HE2: Consultoría Económica e IAR

Clase 3 - Ciclo de Vida de Datos

Catalina Bernal
Santiago Neira

Enero 27 2026
Semana 2

Universidad de los Andes | Facultad de Economía

Contenido

El error de $500 Millones de Amazon

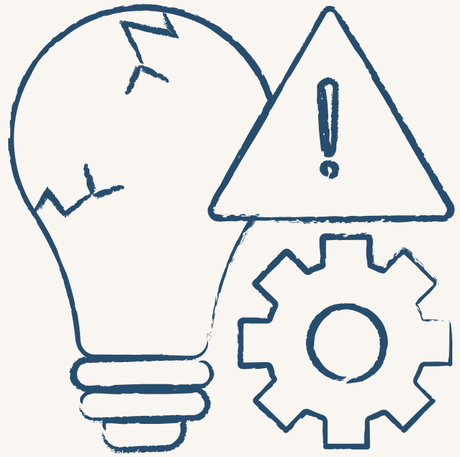
Reuters

En 2014, Amazon invirtió años y millones de dólares desarrollando un sistema de IA para automatizar la contratación de talento. En 2018, tuvieron que descartarlo por completo. ¿La razón? El sistema discriminaba sistemáticamente contra mujeres.

¿Qué salió mal?

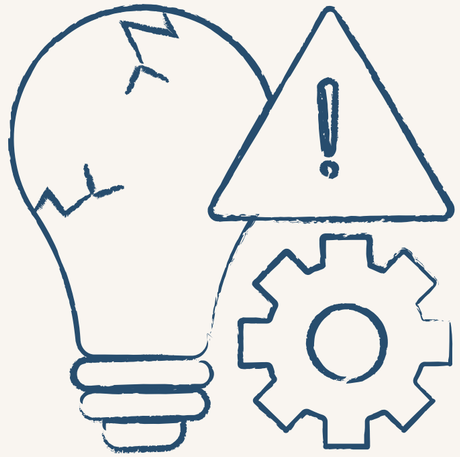
- El modelo fue entrenado con CVs de contrataciones de 2004 a 2014
- La industria tech históricamente contrató más hombres
- El algoritmo “aprendió” que ser hombre era un predictor de éxito
- Penalizaba CVs que incluían palabras como “women’s chess club”

¿De quién es la culpa?



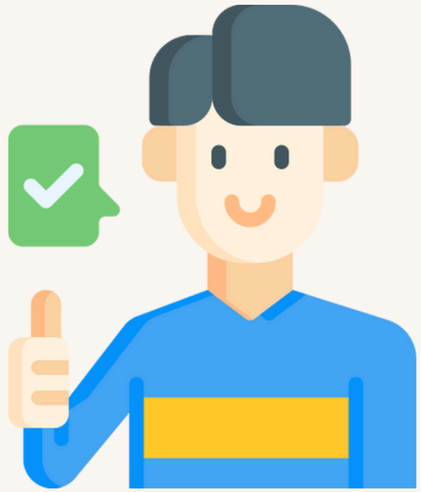
¿De quién es la culpa?

Nadie revisó el *ciclo de vida completo* de los datos ni su *calidad* desde una perspectiva técnica y de equidad.



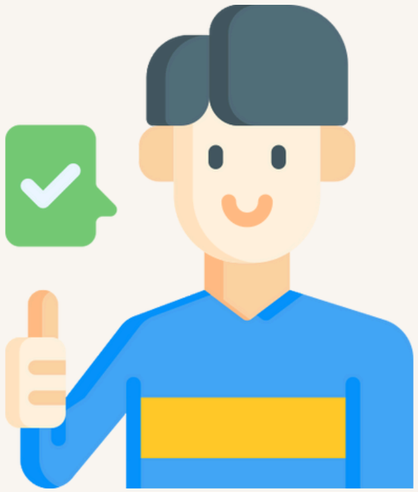
La premisa...

Como consultores en IA responsable, nosotros NO solo implementamos modelos. Somos los *guardianes de calidad* que previenen que proyectos como el de Amazon lleguen a producción con riesgos.



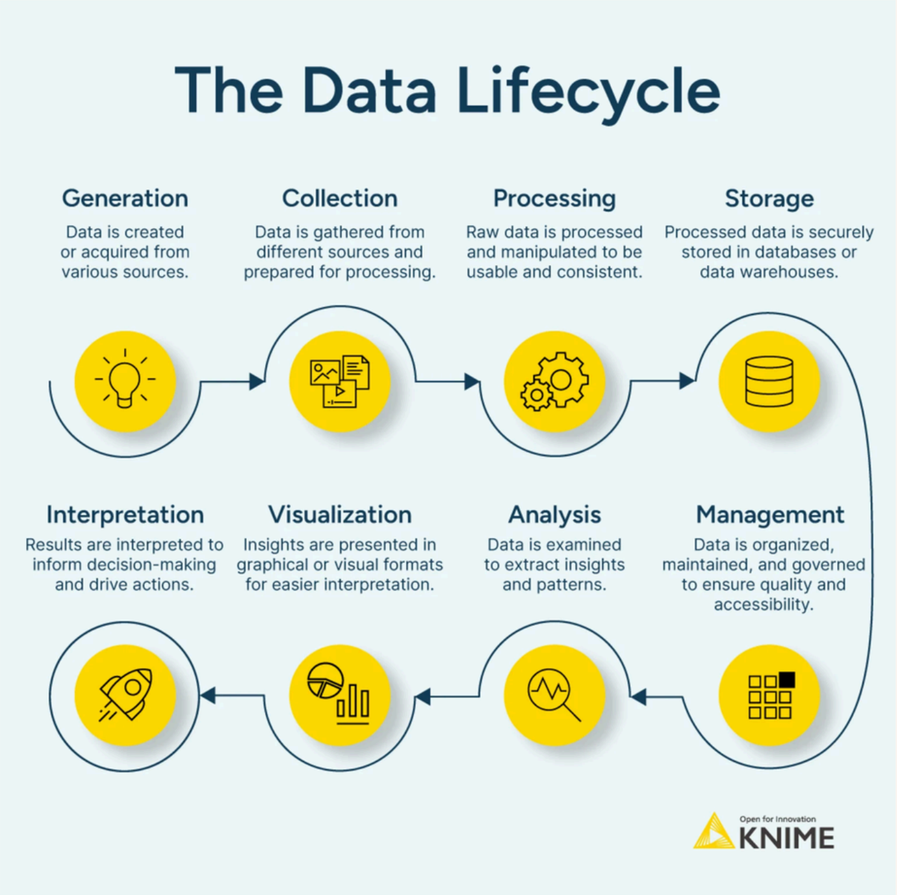
La premisa...

Como consultores en IA responsable, nosotros NO solo implementamos modelos. Somos los *guardianes de calidad* que previenen que proyectos como el de Amazon lleguen a producción con riesgos.



En Colombia, la Ley 1581 de 2012 (Protección de Datos Personales) exige calidad y actualidad de datos

Ciclo de vida de los datos (Data Cycle)

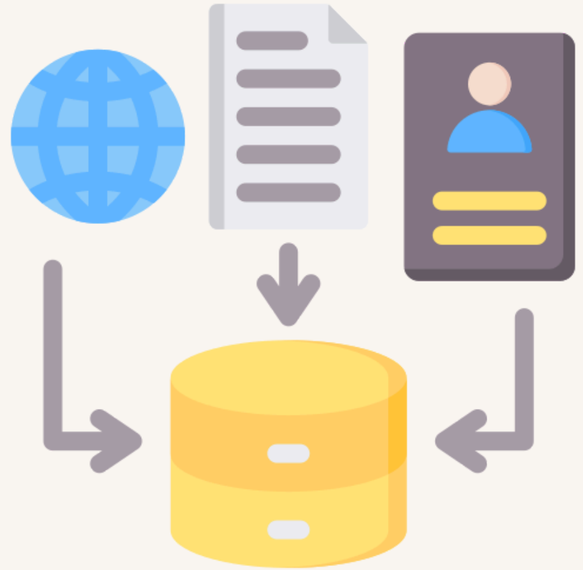


Ciclo de vida - KNime

Data Cycle - Generación y Recolección

¿Qué sucede en esta fase?

- Definición de qué datos se van a usar
- Fuentes (primarias - secundarias)
- Métodos de captura: APIs, scraping, encuestas



Preguntas importantes:

- ¿Estos datos responden a la pregunta de negocio?
- ¿Hay alguien o algo no representado en estos datos?
- ¿Tenemos permiso de captura y uso de los datos?

Data Cycle - Generación y Recolección

¿Qué sucede en esta fase?

- Definición de qué datos se van a usar
- Fuentes (primarias - secundarias)
- Métodos de captura: APIs, scraping, encuestas

Preguntas importantes:

- ¿Estos datos responden a la pregunta de negocio?
- ¿Hay alguien o algo no representado en estos datos?
- ¿Tenemos permiso de captura y uso de los datos?

Caso: App de Microcrédito.

Un cliente quiere predecir riesgo crediticio usando datos de smartphones. ¿Hay riesgos en el uso de estos datos?

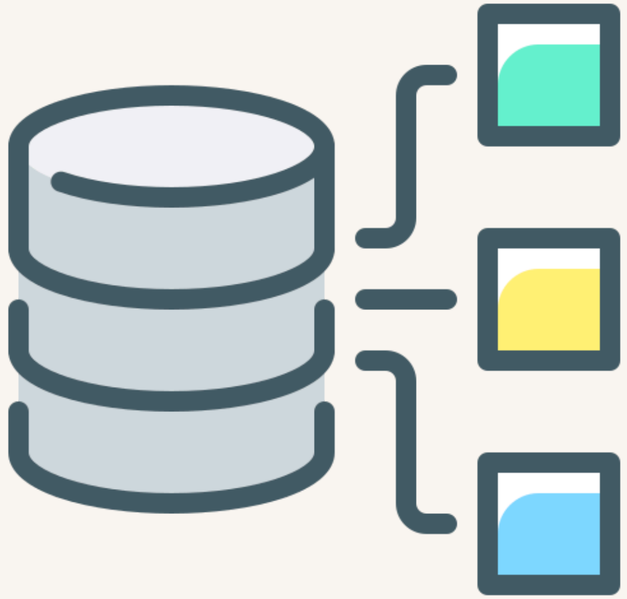
Data Cycle - Almacenamiento y Procesamiento

¿Qué sucede en esta fase? → Más de Data Eng

- Decisión de insfraestructura: Bases de datos, data lakes
- Estructuración: relacional, NoSQL, etc
- Preprocesamiento inicial

Rol del consultor

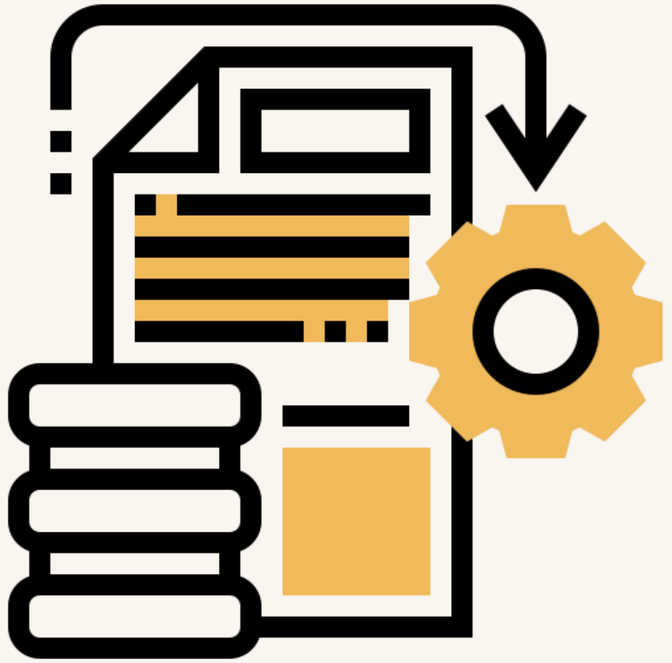
- Recomendar arquitectura según volúmen
- Asegurar trazabilidad ¿ Podemos reproducir resultados?
- **Documentación**



¡Documentación!

Datasheets for Datasets (Gebru et al. (2021))

- Describe origen, construcción, limitaciones y riesgos de un conjunto de datos.
- Responde preguntas clave:
  - ¿Quién recolectó los datos, cuándo y con qué propósito original?
- Su objetivo principal es hacer **trazable** el dataset.



Data Cycle - Análisis y Uso (management)

¿Qué sucede en esta fase?

- Exploración inicial (EDA)
- Modelado + feature engineering

Rol del consultor

- Traducir necesidades del cliente a preguntas analíticas
- Identificar métricas apropiadas
  - ¿accuracy? ¿fairness? ¿ambas?
- Detectar patrones sospechosos que indiquen problemas de **calidad**
- Balancear complejidad técnica con interpretabilidad

¡Calidad! - en 6 dimensiones

Data Quality Asessment (Batini et al. (2021))

Entre alta calidad o basura costosa y cómo explicarle al cliente

Completitud

¿Qué proporción de los datos esperados está presente?

Exactitud - *Accuracy*

¿Los datos reflejan correctamente la realidad que pretenden medir?

Consistencia

¿Los datos son coherentes internamente y entre fuentes?

¡Calidad! - en 6 dimensiones

Data Quality Asessment (Batini et al. (2021))

Entre alta calidad o basura costosa y cómo explicarle al cliente

Actualidad - *Timeliness*

¿Los datos están al día para el uso previsto?

Validez

¿Los datos cumplen con reglas de negocio y formatos esperados?

Unicidad - *Uniqueness*

¿Cada entidad está representada una sola vez?

¡Calidad! - en 6 dimensiones

Data Quality Asessment (Batini et al. (2021))

Entre alta calidad o basura costosa y cómo explicarle al cliente

Decisiones erróneas = pérdidas o mala focalización

Re-trabajo = costos adicionales

Escándalos = daño reputacional

Mini Análisis de Datos

ID	Fecha	Monto	Ciudad
001	2024-13-05	$50.000	Bogotá
002	2024-03-15	-$20.000	Medellin
003	2024-03-15		Bogota D.C.
001	2024-05-13	$50.000	Bogotá

1. Identifique al menos 4 problemas de calidad
2. Clasifíquelos
3. ¿Cuál sería el riesgo de usar estos datos sin limpiar?

HE2: Consultoría Económica e IAR

Clase 3 - Ciclo de Vida de Datos

Santiago Neira
Catalina Bernal

Enero 27 2026
Semana 2

Universidad de los Andes | Facultad de Economía