DATING·SCIENCE

HAZ MATCH CON LA
DISCIPLINA DE LOS DATOS


TE INVITAMOS A LA EXPERIENCIA:
NUCLIO DIGITAL SCHOOL x RICARDO CAVOLO.
DESCUBRE EL FUEGO DEL ARTE, LA
REVOLUCIÓN DEL MUNDO DIGITAL,
Y CUÁL ES TU ROL EN TODO ESTO.

DATING·SCIENCE

HAZ MATCH CON LA DISCIPLINA DE LOS DATOS

NUCLIO DIGITAL SCHOOL × RICARDO CAVOLO

STARTUP YOUR LIFE!

Ofrecer información útil e interesante, proveniente de fuentes confiables e influyentes, a través de la atracción que genera el arte de Ricardo Cavolo; es una excelente manera de promover la digitalización.

Por eso estamos aquí. Para motivarte a descubrir lo más importante acerca de Data Science. Una disciplina que mueve al mundo y que ahora te movilizará a ti.

Sobre **nosotros**

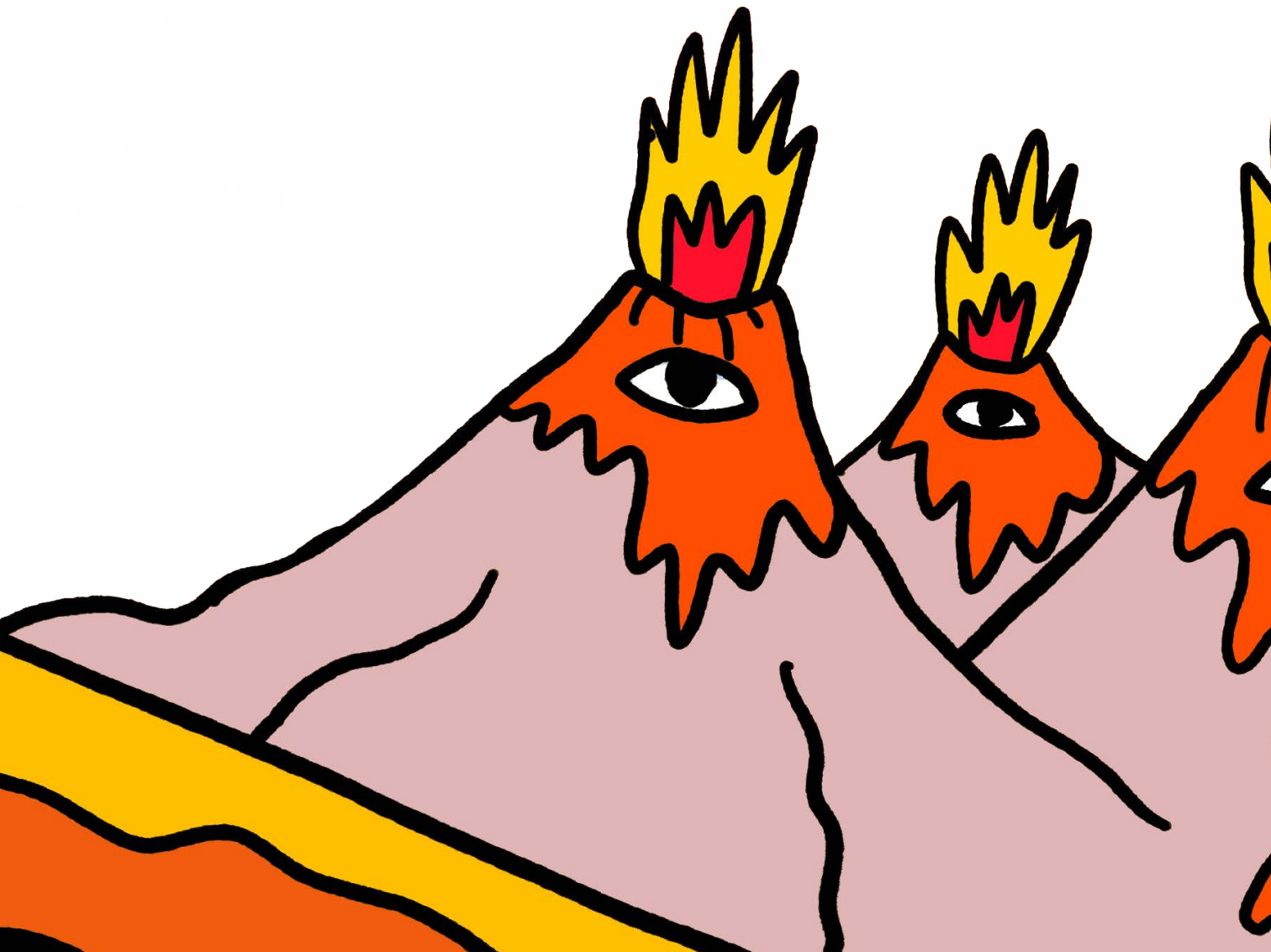
Nuclio
Digital School

Nace de la incubadora de startups Nuclio Venture Builder, con la necesidad de formar a perfiles especializados en el sector tecnológico, con una metodología práctica y centrada en el alumno.

Creación

En 2018, el principal Business Angels de España y presidente de Nuclio, Carlos Blanco, junto al experto en la innovación de modelos educativos, Jared Gil, tuvieron una serie de reuniones que los llevarían a tomar una gran decisión: crear Nuclio Digital School. Una escuela digital enfocada en cubrir la alta demanda de perfiles TIC que tiene la industria.

Para la mayor Venture de España, tenía todo el sentido crear una escuela de másteres con metodología bootcamp especializada en perfiles digitales, no solo para proveer de estos a las startups del grupo, sino a toda la industria.



“Todos los inicios son complicados y siempre inunda la incertidumbre los primeros meses/ años. Pero nosotros hemos contado con un ecosistema y socios que hicieron que el camino sea más seguro” - Palabras de Jared Gil, CEO y Co-Fundador de Nuclio Digital School.

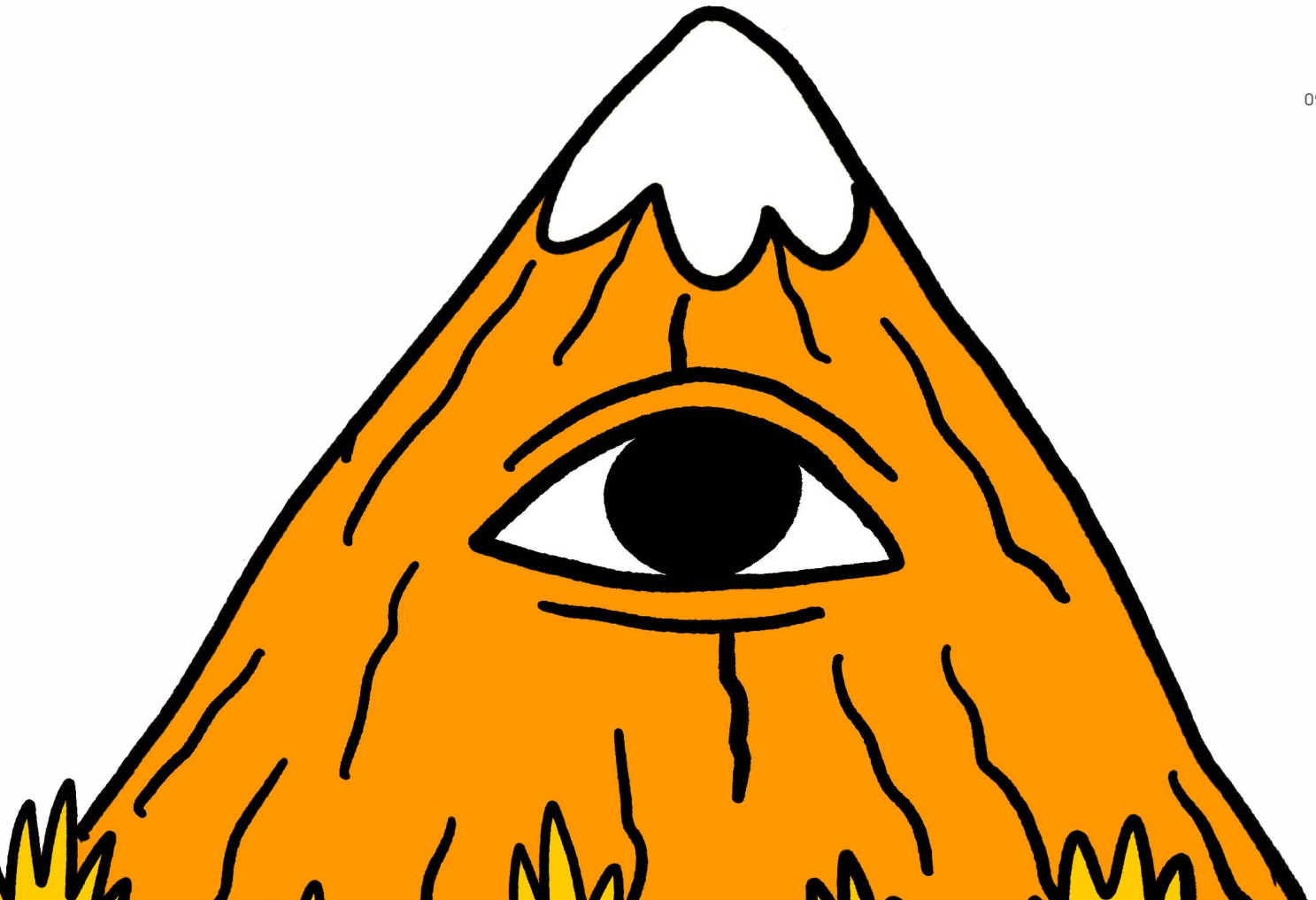
Desarrollo

Establecer una nueva empresa en el mercado y conseguir que sea la #1 no es fácil. ¡Para ello tienes que rodearte de los mejores! Tanto socios como compañeros de trabajo. Y puede que la parte más difícil sea esa: encontrar a las personas adecuadas para llevar a cabo el proyecto con éxito.

En Nuclio Digital School lo logramos y nos convertimos rápidamente en una escuela referente en España. Creando los programas de formación más innovadores y nutriendo a las empresas de perfiles técnicos de alto rendimiento.

La clave del desarrollo fue ofrecer un upskilling y reskilling de los perfiles profesionales, en tan solo 5 meses. Con un gran aporte de valor: networking con profesores en activo de compañías top del mercado.

En NDS preparamos a los estudiantes para generar un gran impacto en las empresas o emprendimientos propios. Dotándolos de conocimientos que no se pueden adquirir en las carreras tradicionales.

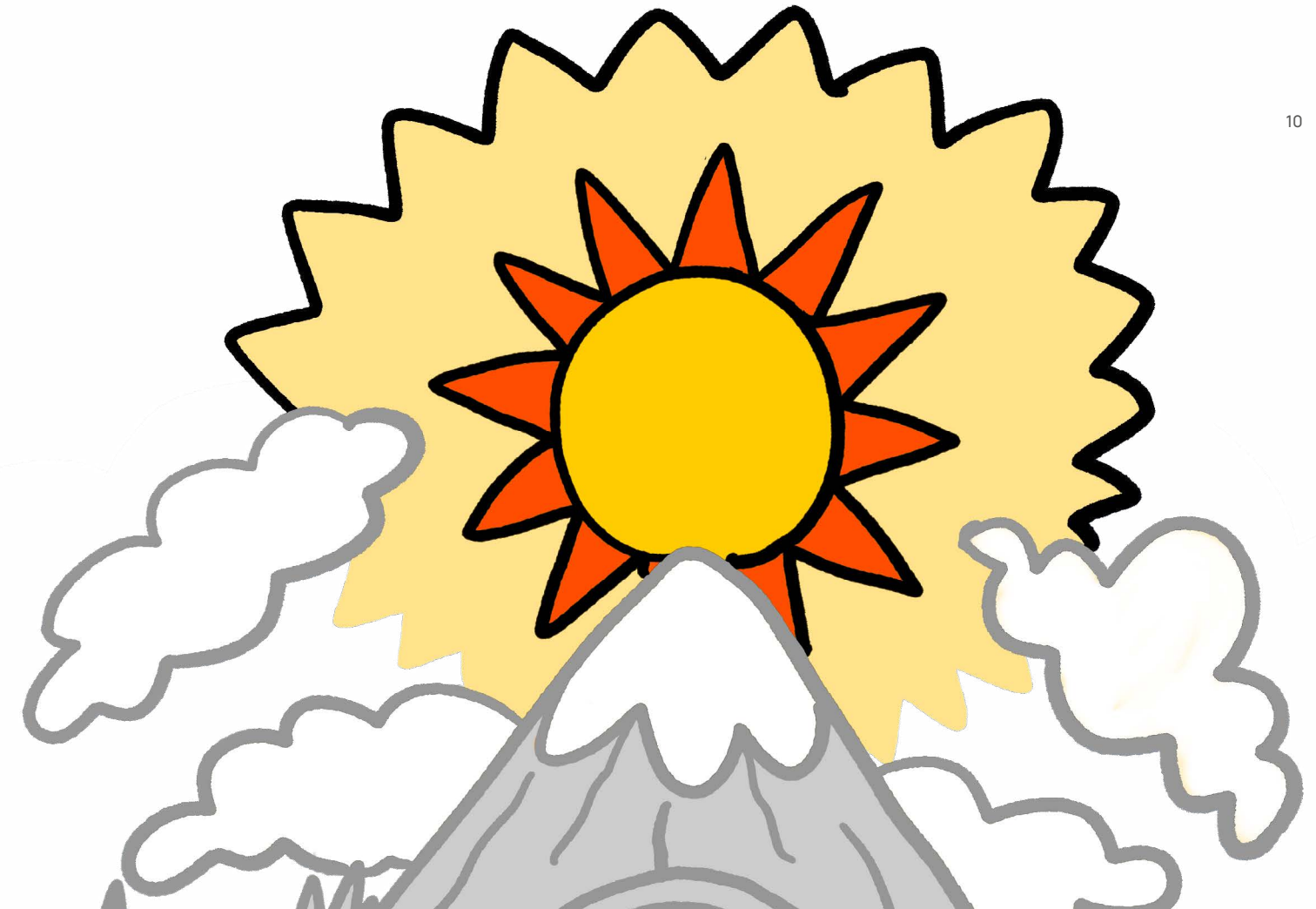


Expansión

En solo algunos años, nos hemos posicionado como la principal escuela de negocios digitales de España. Y ya ha sido galardonada con premios como “European Technology Awards” y “Educational Excellence Awards”.

Con el objetivo claro de digitalizar a todos los profesionales del mundo, hacia 2022 logramos afirmar la expansión, aterrizando en mercados como Latam (México, Brasil, y Chile, entre otros), y Europa (Alemania, Países Bajos, Portugal, y EAU). Sin dejar de lado las ciudades que la vieron nacer y donde se ubican sus tech centers: Barcelona y Madrid.

Nuestra escuela ha logrado también ampliar su oferta de formaciones, y actualmente ofrece másteres especializados en digitalización, en áreas como programación, marketing, diseño, negocios, finanzas y recursos humanos. Brindando las modalidades presencial y streaming.



Dime tu fuente y te diré quién eres

El mundo ya no habla, grita. Cada día escuchamos miles de voces que dicen ser la voz de la verdad. Mientras tanto, tú necesitas estar informado sobre los temas que más te interesan, pero te cuesta decidir en quién confiar.
Lo sabemos y por eso queremos presentarte a nuestros colaboradores.

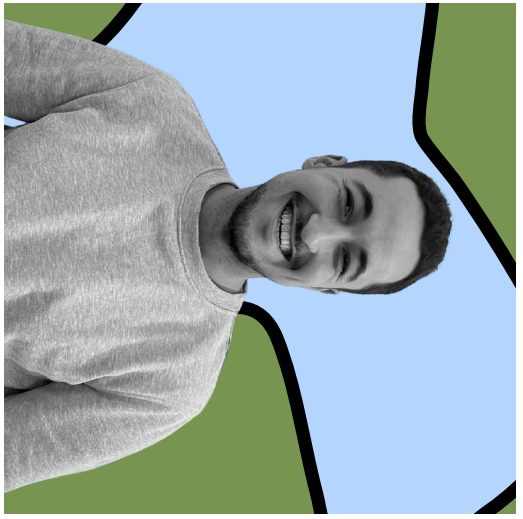
Hemos seleccionado a un grupo de profesionales líderes del sector para que escriban sobre Data Science, tal y como nos hubiera gustado leer tantas veces. Además, mientras te dejas llevar por sus palabras, encontrarás Inteligencia Artificial gracias a las imágenes generadas junto a DALL·E.

Esta es tu oportunidad de escuchar en primera persona a distintos expertos que viven en contacto día y noche con esta disciplina. Perfiles con experiencia de trabajo en reconocidas empresas y emprendimientos propios, que han formado parte de la comunidad de Nuclio Digital School y quieren que también te unas a la revolución digital.

Al igual que ellos, tú puedes hacer la diferencia.
¿Nos acompañas?

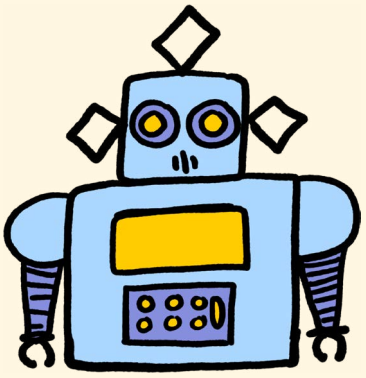
Espartaco Camero _ Responsable de Data Science & Analytics
Jesús Prada _ Responsable de Machine Learning
Massimiliano Brevini _ Data Analyst Senior

Carlos Pérez _ Responsable de Data Science
Toni Badia _ Data Scientist Senior



15 (I) **Big Data**

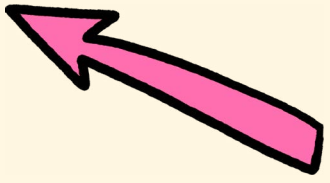
31 (II) **Inteligencia Artificial**

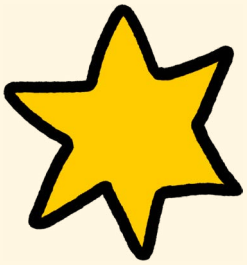


54 (III) **Machine Learning**

70 (IV) **Deep Learning**

93 (V) **Data Science**



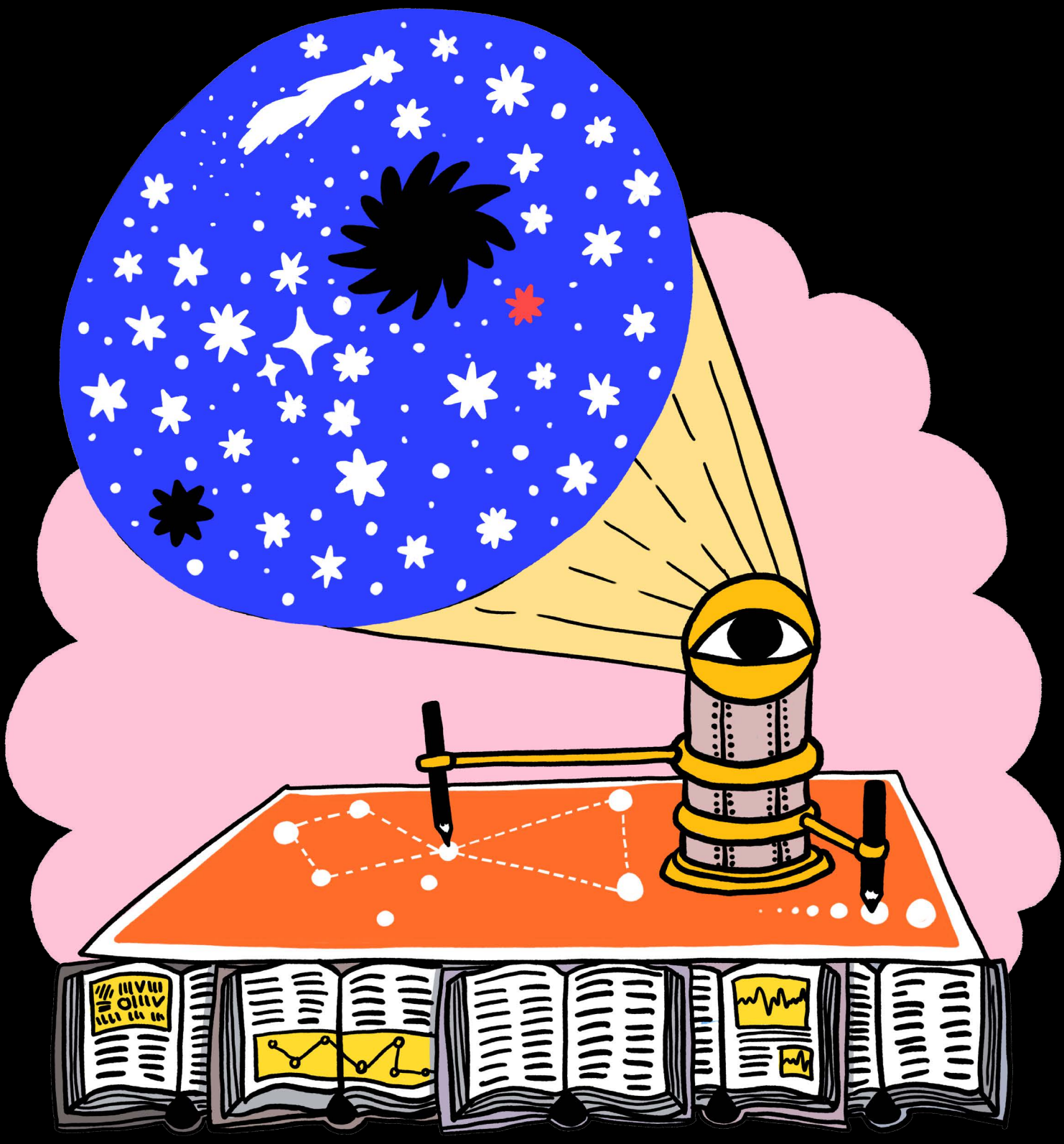


(|) Big Data

Estamos de acuerdo en que hoy los datos son extremadamente grandes y complejos como para ser procesados y analizados utilizando las técnicas tradicionales.

Pero, ¿qué tan confiable es tomar decisiones con las predicciones obtenidas del Big Data?

POR TONI BADIA



¿QUÉ ES?

Se conoce como la recopilación, selección, filtrado y análisis de gran cantidad de datos utilizando algoritmos informáticos. El objetivo es obtener información relevante de forma rápida y dinámica, de manera tal que pueda ser mostrada de forma organizada y preferiblemente visual, a la parte interesada.

El origen de los datos y la relevancia de los mismos es parte fundamental en la obtención de resultados realmente relevantes, que ayuden a extrapolar resultados futuros o ayuden a inferir tendencias de diferente índole.

La definición de Big Data parece haber llegado a un consenso, pero muchas personalidades han realizado sus propias definiciones o comentarios al respecto, y nos ayudarán a entender mejor el concepto:

En la conferencia de Techonomy de California de 2010, Eric Schmidt pronunció la famosa frase (1):

Y las siguientes personalidades han declarado:

"Sin análisis de Big Data las corporaciones son ciegas y sordas. Perdidos en la web como un ciervo en la carretera"

- Geoffrey Moore

"El mundo es un problema de Big Data"

- Andrew Mc Afee. Refiriéndose a que el mundo es como un enorme algoritmo que solo el Big Data es capaz de hacer visible.

"Sin Big Data solo eres otra persona con su opinión"

- W. Edwards

(1) "Hubo 5 exabytes de información creada por el mundo entero entre los albores de la civilización y 2003. Ahora esa misma cantidad se crea en dos días"

“La información es el petróleo del S. XXI, y el Big Data es el motor de combustión”

- Peter Sondergaard

Con el crecimiento exponencial del volumen de datos disponibles, los métodos utilizados para el almacenamiento y procesamiento de datos han quedado obsoletos, siendo sustituidos por otras herramientas y tecnologías, que permiten manejar y procesar grandes cantidades de datos de manera eficiente.

El Big Data se utiliza en una amplia variedad de campos, como la ciencia de datos, la investigación empresarial, la publicidad en línea, la salud y el gobierno, entre otros. Algunas de las características comunes del Big Data son:

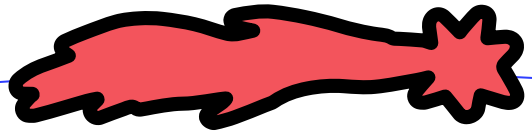
VOLUMEN: el Big Data se caracteriza por tener un gran volumen de datos, que pueden ser de diversos tipos (estructurados, semi-estructurados o no estructurados) y proceder de diferentes fuentes (transacciones, sensores, redes sociales, etc.).

VELOCIDAD: el Big Data se genera y se recopila a una velocidad muy alta, por lo que es necesario contar con herramientas y tecnologías que permitan procesar y analizar los datos de manera rápida y eficiente.

VARIEDAD: el Big Data incluye una gran variedad de datos, que pueden ser de diferentes tipos y proceder de diferentes fuentes.

Estas características forman parte de las llamadas 7 V del Big Data: volumen, variedad, velocidad, veracidad, viabilidad, valor y visualización. Para trabajar con Big Data se suelen utilizar técnicas de análisis de datos como el aprendizaje automático, el procesamiento de lenguaje natural y el análisis de redes sociales, entre otras. Además, se utilizan herramientas y tecnologías de almacenamiento y procesamiento de datos como Hadoop, Spark y NoSQL, que permiten manejar y procesar grandes cantidades de datos de manera eficiente.

NACIMIENTO



A pesar de que no ha sido hasta hoy en día cuando el Big Data ha obtenido mucha importancia, es algo con lo que convivimos desde tiempos inmemoriales. A medida que ha ido avanzando la tecnología se ha convertido en uno de los mayores atractivos para aquellas instituciones que quieran conseguir o preservar el poder, y para las empresas que simplemente buscan encontrar a su público objetivo en un mercado muy diverso.

La humanidad lleva 7000 años recopilando datos de la población para controlar e investigar negocios, inicialmente se trataba solo de registros contables introducidos en Mesopotamia, pero ha habido grandes saltos tecnológicos que han cambiado esa básica recopilación de datos en un sofisticado sistema para obtener e interpretar información.

1663

En 1663, John Graunt, considerado padre de la estadística y pionero dentro de la historia del Big Data, debe su título al libro “Observaciones naturales y políticas”, en el que realizó un análisis exhaustivo de los boletines de mortalidad en Londres, con el objetivo de crear un sistema de alerta para la peste bubónica que los asolaba.

1887

El siguiente avance se produjo en 1887, con la entrada a la era de la información, gracias al invento de Herman Hollerith, una máquina tabuladora que permitía organizar los datos censales. Ese invento le llevó a crear la empresa que cambiaría en unos años su nombre a IBM.

1937

Ya en el año 1937, Franklyn D. Roosvelt decidió realizar el seguimiento de 29 millones de contribuyentes, el responsable de realizarlo sería precisamente IBM, la cual fue contratada para desarrollar una máquina lectora de tarjetas perforadas.

1943

En el año 1943, en plena segunda guerra mundial, los británicos inventaron la primera máquina de procesamiento de datos, dispositivo conocido como “Colossus”, creado para interceptar mensajes del bando nazi, que era capaz de interpretar 5.000 caracteres por segundo, haciendo que el trabajo que antes suponía semanas de esfuerzo, pasara a ser cuestión de horas

1952

Posteriormente, en 1952 se crea la NSA, que en los siguientes 10 años contrataría a 12.000 criptólogos por la sobrecarga de información que recibían debido a la guerra fría con Rusia.

1965

Pero es en 1965 cuando se puede hablar de almacenamiento de datos digitales, teniendo EEUU, en un solo espacio, almacenados 175 millones de huellas dactilares y 742 millones de declaraciones de impuestos, algo que la población estadounidense rechazó por los paralelismos con el personaje de George Orwell “Big Brother”.

1989

Poco después, en 1989, Tim Berners-Lee inventó la World Wide Web (WWW). Este científico británico consiguió facilitar el intercambio de información a un nuevo nivel, sin saber el impacto en la sociedad que eso tendría a partir de los 90, momento en el que cada vez más dispositivos se enlazaban de manera orgánica a internet. Esto, traería un desarrollo tecnológico solo visto en períodos de guerra, apareciendo en 1995 la primera supercomputadora, capaz de hacer cálculos que llevarían años a una persona, en tan solo medio segundo.

2005

Unos años más tarde, en 2005, se acuñó por primera vez el término Big Data. Fue Roger Mougalas, Director de O'Reilly Media (una importante editorial americana), y lo hizo tan solo un año después de que se hablara de Web 2.0, siendo el Big Data una consecuencia de esta. Mougalas se refería al Big Data como un gran conjunto de datos prácticamente imposibles de procesar y administrar con la tecnología de la que disponían.

Con la aparición de las redes sociales y el despegue de la web 2.0, la generación de datos pasa a ser abismal. Una gran oportunidad para pequeñas empresas y startups, pero también para los gobiernos que han emprendido proyectos relacionados con el Big Data. Un ejemplo es el gobierno indio, que en 2009 decide tomar fotografías, escanear el iris y tomar la huella digital de toda su población (1.200 M de habitantes entonces). Lo que supuso un punto de inflexión en la historia del Big Data, ya que comportó la creación de la base de datos biométrica más grande del mundo.

Para el 2022 se estimaba que solo en EEUU hacían falta unos 150.000 Data Scientists, además de unos 1,5 millones de administradores de datos.

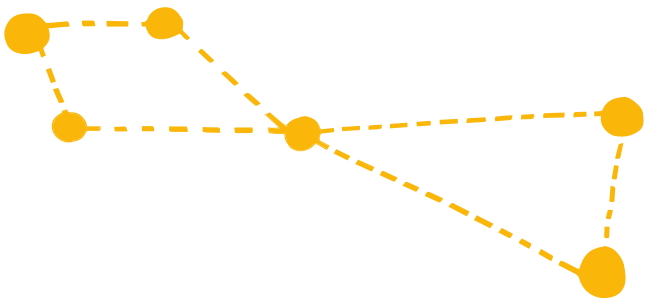
FUNCIONAMIENTO

Para poder trabajar con Big Data se utilizan diversas técnicas y herramientas de análisis de datos y tecnologías de almacenamiento y procesamiento, que permiten manipular y analizar estos grandes grupos de datos.

El Big Data es un proceso totalmente automatizado. El cual trabaja en conjunto con herramientas que buscan una solución a una serie datos que emiten información de relevancia. Se hace uso de aplicaciones analíticas, de aprendizaje, e inclusive, de inteligencia artificial. Sin embargo, es necesario conocer detenidamente cómo funciona, siempre contando con las estructuras necesarias para que sea efectiva.

Según sean las características de la empresa, se tomará en cuenta la integración del tipo de sistema a aplicar. En algunos casos los servidores a automatizar deberán ser bastantes, lo que puede resultar costoso para la misma. Así que determinar el funcionamiento correcto es imprescindible para realizar los debidos presupuestos con antelación.

Para entender el funcionamiento del Big Data, debemos explicar en qué consiste la integración, gestión, análisis de datos, así como las herramientas que se utilizan.



Integración de los datos

La integración del big data es el proceso de combinar datos de diferentes fuentes y en diferentes formatos, para obtener una visión más completa y precisa de una situación o problema en particular. Esto puede ser especialmente útil en el análisis de grandes conjuntos de datos, que pueden ser difíciles de analizar y procesar de otra manera.

La integración del big data puede incluir la recopilación de datos de diversas fuentes, como bases de datos, archivos de registro, sensores, redes sociales y dispositivos móviles. También puede involucrar la limpieza de datos para eliminar redundancias y errores, así como la transformación de datos para que puedan ser utilizados de manera más efectiva.

Una vez que se han integrado los datos, se pueden utilizar herramientas de análisis y visualización de datos para obtener insights y tomar decisiones basadas en los resultados. La integración del Big Data puede ser utilizada en una amplia variedad de campos, como la publicidad, la salud, la financiación y la industria.

El Big Data, como ya hemos mencionado, proviene de gran cantidad de fuentes y el volumen de datos es considerable. Por ende, es necesario el uso de aplicaciones y herramientas tecnológicas que permitan gestionar tales cantidades. Al generar tanta información, es imprescindible recibir los datos, lograr procesarlos y formatear adecuadamente, con el fin de que puedan llegar a ser comprendidos por los profesionales y usuarios.

Gestión Big Data

La gestión eficaz de Big Data es un aspecto importante de los negocios modernos, ya que puede proporcionar información valiosa y permitir que las organizaciones tomen decisiones más inteligentes. Hay varios aspectos clave a considerar cuando se trata de la gestión de Big Data:

RECOPILACIÓN DE DATOS: Juntar y almacenar datos de una variedad de fuentes, incluidas las redes sociales, sensores, bases de datos transaccionales y más.

ALMACENAMIENTO DE DATOS: Big Data requiere sistemas de almacenamiento especializados, como bases de datos Hadoop o NoSQL, para almacenar y procesar grandes volúmenes de datos de manera eficiente.

Procesamiento de datos

Organizar y analizar datos para extraer información valiosa y hacerlos más accesibles para los usuarios comerciales.

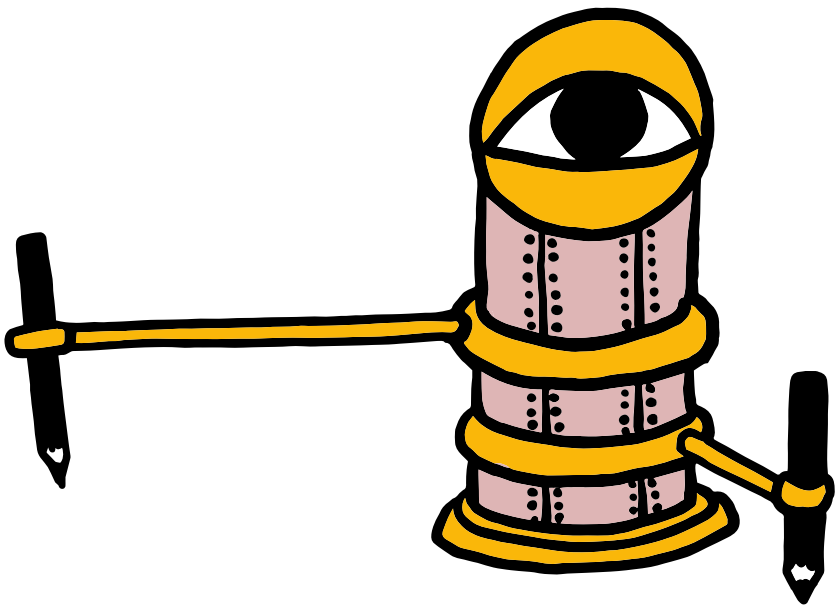
Visualización de datos

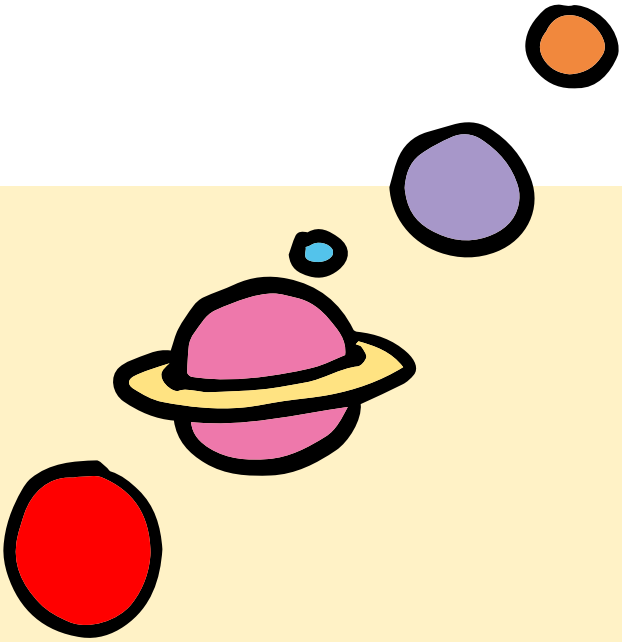
Presentar datos en un formato visual, como cuadros o gráficos, para que sea más fácil de entender e interpretar para las personas.

Seguridad de los datos

Garantizar la seguridad y la privacidad de los grandes datos es fundamental, ya que a menudo contienen información confidencial. Esto incluye la implementación de medidas y protocolos de seguridad apropiados para proteger contra violaciones de datos.

En general, la gestión eficaz de Big Data requiere una combinación de experiencia técnica, toma de decisiones basada en datos y medidas de seguridad sólidas.





Análisis de los datos:

Al analizar todos los datos que han sido almacenados, se determinan las respuestas que estos dan. Tanto aquellas relacionadas con las búsquedas que hacen los clientes, como las necesidades que estos requieren. Lo ideal es sacar el máximo provecho a estos datos al ingresar recursos para el análisis de estos, tanto en infraestructura como en profesionales. Es por esto que el uso de la información es indispensable.

Hay varias formas de analizar Big Data, según los objetivos y necesidades específicas del análisis. Algunos métodos comunes incluyen:

VISUALIZACIÓN DE DATOS: Creación de tablas, gráficos y otras representaciones visuales de los datos, para ayudar a identificar patrones y tendencias.

ANÁLISIS ESTADÍSTICO: Uso de técnicas estadísticas para analizar los datos e identificar patrones y tendencias.

APRENDIZAJE AUTOMÁTICO: Uso de algoritmos y modelos para analizar los datos y hacer predicciones o identificar patrones.

MINERÍA DE TEXTO: Analizar grandes cantidades de datos de texto para identificar tendencias y opiniones.

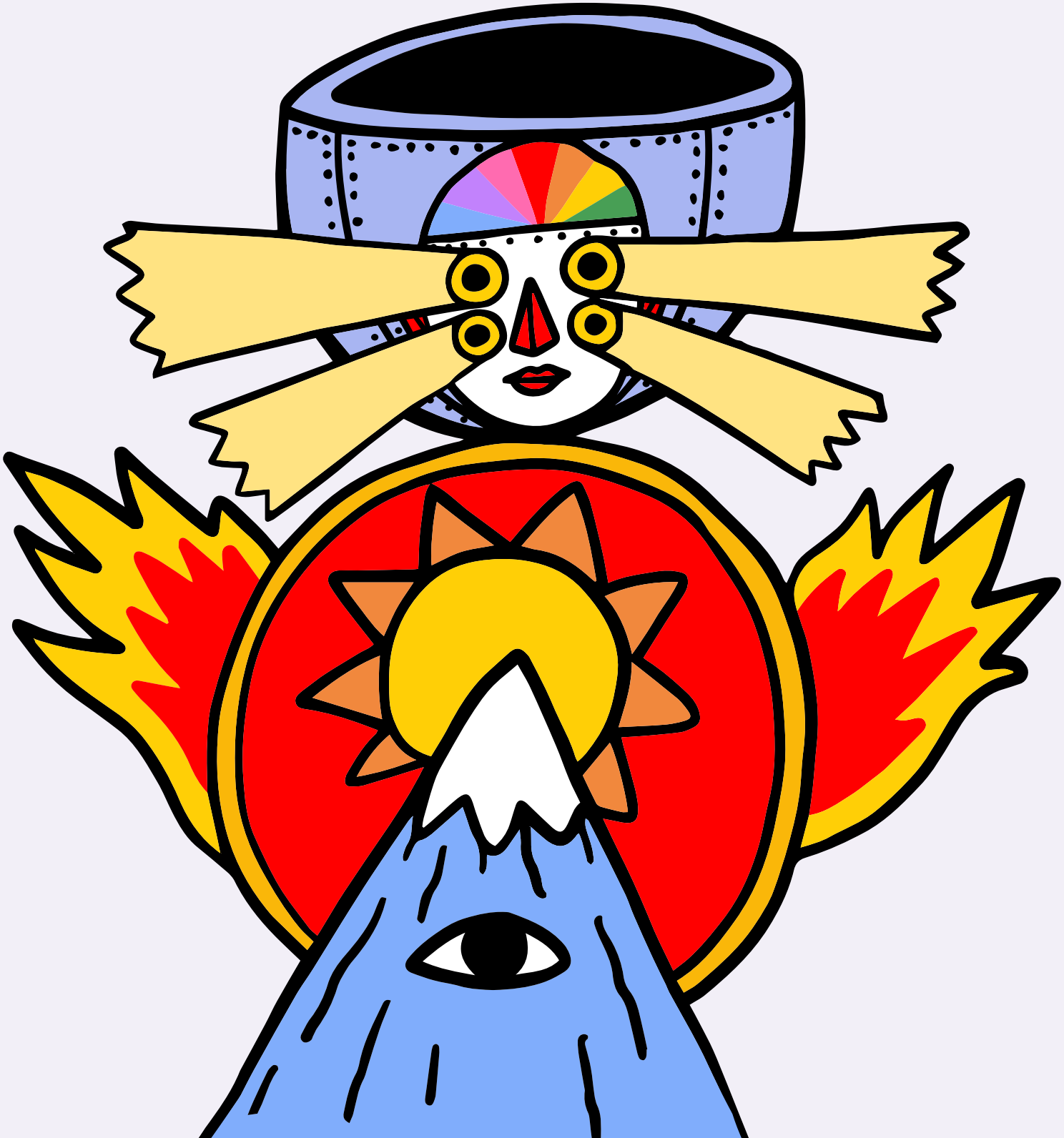
ANÁLISIS DE RED:Analizar datos de redes, como redes sociales o redes de transporte, para identificar patrones y conexiones.

En general, el objetivo de analizar Big Data es extraer información y conocimientos útiles que puedan colaborar con la toma de decisiones, mejorar los procesos e impulsar la innovación.

Si te intriga cada vez más el mundo del Big Data, te interesarán también estas herramientas de análisis:

Hadoop	Motor de procesamiento de datos de código abierto para usar a gran escala.
Spark	Motor de procesamiento de datos de código abierto para el procesamiento de datos a gran escala.
Bases de datos NoSQL	Bases de datos diseñadas para manejar grandes cantidades de datos no estructurados, como MongoDB y Cassandra.
Herramientas de visualización de datos	Herramientas como Tableau y QlikView que le permiten crear visualizaciones de sus datos para ayudar a identificar patrones y tendencias.
Plataformas de aprendizaje automático	Herramientas como TensorFlow y Scikit-learn que le permiten crear y entrenar modelos de aprendizaje automático en grandes conjuntos de datos.
Herramientas de minería de texto	Herramientas como NLTK y GATE que le permiten analizar y extraer información de grandes cantidades de datos de texto.
Herramientas de análisis de red	Herramientas como Gephi y NodeXL que le permiten analizar datos de redes e identificar patrones y conexiones.

En general, la elección de la herramienta dependerá de las necesidades específicas de análisis y del tipo de datos con los que se esté trabajando.



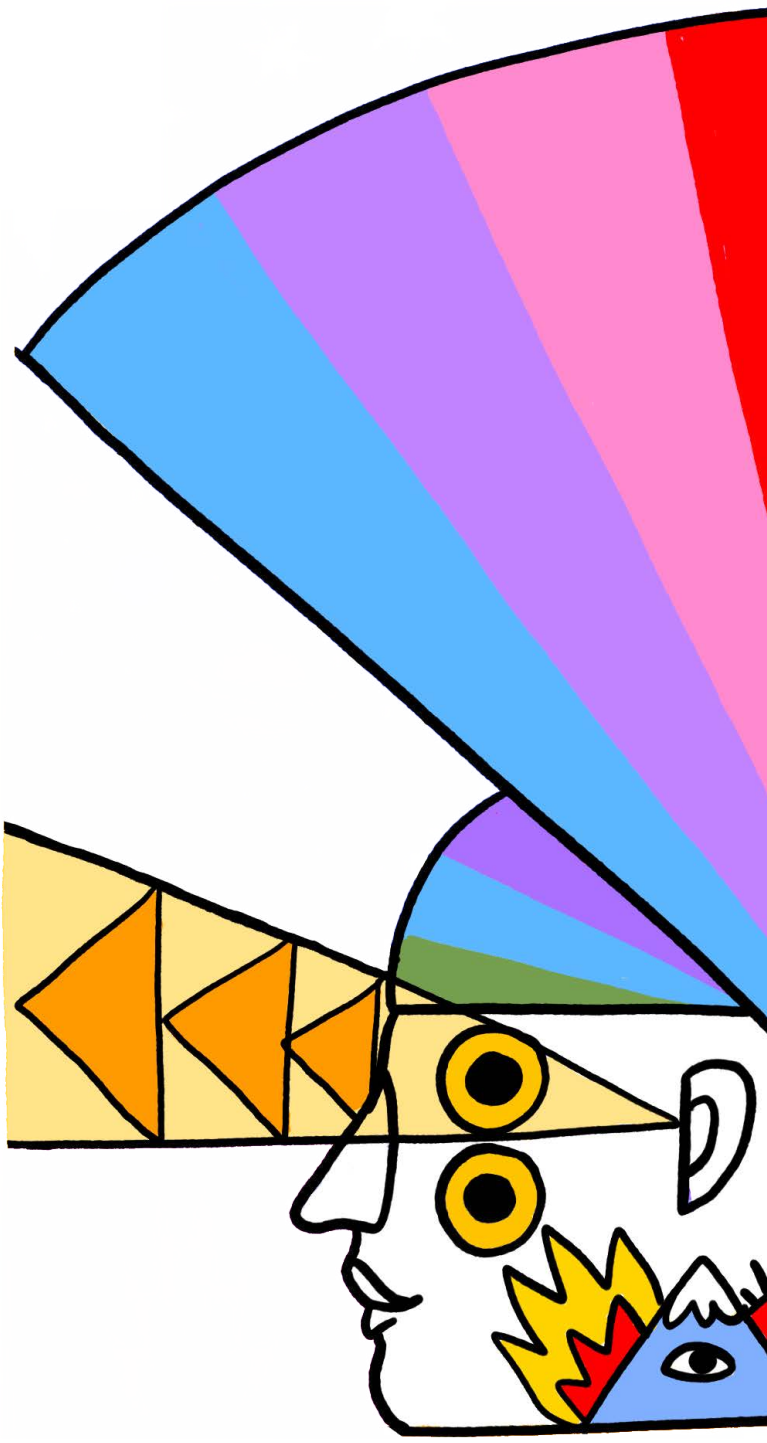
BENEFICIOS EMPRESARIALES

El uso de Big Data en las empresas comporta un conjunto de beneficios para estas, entre los cuales cabe destacar:

- **Mejora la toma de decisiones y el diseño de estrategias** en una empresa, a partir del análisis de grandes cantidades de datos que proporcionan información valiosa.
- **Incrementa la eficiencia**, ayudando a identificar áreas de una compañía, donde se pueden hacer ahorros de tiempo y dinero a través del análisis de patrones y tendencias en los datos.
- **Personaliza el servicio al cliente**, colaborando con las empresas para conocerlos mejor y ofrecerles productos y servicios de alta calidad, acorde a lo que necesitan.
- **Innova y le da herramientas a las empresas**, para identificar nuevas oportunidades de negocio y comprensión del mercado, que impulsen el éxito a largo plazo.
- **Incrementa la seguridad** a la hora de identificar y prevenir posibles riesgos y amenazas sobre la información.

En el caso de Dragon Corp Games, el Big Data sirve sobre todo para identificar patrones en el mercado de los videojuegos, entender de dónde provienen sus fuentes de ingresos, segmentar por países y edades a los usuarios de un determinado juego, y comprobar si la información que deciden proporcionar es veraz y encaja con nuestros datos. Eso nos ha permitido mejorar mucho la toma de decisiones, ya que somos capaces de detectar oportunidades de negocio y disponer de un equipo que comprende el mercado.

DESAFÍOS SOCIALES



El Big Data puede ser útil para mejorar la toma de decisiones y para entender mejor ciertos fenómenos y tendencias. Sin embargo, también plantea algunos desafíos sociales y éticos:

PRIVACIDAD: La recopilación y el análisis de grandes cantidades de datos puede implicar el riesgo de violación de la privacidad de las personas. Es importante garantizar que se respeten las leyes y regulaciones en materia de privacidad y que se adopten medidas de seguridad adecuadas para proteger los datos personales.

DISCRIMINACIÓN: El análisis de datos puede utilizarse para discriminar de manera sutil a ciertos grupos de personas, por ejemplo, en el ámbito laboral o en el acceso a servicios. Es importante asegurarse de que el análisis de datos no se utilice de manera injusta o discriminatoria.

DESIGUALDAD: El acceso a los datos y a las herramientas para analizarlos puede generar o perpetuar desigualdades. Es importante asegurarse de que todas las personas tengan acceso a la información y a las herramientas necesarias para beneficiarse del análisis de datos.

FALTA DE TRANSPARENCIA: A veces, el análisis de datos se realiza de manera opaca y es difícil comprender cómo se están tomando las decisiones basadas en esos datos. Es importante garantizar la transparencia y la claridad en el proceso de análisis de datos.

RESPONSABILIDAD: Es importante establecer quién es responsable de los posibles errores o consecuencias indeseables del análisis de datos.

En resumen, es importante abordar estos desafíos de manera proactiva y garantizar que el análisis de datos se utilice de manera ética y responsable, para el beneficio de todos.

BUENAS Y MALAS PRÁCTICAS

Cuando se está en frente de algo disruptivo, es difícil definir cuáles de los efectos que esto genera son buenos y cuáles malos. La disrupción conlleva cambios y suelen afectar de manera negativa e injusta a mucha gente, a pesar de que a largo plazo puedan ser positivos para el conjunto global de la población.

Hay ciertos eventos e hitos que nos hacen imaginar lo que puede acabar aconteciendo:

PREDICCIÓN LOTERÍA: A pesar de que a día de hoy no podemos saber el número de la lotería que será premiado, gracias al uso de Big Data se puede determinar cuáles son los números con más probabilidades de salir premiados.

*Contra: Aquellos que disponen de la información tendrán cada vez más poder sobre aquellos que no la tienen.

CHAT GPT: Hasta ahora hemos buscado la información desde fuentes como YouTube o Google, que nos indexan contenidos para ayudarnos a encontrar lo que buscamos. Aunque por otro lado, herramientas como Alexa y Siri, han comenzado a responder preguntas más concretas y a tener incluso personalidad. El Chat GPT es la evolución de estos últimos. Su inteligencia es abrumadora, hasta el nivel de poder escribir un poema, hacer los deberes, o detectar y corregir los errores de código de un programador.

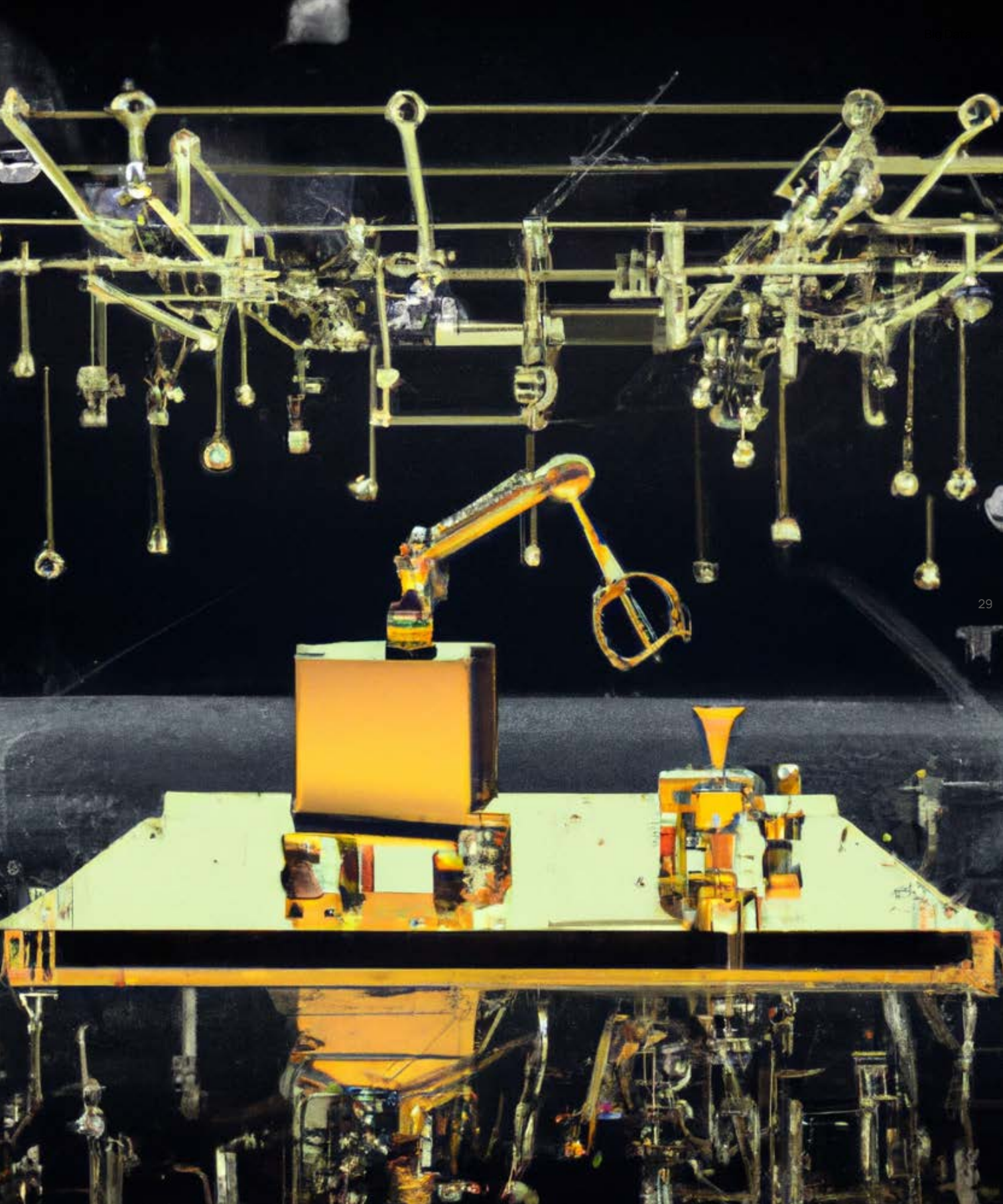
*Contra: Hasta hace poco se creía que los trabajos de transporte serían prácticamente los únicos afectados por la IA a corto plazo, pero ahora vemos que trabajos como el de programador, escritor, dibujante y un sin fin de ejemplos más, están bajo la necesidad de transformarse para perdurar.

SANIDAD: Una de las mayores aplicaciones a día de hoy del Big Data se da en la sanidad. Cruzando datos del historial de los pacientes y sus características físicas, estamos llegando a una sanidad personalizada. Hasta ahora siempre se ha tratado a la salud como unos mantras aplicables al conjunto de la población. No fumar, hacer ejercicio, dormir 8 horas, etc. Pero ahora sabemos que hay gente que necesita dormir solo 6 horas para estar bien, que si sale a correr se lesionará, que tiene una esperanza de vida X, entre otros datos determinados por un conjunto inmenso de información.

*Contra: Si actualmente ya muchos rechazan a las personas por lo que piensan, imaginemos cómo se puede llegar a discriminar en el futuro, en función de parámetros biométricos.

ASEGURADORAS: Las ciencias actuariales se ocupan de las repercusiones financieras de riesgo e incertidumbre. Los actuarios proporcionan evaluaciones de sistemas de garantía financiera, con especial atención a su complejidad, sus matemáticas y sus mecanismos. Y la buena noticia para ellos es que ahora es más fácil y certero determinar las posibilidades de fraude de un asegurado, su posible siniestralidad, etc. Beneficiando a los asegurados que no dan problemas y perjudicando a los posibles estafadores.

*Contra: Habrá consumidores que pagarán más por ser prejuzgados por motivos discriminatorios, incluso teniendo un historial de tráfico sin antecedentes.



FUTURO

El Big Data ha tenido un gran impacto en diversos campos y ha revolucionado la forma en que se recopilan, almacenan y analizan los datos. En el futuro, se espera que el Big Data continúe siendo una herramienta valiosa para mejorar la toma de decisiones y entender mejor ciertos fenómenos y tendencias. El Big Data será integrado dentro de nuestra vida cotidiana, de forma que cada actividad que llevemos a cabo podrá ser contabilizada dentro de un sistema, que manteniendo la privacidad de los usuarios, permita llevar un registro de actividades o preferencias; y que traslade esta información a las instituciones y/o compañías, para la mejora de productos o servicios, beneficiando siempre a la mayoría de los compradores o usuarios.

Algunas de las tendencias futuras del Big Data son:

MAYOR ÉNFASIS EN LA CALIDAD DE LOS DATOS: A medida que se recopilan y analizan más datos, es importante asegurar que estos datos sean precisos y relevantes. Esto puede requerir un mayor énfasis en la limpieza y el procesamiento de los datos para eliminar errores y redundancias.

MAYOR USO DE TECNOLOGÍAS DE ANÁLISIS DE DATOS: Se espera que el uso de tecnologías como la inteligencia artificial, el aprendizaje automático y el análisis de datos en tiempo real, continúen creciendo en el futuro. Esto puede permitir un análisis más rápido y preciso de los datos y ayudar a tomar decisiones más efectivas.

MAYOR IMPORTANCIA EN LA PRIVACIDAD Y LA SEGURIDAD DE LOS DATOS: A medida que se recopilan y analizan más datos, es importante garantizar que se respeten las leyes y regulaciones en materia de privacidad y se adopten medidas de seguridad adecuadas para proteger los datos personales.

MAYOR USO DE DATOS EN LA TOMA DE DECISIONES EMPRESARIALES: Se espera que el análisis de datos se convierta en una parte cada vez más importante en la toma de decisiones empresariales, permitiendo a las empresas tomar decisiones más informadas y adaptarse mejor a los cambios en el mercado.

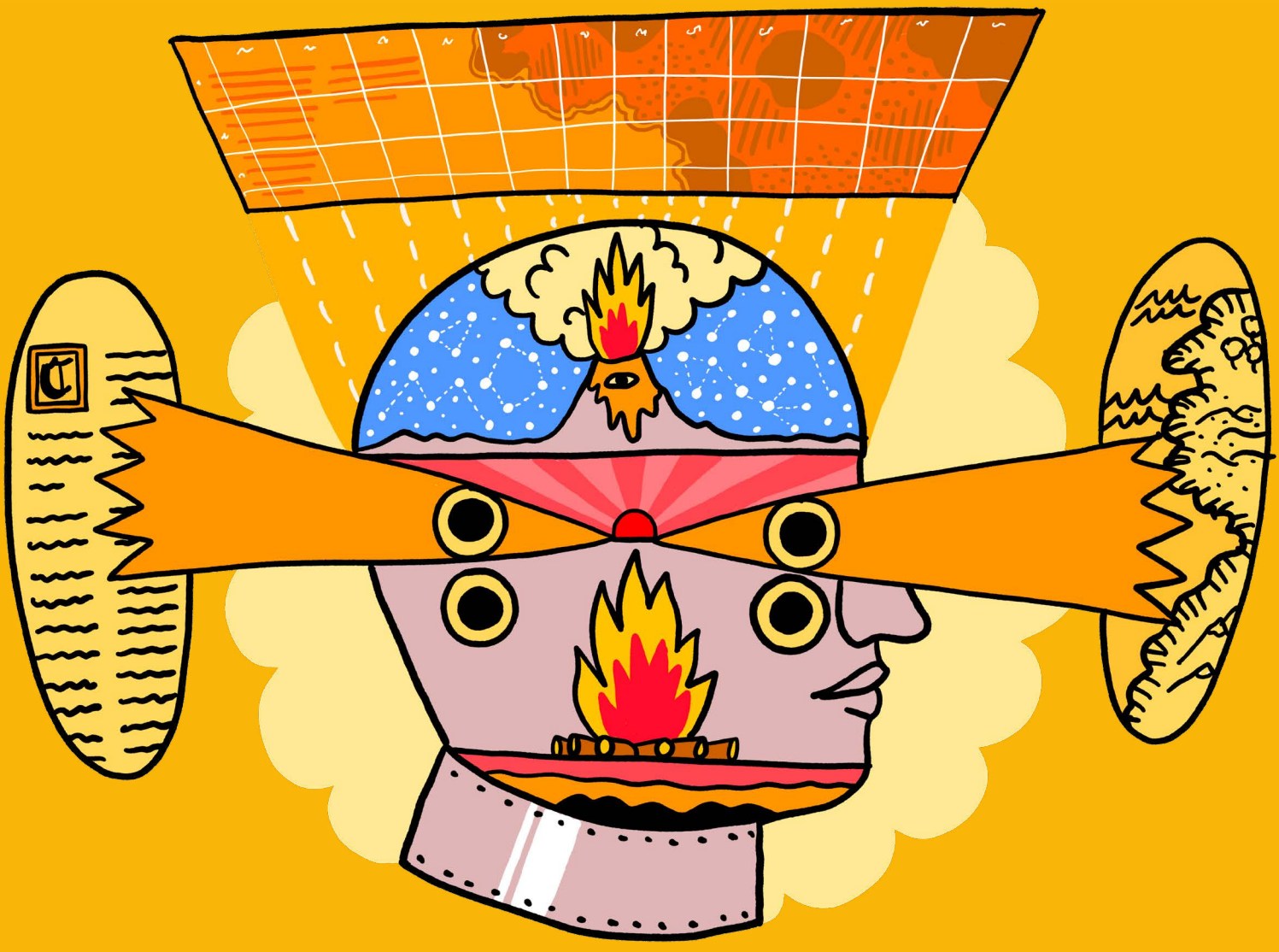
Sin duda alguna, el futuro del Big Data es muy prometedor, al punto de que todas las universidades están adaptando el pénsum o incluyendo alguna materia al respecto. Aunque la necesidad sigue en crecimiento y la velocidad con que se cubre la demanda no va al mismo ritmo, posiblemente sea el momento más idóneo de aprender y especializarse en el tema.

(II) Inteligencia **Artificial**

Seguro que has oído hablar de computadoras que juegan al ajedrez, conducen y producen textos e imágenes. La IA permite que las máquinas realicen tareas y aprendan de la experiencia, como si fuesen seres humanos.

Pero, ¿sabes de qué más son capaces?

POR CARLOS PÉREZ



¿QUÉ ES?

La Inteligencia Artificial (IA) es la simulación de la inteligencia humana mediante máquinas y sistemas computacionales. El término fue acuñado en 1956 por John McCarthy, que la definió como "la ciencia y la ingeniería de hacer máquinas inteligentes". La IA nos puede sonar lejana y misteriosa, pero está presente en nuestro día a día desde hace muy poco tiempo, y la pregunta que muchos se hacen es ¿hasta dónde es capaz de llegar?

Prácticamente todas las industrias están incorporando algún componente de IA, ya que les ayuda a apoyar y justificar sus decisiones, a transformarse digitalmente, o a sacar partido a los datos. Pues si la data es el petróleo del siglo XXI, la inteligencia artificial es quien refina estos datos y los convierte en productos muy valiosos.

La IA aún está en época de crecimiento y necesita madurar, y aunque en algunos sectores parece haberse consolidado, en otros está apenas apareciendo.

TIPOS DE IA

A continuación te contaremos los distintos tipos de Inteligencia Artificial, según su funcionalidad:

Máquinas reactivas

Este tipo de inteligencia se caracteriza porque no tiene capacidad de formar recuerdos y no puede utilizar experiencias pasadas de las que ayudarse para tomar decisiones.

Deep Blue, una supercomputadora creada por IBM, fue capaz de vencer en ajedrez al gran maestro ruso Gary Kasparov a finales de la década de los 90. Pudiendo identificar las piezas del tablero y conocer los movimientos de cada una, para escoger las mejores futuras posibilidades que tenía.

Las máquinas reactivas son capaces de simular millones de combinaciones a una velocidad alucinante.

Memoria limitada

Estas máquinas aprenden lo sucedido en el pasado para adivinar el futuro siendo las que más popularidad han tenido en los últimos años.

Su IA puede almacenar patrones de información de datos del pasado (aunque de manera limitada y temporal) y dar respuesta a eventos para un futuro no lejano.

El ejemplo más claro son los coches autónomos, que utilizan datos recogidos (imágenes y vídeos) en los instantes previos, para tomar decisiones.

Teoría de la mente

¿Has visto películas como Una Odisea en el Espacio (1968), Blade Runner (1982) o Ex-Machina (2014)? La industria del cine nos ha mostrado cómo podría ser el futuro con esta IA.

Según esta teoría, los sistemas serán capaces de entender que las personas tienen emociones, pensamientos y sensaciones que afectan la toma de decisiones. Y serán capaces de ajustar su comportamiento para poder convivir con nosotros, pudiendo razonar y entender patrones de conducta. Alucinante, ¿cierto?

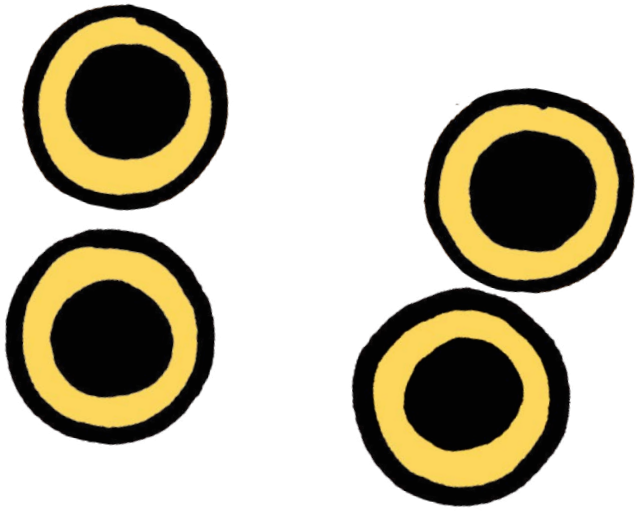
Autoconsciencia

En esta última fase la IA tiene su propia consciencia y percibe sus emociones, pensamientos y sensaciones, y las de los demás. Aún queda tiempo para que se haga realidad, pero no descartamos poder verla más temprano de lo que esperamos, en algún formato.

RAMAS

Por su parte, la Inteligencia Artificial abarca otros dominios como el Machine Learning o el Deep Learning. Estas son las principales ramas:

- MACHINE LEARNING
  - Aprendizaje supervisado
    - Clasificación
    - Regresión
    - Ranking
  - Aprendizaje no supervisado
    - Clustering
    - Reducir la dimensionalidad
- REINFORCEMENT LEARNING
- DEEP LEARNING
  - Natural Language Processing (NLP)
  - Computer Vision (CV)
- ROBÓTICA



Machine Learning

El Machine Learning (ML) se sirve de algoritmos para identificar patrones en los datos, que permitan elaborar predicciones e inferencias. Existen dos categorías importantes:

1. Aprendizaje supervisado

Estos algoritmos trabajan con datos etiquetados. Su propósito es encontrar relaciones y funciones que asocian los datos de entrada con los de salida. El algoritmo se entrena con un histórico de datos y aprende a asignar la etiqueta de salida. El aprendizaje supervisado suele usarse en:

PROBLEMAS DE CLASIFICACIÓN: La variable a predecir es una variable discreta. Por ejemplo, detectar qué email es spam o no, o la detección de fraude de identidad.

PROBLEMAS DE REGRESIÓN: La variable objetivo es una variable continua. Por ejemplo, predecir el precio del Bitcoin o el precio de venta de un inmueble.

PROBLEMAS DE RANKING: Este último problema es una combinación de clasificación y regresión. Por ejemplo, ordenar los productos a recomendar en Amazon en función de las preferencias y compras pasadas, ordenar los sitios publicados en Google tras una búsqueda, de la manera correcta.

2. Aprendizaje no supervisado

En este tipo de aprendizaje no se dispone de datos etiquetados. El objetivo es encontrar algún tipo de organización que simplifique el análisis e interpretación de los datos de entrada. Sus funciones principales son:

AGRUPAR LOS DATOS EN GRUPOS CON CARACTERÍSTICAS SIMILARES (CLUSTERING): Por ejemplo, identificar qué clientes tienen un comportamiento similar para poder impactar con el mismo mensaje.

REDUCIR LA DIMENSIONALIDAD: Disminuir el número de variables que definen los datos de entrada, de manera que eliminamos las variables irrelevantes, reduciendo así la complejidad de futuros modelos de predicción y el rendimiento computacional. Pues, más datos no siempre es mejor (pese a que digan lo contrario).

Nota: Si este tema realmente te interesa, no te preocupes, que en el capítulo III - Machine Learning, profundizaremos con más información.

Reinforcement learning

Este tipo de aprendizaje se basa en mejorar la toma de decisiones de un agente en un ambiente, maximizando el número de aciertos o recompensas acumuladas en el tiempo. El modelo aprende a base de castigos y premios, es decir el sistema aprende a base de ensayo-error. Este tipo de aprendizaje no necesita de datos etiquetados.

Deep Learning

Esta rama utiliza estructuras lógicas que se asemejan al sistema nervioso de los seres humanos, simulando su componente principal, las neuronas. El DL es especialista en detectar características existentes en los objetos o datos percibidos, tal y como lo haría la mente humana. Los campos donde el Deep Learning ha tenido más éxito son:

NATURAL LANGUAGE PROCESSING (NLP): Las redes neuronales (Neural Networks en inglés) entienden el lenguaje humano, tienen en cuenta el contexto, saben construir frases o responder a nuestras preguntas.

COMPUTER VISION (CV). Las redes neuronales identifican patrones y señales en imágenes, lo que permite distinguir un perro de un gato.

Robótica

Esta rama estudia el diseño y construcción de máquinas capaces de desempeñar tareas que realiza el ser humano y que requieran cierto uso de inteligencia. Cada día vemos más procesos automatizados gracias a la robótica.

Los recientes avances y popularidad de la Inteligencia Artificial se deben principalmente a la evolución de los ordenadores y microprocesadores. Como dicta la Ley de Moore, aproximadamente cada 2 años se duplica el número de transistores en un microprocesador. Esto ha permitido poder realizar mayor cantidad de cálculos y cálculos de mayores dimensiones a un menor coste.

Hace ya 40 años, por allá en los ochenta, las compañías más valiosas estaban principalmente en el sector del petróleo y el gas, un par de ellas en informática, telecomunicaciones e incluso alguna fotografía. En el 2020, la lista es encabezada por una empresa del sector petrolífero, pero la siguen empresas tecnológicas estadounidenses como Apple, Microsoft, Amazon, Google, Facebook o la china Alibaba. Empresas que han impulsado y han crecido gracias a la inteligencia artificial y el tratamiento de los datos.

NACIMIENTO

1950

La primera noción de la Inteligencia Artificial surge en los años 50, poco después de la Segunda Guerra Mundial. El matemático Alan Turing, quien con su trabajo acortó la duración de la guerra entre dos y cuatro años tras desencriptar los mensajes cifrados de los nazis alemanes, redactó el conocido Test de Turing en su ensayo "Computing Machinery and Intelligence".

Este test evalúa la capacidad de una máquina para exhibir un comportamiento similar al de un ser humano. La máquina pasa exitosamente el test, luego de que una persona (el entrevistador), tras una conversación con una máquina y otra conversación con una persona, no sepa distinguir cuál fue con el humano y cuál con la máquina.

Cinco años después del ensayo de Turing, en Dartmouth (Estados Unidos) se presentó el primer programa de Inteligencia Artificial organizado por John McCarthy y Marvin Minsky en 1956. En la conferencia de Dartmouth se reunieron los principales investigadores de varias disciplinas para discutir sobre la Inteligencia Artificial. Fue aquí donde se acuñó por primera vez el término Inteligencia Artificial.

1960

Con el desarrollo de los ordenadores, surgió la capacidad de almacenar información y ejecutar procesos más rápido y barato; convirtiéndose con el tiempo en un producto más accesible para el público general.

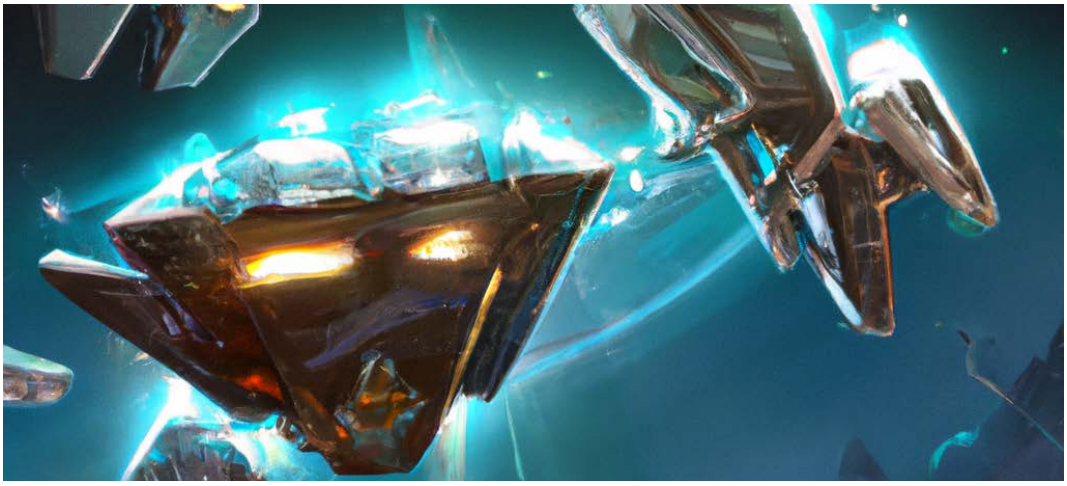
¿Pero cómo se puede avanzar a pasos tan agigantados? Te lo contamos:

Entre 1964 y 1966 en el Massachusetts Institute of Technology (MIT) se desarrolló ELIZA, un programa capaz de procesar el lenguaje natural e interactuar vía texto.

ELIZA era un proyecto embrionario aunque prometedor. Tanto es así que el gobierno de Estados Unidos fundó la Agencia de Proyectos de Investigación Avanzados de Defensa (Defense Advanced Research Projects Agency en inglés, DARPA) para promocionar y fomentar investigaciones en este nuevo mundo. En 1970 Marvin Minsky dijo en la revista Life: "De aquí a 3 u 8 años, tendremos una máquina con la inteligencia media de un ser humano". Sin embargo, la capacidad de los ordenadores aún estaba a años luz de poder exhibir tal inteligencia, y durante los siguientes años la IA dejó de estar en boca de todos. El hype se desvaneció y la gente perdió el entusiasmo por esta nueva tecnología.

1980

Hay que esperar a los inicios de los ochenta, cuando la IA vuelve a renacer gracias a las aportaciones de Edward Feigenbaum. Edward se doctoró en la Universidad de Carnegie Mellon, y en esta misma universidad en 1979 otro profesor creó un programa informático que consiguió vencer al campeón de Backgammon, e introdujo el concepto "Sistemas Expertos" que imita el proceso de decisión de los humanos más expertos. El programa preguntaba a un experto cómo responder ante ciertas situaciones y el sistema era capaz de memorizar y responder a peticiones de personas no tan expertas.



Los Sistemas Expertos fueron utilizados ampliamente en varias industrias, tanto es así que en Japón, los nipones decidieron hacer una fuerte inversión en revolucionar el procesamiento por computador, la lógica programática y mejorar la Inteligencia Artificial. Aunque sus objetivos fueron demasiado ambiciosos y no llegaron a cumplirse, estos esfuerzos inspiraron a muchos ingenieros y científicos jóvenes.

1990

Los nuevos intentos frustrados hicieron que la IA deje de estar, de nuevo, en el foco de atención. Los gobiernos dejaron de invertir en ella y el hype del público se desvaneció. Irónicamente, fue en la década de los 90 y los 2000 cuando más prosperó. En 1997, el gran maestro mundial de ajedrez Gary Kasparov fue derrotado por Deep Blue, un programa de ordenador desarrollado por IBM. La primera partida en 1996, la perdió la máquina, pero una versión mejorada venció al ajedrecista ruso el siguiente año.

En los noventa destacan las evoluciones en el Deep Learning y la consolidación de las redes neuronales. En 1997, Sepp Hochreiter y Jürgen Schmidhuber desarrollan Long Short-Term Memory (LSTM), un tipo de red neuronal recurrente (Recurrent Neural Network en inglés - RNN). Esta arquitectura se volverá muy popular para el reconocimiento del texto y la voz.

Si recordamos a ELIZA, aquel programa con el que se podía hablar con tu ordenador a través de una pantalla y un teclado, en 1996 Richard Wallace desarrolló un chatbot ALICE (Artificial Linguistic Internet Computer Entity). ALICE mejoraba a ELIZA porque añadía un gran corpus de texto (datos) de los que aprender y replicar patrones, tenía una estructura más sencilla y sabía satisfacer limitaciones que ELIZA tenía.

2000

La década de los 2000 destaca por el boom de la robótica y su aparición en el mundo del cine:

En el 2000, Honda lanza ASIMO, un robot humanoide inteligente. La compañía japonesa creó este robot para ayudar a personas que carecen de movilidad completa en sus cuerpos.

En 2004, Will Smith protagonizó la película Yo, Robot. La película se sitúa en Chicago en el año 2035, donde existen robots humanoides que sirven a las personas. Del Spooner, un detective del departamento de Policía de Chicago, investiga el caso del supuesto suicidio del co-fundador de US Robotics. Del Spooner teme que haya sido un robot humanoide el autor del crimen.

No hay que irse al 2035, para ver convivir algunos robots en nuestro día a día, pues en el 2002, iRobot lanzó Roomba. Este robot no será capaz de huir de Will Smith, pero sí de aspirar y barrer nuestra casa evitando sillas y muebles..

2010

Desde 2010 hasta hoy, la IA está en nuestro día a día. Cada día utilizamos smartphones con asistentes de voz, ordenadores y aplicaciones con funciones inteligentes con las que ahora no podríamos vivir.

En 2011 Watson, un ordenador de lenguaje natural creado por IBM, respondió preguntas y venció a 2 ganadores de Jeopardy (un concurso de televisión sobre numerosos temas como historia, lenguas, cultura popular, bellas artes, ciencia, geografía y deportes). Sin embargo, el mayor hito hasta la fecha en términos de juegos de estrategia y concursos, lo ha alcanzado AlphaGo de Google DeepMind, un programa de computadora que compite en el juego de mesa Go, que derrotó a varios campeones entre 2015 y 2017. Si el ajedrez puede parecer un juego complicado, Go es inmensamente más complejo.

En 2011, Apple lanzó Siri, un asistente virtual que utiliza lenguaje natural para entender, responder y recomendar cosas al usuario. Otras empresas en Silicon Valley no tardarían, y en 2014 Microsoft lanzaría Cortana y Amazon a Alexa. El procesamiento del lenguaje natural ha evolucionado con creces en la última década.

2018

En 2018 Google desarrolló BERT, la primera representación bidireccional del lenguaje no supervisado, que se puede usar en una variedad de tareas de lenguaje natural mediante el aprendizaje por transferencia. En 2019, OpenAI, un laboratorio de investigación impulsado por Elon Musk, lanzó GPT-2 un modelo de redes neuronales formado por 1.5 billones de parámetros que generan texto prediciendo palabra a palabra. Pero sin duda es GPT-3 el producto que más revuelos ha causado en el público. El New York Times dijo que GPT-3 no es solo asombrosa, espeluznante y aleccionadora, sino también un poco más que aterradora.

GPT-3 se entrenó con un corpus de más de 1000 millones de palabras y puede generar texto con una precisión en el nivel de los caracteres. Por el momento está en versión beta y no se vende al público, pero ha dejado a muchos expertos con la boca abierta. Tecnologías como esta permitirán grandes avances, pero también abrirán las puertas a nuevas maneras para cometer fraude, terrorismo o suplantación de identidad (deep fake).

2021

Recientemente, OpenAI también ha publicado una versión más ligera de GPT-3 con ChatGPT. Esta IA está entrenada para tener conversaciones, para cualquier tipo de pregunta, la IA te da una respuesta elaborada y acertada. Existen expertos que dicen que podría acabar con Google, reemplazando a Google como buscador si OpenAI se lo propone.

En el ámbito de los coches autónomos, en 2009 Google desarrolló secretamente un coche autónomo, y no mucho más tarde, en 2016 Tesla puso a la venta el suyo. Por el momento, el piloto automático de Tesla permite que su coche gire, acelere y frene automáticamente dentro del carril, y la capacidad de conducción autónoma total se ha diseñado para realizar viajes de corta y larga distancia sin necesitar ninguna acción por parte de la persona en el asiento del conductor. Tesla aún tiene aspectos legales que vencer y en muchos lugares del mundo, este nuevo feature no está disponible. Tesla ha invertido mucho en visión por computador para poder llevar a cabo un proyecto de esta envergadura. La visión por computador es otra disciplina donde el Deep Learning ha ayudado mucho.

Sin duda, Tesla no habría conseguido estos avances, si no fuera por el proyecto de ImageNet en 2010 y de Jeff Dean y Andrew Ng (investigadores de Google) en 2012, que entrenaron una gran red neuronal de 16,000 procesadores para reconocer imágenes de gatos al mostrarle 10 millones de imágenes sin etiquetar de videos de YouTube.

Actualmente convivimos con la IA y esto se puede observar en múltiples industrias. La IA nos está ofreciendo mucho valor en campos como el lenguaje natural, visión por computador, reconocimiento de voz, automatización de procesos, reconocimiento de imágenes, Machine Learning, redes peer-to-peer o agentes virtuales, entre muchos otros.

La historia no acabará aquí y en los próximos años continuarán apareciendo nuevos hitos que superarán lo vivido hasta la fecha. Actualmente, existe un interés en los chatbots y asistentes virtuales, y tienen mucho por avanzar en la experiencia de usuario. Con la aparición de productos como GPT-3, el procesamiento del lenguaje natural ha visto un mundo nuevo de oportunidades; los coches autónomos se consolidarán entre nosotros por los beneficios que ofrecen y tarde o temprano se impondrán a las restricciones legales; y en términos de Machine Learning, ahora se está hablando mucho del AutoML (Automated Machine Learning), y con él, los desarrolladores no tendrán que preocuparse en la creación de los modelos, pudiendo centrarse más en el valor que estos modelos ofrecen al negocio.

FUNCIONAMIENTO

Los sistemas de IA funcionan combinando grandes conjuntos de datos con algoritmos de procesamiento inteligentes e iterativos para aprender patrones y características en los datos que analizan. Cada vez que un sistema de IA ejecuta una ronda de procesamiento de datos, prueba y mide el rendimiento, y genera una experiencia adicional.

El procedimiento usual en la generación de modelos de ML sigue una secuencia de pasos: partición de los datos en entrenamiento y validación, definición de las métricas y selección del modelo.

El primer paso consiste en la partición de los datos. Los datos se dividen en tres bloques: train, test y validación. Con los datos de train, se entrena el modelo de ML o DL. Con los datos de test, se comprueba que el modelo ha aprendido los patrones adecuados en los datos y que no ha memorizado los datos de train. Y finalmente, comprobamos nuestro modelo con los datos de validación.

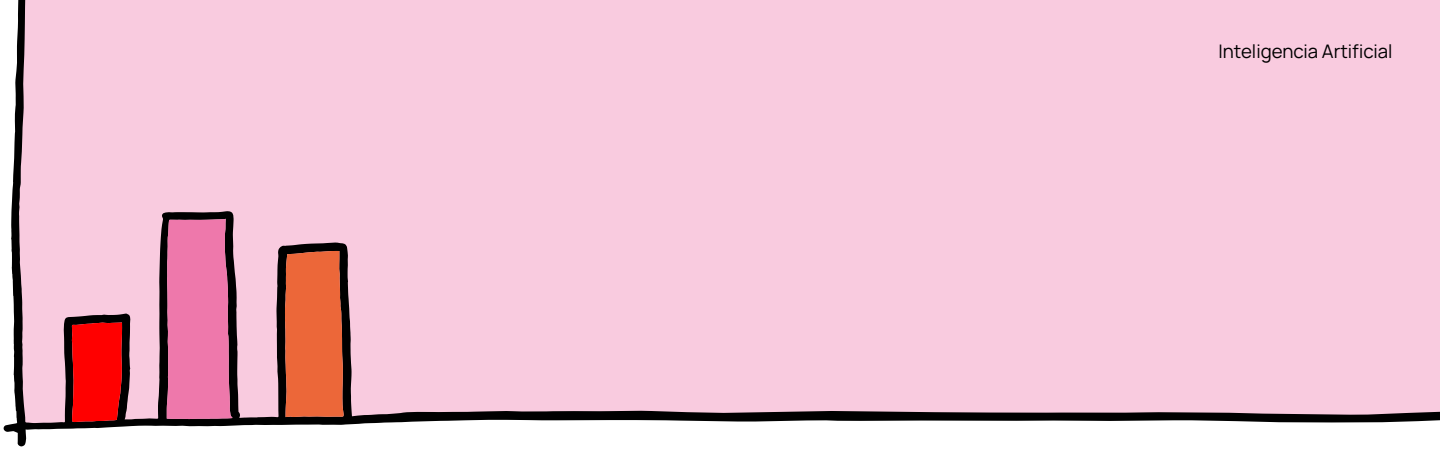
Las particiones suelen hacerse de manera aleatoria, por defecto el train-test corresponde al 90% del cual train corresponde al 80% de los datos existentes y el test al 20%, y la validación al 10%. Aunque si existe algún componente temporal en los datos, las particiones de train y test pertenecen al pasado, y los de validación corresponden a los registros más recientes de las últimas semanas o el último mes. Recordar que el objetivo principal es poner en producción un modelo que capte los patrones del mercado y tendencias actuales. Un modelo que funciona bien en el pasado pero falla en el presente, no es un buen modelo a poner en producción. ¿No lo creen?

Los principales retos a la hora de entrenar un modelo son el underfitting y el overfitting:

El **UNDERFITTING** ocurre cuando el modelo construido falla en capturar los patrones tanto en train como en test. El modelo desarrollado es demasiado simple y/o a las variables generadas les falta capacidad predictiva. Para combatir el underfitting, la solución está en optar por modelos más complejos y aumentar la cantidad de registros, de modo que el modelo pueda estar expuesto a más casos y/o añadir más y mejores variables.

El **OVERFITTING** ocurre cuando el modelo ha memorizado los datos de entrenamiento y falla en las predicciones en test. Su rendimiento en entrenamiento es excelente, pero cuando se evalúa en test, los resultados son muy inferiores. Para combatir el overfitting, se puede reducir la complejidad del modelo utilizado (reduciendo el número de hiperparámetros, seleccionado así un modelo más simple), aumentar la cantidad de registros, reducir el número de variables que describen los datos o utilizar la validación cruzada. La cross validation divide los registros en varias piezas, cada pieza tiene registros seleccionados de manera aleatoria. El modelo será entrenado y evaluado en cada una de estas piezas, ya que de esta manera el modelo generaliza mejor y se evita el sobreentrenamiento.

El underfitting se produce cuando se tiene mucho sesgo y poca varianza. Y el overfitting cuando se tiene poco sesgo y mucha varianza. Existe una batalla entre el sesgo (bias) y la varianza, y el objetivo es encontrar un balance entre ambos.



El segundo paso es la elección de la métrica y la función de coste. Los modelos se entrenan iterativamente con el objetivo de reducir el error que se computa en una variable generalizada que viene definida por la función de coste.

La métrica depende del tipo de problema a resolver:

CLASIFICACIÓN: Podemos querer maximizar la exactitud (accuracy), la precisión (precision), la exhaustividad (recall) o el valor-F (F1-score) que es un balance entre precision y recall.

REGRESIÓN: La métrica más popular es el error cuadrático medio (RMSE). Pero en otras ocasiones, dependiendo del tipo de problema o de la distribución de los datos, nos interesa minimizar el error absoluto medio (MAE) o alguna otra expresión de error ponderada.

RANKING: La métrica más popular es el Discounted Cumulative Gain (DCG). La lógica detrás de esta métrica es premiar la aparición de recomendaciones en los primeros puestos y penalizar la aparición de recomendaciones irrelevantes.

Tras seleccionar la métrica más adecuada, el siguiente paso es escoger el modelo más adecuado. Los modelos podrían clasificarse en 3 grandes grupos:

1. MODELOS GEOMÉTRICOS: Este tipo de modelos encuentra combinaciones entre las variables de entrada que permiten obtener el valor de salida. Los modelos más populares son Support Vector Machine (SVM), Logistic Regression, Linear Regression o K-Nearest Neighbors (KNN). Estos modelos son simples, fáciles de entrenar y almacenar, pero fallan al capturar relaciones no lineales y necesitan que los datos vengan normalizados.

2. MODELOS A BASE DE ÁRBOLES DE DECISIÓN: Este tipo de modelos basan su estructura en los árboles de decisión. Un árbol de decisión define una serie de consultas o pruebas con respuestas de sí o no, que se realizan de manera adaptativa. Los resultados de estas consultas permiten inferir un resultado.
Los modelos más populares son combinaciones de varios árboles de decisión. Destacan Random Forest, y los modelos de Gradient Boosting (CatBoost, LightGBM o XGBoost). Estos últimos modelos se volvieron populares en las competiciones de Kaggle por su gran rendimiento.

3. REDES NEURONALES: Este último tipo de modelos emulan el funcionamiento de las neuronas. Su arquitectura consiste en nodos de entrada y salida. Entre los nodos de entrada y salida, existen unas capas ocultas con otros nodos en los que se producen sumas y multiplicaciones. Las redes neuronales son el tipo de modelo más complejo, requieren el uso de GPU (si se desea que el entrenamiento se agilice y no dure una eternidad), pero en su contra, son capaces de detectar patrones no lineales y ocultos en los datos y son el tipo de modelos que mejores resultados han dado en el campo de la Visión por Computador, el procesamiento del Lenguaje Natural y las secuencias de Series Temporales.

La elección del modelo no suele ser una tarea fácil. A menudo, el modelo más preciso no suele ser el mejor si la predicción tarda mucho tiempo en realizarse o si el coste de poner este modelo en producción es alto. A veces es recomendable combinar el resultado de varios modelos más simples (ensemble). Un tipo de modelo será bueno identificando un tipo de categoría pero quizás flaquee en otra, y otro tipo puede ser bueno generalizando en ambos. La opinión de varios tipos de modelos suele ser lo que da el mejor resultado.

BENEFICIOS EMPRESARIALES

La Inteligencia Artificial ofrece multitud de beneficios a las empresas y aplicar bien IA se traduce en una ventaja competitiva. Las empresas más valiosas y que más han crecido en los últimos años han confiado en algún tipo de IA en alguno de sus procesos y estrategias de negocio. Algunas empresas están más avanzadas y otras menos, pero lo que parece claro es que si no incorporas la IA quedarás obsoleto más pronto que tarde.

Sin embargo, recientemente hemos visto que la IA y la tecnología aún tienen que encontrar su rol, pues tienen muchas virtudes, pero algunos defectos. Empresas como Twitter, Shopify, Meta (Facebook), Netflix o Uber han tenido que hacer un ajuste en el personal, despidiendo hasta al 50% de su plantilla en algunos casos, porque han visto que se estaban dedicando muchos recursos a productos y avances tecnológicos que no estaban reportando un impacto económico positivo.

Invertir en IA porque nuestros competidores lo hacen o para no quedarse atrás, no debería ser la principal palanca que active a las empresas a actuar. A menudo soluciones tecnológicas ya existentes, menos complejas y costosas, ya ofrecen una solución bastante buena a nuestro problema. Otro punto a destacar es que muchas industrias necesitan una transformación digital previa antes siquiera de querer construir un modelo de predicción con la última tecnología del mercado. La realidad es que la IA no es la gallina de los huevos de oro y para algunos propósitos no es la mejor herramienta.

El Machine Learning y la Inteligencia Artificial dan resultados sorprendentes cuando:

- El problema es demasiado complejo para ser resuelto mediante reglas y condiciones. El conocimiento ganado tras años de experiencia en un campo podría permitir conocer los factores determinantes que lideran en el mercado, pero este proceso puede ser muy complicado o incluso imposible. El ML permite capturar estos patrones.

- El problema está cambiando constantemente. Esto lleva a que el trabajo realizado hoy, deje de ser útil mañana.

- Se trata de un fenómeno sin estudiar. Si existen datos sobre este fenómeno, la IA puede llegar a predecir cuándo volverá a ocurrir este fenómeno y poder actuar antes de que pase.

- El problema tiene un objetivo simple. Nuestra variable objetivo es única, no existe ambigüedad.

- Los costes de la alternativa son muy altos. Los costes de contratar a un profesional experto en la materia pueden ser muy altos, y construir un modelo de predicción es la opción más viable económicamente.

Sin embargo, **no se recomienda utilizar IA cuando**:

- Cada acción del sistema debe ser explicable.
- El coste de un error del sistema es muy alto.
- Obtener los datos adecuados es muy complicado o imposible.
- El desarrollo de software tradicional ya ofrece una solución muy buena y a un coste mucho menor.
- Una simple heurística funciona razonablemente bien.
- El fenómeno a predecir tiene muchos posibles resultados y existe mucha ambigüedad.

Lanzar proyectos de Inteligencia Artificial depende de la situación de cada empresa y no es una tarea fácil. Para que todo funcione a la perfección, los sistemas deben ser robustos y adaptarse a la necesidad de los usuarios, los datos deben ser de calidad y el equipo de desarrollo debe tener las capacidades suficientes como para poder poner en producción la solución. El trabajo no acaba ahí, pues habrá que hacer seguimiento e iterar nuestra solución, ya sea entrenando de nuevo los modelos para capturar las nuevas tendencias o añadiendo nuevas variables predictoras.

Algorithmia comentaba en uno de sus estudios que el 55% de las compañías tenía problemas para poner en producción sus modelos de predicción, y es que incorporarlos a los procesos actuales de una empresa para que den soporte o incluso actúen como decisores en negocio, sigue siendo una tarea complicada.

La IA bien usada tiene grandes beneficios porque:

- Automatiza los procesos. La IA permite realizar análisis y tareas repetitivas, optimizando procesos de manera automática, sin la necesidad de la intervención humana.
- Reduce el sesgo y el error humano. Al reducir la intervención de los humanos, se reduce la posibilidad de cometer errores. Un algoritmo no tiene sentimientos, no está enfadado, triste o contento, por lo que no dudará en tomar una decisión difícil si así lo dicen sus estimaciones y predicciones.
- Aporta precisión. Las máquinas pueden llegar a un detalle mayor que el del ser humano, lo que también reduce el error.
- Agiliza la toma de decisiones. La IA es capaz de analizar miles de datos en poco tiempo y ofrecer una recomendación con la que el negocio pueda tomar una decisión.
- En consecuencia, potencia la creatividad. Libera a las personas de trabajos más repetitivos y a menudo pesados, para que sean más creativas y tengan más tiempo para idear.

DESAFÍOS SOCIALES

Las empresas y gobiernos que utilizan IA aún tienen varios desafíos a resolver. Hemos podido ver ya varios casos donde empresas que han lanzado proyectos con una alta dependencia en IA se han visto involucradas en polémicas y escándalos, experimentado algún que otro desastre y viviendo grandes pérdidas económicas.

En 2016 Microsoft publicó Tay, un bot de conversación que se comunicaba a través de Twitter. En menos de 24 horas, los usuarios de Twitter educaron a este bot para que se convirtiera en una persona que comentaba mensajes xenófobos, racistas o con contenido sexual inapropiado. Tay empezó con tweets mencionando lo cool que eran los humanos, pero acabó publicando que Hitler tenía razón sobre los judíos. En tan solo 16 horas, Microsoft dio de baja su creación.

No es el primer caso en el que la IA actúa con comportamientos racistas, machistas y xenófobos. Corregir estos comportamientos y deshacerse del sesgo es un reto aún por resolver. Cabe mencionar que la IA no ha aprendido estos comportamientos porque decida ser así, sino porque ha sido expuesta a datos e interacciones de muchas otras personas que denotan este comportamiento, y la IA ha interpretado este comportamiento como normal.

La existencia de estos sesgos genera dudas en temas mucho más delicados como la salud o el acceso a la financiación que ofrecen los bancos. Este sesgo puede denegar un préstamo a una persona simplemente por pertenecer a cierta etnia.

En algunas ocasiones, tener total confianza en la IA también puede llevar a grandes pérdidas económicas. En 2021, la empresa inmobiliaria estadounidense Zillow declaraba que había perdido 881 millones de dólares debido a su nueva línea de negocio de House Flipping. El House Flipping consiste en un algoritmo de IA y ML que compra inmuebles para venderlos más tarde a un precio más alto. El proyecto resultó ser un completo desastre.

Otra de las incógnitas es qué pasará con los trabajos actuales si la Inteligencia Artificial se impone. Los nuevos avances de la robótica o lanzamientos como GPT-3, DALL·E o ChatGPT pueden reemplazar el trabajo de personal administrativo, desarrolladores y creadores de contenidos; lo que podría llevar a la desaparición de muchos puestos de trabajo.

Sin embargo, como ocurrió en la revolución industrial cuando las máquinas reemplazaron el trabajo manual y repetitivo de millones de personas, se espera que surjan nuevos puestos de trabajo.

La automatización y la IA harán que nuestros trabajos sean más eficientes y rápidos, pero la interacción humana por el momento parece necesaria. Para que la IA funcione correctamente se necesita que los datos estén bien recogidos y sean de calidad, que personas con criterio, entendimiento y experiencia monitoricen qué decisiones está tomando la IA. Tareas como estandarización de los datos, manejo de la seguridad y la integridad, entrenamiento de los modelos, y perfiles como diseñador de sistema de IA, expertos en IA o seguridad de los sistemas, serán necesarios en el futuro.

Otro gran dilema aparece con los coches autónomos. En caso de accidente con peatones o ciclistas, ¿el sistema debe ser diseñado para proteger la vida de los individuos dentro del vehículo o debería proteger a los peatones?

La pregunta clave está en cómo programar el algoritmo que tome la decisión “adecuada” en cada situación. ¿Podrá la IA diferenciar entre lo que está bien y lo que está mal? ¿Y si llega el momento en que no podemos controlar qué decisiones toma la IA?

El estado de completa autonomía se llama Singularidad. Elon Musk, Bill Gates o Stephen Hawking han hablado sobre este momento cuando no podemos manejar a la IA, hasta el punto en que pueda ser una amenaza para la humanidad. Es posible que la IA decida actuar de forma armada ante un conflicto o actuar en contra de nuestros intereses.

Redactar leyes y regularizar la IA parece necesario antes de que sea demasiado tarde. Varias organizaciones también están usando la data para hacer el bien y combatir aquellos casos en los que empresas que emplean IA están siendo injustas.

Los 5 principios de la ética de la IA según la Universidad de Helsinki son los siguientes:

Utilizar la IA para causar el bien y no el mal.

Quién es el responsable cuando la IA causa el mal.

La IA debe ser transparente y debemos entender qué hace y por qué lo hace.

La IA debería ser justa y no debería discriminar.

La IA debería promover y respetar los derechos humanos.

ÉTICA Y LEGALIDAD

La Inteligencia Artificial actuará mal en el pasado, presente y lo hará en el futuro, por esto la sociedad y los gobiernos se han visto forzados a redactar leyes que limitan la actuación de la IA.

La IA suele ser un sistema opaco, con sesgo e intrusivo, que puede llegar a infringir la privacidad de las personas. La Unión Europea lanzó el Reglamento General de Protección de Datos (RGPD) el 25 de mayo de 2018. Su objetivo principal es dar control a los ciudadanos y residentes sobre sus datos personales y simplificar el entorno regulador de los negocios internacionales, unificando la regulación dentro de la UE.

El RGPD **prohíbe la utilización de información personal** como puede ser la etnia, orientación sexual, convicciones religiosas u opiniones políticas, y pide el consentimiento del usuario para poder utilizar información como la dirección, los ingresos o el documento nacional de identidad.

El RGPD debería evitar casos como el escándalo de Cambridge Analytica. En la década de 2010, la consultora británica Cambridge Analytica recopiló datos de millones de usuarios de Facebook sin su consentimiento, principalmente para utilizarlos con un fin de propaganda política. Este tipo de información se utilizó para asistencia analítica a las campañas de Ted Cruz y Donald Trump para las elecciones presidenciales de 2016, para interferir en el referéndum del Brexit y en algunas elecciones de otros países.

Otro aspecto a discutir está en el impacto que puede tener la IA en sectores como el militar, la salud o la educación. **LA IA DEBERÍA HACER EL BIEN Y NO EL MAL,** sin embargo la línea del bien y el mal es muy difusa.

La ley actual no abarca todos los casos y existen huecos donde empresas o la misma IA puede operar realizando acciones en contra de la moralidad común. Es en este punto donde la ética debe imperar. La IA debe ser transparente, no malévola, respetar los derechos humanos, ser justa y en caso de mala actuación, debe haber responsables de esta mala praxis, ya sean personas, gobiernos o empresas.

Para que la IA sea transparente debe ser explicable. Las personas deberían poder entender qué factores se han tenido en cuenta para determinar la aprobación del crédito a un cliente. Sin embargo, los modelos más efectivos y que mejor rendimiento dan suelen ser "cajas negras". A una caja negra se le entran unos datos y devuelve un resultado, pero se desconoce qué se ha razonado para llegar a esta solución.

Los sistemas de IA también tienen que ser robustos, justos y defender la privacidad de los usuarios. Los sistemas deberían estar operativos, evitar el acceso a ataques cibernéticos, ser justos y no tener sesgo que favorece a los grupos privilegiados y que penaliza a los grupos discriminados. Ya existen librerías de programación que mitigan este sesgo y proporcionan interpretabilidad a los modelos de ML y DL.

De esta manera, la IA será más utilizada, porque además de dar resultados geniales, se puede confiar en ella y es posible justificar qué razones han llevado al sistema a tomar esa decisión.

APLICACIONES PRÁCTICAS

En muchas ocasiones, pese a que la IA, el ML o el DL sean herramientas muy potentes, otra alternativa más tradicional puede dar resultados similares o incluso mejores, siendo además una opción mucho menos costosa en términos económicos, de esfuerzo o tiempo.

La gran mayoría de modelos de ML que se utilizan son de aprendizaje supervisado, es decir, necesitan datos etiquetados. Cuando no existen datos etiquetados, los humanos deben categorizar miles de registros con el fin de tener una muestra de la que el modelo pueda aprender patrones. Por ejemplo, compañías que desarrollan coches autónomos tienen a cientos de personas anotando manualmente objetos, vehículos y señales en horas y horas de vídeo.

Aun así puede ocurrir que no se disponga de datos suficientes. En especial, los modelos de DL necesitan miles, incluso millones, de registros para tener un rendimiento similar al del ser humano. Obtener una muestra amplia de datos sobre eventos fortuitos, eventos que ocurren una vez cada miles de casos, puede ser prácticamente imposible.

La otra limitación a tener en cuenta es la **EXPLICABILIDAD**. Los modelos que mejores resultados ofrecen acostumbran a ser complejos, y encontrar la razón que lleva a tomar esa decisión se ha vuelto complicado. Recientemente se ha ido avanzando en este aspecto y muchas librerías pretenden encontrar qué variables son las más significativas, aplicando permutaciones. Un modelo se entrena con todas las variables menos una y se estudia el efecto que tiene considerar o no esta variable. Sin embargo, seguirá habiendo áreas donde la explicabilidad total sea necesaria, lo que limitará el uso de la IA.

Otro punto a considerar es el **ROI** (Return of Investment). ¿Puede la IA añadir valor, aumentar los ingresos y reducir los costes? La primera comparativa siempre debe realizarse con la de un ser humano medio realizando la misma tarea. ¿Da mejores resultados que tener una persona trabajando y resolviendo ese problema? Si la respuesta es no, no tiene sentido utilizar IA.

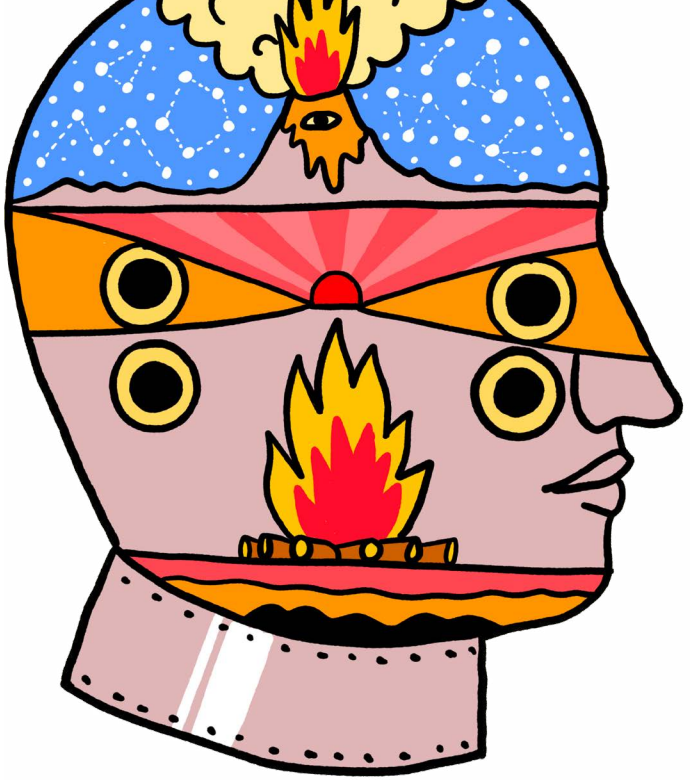
Esta misma pregunta también se traslada a soluciones que ofrece un software normal. Frecuentemente productivizar soluciones complejas resulta tan costoso, que soluciones mucho más sencillas son más fáciles de lanzar al mercado.

No cometer ninguno de los errores anteriores ya te evitará muchos fracasos y decepciones. Esta es la mejor de las recomendaciones. ¡Aunque no acaba aquí! Desarrollar un modelo de predicción para un caso de uso es el primer paso. Hay más trabajo en el despliegue del modelo y en la monitorización. Productivizar el modelo, es decir, hacerlo accesible a nuestros usuarios, tiene mucha interferencia con el mundo del desarrollo de software.

Otro punto importante es conocer qué rendimiento va teniendo el modelo, pues “lo que no se define no se puede medir. Lo que no se mide, no se puede mejorar. Y lo que no se mejora, se degrada siempre”. Conocer que nuestro modelo falla, que la tendencia de los usuarios ha cambiado o que los datos de entrada tienen otro formato, es vital. Es importante desarrollar un plan de contingencia para mitigar cualquier riesgo y responder con una alerta para los responsables, volver a una versión anterior que funcionaba o reentrenar los modelos.

En palabras de William Thomson, lo que no se define no se puede medir. Lo que no se mide, no se puede mejorar. Y lo que no se mejora, se degrada siempre.

FUTURO



El futuro de la IA es incierto, con frentes abiertos e incógnitas por resolver. Existen campos en los que la IA aún no ha demostrado su efectividad y otros donde la IA domina. Los productos que surgen a partir de la IA no dejan de sorprendernos. La gran mayoría de veces para bien, pero en algunas ocasiones para mal.

La evolución de los ordenadores, la mejora en los microprocesadores y los avances en las GPUs y TPUs, que tan potentes son para el entreno de redes neuronales, harán que la IA evolucione a un ritmo más acelerado de lo que lo hace ahora. Si finalmente la computación cuántica se consolida y se vuelve accesible al público general a un coste razonable, este nuevo descubrimiento dispararía los avances en la IA. La computación cuántica permitiría agilizar muchos procesos actuales y resolver problemas mucho más complejos. Industrias como la financiera, logística, transporte, biomedicina y retail, son algunas de las que más se beneficiarían con la aplicación de esta tecnología.

El Deep Learning tiene mucho potencial por explorar y tiene muchas cosas por enseñarnos, como se ha podido ver con el procesamiento del Lenguaje Natural con la aparición de Chat GPT. En otros campos como la Visión por Computador, el procesamiento de voz o el audio también se ha avanzado, pero estos avances no han tenido tanto revuelo mediático. Algunas startups se han especializado en algunos de estos campos y comercializan soluciones a empresas y usuarios mediante el uso del Deep Learning. El Deep Learning ha dejado de estar completamente ligado al mundo de la academia y en la última década la empresa nos ha traído varios productos que se sirven del DL.

En el Machine Learning tradicional, los avances no parecen ser tanto en la mejora del rendimiento, pues los modelos de ensemble y combinaciones de varios modelos dan resultados sorprendentes, sino en la explicabilidad de los modelos. Algunas librerías open source pretenden ya solucionar este problema.

Los productos end-to-end de MLOps están aún por llegar. Existen herramientas, pero no parece haber un claro referente que domine el mercado. El MLOps consiste en el despliegue de modelos de ML o DL en producción mediante software, con el objetivo de poner a disposición del usuario el modelo, monitorizar el rendimiento y mantener los modelos existentes. El MLOps sigue siendo un problema para las empresas, ya que los conocimientos necesarios para poner un modelo en producción escapan de las competencias de los Data Scientists. Es por eso que serán necesarios roles más específicos en este aspecto.

Algunos sectores están al inicio de su transformación gracias a la IA, otros ya son veteranos. En ambos casos, la IA tiene aún avances que explorar. Los campos donde parece que la IA va a explotar son:

TRANSPORTE. Los vehículos autónomos se van a perfeccionar y tarde o temprano serán parte de nuestra realidad.

SALUD. La IA permitirá descubrir medicamentos para enfermedades nuevas y existentes.

EDUCACIÓN. La IA podrá ser capaz de identificar qué alumnos se aburren o a cuáles les cuesta más, en función de sus expresiones faciales y resultados académicos, y así adaptar el nivel y la velocidad a sus necesidades. Mejorando así la educación, algo que parece que sigue en el siglo anterior.

ARTE Y PRENSA. Herramientas como DALL·E o ChatGPT de OpenAI son capaces de entender e incluso generar texto o imágenes. Estas herramientas permitirán generar contenido de manera ilimitada, pero quedará pendiente ver si es material de calidad.

SERVICIO AL CLIENTE. Los sistemas serán capaces de entender las peticiones de los usuarios y elaborar una respuesta en sintonía a los valores y ofertas de las instituciones, tal y como hacen los humanos actualmente.

La IA tiene mucho que aportar, si finalmente el metaverso se convierte en una realidad y el público general empieza a utilizarlo asiduamente. La IA será capaz de generar avatares que podrán interactuar con nosotros de manera perfecta, hasta el punto en que no sabremos quién está detrás de la pantalla, si se trata de otro ser humano o de la IA (superando así el test de Turing, aunque este ya ha sido superado en 2014).

El futuro es incierto porque la IA puede ofrecer grandes beneficios sociales y económicos, aunque mal usada también puede causar mucho daño. Y es que existe un temor global de que no seamos capaces de entender y controlar las decisiones que tome la IA.

Los primeros indicios de AGI (Artificial Generative Intelligence) ya han llegado. Este tipo de inteligencia artificial iguala o excede la inteligencia humana promedio. Esta máquina puede realizar con éxito cualquier tarea intelectual de cualquier ser humano. Los productos de OpenAI (GPT 3, Whisper, DALL·E o ChatGPT) son lo más parecido a AGI que hemos experimentado hasta ahora. La singularidad parece que llegará más pronto de lo que se espera y no estamos preparados para ella.

Para intentar mitigar estos riesgos se deberá avanzar en aspectos legales para poner límites a la IA. Seguimos sin saber si la IA protegerá la privacidad de los usuarios, si actuará a nuestro favor o al suyo propio o cómo actuará en situaciones de peligro. Recientemente, hemos vivido casos de violación de la privacidad con Cambridge Analytica con Facebook o Alexa de Amazon, escuchando conversaciones que no debían. La captación de este tipo de datos puede ser utilizado para predecir nuestro comportamiento, pero inflige nuestra privacidad. Límites legales y establecer unos principios éticos comunes entre todos, son fundamentales en este momento.

La Inteligencia Artificial aún tiene mucho que decidir, lo que hemos visto hasta ahora es un 1% de lo que la IA puede ofrecernos. La realidad supera la ficción. El hype es real. Existen muchas expectativas exageradas, pero lo que la IA necesita es tiempo para consolidarse, para volverse una commodity en nuestro día a día y que no sepamos vivir sin ella.

(III) Machine Learning

Recomendaciones de películas, reconocimiento por voz y asistentes virtuales, son solo algunas de las capacidades de las máquinas para aprender de los seres humanos. Machine Learning está revolucionando la vida que conocemos.

Pero, ¿será la clave para el futuro?

POR MASSIMILIANO BREVINI

¿QUÉ ES?

Para empezar, el aprendizaje automático es una subárea central de la Inteligencia Artificial (IA). Las aplicaciones de ML aprenden de la experiencia (para ser exactos, de los datos) como lo hacen los humanos, sin necesidad de programación directa. Cuando se exponen a nuevos datos, estas aplicaciones aprenden, crecen, cambian y se desarrollan por sí mismas. En otras palabras, el aprendizaje automático consiste en que los ordenadores encuentren información útil sin que se les diga dónde buscar, y es justo aquí donde está la innovación. Por eso sabemos que cuanto más datos mejor, porque lo que hacen estos algoritmos es aprender de los datos en un proceso iterativo.

Las aplicaciones aprenden de cálculos y operaciones anteriores, y utilizan el "reconocimiento de patrones" para producir resultados fiables y fundamentados.

¿Crees que el aprendizaje automático es una de las partes más apasionantes de la Inteligencia Artificial? Nosotros también. Ahora veamos lo siguiente: es importante entender qué es lo que hace que el Aprendizaje Automático funcione y cómo se podrá utilizar en el futuro.

El proceso de aprendizaje automático comienza con la introducción de datos de entrenamiento en el algoritmo seleccionado. Estos pueden ser conocidos o desconocidos. El tipo de datos de entrenamiento (training) que se introduce, influye y es la pieza clave para que el resultado del algoritmo se acerque al resultado esperado.

Los nuevos datos de entrada (test) se introducen para comprobar si funciona correctamente o no. La predicción y los resultados se comparan entre sí. ¿Y entonces qué sucede? Si la predicción y los resultados no coinciden, el algoritmo se vuelve a entrenar con los mismos o diferentes parámetros, varias veces hasta que se obtienen los valores esperados.

Esto permite que el algoritmo de aprendizaje automático, aprenda continuamente por sí mismo y produzca la respuesta óptima, aumentando gradualmente su precisión con el tiempo.

¿Cómo se elige el algoritmo óptimo para un determinado proyecto?

Hay docenas diferentes entre los que elegir, pero no hay una opción mejor ni una que se adapte a todas las situaciones. En muchos casos, hay que recurrir al método científico de prueba y error y entrenar los datos con múltiples algoritmos para establecer cuál de todos ha tenido una mejor performance. Sin embargo, hay algunas preguntas que pueden ayudarnos a reducir las opciones:

¿Cuál es el tamaño de los datos con los que vamos a trabajar?

¿Cuál es el tipo de datos con los que vamos a trabajar?

¿Qué tipo de información buscamos a partir de los datos?

¿Cómo se utilizarán estos datos?

NACIMIENTO

Hoy en día los algoritmos de aprendizaje automático permiten a los ordenadores comunicarse con los humanos, conducir coches de forma autónoma, escribir y publicar informes de partidos deportivos y encontrar sospechosos de terrorismo. Creo firmemente que el aprendizaje automático tendrá un gran impacto en la mayoría de los sectores y en los puestos de trabajo dentro de ellos, por lo que todo directivo debería tener al menos una idea de lo que es el aprendizaje automático y cómo está evolucionando.

A partir de aquí es cuando empezamos un rápido viaje en el tiempo para examinar los orígenes del Machine Learning, así como los hitos más recientes.

1950

1950: Alan Turing crea el "Test de Turing" para determinar si un ordenador tiene inteligencia real. ¿Recuerdas que te lo contamos en detalle en el capítulo de Inteligencia Artificial?

1952: Arthur Samuel escribió el primer programa de aprendizaje para ordenadores. El programa era el juego de las damas y el ordenador de IBM mejoraba en el juego cuanto más jugaba, estudiando qué jugadas constituían estrategias ganadoras e incorporándolas a su programa.

1957: Frank Rosenblatt diseña la primera red neuronal para ordenadores (el perceptrón), que simula los procesos de pensamiento del cerebro humano.

1960-1970

1967: Se escribe el algoritmo del "vecino más cercano" o KNN, que permite a los ordenadores empezar a utilizar un reconocimiento de patrones muy básico. Este algoritmo podía utilizarse para trazar una ruta para los vendedores ambulantes, comenzando en una ciudad al azar, pero asegurándose de que visitaran todas las ciudades durante un breve recorrido.

1979: Los estudiantes de la Universidad de Stanford inventan el "carrito de Stanford", que puede sortear los obstáculos de una habitación por sí solo.

1980-1990

1981: Gerald Dejong introduce el concepto de aprendizaje basado en explicaciones (EBL), en el que un ordenador analiza los datos de entrenamiento y crea una regla general que puede seguir descartando los datos sin importancia.

1985: Terry Sejnowski inventa NetTalk, que aprende a pronunciar las palabras del mismo modo que un bebé.

LOS 90: El trabajo sobre el aprendizaje automático pasa de un enfoque basado en el conocimiento a otro basado en los datos. Los científicos empiezan a crear programas para que los ordenadores analicen grandes cantidades de datos y saquen conclusiones, o aprendan de los resultados.

1997: Deep Blue de IBM vence al campeón mundial de ajedrez (como te contamos en el capítulo anterior).

2000-2014

2006: Geoffrey Hinton acuña el término "aprendizaje profundo" para explicar los nuevos algoritmos que permiten a los ordenadores ver y distinguir objetos y texto, en imágenes y vídeos.

2010: El Microsoft Kinect puede seguir 20 rasgos humanos a una velocidad de 30 veces por segundo, lo que permite a las personas interactuar con el ordenador mediante movimientos y gestos.

2011: Watson, de IBM, vence a sus competidores humanos en Jeopardy (lee nuestro Glosario para saber más).

También se desarrolla Google Brain, permitiendo a su red neuronal profunda aprender a descubrir y categorizar objetos de forma similar a como lo hace un gato.

2012: El X Lab de Google desarrolla un algoritmo de aprendizaje automático que es capaz de explorar de forma autónoma los vídeos de YouTube para identificar los que contienen gatos.

2014: Facebook desarrolla DeepFace, un algoritmo de software que es capaz de reconocer o verificar a los individuos en las fotos al mismo nivel que pueden hacerlo los humanos.

2015-2021

2015: Amazon lanza su propia plataforma de aprendizaje automático.

Ese mismo año, Microsoft crea el kit de herramientas de ML distribuido, que permite repartir de manera eficiente los problemas de aprendizaje automático, en varios ordenadores.

Además, más de 3.000 investigadores de IA y robótica, respaldados por Stephen Hawking, Elon Musk y Steve Wozniak (entre muchos otros), firman una carta abierta en la que advierten del peligro de las armas autónomas que seleccionan y atacan objetivos sin intervención humana.

2016:El algoritmo de inteligencia artificial de Google vence a un jugador profesional en el juego de mesa chino Go.

2020: La publicación del sistema BERT de Google aceleró las técnicas avanzadas de procesamiento del lenguaje natural (PNL). El modelo viene acompañado de una extraordinaria red de PNL que permite comprender lenguajes más sofisticados y compatibles. El modelo de IA de Google está ampliamente difundido y abierto al uso público.

2021: La red neuronal DALL·E es un avance en visión por ordenador (Computer Vision) desarrollado por OpenAI en 2021 que consiste en crear imágenes a partir de contenido textual. Curiosamente, no se basa en las GAN que se utilizan habitualmente para entrenar redes neuronales para la generación de imágenes, lo que lo convierte en un enfoque increíblemente nuevo. Produce versiones antropomorfizadas de diversos objetos, incluidos los animales.

El impacto del Machine Learning es evidente en todo el mundo. Desde las startups hasta las empresas de Fortune 500 han abrazado esta tecnología. El mercado de ML se valoró en 8.000 millones de dólares en 2021, y estas cifras alcanzarán los 117.000 millones de dólares en 2027, con una CAGR del 39%.

FUNCIONAMIENTO

El aprendizaje automático (ML) consiste en codificar programas que ajustan automáticamente su rendimiento, a partir de la exposición a la información codificada en los datos. Este aprendizaje se consigue mediante un modelo basado en parámetros sintonizables, que se ajustan automáticamente según un criterio de rendimiento.

Mucha información, ¿verdad? Iremos por partes.

El aprendizaje automático puede considerarse un subcampo de la inteligencia artificial (IA). Hay tres clases principales de ML:

1. Aprendizaje supervisado

Algoritmos que aprenden de un conjunto de entrenamiento de ejemplos etiquetados (ejemplares) para generalizar todas las entradas posibles. Algunos ejemplos de técnicas de aprendizaje supervisado son la regresión y las máquinas de vectores de apoyo.

El aprendizaje supervisado utiliza un conjunto de datos de entrenamiento para enseñar a los modelos a obtener los resultados deseados. Este grupo de datos de entrenamiento incluye entradas y salidas correctas, que permiten al modelo aprender a lo largo del tiempo. El algoritmo mide su precisión a través de la función de pérdida (loss function), ajustándose hasta que el error se haya minimizado lo suficiente.

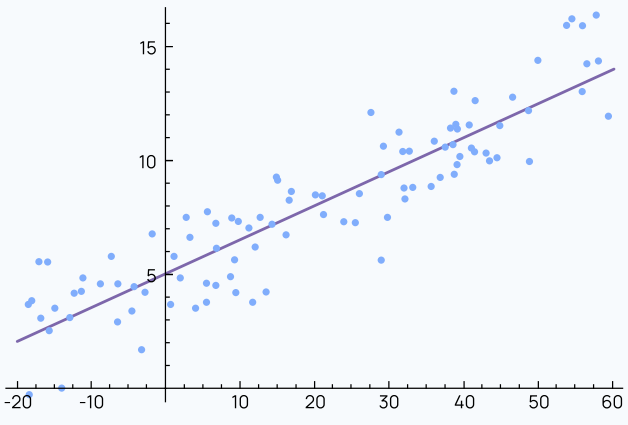
El aprendizaje supervisado puede dividirse en dos tipos de problemas a la hora de extraer datos: clasificación y regresión.

La **clasificación** utiliza un algoritmo para asignar con precisión los datos de prueba a categorías específicas. Reconoce entidades dentro del conjunto de datos e intenta sacar algunas conclusiones sobre cómo deben etiquetarse o definirse esas entidades. Los algoritmos de clasificación más comunes son los clasificadores lineales, las máquinas de vectores de soporte (SVM), los árboles de decisión, los vecinos más cercanos y los bosques aleatorios, que se describen con más detalle a continuación.

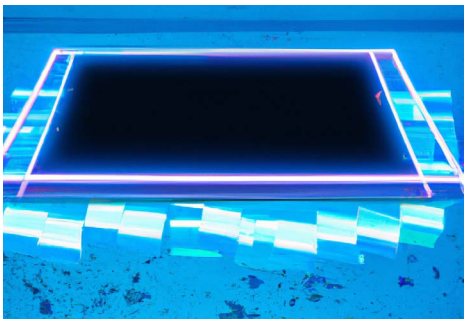
La **regresión** se utiliza para comprender la relación entre variables dependientes e independientes. Se suele utilizar para hacer proyecciones, como por ejemplo de los ingresos por ventas de una empresa determinada. La regresión lineal, la regresión logística y la regresión polinómica, son algoritmos de regresión populares.

Se denomina regresión lineal cuando la función es lineal, es decir, requiere la determinación de dos parámetros: la pendiente y la ordenada en el origen de la recta de regresión. La fórmula de la regresión lineal simple es la siguiente:

$$y=ax+b$$



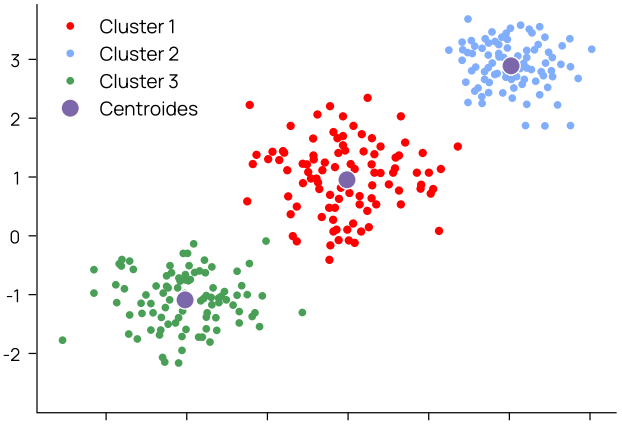
Ejemplo de una regresión lineal con una variable dependiente y una variable independiente.



2. Aprendizaje no supervisado

Algoritmos que aprenden a partir de un conjunto de entrenamiento de ejemplos no etiquetados, utilizando las características de las entradas para categorizarlas juntas según algún criterio estadístico. Algunos ejemplos de aprendizaje no supervisado son la agrupación de K-means y el Kernel Density Estimation.

Para hacer un ejemplo concreto, el K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o clúster. Se suele usar la distancia cuadrática.



Ejemplo algoritmo K-Means (www.unioviedo.es)

3. Aprendizaje por refuerzo (Reinforcement Learning)

Algoritmos que aprenden por refuerzo, basado en recompensar los comportamientos deseados y/o castigar los no deseados. En general, un agente de aprendizaje por refuerzo es capaz de percibir e interpretar su entorno, emprender acciones y aprender por ensayo y error.

El Aprendizaje por refuerzo propone un nuevo enfoque para hacer que nuestra máquina aprenda, para ello, postula los siguientes 2 componentes:

AGENTE: será el modelo que queremos entrenar para que aprenda a tomar decisiones.

AMBIENTE: será el entorno en donde interactúa y "se mueve" el agente. El ambiente contiene las limitaciones y reglas posibles a cada momento.

Entre ellos hay una relación que se retroalimenta y cuenta con los siguientes nexos:

ACCIÓN: las posibles acciones que puede tomar en un momento determinado el agente.

ESTADO (DEL AMBIENTE): son los indicadores del ambiente, de cómo están los diversos elementos que lo componen en ese momento.

RECOMPENSAS (¡O CASTIGOS!): a raíz de cada acción tomada por el agente, podremos obtener un premio o una penalización que orientará al agente en la dirección correcta.

BENEFICIOS EMPRESARIALES

Los innumerables usos del aprendizaje automático indican lo beneficiosa que puede ser esta tecnología para empresas de todo tipo. Las compañías describen sus beneficios de aprendizaje automático en términos de ganancias y mejoras en los diferentes procesos empresariales exponenciales. Te contamos cuáles son algunas de las más comunes:

TOMA DE DECISIONES MÁS RÁPIDA: Al permitir que las empresas procesen y analicen los datos con más rapidez que nunca, el aprendizaje automático permite una toma de decisiones rápida, incluso en fracciones de segundos. Por ejemplo, un software basado en el aprendizaje automático entrenado para identificar anomalías en el entorno de seguridad de una empresa puede detectar automáticamente una violación de datos al instante y notificar al equipo técnico de la organización.

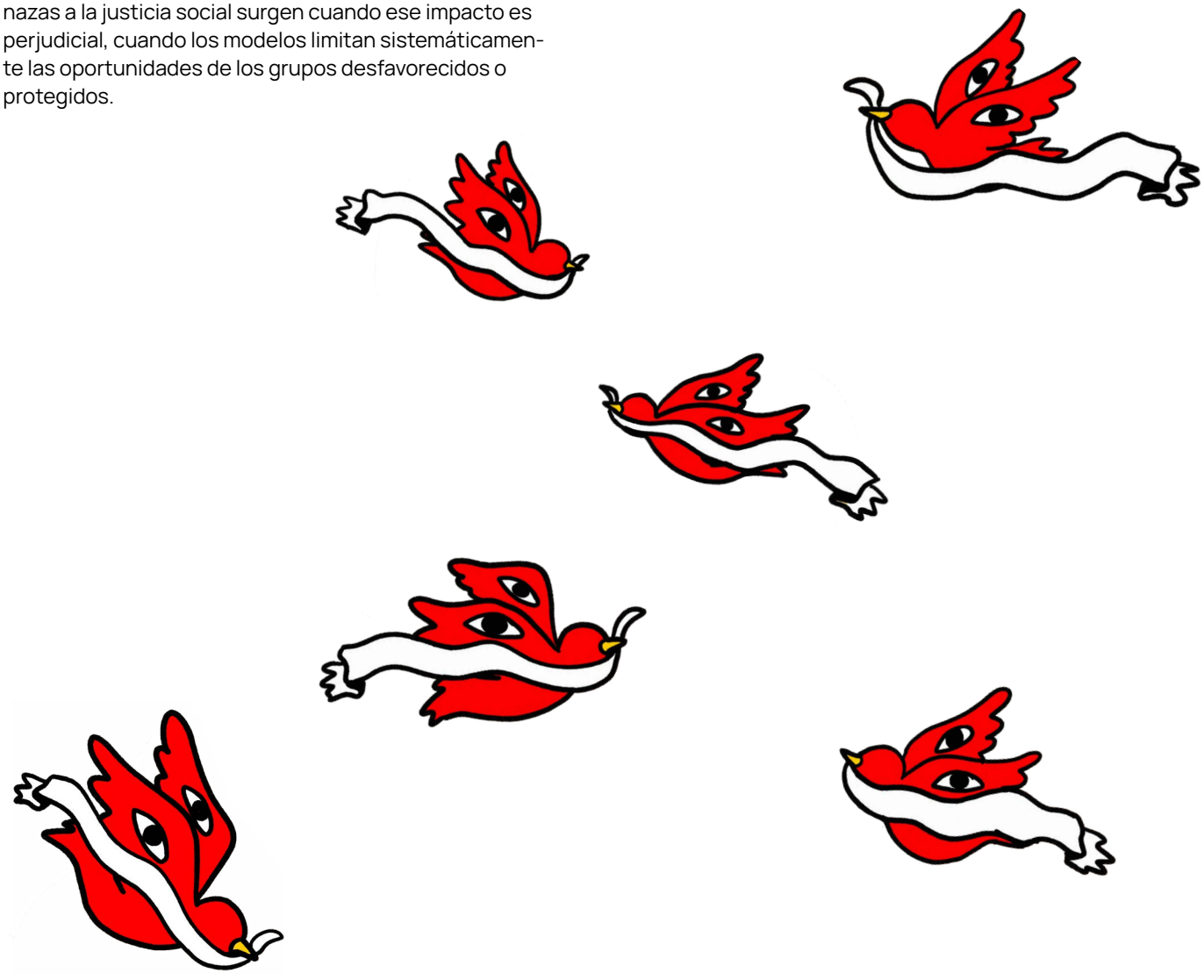
PREVISIÓN DE LA DEMANDA CON MAYOR PRECISIÓN: Para competir en un panorama empresarial que cambia rápidamente, las empresas están sometidas a una presión cada vez mayor para anticipar las tendencias del mercado y el comportamiento de los clientes. Al incorporar modelos de aprendizaje automático a sus análisis de datos, las empresas obtienen capacidades mucho más precisas y potentes para prever la demanda, lo que se traduce en una gestión más eficaz del inventario y un gran ahorro de costes.

PERSONALIZACIÓN DEL COMPROMISO CON EL CLIENTE: La personalización también se ha convertido en una estrategia crítica para competir en el mercado actual. Con las plataformas de aprendizaje automático que analizan el comportamiento del usuario y sugieren productos adicionales en función del historial de compras, los minoristas online interactúan con los clientes de forma más personalizada y consiguen más ventas. El gigante mundial Amazon es un buen ejemplo, ya que utiliza el aprendizaje automático para crear listas de productos recomendados y ofrecer sugerencias a los clientes.

AUMENTO DE LA EFICIENCIA: El uso del aprendizaje automático permite a las empresas acelerar las tareas repetitivas y desplazar los recursos humanos a actividades de mayor valor. Por ejemplo, la tecnología de aprendizaje automático puede realizar búsquedas exhaustivas de documentos en una fracción del tiempo que tardan las personas en realizar tareas de escaneo y referencias cruzadas. Estas capacidades permiten a las empresas reducir los costes de las actividades de recuperación de información relacionadas con el cumplimiento de la normativa y la investigación jurídica, al tiempo que liberan a los empleados para que puedan centrar sus esfuerzos en otros aspectos.

DESAFÍOS SOCIALES

Por la misma razón por la que el aprendizaje automático es valioso, porque impulsa las decisiones operativas con mayor eficacia, también ejerce su poder en el impacto que tiene en la vida de millones de personas. Las amenazas a la justicia social surgen cuando ese impacto es perjudicial, cuando los modelos limitan sistemáticamente las oportunidades de los grupos desfavorecidos o protegidos.



1. Los modelos abiertamente discriminatorios

Son modelos predictivos que basan sus decisiones parcial o totalmente en una clase protegida. Las clases protegidas incluyen la raza, la religión, el origen nacional, el género, la identidad de género, la orientación sexual, el embarazo y el estado de discapacidad. Al tomar una de estas características como entrada, los resultados del modelo (y las decisiones impulsadas por este) se basan, al menos en parte, en la pertenencia a una clase protegida. Aunque los modelos rara vez lo hacen directamente, hay precedentes y apoyo para hacerlo.

2. Inferir atributos sensibles

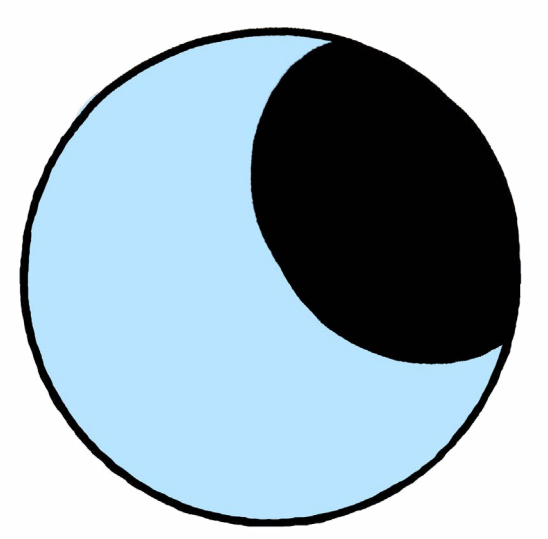
El aprendizaje automático predice información sensible sobre las personas, como la orientación sexual, el embarazo, si alguien va a dejar su trabajo, y hasta si va a morir. En un caso especialmente extraordinario, los funcionarios de China utilizan el reconocimiento facial para identificar y seguir a los Uigures, un grupo étnico minoritario sistemáticamente oprimido por el gobierno. Se trata del primer caso conocido de un gobierno que utiliza el aprendizaje automático para establecer perfiles por etnia. Por su parte, una empresa china valorada en más de 1.000 millones de dólares dijo que su software podía reconocer “grupos sensibles de personas”.

3. Microfocalización depredadora

La impotencia engendra impotencia, y ese ciclo puede ampliarse para los consumidores, cuando el aprendizaje automático aumenta la eficiencia de las actividades diseñadas, maximizando los beneficios de las empresas. La mejora de la microfocalización del marketing y la fijación de precios predictivos de los seguros y los créditos, puede magnificar el ciclo de la pobreza. Por ejemplo, los anuncios altamente segmentados son más hábiles que nunca para explotar a los consumidores vulnerables y separarlos de su dinero.

Y los precios de los seguros pueden llevar al mismo resultado. En el caso de los seguros, el nombre del juego es cobrar más a los que corren más riesgo. Si no se controla, este proceso puede desembocar rápidamente en una tarificación depredadora. Por ejemplo, un modelo de rotación puede descubrir que los asegurados de edad avanzada no tienden a comparar y a cambiar de oferta, por lo que hay menos incentivos para mantener las primas de sus pólizas bajo control. Y la tarificación de las primas en función de otros factores vitales también contribuye a un ciclo de pobreza. Por ejemplo, a las personas con mala calificación crediticia se les cobra más por el seguro del coche. De hecho, una baja puntuación crediticia puede aumentar la prima más que un accidente de coche con culpa.

APLICACIONES PRÁCTICAS



Malas prácticas

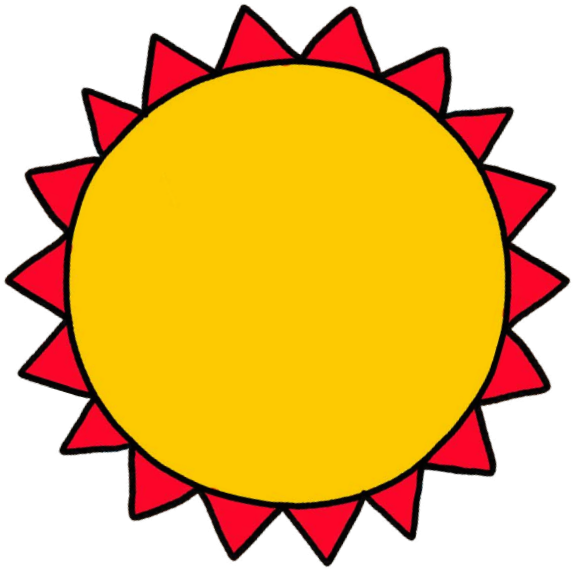
ERROR 1. Una solución de aprendizaje automático que busca un problema.
A menudo se intenta utilizar una solución de Machine Learning cuando el problema de negocio no lo requiere, gastando así recursos y tiempo innecesarios.

ERROR 2. Si no hay datos, no se identifica el problema.
A veces las empresas, sobre todo las más pequeñas, no tienen los datos suficientes para abordar un proyecto de Machine Learning.

ERROR 3. Diseñar una arquitectura monolítica.
La creación de aplicaciones monolíticas provoca el proceso de desarrollo en cascada. Los diferentes pasos y partes de una aplicación son muy dependientes unos de otros.

ERROR 4. Volver a inventar la rueda.
Se han visto muchos ejemplos de proyectos que han tenido que reiniciarse porque no se había investigado el trabajo anteriormente.

ERROR 5. No comunicar el progreso.
Por mucho éxito que tenga el ML, habrá contratiempos en el proyecto a lo largo del camino. Hemos comprobado que un informe semanal de 1-2 páginas sobre el estado del proyecto para el patrocinador y el director del proyecto del cliente, aunque no lo hayan solicitado, elimina la mayoría de los problemas de comunicación en el proyecto.



Buenas prácticas

MEJOR PRÁCTICA 1: Entender el problema de la empresa, definir la solución de aprendizaje automático.
No contrates a empresas que no tengan o no te permitan definir una solución viable de aprendizaje automático.

MEJOR PRÁCTICA 2: El bajo coste significa que la fruta que cuelga es baja. Tus primeros proyectos deben añadir capacidades, no sustituir ni mejorar las existentes.

Te damos un consejo: Hasta que no hayas demostrado el valor añadido del aprendizaje automático, evita añadir aplicaciones de ML al sistema heredado (existente) de una organización. Es más fácil decirlo que hacerlo. ¡Lo sabemos!

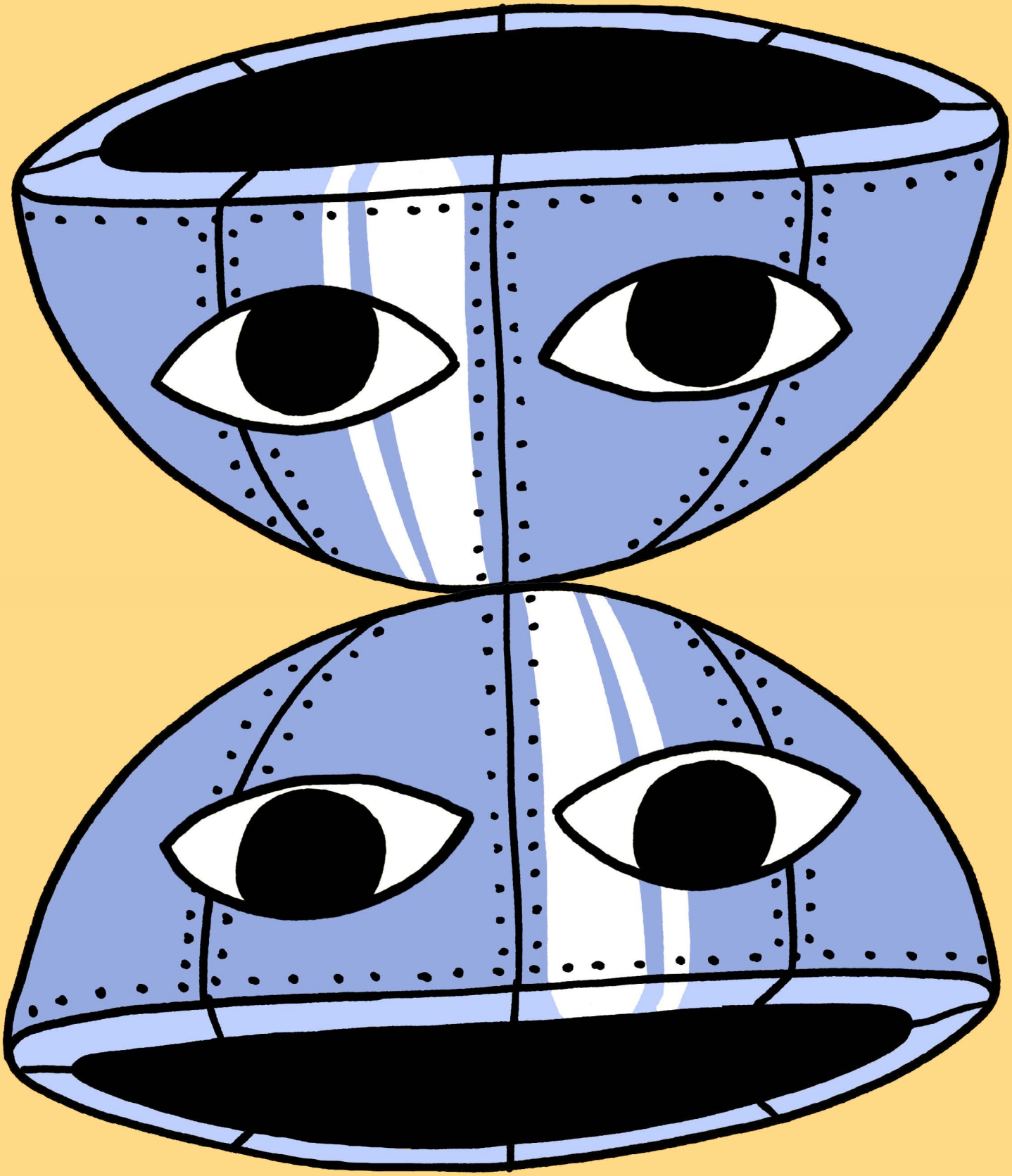
MEJOR PRÁCTICA 3: Encontrar e identificar el caso de negocio, el flujo de procesos y/o los diagramas de flujo de datos. Los KPI (indicadores clave de rendimiento) son un gran indicador de lo que la organización considera importante. La organización puede medir una variedad de métricas de salud de departamentos, procesos y proyectos.

Además, hay que encuestar a los interesados y averiguar qué es lo que más les frustra. Por ejemplo, una cadena hotelera mide la disponibilidad de habitaciones, o un restaurante mide la disponibilidad de mesas. El ML puede predecir las reservas futuras basándose en las pasadas. Sin embargo, el ML más valioso es el que disminuye la disponibilidad.

MEJOR PRÁCTICA 4: Trabajar en un eje de acción a la vez.Una vez identificado el problema, proponer una solución de Aplicación de Machine Learning (AML). Se pueden poner en marcha otros proyectos de AML después de que el primero tenga éxito.

MEJOR PRÁCTICA 5: Si el primer proyecto ha salido con éxito y el cliente está satisfecho, entonces se puede aumentar la inversión poniendo en marcha más proyectos a la vez.

Esta mejor práctica (iterativa) es evidente. Lo que no es tan obvio es que los patrocinadores pueden querer iniciar más proyectos de AML antes de que el primero se ponga en producción, lo que requeriría una gestión del proyecto distinta (por ejemplo, utilizando la metodología Kanban, trabajando en paralelo y ayudándonos a tener una gestión de trabajo más fluida gracias a la visualización del trabajo por fases).



FUTURO

El futuro del aprendizaje automático es excepcionalmente emocionante. En la actualidad, casi todos los ámbitos comunes se nutren de aplicaciones de aprendizaje automático. Por nombrar algunos de ellos: la sanidad, los motores de búsqueda, el marketing digital y la educación, son los principales beneficiarios.

Parece prácticamente imposible trabajar en un dominio desprovisto de esta nueva tecnología para lograr los resultados previstos de forma eficiente. El aprendizaje automático podría ser un mérito para una empresa o una organización, ya sea una multinacional o una empresa privada, ya que las tareas que aún se realizan de forma manual serán ejecutadas en su totalidad por las máquinas, en el futuro.

Según Gartner, la institución líder mundial en investigación, asesoramiento y consultoría, el aprendizaje automático es recordado por casi todas las últimas tendencias y patrones encontrados en los círculos literarios. El aprendizaje automático está preparado para cambiar nuestras vidas de una manera que era imposible décadas atrás. En su resumen de los 10 principales patrones de innovación, Gartner afirma que el razonamiento computarizado y las nuevas técnicas de ML han llegado a un punto de inflexión básico y aumentarán y ampliarán progresivamente a todos los efectos cada asistencia, cosa o aplicación, potenciada por la innovación. La creación de marcos inteligentes avanzados que aprendan, se ajusten y posiblemente actúen de forma autosuficiente, en lugar de limitarse a ejecutar directrices predefinidas, es fundamentalmente un hito para los comerciantes de innovación y los proveedores de tecnología.

Durante el tiempo de la post-industrialización, los individuos han intentado hacer una máquina que actúe y haga cada actividad igual que un humano. Como resultado, el aprendizaje automático se convierte en la mayor bendición de la IA para la humanidad, para la realización efectiva de los objetivos. Por otra parte, las técnicas de máquinas autodidactas han cambiado considerablemente las pautas de empleo de las grandes empresas.

Últimamente, los vehículos automáticos autodirigidos, los ayudantes computarizados, los miembros del personal mecánico, los robots y las áreas urbanas inteligentes han demostrado que las máquinas inteligentes son concebibles y podrían dar resultados tentadores. La inteligencia simulada a semejanza de la mente y el cerebro humano ha cambiado la mayoría de las áreas industriales, como el comercio minorista, la producción, la construcción, la contabilidad, los servicios médicos, los medios de comunicación y la ingeniería. Y sigue ocupando nuevas regiones con un vigor cada vez mayor. Las cinco áreas están pensadas como avances futuristas del aprendizaje automático.

Una de las apuestas para el futuro cercano en este campo es la computación cuántica. Por ahora no hay aplicaciones de hardware o algoritmos cuánticos listos para su comercialización. Sin embargo, para hacerla despegar, varias agencias gubernamentales, instituciones académicas y grupos de reflexión, han invertido millones.

La introducción de la computación cuántica en el aprendizaje automático cambiaría por completo este campo, ya que asistimos a un procesamiento instantáneo, a un aprendizaje rápido, a una ampliación y mejora de las capacidades. Esto implica que en una pequeña fracción de tiempo, se podrán resolver cuestiones complicadas que no podemos abordar con los métodos convencionales y las tecnologías existentes.

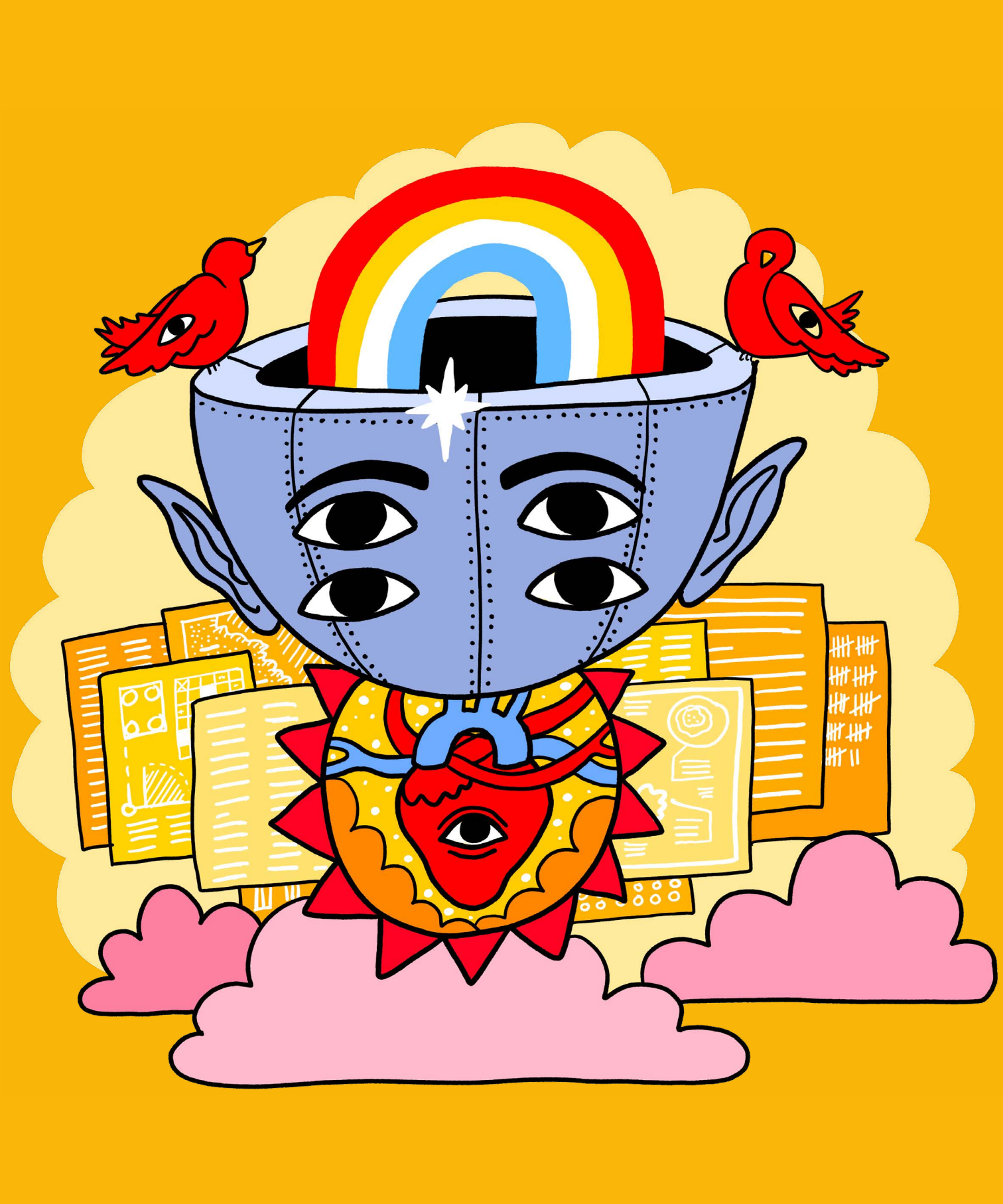
¿Te gustaría ser parte de esto?

(IV) Deep Learning

Los ordenadores se acercan cada vez más al funcionamiento del cerebro humano, e incluso en ciertos aspectos, lo superan.

Pero, ¿cuáles son las ventajas y desventajas de que aprendan por sí mismos y sean capaces de tomar decisiones de manera autónoma?

POR JESÚS PRADA



¿QUÉ ES?

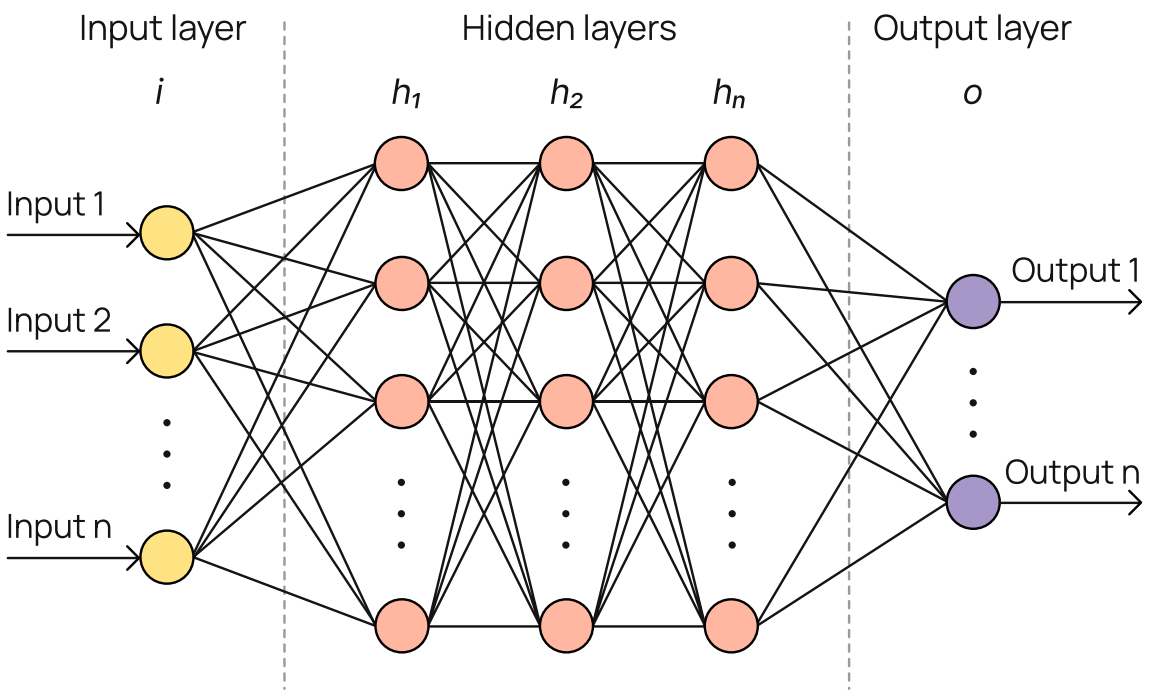
El concepto de aprendizaje profundo, o Deep Learning (DL por sus siglas en inglés), ha tenido diferentes interpretaciones en los últimos años. A menudo DL se emplea simplemente para referirse a un subconjunto específico de Redes Neuronales Artificiales (o ANN por sus siglas en inglés), una familia de modelos de Aprendizaje Automático, o Machine Learning (ML), que pueden utilizarse tanto para tareas de clasificación como de regresión. En concreto, se utiliza para denominar a las ANN con un gran número de lo que se denominan capas ocultas. Ahora bien, ¿estás dispuesto a entrar en detalle en este mundo? Te contaremos lo esencial que debes saber.

Un modelo de ANN está formado por un conjunto de unidades conectadas llamadas neuronas, donde la salida de cada neurona se calcula mediante alguna función no lineal, llamada función de activación, aplicada a la suma ponderada de sus entradas. Las conexiones neuronales tienen pesos o coeficientes asociados, por lo que las activaciones de distintas neuronas pueden tener mayor impacto que otras.

Las neuronas de una capa pueden conectarse a neuronas de las capas anterior y posterior. La capa que recibe los datos externos es la capa de entrada y la última capa, la que produce el resultado final, es la de salida. Entre las de entrada y salida hay cero o más capas ocultas. Cuando el número de estas capas ocultas es grande, se habla de Redes Neuronales Artificiales Profundas. ¿Qué tal esa explicación? Sabemos que no es tan simple de comprender, así que te mostraremos un ejemplo de este tipo de modelos (1).

Sin embargo, la denominación DL también se ha utilizado para referirse a cualquier tipo de marco de modelos de Aprendizaje Automático que consista en un esquema de entrenamiento con varias capas de optimización, cada una de las cuales afecta al resultado del modelo final. Un ejemplo de ello son las Deep Belief Networks, un tipo de modelos de Aprendizaje Automático que se utiliza para el aprendizaje no supervisado y se basa en múltiples capas, que presentan diferencias significativas respecto al esquema estándar de una ANN, que hemos descrito anteriormente.

No obstante, es cierto que el vínculo entre DL y las ANN profundas es fuerte y casi omnipresente en la actualidad. Probablemente han influido en ello varios factores, entre ellos, el hecho de que el esquema de las ANN se adapta casi a la perfección al concepto de Aprendizaje Profundo, y que algunos de los primeros avances pioneros en DL corresponden en efecto a este tipo de estructuras.



(1) Ejemplo de algoritmo K-Means

Teniendo en cuenta lo anterior, debemos tener claro que, aunque a menudo se considera un campo independiente, el Aprendizaje Profundo no es ni más ni menos que otra familia de modelos de Aprendizaje Automático. Sin embargo, es una familia de modelos con algunas propiedades extremadamente relevantes, destacando las dos siguientes:

Potencial Predictivo

Hoy en día se dispone de conjuntos de datos cada vez más grandes para su uso en el entrenamiento de modelos de ML. Con el fin de aprovechar al máximo la información y el potencial predictivo de estos grandes conjuntos de datos, es necesario utilizar métodos complejos, capaces de extraer la máxima información posible de estos datos. Las máquinas de vectores soporte, son uno de los modelos más complejos entre todas las familias estándar de modelos de ML, y esa es la razón principal de su dominio en el pasado durante un largo periodo de tiempo. Sin embargo, presentan importantes problemas de escalabilidad cuando se trabaja con grandes volúmenes de datos.

Con el auge de los marcos de Deep Learning se ha demostrado que estos modelos DL son capaces de lograr un rendimiento superior cuando se entrenan con conjuntos de datos que son suficientemente grandes. Este hecho es probablemente el principal factor por el que esta familia de modelos se está convirtiendo en la elección preferida para resolver una gran variedad de tareas de aprendizaje supervisado.

Aprendizaje de extremo a extremo

En un proyecto de ML estándar, una de las etapas principales a llevar a cabo es el pre-procesamiento de datos. Esta etapa incluye varios pasos, entre ellos lo que suele denominarse feature engineering, es decir, la creación y selección de variables a usar como valores de entrada del modelo predictivo.

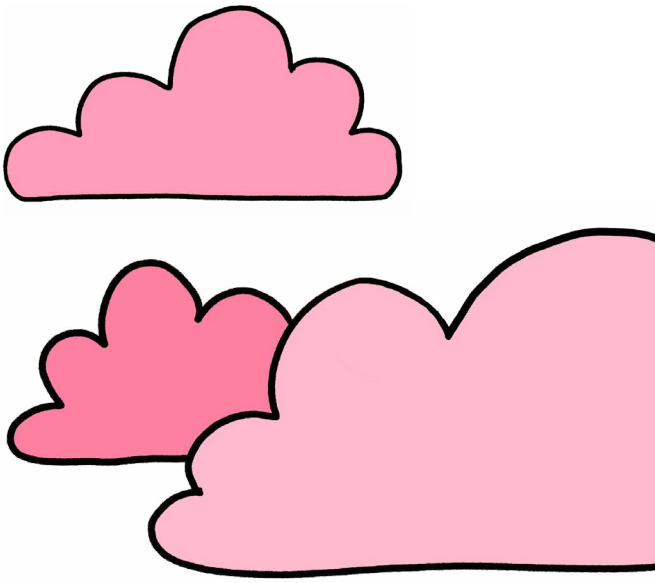
Sin embargo, debido a la naturaleza específica de los marcos de Aprendizaje Profundo, que constan de varias capas que llevan a cabo tareas intermedias necesarias para resolver problemas de ML, estos pasos de preprocesado pueden dejar de ser necesarios al aplicar modelos de DL. Esta propiedad suele denominarse aprendizaje end-to-end y permite a los investigadores y científicos de datos evitar pasos complejos y lentos que antes eran necesarios y que habitualmente requerían de la ayuda de expertos humanos en el campo correspondiente a la tarea en cuestión.

Un ejemplo ilustrativo puede encontrarse en el reconocimiento de voz, donde el objetivo es tomar una entrada, como un clip de audio, y asignarla a una salida, que sería una transcripción del clip de audio. Tradicionalmente, el reconocimiento de voz requería más de una etapa de procesamiento. Primero había que extraer algunas características del audio con métodos de preprocesamiento, como los Coeficientes Cepstrales en las Frecuencias de Mel o MFCCs, que son coeficientes para la representación del habla basados en la percepción auditiva humana. Después, una vez extraídas algunas características de bajo nivel, se podría aplicar un algoritmo de Aprendizaje Automático para encontrar, por ejemplo, los fonemas básicos del sonido en el clip de audio.

Cuando se utilizan marcos DL, este proceso de múltiples etapas puede sustituirse directamente por el entrenamiento de una Red Neuronal Profunda, lo que permite introducir el clip de audio y obtener directamente la transcripción como salida. Sin embargo, es importante señalar que uno de los retos de la DL de extremo a extremo, es que normalmente se necesitan grandes volúmenes de datos antes de que funcione de forma comparable a los marcos clásicos de ML con múltiples pasos de preprocesado, e incluso mayores para poder superar el rendimiento de sus homólogos.

Estas y otras características especiales han hecho de los modelos de Aprendizaje Profundo una de las técnicas más populares en los últimos años dentro del campo del Aprendizaje Automático y, en general, de la Inteligencia Artificial. En particular, este tipo de modelos se ha convertido en referencia en problemas en los que se trabaja con datos no estructurados, es decir, datos que no están representados en forma de tablas o estructuras matriciales, como por ejemplo el reconocimiento de imágenes o el procesamiento de lenguaje natural.

NACIMIENTO



1944

El primer modelo de red neuronal fue propuesto por primera vez en 1944 por Warren McCullough y Walter Pitts. El algoritmo estándar de backpropagation para el cálculo del gradiente durante el entrenamiento de los modelos básicos de ANN, de una sola capa, fue establecido por Frank Rosenblatt en 1958. Por otro lado, la teoría básica correspondiente al perceptrón multicapa, o MLP por sus siglas en inglés, la estructura de Deep Learning más estándar, ya estaba bien establecida en los años 80. De hecho, pueden considerarse como el primer ejemplo de algoritmos modernos de aprendizaje automático que podían utilizarse tanto en problemas de regresión como de clasificación, con variaciones conceptuales mínimas.

Finales 1990

Sin embargo, algunos problemas técnicos, debidos esencialmente a lagunas de conocimiento sobre el entrenamiento de estos modelos, unidos a la falta (en aquella época) de la potencia de cálculo necesaria para manejar grandes volúmenes de datos, provocaron su relativo declive a finales de los 90, y el auge de métodos alternativos, en particular las Máquinas de Vectores Soporte, para clasificación y regresión.

2012

En los últimos años la popularidad de los modelos DL ha aumentado de forma espectacular, debido a la amplia disponibilidad de potentes instalaciones informáticas y a los avances en los fundamentos teóricos de los MLP. Especialmente a partir del año 2012, gracias al trabajo de autores como Hinton, Bengio y LeCun, por varias mejoras en sus procedimientos de entrenamiento y una mejor comprensión de las dificultades relacionadas con las arquitecturas de muchas capas.

Entre los avances, podríamos destacar el desarrollo de nuevos métodos de optimización, como Adam, que han reemplazado a backpropagation como elección estándar de algoritmo de optimización en modelos DL, la propuesta de nuevos métodos de inicialización de coeficientes, con especial mención a la conocida como Xavier initialization, o el empleo de nuevas funciones de activación como la ReLU. A todos estos factores se suma la aparición de múltiples entornos de desarrollo como TensorFlow y Keras, ambos en 2015, que han permitido a los usuarios experimentar con diferentes arquitecturas de DL, activaciones no diferenciables, e incluso, funciones de pérdida no diferenciables.

“Los modelos DL son claramente la opción más recomendable en prácticamente cualquier escenario en el que se empleen datos no estructurados.”

“El primer modelo de red neuronal fue propuesto por primera vez en 1944 por Warren McCullough y Walter Pitts.”

Por otro lado, se ha demostrado que el entrenamiento de estos modelos es un cálculo en tiempo lineal, lo que asegura un buen nivel de escalabilidad frente a grandes volúmenes de datos. Por último, pero no por ello menos importante, se demostró que los modelos DL son capaces de extraer más poder predictivo que otros marcos de ML existentes en esa época, cuando se entrenan con conjuntos de datos suficientemente grandes. Estas características suponen dos de las principales necesidades a cubrir que empujaron el auge de las estructuras de DL: el contar con modelos más escalables en términos de coste computacional que las Máquinas de Vectores soporte y que a la vez proporcionaran igual o mayor potencial predictivo.

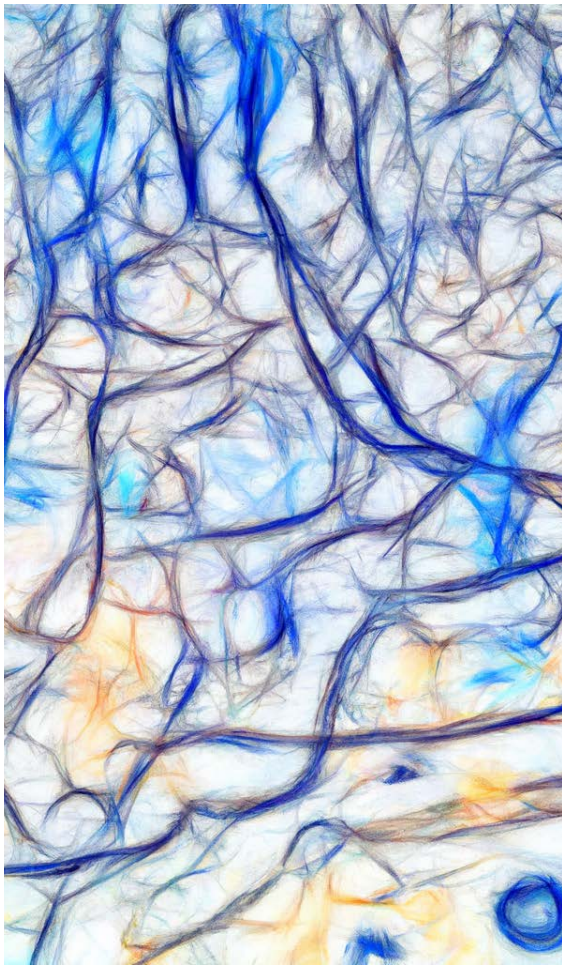
A esto hay que añadir otra de las necesidades que resuelven las estructuras de DL: su aplicación a datos no estructurados, como imágenes o audio. Los modelos clásicos de ML han demostrado su potencial predictivo desde los años 80 en múltiples áreas y aplicaciones. Sin embargo, como se describió en la sección anterior, cuando los datos disponibles no están estructurados, es decir, en un formato tabular, su uso requiere de un preprocesado previo que genere una serie de variables a partir de esa información original no estructurada. Dichas variables se estructuran entonces en un formato tabular que pueda usarse como entrada de los modelos clásicos de ML.

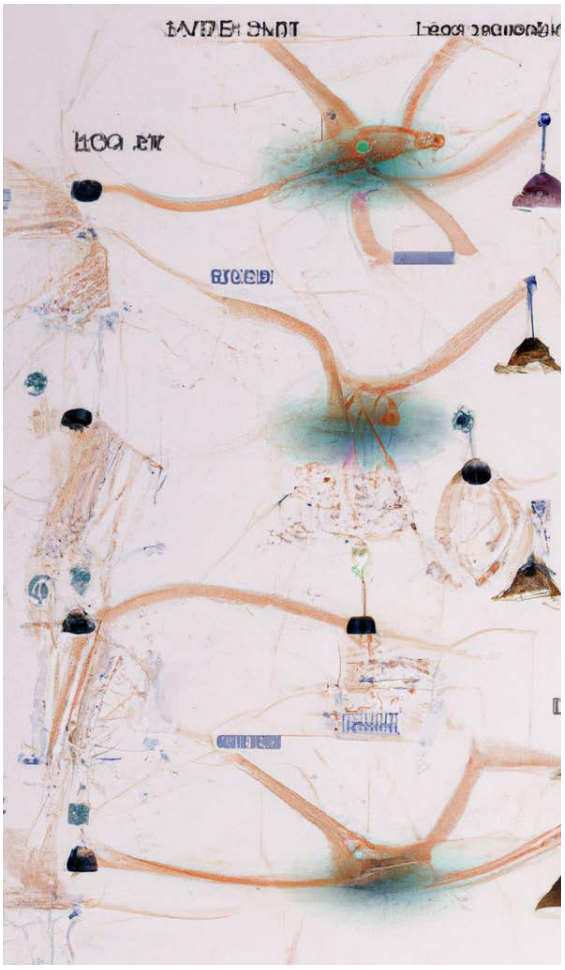
Las estructuras de DL, por el contrario, no requieren de esa fase de preprocesado, ya que son las capas iniciales las que realizarán ese proceso de extracción de patrones relevantes a ser empleados por las capas posteriores. Simplificando, el concepto de profundidad en estas estructuras se basa en el hecho de que la estructura del modelo debe constar de múltiples capas que representen el nivel de abstracción, y cada capa debe adaptarse al entrenamiento del modelo. Las características de los niveles o capas inferiores del modelo, deben combinarse progresivamente para formar características de nivel superior en las capas posteriores.

Como consecuencia de esta aproximación, los modelos DL sobre datos no estructurados no solo permiten un enfoque end-to-end, donde no es necesario el empleo de un preproceso específico de cada área para extraer información en forma de variables estructuradas, sino que han demostrado en múltiples investigaciones proporcionar un mayor potencial predictivo en este tipo de tareas, como el reconocimiento de imágenes. Mientras que en los problemas que emplean datos estructurados, los modelos DL solo son capaces de sobrepasar el potencial de los modelos ML clásicos cuando se proporcionan grandes volúmenes de datos, no siendo útil su empleo cuando las dimensiones no son muy elevadas. Son claramente la opción más recomendable en prácticamente cualquier escenario en el que se empleen datos no estructurados.

TIPOS DE ESTRUCTURA

Cuando se trabaja con datos estructurados, la opción estándar es emplear una Fully Connected Neural Network, FCNN. Sin embargo, existen distintas estructuras de DL especializadas en trabajar con distintos tipos de datos no estructurados, según el problema a resolver:





1 Redes Neuronales Convolucionales, o CNN

Fueron introducidas por primera vez en la década de 1980 por Yann LeCun, un investigador postdoctoral de informática. Estos modelos están especialmente diseñados para un funcionamiento óptimo sobre datos con una estructura espacial, como imágenes. Existen variaciones de estas estructuras para su uso sobre vídeos.

2 Redes Neuronales Recurrentes, o RNN

Las redes neuronales recurrentes se basaron en los trabajos de David Rumelhart en 1986. La famosa arquitectura Long Short-Term Memory, LSTM, que ha sido durante muchos años una referencia en lo relativo a este tipo de estructuras, se inventó en 1997. Estos modelos de DL están especializados en datos temporales, como series numéricas temporales o tareas de procesamiento de lenguaje natural (NLP por sus siglas en inglés), como por ejemplo traducción de un idioma a otro. Pueden emplearse en combinación con una estructura CNN para su empleo sobre vídeos.

3 Transformers

Fueron introducidos en 2017 por un equipo de Google Brain y son cada vez más el modelo de elección para problemas de NLP, sustituyendo a los modelos de RNN. Además, aunque inicialmente no se diseñaron con este objetivo, se ha observado que estas estructuras pueden adaptarse a ser empleadas sobre imágenes, donde han mostrado un gran potencial, incluso sobrepasando los resultados obtenidos por las CNNs, la referencia hasta este momento. Sin embargo, su capacidad de reemplazar a las CNNs como modelo de elección en este tipo de tareas aún requiere de una investigación en más profundidad.

En resumen, la enorme cantidad de datos de que ya se dispone y que se prevé que aumente en el futuro, significa que el Aprendizaje Profundo es la solución óptima para muchas tareas de aprendizaje automático, ya que la mayoría de los demás métodos de aprendizaje automático aprovechan los datos hasta cierto nivel y luego dejan de aprender más, además de sufrir de problemas de escalabilidad en algunos casos, como por ejemplo las Máquinas de Vectores Soporte. En segundo lugar, la potencia de cálculo, especialmente la computación en la nube, proporciona la plataforma para el rápido entrenamiento de modelos de Aprendizaje Profundo. Por último, la investigación o el desarrollo de algoritmos que se está llevando a cabo en el ámbito del Deep Learning es ahora fácilmente comercializable, lo que significa que se invierten más recursos en términos de capital y mano de obra.

FUNCIONAMIENTO

Existen tres tipos de técnicas de DL:

Aprendizaje supervisado

En el aprendizaje supervisado, la red se alimenta de entradas de ejemplo, cada una con su correspondiente etiqueta o target a predecir como salida esperada. Por ejemplo, la entrada podrían ser ciertas características de un paciente cuando ingresa al hospital, y la salida o etiqueta a predecir, sería si dicho paciente finalmente padece una determinada patología. El objetivo principal de este método es obtener un modelo que sea capaz de generalizar la capacidad de predecir cáncer de colon a datos de nuevos pacientes no vistos durante el proceso de entrenamiento. En esta área incluiríamos distintas estructuras DL según el tipo de datos de entrada empleados, como por ejemplo FCNN para datos tabulares, CNN para imágenes, RNN para series temporales, y Transformers para texto en lenguaje natural.

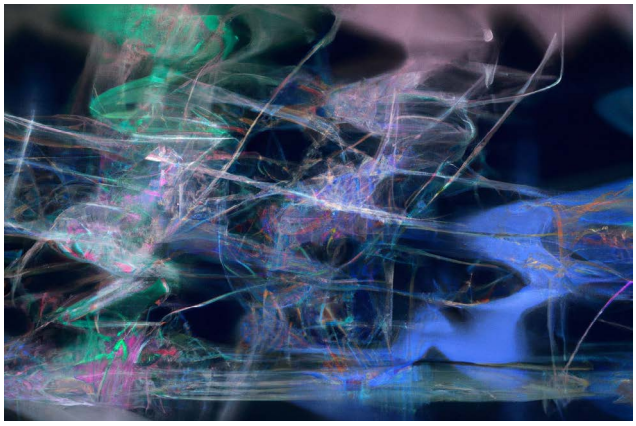
Aprendizaje no supervisado

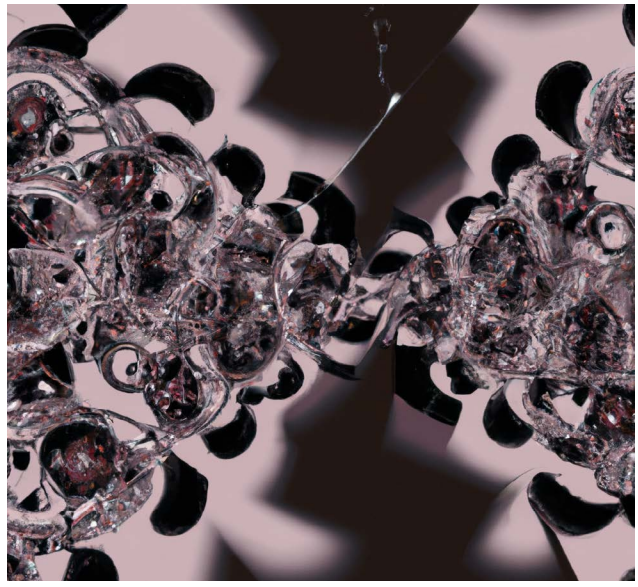
En el aprendizaje no supervisado, la red recibe datos de entrada no etiquetados; es decir, no existe un target o salida a predecir, y por lo tanto, la red tiene que aprender por sí misma los patrones ocultos de los datos de entrada para producir una salida generalizada. En el aprendizaje no supervisado, el modelo o la red tiene como objetivo encontrar patrones o representaciones en los datos de entrada. En el Aprendizaje Profundo, las Máquinas de Boltzmann Restringidas (RBM por sus siglas en inglés), los autoencoders y las Redes Generativas Adversarias (GANs por sus siglas en inglés), son algunos ejemplos de técnicas no supervisadas.

Aprendizaje por refuerzo

En el aprendizaje por refuerzo, el agente alcanza el objetivo interactuando con su entorno. Aunque el agente no tiene un conocimiento exacto del objetivo o salida óptima, cuenta con una función de recompensa cuyo valor depende de las acciones llevadas a cabo. De esta manera, el algoritmo aprenderá a realizar las acciones que maximizan la recompensa y así alcanzar el objetivo. Un ejemplo de aprendizaje por refuerzo sería la creación de un jugador de ajedrez artificial basado en técnicas de ML, donde la función de recompensa será mayor cuando las acciones impliquen la victoria del jugador artificial, pudiendo dar mayores recompensas cuando las victorias sean por mayor margen.

De este modo, los modelos de aprendizaje por refuerzo aprenderán a realizar los movimientos de ajedrez que hayan llevado a victorias, cuánto más claros mejor, en las partidas empleadas para entrenar al modelo. El Aprendizaje Profundo por Refuerzo, DRL por sus siglas en inglés, es la técnica de aprendizaje por refuerzo del DL.





Por motivos de espacio, nos centraremos aquí en la descripción técnica de la versión fully connected de las Redes Neuronales Profundas, la estructura de Aprendizaje Profundo empleada en problemas estándar donde se trabaja con datos tabulares. Sin embargo, como se comentó anteriormente, es importante tener en mente que existen estructuras de Deep Learning especializadas en problemas de reconocimiento de imágenes, como las Redes Neuronales Convolucionales, o CNNs por sus siglas en inglés, series temporales, como las Redes Neuronales Recurrentes, o RNN, y en tareas de Procesamiento del Lenguaje Natural, como los llamados Transformers.

En las FCNN, dada una muestra de entrenamiento y un target o etiqueta a predecir, una ANN calculará todas las funciones de activación, desde la capa de entrada hasta la capa de salida, obteniendo como resultado una predicción final. A esto lo llamamos paso hacia delante o forward pass.

Una vez realizado este forward pass, podemos calcular un error entre su salida y el target o etiqueta real, utilizando la función de error seleccionada. Utilizando la técnica de descenso por gradiente sobre este error, se podrían obtener nuevos valores para los pesos de las unidades de la capa de salida, para intentar mejorar su predicción. Sin embargo, esto solo modificaría los pesos de la capa de salida y no los de todas las capas precedentes, que también tienen un impacto en la salida resultante, y por lo tanto este proceso de optimización tendrá un efecto lejos de ser óptimo.

Por lo tanto, necesitamos un algoritmo para propagar hacia atrás el error de las unidades en la capa de salida a las unidades en las capas anteriores. Este algoritmo se llama backpropagation y se utiliza para optimizar las ANN.

El objetivo del método de backpropagation es poder extender el descenso de gradiente a todas las capas de la red. En backpropagation se define el error generalizado asociado a una unidad oculta como una media ponderada de los errores de las unidades de la capa adyacente. De este modo, mediante backpropagation podemos calcular el error de las unidades de salida, luego el error generalizado de las unidades de la última capa oculta, y sucesivamente, los de todas las capas ocultas anteriores. Esto se denomina paso hacia atrás o backward pass y permite la posibilidad de optimización de las ANN, incluidas sus versiones profundas como las FCNN.

En la actualidad existen modificaciones y mejoras del algoritmo de optimización de descenso por gradiente, como el método de optimización Adam comentado anteriormente, pero la intuición detrás de su funcionamiento es muy similar a la explicada aquí para la estructura clásica basada en descenso por gradiente.

BENEFICIOS EMPRESARIALES

Algunas de las áreas de aplicación de éxito del Deep Learning son las siguientes:

1. **Imágenes:** Reconocimiento facial de imágenes, búsqueda de imágenes, visión artificial, creación de imágenes artificiales.

2. **Texto:** Análisis de texto-sentimiento, búsqueda aumentada, traducción de un idioma a otro, ayuda a la programación, chatbots, generación de textos automáticos.

3. **Series temporales**: Detección de riesgos, predicción meteorológica, análisis económico.

4. **Audio**: Detección de sonido-voz, reconocimiento de hablantes, análisis de sentimientos.

5. **Vídeo**: Detección de vídeo-movimiento, detección de amenazas en tiempo real.

El número de sectores donde el Aprendizaje Profundo ha proporcionado beneficios empresariales crece día a día y se espera que vaya a transformar muchas industrias. En algunas de ellas ya lo ha hecho, así como la sociedad misma, a través de desarrollos en la asistencia sanitaria, la educación y los negocios. El Deep Learning se ha aplicado en campos como la bioinformática, medicina, información espacial y predicción meteorológica, educación, tráfico y transporte, agricultura, robótica y juegos. Y también se ha aplicado ampliamente en redes móviles e inalámbricas, la clasificación del tráfico, la minería de registros detallados de llamadas y la calidad de la experiencia, así como en la asistencia a la programación.

En la actualidad es difícil pensar un área en que no se pueda encontrar una problemática en la que aplicar técnicas de ML, y en concreto, en la que DL no pueda proporcionar claros beneficios empresariales. Dichos beneficios pueden ser de varios tipos:

Ahorro de costes

Las técnicas de DL se pueden aplicar para optimizar procesos reduciendo costes, como por ejemplo encontrar la ruta de transporte óptima para evitar pérdidas monetarias indeseadas, o a través de la creación de modelos que detecten las transacciones fraudulentas en un e-commerce.

Optimización de ingresos

También se pueden aumentar los beneficios de una empresa mediante el uso de DL atacando a la otra rama económica, los ingresos. Existen múltiples ejemplos del uso de DL para aumentar los ingresos de una compañía, que van desde la estimación del precio óptimo de un producto, que será aquel que maximice el producto, multiplicando la probabilidad de compra, dado un precio por el ingreso obtenido al aplicar ese precio, hasta la recomendación de los productos de mayor interés para un potencial cliente. O por ejemplo, mediante webs personalizadas como es el caso de Netflix, o a través de diferentes campañas de marketing, como puede ser el caso de las campañas de descuentos por email llevadas a cabo por algunas aerolíneas.

Mejora del producto/servicio

En las tareas indicadas en los dos puntos anteriores, los beneficios empresariales obtenidos mediante la aplicación de técnicas de DL pueden traducirse de una forma directa en ganancia monetaria. Sin embargo, esto no ocurre con la mayoría de las aplicaciones de DL, donde el objetivo es la mejora del producto o servicio ofrecido a los clientes o usado internamente para la mejora del funcionamiento de una empresa. Estas aplicaciones también tienen un impacto positivo en la empresa, la mayoría llevando también a mayores beneficios, pero en muchos casos la cuantificación de dicha ganancia no es directa o sencilla. Ejemplos de este tipo de usos de los modelos de DL son el reconocimiento facial llevado a cabo para desbloquear un smartphone, la creación de un chatbot asistencial automático en la app y/o web de un e-commerce, o el desarrollo de un servicio que permite traducir textos de un idioma a otro de forma automática en un editor de texto.

Mejora del entorno laboral

En la actualidad muchas empresas están llevando a cabo acciones para mejorar la calidad del entorno laboral y la salud, tanto física como mental, del trabajador. En este campo, los métodos de DL también pueden proporcionar beneficios significativos. Por ejemplo, se pueden entrenar modelos de DL para predecir la probabilidad de un trabajador de poder sufrir burnout o agotamiento laboral en el futuro, permitiendo llevar a cabo acciones preventivas. También puede emplearse para detectar posturas incorrectas y proporcionar alertas automáticas cuando el trabajador lleve demasiado tiempo sin realizar descanso. Más allá de la salud del trabajador, también podrían ofrecerse formaciones personalizadas según los conocimientos e intereses de cada trabajador mediante sistemas de recomendación entrenados mediante modelos de DL.

Seguridad

La seguridad de una empresa, especialmente todo lo relacionado con la ciberseguridad, también puede verse mejorada mediante el empleo de DL. Por ejemplo, algoritmos de reconocimiento facial basados en CNNs pueden emplearse para controlar el acceso y salida de la oficina. En cuanto a la ciberseguridad, las aplicaciones son varias:

DETECCIÓN de virus y **DEFENSA** frente a distintos tipos de ataques cibernéticos.

DETECCIÓN DE ANOMALÍAS en los accesos y peticiones realizados a una e-commerce, que pueden implicar accesos realizados de forma automática con efectos perniciosos para la empresa. Ya sea de forma intencionada, como los llevados a cabo por hackers, o de forma indirecta como el llevado a cabo mediante web scrapping (técnicas que simulan la navegación de un humano con el objetivo de extraer información de sitios web).

Los ejemplos descritos anteriormente no pretenden ser una lista exhaustiva de las posibles aplicaciones que pueden proporcionar beneficios a una empresa, ya que esto es ilimitado. Sin embargo, proporcionan un contexto variado que da una perspectiva global de las posibles aplicaciones de DL en el mundo empresarial. En cualquier caso, no debemos perder de vista que este es solo un subconjunto del panorama completo, y conviene tener en mente que prácticamente en cualquier área empresarial un especialista en técnicas de ML y DL podrá proponer aplicaciones de estas técnicas para la mejora del desempeño y consecución de los objetivos de una empresa.

DESAFÍOS SOCIALES

En las secciones anteriores nos hemos centrado en los múltiples beneficios, tanto puramente empresariales como globales para el conjunto de la sociedad, que puede tener la aplicación de técnicas de DL en distintas áreas. Sin embargo, no debemos olvidar que como toda revolución tecnológica, el auge del ML (incluyendo las técnicas de Aprendizaje Profundo), también conlleva una serie de desafíos sociales. Aunque dichos problemas potenciales no nacen de la tecnología en sí misma, sino del uso humano dado a las mismas, como puede ocurrir con otras tecnologías disruptivas como internet o las redes sociales, conviene tenerlas en cuenta para minimizar su posible impacto negativo.

En primer lugar, tenemos el potencial problema de crear aplicaciones basadas en DL que tomen decisiones con un bias o sesgo. Este sesgo implicaría que nuestro modelo dé predicciones con un ajuste mucho menor para determinados grupos de población. Este efecto puede ser especialmente dañino cuando se trata de aplicaciones críticas, como las empleadas para la ayuda a la decisión clínica en hospitales, o para los grupos de población discriminados. Existen múltiples ejemplos de aplicaciones que han sufrido de esta problemática, siendo famoso el empleo en E.E.U.U. de técnicas de ML en un modelo que estimaba la probabilidad de un delincuente de reincidir, para ayudar a la decisión de si proporcionar libertad condicional o no, y que mostró un claro sesgo negativo no justificado hacia la población afroamericana debido a un error en la elección de los datos de entrenamiento del modelo.

Este ejemplo es bastante interesante porque muestra, no solo que las implicaciones de un problema de sesgo en este tipo de aplicaciones pueden tener un impacto negativo de gran entidad sobre el ciudadano según el área de aplicación, sino por la causa de este tipo de sesgos. Es importante clarificar aquí que el sesgo no proviene de las tecnologías de DL en sí mismas, sino de los datos empleados para entrenar dichos modelos. Estos datos pueden estar sesgados, bien por la forma en la que fueron recolectados y seleccionados, o bien por reflejar sesgos que ya existen en la sociedad, como puede ocurrir por ejemplo cuando se obtienen datos directamente de redes sociales.

El problema del sesgo en los modelos de DL, si bien puede tener un impacto negativo, puede ser detectado, mediante un análisis del funcionamiento del modelo para distintos grupos poblacionales, así como solucionado mediante una correcta selección del dataset de entrenamiento a emplear.

El problema de los modelos con sesgo suele ser un efecto no intencionado. Sin embargo, existen también desafíos sociales relacionados con un uso inadecuado voluntario de estas tecnologías. Cobran especial relevancia aquí el empleo de aplicaciones de DL en las redes sociales, donde estas tecnologías pueden contribuir a aumentar exponencialmente el impacto de dos fenómenos ya existentes. Por un lado, el uso abusivo de este tipo de algoritmos para maximizar el tiempo de permanencia en estas aplicaciones, incluyendo el uso de clickbait o recomendaciones personalizadas para atraer la atención del usuario sin prestar atención al beneficio de este, que puede dar lugar a comportamientos que han sido definidos por expertos como un nuevo tipo de adicción. Por otro lado, el auge de las fake news se ha visto ayudado por el empleo de estas tecnologías y contribuye a una sociedad más desinformada, habiendo sido por ejemplo estudiado su impacto en diversos procesos electorales.

Por último, hemos de mencionar también como un importante desafío social ligado al empleo de técnicas de DL, aunque no exclusivo de las mismas, el uso inapropiado de datos personales. En este aspecto entran en juego aspectos esenciales como el consentimiento del usuario para que sus datos sean empleados, la anonimización de los datos personales, etc. Haremos mención a estos factores en la siguiente sección, dedicada a los aspectos legales y éticos de las técnicas de Aprendizaje Profundo.

ÉTICA Y LEGALIDAD

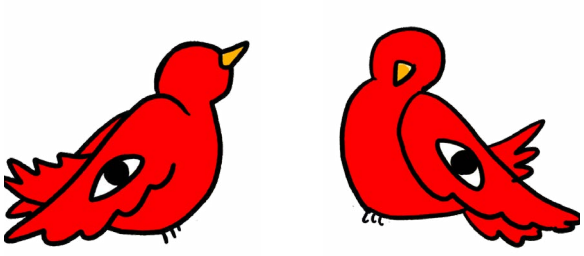
Existen actualmente distintas regulaciones que afectan, directa o indirectamente, al empleo de técnicas de Aprendizaje Profundo. En primer lugar, podemos hacer mención a la General Data Protection Regulation, RGPD, un reglamento de la Unión Europea sobre protección de datos y privacidad en la Unión Europea, UE, y el Espacio Económico Europeo, EEE. El RGPD es un componente importante de la legislación sobre privacidad de la UE y de la legislación sobre derechos humanos. También aborda la transferencia de datos personales fuera de la UE y de las zonas del EEE.

El objetivo principal del RGPD es mejorar el control y los derechos de las personas sobre sus datos personales y simplificar el entorno normativo para las empresas internacionales. Este reglamento, que sustituye a la Directiva 95/46/CE sobre protección de datos, contiene disposiciones y requisitos relacionados con el tratamiento de los datos personales de quienes se encuentran en el EEE, y se aplica a cualquier empresa, independientemente de su ubicación y de la ciudadanía o residencia de las personas afectadas, que procese información personal de personas dentro del EEE.

Es interesante indicar que el GDPR contiene un apartado que afecta de forma directa a las aplicaciones de ML, incluido el uso de técnicas de DL. Este reglamento cuenta con el Considerando 71, que proporciona el derecho a una explicación sobre la toma de decisiones automatizadas de la Directiva de Protección de Datos de 1995.
Esto es:

> "El interesado debe tener derecho a no ser objeto de una decisión, que puede incluir una medida, que evalúe aspectos personales relativos a él, y que se base únicamente en el tratamiento automatizado y produzca efectos jurídicos en él o le afecte significativamente de modo similar, como la denegación automática de una solicitud de crédito en línea o los servicios de contratación en red en los que no medie intervención humana alguna."

> "En cualquier caso, dicho tratamiento debe estar sujeto a las garantías apropiadas, entre las que se debe incluir la información específica al interesado y el derecho a obtener intervención humana, a expresar su punto de vista, a recibir una explicación de la decisión tomada después de tal evaluación y a impugnar la decisión."



"Se impidan, entre otras cosas, efectos discriminatorios en las personas físicas por motivos de raza u origen étnico, opiniones políticas, religión o creencias, afiliación sindical, condición genética o estado de salud u orientación sexual, o que den lugar a medidas que produzcan tal efecto."

Sin embargo, la medida en que estos reglamentos establecen un "derecho a explicación" es objeto de un intenso debate. Por un lado, existen importantes problemas jurídicos, ya que los considerandos no son vinculantes y el derecho a una explicación no se menciona en los artículos vinculantes del texto, ya que se eliminaron durante el proceso legislativo. Además, existen importantes restricciones sobre los tipos de decisiones automatizadas que están cubiertas, que deben basarse "únicamente" en el tratamiento automatizado y tener efectos jurídicos o similares, lo que limita significativamente la gama de sistemas y decisiones automatizados a los que se aplicaría el derecho.

El RGPD se adoptó el 14 de abril de 2016 y entró en vigor a partir del 25 de mayo de 2018. Como el RGPD es un reglamento, no una directiva, es directamente vinculante y aplicable, y ofrece flexibilidad para que los Estados miembros adapten determinados aspectos del reglamento. Este es el caso de España, donde la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales, LOPD-GDD, es una ley orgánica aprobada por las Cortes Generales que tiene por objeto adaptar el derecho interno español al Reglamento General de Protección de Datos. Esta ley orgánica sustituye a la anterior Ley Orgánica 15/1999 de Protección de Datos de Carácter Personal, conocida como LOPD.

Por último, es importante mencionar que el empleo de técnicas de ML y DL en el ámbito sanitario cuenta con su propia regulación, incluida como parte de la Medical Device Regulation, MDR, de la Unión Europea. Según esta regulación, cuando una aplicación de DL realiza una acción sobre los datos distinta del almacenamiento, el archivo, la comunicación o la búsqueda, y el objetivo es el diagnóstico, la prevención, el seguimiento, la predicción, el pronóstico, el tratamiento o el alivio de una enfermedad, dicha aplicación debe considerarse un dispositivo médico y ajustarse a lo establecido en dicha regulación.

Todas estas aplicaciones requerirán de la obtención de un Certificado Europeo, CE, otorgado por organizaciones externas conocidas como organismos notificados. En particular, dicha regulación indica distintos tipos de criticidad, que implicarán tener que pasar por procesos de validación más o menos exigentes.

Sin embargo, es importante recalcar que la ley no exige el certificado CE para los productos a medida, y los productos destinados a la investigación clínica.

BUENAS PRÁCTICAS

Mejora de la atención sanitaria

El uso de modelos de Aprendizaje Profundo para resolver tareas de reconocimiento de imágenes tiene un gran potencial en el mundo sanitario. Por ejemplo, se pueden aplicar modelos de este tipo para el diagnóstico precoz de enfermedades, estimar la prognosis de un paciente, o localizar áreas de interés en distintas pruebas de imagen médica.

Extracción de interpretabilidad

Se suele decir, erróneamente, que los modelos de Aprendizaje Profundo, y en general los modelos de ML, son una "caja negra". Se sabe lo que entra y lo que sale, pero no lo que pasa entre medias. Sin embargo, la extracción de interpretabilidad en este tipo de modelos es una línea de investigación muy popular y ya se han desarrollado algunos algoritmos que han mostrado muy buenos resultados en distintas aplicaciones. Esta extracción de interpretabilidad puede ayudarnos a crear modelos con menos dudas a nivel ético y a generar una mayor confianza por parte de la población general.

MALAS PRÁCTICAS

Uso de datos sesgados

Entrenar modelos con datasets que solo representen a un subconjunto de la población puede dar lugar a modelos con un sesgo indeseado, cuyo rendimiento sea muy inferior para determinados grupos o minorías poblacionales. Diversos estudios han demostrado que, por ejemplo, esto era un problema en múltiples aplicaciones de Aprendizaje Profundo para el mundo sanitario, dado que se habían utilizado conjuntos de datos con una sobrerrepresentación de pacientes caucásicos.

Modelos no equitativos

Ya sea por el empleo de datos sesgados o por otro tipo de problemas en la creación de los modelos de Deep Learning, las predicciones dadas por estos modelos pueden mostrar distintos grados de ajuste para diferentes grupos de poblaciones. Por ejemplo, podemos encontrarnos con modelos con un mayor grado de ajuste, y por tanto más justos en las predicciones dadas para el género masculino con respecto al femenino. Existen técnicas para evitar este tipo de desajustes, y su aplicación debe ser contemplada en el diseño de cualquier modelo de Aprendizaje Profundo.

FUTURO

El DL tiene actualmente un alto grado de penetración en múltiples áreas, desde aplicaciones sanitarias hasta la optimización de ventas en un e-commerce. Las estructuras de Aprendizaje Profundo son actualmente una referencia en ML, cuando se emplean datos no estructurados, como imágenes, mientras que los conocidos como modelos de Boosting, son la referencia en la aplicación de ML a datos tabulares.

En los últimos años, los grandes avances en DL han estado dominados por su aplicación sobre imágenes, en primer lugar, y por aplicaciones de procesamiento de lenguaje natural, tras la aparición de los Transformers en 2017. Este tipo de estructuras ha revolucionado el mundo de las aplicaciones de Aprendizaje Automático en NLP, desde la traducción de texto, la creación de chatbots o la generación automática de texto realista. Aunque no creados inicialmente con ese objetivo, los Transformers han empezado también recientemente a mostrar su utilidad en aplicaciones sobre imágenes, donde las CNNs eran la clara referencia, hasta la irrupción de estas nuevas estructuras.

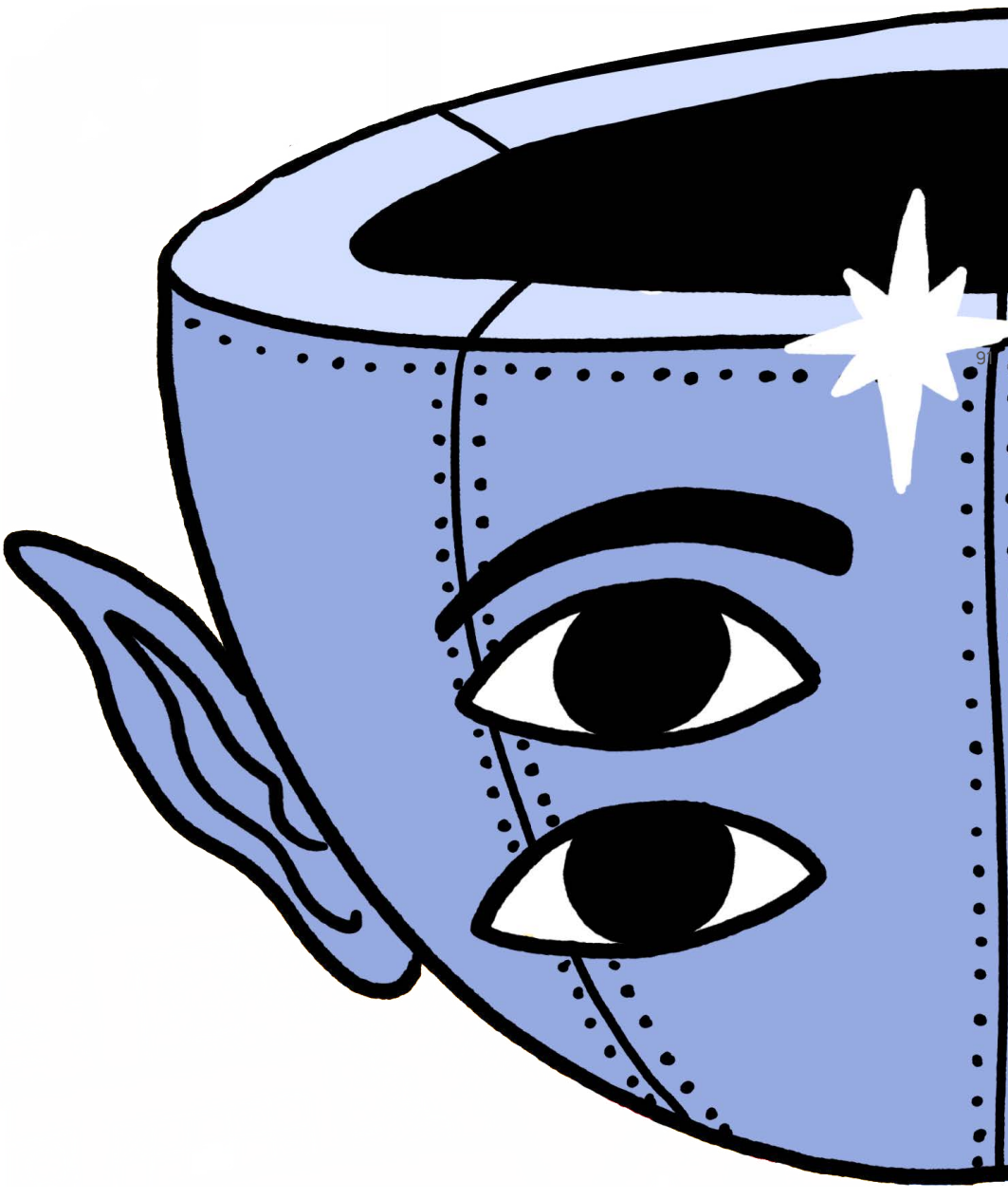
La aparición de estructuras de Transformers de gran complejidad, lideradas por grandes entidades como Google (BERT), OpenAI (GPT-2/GPT-3), Facebook (RoBERTa) y Microsoft (DeBERTa), ha empujado no solo el alcance e impacto de la aplicación de este tipo de tecnologías, sino también su difusión a gran escala; incluso entre segmentos de la población no especializada en el sector. Esto es un factor nada desdeñable, que ha contribuido a la popularidad y desarrollo, aunque quizá también a un cierto efecto hype asociado, de estas tecnologías. La velocidad de aparición de nuevas estructuras, sumada a sus correspondientes campañas de divulgación y publicidad, ha alcanzado un ritmo y ha generado un volumen de noticias inusitado, incluso dentro del campo del Machine Learning, un área que ya destacaba en ambos factores.

Es de esperar que esta tendencia se mantenga, y que la "competición" entre estas grandes entidades tecnológicas contribuya al impulso de las técnicas de DL. Esto sin duda tiene implicaciones positivas muy significativas, ya que hasta ahora el área del Aprendizaje Profundo se ha mantenido en un formato abierto, en el que en la gran mayoría de las nuevas estructuras desarrolladas se hacen públicos los detalles teóricos y en muchos casos se proporcionan librerías que permiten el empleo y entrenamiento de dichos modelos.

Sin embargo, esto último no siempre es así, como el reciente caso de OpenAI con GPT-3, que optó por publicar un artículo científico con detalles teóricos, pero no el algoritmo propiamente dicho, lo cual ha sido criticado afirmándose que va en contra de principios científicos básicos y hace que las afirmaciones de la empresa sean más difíciles de verificar. La justificación de OpenAI para tomar esta decisión es, en primer lugar, que esta aplicación es demasiado peligrosa para ser difundida, debido a su capacidad potencial para generar desinformación o noticias falsas. Por otro lado, también afirmaron que los algoritmos son demasiado grandes y caros de ejecutar.

Más allá de si estas razones son la verdadera causa detrás de la decisión tomada por OpenAI o es una manera de justificar una decisión basada únicamente en objetivos de negocio y monetización, ya que los servicios son ofrecidos a través de una API de pago, este caso particular plantea un interesante debate sobre dos posibles efectos adversos de la reciente evolución del Aprendizaje Profundo. Por un lado, tenemos la cada vez mayor peligrosidad de un uso indebido de este tipo de aplicaciones. Hemos hablado anteriormente de las regulaciones que existen actualmente a nivel europeo para proteger de este uso indebido, pero, además de existir regiones con una menor protección regulatoria, es difícil no tener la impresión de que estas regulaciones no evolucionan a la misma velocidad vertiginosa que las tecnologías.

Las estructuras de Aprendizaje Profundo son actualmente el claro estado del arte en ML cuando se emplean datos no estructurados, como imágenes, mientras que los conocidos como modelos de Boosting, son la referencia en la aplicación de ML a datos tabulares.



Por otro lado, los recientes avances en estas tecnologías se están basando en gran parte en el principio de que "cuanto más grande, mejor". Esto es en el sentido de que, en buena medida, la evolución se está centrando en usar tecnologías similares, pero en mayores estructuras de computación. Lo que permite entrenar modelos con más datos y más unidades de aprendizaje; y esto nos lleva a otro debate: ¿Está el enfoque actual amenazando la "democratización" de la inteligencia artificial (una idea según la cual el acceso a la Inteligencia Artificial debería estar al alcance de cualquiera)?

Esta democratización implica el acceso a la potencia de cálculo, los conjuntos de datos y los propios algoritmos. Los marcos de código abierto facilitan la creación y el intercambio de algoritmos, y existen muchos conjuntos de datos de código abierto. Pero la potencia de cálculo procede del hardware, un recurso físico limitado al que pueden acceder sobre todo grandes empresas y organizaciones bien financiadas. Incluso aunque OpenAI hubiera decidido hacer público su código, solo grandes entidades que tuvieran acceso a enormes capacidades de computación habrían sido capaces de usarlo. Si los experimentos de OpenAI resultan ser el camino a seguir, y algoritmos más grandes se traducen en un mayor rendimiento, entonces el DL de vanguardia se vuelve inaccesible para quienes no pueden permitírsela.

Sin embargo, conviene indicar que no todos los expertos están convencidos de que "el método más grande es mejor" sea lo correcto. Aunque GPT-3 obtuvo buenos resultados en muchas pruebas, se ha descubierto que no era capaz de captar algunos conceptos sencillos que otros algoritmos dominan desde hace décadas, como es el caso de una prueba de "imitación", en la que se pedía al algoritmo que identifique patrones en la forma de cambiar determinadas series de letras. La creación de modelos gigantes que intentan usarse de forma general para cualquier problema relacionado con NLP tiene un gran potencial, pero de momento parece que sigue siendo necesario una especialización de dichos algoritmos para resolver de manera adecuada ciertos problemas concretos.

Si los algoritmos se hicieran públicos, este proceso de especialización de un modelo general a otro específico podría realizarse mediante un conjunto de técnicas llamadas Transfer Learning, sobre un conjunto de datos etiquetados ajustados al problema concreto a resolver, aunque solo por aquellos que dispusieran de la capacidad de computación necesaria. Pero esto se hace imposible cuando solo se puede acceder a la funcionalidad a través de una API.

Por otro lado, es interesante hacer hincapié en una aplicación de DL dentro del NLP: La generación automática de código. Un ejemplo de ello es Copilot, integrado en GitHub. Más allá del acierto conseguido por esta herramienta, la sola posibilidad de crear una aplicación que es capaz de generar código de forma automática podría abrir las puertas a nuevos horizontes teóricos que, aunque lejanos, hasta ahora eran directamente irrealizables. En concreto, estamos hablando del concepto de singularidad, que define un punto en el que un agente inteligente mejorable acabará entrando en un proceso de ciclos de autosuperación, apareciendo cada vez más rápidamente una nueva generación más inteligente.

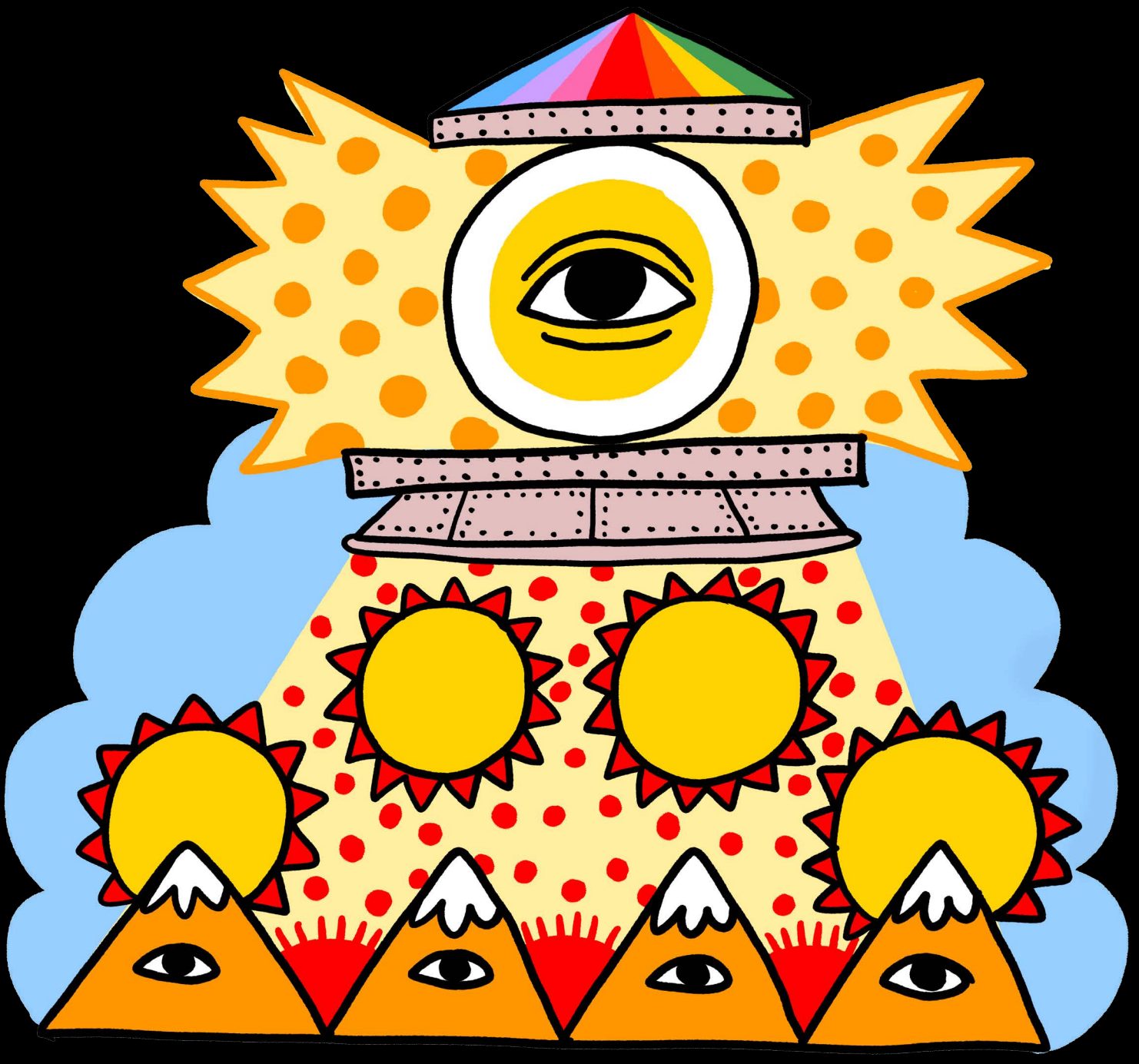
En el momento en el que se crea un algoritmo capaz de programar, se abre la puerta, aunque lejana y solo a nivel teórico, de crear un algoritmo que aprendiera a crear mejores versiones de sí mismo. Aún suena algo muy distante, pero es la primera vez que esta posibilidad se abre a nivel teórico y que incluso ha empezado a ser discutida en algunos artículos académicos.

(V) Data Science

Es cierto que en un mercado tan competitivo, aquellos que puedan entender a las personas y sus necesidades, serán capaces de tomar las mejores decisiones para sus negocios.

Pero, ¿realmente es posible lograrlo todo a partir de los datos?

POR ESPARTACO CAMERO



¿QUÉ ES?

Data Science es un campo muy dinámico, pero si tuviéramos que definirlo de una manera sencilla, diríamos que es el conjunto de metodologías para trabajar datos en cualquier forma que se nos presenten (imágenes, clics en una página web, transacciones, etc.), y tomar decisiones basadas en ellos. Estas decisiones pueden ser para entender el pasado, describir el presente o tratar de predecir el futuro, siempre con la intención de obtener valor de los mismos.

Si bien el Data Science es la intersección de diferentes disciplinas (1), existen dos tipos de Data Scientist: A y B.

A: ANALYST (ANALISTA): Aquellos que buscan conseguir valor e insights en los datos a través de análisis.

B: BUILDING (CONSTRUCTORES): Cuyo foco principal es la construcción de modelos estadísticos para la resolución de problemas.

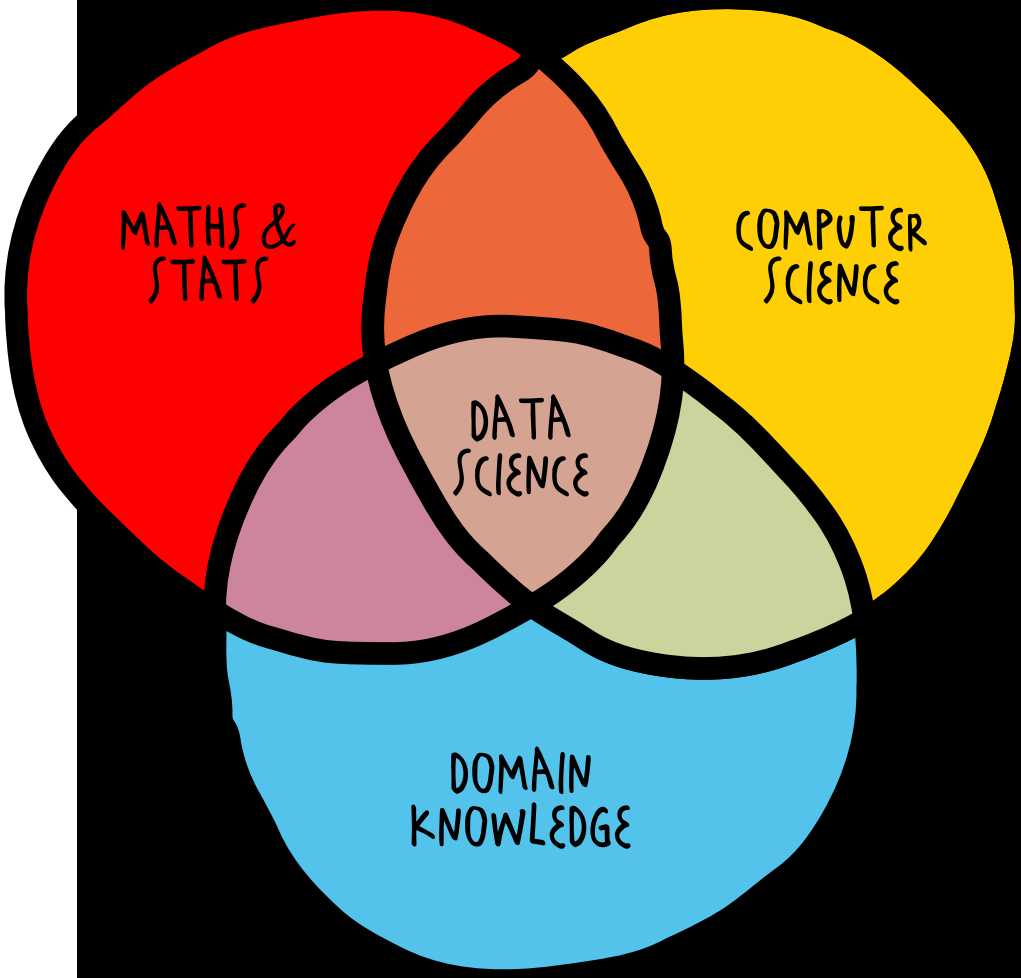
Uno de los principales atractivos del científico de datos es que tiene estas 3 importantes skills:

HABILIDADES TECNOLÓGICAS, ya que deben saber programar y manipular datos usando herramientas como Python, R o SQL.

CONOCIMIENTOS MATEMÁTICOS Y ESTADÍSTICOS, para darle valor a esos datos.

HERRAMIENTAS para la resolución de problemas **DE NEGOCIO**.

No es tarea sencilla conseguir una persona con estas 3 habilidades. Es por ello que el Data Scientist sea un perfil tan demandado, bien remunerado y que, en ocasiones, puede llegar a convertirse en un unicornio para muchos reclutadores.



(1) Data Science como actividad multidisciplinaria.

NACIMIENTO

1962

Aunque el término Data Science empezó a coger auge alrededor de 2012, cuando Tom Davenport y D.J Patil publicaron en la Harvard Business Review: "Data Scientist: The Sexiest Job of the 21st Century", su origen se remonta a 1962 cuando el estadístico John W. Turkey comienza a comentar sobre el futuro de la estadística como ciencia empírica en su libro The Future of Data Analysis.

1974

Luego, en 1974, se presenta por primera vez el término Data Science de la mano de Peter Naur, quien lo definió como "La ciencia de tratar con datos, una vez que se han establecido".

1990

Posteriormente, durante los años 90 y 2000, se empieza a usar en conferencias, revistas y otros campos como la computación, de la mano del Data Mining, siempre asociado a la disciplina de usar los datos para generar conocimiento e información de valor como se conoce hoy en día.

2000

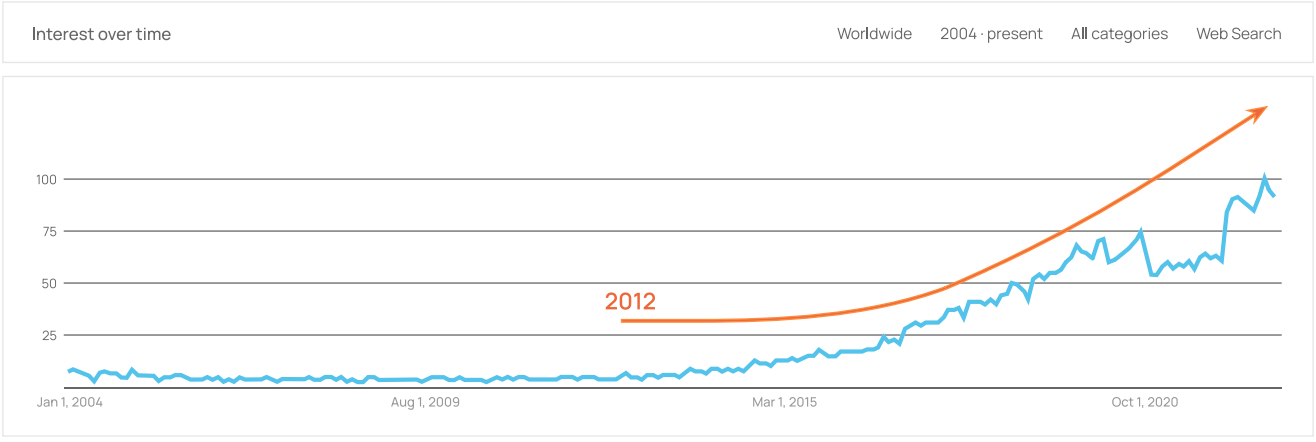
Desde la primera década del año 2000, el área de Data Science también se ha transformado bastante en función de las necesidades y/o cambios de la tecnología. El término venía frecuentemente asociado a lo que se conoce como Big Data, concepto introducido por Doug Laney en el 2001, donde se hablaba de las 3 V: **Velocidad**, **Volumen** y **Variedad**. La primera de ellas hace referencia a lo rápido que se generaban los datos. Fuentes como móviles, internet y redes sociales, generaban una cantidad de datos por segundo nunca antes vista. Esto ocasionó que se tuvieran que almacenar muchos más datos que antes: Volumen.

Y adicionalmente no eran datos de una sola forma, sino de muchos tipos: Variedad (imágenes, música, texto, sensores, entre muchos otros). Y el área que se encargaba de analizarlos era la de la ciencia del dato. De aquí surgieron herramientas de grandes empresas tecnológicas como Google y Yahoo! que permitieron manipular esa gran cantidad de datos de una forma más eficiente, y posteriormente otras empresas se subieron a esta ola del Big Data y Data Science para explotarlos usando los mismos conceptos y herramientas.

Actualidad

Actualmente, disponer de herramientas tecnológicas como la nube ha hecho que la manera de procesar esta cantidad de datos tan grandes sea algo mucho más "trivial" de lo que fue en su momento. A su vez, mucho más económico, por disponer de almacenaje casi infinito de información a precios muy bajos, y cientos de computadoras en la nube con tan solo hacer unos clics.

Esto ha hecho que el foco en Data Science también haya cambiado con el paso del tiempo. Cuando antes el rol principal de un científico de datos era manipular esta cantidad gigantesca de información de una manera eficiente con el uso de esas tecnologías, hoy en día el rol está más orientado a sacar provecho de esos datos con el fin de obtener un resultado que pueda añadir valor.



(2) Búsqueda en Google del término "Data Science" a través de los años.

FUNCIONAMIENTO

Dentro de esta área existe un framework bastante generalizado, aunque con matices, según algunos especialistas, que podríamos resumir en los siguientes pasos:

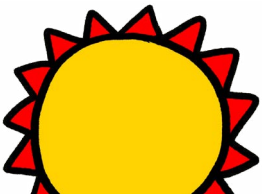
1. Entendimiento del problema/ caso de negocio

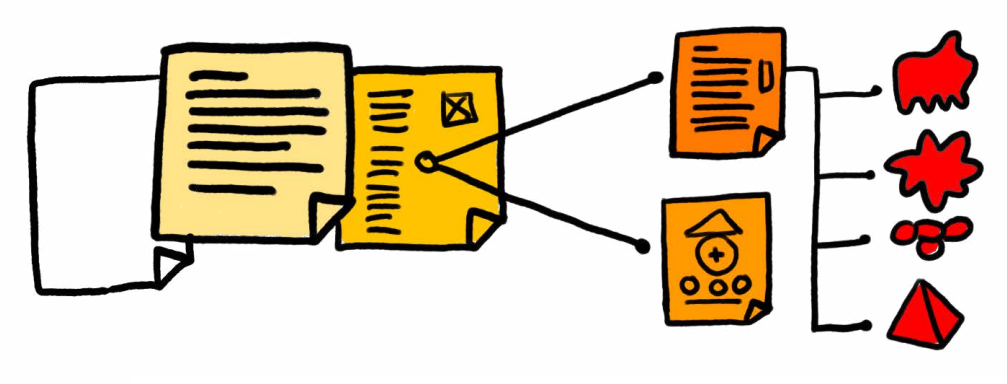
Es la primera fase y la más importante, donde en conjunto con el experto del área (ya sea del área de negocio o quien conoce el problema), se analiza el caso a estudiar y cuáles serán las preguntas a responder, así como el objetivo del estudio. Será el Data Scientist quien ponga sobre la mesa los posibles enfoques que se puedan utilizar de una manera sencilla y práctica, para que puedan dar respuesta al problema en cuestión.

2. Recopilación y limpieza de datos

Sabiendo cuáles son esas variables de interés para el problema, se inicia la extracción de los datos asociados al mismo. Estos datos pueden venir de distintas fuentes como bases de datos relacionales (Data Warehouse, Data Lakes), APIs, ficheros de Excel, Data Marts, etc., y es tarea del Data Scientist recopilarlos, juntarlos y analizarlos.

Previo al análisis, es muy común que los datos vengan con algún formato indeseado, o con valores perdidos (missing data) o atípicos (outliers), entre otros, por lo que se deben limpiar y procesar antes de analizarlos de forma detallada. De todos los pasos, este puede ser uno de los que más tiempo le lleve completar.





3. Análisis Exploratorio de los Datos (EDA)

Cuando tenemos recopilado y limpio nuestro set de datos, se inicia la exploración estadística de las distintas variables (cálculo de medias, medianas, distribuciones, etc.) y la relación que pueda existir entre ellas. Aquí se juega mucho con la parte artística del científico de datos, ya que debe encontrar la forma de visualizarlos de manera gráfica, para una mejor descripción y entendimiento de los mismos, porque como sabemos, una imagen vale más que mil palabras.

A su vez, es en esta etapa donde además de utilizar las variables que extrajimos durante el paso 2, crearemos nuevas variables en función del conocimiento del negocio y del problema en cuestión. Este proceso se conoce como Feature Engineering, y es de los que más valor aporta al siguiente paso del ciclo de un problema de Data Science.

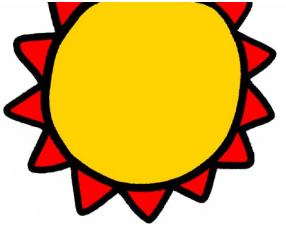
A modo de ejemplo: una empresa quiere saber cuándo será la próxima compra de un cliente. En nuestro dataset tenemos solo fechas de transacciones para cada usuario; una variable que surge del Feature Engineering es la diferencia en días entre la 1ª y 2ª compra de clientes pasados. Una vez que ambos estos datos sean unidos en un mismo gráfico, se podrán explorar sacando conclusiones acordes a las necesidades del negocio.

4. Modelado de datos

Seguramente una de las partes que más entretiene al científico de datos es probar y tunear distintos tipos de algoritmos, según el problema a resolver, y encontrar aquel que tenga la mejor performance. Existen dos approaches, según el problema a resolver: los modelos supervisados y los no supervisados. Si te olvidaste de qué se tratan, ¡te lo recordamos brevemente!

El primero es un tipo de modelo donde conocemos una variable target que queremos predecir, por ejemplo, la fecha de la próxima compra de un cliente (ejemplo anterior) o si una transacción es fraudulenta o no. Esta variable target se conoce como el Ground Truth y puede ser del tipo numérica: ¿Cuántas compras hizo la persona? O del tipo categórica: ¿Compró o no compró? ¿Es una transacción fraudulenta o no?

El otro tipo de approach es el modelo no supervisado, donde no tenemos una variable target, sino que utilizamos los mismos datos para generar información de valor. Por ejemplo, si de un grupo de usuarios tenemos información sobre las transacciones que hacen en un comercio (frecuencia, valor de la compra, antigüedad, etc.), podríamos utilizar esas variables para segmentarlas en clientes recurrentes, de alto valor, esporádicos, etc. Y con ello crear acciones de marketing para impulsar las ventas, hacer que regresen, consentir a los de alto valor, etc.



5. Despliegue, evaluación y monitorización del modelo

Cuando ya sabemos qué algoritmo usaremos, lo normal es poner el modelo en producción, evaluar si está arrojando los resultados esperados y trackear su performance. Estas tareas se denominan MLOps o Machine Learning Operations. Dependiendo del nivel técnico del Data Scientist, puede ser una tarea que realice él solo, o donde necesite ayuda de un Machine Learning Engineer o de un Data Engineer, si por ejemplo, necesitamos hacer predicciones en tiempo real.

Imaginemos un caso de detección de fraude. La predicción no puede esperar a que la transacción ya se haya hecho porque puede ocasionar pérdidas monetarias. Aquí la inferencia se debe hacer justo al momento en que se está realizando la operación y antes de decir si fue efectiva o no, por lo que se debe dar respuesta en cuestión de milisegundos.

6. Comunicación de resultados/ outputs

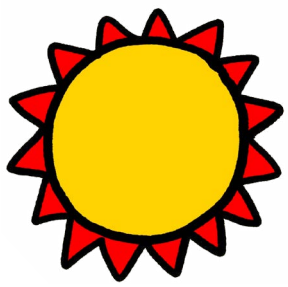
Tenemos una solución matemática al problema en nuestra “blackbox” o algoritmo de Machine Learning, pero debemos traducir esta solución a un lenguaje que sea fácil de transmitir a los stakeholders.

Por ejemplo, si tenemos un modelo que previene el fraude en un 80%, ¿qué representa esto en términos de ahorro a la empresa?, ¿cuánto será el ahorro mensual?, ¿de llegar a entrar fraude, cuánto sería? Entre otras preguntas relacionadas con el negocio o problema inicial.

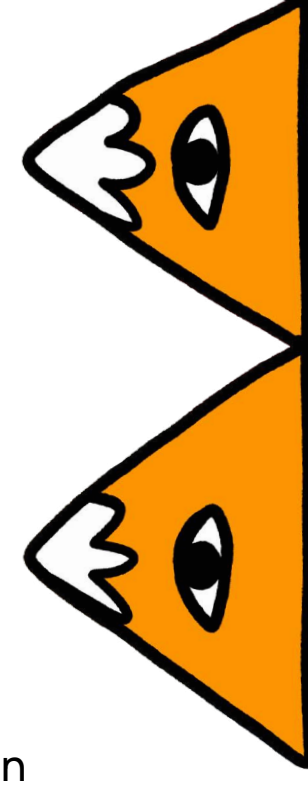
Entre las habilidades de un Data Scientist, hemos mencionado el Domain Knowledge o conocimiento de negocio. Es tarea de él transmitir estas ideas y beneficios del modelo, de una manera sencilla, práctica y que pueda llegar a cualquier tipo de público.

Respecto a este ciclo de Data Science, existen algunas metodologías bastante conocidas y relacionadas con este proceso como lo son CRISP-DM de IBM o TSDP de Microsoft, que como mencionamos anteriormente, pueden tener sus matices en comparación con los procesos o pasos a realizar, pero en general, son bastantes parecidos.

Cabe destacar que estos pasos aplican para el Data Scientist tipo B, quien se encarga de la construcción de modelos. Para el tipo A, simplemente es omitir los pasos 4 y 5. Adicionalmente, un Data Scientist tipo A puede desempeñar otro tipo de tareas, como por ejemplo, realizar A/B test, análisis estadísticos de datos, cálculo de matrices de cohortes, LifeTime Value del cliente (LTV), entre otras métricas de interés.



BENEFICIOS EMPRESARIALES



Son numerosos los beneficios que puede traer el Data Science a una organización y es por esto, que ha generado todo un boom en la industria. Entre algunos de ellos se pueden mencionar:

Segmentación de clientes

Para la correcta aplicación de campañas de marketing y/o retención de los mismos: Por ejemplo, identificar quiénes son mis clientes VIP, los que están más enganchados con mis productos, los que realizan más compras de manera frecuente, etc.

Optimización de precios

En servicios y productos (dynamic pricing): Cuál es el precio que debo asignar a mi producto para sacar el mayor beneficio del mismo, según demanda y oferta.

Creación de modelos de recomendación

Para impulsar la venta cruzada (cross-selling o up-selling): El clásico ejemplo de Amazon de quién compró este artículo, también vio y/o compró estos otros, por tanto se lo recomendaremos a otros clientes con intereses similares.

Detección y prevención de fraude

Por ende, ahorro de dinero.

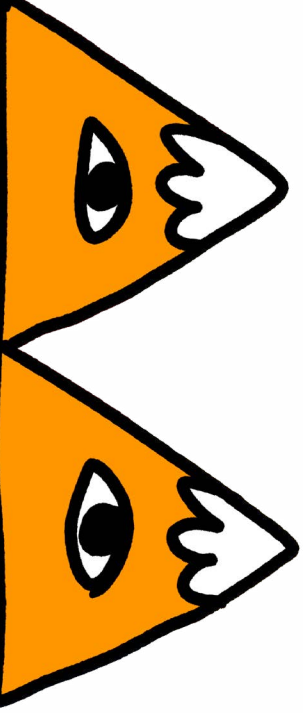
A/B tests

Para mejorar la experiencia de usuarios en páginas web: Dónde y de qué color agregar el botón de “comprar” en mi página web para obtener más clics.

Predicción del Lifetime Value (LTV)

De un cliente, o lo que es lo mismo, cuánto dinero me proporcionará un cliente al contratar un servicio con la empresa.

Entre muchas otras aplicaciones que dependen del sector y empresa que los necesite.



DESAFÍOS SOCIALES

No todo ha sido color de rosa con la llegada del Data Science a la sociedad. La gran potencia que tiene, unida al uso masivo de los datos capturados de manera "inconsciente", casi que a cualquier nivel de nuestras vidas (lo que vemos en internet, el aceptar las condiciones de una app, etc.), ha ocasionado que los gobiernos y entes sociales hayan tenido que tomar acciones para proteger al usuario, a fin de evitar el uso desinformado de sus datos personales.

Un claro ejemplo de esto es la creación en Europa del Reglamento General de Protección de Datos (RGPD o GPDR por sus siglas en inglés), tal como te hemos comentado en la sección de Inteligencia Artificial. El reglamento surgió porque la Unión Europea quería dar más control a los usuarios sobre sus datos para que no sean manipulados por las empresas para uso comercial. De igual forma, otros entes gubernamentales han aplicado leyes similares, como es el caso de California, Estados Unidos, con su CCPA o Ley de Privacidad del Consumidor de California.

Pero el impacto social no ha sido solo a nivel de cómo persistir y usar los datos. También ha sido en la manera en cómo esta ciencia ha desplazado el trabajo manual de operadores por ordenadores que pueden realizar el mismo trabajo de una manera más óptima y a un coste mucho menor. Por ejemplo, una máquina con un algoritmo de Machine Learning es capaz de detectar y analizar patrones de fraude en transacciones de un comercio electrónico mucho más rápido y eficiente de lo que un operador manual lo podría hacer. Además, puede procesar millones de datos de una forma precisa.

Otro ejemplo más reciente y que podría tener impacto a futuro en el uso de árbitros y asistentes en un partido de fútbol, es el uso de Inteligencia Artificial (IA) en el mundial de Fútbol Qatar 2022, donde un algoritmo de Machine Learning desempeñó un papel fundamental para identificar qué jugadores estaban fuera de juego de una manera mucho más precisa que lo que el ojo humano podría.

Entre otros desafíos sociales, está la adaptación de carreras universitarias a estas nuevas y demandadas profesiones asociadas al mundo del Data y en especial del Data Science. Al día de hoy, ya se han creado carreras de pregrado y postgrado a fin de cumplir con las necesidades que el ámbito profesional y social está necesitando.

Además de estos, hay muchos otros desafíos, más los nuevos que vendrán a medida que la tecnología avance y nos enfrentemos a nuevos problemas que la ciencia del dato resolverá de una manera ingeniosa y óptima, con sus ciertas desventajas que la sociedad buscará adaptar o regular, de ser necesario.

Aplicaciones prácticas

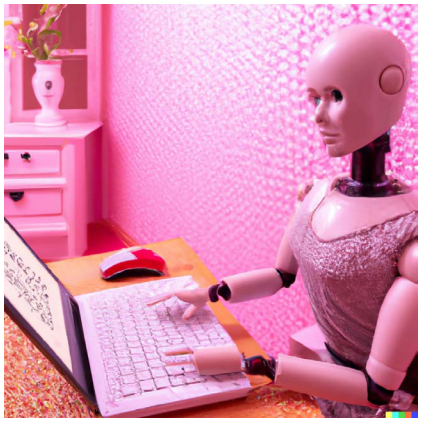
Comencemos con las llamadas malas prácticas:

- Durante 2010, Facebook fue duramente criticado por todo el escándalo de Cambridge Analytica y la manipulación de datos de usuarios sin su consentimiento.
- En 2018, un algoritmo de Amazon asociado a temas de Recursos Humanos, tenía preferencias para escoger hombres sobre mujeres para ciertos puestos de la empresa.
- Un año después, la tarjeta de Apple fue criticada por crear ciertas desventajas al momento de otorgar préstamos a mujeres.
- El sistema de reconocimiento de imágenes de Google, fue clasificado como racista al etiquetar automáticamente a una persona de piel negra como un gorila.

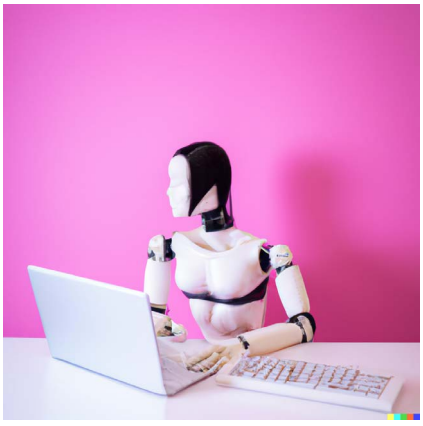
Como estos, hay muchos otros casos. Esto ocurre por la manera de entrenar y aplicar el modelo, que no entiende de contexto y que puede tener un conjunto de datos erróneos durante su entreno. Aparte de estos casos, esta tecnología en las manos equivocadas puede ocasionar actos delictivos como fraude, violación de sistemas de seguridad, irrupción de datos de usuarios, entre otros.

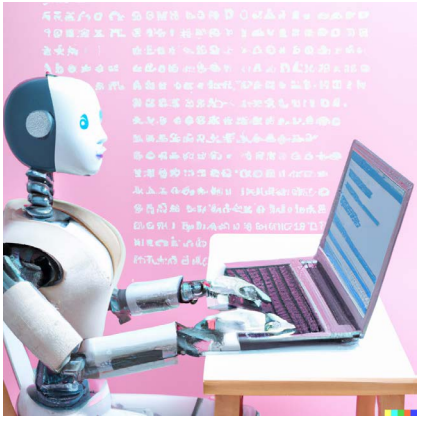
Pero no todo ha sido negativo en temas sociales, al mismo tiempo estas tecnologías han aportado grandes ideas que son adaptadas a gran escala, muchas veces sin saber lo que hay detrás de ellas. Por ejemplo:

- La recomendación de texto cuando escribimos un correo en Google, viene de un modelo que predice la siguiente palabra al escribir.
- La detección de tumores cancerígenos con mayor eficacia que como lo hacen los humanos.
- La identificación de texto en una foto y su traducción automática.
- La detección de ataques cibernéticos, tanto en entes gubernamentales como empresariales.
- El chatbot GPT-3 que simula conversaciones humanas con bastante similitud y que puede revolucionar desde el sistema educativo hasta el empresarial, ya que escribe código por sí solo.



Le dimos a DALL·E estas instrucciones: “Humanoid machine programming on a computer, in a pink room”.
¿Notas algún sesgo en su algoritmo?





FUTURO

Aún queda mucho que descubrir y explotar dentro de esta área y es por eso que siguen habiendo cientos de ofertas laborales, nuevas carreras dedicadas explícitamente a esta ciencia y mucho, pero mucho, que aportar dentro de las distintas unidades de negocio o a la sociedad en sí.

Ya vimos cómo el término Data Science sigue una tendencia creciente a nivel mundial (2) y esto es apenas el comienzo. Porque a medida que la tecnología avance y se puedan resolver aún más problemas usando la ciencia de datos, la proliferación de esta carrera irá a más.

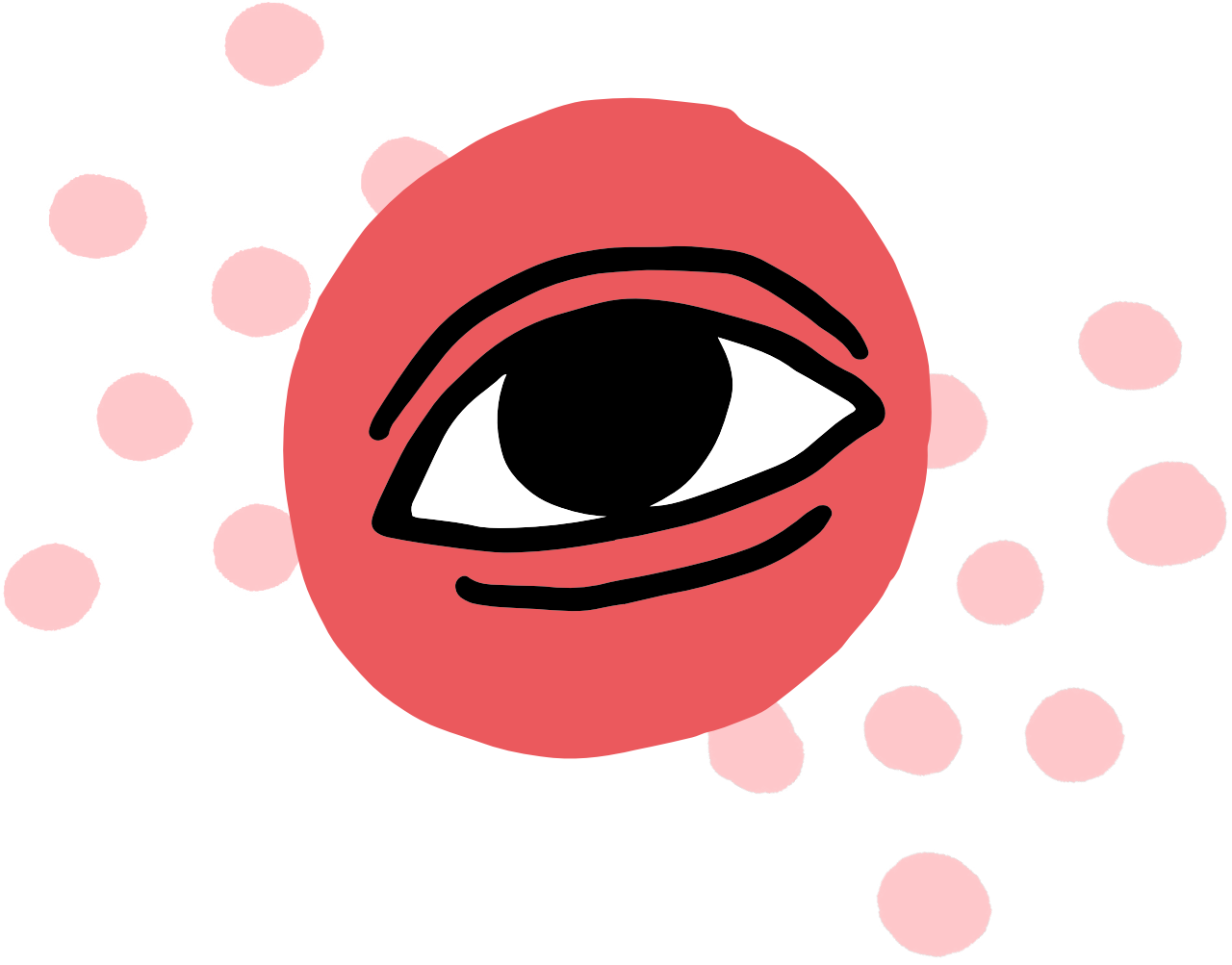
Si hubiera que decantarse por cómo será el Data Science a corto plazo, diría que se seguirán abriendo posiciones en empresas, tanto tradicionales como startups. Y que la demanda del Data Scientist tipo A será inmensa porque ataca el primer problema básico: añadir valor con insights de los datos, y que luego se dará valor añadido con lo que pueda resolver y construir el Data Scientist tipo B con sus modelos de Machine Learning.

A mediano plazo, diría que iremos un paso más allá del ML, y es aquí donde el Deep Learning será el que tenga mayor impacto a nivel social y empresarial, por lo que se espera un mayor foco en esta área de la Inteligencia Artificial.

A largo plazo, sería importante que el Data Science no solo sea un área donde algunos se especializan, sino que sea parte básica de la formación de cualquier persona. Ya que los datos serán las respuestas a muchas de nuestras preguntas.

Por último, siendo los datos el nuevo petróleo, existe cada vez más la necesidad de disponer de talento que no solo se enfoque en el uso de datos, sino que sepa tratarlos y manipularlos de una manera eficiente. Y es aquí donde los Data Engineers jugarán un papel muy importante para poder habilitar esa información, los Data Analyst sacarán insights de ella y entenderán el pasado de esos datos, y los Data Scientists la utilizarán para predecir el futuro.

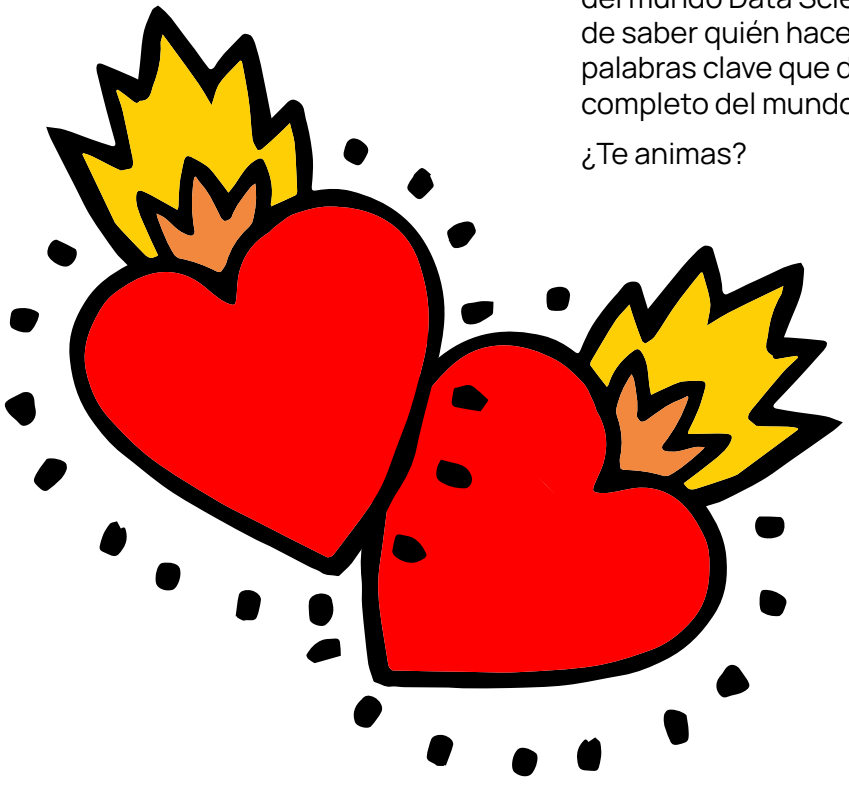
“Los datos son el nuevo petróleo”, existe cada vez más la necesidad de disponer de talento que no solo se enfoque en el uso de datos, sino que sepa tratarlos y manipularlos de una manera eficiente.



AHORA SÍ ¡HAZ MATCH!

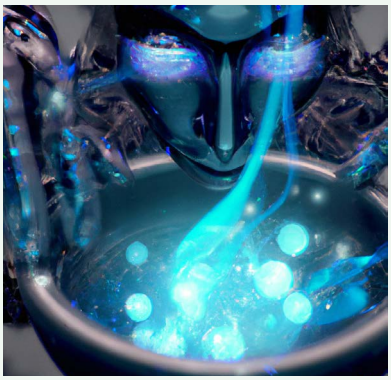
Ya has tenido una cita con los conocimientos infaltables del mundo Data Science. Pero... ¿y ahora qué? Es momento de saber quién hace qué en esta relación y cuáles son las palabras clave que debes entender para enamorarte por completo del mundo de los datos.

¿Te animas?



Si te interesa el mundo de los datos, esta será una gran noticia para ti: hay un inmenso abanico de roles que se pueden desempeñar, dependiendo de las habilidades que se desarrollen y la especialización profesional que se desee elegir.
¡Te contaremos algunas!

EL EQUIPO SOÑADO



Data Analyst

"Tortura a los datos
y lo confesarán todo"

- Ronald Coase

Funciones:

Da valor al negocio en función de los insights, tendencias y patrones que consigue en los datos, así como el reporte de los mismos para seguimiento y control de las diferentes unidades de negocio, buscando darle sentido y explicaciones.
Un analista de datos responde a preguntas sobre el presente como: ¿Qué está pasando ahora? ¿Cuáles son las causas? ¿Puede mostrarme XYZ? ¿Qué debemos hacer para evitar/conseguir ABC? ¿Cuál es la tendencia de las ventas en los últimos 5 años?

El trabajo de un analista de datos incluye 3 partes principales: Entender el problema de las métricas del negocio, es decir, hacer las preguntas correctas; averiguar las respuestas o conocimientos a partir de los datos; y saber comunicar. Esto incluye la creación de cuadros de mando con las visualizaciones adecuadas y la explicación de los mismos de una manera fácil de entender para las partes interesadas no tecnológicas o "Business".

Herramientas:

- **SQL**: Es la "base" de un analista de datos y entonces esencial para comunicarse con la base de datos empresarial.
- **VISUALIZACIÓN DE DATOS**: La mayoría de las empresas tienen licencias de herramientas de Business Intelligence como Power BI, Tableau, Looker, Qlik, etc.
- **CONOCIMIENTO DEL DOMINIO**: ¿Qué significan las métricas? ¿Cómo interactúan entre sí? ¿Qué es lo que mueve la aguja?

Data Scientist

"Resolver grandes problemas es más fácil que resolver pequeños problemas"

- Sergey Brin

Funciones:

En lugar de responder a preguntas sobre el presente, tratan de encontrar patrones en los datos y responder a las preguntas sobre el futuro, es decir, la predicción.

Esta técnica existe desde hace mucho tiempo y seguro has oído hablar de ella: se llama estadística. El aprendizaje automático (ML) y el aprendizaje profundo (DL) son las dos formas más populares de utilizar el poder de los ordenadores para encontrar patrones en los datos.

Los científicos de datos también construyen productos basados en esas predicciones. Por ejemplo, un sistema de recomendación predice lo que te gusta, un sistema de clasificación predice el orden de popularidad, la NLP predice lo que significa una frase. Los científicos de datos construyen estos productos, no para ayudar a tomar decisiones empresariales, sino para resolver sus problemas.

Herramientas:

- **SQL**: Es esencial en este tipo de roles, para interactuar con las bases de datos.
- **COMUNICACIÓN**: La investigación debe transmitirse de forma eficaz, tanto al público técnico como al no técnico.
- **ESTADÍSTICAS/MATEMÁTICAS**: Hay que dominar los conocimientos de estadística, como las teorías que hay detrás de cada método de aprendizaje automático, para resolver problemas más complejos.
- **HABILIDADES DE PROGRAMACIÓN**: Actualmente, Python y R son los lenguajes de programación más populares.
- **DESARROLLO DE SOFTWARE**: El flujo de trabajo de Git, CI/CD, DevOps, etc. son básicos en el arsenal de un científico de datos.

Data Engineer

"No he fracasado, sino que he encontrado 1.000 maneras de no fabricar una bombilla"

- Thomas Edison

Funciones:

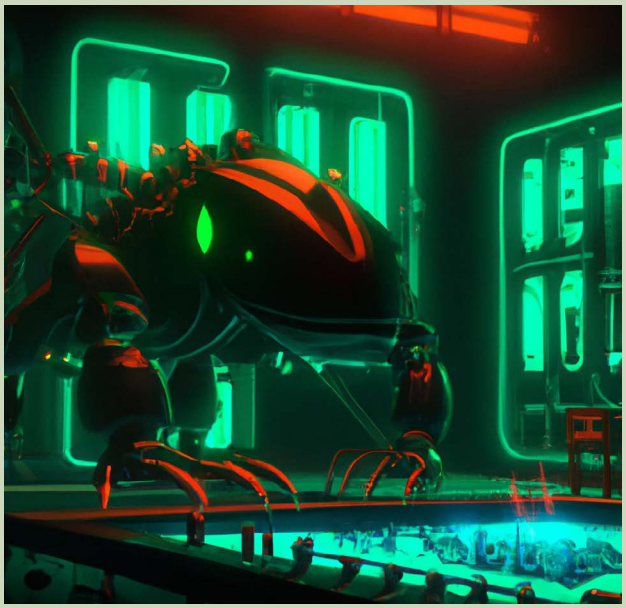
Es la persona encargada de desarrollar, implementar y mantener sistemas que generan datos (a menudo desordenados y en formatos distintos) y producir con ellos información de alta calidad, consistente y fácil de analizar para otras unidades de negocios.

Los consumidores de datos no pueden realizar su trabajo sin que los ingenieros de datos configuren toda la estructura. Por decirlo de forma sencilla, los ingenieros de datos se encargan de todo lo que ocurre con los datos antes de llegar a la base de datos:

- Asegurarse de que el conducto de datos, el almacenamiento y la estructura estén optimizados y sean lo más rentables posible para la empresa.
- Asegurarse de que los datos que utilizan los analistas y científicos son los más actualizados, validados y responsables.

Herramientas:

- **SQL**: Debe conocer los entresijos de cada una de las diferentes bases de datos, cuándo utilizar cada una, cuáles son sus aristas.
- **COMPUTACIÓN EN LA NUBE**: AWS (Amazon), Azure (Microsoft) y GCP (Google) son los tres servicios en la nube más populares del mercado. Esto también incluye la aplicación de la computación paralela (Hadoop, Spark) y el big data.
- **DESARROLLO DE SOFTWARE**: Lo mismo que lo anterior, en el caso de Científico de Datos.
- Gran conocimiento sobre el funcionamiento de los **DATA LAKES**.
- Extract Transform Load (**ETL**).



> “Los ordenadores superarán a los humanos.
> Cuando eso ocurra, tenemos que asegurarnos de que tengan objetivos alineados a los nuestros”
>
> - Stephen Hawking

Machine Learning Engineer

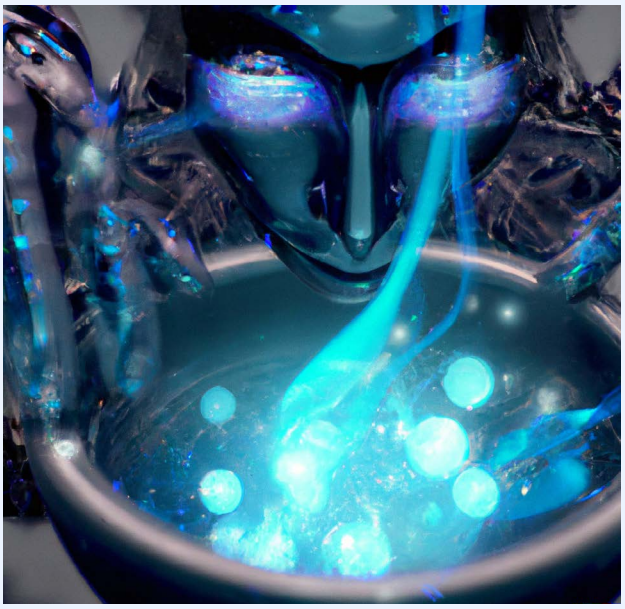
Funciones:

Es la persona que pone en producción los modelos desarrollados por el Data Scientist, de una manera óptima, escalable y mantenible. Son especialistas en mejorar los modelos, encontrando el que mejor rendimiento da. Tiene un perfil más de Software Engineer, pero conoce bien el ciclo de vida de los modelos y su operación (término conocido como MLOps).

La principal diferencia con otros perfiles es en qué se centran. Los ingenieros de aprendizaje automático se centran exclusivamente en este. Su objetivo es crear componentes de software o productos que puedan trabajar con un mínimo de supervisión humana y que ayuden a obtener información de los datos que se les proporcionan. Por eso, un ingeniero de aprendizaje automático se centra tanto en los fundamentos de la informática como en el desarrollo de software, ya que esa es su especialización.

Herramientas:

- Lenguajes de programación: **PYTHON, JAVA** y/o **C++**.
- **SQL**: Esto es esencial para todos los roles relacionados con los datos.
- **CONOCIMIENTOS DE INGENIERÍA DE SOFTWARE:** Escritura de algoritmos que puedan buscar, ordenar y optimizar, comprensión de las estructuras de datos y conocimiento de la arquitectura de los ordenadores.
- **CONOCIMIENTOS PROFUNDOS DE APRENDIZAJE PROFUNDO**: Programación dinámica, arquitecturas de redes neuronales, procesamiento del lenguaje natural, procesamiento de audio y vídeo, aprendizaje de refuerzo, técnicas avanzadas de procesamiento de señales y optimización de algoritmos de aprendizaje automático.
- **DESARROLLO DE SOFTWARE**: El flujo de trabajo de Git, CI/CD, DevOps, etc. son básicos en el arsenal de un científico de datos.



> “La función de un buen software es hacer que lo complejo aparente ser simple”
>
> - Grady Booch

MLOps Engineer

Funciones:

Profesional especialista en el despliegue de los modelos de ML a producción.

MLOps son las siglas de Machine Learning Operations, una extensión de la metodología DevOps que tiene como objetivo incluir los procesos de aprendizaje automático y ciencia de datos en la cadena de desarrollo y operaciones, de forma que el Machine Learning resulte más productivo y confiable.

Esta área es la encargada de toda la operatividad de un modelo de Machine Learning que está en producción, asegurando la continuidad y desarrollo del mismo de manera continua.

En un mundo en el que los datos tienen un nivel tan alto de importancia, los modelos MLOps surgen con la intención de facilitar y agilizar los proyectos de Machine Learning e Inteligencia Artificial dentro de una empresa. Gracias a estos modelos entrenados, se ha conseguido una mayor optimización de procesos.

Los tiempos han cambiado y el flujo de datos que necesita procesar una empresa a día de hoy, requiere de herramientas que permiten hacerlo de forma automática.

Herramientas:

- **MLFLOW**: Plataforma de código abierto para administrar el ciclo de vida completo del aprendizaje automático.
- **KUBERNETES**: Plataforma portable y extensible de código abierto para administrar cargas de trabajo y servicios.
- **DOCKER**: Es un sistema operativo (o runtime) para contenedores.
- **AIRFLOW**: Es una plataforma para crear, programar y monitorear flujos de trabajo mediante código.

Deep Learning Expert

"Los últimos avances ya han dado lugar a inventos que antes vivían en el reino de la ciencia ficción, y solo hemos arañado la superficie de lo que es posible"

- Jeff Bezos

Funciones:

El perfil de un buen experto en Deep Learning es una combinación de tres conjuntos de habilidades.

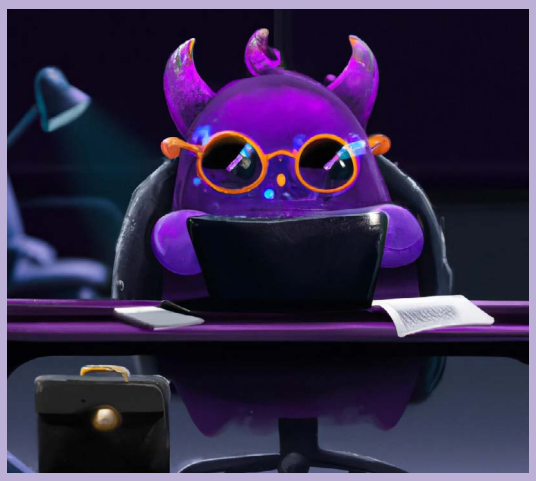
Por un lado, las matemáticas, para entender el funcionamiento de los modelos se requieren ciertas bases de teoría matemática, con especial énfasis en álgebra, estadística y teoría de optimización.

Además, la informática, para poder implementar los modelos de Deep Learning son necesarios altos niveles de programación. También son recomendables ciertos conocimientos de manejo de servidores y arquitectura de datos.

Y por último, conocimiento de negocios. La teoría de Deep Learning no es útil si no se es capaz de traducirla a cómo resolver problemáticas reales de la empresa. Por ello, es recomendable que el experto en estas técnicas tenga ciertos conocimientos de negocio, para ser capaz de "traducir" los problemas existentes, a soluciones de Deep Learning.

Herramientas:

- **PROGRAMACIÓN:** Python y R. Dentro de Python, las librerías más ampliamente utilizadas en este campo son Keras, Tensorflow y Pytorch.
- **SISTEMAS OPERATIVOS**: Las distribuciones Linux son las más utilizadas en los proyectos de Deep Learning.
- **ARQUITECTURA**: Amazon Web Services, Azure y Google Cloud para computación en la nube. Kubernetes y MLflow para la automatización del despliegue, el escalado y la gestión de aplicaciones.



"En Dios confiamos. Todos los demás deben traer datos"

- Edwards Deming

Business Analyst

Funciones:

Profesional capaz de extraer datos y realizar análisis para satisfacer peticiones de negocio, traducirlas, interpretarlas y calcular KPIs.

Es quien amplía la productividad de un proceso empresarial. También actúa como vínculo entre la dirección de la empresa y el equipo informático.

Herramientas:

- SQL
- Excel
- Python

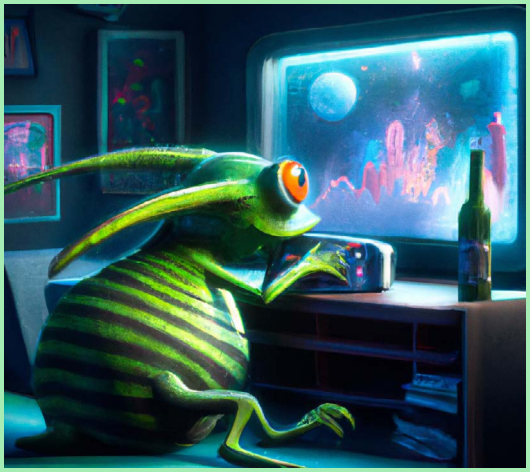
Visualization Tool Developer

Funciones:

Profesional especialista en la generación de dash-boards para la visualización de valores y métricas re-levantes para el negocio. Tiene conocimientos sobre Structured Query Language (SQL) y su arquitectura.

Herramientas:

- Tableau
- PowerBI
- Looker



"La información solo es útil cuando es comprendida"

- Muriel Cooper

"Todo ser humano puede ser, si se lo propone, escultor de su propio cerebro"

- Ramón y Cajal

Deep Learning Engineer

Funciones:

Profesional especialista en el desarrollo de modelos de Deep Learning con redes neuronales. Tiene gran entendimiento sobre la arquitectura del sistema de las redes neuronales. Son responsables del desarrollo de modelos para reconocimiento de imágenes, reconocimiento de la voz o procesamiento del lenguaje natural.

Herramientas:

- Python
- Tensorflow
- Pytorch

Analytics Engineer

Funciones:

Profesional especialista en realizar queries (peticiones precisas para obtener información en una base de datos o sistema de información) en Data Lakes, como también del mantenimiento de estas. Se trata de un rol nuevo y deben tener conocimientos de la arquitectura del Data Lake, como funcionan las ETLs y saber traducir las peticiones de negocio a queries.

Herramientas:

- Snowflake
- Bigquery
- Redshift
- DBT

"Nunca inviertas en un negocio que no puedes entender"

- Warren Buffett

“Dos cosas dan igualdad en la vida: Internet y la educación”

- John T. Chambers

Cloud Engineer

Funciones:

Se trata de un perfil polivalente que se dedica a identificar e integrar servicios y soluciones de computación en la nube, con el objetivo de ayudar a las organizaciones a funcionar con mayor eficiencia, seguridad y atención al detalle.

Herramientas:

- CompTIA A+.
- Systems Security Certified Practitioner (SSCP).
- AWS Certified Solutions Architect Associate.
- Plataformas Cloud: Amazon Web Services, Google Cloud o Microsoft Azure.
- Componentes de redes comunes (firewall, router, switch).
- Lenguajes de programación (Python, Java, Go, R).
- Sistemas operativos (Linux, UNIX, Windows, macOS).
- Protocolos TCP/IP y comunes (DNS, HTTP).

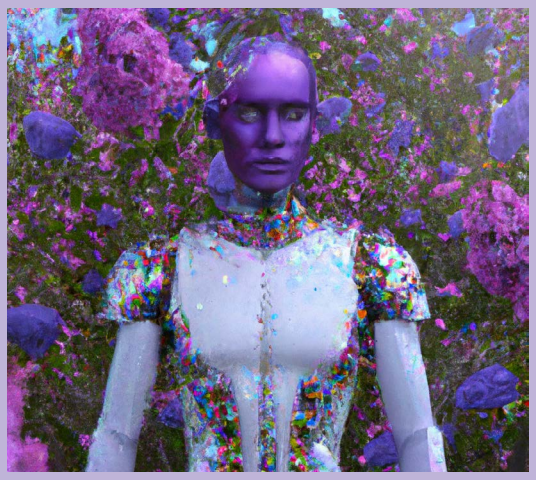
Big Data Architect

Funciones:

Entiende y domina toda la infraestructura de las bases de datos. Trabaja con los Data Engineers para optimizar el rendimiento de los workflows de datos y también debe probar y diseñar nuevos prototipos de bases de datos para satisfacer las necesidades de la empresa.

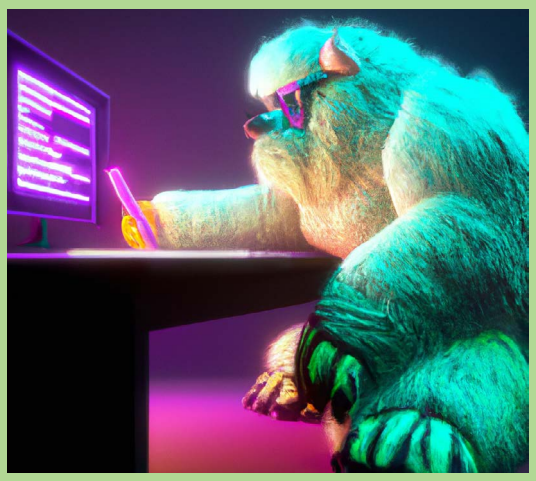
Herramientas:

- Matemáticas, estadística y técnicas de análisis avanzado
- Python y R
- Structured Query Language (SQL) y NoSQL
- Hadoop
- Apache Spark
- Sistemas Cloud
- Looker



“Los arquitectos no inventan nada, solo transforman la realidad”

- Álvaro Siza



"La acción es la clave fundamental de todo éxito"

- Pablo Picasso

Database Manager

Funciones:

Dirige todo el equipo de "database" y es responsable de las bases de datos de la empresa. Debe supervisar el presupuesto y las necesidades de personal, y procesar las solicitudes de datos de la empresa. Su función es revisar el uso de los datos y evaluar estas fuentes para su optimización. Por tanto, es un experto con capacidad de liderazgo y gestión.

Herramientas:

- Power BI
- Tableau Public
- Panoply
- Excel
- Looker

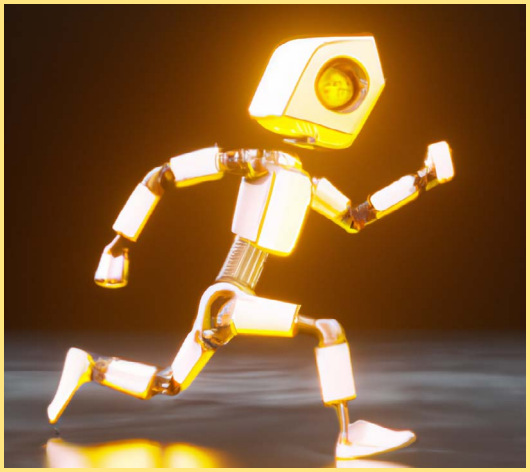
Data Warehouse Manager

Funciones:

Tiene una función similar al del gestor de la base de datos. También tiene que desarrollar e implementar nuevas estrategias de gestión de datos. Su misión es coordinar todas las soluciones de gestión de la información. Tiene que realizar tareas de sourcing, migración de datos, diseño e implementación.

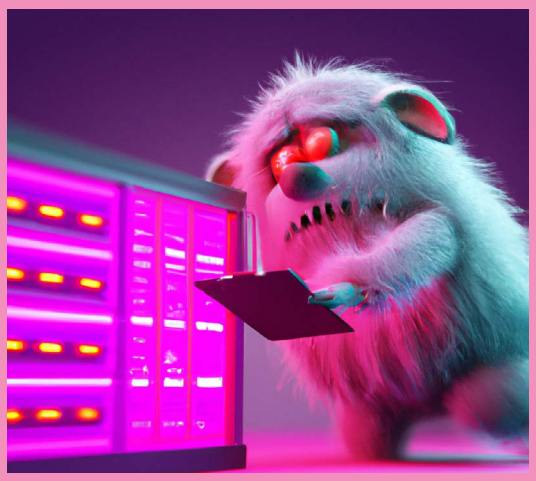
Herramientas:

- Digital Dashboards o paneles de control digital
- OLAP (Procesamiento analítico en línea, por sus siglas en inglés): HOLAP, ROLAP y MOLAP
- Aplicaciones de informes
- Minería de datos



"Si quieres cambiar el futuro, empieza a vivir como si ya estuvieras allí"

- Lynn Conway



"La tecnología es importante. Pero lo único que realmente importa es qué hacemos con ella"

- Muhammad Yunus

Chief Data Officer

Funciones:

El Chief Data Officer es el responsable de todos los equipos especializados en Big Data de la organización. Su función es la de liderar y gestionar datos y analítica asociados con el negocio y asegurarse de que la empresa sea data-driven. Es decir, es el encargado de la explotación de los activos de datos para crear valor de negocio.

Herramientas:

- Apache Hadoop
- Spark
- Scala
- Python
- PySpark
- TensorFlow

Data Artist

Funciones:

Son científicos de datos que también son expertos en el procesamiento gráfico y textual de flujos de datos grandes y complejos. Analizan las fuentes de datos y, tras sacar sus propias conclusiones de ellas, las preparan para la comunicación y gestión corporativa interna o externa.

Herramientas:

- Visme
- Tableau
- Infogram
- Datapine
- Google Charts

"Lo más revolucionario que una persona puede hacer, es decir siempre en voz alta lo que realmente está ocurriendo"

- Rosa Luxemburgo

¿Crees en el amor a primera vista?

Te desafiamos a que elijas el perfil profesional con el que te hayas sentido más identificado y luego le pidas a una persona de tu confianza, que te diga si cree que podrías dedicarte a eso.

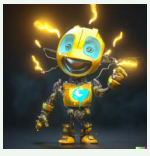
¡Ten cuidado! Quizás en el fondo te pareces a uno de los avatares...

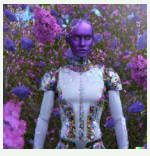
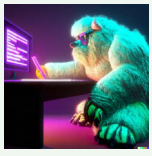
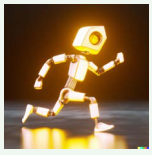






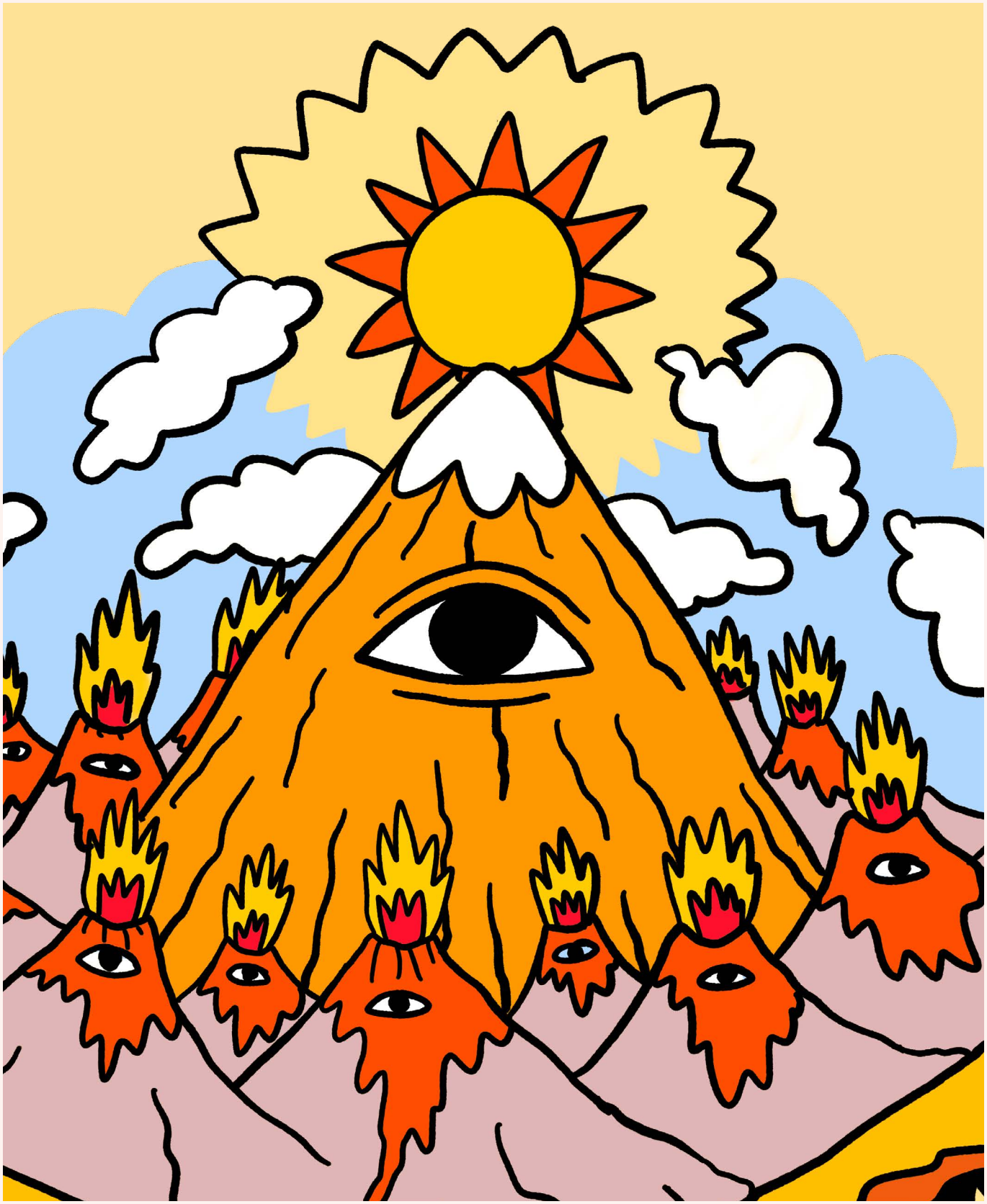




El Máster en Data Science de Nuclio Digital School proporciona a los alumnos una gran variedad de competencias ajustadas a la realidad del mercado laboral, que les permiten mejorar su posición o buscar nuevas oportunidades en el ámbito del Big Data y Analytics.

LEARN [TO BE] THE FUTURE



Máster en Data Science

Data Scientist es uno de los perfiles con más potencial del presente y será una de las posiciones más demandadas del futuro.

Desde Nuclio Digital School hemos diseñado un Máster en Data Science, teniendo en cuenta la necesidad de los alumnos de contar con una experiencia práctica completa de todas las etapas del proceso de análisis de datos.

El programa aborda temas desde la adquisición e integración de datos hasta la productividad de modelos matemáticos basados en técnicas de Inteligencia Artificial, Machine Learning, y el aprendizaje de programación en Python y SQL. Además, proporciona a los alumnos una gran variedad de competencias ajustadas a la realidad del mercado laboral, que les permiten mejorar su posición actual o buscar nuevas oportunidades en el ámbito de los datos.

Gracias a nuestro máster con metodología bootcamp, en solo 5 meses y a través de la modalidad Learning by doing, quien realmente lo quiera, podrá convertirse en un Data Scientist de éxito.

¿Qué sucede si alguien sin conocimientos previos quiere hacer un máster de este nivel?

No sería un problema, porque hemos diseñado un pre-curso que hace la vía de aprendizaje más fácil para aquellos que no han tenido un gran acercamiento con la programación.

¿Cuánto tiempo de dedicación supone?

Buscamos adaptarnos al estilo de vida de nuestra comunidad mediante programas part-time, con un enfoque 100% hands-on y personalizado.

¿Cuál es la salida laboral?

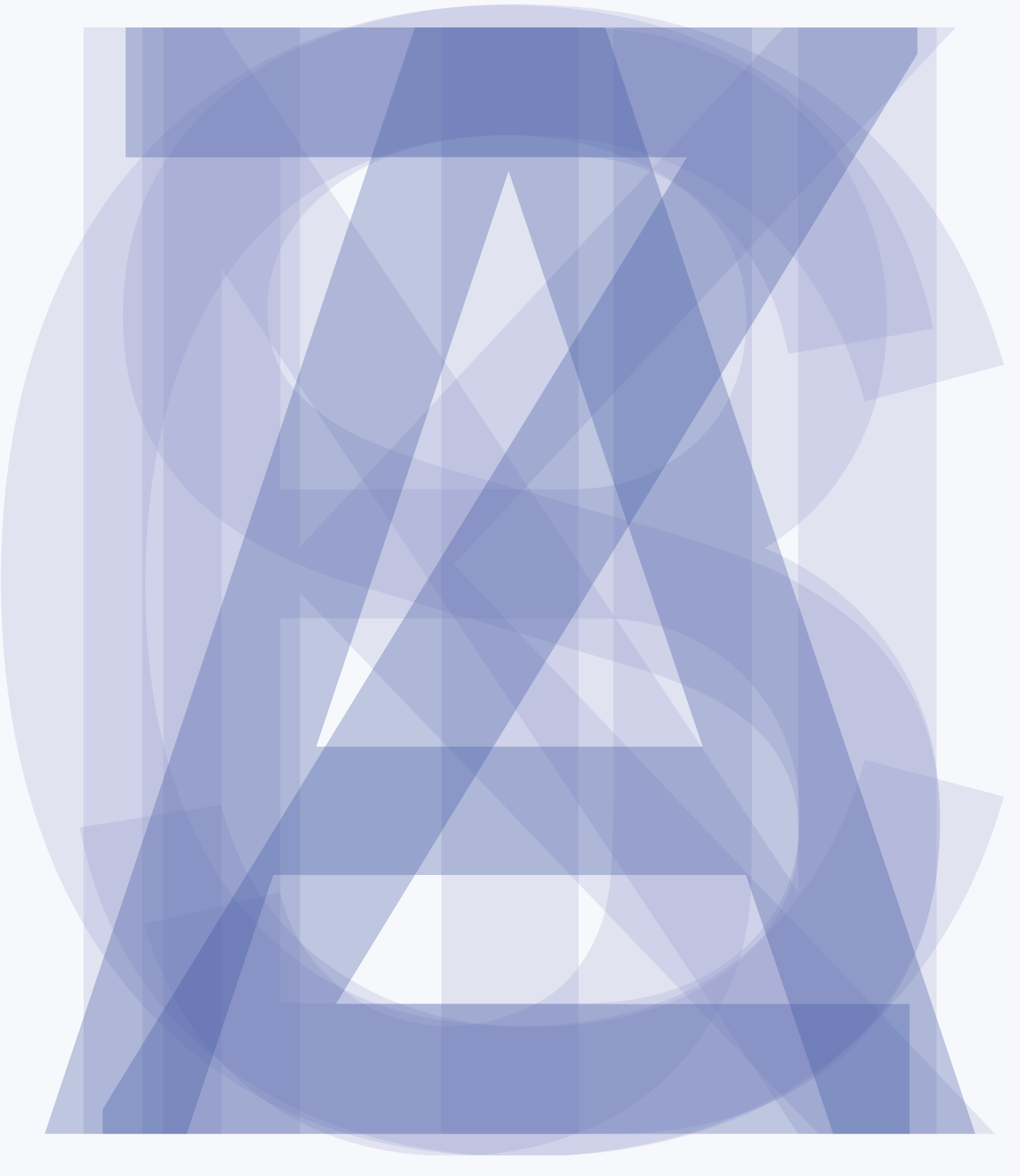
Tras finalizar el Máster en Data Science, se cuenta con los conocimientos necesarios para optar por puestos como Data Analyst, Data Scientist, Business Intelligence Analyst o Data Translator.

Según nuestras predicciones, esta disciplina te interesa más de lo que pensabas. ¿Es así?

Te invitamos a que contactes con nosotros y descubras todo lo que Data Science le puede dar a tu vida, antes de tomar la gran decisión.

Si estás aquí es porque ya te has sumergido en el mundo de los datos, pero tienes dudas sobre algún concepto. ¡Felicidades! Este es el camino correcto para el éxito. Vuelve sobre tus pasos cuando sea necesario y reafirma la información para avanzar certeramente.

GLOSARIO



Aa

A/B Test: Experimento aleatorio con el fin de testear diferentes versiones de un mismo contenido, logrando hacer una comparativa. El resultado será encontrar la versión que sea más eficiente.

Algoritmo: Conjunto de instrucciones o reglas definidas y no-ambiguas, ordenadas y finitas, que permiten solucionar un problema, realizar un cómputo, procesar datos y llevar a cabo otras tareas o actividades.

Análisis Predictivo: Utiliza los datos para determinar qué puede pasar en el futuro y descubrir relaciones entre los datos que normalmente no son detectadas con un análisis menos sofisticado.

Aprendizaje Supervisado: Técnica que utiliza un conjunto de datos etiquetados o conjunto de muestras (train), para entrenar un modelo de Machine Learning. ¿El objetivo? Predecir la etiqueta que tendrán nuevas muestras (test), que el modelo no ha visto en su entrenamiento.

Autoencoders: Son un tipo de arquitectura de redes neuronales que pertenece al grupo de métodos de Aprendizaje No Supervisados. Esta arquitectura extrae las características más importantes del input, eliminando el resto de poca relevancia.

AWS (Amazon Web Services): Es una plataforma en la nube, que cuenta con un conjunto de productos como aplicaciones de informática, almacenamiento, bases de datos, IoT (Internet of Things) y análisis.

Bb

Black Box: En ciencia, informática e ingeniería, una caja negra es un sistema que se puede ver en términos de sus entradas y salidas, sin ningún conocimiento de su funcionamiento interno.

Business Analytics: Permite conseguir los objetivos empresariales, a partir del análisis de datos. Utilizando los modelos predictivos para detectar tendencias, realizar pronósticos y optimizar los procesos del negocio.

Business Intelligence: Conjunto de estrategias con las que se analizan todos los datos que puede manejar un negocio, de forma inteligente. Se trata de trabajar a partir de la información que los datos aportan y aprovecharla para mejorar las estrategias empresariales.

Cc

Centroide: En el universo del Machine Learning, un centroide es la ubicación real o imaginaria que representa el centro del grupo.

CI/CD: Prácticas combinadas de integración continua y entrega continua o despliegue continuo. ¿No lo has entendido? Estos servicios permiten que el equipo encargado del desarrollo de software satisfaga los requisitos de las empresas y brinde mayor atención al código y la seguridad del mismo.

Clúster de Servidores: Un servidor en clúster es la unión de varios sistemas informáticos (servidores) que funcionan como si fueran uno solo, con el objetivo de ofrecer velocidad y alta disponibilidad ante fallos.

Dd

Data-driven: Cuando una empresa toma decisiones estratégicas basadas en análisis e interpretación de datos. Este enfoque permite que examine y organice sus datos con el fin de atender mejor a sus clientes.

Data Lake: Es el lago de datos en el que el profesional se baña para conseguir todas las respuestas a las preguntas que ofrece el Big Data. Es el almacenamiento de toda la información recogida en bruto y que trabaja con una arquitectura plana.

Data Mart: Es una versión específica del almacén de datos, centrados en un tema o un área de negocio dentro de una organización. Son subconjuntos de datos con el propósito de ayudar a que un área específica pueda tomar mejores decisiones.

Data Mining: Es el proceso para descubrir patrones útiles o conocimientos, a partir de fuentes tales como bases de datos, textos, imágenes, internet, etc. Los patrones deben ser válidos, potencialmente útiles y entendibles.

Data Sourcing: Es el conjunto de fuentes utilizadas para obtener información, generalmente de una base de datos. Se sirve de una serie de conectores hacia diferentes medios, canales o soportes cruzados, para extraer datos y proporcionar información relevante.

Data Warehouse: Repositorio de datos desde distintas fuentes, de una manera eficiente y útil, a fin de que sea utilizado para responder a preguntas de negocio y ayudar a la toma de decisiones.

Datos Etiquetados: Son los datos para los que el Aprendizaje Supervisado ya conoce la respuesta de destino.

Datos No Estructurados: Son aquellos que carecen de una estructura o arquitectura identificable. Esto significa que no se ajustan a un modelo de datos predefinidos, son más cualitativos y conllevan una administración más complicada (menos convencional).

Datos Relacionales: Una base de datos relacional almacena y proporciona acceso a puntos de datos relacionados entre sí. Son una forma intuitiva y directa de representar datos en tablas.

Data Set: El término hace referencia a una única base de datos de origen. Representa un conjunto completo de datos, incluyendo las tablas que los contienen, ordenan y restringen, así como las relaciones entre ellas.

Datos Tabulares: Son elementos representados por marcadores diferentes, donde cada campo de marcas representa la presencia o ausencia de un elemento específico. Cada registro representa un conjunto completo de elementos asociados.

Deep Fake: Es un vídeo en el que se muestran imágenes falsas, habitualmente del rostro de una persona, que parecen ser reales y que se han producido utilizando inteligencia artificial.

Deep Learning (DL): Marco de modelos de aprendizaje automático que consiste en un esquema de entrenamiento que contiene varias capas de optimización. Comúnmente está asociado a las ANN con un gran número de capas ocultas.

DevOps: Conjunto de prácticas que combina el desarrollo de software y las operaciones informáticas. Promueve un mejor desarrollo de aplicaciones en menos tiempo y la rápida publicación de nuevas o revisadas funciones de software y productos.

Discounted Cumulative Gain (DCG): La ganancia acumulada descontada es una medida de la calidad del ranking. En la recuperación de información, a menudo se usa para medir la eficacia de los algoritmos de los motores de búsqueda web o aplicaciones relacionadas.

Ee

Exabytes: Es una unidad de medida de almacenamiento de datos cuyo símbolo es el EB. Equivale a 1018 bytes.

Extract, Transform and Load (ETL): Es un tipo de integración de datos que hace referencia a los tres pasos (extraer, transformar, cargar) que se utilizan para mezclar datos de múltiples fuentes. Sirven, a menudo, para construir un almacén de datos.

Ff

Fully Connected Neural Network (FCNN): Consta de una serie de capas totalmente conectadas que unen cada neurona de una capa con cada neurona de la otra capa. La principal ventaja es que son "independientes de la estructura", es decir, no es necesario hacer suposiciones especiales sobre la entrada.

Funciones de Activación: Se utiliza en Deep Learning y significa que en la salida de la neurona, puede existir, un filtro, función limitadora o umbral, que modifica el valor resultado o impone un límite que se debe sobrepasar para poder proseguir a otra neurona.

Funciones de Pérdida No Diferenciables: Las máquinas aprenden mediante una función de pérdida. Es un método para evaluar qué tan bien un algoritmo específico modela los datos otorgados.

Gg

Git: Sistema de control de versiones distribuido y de código abierto. Rastrea los cambios en cualquier conjunto de archivos. Es un clon local del proyecto, un repositorio de control de versiones completo.

Google Cloud Platform (GCP): Es un conjunto de servicios de computación en la nube que se ejecuta en la misma infraestructura que Google utiliza internamente para sus productos.

GPT-3 (Generative Pre-Trained Transformer): Modelo de lenguaje autorregresivo, de Deep Learning, utilizado para producir textos que simulan la redacción humana, basándose en una entrada de datos recibidos (input).

GPU: Es un procesador formado por muchos núcleos más pequeños y especializados. Al trabajar conjuntamente, los núcleos ofrecen un rendimiento masivo cuando una tarea de procesamiento se puede dividir y es procesada por muchos núcleos.

Hh

Hadoop: Apache Hadoop es una colección de software de código abierto que facilita el uso de una red de muchos ordenadores para resolver problemas que implican cantidades masivas de datos y cálculos.

Ii

Imputación de Datos: Es la sustitución de valores no informados en una observación, por otros, para poder continuar con el análisis. Existen diversas técnicas para hacerlo.

Inteligencia Artificial (IA): ¿Una que todos sabemos? Básicamente es la combinación de algoritmos, con el propósito de crear máquinas que muestren las mismas capacidades que el ser humano.

Interfaz de Programación de Aplicaciones (APIs): Es un conjunto de protocolos que permite desarrollar y comunicar aplicaciones de software entre ellas. Por ejemplo, un motor de búsqueda que permite integrar en una sola app, datos de distintas páginas web.

Kk

Kaggle: Kaggle es una comunidad en línea de científicos de datos y profesionales de la Inteligencia Artificial y el Aprendizaje Automático. Esta plataforma hace que el entorno sea competitivo mediante el otorgamiento de premios y rankings para ganadores y participantes.

Ll

Ley Moore: Esta predicción expresa que aproximadamente cada 2 años se duplica el número de transistores en un microprocesador. Teniendo en cuenta que cuantos más transistores o componentes haya en un dispositivo, el coste por dispositivo se reduce, mientras que el rendimiento por dispositivo aumenta.

Librerías de Python: Proveen soluciones estandarizadas para los diversos problemas que pueden ocurrir en el día a día en la programación. Responden al conjunto de implementaciones que permiten codificar este lenguaje, con el objetivo de crear una interfaz independiente.

Librerías Open Source: Son librerías donde se pueden encontrar códigos diseñados de manera que sean accesibles al público: todos pueden verlos, modificarlos y distribuirlos de la forma que consideren conveniente.

Mm

Máquinas Boltzmann Restringidas (RBM por sus siglas en inglés): Es una red neuronal artificial de dos capas (capa de entrada y capa oculta) que aprende una distribución de probabilidad basada en un conjunto de entradas. Ayuda a resolver diferentes problemas basados en combinaciones.

Máquinas de Vectores Soporte o Apoyo: Son una técnica de Machine Learning que encuentra la mejor separación posible entre clases. Con dos dimensiones es más simple entender lo que está haciendo, ya que normalmente los problemas de aprendizaje automático tienen muchísimas dimensiones.

Metadatos: Son la mínima información indispensable para identificar un recurso, como puede ser un archivo en el ordenador o una información extra sobre un tipo de dato.

Metodología Kanban: Se trata de un método visual de gestión de proyectos que permite a los equipos visualizar sus flujos y carga de trabajo, a través de un tablero organizado por columnas.

Microprocesador: Es la unidad de procesamiento principal de un ordenador, su "cerebro".

Microsoft Azure: Es una plataforma de pago por uso que integra servicios completos en la nube pública para que desarrolladores y equipos de TI administren e implementen aplicaciones y otros recursos, a través de un gran centro de datos mundial.

Migración de datos: Es el proceso de mover datos de una ubicación, formato o aplicación, a otra. Normalmente, esto es resultado de la introducción de un nuevo sistema o ubicación.

Missing Data: Término usado cuando la data trae con ella datos perdidos o no registrados durante la recopilación de los mismos. Se debe a fallas en la recopilación o momentos ocasionales en que no se registró, entre otras.

MLOps: Área encargada de toda la operatividad de un modelo de Machine Learning que está en producción, asegurando la continuidad y desarrollo del mismo, de manera continua.

Modelo: Representación matemática de las relaciones en un conjunto de datos. Es una forma simplificada y matemáticamente formalizada de aproximarse a la realidad y hacer predicciones.

Nn

Natural Language Processing (NLP): Es una tecnología de aprendizaje automático que brinda a las computadoras la capacidad de interpretar, manipular y comprender el lenguaje humano.

Neuronas: Unidades de computación de las ANN (Redes Neuronales Artificiales). La salida de cada neurona se calcula mediante alguna función no lineal, llamada función de activación, aplicada a la suma ponderada de sus entradas.

Nodo: Es un punto de intersección, conexión o unión de varios elementos que confluyen en el mismo lugar. Dentro de la informática, puede referirse a conceptos diferentes según el ámbito particular.

Nube: Se conoce como computación en la nube (cloud computing en inglés) o simplemente «la nube», al uso de una red de servidores remotos conectados a internet para almacenar, administrar y procesar datos, servidores, bases de datos, redes y software.

Oo

Open Data: Práctica que dispone de unos tipos de datos de forma libre para todo el mundo, sin restricciones de derecho de autor, patentes u otros mecanismos. Su objetivo es que estos datos puedan ser consultados, redistribuidos y reutilizados libremente.

Outliers: Se dice que un dato es un outlier o dato aislado, cuando se encuentra fuera de lo que sería la distribución normal. Estadísticamente, se diría que esto sucede si se encuentra muy distanciado del resto de datos.

Pp

Precisión: Es una métrica para determinar la calidad del modelo de Machine Learning en tareas de clasificación. Por ejemplo: ¿qué porcentaje de los clientes que contactemos estarán interesados?

Procesamiento del Lenguaje Natural (PLN): Hace posible la compresión y procesamiento asistidos por ordenador de información expresada en lenguaje humano, o lo que es lo mismo, hace posible la comunicación entre personas y máquinas.

Python: Lenguaje de programación de alto nivel y propósito general. El lenguaje que básicamente todos deberían conocer para dar el primer paso en el mundo de la programación. ¿Has oído hablar de él? Ten cuidado, ¡te puedes enamorar!

Rr

R: Es un entorno y lenguaje de programación con un enfoque al cálculo y análisis estadístico. R nació como una reimplementación de software libre del lenguaje S, añadiendo soporte para ámbito estático.

Raíz del Error Cuadrático Medio (RMSE): Es una medida de uso frecuente de las diferencias entre los valores (valores de muestra o de población) predichos por un modelo o un estimador y los valores observados.

Razonamiento Computarizado: Es trasladar el sistema de pensamiento que utilizaría un científico informático a la resolución de un problema: aplicar procesos de pensamiento lógico, sistémico, algorítmico, para lograr representar las soluciones a un problema como secuencias de instrucciones y algoritmos.

Recall: Esta métrica informa sobre la cantidad que un modelo de Machine Learning es capaz de identificar. Por ejemplo: ¿qué porcentaje de los clientes que están interesados, somos capaces de identificar?

Redes Generativas Adversariales (GANs por sus siglas en inglés): Son una nueva forma de utilizar Deep Learning para generar, por ejemplo, imágenes que parezcan reales, música y predicciones futuras, entre otros.

Redes Neuronales Artificiales (ANN): Familia de modelos de Aprendizaje Automático formados por un conjunto de unidades conectadas llamadas neuronas. Pueden utilizarse tanto para tareas de clasificación como de regresión.

Redes Peer to Peer: Son redes de ordenadores en las que todos o algunos aspectos funcionan sin clientes ni servidores fijos, pero sí con una serie de nodos que se comportan como iguales entre sí. Las redes P2P permiten el intercambio directo de información, en cualquier formato, entre los ordenadores interconectados.

Regresión: Problemas de aprendizaje supervisado en los que las etiquetas o target son numéricas, es decir, valores continuos, e indican un valor asociado a cada muestra.

Ss

Spark: Apache Spark es un motor de análisis unificado de código abierto para el procesamiento de datos que permite a los programadores realizar operaciones sobre un gran volumen de datos en clústeres de forma rápida y con tolerancia a fallos.

Structured Query Language (SQL): Lenguaje específico de programación, diseñado para gestionar bases y flujos de datos. Estos son algunos de los programas que usan SQL: Oracle, MySQL, Microsoft SQL Server, Access, Ingres, etc.

Tt

Target: Variable objetivo que se quiere predecir mediante el empleo de técnicas de ML. Básicamente lo que todos conocemos por "target", pero aplicado a las IT.

TensorFlow: Biblioteca de código abierto desarrollada por Google para llevar a cabo proyectos de Machine Learning.

Test de Turing: Es un experimento en el que un humano mantiene una conversación con una computadora y otra persona, pero sin saber quién de los dos conversadores es realmente una máquina. El objetivo es determinar si la inteligencia artificial puede imitar las respuestas humanas.

Transistor: Es el dispositivo electrónico semiconductor que permite el paso de una señal en respuesta a otra.

Uu

User Experience (UX): Conjunto de factores y elementos que intervienen en la vivencia de una persona con una empresa, lo cual ayuda a determinar cómo se sienten con respecto a la marca desde el primer hasta el último contacto.

Vv

Valor-F (F1-score): Se utiliza para combinar las medidas de precisión y recall en un solo valor. Esto es práctico porque hace más fácil el poder comparar el rendimiento combinado de la precisión y la exhaustividad entre varias soluciones.

Ww

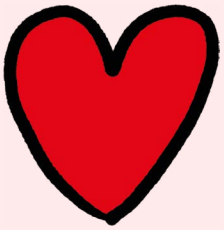
Watson: Es un sistema basado en Inteligencia Artificial capaz de responder a preguntas formuladas en lenguaje natural, desarrollado por la empresa estadounidense IBM.

World Wide Web (WWW): Red informática mundial, sistema lógico de acceso y búsqueda de la información disponible en Internet, cuyas unidades informativas son las páginas web.

Hemos querido regalarte una cita entre arte y tecnología, y para eso elegimos el arte de uno de los ilustradores españoles más reconocidos a nivel internacional. Sus creaciones tratan de historias, personajes y experiencias a lo largo del tiempo; y se basan en las relaciones con el arte popular, la cultura del tatuaje tradicional y moderno, el imaginario religioso europeo y las artes tribales.

RICARDO CAVOLO

El corazón de Ricardo Cavolo



Nacido en Salamanca en 1982 y bajo la influencia de la pintura de su padre, Ricardo Cavolo aprendió desde pequeño a coger bien el lápiz, "para no soltarlo desde entonces".

Tras su paso por Bellas Artes, se destacó por su estilo naïf con colores fuertes y vibrantes, que reflejan variadísimas referencias estéticas como videojuegos, cómics, dibujos animados, superhéroes, tarot, arte popular y todo tipo de universos místicos. Y, cómo no, por los 4 ojos de sus personajes, tan potentes por su diseño como por su historia.

"Un día le pregunté a mi padrastro, que era gitano: ¿cómo es que sabes tanto, si no sabes ni leer ni escribir? Y él contestó: sé muchas cosas porque he vivido muchas cosas. Cuanto más viajas, más ves y más sabes. Es por eso que los 4 ojos son un mimo que le hago a mis personajes, para que sean especiales".

El amor inclusivo

¿Quién mejor que un artista para combinar la fantasía con una visión crítica de la sociedad? Con un estilo directo, formas sencillas y una paleta de color llamativa y cálida, Ricardo Cavolo crea composiciones cargadas de detalles y simbología, que expresan más cuanto más se las explora.

Sus diseños abarcan heterogéneos formatos como pinturas para exposiciones y murales, e ilustraciones editoriales para campañas publicitarias. Trabajos que lo han llevado a participar en diversos festivales como Mural (Canadá), Glastonbury Festival (UK), Cut Out Fest (México) o Mulafest (España); y exposiciones en galerías de Madrid, Londres, Nueva York, Montreal, Oporto y Milán.

Por su parte, ofrece talleres de ilustración a lo largo del mundo, y ya ha colaborado con diversas obras de arte para marcas como Gucci, Apple, Zara, Starbucks, Alexander McQueen, Bally, Nike, Converse, Coca-Cola, Levi's y Circo del Sol, entre otras.

Sun & Moon _ Ricardo Cavolo

La magia de la singularidad

¿Cómo destacar el encanto de su arte? Sin duda en el caso de Ricardo Cavolo la respuesta está en el esfuerzo diario y la fe que tiene en sí mismo.

Es cierto que el mundo está lleno de trampas y fosos, y la carrera a veces puede ser confusa. Quizás la mayor magia está en meterse a fondo en lo que uno ama. “Va a sonar muy cursi, pero el gran hito pasa cada día que me levanto y voy al estudio a trabajar en lo que más me gusta en el mundo. Así día tras día, ya hace 12 años”. Y en estar siempre dispuesto a dar un paso más.

Es el caso de este e-book. Es la primera vez que Cavolo trabaja en un proyecto cuyo fin principal es ofrecer contenido de actualidad para enseñar a través de la singularidad de sus ilustraciones.

El poder de la diferencia

¿Qué tiene de especial este libro? En todo su proceso de creación se aplicó Learning by doing. Una metodología clave para el éxito, que se basa en el desarrollo de habilidades a través de la experimentación. Aprendiendo a resolver los verdaderos desafíos y adquiriendo nuevos conocimientos y habilidades, partiendo de los errores y aciertos propios.

Esta es la metodología de enseñanza de Nuclio Digital School y el procedimiento que llevó adelante Cavolo a la hora de lanzarse a este ambicioso proyecto.

En palabras del artista: “Traté de traducir la idea que debía desarrollar, a mi lenguaje de fantasía. Una vez que encontré el lugar en ese universo de inspiraciones, lo comencé a envolver con detalles y símbolos que lograran cerrar con el significado”.

Un significado que, fundamentalmente, plasma la íntima relación entre diseño y tecnología.

Wall of friends _ Ricardo Cavolo
Fantasy _ Ricardo Cavolo

AGRADECI-MIENTOS


ean Code

Toni Badia (I)

Mi nombre es Toni Badia, soy un joven emprendedor de Barcelona apasionado y dedicado a todo aquello en lo que confluyen economía y tecnología. Creo que la atracción que siento por el mundo de los datos viene de los hábitos de mi familia, que siempre ha compartido los conocimientos de forma abierta y sin tabúes. A pesar de que mi formación académica sea en empresariales y derecho, con cursos especializados y de manera autodidacta, he aprendido todos esos conocimientos que no se encuentran en las universidades y que me despertaban un gran interés. Es el caso del Big Data y la Inteligencia Artificial, que llamaron mi atención hace ya muchos años cuando diseñaba campañas de SEM para una de mis startups.

Después de unos años emprendiendo distintos negocios, entré por completo en el universo de la blockchain, primero como trader y posteriormente como co-fundador de Dragon Corp Games. A partir de esto, hemos recibido un montón de información de proyectos NFT Gaming que clasificamos y evaluamos, siendo necesario entrar de lleno en el apasionante mundo del Big Data y la IA. Aun así, todavía no he sido capaz de hacerle entender a mi abuela a qué me dedico.

Carlos Pérez (II)

Mi nombre es Carlos Pérez Ricardo, soy hijo de una familia trabajadora de origen cubano y desde pequeño desarrollé un gran interés por la tecnología, gracias a las series de televisión y películas que veía, y a la profesión de mis padres. Decidí estudiar Ingeniería Aeroespacial en la Universitat Politècnica de Catalunya, entre otras cosas, porque necesitaba responderme cómo es posible que un amasijo de metal consiga volar. Durante la carrera me acabaron interesando más las materias que tenían una componente de programación y optimización, y actualmente me fascina cómo puedo experimentar e innovar desde mi propio ordenador.

Al acabar el máster, después de trabajar como Ingeniero en SEAT, decidí adentrarme en el Data Science y la Inteligencia Artificial de la mano de Nuclio Digital School. A partir de eso empecé a trabajar como Data Scientist en un grupo hotelero internacional, Grupo Hotusa. Y junto a nuestro equipo desarrollamos proyectos y herramientas que sacan partido a los datos. El principal objetivo es atender las necesidades de nuestros clientes, maximizar los ingresos del grupo y optimizar procesos internos. Para ello, necesitamos un gran entendimiento del negocio, obsesión por el detalle y pasión por el análisis.

Massimiliano Brevini (III)

Mi nombre es Massimiliano Brevini y soy italiano de origen, aunque llevo unos 5 años viviendo en España, principalmente entre Barcelona y Valencia. Trabajo como Senior Data Analyst para una empresa española del sector tecnológico y estoy cursando dos másteres afines al mundo de la ciencia de datos, uno con especialización en la ingeniería de datos y otro en la ingeniería matemática.

(I)

(II)

Mi pasión por la tecnología y los datos se desarrolló cuando era pequeño, concretamente cuando me regalaron mi primer ordenador, con Windows 95. Aunque estoy orientado al mundo Data Science/IA, empecé trabajando en el sector de logística y operaciones en Glovo (Barcelona) y luego en el mundo de micromovilidad eléctrica, de la mano de Voi. Lo que más me gusta del universo Data Science es que nunca paras de estudiar y aprender cosas nuevas, además de utilizar múltiples disciplinas interconectadas como matemáticas, informática y negocios, para generar valor, extraer conocimiento y dar respuestas precisas dentro de la organización.

Jesús Prada Alonso (IV)

Mi nombre es Jesús Prada Alonso y soy Doctor en Machine Learning e inconformista por naturaleza. Siempre me han atraído mucho las matemáticas, pero a la hora de decidir qué estudiar tenía algo de miedo por las reducidas salidas laborales que tenía la carrera en ese entonces. Por este motivo, y dado que la informática siempre ha estado muy presente en mi casa, decidí estudiar un plan doble de informática-matemáticas y un doble máster de Inteligencia Computacional y Matemáticas Aplicadas. Fue en ese momento cuando escuché hablar por primera vez sobre Machine Learning, ¿y qué decir? Creo que fue amor a primera vista.

Desde entonces he cursado un doctorado centrado en las técnicas de Machine Learning y un máster en Bioestadística y Biología Computacional. Me convertí en freelancer y he colaborado con compañías como Iberia Express y en todo tipo de proyectos. Luego emprendí mi propia empresa, Horus ML, dedicada al desarrollo de proyectos innovadores de Machine Learning en el ámbito sanitario, dado que crear aplicaciones en sanidad siempre ha sido mi objetivo a nivel laboral por su gran potencial de impacto positivo.

Espartaco Camero (V)

Hola, mi nombre es Espartaco Camero y soy italo-venezolano. Estudié Matemática en la Universidad Central de Venezuela, y aunque la carrera allí era bastante teórica, siempre me inclinaba por la parte aplicada de la misma, enfocándome en estadística y probabilidad, así como en programación. Posteriormente hice una Especialización en Estadística, donde tuve mayor contacto con temas asociados a Machine Learning y Data. Aunque ya sabía que el mundo de los datos era lo que me apasionaba, comencé dando clases de Matemáticas en la universidad por 3 años, hasta que encontré una oferta para aplicar mis conocimientos en el mundo de la investigación de mercados. Fue allí cuando inicié mi carrera como Data Analyst.

Durante más de una década he pasado por empresas como eDreams Odigeo, Ymedia y Telefónica, trabajando como Data Scientist en el desarrollo de modelos de Machine Learning que añadieran valor a diferentes unidades de negocio como marketing, fraude, User Retention, entre otros. Actualmente lidero el equipo de Data Science y Analytics para Mad Collective, donde usamos data para responder preguntas complejas de la manera más acertada, para partners estratégicos de negocio.

(III)

(IV)

(V)

CREATIVE TEAM

(I)

(VI)

(III)

(V)

(II)

(IV)

Guillem Sánchez (I)

Responsable de proyectos y comunicación, me he encargado de dar vida a todo lo que rodea a este libro digital. Nacido en Barcelona, el mundo del marketing y la comunicación creativa son lo mío, siempre potenciado por un mismo objetivo: romper los esquemas de una construcción social impuesta.

Aquí nadie se va a inventar nada

Victoria Bonifacino (II)

Encargada de la redacción creativa y edición del contenido técnico del e-book. Nací en Argentina, donde estudié Comunicación Social, Publicidad y Creatividad. En busca de nuevas aventuras y aprendizajes me mudé a España y trabajo como Content Manager para Nuclio Digital School.

La vida es el arte de escribir sin borrar

Sandra Párraga (III)

Responsable del diseño editorial y la línea gráfica del libro digital e impreso. Nacida en Barcelona, soy diseñadora gráfica especializada en diseño de interfaz y experiencia de usuario. Este proyecto ha sido todo un reto para mí, ¡y lo he disfrutado mucho!

Ningún proyecto es imposible si te rodeas de las personas adecuadas

Melina Belén Delgado (IV)

Encargada de la Traducción y transcreación del e-book. Soy de Argentina, me gradué en traducción y en el 2021 aterricé en Barcelona para estudiar Marketing. Hoy me dedico a la comunicación creativa en inglés y español.

Las palabras son nuestra fuente más inagotable de magia

Julieta Pandiani (V)

Manager del Departamento Creativo, encargada de la consistencia visual del e-book, coherencia comunicativa visual y escrita, y coordinación del equipo para lograr los objetivos. Nací en Argentina y llevo más de 8 años de experiencia como diseñadora gráfica, especializada en branding.

Las cosas pasan por algo

Esther García (VI)

Diseñadora gráfica y creadora audiovisual, he participado como maquetadora y en la estrategia de difusión de redes sociales del libro digital. Nací en Barcelona y llevo más de 4 años trabajando en el mundo creativo.

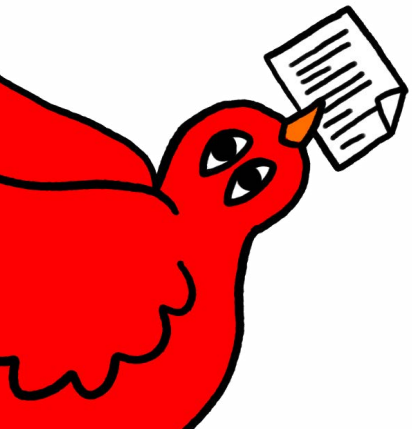
La creatividad es la inteligencia divirtiéndose

Si estás aquí es porque tenemos algo en común. Nos une el fuego imparable de la revolución digital. No nos basta ser parte, queremos ser protagonistas, líderes.

El apasionante mundo de los datos tiene mucho más por explorar.

Dar el siguiente paso depende de ti.

Te invitamos a que conozcas más sobre Nuclio Digital School. Estaremos a tu lado asegurando tu impulso.



NUCLIO DIGITAL SCHOOL