Introducción a Ciencia de Datos

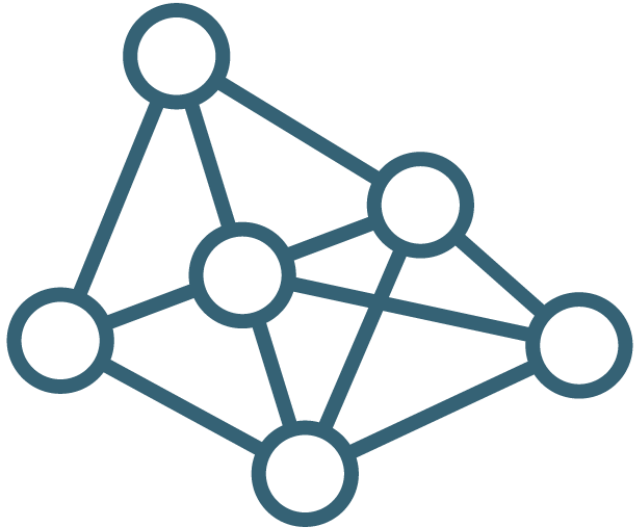
Váquez Pérez Ignacio David

ignacio.vperez@alumnos.udg.mx

¿Qué es la ciencia de datos?

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.

Tiene como objetivo la producción de creencias informadas por datos y que se utilicen como base para la toma de decisiones.



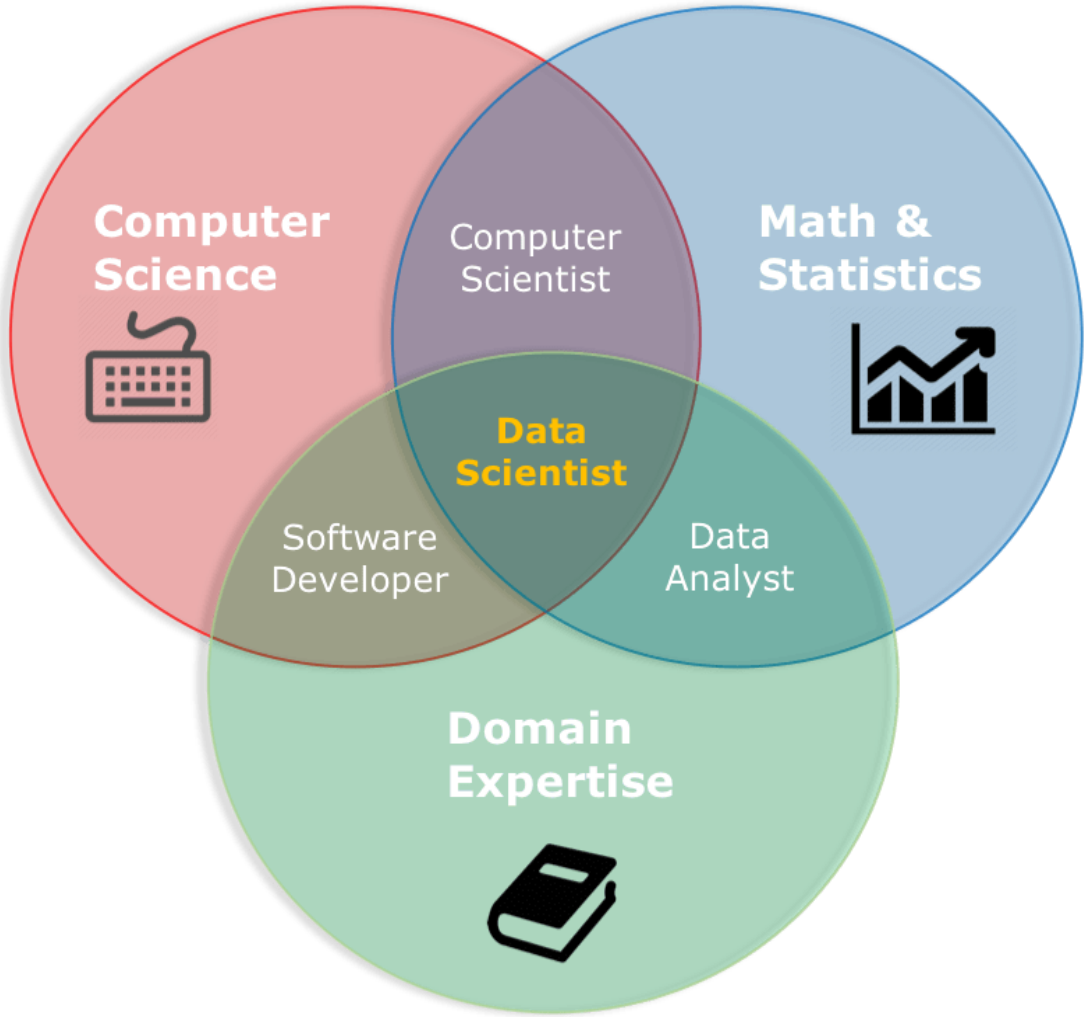
¿Que oportunidades ofrece?

- De 2010 a 2020, la cantidad de datos creados, capturados, copiados y consumidos en el mundo aumento de 1,2 billones de gigabytes a 59 billones de gigabytes, un crecimiento de casi 5,000%. [1]
- La cantidad de datos creados durante los próximos tres años (2021-2024) será mayor que todos los datos creados durante los últimos 30 años. [2]
- $110 mil millones de gasto global en inteligencia artificial (AI) en 2024, frente a $50,1 mil millones en 2020. [3]

[1] G. Press, "54 Predictions About The State Of Data In 2021," *Forbes*, 2020. https://www.forbes.com/sites/gilpress/2021/12/30/54-predictions-about-the-state-of-data-in-2021/?sh=71bf700397d3

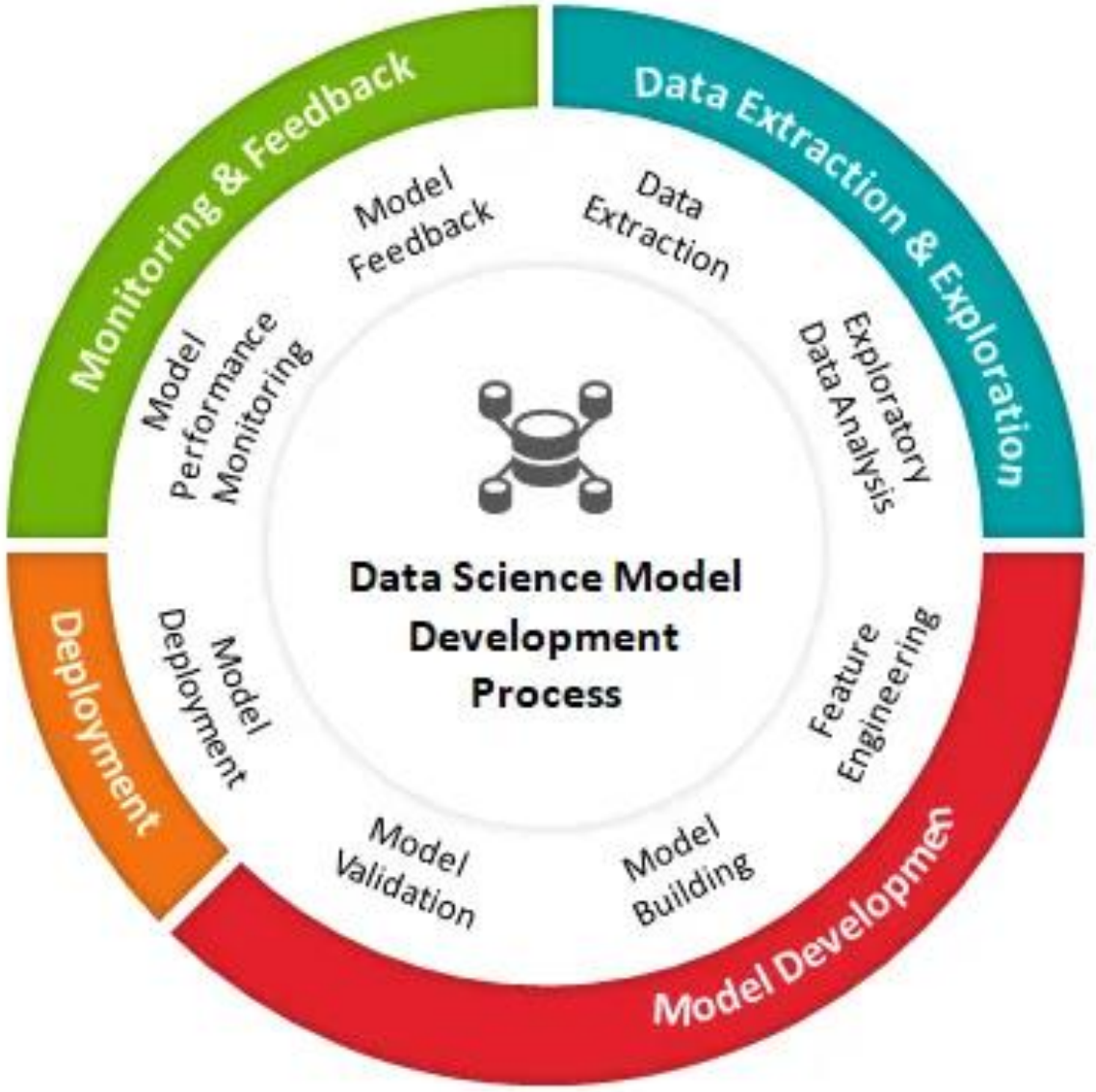
[2] "Data Creation and Replication Will Grow at a Faster Rate than Installed Storage Capacity, According to the IDC Global DataSphere and StorageSphere Forecasts," *IDC,* 2021. https://www.idc.com/getdoc.jsp?containerId=prUS47560321

[3] S. Brennan, "Worldwide AI spending expected to reach $110 billion in 2024," *IT PRO*, 2020. https://www.itpro.com/technology/artificial-intelligence-ai/356880/worldwide-ai-spending-expected-to-reach-110b-in-2024#:~:text=Global%20spending%20on%20artificial%20intelligence,Worldwide%20Artificial%20Intelligence%20Spending%20Guide.



Un campo interdisciplinario

Ciclo de vida



La ciencia de datos nos permite adoptar cuatro estrategias diferentes

- Sondeo de la realidad
- Descubrimiento de patrones
- Predección de eventos futuros
- Comprensión de las personas y el mundo

Sondeando la realidad

- Los datos representan la respuesta del mundo a nuestras acciones.
- El análisis de estas respuestas puede ser extremadamente valioso cuando se trata de tomar decisiones sobre nuestras acciones posteriores.
- La mejor respuesta solo se puede encontrar investigando el mundo.

Descubrimiento de patrones

- Los problemas identificados se pueden analizar automáticamente para descubrir patrones útiles y agrupaciones naturales que pueden simplificar enormemente las soluciones.
- El uso de esta técnica para perfilar a los usuarios se utiliza como la publicidad programática o el marketing digital.

Predecir eventos futuros

- La analítica predictiva permite tomar decisiones en respuesta a eventos futuros, no solo de manera reactiva. Por supuesto, no es posible predecir el futuro en ningún entorno y siempre habrá eventos impredecibles; pero la identificación de eventos predecibles representa un conocimiento valioso.
- Se puede utilizar para optimizar las tareas , analizando datos como el clima, las ventas históricas, las condiciones del tráfico, etc.

Comprender a las personas y al mundo

- Las grandes empresas y los gobiernos están invirtiendo cantidades considerables de dinero en áreas de investigación como la comprensión del lenguaje natural, la visión por computadora, la psicología y la neurociencia.
- Para tomar decisiones óptimas, es necesario conocer los procesos reales que impulsan las decisiones y el comportamiento de las personas.
- Un ejemplo es el desarrollo de métodos de aprendizaje profundo para la comprensión del lenguaje natural y el reconocimiento visual

Proceso de ciencia de datos

Paso 1: Definición del problema

Paso 2: Recopilación de datos

Paso 3: Verificación y corrección de la calidad de los datos

Paso 4: Análisis de datos exploratorios

Paso 5: Modelado de datos

Paso 6: Comunicación de los datos

Paso 7: Implementación y monitoreo

Paso 1: definir el enunciado del problema

- La creación de una declaración de problema bien definida
- Una breve descripción del problema que vas a resolver
- Incorporar una comprensión de la funcionalidad del producto y la perspectiva del cliente para brindar contexto a esos problemas.

Ejemplo:

- *Quiero aumentar los ingresos*
- *Quiero predecir el incumplimiento del préstamo para mi departamento de crédito,*
- *Quiero recomendar el trabajo a mis clientes.*

Paso 2: Recopilación de datos

- Recopilar los datos que pueden ayudar a resolver el problema
- El método de recopilación de datos se clasifica ampliamente en dos categorías:
    1. Recopilar los datos a través de varios métodos, como encuestas, entrevistas y monitoreo.
    2. Utilizar los datos que están fácilmente disponibles o recopilados por un tercero.

Paso 3: Verificación y corrección de la calidad de los datos

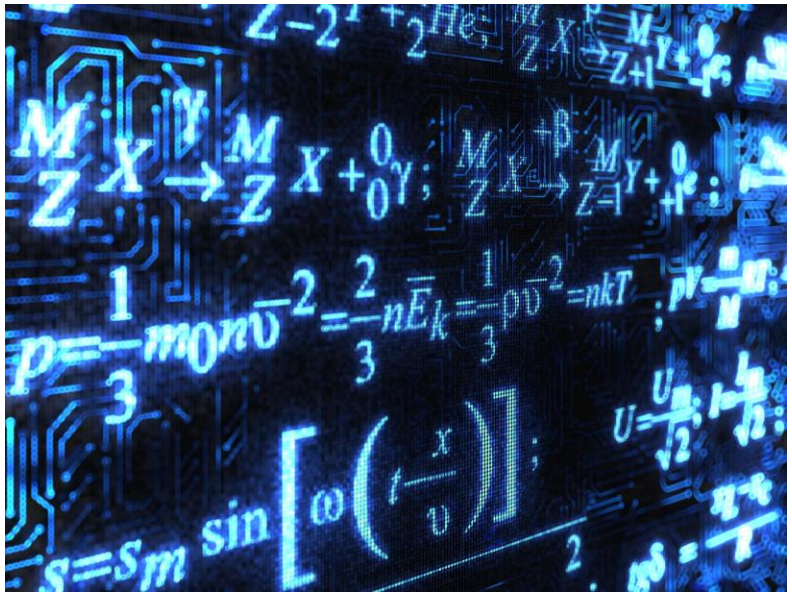
- Garantizar que los datos que se utilizan para el análisis y la interpretación sean de buena calidad.
- Hacer una verificación de cordura de los datos.
- Limpieza de basura o información no deseada
- Manipular información a un formato estructurado y organiz ado.

Paso 4: Análisis de datos exploratorios

- Familiarizarse con los datos y extraer información útil
- Estadísticas descriptivas como medidas de valor central y medidas de variabilidad
- Métodos de visualización como gráficos y diagramas
- Es un proceso iterativo que le ayuda a acercarse a la solución

Paso 5: modelado de datos

- Modelar significa formular cada paso y reunir las técnicas necesarias para lograr la solución.
- Existen varias técnicas en estadísticas y aprendizaje automático que puede elegir según los requisitos.



Paso 6: Comunicación de datos

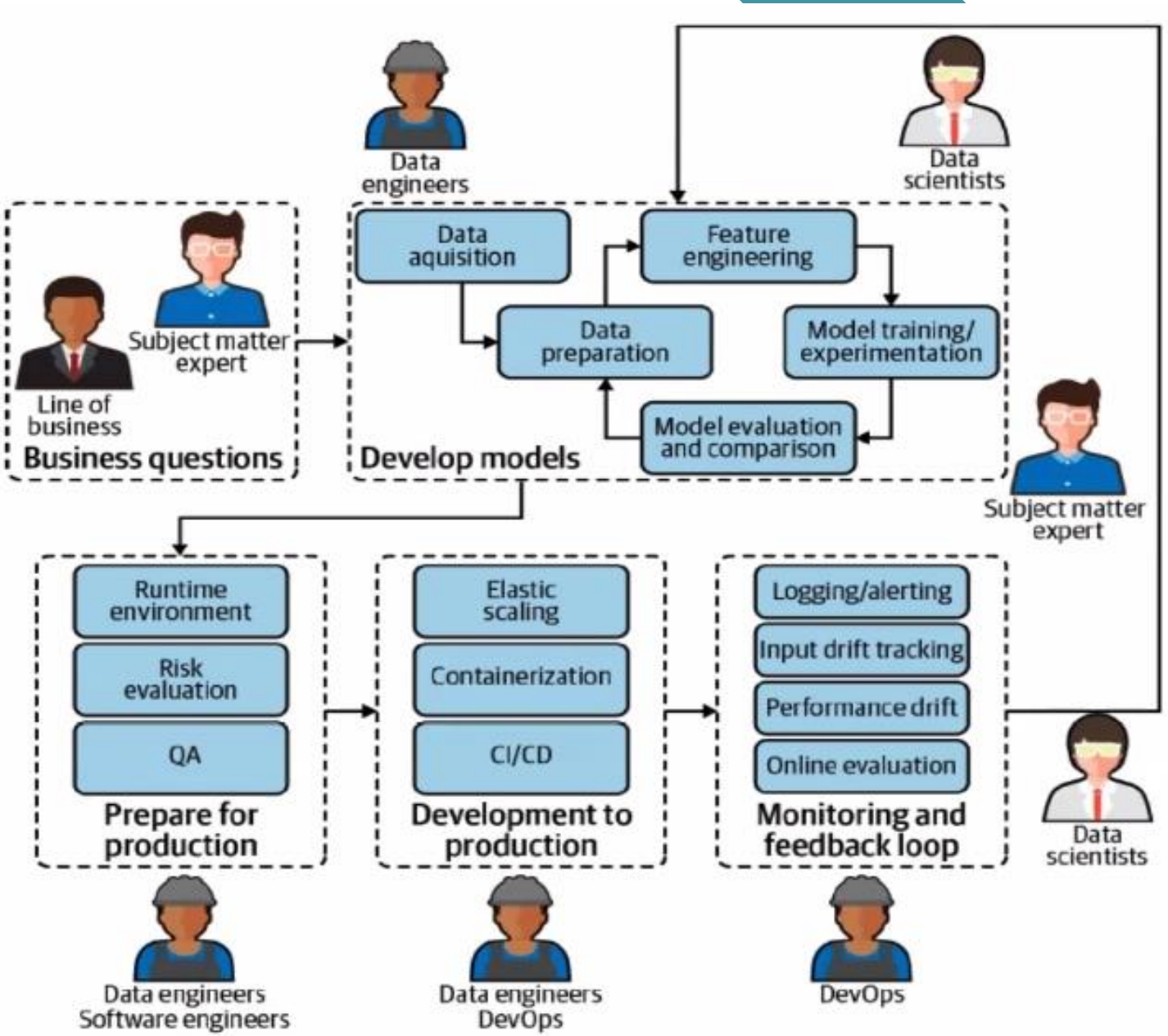
- Presenta los resultados de su análisis a las partes interesadas. Les explica cómo llegó a una conclusión específica y sus hallazgos críticos.
- Debe comunicar los resultados de una manera sencilla de comprender. Y las partes interesadas deberían poder elaborar un plan viable a partir de él.

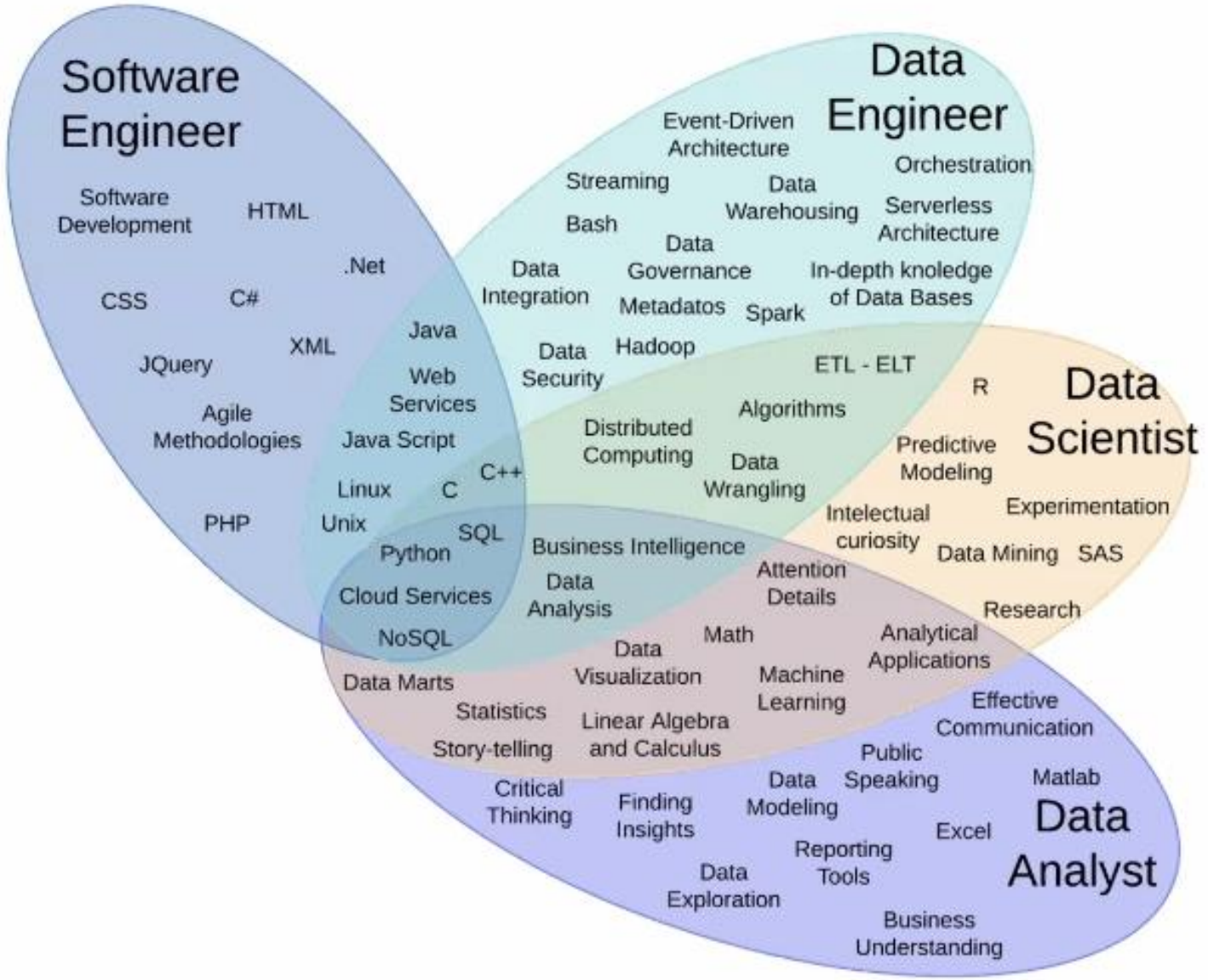
Al comunicar los resultados:

- Conoce a tu audiencia y habla su idioma
- Centrarse en los valores y los resultados
- Comunicar suposiciones y limitaciones

Paso 7: Implementación y monitoreo

- Supervisión del despliegue en el proceso de toma de decisiones en tiempo real.
- Validar que funciona como se esperaba.
- Monitorear, analizar e interpretar para mejorar continuamente.
- Considerar el rendimiento computacional del ETL y de los modelos para que sea eficaz, optimizado, y flexible a futuras anormalidades en los datos.


Data engineers
Data scientists
Data aquisition
Feature engineering
Data preparation
Model training/ experimentation
Model evaluation and comparison
Develop models
Subject matter expert
Line of business
Business questions
Subject matter expert
Runtime environment
Risk evaluation
QA
Prepare for production
Elastic scaling
Containerization
CI/CD
Development to production
Logging/alerting
Input drift tracking
Performance drift
Online evaluation
Monitoring and feedback loop
Data scientists
Data engineers
Software engineers
Data engineers
DevOps
DevOps


Software Engineer
Software Development
HTML
.Net
CSS
C#
XML
Java
JQuery
Web Services
Agile Methodologies
Java Script
C++
Linux
C
PHP
Unix
SQL
Python
Cloud Services
NoSQL
Data Engineer
Event-Driven Architecture
Orchestration
Streaming
Data Warehousing
Serverless Architecture
Bash
Data Governance
Data Integration
In-depth knoledge of Data Bases
Metadatos
Spark
Data Security
Hadoop
ETL - ELT
Distributed Computing
Algorithms
Data Wrangling
Business Intelligence
Data Analysis
Data Marts
Data Visualization
Math
Machine Learning
Statistics
Linear Algebra and Calculus
Story-telling
Data Scientist
R
Predictive Modeling
Intelectual curiosity
Experimentation
Data Mining
SAS
Attention Details
Research
Analytical Applications
Effective Communication
Public Speaking
Matlab
Critical Thinking
Finding Insights
Data Modeling
Excel
Reporting Tools
Data Exploration
Business Understanding
Data Analyst