

**DE CERO A CIENCIA DE DATOS**

# **CIENCIA DE DATOS**

# AGENDA

- Inteligencia Artificial y Aprendizaje Automático (ML)
- Análisis de Datos
- Ciencia de Datos
- Científicos de Datos
- Proceso que sigue un Científico de Datos
- Modelos
- Términos más usados
- El lado humano
- Para tener un mejor proceso
- Ejemplo rápido :D

# INTELIGENCIA ARTIFICIAL Y APRENDIZAJE AUTOMÁTICO



# ANÁLISIS DE DATOS

## ■ Por su definición:

**El análisis de datos es un proceso que consiste en inspeccionar, limpiar y transformar datos con el objetivo de resaltar información útil, para sugerir conclusiones y apoyo en la toma de decisiones.**



# CIENCIA DE DATOS

## ■ Por definición:

**La ciencia de datos es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados, lo cual es una continuación de algunos campos de análisis de datos como la estadística, la minería de datos, el aprendizaje automático, y la analítica predictiva.**



# CIENTIFICO DE DATOS

**Se les define como una mezcla de estadísticos, computólogos y pensadores creativos, con las siguientes habilidades:**

- Recopilar, procesar y extraer valor de las diversas y extensas bases de datos.
- Imaginación para comprender, visualizar y comunicar sus conclusiones a los no científicos de datos.
- Capacidad para crear soluciones basadas en datos que aumentan los beneficios, reducen los costos.
- Los científicos de datos trabajan en todas las industrias y hacen frente a los grandes proyectos de datos en todos los niveles.



# EL PROCESO QUE SIGUE UN CIENTÍFICO DE DATOS:

1. **Extraer datos**, independientemente de la fuente y de su volumen.
2. **Limpiar los datos**, para eliminar lo que pueda sesgar los resultados.
3. **Procesar los datos** usando métodos estadísticos como inferencia estadística, modelos de regresión, pruebas de hipótesis, etc.
4. **Diseñar experimentos** adicionales en caso de ser necesario.
5. **Crear visualizaciones gráficas** de los datos relevantes de la investigación



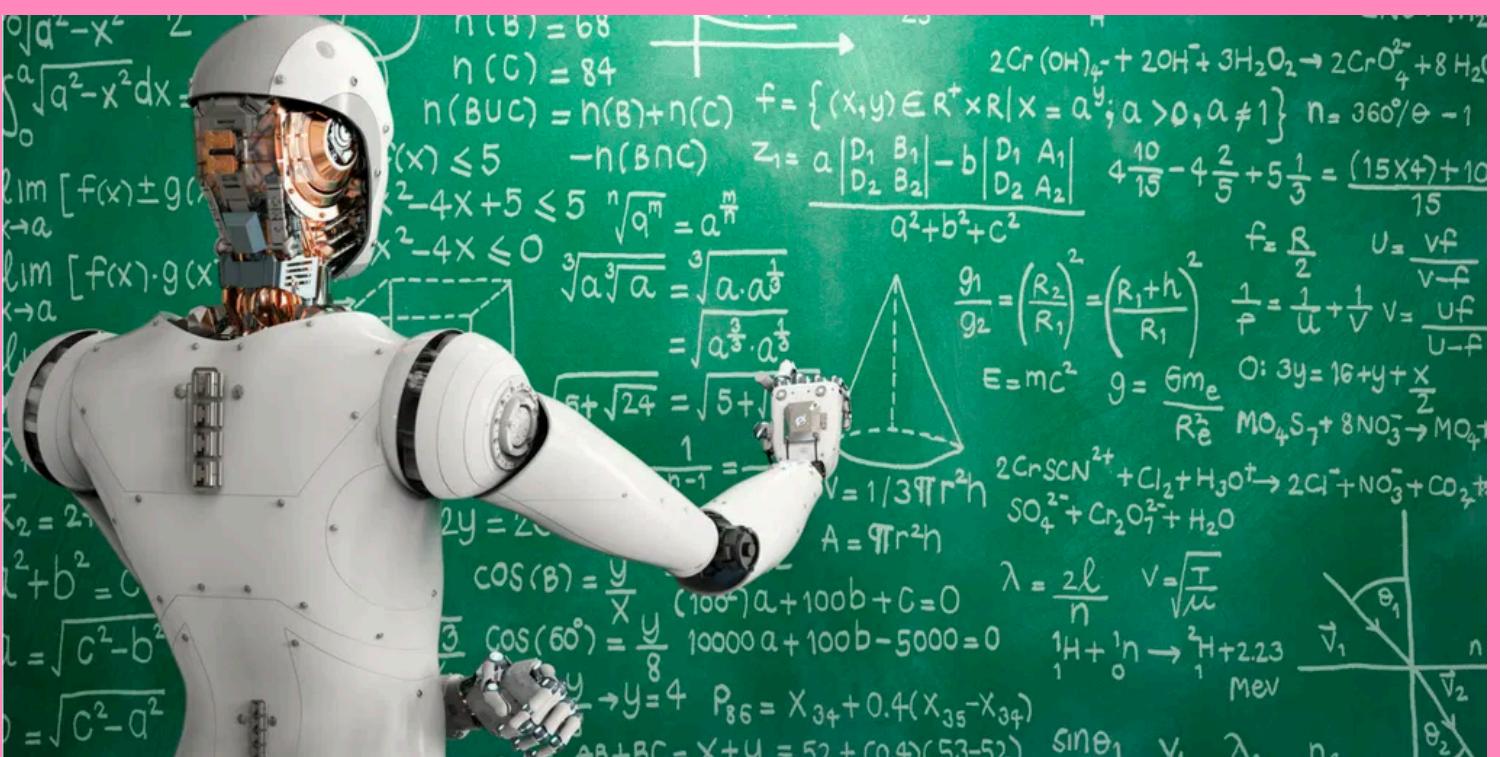
## Data Engineer



## Machine Learning Engineer



## Data Scientist



## Data Analyst



# MODELOS

- En ciencias aplicadas y en tecnología, un modelo matemático es uno de los tipos de modelos científicos que emplea algún tipo de formulismo matemático para expresar relaciones, proposiciones sustantivas de hechos, variables, parámetros, entidades y relaciones entre variables de las operaciones, para estudiar comportamientos de sistemas complejos ante situaciones difíciles de observar en la realidad.

A blue-toned graphic containing mathematical equations and geometric diagrams. The equations include:  
1.  $y = a \cdot x^2 + b \cdot x + c$   
2.  $x_{1/2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$   
3.  $V = \frac{\pi \cdot r^2 \cdot h}{3}$   
The background features faint geometric drawings of circles, triangles, and a coordinate system.

Label, Target, Objetivo

## Terminos más usados

Variables

Precios Casas			
M2	#cuartos	Precio	
100	1	10000	
200	2	20000	
300	3	X	



Dataset



Datapoints



# EL LADO HUMANO

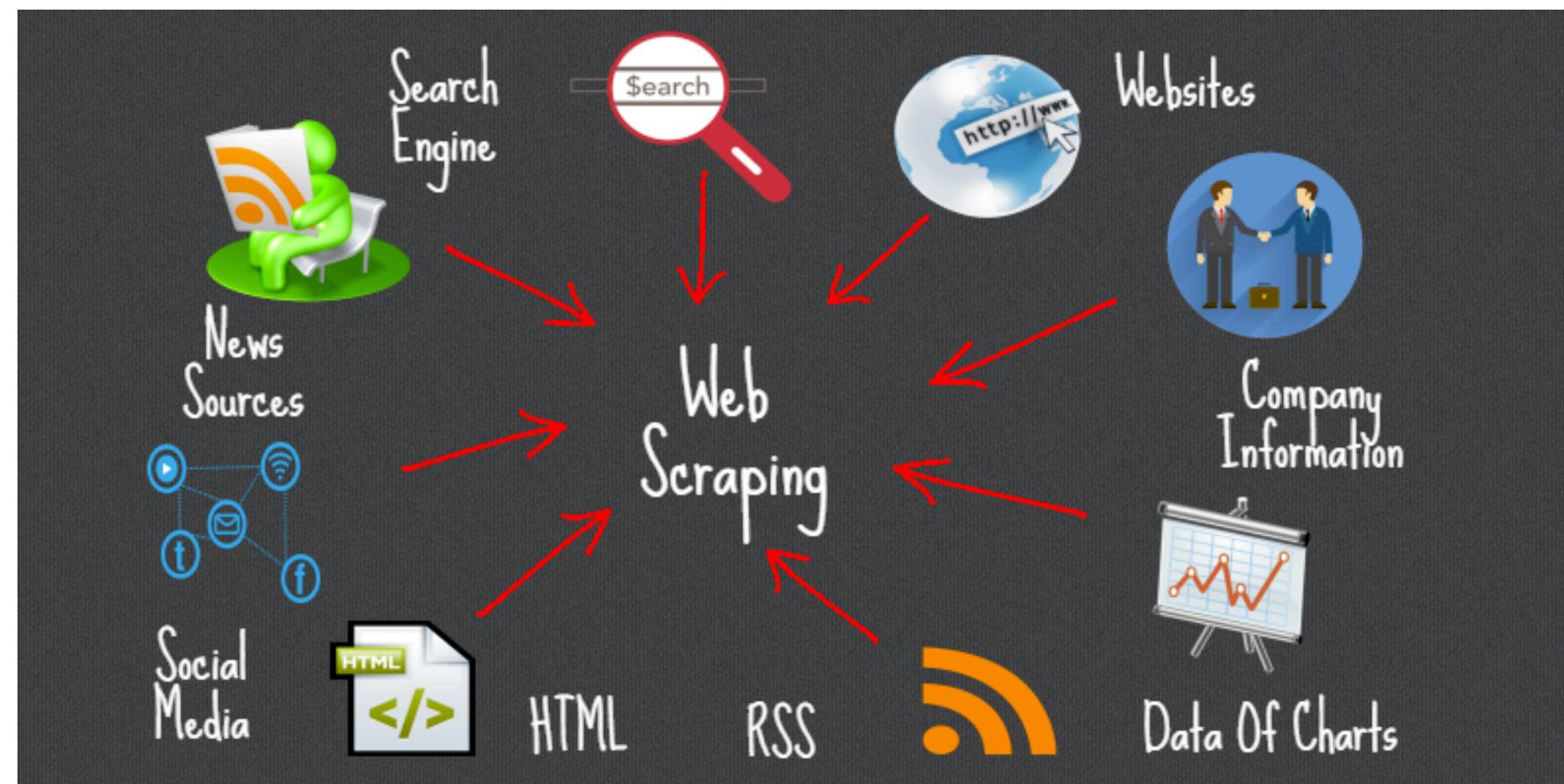
**NO ES MAGIA...**

# EXTRACCIÓN DE DATOS

Hoy en día existen numerosas herramientas o procesos mediante los cuales se pueden extraer datos de formatos complejos como un PDF o bien de una o varias páginas web, lo que se conoce como web scraping.

**El objetivo es tener los datos para poder visualizar y entender.**

**Web scraping** es la técnica con la que se es capaz de rascar, escrapear o liberar datos de páginas web de gobiernos, instituciones públicas u organizaciones para acceder a datos privados o públicos que puedan ser publicados o distribuidos en formato abierto.



# EXTRACCIÓN DE DATOS

El problema es que la mayoría de los datos de interés están en formatos **no reutilizables y poco transparentes como un PDF, por ejemplo.**

Para acceder y distribuir este tipo de información existe una gran cantidad de herramientas o procesos mediante el uso de lenguajes de programación.



Google  
Sheets

ImportHTML



Table Capture

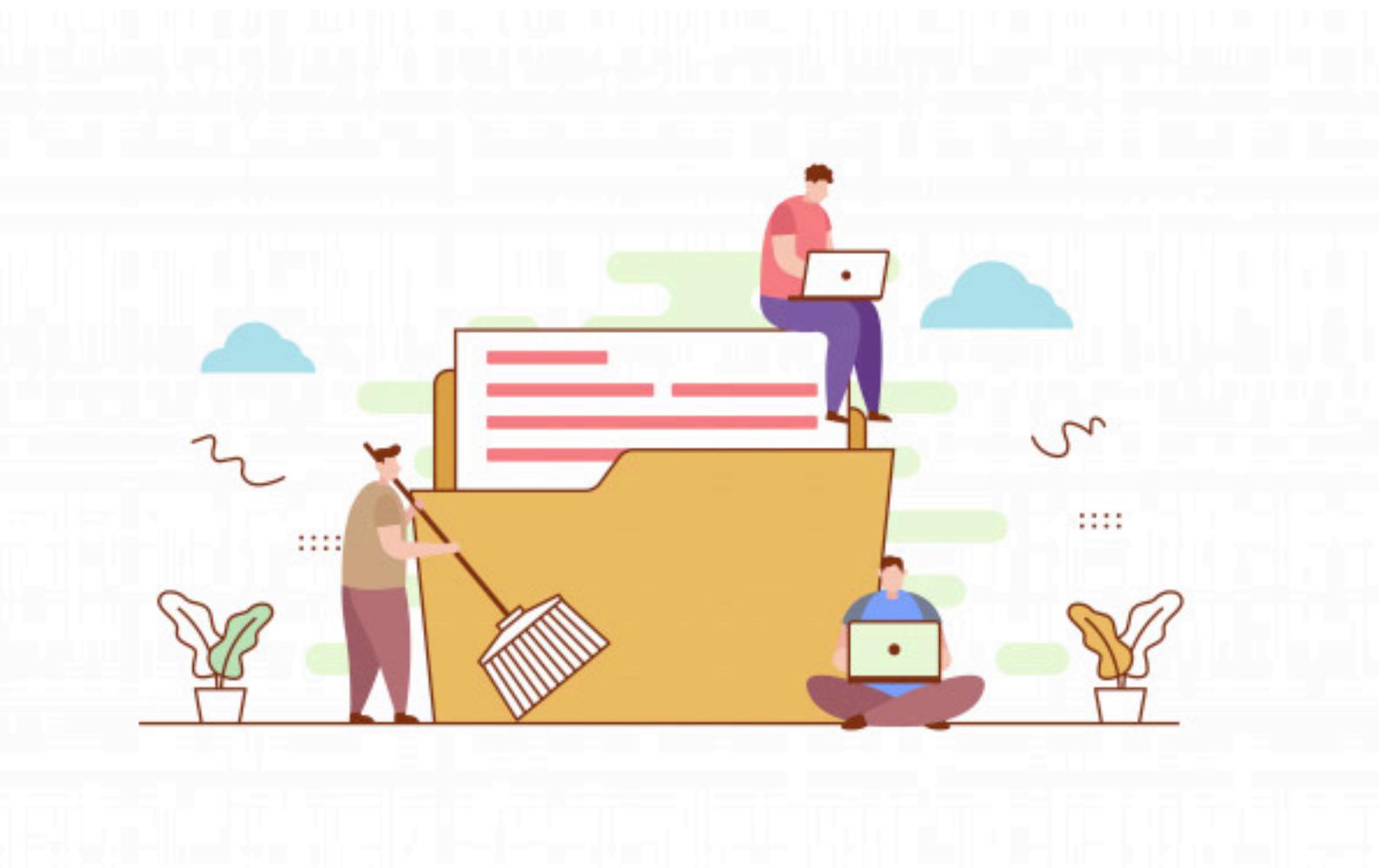
Ofrecido por: [www.georgemike.com](http://www.georgemike.com)



# LIMPIEZA DE DATOS

Este es uno de los pasos que pueden llegar a tomar más tiempo en el proceso, en este paso se puede considerar lo siguiente:

- Aplicar reglas de unificación de datos: Por ejemplo, poner en la fila correspondiente al sexo la misma letra identificativa, como podría ser “M” para masculino y “F” para femenino. En este caso, también se tendrían que identificar o corregir posibles errores, como que algún usuario haya puesto la “M” como mujer.
- Validaciones: Como por ejemplo, comprobar que en todos los registros de datos de los clientes de un banco esté introducida la dirección postal completa, saltando una alarma si falta alguno.



# LIMPIEZA DE DATOS

- Estandarización de datos: El objetivo es que todos los datos del mismo tipo estén introducidos de idéntica forma. Un ejemplo sería el RFC con homoclave.
- Normalización: El objetivo es que todos los datos numéricos estén en la misma escala lo cual es necesario en el uso de ciertos algoritmos.
- Categorización: Algunos modelos funcionan mejor con datos categóricos y podemos agrupar datos numéricos, así en lugar de tener las edades independientes, empezamos a clasificar en grupos: “De 18 a 25”, “De 26 a 30”, etc.



# LIMPIEZA DE DATOS

— ¿Qué pasa cuando los datos si existen pero no están etiquetados, clasificados?

The grid contains the following templates:

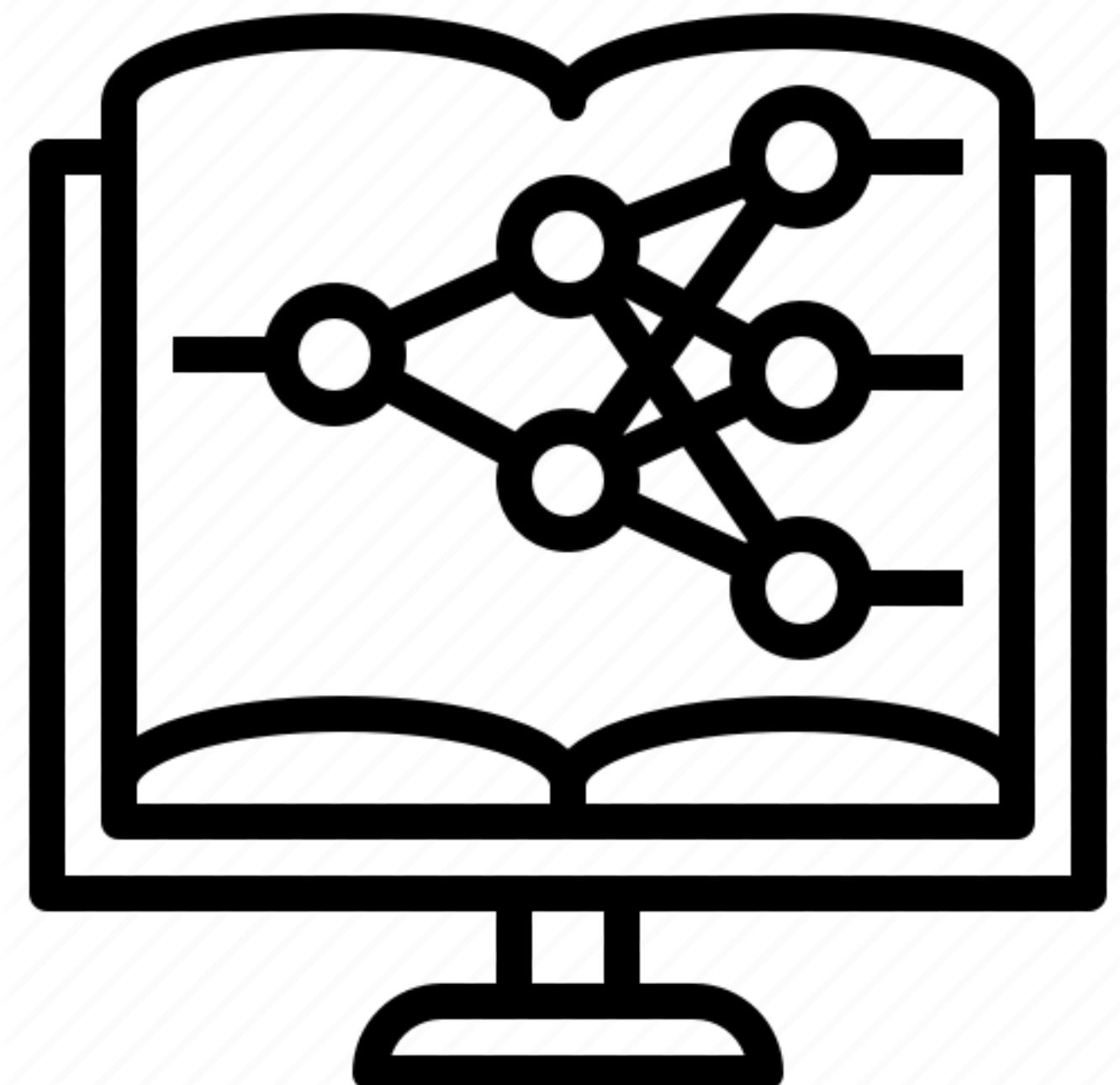
- Finding Company Website and Description** (Data Mining)
- Website Classification** (Categorization)
- Image Annotation (Hair Segmentation)** (Data Annotation)
- Sequencing of Events** (Categorization)
- Collect Contact Information (Spreadsheet)** (Data Mining)
- Video Comparison** (Content Moderation)
- Image Annotation (Facial Spots Detection)** (Data Annotation)
- Match Destinations** (Content Moderation)
- Building Words Database** (Data Mining)
- Take a Survey (Dynamic - 5 pages)** (Survey)
- Translation of Sentences** (Translation)
- Dresses Classification** (Categorization)
- Tweet Emotion** (Sentiment Analysis)
- Product Tagging (Open Fridge Inventory)** (Image Tagging)
- Image Transcription (Handwritten Name)** (Transcription)
- Ads Monitoring Task** (Content Moderation)
- Label the Emotion of Message in a Short Conversation** (Sentiment Analysis)
- Choose Preferred Meta Description** (Survey)



Amazon Mechanical Turk

# PROCESAMIENTO DE DATOS

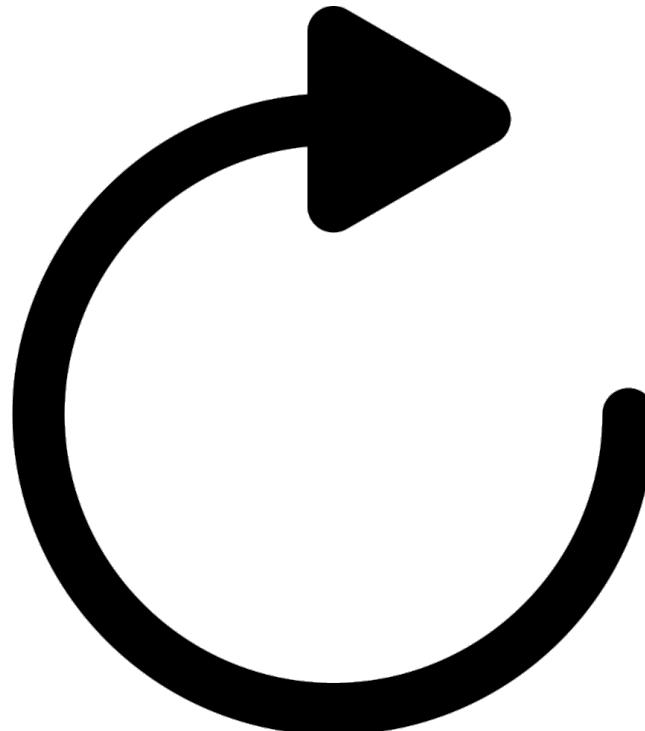
1. Limpieza de datos
2. Selección y partición de instancias
3. Ajustes de los atributos
4. Transformación de la representación
5. Extracción de atributos
6. Selección de atributos
7. Creación de atributos
8. Entrenamiento del modelo



# DISEÑO DE EXPERIMENTOS ADICIONALES

¿Funciona a la primera? ¿Qué pasa si no funciona a la primera?

1. Limpieza de datos
2. Selección y partición de instancias
3. Ajustes de los atributos
4. Transformación de la representación
5. Extracción de atributos
6. Selección de atributos
7. Creación de atributos
8. Entrenamiento del modelo



# CREAR VISUALIZACIONES GRÁFICAS

**Una gráfica o una representación gráfica o un gráfico, es un tipo de representación de datos, generalmente cuantitativos, mediante recursos visuales, para que se manifieste visualmente la relación matemática o correlación estadística que guardan entre sí.**



# **PARA TENER UN MEJOR PROCESO:**

- 1. ASEGUREMONOS DE TENER BIEN DEFINIDOS TANTO LOS REQUERIMIENTOS COMO EL OBJETIVO DEL MODELO.**
- 2. ES UN TRABAJO INTERDISCIPLINARIO, SI TRABAJAMOS CON UN EXPERTO EN EL TEMA DE LOS DATOS QUE ESTAMOS UTILIZANDO PUEDE HACER MÁS RÁPIDO TODO EL PROCESO, ASÍ COMO TRABAJAR CON DISEÑADORES DE EXPERIENCIA E INTERFACES MEJORA EL PROCESO.**
- 3. NUNCA OLVIDARNOS DE NUESTRO PROPIO SESGO**
- 4. PREPARACIÓN MENTAL PARA MUCHO TRABAJO MANUAL**
- 5. ES IMPORTANTE TENER EN CUENTA EL CONTEXTO GENERAL PARA SABER DIFERENCIAR LO QUE SE PUEDE MEDIR DE LO QUE NO SE PUEDE.**

**GRACIAS!**