Fundamentos de Análisis de Datos

Unidad 2

Naturaleza, origen y estructura de los datos

Agenda

- Datos, información y conocimiento
- Tipos de datos
- Tipos de variables
- Ciclo de vida de los datos
- Calidad y gestión de los datos.

Sociedad del Conocimiento y Ciencia de Datos

¿Qué es la sociedad del conocimiento?

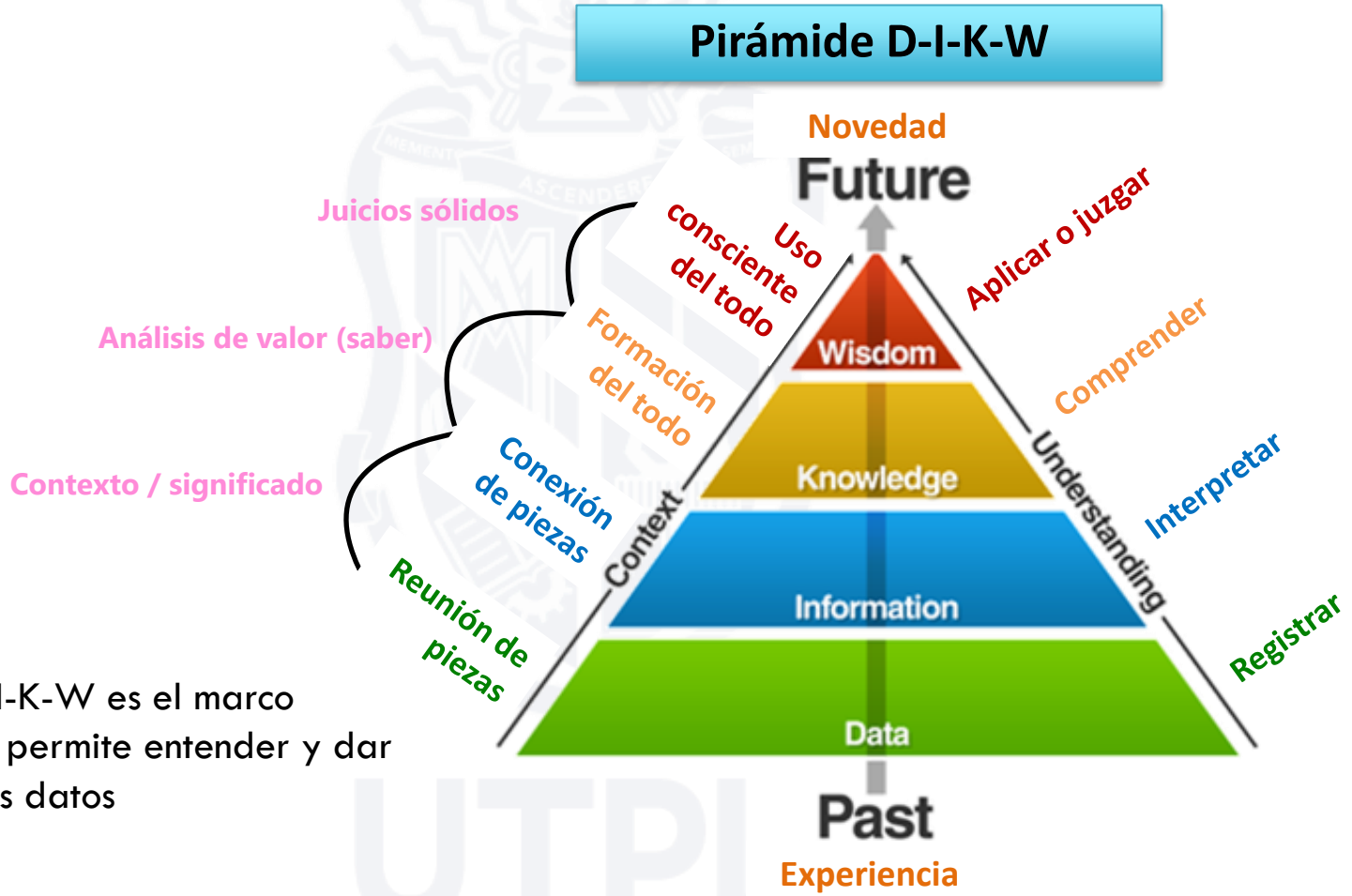
Es una etapa del desarrollo social y económico donde **el conocimiento es el principal motor de crecimiento, innovación y bienestar**.
Se caracteriza por la **producción, distribución y uso intensivo del conocimiento**, más allá del simple acceso a la información.

Rol de la ciencia de datos en la sociedad del conocimiento:

- **Transforma datos en decisiones**: convierte grandes volúmenes de datos en **conocimiento útil y accionable**.
- **Impulsa la innovación**: permite descubrir patrones, optimizar procesos y predecir comportamientos.
- **Conecta saberes diversos**: combina estadística, computación, ética y dominio de negocio.
- **Genera impacto social y económico**: desde la salud pública hasta la educación, el transporte o el medio ambiente.

"En la sociedad del conocimiento, los datos son el nuevo recurso; pero es el **conocimiento generado con ciencia de datos** lo que crea verdadero valor."

Pirámide del conocimiento

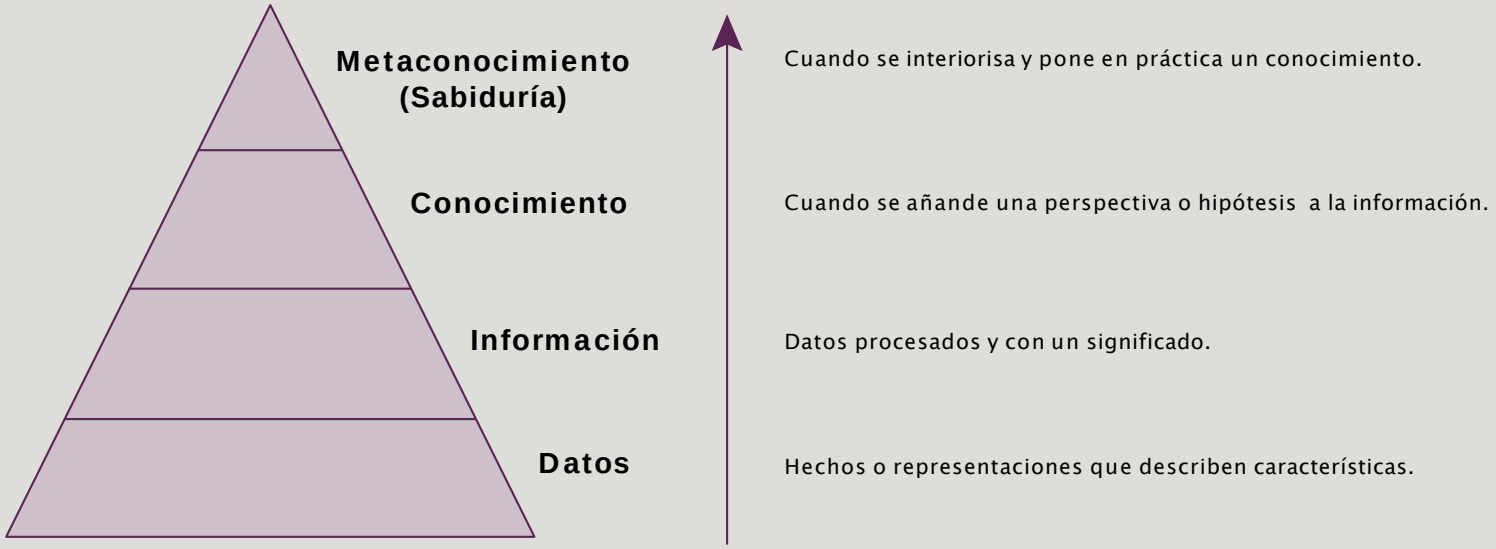


La pirámide D-I-K-W es el marco conceptual que permite entender y dar significado a los datos

Fuente: http://vishalkumarg325.blogspot.com.es/2013/03/dikw-pyramid-theory.html

Pirámide del conocimiento

Datos vs Información vs Conocimiento vs Sabiduría



Pirámide del conocimiento

- **Dato**
  - Mínima unidad semántica
  - Por si solos solo describen algo pero carecen de relevancia para toma de decisiones
  - Elementos puros, sin contexto ni interpretación. Son registros, cifras, hechos aislados
- **Información**
  - Datos procesados, organizados o contextualizados que permiten responder preguntas básicas y ya tienen alguna relevancia para toma de decisiones
  - Son datos con valor añadido: contextualizados, categorizados, calculados, corregidos, condensados (resumidos, agregados, agrupados)

Pirámide del conocimiento

- **Conocimiento:**
  - Comprensión profunda de patrones, causas y consecuencias. Permite analizar e interpretar información, y tomar de decisiones
  - Se deriva del análisis de la información, cuyas acciones pueden ser:
    - Comparación con otros elementos
    - Predicción de consecuencias
    - Identificación de conexiones
- **Sabiduría**
  - Inteligencia generada a partir del aprovechamiento del conocimiento para tomar la mejores decisiones
  - Capacidad de aplicar el conocimiento con juicio, ética y experiencia para tomar decisiones acertadas

139

139 miles de dólares

Ventas: 139 miles de dólares

Ventas: 139 miles de dólares

Ámbito: Internacional

Fecha: Agosto 2009

Ventas: 139 miles de dólares

Ámbito: Internacional

Fecha: Agosto 2009

Ventas: 2634 miles de dólares

Ámbito: Doméstico

Fecha: Agosto 2009

Ventas: 599 miles de dólares

Ámbito: Internacional

Fecha: Septiembre 2009

…

Ventas: 139 miles de dólares

Ámbito: Internacional

Fecha: Agosto 2009

Ventas: 2634 miles de dólares

Ámbito: Doméstico

Fecha: Agosto 2009

Ventas: 599 miles de dólares

Ámbito: Internacional

Fecha: Septiembre 2009

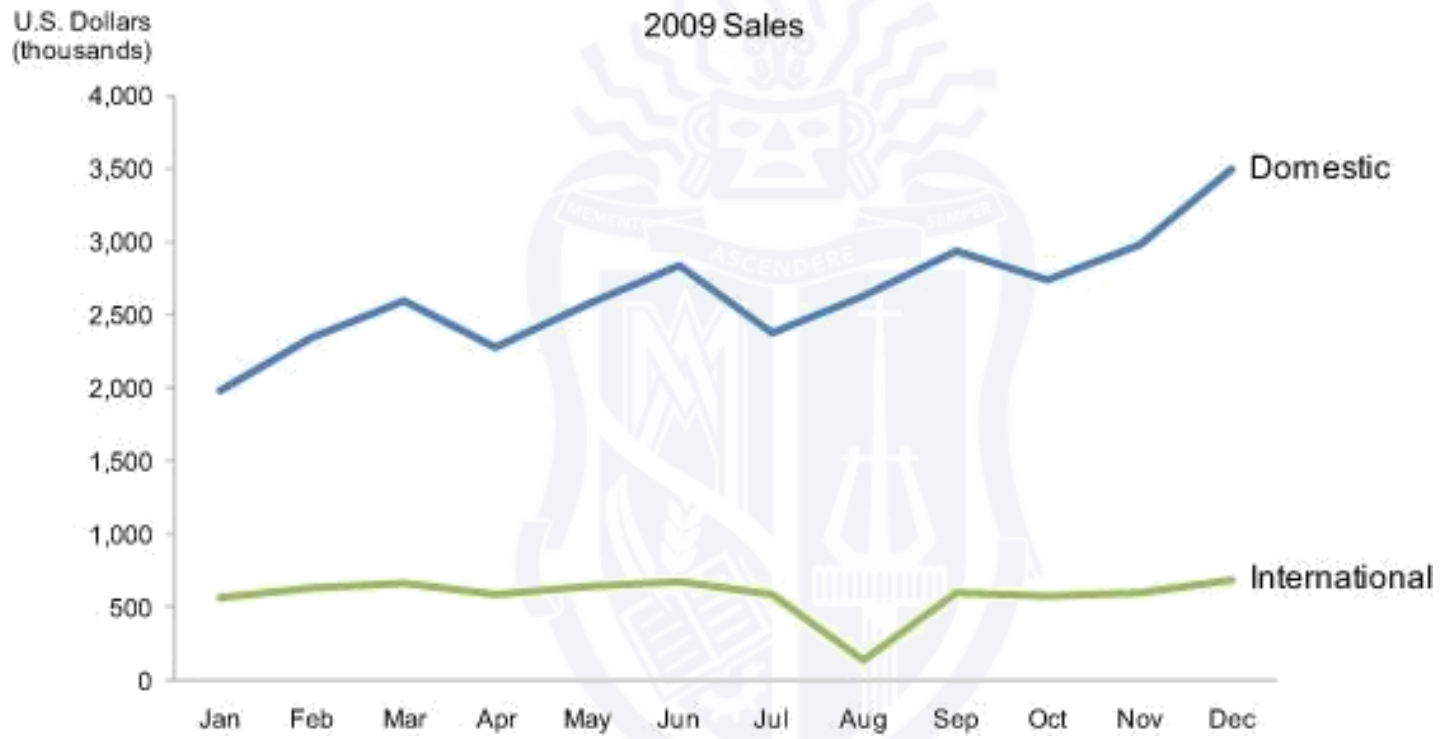
…

A nivel internacional en septiembre de 2009 las ventas aumentaron en 330 %

2009 Sales (thousands of U.S. $)													
Region	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
Domestic	1,983	2,343	2,593	2,283	2,574	2,838	2,382	2,634	2,938	2,739	2,983	3,493	31,783
International	574	636	673	593	644	679	593	139	599	583	602	690	7,005
Total	2,557	2,979	3,266	2,876	3,218	3,517	2,975	2,773	3,537	3,322	3,585	4,183	38,788

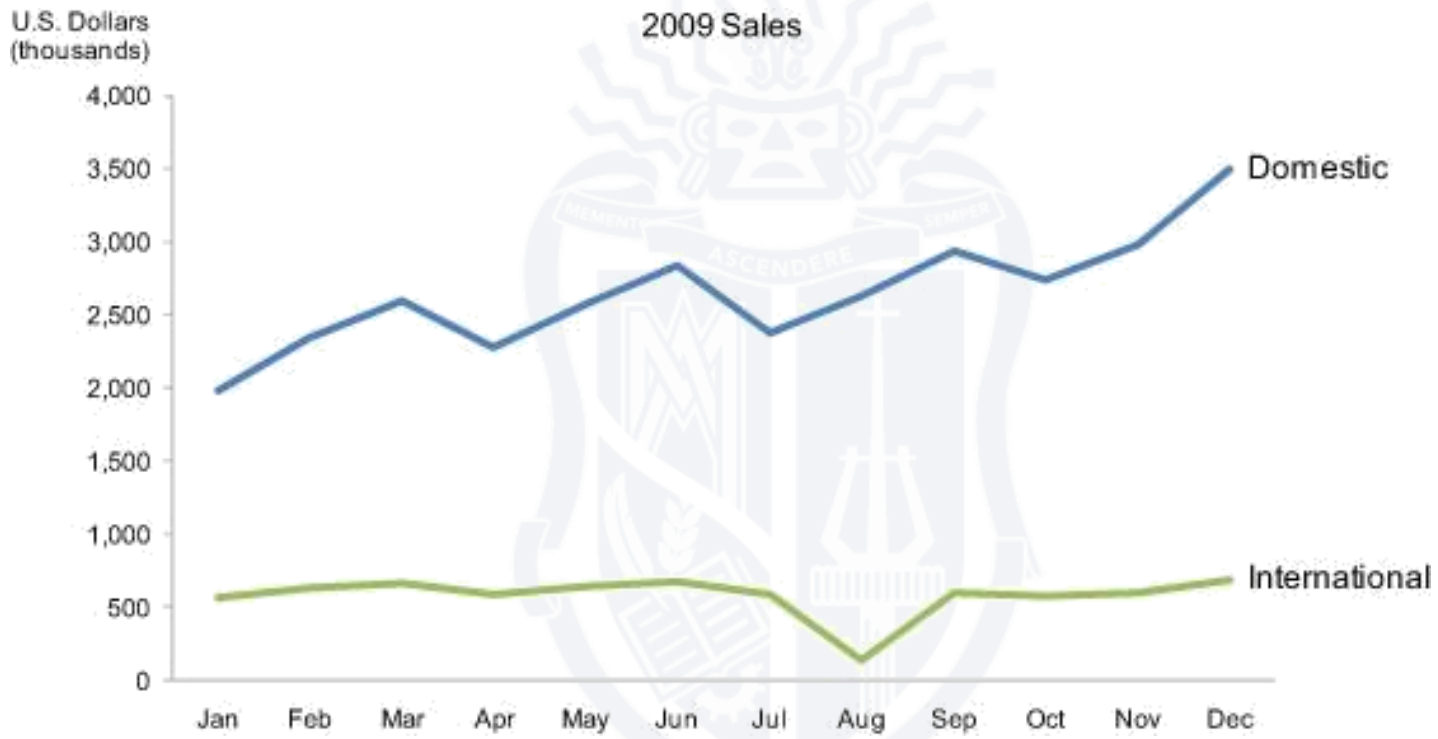
Few, S. Data Visualization for Human Perception

Tomado de UNIR (Yusef Hassan Montero)



Few, S. Data Visualization for Human Perception

Tomado de UNIR (Yusef Hassan Montero)



Lo que en realidad ha ocurrido es que a nivel internacional, en agosto de 2009, las ventas cayeron un 78% respecto al promedio de los meses anteriores

En agosto de 2009, las ventas internacionales cayeron un 78% debido a la interrupción del canal logístico marítimo desde el puerto principal de exportación, causada por una huelga portuaria registrada entre el 3 y el 25 de agosto. El cruce de datos operacionales, registros aduaneros y tiempos de despacho mostró que el 82 % de los pedidos internacionales programados para ese mes no salieron del país a tiempo. Además, el análisis de correlación y series temporales detectó que las rutas de distribución afectadas representaban históricamente el 65 % del volumen exportado. Esta disrupción logística, no advertida a tiempo por el área comercial, generó cancelaciones masivas y aplazamientos que impactaron directamente en los ingresos del mes

A partir del conocimiento adquirido sobre el impacto de disrupciones logísticas en las ventas, se implementó un sistema de alertas tempranas que, mediante modelos de predicción de riesgo operativo y aprendizaje automático, detecta eventos como huelgas, bloqueos o anomalías en tiempos de despacho. El sistema prescribe acciones automáticas como: diversificar rutas de exportación, activar almacenes temporales cercanos a otros puertos secundarios, y notificar a clientes críticos sobre retrasos previstos para prevenir cancelaciones. Además, se definió una política de contingencia comercial para ajustar precios y promociones en mercados menos afectados

Ejemplos D-I-K-W

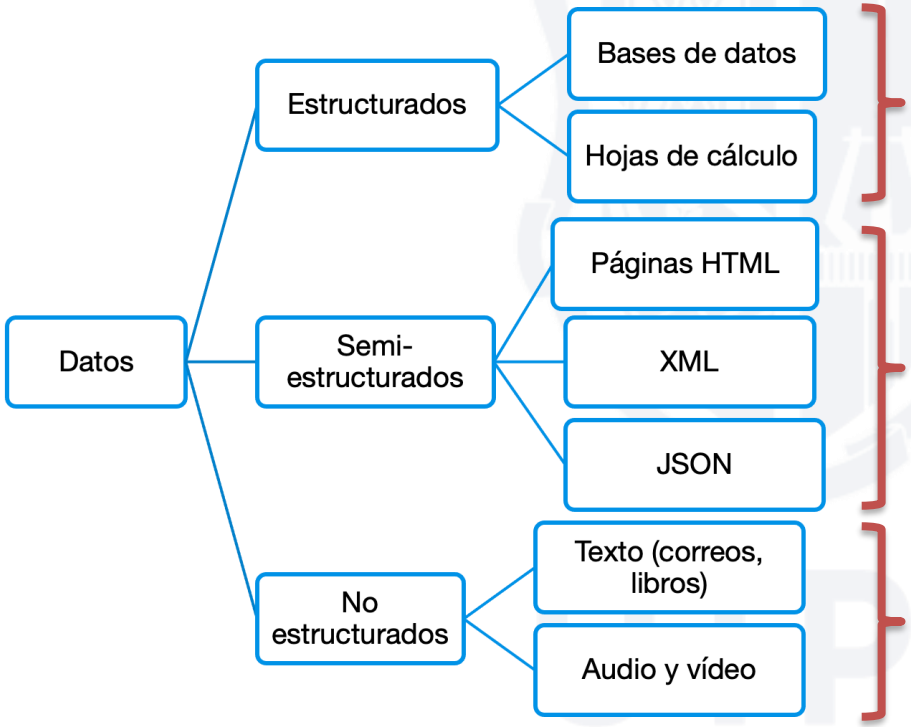
Nivel	Ejemplo 1 *Rendimiento académico*	Ejemplo 2 *Contaminación ambiental*
Datos	Calificaciones de 1000 estudiantes en diferentes materias del semestre. (*Datos sin procesar*)	Mediciones diarias de dióxido de nitrógeno (NO_2) en distintas estaciones de monitoreo
Información	El 35% de los estudiantes reprobó Matemáticas. (*Datos resumidos y organizados*)	En ciertas zonas los niveles de NO_2 superan los límites recomendados por la OMS
Conocimiento	Los estudiantes con más ausencias tienen mayor probabilidad de reprobar. (*Se identifican patrones y relaciones*)	El aumento coincide con horarios pico de tráfico vehicular y bajas temperaturas
Sabiduría	Se recomienda implementar tutorías obligatorias para estudiantes con bajo rendimiento y más de 3 faltas. (*Acción estratégica basada en el conocimiento*)	Se sugiere restringir la circulación de vehículos en horas punta y promover el transporte público limpio

Fuentes de datos

- Múltiples
- Heterogéneas
- Internas y externas
- Estructuradas, semiestructuradas, no estructuradas
- Datos abiertos, datos propietarios

Tipos de datos según su estructura

Nivel de estructura: Forma en que se organizan los datos para facilitar su procesamiento por un computador.

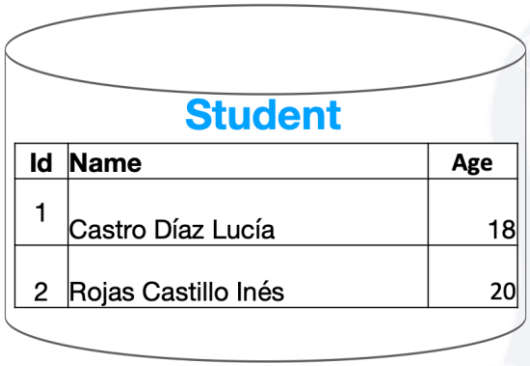


Tienen una estructura predefinida que permite asociar valores a atributos que describen un individuo, cosa o evento

Tienen una estructura flexible, no estricta. Usan marcadores o etiquetas no rígidas.

No tienen una estructura definida. Requieren técnicas avanzadas de procesamiento de lenguaje natural, procesamiento de imágenes, etc.

Datos estructurados



Student

Id	Name	Age
1	Castro Díaz Lucía	18
2	Rojas Castillo Inés	20

Semana	Fecha inicio	Fecha fin
1	11-abr-2023	14-abr-2023
2	17-abr-2023	21-abr-2023

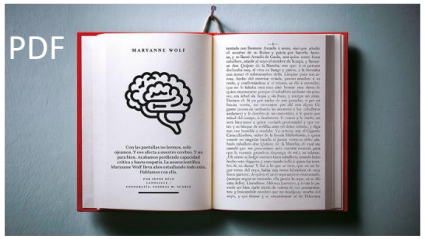
- Nivel más alto de estructura: procesamiento eficiente y sencillo
- Los datos se almacenan con una estructura bien definida
- Bases de datos: la información se almacena en tablas las cuales se componen de filas (tuplas) y columnas (campos o atributos).
- Ejemplos: Bases de Datos Relacionales, Hojas de cálculo, ERP, CRM

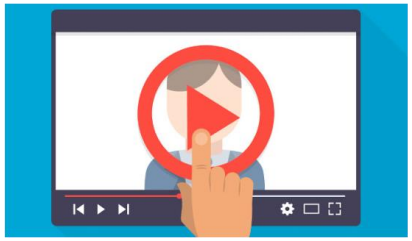
Datos semi-estructurados

Ejemplo de archivo JSON

- Los datos se almacenan conforme a conjunto de reglas menos estrictas y más flexibles.
- El nivel de estructura puede variar según su aplicación y, por tanto, también la dificultad de procesamiento.
- Algunos de los formatos semi-estructurados más usados: HTML, XML, JSON
- Ejemplos: Páginas web, Publicaciones twitter y Facebook obtenidas mediante APIs o crawling, Emails (metadatos), Base de datos documentales (incluyen metadatos y contenido), Logs de sistemas

Datos no estructurados





Ejemplos

- No tienen una estructura definida de forma explícita.
- Pueden tener algún tipo de estructura implícita: párrafos de un texto, escenas de una película.
- Para un computador puede llegar a ser muy difícil de interpretar.
- Es el tipo de dato más abundante en la Web.
- Ejemplos: Documentos texto, emails, documentos ofimática, PDFs, Imágenes, Audios, Videos

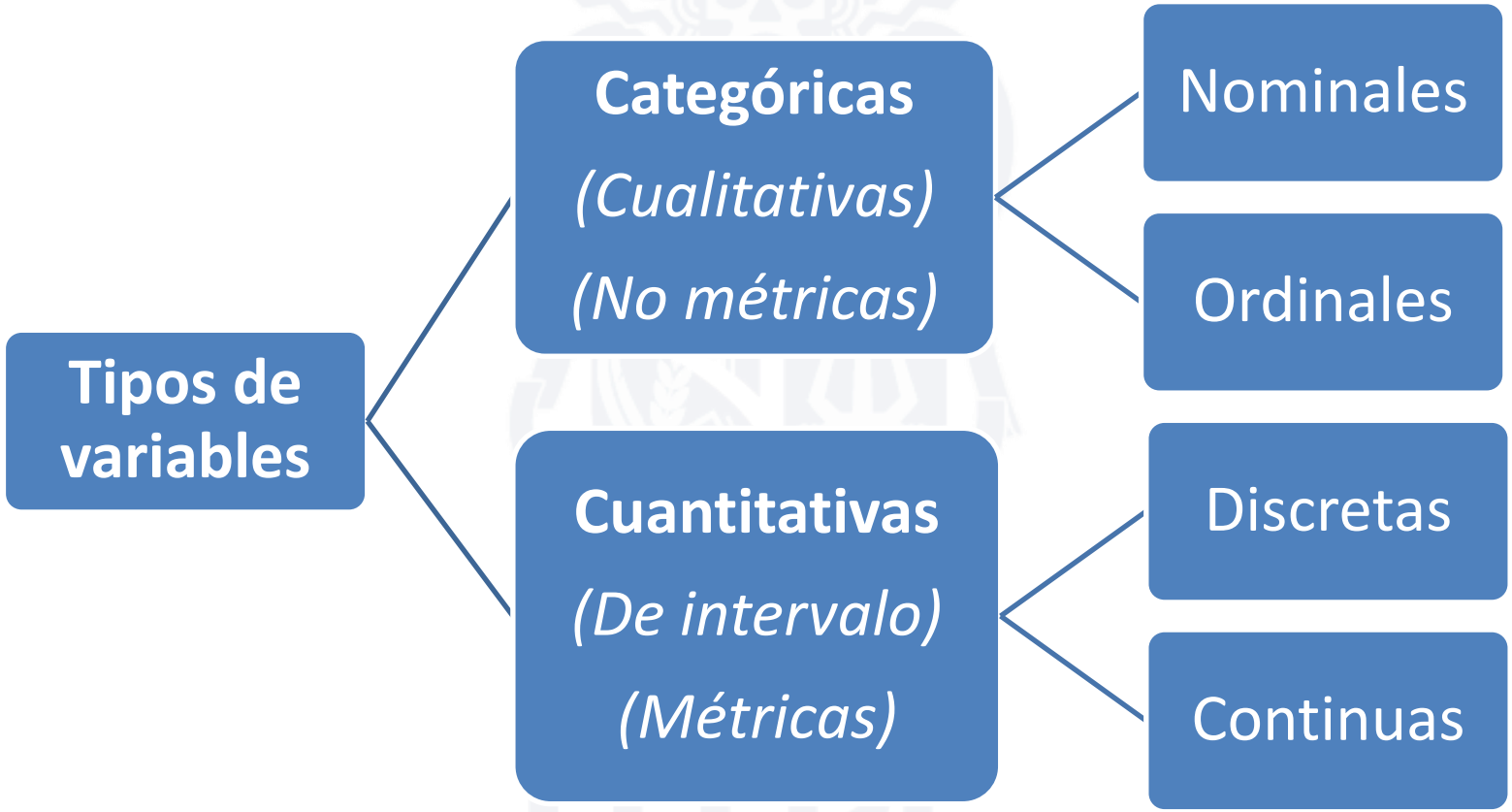
Datos en el contexto estadístico

- Variable
  - Característica de un objeto, individuo o caso
  - Tiende a variar según el caso
- Valor
  - Conjunto de números y/o símbolos que determinan el estado de una variable
- Medición o métrica
  - Asociación de un valor a una característica de un objeto, según reglas preestablecidas

Datos en el contexto estadístico

- Escala o dominio:
  - Conjunto de valores de puede tomar una variable
  - Ejemplo:
    - Precio de un artículo: números decimales mayores a cero
    - Genero de una persona: “Masculino”, “Femenino”
    - Tasa de interés: números de 0 a 100
- Definir y medir bien las variables es fundamental para el éxito de un análisis de datos

Tipos de variables



Tipos de variables

- Variables categóricas
  - También llamadas “Cualitativas” o “No métricas”
  - Los valores son etiquetas que determinan la categoría de cada individuo
  - Las categorías son mutuamente excluyentes
  - Los valores son diferentes por cualidad, no por cantidad
  - Dos subtipos
    - **Nominales**
    - **Ordinales**

Tipos de variables

- Variables nominales
  - Categorías que representan una cualidad no jerárquica
  - Ejemplos: tipo de animal, color de vehículo
- Variables ordinales
  - Los valores o categorías determinan orden o jerarquía
  - Ejemplos: rango militar, posición en un campeonato

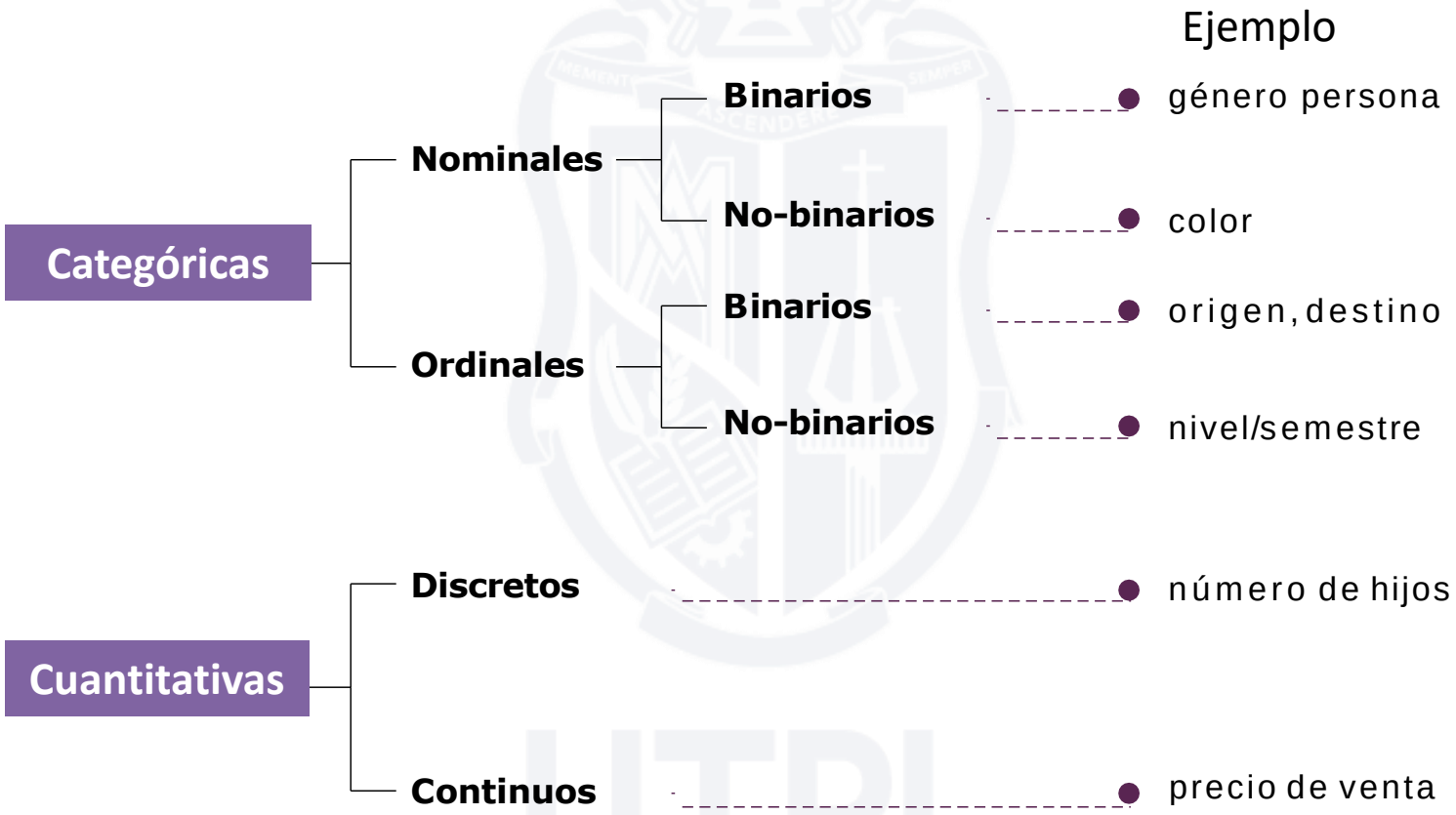
Tipos de variables

- Variables cuantitativas
  - También llamadas “De intervalo” o “Métricas”
  - Los valores son números cuyas diferencias tienen sentido
  - Los valores son diferentes por la cantidad
  - Se pueden establecer jerarquías
  - Dos subtipos
    - **Discretas**
    - **Continuas**

Tipos de variables

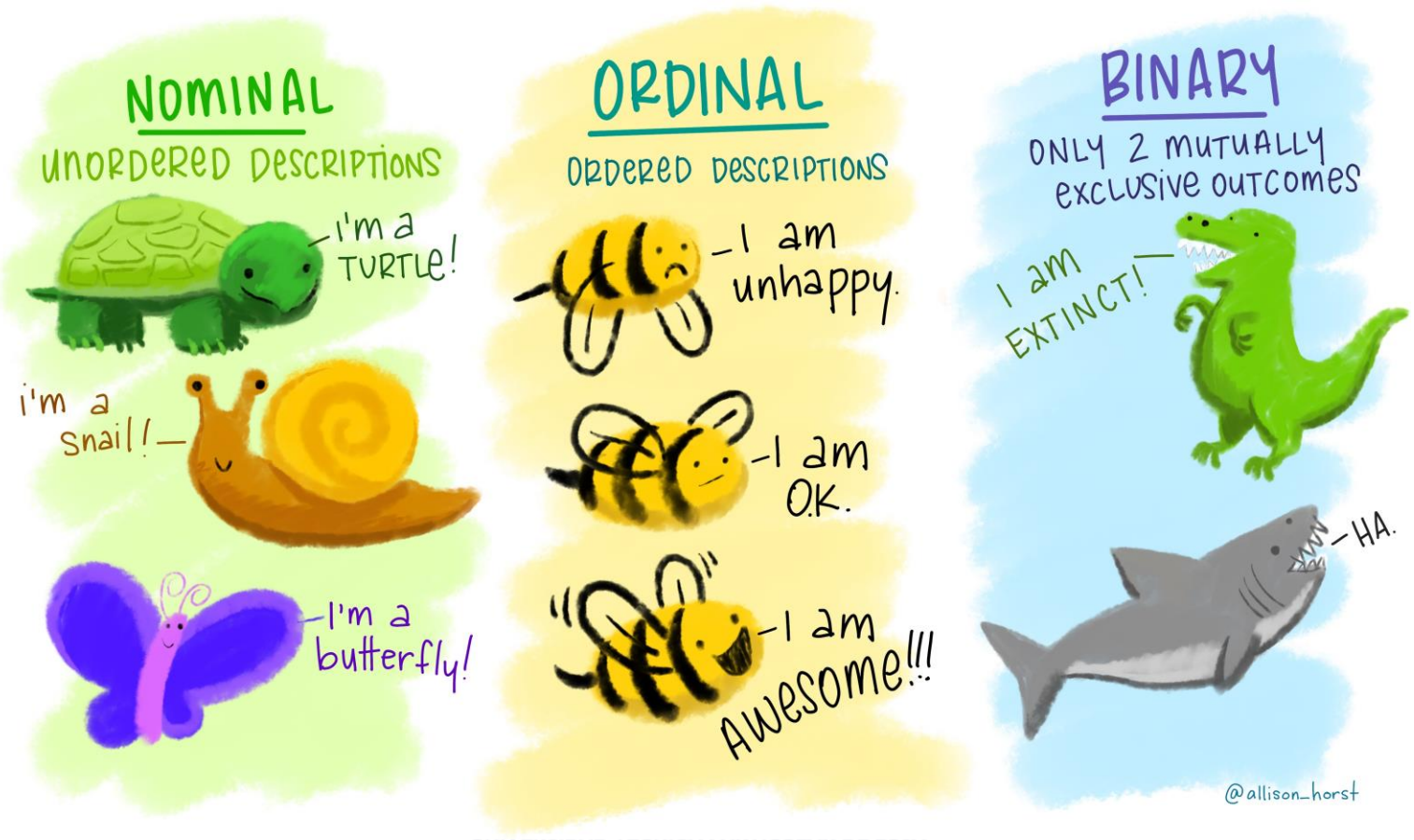
- Variables discretas
  - Valores contados, no medidos
  - Solo números enteros (no fraccionarios o decimales)
  - El conteo se basa en “saltos” fijos. Ej: de uno en uno
  - Ejemplos: goles anotados por un equipo, número de hijos de una persona, número de empleados
- Variables continuas
  - Valores medidos
  - Cualquier valor numérico, incluidos fraccionarios o decimales
  - Pueden llegar a ser muy precisos
  - Puede tomar cualquier valor en un intervalo.
  - Ejemplos: temperatura en ^{o}C, superficie de un país

Tipos de variables



Tipos de variables

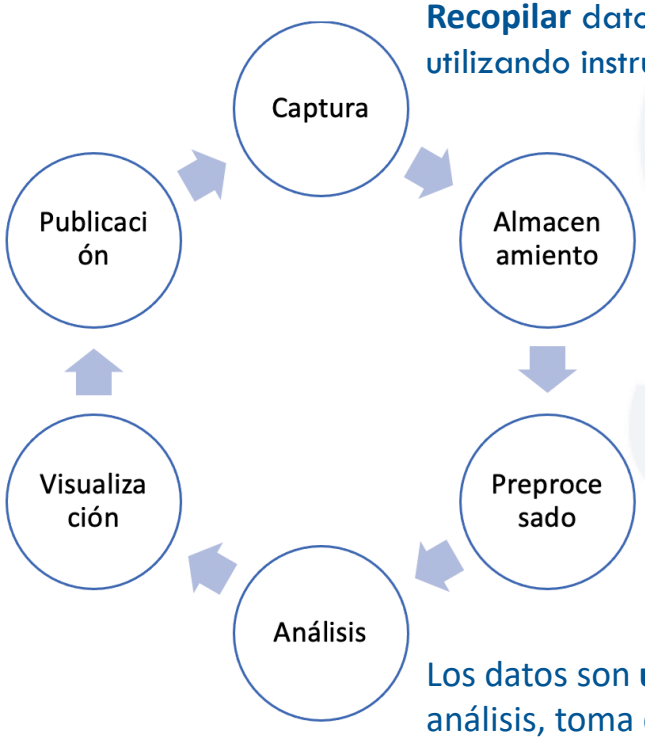
Tipos de datos (Categóricos)



Tipos de variables

Tipos de datos (Cuantitativos)

Ciclo de vida de los datos



Recopilar datos que se generen durante un proceso (repositorio, API, sraping, utilizando instrumentos de recolección de datos)

Almacenar datos en un formato que permitan su posterior manipulación (ficheros, bases de datos) y actualización (guardar metadatos de origen).

Durante esta etapa, es importante realizar tareas de **mantenimiento** que incluyen la validación, limpieza, clasificación y actualización de datos para asegurar su calidad y precisión.

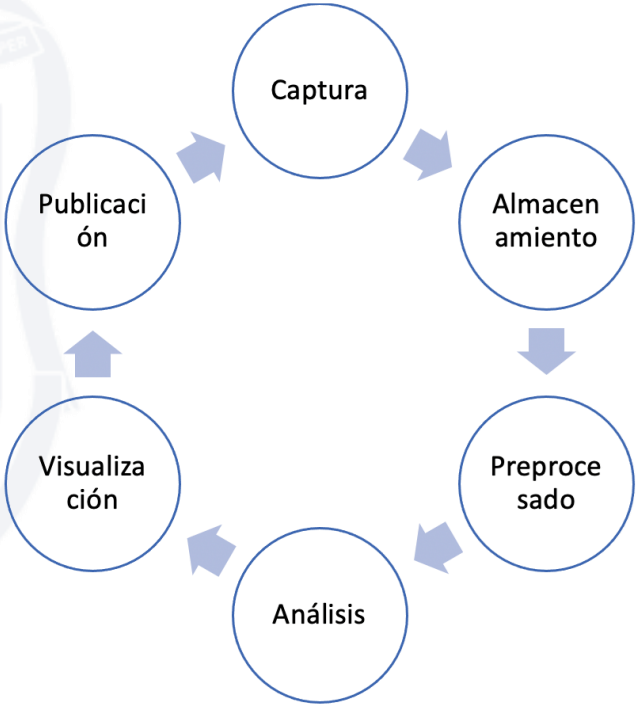
Los datos son **usados** para los fines para los cuales fueron recolectados, como análisis, toma de decisiones, y operaciones del día a día. Esta fase puede involucrar la manipulación de los datos mediante consultas, reportes, y otros métodos de análisis como estadísticas y machine learning.

Source: Fundamentos de data science (Julià Minguillón, UOC)

Ciclo de vida de los datos

Los datos pueden necesitar ser **compartidos** con otras partes, ya sean internas (entre departamentos) o externas (con socios comerciales, reguladores, etc.). La compartición debe cumplir con políticas de seguridad de datos y regulaciones de privacidad para asegurar que solo los usuarios autorizados tengan acceso a los datos.

La visualización es como la interfaz de **navegación** de los datos: para ver, ampliar, filtrar, condensar, comparar, exporar datos seleccionados.



Source: Fundamentos de data science (Julià Minguillón, UOC)

Ciclo de vida de los datos

- Dependiendo de la característica y vigencia del dato, éste, luego de la publicación podría pasar a fases de archivo y eliminación:
  - **Archivo**: preservar datos ya usados, por temas legales
  - **Eliminación**: eliminar datos que ya no son requeridos ni administrativa, ni legalmente

Calidad de los datos

No existe un buen sistema de inteligencia de negocio sin datos de calidad que se gestionen adecuadamente y tengan sentido para el negocio (Rodriguez, s.f.)

El mayor fracaso de un proyecto de DS es trabajar con información de baja calidad, pobre o que no tiene significado para los ejecutivos y mandos intermedios (Rodriguez, s.f.)

La gestión de datos garantiza que la compañía dispone de la información correcta y que es utilizada de forma apropiada (Davenport y Harris, 2007)

¿Dónde podemos obtener datos de calidad?

Rodriguez (s.f.)

Calidad de los datos

Realizar consulta para responder a preguntas clave sobre calidad de datos:

1. ¿Por qué es importante la calidad de los datos?

2. ¿Cuáles son los problemas de calidad de datos más comunes?

3. ¿Cuáles son los principales atributos (métricas) de los datos de calidad?

4. ¿Cuáles son los métodos que podemos usar parra mejorar/corregir la calidad de los datos?

5. ¿Qué es Data Wrangling?

6. ¿Qué herramientas/funciones/paquetes R podemos usar para gestionar la calidad de los datos?

7. ¿Cómo gestionar la calidad de datos en la organización?