

Big Data

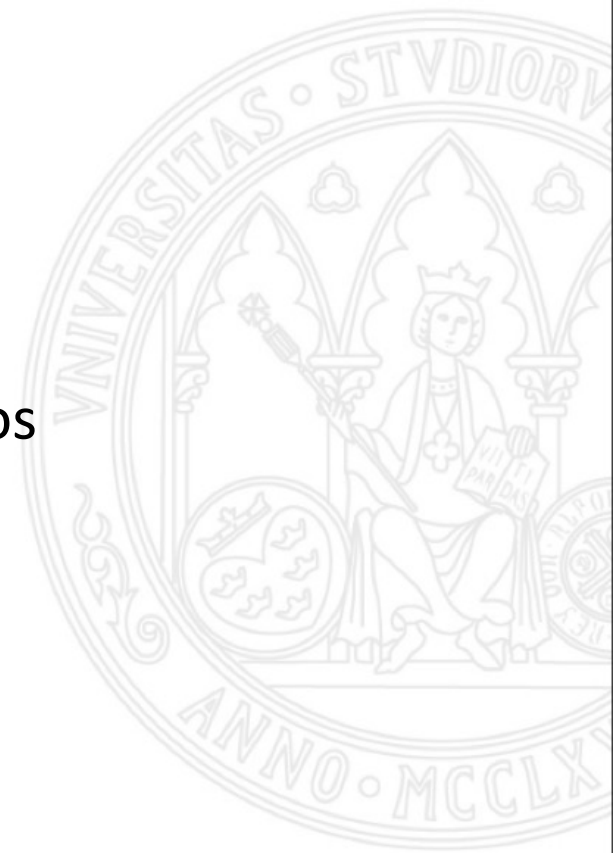


Unidad 1: Introducción a Big Data y Minería de Datos

UNIVERSIDAD DE
MURCIA

Unidad 1. Introducción a Big Data y Minería de Datos

BIG DATA



- Big Data hace referencia a conjuntos de datos cuyo **tamaño**, **velocidad** de generación y **diversidad** superan lo que resulta factible gestionar con herramientas tradicionales.
- **Crecimiento continuo** de la actividad digital, la sensorización, la nube y la automatización de procesos.
- No se trata solo de almacenar muchos datos, sino de capturarlos, organizarlos, procesarlos y convertirlos en valor.
- Evolución hacia **arquitecturas** que puedan soportar el almacenamiento, procesamiento y compartición de grandes volúmenes de datos

Data Collection and Database Creation
(1960s and earlier)

- Primitive file processing

Database Management Systems
(1970s to early 1980s)

- Hierarchical and network database systems
- Relational database systems
- Data modeling: entity-relationship models, etc.
- Indexing and accessing methods
- Query languages: SQL, etc.
- User interfaces, forms, and reports
- Query processing and optimization
- Transactions, concurrency control, and recovery
- Online transaction processing (OLTP)

Foco exclusivo en
almacenamiento
y acceso a datos

Advanced Database Systems
(mid-1980s to present)

- Advanced data models: extended-relational, object relational, deductive, etc.
- Managing complex data: spatial, temporal, multimedia, sequence and structured, scientific, engineering, moving objects, etc.
- Data streams and cyber-physical data systems
- Web-based databases (XML, semantic web)
- Managing uncertain data and data cleaning
- Integration of heterogeneous sources
- Text database systems and integration with information retrieval
- Extremely large data management
- Database system tuning and adaptive systems
- Advanced queries: ranking, skyline, etc.
- Cloud computing and parallel data processing
- Issues of data privacy and security

Nuevos formatos,
múltiples fuentes
y gestión de
almacenamiento
escalable

Advanced Data Analysis
(late-1980s to present)

- Data warehouse and OLAP
- Data mining and knowledge discovery: classification, clustering, outlier analysis, association and correlation, comparative summary, discrimination analysis, pattern discovery, trend and deviation analysis, etc.
- Mining complex types of data: streams, sequence, text, spatial, temporal, multimedia, Web, networks, etc.
- Data mining applications: business, society, retail, banking, telecommunications, science and engineering, blogs, daily life, etc.
- Data mining and society: invisible data mining, privacy-preserving data mining, mining social and information networks, recommender systems, etc.

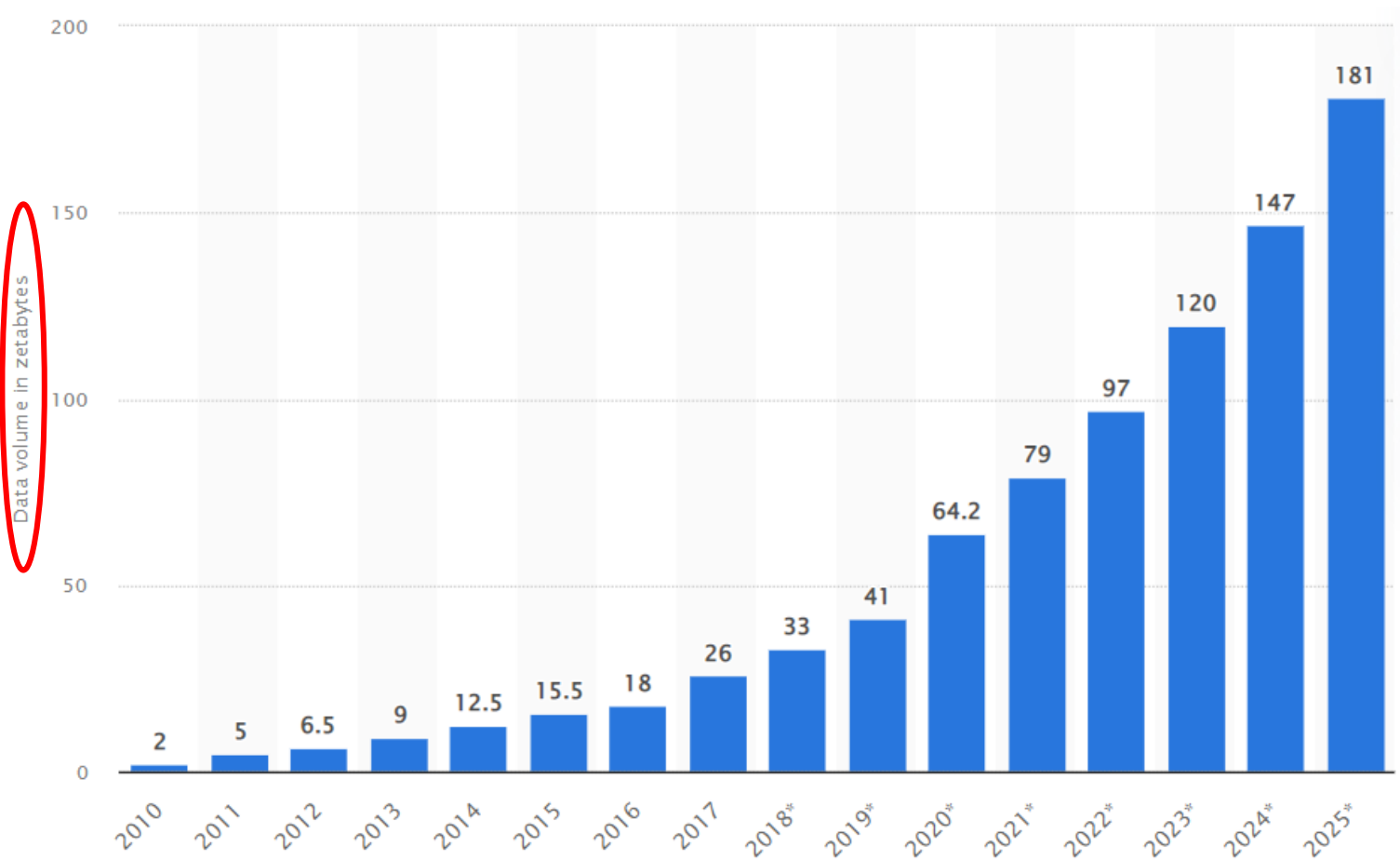
Paso de gestionar
datos a extraer
conocimiento útil

Future Generation of Information Systems
(Present to future)

Generación de datos continua, sistemas
distribuidos, inteligencia artificial, etc

¿Por qué hablamos hoy de Big Data?

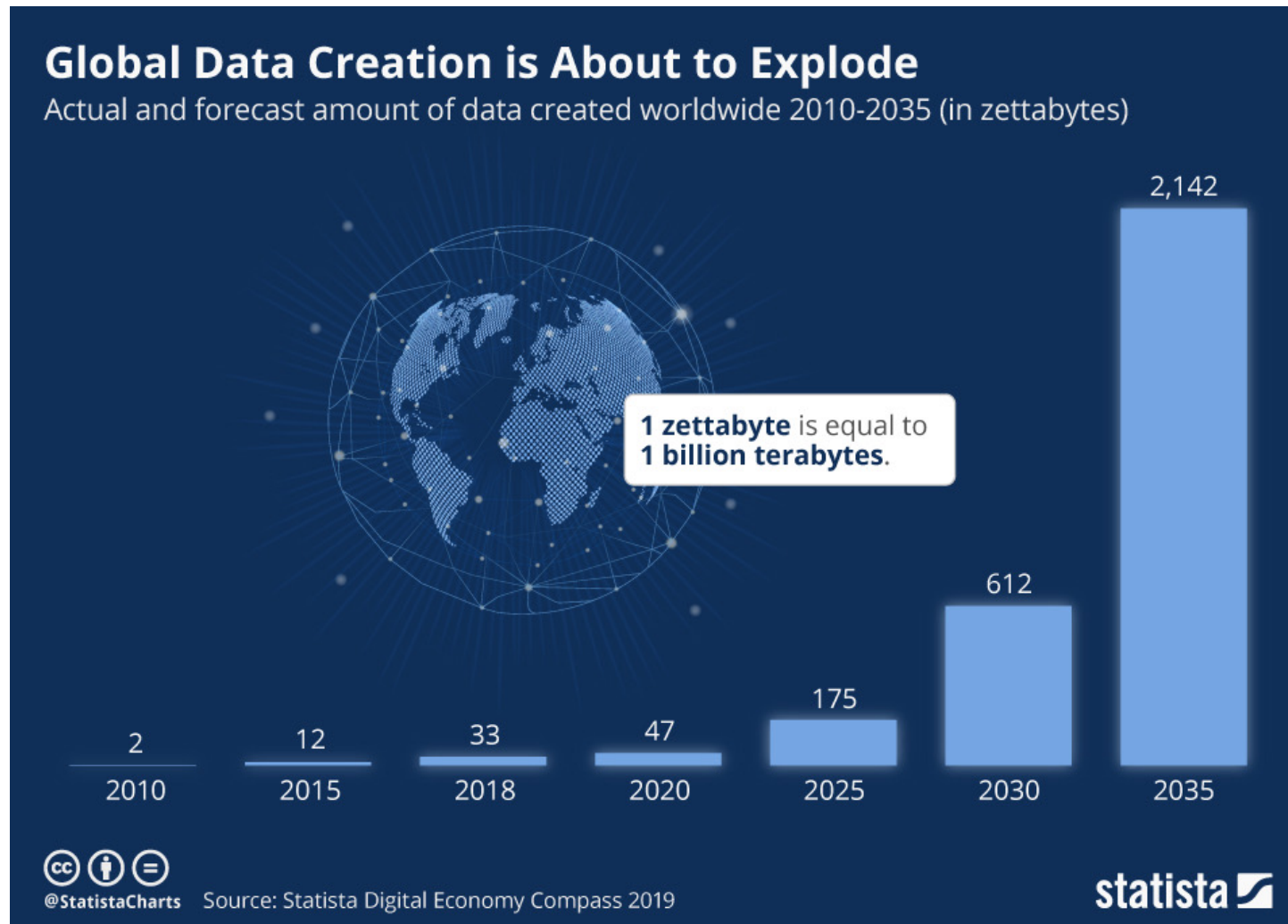
- La cantidad de datos generados a nivel mundial no para de crecer: servicios digitales, cloud, IoT, IA...
- Las bases de datos tradicionales ya no son suficiente



Fuente y gráfico <https://www.statista.com/statistics/871513/worldwide-data-created/>

¿Por qué hablamos hoy de Big Data?

- La previsión a diez años es preocupante, especialmente por el auge de la inteligencia artificial



DATA NEVER SLEEPS



Every minute of every day, the world generates a dizzying amount of data, and how we interact with it is constantly changing. AI tools are now answering millions of questions in real time and are transforming how we work, shop, and connect. Digital platforms are seeing explosive usage, with billions of emails, texts, and reels shared every day. Entertainment continues to drive engagement across streaming, gaming, and social media while e-commerce is setting new benchmarks as digital habits evolve and expand at an unprecedented pace.

In Domo's 12th edition of Data Never Sleeps, we capture a snapshot of this world powered by the rapid rise of data, AI, and digital activity, shaping every moment of modern life.

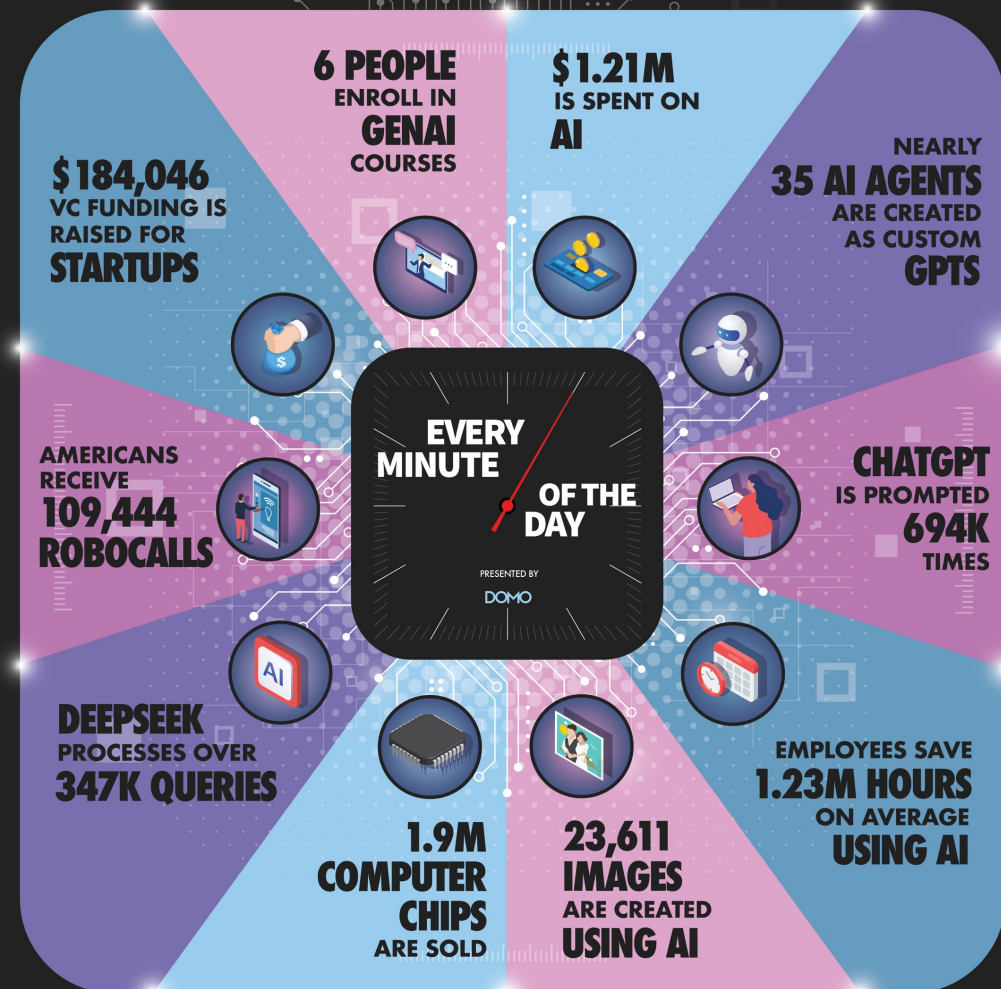
12



DATA NEVER SLEEPS AI EDITION 2025



Domo has tracked minute-by-minute digital engagement for over a decade, and AI is moving from experimental to essential at unprecedented speed. Spending on artificial intelligence has more than tripled since 2024. From boardrooms to living rooms, this special second Data Never Sleeps AI edition reveals how AI is reshaping the global economy, 60 seconds at a time.



Sources: Coolest Gadgets, Coursera, Cropink, Entrepreneur, Everypixel, Gold IRA Companies, OpenAI, Precedence Research, SIA, Trading Economics

UNIDAD	ÓRDENES DE MAGNITUD
bit (b)	0 o 1, sí o no
byte (B) = 8 bits	Una letra: a, b, c, d, ...
kilobyte (kB) = 10^3 bytes	
megabyte (MB) = 10^6 bytes $\cong$ 1 millón bytes	5 MB = una canción (formato mp3)
gigabyte (GB) = 10^9 bytes $\cong$ 1 mil millones bytes	256 GB = almacenamiento smartphone
terabyte (TB) = 10^{12} bytes $\cong$ 1 billón bytes	1 TB = almacenamiento típico portátil
petabyte (PB) = 10^{15} bytes $\cong$ 1 mil billones bytes	Grandes repositorios empresariales y científicos
exabyte (EB) = 10^{18} bytes $\cong$ 1 trillón bytes	Infraestructuras globales de gran escala
zettabyte (ZB) = 10^{21} bytes	Escala del ecosistema mundial de datos
yottabyte (YB) = 10^{24} bytes	
ronnabyte (RB) = 10^{27} bytes	
quettabyte (QB) = 10^{30} bytes	

- En almacenamiento comercial (SSD, USB) los tamaños se usan en prefijos decimales
- En algunos contextos informáticos se usa base 2, con medidas basadas en 1024 (KiB, MiB, GiB...)

- Big Data no consiste solo en tener muchos datos, sino en gestionar datos masivos, rápidos, diversos, imperfectos y útiles
- **Volumen:** gran cantidad de datos generados y almacenados.
- **Velocidad:** los datos se crean, transmiten y procesan cada vez más rápido.
- **Variedad:** datos proceden de múltiples fuentes y formatos distintos.
- **Veracidad:** no todos los datos son igual de fiables, completos o precisos.
- **Valor:** el objetivo final es transformar los datos en información útil para la toma de decisiones.

- Datos estructurados
 - Siguen un esquema fijo y organizado.
 - Se almacenan normalmente en tablas con filas y columnas.
 - Ejemplos: bases de datos relacionales, hojas de cálculo, archivos CSV.
- Datos semiestructurados
 - Tienen cierta organización, pero no un esquema rígido.
 - Permiten almacenar información más flexible.
 - Ejemplos: JSON, XML, logs.
- Datos no estructurados
 - No siguen una estructura tabular fija.
 - Suelen requerir técnicas específicas de procesamiento y análisis.
 - Ejemplos: texto libre, imágenes, audio, vídeo.

Ejemplo de datos estructurados: fichero CSV

The diagram shows a CSV file structure with the following components:

- Columns:** Indicated by purple arrows pointing to the header row (Surname, Country, City, Occupation, Age).
- Rows:** Indicated by yellow arrows pointing to the data rows (0 to 6).
- Data:** Indicated by red arrows pointing to the individual data cells within the rows.

	Surname	Country	City	Occupation	Age
0	Schmidt	Germany	Munich	Carpenter	43
1	Kelly	Ireland	Galway	Musician	25
2	Johnson	USA	Houston	Doctor	37
3	Hayashi	Japan	Kyoto	Chef	30
4	Baraka	Syria	Damascus	Professor	61
5	Johansson	Sweden	Malmö	Driver	45
6	Park	South Korea	Busan	Student	21

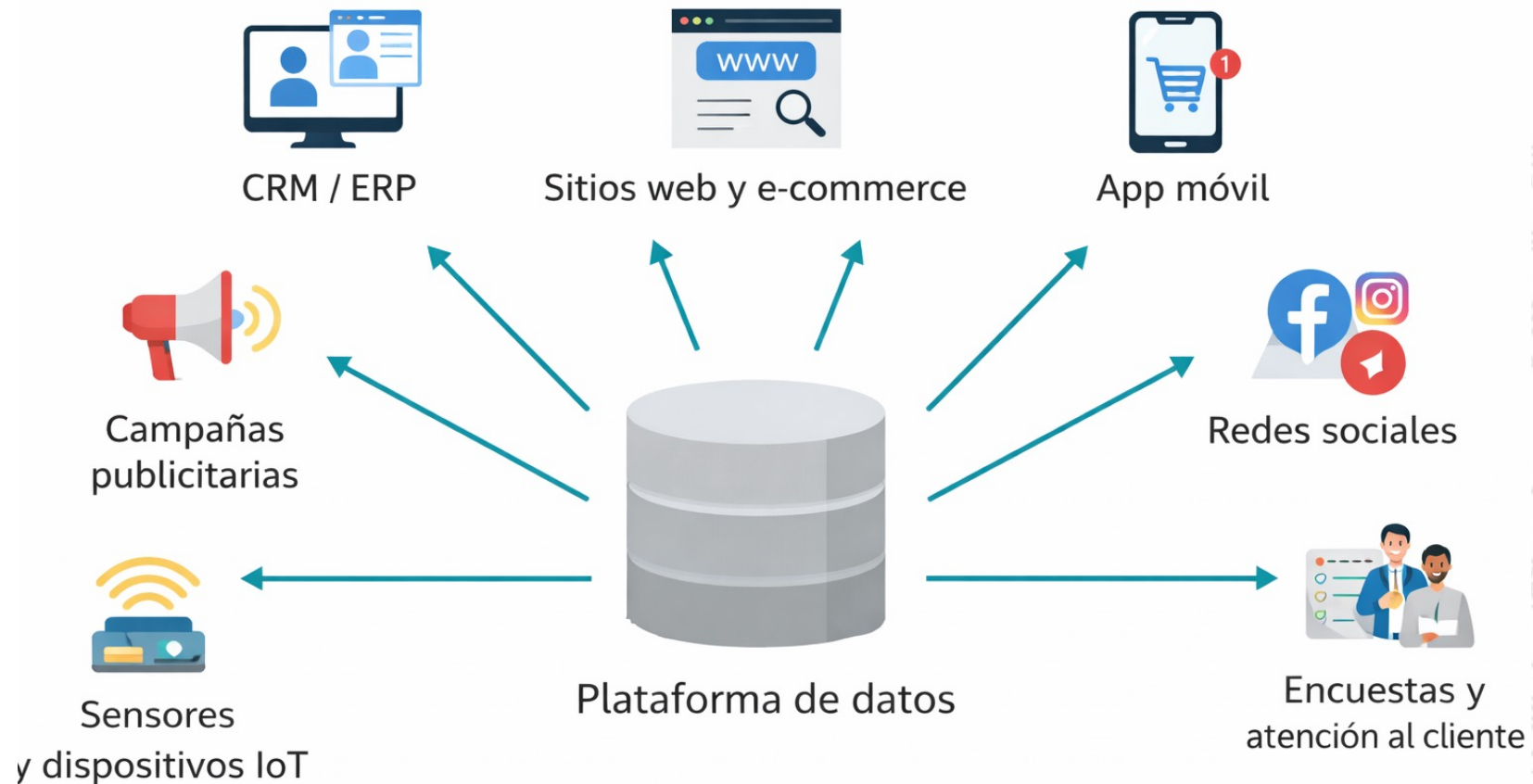
Ejemplo de datos semiestructurados: fichero JSON

```
{  
  "libro": [  
    {  
      "id": "01",  
      "lenguaje": "Java",  
      "edición": "tercera",  
      "autor": "Herbert Schildt"  
    },  
    {  
      "id": "07",  
      "lenguaje": "C++",  
      "edición": "seguna",  
      "autor": "E.Balagurusamy"  
    }  
  ]  
}
```

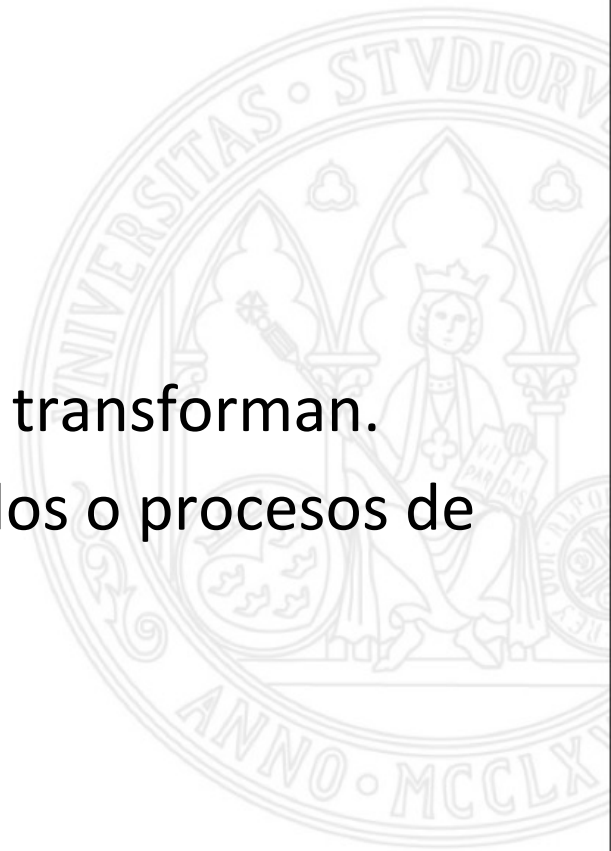


Ejemplo de datos no estructurados: imagen



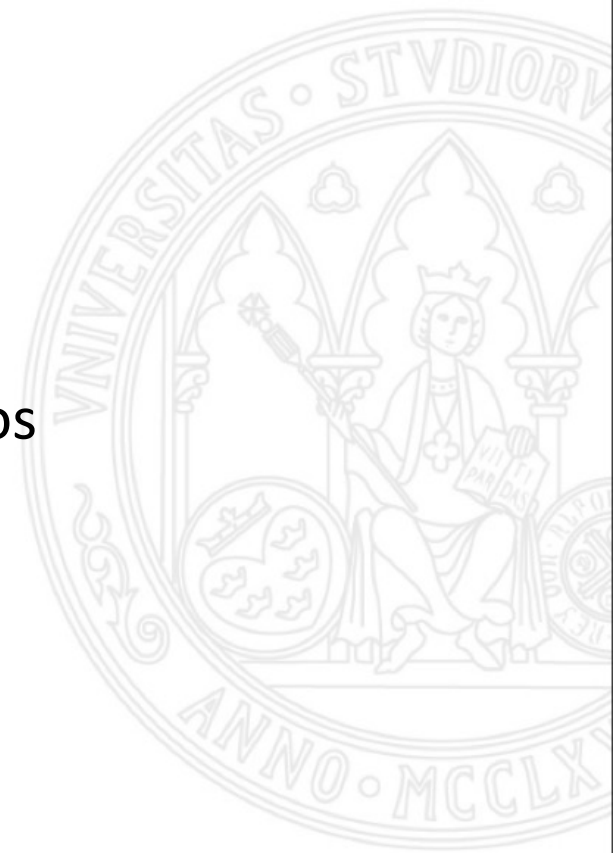


- Los datos no se analizan directamente en el mismo lugar donde se generan.
- Normalmente siguen un flujo general formado por varias etapas:
 - **Fuentes:** dónde se generan los datos.
 - **Ingesta:** cómo se recogen y trasladan.
 - **Almacenamiento:** dónde se guardan.
 - **Procesamiento:** cómo se limpian, integran y transforman.
 - **Consumo:** cómo se usan en informes, modelos o procesos de toma de decisiones.



Unidad 1. Introducción a Big Data y Minería de Datos

MINERÍA DE DATOS

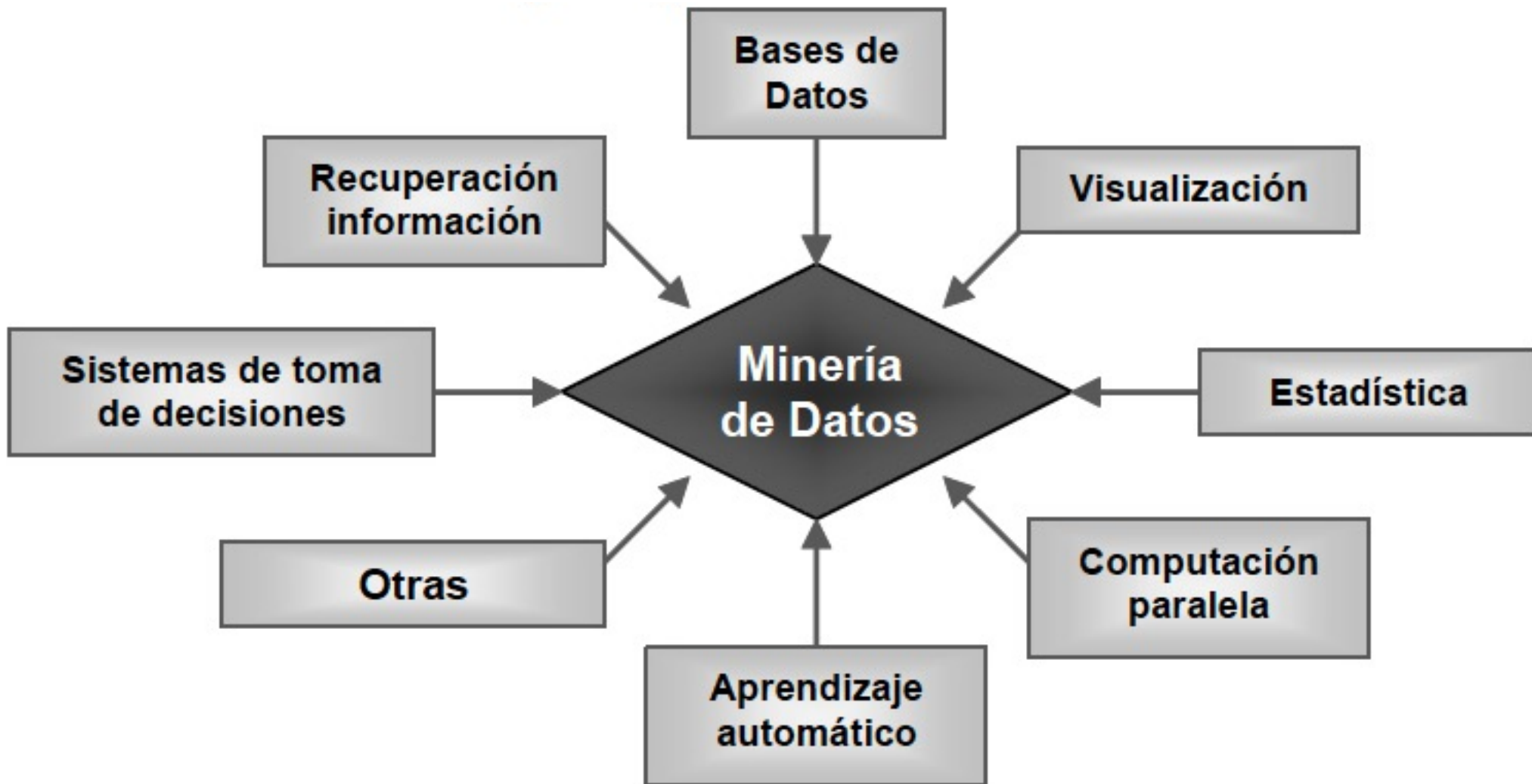


¿Qué es la minería de datos?

- [Witten & Frank 2000] se define la minería de datos como “*el proceso de **extraer conocimiento útil y comprensible**, previamente desconocido, desde grandes cantidades de **datos** almacenados en **distintos formatos**.*”
- Su objetivo no es solo describir datos, sino apoyar la comprensión y la toma de decisiones
- Permite descubrir:
 - Patrones frecuentes
 - Relaciones entre variables
 - Grupos similares
 - Valores atípicos
- El usuario final no tiene por qué ser un experto en minería de datos (la comunicación de información es clave)



La minería es un campo multidisciplinar



Proceso de descubrimiento de conocimiento (KDD)

- La minería de datos forma parte de un proceso más amplio
- KDD describe el recorrido completo desde los datos iniciales hasta el conocimiento útil.
- Las dos primeras fases se engloban como preparación (o procesamiento) de datos
- Fase previa de necesidades y requisitos

