

# BigDataCo

2020

## Introducción al análisis de datos

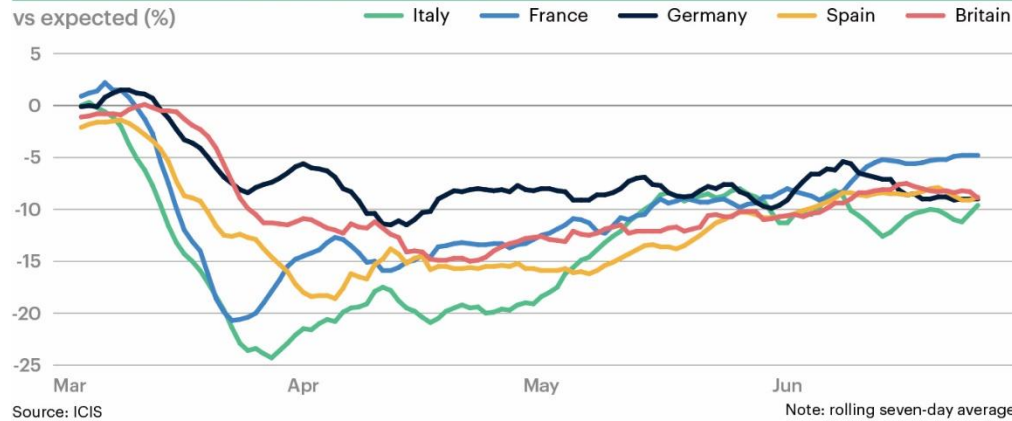
Andrés García-Suaza  
[andres.garcia58@eia.edu.co](mailto:andres.garcia58@eia.edu.co)

*Dic 5 de 2020*

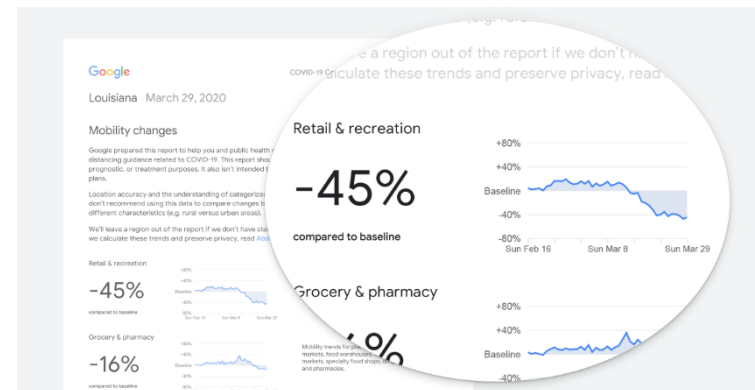
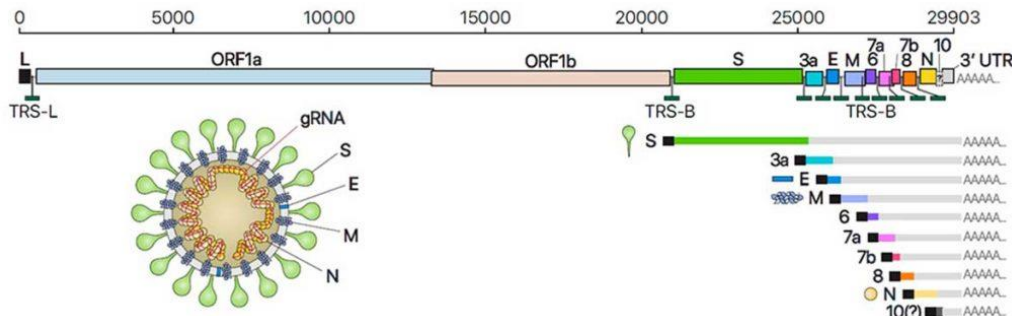
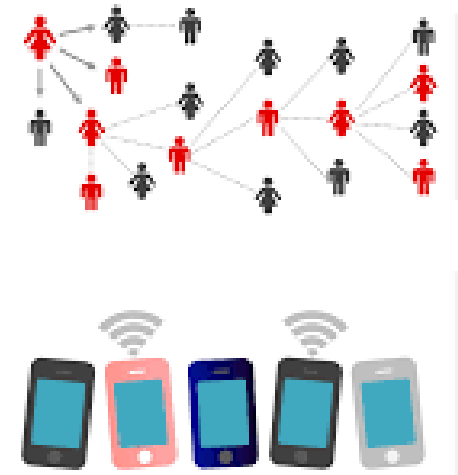


# ¿Cómo ha ayudado el Big Data en la coyuntura del COVID-19?

European electricity demand struggling to return to normal



## What is contact tracing?




# **Múltiples conceptos asociados a la análisis de datos**

- Big Data
- Data mining
- Machine learning (Deep learning)
- Ciencia de datos
- Inteligencia artificial

# Múltiples conceptos asociados a la analítica de datos

**Inteligencia artificial:** Disciplina amplia que tiene como objetivo crear máquinas inteligentes capaces de simular el comportamiento humano.

**Machine learning:** Es un subconjunto de técnicas estadísticas para entrenar a las máquinas a desempeñar tareas específicas -> 

**Deep learning:** Es un área de Machine Learning que busca replicar la capacidad cognitiva de los humanos.



# Múltiples conceptos asociados a la analítica de datos

**Big**

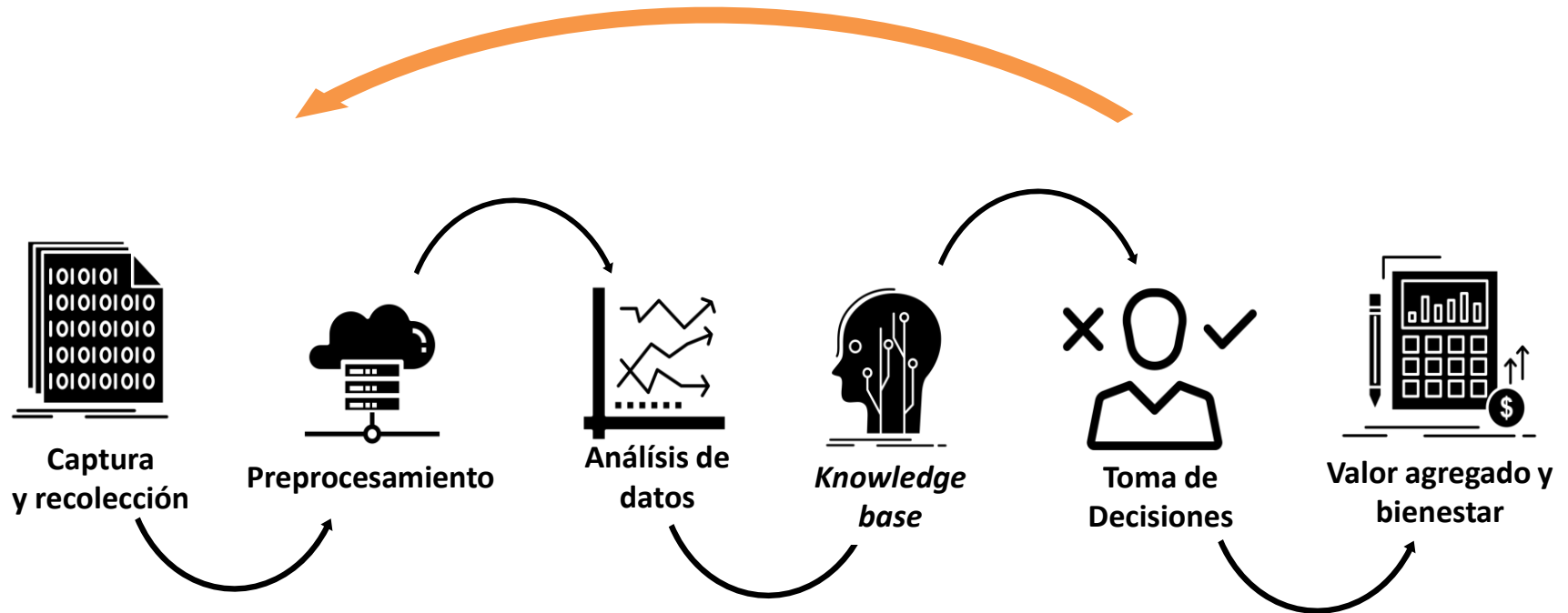
tecnología  
datos,  
generar  
organiz

procesos,  
procesar  
valor y  
y las

Big Data is like teenage sex:  
everyone talks about it,  
nobody really knows how to  
do it, everyone thinks  
everyone else is doing it, so  
everyone claims they are  
doing it.

Dan Ariely

# Cadena de valor de los datos





# ¿Qué se necesita para generar valor a través de los datos?

El ingrediente secreto 1...**Estadística**

**Cuando haces Data Science.**



**Pero No tienes bases estadísticas.**



# ¿Qué se necesita para generar valor a través de los datos?

El ingrediente secreto 2....**Programar**



TensorFlow



SciPy



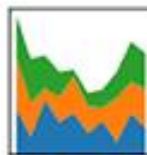
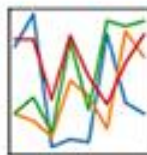
scikit  
learn



Keras

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



matplotlib



# ¿Qué se necesita para generar valor a través de los datos?



“Don't look for unicorns, build a data science team”  
*Bob Rogers, chief data scientist with Intel's Big Data Solution*

# **Knowledge Discovery in Datasets (KDD) y Data Mining (DM)**

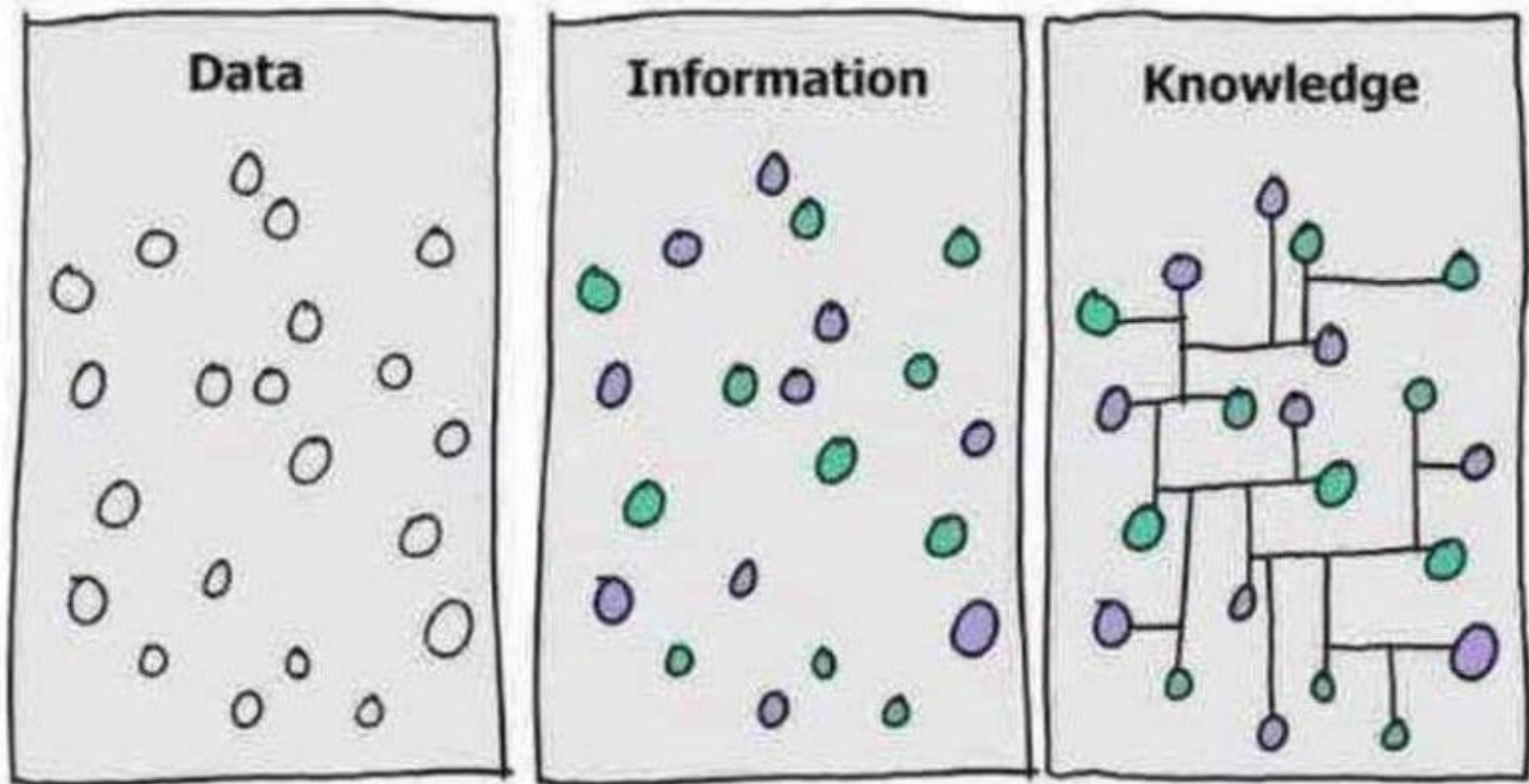
# KDD y DM

*“There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD).”*

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37-37.

# KDD y DM

- KDD es la extracción (no trivial) de conocimiento **implícito**, **descubierto automáticamente** a partir de los datos



OS

.es

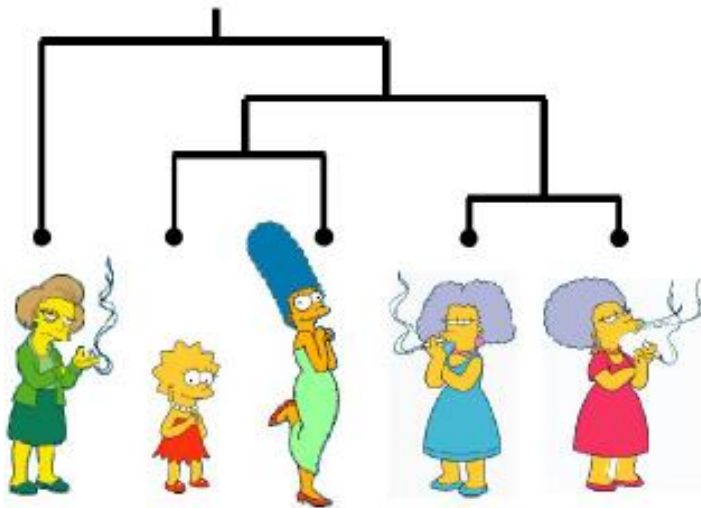
# **Analítica de datos y objetivos de análisis**

- Analítica descriptiva: proceso de análisis que permite resumir la información contenida en datos así como descubrir nuevos patrones
- Analítica predictiva: estrategias, técnicas y tecnologías que permite clasificar o predecir fenómenos de diferente naturaleza
- Analítica prescriptiva: hacer referencia procesos de análisis cuyo resultado brinda recomendaciones óptimas para la toma de decisiones

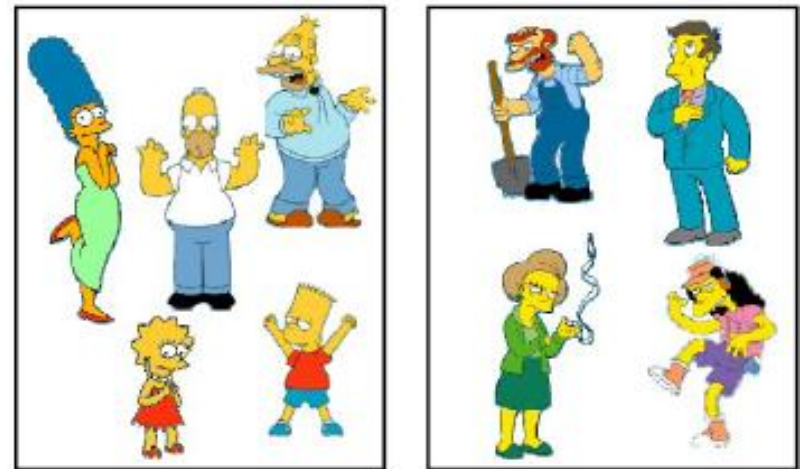
# Métodos de asociación (clustering)

- Los métodos de clustering se dividen en:
  - Jerárquicos: represent relaciones de similitud entre los objetos (e.g., aglomerativos y de división)
  - De partición: distribuye los elementos entre un número determinado de grupos (e.g., k-means)

**Clustering Jerárquico**



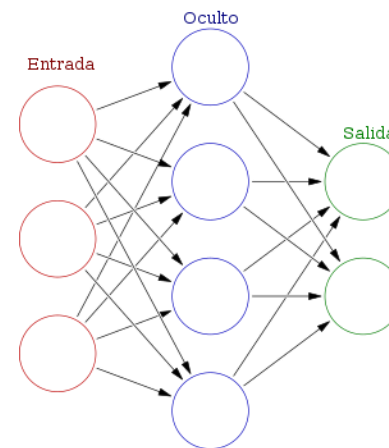
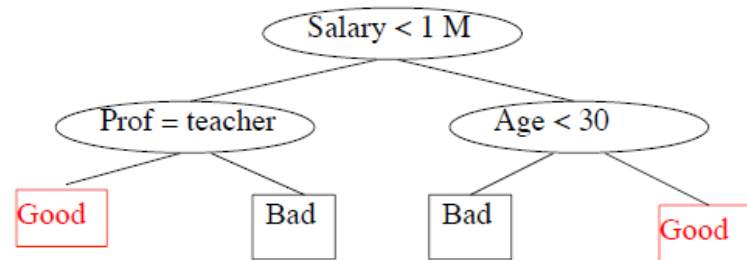
**Clustering de Partición**





# Métodos de clasificación

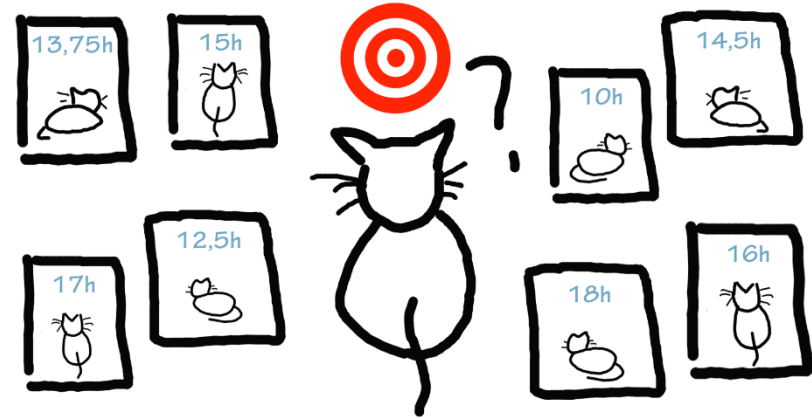
- Vecinos cercanos (nearest neighbor): define proximidad entre los objetos a partir de criterios de distancia
- Árboles de decisión (decision tree): divide la decisión/clasificación en reglas simples sobre los atributos
- Redes neuronales (neural networks): utiliza particiones no lineales a través de la conexión de atributos y valores de salida



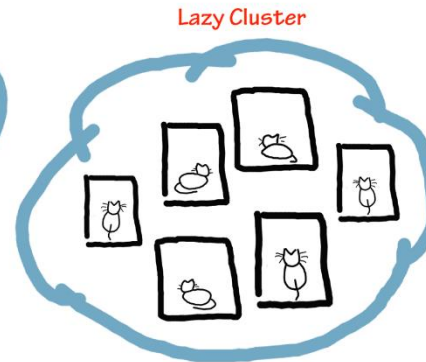
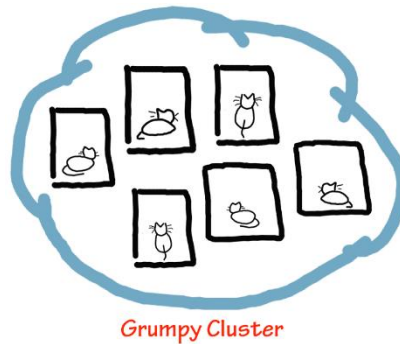
# Regresión, clasificación y clustering



¿Está el gato despierto?



¿Cuánto tiempo dormirá el gato?

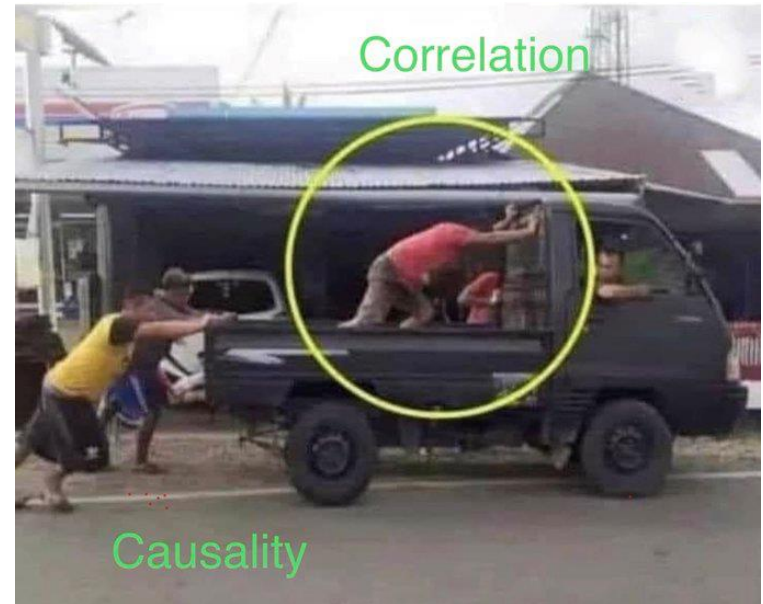
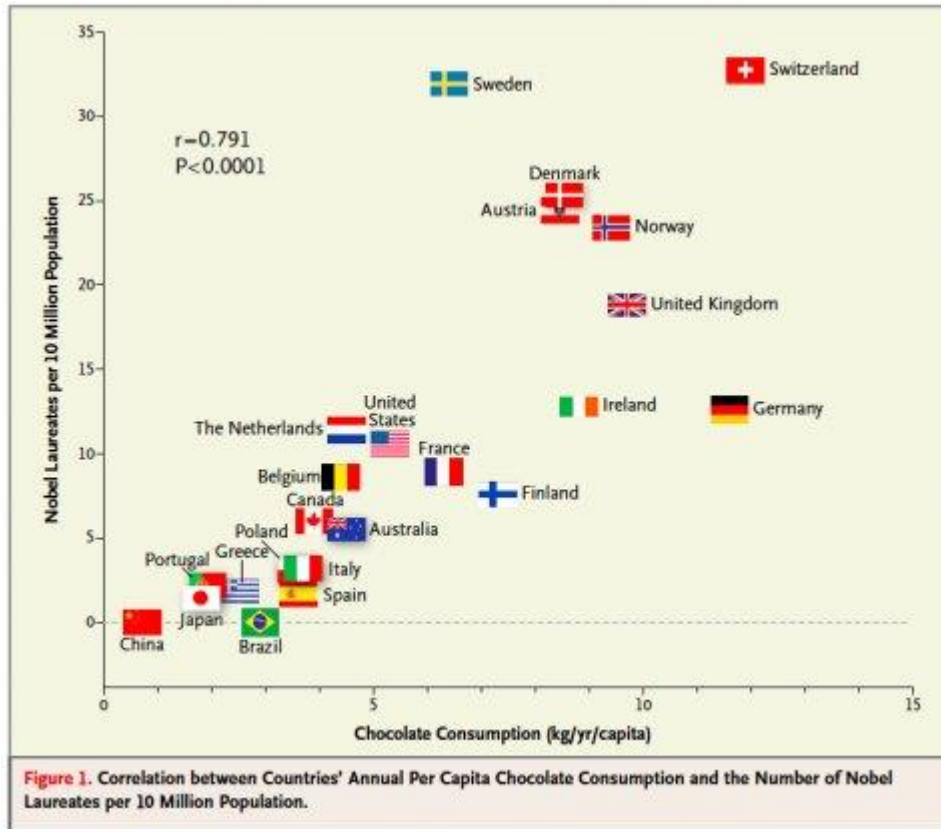


¿Como agrupar los gatos?

<https://medium.com/@csofiamsouza/learning-machine-learning-0-1-d4e199bb044e>

*Estos tipos de análisis no son mutuamente excluyentes*

# Correlación – Dependencia (Causalidad?)



# Tipos de datos

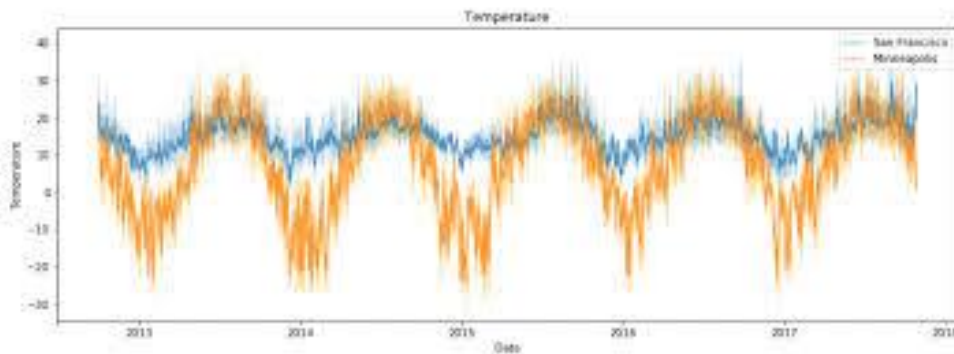
- Datos estructurados

features

year	model	price	mileage	color	transmission
2011	SEL	21992	7413	Yellow	AUTO
2011	SEL	20995	10926	Gray	AUTO
2011	SEL	19995	7351	Silver	AUTO
2011	SEL	17809	11613	Gray	AUTO
2012	SE	17500	8367	White	MANUAL
2010	SEL	17495	25125	Silver	AUTO
2011	SEL	17000	27393	Blue	AUTO
2010	SEL	16995	21026	Silver	AUTO
2011	SES	16995	32655	Silver	AUTO

examples

Colegios en Bogotá

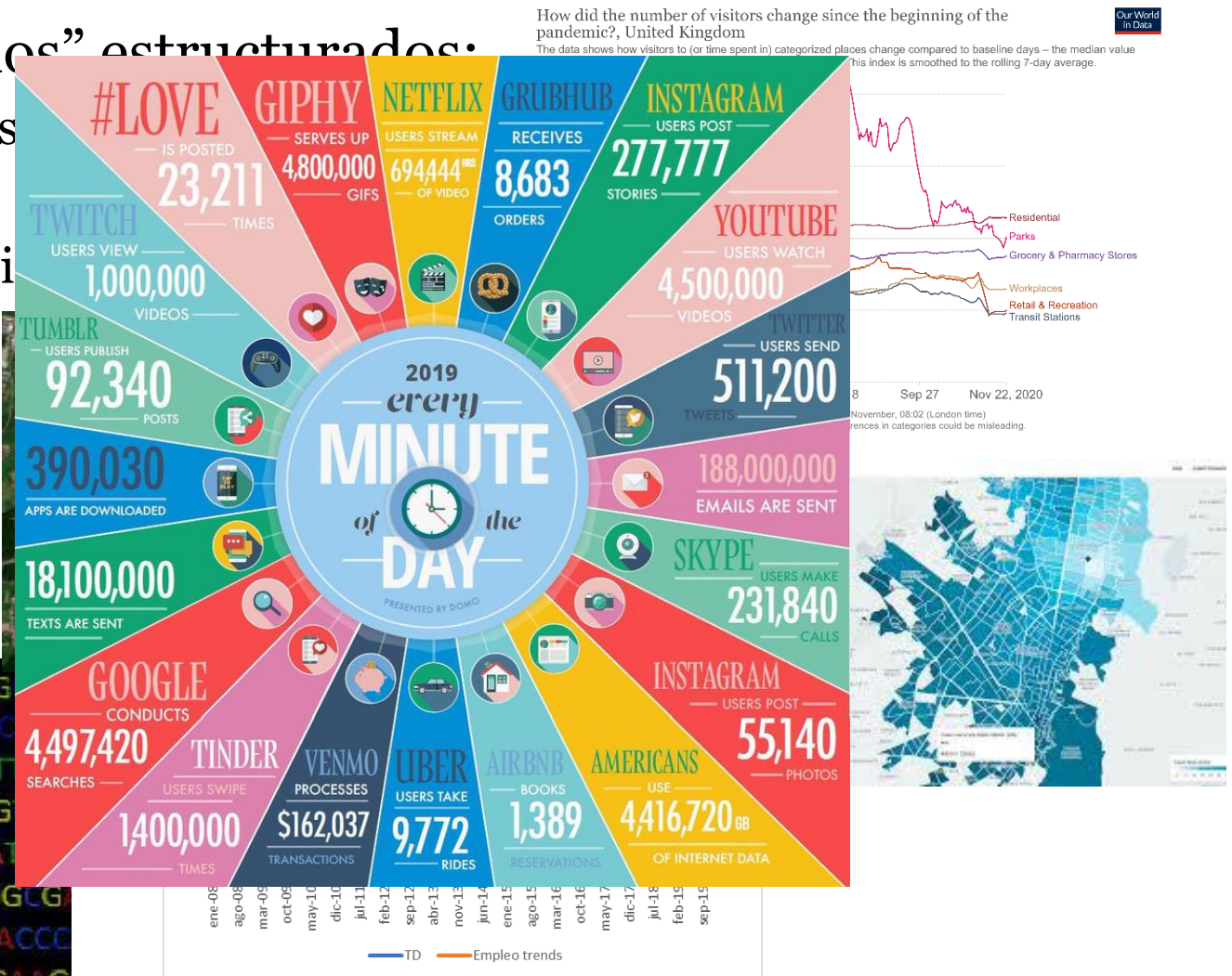




# Datos: Naturaleza y tipos

- Datos “menos” estructurados:

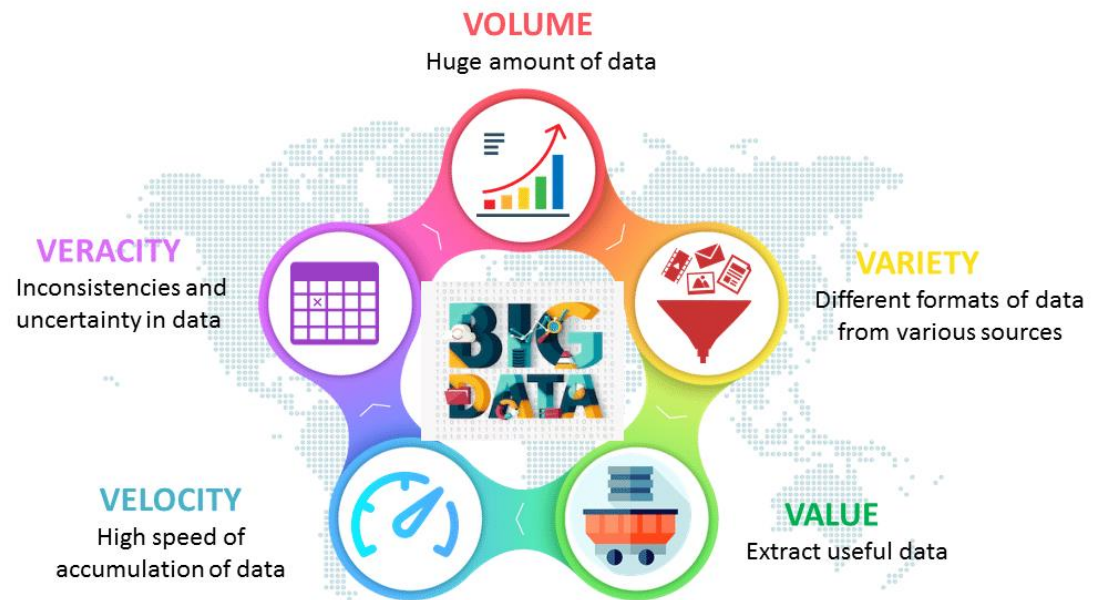
- Imágenes s
- Texto....
- Huella digi



# ¿Por qué hay una revolución alrededor de la analítica de datos?

Conceptualmente podemos pensar en 5Vs

- **V**olumen
- **V**elocidad
- **V**eracidad
- **V**ariiedad
- **V**alor





# Retos en el uso de datos

- Representatividad
- Integración de fuentes de información
- Definición de unidades de observación
- Errores de medición
- Rezago en el reporte



# KDD y DM: elementos clave

Datos -> Información -> modelo -> Conocimiento



<https://www.inverse.com/article/11456-microsoft-dog-recognition>

# Pronostico gasto en turismo en España





# Predecir...una pintura?

A



B



C



D



# Crear escenarios no conocidos

