¿Qué es la ciencia de datos?

Por Animikh Aich | 24 de mayo de 2019 | Categoría: Blog | 7291 | 857

¿Qué es la ciencia de datos?

La ciencia de datos es un campo multidisciplinario que utiliza inferencias científicas y algoritmos matemáticos para extraer conocimientos y perspectivas significativas de una gran cantidad de datos estructurados y no estructurados. Estos algoritmos se implementan a través de programas de computadora que generalmente se ejecutan en hardware potente, ya que requiere una cantidad significativa de procesamiento. La ciencia de datos es una combinación de matemáticas estadísticas, aprendizaje automático, análisis y visualización de datos, conocimiento del dominio e informática.

Como se desprende del nombre, el componente más importante de la ciencia de datos son los propios "datos". Ninguna cantidad de cálculo algorítmico puede extraer información significativa de datos incorrectos. La ciencia de datos involucra varios tipos de datos, por ejemplo, datos de imágenes, datos de texto, datos de video, datos dependientes del tiempo, etc.

Historia de la ciencia de datos

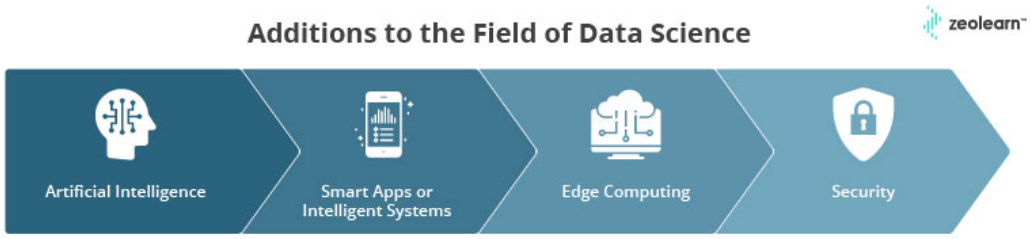
El término "ciencia de datos" se ha mencionado en varios contextos durante los últimos treinta años, pero solo recientemente se ha establecido y reconocido internacionalmente. Más recientemente, el término se convirtió en una palabra de moda cuando Harvard Business Review lo llamó "El trabajo más sexy del siglo XXI" en 2012.

Origen del concepto

Aunque no está claro cuándo y dónde se desarrolló originalmente el concepto, William S. Cleveland acuñó el término "Ciencia de datos" en 2001. Poco después, en abril de 2002 y enero de 2003, las publicaciones del "CODATA Data Science Journal" de la International Council for Science: el Comité de Datos para la Ciencia y la Tecnología y el "Journal of Data Science" de la Universidad de Columbia, respectivamente, iniciaron el viaje de Data Science.

Además, también fue en esta época cuando la burbuja de las "punto com" estaba en pleno apogeo, lo que llevó a la adopción generalizada de Internet y, a su vez, a la generación de una enorme cantidad de datos. Esto, sumado al avance de la tecnología, que condujo a una computación más rápida y barata, fue responsable del lanzamiento del concepto de "Ciencia de datos" al mundo.

Adiciones recientes al campo de la ciencia de datos



El campo de la ciencia de datos se ha expandido desde su inicio a principios de la década de 2000. Con el tiempo, se están incorporando cada vez más tecnologías de vanguardia en el campo. Algunas de estas adiciones más recientes se enumeran a continuación:

1. **Inteligencia artificial: el** aprendizaje automático ha sido uno de los elementos centrales de la ciencia de datos. Sin embargo, con el aumento de las capacidades de cómputo paralelo, Deep Learning ha sido la última y una de las adiciones más importantes al campo de la ciencia de datos.
2. **Aplicaciones inteligentes o sistemas inteligentes:** el desarrollo de aplicaciones inteligentes basadas en datos y su accesibilidad en un factor de forma portátil ha llevado a la inclusión de una parte de este campo en la ciencia de datos. Esto se debe principalmente a que una gran parte de la ciencia de datos se basa en el aprendizaje automático, que es también en lo que se basan las aplicaciones inteligentes y los sistemas inteligentes.
3. **Edge Computing:** Edge Computing es un concepto desarrollado recientemente y está relacionado con IoT (Internet of Things). La computación perimetral básicamente acerca la canalización de recopilación, entrega y procesamiento de información de la ciencia de datos a la fuente de información. Esto se puede lograr a través de IoT y recientemente se agregó para ser parte de Data Science.

4. **Seguridad: la** seguridad ha sido un desafío importante en el espacio digital. La inyección de malware y el concepto de piratería son bastante comunes y todos los sistemas digitales son vulnerables a ella. Afortunadamente, ha habido pocos avances tecnológicos recientes que apliquen técnicas de ciencia de datos para prevenir la explotación de sistemas digitales. Por ejemplo, las técnicas de aprendizaje automático han demostrado ser más capaces de detectar virus informáticos o malware en comparación con los algoritmos tradicionales.

Desdibujando las líneas entre ciencia de datos y análisis de datos

Las palabras de moda "Ciencia de datos" y "Análisis de datos" se utilizan a menudo indistintamente. Aunque estos dos campos están estrechamente relacionados, no significan lo mismo. En resumen, la ciencia de datos es un término general que consta de los campos de aprendizaje automático, análisis de datos y minería de datos.

En términos de descripción del puesto, un "científico de datos" y un "analista de datos" también trabaja en tecnologías diferentes pero relacionadas.

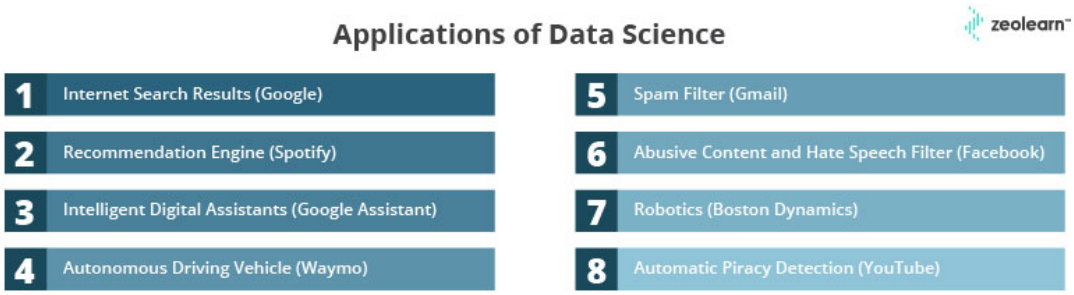
Parámetros	Científico de datos	Analista de datos
Definición	Una persona que tiene la habilidad de manejar una gran cantidad de datos para construir modelos y extraer información significativa de ellos con la ayuda de algoritmos estadísticos y de aprendizaje automático que utilizan conceptos informáticos.	Una persona cuyo trabajo principal es examinar una gran cantidad de datos, discutirlos y visualizarlos y determinar qué conocimientos ocultan los datos.
Habilidades	Aprendizaje automático, estadísticas, visualización de datos, bases de datos, ingeniería de software, minería de datos, conocimiento del dominio	Estadísticas, visualización de datos, manipulación de datos, bases de datos, minería de datos
Tecnologías	Python, R, SQL, AWS, bibliotecas de aprendizaje automático,	Java, Hadoop, Hive, Spark, AWS, SQL, Tableau

Papel de Big Data en la ciencia de datos

El término "Big Data" se refiere a una gran colección de datos heterogéneos estructurados, semiestructurados o no estructurados. Las bases de datos generalmente no son capaces de manejar conjuntos de datos tan voluminosos.

Como se mencionó anteriormente, el componente clave de Data Science son los datos. Como regla general, "a mayor cantidad de datos, mejor información". Por lo tanto, Big Data juega un papel muy importante en el campo de la ciencia de datos. Big Data se caracteriza por su variedad y volumen, ambos fundamentales para la ciencia de datos. Data Science captura los patrones complejos de Big Data mediante el desarrollo de algoritmos y modelos de aprendizaje automático.

Aplicaciones de la ciencia de datos



La ciencia de datos es un campo que se puede aplicar a casi todas las industrias para resolver problemas complejos. Cada empresa aplica la ciencia de datos a una aplicación diferente con el fin de resolver un problema diferente. Algunas empresas dependen completamente de las técnicas de Data Science y Machine Learning para resolver un cierto conjunto de problemas que, de otra manera, no se podrían haber resuelto. Algunas de estas aplicaciones de Data Science y las empresas que las respaldan se enumeran a continuación.

1. **Resultados de búsqueda en Internet (Google):** cuando un usuario busca algo en Google, los complejos algoritmos de aprendizaje automático determinan cuáles son los resultados más relevantes para los términos de búsqueda. Estos algoritmos ayudan a clasificar las páginas de modo que la información más relevante se proporcione al usuario con solo hacer clic en un botón.
2. **Motor de recomendación (Spotify):** Spotify es un servicio de transmisión de música que es bastante popular por su capacidad para recomendar música según el gusto del usuario. Este es un muy buen ejemplo de ciencia de datos en juego. Los algoritmos de Spotify utilizan los datos generados por cada usuario a lo largo del tiempo para conocer el gusto musical del usuario y recomendarlo con música similar en el futuro. Esto permite a la empresa atraer más usuarios ya que es más conveniente para el usuario utilizar Spotify ya que no demanda mucha atención.
3. **Asistentes digitales inteligentes (Asistente de Google):** el Asistente de Google, similar a otros asistentes digitales basados en voz o texto (también conocidos como chatbots) es un ejemplo de algoritmos avanzados de aprendizaje automático puestos en uso. Estos algoritmos son capaces de convertir el habla de una persona (incluso con diferentes acentos e idiomas) en texto, comprender el contexto del texto / comando y proporcionar información relevante o realizar una tarea deseada, todo con solo hablar con el dispositivo.
4. **Vehículo de conducción autónoma (Waymo): los** vehículos de conducción autónoma son una de las **últimas** tecnologías. Empresas como Waymo utilizan cámaras de alta resolución y LIDAR para capturar videos en vivo y mapas 3D de los alrededores con el fin de alimentarlos a través de algoritmos de aprendizaje automático que ayudan a conducir el automóvil de forma autónoma. Aquí, los datos son los videos y mapas 3D capturados por los sensores.
5. **Filtro de correo no deseado (Gmail):** otra aplicación clave de la ciencia de datos que utilizamos en nuestro día a día son los filtros de correo no deseado en nuestros correos electrónicos. Estos filtros separan automáticamente los correos electrónicos no deseados del resto, lo que le brinda al usuario una experiencia de correo electrónico mucho más limpia. Al igual que las otras aplicaciones, la ciencia de datos es el componente clave aquí.
6. **Filtro de contenido abusivo y discurso de odio (Facebook):** similar al filtro de correo no deseado, Facebook y otras plataformas de redes sociales utilizan algoritmos de ciencia de datos y aprendizaje automático para filtrar el contenido abusivo y con restricción de edad de la audiencia no deseada.

7. **Robótica (Boston Dynamics):** un componente clave de la ciencia de datos es el aprendizaje automático, que es exactamente lo que impulsa la mayoría de las operaciones de robótica. Empresas como Boston Dynamics están a la vanguardia de la industria de la robótica y desarrollan robots autónomos que son capaces de realizar movimientos y acciones humanoides.
8. **Detección automática de piratería (YouTube): la** mayoría de los videos que se cargan en YouTube son contenido original creado por creadores de contenido. Sin embargo, con bastante frecuencia, los videos pirateados y copiados también se cargan en YouTube, lo que va en contra de su política. Debido al gran volumen de cargas diarias, no es posible detectar y eliminar manualmente dichos videos pirateados. Aquí es donde se usa Data Science para detectar automáticamente videos pirateados y eliminarlos de la plataforma.

El ciclo de vida de la ciencia de datos

El campo de la ciencia de datos no es un proceso de un solo paso. Tiene muchos pasos involucrados. Estos pasos se enumeran a continuación.

1. **Análisis de proyectos:** este paso se inclina más hacia la gestión de proyectos y la evaluación de recursos que una implementación directa de algoritmos. En lugar de iniciar un proyecto a ciegas, es crucial determinar los requisitos del proyecto en términos de la fuente de datos y su disponibilidad, la cantidad de recursos humanos disponibles y si el presupuesto asignado al proyecto es suficiente para completarlo con éxito.
2. **Preparación de datos:** en este paso, los datos sin procesar se convierten en datos estructurados y se limpian. Esto implica análisis de datos, limpieza de datos, manejo de valores perdidos, transformación de datos y visualización. A partir de este paso, se utilizan lenguajes de programación como R y Python para lograr resultados para grandes conjuntos de datos.
3. **Análisis de datos exploratorios (EDA):** este es un paso crucial en la ciencia de datos, donde el científico de datos explora los datos desde varios ángulos e intenta sacar conclusiones iniciales de los datos. Esto incluye visualización de datos, creación rápida de prototipos, selección de características y, finalmente, selección de modelos. En este paso se utiliza un conjunto diferente de herramientas. Los más utilizados son R o Python para scripting y manipulación de datos, SQL para interactuar con bases de datos y diferentes bibliotecas para manipulación y visualización de datos.

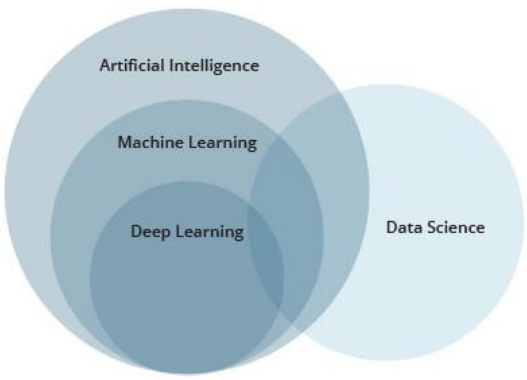
4. **Construcción del modelo:** Una vez determinado el tipo de modelo a utilizar a partir de la EDA, la mayoría de los recursos se canalizan hacia el desarrollo del modelo con hiperparámetros ideales (parámetros modificables), de manera que pueda realizar análisis predictivos sobre datos similares pero no vistos. Varias técnicas de Machine Learning aplicadas a los datos, como Clustering, Regression, Classification o PCA (Main Component Analysis) para extraer información valiosa de ellos.
5. **Implementación: una** vez que el modelo se ha creado correctamente, es el momento de llevarlo al mundo real desde su entorno de pruebas. Aquí es donde entra en juego la implementación del modelo. Hasta ahora, todos los pasos estaban dedicados a la creación rápida de prototipos. Sin embargo, una vez que el modelo se ha construido y entrenado con éxito, la aplicación principal del mismo es en el mundo real, donde se implementa. Esto puede ser en forma de una aplicación web, aplicación móvil o puede ejecutarse en el back-end del servidor para procesar datos de alta frecuencia.
6. **Pruebas y resultados** del mundo real: una vez implementado el modelo, se enfrenta a datos invisibles del mundo real en tiempo real. El modelo puede funcionar muy bien en la zona de pruebas, pero no funciona adecuadamente después de la implementación. Esta es la fase en la que se requiere un monitoreo constante de la salida del modelo para detectar escenarios donde el modelo falla. Si falla en algún momento, el proceso de desarrollo vuelve al Paso 1. Si el modelo tiene éxito, los hallazgos clave se anotan y se informan a las partes interesadas.

¿Dónde encaja la ciencia de datos en comparación con las otras palabras de moda? IA, aprendizaje automático, aprendizaje profundo

"Ciencia de datos" parece ser una palabra bastante confusa, que no tiene una definición o límites claros. Las palabras de moda "Inteligencia artificial", "Aprendizaje automático" y "Aprendizaje profundo" a menudo se usan indistintamente con "Ciencia de datos" o en asociación con ella. Definamos claramente los límites de cada uno de estos términos.

Como se mencionó anteriormente, el aprendizaje automático es parte de la ciencia de datos. Como se muestra en la siguiente figura, el aprendizaje profundo es parte del aprendizaje automático y el aprendizaje automático es a su vez parte de la inteligencia artificial.

Aunque la ciencia de datos incluye una parte de cada uno de inteligencia artificial, aprendizaje automático y aprendizaje profundo, contiene más que solo estos tres subdominios en su interior. Data Science también contiene programación estadística, análisis de datos, minería de datos, Big Data y adiciones más recientes como IoT, Edge Computing y seguridad.



Por lo tanto, la ciencia de datos es un campo complejo del estudio científico de datos, que contiene una parte significativa de algunos de los avances más recientes en ciencias de la computación y matemáticas.

Habilidades necesarias para convertirse en científico de datos

Como se mencionó en la sección anterior, la ciencia de datos es un campo complejo. Por lo tanto, requiere el dominio de múltiples subcampos, que juntos suman el conocimiento completo requerido para ser un científico de datos.

1. Matemáticas: El primer y más importante campo de estudio para convertirse en un científico de datos son las matemáticas; más específicamente, probabilidad y estadística, álgebra lineal y algo de cálculo básico.

- **Estadísticas:** Es esencial en EDA y en el desarrollo de algoritmos para realizar inferencias estadísticas sobre los datos. Además, la mayoría de los algoritmos de aprendizaje automático utilizan estadísticas como sus bloques de construcción fundamentales.
- **Álgebra lineal:** Trabajar con una gran cantidad de datos significa trabajar con matrices de alta dimensión y operaciones matriciales. Los datos que toma el modelo y el que da como salida están en forma de matrices y, por lo tanto, cualquier operación que se lleve a cabo en ellos utiliza los fundamentos del álgebra lineal.
- **Cálculo:** dado que la ciencia de datos incluye el aprendizaje profundo, el cálculo es de inmensa importancia. En Deep Learning, el cálculo del gradiente es muy importante y se realiza en cada paso del cálculo en las redes neuronales. Esto requiere un conocimiento sólido del cálculo diferencial e integral.

2. Conocimiento algorítmico: aunque la ciencia de datos normalmente no implica el desarrollo y diseño de algoritmos como lo hace cualquier otra aplicación de la ciencia de la computación, sigue siendo imperativo que un científico de datos tenga un conocimiento sólido de los algoritmos. Esto se debe a que, al final del día, los científicos de datos son programadores que se espera que desarrollen programas que deriven información significativa de los datos. Tener conocimiento algorítmico permite al científico de datos escribir código significativo y eficiente, lo que ahorra tiempo y recursos y, por lo tanto, es muy valorado.

3. Lenguajes de programación (R y Python): aunque, cualquier lenguaje de programación puede usarse para cualquier tipo de caso de uso lógico, que por supuesto, incluye la ciencia de datos; pero los lenguajes más utilizados son R y Python. Ambos lenguajes son de código abierto y, por lo tanto, tienen un gran apoyo de la comunidad, tienen múltiples bibliotecas desarrolladas teniendo en cuenta la ciencia de datos y son relativamente fáciles de aprender y usar. Sin el conocimiento de lenguajes de programación, un científico de datos no puede aplicar ningún tipo de conocimiento algorítmico o matemático a los datos.

4. Entorno de programación adecuado: dado que un conocimiento sólido de programación es uno de los requisitos clave para la ciencia de datos, es necesario que exista una plataforma conveniente para escribir y ejecutar el código. Esta plataforma se llama IDE o entorno de desarrollo integrado. Hay varios IDE para elegir y algunos de ellos se han desarrollado específicamente para la ciencia de datos. Este artículo habla sobre los 10 mejores IDE de Python.

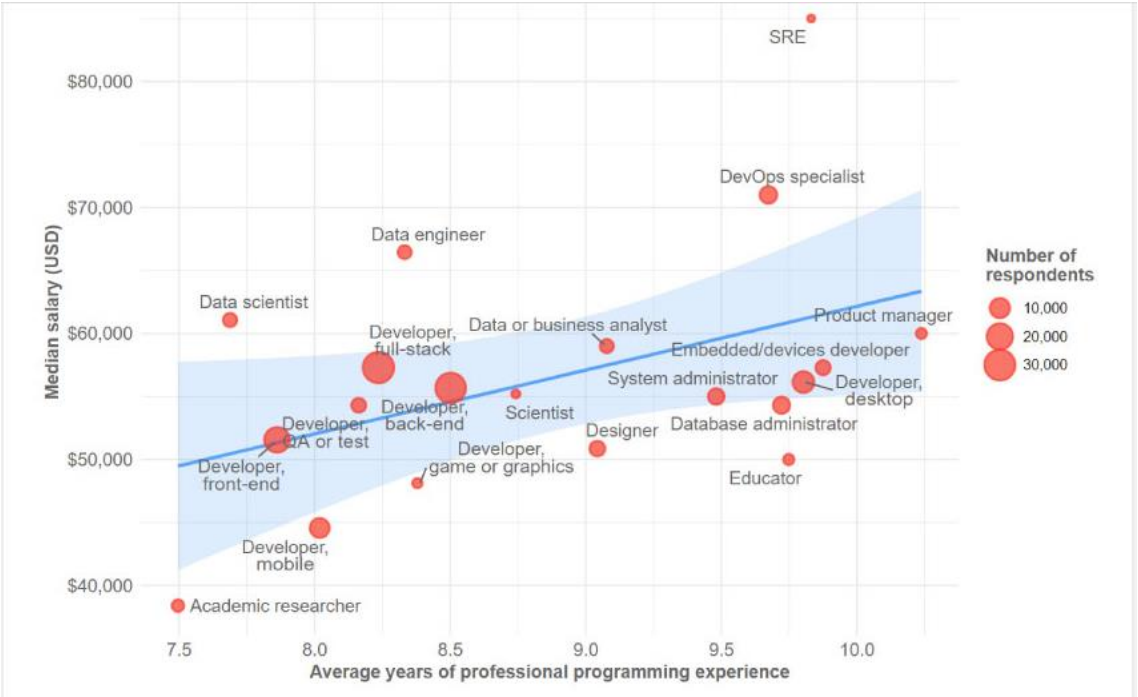
5. Marcos de aprendizaje automático: el aprendizaje automático es una parte importante de la ciencia de datos y su implementación involucra ciertas bibliotecas y marcos, cuyo conocimiento es esencial para cualquier científico de datos. A continuación, se enumeran algunos de los marcos de aprendizaje automático más utilizados.

- **Numpy:** Esta es una biblioteca que permite la fácil implementación de álgebra lineal y manipulación de datos.
- **Pandas:** esta biblioteca se utiliza para cargar, modificar y guardar datos. Esto también se utiliza en la manipulación de datos.
- **Matplotlib:** esta es una de las bibliotecas más utilizadas para la visualización de datos.
- **Seaborn:** este es un contenedor sobre Matplotlib, que se utiliza para visualizar datos más complejos.
- **Sklearn:** se utiliza para aplicar e implementar la mayoría de los algoritmos de aprendizaje automático y las técnicas de preprocesamiento de datos.
- **Tensorflow:** este es un marco de aprendizaje profundo respaldado por Google y permite una fácil implementación de varios tipos de redes neuronales.
- **PyTorch:** similar a tensorflow, este también es un marco de aprendizaje profundo que se usa con frecuencia.
- **Keras:** este es un contenedor que funciona junto con tensorflow y permite una implementación relativamente fácil de técnicas de aprendizaje profundo.
- **OpenCV:** este es un marco de visión por computadora y generalmente se usa para procesamiento de imágenes y manipulación de imágenes. Se utiliza para datos de video o imágenes.

6. SQL: Las bases de datos son de inmensa importancia en el campo de la ciencia de datos ya que son el método más adecuado para almacenar datos. También es importante un conocimiento profundo de una o más tecnologías de bases de datos como MySQL, MariaDB, PostgreSQL, MS SQL Server, MongoDB, Oracle NoSQL, etc.

Salarios de un científico de datos

El campo de la ciencia de datos es uno de los trabajos mejor pagados en el dominio del software. También es el que paga más con la menor cantidad de experiencia laboral relevante en comparación con cualquier otro campo en el dominio del software, como se muestra en la siguiente figura. Estos datos provienen de la Encuesta para desarrolladores de Stack Overflow 2019 .



Algunos de los salarios ofrecidos se enumeran a continuación.

- Según DataJobs, el rango salarial para los científicos de datos en EE. UU. Es de **$ 85,000** a **$ 170,000.**
- Según PayScale, el rango salarial en la India es de **₹ 305,000** a **₹ 2,000,000** y el salario medio es de **₹ 620,000.**
- Glassdoor establece que el salario base promedio para los científicos de datos en India es de **₹ 947,698** por año.

Futuro de la ciencia de datos

La ciencia de datos es un campo en constante crecimiento y se espera que aumente la demanda en el futuro previsible. Algunos de los cambios clave se enumeran a continuación.

- **Datos:** con el aumento radical de la generación de datos, el rendimiento de los algoritmos predictivos mejorará con el tiempo a medida que haya más datos estructurados disponibles para hacer inferencias. Este fenómeno está impulsado por el crecimiento de los dispositivos basados en las redes sociales y la IoT, que generan muchos más datos estructurados.
- **Algoritmos:** se espera que los algoritmos de aprendizaje automático como los algoritmos genéticos y los algoritmos de aprendizaje por refuerzo mejoren con el tiempo y provoquen sistemas más inteligentes.
- **Computación distribuida:** con los avances de la tecnología blockchain, el desarrollo de TPU (Unidad de procesamiento de tensor) y una GPU (Unidad de procesamiento de gráficos) más rápida disponible en la nube, la ciencia de datos ve un futuro en el que el hardware computacional más potente ayuda a los algoritmos de complejidad creciente.

Se espera que más datos y algoritmos y hardware mejorados juntos traigan mejoras significativas en el campo de la ciencia de datos en un futuro próximo.

Conclusión

La ciencia de datos es un campo de estudio complejo y promocionado. En su mayor parte, el bombo publicitario es cierto y ofrece soluciones a los problemas según lo prometido. Algunos campos de la ciencia de datos incluso han comenzado a superar a los humanos y se espera que esa tendencia aumente en el futuro cercano. Puede realizar una formación en ciencia de datos para mejorar su carrera.

La ciencia de datos es definitivamente el trabajo más "sexy" del siglo XXI **.** Define la vanguardia de la tecnología en la actualidad y promete nuevos avances tecnológicos en un futuro próximo. También es uno de los trabajos más demandados y mejor pagados de la industria. Por lo tanto, ¡no hay mejor momento para ser un científico de datos que ahora!