

Bases de Datos

Datos no estructurados



Datos estructurados

- relacionales
- objetos: más flexibilidad, pero la misma estructura
- ontologías: mucha precisión

Datos no estructurados

Bolsas de gatos

Cajón desastre

- Difíciles de catalogar
- Creemos que tienen algo valioso, indefinido, para un futuro incierto

Datos no estructurados

Bolsas de gatos

Cajón desastre

Documentos

- Difíciles de catalogar
- Creemos que tienen algo valioso, indefinido, para un futuro incierto

Datos no estructurados

- Información sin un modelo de datos predefinido
- Información no organizada
- Texto, también datos referenciales (fechas, números, hechos)
- Irregularidades, ambigüedades

https://en.wikipedia.org/wiki/Unstructured_data

¿Un problema menor?

80% - 90% de los datos generados y recolectados son no estructurados, y creciendo más que los estructurados

https://en.wikipedia.org/wiki/Unstructured_data

Un problema con valor

- Los datos no estructurados se pueden usar para tomar decisiones
- Difícil de analizar
- El aprendizaje automático nos ayuda a analizar datos no estructurados

<https://www.mongodb.com/unstructured-data>

Ejemplos

Documentos

Mails

Chats

Comentarios a fotos, videos

Emoticones

Ejemplos

Fotos

Videos

Audio

Ejemplos

Redes sociales

- Textos
- Republicaciones
- Reacciones (y falta de reacciones)
- Follow, Unfollow
- Silencio, Block, Denuncia

Ejemplos

Uso de webs y aplicaciones

- Qué partes de la web se visitan
- Dónde hacen click
- Qué acciones hacen
- Usabilidad: idas y venidas

Ejemplos

Uso de aplicaciones

- Teléfono
- Smart cosas: televisión, reloj, heladera, aspiradora...

Ejemplos

- Imágenes de satélites, telescopios
- Vigilancia: cámaras, micrófonos
- Datos científicos: datos de experimentos, datos de sensores (p.ej. meteorológicos, contaminación)

También hay datos semi-estructurados

- Historia clínica: datos numéricos y texto libre
- Reportes meteorológicos
- Documentos con metadatos
- Documentos con marcas: títulos, links, infoboxes

**Necesitamos que sean
estructurados**

Yo te los estructuro!

un documento como tabla

Yo te los estructuro!

un documento como tabla

- cada palabra es un atributo (una columna)

el valor es la cantidad de veces que ocurre la palabra

Yo te los estructuro!

un documento como tabla

- sólo las palabras frecuentes
o frecuentes en el documento pero no
en la colección?
- sólo las palabras canónicas
- sólo palabras clave

Yo te los estructuro!

un documento como tabla

- expresiones multipalabra! Cuáles?
- son importantes los títulos?
- y los links?
- esto tiene imágenes!
- y tablas!

Yo te los estructuro

una imagen

Yo te los estructuro

una imagen

- palabras clave
- bolsas de palabras visuales
- características de redes neuronales

Yo te los estructuro

una imagen

- palabras clave
- bolsas de palabras vectoriales
- características

A large, multi-pointed pink starburst graphic with a white outline, centered on the slide. It contains the text 'Ke dice la profe' in white.

Ke dice la profe

Paréntesis: la magia del deep learning

Esto es magia, o marketing?

Minería de datos (data mining)

Machine learning (aprendizaje automático)

Big Data

Inteligencia Artificial

IoT

Deep Learning / Redes Neuronales

Aprendizaje automático

En lugar de programar un programa,
Hay programas que infieren programas
a partir de casos de ejemplo

<http://www.r2d3.us/una-introduccion-visual-al-machine-learning-1/>

Aprendizaje automático no supervisado

En lugar de crear atributos,

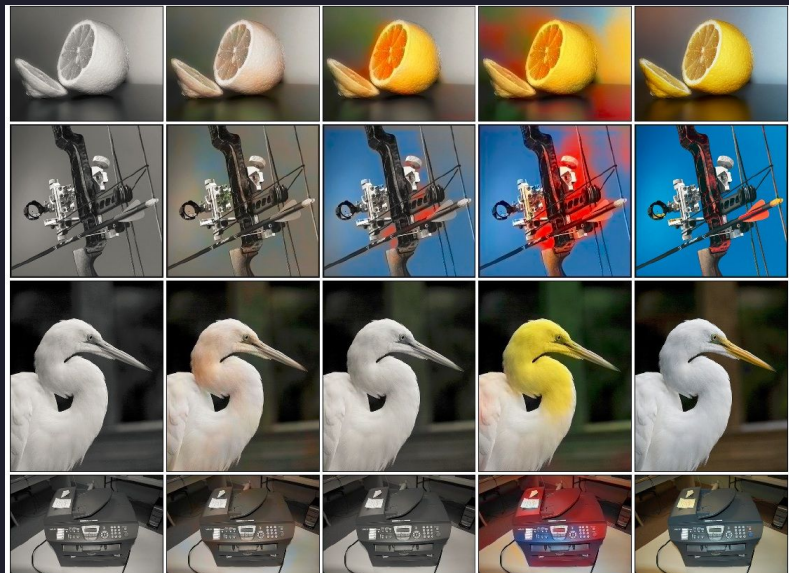
Hay programas que descubren información relevante a partir de grandes cantidades de datos, con una tarea de pretexto

(suelen ser redes neuronales)

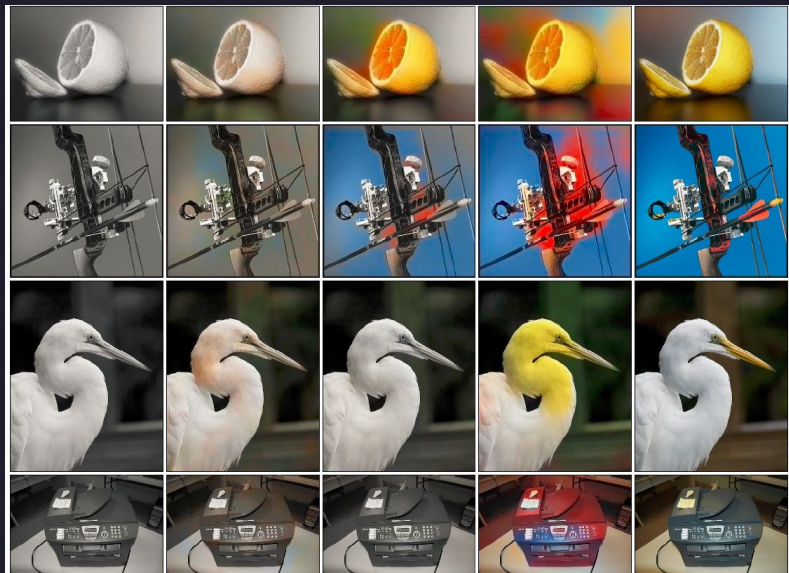
Tarea de pretexto



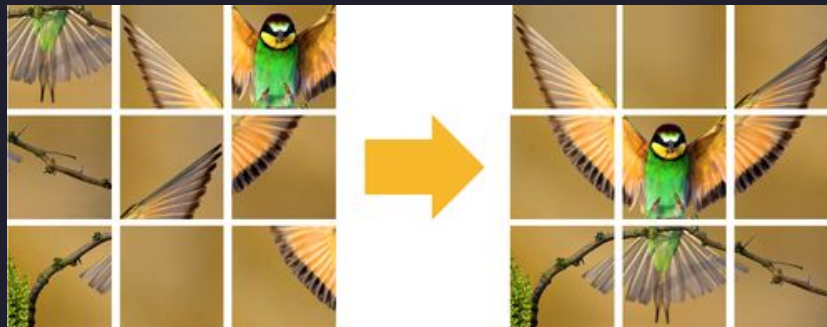
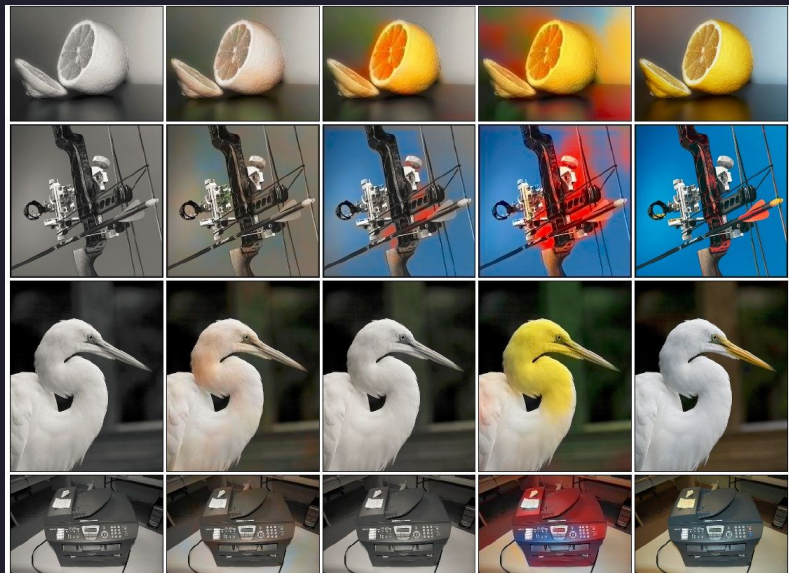
Tarea de pretexto



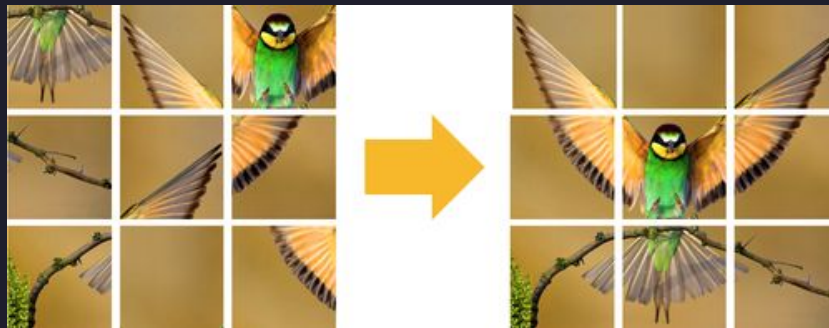
Tarea de pretexto



Tarea de pretexto

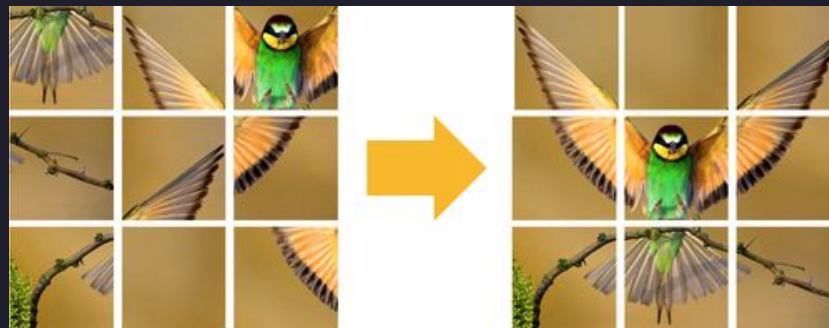
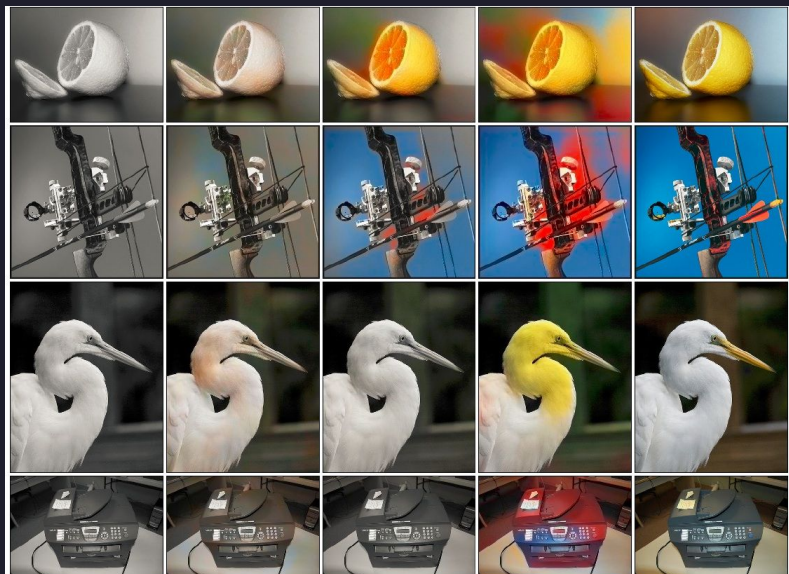


Tarea de pretexto



el gato come pescado

Tarea de pretexto



el gato come pescado

__	gato	come	pescado	?
el	__	come	pescado	?
el	gato	__	pescado	?
el	gato	come	__	?

¿Aprendizaje Automático?



<https://www.fhuc.unl.edu.ar/veracartonera/portfolio/aprendizaje-automagico/>

Vocabulario de machine learning / data mining

Atributo = *feature* / característica

Registro = caso, ejemplo

Fin del Paréntesis

¿Cómo hago una BD de datos no estructurados?

- Guardar toda la información
- Guardar toda la información que pueda guardar
- Crear atributos específicos para una necesidad de información específica
- Mucha creatividad

Cuestiones éticas

Qué implica juntar y persistir todos esos datos?

**Pero en concreto,
¿cómo hago?**

De datos no estructurados a estructurados

- Identificar la estructura subyacente
 - Manualmente
 - Automáticamente
- Análisis morfosintáctico, semántico
- Reconocimiento de objetos
- Reconocimiento temprano de eventos

Reconocimiento temprano de eventos

- Epidemias
- Crisis económicas
- Grooming
- Enfermedades latentes
- Abandono escolar
- Fallos de hardware (disco)

**Los datos no estructurados
suelen ser
Big Data**

Grandes volúmenes de datos (Big Data)

Almacenamiento

Recuperación

- Índices invertidos
- Paralelización (y distribución)
- Map reduce, hadoop, spark

Mapa de ruta

- ~~1. Orientación a objetos~~
- ~~2. Datos no estructurados~~
3. Recuperación de Información
4. Data Mining
5. Data Warehousing

/THANKS!

/DO YOU HAVE ANY QUESTIONS?

youremail@freepik.com

+91 620 421 838

yourwebsite.com



CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

> Please keep this slide for attribution

