Introducción a BigData y la Ciencia de Datos

Introducción a BigData y la Ciencia de Datos

Contenidos

- El mundo gira en torno a los Datos

El mundo gira en torno a los Datos

Ciencia:

- Bases de datos de astronomía, genómica, datos medio-ambientales, datos de transporte, ⋯

Ciencias Sociales y Humanidades:

- Libros digitales, documentos históricos, datos sociales, ⋯

Negocio y Comercio:

- Ventas de corporaciones, transacciones de mercados, censos, tráfico de aerolíneas, ⋯

Entretenimiento y Ocio:

- Internet, películas, vídeos, música, ⋯

El mundo gira en torno a los Datos

Medicina:

- Datos de pacientes, datos de escaner, radiografías, Telemedicina, Teleconsulta, Telediagnostico ···

Industria:

- Energía, Sensores, ···

Somos ricos en datos pero pobres en información.

Debemos aprovechar la data disponible y trascender la dimensión de la información hacia el conocimiento.

Minería de Datos

La ***Minería de datos*** (MD) es el proceso de extracción de patrones de información (implícitos, no triviales, desconocidos y potencialmente útiles) a partir de grandes cantidades de datos.

También se conoce como:

- Descubrimiento de conocimiento en bases de datos (KDD).
- Extracción del conocimiento.
- Análisis inteligente de datos.
- Descubrimeinto de patrones.
- ...

Mineria de datos

Muchas de las técnicas utilizadas en la ***Mineria de Datos*** ya se conocían previamente, pero en la actualidad convergen los siguientes factores:

- Los datos se están produciendo masivamente
- Los datos se están almacenando
- La potencia computacional necesaria está disponible
- Existe una gran presión competitiva a nivel empresarial
- Las herramientas software están disponibles.

¿Para qué se utiliza el 'conocimiento' obtenido?

- Predicciones sobre nuevos datos
- Explicar los datos existentes
- Resumir datos masivos para facilitar la toma de decisiones
- Visualizar datos altamente dimensionales, extrayendo estructura local simplificada
- ...

Big Data

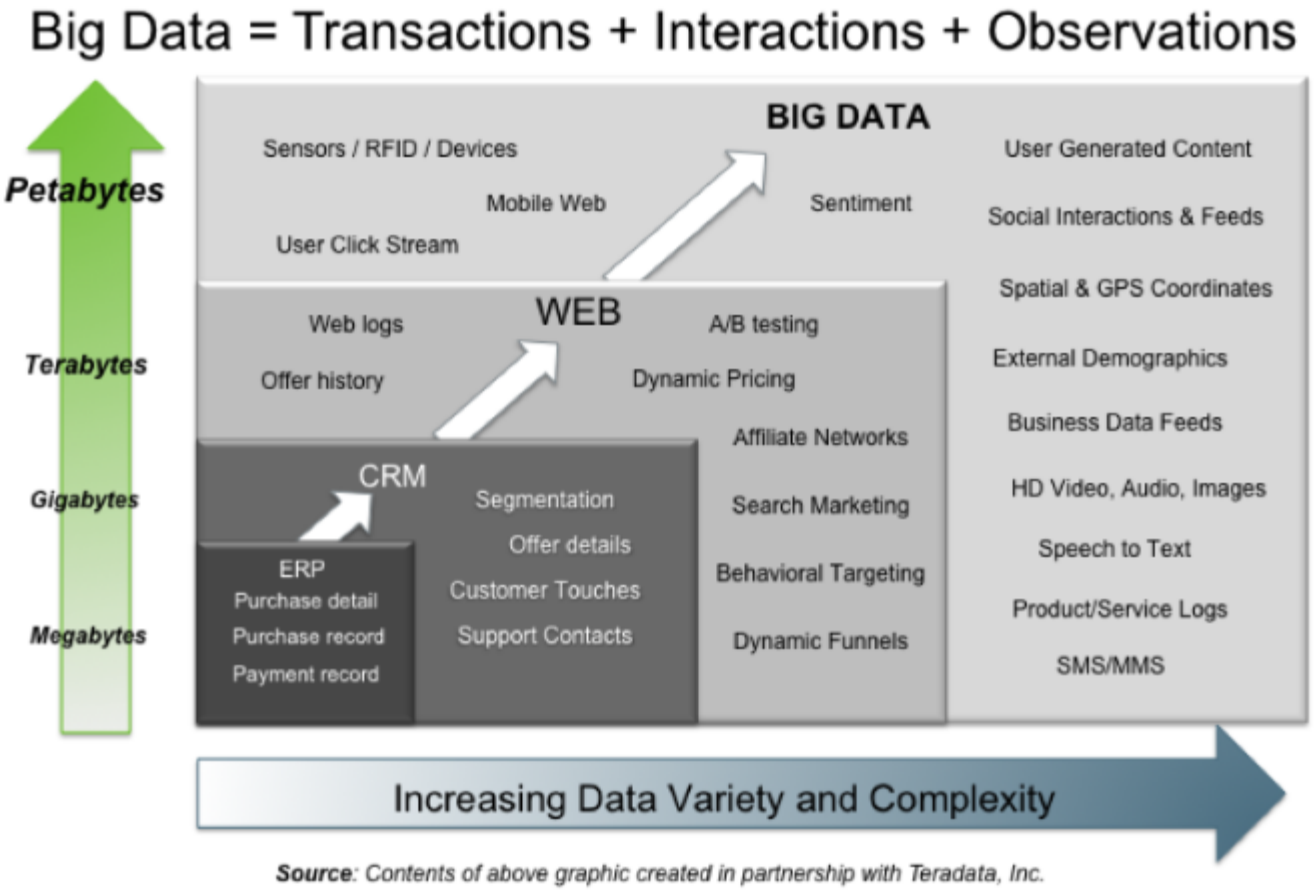
"El concepto de big data se puede definir como las múltiples fuentes de información de alto volúmen, alta velocidad y alta variedad que exigen de formas innovadoras y costo efectivas para ser procesadas con el fin de generar descubrimientos, procesos de desición y automatización de procesos" (Gartner, 2018)

- Big data es una colección de datos grande, complejos, muy difícil de procesar a través de herramientas de gestión y procesamiento de datos tradicionales.

Las 5v de BigData

- Volumen.
- Velocidad.
- Variedad.
- Veracidad.
- Valor.

Big Data



***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Big Data

- Representa una ***oportunidad:*** Tomar decisiones basadas en el uso intensivo de los datos.
- Representa un ***reto:***
  - *Hay que manejar inconsistencias, datos incompletos, escalabilidad, corriente continua de datos, problemas de seguridad.*
  - *Se requieren nuevas tecnologías para el almacenamiento, operaciones de entrada/salida de datos y procesamiento.*
- ***Obliga*** a romper con el enfoque relacional de las bases de datos abordando modelos ***NoSQL*** mas acordes con las dimensiones y naturaleza de los datos.
- ***Requiere*** otro enfoque para la programación paralela.
- ***Rompe*** con los conceptos clásicos de seguridad de los datos.

Big Data

- ***Obliga*** a trabajar con mucha informacion privada (Data anonymization).
- ***Obliga*** a manipular enormes cantidades de datos no estructurados.
- ***Requiere*** intercambio y cooperacion internacional.

Big Data

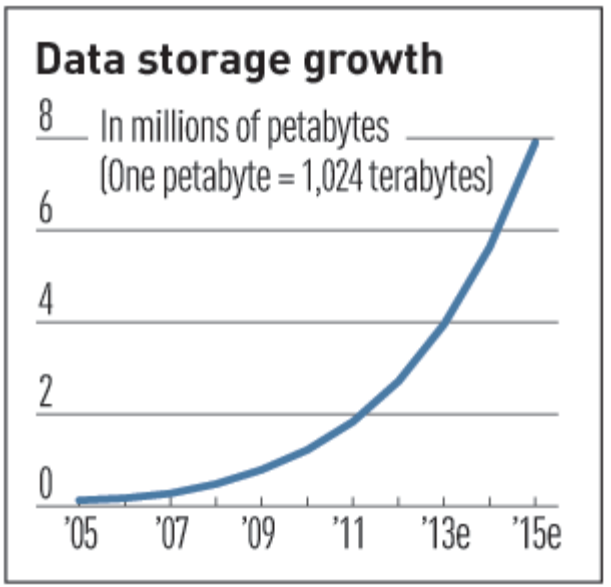
¿Que tan grande es Big Data?

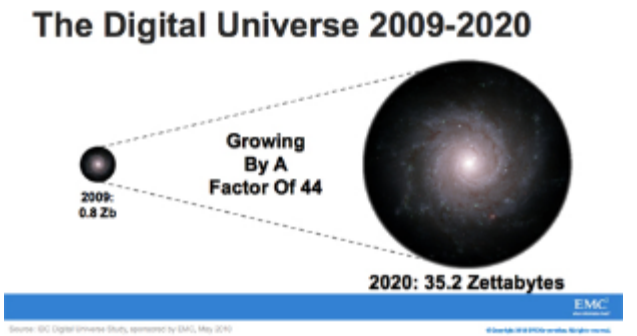
Algunas estadísticas tomadas de https://seedscientific.com/

- The amount of data in the world was estimated to be **44 zettabytes** at the dawn of 2020.
- By 2025, the amount of data generated each day is expected to reach **463 exabytes** globally.
- Google, Facebook, Microsoft, and Amazon store at least **1,200 petabytes** of information.
- The world spends almost **$1 million** per minute on commodities on the Internet.
- Electronic Arts process roughly **50 terabytes** of data every day.
- By 2025, there would be **75 billion** Internet-of-Things (IoT) devices in the world
- By 2030, nine out of every ten people aged six and above would be digitally active.

Big Data

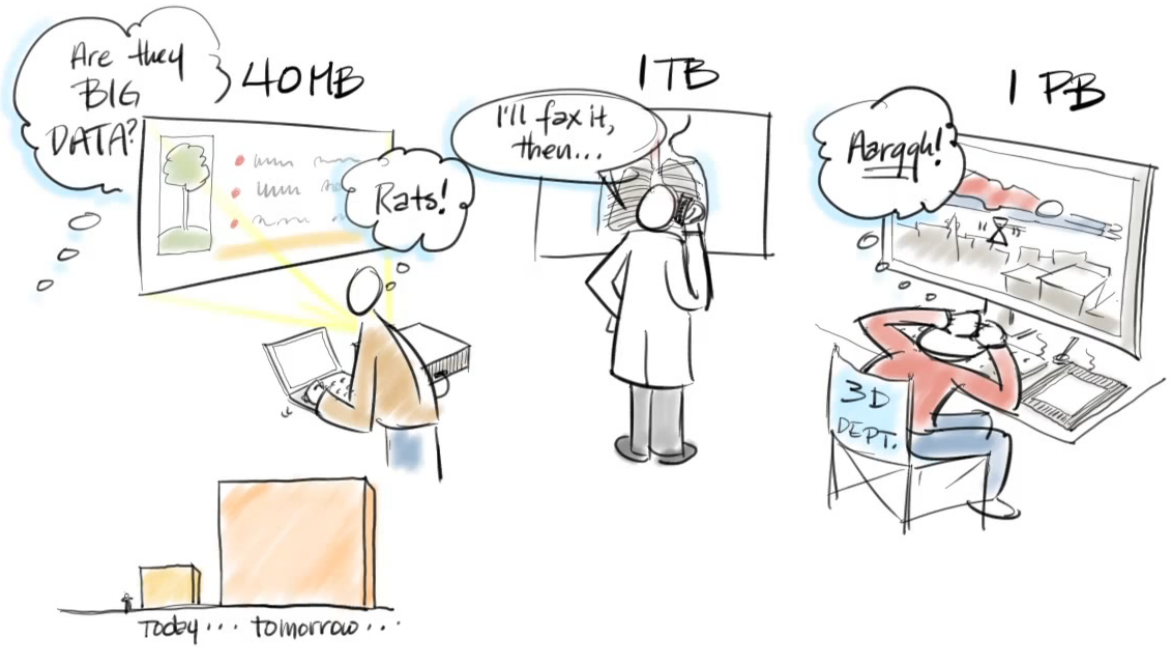
- El volumen de datos crece ***exponencialmente***





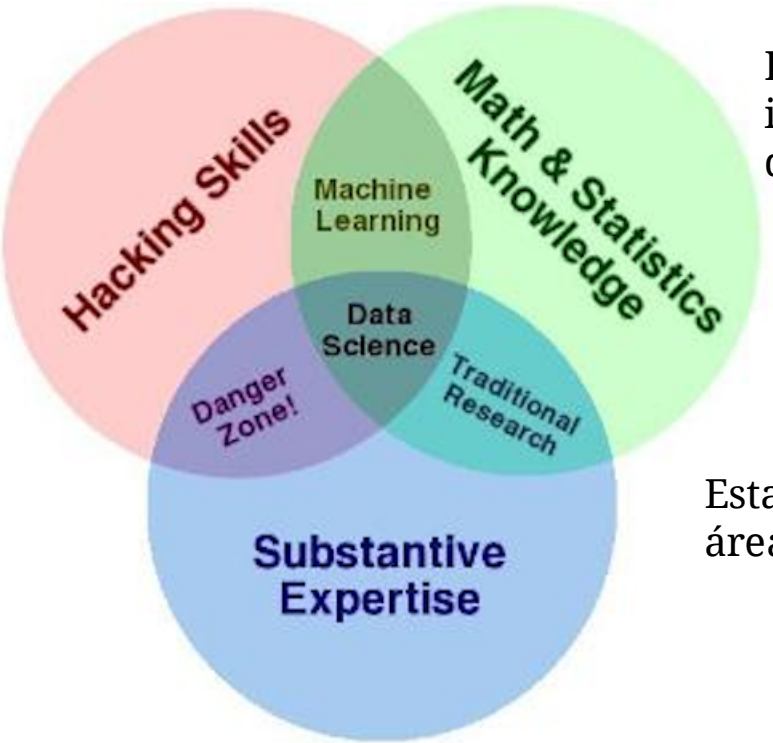
- Crecimiento x 44 de 2009 a 2020
- De 0.8 zettabytes a 35ZB

Big Data



Big data se peuede dar en diferentes escalas y se puede ver como cualquier característica sobre los datos que represente un reto para las funcionalidades de un sistema.

Ciencia de datos

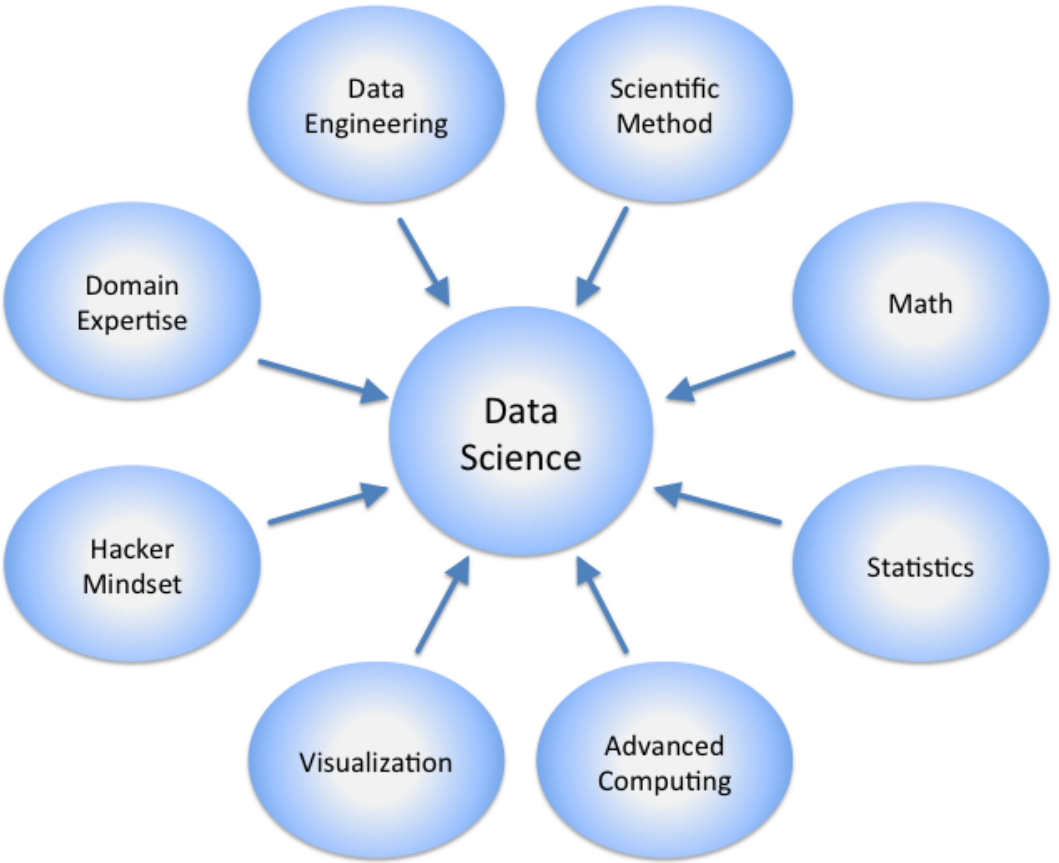


La ciencia de datos es un campo interdisciplinario que aplica técnicas de:

- Matemáticas
- Estadísticas
- Computación

Estas técnicas se aplican a diferentes áreas del conocimiento.

Ciencia de datos



Ciencia de datos

la Ciencia de Datos incorpora diferentes campos del conocimiento como:

- matemáticas
- estadística
- ingeniería de datos
- reconocimiento de patrones y aprendizaje
- computación avanzada
- visualización
- modelado de incertidumbre
- almacenamiento de datos
- informática de alto rendimiento

con el objetivo de extraer el significado de datos y la creación de productos de datos.

Ciencia de datos

La ciencia de datos tiene como objetivos modelar, analizar, visualizar y extraer conocimiento a partir de los datos.

La ciencia de datos abarca la preparación de los datos para el análisis, incluida la limpieza, la agregación y la manipulación de los datos para realizar análisis avanzados.

Las aplicaciones analíticas y los científicos de datos pueden revisar los resultados para descubrir patrones y permitir la toma de decisiones basada en información fundamentada.

La ciencia de datos busca utilizar todos los datos disponibles y relevantes para *“extraer conocimiento”* que pueda ser fácilmente comprendido por los expertos en el área de aplicación.

¿Por qué es tan importante la ciencia de datos?

- Tenemos un tesoro de datos sin aprovechar.
- La tecnología ha permitido la creación y el almacenamiento de cantidades cada vez mayores de datos.
- Se estima que el 90% de los datos en el mundo se crearon en los últimos dos años.
- Los datos frecuentemente solo están inmóviles en las bases de datos y los data lakes.
- La gran cantidad de datos recopilados y almacenados puede generar beneficios transformadores para las organizaciones y sociedades solo si sabemos interpretarlos. ***oportunidad para la ciencia de datos***.
- La ciencia de datos revela tendencias y genera información que las empresas pueden utilizar para tomar mejores decisiones.

¿Por qué es tan importante la ciencia de datos?

- Permite que los modelos de aprendizaje automático (ML) aprendan de las grandes cantidades de datos que se les suministran en vez de depender principalmente de los analistas de negocios.
- Los datos son la base de la innovación, cuando de ellos extraemos información.

STAMFORD, Conn., January 25, 2018

Gartner Says Self-Service Analytics and BI Users Will Produce More Analysis Than Data Scientists Will by 2019

Analysts to Discuss How to Implement Self-Service Analytics and BI at Gartner Data & Analytics Summit, March 19-21, 2018 in London, U.K.

Organizations are embracing self-service analytics and business intelligence (BI) to bring these capabilities to business users of all levels. This trend is so pronounced that Gartner, Inc. predicts that by 2019, the analytics output of business users with self-service capabilities will surpass that of professional data scientists.

Muchas empresas han hecho de la ciencia de datos una prioridad y están realizando grandes inversiones en ella. Los directores de informática ven las tecnologías asociadas a la ciencia de datos como las más estratégicas para sus empresas y están realizando las inversiones correspondientes.

¿Por qué es tan importante la ciencia de datos?

Datos de MinTic

Comunidad

Si quieres conocer más sobre los proyectos de ciencia de datos o los científicos de datos formados por MinTIC, escribe a minticresponde@mintic.gov.co

Científicos de datos
1713

Departamentos
23

Municipios
81

Ciencia de datos, la inteligencia artificial y el aprendizaje automático

- **IA** significa hacer que una computadora imite de alguna manera el comportamiento humano (*inteligencia*).
- La **ciencia de datos** es un subconjunto de la **IA** que se refiere más a las áreas superpuestas de las estadísticas, los métodos científicos y el análisis de datos, que se utilizan todas para extraer significado y conocimientos de los datos.
- El **aprendizaje automático** es otro subconjunto de la **IA** y consiste en las técnicas que permiten que las computadoras descubran cosas a partir de los datos y realicen aplicaciones de IA.
- El **aprendizaje profundo**, es un subconjunto del aprendizaje automático que permite que las computadoras resuelvan problemas más complejos.

¿Que nos permite la Ciencia de Datos?

- Determinar la fuga de clientes analizando los datos que se recopilan de los centros de llamadas, para que el departamento de Marketing pueda tomar medidas a fin de retenerlos.
- Mejorar la eficiencia al analizar los patrones de tráfico, las condiciones climáticas y otros factores para que las empresas de logística puedan mejorar los tiempos de entrega y reducir los costos.
- Mejorar los diagnósticos de los pacientes mediante el análisis de los exámenes médicos y los síntomas informados para que los médicos puedan diagnosticar antes las enfermedades y tratarlas de manera más eficaz.
- Optimizar la cadena de suministro al predecir cuándo se producirán fallos en los equipos.
- Detectar los fraudes en los servicios financieros mediante el reconocimiento de los comportamientos sospechosos y las acciones anómalas.
- Mejorar las ventas al crear recomendaciones para los clientes basadas en las compras anteriores.

El proceso de la ciencia de datos

El proceso de analizar y utilizar los datos es iterativo más que lineal, e incluye etapas como:

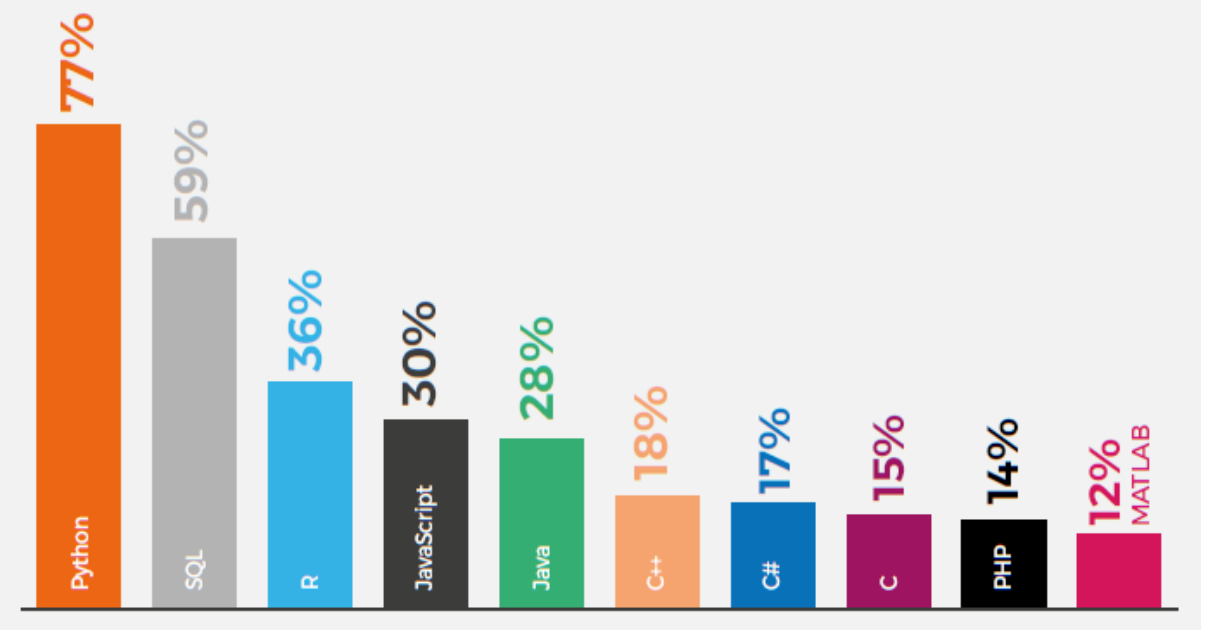
- ***Planificación:*** Definición del proyecto y sus resultados.
- ***Construcción del modelo de datos:*** Partiendo de una variedad de bibliotecas de código abierto o herramientas para construir los modelos.
- ***Evalución del modelo:*** En pro de lograr un alto porcentaje de exactitud en sus modelos antes de poder implementarlos con confianza.
- ***Explicar los modelos:*** Retroalimentación automática sobre el modelo.
- ***Implementción del modelo:*** Tomar un modelo de aprendizaje automático entrenado e implementarlo en los sistemas correctos.
- ***Monitorear los modelos:*** Los modelos siempre deben monitorearse después de la implementación para garantizar el funcionamiento correcto.

El científico de datos

El científico de Datos es aquel que puede crear puentes entre los datos crudos y el análisis haciéndolos accesibles . Es un rol democratizarte en la medida que lleva los datos a la gente común, haciendo el mundo un poco mejor paso a paso.

Los científicos de datos están involucrados con el agrupamiento de datos desde distintas fuentes, su edición en formas mas tratables y entendibles de forma que cuenten una historia que pueda ser presentada por ellos para ser entendida por todos

Herramientas de tecnología



Lenguajes de programación

Herramientas de tecnología

Librerías o frameworks