

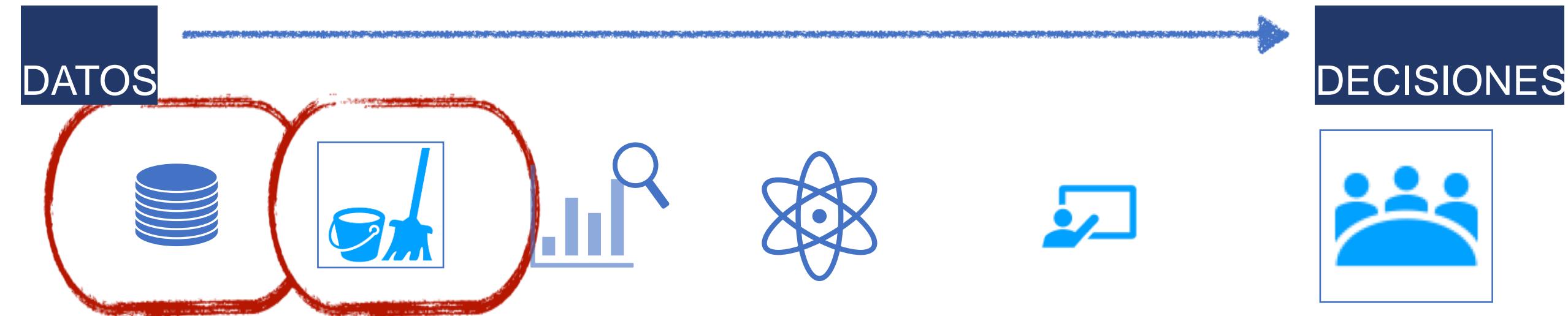
Unidad 4

Sobre los datos, su limpieza y exploración

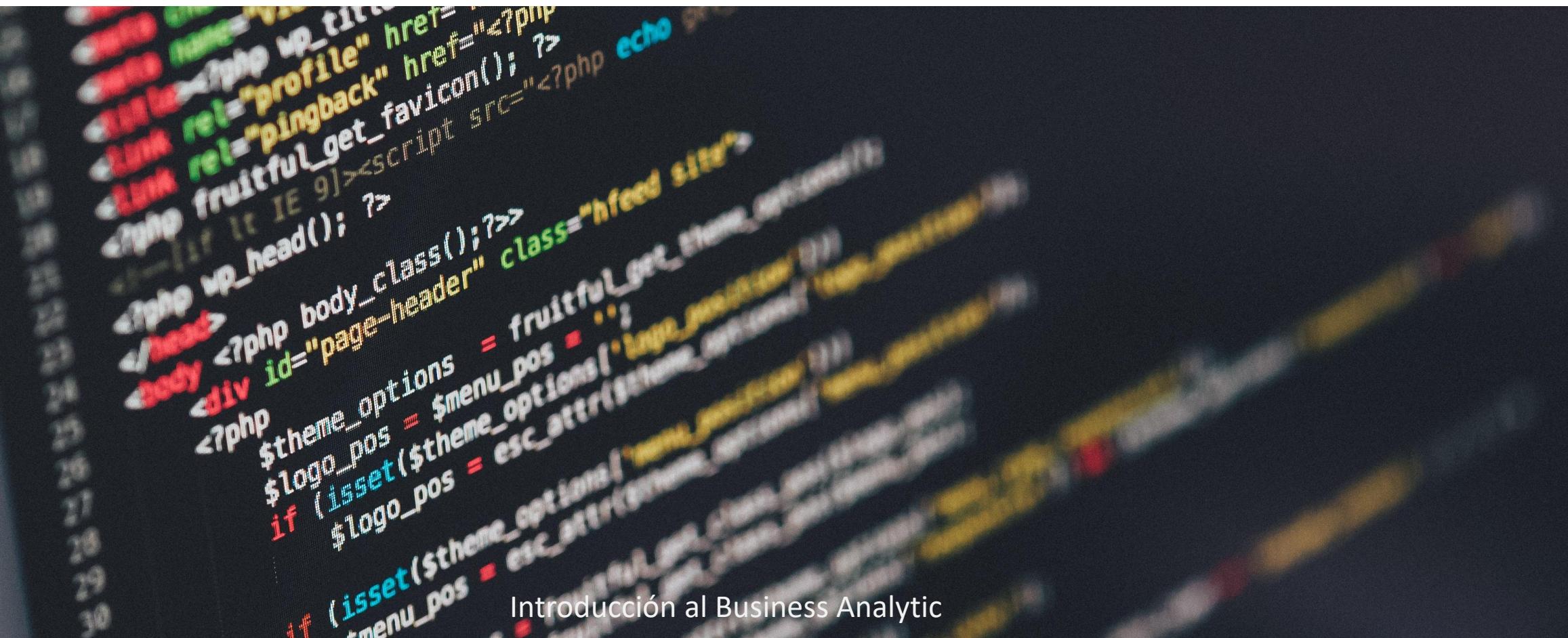
Objetivos

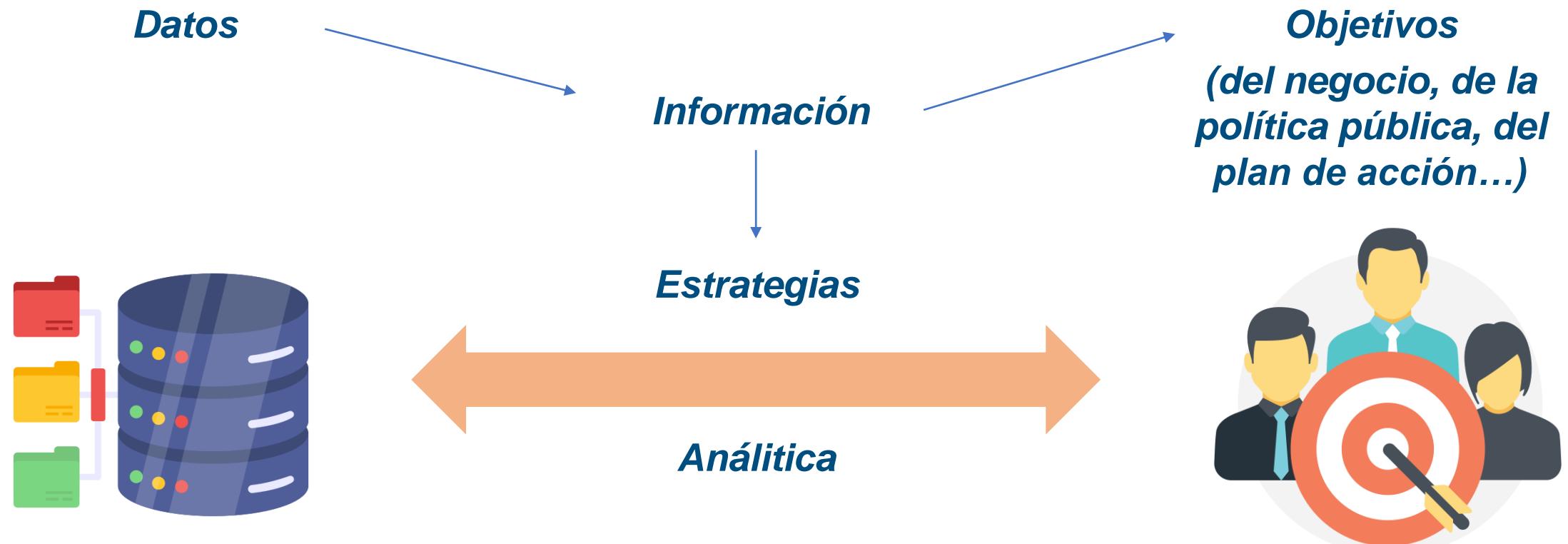
Al finalizar esta unidad el estudiante estará en capacidad de:

- Explicar en sus propias palabras las posibles fuentes de los datos que emplean las organizaciones para el proceso de ***business Analytics***.
- Explicar en sus propias palabras los conceptos de: base de datos, data warehouse, Data lake, ETL y Metadata.
- Realizar un proceso de limpieza inicial de una base de datos en R.
- Explicar en sus propias palabras que tipos de problemas se pueden encontrar en una base de datos



Un poco de vocabulario antes de entrar en el detalle técnico





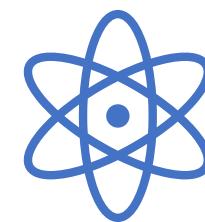
Enlaza los datos con los objetivos del negocio permitiendo resolver problemas específicos o apoyar la toma de decisiones.

Work flow

DATOS



DECISIONES





Fuentes de Datos
en este mundo

Fuentes de datos

De la organización

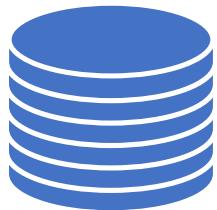
Eventos Web

Datos de clientes

Datos de logística

Transacciones financieras

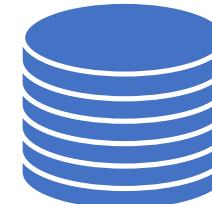
Datos abiertos o públicos



Eventos Web

event_name	timestamp	user_id
homepage_visit	2019-01-01 12:01:01	1234

Grandes volúmenes de información



Se refiere a los datos recopilados a partir de interacciones de usuarios en sitios web y aplicaciones.

Registros de navegación: Rutas que toman los usuarios a través de un sitio web.

Tasas de clics (CTR): Información sobre enlaces que los usuarios han hecho clic.

Datos del consumidor (solicitados)

¿Para qué?

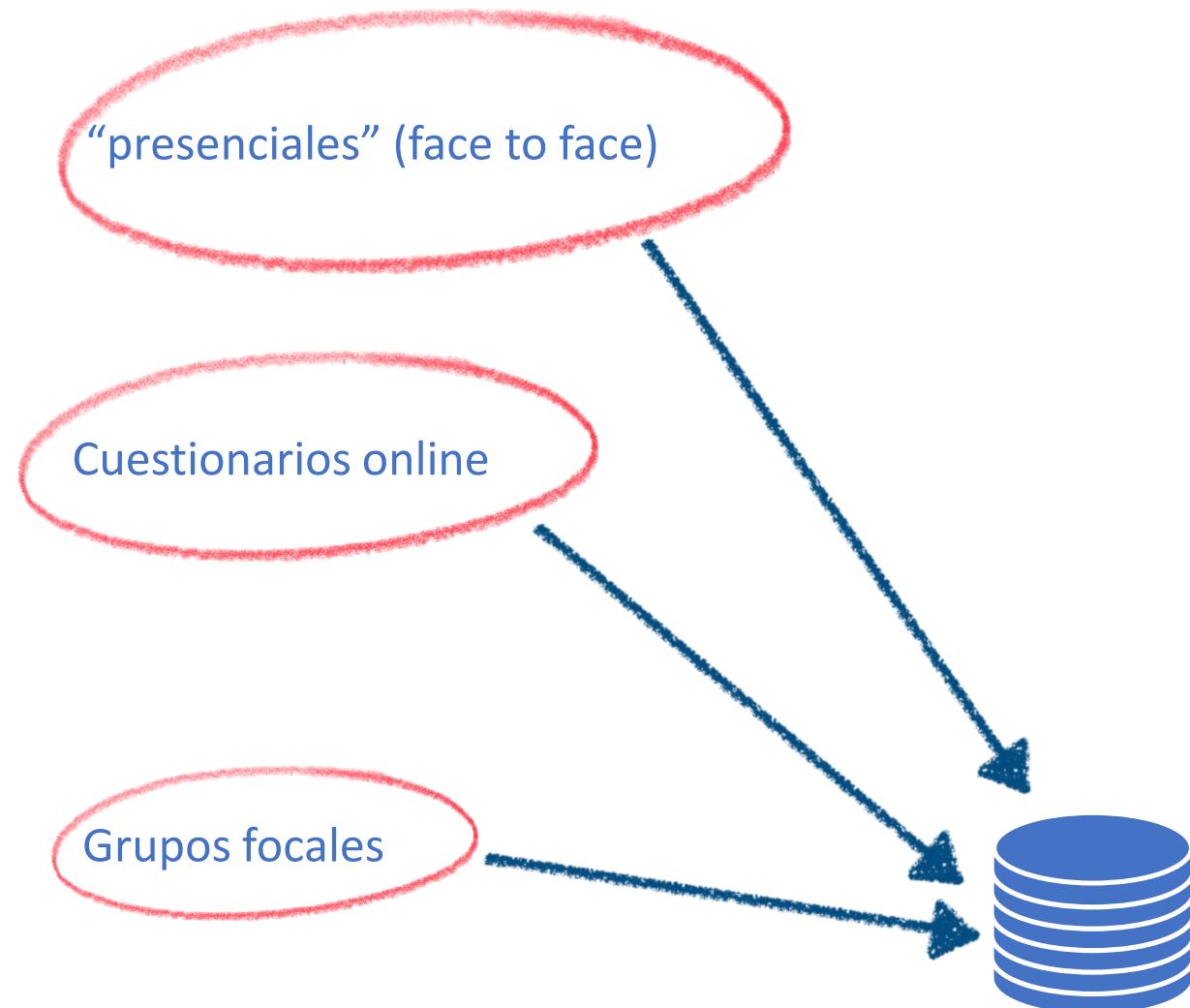
Crear “marketing collateral”

- **Encuestas de satisfacción:** Opiniones y valoraciones directas de productos o servicios.
- **Datos de compra:** Historial de compras y preferencias de los clientes.
- **Comentarios y valoraciones**

Tomar decisiones sin riesgos

Monitorear la calidad

Fuentes de datos solicitados



Fuentes de datos

De la organización

Eventos Web

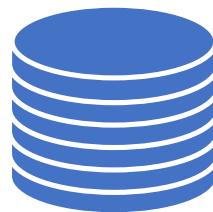
Datos de clientes

Datos de logística

Transacciones financieras

Datos abiertos o públicos

APIs (Application Programming Interface)



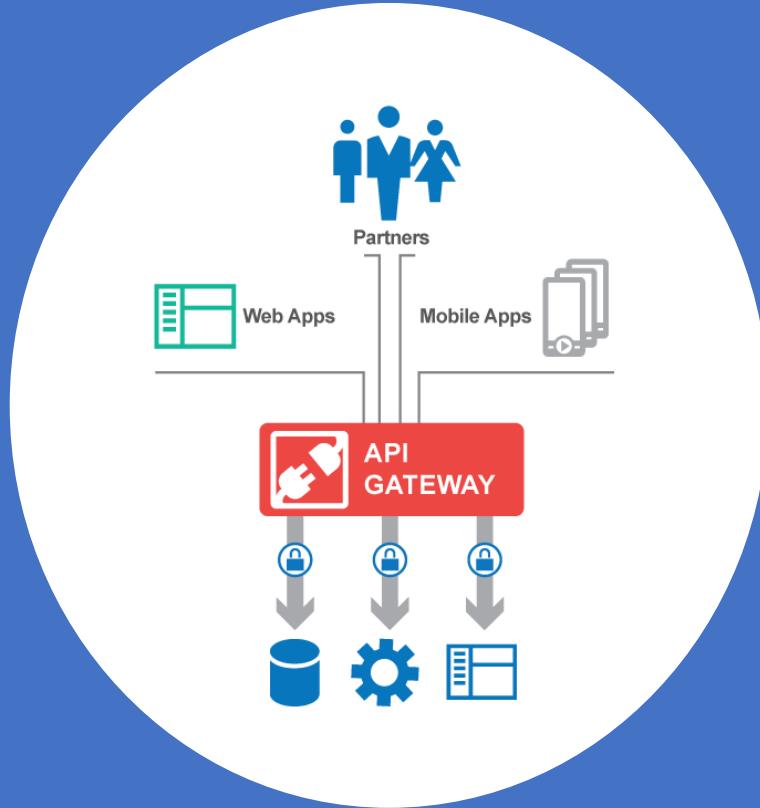
APIs

Permiten “chupar” información de internet

- Twitter
- Wikipedia
- Yahoo! Finance
- Google Maps

Las APIs permiten por ejemplo monitorear un hashtag #icesi





Una API (Application Programming Interface) es un conjunto de código que permite la transmisión de datos entre un producto de software y otro. (contiene las condiciones de este intercambio de datos).

Fuentes de datos

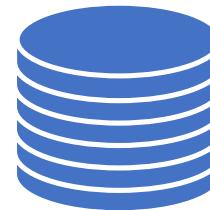
De la organización

Eventos Web

Datos de clientes

Datos de logística

Transacciones financieras

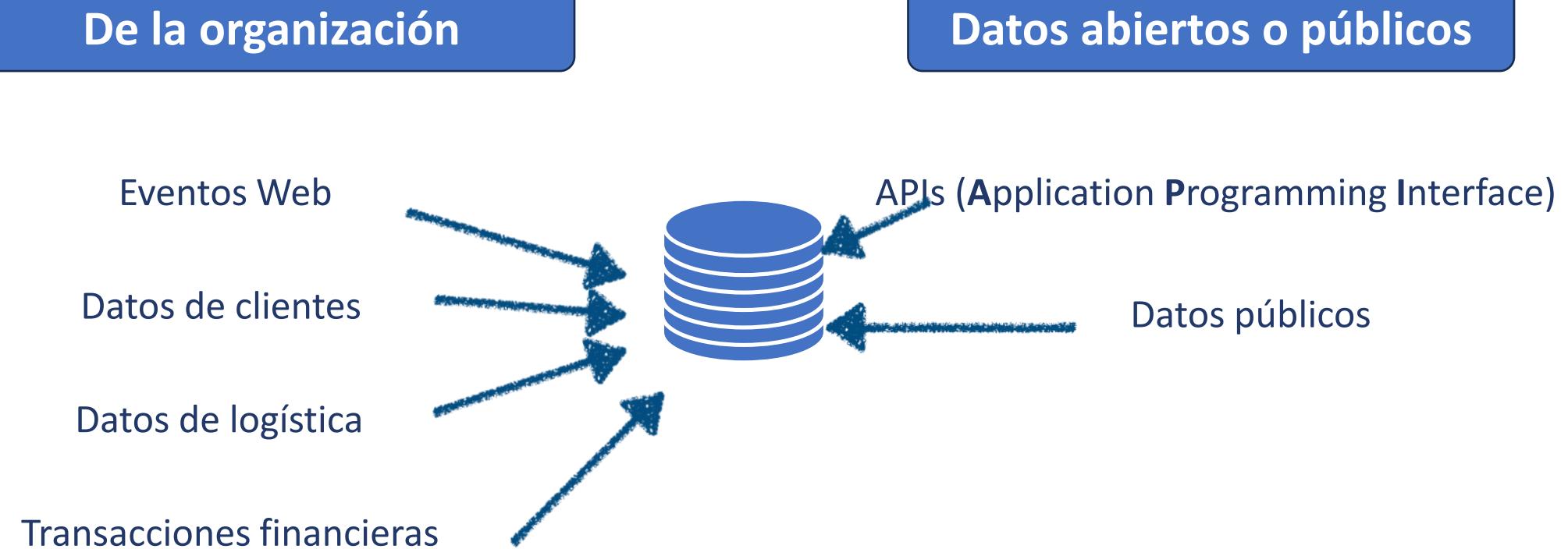


Datos abiertos o públicos

APIs (Application Programming Interface)

Datos públicos







Tipos de Datos en este mundo

¿Por qué nos importan los tipos de datos?

Visualización y análisis de los datos

Almacenamiento de los datos

Datos cuantitativos vs cualitativos

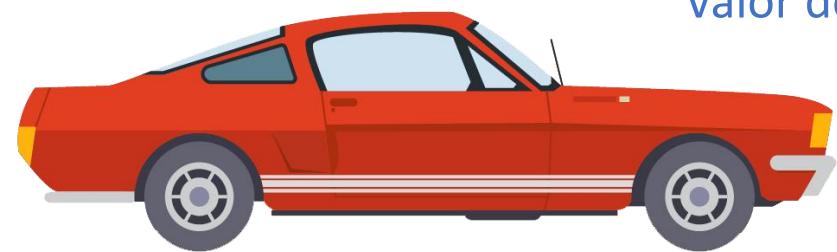
Cuantitativos

- Tienen un número naturalmente asociado
- Puede ser medido
- El número tiene un significado natural

Cualitativos

- Normalmente relacionado con descripciones
- Los datos son observables pero no medibles
- Si se asigna un número, no tiene significado natural

Datos cuantitativos



Valor de compra \$ 60 millones

2 ruedas

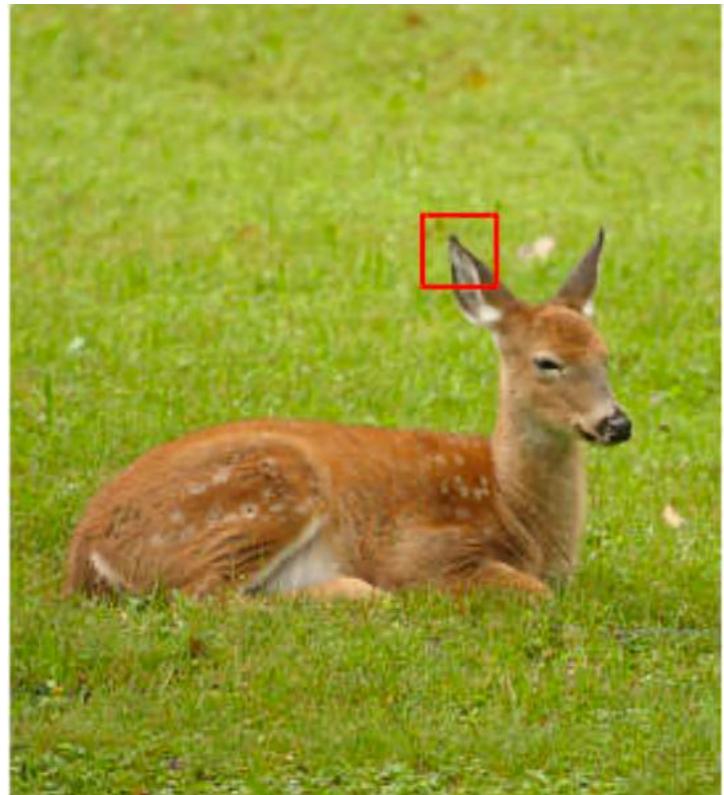
Velocidad máxima 120 km/h

Datos cualitativos



Pero hay más tipos de
datos en el mundo del
business analytics

Datos de imagen



Datos que se presentan en forma visual y se utilizan para análisis de imágenes o procesamiento visual.

- Fotografías de productos para un catálogo en línea.
- Imágenes médicas (como resonancias magnéticas o radiografías).
- Gráficos y diagramas utilizados en presentaciones.

Datos de Texto

 5 months ago

Estación de gasolina con servicio ágil, oportuno, confiable y con precios ajustados al mercado. Cuenta con un personal servicial, amable (Se esmeran en los detalles y te recuerdan aspectos importantes de revisión periódica, tales como nivel ... [More](#)



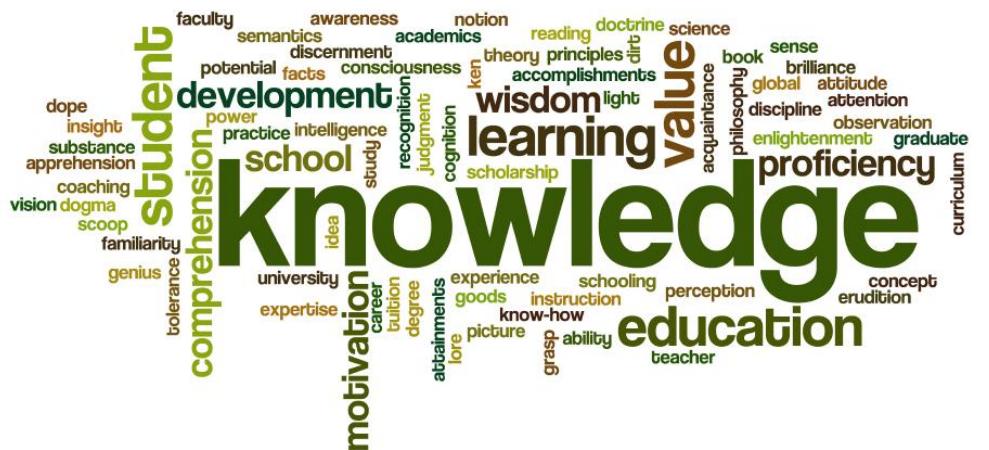
5

Response from the owner 5 months ago

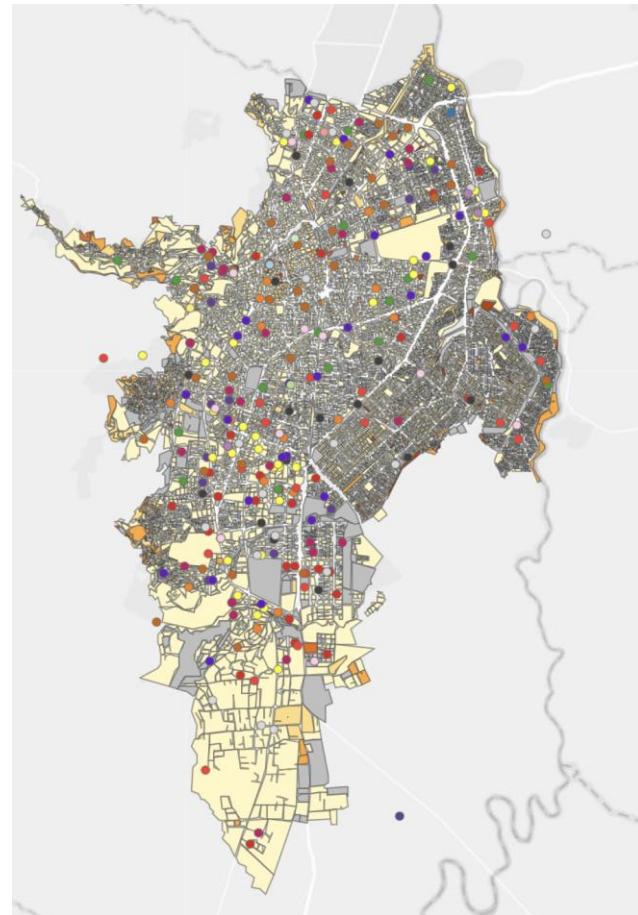
Hola, estamos muy contentos por tu visita y que haya decidido contarnos a todos la experiencia que tuviste en nuestras instalaciones. ¡Gracias por visitarnos y te esperamos pronto para que sigas eligiendo lo mejor para tu vehículo en un solo lugar!

Datos en forma de cadenas de caracteres, que pueden ser analizados para extraer información relevante, como patrones o sentimientos.

- Comentarios de clientes en redes sociales. Tweets.
 - Artículos de noticias o blog posts.
 - Descripciones de productos en un sitio web.



Datos Geoespaciales



Datos que contienen información vinculada a ubicaciones geográficas y se utilizan para análisis espaciales y visualización en mapas.

- Coordenadas GPS de una tienda (por ejemplo, Latitud: 34.0522, Longitud: -118.2437).
- Datos de tráfico en una carretera en tiempo real. – waze, Google maps
- Mapas de distribución de clientes en una región determinada (o áreas geográficas específicas).

 Techopedia

[IBM y la NASA están aplicando la IA geoespacial a los grandes problemas](#)

Este artículo presenta a la IA geoespacial de IBM y la NASA con su aporte para mejorar la respuesta ante catástrofes y supervisión...

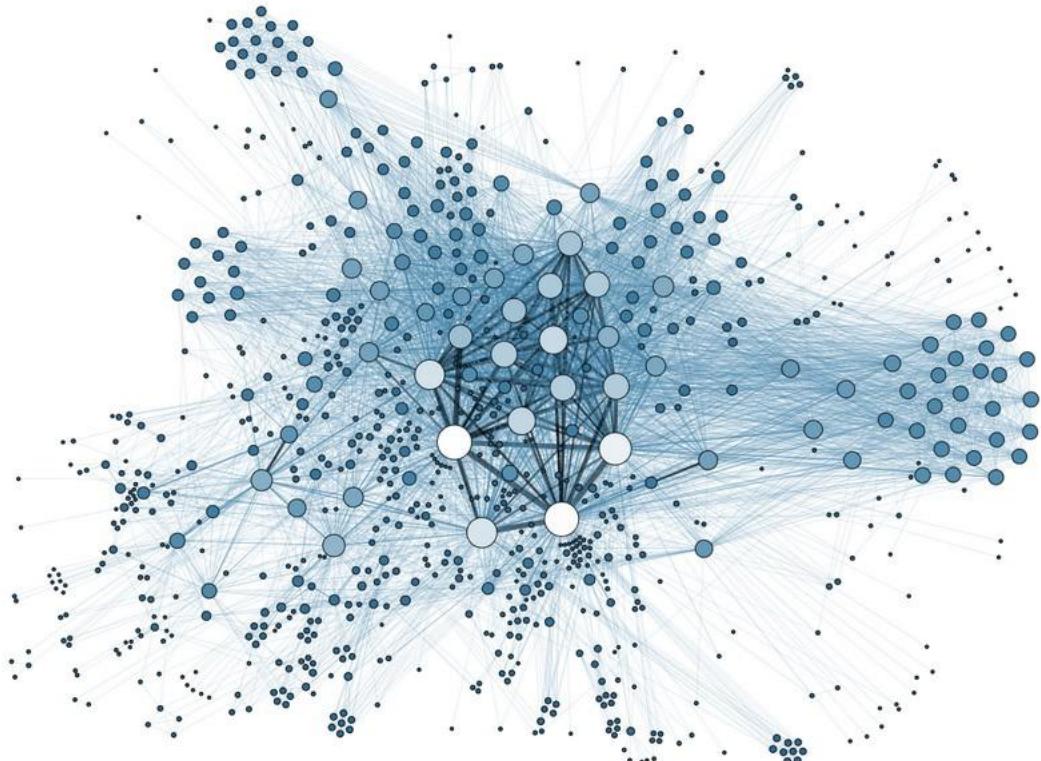
3 abr 2024



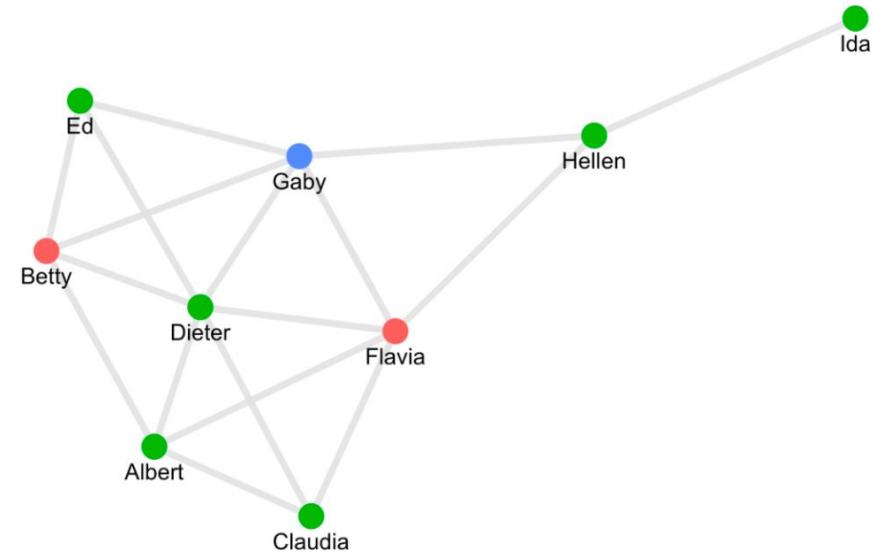
...La colaboración de IBM con la NASA ha dado como resultado un modelo de IA geoespacial fundacional que aborda los retos de la gestión de catástrofes, la supervisión medioambiental y la planificación urbana...

<https://hls.gsfc.nasa.gov/>

Datos de redes



Datos que representan relaciones y conexiones entre diferentes entidades, que pueden ser personas, organizaciones o dispositivos.

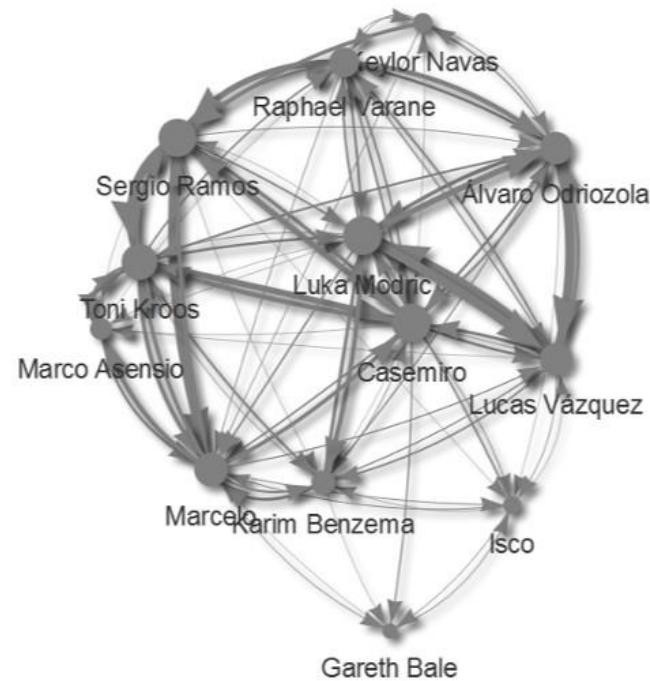


M Marca.com

Todas a Will. Analizando el juego de los equipos de LaLiga

Para realizar este análisis, se han utilizado los datos de Opta, de la temporada 2018-2019. De la aplicación del big data al fútbol tenemos...

6 jun 2019



El Análisis de Redes Sociales (SNA) permite estudiar las conexiones entre jugadores y los patrones de pase en el campo como grafos

<https://www.marca.com/blogs/master-big-data-deportivo/2019/06/06/todas-a-will-analizando-el-juego-de-los.html>

Tipos de datos

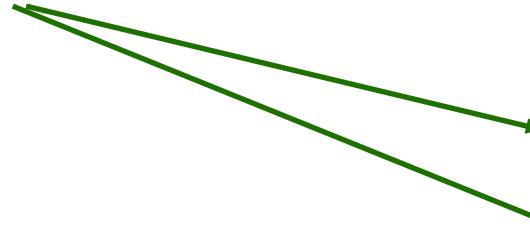
- Cuantitativos
- Qualitativos
- Imágenes
- Textos
- Geoespacial
- Redes
- Y muchos más

Número de veces que aparece el nombre de la organización en tweets sobre un mal servicio corresponde a

Las respuestas a las preguntas abiertas de una encuesta de satisfacción aplicada por el área de mercadeo corresponde a

¿Qué tan satisfecho se sintió con el producto ? La respuesta está en una escala de 1-5, siendo 1 muy insatisfecho y 5 muy satisfecho.

Tipos de datos

- Cuantitativos
 - Cualitativos
 - Imágenes
 - Textos
 - Geoespacial
 - Redes
 - Y muchos más
- 

Y en R

Variables que se encuentran en objetos de clase *data.frame* o *tbl_df*

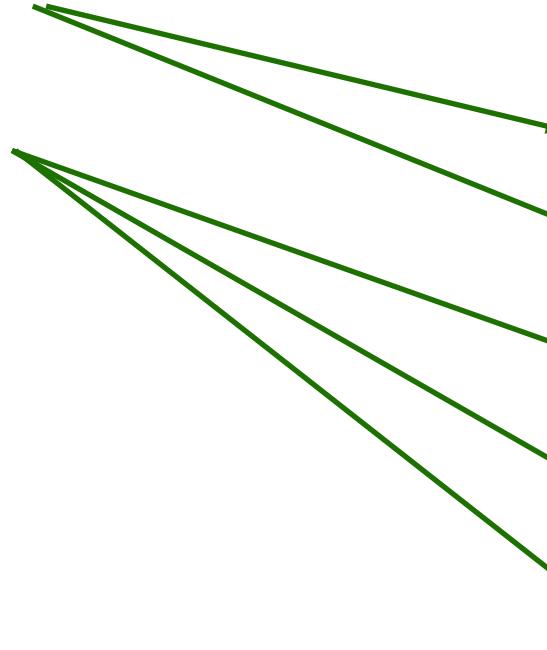
numeric: números reales o decimales
integer: números enteros

Tipos de datos

- Cuantitativos
- Cualitativos
- Imágenes
- Textos
- Geoespacial
- Redes
- Y muchos más

Y en R

Variables que se encuentran en objetos de clase *data.frame* o *tbl_df*

- 
- numeric: números reales o decimales
 - integer: números enteros
 - character: caracteres
 - logical: resultados lógicos (TRUE o FALSE)
 - factor: categórica (puede ser ordenada)

Estructura de los datos



Los datos pueden estar en diferentes formatos:

1. Datos estructurados,
2. Datos semiestructurados y
3. Datos no estructurados.

Datos estructurados:
se pueden almacenar en
filas y columnas (una
tabla)

Datos estructurados: “bien organizados”

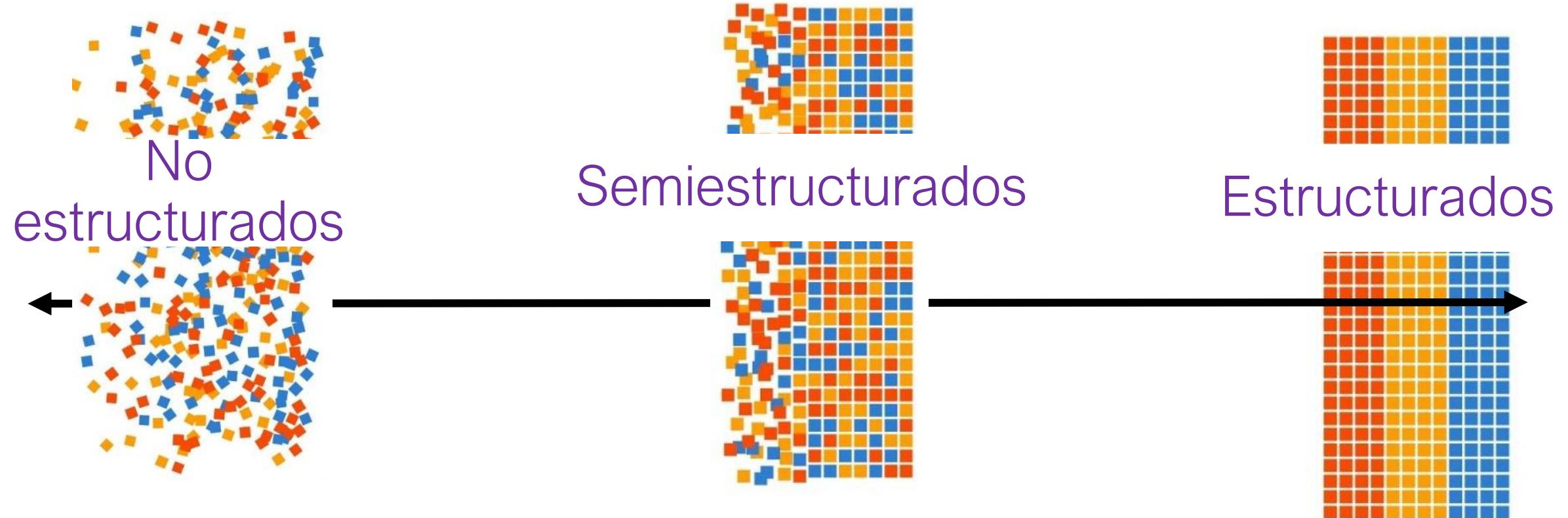
Datos no estructurados:
no están organizados de
forma predefinida o no
tienen un modelo de
datos predefinido.

Ejemplo: un archivo de
Word, PDF, videos, audios

Datos no estructurados:
son datos en bruto y no
organizados. Y a veces
dificilmente se podrán
organizar

Datos semi-
estructurados:
son datos poco
organizados.

Por ejemplo: tweets con
hashtags, información de
logs a un servidor





Almacenamiento
de los datos

Datos de sistemas internos



En una organización los datos tienen diferente origen

Aplicaciones o API's



Una base de datos es cualquier colección de datos organizada para su almacenamiento, accesibilidad y recuperación.

Otros orígenes



Datos de sistemas internos



data warehouse (DW)
(enterprise data warehouse (EDW))
(almacén de datos)



Aplicaciones o API's



Otros orígenes



Un **Data Warehouse** es un repositorio de datos donde se almacenan después de ser extraídos de las bases de datos origen y transformados para evitar duplicidades, espacios en blanco e incoherencias. Su finalidad es crear informes y análisis de datos para la toma de decisiones

Datos de sistemas internos



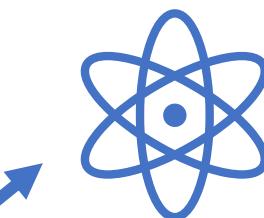
data warehouse (DW)
(enterprise data warehouse (EDW))
(almacén de datos)



Aplicaciones o API's



Otros orígenes



Business Analytics
(Data Science)



Business Intelligence
(BI)



Toma de decisiones

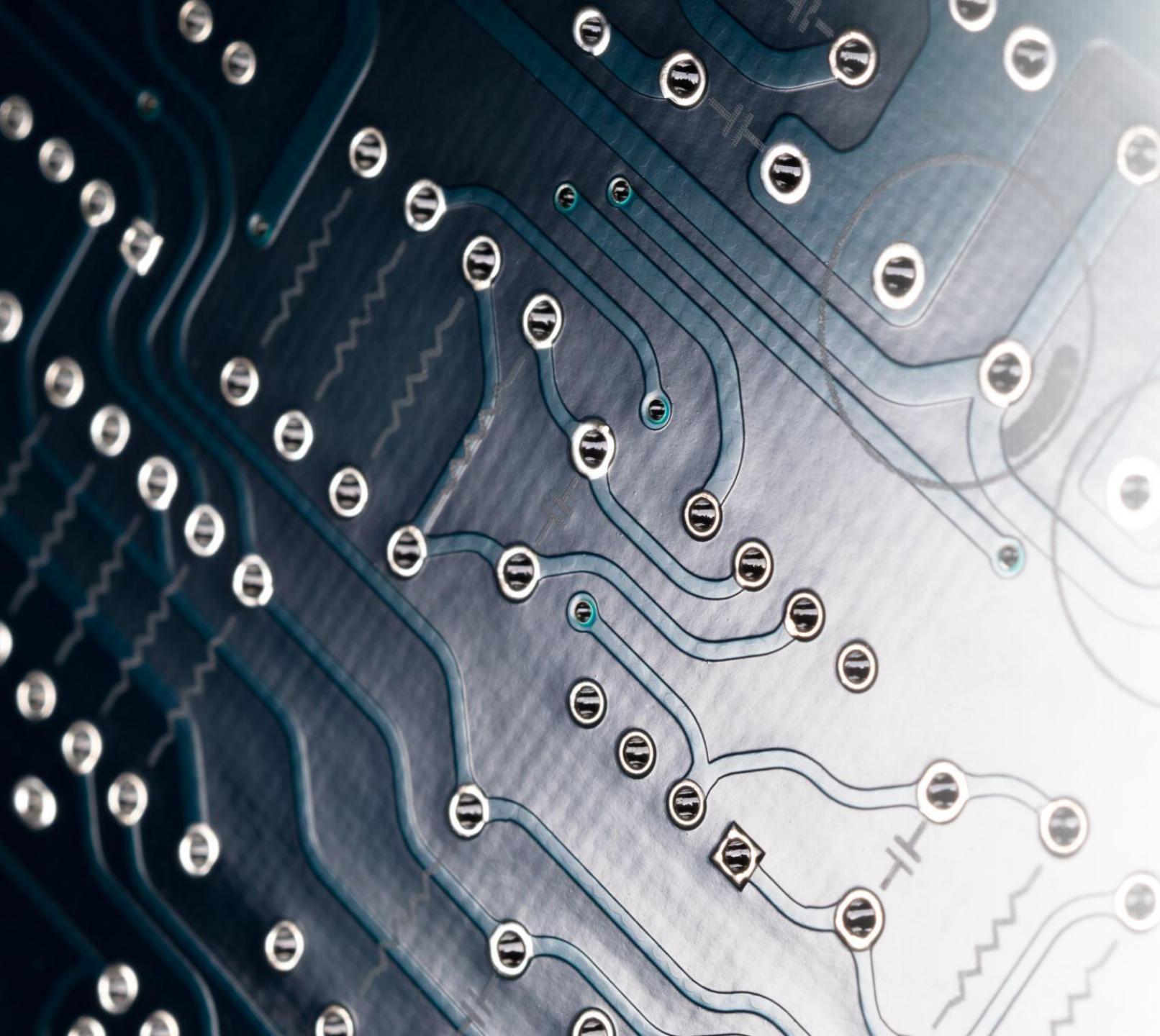
Una **base de datos** es cualquier colección de datos organizada para su almacenamiento, accesibilidad y recuperación.

Un **Data Warehouse** es un tipo de base de datos que integra copias de datos procedentes de sistemas de origen dispares y los guarda para hacer BI o B Analytics

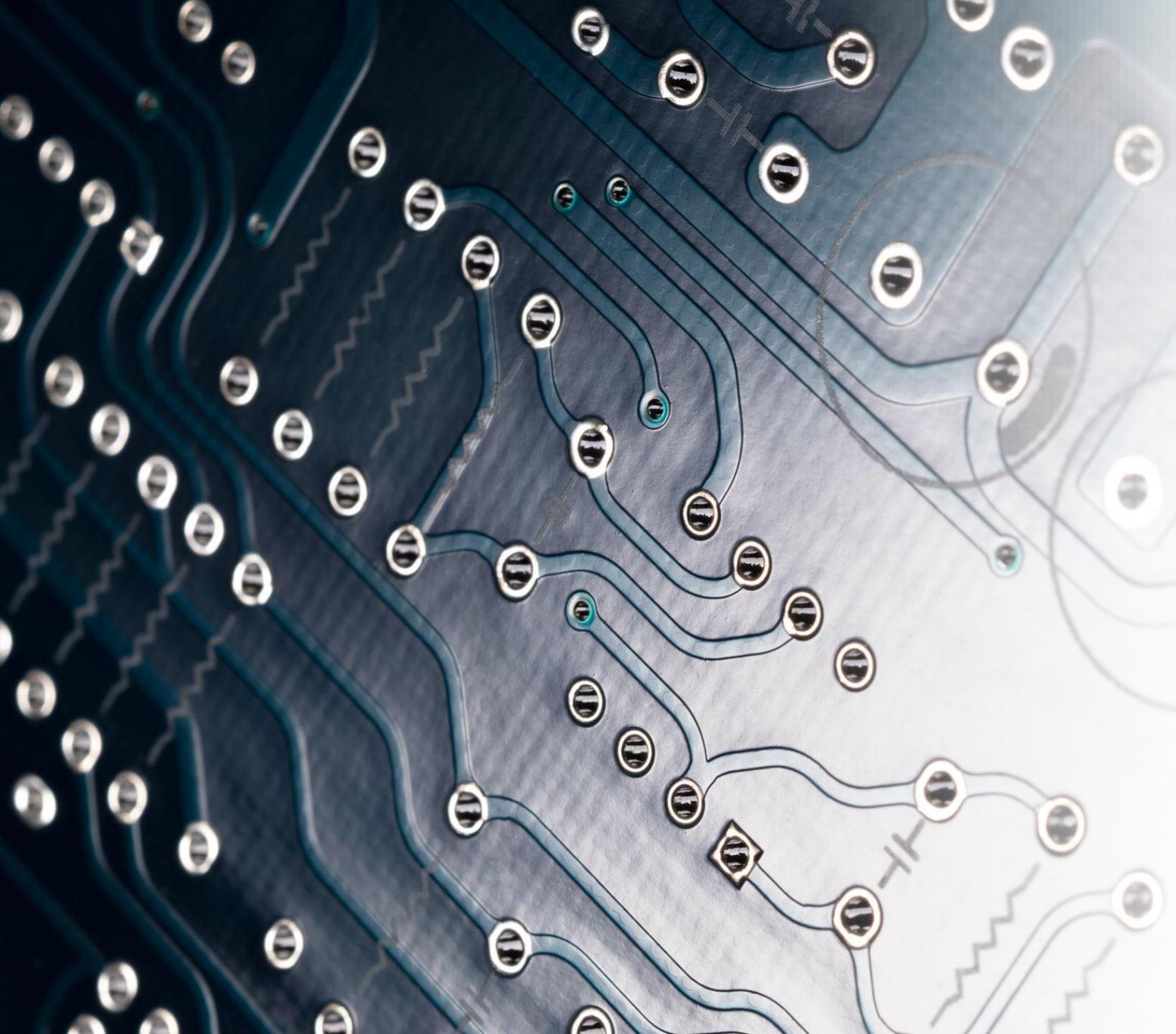
Un **Data Warehouse** es un repositorio de datos donde se almacenan después de ser extraídos de las bases de datos originales y transformados para evitar duplicidades, espacios en blanco e incoherencias. Finalmente, también es empleada para crear informes y análisis de datos para la toma de decisiones.

Tradicionalmente data estructurada

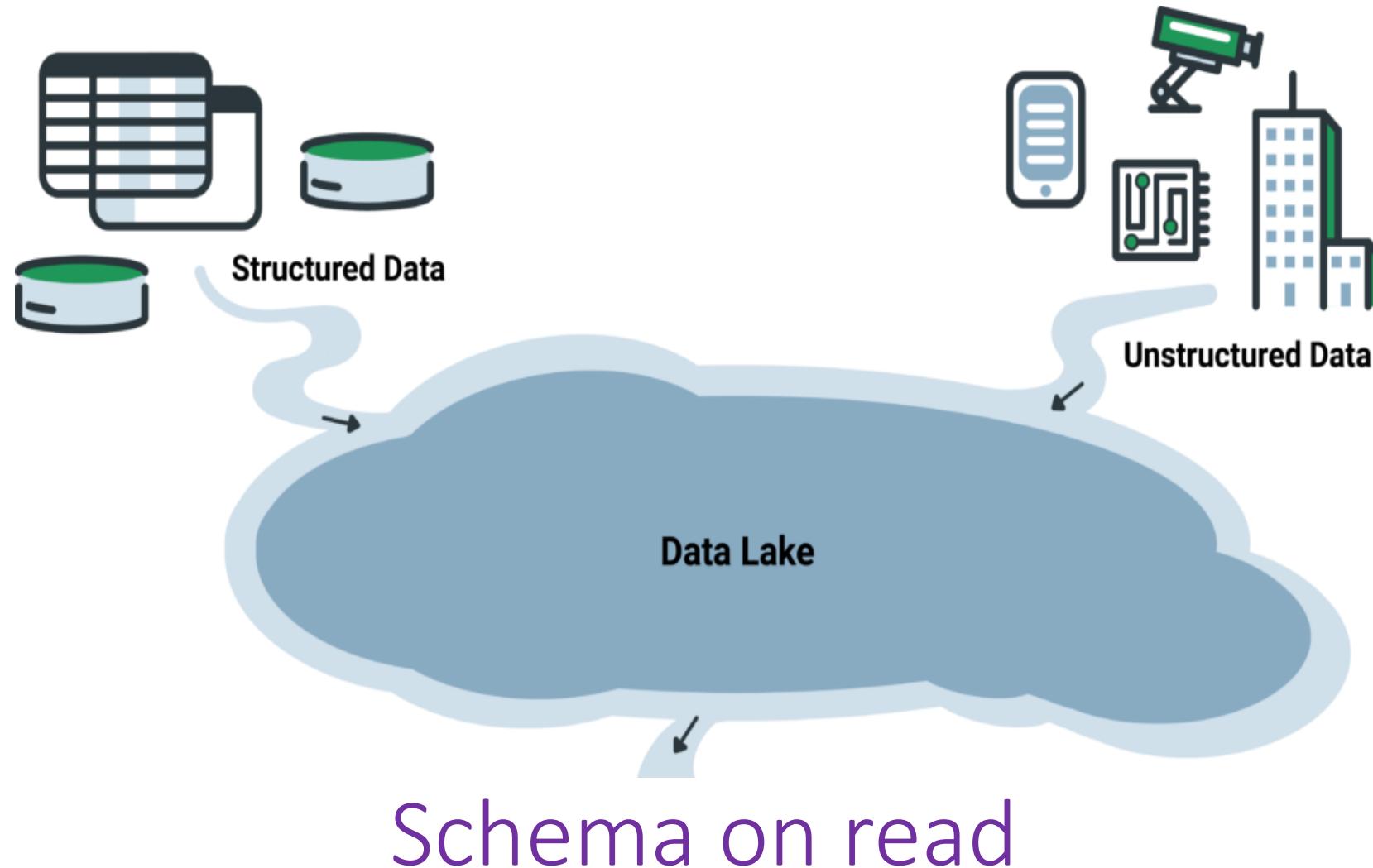
Schema on write



Un *data lake* es un repositorio de almacenamiento central que contiene *big data* de muchas fuentes en un formato crudo y granular.



Puede almacenar datos estructurados, semiestructurados o no estructurados, lo que significa que los datos pueden conservarse en un formato más flexible para su uso futuro.



Estos enfoques se relacionan a cómo se define y aplica la estructura de los datos en un sistema de gestión de datos y afecta tanto la ingesta como la consulta de los mismos

Aspecto	Schema on Write	Schema on read
Definición del Esquema	Se define antes de almacenar datos	Se define al momento de la lectura de datos
Estructura	Rígida y predefinida	Flexible y adaptable
Optimización	Eficiente para consultas rápidas	Eficiente para análisis exploratorio
Ejemplo Común	Bases de datos relacionales (SQL)	Sistemas de big data y NoSQL



ETL (extract,
transform, load):

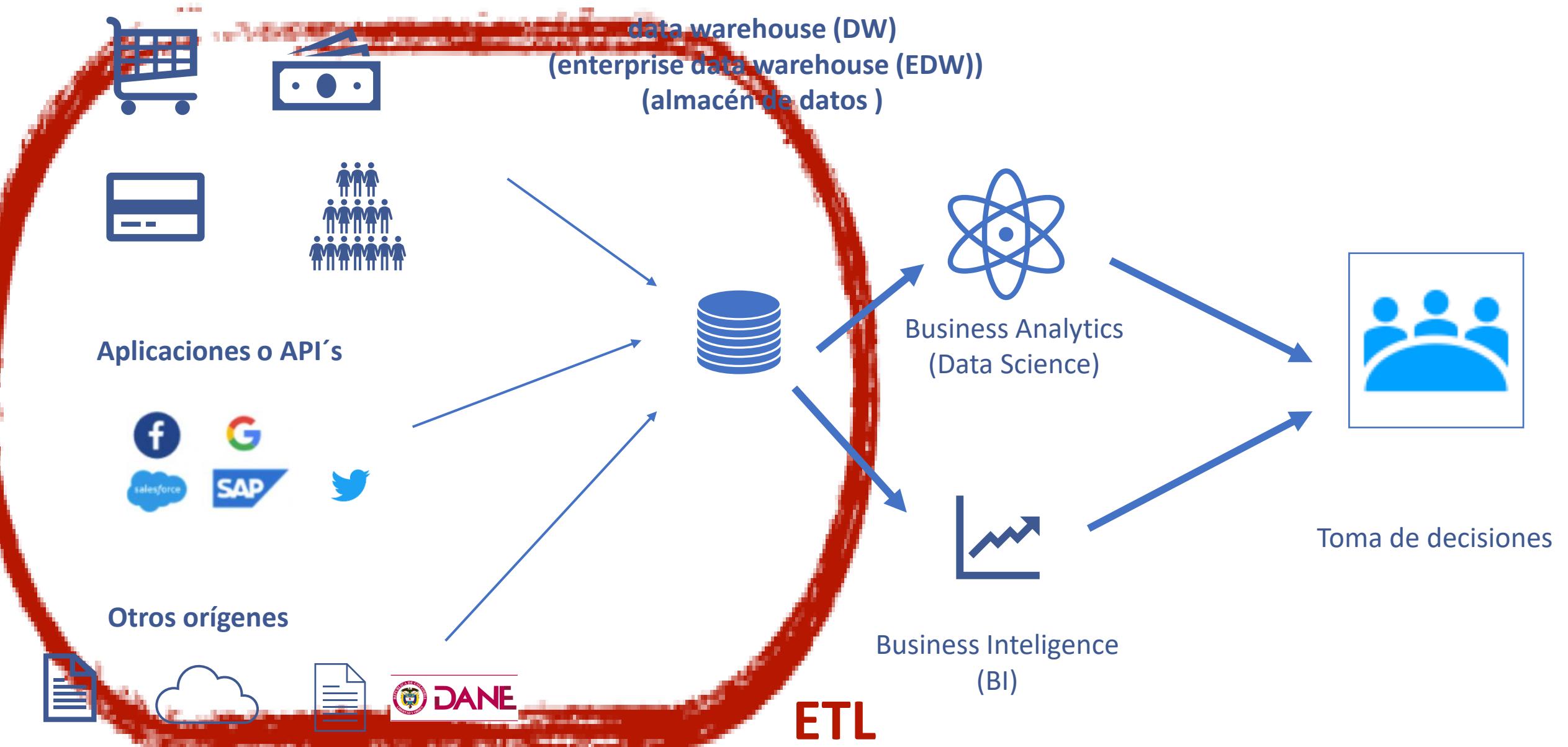
ETL (extract, transform, load):

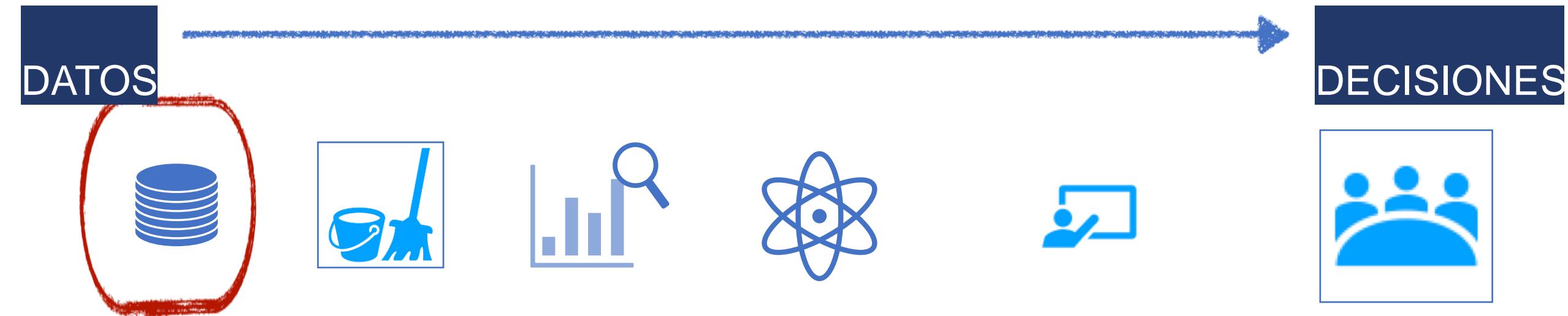
Extracción es el proceso de lectura de datos de una base de datos

Transformación es el proceso de convertir los datos extraídos de su forma anterior a la forma que deben tener para poder colocarlos en otra base de datos (acceder y manipular los datos)

La carga es el proceso de escribir los datos en la base de datos de destino.

Datos de sistemas internos





Ingeniero de Datos



Nosotros nos
concentraremos en la
limpieza y exploración

Limpieza de datos - fases



The image shows a blurred background of a computer monitor displaying a large amount of code. The code is written in a programming language, likely PHP, and includes various HTML tags, CSS classes, and PHP functions like bloginfo(), get_header(), and esc_attr(). The text is in a monospaced font with color-coded syntax highlighting.

```
1 // Crea el header
2
3 <?php wp_head(); ?>
4
5 </head>
6 <body <?php body_class(); ?>>
7 <div id="page-header" class="hfeed site">
8     $theme_options = fruitful_get_theme_settings();
9     $logo_pos = $menu_pos = '';
10    if (isset($theme_options['logo_pos'])) {
11        $logo_pos = esc_attr($theme_options['logo_pos']);
12    }
13    if (isset($theme_options['menu_pos'])) {
14        $menu_pos = esc_attr($theme_options['menu_pos']);
15    }
16    $logo_pos_class = fruitful_get_logo_class();
17    $menu_pos_class = fruitful_get_menu_class();
18    $responsive_menu_type = fruitful_get_responsive_menu_type();
19    $responsive_menu_class = fruitful_get_responsive_menu_class();
20
21    <header class="site-header" role="banner">
22        <div class="inner">
23            <div class="site-branding">
24                <?php bloginfo('name'); ?>
25                <?php bloginfo('description'); ?>
26            </div>
27            <div class="nav-menu">
28                <?php wp_nav_menu(); ?>
29            </div>
30            <div class="social-links">
31                <?php echo get_header_social_links(); ?>
32            </div>
33            <div class="search-form">
34                <?php echo get_search_form(); ?>
35            </div>
36        </div>
37    </header>
38
```

Limpieza de datos - fases

Limpiar una base de datos es una tarea difícil de automatizar

Pero es vital para que el proceso de BA tenga éxito.

Análisis de los datos

Definición y registro de los pasos de limpieza

Verificación

Transformación

Reemplazo de datos sucios con datos limpios

Veamos unos ejemplos
y cómo deberían
limpiarse los datos



No Correspondencia entre el formato de la variable y su tipo

Edad	Edad en R	Estado Civil	Estado Civil
4	"4"	1	1
5	"5"	4	4
6	"6"	5	5
8	"8"	2	2
12	"12"	3	3

Siempre que encontrremos no Correspondencia entre el formato de la variable y su tipo



1=Casado
2=Unión Libre
3=Soltero
4=Viudo
5=Divorciado

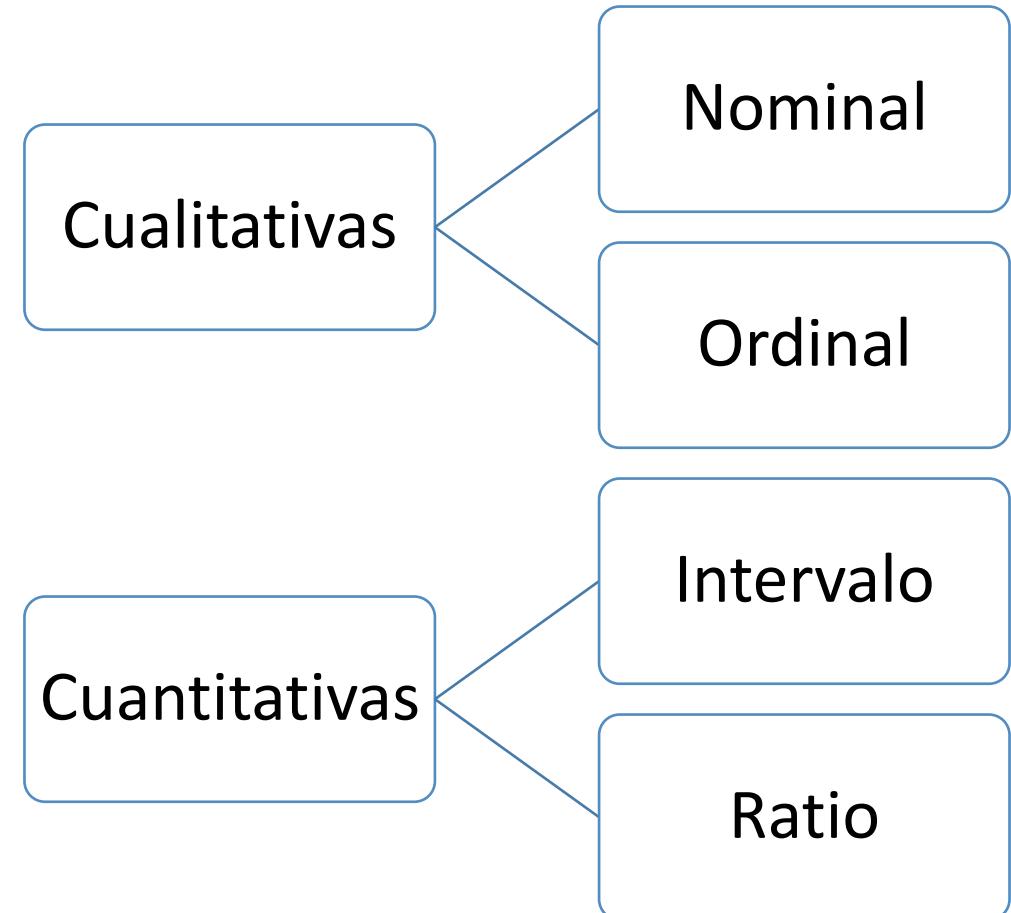
No Correspondencia entre el formato de la variable y su tipo

Solución

Convertir los datos a su respectivo formato.

Funciones

- `as.character`
- `as.numeric`
- `as.integer`
- `as.factor`
- `as.Date`



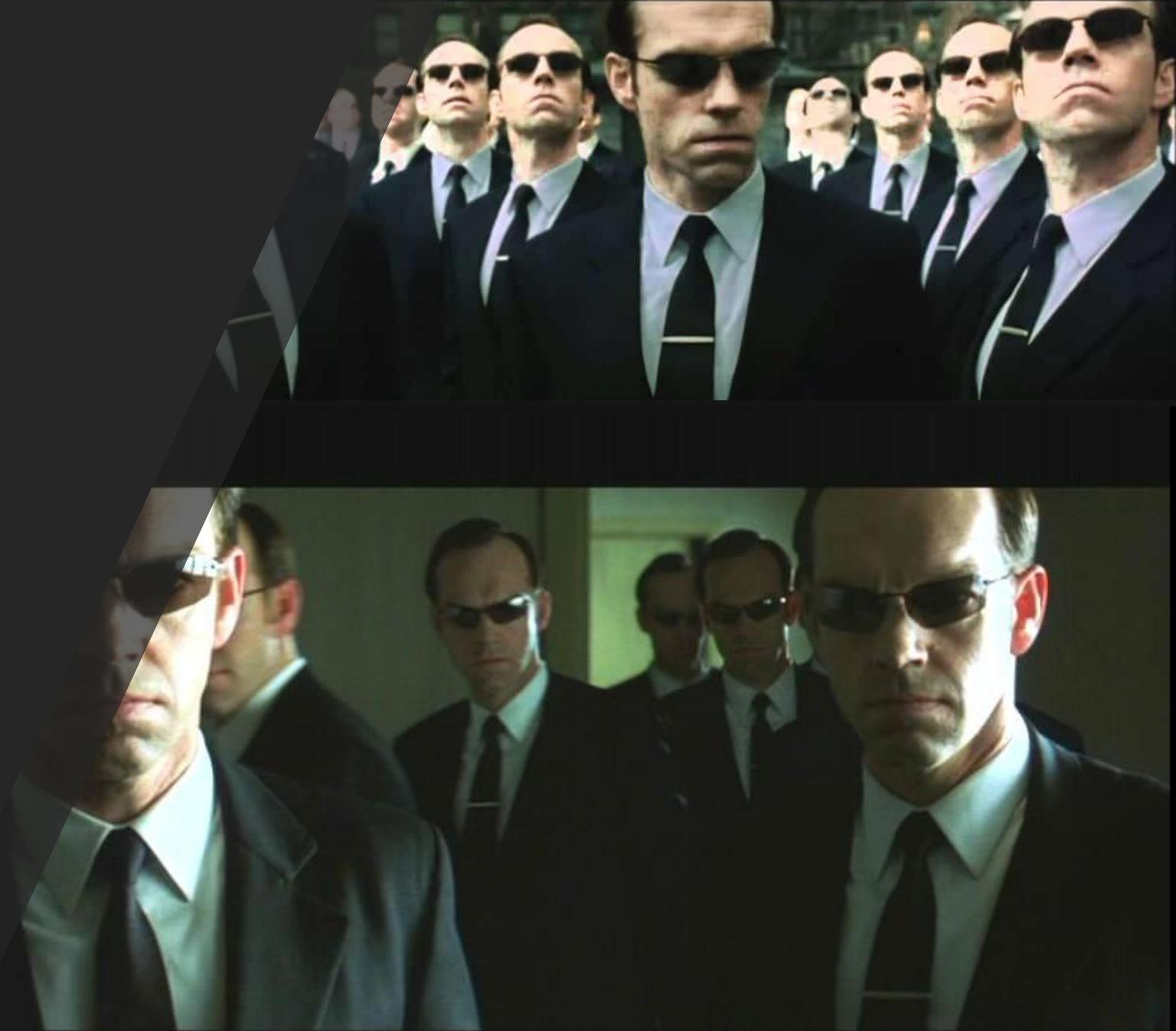
Observaciones duplicadas

¿Por qué este paso es importante?

- Para entender mejor la base de datos
- Evitar la doble contabilización

Solución

- Revisar si en el proceso de carga de datos o de extracción de datos se cometió algún error.
- Al final del proceso
- Eliminar las observaciones duplicadas* se debe tener cuidado con esta solución. No recomendada si no se es experto.



Valores perdidos



“ ”

99



NA

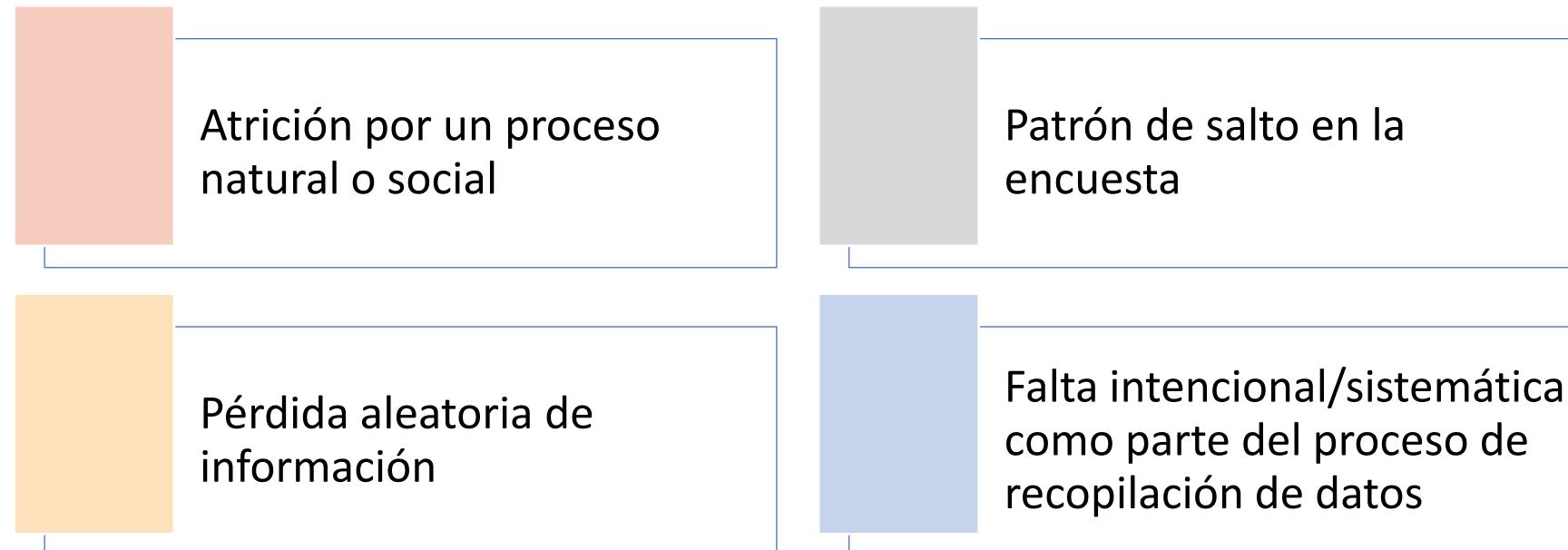
1. Identifique donde se encuentran los valores perdidos y si es necesario recodifiquelos correctamente.

Tip: Maneje el mismo código para indicar que falta un valor en un campo de datos.

2. Intente entender la naturaleza de los datos perdidos.

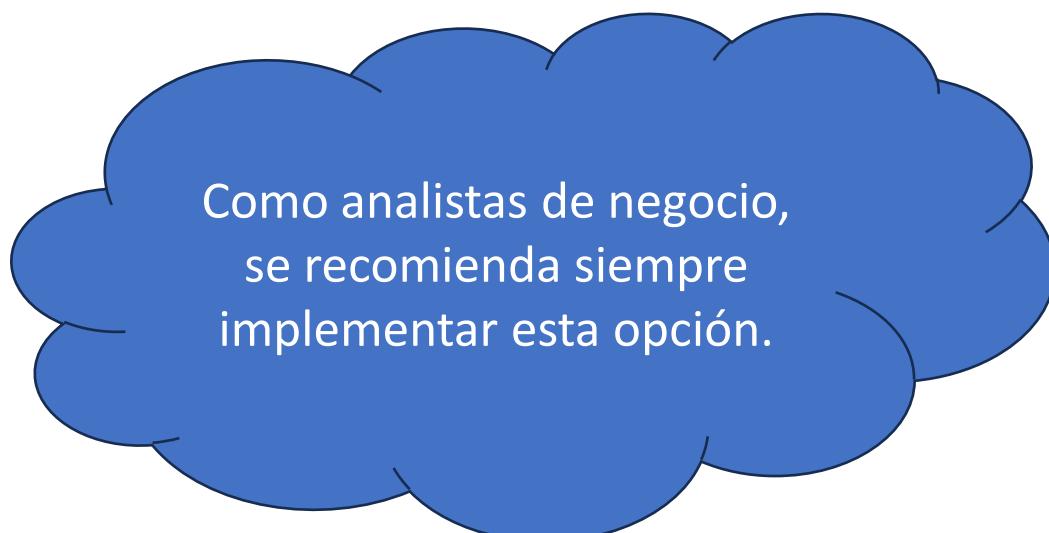
3. Determine cuál es el mejor método para tratarlos.

Entender el porqué hay valores perdidos



Tratamiento de los valores perdidos

1. Reconocer su existencia y no efectuar tratamiento alguno. Pero tenerlo en cuenta en la aplicación de los modelos y el análisis.



Como analistas de negocio,
se recomienda siempre
implementar esta opción.

Otra opciones NO recomendadas

2. Eliminar observaciones con valores perdidos

Gender	8 th grade math test score	12 th grade math score
F	45	.
M	.	99
F	55	86
F	85	88
F	80	75
.	81	82
F	75	80
M	95	.
M	86	90
F	70	75
F	85	.

Gender	8 th grade math test score	12 th grade math score
F	45	.
M	.	99
F	55	86
F	85	88
F	80	75
.	81	82
F	75	80
M	95	.
M	86	90
F	70	75
F	85	.

Fuente: https://liberalarts.utexas.edu/prc/_files/cs/Missing-Data.pdf

Otra opciones NO recomendadas

3. Imputación de datos: sustituir valores perdidos en un campo por otros empleando:

- Media
 - Mediana
 - Moda
-
- Predicción de esos valores faltantes empleado modelos estadísticos



Solo para científicos
de datos
experimentados

Errores de digitación

Convenciones de denominación

- NYC vs New York

Representaciones diferentes

- Si, si, Sí

Espacios vacíos

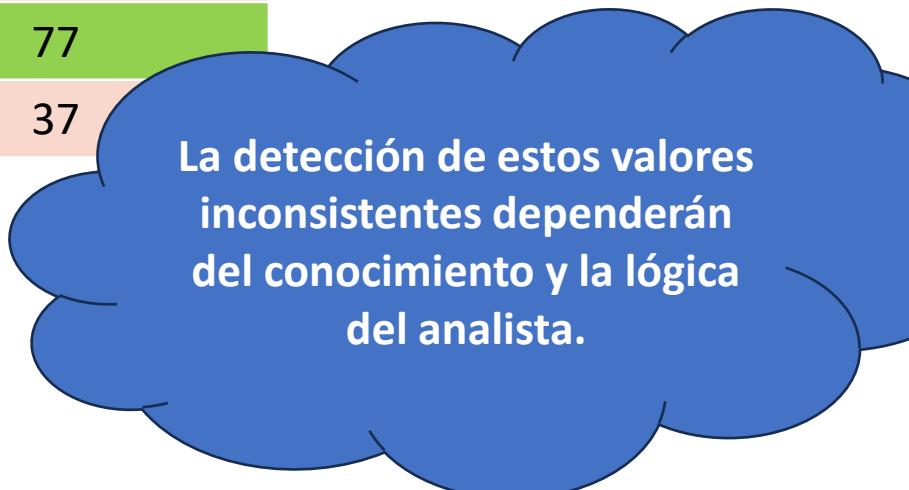
- “Mujer” vs “ Mujer ”

¿Soluciones?

Valores inconsistentes

Estos valores se hallan analizando la consistencia de las respuestas entre variables que se encuentran relacionadas.

Edad	Fecha de nacimiento	Edad calculada
27	1990/01/17	27
21	1995/05/14	21
60	11/11/1956	60
47	7/04/1940	77
37	12/08/1980	37



La detección de estos valores inconsistentes dependerán del conocimiento y la lógica del analista.

Valores sin referencia en el diccionario de variables

Nombre	Estado Civil
Paola	1
Esteban	4
Cesar	5
Carlos	8
María	3
Mayra	2
Roberto	1

Código Estado Civil	
1	Soltero
2	Casado
3	Separado o divorciado
4	Unión libre
5	Viudo

¿Soluciones?

sustituir el
valor por un
"NA"

Importancia del diccionario de variables

Metadata

Datos que describen las características de otros datos

employee_id	first_name	last_name	nin	department_id
44	Simon	Martinez	HH 45 09 73 D	1
45	Thomas	Goldstein	SA 75 35 42 B	2
46	Eugene	Comelsen	NE 22 63 82	2
47	Andrew	Petculescu	XY 29 87 61 A	1
48	Ruth	Stadick	MA 12 89 36 A	15
49	Barry	Scardelis	AT 20 73 18	2
50	Sidney	Hunter	HW 12 94 21 C	6
51	Jeffrey	Evans	LX 13 26 39 B	6
52	Doris	Berndt	YA 49 88 11 A	3
53	Diane	Eaton	BE 08 74 68 A	1
54	Bonnie	Hall	WW 53 77 68 A	15
55	Taylor	Li	ZE 55 22 80 B	1

Metadata

Column	Data Type	Description
employee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current position title, e.g. Secretary
department_id	int	Employee department. Ref: Departments
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date. Null if employee still employed.

Diccionario de variable (*Codebook*)

Un Diccionario de variable es una **descripción técnica de los datos**. Describe **cómo se organizan los datos** en el archivo, qué significan los diferentes números y letras, y cualquier instrucción especial sobre cómo usar los datos correctamente.

Características de un buen diccionario de variables:

- Descripción del estudio: quién lo hizo, por qué lo hicieron, cómo lo hicieron.
- Información muestral: cuál fue la población estudiada, cómo se extrajo la muestra, cuál fue la tasa de respuesta.
- Información técnica sobre los archivos en sí: número de observaciones, longitud de registro, número de registros por observación, etc.
- Estructura de los datos dentro del archivo: jerárquico, etc.
- Detalles sobre los datos: columnas en las que se pueden encontrar variables específicas, ya sean de carácter o numéricas, y si son numéricas, qué formato.
- Texto de las preguntas y respuestas: algunos incluso tienen cuántas personas respondieron de una manera particular.

A dark blue background filled with a dense grid of binary digits (0s and 1s) in a light blue color, creating a digital or data-oriented aesthetic.

Big Data

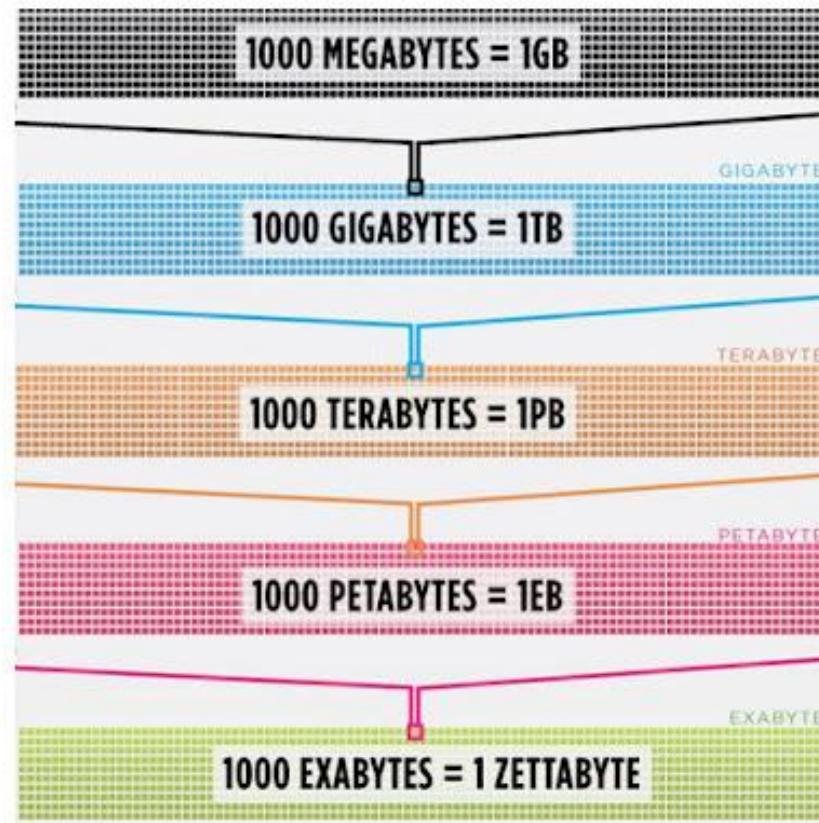
grandes volúmenes de
datos que son muy
variados y veloces

Es muy complicado
capturarlos y procesarlos
con métodos tradicionales

los datos deben cumplir
con las 5V (8V)



El *volumen* de datos
tiende a ser inmenso



How big is a Yottabyte?

TERABYTE

Will fit 200,000 photos or mp3 songs on a single 1 terabyte hard drive.



PETABYTE

Will fit on 16 Backblaze storage pods racked in two datacenter cabinets.



EXABYTE

Will fit in 2,000 cabinets and fill a 4 story datacenter that takes up a city block.



ZETTABYTE

Will fill 1,000 datacenters or about 20% of Manhattan, New York.



YOTTABYTE

Will fill the states of Delaware and Rhode Island with a million datacenters.





La velocidad a la que se generan es muy alta

Cantidad de reproducciones en youtube



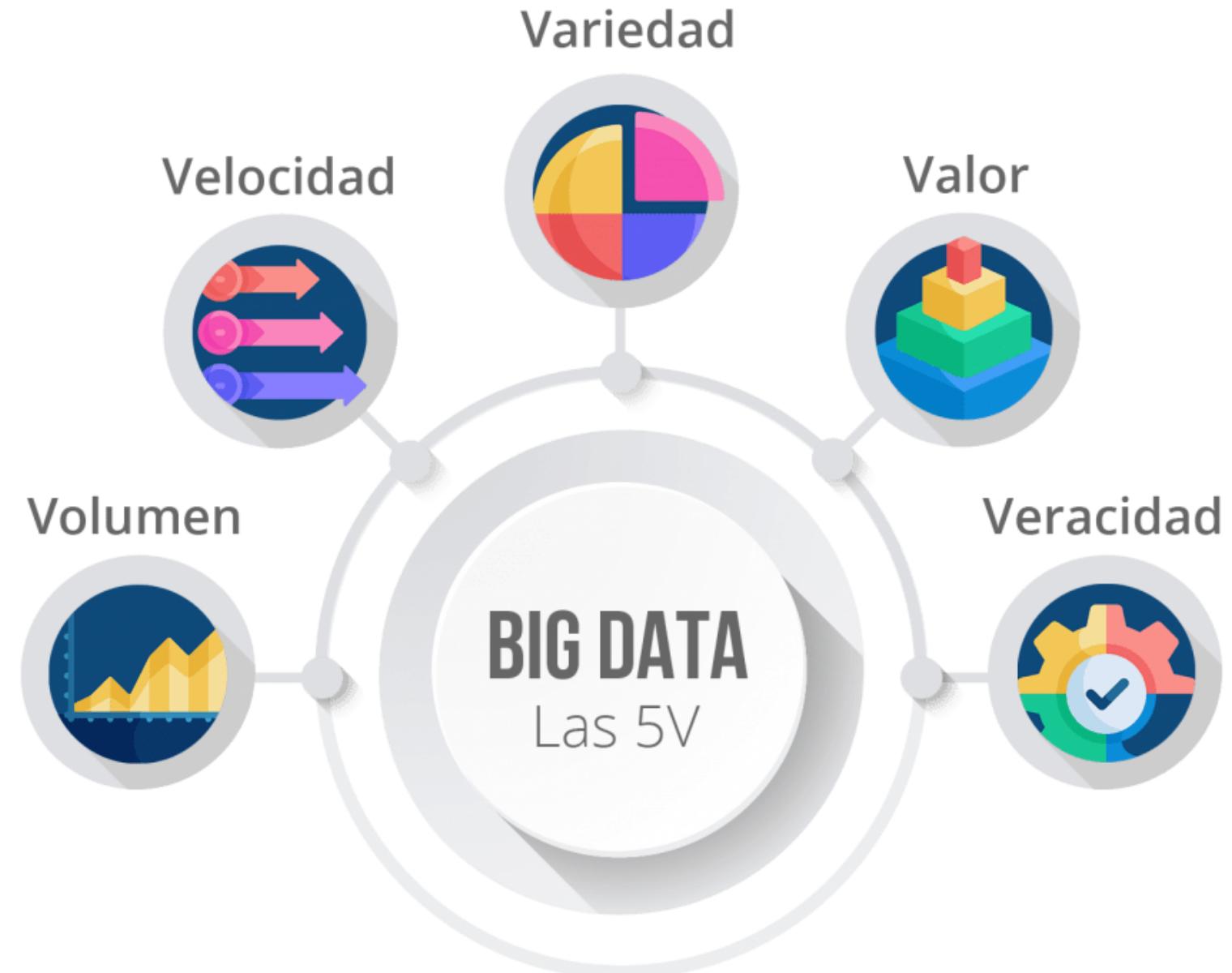
Todo tipo de datos, ya sea estructurados o no estructurados. (tablas, texto, imágenes, videos, audio) (*Variedad*)



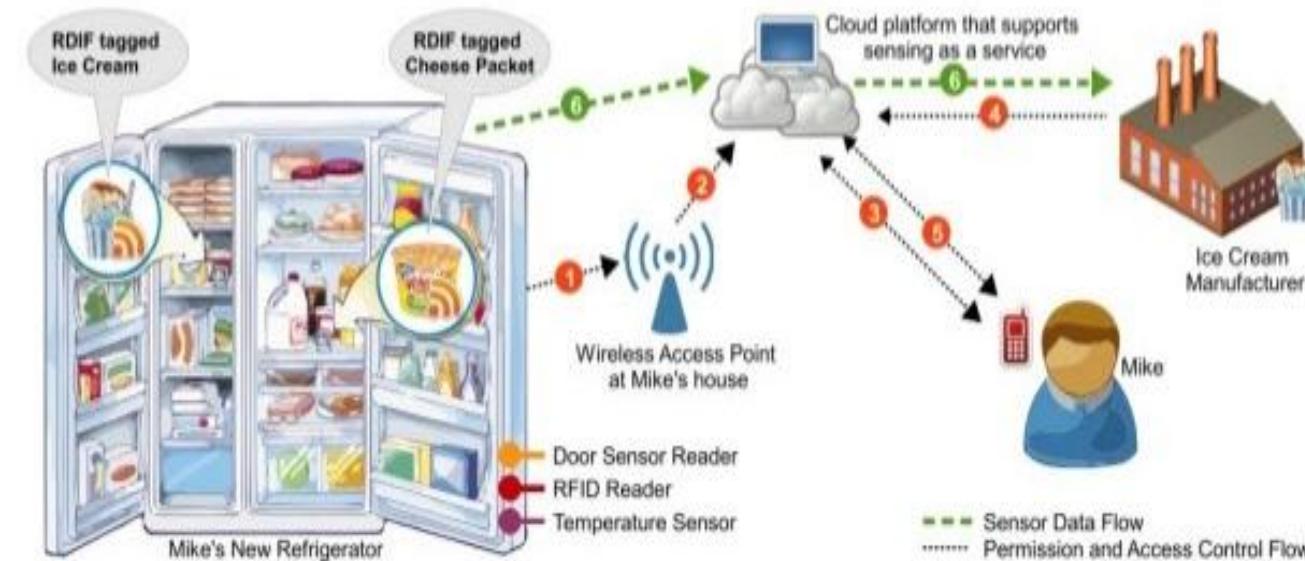
Los datos deben poder proporcionar un *valor* o beneficio a la organización.



Se debe garantizar la calidad y confiabilidad de los datos. El Big Data debe alimentarse con datos relevantes y verdaderos. (**veracidad**)

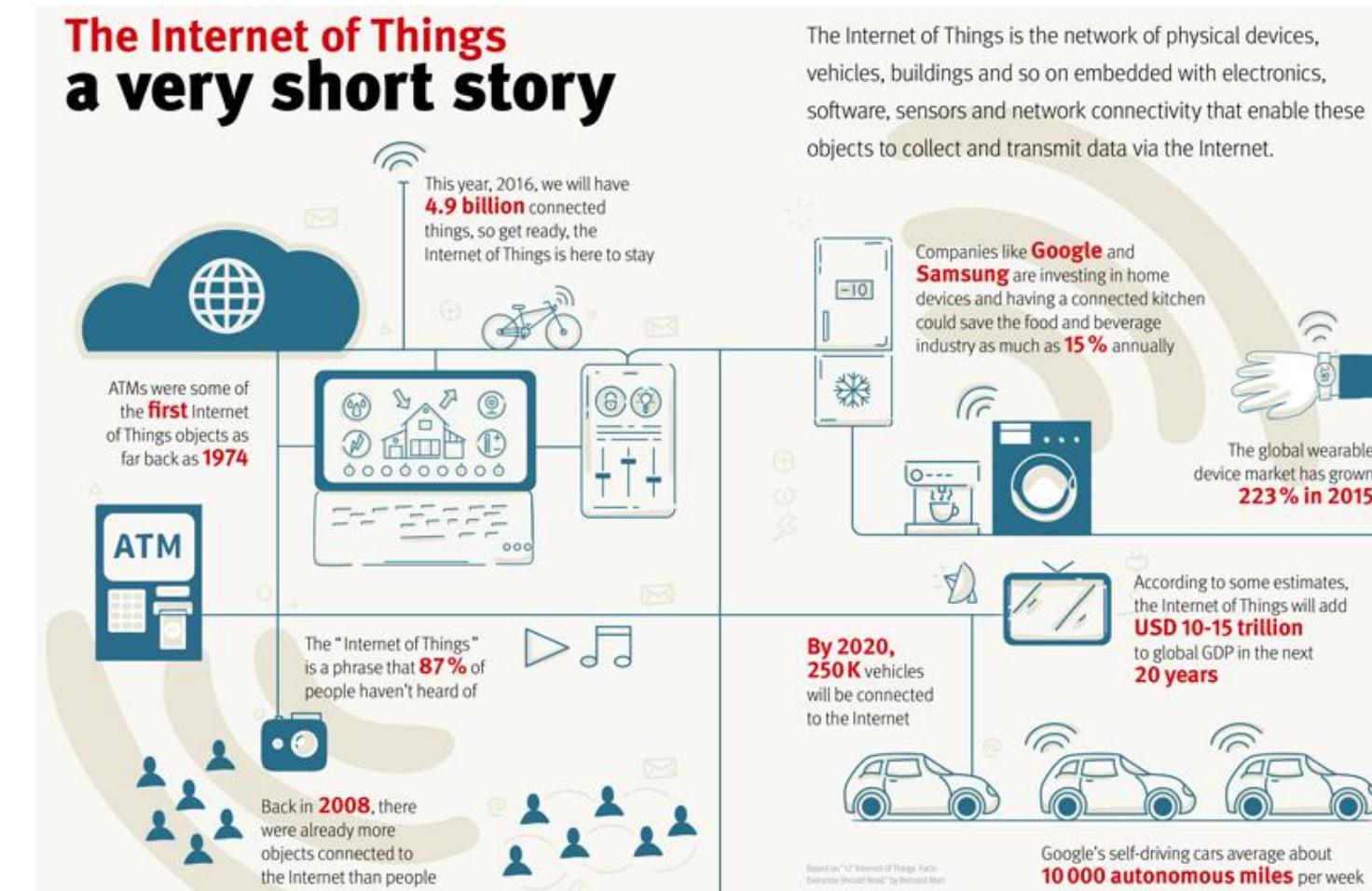


¿Qué sabemos del IoT?



[Source: "Sensing as a Service Model for Smart Cities Supported by Internet of Things", Charith Perera et. al., Transactions on Emerging Telecommunications Technology, 2014]

¿Qué sabemos del IoT?



¿Qué sabemos del IoT?

