

## ESTRUCTURA

# Introducción a la ciencia de datos

1. *¿Qué es la ciencia de datos?*
2. *Historia*
3. *Evolución de la ciencia de datos*
4. *Componentes de la ciencia de datos*

### ***¿Que es la ciencia de datos?***

#### ***Definición***

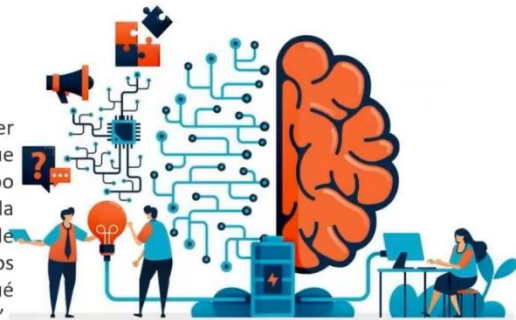
*Ciencia: Conjunto de conocimientos obtenidos mediante la observación y el razonamiento , sistemáticamente estructurados y de los **que** se deducen principios y leyes generales con capacidad predictiva y comprobables experimentalmente .*



### ***¿Que es la ciencia de datos?***

#### ***Definición***

“La ciencia de datos es el estudio de datos con el fin de extraer información significativa para empresas. Es un enfoque multidisciplinario que combina principios y prácticas del campo de las matemáticas, la estadística, la inteligencia artificial y la ingeniería de computación para analizar grandes cantidades de datos. Este análisis permite que los científicos de datos planteen y respondan a preguntas como “qué pasó”, “por qué pasó”, “qué pasará” y “qué se puede hacer con los resultados”.





## Precepto



## Historia

Establecer la historia para la ciencia de datos es una labor compleja, quizás es un termino que viene cobrando una alta relevancia en los últimos 15 – 20 años, pero si nos damos cuenta y hacemos el análisis consciente de la información con la que contamos referente a la ciencia de datos, el resultado es que la ciencia de datos es tan antigua como la humanidad. Con el nacimiento del Homo Sapiens Sapiens y unas generaciones anteriores ya estabamos usando la ciencia de datos para suplir necesidades.

## Historia

Un hito importante y quizás el mas importante dentro de la era de ciencia de datos es la creación del computador, en 1941 se crea el primer computador gracias al ingeniero civil Konrad Zuse, en ese momento el computador lograba almacenar 64 palabras.

En la segunda guerra mundial, los aliados en cabeza de EEUU logran minimizar el derribamiento de sus aeronaves gracias al análisis de la información existente.

**Lo letal es aquello que no se observa**



## Historia

De acuerdo con el informe "Data Never Sleeps", para el año 2020, se tenia un estimado de que una persona genera 1.7 MB de información por segundo es decir en un día una persona genera aproximadamente 147 gb de información.

La anterior estimación esta dimensionada única y exclusivamente para personas, ¿pero cuanta mas información se genera?, de acuerdo a estimaciones de portales especializados, para el año 2025 se crearan 463nexasbytes de información diariamente. Para guardar esta información se requieren aproximadamente 213 millones de DVD.

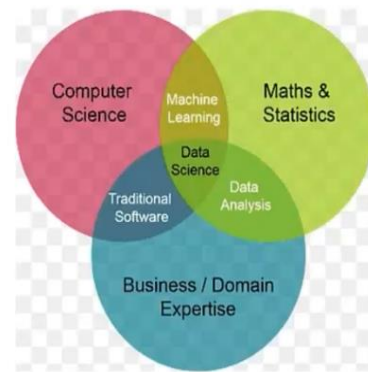
**Lo letal es aquello que no se observa**

## Componentes

El panorama a nivel macro nos deja ver que la ciencia de datos es la intersección de varios campos:

- Ciencias computacionales.
- Estadística y Matemáticas.
- Áreas y expertos de negocio.

La interacción entre estas áreas siempre va a ser en dos vías, es decir se deben crear o reforzar los canales de comunicación entre las mismas para que el proceso asociado a la ciencia de datos sea un éxito.

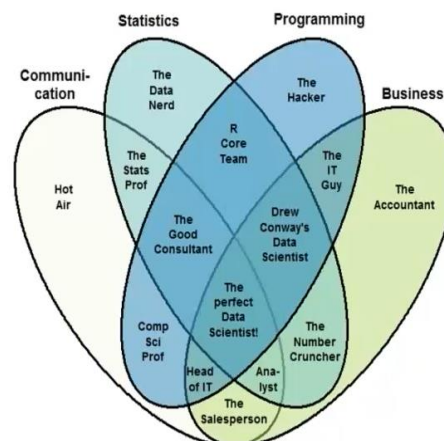


## Componentes

Según las estadísticas de Kaggle, entre el 70 y 85% de los proyectos asociados a ciencia de datos fracasan, si esta tasa de fracaso se presentara en las primeras fases de los proyectos quizás no seria tan preocupante, pero cuando la tasa de fracaso se presenta en fases avanzadas, quizás el menor de los problemas es dar cierre definitivo al proyecto, los problemas relevantes son:

- ¿Cuanto dinero me gaste durante el proceso?.
- ¿Cuanto dinero perdí por no alcanzar el objetivo?.
- ¿Que talento agregue al equipo para el desarrollo del proyecto?
- ¿El equipo que agregue pudo haber ejecutado alguna otra tarea que se desarrollara con éxito?

The Data Scientist Venn Diagram

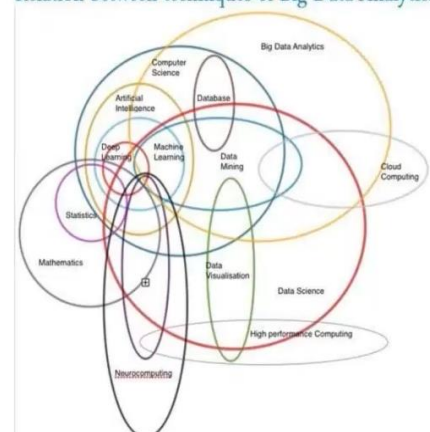


## Componentes

Como podemos observar las componentes asociadas a un proceso de ciencia de datos muchas y cada vez que hacemos ZOOM se va haciendo mas grande, entonces la pregunta es ¿Qué requiero para dar ese primer paso?

- Objetivo claro.
- Datos acorde al objetivo planteado, yo no puedo crear un cohete con datos relacionados a vías terrestres.
- Experto en negocio, es decir aquella persona que puede o no tener una formación técnica en el área, pero lo que si debe tener es un conocimiento pleno del negocio.
- Iniciemos por el ¿que paso? Y el ¿que esta pasando?, para esto necesitamos KPIs claros y personas con conocimiento en análisis descriptivo de información (BI).
- ¿Qué va a pasar?, ahora si entra el grupo de personas especialistas en predicción.
- ¿Qué se puede hacer?, especialistas en negocio, en ciencia de datos.

Relation between techniques of Big Data Analytics





## Componentes

*¿Qué paso y que esta pasando?, es la fase inicial dentro del camino de la ciencia de datos, típicamente se apoya en análisis de tipo descriptivo, por ejemplo: cuanto es la factura promedio en productos de aseo, cada cuanto tiempo mi cliente hace una recompra, cual es la sintomatología de un paciente que sufrió un desmayo, hacer un buen ejercicio descriptivo, no solo nos ayuda a comprender el problema que estamos abarcando, también ayuda a tomar decisiones y quizás encontrar soluciones menos complejas y costosas.*



## Componentes

*¿Qué va a pasar?, es la segunda fase dentro del camino de la ciencia de datos, es la fase donde una vez se ha planteado un objetivo claro se busca “predecir” ese futuro, sin embargo, no apunta únicamente a la predicción del futuro, este modelo puede ir enfocado en recomendar un nuevo producto a clientes pertenecientes a un nicho de mi mercado, enviar una serie de cupones a clientes donde sabemos que el cliente requiere los productos y adicional son los productos mas rentables para la organización o evidentemente predecir el futuro del precio del dólar al cierre del año y la acción de ecopetrol.*

*En esta fase también encontramos modelos preventivos, como por ejemplo determinar el riesgo de infarto de una persona con características puntuales, estimar el riesgo de no pago de una hipoteca.*



## Componentes

*¿Qué se puede hacer?, es la tercera fase dentro del camino de la ciencia de datos, es la fase donde una vez se han construido los diferentes modelos, corroborado y desplegado, se toman las acciones necesarias que persiguen un nuevo objetivo, es decir son las decisiones y estrategias que se plantean a partir de los modelos descriptivos o predictivos.*



## Componentes

*Minería de datos, esta es una componentes transversal a los procesos anteriores y cíclica, es un proceso que nunca se puede ni se debe dejar de lado y cobra una relevancia muy alta en la salida avante de cualquier proyecto enfocado en ciencia de datos.*

