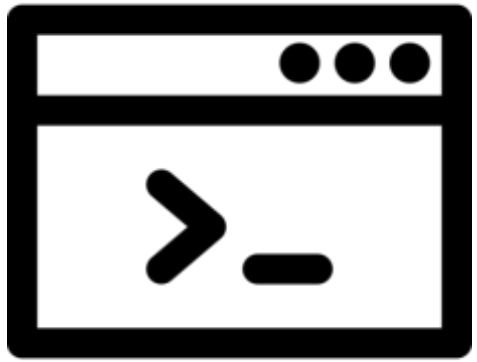
CREATIC

# CENTRO DE DESARROLLO TECNOLÓGICO

# 2.

# Uso de los datos en DL



CREATIC

# ¿Qué es la ciencia de datos?

## ¿Qué es un dato?

Cualquier información generada por una acción.

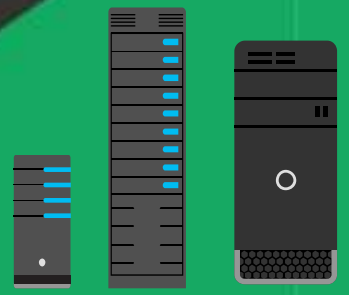
## ¿Por qué se crea la ciencia de datos?

- Gran cantidad de datos generados y almacenados.
- Necesidad de entender los datos.
- Desconocimiento de como interpretar datos.

“La extracción, exploración y análisis de datos estructurados y no estructurados, con el fin de entender, desarrollar, adquirir nuevo conocimiento y formular resultados accionables.”

# ¿De donde proviene los datos?

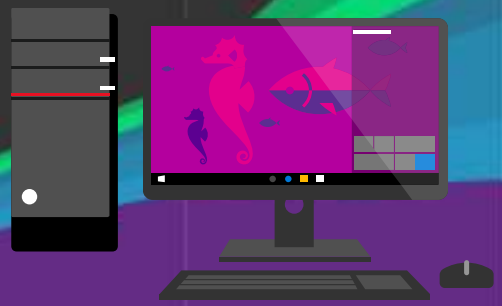
Internet
Digital
Análogo



Nube

30 años

Móvil

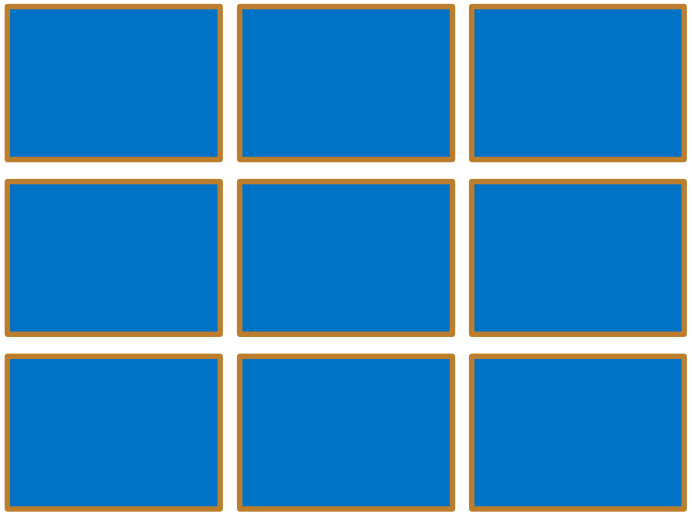


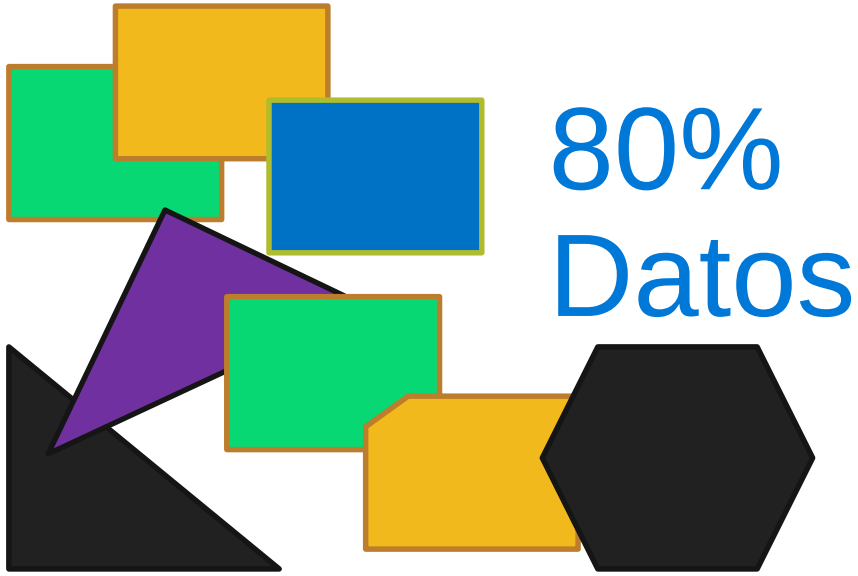
PC

1985 1990 1995 2000 2005 2010 2015 2020

# Tipo de Datos

| Estructurados | No Estructurados |
| --- | --- |

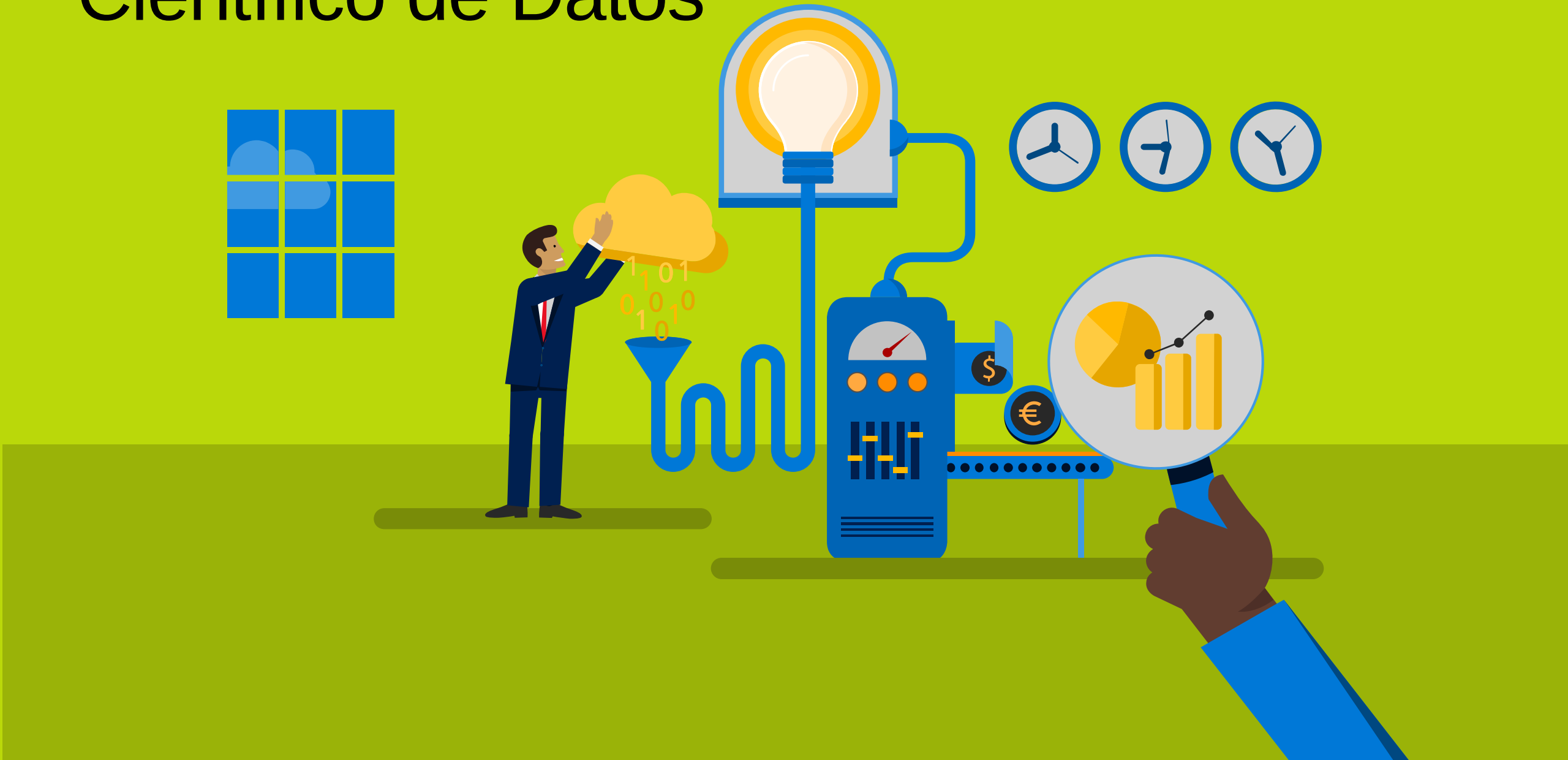




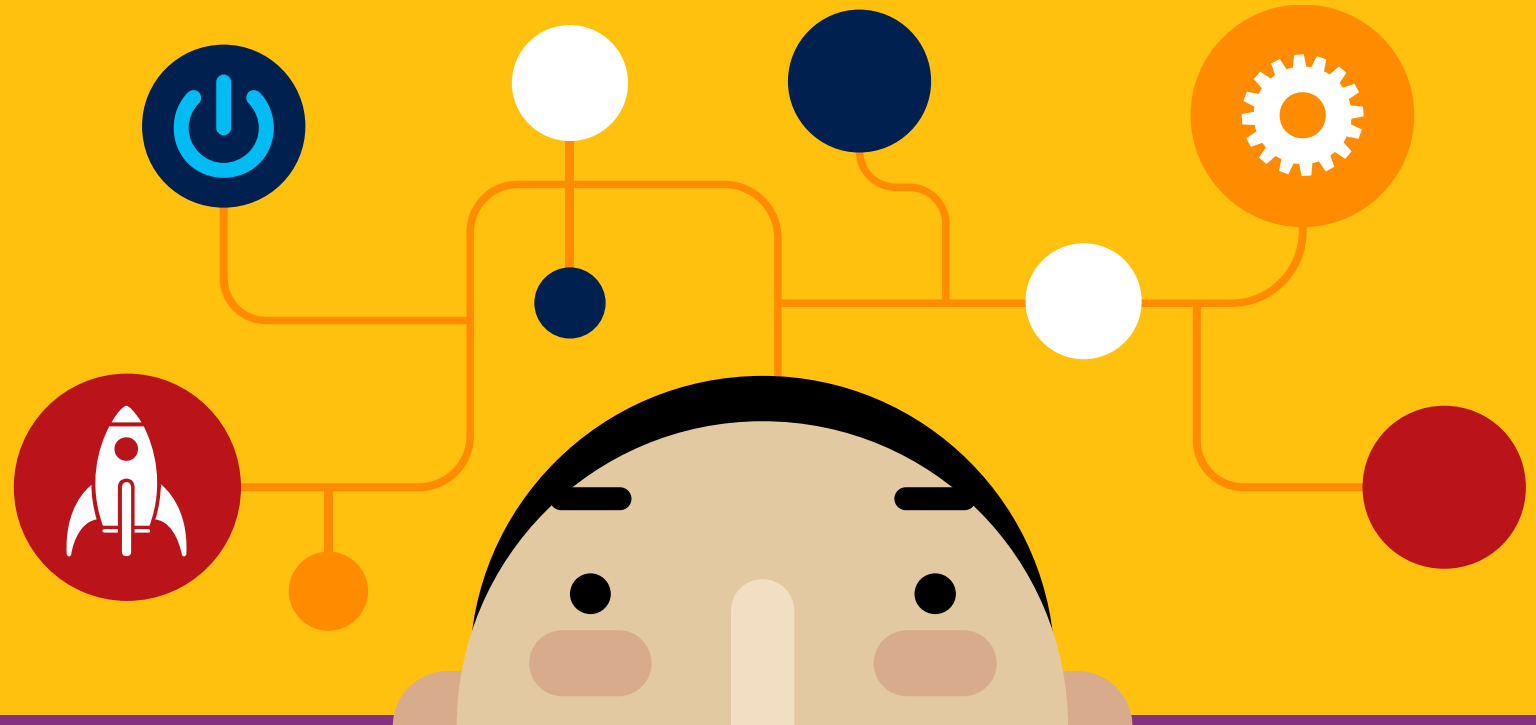
80%
Datos

# ¿Para que sirven los datos?

# Científico de Datos



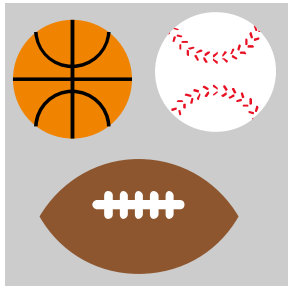
*El científico de datos, es una atractiva combinación entre visión de liderazgo, conocimientos en ciencias de la computación, estadística y habilidad de identificar respuestas entre masivos cúmulos de información.*



# ¿Para que se usa Machine Learning?

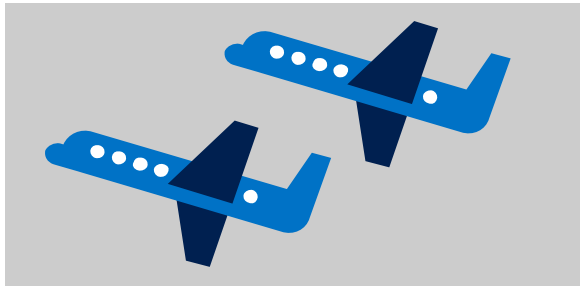
## Spam

Controlar los mensajes no deseados de nuestro correo electrónico



## Elecciones

Predecir quien ganará las elecciones electorales de un país.



Telemetría y Análisis

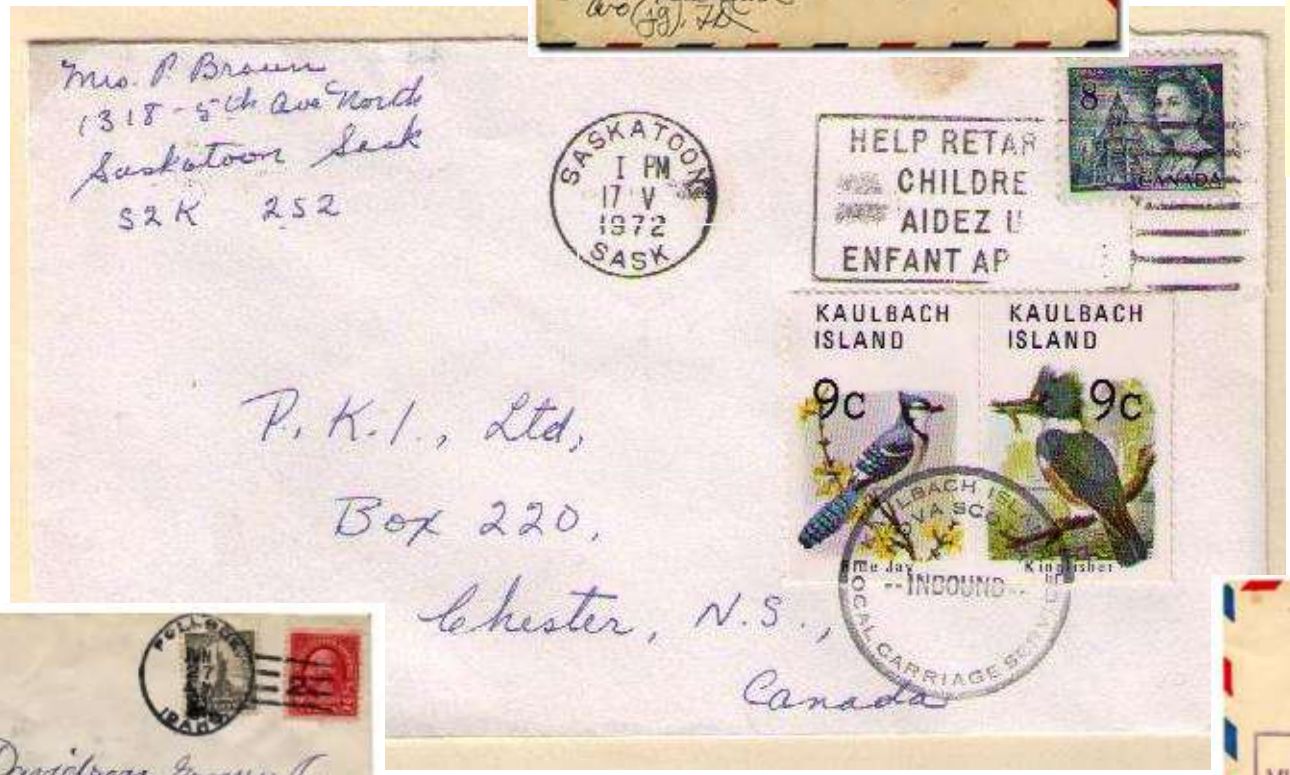
Red Social


181
RICHUR
Registered
Mr. Eugene Zimmerbach
18, 17 South Broadway
42251
St Louis Mo.
U.S.A.


PASSED BY
U.S. ARMY POSTAL SERVICE
Mr George A Hughes
1135 Clarendon Crescent
Oakland
Calif.

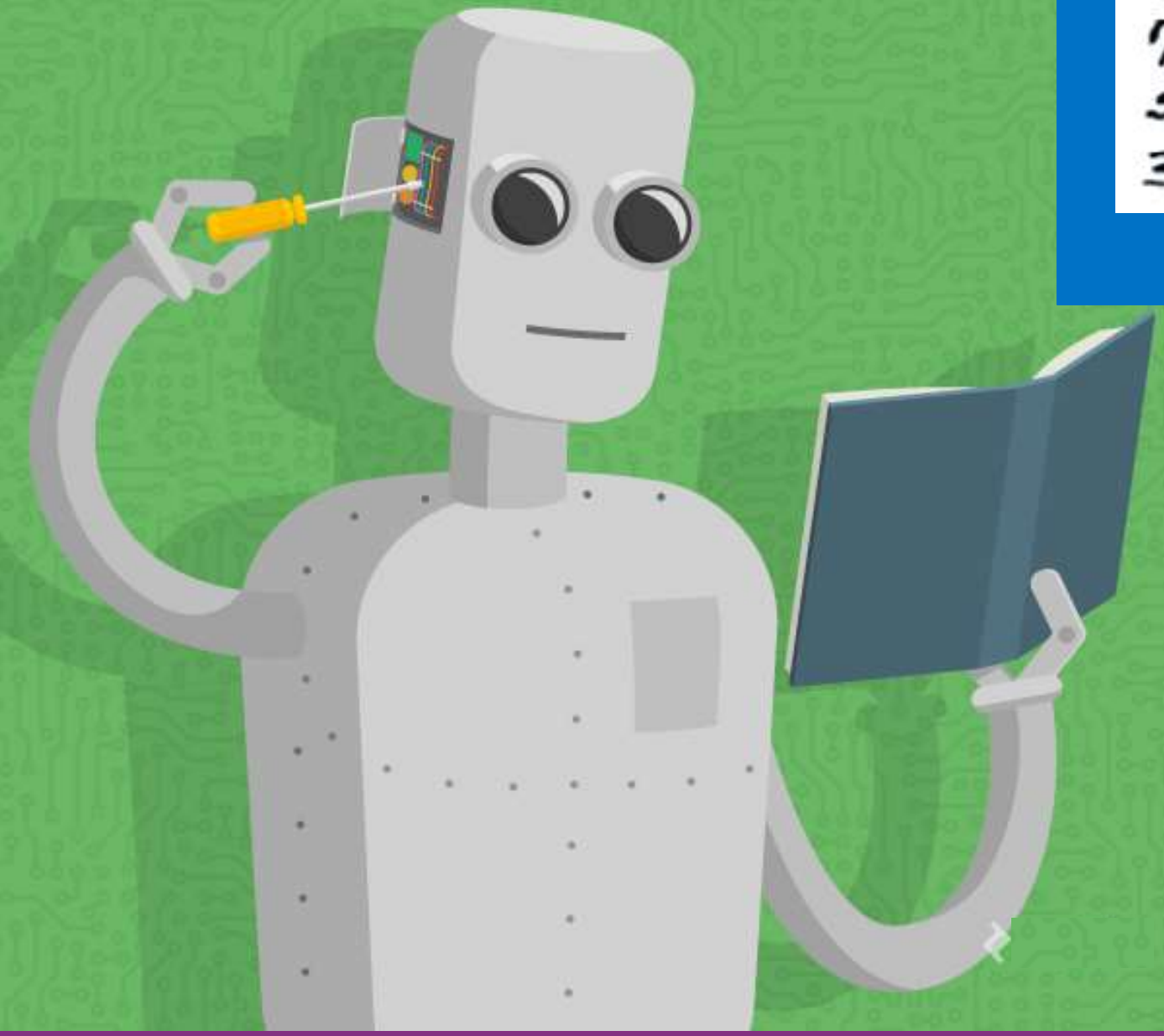

W.C.
D. Fabian, Esq.
208 Columbia Street
Portland
ORE.

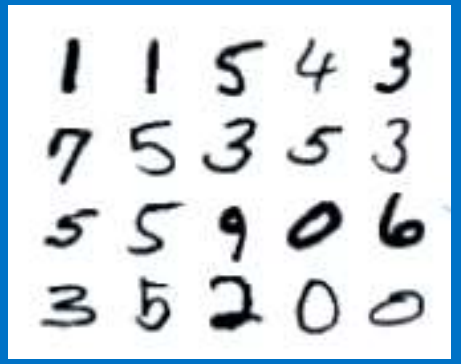

W.Moore, Esq.
Brockport,
New York.


Mrs. P Braun
1318 - 5th Ave North
Saskatoon Sask
S2K 252
SASKATOON
1 PM
17 V
1972
SASK
HELP RETAR
CHILDRE
AIDEZ L
ENFANT AP
CANADA
KAULBACH ISLAND
KAULBACH ISLAND
9c
9c
P. K. I., Ltd,
Box 220,
Chester, N.S.,
Canada
--INDOUND--


H. J. POTTER
POLLOCK, IDAHO
GENERAL STORE
We can Supply you
POLLOCK
IDAHO
REGISTERED
No. 172
Davidson Grocery Co.
Boise,
9391
Idaho.


VIA AIR MAIL
AIR MAIL
VIA NYRBA
Par Avion
By Air Mail

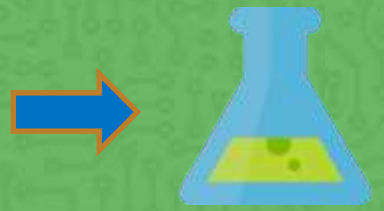




Entrenamiento

| | | | | |
|---|---|---|---|---|
| 1 | 1 | 5 | 4 | 3 |
| 7 | 5 | 3 | 5 | 3 |
| 5 | 5 | 9 | 0 | 6 |
| 3 | 5 | 2 | 0 | 0 |

Parámetros



2

# Trabajando con datos

- Relevantes
- Conectados
- Precisos
- Suficientes
- Contestan preguntas definidas

## Irrelevantes

| Precio entrada al cine | # equipos de fútbol | % contaminación |
|---|---|---|
| 4.00 | 4 | 84.0 |
| 3.50 | 2 | 1.7 |
| 4.00 | 1 | 0.2 |
| 4.50 | 3 | 11.7 |

## Relevantes

| Automotores | Ton/año | % contaminación |
|---|---|---|
| 4449 | 9270.6 | 84.0 |
| 511 | 187.4 | 1.7 |
| 396 | 18.4 | 0.2 |
| 164 | 1286.6 | 11.7 |

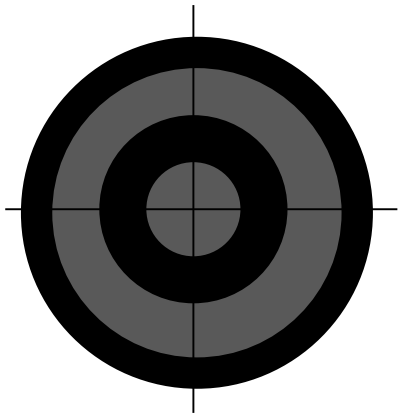
## Desconectados

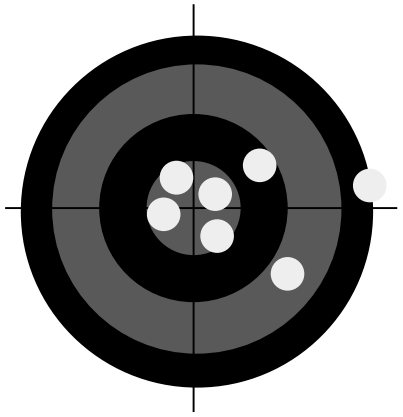
| Automotores | Ton/año | % contaminación |
|---|---|---|
| | 9270.6 | 84.0 |
| 511 | | 1.7 |
| | 18.4 | 0.2 |
| 164 | 1286.6 | 11.7 |

## Conectados

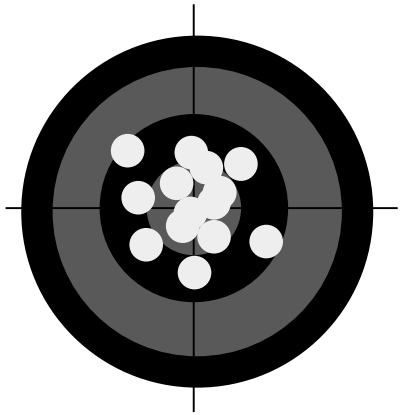
| Automotores | Ton/año | % contaminación |
|---|---|---|
| 4449 | 9270.6 | 84.0 |
| 511 | 187.4 | 1.7 |
| 396 | 18.4 | 0.2 |
| 164 | 1286.6 | 11.7 |

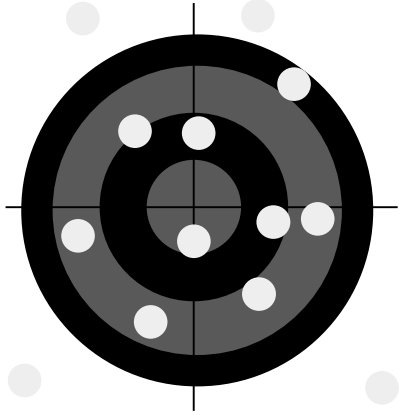
## No Precisos





## Precisos

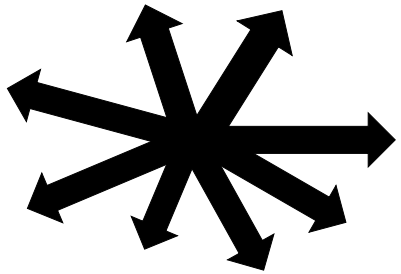




Insuficiente

Suficiente

Pregunta abierta vs Pregunta definida



No puede ser contestada con un número o un nombre

Puede ser contestada con un número o un nombre

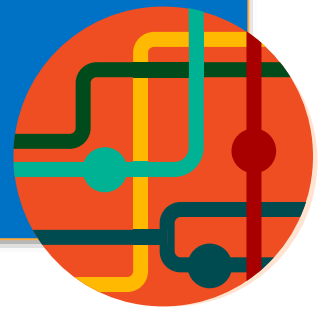
# Tipos de Machine Learning

- Predicción.
- Se dispone de los valores de los datos.
- Modelo entrenado para predecir datos.

Supervisada

- Identificar clúster de datos.
- Encontrar el valor de los datos.
- Obtener clúster de datos del modelo.

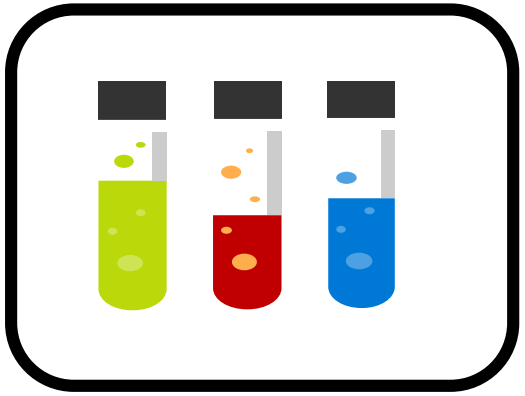
No Supervisada



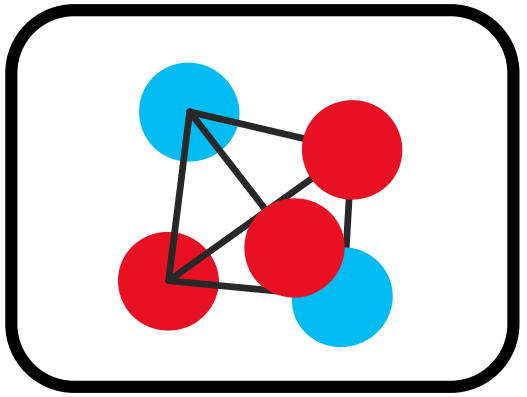
# Rol del algoritmo

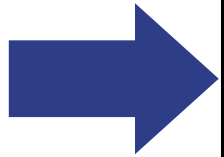
## Entrenar

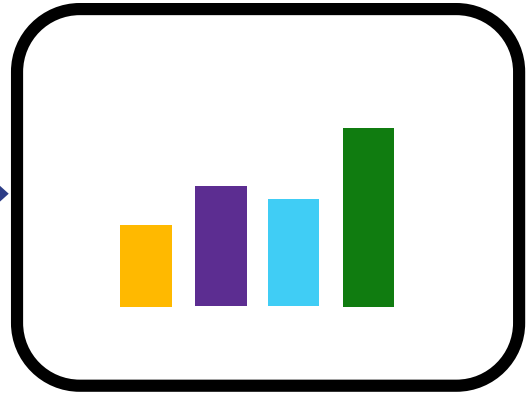




## Predecir

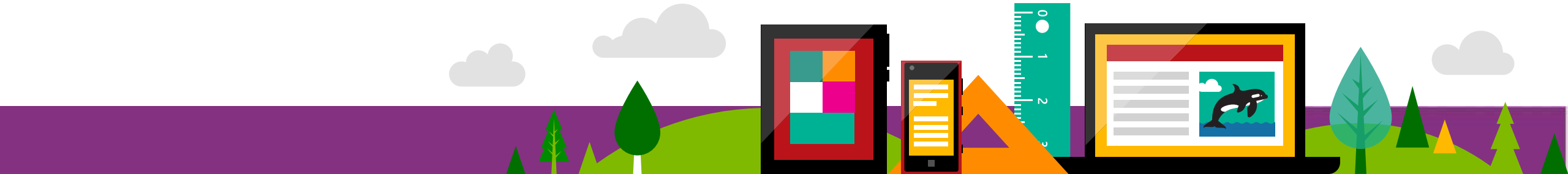






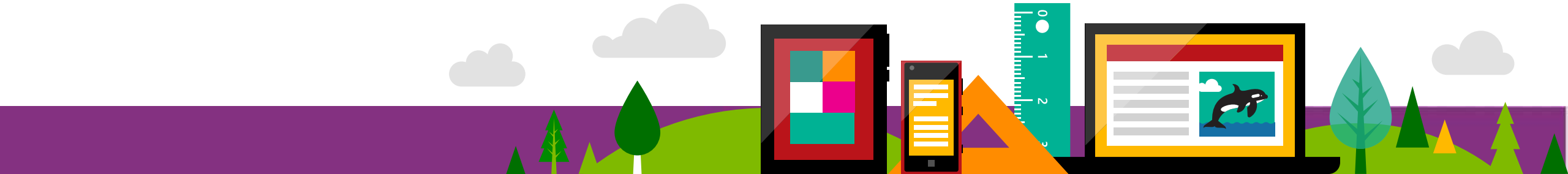
# Tipos de algoritmo

- Clasificación
- Detección de anomalías
- Regresión
- Clusterización
- Reforzar Aprendizaje



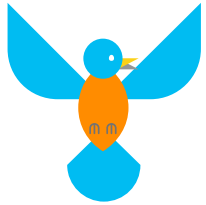
# Preguntas a contestar

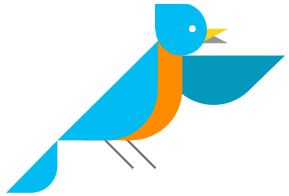
- ¿Es A o B?
- ¿Qué está fuera de lo común?
- ¿Cuanto o cuantos?
- ¿Cómo está organizado?
- ¿Qué debo hacer después?



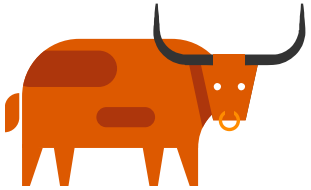
¿Es A o B?

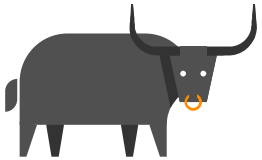
A





B

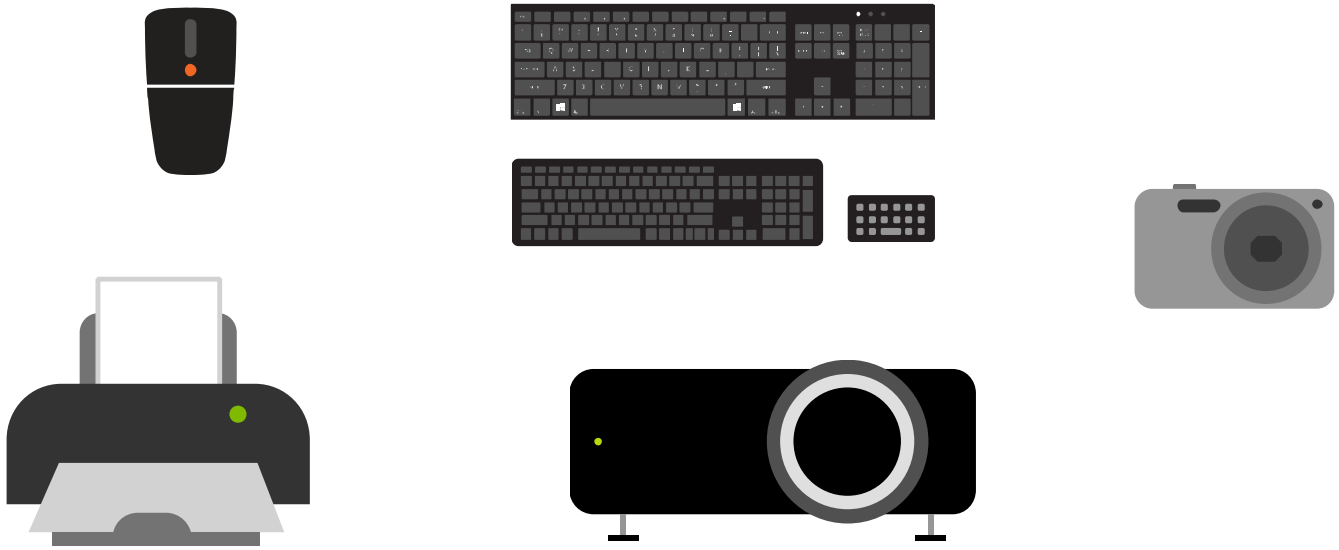




C D F G

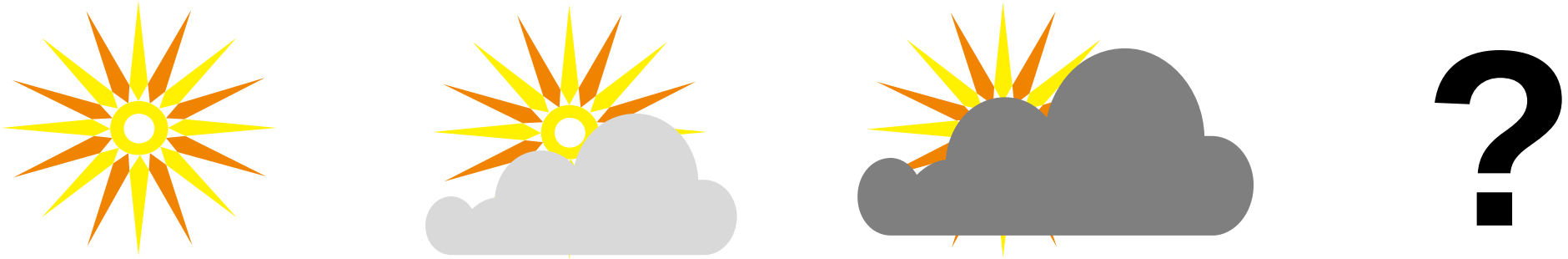
# Clasificación

# ¿Qué está fuera de lo común?



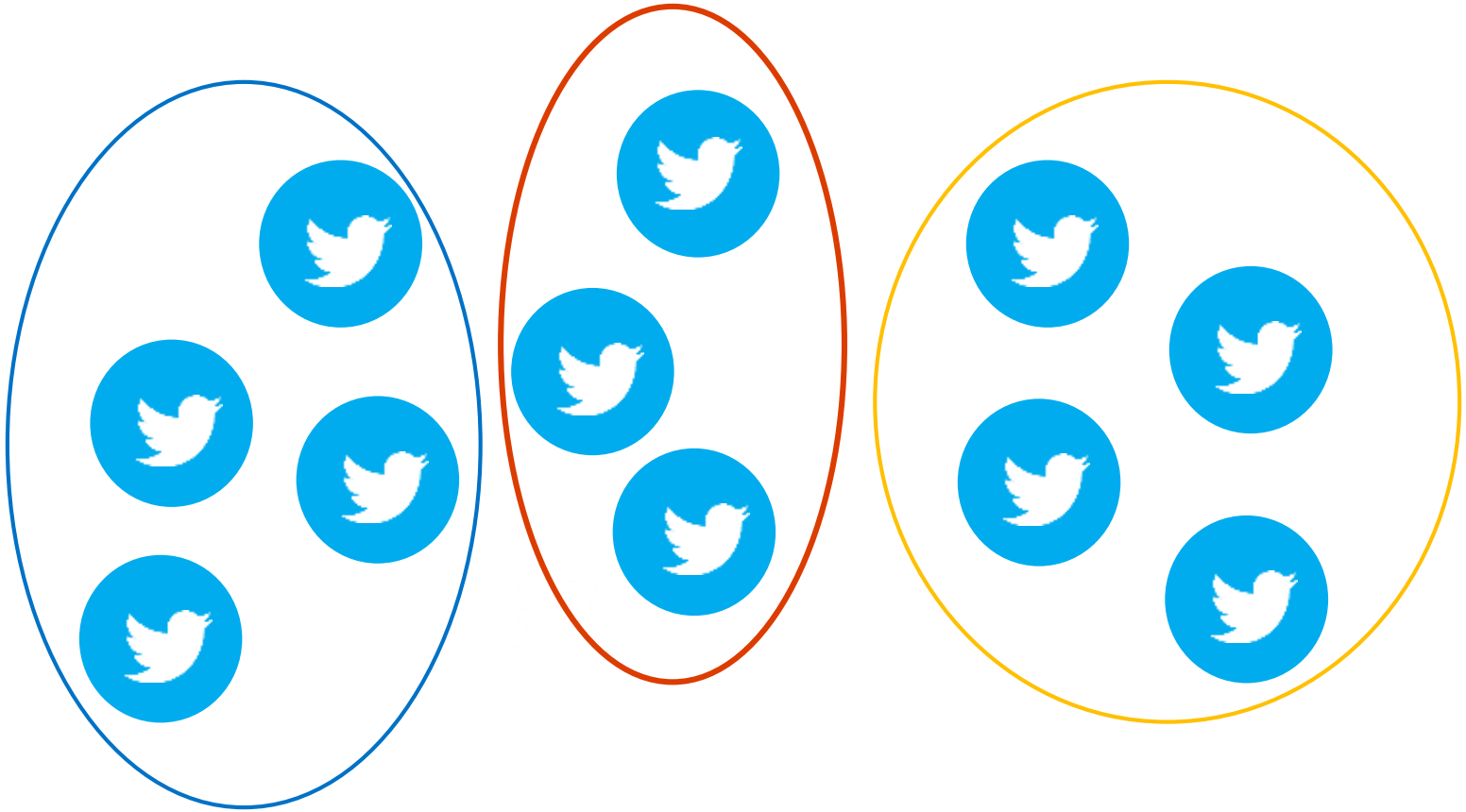
# Detección de anomalías

# ¿Cuanto o cuantos?



## Regresión

¿Cómo está organizado?

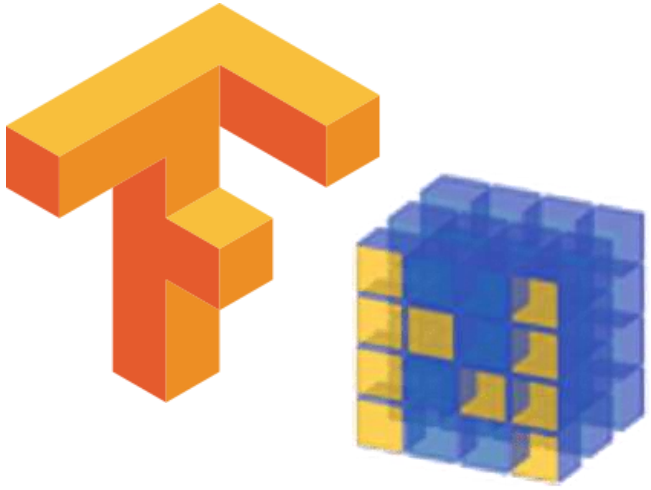


# Clusterización

# Configura tu ambiente de trabajo

- Crea tu proyecto en *Github*

- Usar *virtualenv* para instalar las librerías en Python
  - pip install virtualenv
  - virtualenv deep_env
  - source deep_env /Scripts/activate
  - cd deep_learning_project
  - pip list
  - pip install -r requirements.txt
  - pip list



# Herramientas

- Librerías de algebra lineal y análisis de datos
  - NumPy (http://www.numpy.org/)
  - Pandas (https://pandas.pydata.org/)

- Frameworks de Machine Learning (redes neuronales y optimizadores matemáticos)
  - Tensorflow (https://www.tensorflow.org/)
  - Scikit-learn (http://scikit-learn.org/)


# Jupyter

- Agrega el kernel de Jupyter al virtualenv
  - /> ipython kernel install --user --name=deep_env
  - /> Jupyter notebook

# Repositorios

enigma public

**https://public.enigma.com**

kaggle

**https://www.kaggle.com**

El futuro digital es de todos | MinTIC

**https://www.datos.gov.co**

# Cargando datos en Jupyter Notebook


CREATIC

# Convirtiendo datos categóricos a numéricos

- Get_dummies
- One hot encoding

| | name | country_australia | country_germany | country_korea | country_russia |
|---|---|---|---|---|---|
| 0 | josef | 0 | 0 | 0 | 1 |
| 1 | michael | 0 | 1 | 0 | 0 |
| 2 | john | 1 | 0 | 0 | 0 |
| 3 | bawool | 0 | 0 | 1 | 0 |
| 4 | klaus | 0 | 1 | 0 | 0 |

| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 44 | 72000 |
| 1 | 0 | 0 | 1 | 27 | 48000 |
| 2 | 0 | 1 | 0 | 30 | 54000 |
| 3 | 0 | 0 | 1 | 38 | 61000 |
| 4 | 0 | 1 | 0 | 40 | 63777.8 |
| 5 | 1 | 0 | 0 | 35 | 58000 |
| 6 | 0 | 0 | 1 | 38.7778 | 52000 |
| 7 | 1 | 0 | 0 | 48 | 79000 |
| 8 | 0 | 1 | 0 | 50 | 83000 |
| 9 | 1 | 0 | 0 | 37 | 67000 |

# Titanic: Data process from Disaster