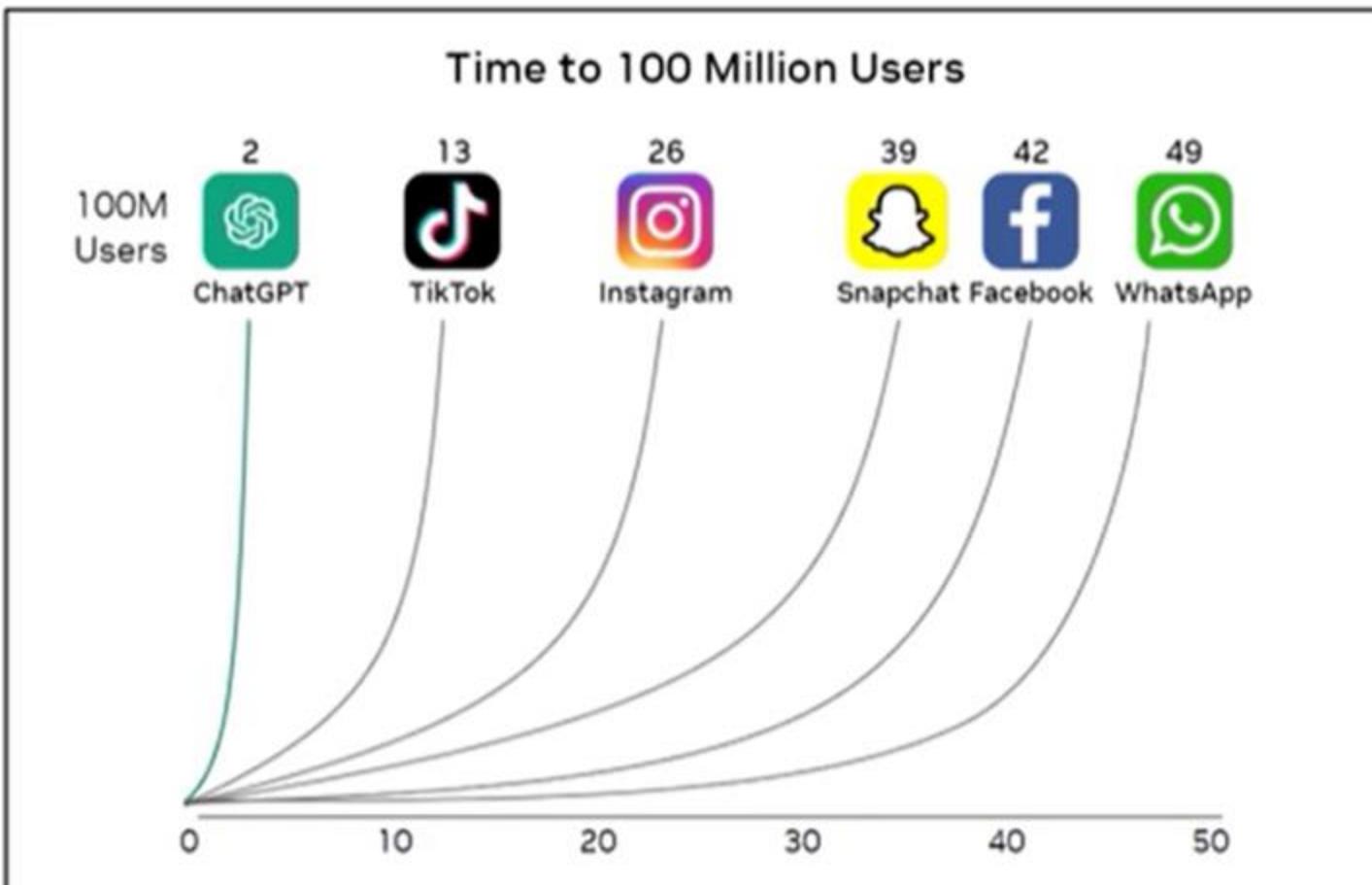


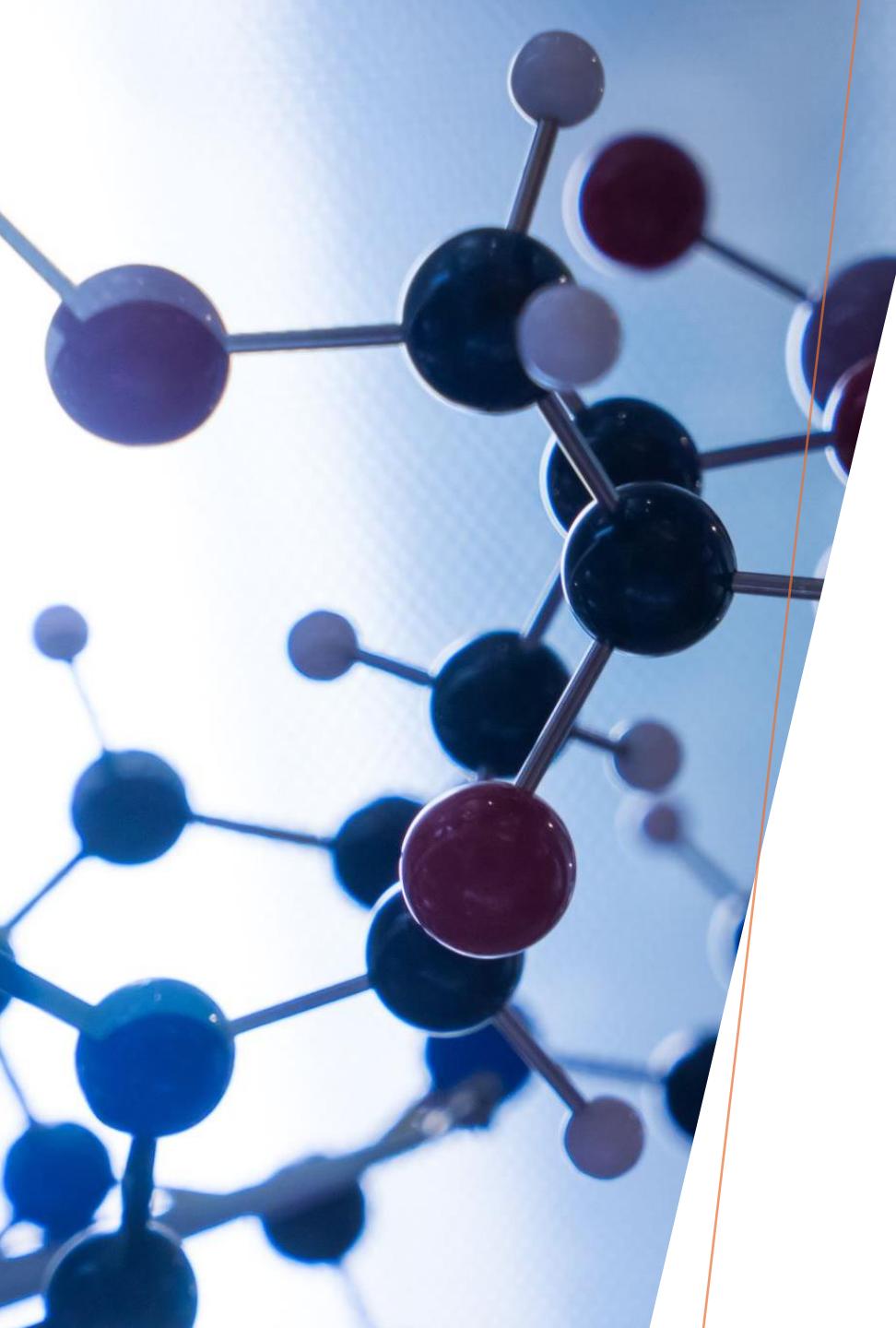
APRENDIZAJE AUTOMÁTICO Y APRENDIZAJE PROFUNDO



Key Data Center Trends

As data centers become power-limited, the demand for compute grows





ORDEN DEL DÍA

- Introducción a la IA
- Algoritmos de Aprendizaje Automático
- Regresión Lineal
- Introducción al Aprendizaje Profundo
- Ejemplo de Funcionamiento de una Red Neuronal
- Tipos de Modelos de IA Generativa
- Limitaciones de la IA Generativa
- Casos de Uso de la IA Generativa
- Modelos de Lenguaje Grande (LLM)
- Limitaciones de LLM
- Comparación CPU vs GPU
- Servidores de Centro de Datos



INTRODUCCIÓN A LA IA

IA

Sistemas informáticos capaces de aplicar razonamiento

Imitan las capacidades cognitivas del humano

Aprendizaje automático

Permite a los sistemas predecir y clasificar datos

Responde a datos en constante cambio

Aprendizaje profundo

Tipo de aprendizaje automático

Realiza tareas complejas inspiradas en redes neuronales

ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

- Algoritmos binarios
 - Producen resultados de 0 o 1
- Algoritmos clásicos
 - Árbol de decisión
 - Regresión lineal
 - Regresión logística



ÁRBOL DE DECISIÓN



- Árbol de decisión como diagrama de flujo
 - Incluye nodos, ramas y hojas
- Nodos
 - Representan las preguntas
- Ramas
 - Posibles respuestas
- Hojas
 - Representan la decisión

[¿Es la fruta roja?]

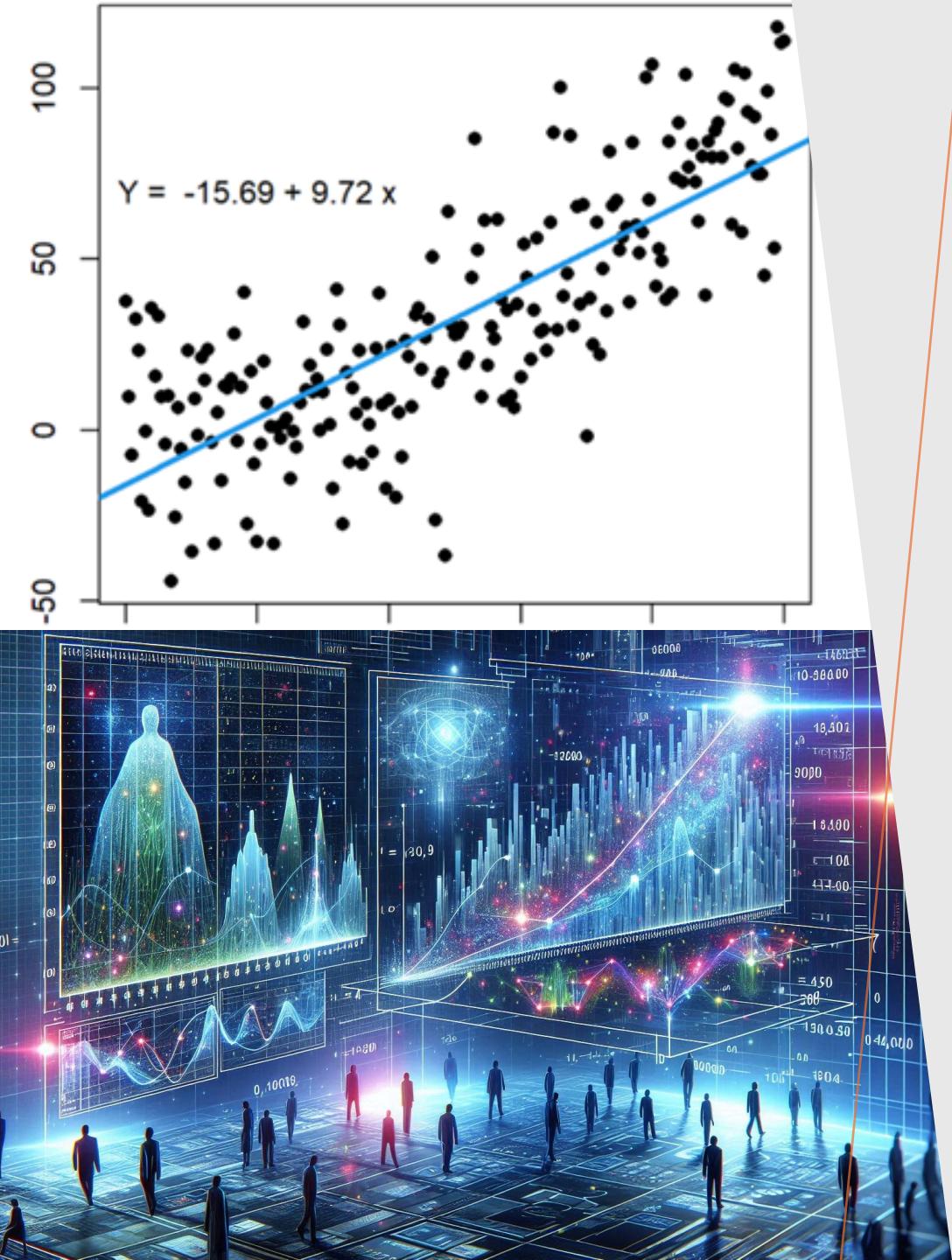
/ \
 Sí No

[¿Es la fruta pequeña?] [¿Es la fruta grande?]

/ \ / \
 Sí No Sí No
/ \ / \
Manzana Cereza Sandía Plátano

REGRESIÓN LINEAL

- Método estadístico
 - Modela más de una variable dependiente
 - Utiliza una o más variables independientes
- Objetivo
 - Encontrar una línea recta que mejor se ajuste a los datos
 - Predecir el valor de la variable dependiente a partir de la variable independiente
- Ejemplo
 - Relación entre ventas y publicidad
 - Cómo suben las ventas con el aumento de publicidad



REGRESIÓN LOGÍSTICA

Definición

Técnica estadística para modelar problemas de clasificación binaria

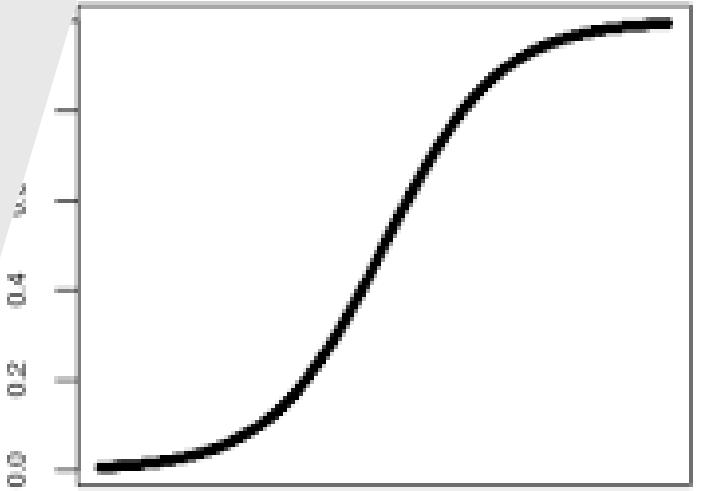
Diferencias con regresión lineal

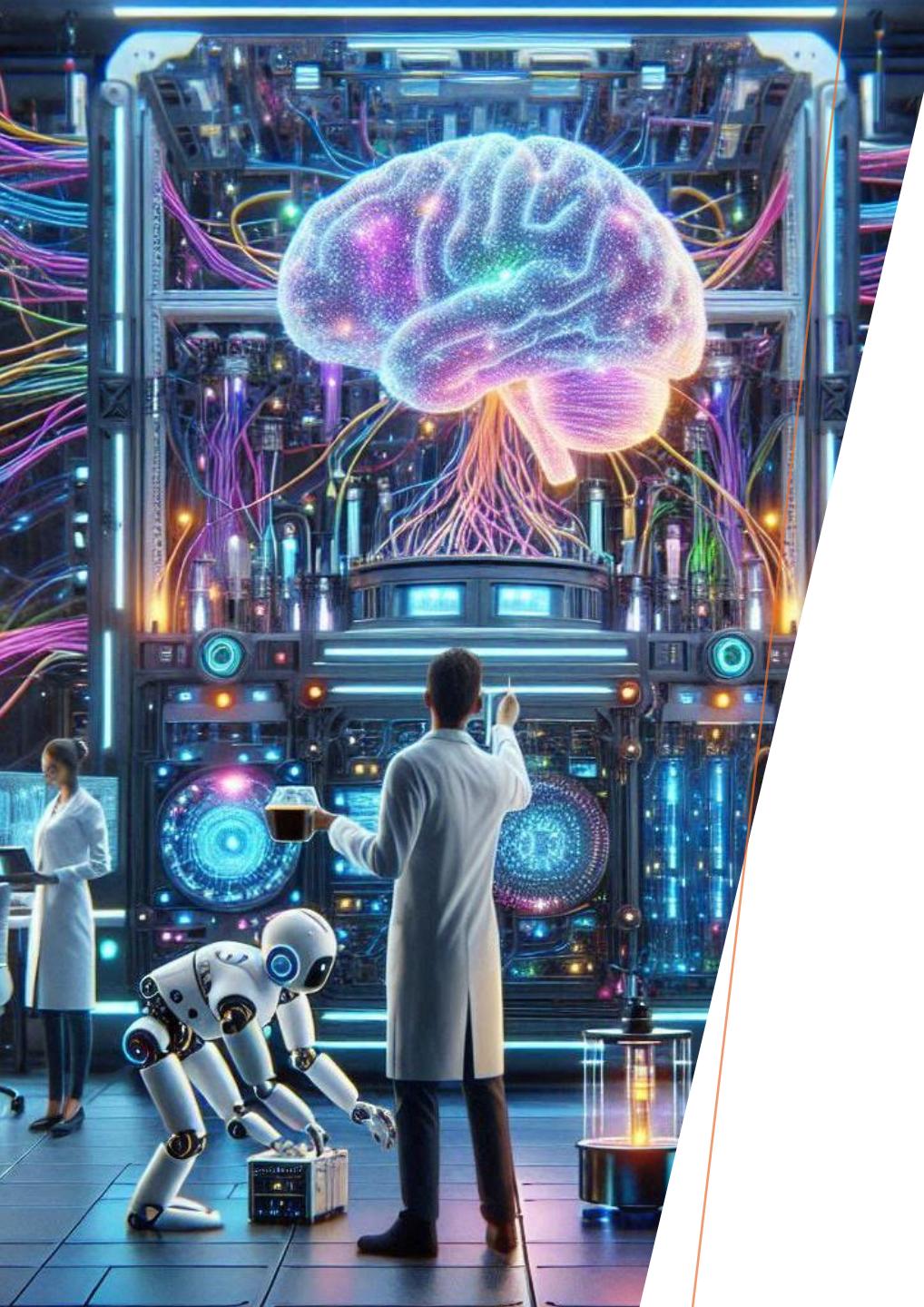
Predice probabilidades en lugar de valores continuos

Aplicaciones

Predicción de pertenencia a una clase

Ejemplo: Predicción de compra de un producto



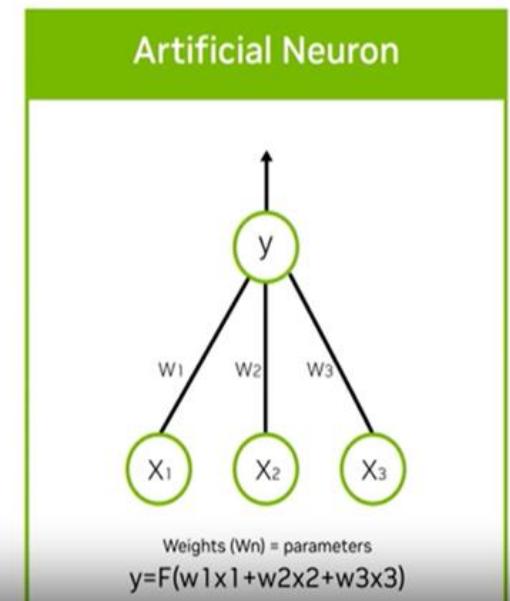
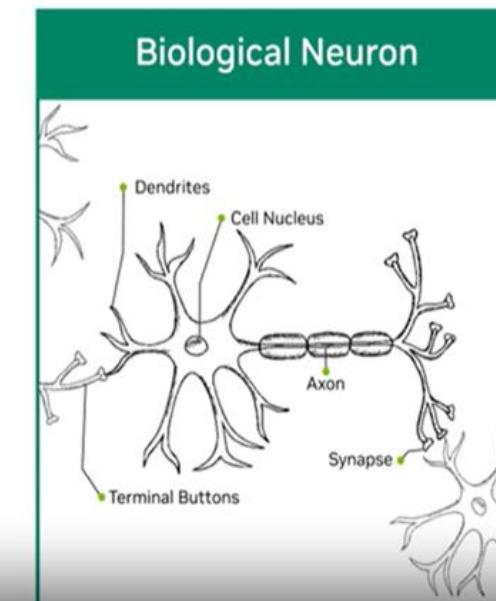
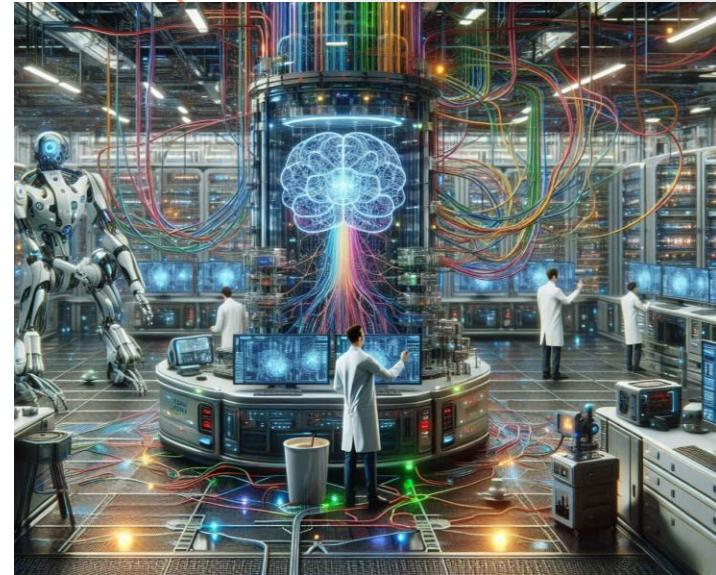


INTRODUCCIÓN AL APRENDIZAJE PROFUNDO

- Inspiración en redes neuronales
 - Basado en el funcionamiento del cerebro humano

FUNCIONAMIENTO DEL CEREBRO HUMANO

- Interconexión de miles de neuronas
 - Permiten ver, entender, sentir, crear, imaginar y tomar decisiones
- Estímulos eléctricos
 - Cada neurona recibe un estímulo eléctrico de otra neurona
 - Las neuronas receptoras , se activan y generan estímulos a otras
- Red neuronal
 - Esta dinámica va generando un proceso de aprendizaje
- Proceso de aprendizaje
 - Conexiones neuronales se fortalecen con la práctica
 - Neuronas activadas se van acostumbrando



APRENDIZAJE PROFUNDO

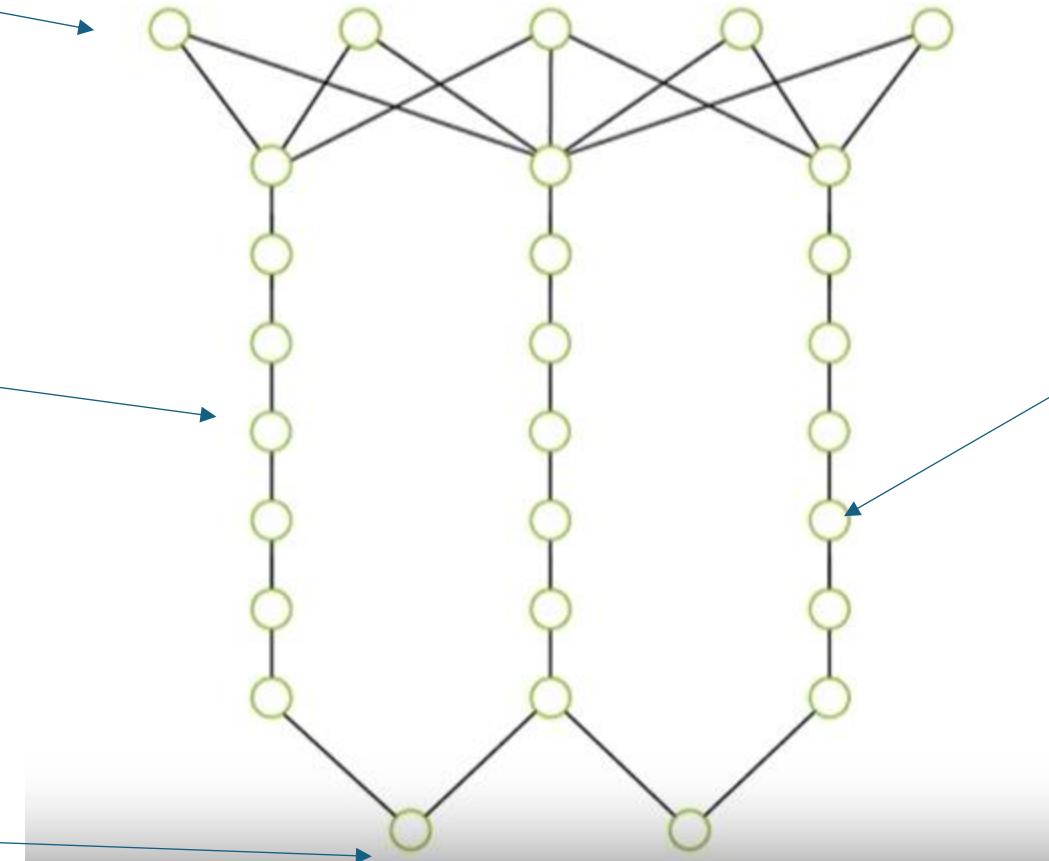
ENTRADA

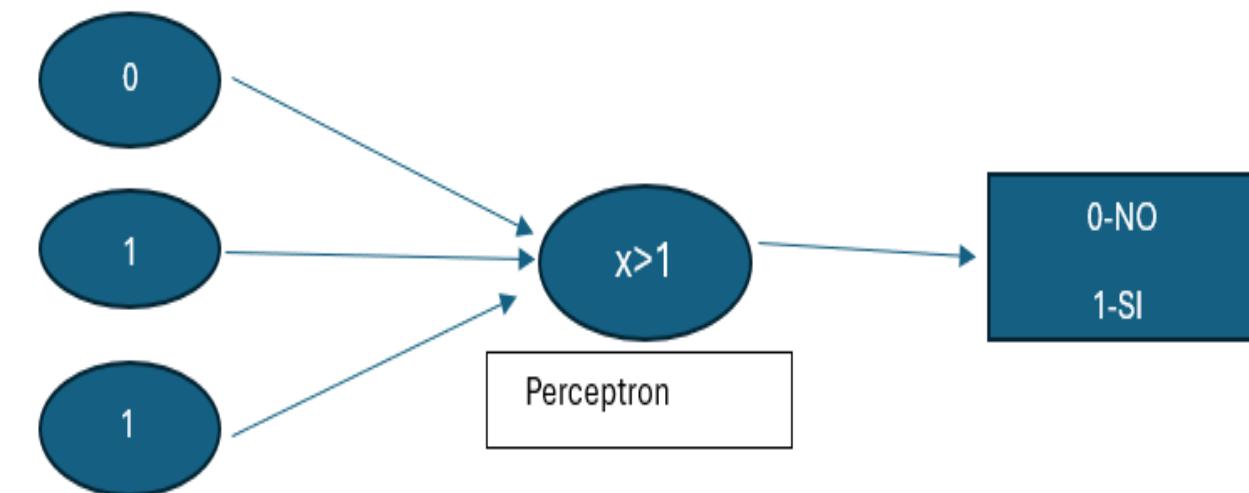
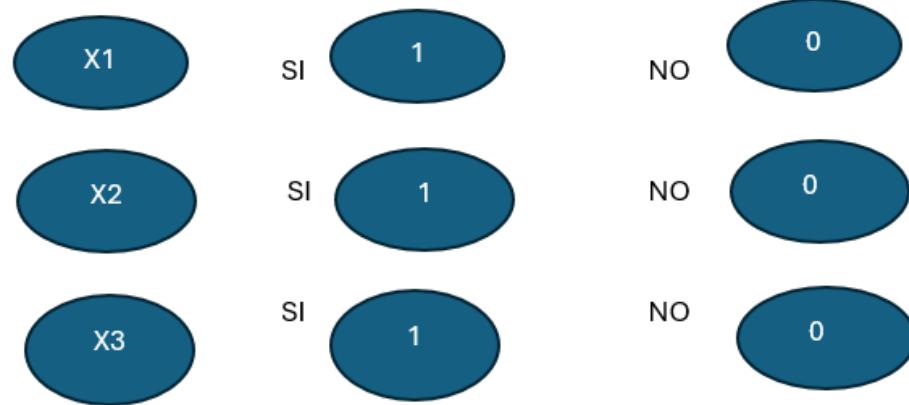
NODOS OCULTOS

SALIDA

NEURONAS O PERCEPTRONES

PERROS O GATOS O MAPACHES

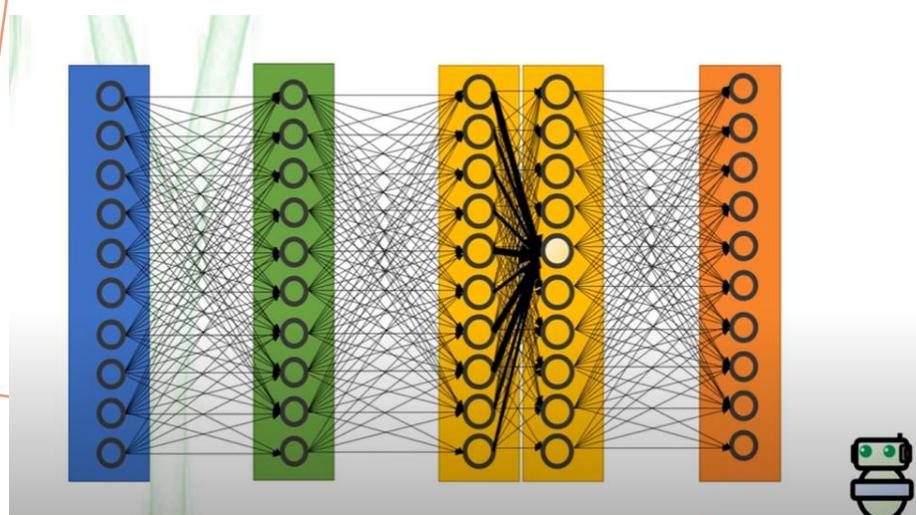
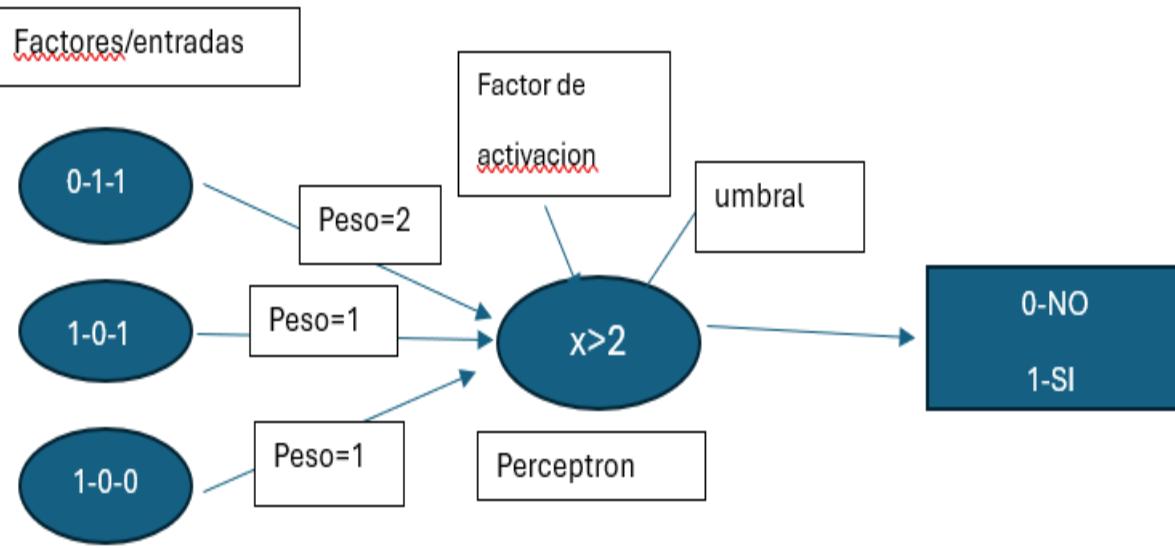




EJEMPLO QUIERO IRME DE VIAJE

- Factores a considerar
 - Tienes dinero suficiente (x_1)
 - Tu pareja quiere ir (x_2)
 - El lugar tiene clima agradable (x_3)
- Respuestas posibles
 - SI -1
 - NO-0
- Decisiones básicas
 - Establecemos un umbral 1

TOMA DE DESICIONES

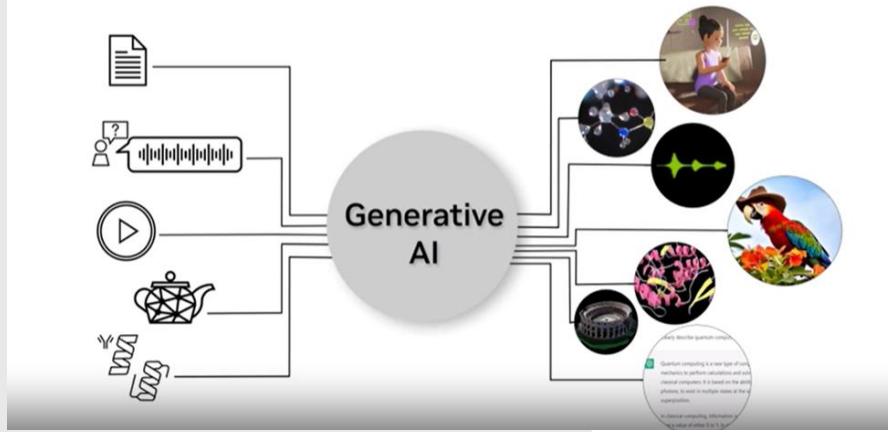


Decisiones más complejas

- Si no tengo dinero, no viajo
- Sino va mi pareja no viajo
- Sino hay buen clima no viajo
- Sino no tengo plata , pero mi pareja si viaja y el clima es bueno
- Si tengo plata, no viaja mi pareja pero es clima es bueno

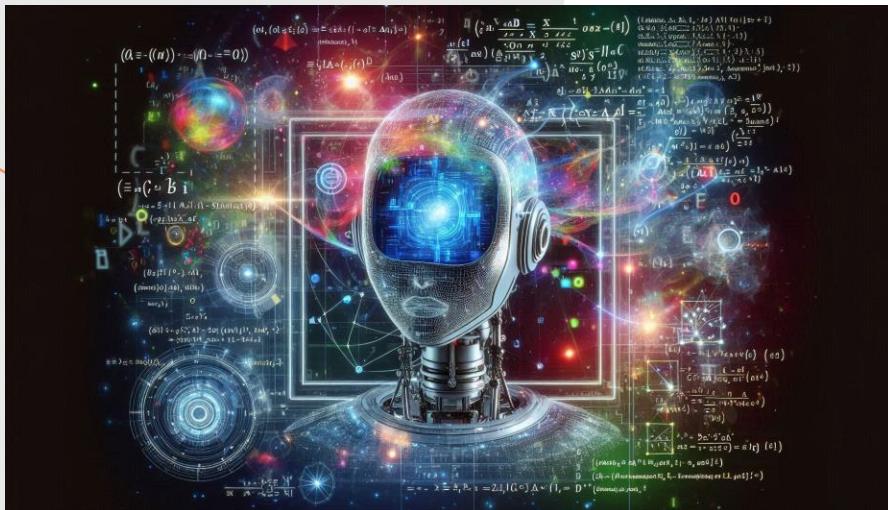
Realidad más compleja

What is Generative AI?



INTELIGENCIA ARTIFICIAL GENERATIVA

- Sistemas de IA de aprendizaje profundo
 - Generan contenido en función de un mensaje
 - Contenido puede ser video, imagen o texto



TIPOS DE LLM DE IA GENERATIVA

GPT (Generative Pre-trained Transformer)

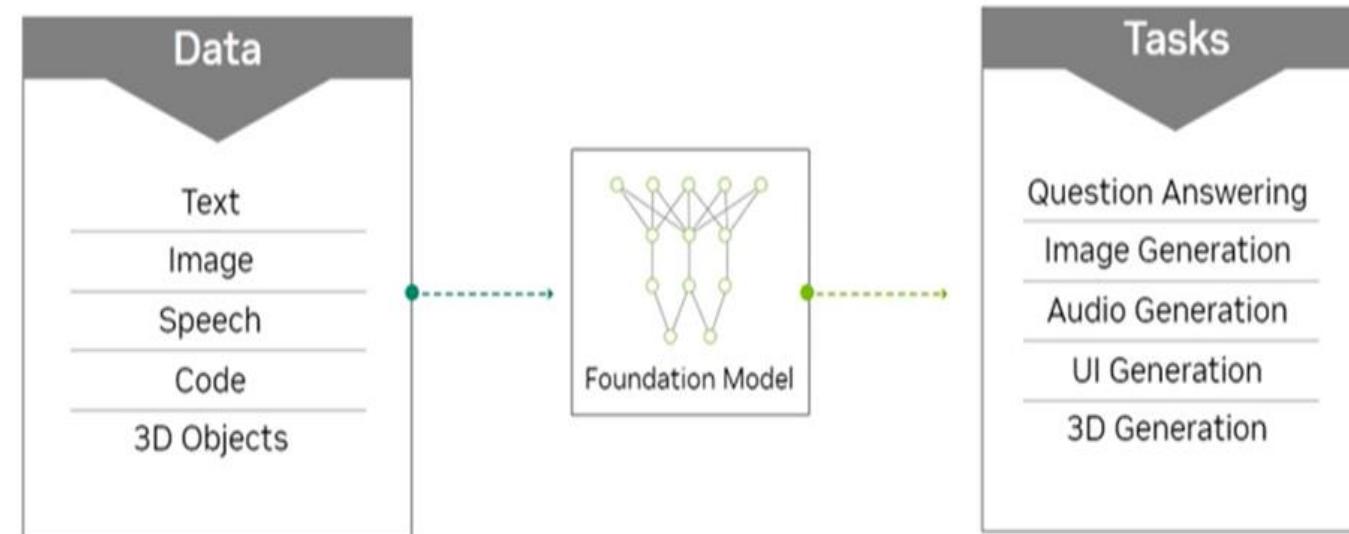
- **Modelo preentrenado generativo**
 - Ejemplo:chatboot-github-copilot

BERT (Bidirectional Encoder Representations from Transformers)

- Representaciones de codificadores bidireccionales
 - Ejemplo google-QA

T5 (Text-To-Text Transfer Transformer)

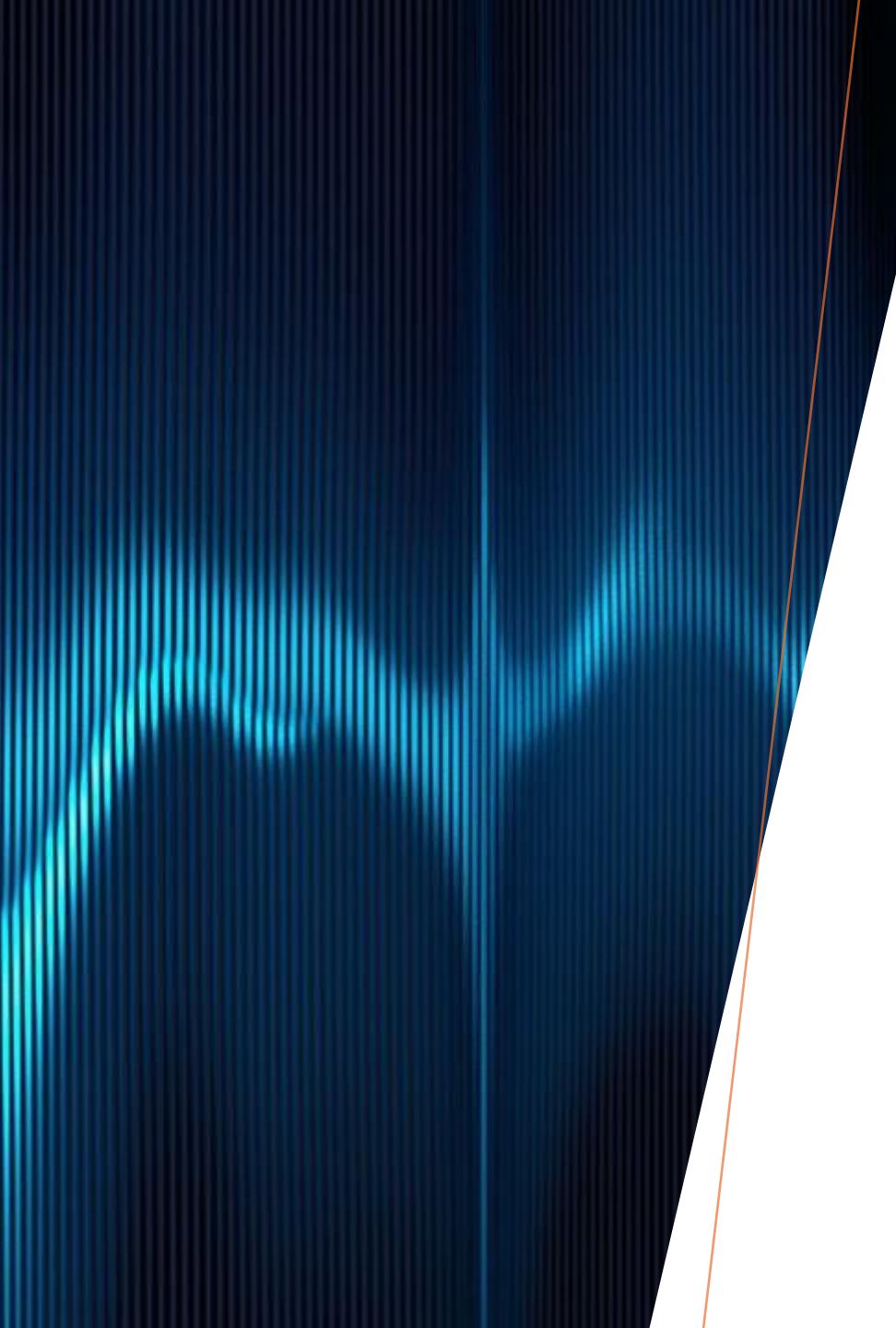
- Transformador de transferencia de texto a texto
 - Ejemplo: corrección gramatical, correctores ortográficos





TIPOS DE MODELOS DE IA GENERATIVA: MODELOS DE IMÁGENES

- GANs (Generative Adversarial Networks)
 - Redes neuronales que generan imágenes nuevas a partir de un conjunto de datos.
 - Deepart-stylegan
- VAEs (Variational Autoencoders)
 - Modelos que codifican y decodifican datos para generar imágenes.
 - Herramientas que modelan caras
- DALL-E
 - Modelo que crea imágenes a partir de descripciones textuales.
 - Prototipado y diseño rápido



TIPOS DE MODELOS DE IA GENERATIVA: MODELOS DE AUDIO

- WaveNet
 - Modelo generativo de audio
 - Ejemplo google assistent
- Jukedeck
 - Plataforma de creación musical
 - Tik tok
- Amper Music
 - Herramienta de composición musical
 - Suno



TIPOS DE MODELOS DE IA GENERATIVA: MODELOS DE VIDEO

- **GANs aplicados al video**
 - Generative Adversarial Networks (GANs) se utilizan para generar videos realistas
- **Neural Radiance Fields (NeRFs)**
 - NeRFs se emplean para la representación y generación de escenas 3D

TIPOS DE MODELOS DE IA GENERATIVA: MODELOS MULTIMODALES

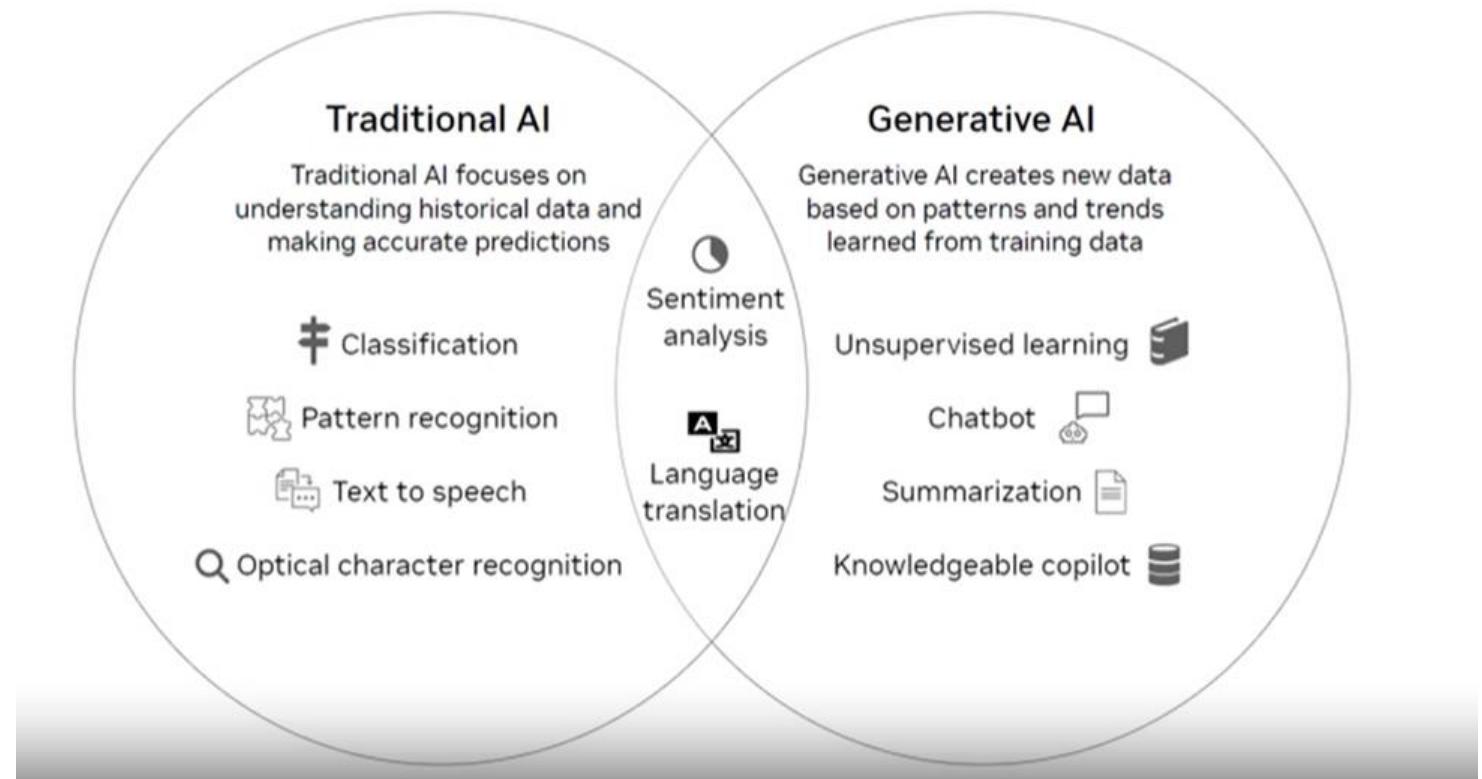
- Modelos Generativos Multimodales
- CLIP (Contrastive Language-Image Pre-training)
- Dall-E
- CLIP
- IMAGEN
- MUSE
- GPT-4



COMPARACIÓN IA TRADICIONAL VS IA GENERATIVA

IA TRADICIONAL
CLASIFICA
RELACIONA PALABRAS
ENCUENTRA PATRONES
IA GENERATIVA
APRENDE
ENTIENDE
RESUME

When to Use Generative AI to Solve Enterprise Challenges





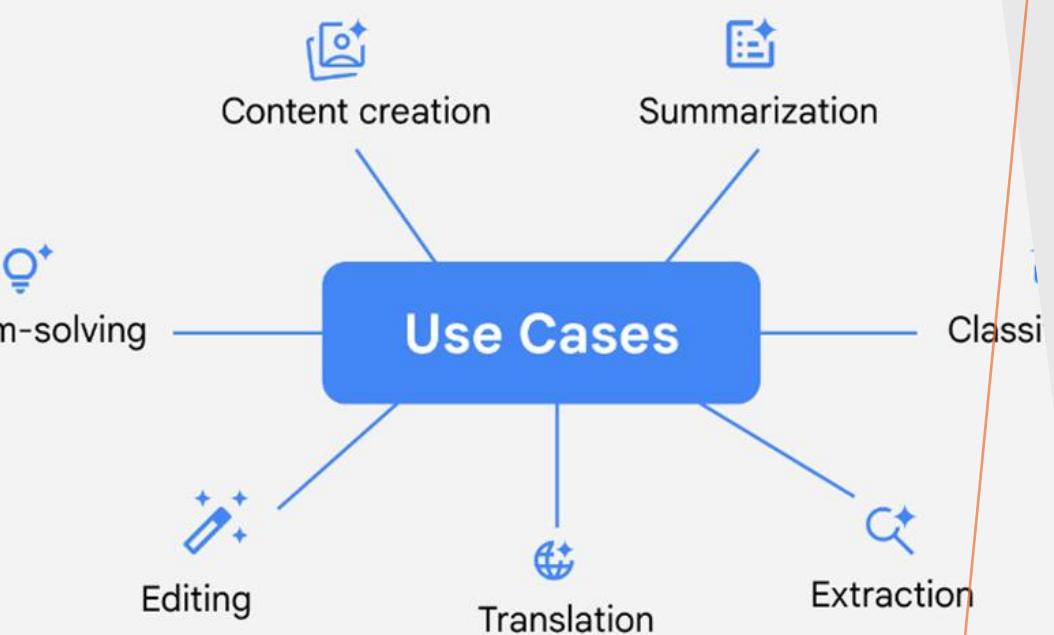
LIMITACIONES DE LA IA GENERATIVA

- Falta de Originalidad
 - Genera contenido repetitivo o predecible
- Datos incompletos
 - Puede producir resultados inexactos
- Sesgos
 - Refleja prejuicios presentes en los datos de entrenamiento SISTEMICOS Y DE DATOS
- Discriminatorios
 - Puede perpetuar estereotipos negativos

ÉTICA EN LA IA GENERATIVA

- Deepfakes
 - Contenido falso
- Propiedad intelectual
 - Derechos de autor
- Privacidad
- Pérdida del toque humano
- Desempleo y desplazamiento laboral





CASOS DE USO DE LA IA GENERATIVA

Crear contenido

- Generar textos originales y creativos

Resumir

- Condensar información extensa en puntos clave

Clasificar

- Organizar datos en categorías específicas

Extraer

- Obtener información relevante de grandes volúmenes de datos

Traducir

- Convertir texto de un idioma a otro

Editar

- Revisar y mejorar la calidad del texto



INGENIERÍA DE INSTRUCCIÓN EN LLM

- Ingeniería de Instrucción
 - Se divide en tres partes
 - La Instrucción
 - La iteraccion
 - La Satisfaccion

MODELOS DE LENGUAJE GRANDE (LLM)

Capacidades del LLM

- Recibe gran cantidad de texto
- Genera respuestas

Identificación de patrones

- Predicción de palabras

Uso de estadísticas de palabras

- Coloca la palabra con mayor probabilidad

CÓMO TRABAJAR CON LLM

Proporciona el contexto adecuado

- Identifica el público objetivo
- Determina el tono adecuado
- Estructura los resultados correctamente
- Define la finalidad del resultado

Proporciona referencias

Evalúa tus resultados

- Verifica la precisión
- Comprueba la imparcialidad
- Asegúrate de que la información sea suficiente
- Revisa la pertinencia

Adopta un enfoque iterativo

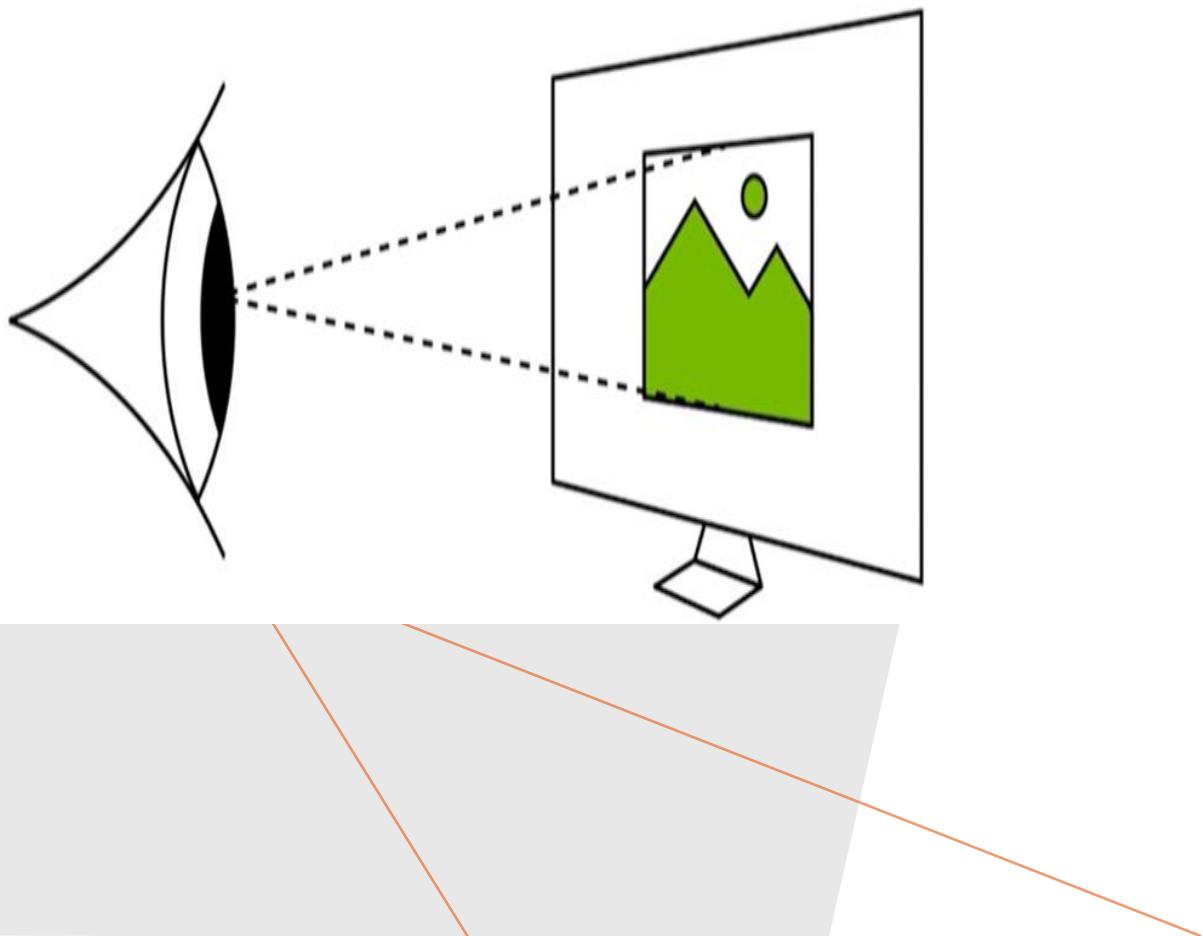


LIMITACIONES DE LLM



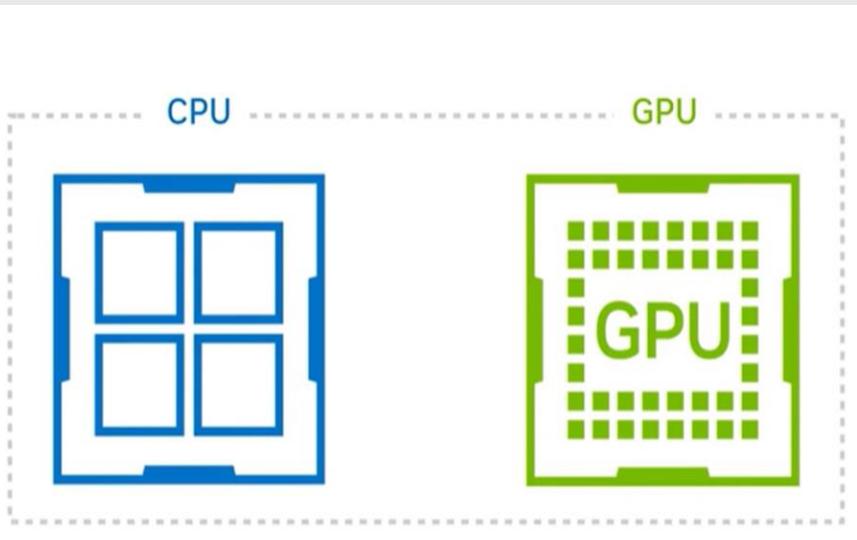
- Sesgado
 - Prejuicios injustos impuestos por la sociedad
 - Ejemplo: cuestiones de género
- Insuficiencia
 - Resultados no son suficientes
 - Entrenamiento con información insuficiente
- Alucinaciones
 - Resultados no ciertos
 - IA arroja información incorrecta
- Verificación
 - Siempre verificar los resultados

INFRAESTRUCTURA



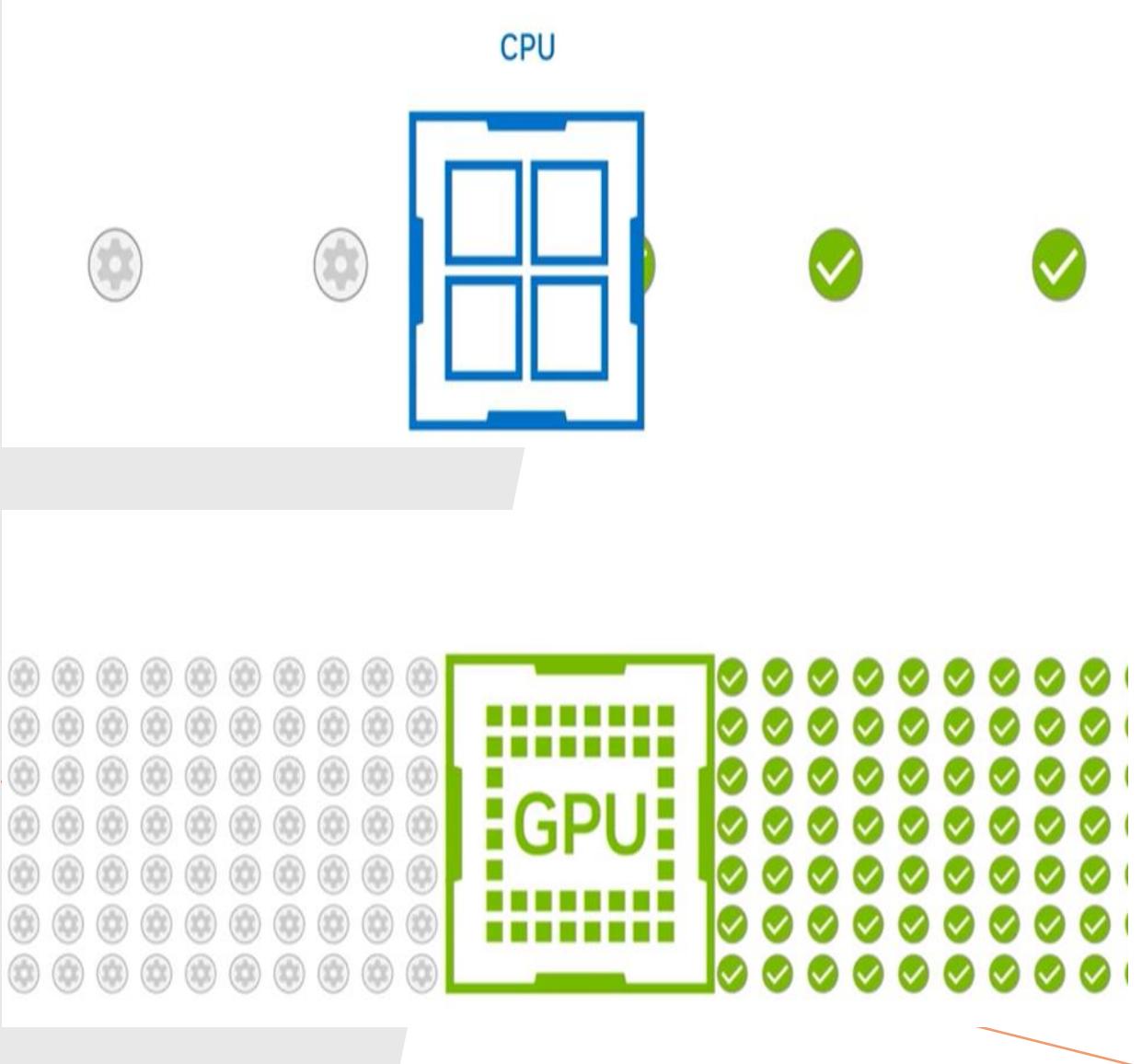
- Resolución de pantalla y pixeles
 - La imagen en un monitor se compone de pixeles
 - Cada pixel tiene posición, brillo y color
 - La resolución de pantalla representa la densidad de pixeles
 - A mayor resolución, mayor cantidad de pixeles a procesar
 - El procesamiento lo realiza el GPU
- Arquitecturas de GPU
 - En el centro del chip hay miles de núcleos GPU
 - Cada núcleo procesa instrucciones simples de datos
 - Tiene memoria cache integrada
 - Procesamiento paralelo gracias a los miles de núcleos
- Escalamiento de centro de datos

COMPARACIÓN CPU VS GPU



- Limitaciones de las CPU
 - Ejecutan instrucciones complejas una a una
 - Hasta 128 núcleos
 - Memoria principal grande pero ancho de banda bajo
- Avances en tecnología de CPU
 - Desarrollo de chips multinúcleos
 - Permite ejecutar varias instrucciones simultáneamente
- Características de las GPU
 - Procesan funciones simples
 - Más de 10000 núcleos por GPU
 - Ejecutan múltiples instrucciones sencillas en paralelo
- Funcionamiento conjunto de CPU y GPU

FUNCIONAMIENTO DE UNA GPU



- Transferencia de datos
 - Los datos se pasan de la CPU a la GPU
- Copia de código de ejecución
 - El código se copia del CPU al GPU
- Ejecución en la GPU
 - La GPU ejecuta y obtiene resultados
- Transferencia de resultados
 - El resultado se pasa a la CPU para procesamiento adicional
- Uso de caché en la GPU
 - La GPU utiliza caché para ejecutar más rápido
- Interconexión de datos
 - Los datos se mueven por buses como PCI o NVLINK

Acceleration Takes a Full Stack

APPLICATION

ACCELERATION LIBRARY

SYSTEM



DPU



GPU



CPU

SERVIDORES DE CENTRO DE DATOS

- Principales servidores de IA
 - Utilizan GPU
- Configuración de servidores
 - Cada servidor utiliza de 2 a 16 GPU
 - Trabajan en un servidor con varias GPU o en un clúster de servidores en paralelo
- Carga de trabajo de una GPU

		Hewlett Packard Enterprise			

7 PRINCIPIOS DE DESARROLLO DE IA

1-La IA debe ser beneficiosa a Nivel Social

2-La IA debe evitar crear o reforzar sesgos Injustos

3-la IA debe ofrecer un diseño con Seguridad Comprobada

4-La IA debe ser Responsable ante la Sociedad

5-La IA debe incorporar diseños de Privacidad

6-La IA debe tener altos estandares de Excelencia Cientifica

7-La IA debe estar disponible para usos que concurden con estos Principios

APLICACIONES DE IA QUE NO DESARROLLAR

1- Desarrollar IA con tecnologías que causen o puedan causar daño general

2-Desarrollar IA que promueva el uso de armas o cualquier tecnología que provoque lesiones a las personas

3-Desarrollar IA que recopilen o usen información para vigilancia

4- Desarrolla IA que se opongan a las leyes Internacionales y de derechos humanos



QUE HAY DETRÁS DE IA

Hardware

Servidores

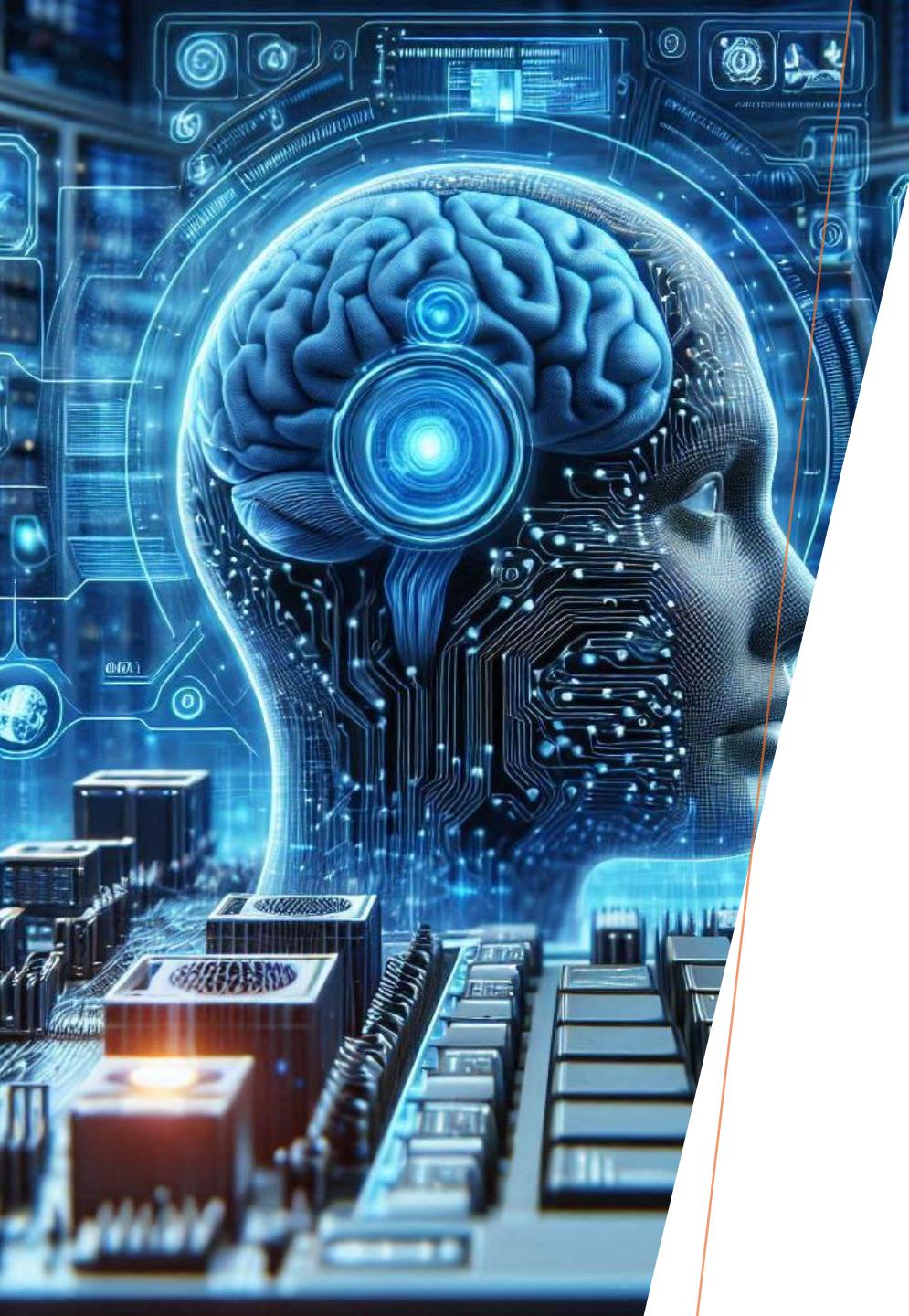
Servidores equipados con CPU-GPU-DTU-TSU

Almacenamiento

Para la inferencia de grandes cantidades de datos SSD de alta velocidad, almacenamiento en la nube y sistemas de archivo HDFS

Redes

Tecnología de redes definidas por software SDN , con alta velocidad y baja latencia



QUE HAY DETRÁS DE IA

Software

Virtualizacion y Contenedores

IA se ejecuta en entorno virtualizados con software de orquestacion de contenedores lo que falcita el equilibrio de trabajo y el escalado

Sistemas de Gestión de Datos

IA utiliza sistemas de datos distribuidos y gestión de datos masivos (hadoop o sistemas de bases de datos no sql)

Plataforma Machine Learning

Como Tensor Flow y Pytorch para el entrenamiento de IA



QUE HAY DETRÁS DE IA

Escalabilidad y Redundancia

Esalabilidad Horizontal

es un cluster de servidores que se puede ir agregando a medida que se necesitas aumentar la capacidad

Redundancia de alta Disponibilidad

Copias de seguridad de datos, alimentacion redundante,mecanismos de recuperacion ante desastres, fault tolerance para que los servidores esten disponibles aun en caso de fallo

QUE HAY DETRÁS DE IA

Seguridad

Seguridad Fisica los centros de datos donde se aloja la IA suele estar controlado con seguridad fisica avanzada

Seguridad de datos, se encriptan datos en reposo-
Transito para proteger los datos, se utilizan Firewall,
IPS, IDS, Tecnologias avanzadas de ciberseguridad
para proteger los datos

Control de Acceso, se utilizan sistema de control estrictos y autenticacion multifactor, AAA para garantizar que solo las personas autorizadas Ingresen

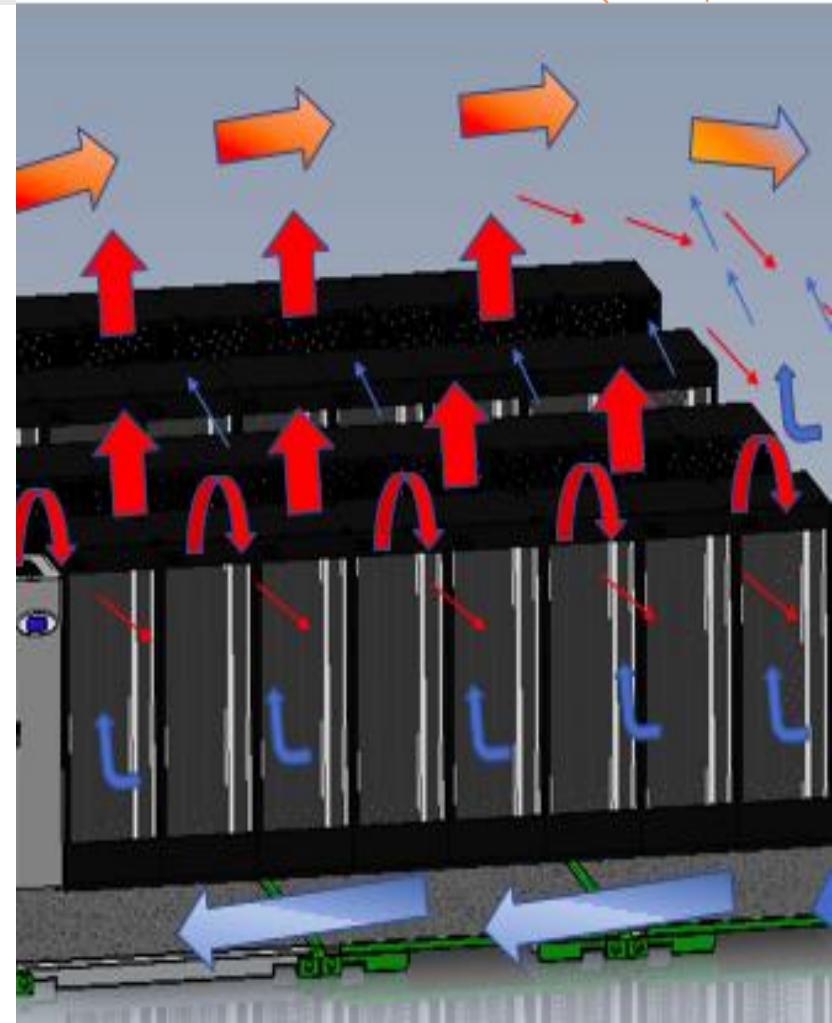


QUE HAY DETRÁS DE UNA IA

Energia y enfriamiento

los centro de datos utilizan fuentes de alimentacion redundante , en algunos casos energia renovable para minimizar el impacto ambiental

debido a la gran cantidad de calor que generan, se utilizan sistemas avanzados de enfriamiento por aire con control por tempertura y humedad para mantenerse en condiciones optimas



Gestion y Mantenimiento

Monitoreo y automatizacion

el sistema de IA esta constantemente monitoreado para detectar problemas potenciales antes de que ocurran

Actualizaciones y Parches

El software y el hardware se mantienen actualizados con los ultimos parches para evitar vulnerabilidades

PREGUNTAS

APRENDIZAJE AUTOMÁTICO-
APRENDIZAJE PROFUNDO

