

Introducción a Ciencia de Datos usando Python

Prof. Dr. Jorge Zavaleta

Departamento de Ciencias Ambientales

Universidade Federal Rural de Rio de Janeiro (UFRRJ)

Investigador de posdoctorado (PDJ/CNPq)

Brasil

Agenda

- Introducción
- ¿Qué es ciencia de datos?
- Aplicaciones
- ¿Qué es un científico de datos?
- Importancia de la ciencia de datos
- Herramientas
- Manos a la masa

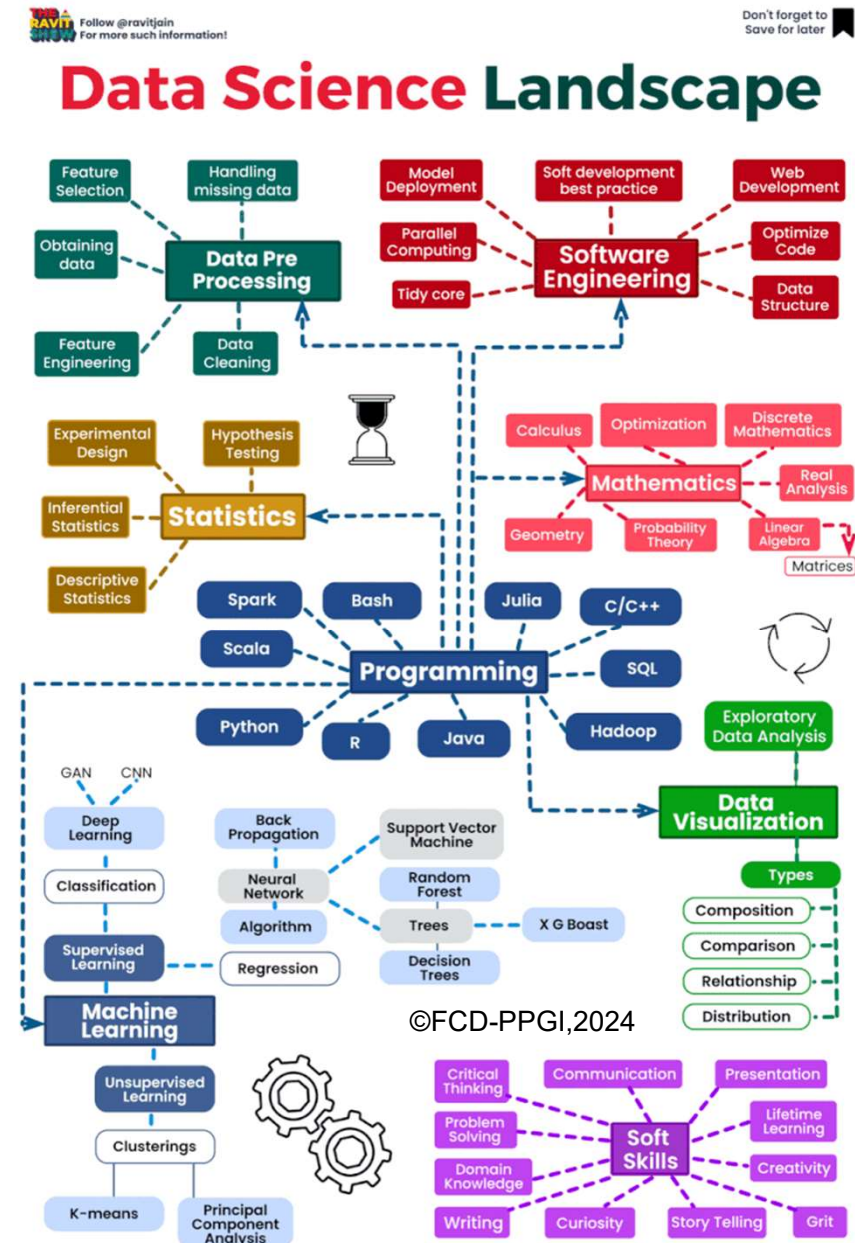
Introducción

- **Datos:** Flujos de hechos recopilados (en bruto) que representan eventos del dominio. Cualquier evento que se pueda almacenar en formato digital, incluidos texto, números, imágenes, vídeos o películas, audio, software, algoritmos, ecuaciones, animaciones, modelos, simulaciones, etc.
- **Información:** Conjuntos de datos que son significativos y útiles para los seres humanos en procesos como la toma de decisiones
- **Conocimiento:** Información interrelacionada no estructurada de reglas que guían la toma de decisiones.

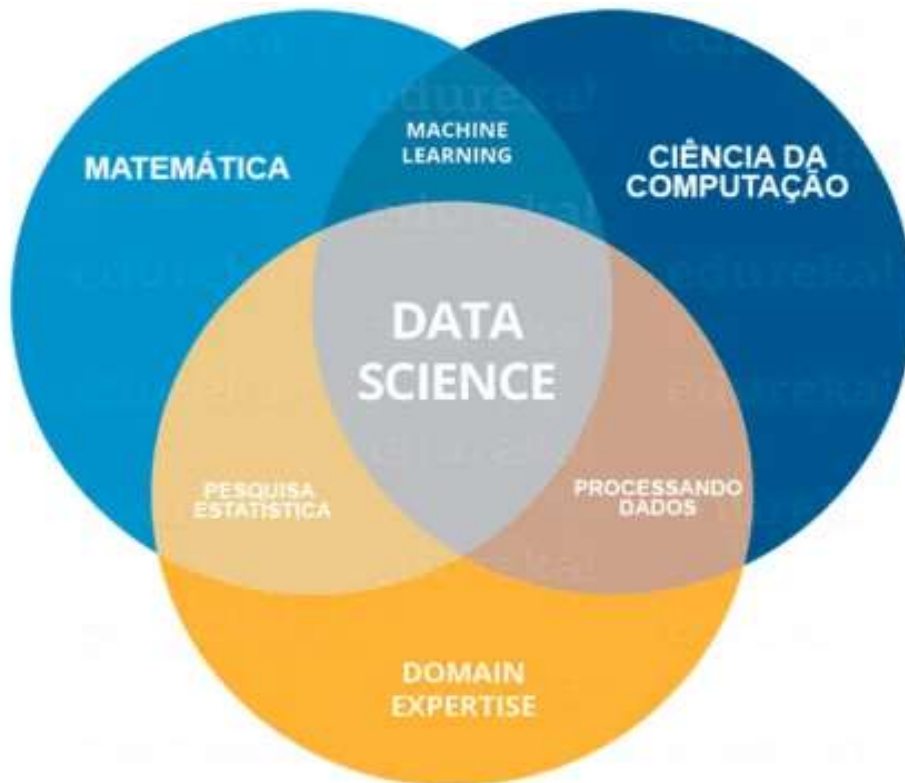
¿Qué es la ciencia de datos?

- La ciencia de datos es el estudio de datos con el fin de extraer información **significativa** de los datos en varias formas
- Es un enfoque **multidisciplinario** que combina principios y prácticas del campo de la **matemática**, la **estadística**, la **inteligencia artificial** y la **ingeniería de computación** para analizar grandes cantidades de datos (**Big Data**).
- 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**”.

Zavaleta, J. Introducción a Ciencia de Datos Usando Python. UNP, 2024



¿Qué es la ciencia de datos?



- Computación
 - Programación
 - Tecnologías Big Data
- Matemáticas y Estadística
 - Machine Learning
 - Detección de anomalías
- Dominio
 - Conocimiento del contexto

<https://harve.com.br/blog/data-science-blog/o-que-e-data-science-guia-iniciantes/>

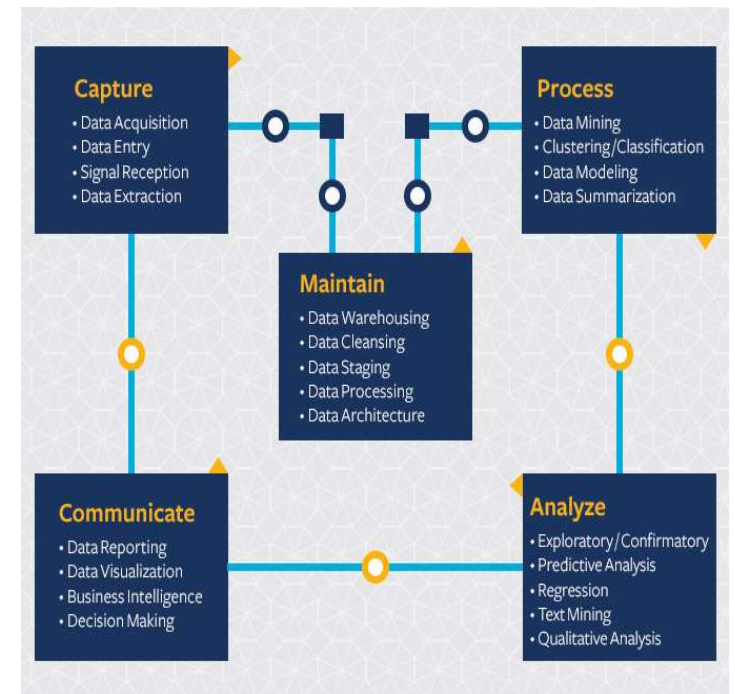
Zavaleta, J. Introducción a Ciencia de Datos Usando Python. UNP, 2024

Aplicaciones

- Búsquedas en internet
- Sistemas de recomendación
- Comparadores de precios
- Logística de entrega
- Planificación de rutas aéreas
- Fraude y riesgo
- Publicidad digital
- Reconocimiento de imagen y voz.
- Juegos
- Finanzas
- Educación
- Agricultura
- Medicina
- Recursos humanos
- Deportes
- Ciencias sociales

¿Qué es un científico de datos?

- Profesional que utiliza los **principios de la ciencia de datos para resolver problemas mediante el análisis y la interpretación de datos para encontrar información y patrones**
- Los científicos de datos a menudo trabajan con analistas y empresas para convertir la información de los datos en acción.
- Cree y ajuste modelos para predecir tendencias futuras, haga diagramas, gráficos y tablas para representar tendencias y predicciones, y resuma datos para ayudar a las partes interesadas a comprender e implementar resultados.
- Los científicos de datos pueden trabajar en una variedad de áreas, que incluyen: finanzas, academia, investigación científica, salud, comercio minorista, tecnología de la información, gobierno y comercio electrónico.
- Los científicos de datos deben ser comunicadores eficaces, líderes, miembros del equipo y pensadores analíticos de alto nivel. A menudo necesitan habilidades en lenguajes de programación como **Python**, **R** y técnicas de aprendizaje automático.

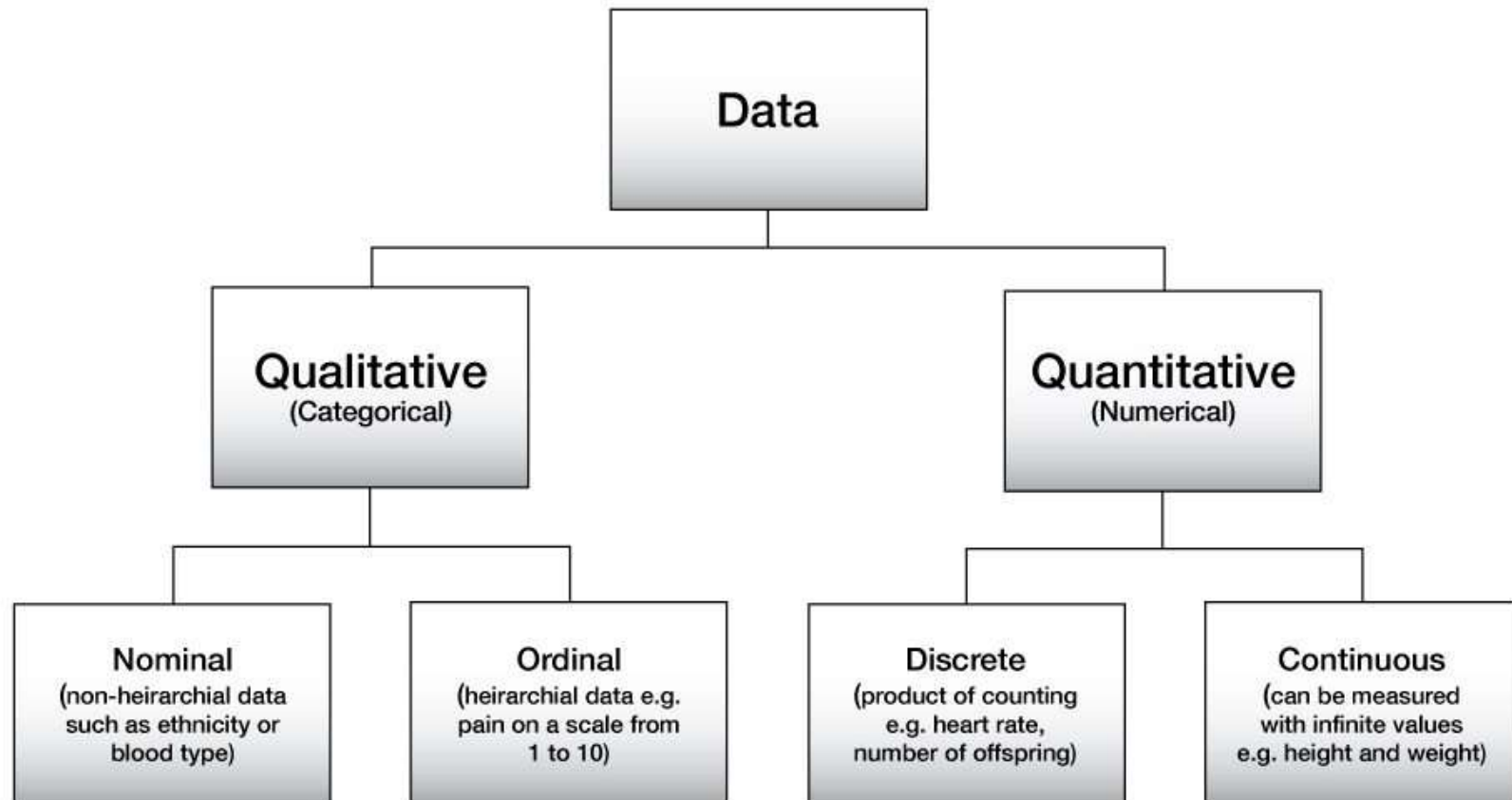


Fuente: ©FCD-PPGI,2024

¿Por qué es importante la ciencia de datos?

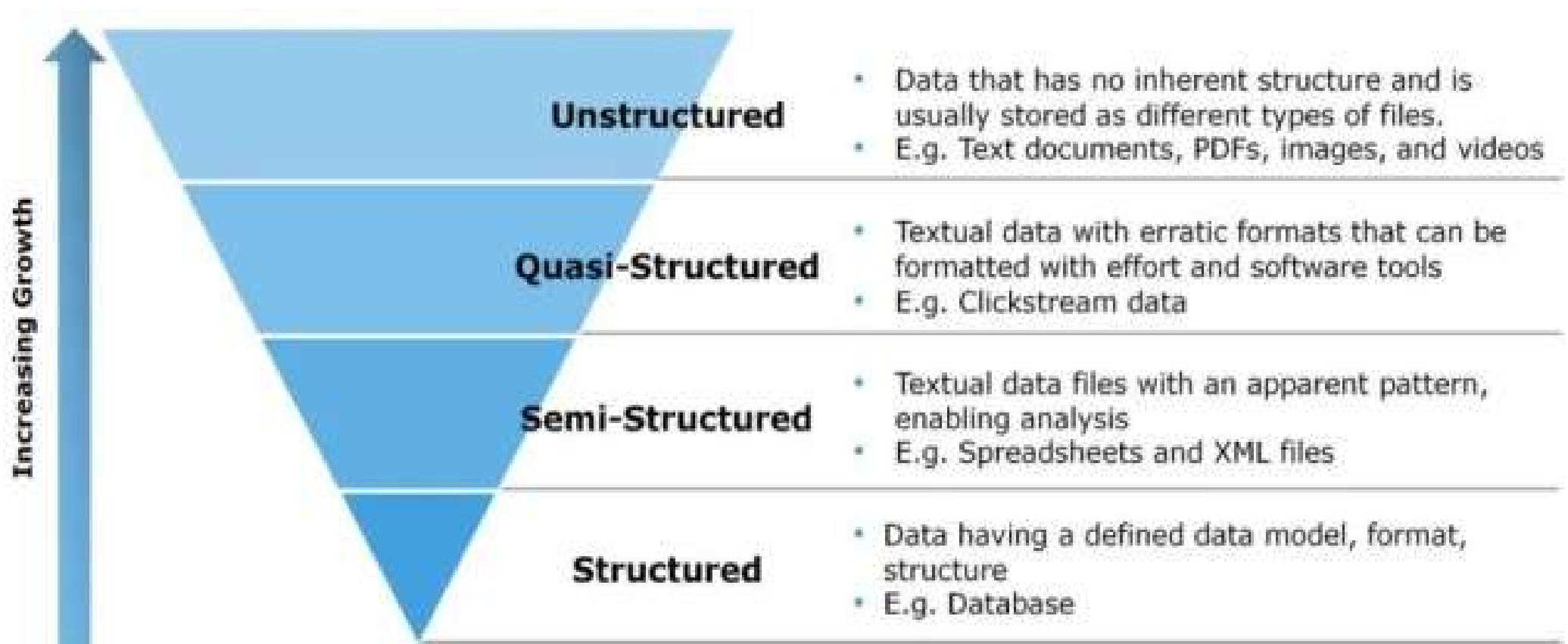
- La ciencia de datos es importante porque **combina herramientas, métodos y tecnología para generar significado a partir de los datos.**
- Las organizaciones modernas están inundadas de datos; hay una proliferación de dispositivos que pueden recopilar y almacenar información de manera automática.
- Los sistemas en línea y los portales de pago capturan más datos en los campos del comercio electrónico, la medicina, las finanzas y cualquier otro aspecto de la vida humana.
- Disponemos de grandes cantidades de **datos de texto, audio, video e imágenes.**

Tipos de datos



Fuente: ©FCD-PPGI,2024

Tipos de datos digitales

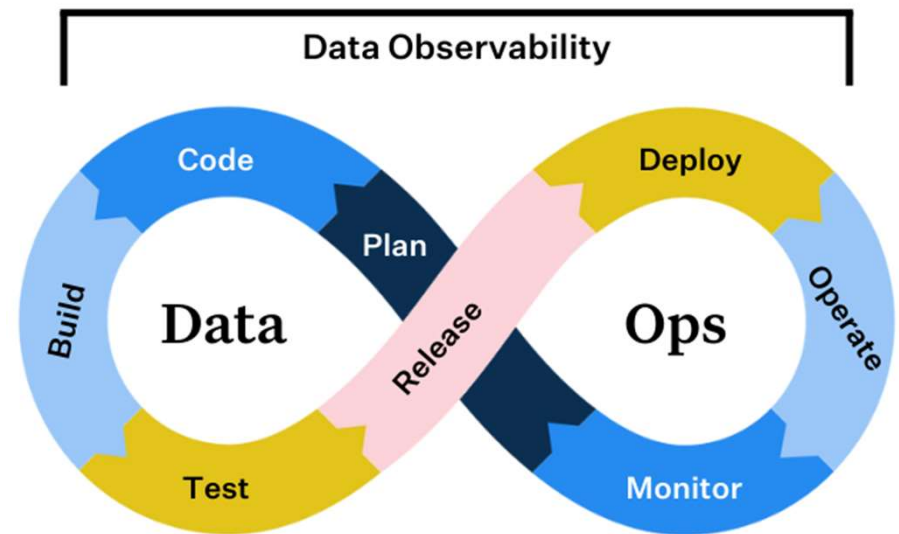


Fonte: <https://mycloudwiki.com/san/data-and-information-basics/>

Zavaleta, J. Introducción a Ciencia de Datos Usando Python. UNP, 2024

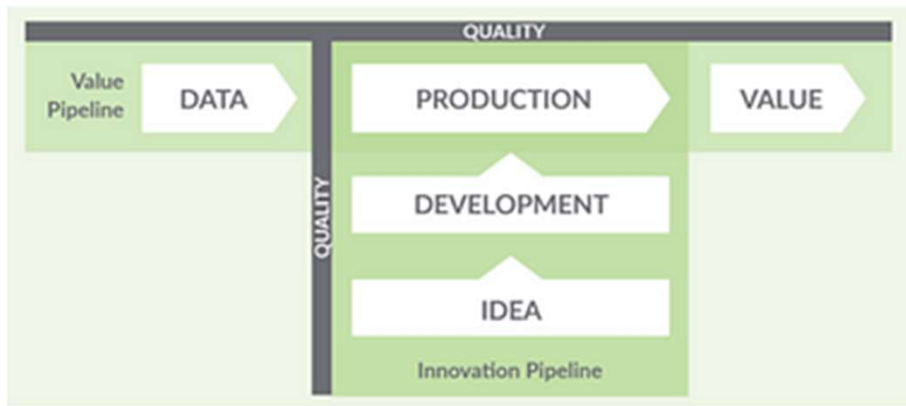
DataOps

- **DataOps** es una disciplina que fusiona equipos de **ingeniería de datos y ciencia de datos** para respaldar las necesidades de datos de una organización, de manera similar a cómo DevOps ayuda a las organizaciones a escalar la ingeniería de software.
- **DataOps** es un conjunto de **prácticas, procesos y tecnologías** que combina una perspectiva integrada y orientada a procesos sobre los datos con la automatización y los métodos de la ingeniería de software ágil para mejorar la calidad, la velocidad y la colaboración y promover una cultura de mejora continua en el área de análisis de datos.



Fuente: ©FCD-PPGI,2024

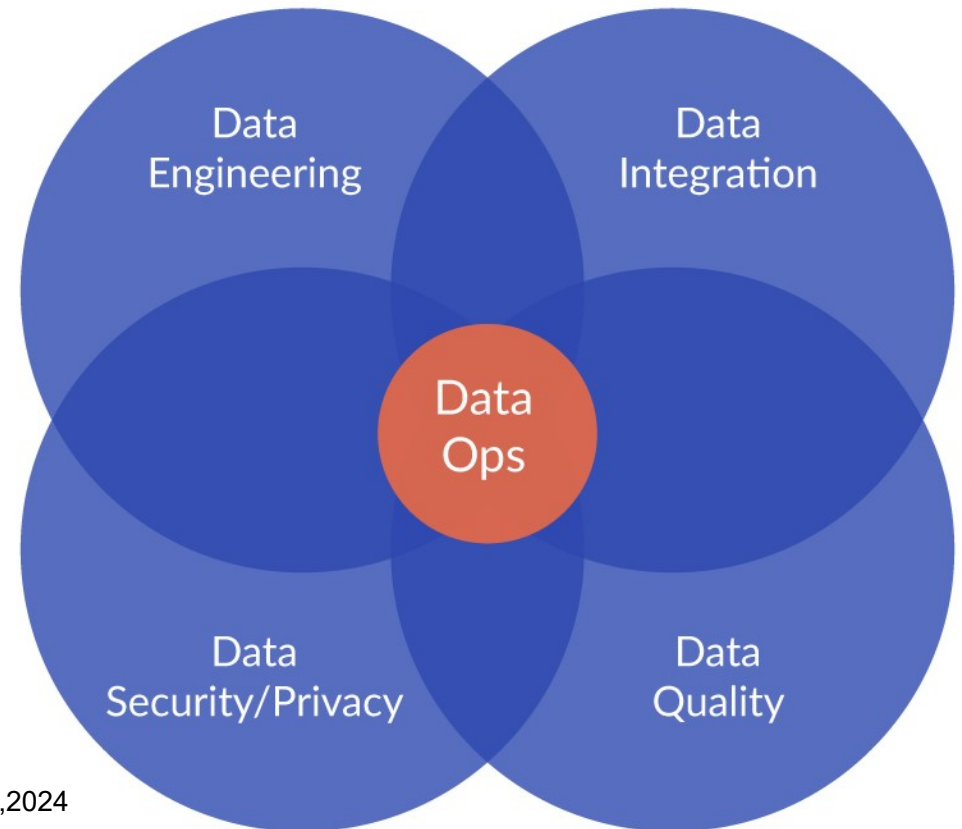
DataOps



DataOps

<https://medium.com/data-ops/dataops-is-not-just-devops-for-data-6e03083157b7>

Fuente: ©FCD-PPGI,2024



Fonte: <https://esimplicity.com/works/data-ops-for-field-agents/>

Ingeniería de datos

Los **ingenieros de datos** son el vínculo entre la estrategia de Big Data de la gerencia y los científicos de datos que necesitan trabajar con datos.

Lo que hacen es construir las plataformas que permiten a los científicos de datos hacer su magia.

Estas plataformas se suelen utilizar de cinco formas diferentes:

- **Ingestión y almacenamiento** de grandes cantidades de datos
- **Creación de algoritmos** por parte de los científicos de datos
- **Automatización de los modelos y algoritmos de aprendizaje automático** de los científicos de datos para su uso en producción
- **Visualización de datos** para empleados y clientes
- La mayoría de las veces, estos jóvenes comienzan como arquitectos de soluciones tradicionales para sistemas que involucran bases de datos SQL, servidores web, instalaciones de SAP y otros sistemas "estándar"

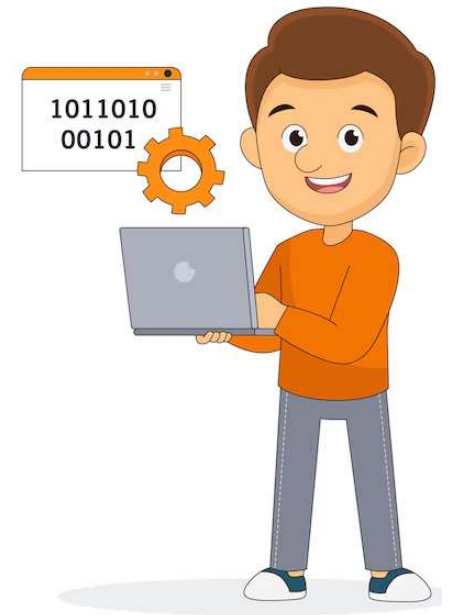


Ingeniería de datos

Para crear plataformas de **big data**, el ingeniero debe ser experto en la especificación, configuración y mantenimiento de tecnologías de big data como:

- **Hadoop, Spark, HBase, Cassandra, MongoDB**, Kafka, Redis y más.

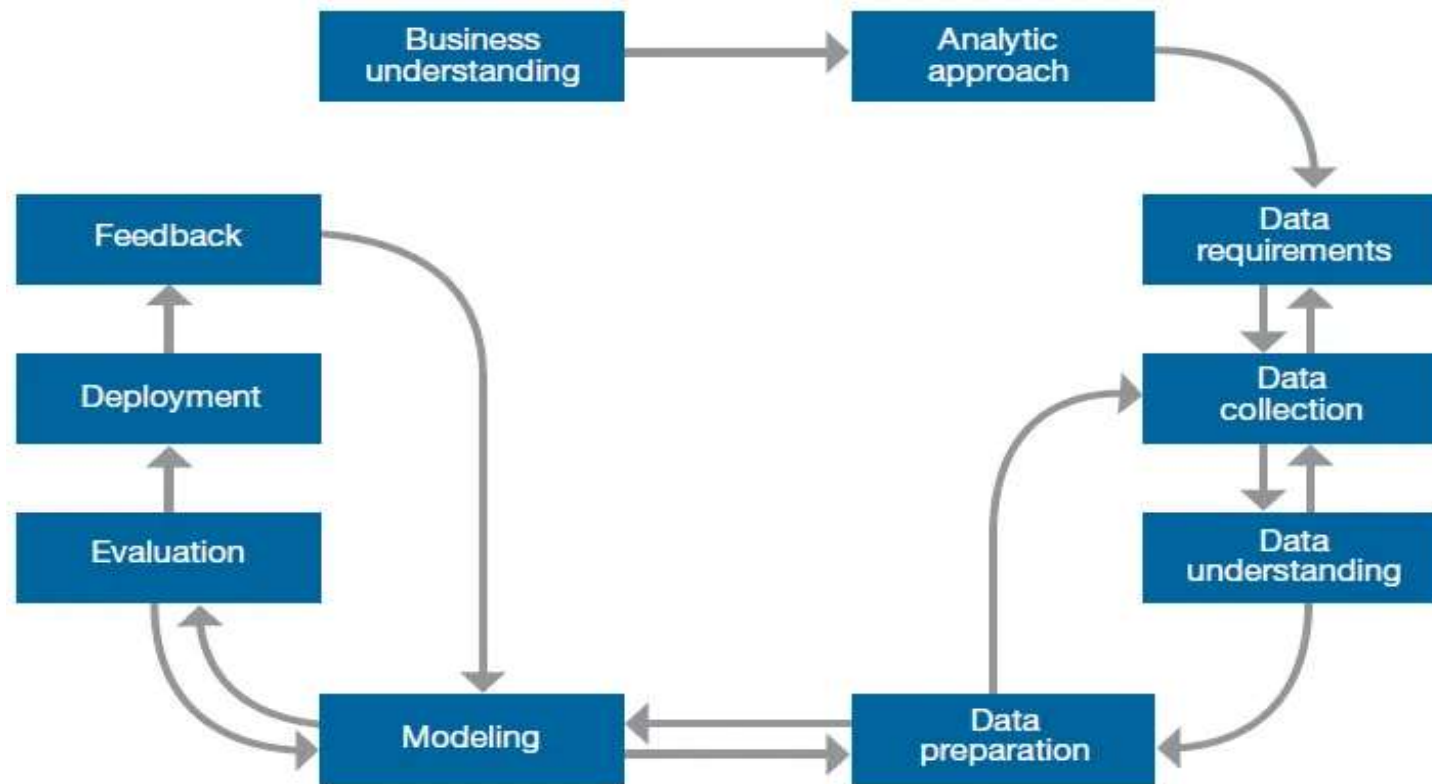
También necesita experiencia en cómo implementar sistemas en infraestructuras en la nube como Amazon o Google o en hardware local.



Ciclo de vida de la ciencia de datos

1. Comprender el problema y establecer objetivos: ¿qué problema estoy resolviendo?
2. Recopilar y analizar los datos: ¿qué información necesito?
3. Preparar los datos: ¿cómo debo procesarlos?
4. Construir el modelo: ¿cuáles son los patrones en los datos que conducen a soluciones?
5. Evaluar y criticar el modelo: ¿el modelo resuelve mi problema?
6. Presentar resultados: ¿cómo puedo resolver el problema?
7. Implementar el modelo: ¿cómo resuelvo el problema en el mundo real?

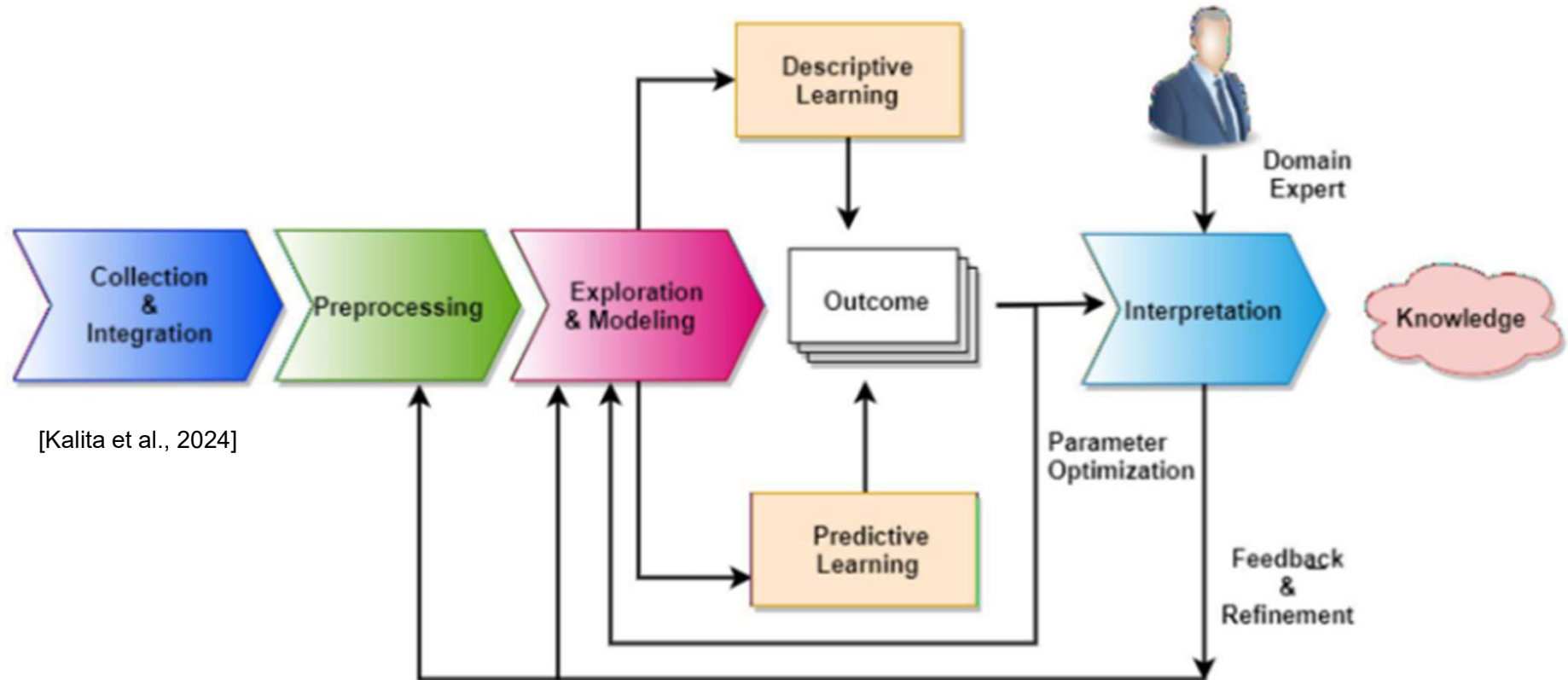
Ciclo de vida de la ciencia de datos



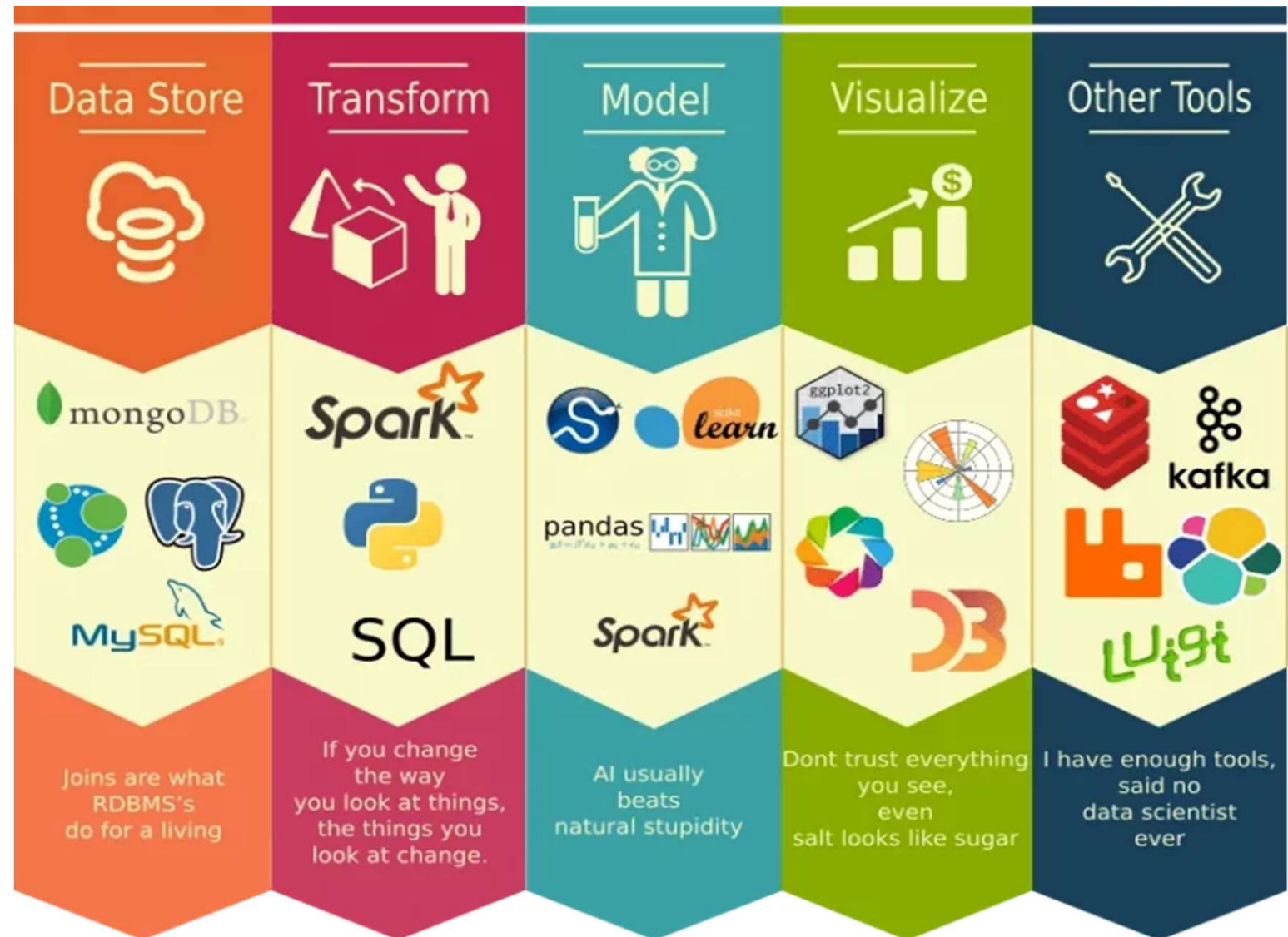
Fuente: <https://medium.com/@applying.pe/10-etapas-para-la-ciencia-de-datos-b4689181d0a2>

Zavaleta, J. Introducción a Ciencia de Datos Usando Python. UNP, 2024

Pipeline de Ciencia de datos



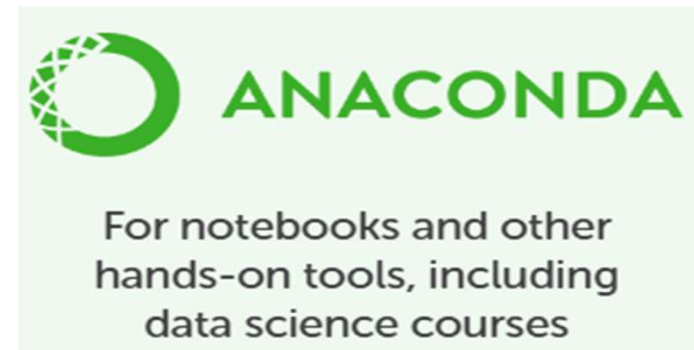
Herramientas



Fuente: <https://research.aimultiple.com/data-science-tools/>





IDEs







- Jupyter notebook: <https://jupyter.org/>
- JupyterLab: : <https://jupyter.org/>
- Deepnote: <https://deepnote.com/>
- Google Colab: [Colaboratory \(colab\)](https://colab.research.google.com/)
- GitHub: https://github.com/zavaleta/Intro_DS_UNP








IDEs

 main ▾ Intro_DS_UNP / s1_1.ipynb 

Go to file  

2024v.1  bfd979e · 18 minutes ago  History

Preview Code Blame 717 lines (717 loc) · 23.5 KB    



Introducción a Ciencia de Datos

Prof. Dr. Jorge Zavaleta - [e-mail](#)

Módulo 1 - Introducción a Markdown

¿Qué es Markdown?

Markdown es un lenguaje de marcado ligero que se puede utilizar para agregar elementos de formato a documentos de texto sin formato. Creado por John Gruber en 2004, Markdown es ahora uno de los lenguajes de marcado más populares del mundo.



Zavaleta, J. Introducción a Ciencia de Datos Usando Python. UNP, 2024

Bibliografía

- Kalita, J. K., Bhattacharyya, D. K., & Roy, S. (2024). Fundamentals of data science: theory and practice. Academic Press.
- Igual, L., & Seguí, S. (2024). Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications (2 ed.). Springer International Publishing. <https://doi.org/10.1007/978-3-031-48956-3>
- Lau, S., Gonzalez, J., & Nolan, D. A. (2023). Learning data science: data wrangling, exploration, visualization, and modeling with Python (First edition). O'Reilly Media, Inc.
- Tyagi, A. K. (2022). Data Science and Data Analytics: Opportunities and Challenges. CRC Press.
- Rengaswamy, R., & Suresh, R. (2022). Data Science for Engineers (1º ed). CRC Press. <https://doi.org/10.1201/b23276>
- Prakash, K. B. (2022). Data Science Handbook: A Practical Approach. Scrivener Publishing LLC.