

Bases de Datos

Minería de Datos



Lecturas

https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico

Otros nombres para el área

Data Mining

Big Data, Grandes Datos

Aprendizaje Automático, Machine Learning

Modelos predictivos

Inteligencia de Negocios, BI

Inteligencia Artificial

¿Qué es la minería de datos?

descubrir patrones

para predecir el futuro

¿Qué cosas se pueden predecir?

si va a llover mañana

si me va a gustar una película

si un mail es spam

si hay un tumor en una imagen

si voy a comprar cerveza

¿Cómo es un patrón?

Correlación entre valores de atributos

- significativa
- representativa
- interesante

Reglas de Asociación

Intuición

La probabilidad condicional hecha regla

¿Qué nos suma este formato?

- Más fácil de inspeccionar
- Se pueden insertar métricas: novedad, sorpresa, valor económico, clase

→ Más accionable!

De intuición a producción hay un buen trecho!

Contexto

- El algoritmo más popular es Apriori (Agrawal et al 1993)
- Todos los datos tienen que ser categóricos
- Inicialmente se usó para Análisis del Carrito de la Compra (Market Basket Analysis)

Pan → Leche [sop = 5%, conf = 100%]

Terminología

I = $\{i_1, i_2, \dots, i_m\}$: un conjunto de **items**.

Transacción **t** :

t es un conjunto de items sin orden, y $t \subseteq I$.

Base de datos de transacciones: un conjunto de transacciones $T = \{t_1, t_2, \dots, t_n\}$.

Ejemplo

Transacciones de compra de mercado:

t1: {pan, queso, leche}

t2: {manzana, huevos, sal, yogur}

... ..

tn: {bizcocho, huevos, leche}

Definiciones:

- **item**: un item/artículo en el carrito de la compra
- **I**: todos los items que se venden en el negocio
- **transacción**: items comprados en un ticket (*basket*)

Ejemplo

Transacciones de compra de mercado:

t1: {pan, queso, leche}

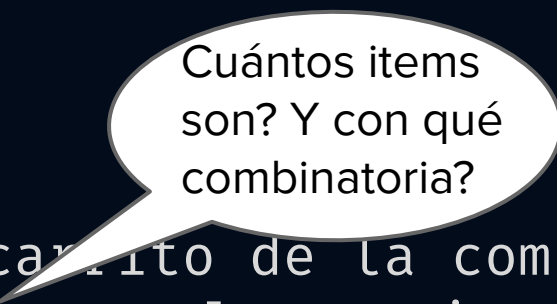
t2: {manzana, huevos, sal, yogur}

... ..

tn: {bizcocho, huevos, leche}

Definiciones:

- **item**: un item/artículo en el carrito de la compra
- **I**: todos los items que se venden en el negocio
- **transacción**: items comprados en un ticket (*basket*)



Cuántos items son? Y con qué combinatoria?

Ejemplo

Un dataset de documentos de texto

doc1: Estudiante, Enseñar, Escuela

doc2: Estudiante, Escuela

doc3: Enseñar, Escuela, Ciudad, Partido

doc4: Beisbol, Basket

doc5: Basket, Player, Espectador

doc6: Beisbol, Entrenador, Partido, Equipo

- **item**: una palabra en un documento
- **I**: todas las palabras del conjunto de documentos
- **transacción**: las palabras de un documento

Ejemplo

Un dataset de documentos de texto

doc1: Estudiante, Enseñar, Escuela

doc2: Estudiante, Escuela

doc3: Enseñar, Escuela, Ciudad, Partido

doc4: **Qué queremos saber?**

doc5: Fútbol, Fútbol, Espectador

doc6: Beisbol, Entrenador, Partido, Equipo

- **item**: una palabra en un documento
- **I**: todas las palabras del conjunto de documentos
- **transacción**: las palabras de un documento

Ejemplo

alumno, inscripto,
becario, alumnas

Un dataset de documentos

doc1: Estudiante, Enseñar, Escuela

doc2: Estudiante, Escuela

doc3: Enseñar, Escuela, Ciudad, Partido

doc4: Beisbol, Basket

doc5: Basket, Player, Espectador

doc6: Beisbol, Entrenador, Partido, Equipo

- **item**: una palabra en un documento
- **I**: todas las palabras del conjunto de documentos
- **transacción**: las palabras de un documento

Ejemplo

Un dataset de documentos

doc1: Estudiante, Enseñar, Escuela

doc2: Estudiante, Escuela

doc3: Enseñar, Escuela

doc4: Beisbol, Basket

doc5: Basket, Player,

doc6: Beisbol, Entrer

alumno, inscripto,
becario, alumnas

- Pre-procesos
- Conocimiento de dominio (traductores, sinónimos)
- Embeddings!

- **item**: una palabra en un documento
- **I**: todas las palabras del conjunto de documentos
- **transacción**: las palabras de un documento

Ejemplo

Un conjunto de historias clínicas.

paciente1:

consulta1:deshidratación, fiebre38.5, ibuprofeno

consulta2:gastritis, protector_gástrico

paciente2:

consulta1:dolor_articular, fiebre39, antibiótico

consulta2:dolor_articular, febrícula37.5, ibuprofeno

consulta3:gastritis, protector_gástrico

- **item**: un evento en una historia clínica
- **I**: todos los eventos en todas las historias clínicas
- **transacción**: consulta? historia clínica? período?

Ejemplo

Un conjunto de historias clínicas.



discretizar

paciente1:

consulta1:deshidratación, fiebre38.5, ibuprofeno

consulta2:gastritis, protector_gástrico

paciente2:

consulta1:dolor_articular, fiebre39, antibiótico

consulta2:dolor_articular, febrícula37.5, ibuprofeno

consulta3:gastritis, protector_gástrico

- **item**: un evento en una historia clínica
- **I**: todos los eventos en todas las historias clínicas
- **transacción**: consulta? historia clínica? período?

Ejemplo

Un conjunto de historias clínicas.

discretizar

paciente1:

consulta1:deshidratación, fiebre38.5, ibuprofeno

consulta2:gastritis, protector_gástrico

paciente2:

clases de equivalencia
semántica

consulta1:dolor_articular, fiebre37.5, antibiótico

consulta2:dolor_articular, febrícula37.5, ibuprofeno

consulta3:gastritis, protector_gástrico

- **item**: un evento en una historia clínica
- **I**: todos los eventos en todas las historias clínicas
- **transacción**: consulta? historia clínica? período?

Ejemplo

- Patrones de navegación de usuarios en la web
- Patrones de aprendizaje en plataformas on-line
- Patrones de fallo de discos rígidos
- Esperanza de vida de animales
- ...

Una regla de asociación $X \rightarrow Y$ es un patrón que dice que cuando ocurre X , ocurre Y con una cierta probabilidad.

Una transacción t contiene X , un conjunto de items (itemset) en I , si $X \subseteq t$.

Una regla de asociación es una implicación:

$$\mathbf{X} \rightarrow \mathbf{Y}, \text{ donde } X, Y \subset I, \text{ y } X \cap Y = \emptyset$$

Un itemset es un conjunto de items.

$$X = \{\text{leche}, \text{pan}, \text{cereal}\}$$

Un k -itemset es un itemset con k items.

$\{\text{leche}, \text{pan}, \text{cereal}\}$ es un 3-itemset

Métricas

Soporte: La regla $X \rightarrow Y$ tiene Soporte sup en T (el dataset de transacciones) si $sup\%$ de las transacciones contienen $X \cup Y$.

$$sup = Pr(X \cup Y).$$

Confianza: La regla $X \rightarrow Y$ tiene Confianza $conf$ en T si $conf\%$ de las transacciones que contienen X también contienen Y .

$$conf = Pr(Y \mid X)$$

Métricas

Soporte: La regla $X \rightarrow Y$ tiene Soporte sup en T (el dataset de transacciones) si $sup\%$ de las transacciones contienen X y Y .

$sup = |$

Confianza

si $conf\%$

también contienen Y .

$conf = Pr(Y \mid X)$

¿Qué van a priorizar estas métricas?

¿Responden a nuestras preguntas?

¿Nos aportan información valiosa?

$conf$ en T

enen X

Objetivo de las reglas de asociación

Encontrar todas las reglas que satisfacen un soporte mínimo y confianza mínima

- Todas las reglas
- No hay items objetivo

Una visión simplista de los datos, porque no incluye:

- cantidad
- precio
- promociones

Algoritmo Apriori

Apriori(T, ϵ)

$L_1 \leftarrow \{\text{large 1-itemsets}\}$

$k \leftarrow 2$

while $L_{k-1} \neq \emptyset$

$C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a\} - \{c \mid \{s \mid s \subseteq c \wedge |s| = k-1\} \not\subseteq L_{k-1}\}$

for transactions $t \in T$

$C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\}$

for candidates $c \in C_t$

$count[c] \leftarrow count[c] + 1$

$L_k \leftarrow \{c \mid c \in C_k \wedge count[c] \geq \epsilon\}$

$k \leftarrow k + 1$

return $\bigcup_k L_k$

Pasos

1. Encontrar todos los itemsets con soporte mínimo (itemsets frecuentes)

{pollo, ropa, leche} [sop = 3/7]

2. Usar los itemsets para generar reglas

ropa \rightarrow leche, pollo [sop = 3/7, conf = 3/3]

Encontrar itemsets frecuentes

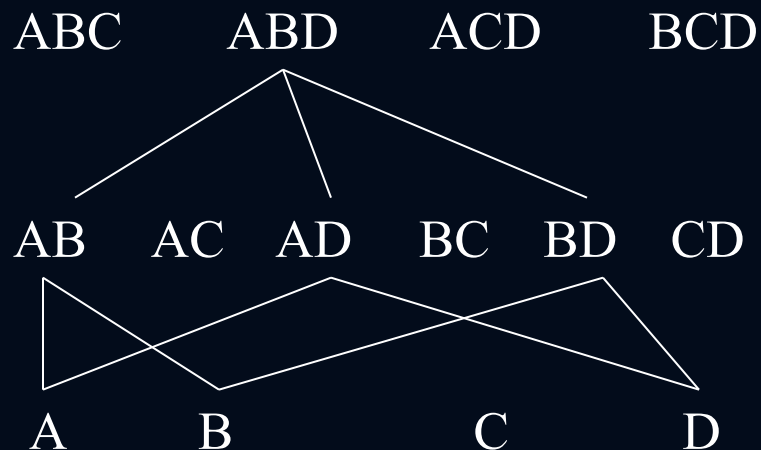
Iterativo (por niveles)

1. Encontrar todos los itemsets frecuentes de 1 item, entonces todos los itemsets frecuentes de 2 items, y así sucesivamente
2. → en cada iteración k , considerar solamente los itemsets que contienen un $(k-1)$ -itemset frecuente (descartar de entrada los itemsets que no contienen un $(k-1)$ -itemset frecuente)
3. Los items están ordenados, para evitar repeticiones

Encontrar itemsets frecuentes

Itemset frecuente \rightarrow Soporte \geq minsup

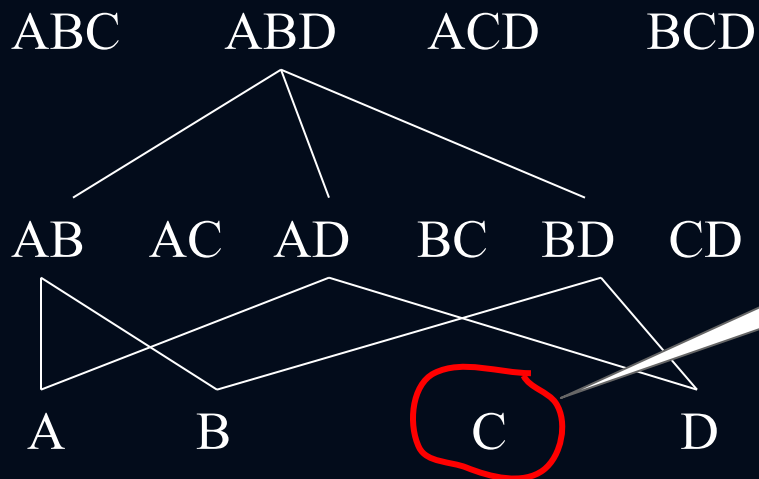
propiedad apriori (downward closure): todos los subconjuntos de un itemset frecuente también son itemsets frecuentes



Encontrar itemsets frecuentes

Itemset frecuente \rightarrow Soporte \geq minsup

propiedad apriori (downward closure): todos los subconjuntos de un itemset frecuente también son itemsets frecuentes

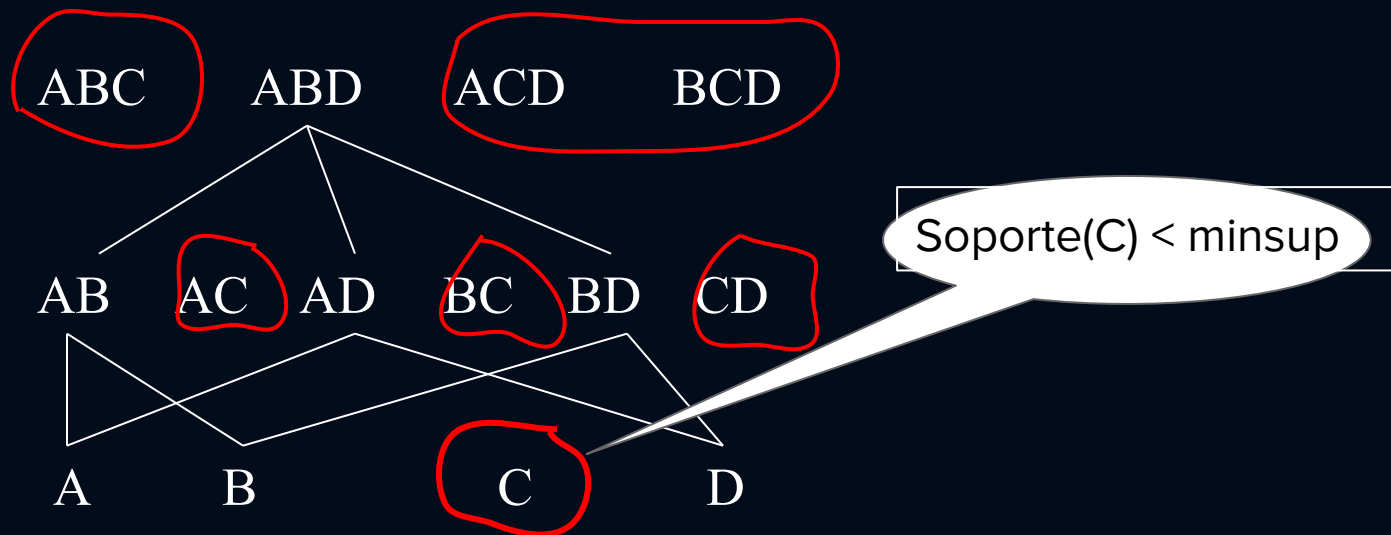


Soporte(C) < minsup

Encontrar itemsets frecuentes

Itemset frecuente \rightarrow Soporte \geq minsup

propiedad apriori (downward closure): todos los subconjuntos de un itemset frecuente también son itemsets frecuentes



Encontrar confianza

Para cada itemset frecuente X ,

Para cada subconjunto no vacío $A \subset X$,

Sea $B = X - A$

$\text{Soporte}(A \rightarrow B) = \text{Soporte}(A \cup B) = \text{Soporte}(X)$

$\text{Confianza}(A \rightarrow B) = \text{Soporte}(A \cup B) / \text{Soporte}(A)$

$A \rightarrow B$ es una regla de asociación si

$\text{Confianza}(A \rightarrow B) \geq \text{minconf}$

Esta información ya se obtuvo en el momento de generación de itemsets, no hay que recorrer el dataset de vuelta

Ejemplo

Supongamos $\{2,3,4\}$ es frecuente, con $\text{sop}=50\%$

Subconjuntos propios no vacíos: $\{2,3\}$, $\{2,4\}$, $\{3,4\}$, $\{2\}$, $\{3\}$, $\{4\}$, con $\text{sop}=50\%$, 50% , 75% , 75% , 75% , 75% respectivamente

Generan estas reglas de asociación:

$2,3 \rightarrow 4$, Confianza=100%

$2,4 \rightarrow 3$, Confianza=100%

$3,4 \rightarrow 2$, Confianza=67%

$2 \rightarrow 3,4$, Confianza=67%

$3 \rightarrow 2,4$, Confianza=67%

Consideraciones sobre Apriori

Parece muy caro pero...

- Búsqueda por niveles, explotando downward closure
- El parámetro k (tamaño del itemset) limita el coste
- Escalable!
- El espacio de todas las reglas de asociación es exponencial, $O(2^m)$, donde m es el número de items en I .
- Explota la sparseness de los datos, los valores altos de Soporte y Confianza.
- Igualmente: un número enorme de reglas!!!

Clustering

Cómo funciona clustering

Agrupar objetos semejantes

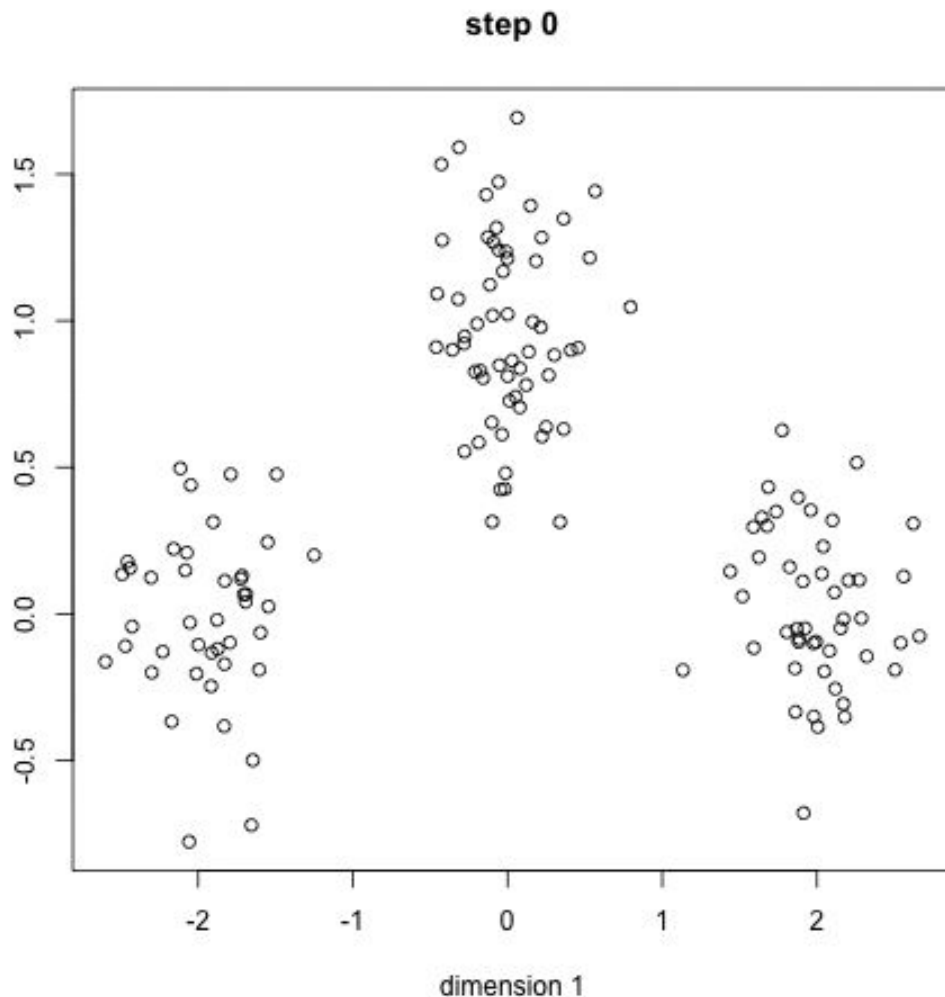
- Entrada: vectores n-dimensionales
- Salida: grupos (clusters) de vectores semejantes → cercanos en el espacio
 - Se minimiza la distancia entre los objetos de un mismo grupo
 - Se maximiza la distancia entre los objetos de distintos clusters

Cómo fu

Agrupar

- Entra
- Salic
- semej
- Se
- ob
- Se
- ob

dimension 2

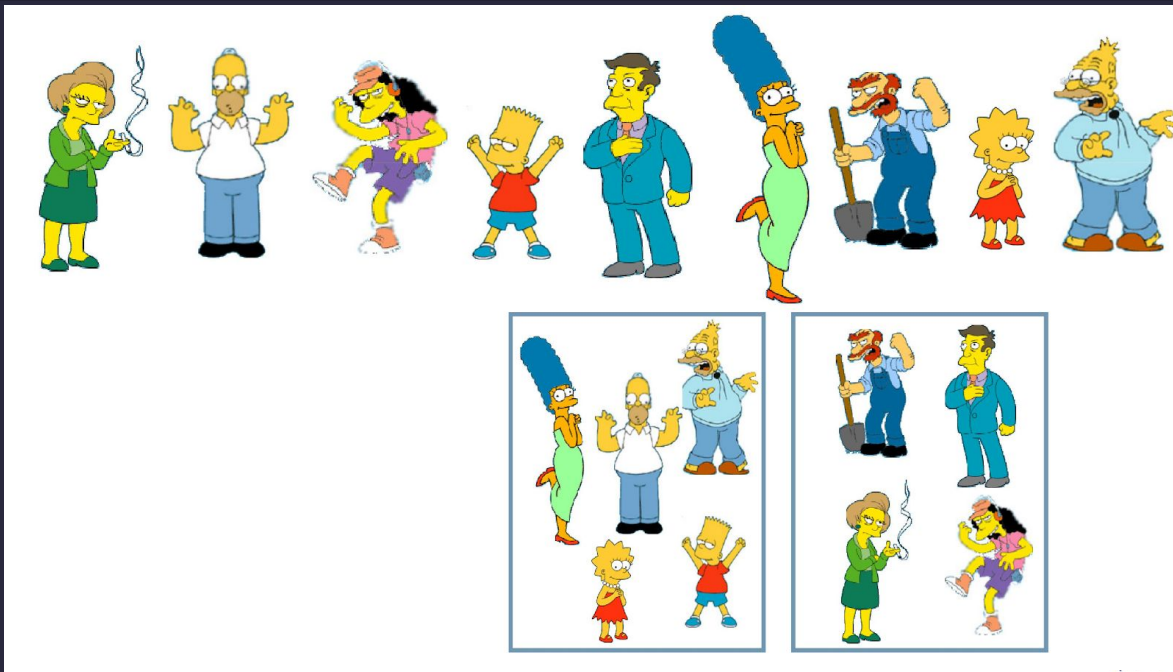


es

OS

OS

Datos



Datos

| | id | sexo | fechnac | educ | catlab | salario | salini | T.emp | expprev | minoría |
|---------|-----|--------|-------------|------|----------------|----------|----------|-------|---------|---------|
| Grupo 1 | 121 | Mujer | 6-ago-1936 | 15 | Administrativo | \$18.750 | \$10.500 | 90 | 54 | No |
| | 122 | Mujer | 26-sep-1965 | 15 | Administrativo | \$32.550 | \$13.500 | 90 | 22 | No |
| | 123 | Mujer | 24-abr-1949 | 12 | Administrativo | \$33.300 | \$15.000 | 90 | 3 | No |
| | 124 | Mujer | 29-may-1963 | 16 | Administrativo | \$38.550 | \$16.500 | 90 | Ausente | No |
| | 125 | Hombre | 6-ago-1956 | 12 | Administrativo | \$27.450 | \$15.000 | 90 | 173 | Sí |
| Grupo 2 | 126 | Hombre | 21-ene-1951 | 15 | Seguridad | \$24.300 | \$15.000 | 90 | 191 | Sí |
| | 127 | Hombre | 1-sep-1950 | 12 | Seguridad | \$30.750 | \$15.000 | 90 | 209 | Sí |
| Grupo 3 | 128 | Mujer | 25-jul-1946 | 12 | Administrativo | \$19.650 | \$9.750 | 90 | 229 | Sí |
| | 129 | Hombre | 18-jul-1959 | 17 | Directivo | \$68.750 | \$27.510 | 89 | 38 | No |
| | 130 | Hombre | 6-sep-1958 | 20 | Directivo | \$59.375 | \$30.000 | 89 | 6 | No |
| | 131 | Hombre | 8-feb-1962 | 15 | Administrativo | \$31.500 | \$15.750 | 89 | 22 | No |
| | 132 | Hombre | 17-may-1953 | 12 | Administrativo | \$27.300 | \$17.250 | 89 | 175 | No |
| | 133 | Hombre | 12-sep-1959 | 15 | Administrativo | \$27.000 | \$15.750 | 89 | 87 | No |

¿Y el aprendizaje automático?

minería = aprendizaje NO supervisado

Aprendizaje supervisado:

- desde un subconjunto de los atributos X a otro subconjunto Y
- a partir de ejemplos con X e Y
- para predecir Y en nuevos casos

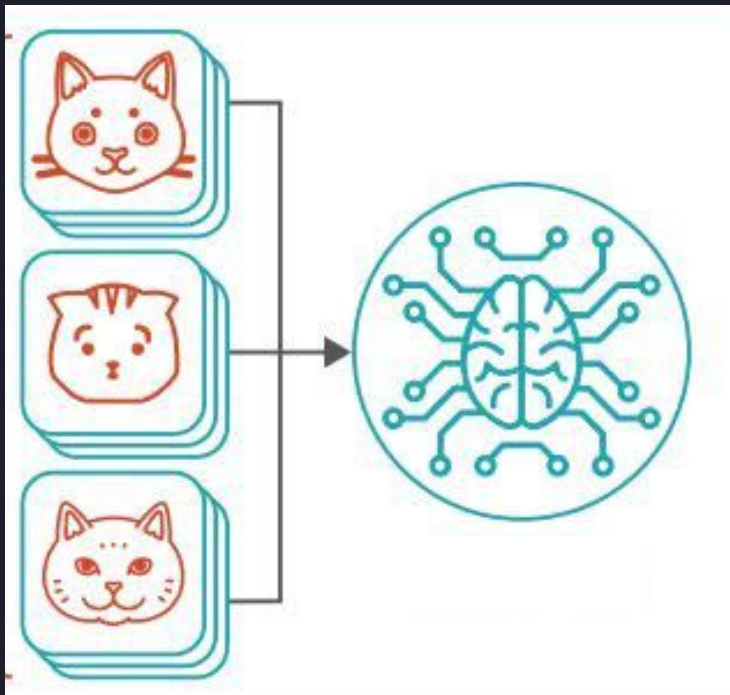
¿cómo funciona?



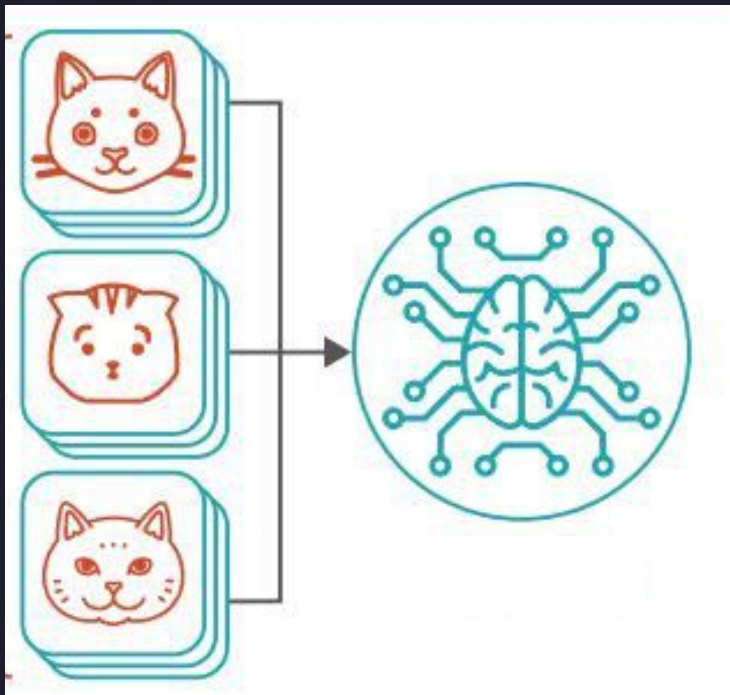
¿cómo funciona?



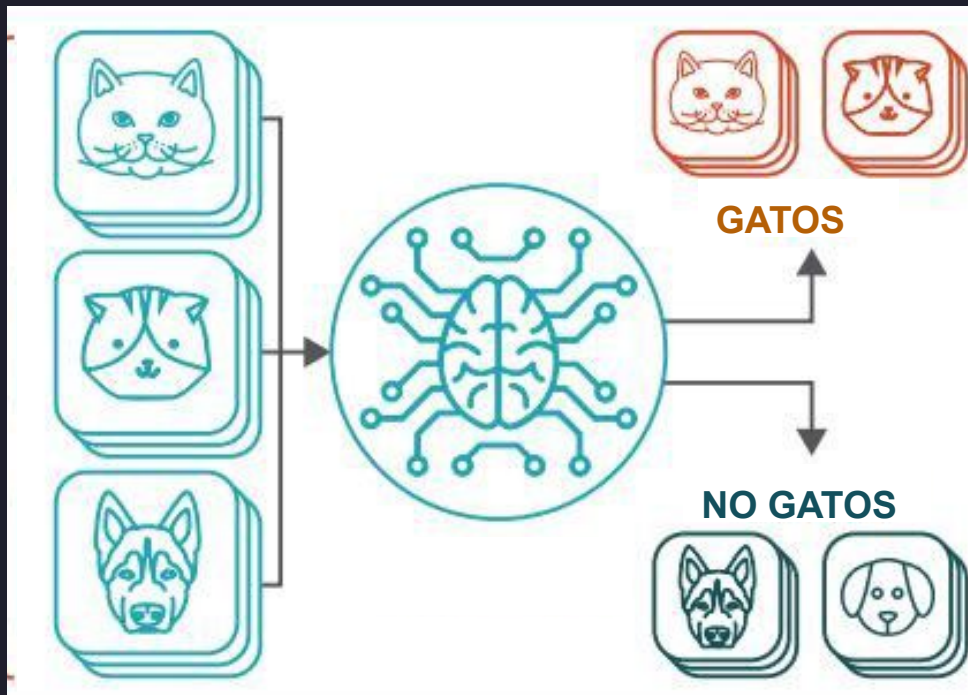
¿cómo funciona ahora?

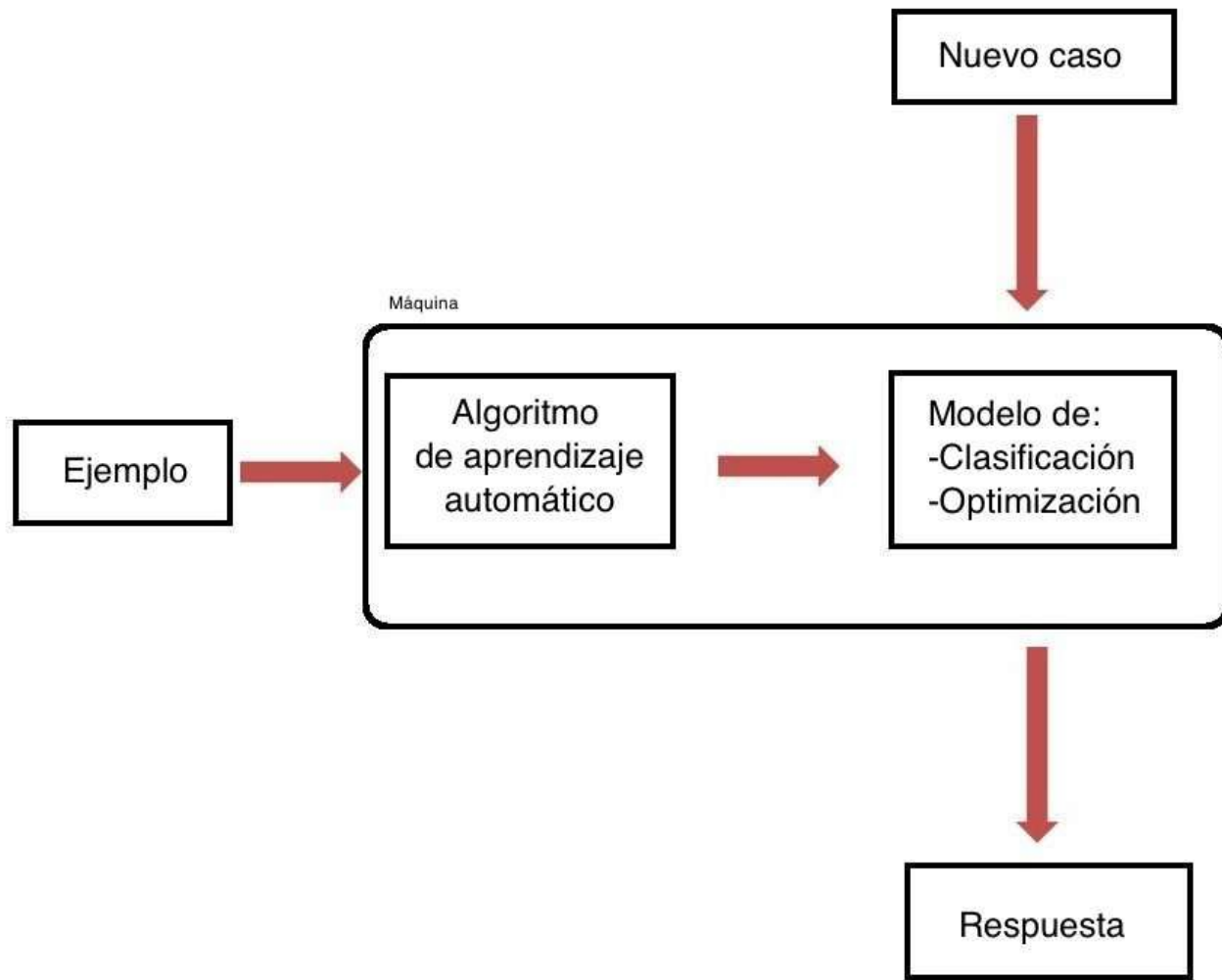


¿cómo funciona ahora?

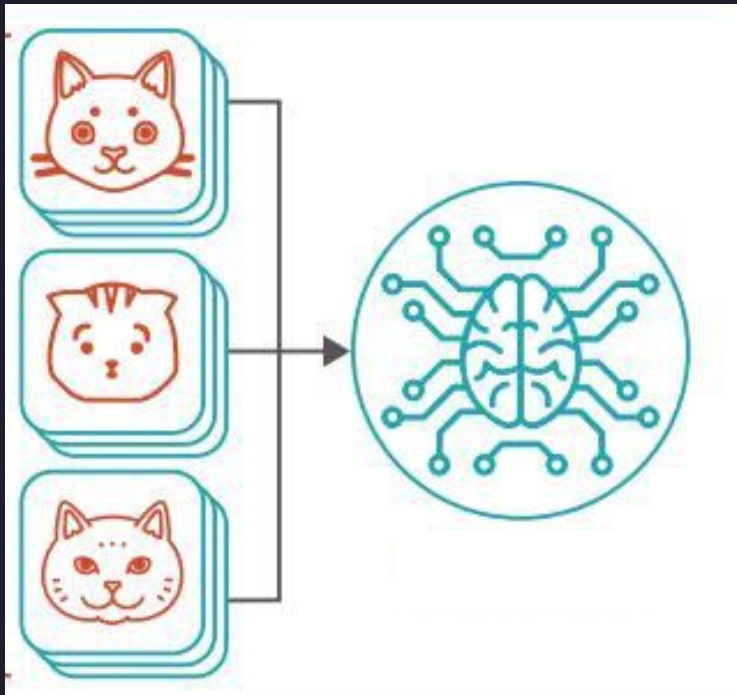


¿cómo funciona ahora?

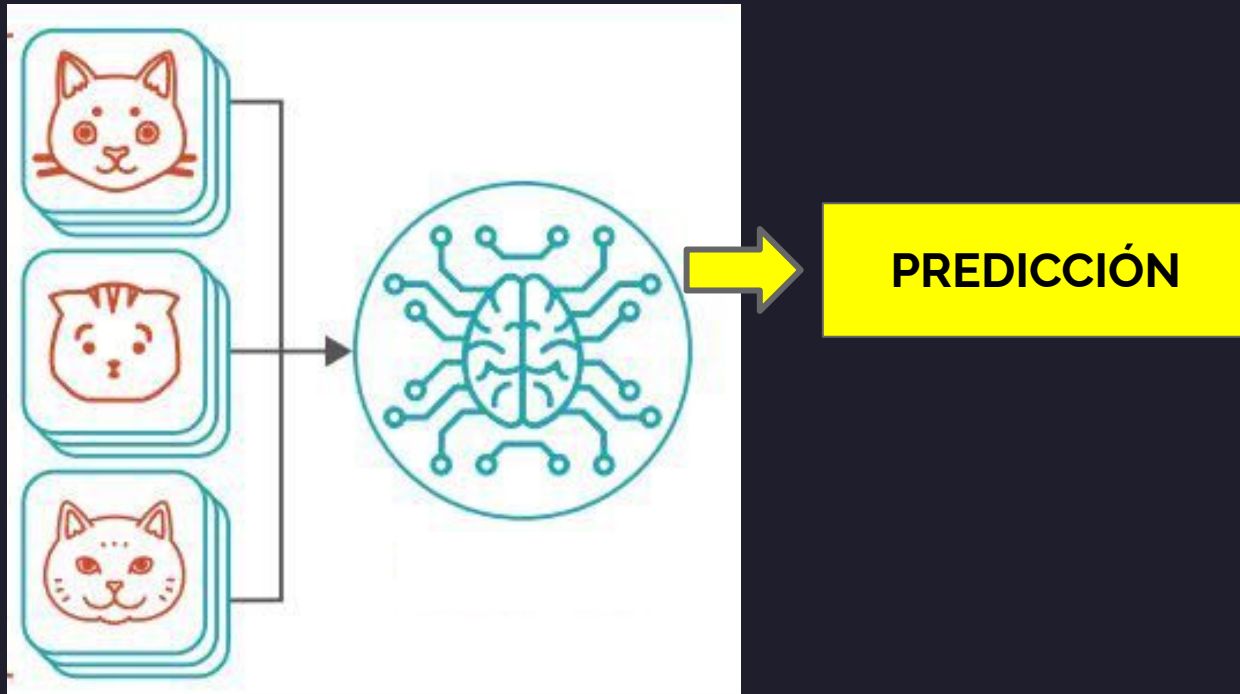




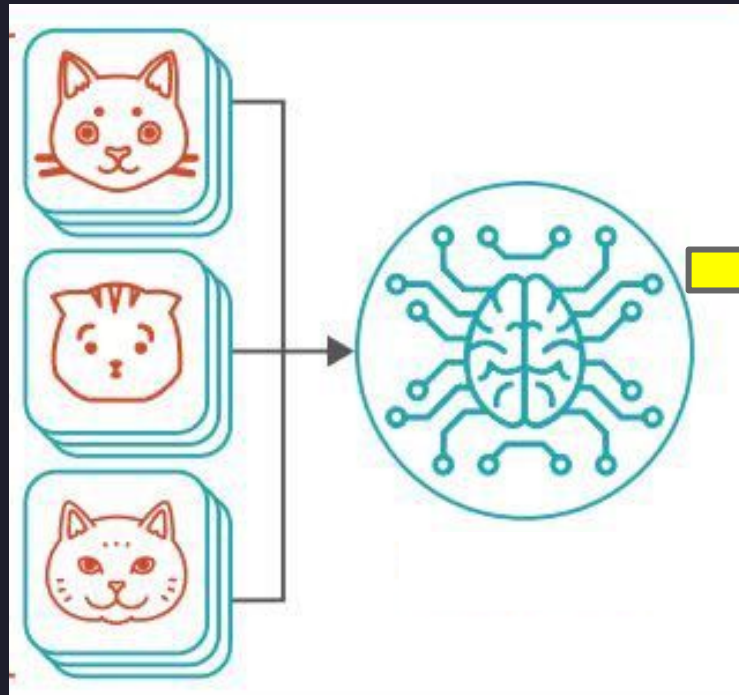
¿cómo funciona ahora?



¿cómo funciona ahora?



¿cómo funciona ahora?

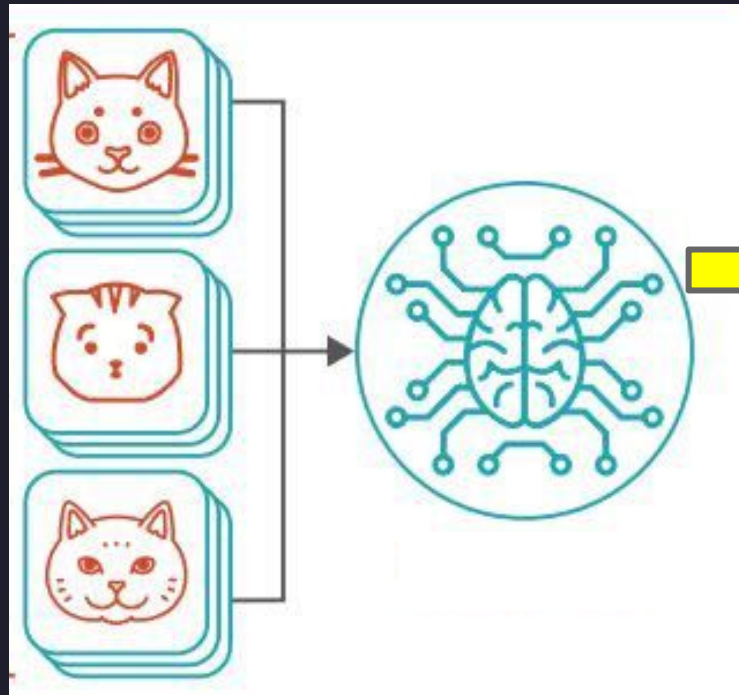


PREDICCIÓN

**99% de
acierto!**



¿cómo funciona ahora?



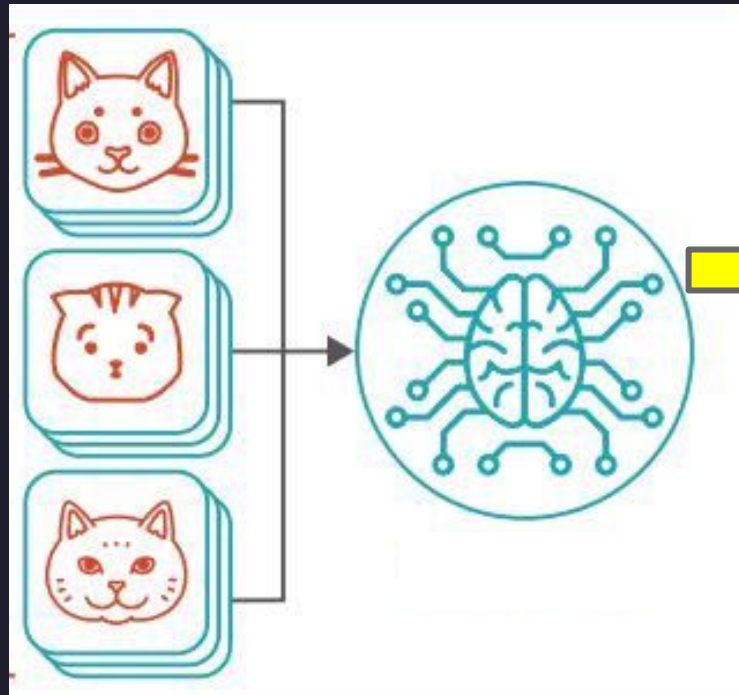
PREDICCIÓN

**100%
no cáncer**

**99% de
acierto!**



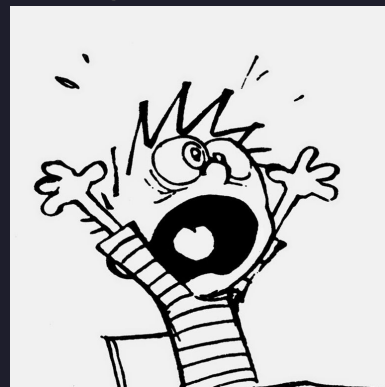
¿cómo funciona ahora?



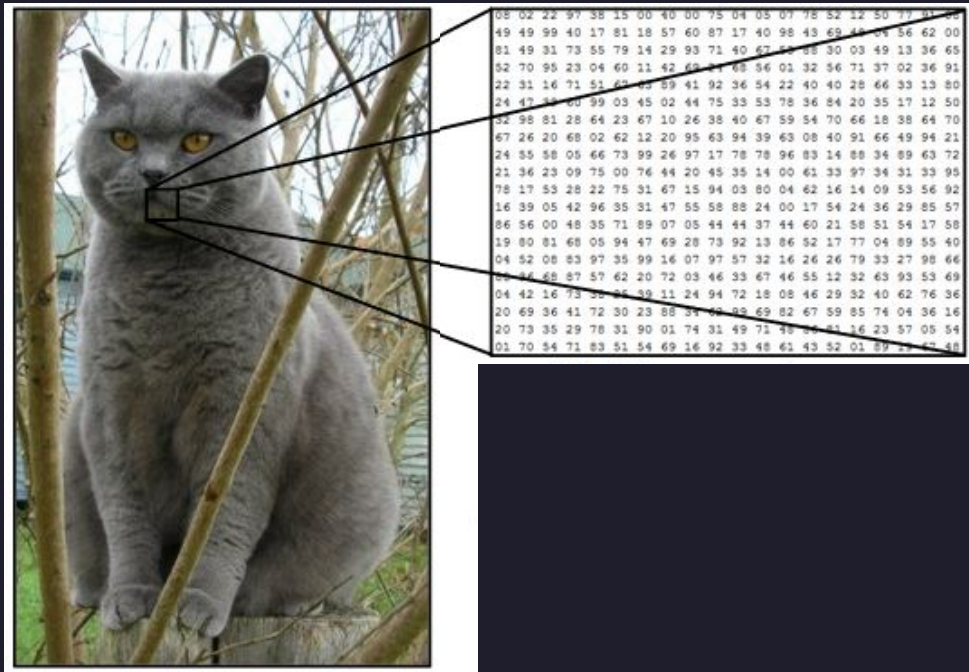
PREDICCIÓN

**1% de
cáncer**

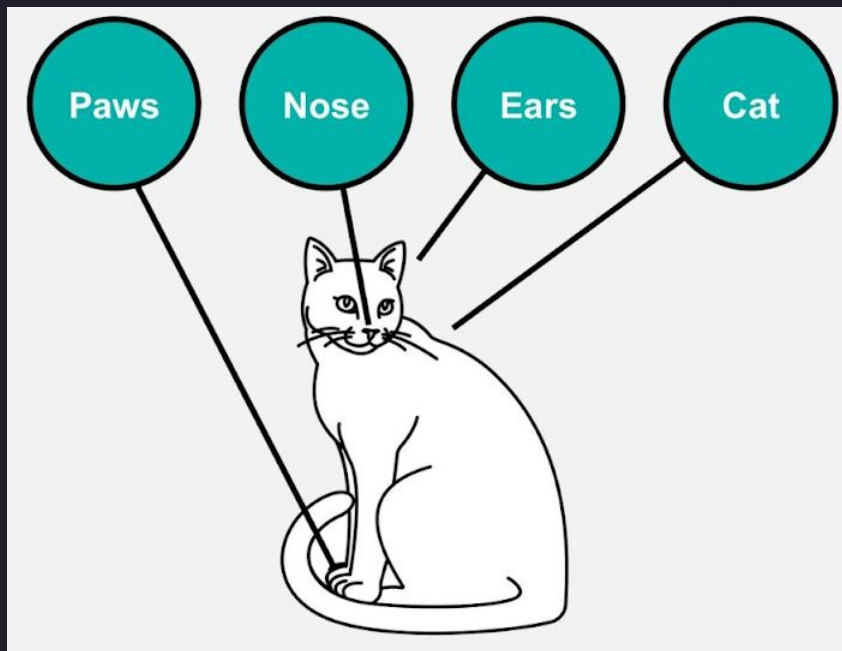
**100%
no cáncer**



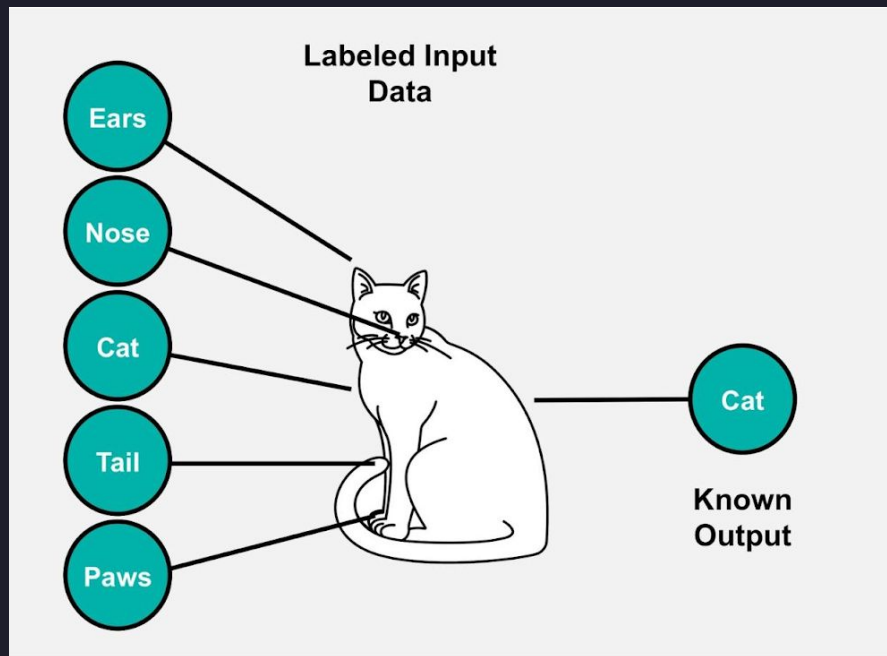
¿cómo funciona ahora?



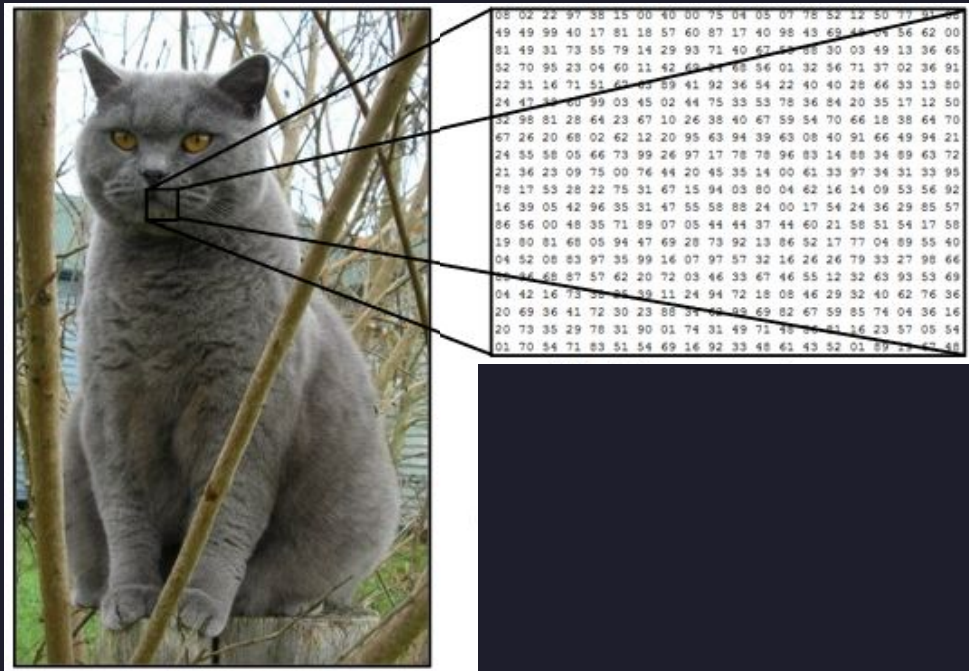
¿cómo funciona ahora?



¿cómo funciona ahora?



¿cómo funciona ahora?



¿por qué ahora?



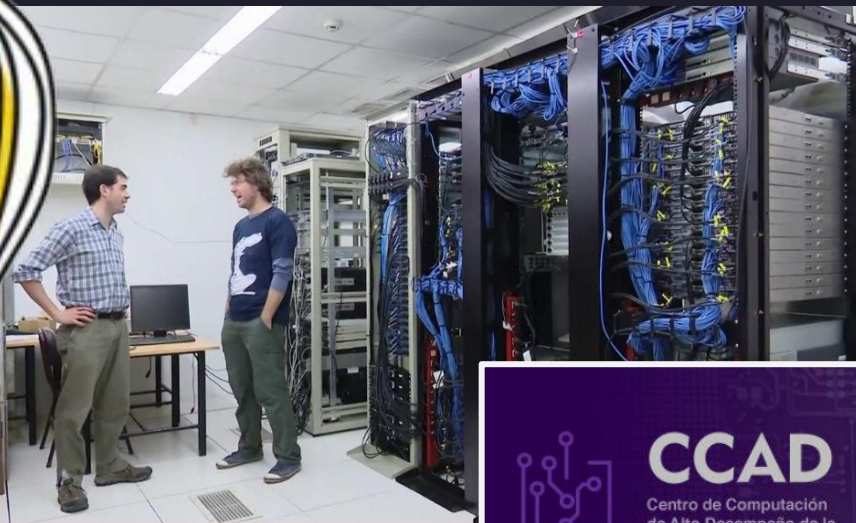
¿por qué ahora?



CCAD

Centro de Computación
de Alto Desempeño de la
Universidad Nacional de Córdoba

¿por qué ahora?



CCAD

Centro de Computación
de Alto Desempeño de la
Universidad Nacional de Córdoba

¿Y la inteligencia artificial?

Cualquier cosa con un comportamiento
inteligente

/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

