

# ESTRUCTURAS Y FORMATOS DE DATOS



# ¿Qué son los datos?



Datos es cualquier **representación simbólica** de un atributo o variable, que describe una entidad del mundo real.

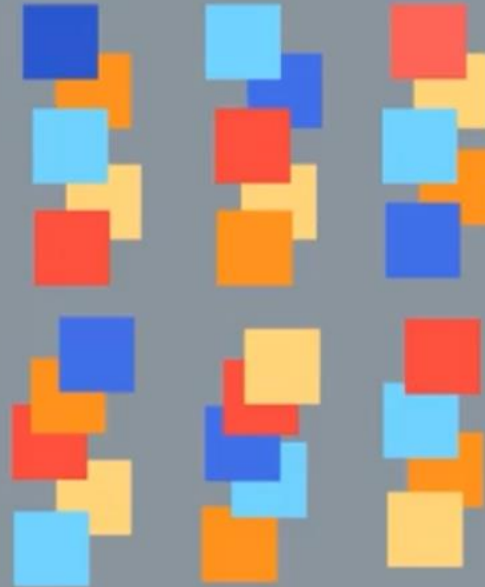


Sirve como base para un cálculo, razonamiento o discusión.

## Datos estructurados



## Datos semi-estructurados

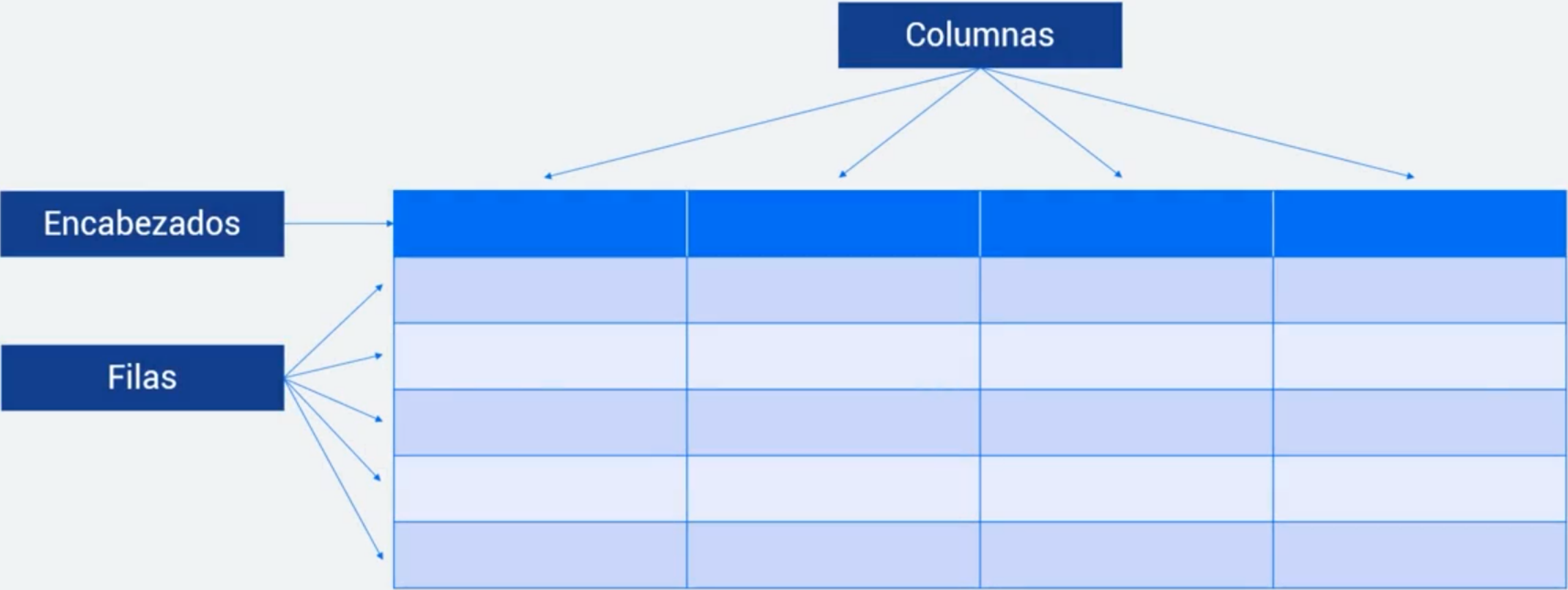


## Datos no estructurados



 **humio**  
A CrowdStrike  
Company

# Datos estructurados



# Datos estructurados: formatos

---



**Archivos de texto  
simple**



**Planillas de  
cálculo**



**Bases de datos  
relacionales**

# Ejemplo

## Tabla

Producto	Marca	CodigoSKU	Costo	PrecioVenta	Unidades
Cuaderno lineas	SVLUS	1003789	1210	1452	116
Cuaderno cuadros	SVLUS	1929835	1210	1452	165
Lapiz graffito	Lappau	1366925	470	564	194
Goma de borrar	Lappau	1461758	285	342	131
Corchetera	Officia	1835948	3250	3900	190
Pincel fino	ArtM	1992659	990	1188	139
Pincel grueso	ArtM	1959137	1100	1320	198
Acilico 6 colores	ArtM	1124149	7320	8784	45
Mezclador 6	ArtM	1954385	890	1068	138
Toner	TechOf	1301317	15230	18276	145
Resma carta	MultiOffice	1987180	4580	5496	7



# Ejemplo

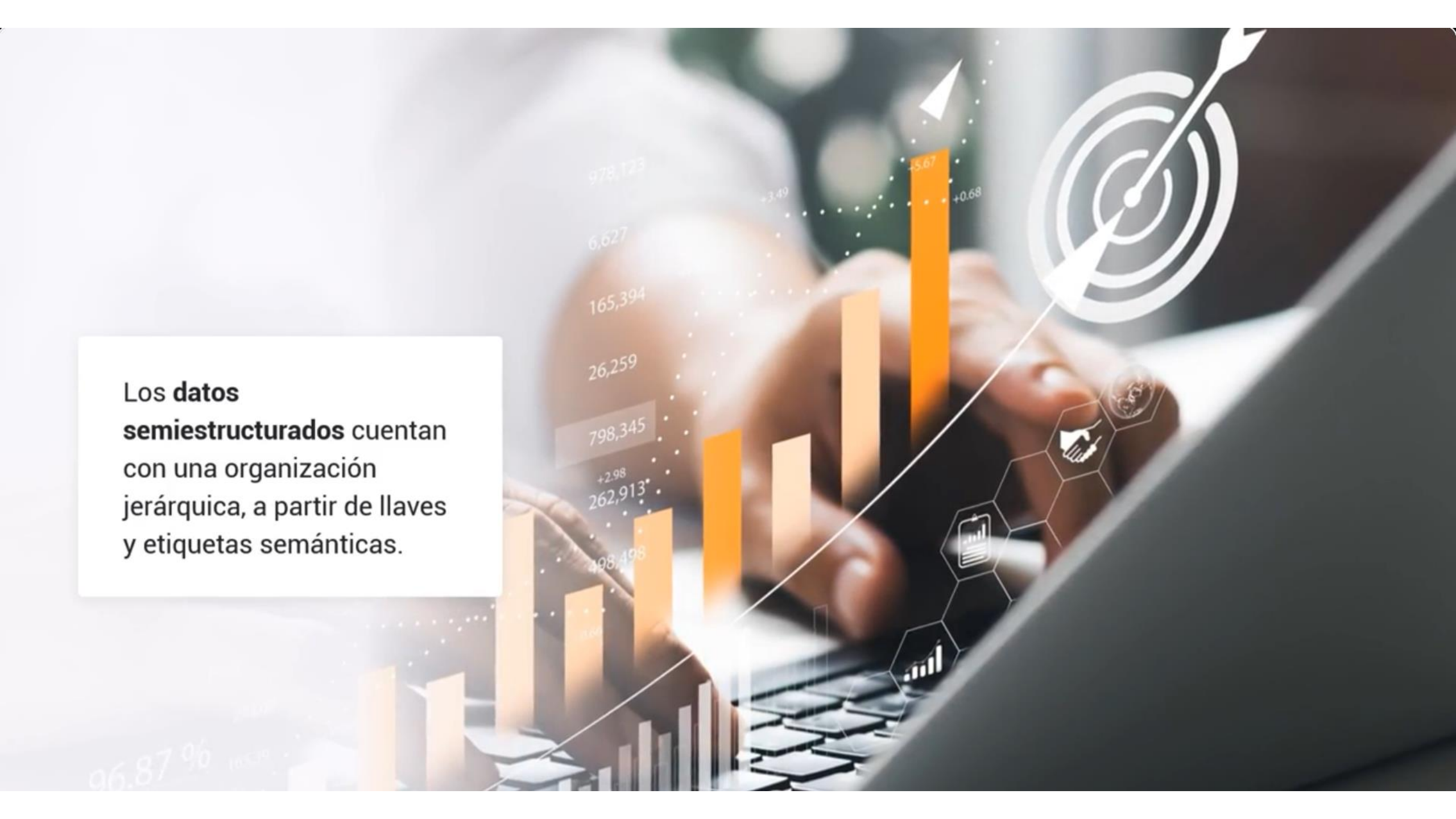
## Tabla

Producto	Marca	CodigoSKU	Costo	PrecioVenta	Unidades
Cuaderno lineas	SVLUS	1003789	1210	1452	116
Cuaderno cuadros	SVLUS	1929835	1210	1452	165
Lapiz graffito	Lappau	1366925	470	564	194
Goma de borrar	Lappau	1461758	285	342	131
Corchetera	Officia	1835948	3250	3900	190
Pincel fino	ArtM	1992659	990	1188	139
Pincel grueso	ArtM	1959137	1100	1320	198
Acrilico 6 colores	ArtM	1124149	7320	8784	45
Mezclador 6	ArtM	1954385	890	1068	138
Toner	TechOf	1301317	15230	18276	145
Resma carta	MultiOffice	1987180	4580	5496	7

## Formato .csv

```
Producto,Marca,CodigoSKU,Costo,PrecioVenta,Unidades
Cuaderno lineas,SVLUS,1003789,1210,1452,116
Cuaderno cuadros,SVLUS,1929835,1210,1452,165
Lapiz grafito,Lappau,1366925,470,564,194
Goma de borrar,Lappau,1461758,285,342,131
Corchetera,Officia,1835948,3250,3900,190
Pincel fino,ArtM,1992659,990,1188,139
Pincel grueso,ArtM,1959137,1100,1320,198
Acrilico 6 colores,ArtM,1124149,7320,8784,45
Mezclador 6 ,ArtM,1954385,890,1068,138
Toner,TechOf,1301317,15230,18276,145
Resma carta,MultiOffice,1987180,4580,5496,7
```

Los **datos semiestructurados** cuentan con una organización jerárquica, a partir de llaves y etiquetas semánticas.





# Ejemplo HTML

## Lenguaje de Marcado de Hipertexto

### Resumen de clases.

Un ejemplo sencillo de datos para conocer el formato HTML.

```
<!DOCTYPE html>
<html>
<head>
<title>Página del Curso MDS3020</title>
</head>
<body>

<h1>Resumen de clases.</h1>

<p>Un ejemplo sencillo de datos para conocer
el formato HTML.</p>

</body>
</html>
```

# Ejemplo HTML

## Lenguaje de Marcado de Hipertexto

### Resumen de clases.

Un ejemplo sencillo de datos para conocer el formato HTML.

```
<!DOCTYPE html>
<html>
<head>
<title>Página del Curso MDS3020</title>
</head>
<body>

<h1>Resumen de clases.</h1>

<p>Un ejemplo sencillo de datos para conocer
el formato HTML.</p>

</body>
</html>
```

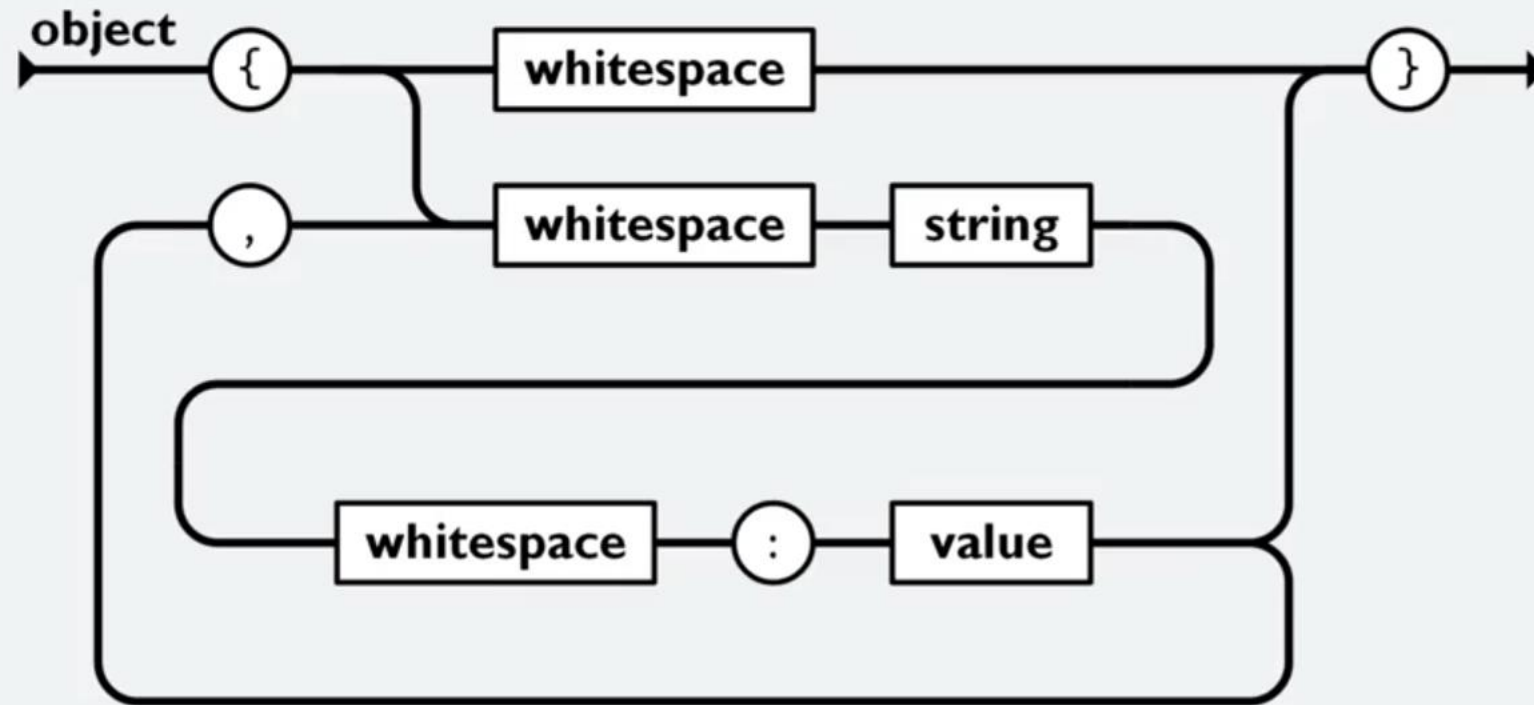
# Ejemplo formato XML

## Lenguaje de Marcado Extensible

```
<?xml version="1.0" encoding="UTF-8"?>
<datos>
<fechaBoleta>20210915</fechaBoleta>
<impuestoVenta>19%</ impuestoVenta >
<montoVenta>226192</ montoVenta >
<mailReceptor>vsolar@mail.com</mailReceptor>
<nombreReceptor>Victoria Solar
</nombreReceptor>
<numeroBoleta>275</numeroBoleta>
</datos>
```

# Ejemplo formato JSON

## JavaScript Object Notation



# Ejemplo formato JSON

## JavaScript Object Notation

```
{  
  'nombre': 'Luis',  
  'apellido': 'Perez',  
  'mail': 'luisperez@mail.com',  
  'comprasMes': 178540  
},  
{  
  'nombre': 'Victoria',  
  'apellido': 'Solar',  
  'mail': 'vsolar@mail.com',  
  'compras': {'Cuaderno cuadros':5, 'Lapiz grafito':10, 'Goma de borrar':8},  
  'telefono':'5687654325'  
}
```



# Datos no estructurados

---



No siguen una organización o jerarquía interna clara.



Pueden contener mucha información cualitativa.



Requieren de herramientas de análisis distintas como PLN.



**Ideas finales**





**Datos es cualquier representación simbólica de un atributo o variable, que describe una entidad del mundo real.**



**Existen tres formas en las que se pueden presentar: datos estructurados, semi estructurados, y no estructurados.**



# FUENTES DE DATOS



# Contenidos



Tema 1

**Clasificación de  
las fuentes de  
datos**

Tema 2

**Sesgos en los  
datos**



# Desde una pregunta de investigación

¿Qué se quiere **resolver**?

¿Qué variable se requiere estimar o predecir?



## Ejemplo:

Realizar un análisis del mercado inmobiliario, para predecir el valor de venta de una propiedad

**¿Cuáles son estas potenciales fuentes, y en qué se diferencian?**

# Desde una pregunta de investigación

---

¿Qué variables influyen en el precio por metro cuadrado de una propiedad?



## Ejemplo:

Realizar un análisis del mercado inmobiliario, para predecir el valor de venta de una propiedad

# Datos relevantes para el análisis

---



## Datos censales

Demografía



## Datos servicio de impuestos

Contribuciones y tasaciones fiscales.



## Páginas web

De corredoras inmobiliarias.



## Encuesta directa

A personas que han vendido sus propiedades.



## Datos servicios

Escuela, transporte, áreas verdes, etc.



## Datos de redes sociales

Opiniones de mejores o peores barrios.



# Clasificación por origen



## Fuentes internas

Recolectados o generados sistemáticamente por la misma entidad.



## Fuentes externas

Proveen datos relativos a otras personas u organizaciones. La red es la principal fuente.





OPEN  
DATA

## Fuentes abiertas

Permiten el acceso **libre y gratuito** a datos



Pueden ser utilizados, reutilizados y redistribuidos.

# OPEN DATA

## Fuentes abiertas

Permiten el acceso **libre y gratuito** a datos



Pueden ser utilizados, reutilizados y redistribuidos.



Cumpliendo los requerimientos de atribución e integridad de la información.

# **Ejemplos de repositorios de datos abiertos**



# Bienvenido a la Plataforma Nacional de Datos Abiertos de Colombia

Impulsamos la transparencia y la toma de decisiones basada en datos públicos. Accede y reutiliza nuestros datos sin costo.



DESCUBRE



PUBLICA



CONOCE



<https://www.datos.gov.co/>





# Fuentes privadas o de propiedad de una organización

Acceso **limitado** a usuarios autorizados





# Fuentes privadas o de propiedad de una organización

Acceso **limitado** a usuarios autorizados

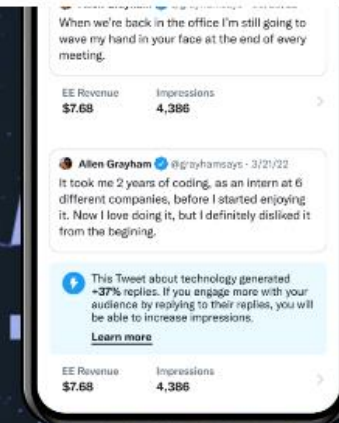
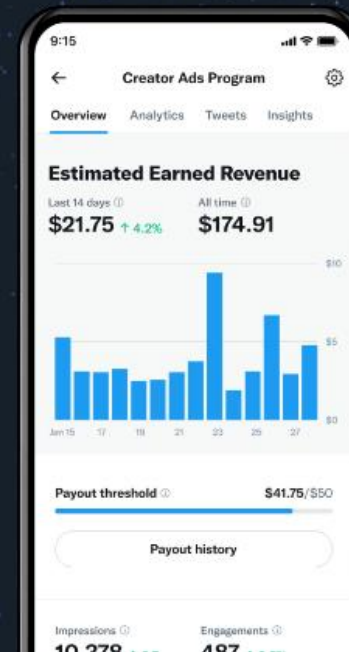


Que cumplen con algún criterio de pertenencia o de perfil de usuario.

## Ejemplos de fuentes de datos de acceso limitado



# Twitter Analytics







## Maps Datasets API

[Comenzar](#)[Comunicarse con Ventas](#)[Guías](#)[Referencia](#)[Asistencia](#) Filtrar

## Trabaja con conjuntos de datos

[Realiza una solicitud a la API](#)[Crea y modifica conjuntos de datos](#)[Antes de comenzar](#)[Cómo crear un conjunto de datos](#)[Obtén un conjunto de datos](#)[Actualiza un conjunto de datos](#)[Borra un conjunto de datos](#)

## Prácticas recomendadas

[Prácticas recomendadas para los servicios web](#)[Bibliotecas cliente](#)[Centro de arquitectura !\[\]\(241407ae374027aec4b030ca93d07b05\_img.jpg\)](#)translated by Google Se usó la [API de Cloud Translation](#) para traducir esta página.[Switch to English](#)[Página principal](#) > [Productos](#) > [Google Maps Platform](#) > [Documentación](#) > [Maps Datasets API](#)¿Te resultó útil?  

## Obtén un conjunto de datos

[Enviar comentarios](#)

Después de crear un conjunto de datos y subir datos a él, puedes usar solicitudes HTTP GET para acceder al conjunto de datos. En esta página, se describe cómo enumerar todos tus conjuntos de datos, cómo obtener información sobre un conjunto de datos específico y cómo descargar los datos de un conjunto de datos.

### Acerca de las versiones de los conjuntos de datos

Después de subir los datos correctamente, el estado del conjunto de datos se establece en `STATE_COMPLETED` y ese conjunto de datos se convierte en la versión *activa*. Eso significa que el conjunto de datos está listo para usarse en tu app. Para determinar el `state` del conjunto de datos, puedes enumerar todos los conjuntos de datos o obtener uno específico.


Apoyo

# Biblioteca API

## Notas

**Versiones compatibles** : estas versiones han sido certificadas por el equipo de productos empresariales de Bloomberg para su uso por parte de los clientes de Bloomberg.



An illustration of a diverse crowd of stylized human figures. A large, dark grey magnifying glass is positioned over the center of the image, focusing on a woman with short blonde hair wearing a dark blue t-shirt and a red pleated skirt. The background is a light grey gradient with faint, larger-scale patterns. Other figures in various outfits are scattered around, some partially visible or faded.

**Debemos identificar y  
sopesar los sesgos en  
nuestro análisis y  
conclusiones.**

## Población objetivo

Los datos corresponden  
a una muestra de la  
población o conjunto  
completo de los eventos  
a estudiar.

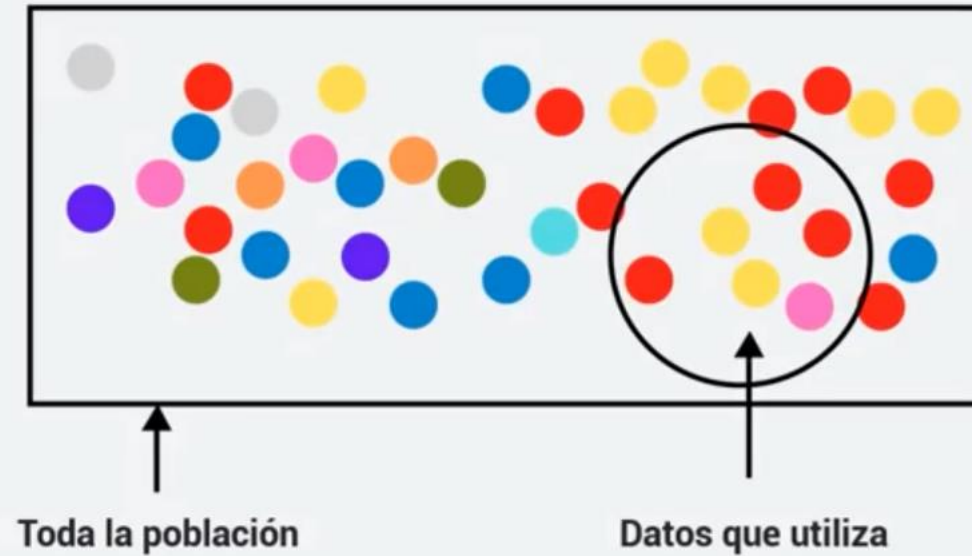


Ejemplo

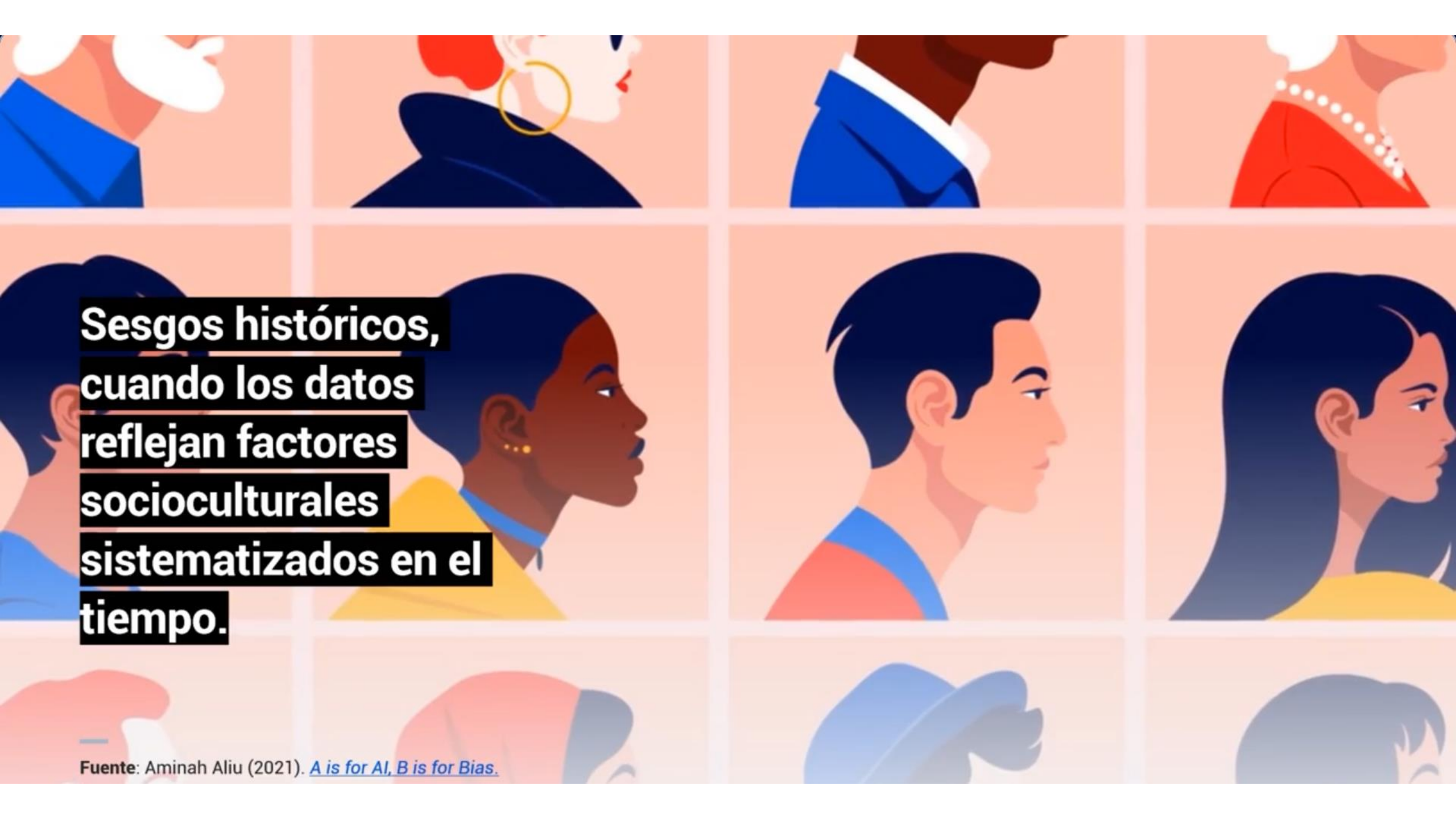
# SESGO DE REPRESENTACIÓN

- Cuando la muestra obtenida de la población, no es representativa.

# Sesgos de representación



La mayoría son rojos o  
amarillos



**Sesgos históricos,  
cuando los datos  
reflejan factores  
socioculturales  
sistematizados en el  
tiempo.**



Población objetivo

**Los sesgos de medición  
emergen de la forma en  
que elegimos, usamos y  
medimos ciertas  
variables.**




Ejemplo




**El sesgo de agregación se produce cuando se usan datos de una población para extraer conclusiones sobre un individuo en particular.**




A stylized illustration of a person with short, wavy purple hair. They are wearing a blindfold made of a dark blue band with white binary code (0s and 1s) on it. The background is a large, soft orange circle on a light grey background.

**Sesgo de omisión,  
cuando una o más  
variables importantes  
son excluidas de los  
datos.**



Suponga que una cierta consultora política realiza un modelo predictivo para la próxima elección presidencial, a partir del análisis de contenido de cerca de un millón de comentarios y mensajes publicados por los usuarios nacionales en la red social Twitter. ¿Cuál es el principal tipo de sesgo que debería tenerse en consideración al utilizar este conjunto de datos?

- Sesgo de agregación
- Sesgo de medición
- Sesgo político
- Sesgo de representación



Suponga que una cierta consultora política realiza un modelo predictivo para la próxima elección presidencial, a partir del análisis de contenido de cerca de un millón de comentarios y mensajes publicados por los usuarios nacionales en la red social Twitter. ¿Cuál es el principal tipo de sesgo que debería tenerse en consideración al utilizar este conjunto de datos?

- Sesgo de agregación
- Sesgo de medición
- Sesgo político
- Sesgo de representación



## ALGUNAS LECTURAS ...

- **A Survey on Bias and Fairness in Machine Learning**  
([Enlace](#))
- **Doing Data Science: A Framework and Case Study**  
([Enlace](#))

