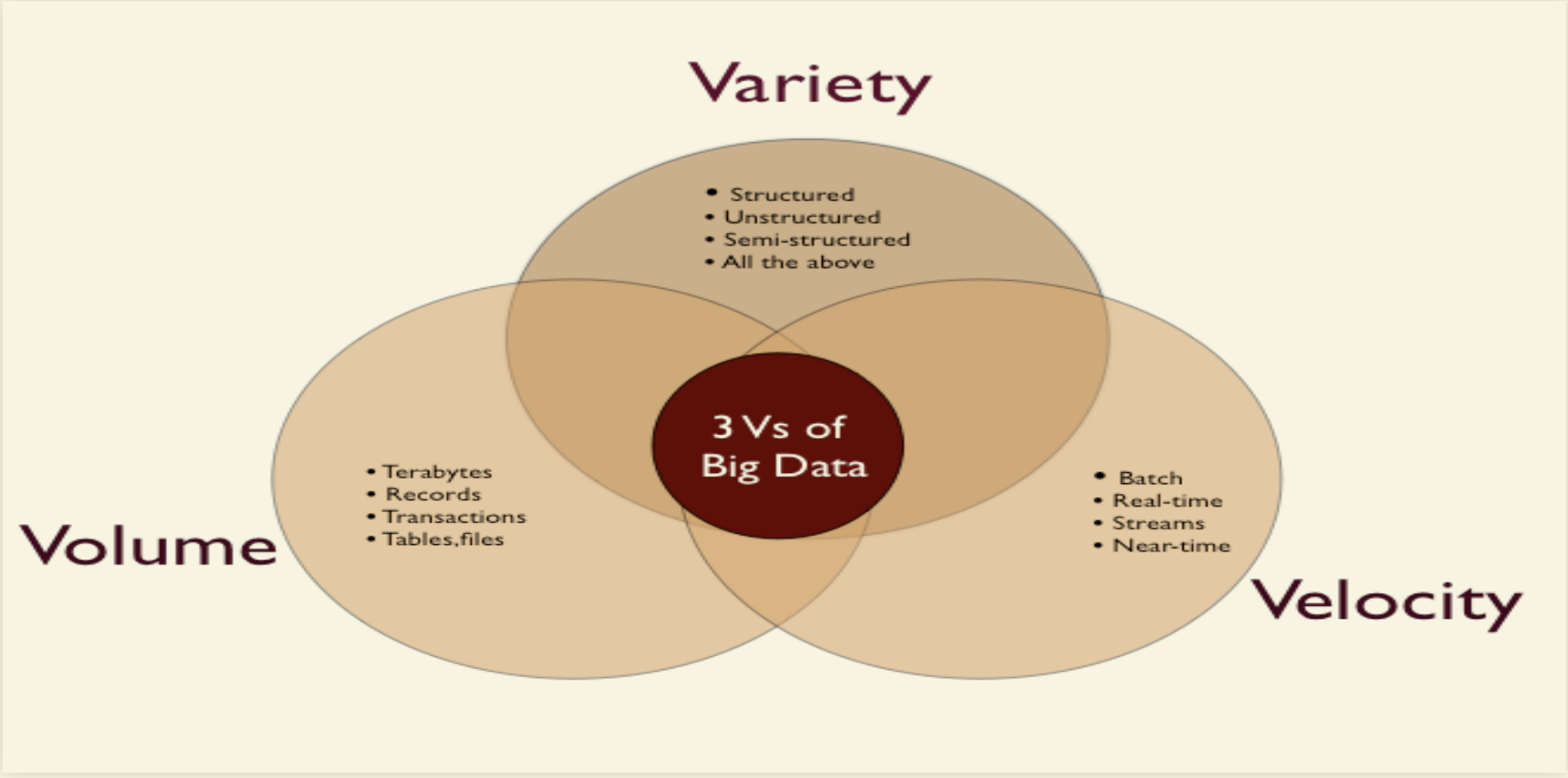
APROXIMACIÓN AL *BIG DATA*

Carlos G. Figuerola
Universidad de Salamanca
figue@usal.es

SEDIC, Jornada ACTUALÍZATE 2020

QUÉ ES EL *BIG DATA*

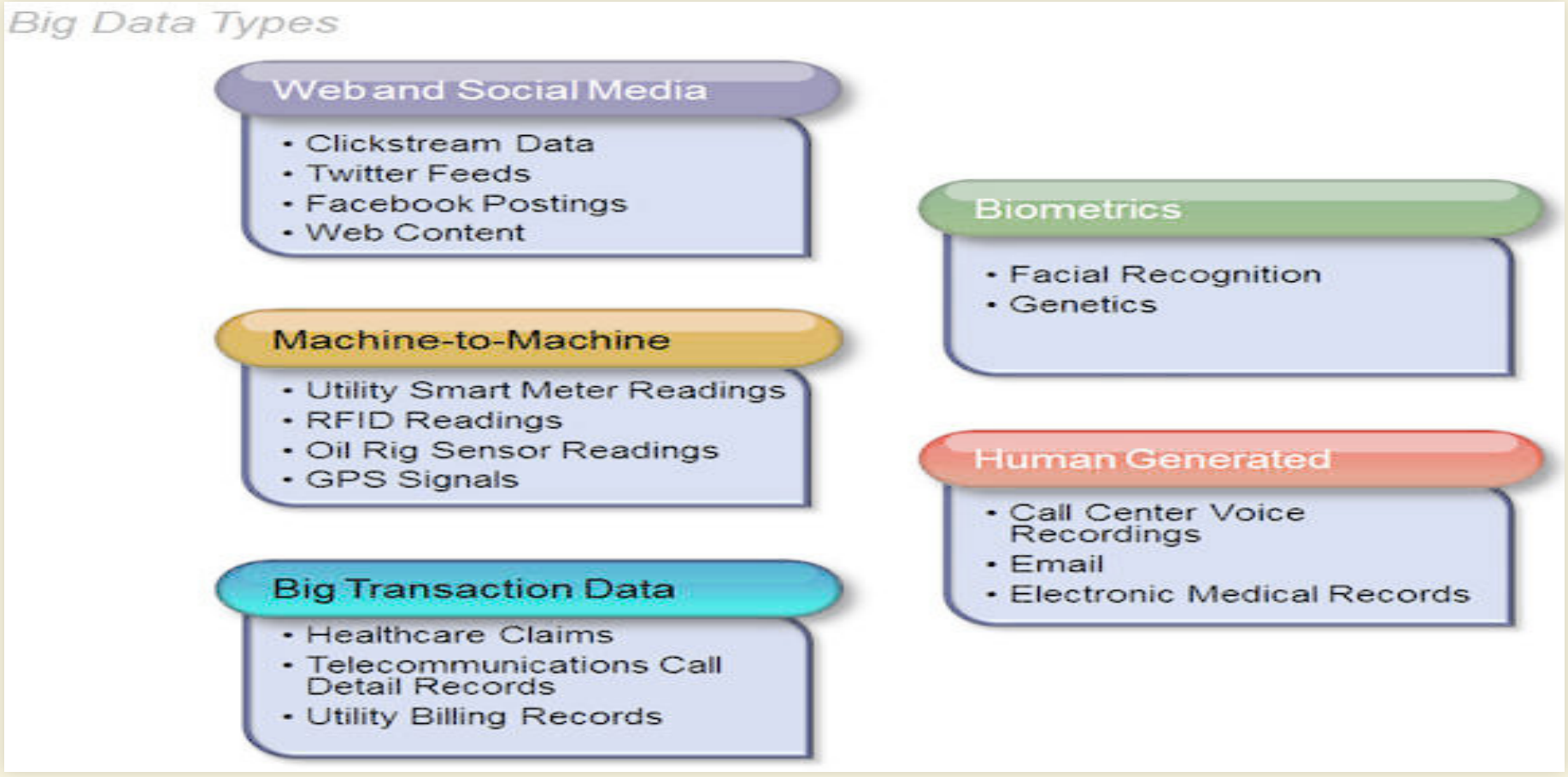


VOLUMEN Y VELOCIDAD

plantean retos técnicos

- almacenamiento
- disponibilidad
- tiempo de procesamiento
- ancho de banda
- etc

VARIEDAD DE DATOS



BIG DATA EN BIBLIOTECAS Y CENTROS DE DOCUMENTACIÓN

- catálogos crecen
  - además muchos tienen datos enlazados
  - bibliotecas digitales y repositorios
- datos de utilización
  - consulta y préstamo
  - descargas y visualizaciones
  - citas, enlaces recibidos
  - navegación y búsquedas

HETEROGENEIDAD DE DATOS

- gran número de fuentes (producen aumento en volumen)
- diversos tipos de datos
- datos estructurados, semiestructurados y no estructurados
- datos *limpios*, datos que necesitan ser preparados (ruido, datos incompletos, ambiguos ...)

NUEVOS ALGORITMOS

- permiten trabajar con esos tipos de datos
- van más allá de la estadística analítica *clásica*
  - mejores análisis
  - obtienen nuevo conocimiento
- más campos de aplicación
- fuerte componente de **interdisciplinaridad**

TÉCNICAS DE ANÁLISIS Y PROCESAMIENTO

- estrechamente vinculadas con la *Inteligencia Artificial* - un elemento importante es el **aprendizaje automático** (*machine learning*)

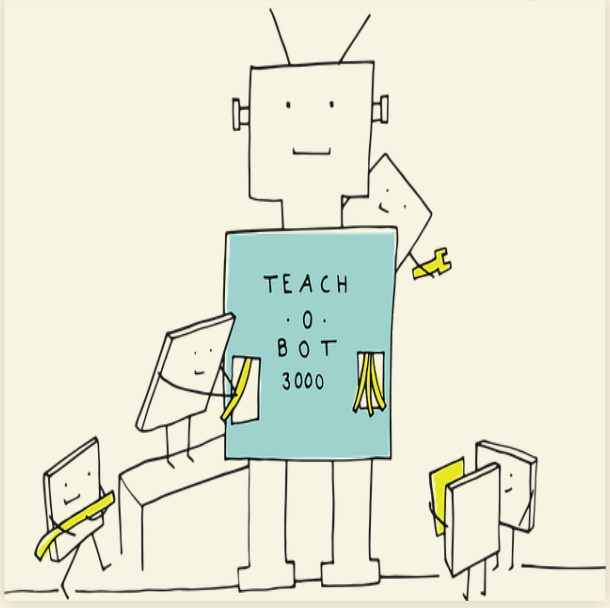


Image by Manfred Steger from Pixabay

QUÉ ES EL APRENDIZAJE AUTOMÁTICO?

- los ordenadores aprenden cosas aparentemente simples
- con esos elementos simples se construyen universos complejos
  - posibilidad de automatizar operaciones complejas
    - toma de decisiones
    - gestión automática
    - organización automática
    - predecir comportamientos
    - etc

TIPOS DE APRENDIZAJE AUTOMÁTICO

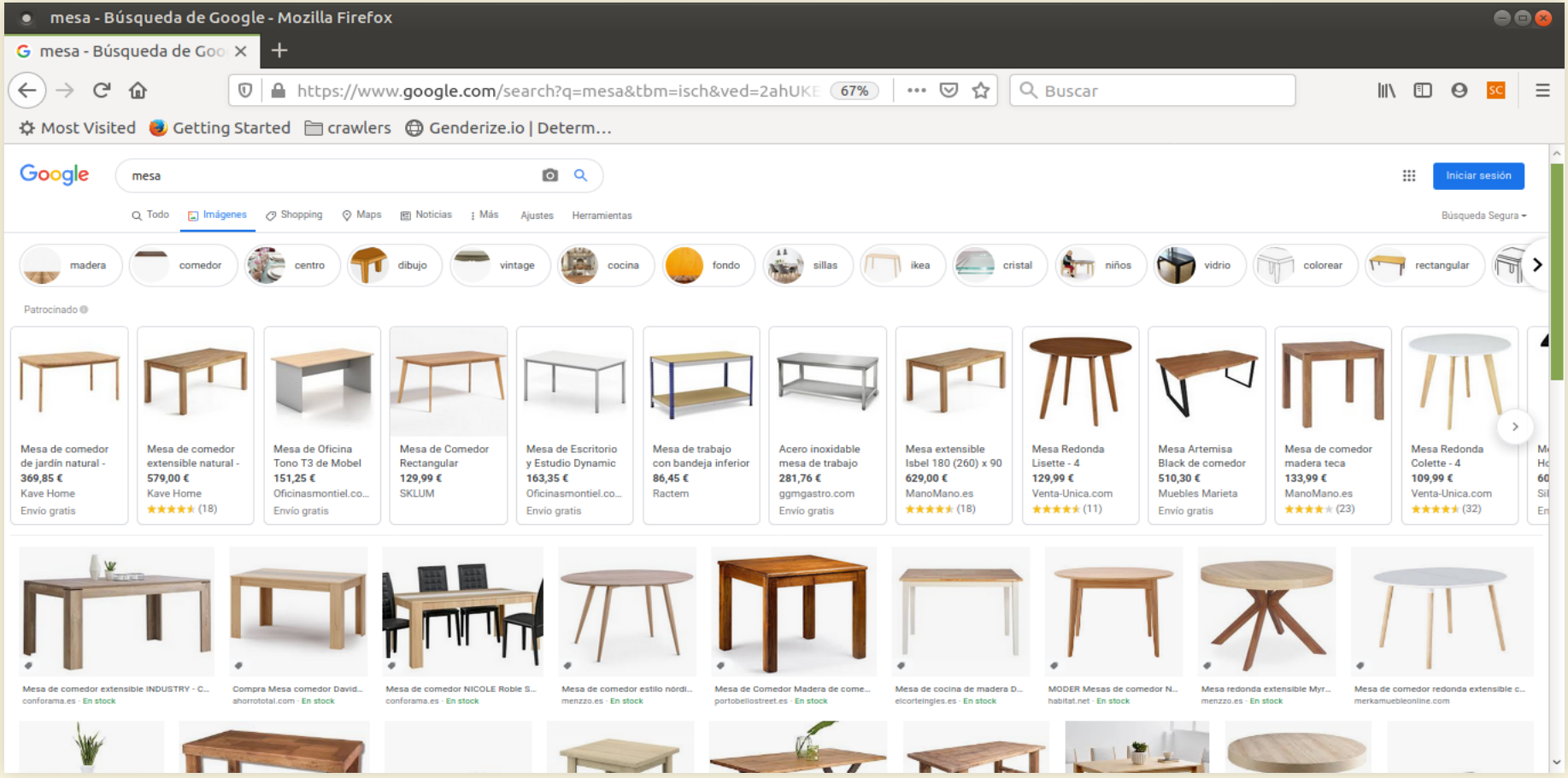
- diversas clasificaciones en base a diversos criterios
- aprendizaje **heurístico** vs. **aprendizaje estadístico**
- aprendizaje **supervisado** vs. **no supervisado**

APRENDIZAJE HEURÍSTICO VS. ESTADÍSTICO

Ejemplo: aprehensión del concepto *mesa*

Imagen de Clker-Free-Vector-Images en Pixabay

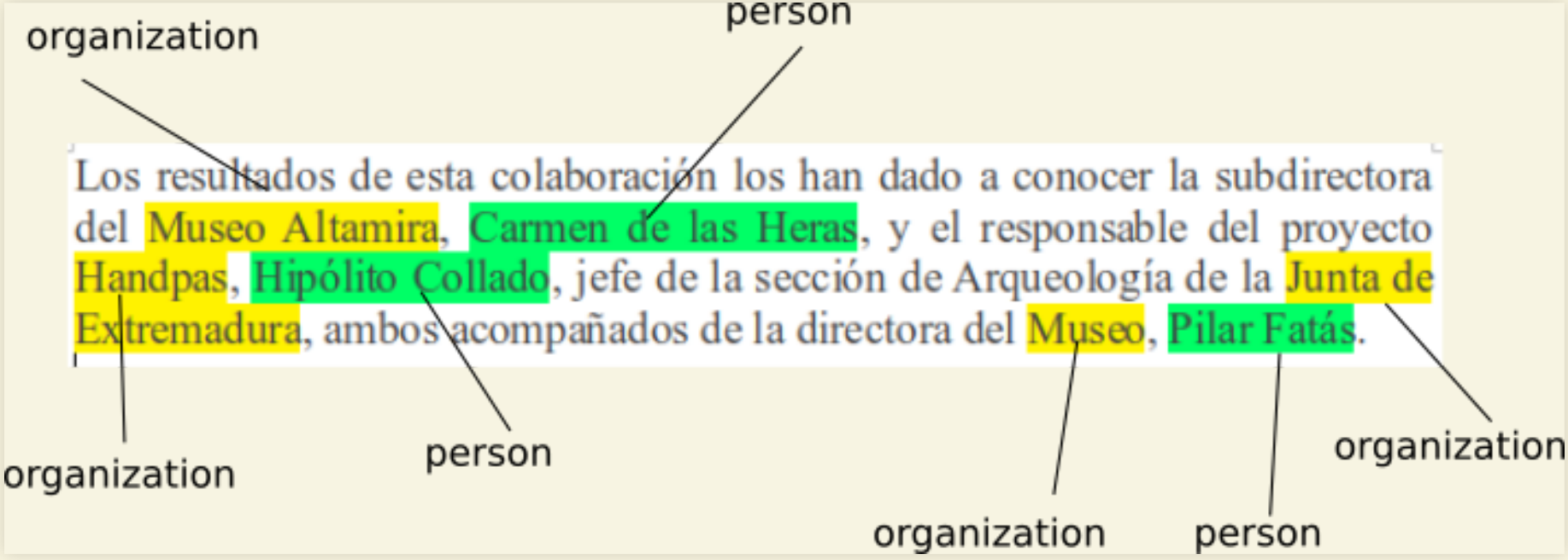
APRENDIZAJE HEURÍSTICO VS. ESTADÍSTICO



HEURÍSTICA Y ESTADÍSTICA PUEDEN COMBINARSE

Pueden combinarse ambos tipos

Ejemplo: Reconocimiento de entidades (NER)



APRENDIZAJE SUPERVISADO VS. NO SUPERVISADO

- muy relacionado con clasificación automática
- supervisado:
  - se *enseña* un modelo (o varios)
  - el sistema es capaz de medir el parecido de nuevos casos con los modelos aprendidos

EJEMPLO

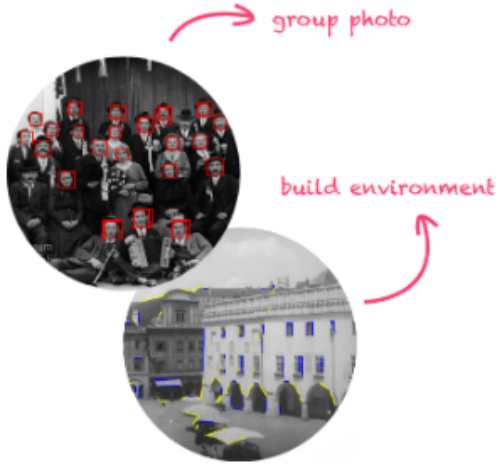
Diagnósticos médicos

EJEMPLO

Clasificación de Fotografías

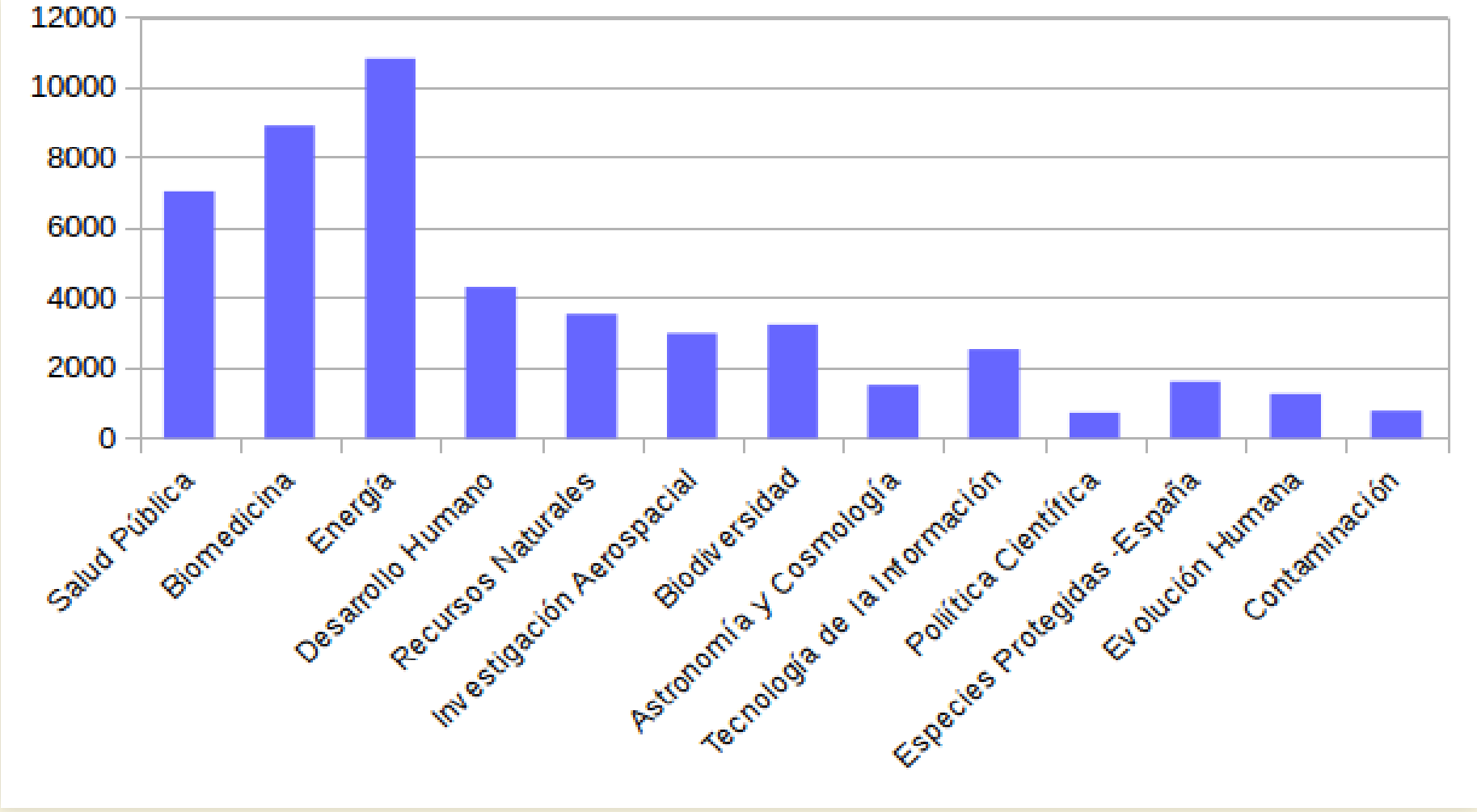
DEVELOPMENT OF AN AUTOMATED IMAGE ANALYSIS PROCESS

Through imaging processes, historical photos can be examined for objects, buildings, text, etc., and individual categories can be assigned automatically.



EJEMPLO

Clasificación de documentos



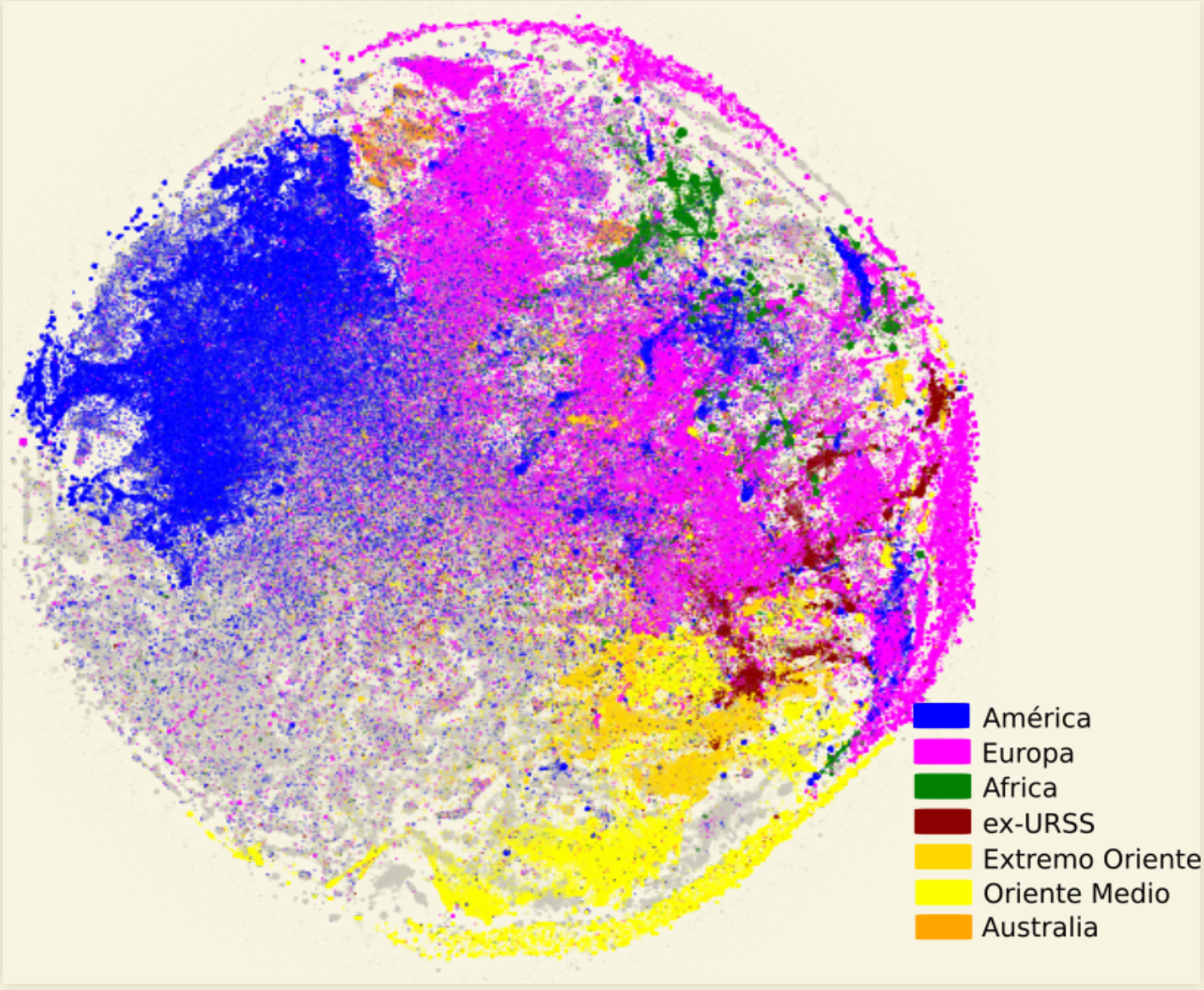
APRENDIZAJE NO SUPERVISADO

- el sistema aprende él solo
- descubre patrones, regularidades, características diferenciadoras ...
- permite organizar automáticamente la información (p. ej.: documentos)
- permite descubrir y extraer información no explícita

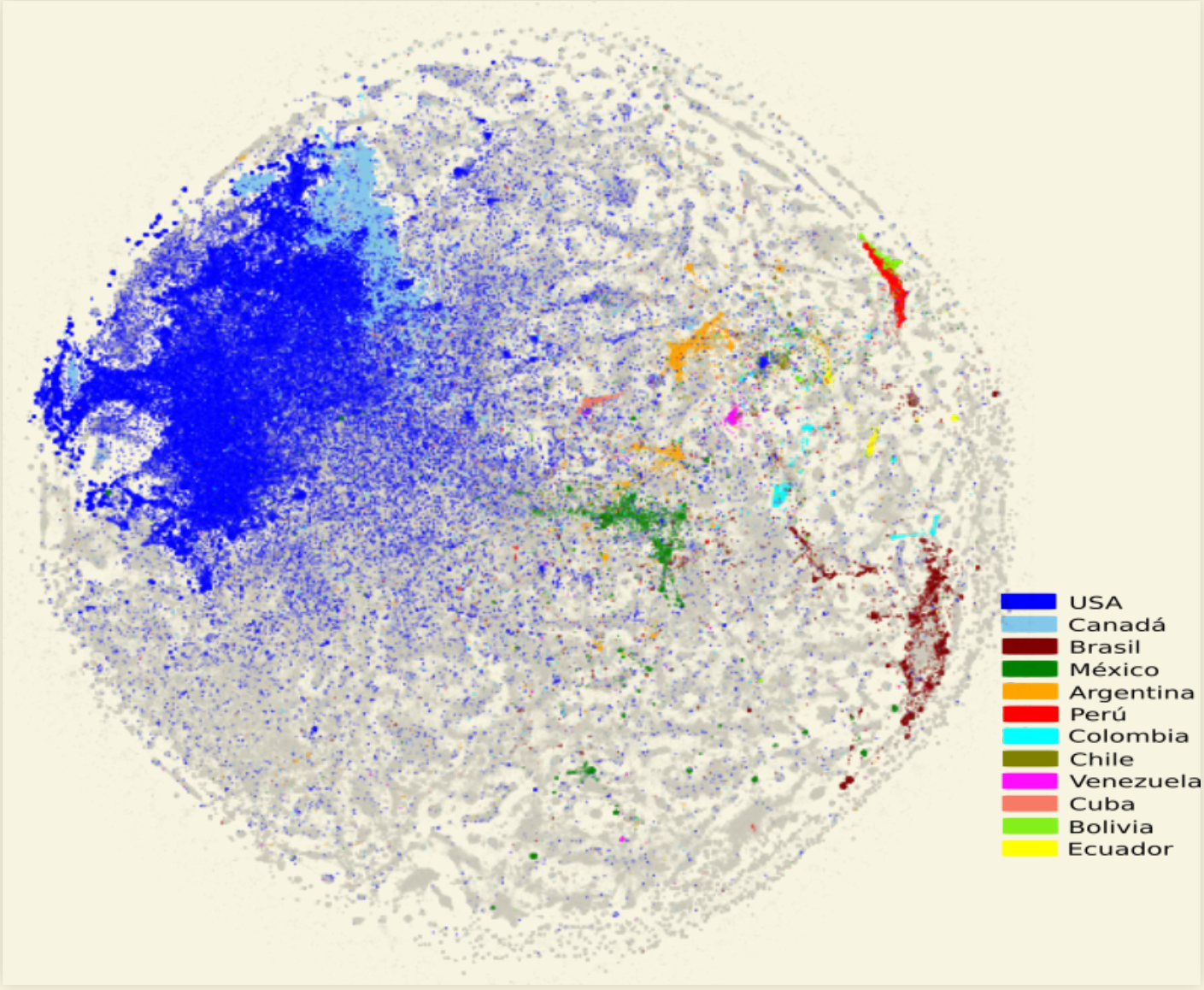
EJEMPLO: CLUSTERS DE DOCUMENTOS

```
1,1,0.00456697,pais19770226-177,"Profesores no numerarios: sigue la huelga"
1,2,0.00422373,pais19770123-135,"La huelga de profesores continuará esta semana"
1,3,0.00417448,pais19770127-135,"Se acentúa el conflicto de la enseñanza"
1,4,0.00399125,pais19770222-153,"El Ministerio amenaza con sanciones económicas a los PNN de institu
1,5,0.00385055,pais19770204-132,"Los enseñantes piden estabilidad laboral y control democrático de l
...
...
2,1,0.00358621,pais19771008-136,"Los partidos mayoritarios y la Federación Católica de Padres se pro
2,2,0.00348011,pais19770614-099,"Ultimas intervenciones de los líderes en Televisión"
2,3,0.00339085,pais19770216-177,"Sobre los PNN de Universidad"
2,4,0.00274552,pais19770518-061,"Al día siguiente"
2,5,0.00257922,pais19771222-117,"Los problemas concretos y la autocrítica"
2,6,0.00250335,pais19770423-129,"Entrevista con Arias Navarro"
2,7,0.0024123,pais19770720-067,"El voto por la libertad y la democracia"
...
...
13,1,0.00180532,pais19770330-009,"Los estudiantes de Educación Física se manifiestan ante el ministe
13,2,0.00177176,pais19770212-013,"Ministerio de Educación y DND tratarán el lunes los problemas del
13,3,0.00137942,pais19770326-020,"Los estudiantes de Educación Física reivindican un rango universit
13,4,0.00134001,pais19770218-016,"La educación física, un problema académico, laboral y político"
13,5,0.000893123,pais19771227-021,"La figura del profesor, centro de atención"
13,6,0.000752091,pais19771112-149,"Las instalaciones deportivas oficiales, abiertas al público"
13,7,0.000743549,pais19770430-026,"Los educadores físicos, pilares fundamentales del deporte alemán"
...
...
```

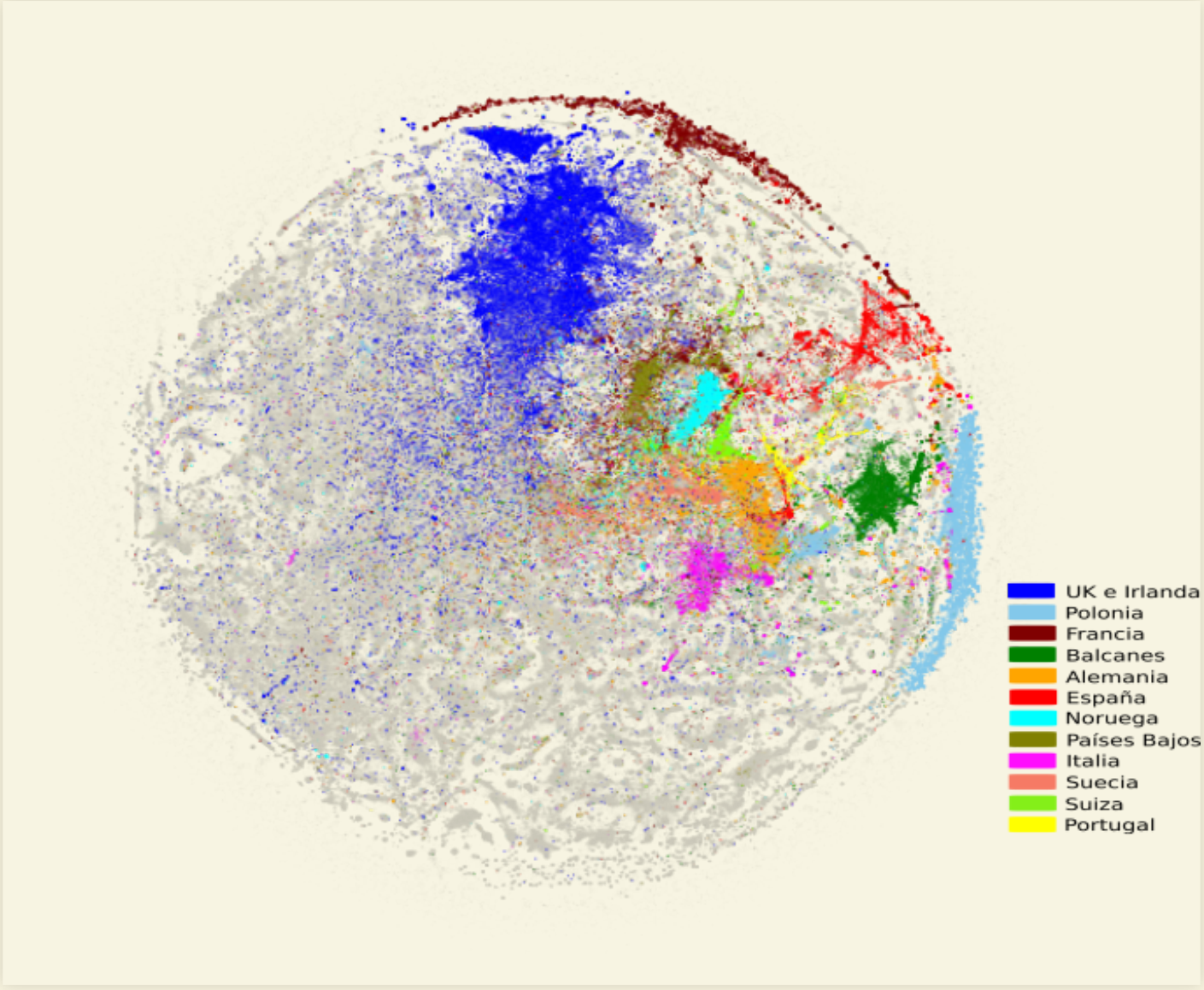
EJEMPLO: ORGANIZACIÓN AUTOMÁTICA DE DOCUMENTOS



EJEMPLO


USA
Canadá
Brasil
México
Argentina
Perú
Colombia
Chile
Venezuela
Cuba
Bolivia
Ecuador

EJEMPLO


UK e Irlanda
Polonia
Francia
Balcanes
Alemania
España
Noruega
Países Bajos
Italia
Suecia
Suiza
Portugal

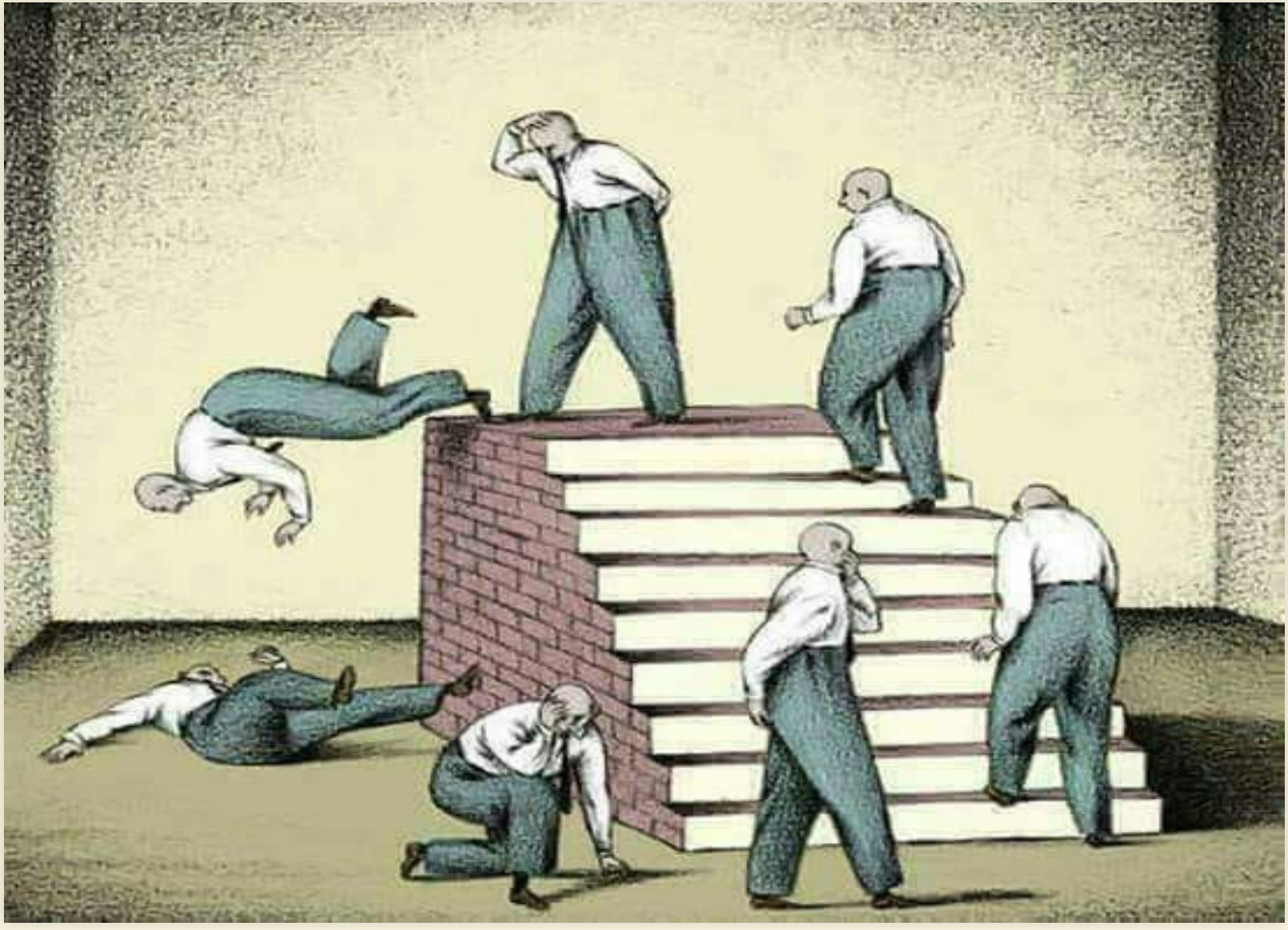
APRENDIZAJE AUTOMÁTICO Y BIG DATA

- la disponibilidad de muchos datos sobre muchas cosas (**BIG DATA**) permite la aplicación práctica del aprendizaje automático de base estadística

BIG DATA = BIG LEARNING

AMENAZA: LA PRIVACIDAD

AMENAZA: CÍRCULOS VICIOSOS



AMENAZA: LA LIBERTAD PERSONAL

MUCHAS GRACIAS POR SU ATENCIÓN ...

CARLOS G. FIGUEROLA

figue@usal.es