INSTITUCIÓN TECNOLÓGICA CEDESISTEMAS TECNOLOGÍA Y FUTURO

La Formación cambia tu Vida

cetia Formación y Consultoría

Bienvenidos

Machine Learning

Introducción a Python (Ciencia de Datos)

Oscar Andres Gaspar Alvarez

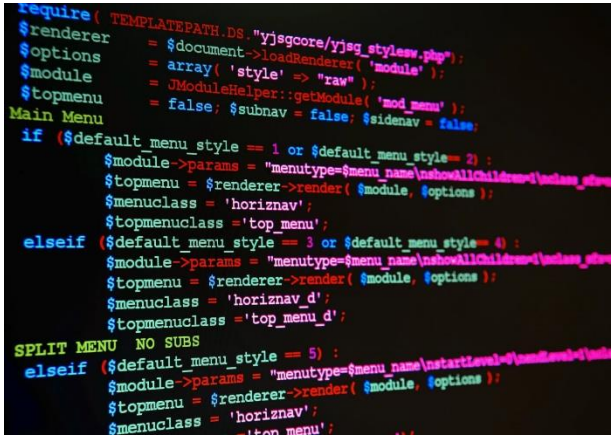
oscar.gaspar@cedesistemas.edu.co

Metodología de clase

Exponer los objetivos de la sección.

Teoría : Fundamentos básicos y conceptuales

Practica: Practica en Python.



Presentación del Modulo

1) Introducción a ciencia de datos
- El rol del científico de datos en las compañías
- Metodologías de proyectos de Analítica en las empresas
- Inteligencia artificial

2) Introducción a Python
- Fundamentos básicos de programación
- Librerías para ciencia de datos

3) Análisis descriptivo y exploratorio
- Tipos de Variables
- Valores Atípicos
- Gráficas exploratorias
- Inferencia de parámetros poblacionales
- Pruebas de Hipótesis
- Bondad de Ajuste

Presentación del Modulo

4) Aprendizaje Automatico

- Modelos de clasificación
- Modelos de Regresión
- Redes Neuronales
- Metodos de Ensambles
- Modelos No supervisados

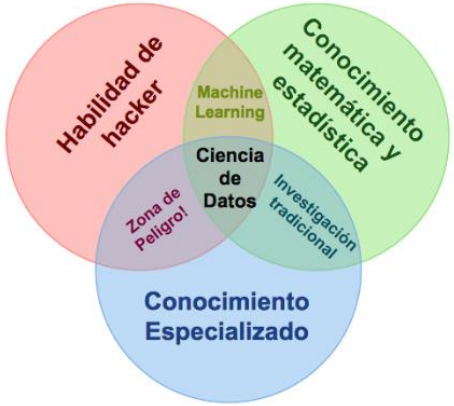
5) Modelos de Series de Tiempo

6) Procesamiento lenguaje Natural

¿Qué es Ciencia de Datos?

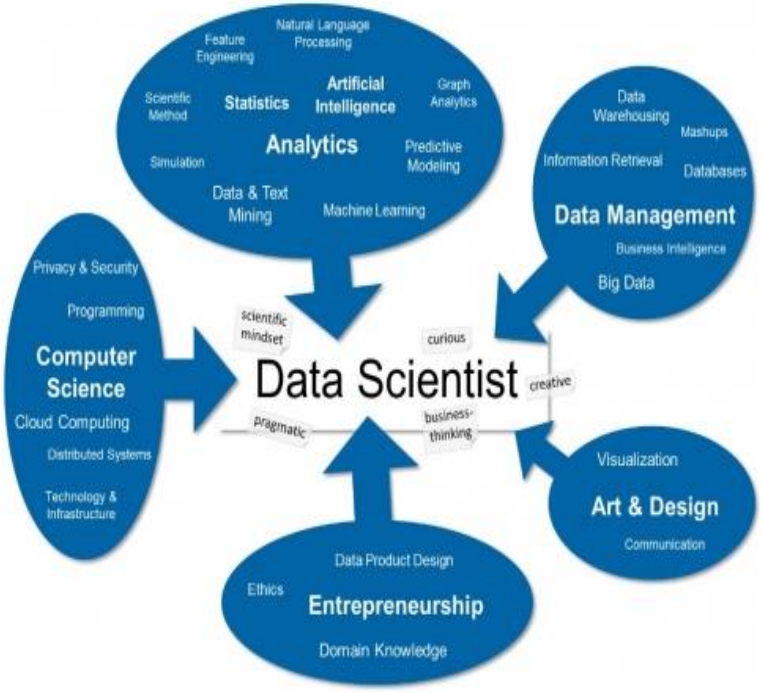
¿ Que es ciencia de Datos?

- Un término que realmente nadie definió.
- “Un científico de datos es un estadístico que puede programar”
- Journal of Data Science del 2003, en donde hacen una definición “muy precisa” diciendo: “Por 'Ciencia de Datos' queremos referirnos a casi todo lo que tiene algo que ver con los datos”
- Diagrama de Venn de Conway.
- Wikipedia : La ciencia de datos es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados,1 lo cual es una continuación de algunos campos de análisis de datos como la estadística, la minería de datos, el aprendizaje automático, y la analítica predictiva.
- “La ciencia de datos es la disciplina de hacer que los datos sean útiles.” Cassie Kozyrkov Head of Decision Intelligence, Google.



Conocimientos necesarios para el mundo empresarial

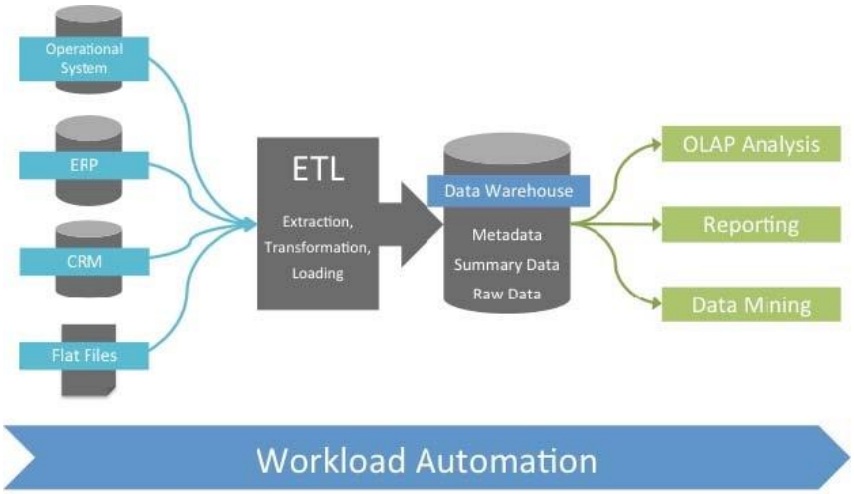
- Un Data Scientist es un experto en Data Science (Ciencia de datos), su trabajo consiste en extraer conocimiento a partir de los datos.




Are data scientists unicorns?

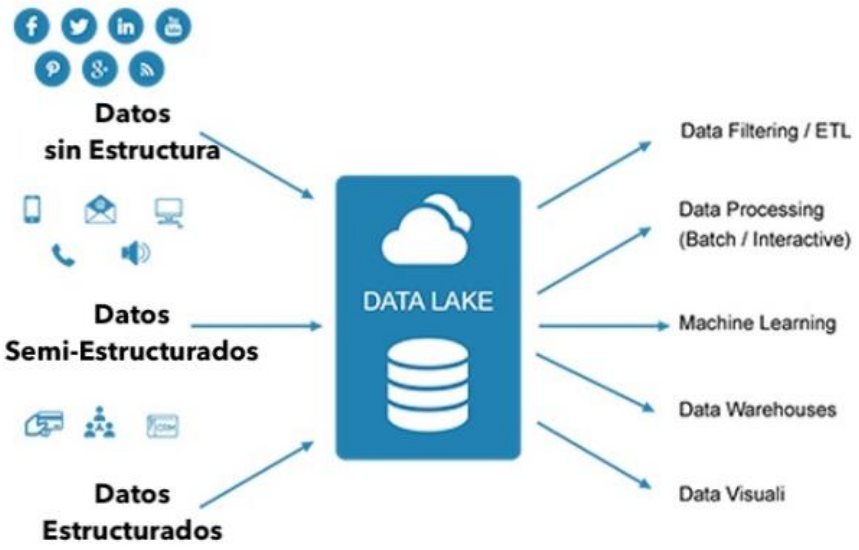
Gestión de Datos

- "**La Inteligencia de Negocio (BI)** es un término genérico que incluye las aplicaciones, la infraestructura y las herramientas, y las mejores prácticas que permiten el acceso y el análisis de la información para mejorar y optimizar las decisiones y rendimiento."
- **Un data warehouse** se aloja en un servidor corporativo o cada vez más, en la nube. Los datos de diferentes aplicaciones de procesamiento de transacciones Online (OLTP) y otras fuentes se extraen selectivamente para su uso por aplicaciones analíticas y de consultas por usuarios.
- **Datos estructurados:** Son archivos de tipo texto que se suelen mostrar en filas y columnas con títulos



Business Analytics (BA)

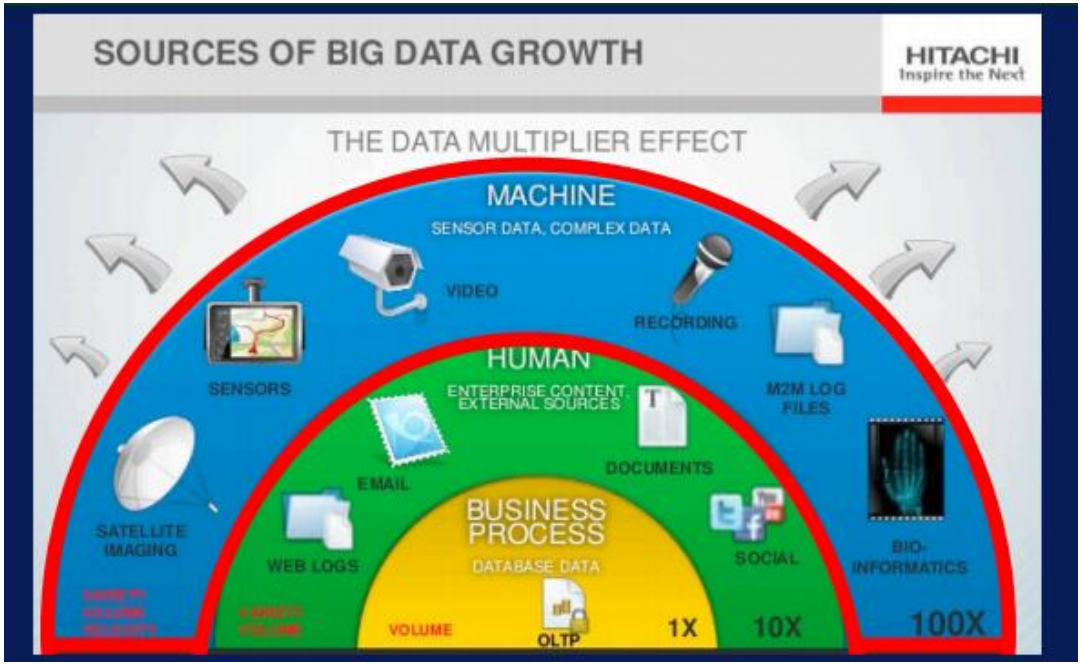
- **Business Analytics** es el proceso de recopilación, clasificación, procesamiento y estudio de datos comerciales, y el uso de modelos estadísticos y metodologías iterativas para transformar los datos en información comercial.
- Datos no estructurados y semi estructurados.
- Modelo en time real.
- Modelos dinámicos



Ciencia Computacional en Big Data

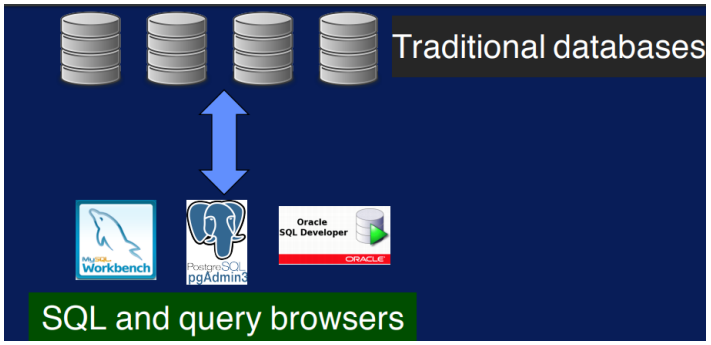
Este campo es distinto a la informática, y a la teoría y experimentación, que son las formas tradicionales de la ciencia y la ingeniería. El enfoque de la computación científica es para ganar entendimiento, principalmente a través del análisis de modelos matemáticos implementados en computadores.

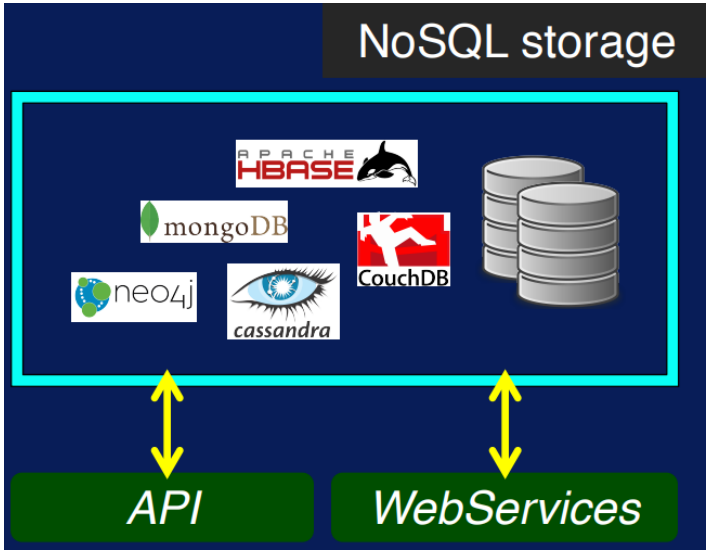
Científicos e ingenieros desarrollan software, aplicaciones informáticas para modelar sistemas que están siendo estudiados, y correr estos programas con diferentes conjuntos de entradas.

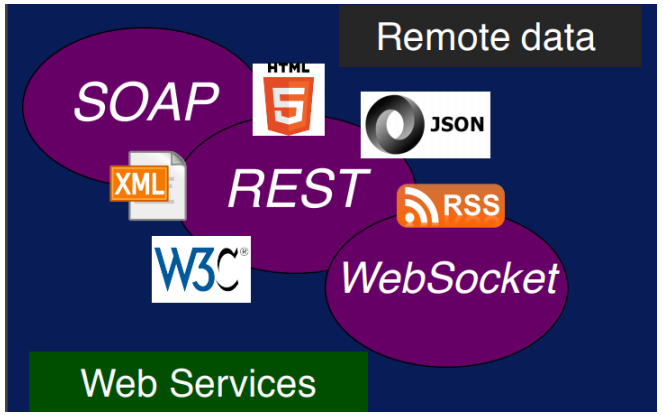


Ciencia de Datos para dar valor al Big Data

Captando datos

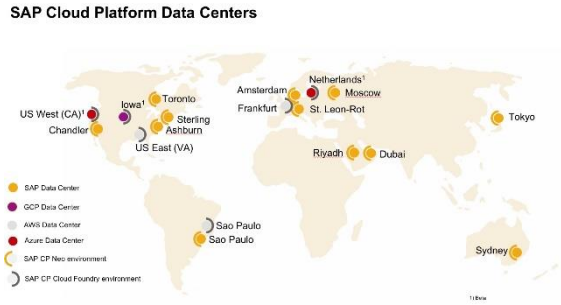




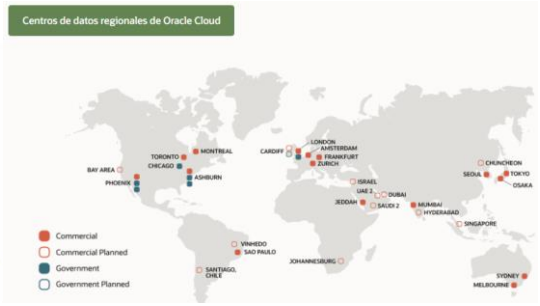


Ejemplos en la actualidad

SAP



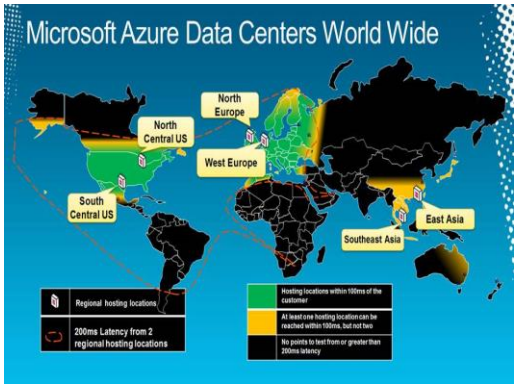
Oracle



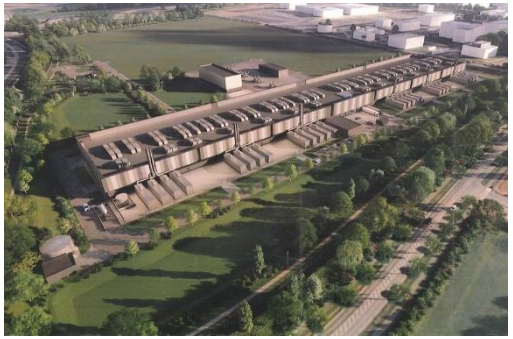
IBM

Google

Microsoft



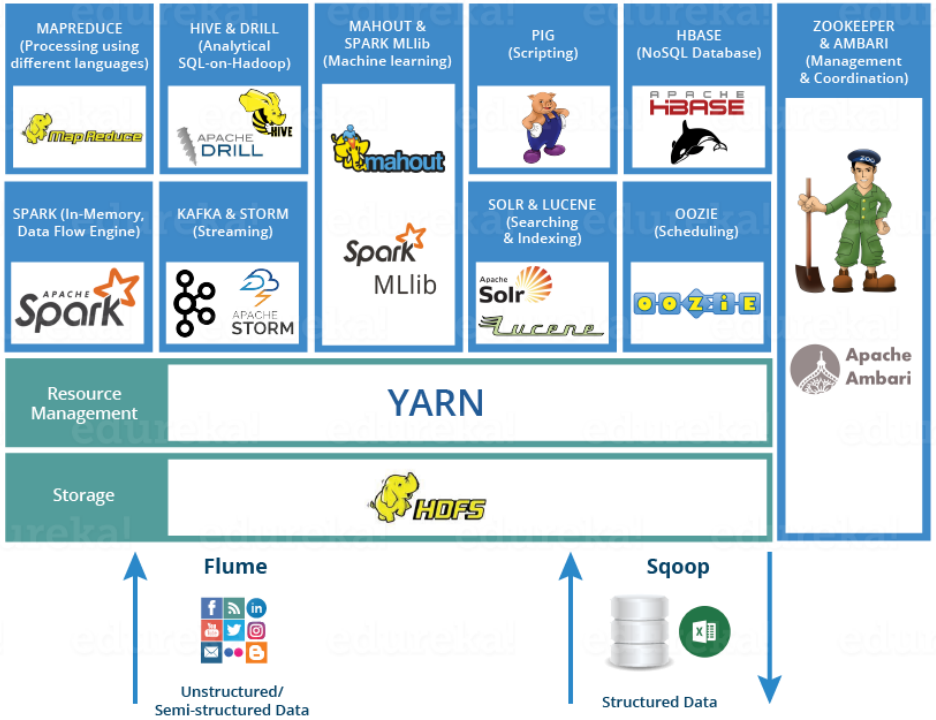
Amazon



Hadoop

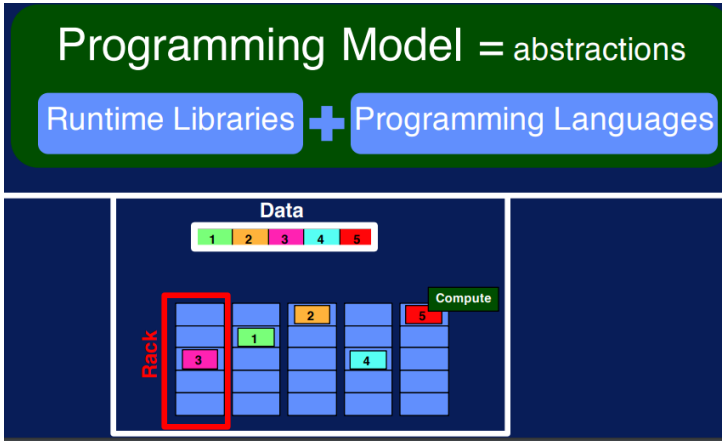
- Apache Hadoop es un framework de software que soporta aplicaciones distribuidas bajo una licencia libre.1 Permite a las aplicaciones trabajar con miles de nodos y petabytes de datos. Hadoop se inspiró en los documentos Google para MapReduce y Google File System (GFS).

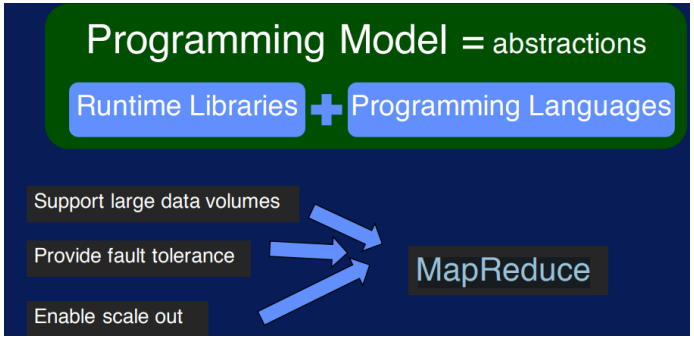
- Hadoop es un proyecto de alto nivel Apache que está siendo construido y usado por una comunidad global de contribuyentes,2 mediante el lenguaje de programación Java. Yahoo! ha sido el mayor contribuyente al proyecto,3 y usa Hadoop extensivamente en su negocio.4

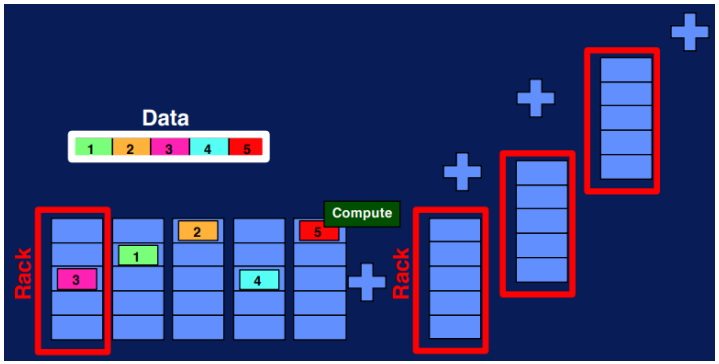


Programando Modelos Big Data

- Programabilidad de Sistemas de archivos distribuidos.
- Dividir Volumen
- Acceso Rápido de Datos
- Distribuir cálculos a nodos.
- Tolerancia a Fallas
- Replicar particiones de datos.
- Recupere archivos cuando sea necesario.



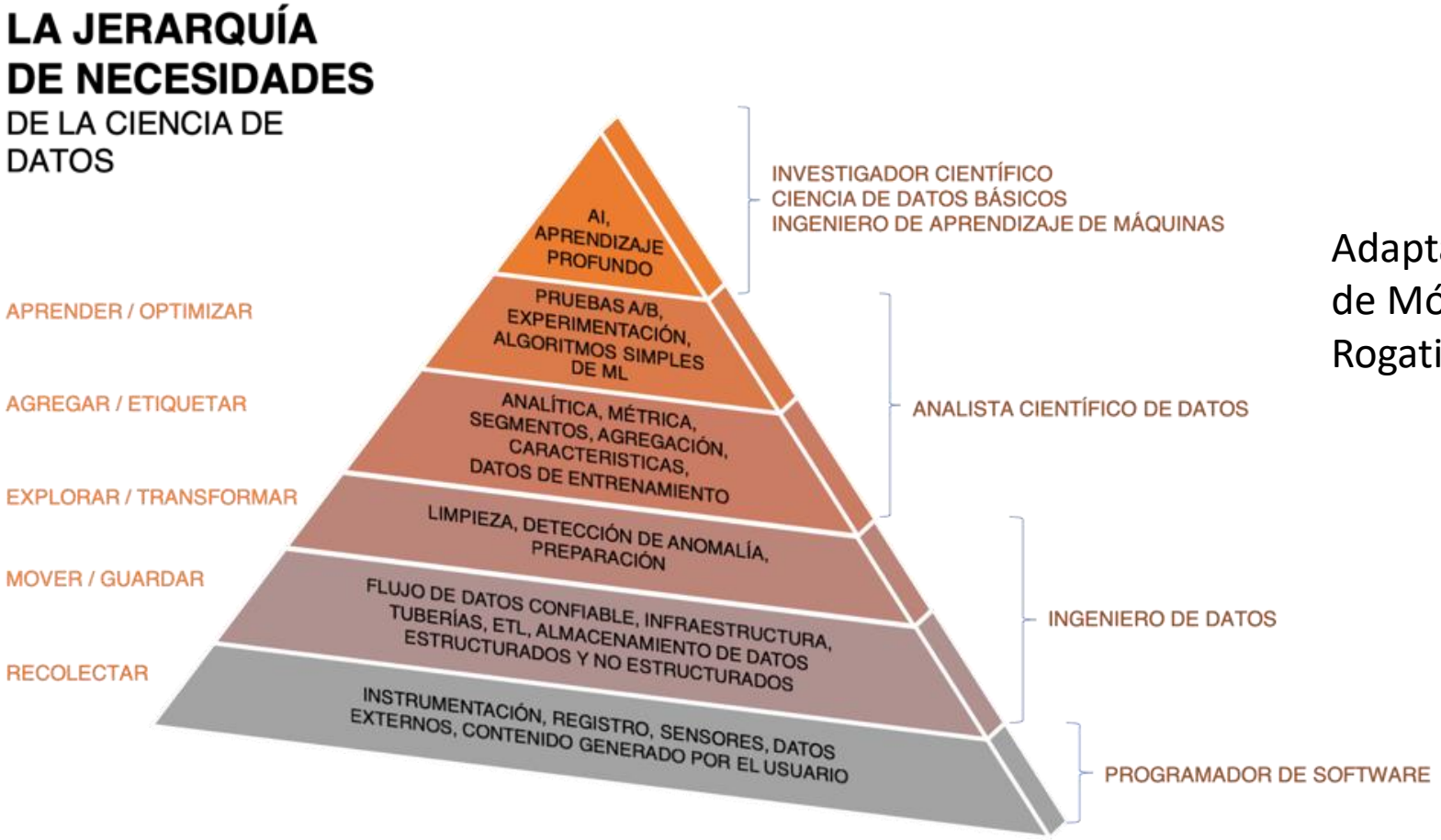




El rol del científico de datos en las compañías

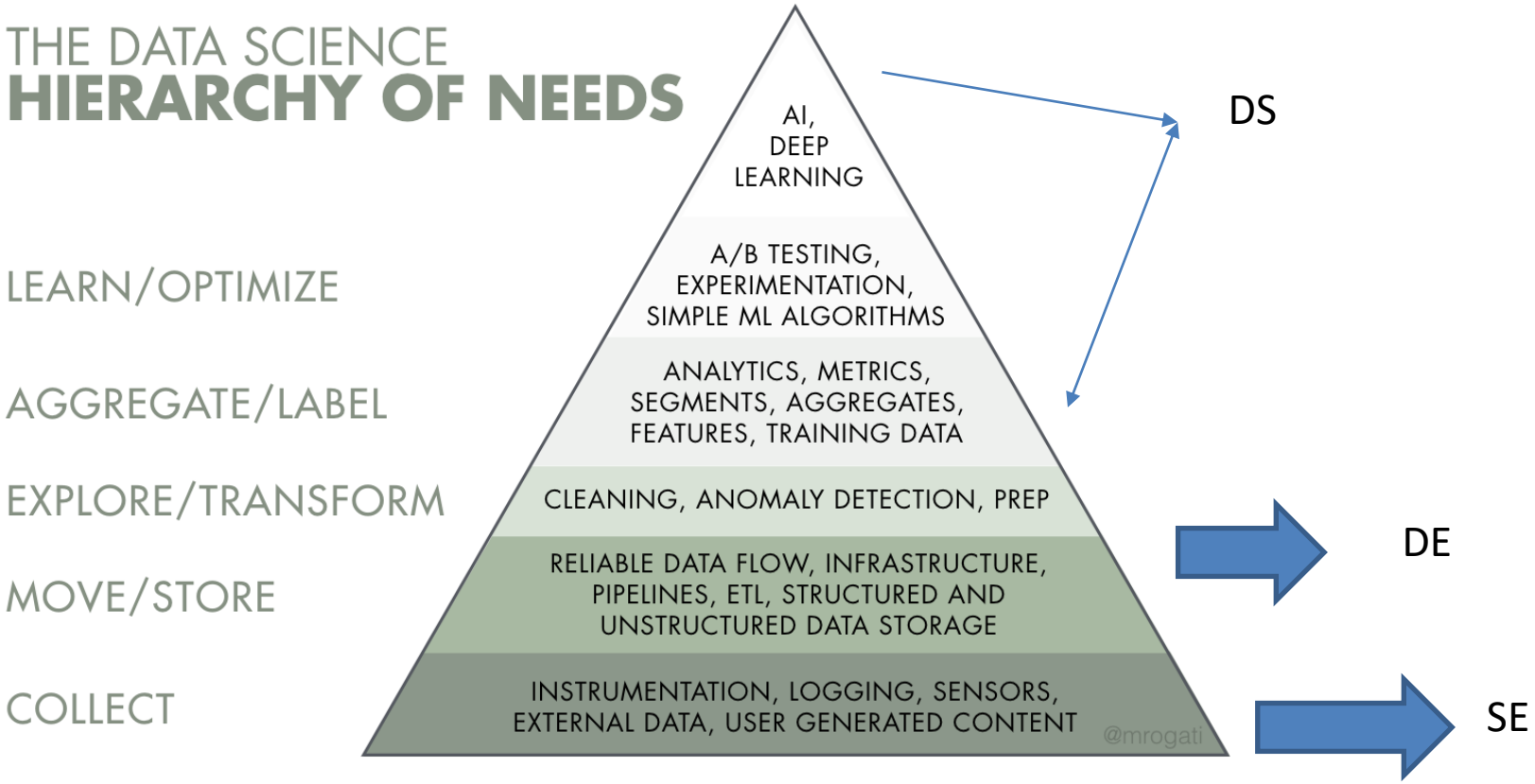
- Ser un Científico de datos no se trata de cuan avanzados son tus modelos, se trata de cuanto impacto puede tener tu trabajo.
- Eres un solucionador de Problemas. Un estratega de la empresa.

Grandes Empresas

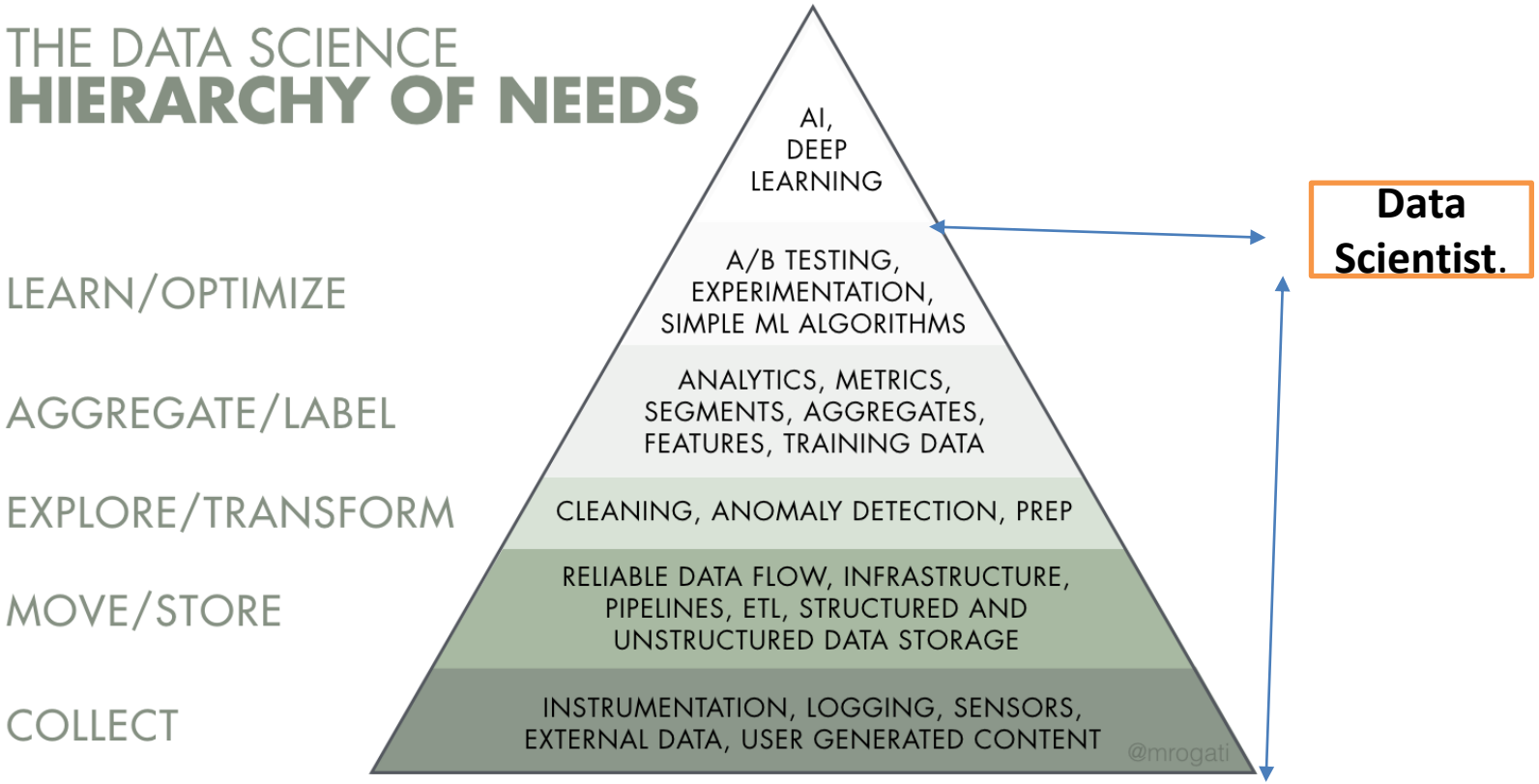


Adaptación de Mónica Rogati

Empresas Medianas

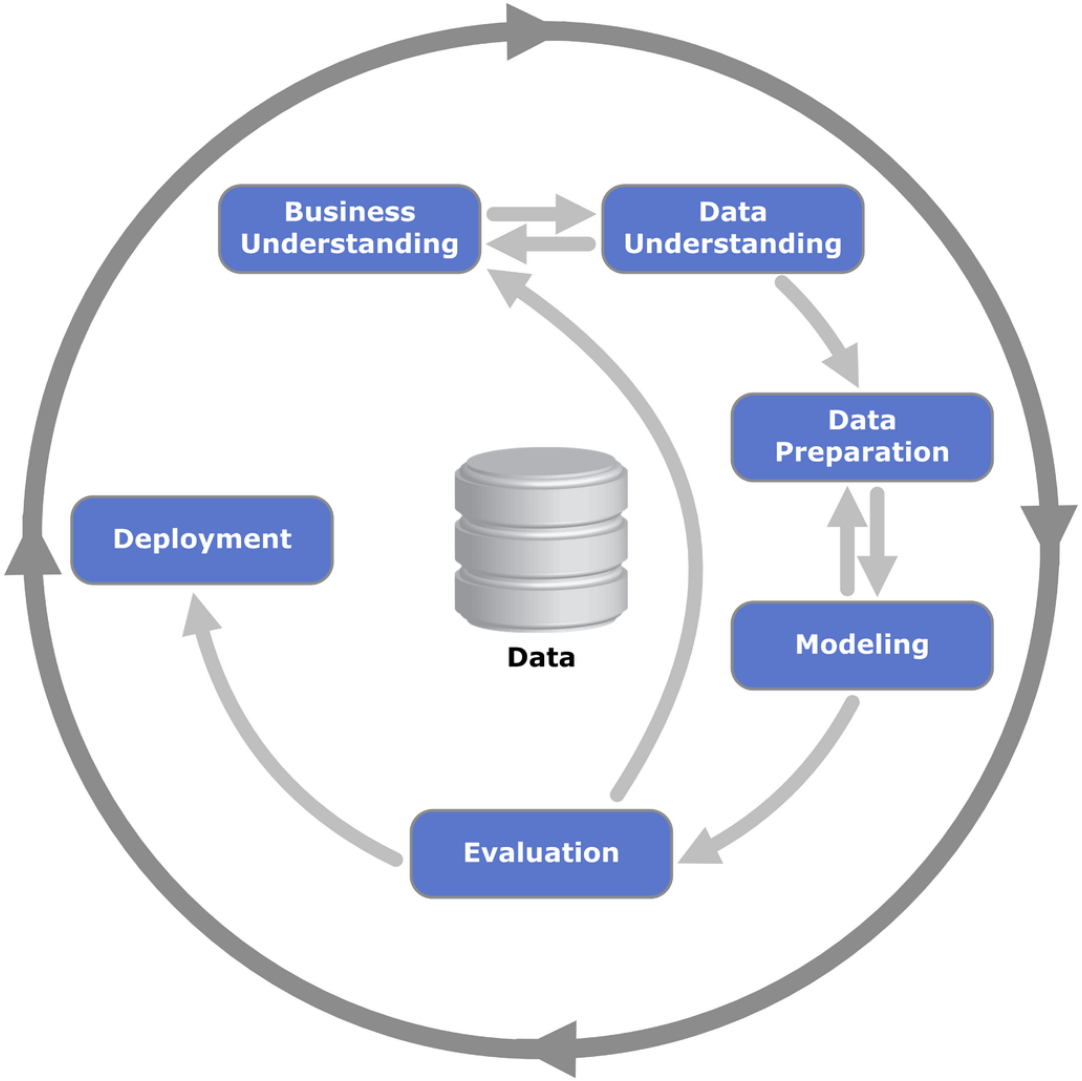


Empresas Pequeñas



Metodología Trabajo

CRISP-DM fue concebido en 1996. En 1997 se puso en marcha como un proyecto de la Unión Europea bajo la iniciativa de financiación ESPRIT. El proyecto fue dirigido por cinco empresas: SPSS, Teradata, Daimler AG, NCR y Ohra, una compañía de seguros.



Referencias

- https://www.coursera.org/learn/big-data-introduction/home/welcome
- https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007