2. Conceptos básicos: **Ciencia de Datos**

Parte 1

2.1 Situación actual: Crecimiento en los datos (IDC, 2018)

El uso de datos está transformando la forma en la que vivimos.

Transformación Digital

Es la integración de datos inteligentes en todo lo que hacemos.

Evolución de dispositivos.

Los datos son el corazón de la transformación digital.

Dependencia en datos aumentará a medida que las empresas: capturen, cataloguen y saquen provecho de los datos en cada paso de su cadena de suministro.

Datos

Empresa

- Mejorar la experiencia del cliente.
- Encontrar nuevos mercados.
- Nuevos niveles de eficiencia.
- Productos adaptables a cada perfil.
- Procesos más productivos.
- Empleados más productivos.
- Nuevas fuentes de ventaja competitiva.

Clientes

- Construir relaciones más profundas.
- Acceder a productos y servicios más rápido (cuando sea y donde sea)

2.1 Situación actual: Crecimiento en los datos (IDC, 2018)

El mundo basado en datos estará siempre: activo, rastreando, monitoreando, escuchando y observando, porque siempre estará aprendiendo.

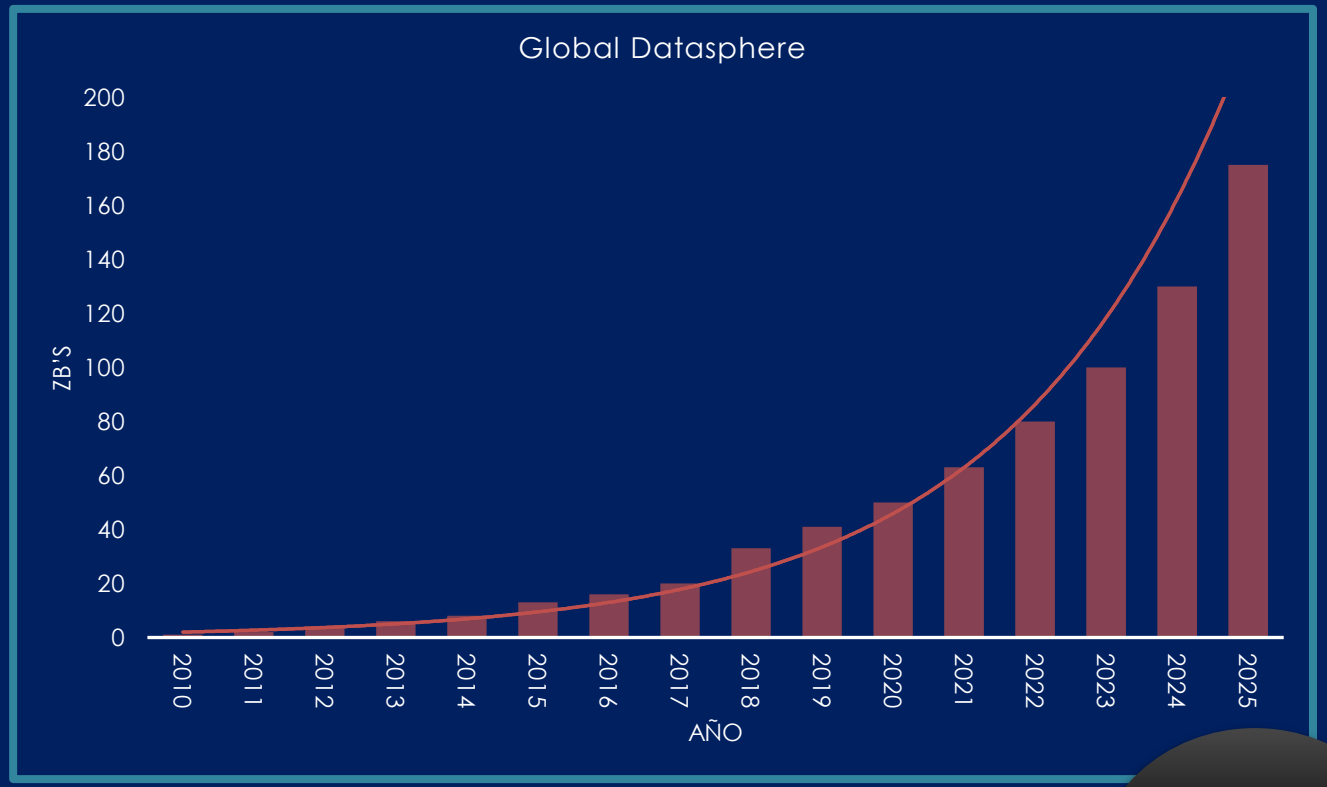
Estadísticas

Datos que son creados, capturados y replicados, provenientes de diversas fuentes.

Global Datasphere

Predicción:
2018 - 33 Zettabytes (ZB)
2025 - 175 ZB

1 ZB = 1e+12 Gigabytes
DVD's = 222 vueltas al mundo



Sólo el 0.5% de los datos se usa y analiza.

2.2 ¿Qué es Ciencia de Datos?

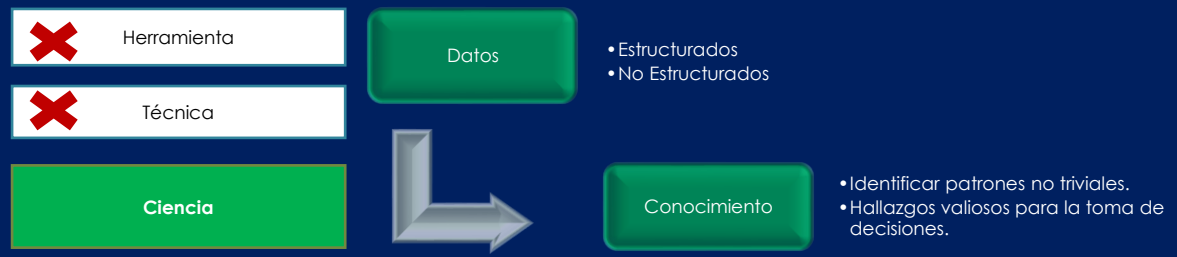
Data Science ≠ Big Data

Data Science (Ciencia de Datos)

La Ciencia de Datos abarca un conjunto de principios, definiciones de problemas, algoritmos y procesos para extraer patrones no obvios y útiles de grandes conjuntos de datos. (John D. Kelleher y Brendan Tierney, 2018)

La práctica de convertir datos en bruto en hallazgos valiosos que habiliten acciones informadas. – Mad Gee (DSaPP)

Estudio fenomenológico de Sistemas Complejos Adaptativos, con el propósito de construir productos de datos que ayuden / soporten a la toma de decisiones y acciones sobre el sistema. –Adolfo de Unanue (ITAM)



Big Data

Enormes volúmenes de datos que no pueden ser tratados de manera convencional, ya que superan los límites y capacidades de las herramientas tradicionales para la captura, gestión y procesamiento de datos

Descripción en términos de:

UTILIZAR CIENCIA DE DATOS EN BIG DATA.
Data Science ≠ Big Data

2.2 ¿Qué es Ciencia de Datos?

Todo empieza con una Pregunta

La ciencia de datos sólo es útil cuando se utilizan los datos para responder una pregunta (Mármol, 2017).

- ¿Quién? — ¿Quién es más propenso a no pagar?
- ¿Qué? — ¿Qué productos debo de vender?
- ¿Dónde? — ¿Dónde hay que abrir una nueva sucursal?
- ¿Cuándo? — ¿Cuándo debo de hacer nuevos pedidos?
- ¿Por qué? — ¿Por qué las ventas se comportan así?
- ¿Cuánto? — ¿Cuántos productos debo de vender para que no haya pérdidas?

2.2 ¿Qué es Ciencia de Datos?

Tipos de Tareas

Regresión

•Predicción de un valor numérico.

Clasificación

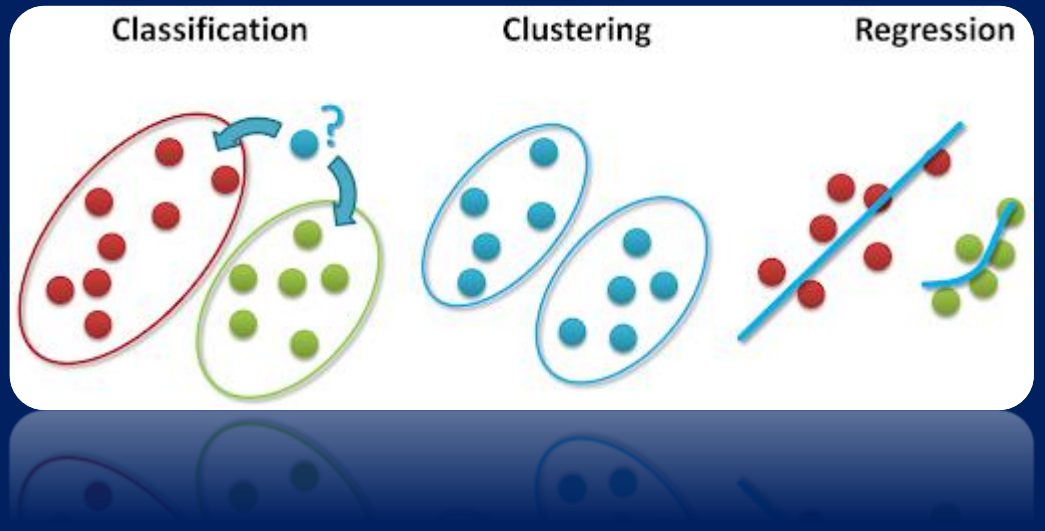
¿Dónde pertenece un individuo?
(Dar ejemplo de Walmart)

Scoring

Probabilidad de que pertenezca a una clase.

Agrupamiento (Clustering)

Agrupar individuos por sus similitudes



2.2 ¿Qué es Ciencia de Datos?

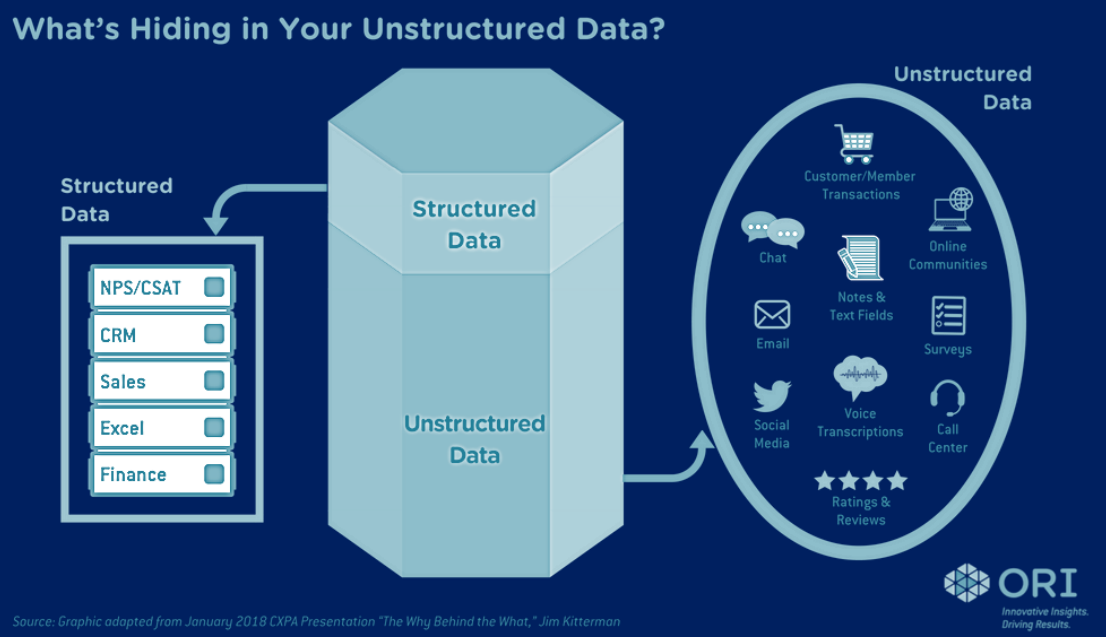
Tipos de Datos

Estructurados

- Datos que pueden ser almacenados en una tabla.
- Fácil (normalmente) de almacenar, organizer, buscar, reordenar y de combinar con más datos en esta misma estructura.
- Aplicar Ciencia de Datos a este tipo de datos es, relativamente, fácil, poque ya se encuentra en un formato adecuado para su registro analítico.

No Estructruados

- Correos, imágenes, videos, audio, etc.
- A menudo podemos extraer datos estructurados de datos no estructurados utilizando técnicas de inteligencia artificial.



¿Los datos están disponibles?
¿Los datos me ayudan a resolver mi pregunta?
¿Los datos son suficientes?
¿Calidad? ¿Buena? ¿Mala?

2.3 Perfil de un Científico de Datos

Data People (Personas de Datos)

Científico de Datos

Expertos en:
Programación, matemáticas, estadísitica y comunicación.

Actividades:
Colectan y limpian datos, manejan grandes cantidades de datos, crean modelos matemáticos, e interpretan sus resultados en soluciones comerciales.

Herramientas / Conocimientos generales:
Python, R, SQL y Algoritmos de Aprendizaje de máquina

Extra:
Excelentes Habilidades de colaboración y comunicación.

Ingeniero de Datos

Expertos en:
Desarrollo de software, bases de datos y programación

Actividades:
Diseñan, construyen, integran y mantienen los datos de múltiples fuentes. También, extraen, transforman y cargan datos.

Herramientas /Conocimientos generales:
SQL, NoSQL Apache Spark y Hadoop, Python, R, Java y C / C ++.

Extra:
No tienen ningún papel en el análisis de datos; propósito es hacer que los datos estén disponibles para uso interno.

.

Analista de Datos

Expertos en: Negocios

Conocimientos solidos en: Programación y estadísitica.

Actividades:
Crear informes de inteligencia empresarial (para uso interno y clientes). Principalmente, son los encargados de ayudar a aquellos que necesitan comprender consultas específicas con gráficos.

Heramientas/ / Conocimiento:
Microsoft Excel, SQL y Tableau, así como Python y / o R, y conocimiento de la organización,

Ingeniero de Software

Expertos: Programación.

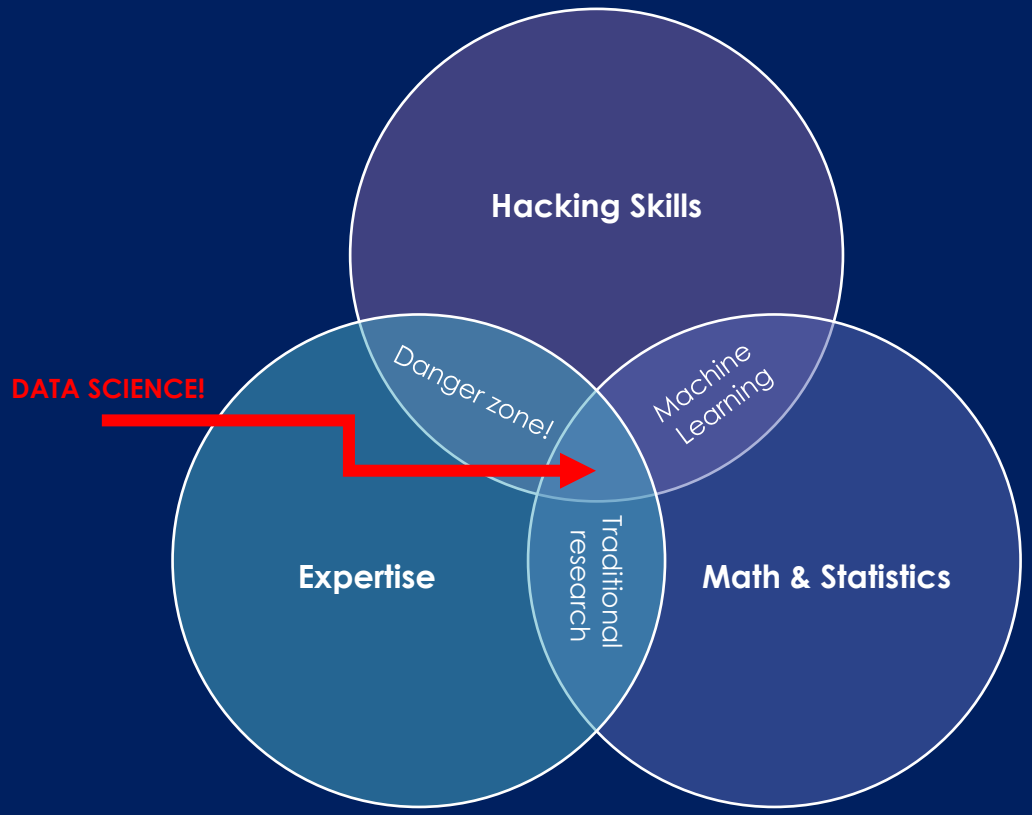
Actividades:
Desarrollar sistemas operativos, diseñar de software, y desarrollar aplicaciones móviles.
Son responsables de crear el sistema que importa y almacena los datos, ya sea un sitio web, un software especializado o una aplicación.

Herramientas / Conocimientos generales:
SQL, diferentes lenguajes de programación (Python, R, Java y C / C ++), cómputo en la nube, etc.

Extra:
Los ingenieros de software también se denominan a veces ingenieros de infraestructura o plataforma.

2.3 Perfil de un Científico de Datos

Perfil de un Científico de Datos



Conocimiento/Habilidades

Experiencia:
Negocios: Comprensión del negocio.

Habilidades Computacionales:
Programación (diferentes lenguajes), administración de sistemas, computo en la nuebe.

Matemáticas:
Cálculo, algebra, optimización, simulación.

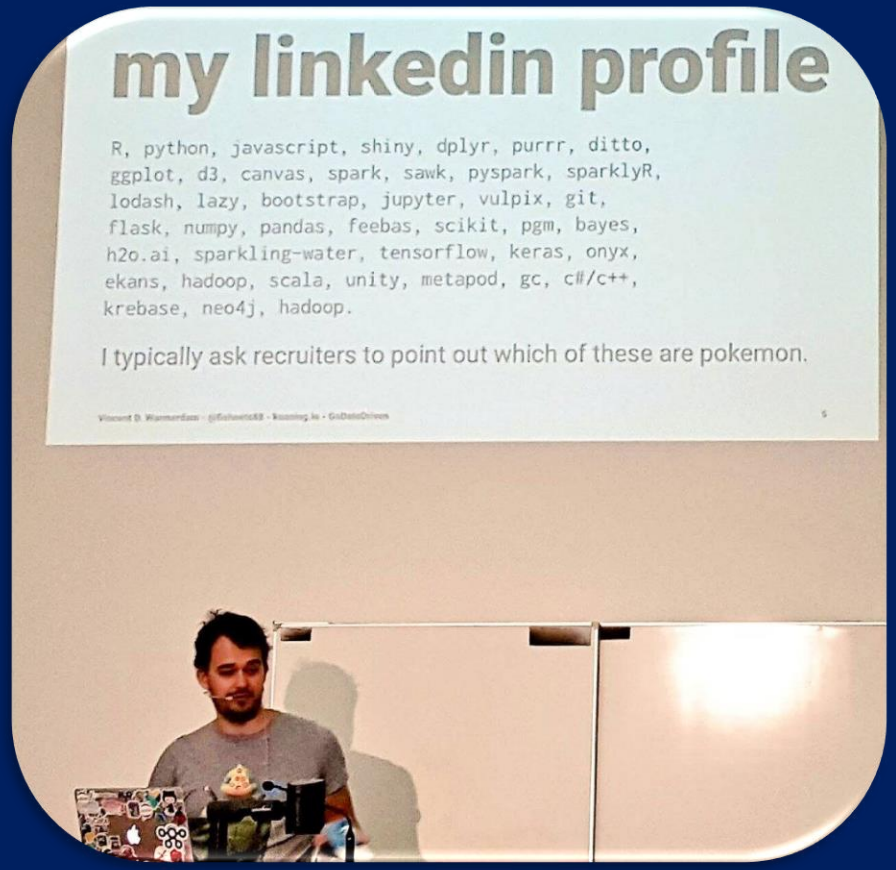
Aprendizaje de máquina/Big Data:
Tipos de datos, distribuir datos, algoritmos de aprendizaje de máquina (supervisados, no supervisados).
Cálculo, algebra, optimización, simulación.

Estadística::
Probabilidad, series de tiempo, muestreo, estadística tradicional, estadística bayesiana.

Storytelling:
Compartir resultados.

2.3 Perfil de un Científico de Datos

Herramientas



Shiny
by RStudio

Pokémon

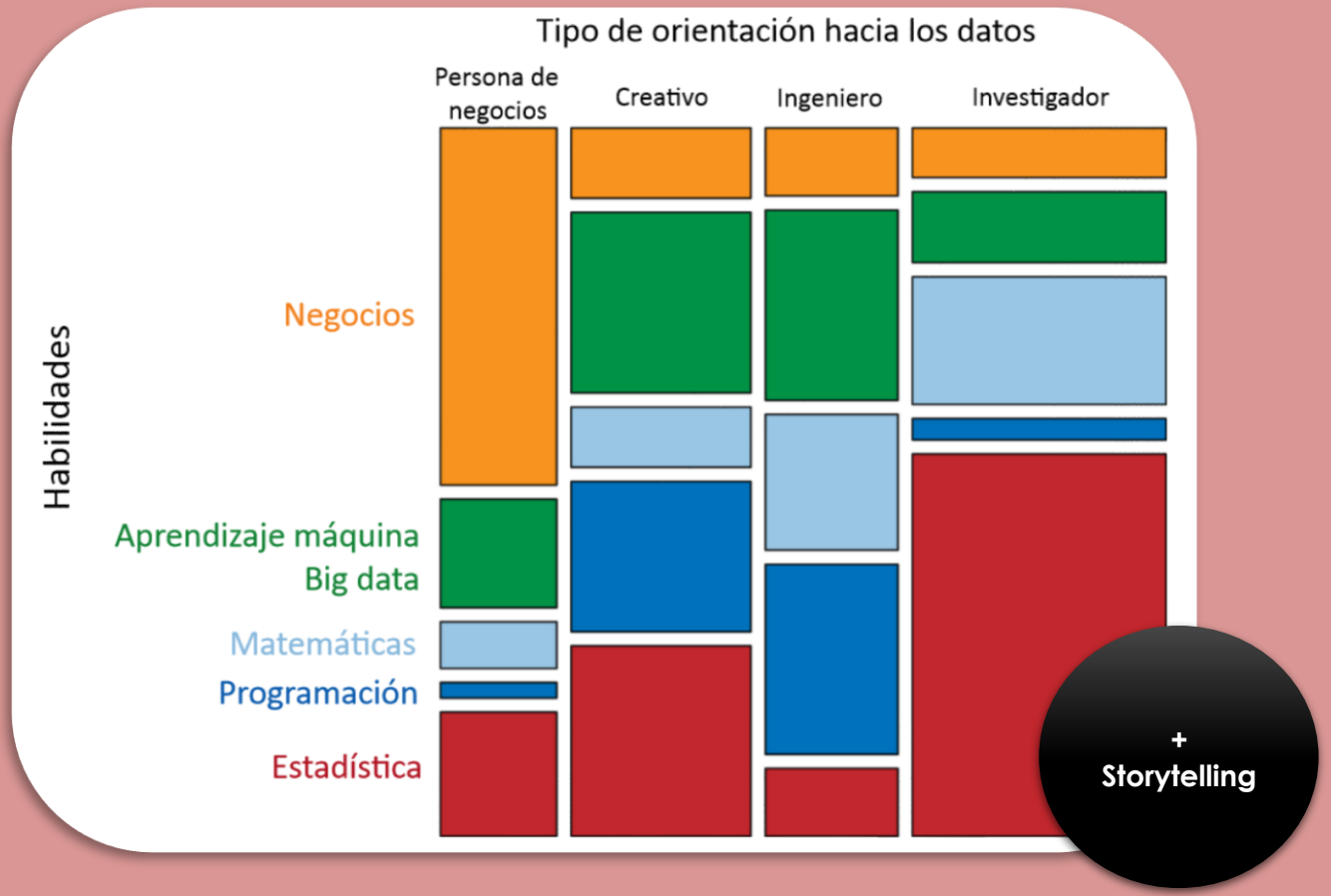
2.3 Perfil de un Científico de Datos

Actividad 2: Identificar Fortalezas dentro del Perfil de un Científico de Datos.

Fecha límite: Miércoles 23:59 hrs.

Instrucciones:

1. Hacer una visualización que permita mostrar la distribución de sus aptitudes en las áreas de: Negocios, Aprendizaje de máquina (Big Data), Matemáticas, Programación, Estadística y Storytelling.
2. Subir la visualización a PowerPoint:
3. Subir la visualización a Canvas. (Actividad para calificación)

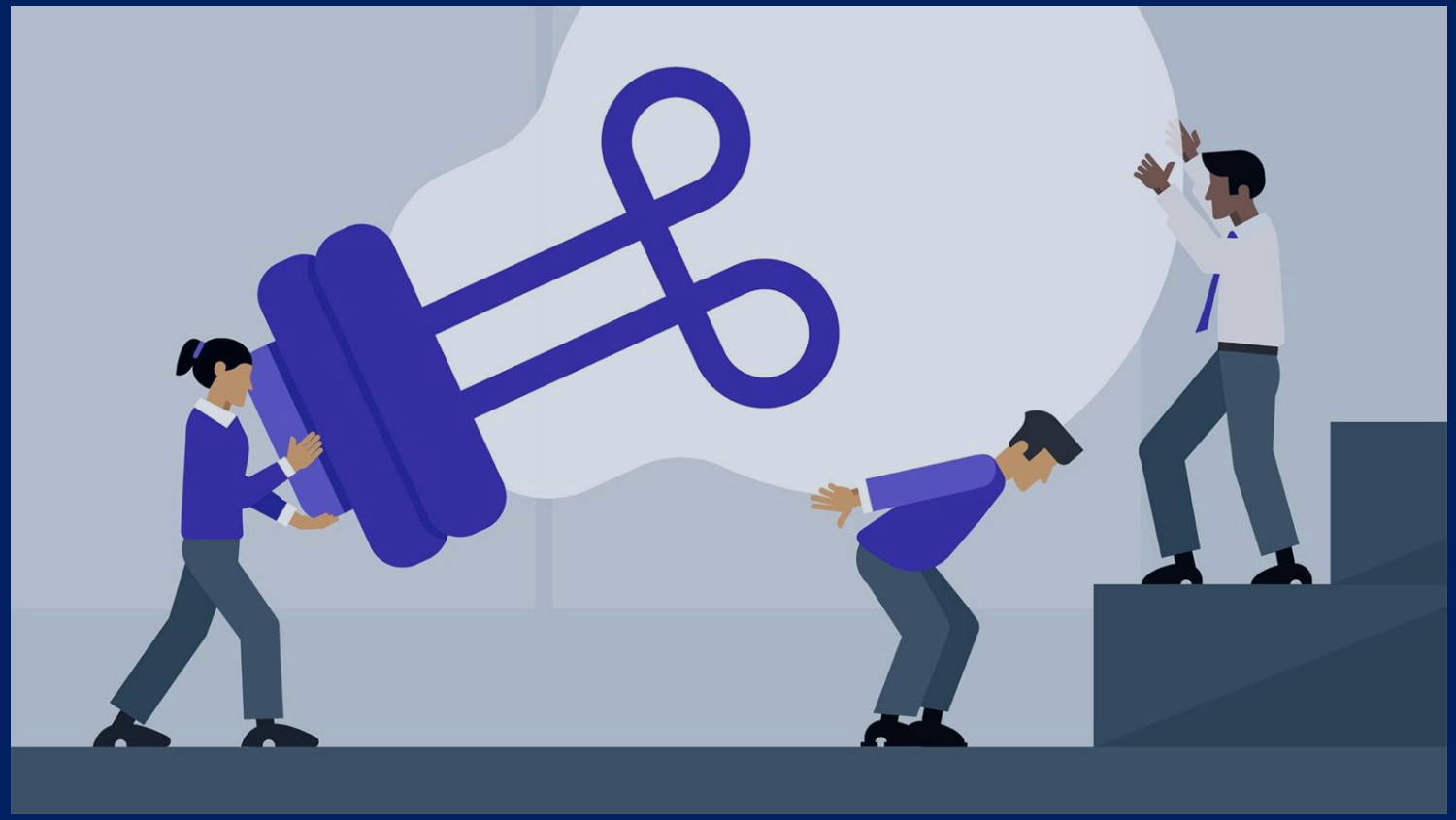


2.3 Perfil de un Científico de Datos

Trabajo en Equipo

Lo mejor es contar con un equipo diverso:

- Ingenieros en Sistemas
- Ingenieros Industriales
- Economistas
- Financieros
- Físicos
- Expertos en el tema



2.3 Perfil de un Científico de Datos

Actividad 3: Selección de Equipos para el Proyecto Final.

Fecha límite: Miércoles 23:59 hrs.

Instrucciones:

1. Considerando la información que subieron tus compañeros en la actividad 2, crea tu equipo para el trabajo final (2-3 integrantes).
2. En el siguiente documento subir el nombre de los integrantes de su equipo:
3. Subir a canvas un documento en la "Actividad 3" con la siguiente información:
   - Nombre completo de los integrantes .
   - Matrícula de los integrantes .
   - Correo institucional de los integrantes .
   - ¿Cómo es que se complementan con respecto a sus habilidades?

Bibliografía

- Kelleher, J. D., & Tierney, B. , Data Science, United States : MIT Press , 2018.
- El Economista. (2016). Big Data. 15 de agosto de 2020, de El Economista Sitio web: https://www.eleconomista.es/diccionario-de-economia/big-data
- Mármol, J. (2018). Introducción. 15 de agosto de 2020, de ITAM Sitio web: https://github.com/AnaLuisaMasetto/intro-to-data-science-2018
- Reinsel, D. Gantz, J and Rydning, J. (2018). The Digitization of the WorldFrom Edge to Core. 15 de agosto de 2020, de Seagate Sitio web: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf
- Significados. (s.f.). ¿Qué es ciencia? . 15 de agosto de 2020, de Significados Sitio web: https://www.significados.com/ciencia/
- Yuste, B. (2018). En 2025 el volumen de datos en el mundo será 175 veces más que en 2011. 15 de agosto de 2020, de Bankinter Sitio web: https://www.fundacionbankinter.org/blog/noticia/en-2025-el-volumen-de-datos-en-el-mundo-sera-175-veces-mas-que-en-2011
- DeSantis, G. (2019). Data Scientist vs. Data Analyst vs. Data Engineer. 15 de agosto de 2020, de Medium Sitio web: https://medium.com/@gdesantis7/data-scientist-vs-data-analyst-vs-data-engineer-bd4868f9b31e
- Simplilearn. (2020). Data Scientist vs Data Analyst vs Data Engineer: Job Role, Skills, and Salary. 15 de agosto de 2020, de Simplilearn Sitio web: https://www.simplilearn.com/tutorials/data-science-tutorial/data-scientist-vs-data-analyst-vs-data-engineer
- Thinkful. (s.f.). Data Engineer vs Software Engineer. 15 de agosto de 2020, de Thinkful Sitio web: https://www.thinkful.com/blog/data-engineer-vs-software-engineer/