

# Exploración y Curación de Datos

---

Clase 1

Ariel Wolfmann

Mayo 2025

# Sobre mi

- Data & Software Engineering
  - + 10 años exp en Data & ML
- Ecosistema de startups
- Cs de la Computación, Famaf
- **Construyendo puentes entre Ciencia y Negocios**



yalo



Rappi

Addi

AstroPay

PyData

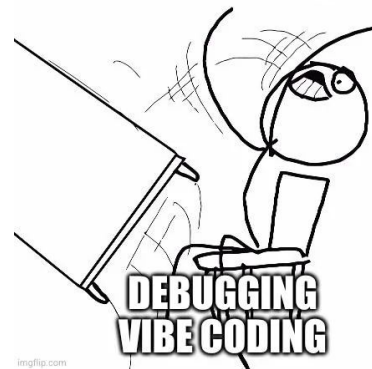
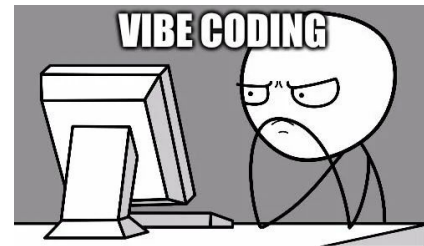
# Porque decidieron hacer la diplomatura?

Con todo el contenido disponible gratis en internet...

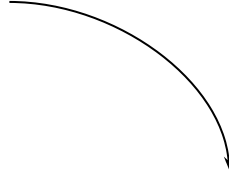
¿Por qué están acá hoy?

# Aprender hoy: Programar, Buscar, Preguntar

- No todo es memorizar: saber buscar y preguntar.
- Aprender a hacer las preguntas correctas.
  - No reinventar la rueda
  - Prompt Engineering
  - LLMs te pueden resolver una gran parte pero no el 100% (aun).
- Entender mensajes de error
  - Manejar la frustración como parte del proceso de aprendizaje
- Leer documentacion, ej: <https://pandas.pydata.org/docs/>

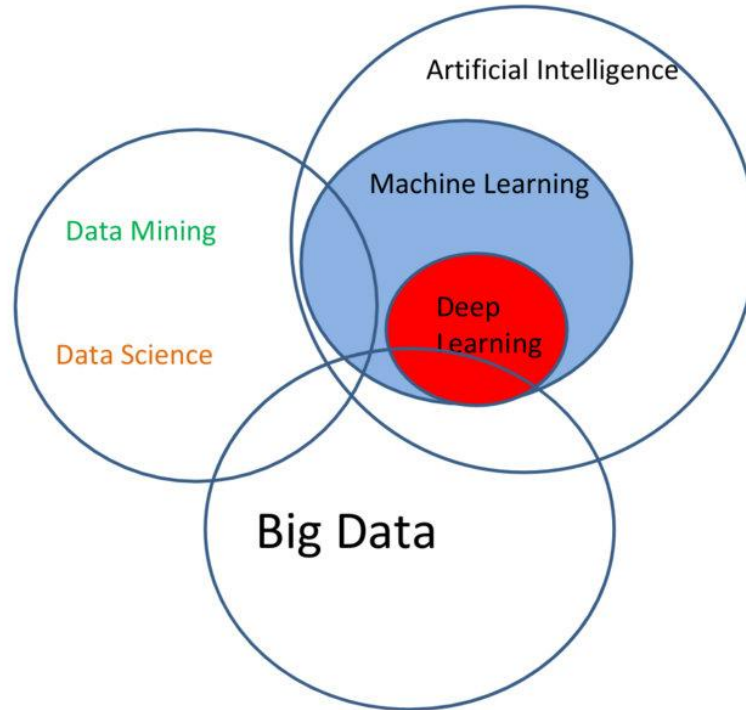


Analytics, Data Science,  
Machine Learning,  
Inteligencia Artificial



**¿Cómo los definirían? ¿Son sinónimos?**  
**¿Cuáles son las diferencias?**

# Conceptos relacionados



# Conceptos - Ecommerce que quiere mejorar sus ventas

Concepto	¿Qué es?	Ejemplo
Analytics	Análisis para obtener insights	Métricas de ventas
Data Science	Transformar datos en decisiones	Predicción de demanda
Machine Learning	Modelos que aprenden patrones	Predecir churn de usuarios
Inteligencia Artificial	Imitar comportamientos humanos	Chatbot que responde consultas

# Conceptos Relacionados

- **Analytics:** Análisis de datos para obtener insights (revelar info util).
- **Inteligencia Artificial:** Computadoras emulando comportamientos humanos específicos.
- **Big Data:** Disponibilidad de grandes volúmenes de datos y la capacidad de procesarlos.
- **Machine Learning:** Predecir un comportamiento basado en métodos estadísticos, encontrando y aprendiendo patrones.
- **Deep Learning:** Subconjunto de ML, sin necesidad de definir características pero requiere de mucho mayor volumen de datos.
- **Business Intelligence:** Orientado a medir y reportar eventos del pasado
- **Data Science:** Recolección, limpieza, transformar datos e hipótesis en información para la toma de decisiones o para predecir acciones.
- **NLP:** Procesamiento de lenguaje natural (LLMs).



# Conceptos - Ecommerce que quiere mejorar sus ventas

- **Analytics:** Medir visitas, conversión, ticket promedio.
- **Big Data:** recopilar datos comportamiento de usuarios.
- **Data science:** Detectar patrones de compra, optimizar carrito abandonado.
- **Machine Learning:** Predecir abandono de compra, entrenar modelo usando datos históricos -> activar estrategias como descuentos, recordatorios, etc.
- **Inteligencia artificial:** chatbot servicio al cliente, usando modelos **NLP**, recomendar productos similares.
- **Business Intelligence:** armar tablero de reporte de las métricas,
  - como mejoro el experimento de carrito abandonado?
  - interpretar resultados

# Ingeniería de datos

- Recolección, almacenamiento y procesamiento **eficiente** de datos.
  - ETL/ELTs
  - Infraestructura de datos
- **Calidad de datos:** datos precisos y confiables para ofrecer resultados útiles.
- **Escalabilidad:** manejar la carga de trabajo
- **Seguridad y privacidad:** manejar datos sensibles y confidenciales. Ej GDPR.



# Data Product

**Producto que para lograr su objetivo lo hace  
a través del uso de datos.**

***Sin datos no sería posible***

# Data Product

## ***“Los datos son el nuevo petróleo”***

- Recolectarlos a gran escala no es fácil.
- Necesitan ser **transformados** para tener valor.
- El petróleo se usa una sola vez, en cambio los datos son persistentes, se pueden **volver a utilizar**.
- Los datos son como el agua: son esenciales, necesitan ser **limpiados y accesibles** para todos
  - *Democratizar el acceso a datos, deben ser interpretables.*
  - *Calidad de datos.*

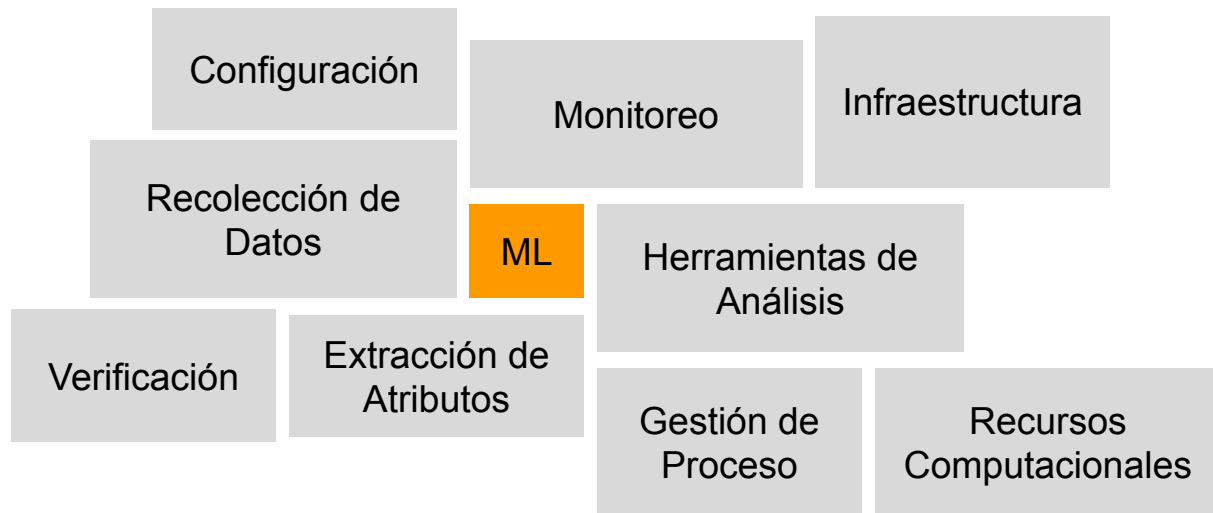
# Data Product - *Desafíos*

**Product - Market fit:** Desafío de producto tecnológico tradicional

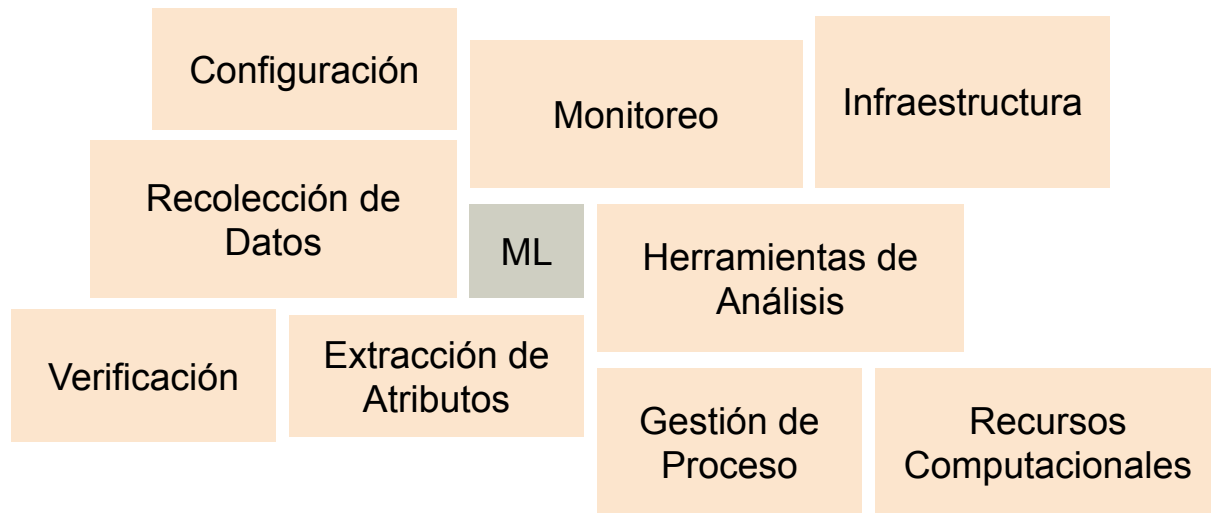
**DATOS** - Un producto basado en datos tiene a su vez otra fuente de **incertidumbre**:

- La predicciones nunca van a ser 100% correctas
- Es muy difícil garantizar la performance de un modelo sin desarrollarlo.
  - Quizás no haya datos o señal suficiente.
- La performance de un modelo en desarrollo puede variar significativamente al ponerlo productivo.

# Producto de datos



# Producto de datos

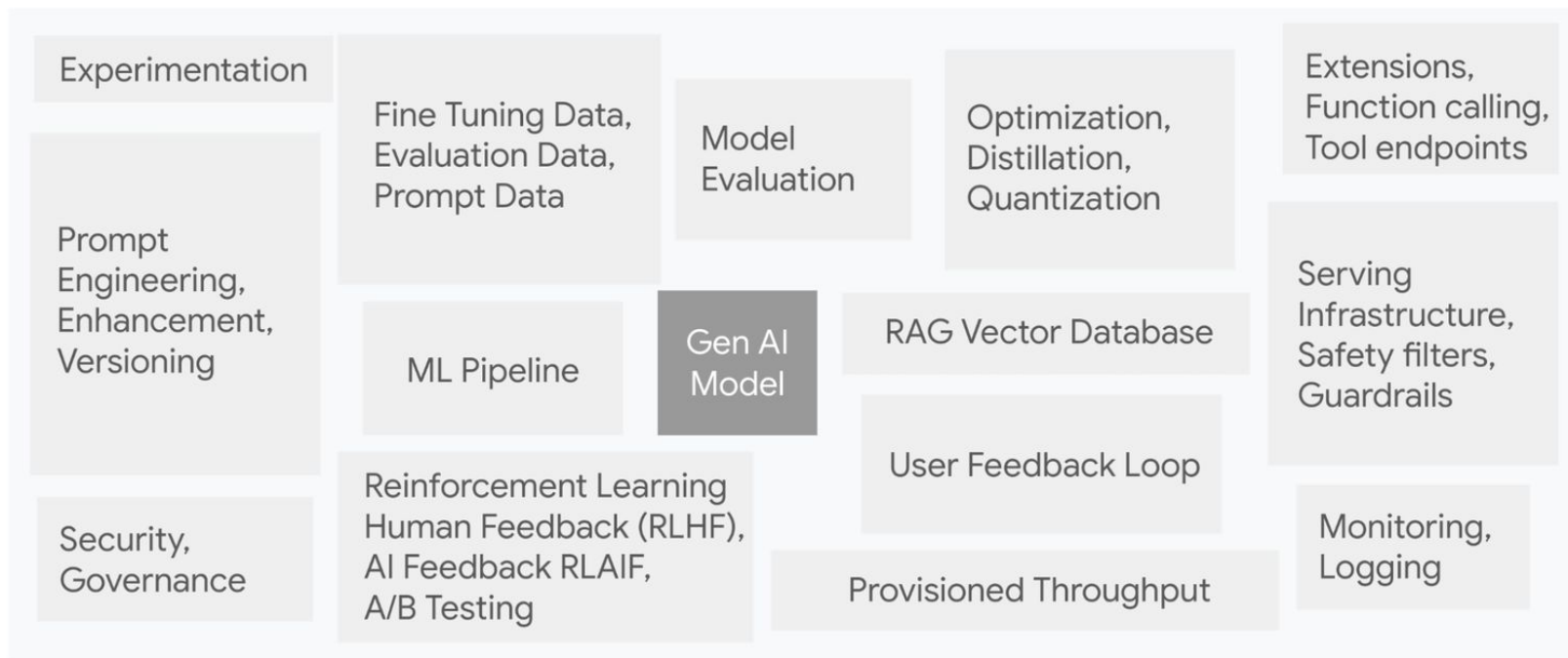


## **Tendencia:**

Modelo de ML -> Commodity

Data pipelines y analisis -> ventaja competitiva

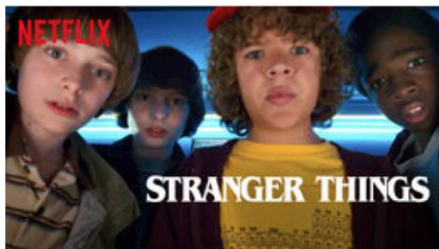
# Gen AI



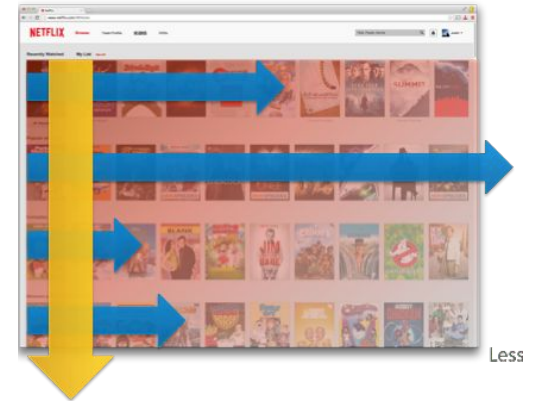
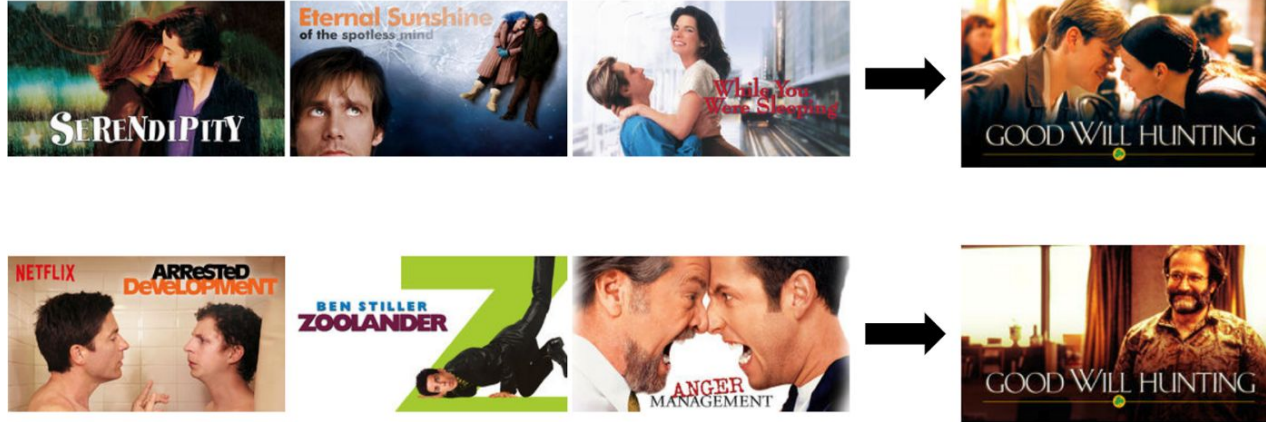
<https://cloud.google.com/blog/products/ai-machine-learning/learn-how-to-build-and-scale-generative-ai-solutions-with-genops>



# Netflix - *Recomendaciones*



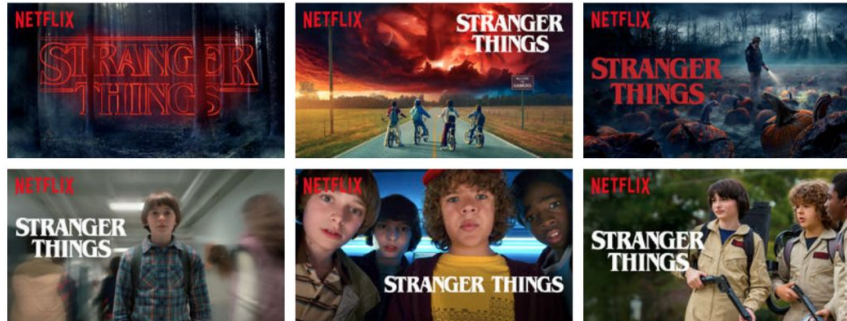
# Netflix - Recomendaciones



# Actividad - Netflix

## Contesten a las siguientes preguntas:

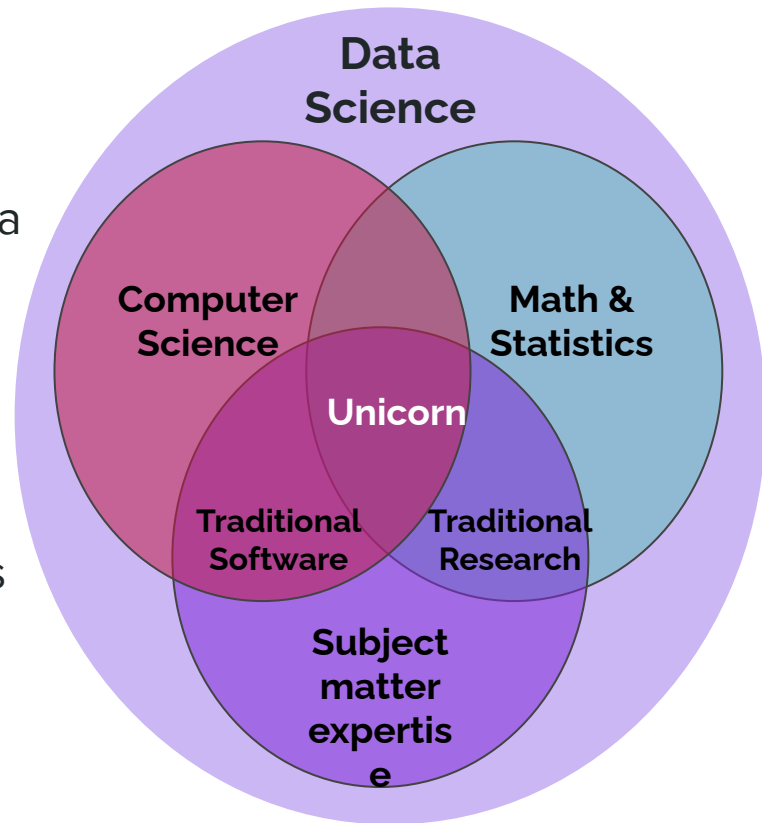
- ¿Cómo hace Netflix para recomendarte películas?
- ¿Por qué a algunos muestra letras rojas y a otros blancas, con distintas letras de fondo?
- ¿Cómo los ordena / rankea?
- ¿Qué datos utiliza para recomendarte?
- ¿Cuáles son los componentes principales?



# Ejecución de un proyecto

## Equipos interdisciplinarios:

- **Data Scientists:** Genera hipótesis, experimenta, modela.
- **Data Engineers:** Construye pipelines, asegura acceso a datos limpios.
- **Analytics Engineers:** Crea datasets listos para análisis / BI.
- **Data Analysts:** provee insights al negocio.
- **Machine Learning Engineer:** Lleva modelos a producción.



# Pasos en un proyecto de Datos

Situación problemática

Recolección de datos

**Análisis y exploración**

- ¿Qué variables están disponibles?
- ¿Qué distribución tienen?
- ¿Cómo se relacionan las distintas variables?

**Definición de la tarea**

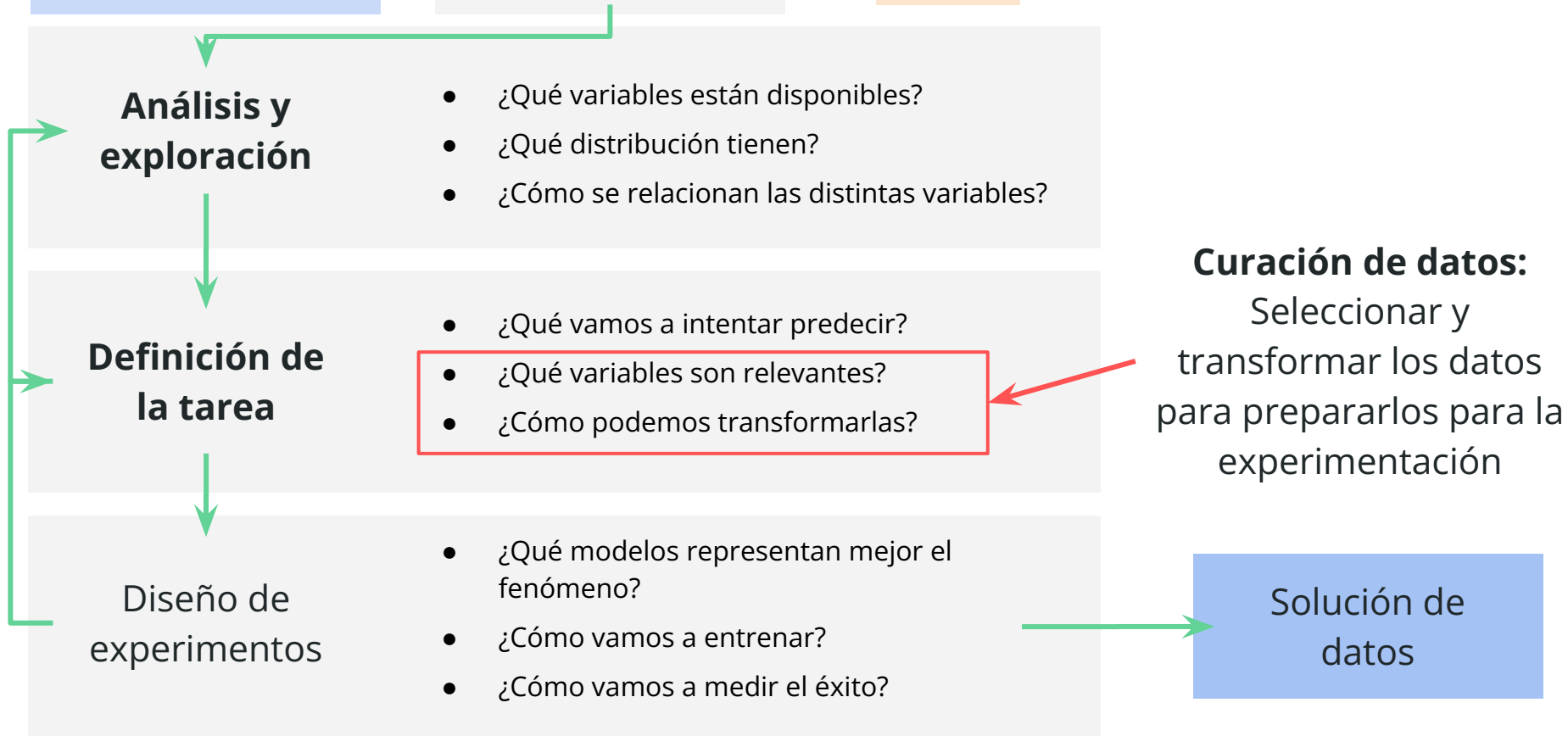
- ¿Qué vamos a intentar predecir?
- ¿Qué variables son relevantes?
- ¿Cómo podemos transformarlas?

**Curación de datos:**  
Seleccionar y transformar los datos para prepararlos para la experimentación

Diseño de experimentos

- ¿Qué modelos representan mejor el fenómeno?
- ¿Cómo vamos a entrenar?
- ¿Cómo vamos a medir el éxito?

Solución de datos



# Identificar el problema - Definir Casos de Uso

- Hablar con los especialistas/stakeholders y entender objetivos
- Definir métricas de **éxito** del proyecto.
- ¿Qué restricciones tenemos (datos, tiempos, negocio)?
- ¿Cuáles son los beneficios de priorizar este problema?

***Ej: Reducir la tasa de abandono de carritos en ecommerce.***

# Recolectar datos

- ¿Qué datos existen ya?
- ¿Qué datos faltan?
- ¿Qué calidad tienen esos datos?
- ¿Existen fuentes externas (APIs, Open Data)?
  - a. ¿Qué fuentes de datos conocen los especialistas?
  - b. ¿Hay datos en Internet que puedan servir para resolver el problema?
- Definir pipeline e infraestructura de datos
- Pensar en la **privacidad**, los **sesgos** y el impacto de nuestras decisiones.

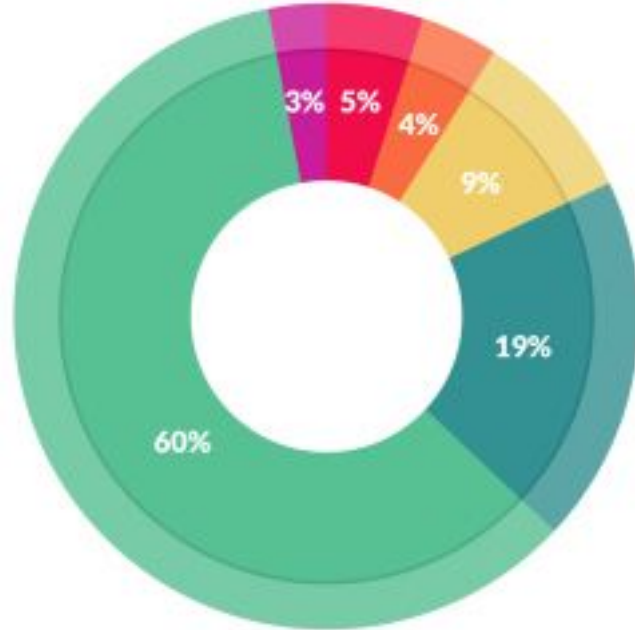
*Ej: Logs de navegación web, historial de compras.*

# Exploración de datos

- Análisis exploratorio inicial (EDA).
  - a. Identificar patrones y relaciones significativas
  - b. Estadística descriptiva
- Para decidir los procesos de curación, tenemos que entender nuestros datos **en conjunto**. Incluye:
  - a. Técnicas más complejas para el análisis de datos que permiten relacionar múltiples variables.
  - b. Técnicas de visualización de datos no estructurados



# Limpieza y Exploración de datos



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

***La mayoría de los proyectos fallan no porque el modelo sea malo, sino porque los datos no estaban listos o entendidos.***

# Curación de datos

- Detección y tratamiento de:
  - a. Valores faltantes.
  - b. Outliers.
  - c. Variables redundantes o poco informativas.
- Transformaciones necesarias:
  - a. Normalización.
  - b. Codificación de categorías.
  - c. Ingeniería de features.

***(Ej: Identificar usuarios activos vs inactivos.)***

# Curación de datos

Problema	Datos	Decisiones de curación
Predecir los salarios de los programadores en Argentina en 2020	Encuesta voluntaria con columnas edad, género, años de experiencia y salario	<ul style="list-style-type: none"><li>• Eliminar edades menores que 18 y mayores que 99</li><li>• Eliminar salarios mayores que 1 millón de pesos</li><li>• Estandarizar los años de experiencia de tal forma que la media sea 0.</li><li>• Re-escalar las edades en un rango de 1 a 0, tal que 18 años o menos corresponda a 0 y 70 años o más corresponda a 1.</li><li>• Eliminar la columna género.</li></ul>
Predecir el precio de una propiedad	Base de datos gubernamental con registros de transacciones inmobiliarias. Tiene precio, fecha y ubicación.	<ul style="list-style-type: none"><li>• Eliminar día y mes de la transacción.</li><li>• <i>Escrapear</i> sitios de compra/venta para extraer información adicional sobre cada propiedad.</li><li>• Imputar los valores faltantes utilizando estimaciones en base a ejemplos parecidos.</li></ul>

# Desarrollar y entrenar modelo

- Entrenar modelos predictivos, seleccionar algoritmos y variables con las que entrenar, analizar su rendimiento
- ¿Tenemos una hipótesis de experimentación?
- ¿Si desarrollamos un modelo predictivo o de machine learning, ¿es el modelo trivial?  
¿se basa el modelo únicamente en datos anteriores a lo que queremos predecir?  
¿entendemos cómo funciona el modelo? ¿es ético?

***Ej: Modelo de predicción de abandono usando Random Forest.***

***Error común:*** *querer aplicar modelos a casos de uso donde no necesariamente aplican trivialmente (LLMs para detección de objetos o recomendación).*

# Modelar sin curar

- Error común: saltar a modelar antes de explorar
- "garbage in, garbage out".

ID Casa	Superficie (m <sup>2</sup> )	Habitaciones	Precio
1	120	3	150,000 USD
2	-80	2	80,000 USD
3	300	5	2 BTC
4	0	1	20,000 USD
5	500	1000	500 millones ARS
6	85	2	(vacío)

# Evaluar modelo

- Qué tan bien predice o clasifica el modelo?
- ¿Qué métricas usamos? (Precisión, recall, F1-score, RMSE, etc.)
- ¿El modelo generaliza o está sobreajustado?

***Ej: Validar que el modelo predice mejor que el azar.***

En modelos de AI / LLMs, este paso es mucho más desafiante ya que son modelos no determinísticos y la evaluación se hace más compleja.

# Humanos detrás del ML / AI

EL PAÍS | ECONOMÍA | SOCIEDAD | CULTURA Y ESPECTÁCULOS | EL MUNDO | DEPORTES | PSICOLOGÍA | CONTRATAPA SECCIONES

**Página12** Edición Impresa | 07 de junio de 2018  
Hoy: UNIVERSIDAD NO

CULTURA Y ESPECTÁCULOS  
10 de abril de 2018

Un trabajo soñado para los fanáticos del streaming

## Netflix paga por ver y etiquetar series y películas

Twitter Facebook YouTube Print



Arts and Entertainment

## Netflix tagging: Yes, it's a real job

By Jhaan Elker June 11, 2015 Email the author



Josh Garrell is a Netflix tagger. His job? He is paid to watch television shows and movies for hours a day. (Jhaan Elker/The Washington Post)

Starting in the wee hours of June 12, hundreds of thousands of fans are  
goodreads.com/page/ElkerJhaan/2015 Netflix's original production "Orange Is

## Puesta en producción - Despliegue

- Modelo validado debe implementarse en un entorno productivo para uso en el mundo real
- Integrar el modelo en un sistema existente
- Implementar interfaces de usuario para ser consumido (API).
- Definir alertas de fallos o errores.

***Ej: Activar un modelo de recomendación en el sitio web.***



# Netflix - Competencia

- Premio U\$S 1 Millón.
- Sistema de recomendaciones.
- Predecir películas que le podría interesar a cada usuario.
- Netflix nunca pudo utilizar la solución ganadora debido a los costos de ingeniería requeridos.



## Visualizar y comunicar resultados

- Evaluar, reportar y comunicar resultados.
- Facilitar la interpretación de los datos
- Promover la toma de decisiones informada / basada en datos.
- Storytelling
  - a. ¿Entendemos a quién le estamos comunicando, cuáles son sus conocimientos e intereses y cómo difieren de los nuestros?

# Visualización de datos: Dashboards



**Superset:** Open Source <https://superset.apache.org/>

# Monitoreo y mantenimiento - Mejora continua



Tareas:

- Seguir la performance del modelo en producción.
- Detectar degradaciones.
- Actualizar datos, reentrenar modelos periódicamente.
- Incorporar feedback de usuarios.

***Ej: Si cambian los patrones de compra, ajustar el modelo.***

*Un proyecto de ciencia de datos no termina cuando entrena un modelo. Termina cuando **genera valor real**, en producción, y se mantiene vivo en el tiempo.*

***ITERAR!***

*Las personas con mejor entendimiento de problemas, claridad para resolver, capacidad de aprender, y obsesión por iterar serán las que marcarán la diferencia.*

# Recolección de Datos

---

**¿Qué datos generamos  
cuando salimos a correr  
escuchando música?**

# Fuentes de datos

- Internas / Externas / Comerciales / Públicas
- Transaccionales: Compras, ventas, cotización de acciones en la bolsa, etc.
- Conversacionales: chats, mails, grabaciones telefónicas.
- Fotos y videos: subir a las redes sociales, etc.
- Sensores: posición del gps, acelerómetro, etc.
- IoT: smart tv, smart watch, alexa, etc.
- registro de actividades: escuchar música, leer un libro, buscar en google, comprar un artículo en un ecommerce (metadatos).

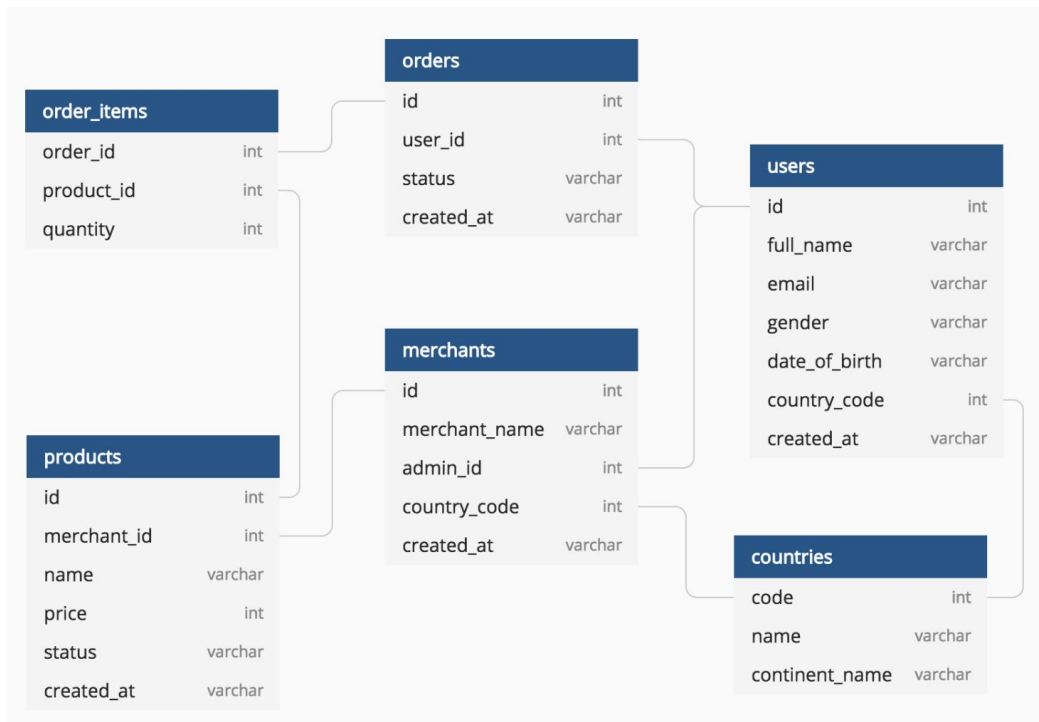


# Estructura de los datos

- Llamamos “datos” a un conjuntos de registros.
- Cada registro tiene asociado un conjunto de características, y las características pueden relacionarse de manera compleja.
- Distintas estructuras suelen almacenarse con formatos de archivo particulares
  - La estructura de los datos no es lo mismo que el tipo de base de datos o archivos en los que se almacena

# Datos estructurados

- Todos los registros tienen las mismas características con el mismo tipo
- Filas y columnas, formato consistente
- Fácil de almacenar, consultar y analizar.



- Archivos en formato CSV, parquet, etc.
- Bases de datos relacionales como MySQL, Postgres

# Datos semi-estructurados

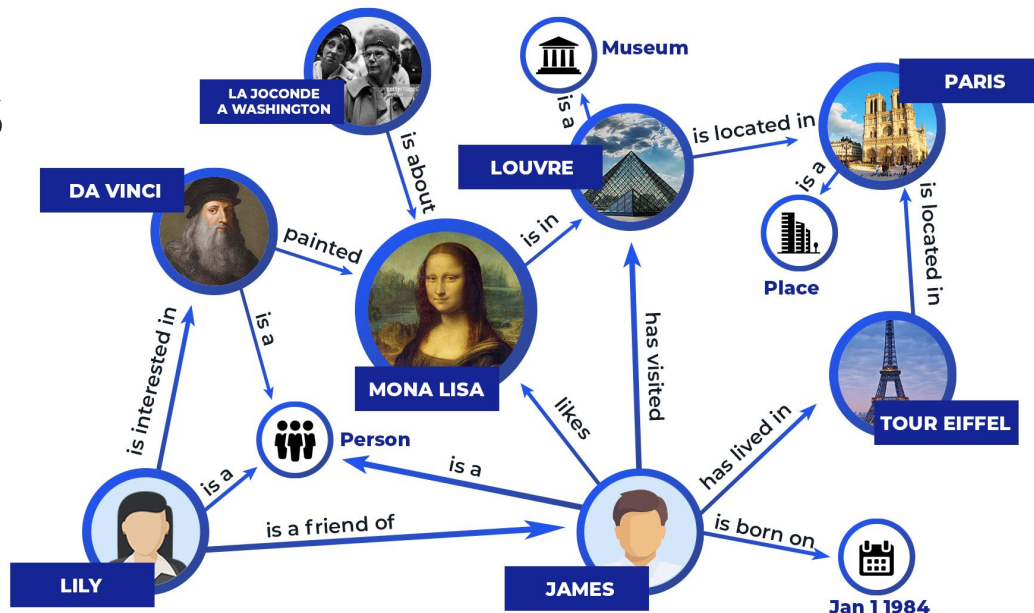
- Cada registro tiene un conjunto distinto de características
- Los registros pueden estar anidados
- Formato flexible, pero requiere procesamiento adicional para analizarlo

```
{ "orders": [  
  {  
    "client_id": 1458,  
    "items": [  
      { "description": "Empanadas" , "amount": 12},  
      { "description": "Salsa picante" , "amount": 1}  
    ],  
    "total": 950,  
    "payment_method": "cash"  
  },  
  {  
    "client_id": 985,  
    "items": [  
      { "description": "Lomito Completo" , "amount": 2,  
        "observations": "Uno sin huevo" }  
    ],  
    "total": 1400,  
    "payment_method": "debit",  
    "debit_card": "Maestro"  
  }  
]}
```

- Archivos en formato JSON
- Bases de datos no relacionales como MongoDB

# Datos semi-estructurados

- Los registros pueden tener relaciones complejas
  - Jerarquías
  - Estructura de grafo (Twitter)



- Triplas RDF
- Graph-oriented databases

# Datos no estructurados

- Colecciones de distintos tipos:
  - Documentos de texto
  - Imágenes, Audio
- Pueden o no tener metadatos asociados
- Requiere técnicas especiales para su procesamiento y análisis



# Tipos de datos

<b>Tipo de Datos</b>	<b>Ejemplos</b>	<b>Dificultades</b>
Estructurados	Bases SQL, hojas de cálculo	Duplicados, columnas mal definidas
Semiestructurados	JSONs, logs de app	Inconsistencias de formatos
No estructurados	Texto, imágenes, videos	Procesamiento y almacenamiento

# Combinar fuentes de datos

Problemas al juntar distintas fuentes:

- Distintas unidades (ej: peso en kg vs libras).
- Diferentes convenciones de nombres.
- Tiempos desincronizados (ej: eventos que no matchean).

***Ej Ecommerce: integrar pedidos + clientes + envíos + soporte.***

# Data Integration

proceso de **unir datos de distintas fuentes** para que estén disponibles en un **formato coherente y unificado**, listo para ser analizado o usado en modelos.

- En el mundo real, los datos vienen de muchos lados: bases de datos internas, APIs, archivos Excel, sensores, logs.
- Cada fuente puede tener su propio formato, estructura, calidad y tiempos.
- Sin integración, es imposible analizar todo junto.

***Sin integración, cada fuente de datos te cuenta solo una parte de la historia.***



# Desafíos comunes en recolección de datos

## ¿Qué puede salir mal? ⚠️

- Datos incompletos o inconsistentes.
- Fuentes de datos que cambian con el tiempo (versionado de APIs).
- Falta de estándares (nombres, unidades, formatos).
- Duplicados o datos corruptos.
- Problemas de permisos y acceso.

***Recolectar datos no es solo copiarlos: es entender, validar y controlar la calidad desde el inicio.***

# Privacidad de Datos

## ¿Qué implica?

- Proteger la información personal y sensible de los usuarios.
- Asegurar que los datos se usen sólo con consentimiento y para fines legítimos.
- Ejemplos de datos sensibles:
  - Nombre, DNI, dirección, datos biométricos, historial médico, ubicación en tiempo real.

## Principales leyes que deben conocer:

- GDPR (Europa): Derecho a ser olvidado, portabilidad de datos, consentimiento explícito.
- Ley de Protección de Datos Personales (Argentina - Ley 25.326).

*¿Alguna vez completaron un formulario online y se preguntaron qué hacen con esos datos?*

*Le preguntaron a ChatGPT que saben de uds?*

# Gobierno de Datos

## ¿Qué implica?





- Definir quién es responsable de los datos, cómo se almacenan, usan y protegen.
- Crear reglas claras para la calidad, accesibilidad, seguridad y privacidad.

## Objetivos:

- Disponibilidad: Los datos correctos a la persona correcta, en el momento correcto.
- Calidad: Asegurar que los datos sean válidos, completos, consistentes.
- Seguridad: Proteger contra accesos no autorizados o pérdida de información.

***Gobernar los datos no es frenar la innovación, es hacerla posible de forma responsable.***

# Calidad de datos

- **Compleitud**  — Tenemos toda la información necesaria, valores faltantes?
- **Validez** — ¿El dato cumple con el formato y las reglas esperadas?
  - (ej: fechas válidas, números positivos)
- **Precisión**  — ¿Los datos reflejan correctamente la realidad?
  - (ej: coordenadas GPS exactas, medidas correctas)
- **Integridad**  — ¿Las relaciones entre los datos son consistentes?
  - (ej: una venta debe estar asociada a un cliente existente)
- **Consistencia**  — ¿Hay contradicciones o duplicados entre fuentes?
  - (ej: mismo usuario con dos fechas de nacimiento diferentes)
- **Temporalidad** — ¿Los datos están disponibles a tiempo para el análisis? ¿Son recientes y relevantes?