



Datos, Machine Learning e Inteligencia Artificial

Diplomado en Ciencia de Datos:
Aplicaciones con Machine Learning





**EL FUTURO: LA IA
EL ACTIVO MÁS VALIOSO: LOS DATOS**



CONSEJOS PARA ESTE DIPLOMADO:

- 1. Los mejores trabajos IT requieren inglés**
- 2. Crea tu LinkedIn, muestra tu experiencia**
- 3. Crea tu repositorio en GitHub**
- 4. Trabaja en equipo**
- 5. Acepta la derrota y continúa**



TIPS DE VALOR

1. Nunca pares de aprender.
2. Toda empresa que no tenga IA o trabaje con datos, está mal (sea ahora o en un futuro)

¡AHORA SÍ!



¿Qué son los datos?

Hechos
Cifras
Observaciones
Mediciones



igerencia
THE DATA ANALYSIS COMPANY

**Los datos
son el nuevo
PETRÓLEO**

Guard Hawk

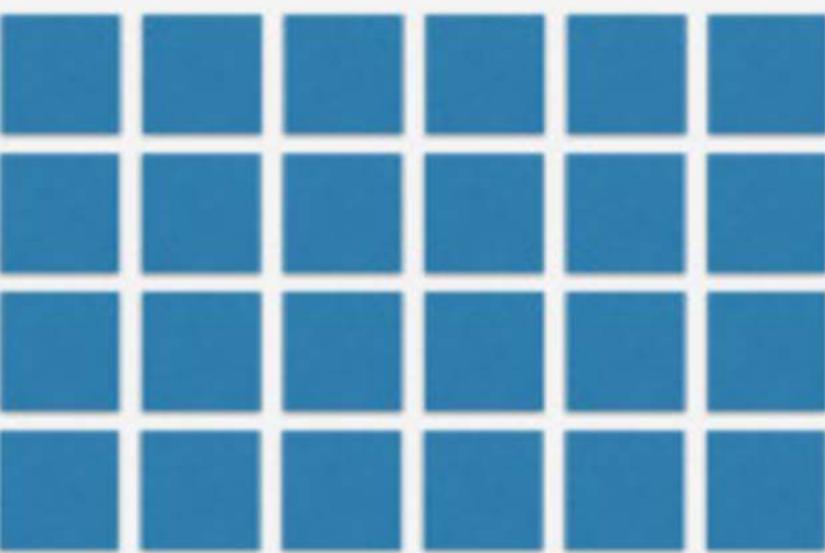
TOMA EL CONTROL DEL
ACTIVO MÁS VALIOSO DEL SIGLO XXI:
TUS DATOS DIGITALES

Guard Hawk

AI

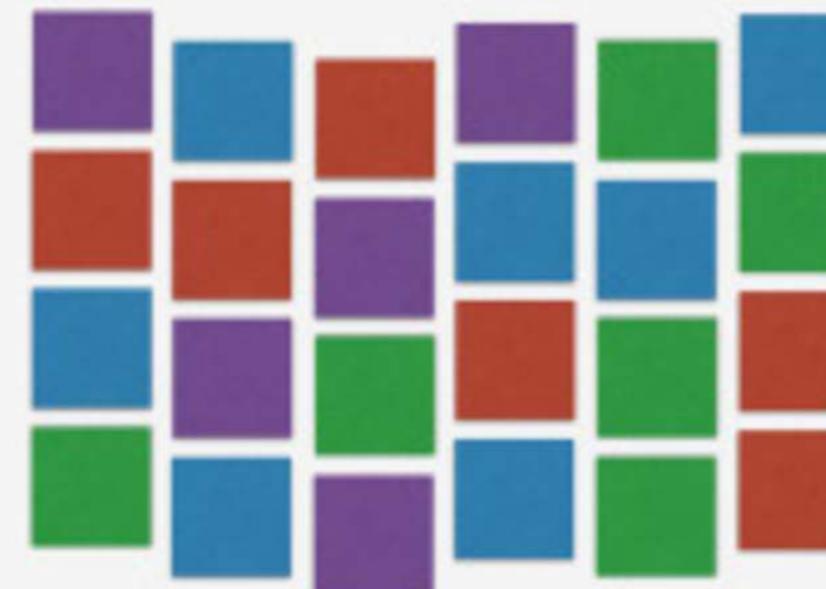
Tipos de datos

Datos estructurados



Lo que encuentras en una base de
datos (usualmente)

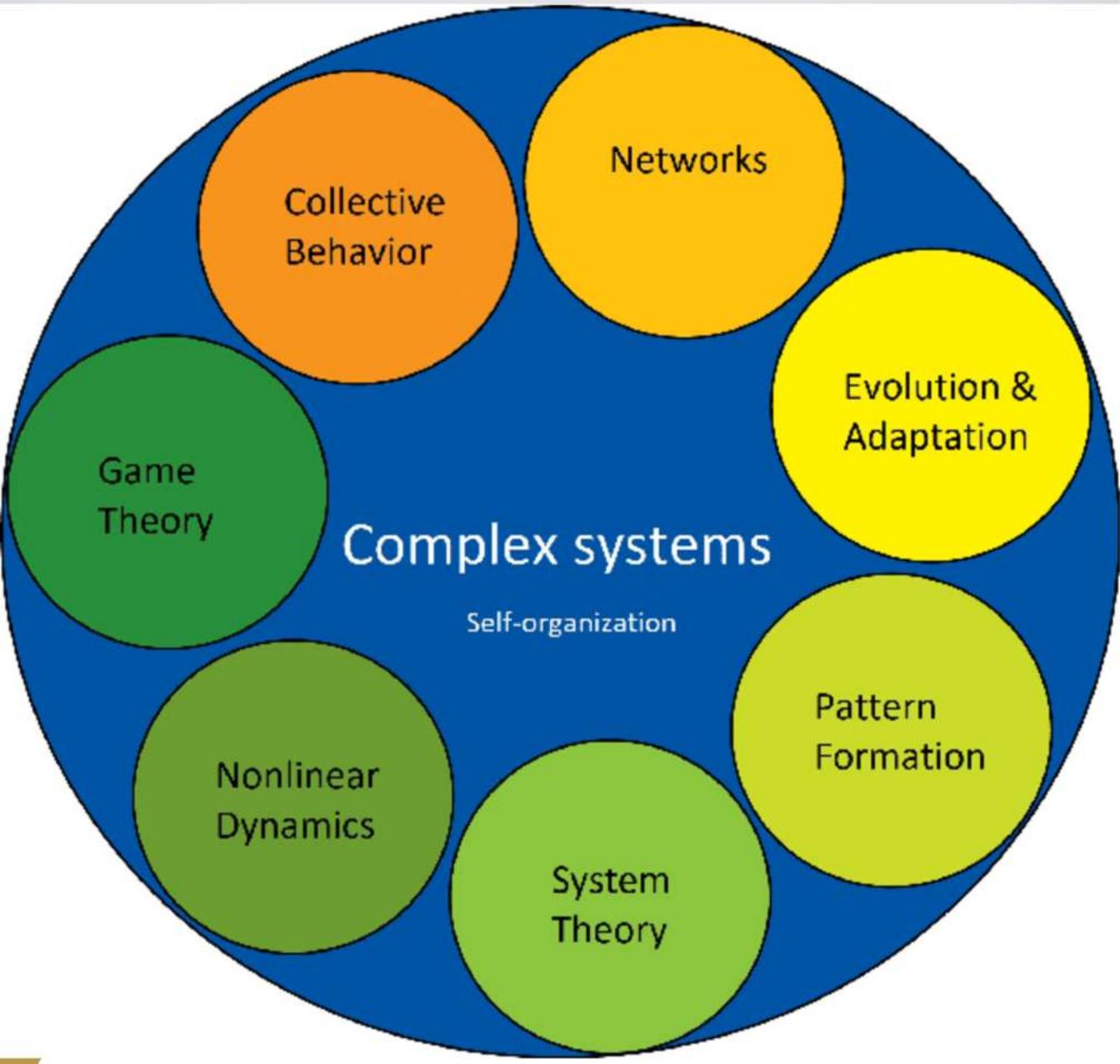
Datos No estructurados



Lo que tu encuentras fuera da la base
(texto, imagen, audio, video)



BIG DATA



¿Para qué nos sirven los datos?

Análisis

Interpretación

Uso

Entendimiento

Descripción

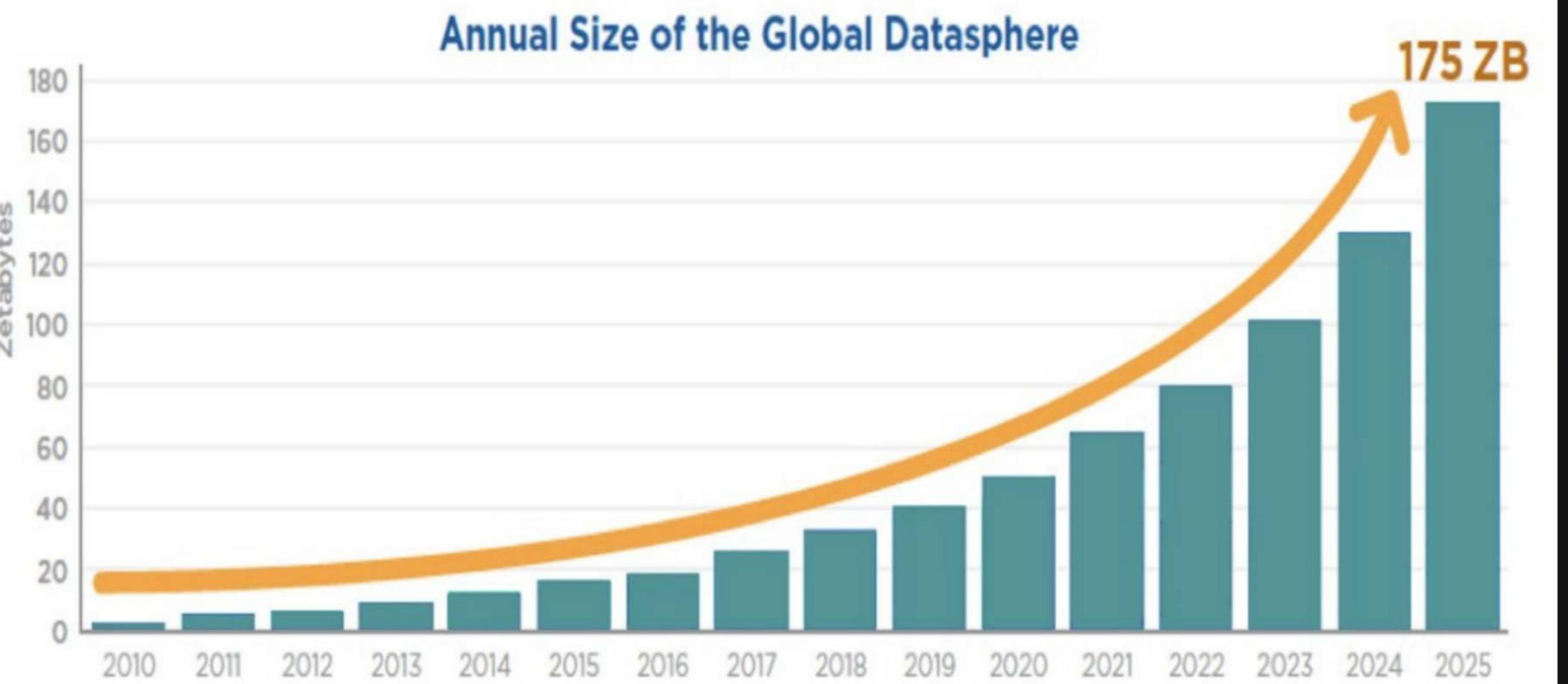
Predicción



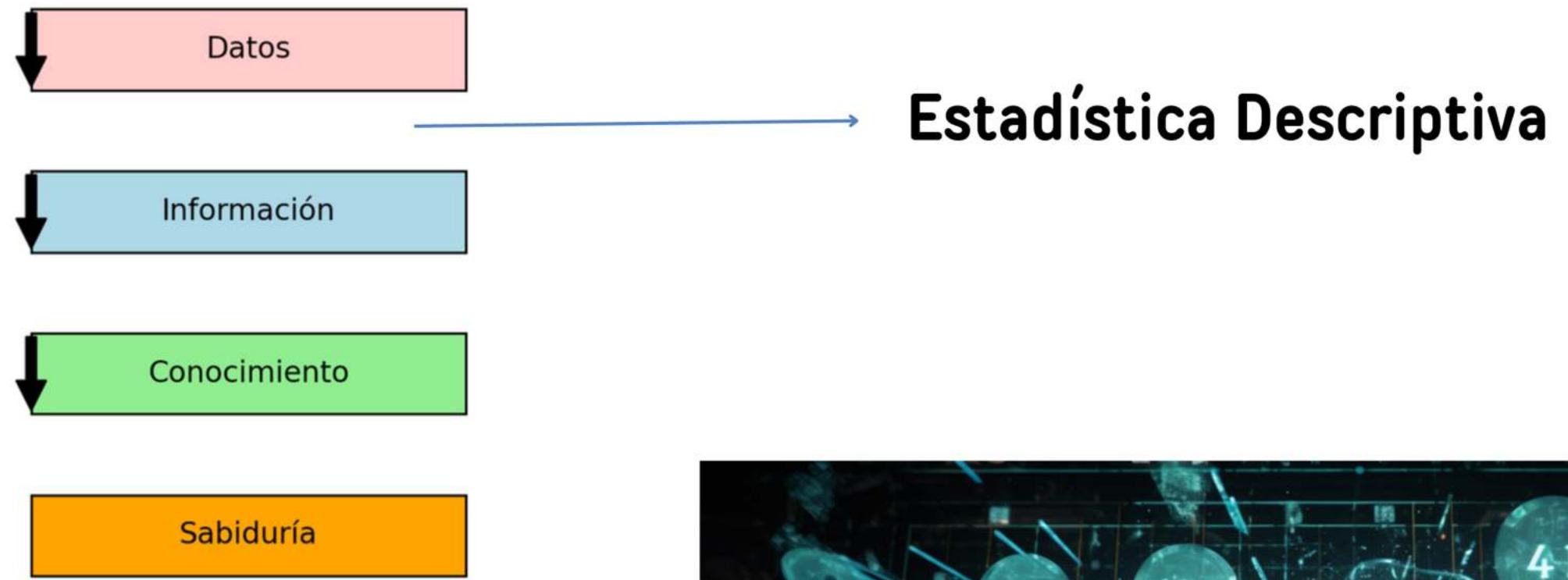
TOMA DE DECISIONES



Fenómenos
Comportamientos



Toma de decisiones



¿Qué es la ciencia de datos?

Es la ciencia de extraer patrones ocultos a partir de conjuntos de datos grandes

Los patrones ocultos pueden aparecer en forma de tendencias, ciclos, asociaciones, reglas y grupos

El término “ciencia” se refiere a las herramientas **estadísticas** y técnicas empleadas para entender los datos y la confiabilidad de los patrones identificados



¿Por qué la estadística?

Estadística Descriptiva: Entendimiento de los datos a partir de ideas vitales en términos de los valores centrales, la dispersión y la forma de distribución de los datos.

- Interpretabilidad a partir de resumen de los datos
- Gráficas, mapas y tablas

Estadística Inferencial: Establecer la confiabilidad de los patrones potenciales identificados.

- Interpretabilidad de parámetros poblacionales a partir de una muestra

Inteligencia Artificial

¿Qué entendemos por “**Inteligencia**”?

“Capacidad de entender, pensar, aprender y adaptarse a nuevas situaciones”

ChatGPT, 2023

- Resolver problemas
- Aprender de la experiencia
- Optimización
- Los sentidos

DATOS como activo



¿Inteligencia Artificial?

La “IA” se refiere a la simulación de procesos de inteligencia humana por parte de sistemas computacionales.

- Aprendizaje
- Razonamiento
- Auto-corrección

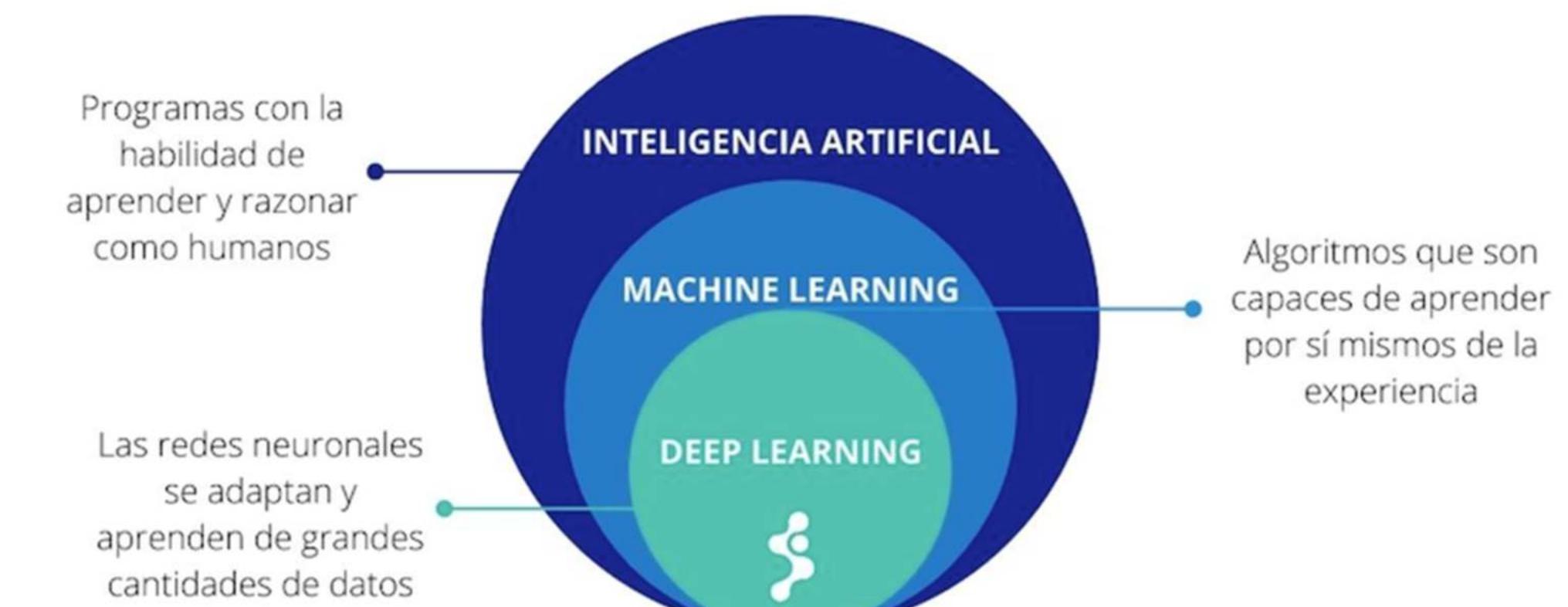


Imagen extraída de: Qué es la inteligencia artificial, Solver, 2020.

Inicios de la IA

- **1936:** Alan Turing y los fundamentos de la informática moderna
- **1941:** Konrad Zuse crea el primer computador programable
- **1943:** Warren McCulloch y Walter Pitts sientan las bases de las redes neuronales
- **1949:** Warren Weaver propone la computación para el procesamiento (traducción) de lenguajes
- **1950:** Alan Turing publica “Computing Machinery and Intelligence”: **¿Pueden pensar las máquinas?**
- **1956:** Conferencia de Dartmouth, donde John McCarthy acuña el término **“Inteligencia Artificial”**

Conferencia en Dartmouth College (1.956)

Padres fundadores de la I.A.



John McCarthy

Marvin L. Minsky

Claude E. Shannon

Nathaniel Rochester

Ray Solomonoff

Herbert A. Simon

Arthur Samuel

Oliver Selfridge

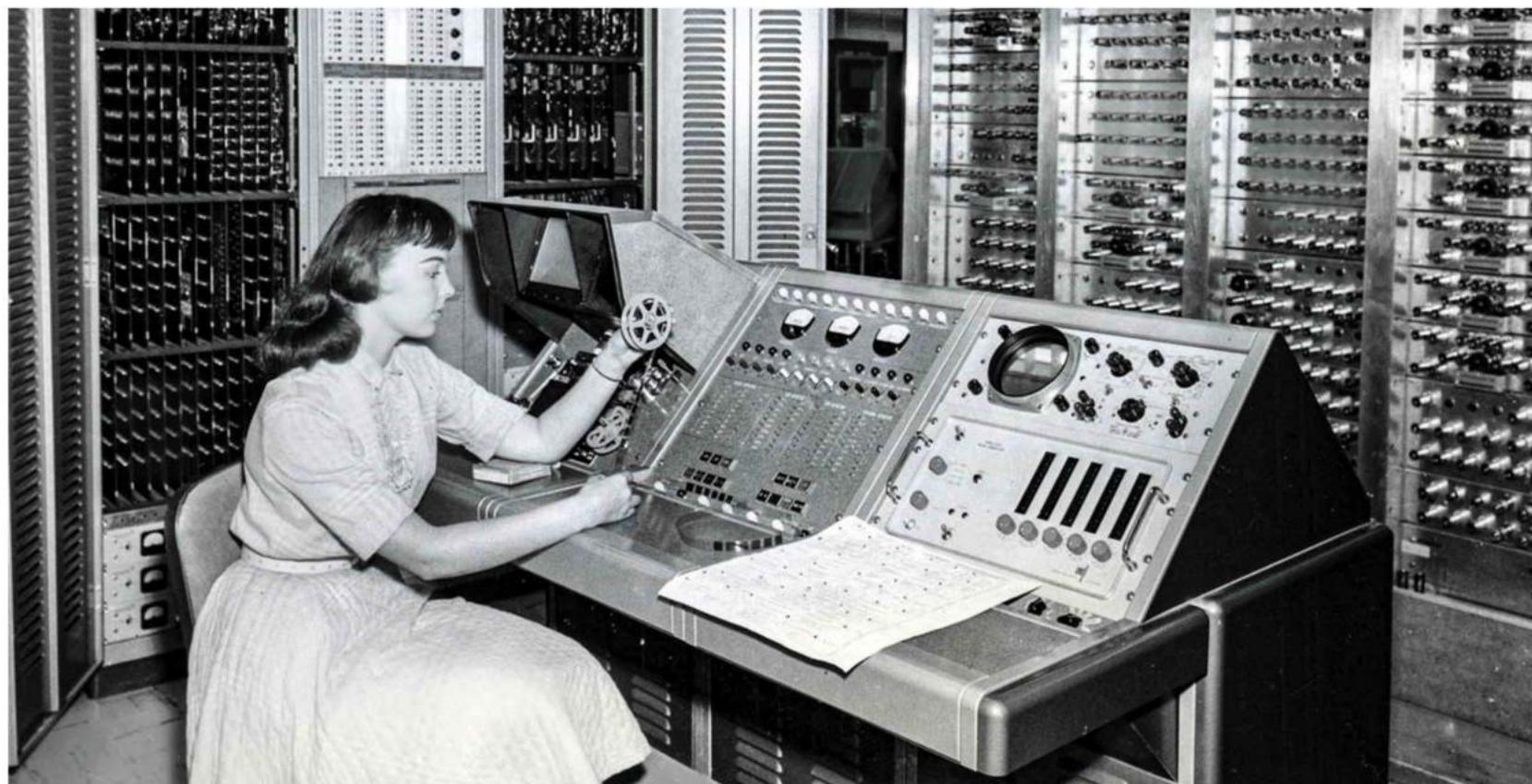
Allen Newell

Trenchard More

Imagen extraída de: Confusiones artificiales, Collateral Bits, 2019

1960s: La IA simbólica

- **1966:** **ELIZA**, un programa de procesamiento de lenguaje natural creado por Joseph Weizenbaum, simula una conversación. **Primer chatbot.**
- **1969:** Shakey, el robot que podía moverse y tomar decisiones de manera autónoma, es desarrollado por la Stanford Research Institute.



1970s - Primer invierno de la IA

- **1970:** Se propone la representación del conocimiento: ontologías.
- **1974-1980:** Reducción en el financiamiento y el interés en IA debido a expectativas no cumplidas.

**Fracaso del primer
modelo de GPT**



1980s - Resurgimiento de la IA

1980: La IA recupera atención con sistemas expertos como XCON.

1986: Se formula el algoritmo de retropropagación como herramienta de entrenamiento de las redes neuronales por parte de Geoffrye Hinton.

1989: Yan Le Cunn propone las redes convolucionales.



GEOFRYE HINTON

Creador del algoritmo de retropropagación.



YAN LE CUNN

Creador de las redes convolucionales.

1990s - Auge de la IA en la Red

- **1990:** Jeffrey L. Elman formula las redes neuronales recurrentes (RNN).
- **1997:** Deep Blue de **IBM** derrota al campeón mundial de ajedrez, Garry Kasparov
- **1998:** Lanzamiento PageRank, el poderoso algoritmo de Google

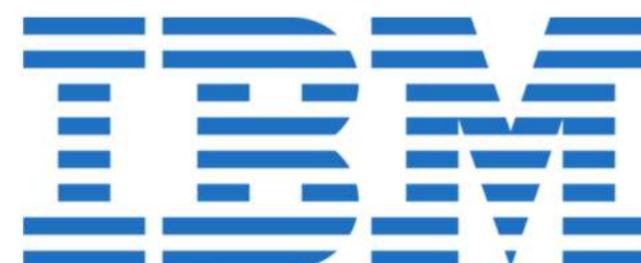


Computer Networks and ISDN Systems 30 (1998) 107–117

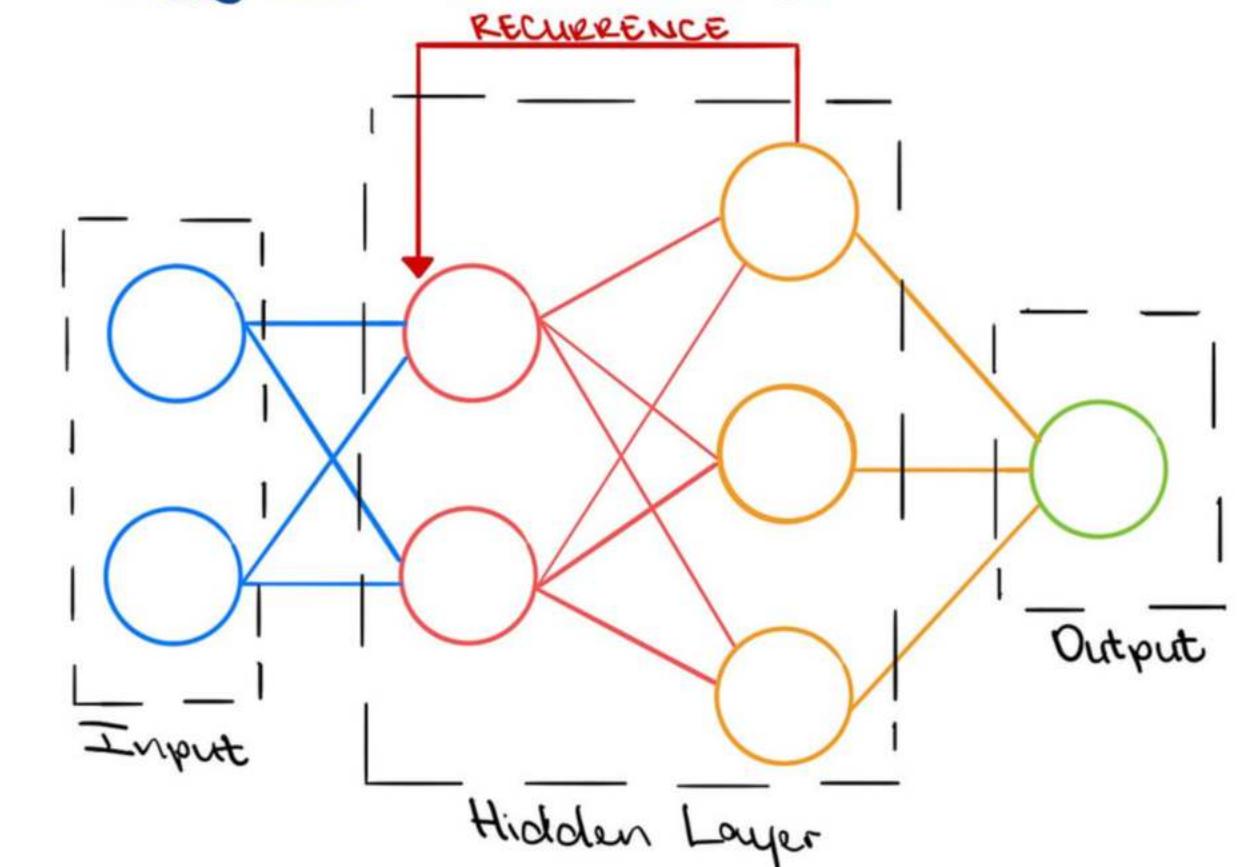
The anatomy of a large-scale hypertextual Web search engine¹

Sergey Brin², Lawrence Page^{*2}

Computer Science Department, Stanford University, Stanford, CA 94305, USA



RECURRENT
NEURAL NETWORKS



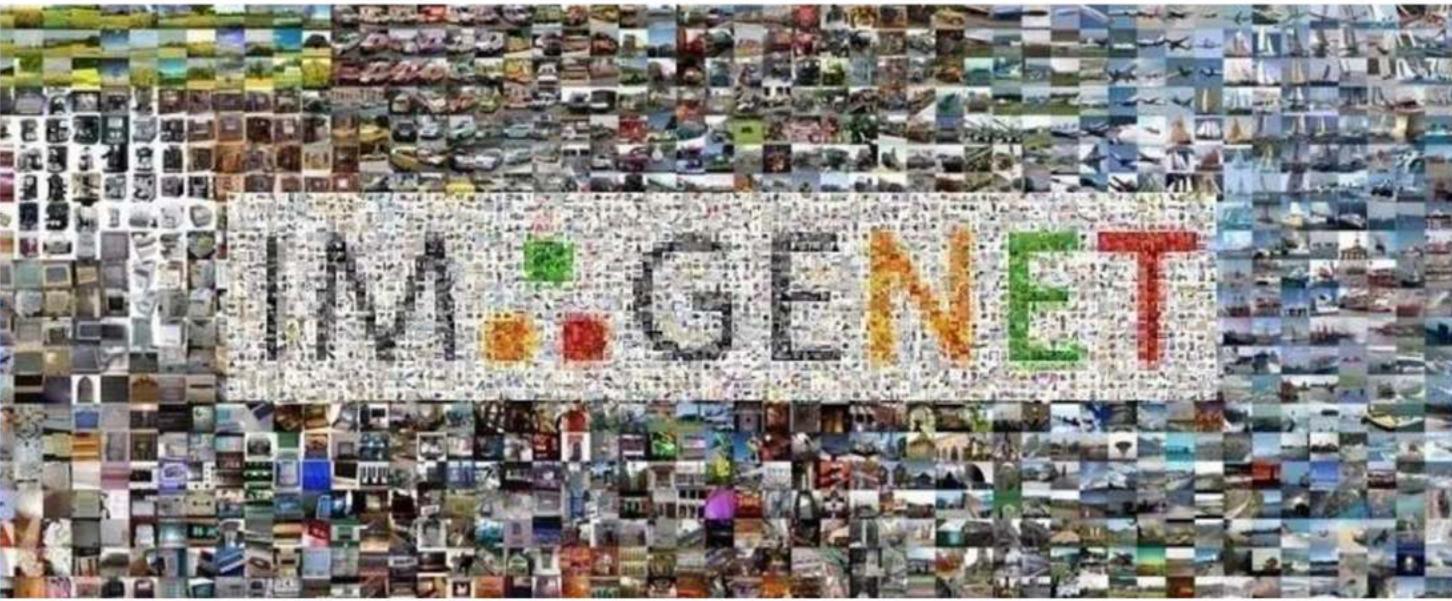
2000s - IA integrada

- **2005:** Stanley, el vehículo autónomo, gana el Gran Desafío DARPA.
- **2009:** Microsoft lanza Bing, integrando capacidades de IA en su motor de búsqueda.



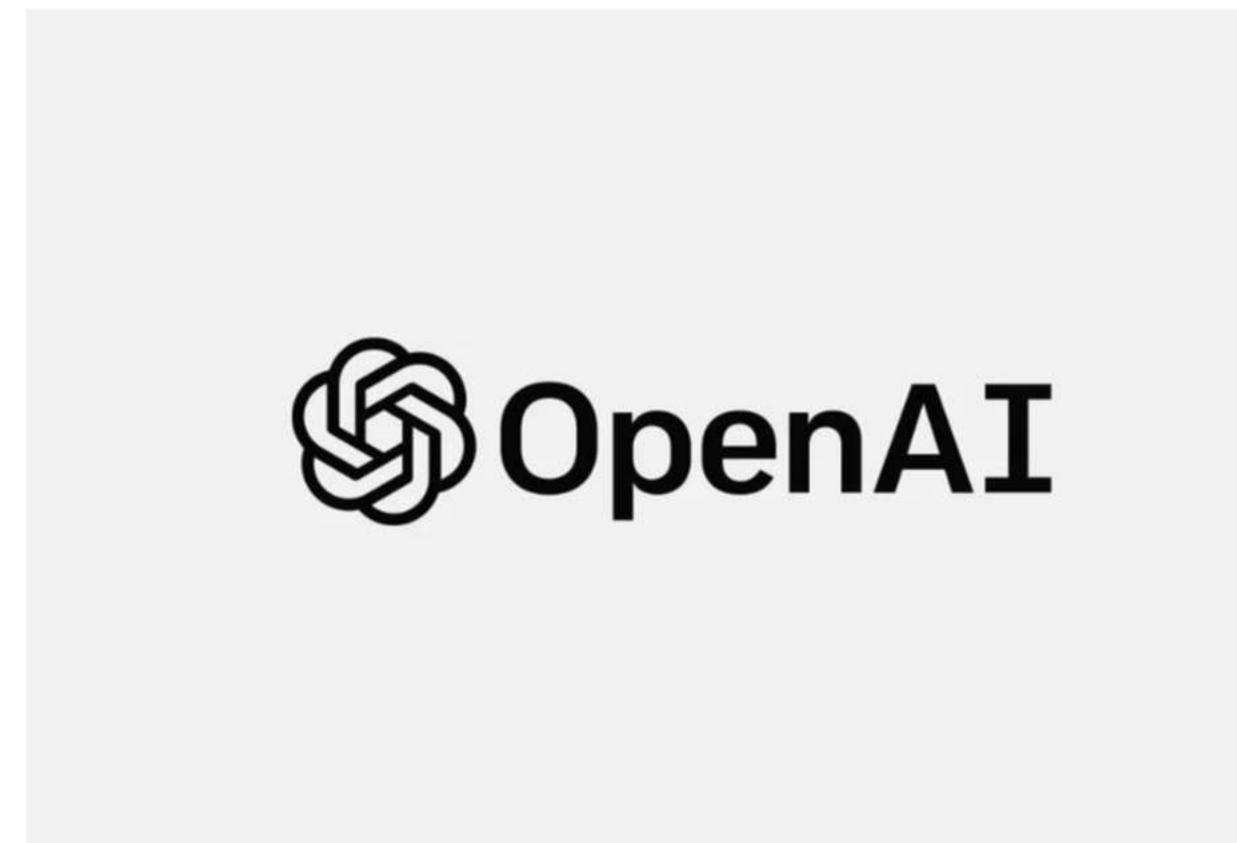
2010s - Aprendizaje Profundo

- **2011:** IBM Watson gana en Jeopardy!, un concurso de televisión
- **2012:** AlexNet, un modelo de red neuronal convolucional, domina la competencia ImageNet.
- **2014:** Google adquiere DeepMind y AlphaGo derrota al campeón mundial de Go.

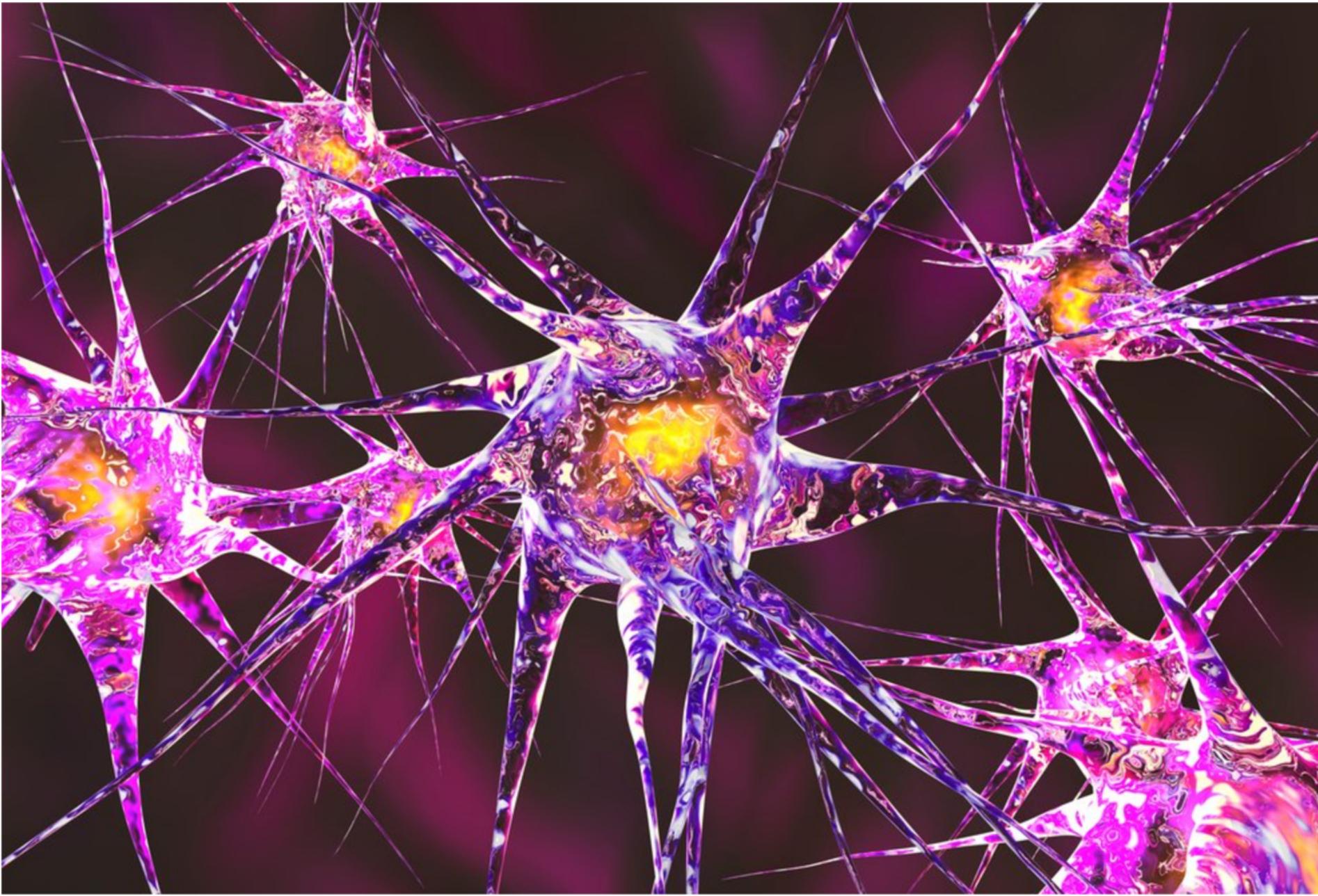


2020s - IA Avanzada y Ética

- **2021:** Lanzamiento de modelos de lenguaje avanzados como GPT-3
- **2023:** Lanzamiento de GPT-4 y poco después se inicián discusiones y legislaciones crecientes sobre la ética de la IA y la necesidad de regulaciones globales
- **2024:** Modelo Bard de Google evoluciona a Gemini, el modelo de lenguaje avanzado superior a GPT-4



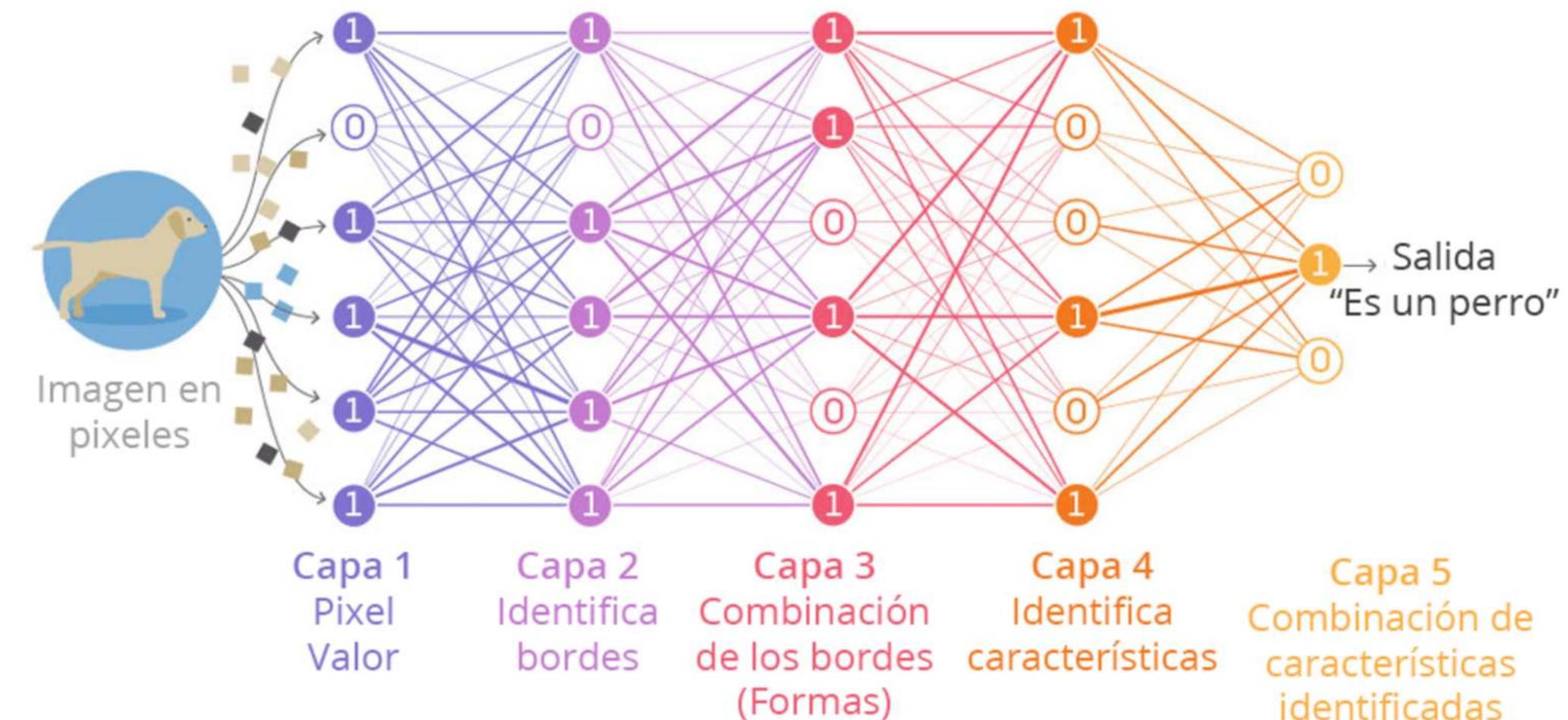
Cómo funciona la IA en sí?



Deep Learning

- ¿Qué es?
- ¿Por qué es tan poderoso?
- ¿Por qué funcionan?

- El **DeepLearning** es una forma de aprendizaje automático que emplea redes neuronales artificiales de múltiples capas interconectadas para aprender y realizar tareas complejas.

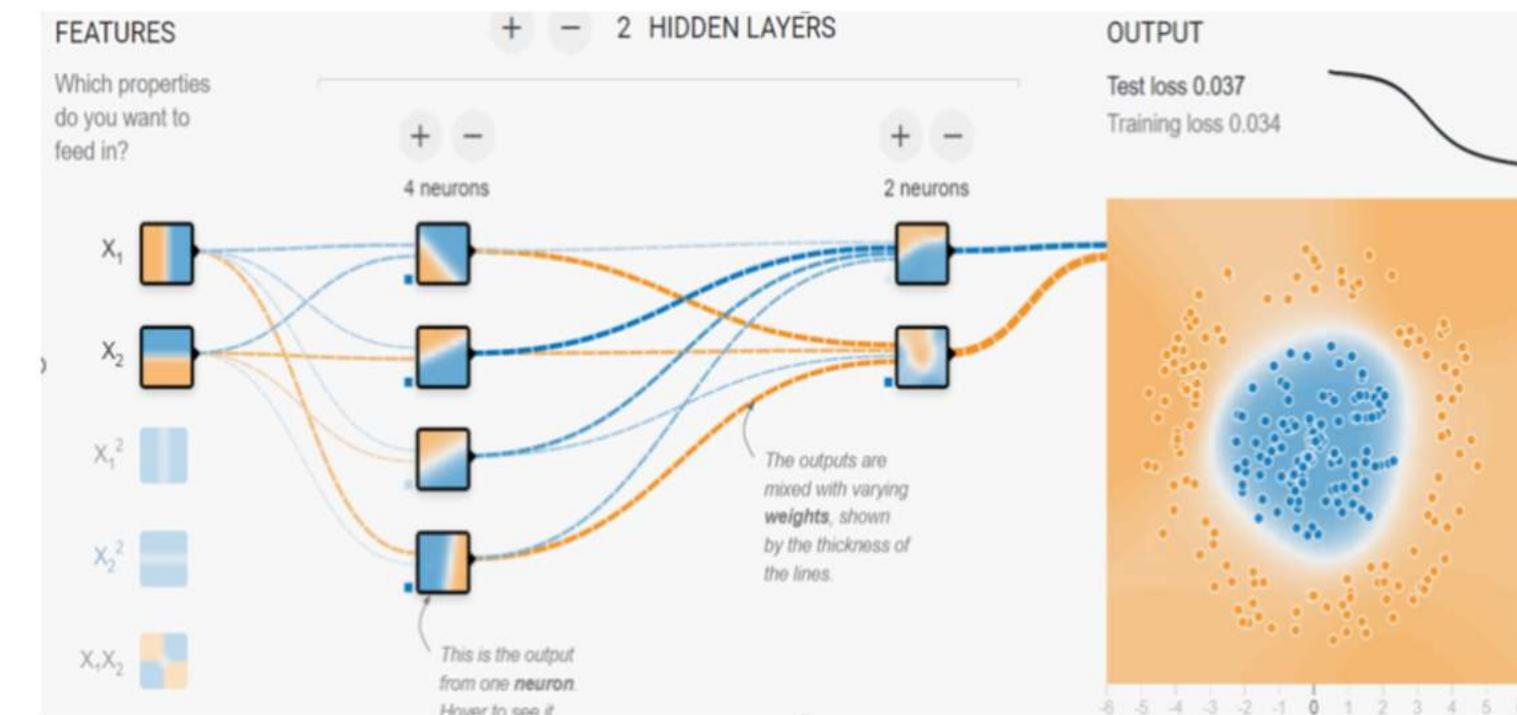
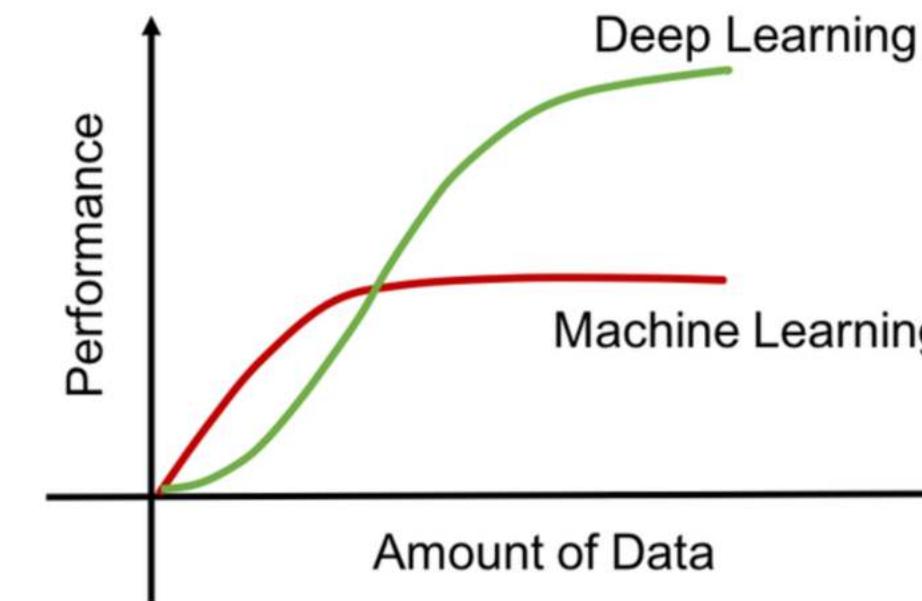


Fuente: <https://www.quantamagazine.org/>

Deep Learning

- ¿Que es?
- ¿Por que es tan poderoso?
- ¿Por que funcionan?

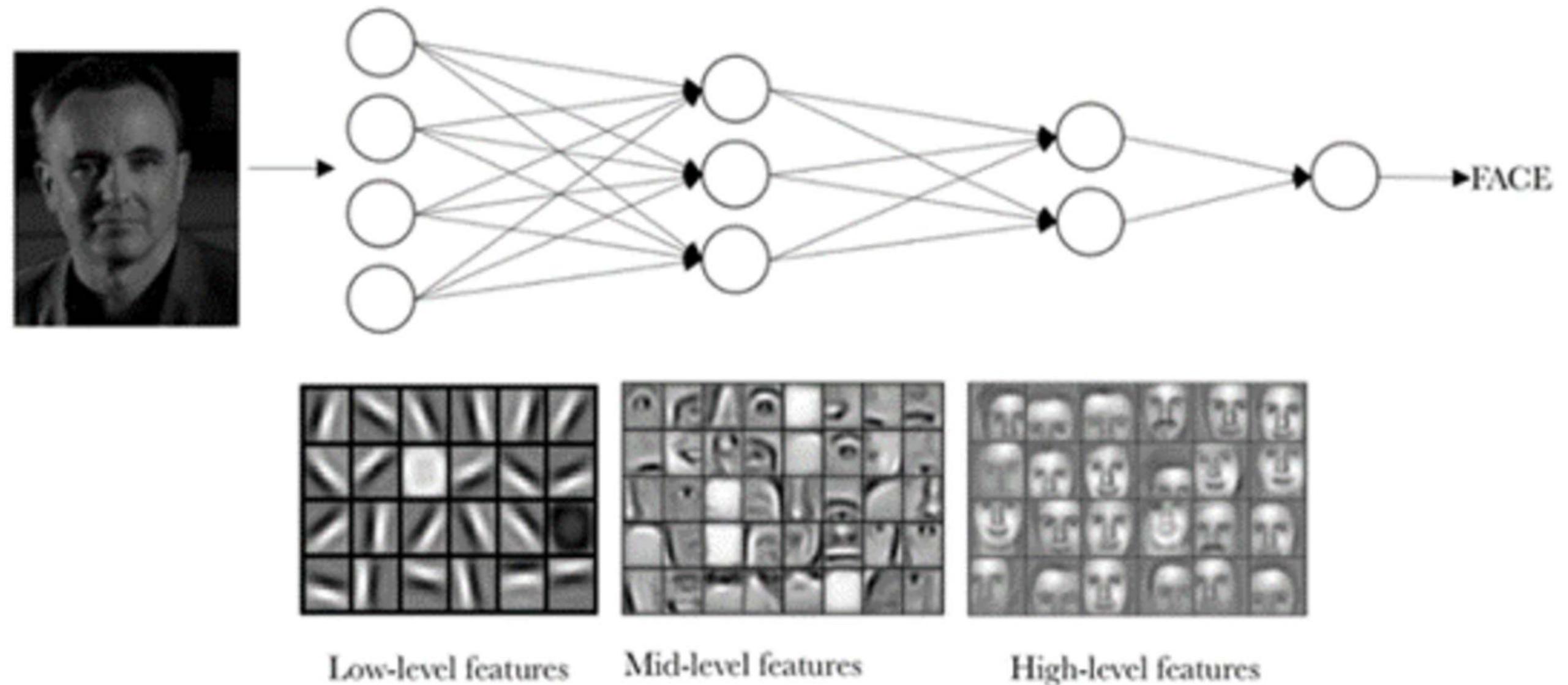
- Su poder radica en su capacidad para deducir **automáticamente la mejor codificación / representación** de la información, en su capacidad para aprender **características sofisticadas** en cada capa de la red, y en su capacidad de **generalización y adaptación al problema en cuestión**.



Deep Learning

- ¿Que es?
- **¿Por que es tan poderoso?**
- ¿Por que funcionan?

- Su poder radica en su capacidad para deducir **automáticamente la mejor codificación / representación** de la información, en su capacidad para aprender **características sofisticadas** en cada capa de la red, y en su capacidad de **generalización y adaptación al problema en cuestión**.

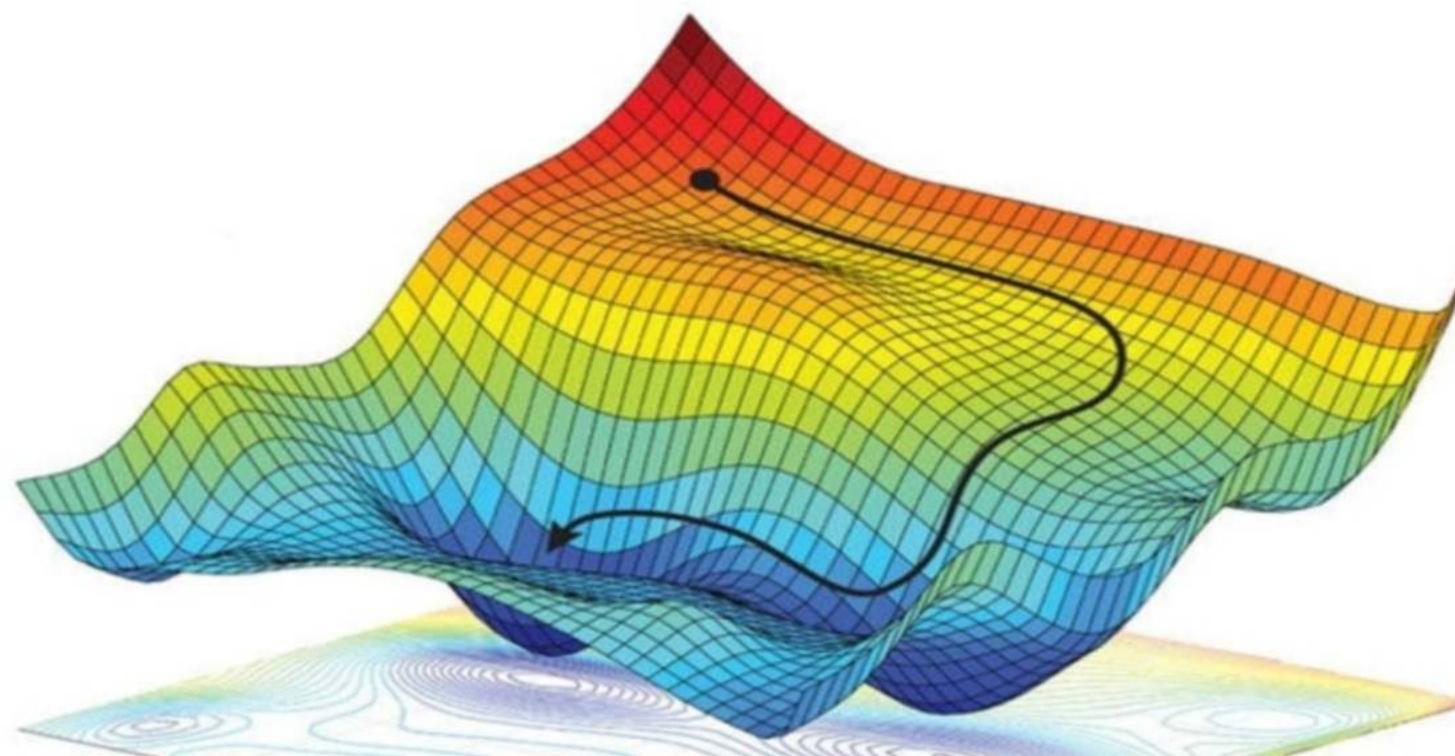


Deep Learning

- ¿Que es?
- ¿Por que es tan poderoso?
- ¿Por que funcionan?

- Su funcionalidad radica en la capacidad de **aprender a partir del ajuste de parámetros** en una **consecución se procesamientos** de información cuyas etapas **capturan patrones específicos** dentro de los datos. Son **modelos increíblemente flexibles** que permiten representar casi cualquier patrón si la **data, la arquitectura y la función de costo es la apropiada.**

- Parámetros ajustables
- Función de costo optimizable.
- Arquitectura flexible.
- Ensamble de modelos “simples”.
- Etapas de procesamiento especializadas.



Deep Learning

• ¿Por que Hoy?

- A pesar de que los fundamentos matemáticos de las redes neuronales artificiales se plantean en las primeras décadas del siglo pasado, las redes neuronales artificiales encuentran su éxito en la era contemporanea gracias a:

Disponibilidad de enormes cantidades de **datos**.



Software especializado, **tecnicas** establecidas, **marcos experimentales** claros.

 **TensorFlow**

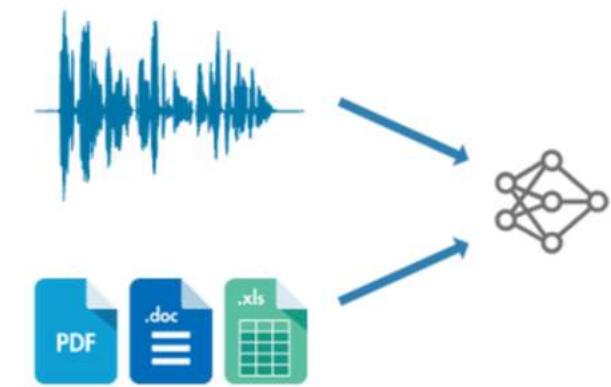
 **PYTORCH**

Mayor capacidad de computo, creación de los **procesos paralelizables** y las **GPU**.

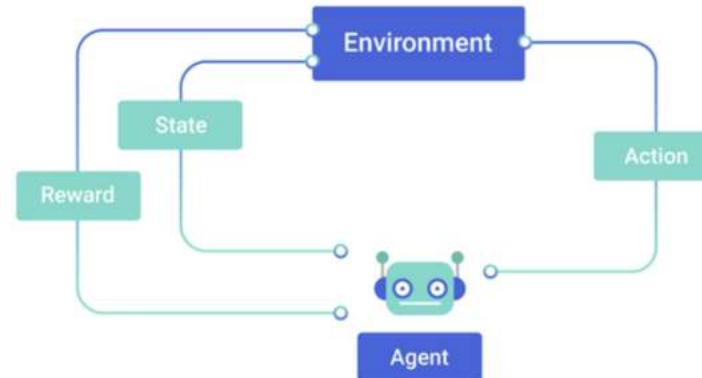
Ramas de la IA



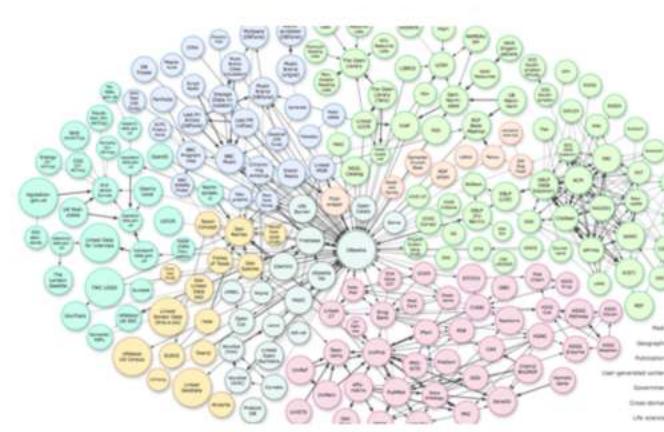
Computer Vision



Natural Language Processing

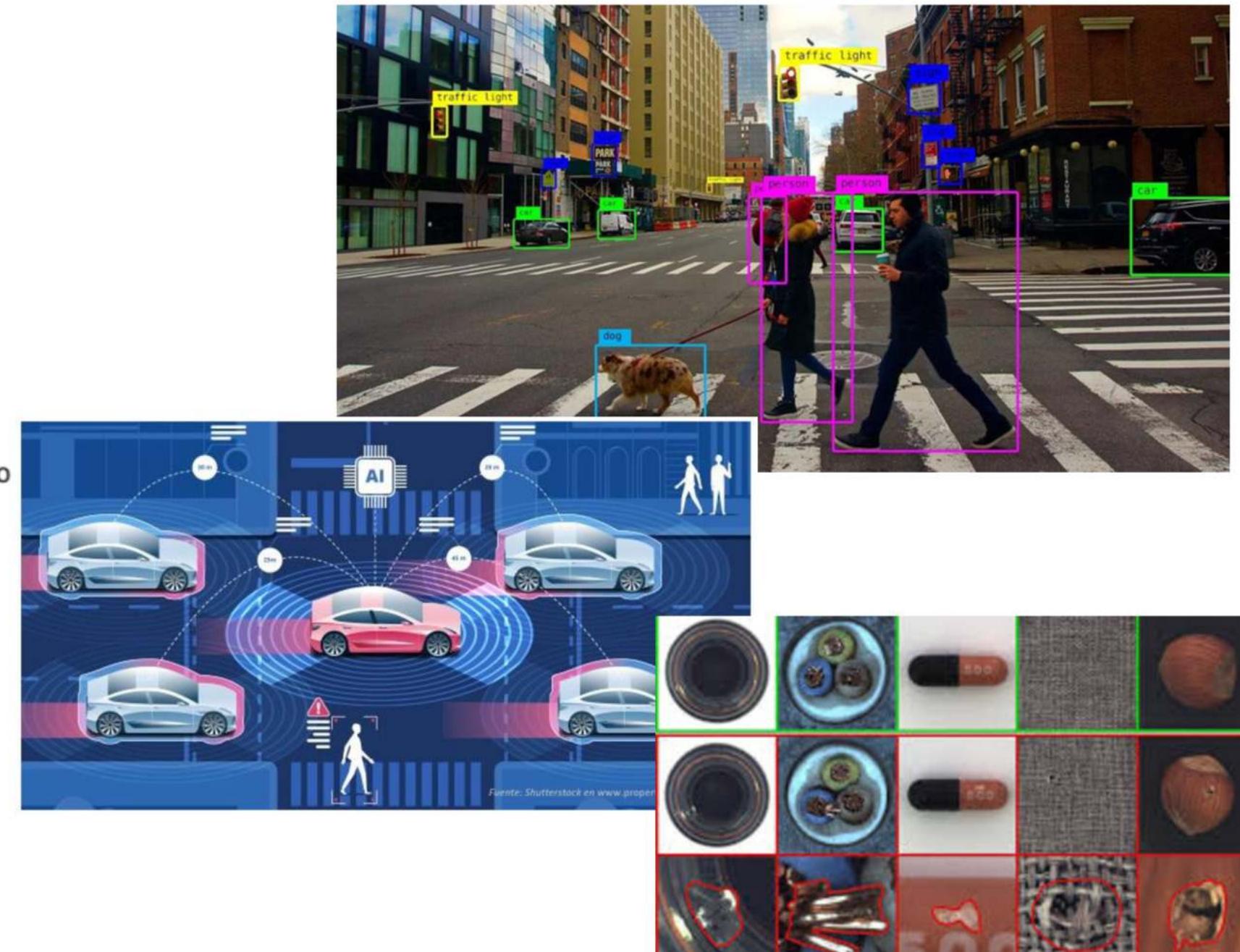
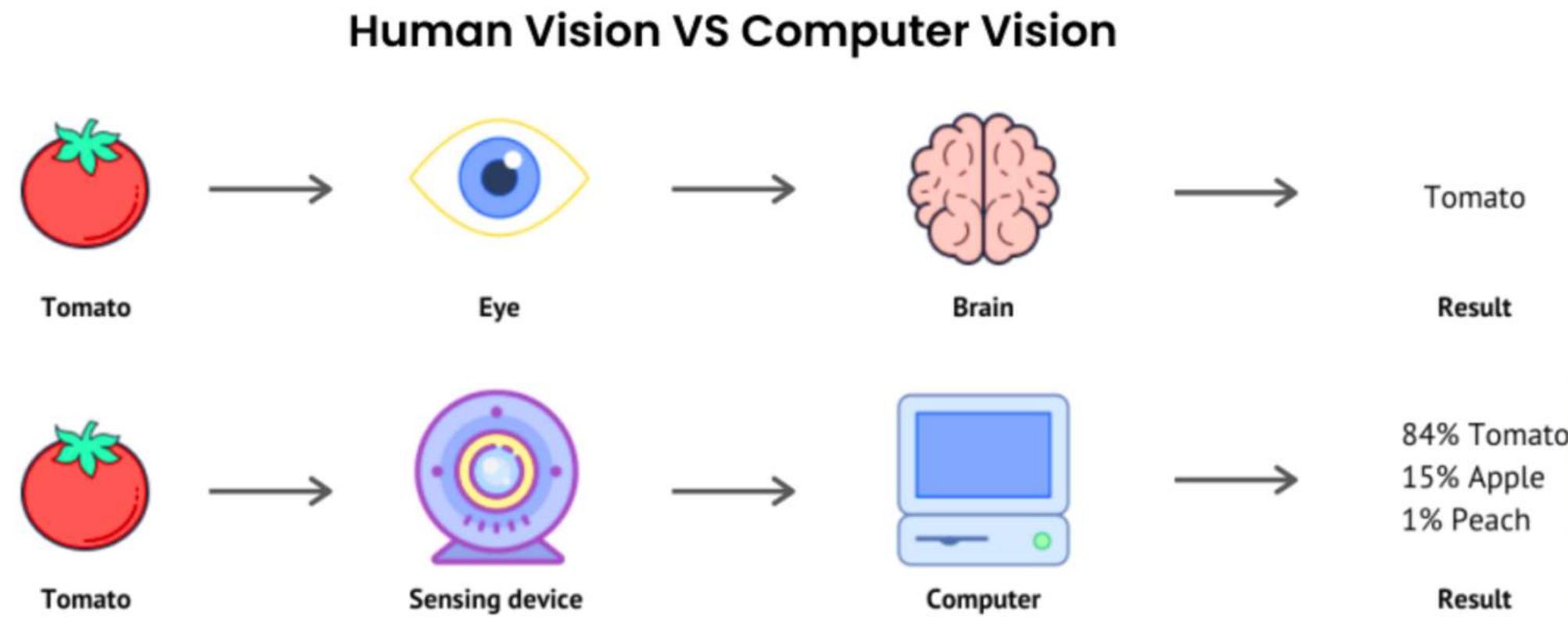


Aprendizaje por Refuerzo



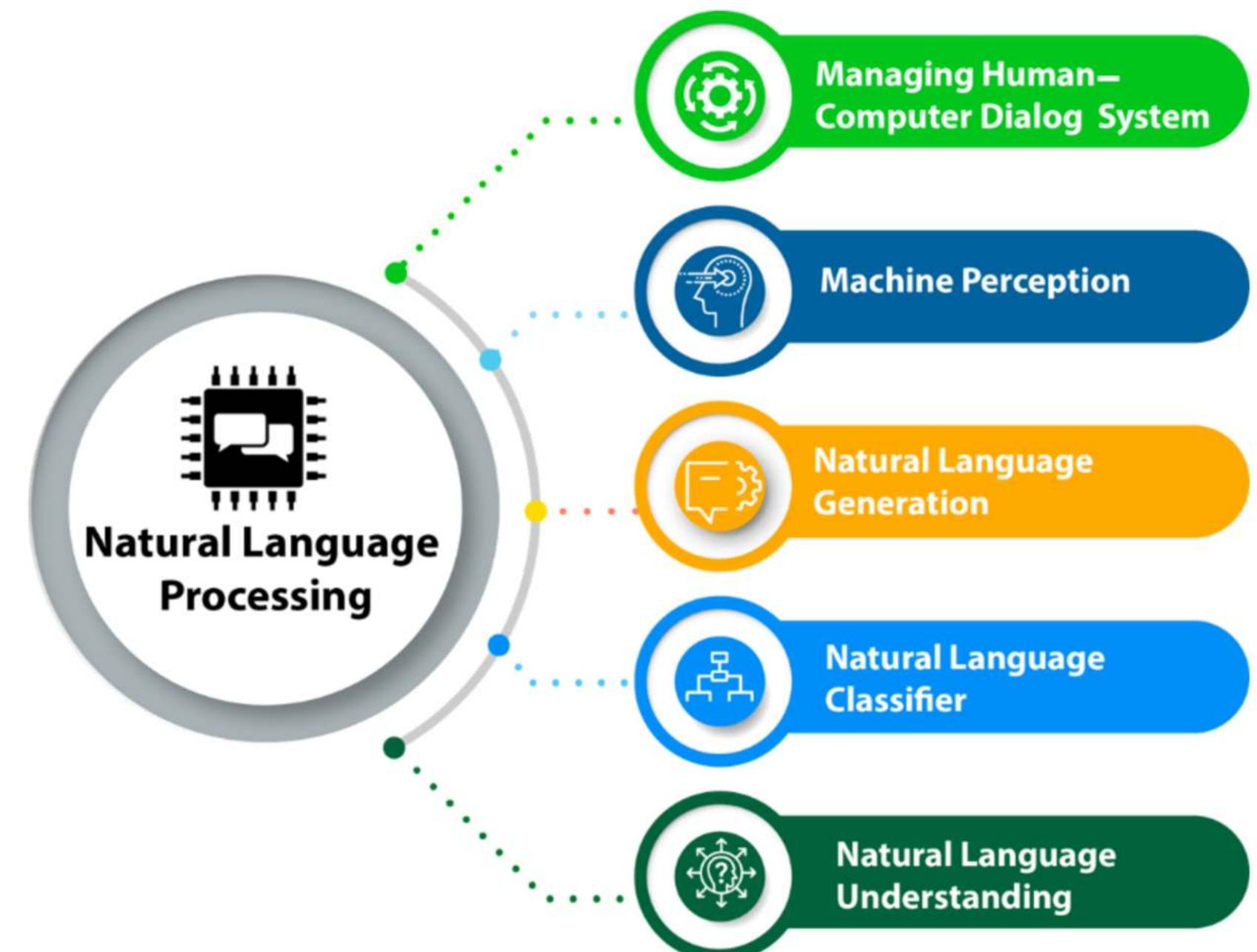
Visión por Computadora: Entendiendo el mundo visual a través de la IA

La visión por computadora es una rama de la inteligencia artificial que permite a las máquinas 'ver' y analizar contenido visual a partir de imágenes y videos digitales.



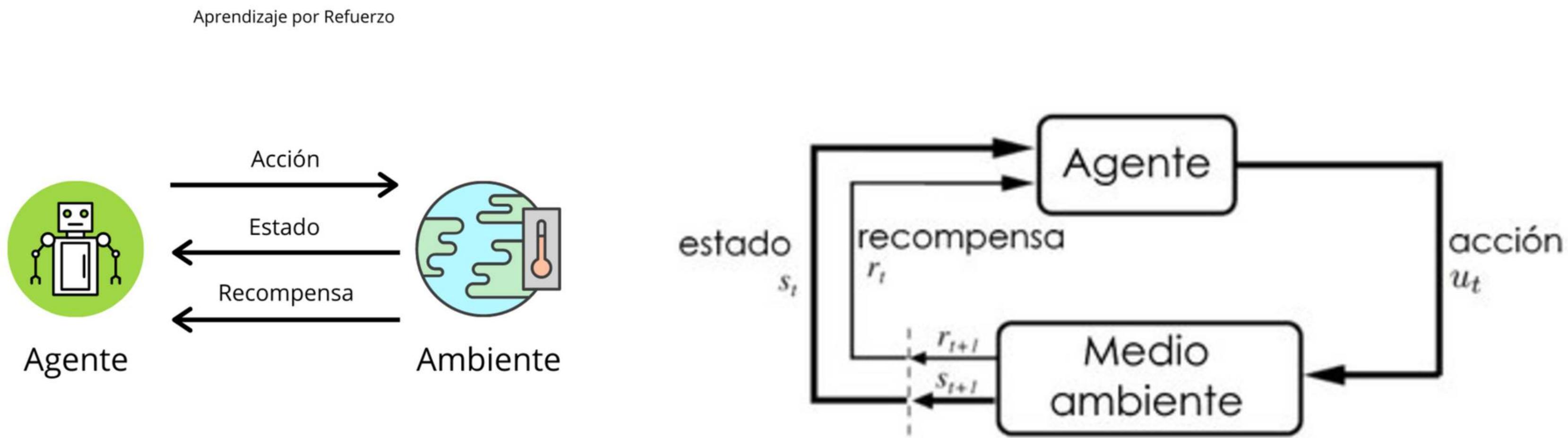
Procesamiento de Lenguaje Natural (NLP): La Interfaz Humano-Máquina

El procesamiento de lenguaje natural es una subdisciplina de la inteligencia artificial que se centra en la interacción entre computadoras y humanos a través del lenguaje natural.



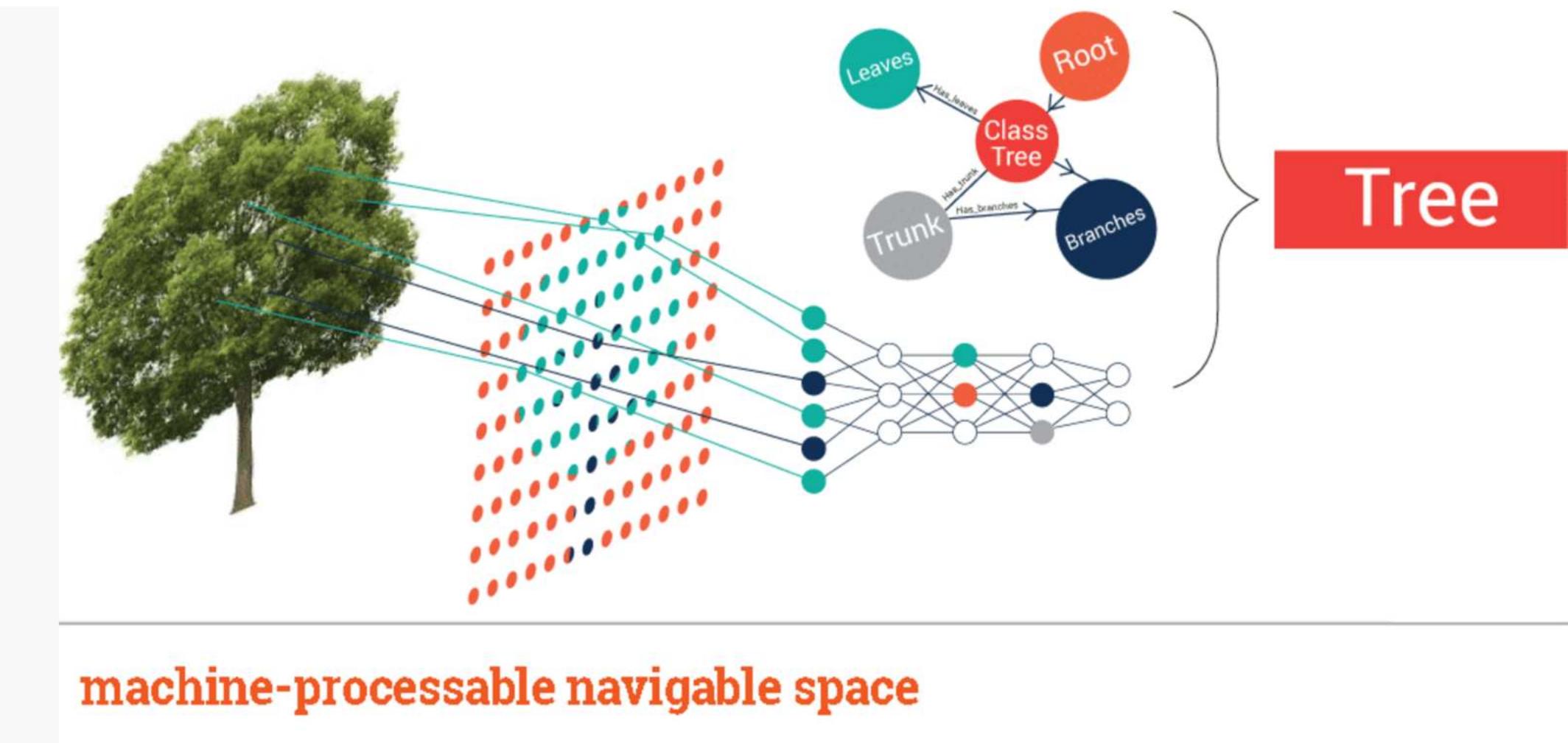
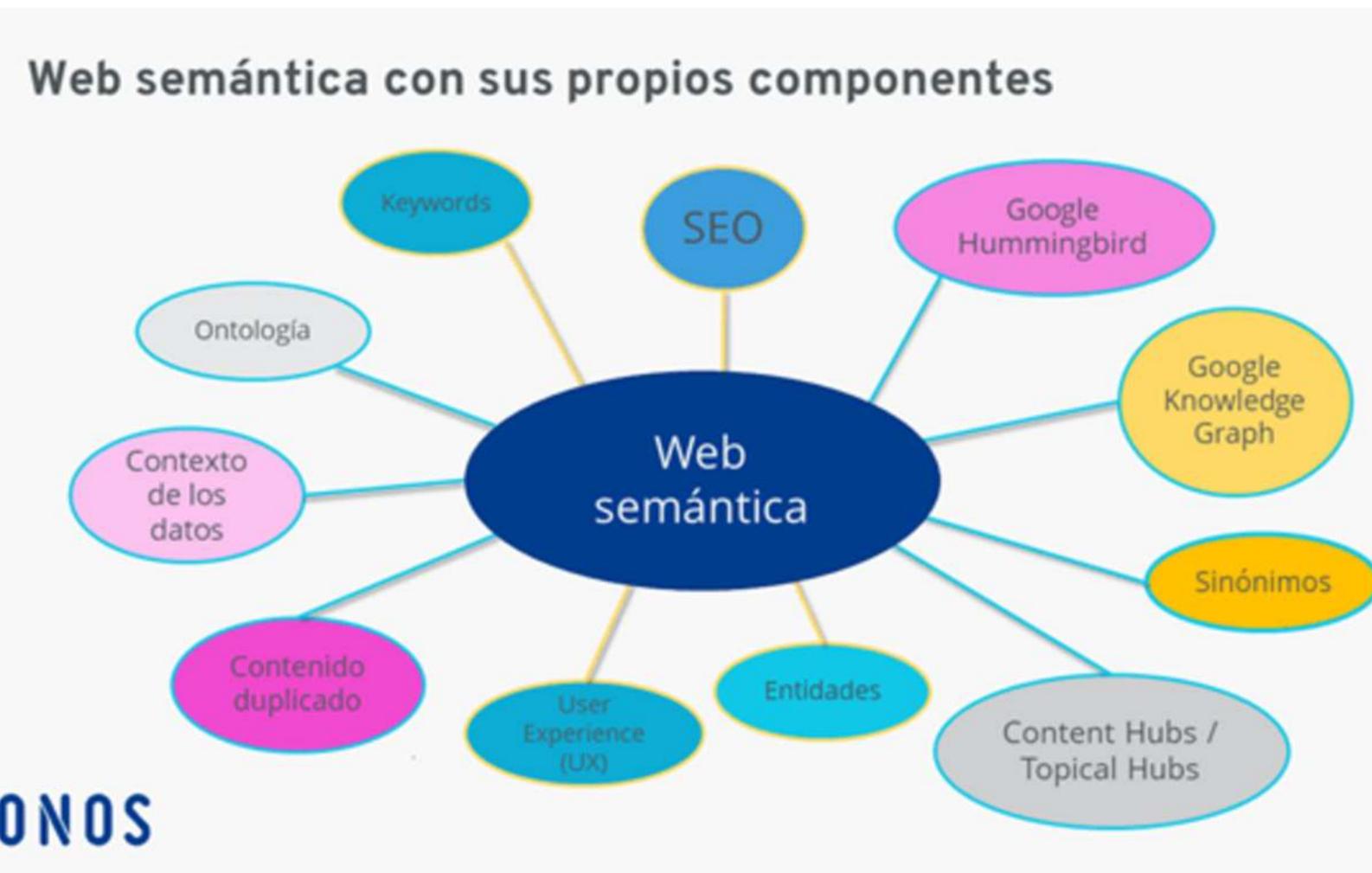
Aprendizaje por Refuerzo: Entrenando Modelos con Recompensas

El aprendizaje por refuerzo es un tipo de aprendizaje automático donde un agente aprende a tomar decisiones optimizando acciones basadas en recompensas y penalizaciones para alcanzar un objetivo.



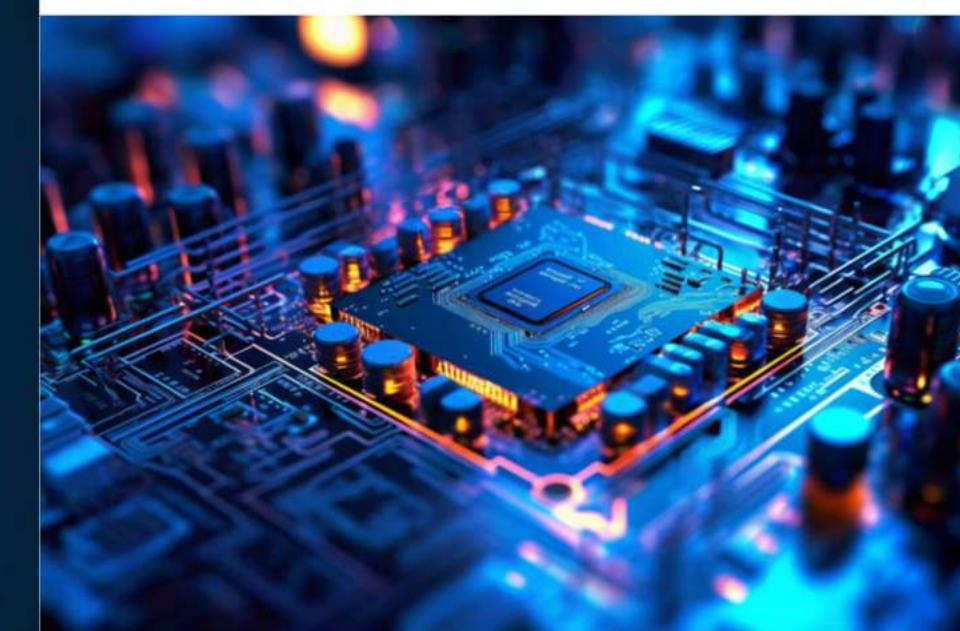
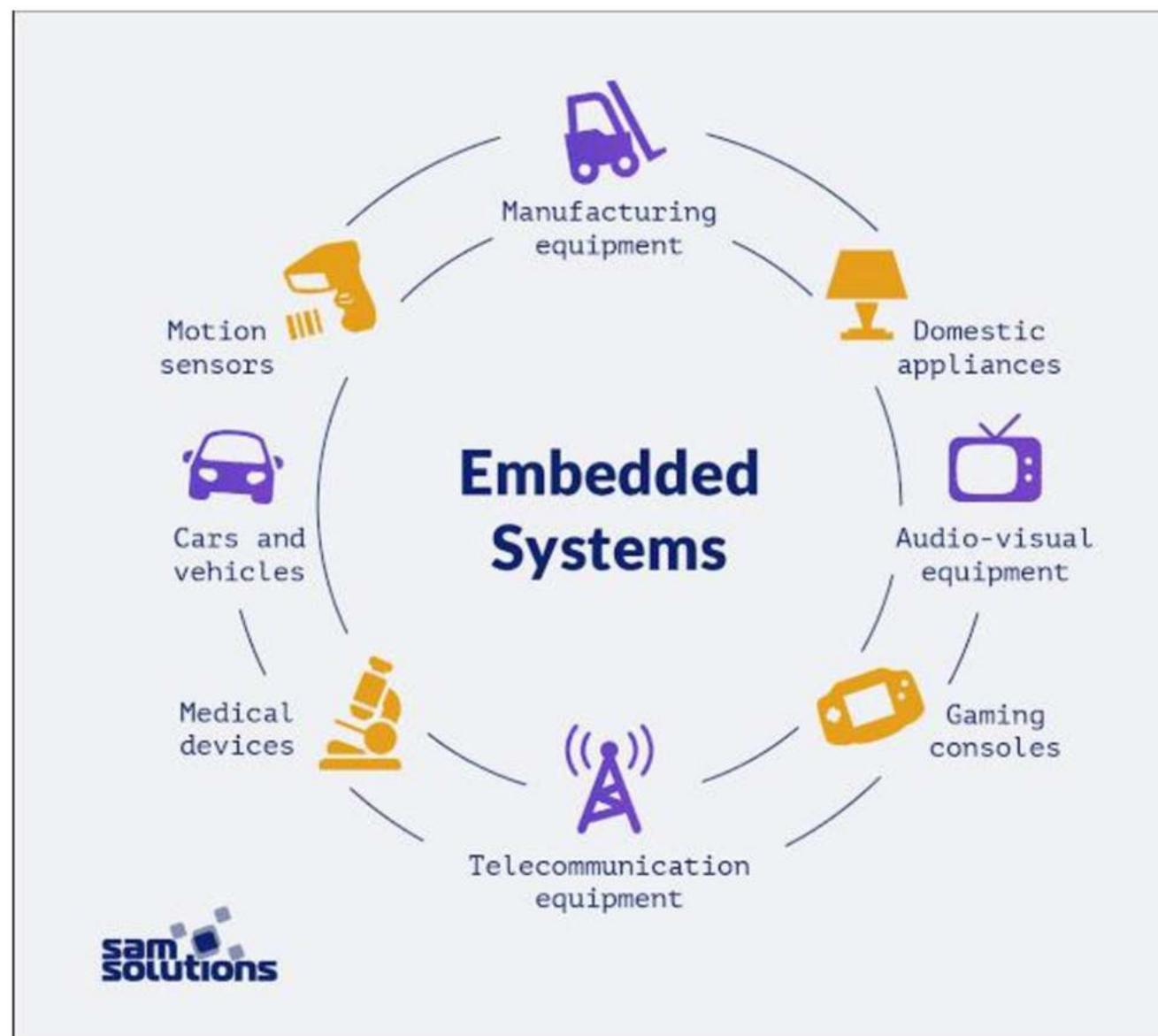
Web Semántica: Enriqueciendo la Web con Significado

La web semántica es una extensión de la web actual que permite a los datos ser compartidos y reutilizados a través de aplicaciones, empresas y comunidades. Utiliza estándares para proporcionar un marco común que permita compartir e interpretar datos con un significado bien definido.



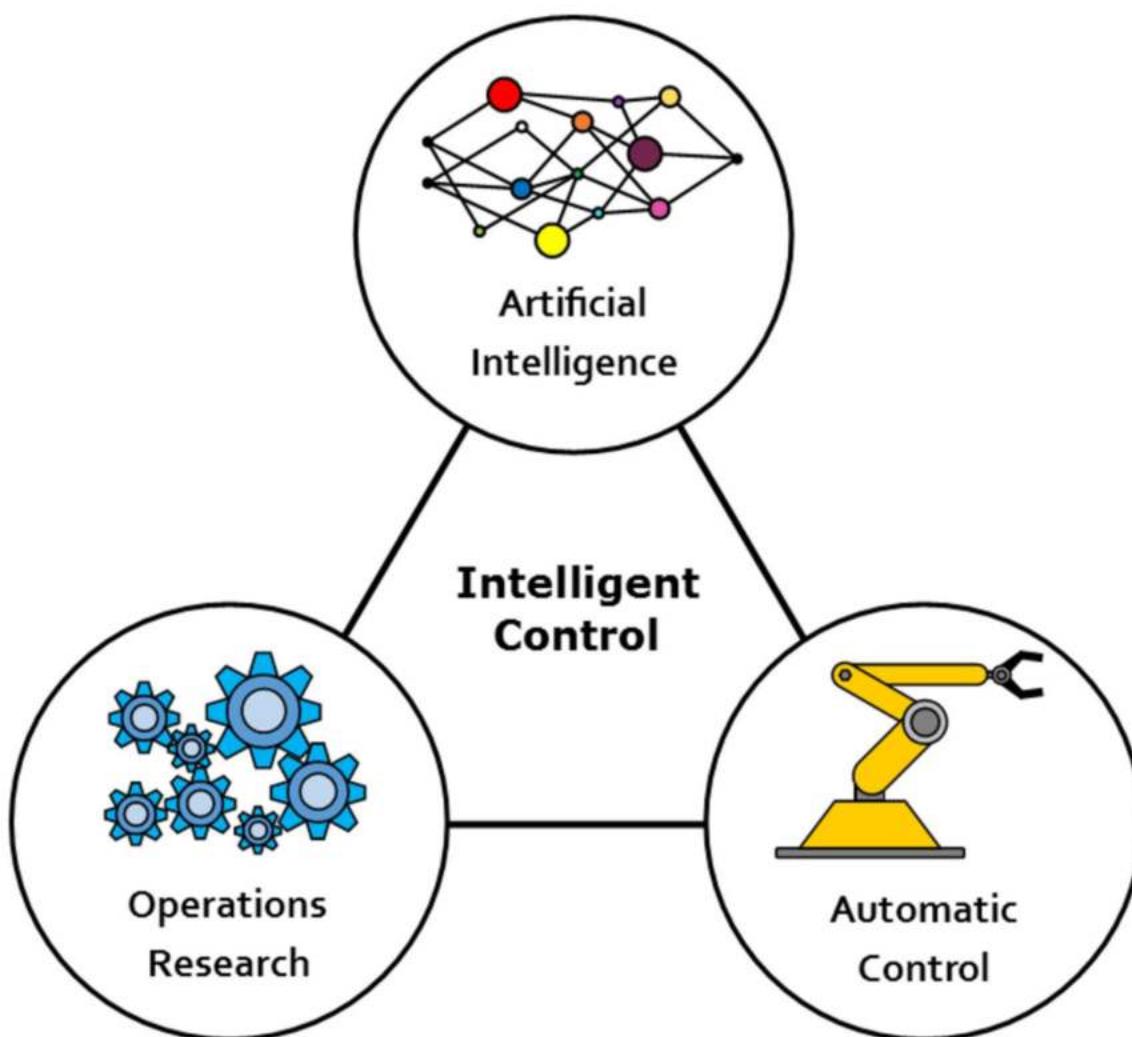
Sistemas Embebidos: Tecnología Integrada en Nuestro Día a Día

Un sistema embebido es un sistema de computación diseñado para realizar una o unas pocas funciones dedicadas, a menudo en un sistema de computación en tiempo real. Son dispositivos integrados como parte de un sistema completo y están destinados a operar de forma autónoma.



Control Inteligente: Optimización y Automatización en Sistemas Dinámicos

El control inteligente se refiere al uso de técnicas de inteligencia artificial para mejorar o reemplazar los métodos tradicionales de control en sistemas dinámicos, permitiendo una respuesta más adaptativa y eficiente.

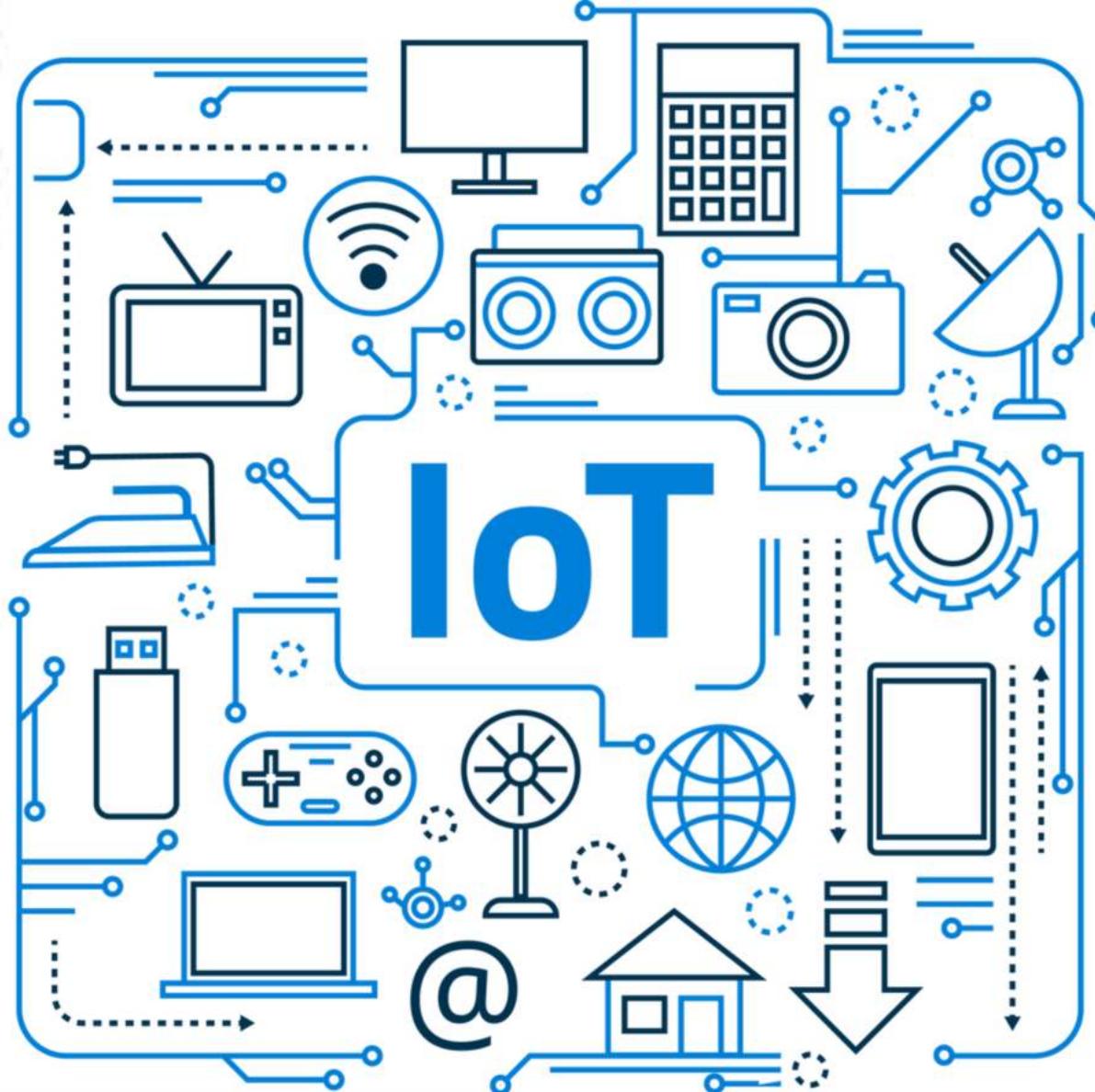
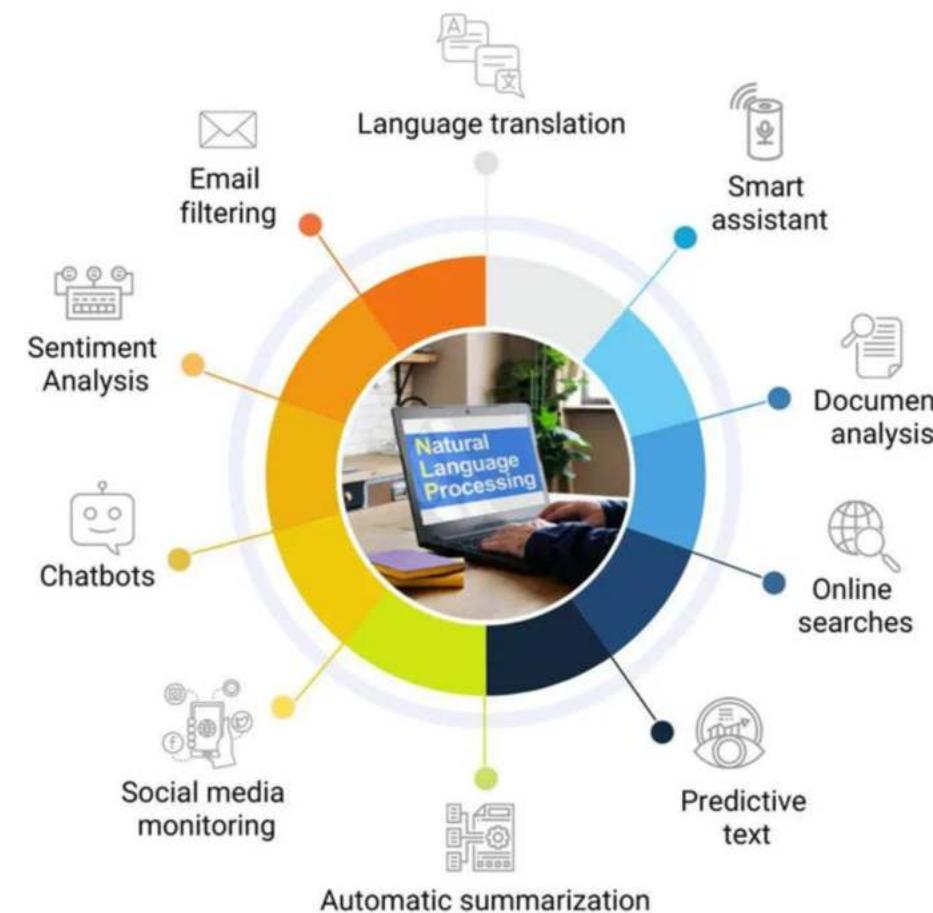


Tendencias Actuales en Inteligencia Artificial (IA) para 2024

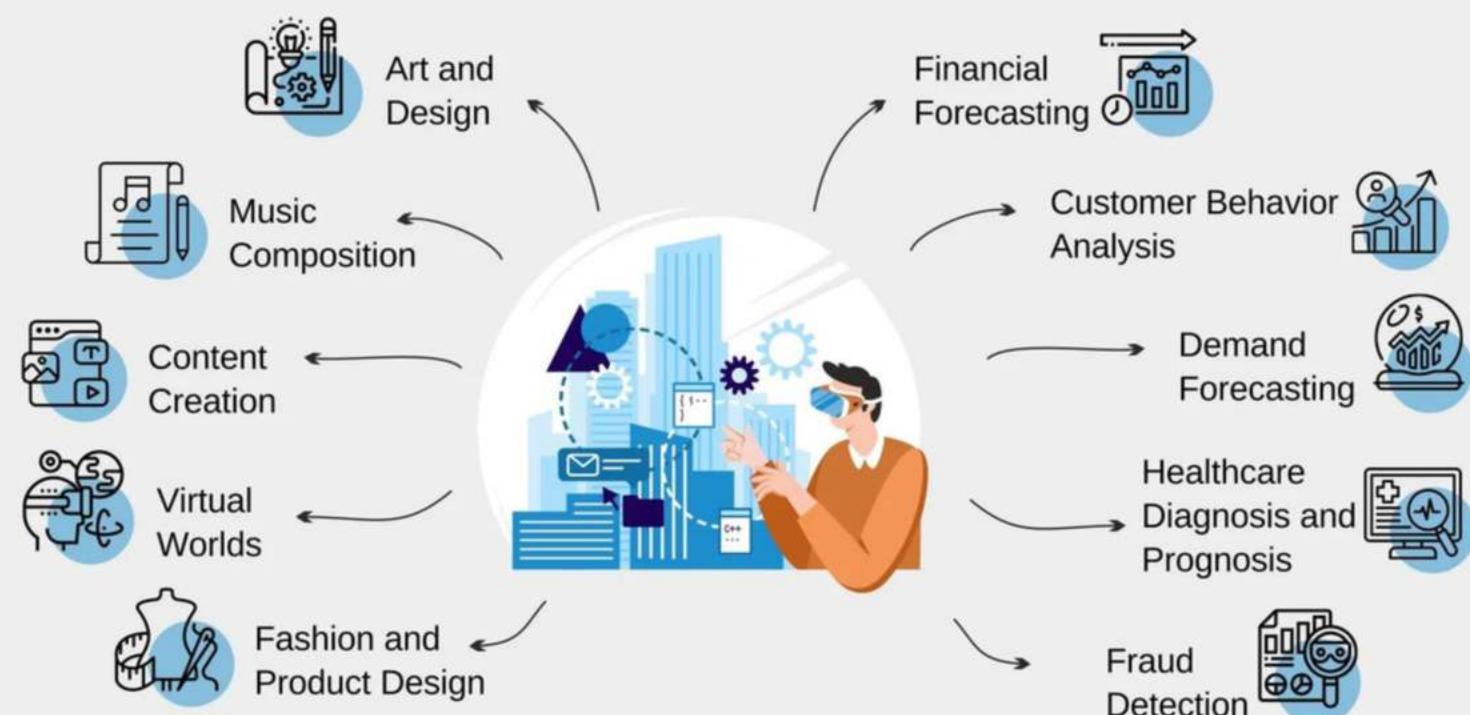
Descubre las últimas tendencias en IA



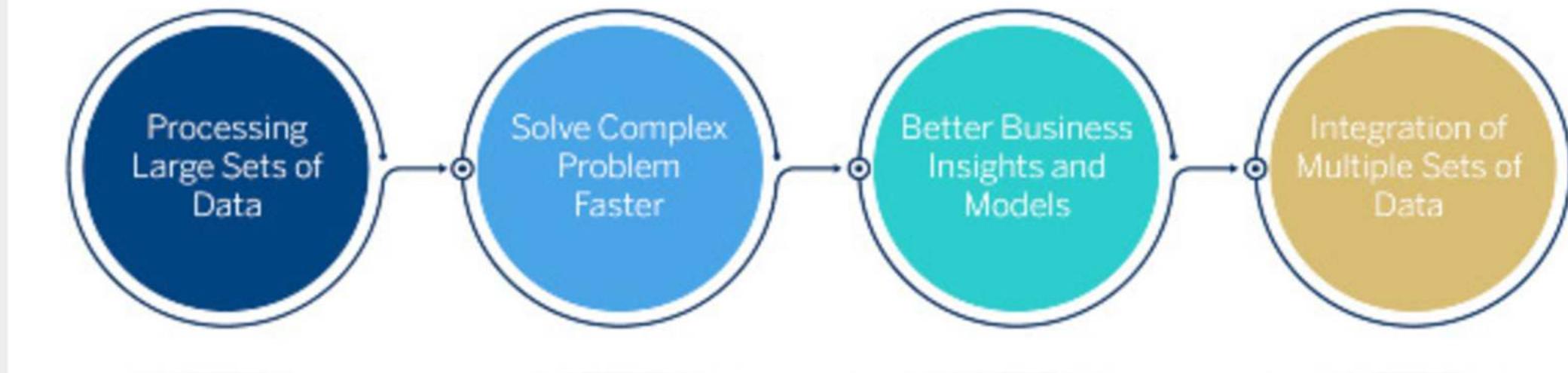
Applications of Natural Language Processing



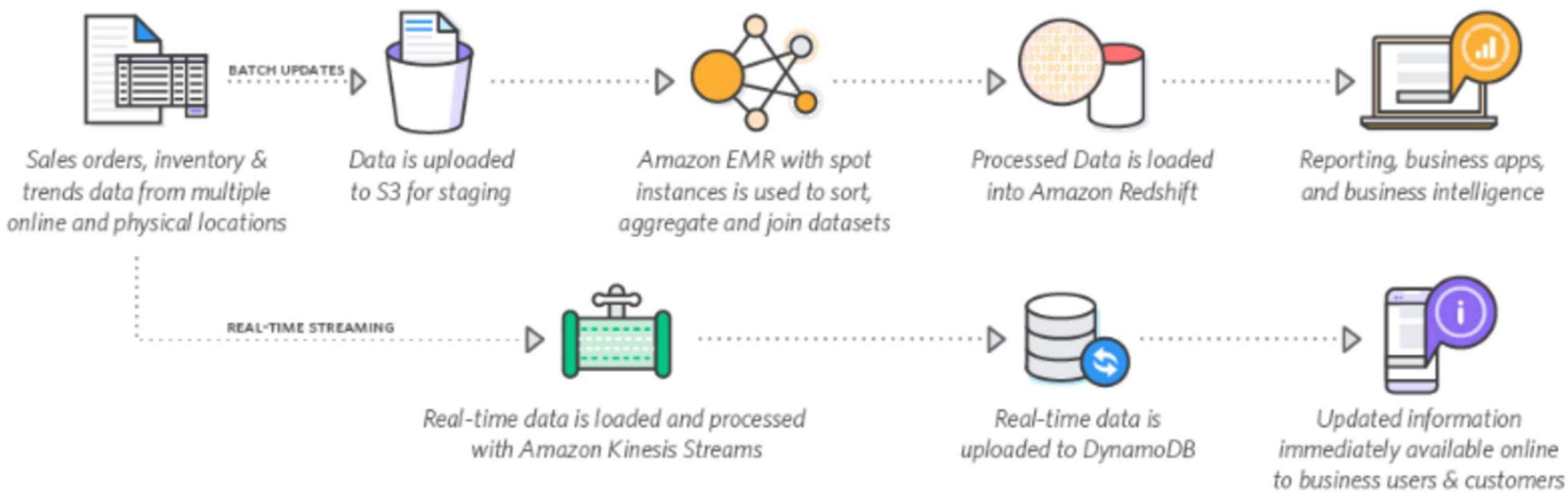
Generative AI Applications



Applications of Quantum Computing and AI



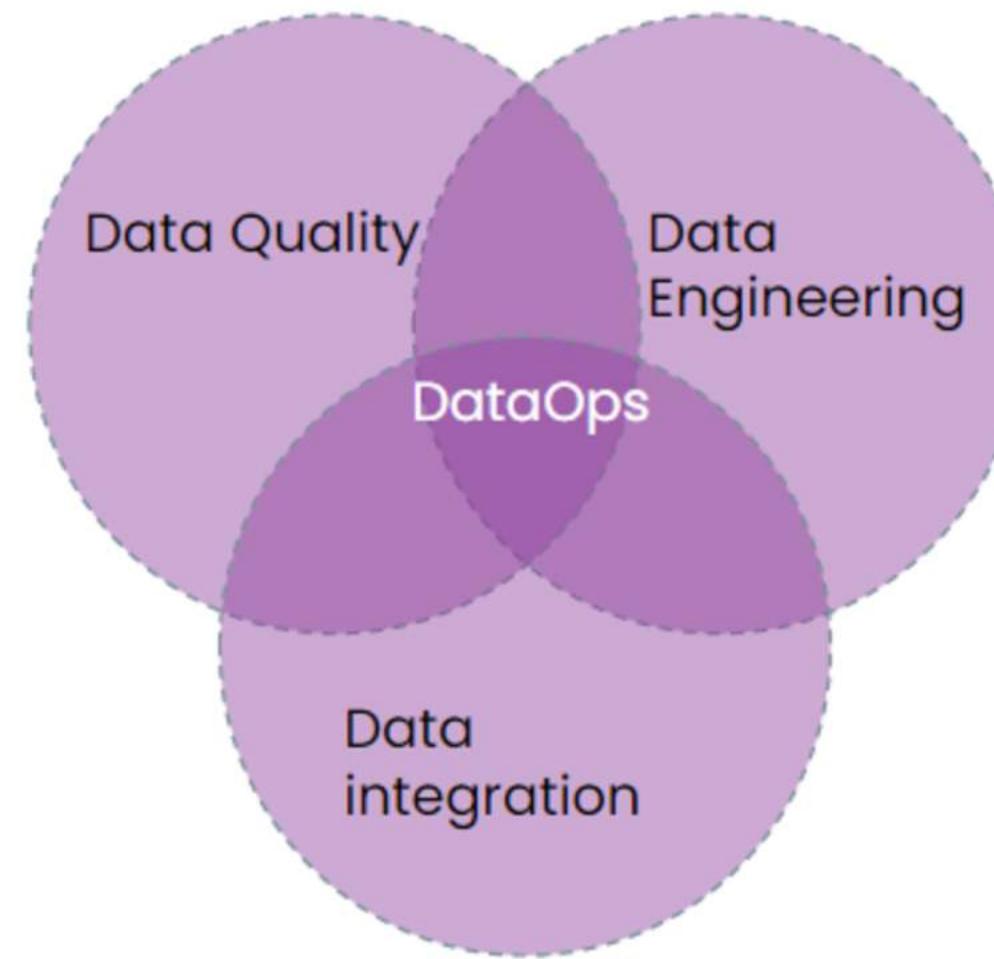
CIENCIA DE DATOS EN LA NUBE:



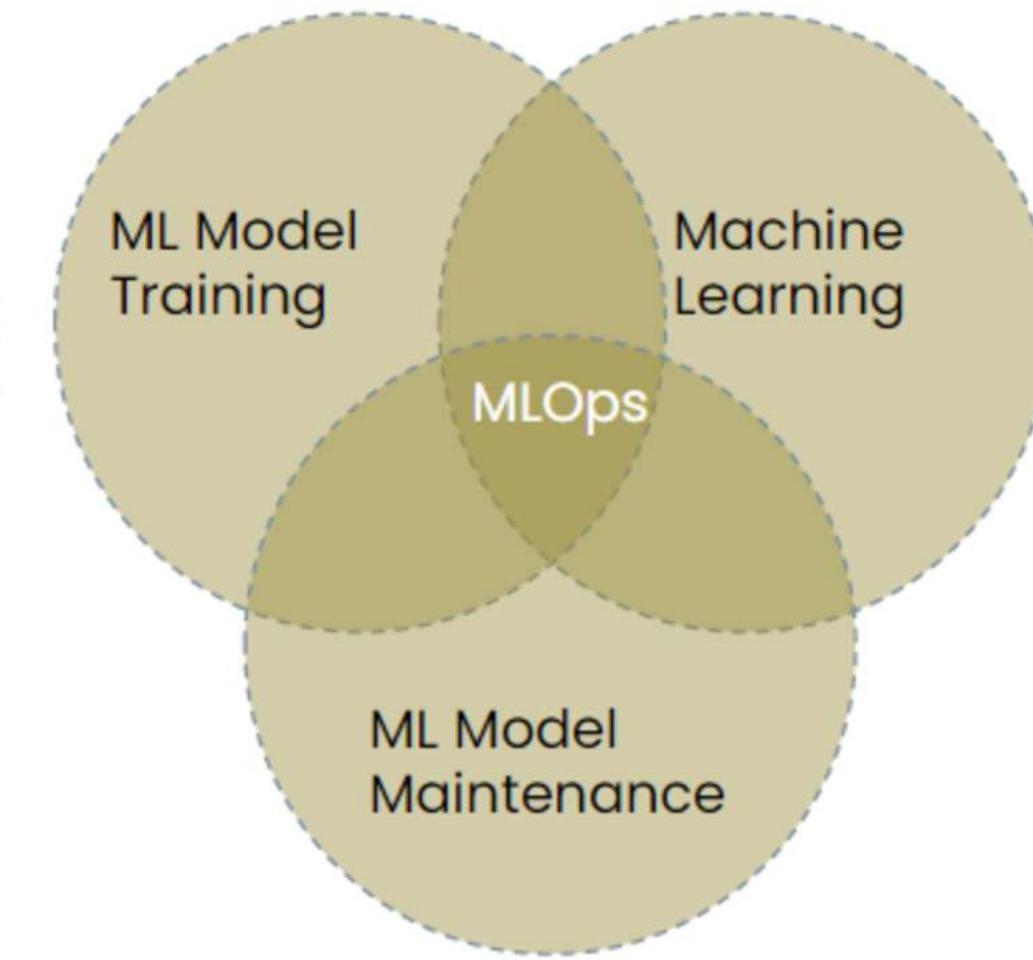
DATAOPS Y MLOPS



DevOps



DataOps



MLOps

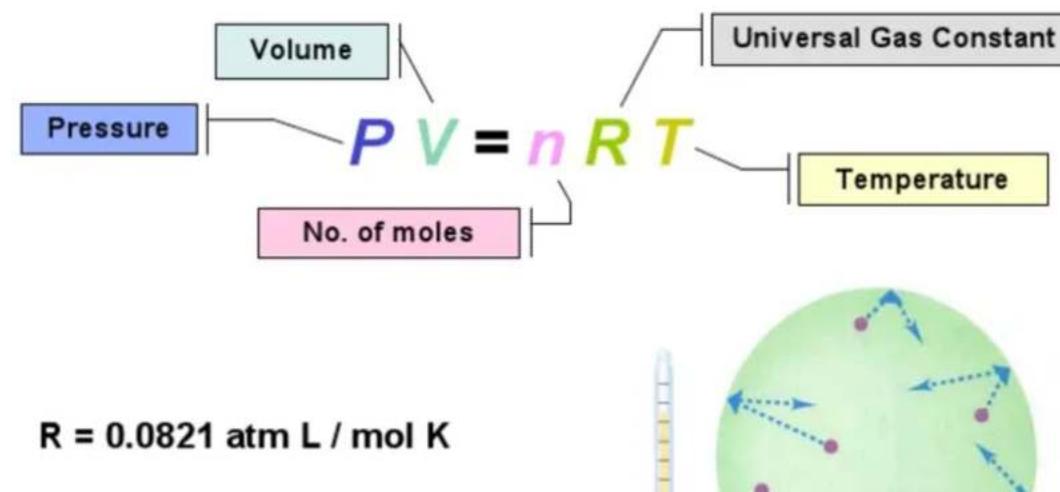
Ideal Gas Law

$$PV = nRT$$

Brings together gas properties.

Can be derived from *experiment and theory*.

Ideal Gas Equation



Estadística:

P multiplica a V y es proporcional a T
(Interpretabilidad a partir de pruebas empíricas a las hipótesis)

Machine Learning:

P y V pueden predecir T
(Predicciones precisas pero con casos de error)

Artificial Intelligence:

Mantener T deseable y variar P,V
(Control de automatización)

Lo que aprenderás...

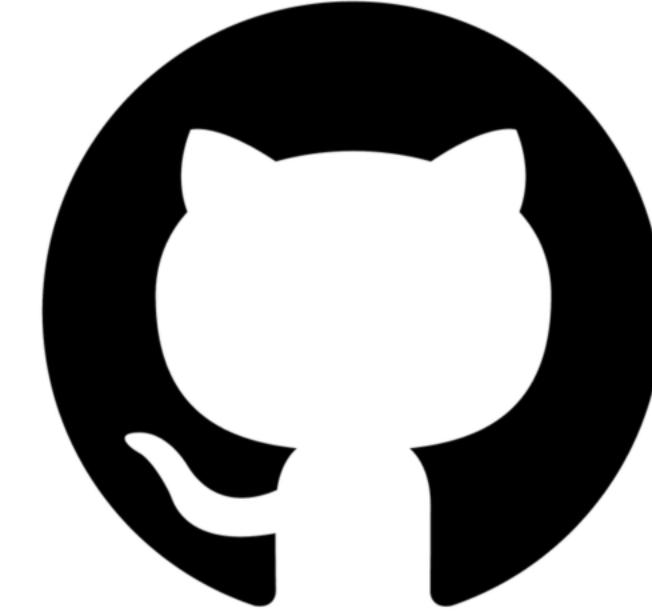


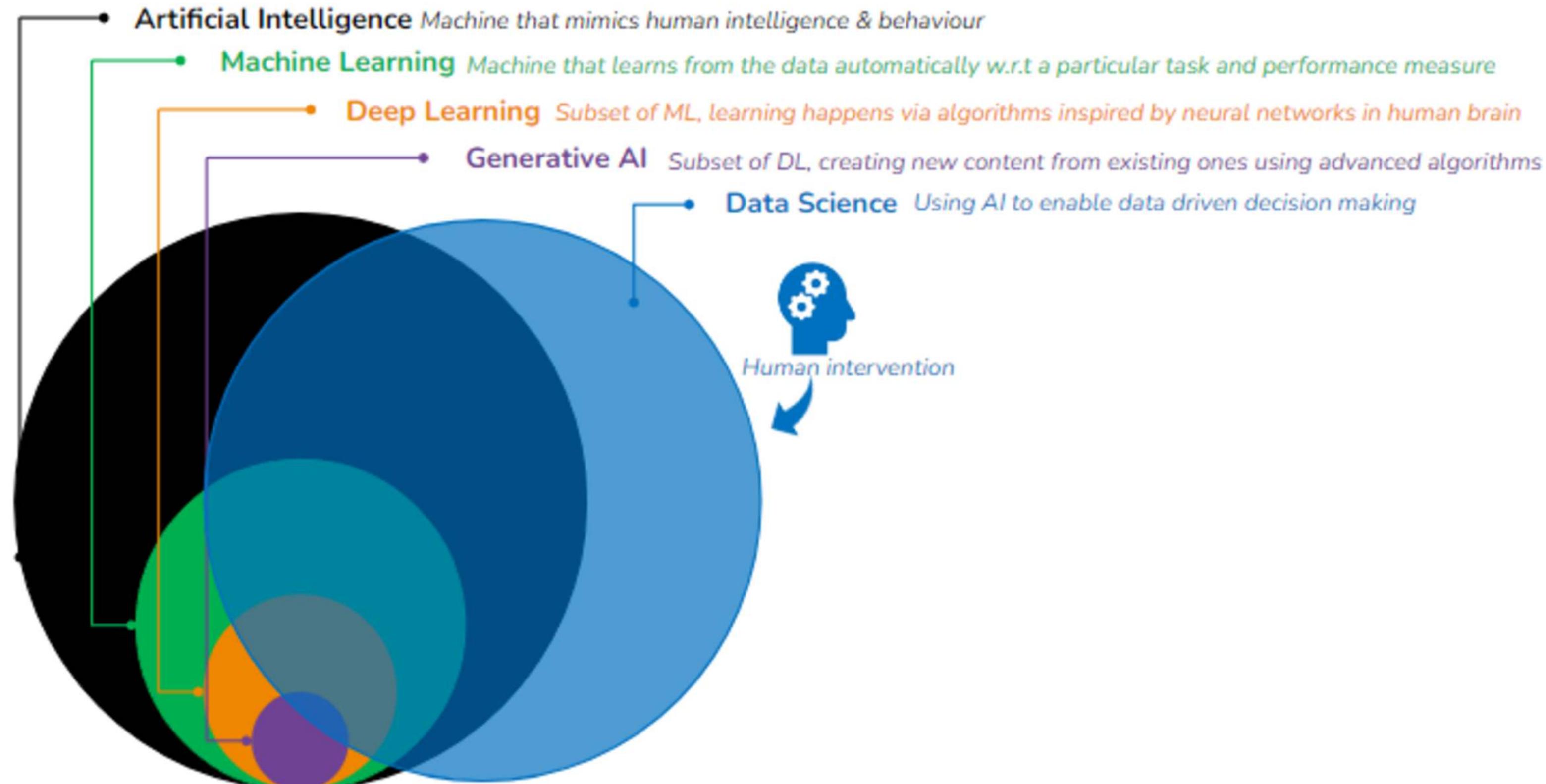
TensorFlow

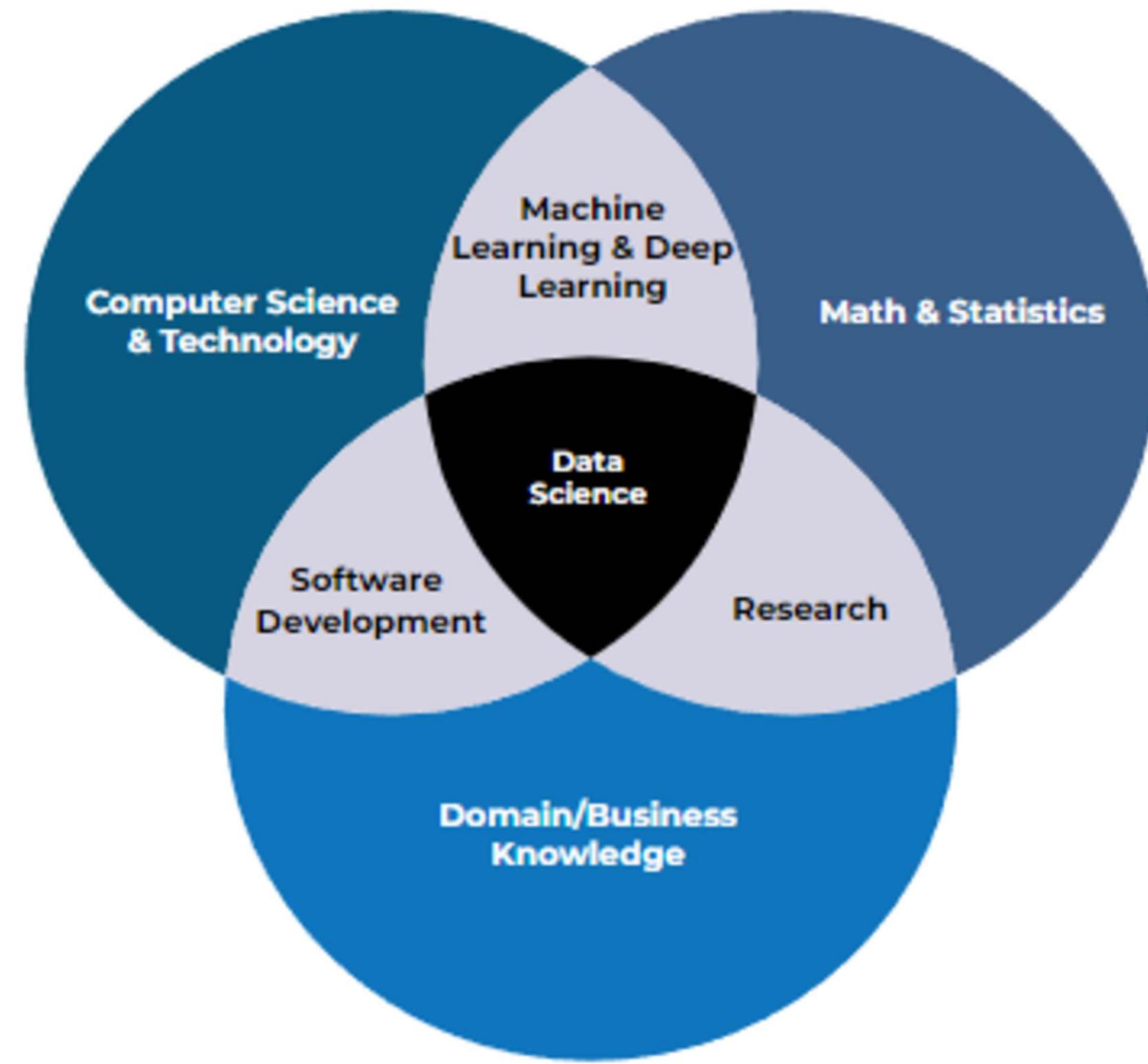
K Keras



ENTORNOS DE INTERÉS







CASOS DE ÉXITO EN CIENCIA DE DATOS

Detección de Fraude Financiero: Empresas como PayPal utilizan modelos de ciencia de datos para detectar actividades fraudulentas en tiempo real. Esto se logra analizando millones de transacciones y utilizando algoritmos que identifican patrones sospechosos.

Recomendaciones Personalizadas: Empresas como Netflix y Amazon utilizan sistemas de recomendación basados en ciencia de datos para sugerir productos o contenidos. Estos sistemas analizan el comportamiento de compra o visualización del usuario y los comparan con grandes conjuntos de datos para predecir lo que podría gustar al usuario.

Descubrimientos en Biomedicina: En el campo de la biomedicina, la ciencia de datos ha permitido avances significativos, como en la secuenciación del genoma humano y la identificación de biomarcadores para enfermedades. Esto ha llevado al desarrollo de terapias personalizadas en medicina.

Optimización de la Cadena de Suministro: Empresas como UPS y FedEx utilizan ciencia de datos para optimizar rutas de entrega y gestionar sus cadenas de suministro. Esto no solo reduce costos, sino que también mejora la eficiencia y la satisfacción del cliente.

CASOS DE ÉXITO EN CIENCIA DE DATOS

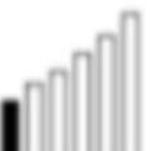
Predicción Meteorológica y Cambio Climático: Los científicos utilizan grandes conjuntos de datos y modelos avanzados para predecir el clima y entender mejor los patrones del cambio climático. Esto es crucial para la planificación en sectores como la agricultura y la gestión de desastres.

Análisis de Sentimientos en Redes Sociales: Las empresas utilizan la ciencia de datos para analizar opiniones y tendencias en redes sociales, lo que les permite entender mejor la percepción del cliente y ajustar sus estrategias de marketing y producto.

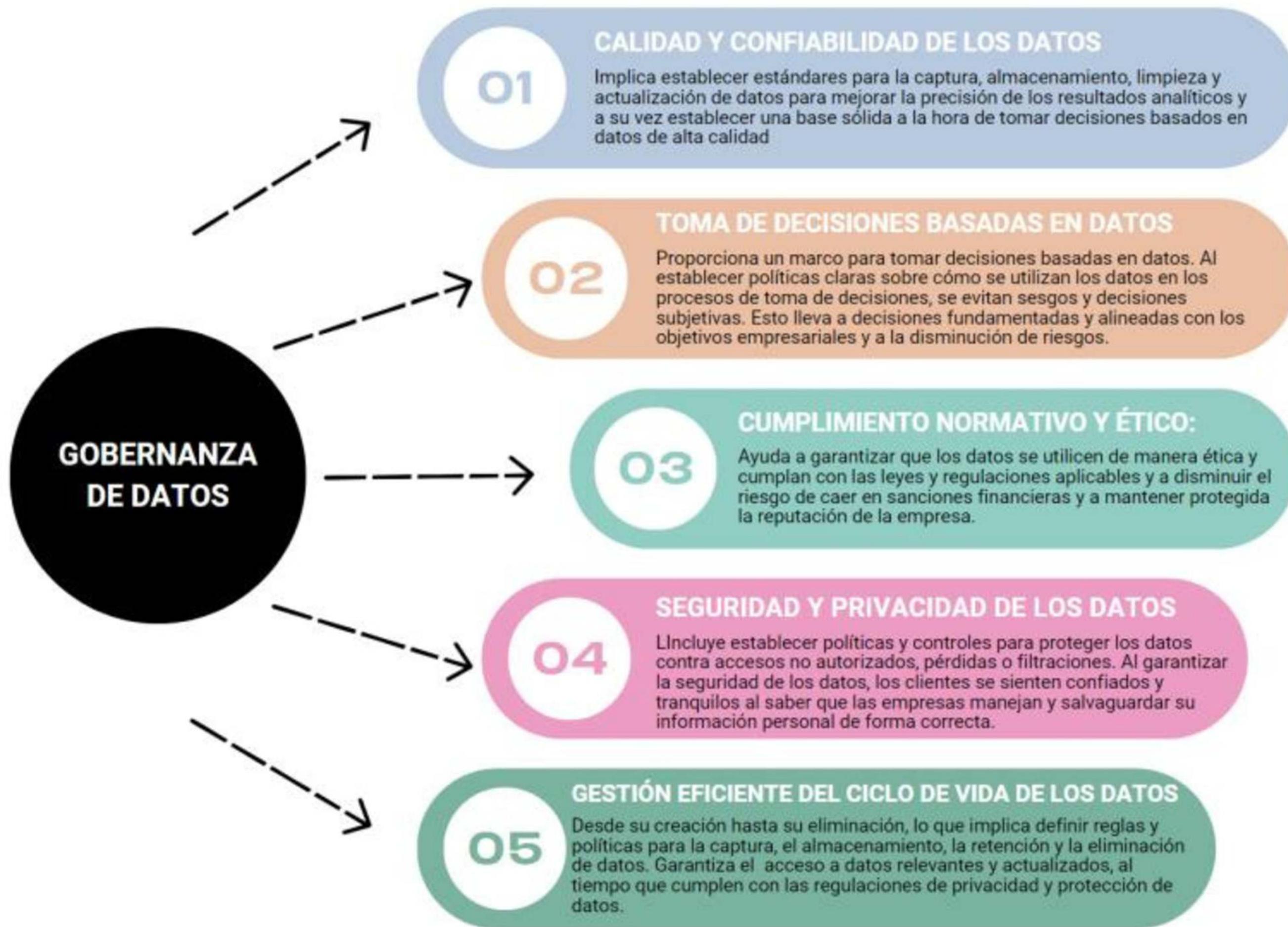
Mejora del Rendimiento Deportivo: En el deporte, la ciencia de datos se utiliza para analizar el rendimiento de los atletas, mejorar las estrategias de entrenamiento y prevenir lesiones.

Automatización y Vehículos Autónomos: Empresas como Tesla y Waymo están utilizando la ciencia de datos en el desarrollo de vehículos autónomos. Esto implica el análisis de enormes cantidades de datos de sensores para mejorar la seguridad y la eficiencia de estos vehículos.

Data warehouse vs Data mart vs Data lake

	Most Important Use Group & Use-Cases	Time-to-Market Questions & Solutions	Cost Implementation & Ownership	Users (# & Types)	Data Growth Volume & Variety
Data Lake	Predictive & Advanced Analytics	 Weeks - Months	\$\$\$\$\$		
Data Warehouse	Multi-Purpose Enabler of Operational & Performance Analytics	 Hours - Days	\$\$\$\$		
Data Mart	Line of Business Specific Reporting & Analytics	 Minutes - Hours	\$\$\$\$\$		

ÉTICA Y GOBERNANZA DE DATOS:



GOBIERNO DE LOS DATOS

- Los datos son un activo de gran valor que permiten tomar decisiones e implementar estrategias de negocio, también permiten comprender hábitos de consumo y la forma en cómo se organizan las sociedades. El orden de los datos exige una estrategia para el gobierno de los datos.
- Se requiere una gestión global de la disponibilidad, facilidad de uso, la integridad y seguridad de los datos empleados en una empresa.
- Es necesario organizar e implementar políticas, procedimientos y normas para el uso eficaz de los activos de información, ya sean estructurados o no estructurados de una organización.

Data Quality

Perfilado de datos, políticas y guías de calidad de datos y monitorización de la calidad del dato

Metadata

Definición del modelo de metadata, incluyendo tanto su descripción técnica como de negocio

Data Warehouse & BI

Definición de arquitectura de DWH y sistemas de reporting para asegurar el uso correcto de los datos

Reference & Master Data

Definición de requisitos y modelos de datos maestros críticos para la organización

Documents & Content

Políticas y actividades para el control del ciclo de vida de los datos sobre cualquier tipo de medio



Data Architecture

Es el diseño de la estructura tanto física como lógica de los sistemas que manejan los datos

Data Modeling & Design

Es el modelo lógico de los datos y como se implementa en nuestra organización

Data Storage & Ops

Diseño de bases de datos, entornos tecnológicos entornos y planes de continuidad asociados

Data Security

Diseño y desarrollo de políticas, estándares y auditoria de la seguridad y cumplimiento regulatorio

Data Integration

Diseño e implementación de arquitecturas y estándares de interoperabilidad e integración de datos

Un adecuado gobierno de los datos al interior de las organizaciones resulta esencial para:

- Obtener datos consistentes y fiables
- Potenciar la integración del negocio
- Guiar los procesos de análisis y evitar problemas de reporting
- Prevenir conflictos entre los distintos conjuntos de datos
- Minimizar el riesgo en la toma de decisiones
- Garantizar que todas las operaciones cumplen con los requisitos legales mínimos exigibles
- Armonizar los procesos de negocio a nivel global
- Ahorrar costes

Riesgos asociados al Big Data

PRIVACIDAD SEGURIDAD

- Información sensible expuesta
- Accesos no autorizados o violaciones de seguridad

CALIDAD

- Las inconsistencias en los datos pueden llevar a decisiones erróneas
- Proceso de limpieza y validación



NORMATIVA

- Regulaciones gubernamentales
- Protección de los datos

GESTIÓN

- Complejidad
- Costo

SESGOS

- Decisiones discriminatorias e injustas

TRANSPARENCIA

- Confianza
- Partes honestas

DEPENDENCIA

- Fallas en el sistema, interrupciones de servicios o pérdida de datos

ÉTICA

- Mitigación de impactos
- Individuos y grupos con derechos

CONSIDERACIONES LEGALES

Legalmente, ¿Qué hay detrás de todo esto?

Consideraciones legales y reglamentarias que las organizaciones deben tener en cuenta para garantizar el cumplimiento y evitar problemas legales.

PRIVACIDAD

Reglamento General de Protección de Datos (GDPR) en la Unión Europea Ley de Privacidad del Consumidor de California (CCPA) en los Estados Unidos Restricciones en recopilación, procesamiento almacenamiento de datos personales

SEGURIDAD

Confidencialidad, integridad y disponibilidad de los datos

TRANSPARENCIA

Responder a la pregunta respecto a Cómo se toman las decisiones basadas en algoritmos de Big Data.

TRANSFERENCIA INTERNACIONAL

Cláusulas Contractuales Estándar (SCC) o los Códigos de Conducta pueden ser necesarios para garantizar la legalidad de la transferencia de datos.

RESPONSABILIDAD

Prácticas de manejo de datos: implementación de políticas y procedimientos adecuados, así como la designación de responsables de protección de datos.

DERECHOS

Acceso, rectificación, eliminación y portabilidad

Normativa Nacional Colombiana

Cómo estamos a nivel de Colombia?

LEY 1581 DE 2012 - LEY DE PROTECCIÓN DE DATOS PERSONALES

- Esta ley establece los principios y normas para la protección de datos personales en Colombia.
- Define conceptos clave como datos personales, tratamiento de datos, responsables del tratamiento, y titulares de los datos personales.
- Establece los derechos de los titulares de datos, incluyendo el derecho de acceso, rectificación, actualización y supresión de datos personales.
- Impone obligaciones a los responsables y encargados del tratamiento de datos, incluyendo la adopción de medidas de seguridad, obtención de consentimiento, y notificación de brechas de seguridad.
- Contempla sanciones por el incumplimiento de las disposiciones de la ley.

DECRETO 1377 DE 2013

- Reglamenta algunos aspectos de la Ley 1581, incluyendo el registro de las bases de datos ante la Superintendencia de Industria y Comercio.
- Establece el contenido y formato del aviso de privacidad que debe ser proporcionado a los titulares de datos.
- Proporciona directrices sobre el manejo de solicitudes de los titulares de datos para el ejercicio de sus derechos.

SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO (SIC)

- La SIC es la entidad encargada de supervisar y controlar el cumplimiento de la normativa de protección de datos en Colombia.
- La Superintendencia puede imponer sanciones por el incumplimiento de las disposiciones de la Ley 1581.

¿Cómo de relevante es una regulación? ¿Por qué es importante para la organización?

¿Cómo la interpretamos? ¿Qué políticas y procedimientos requiere?

¿Ya se están cumpliendo sus preceptos? ¿En qué medida? ¿Es suficiente para el futuro?

¿Cómo puede la organización demostrar el cumplimiento?

¿Cómo monitorizar el ajuste a las leyes aplicables? ¿Con qué frecuencia se revisa el cumplimiento?

¿Cómo se informa acerca de cuestiones relacionadas con el cumplimiento regulatorio? ¿Cómo se gestionan y se corrigen las desviaciones?

ROLES EN DATOS

¿Quién hace qué?

Analista de Datos (Data Analyst):

Alcance:

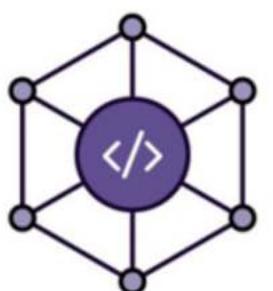
Se centra en el análisis de datos para ayudar en la toma de decisiones.

Utiliza herramientas estadísticas y de visualización de datos para interpretar conjuntos de datos, identificar tendencias y patrones, y presentar resultados a los stakeholders.

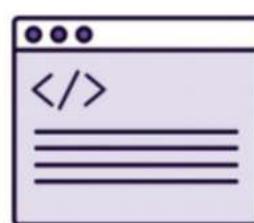
Herramientas comunes:

SQL, Excel, R, Python, Tableau, Power BI.

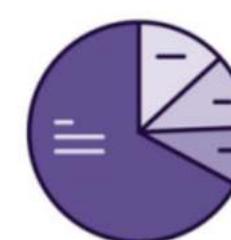
7 Essential Data Analyst Skills



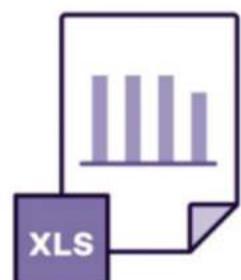
SQL and
NoSQL



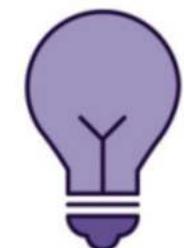
R, Python,
and MATLAB



Data
Visualization



Microsoft
Excel



Critical
Thinking



Math and
Statistics



Communication

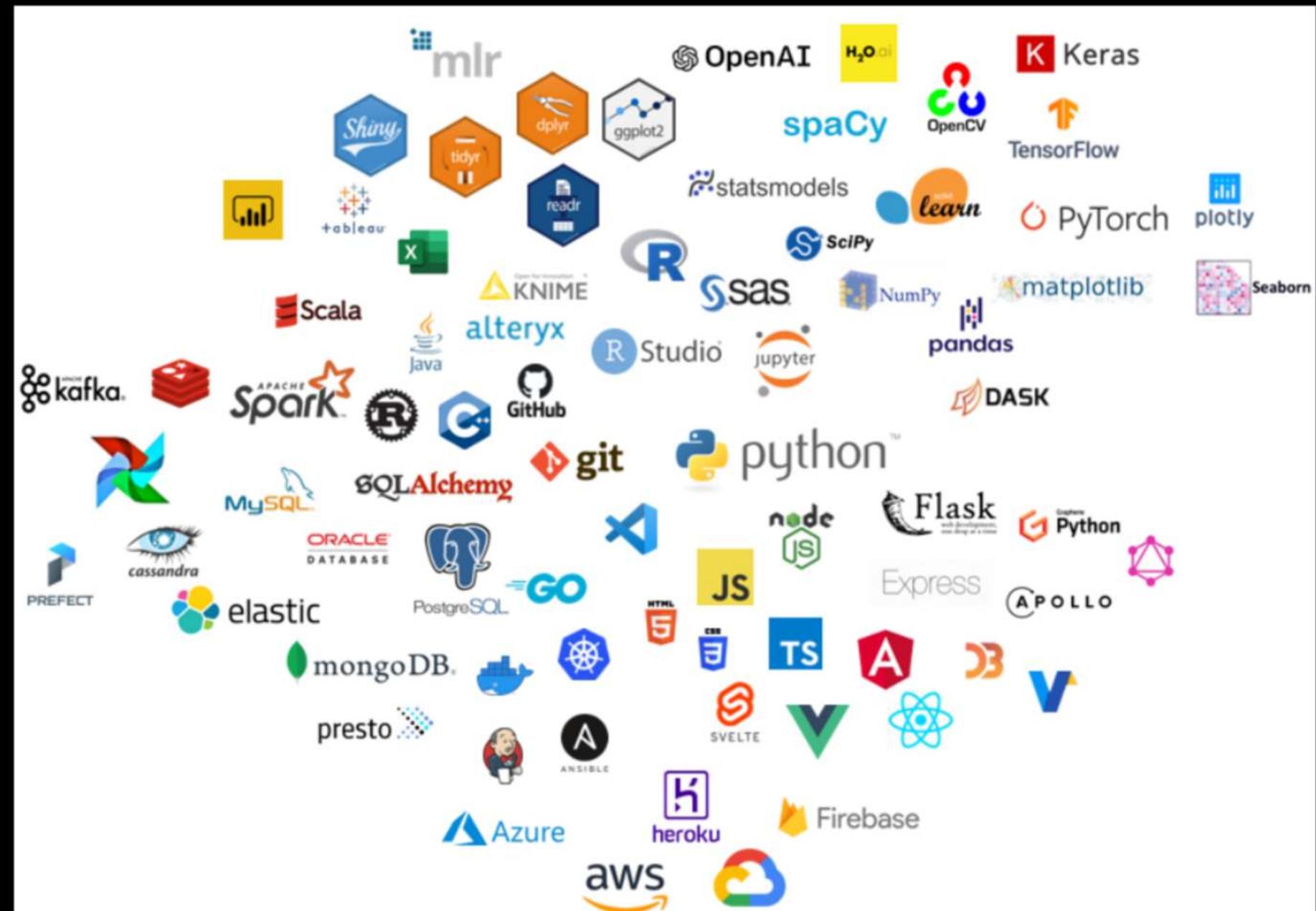
Científico de Datos (Data Scientist):

Alcance:

Va más allá del análisis básico de datos para incluir técnicas de aprendizaje automático, minería de datos y modelado predictivo.

Desarrolla algoritmos avanzados para interpretar grandes volúmenes de datos y extraer insights más profundos.

Herramientas comunes: Python, R, TensorFlow, scikit-learn, Spark.



Ingeniero de Datos (Data Engineer):

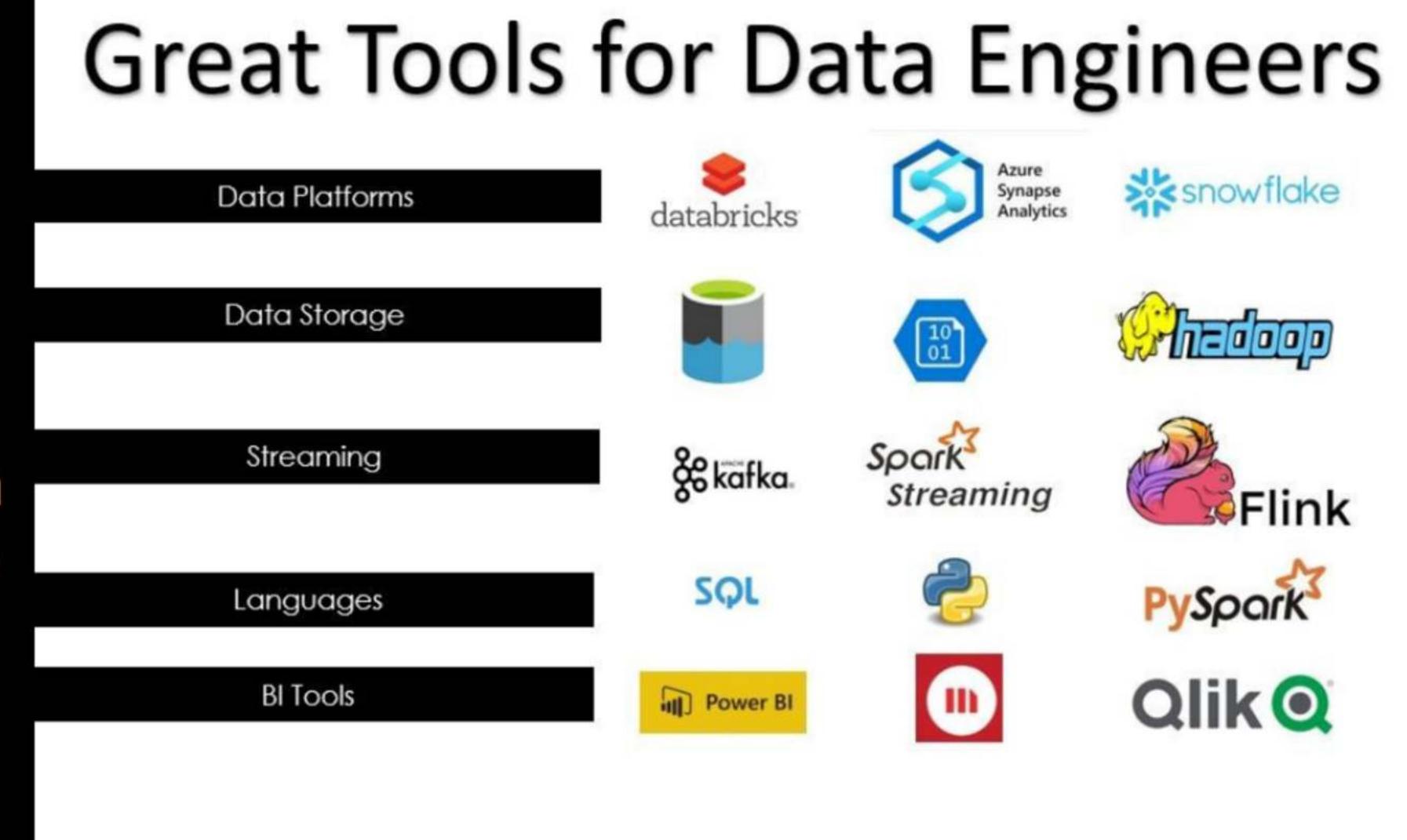
Alcance:

Se enfoca en la preparación y arquitectura de sistemas de datos.

Construye y mantiene la infraestructura de datos (como bases de datos, data warehouses, y pipelines de datos) que permite el almacenamiento, la limpieza y la manipulación eficiente de grandes volúmenes de datos.

Herramientas comunes:

Hadoop, Spark, Kafka, SQL, AWS, Azure, Google Cloud.



Administrador de Bases de Datos (Database Administrator, DBA):

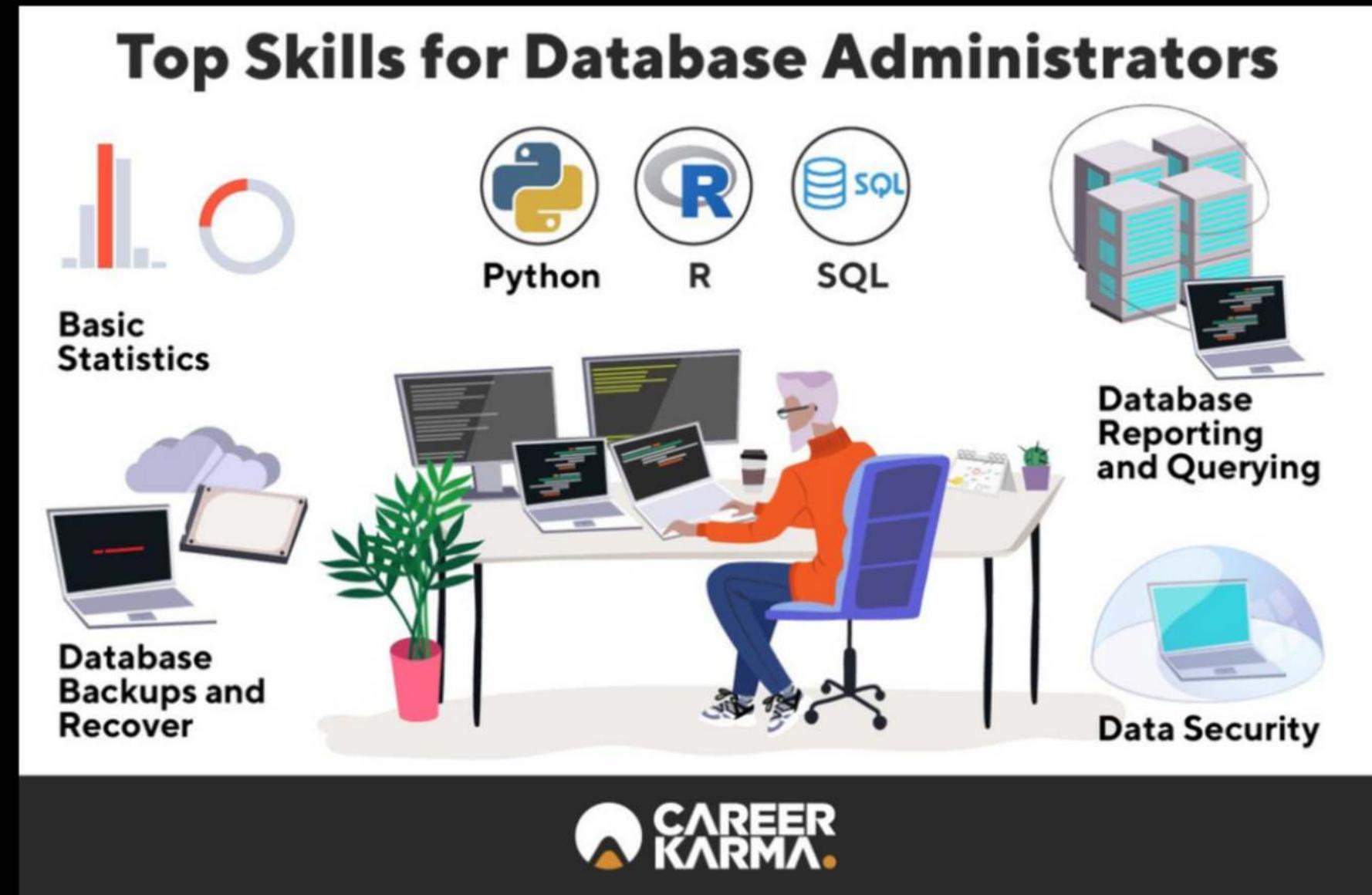
Alcance:

Se especializa en el diseño, la implementación, el mantenimiento y la reparación de la base de datos de una organización.

Garantiza la disponibilidad, rendimiento y seguridad de las bases de datos.

Herramientas comunes:

MySQL, PostgreSQL, Oracle, Microsoft SQL Server



Analista de Negocio (Business Analyst):

Alcance:

Combina el análisis de datos con conocimientos empresariales.

Ayuda a la organización a entender las tendencias del mercado y las necesidades del negocio, y cómo los datos pueden ser utilizados para lograr objetivos empresariales.

Herramientas comunes:

Excel, SQL, herramientas de BI como Tableau o Power BI.

