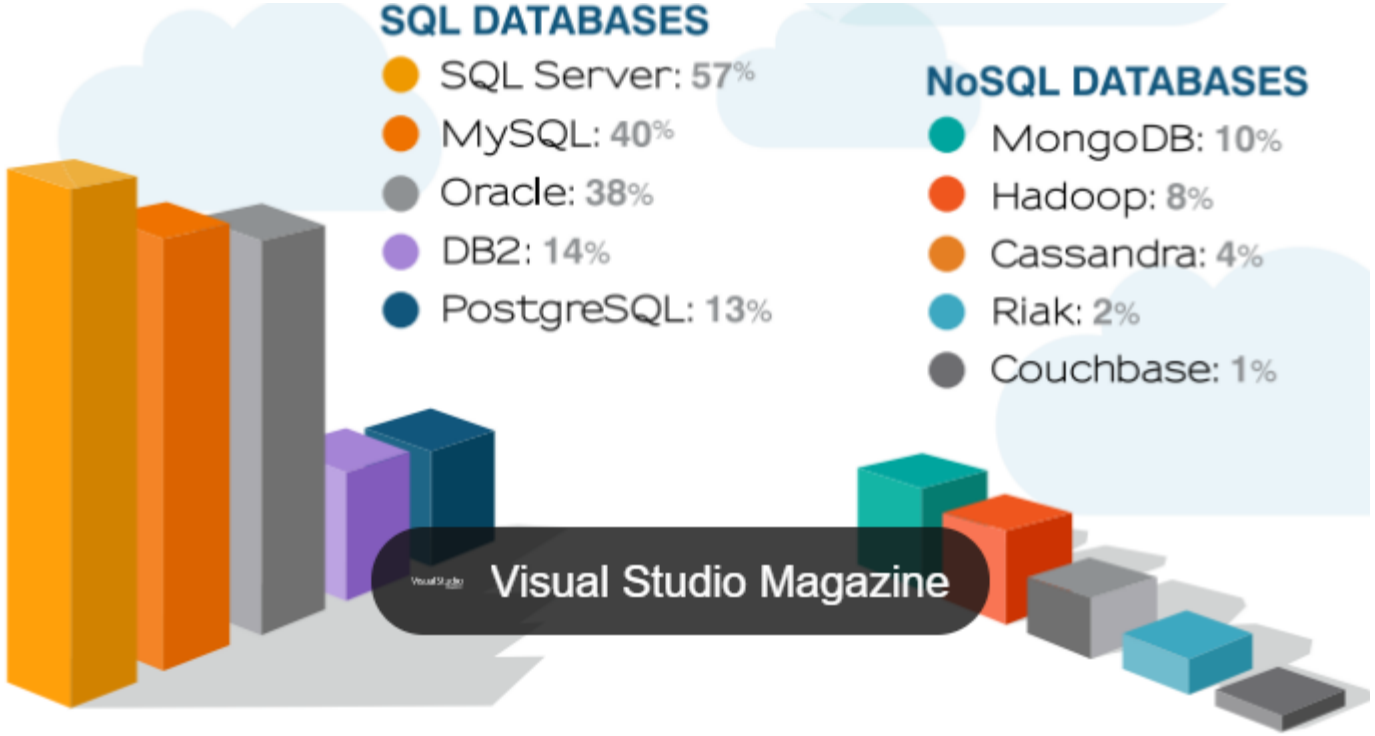
Tema 1
Introducción a la asignatura

Contenido

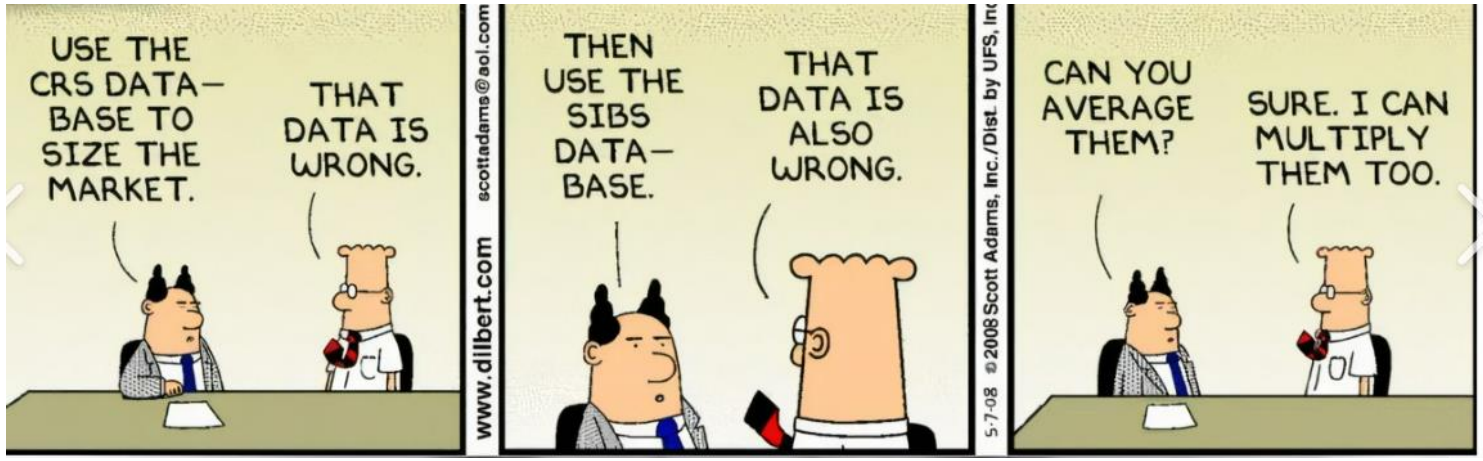
Origen y calidad de los datos

Organización de los datos

Casos de estudio



Contenido



Origen y calidad de los datos

Evaluación de la 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

Fuentes de datos

Organización de los datos

Casos de estudio

Contenido

Origen y calidad de los datos

Evaluación de la calidad

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

Organización de los datos

Casos de estudio

Contenido

Origen y calidad de los datos

- Evaluación de la calidad
- Niveles de abstracción
- Fuentes de datos
  - **Captura manual**: encuestas y observaciones.-
  - **Análisis de documentos estructurados**: estructurados (HTML) y sin formato (lenguaje natural).-
  - **Salida de aplicaciones**: *logs* bases de datos.-
  - **Sensores**: dispositivos de medición.-
  - **Datos de acceso público**: gubernamentales y servicios web públicos

Organización de los datos

Casos de estudio

Contenido

Origen y calidad de los datos

Organización de los datos

- Archivos planos
- Bases de datos

Casos de estudio

Contenido

Origen y calidad de los datos

Organización de los datos

- Archivos planos
  - **CSV:** RFC 4180. Define un registro por línea y separa 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

Casos de estudio

Contenido

Origen y calidad de los datos

Organización de los datos

- Archivos planos
- **Bases de datos**
  - Conjunto de datos persistentes, utilizados por sistemas de aplicación.-
  - Diferentes modelos de datos acorde con los tipos de aplicación

Casos de estudio

Contenido

Origen y calidad de los datos

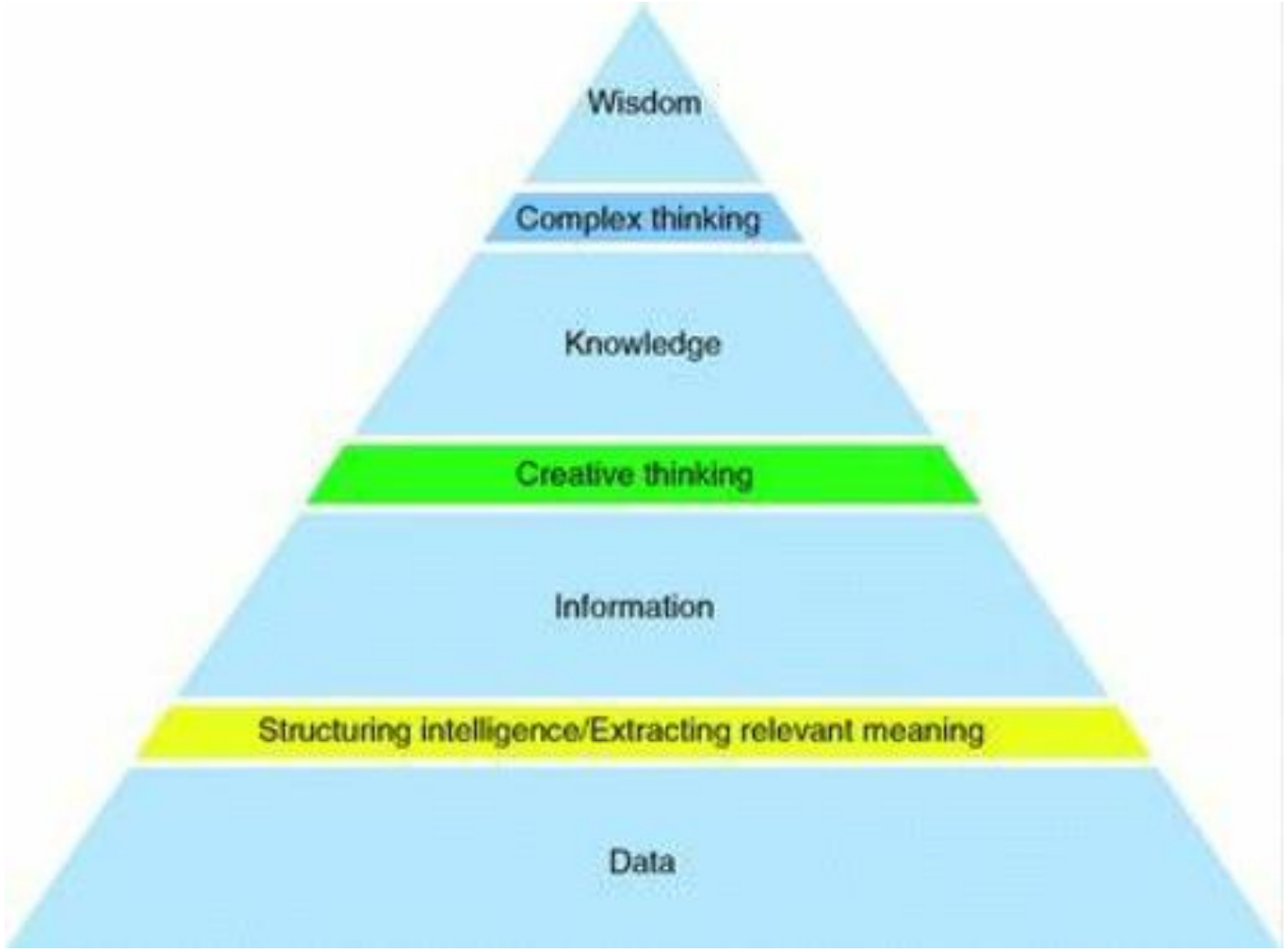
Organización de los datos

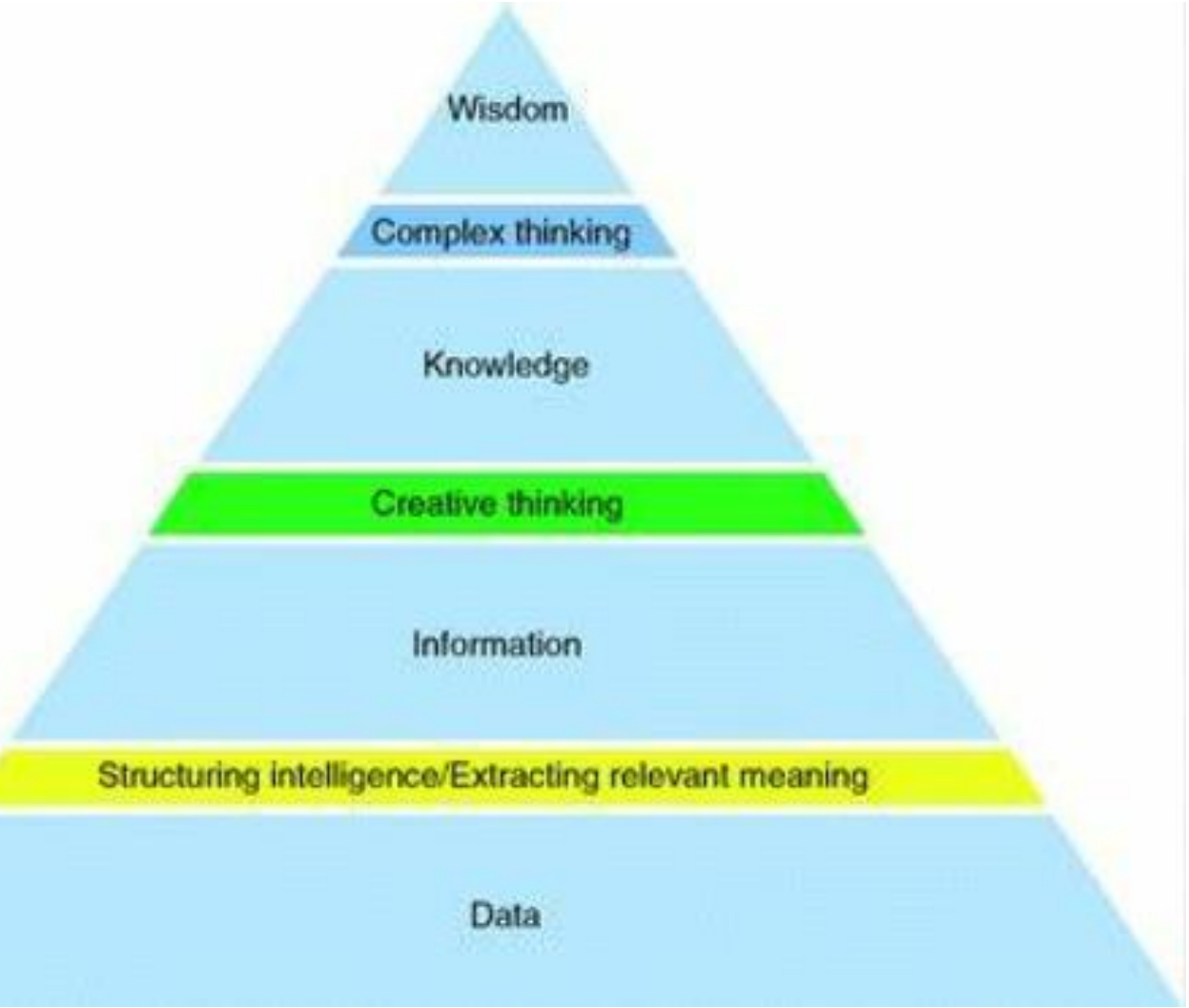
Casos de estudio

- Procesamiento de sitio web sobre cursos *online*
- Procesamiento de *logs* de un servidor web
- API de acceso a transacciones bancarias
- Almacenamiento de información sobre productos en un archivo CSV.
- Almacenamiento de información sobre clientes de una base de datos relacional

La lectura de los casos de estudio también será de ayuda, ya que proporcionan información de las decisiones tomadas para la captura de datos en un entorno determinado.

Dato, información, conocimiento y sabiduría





JSON (JavaScript Object Notation)		CSV (Comma Separated Values)
File Size		
Larger file size.		Compact file size.
Hierarchy		
Supports hierarchical and relational data.		Errors when displaying hierarchical data.
Scalability		
Allows scalability and integrates with APIs easily.		Not easily scalable and difficult to integrate.
Best For		
Works best for complex and large-scale datasets.		Convenient for small datasets.

Un dato puede definirse como un hecho concreto y discreto acerca de un evento. La característica de ser discreto significa que, **semánticamente, es la unidad mínima que puede comunicarse o almacenarse**. Por sí solos, los datos no brindan detalles significantes del entorno del que fueron obtenidos.

mongoDB

cassandra Apache Cassandra

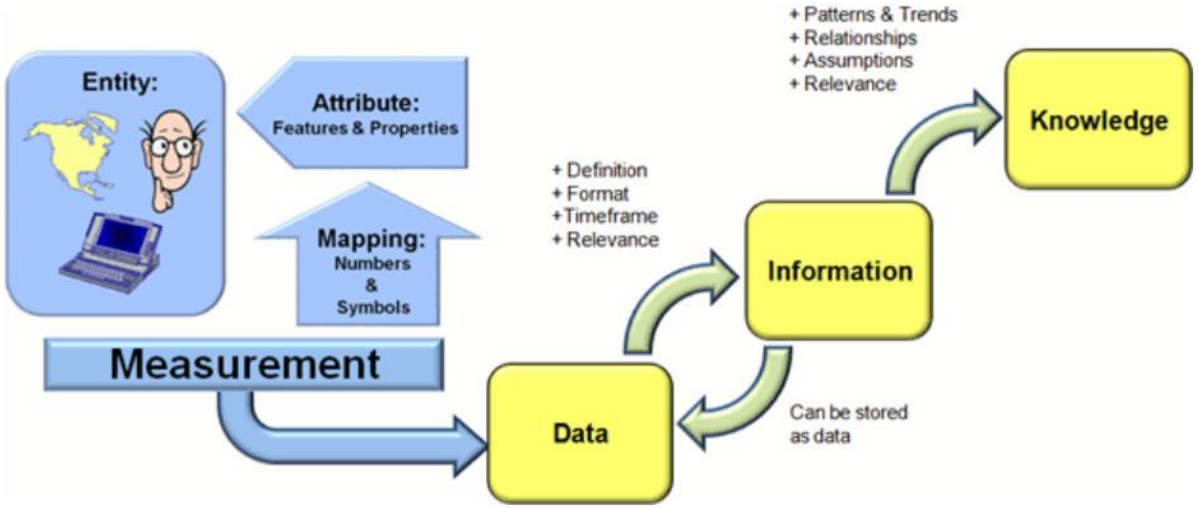
neo4j

La **información** puede distinguirse simplemente como un **mensaje** formado por la **composición de varios datos**. Esto significa que, a diferencia del dato, la información sí posee un significado para un receptor u observador.

De los datos a la información
information is data in context

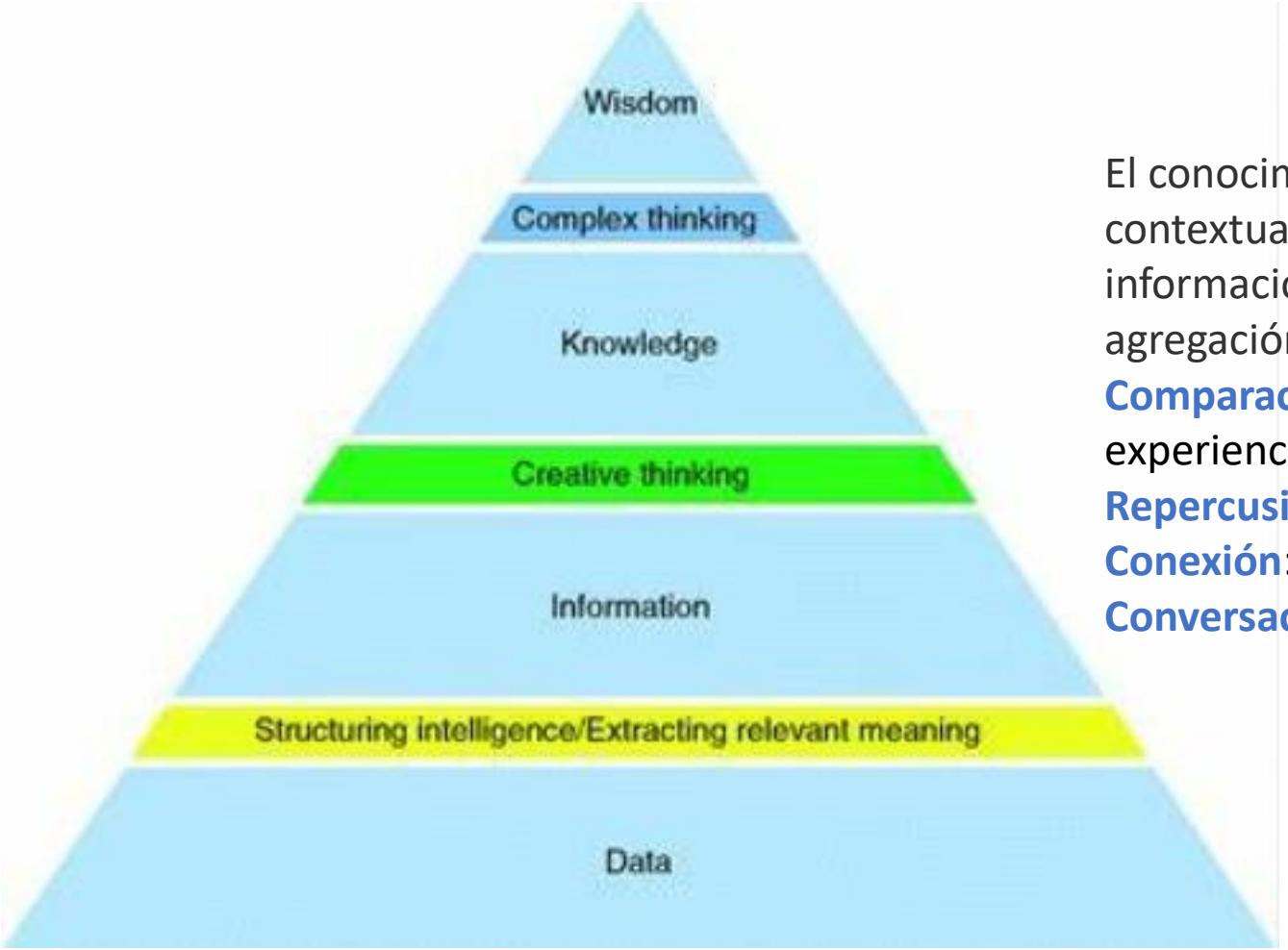
Los datos deben ser transformados para añadirles valor y convertirlos en información.

- **Contextualización**: conocer el propósito del dato obtenido.
- **Categorización**: conocer la unidad de medida y los componentes del dato.
- **Cálculos**: realizar operaciones matemáticas sobre el dato.
- **Corrección**: eliminar errores del dato.
- **Agregación**: resumir o minimizar un dato de forma más concisa.



Data to Information to Knowledge | Welcome to Westfallteam.com

Conocimiento



El conocimiento implica una combinación de experiencias, información contextual y relevancia sobre cierta información. Así como la información se genera a partir de datos, el conocimiento surge de la agregación de información.

Comparación: relación entre información obtenida en distintas experiencias.

Repercusión: implicación de la información en decisiones y acciones.

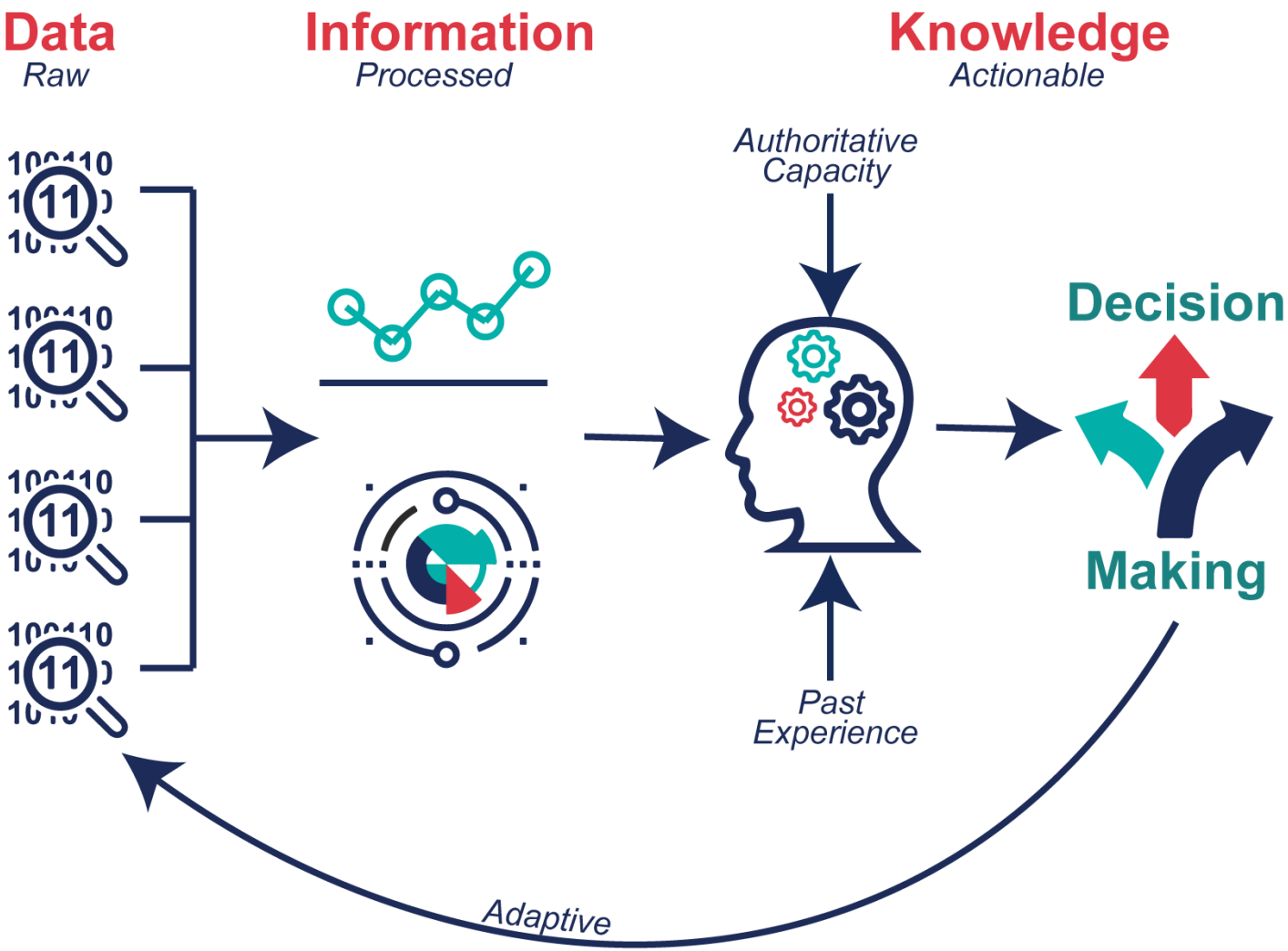
Conexión: relación entre distintos tipos de información.

Conversación: opinión de otras personas sobre la información.

Wisdom

Para que el conocimiento resulte en acción, un individuo debe tener la autoridad y la capacidad para tomar e implementar una decisión. Se necesita conocimiento (y autoridad) para producir información procesable que pueda generar impacto.

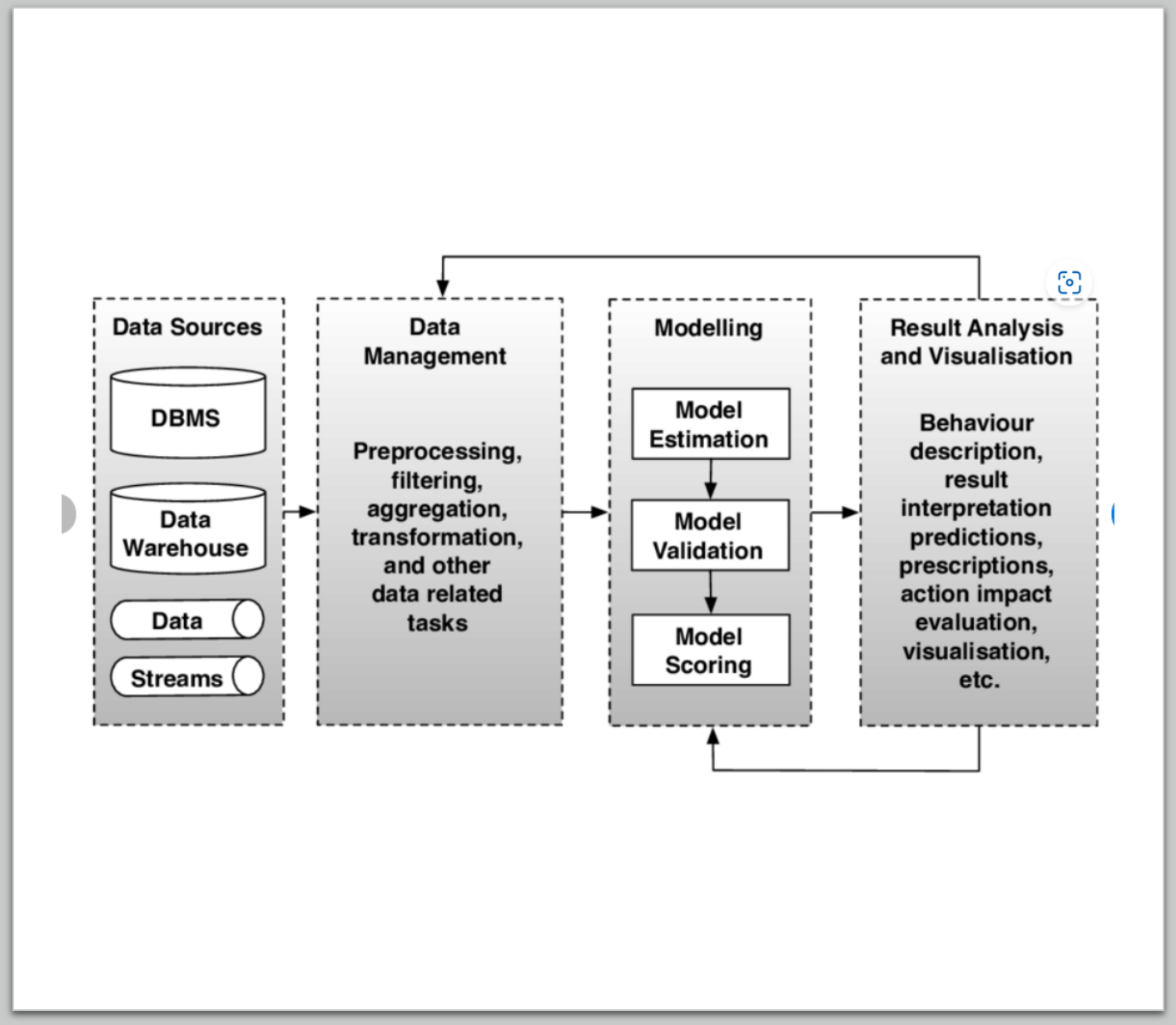
Es un proceso iterativo



What are data, information, and knowledge? (internetofwater.org)

Evaluación de la calidad

- Los datos, provenientes de diversas fuentes, se almacenan en almacenes de datos.
- ¿Ejemplos de almacenes de datos?
- La calidad de los datos es obligada para obtener resultados correctos.
- La limpieza de datos (data cleaning) es una actividad previa al workflow de analítica de datos.
- En la primera actividad limpiarán un dataset en formato csv con bastantes errores.



Algunas dimensiones para evaluar la integridad de los datos

- Completitud
- Credibilidad
- Precisión
- Consistencia
- Interpretabilidad

Dimensión Completitud

- ¿Qué tan completo es un “registro”? Esquemas estrictos versus esquemas flexibles.
- Un registro puede considerarse completo aunque falten valores para algunos de sus atributos. Todo depende de la aplicación.
- Cobertura de los datos. La porción de datos de la realidad que se encuentra contenida en el data set.
- Densidad de los datos. Cantidad de información contenida y faltante acerca de las entidades.

Dimensión completitud(cont.)

¿Qué tan completo es un “registro”? Esquemas estrictos versus esquemas flexibles.

Un registro puede considerarse completo aunque falten valores para algunos de sus atributos. Todo depende de la aplicación.

Adelantemos un ejemplo de esquema flexible con “registros” con campos faltantes. Es un adelanto de mongo

Dimensión completitud(cont.)

```
video> show collections
movieDetails
moviesScratch
video> fields = Object.keys(db.movieDetails.findOne())
[
  '_id',          'title',
  'year',         'rated',
  'runtime',      'countries',
  'genres',       'director',
  'writers',      'actors',
  'plot',         'poster',
  'imdb',         'tomato',
  'metacritic',   'awards',
  'type'
]
video> db.movieDetails.countDocuments()
2296
```

Cantidad de documentos en la colección (tabla)

```
video> db.movieDetails.find({tomato:{$exists:1}}).count()
376
```

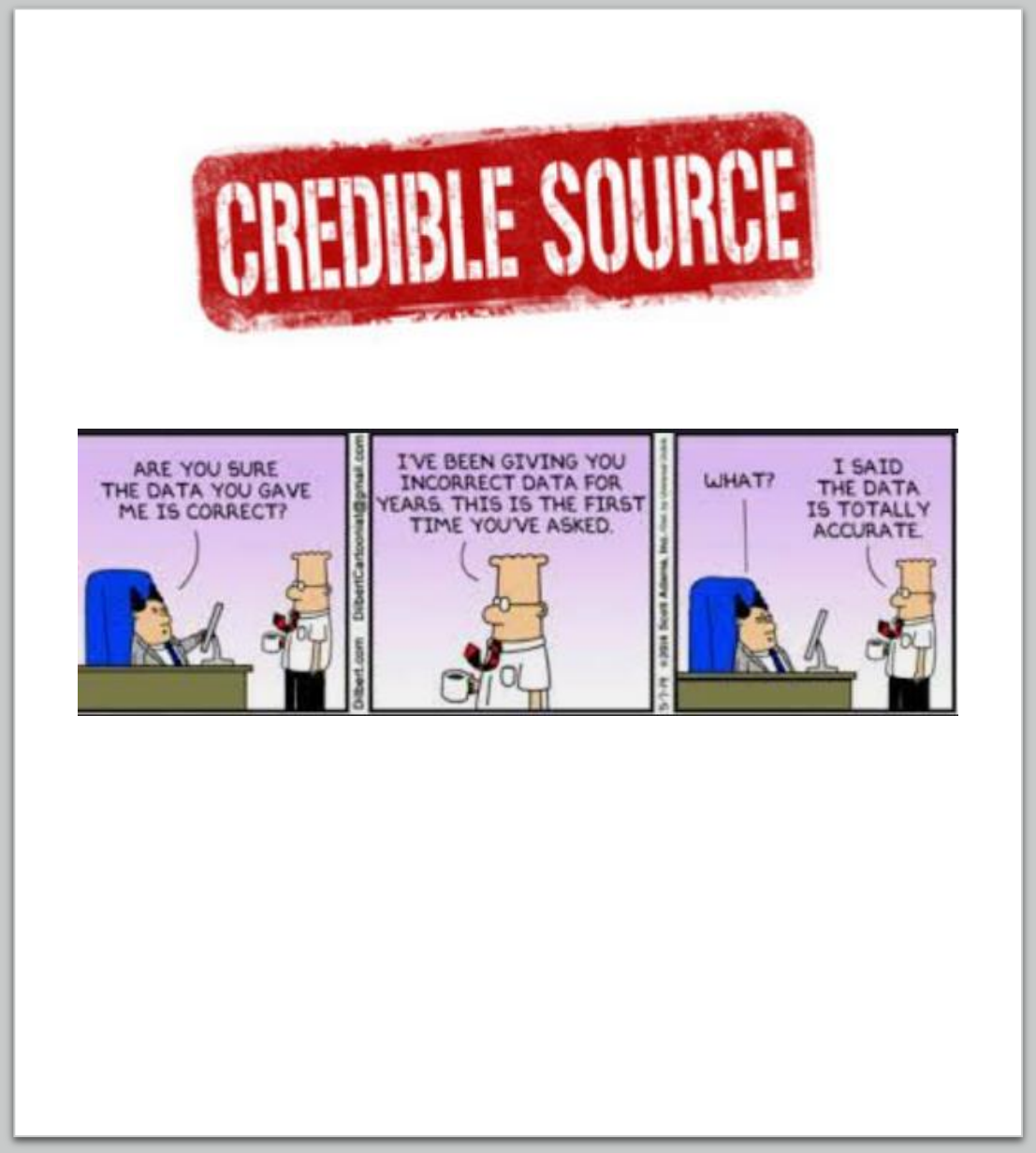
No todos los documentos o "records" tienen el campo tomato

There are 132302 **open data** datasets available on data.world.

Find open data about open data contributed by thousands of users and organizations across the world.

Credibilidad

- La credibilidad representa la fiabilidad que se le brinda al organismo que proporciona el conjunto de datos.
- Esta métrica puede reflejarse en el conjunto de datos por aquellas características que presenten un valor por defecto.



THE CREDIBILITY CRISIS

2022 February Wallace G. Smith

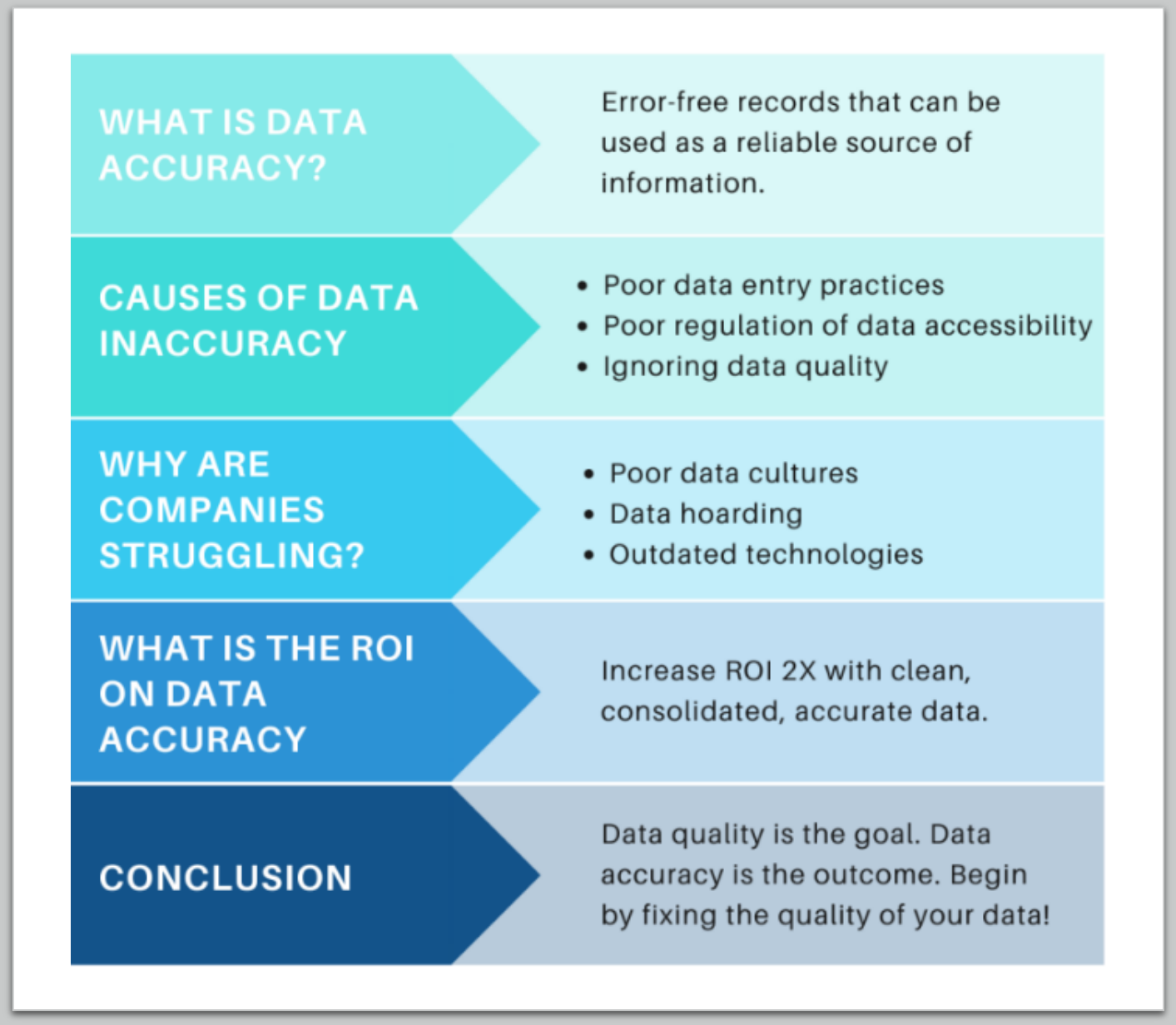
Comment on this article

00:00 00:00

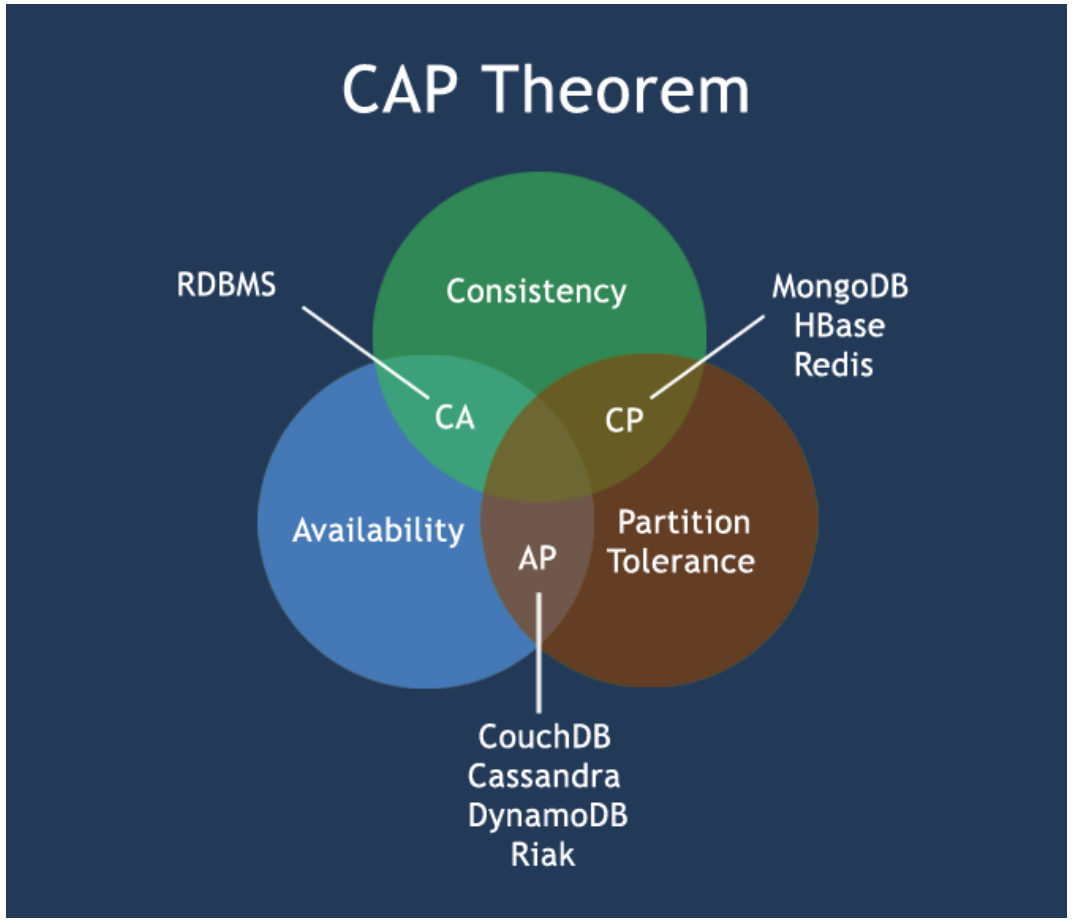
Download Audio MP3

Precisión (Accuracy)

- Datos correctos versus datos disponibles.
- ¿Qué tanto de esos datos son correctos?
- ¿Cómo lo podemos determinar?
- ¿De qué fuentes provienen?



Consistencia



[CAP Twelve Years Later: How the "Rules" Have Changed (infoq.com)]

Fuentes de datos

Fuentes de datos

Captura manual de datos

Procesamiento de documentos

Salida de aplicaciones

Sensores

Datos públicos

Organización de los datos

- Archivos planos
  - Aquellos que pueden editarse con un editor de texto. Legibles por los humanos. Requieren de estándares de conjuntos de caracteres para su almacenamiento.
  - Formatos típicos de archivos planos: CSV, JSON y XML
- Archivos binarios.
  - Se requiere especificar la longitud de los tipos de datos para su almacenamiento, así como estándares.
- Bases de Datos
  - Diversos modelos de datos: En el curso estudiamos modelos de datos no tabulares.

Todo es binario en un archivo

```java
import java.io.DataOutputStream;
import java.io.FileOutputStream;
public class unir {
    public static void main(String[] args) throws Exception {
        FileOutputStream fout = new FileOutputStream ( name: "data.txt");
        DataOutputStream dout = new DataOutputStream(fout);
        dout.writeChar( v: 'A');
        dout.writeByte( v: 5);
        dout.writeInt( v: 5);
        dout.writeLong( v: 10);
        String str = "Hello";
        dout.writeChars(str);
        dout.close();
    }
}
```

[The Unicode Standard, Version 14.0]

data.txt - Hex

```
0:  00 41 05 00 00 00 05 00  00 00 00 00 00 00 0A 00  .A..............
10: 48 00 65 00 6C 00 6C 00  6F                       H.e.l.l.o
```

Formatos típicos de archivos planos

- XML Tutorial (tutorialspoint.com)
- JSON Tutorial (tutorialspoint.com)
- Working With JSON Data in Python – Real Python
- Python - Processing CSV Data (tutorialspoint.com)