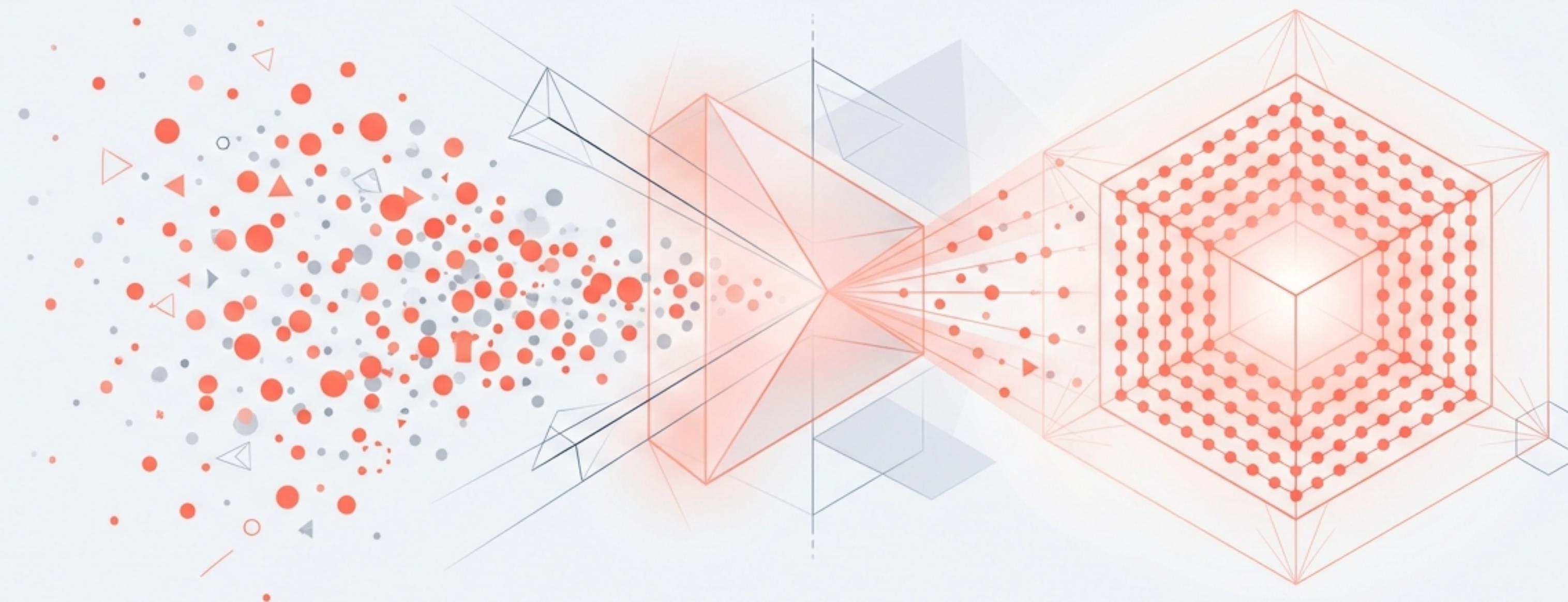


La Ciencia de Datos en el Mundo Real

El Viaje del Dato al Valor

Metodologías, Procesos y la Historia de una Transformación Empresarial



Un recurso de aprendizaje sobre metodología y aplicación práctica

Los datos son el activo. La ciencia es la llave.

“

El objetivo es convertir los datos en información y la información en conocimientos.

— Carly Fiorina
(Ex-CEO HP)

“

Los datos son cosas valiosas que durarán más que los propios sistemas.

— Tim Berners-Lee
(Inventor de la WWW)

“

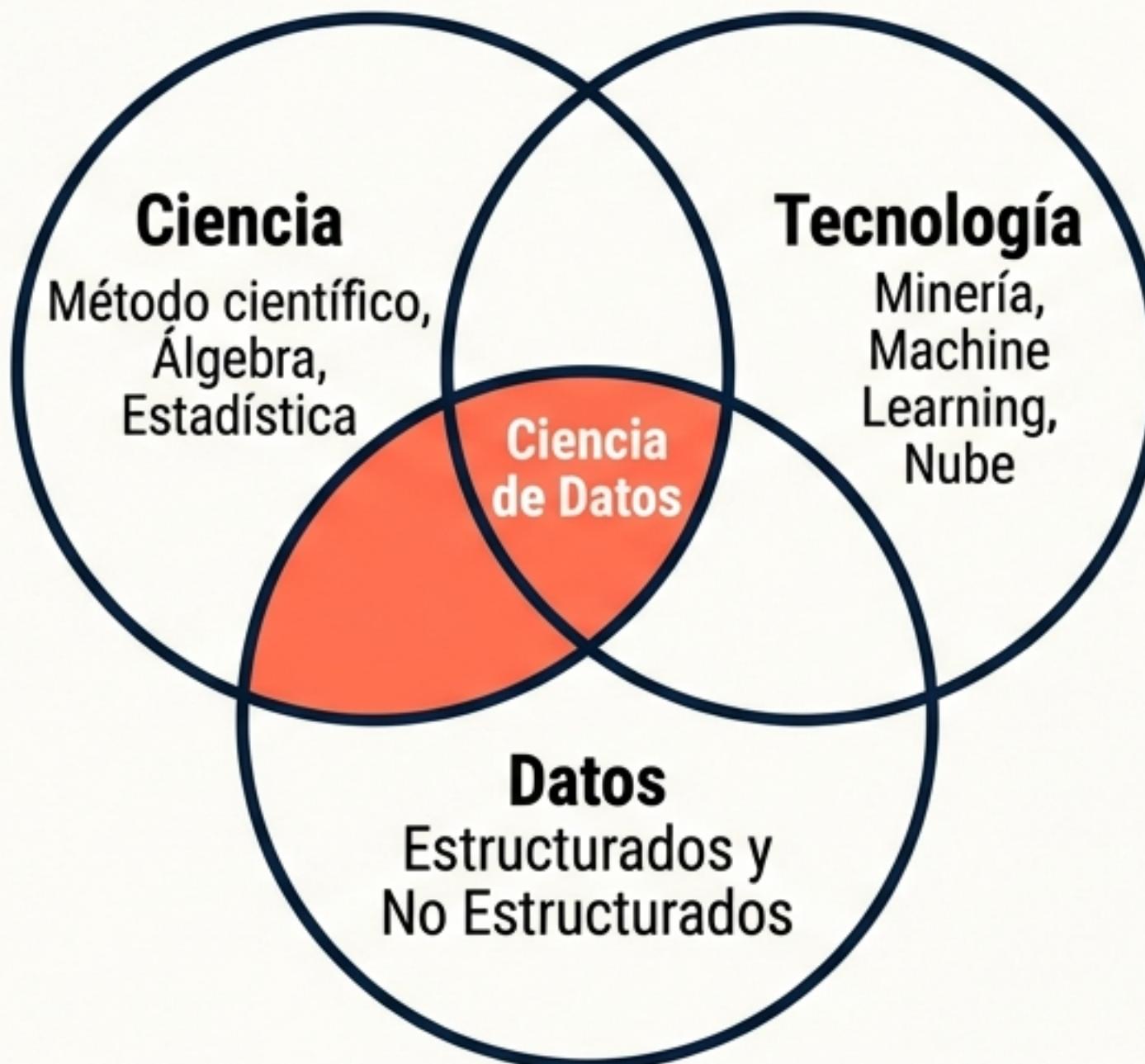
Es un requisito previo para resolver muchos de los problemas que afronta la humanidad.

— Robert Cailliau
(Co-desarrollador de la WWW)

Vivimos en un mundo digital creciente donde las empresas utilizan el método científico para descubrir información oculta.

¿Qué es realmente la Ciencia de Datos?

La intersección de disciplinas para comprender el mundo digital.



Diferenciación Clave



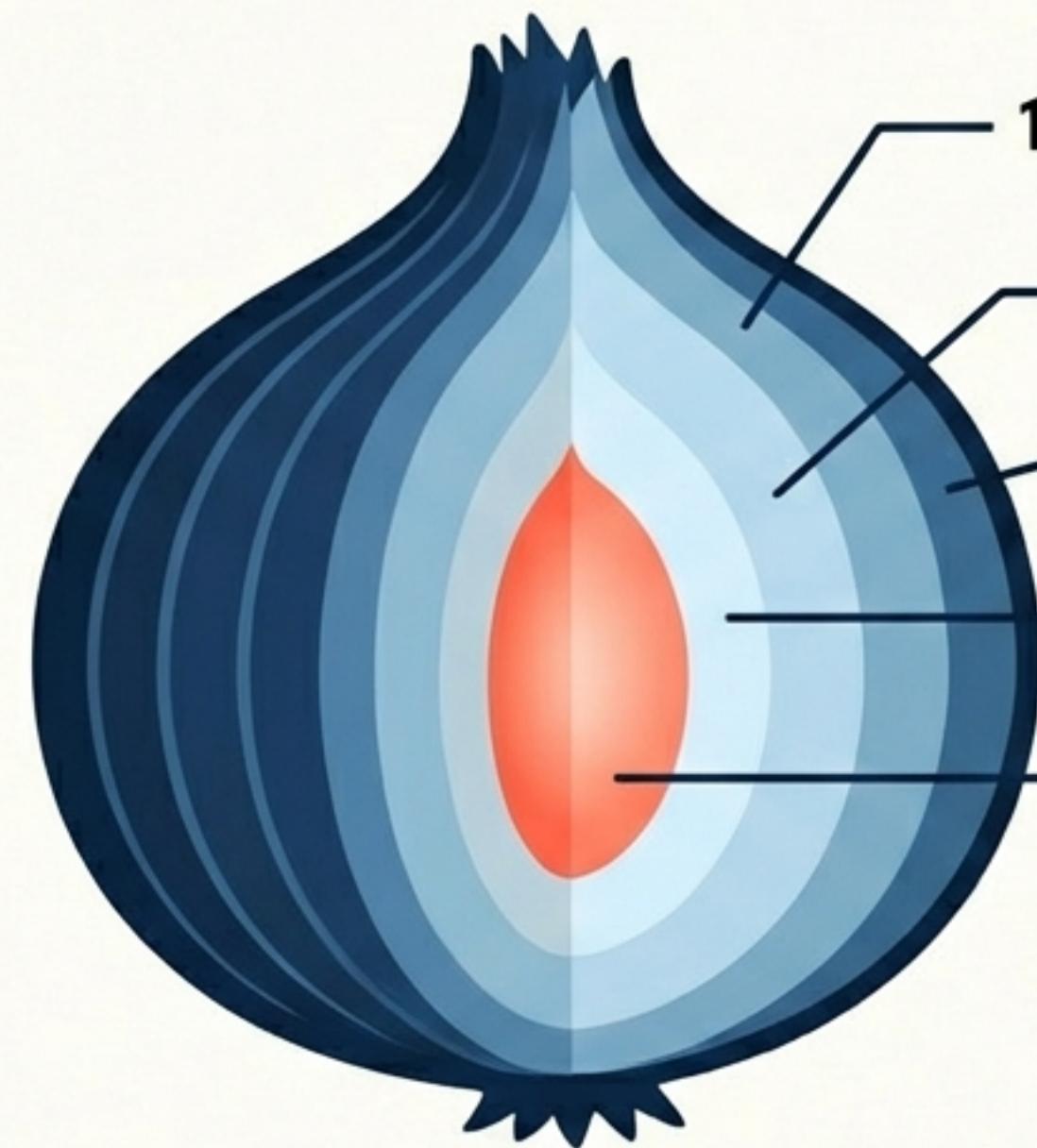
Análisis de Datos:
Examina el pasado. Busca tendencias y narra una historia con lo existente.



Ciencia de Datos: Modela el futuro. Crea nuevos procesos, usa algoritmos predictivos.

El Motor de la Ciencia: La Curiosidad Infinita

Un científico de datos no asume; pregunta.



1. ¿Por qué? (Surface Problem)

2. ¿Por qué?

3. ¿Por qué?

4. ¿Por qué?

5. ¿Por qué?
(Causa Raíz)

La Regla de Oro

Evitar suposiciones.
Encontrar la causa
raíz es el primer
paso para resolver el
problema real.

El Mapa: Metodologías Clásicas

Una metodología es la estrategia general, no la herramienta.

CRISP-DM

Cross-Industry Standard Process

- Enfoque de Negocio
- 6 Fases
- La más flexible y popular

KDD

Knowledge Discovery in Databases

- Enfoque en Datos
- 5 Pasos
- Refinamiento y Minería

SEMMA

Sample, Explore, Modify, Model, Assess

- Enfoque SAS
- Modelado Técnico
- Muy secuencial

Todas son Iterativas: El ciclo se repite para reciclar el conocimiento.

Paso 1: Comprensión del Negocio



Patrocinador Empresarial

Inicia el proyecto.
Identifica el punto débil.

Design Thinking

Empatizar con el usuario.
Definir objetivos.
Crear “personas”.

Enfoque Analítico

Traducción técnica.
(Ej: Problema de predicción
= Modelo de regresión).

Objetivo: Definir claramente el problema antes de tocar los datos.

Definiendo el Problema en GAXR

Empresa de Software de Videojuegos | 1000+ Empleados

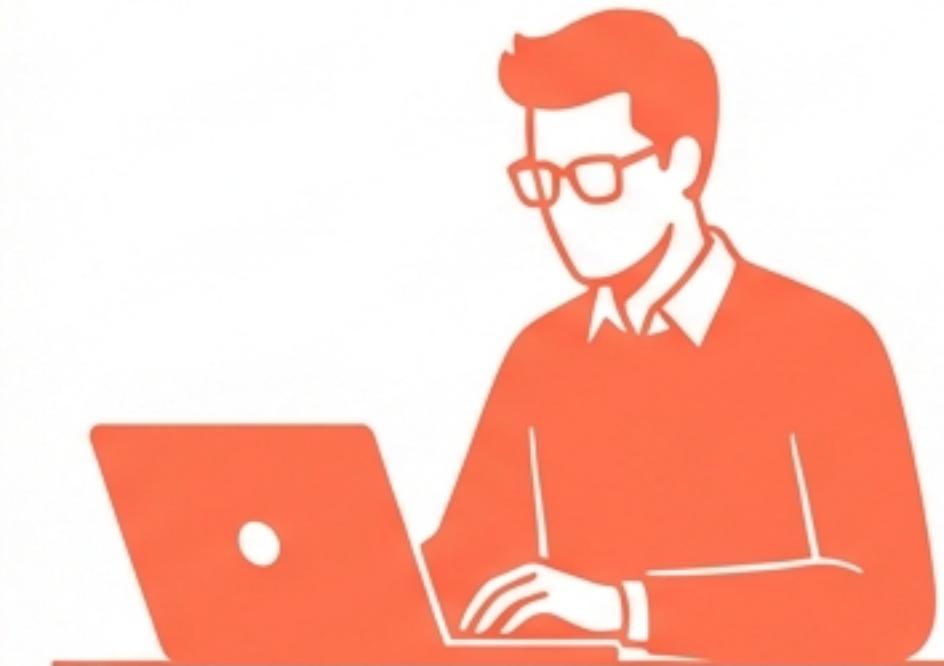
Marilyn Shah (RR.HH.)



Nuestros empleados renuncian pese a los buenos sueldos.
¿Por qué se van?
(La Deserción).

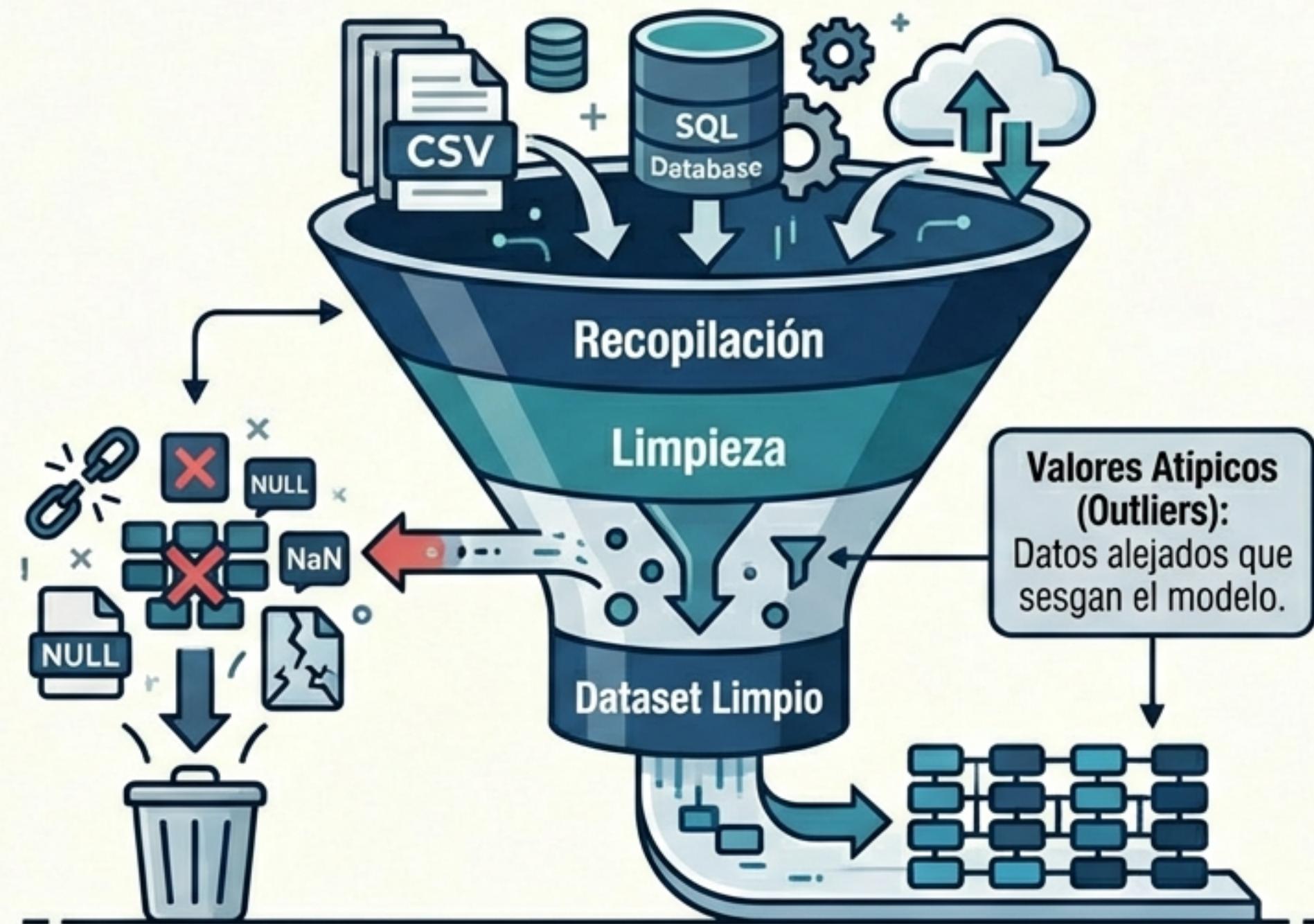


Scott Hill (Científico de Datos)



→ **Define el enfoque:** Usar datos históricos para probar la hipótesis del sueldo.

Paso 2: Exploración y Preparación



El 80% del trabajo de un científico de datos ocurre aquí.

Limpieza de Datos en GAXR



Datos Crudos (SQL Extract)

- 👤 Nombre / Apellido
(Privacidad)
- 👤 Número de empleado
(Redundante '1')
- 📞 Teléfono
(Irrelevante)
- 💰 Sueldos con decimales



CSV Final (1470 Registros)

- ✓ Edad
- ✓ Departamento
- ✓ Distancia desde casa
- ✓ Ingresos (Redondeados)
- ✓ Estado Civil

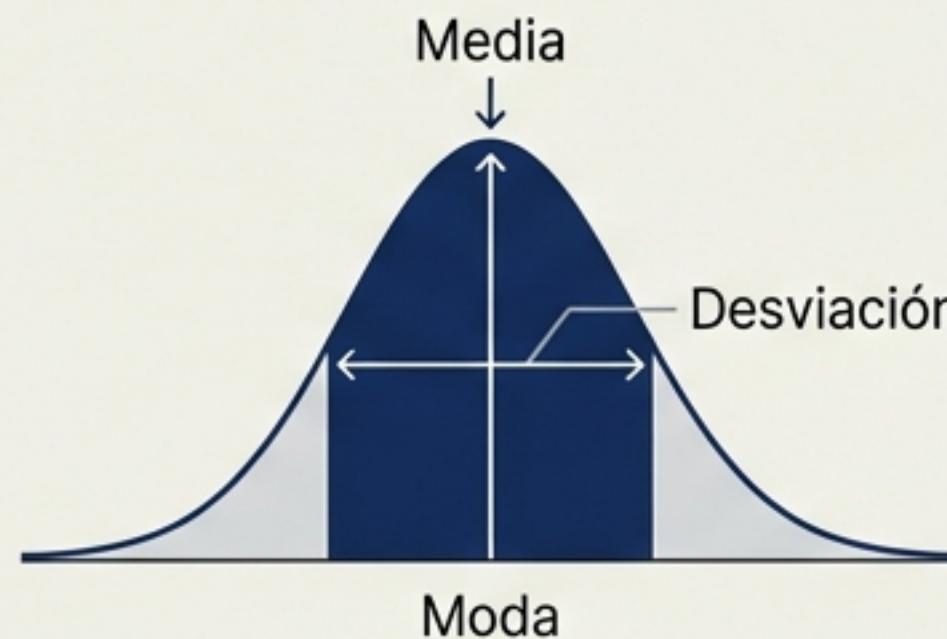
```
SQL Query
SELECT * FROM employees
ORDER BY hiredate DESC LIMIT 10
```



Paso 3: Representación y Transformación

Estadística Descriptiva y Normalización

Estadística Descriptiva



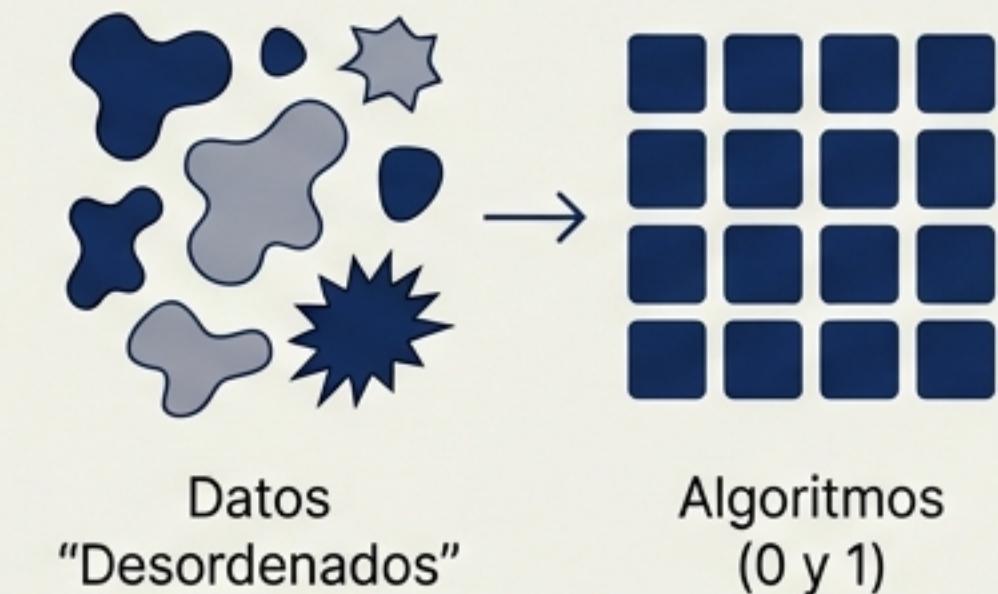
¿Qué ocurre? Entendiendo el centro y la distribución.

Tokenización



Convertir texto en elementos contables.

Normalización



Datos "Desordenados"

Algoritmos (0 y 1)

Convertir datos "desordenados" en formatos para algoritmos (0 y 1).

La Transformación Clave

	A	B	C	D	F	G	H
1		Deserción (Attrition)					
2		Sí					
3		No					
4		Sí					
5		No					

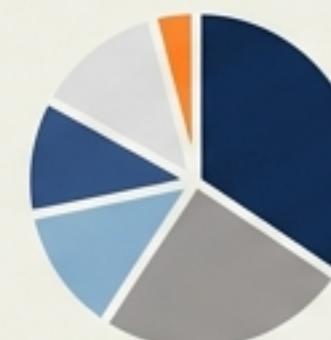
Datos Dicotómicos:
Esta variable es la clave.
Servirá como "Etiqueta" para
entrenar el modelo supervisado.

Scott fusiona esta nueva fuente de datos para permitir que el modelo aprenda quién se fue y quién se quedó.

Paso 4: Visualización de Datos

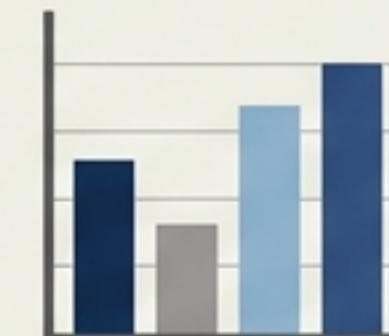
Elegir la herramienta correcta para la historia.

Gráfico Circular (Pie)



Proporciones **relativas**
(Partes de un todo).

Gráfico de Barras



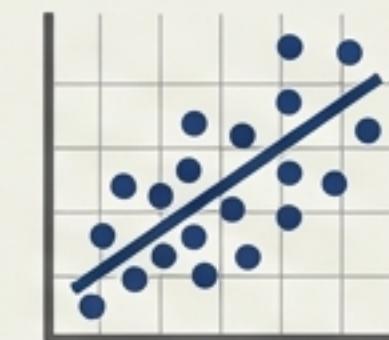
Comparación entre categorías.

Gráfico de Líneas



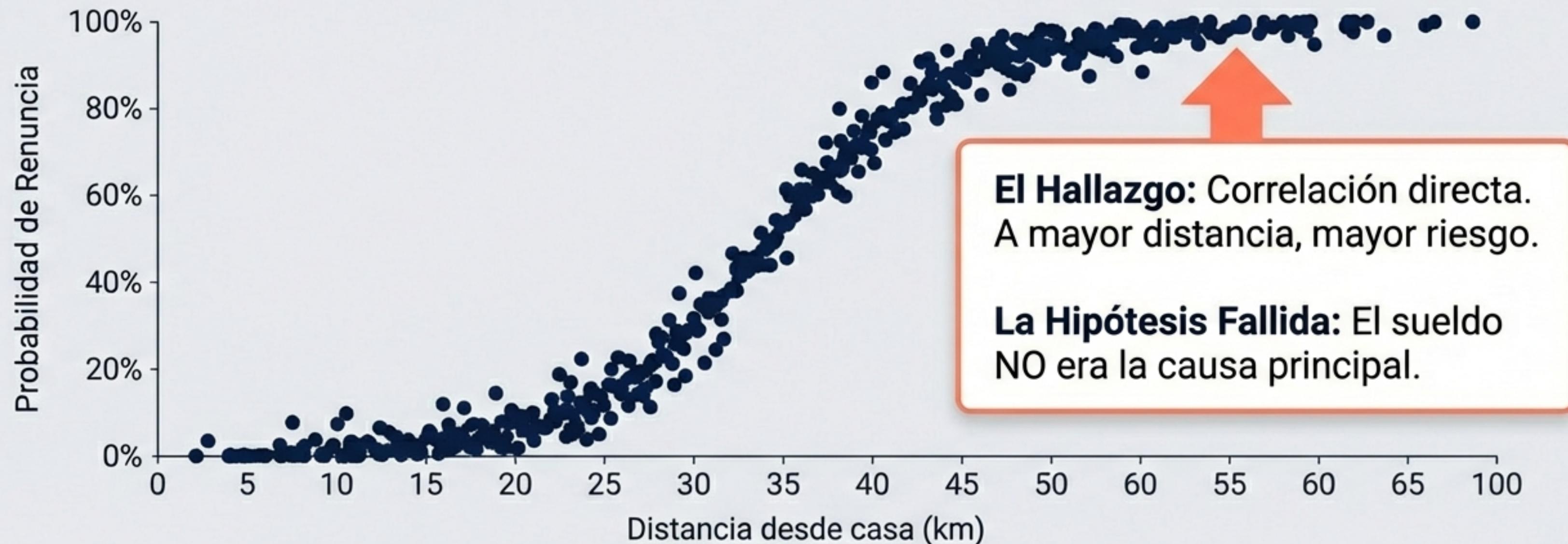
Tendencias a lo largo del tiempo.

Gráfico de Dispersión (Scatter)



Relación entre variables y detección de patrones.

El Momento "¡Ajá!"

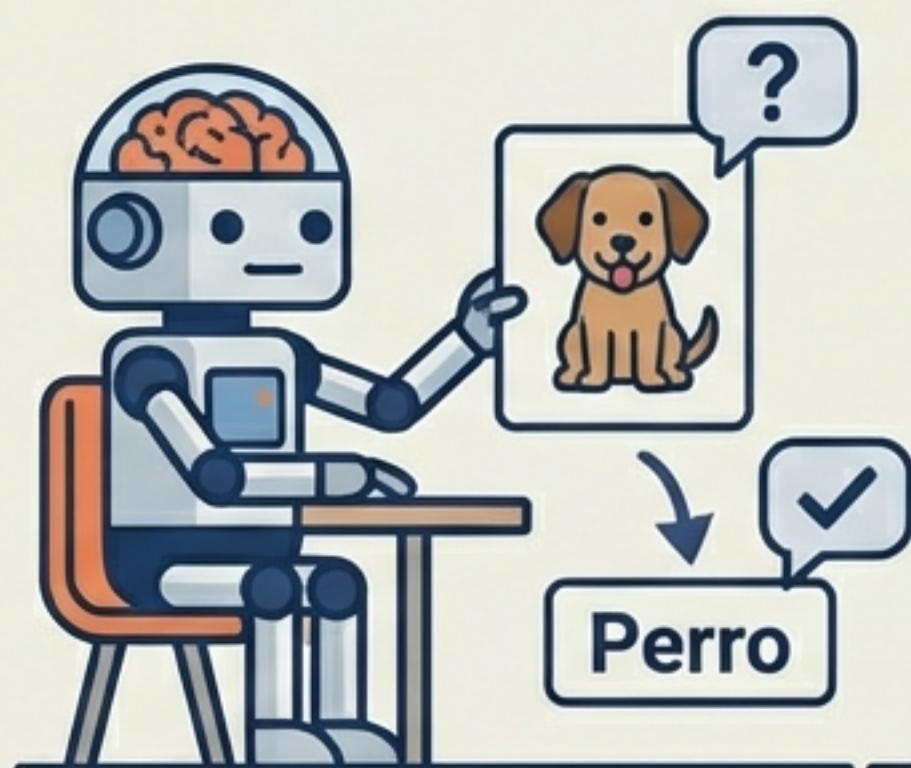


Propuesta: Marilyn propone Trabajo Remoto
(2 días/semana) como solución inmediata.

Paso 5: Modelado de Datos

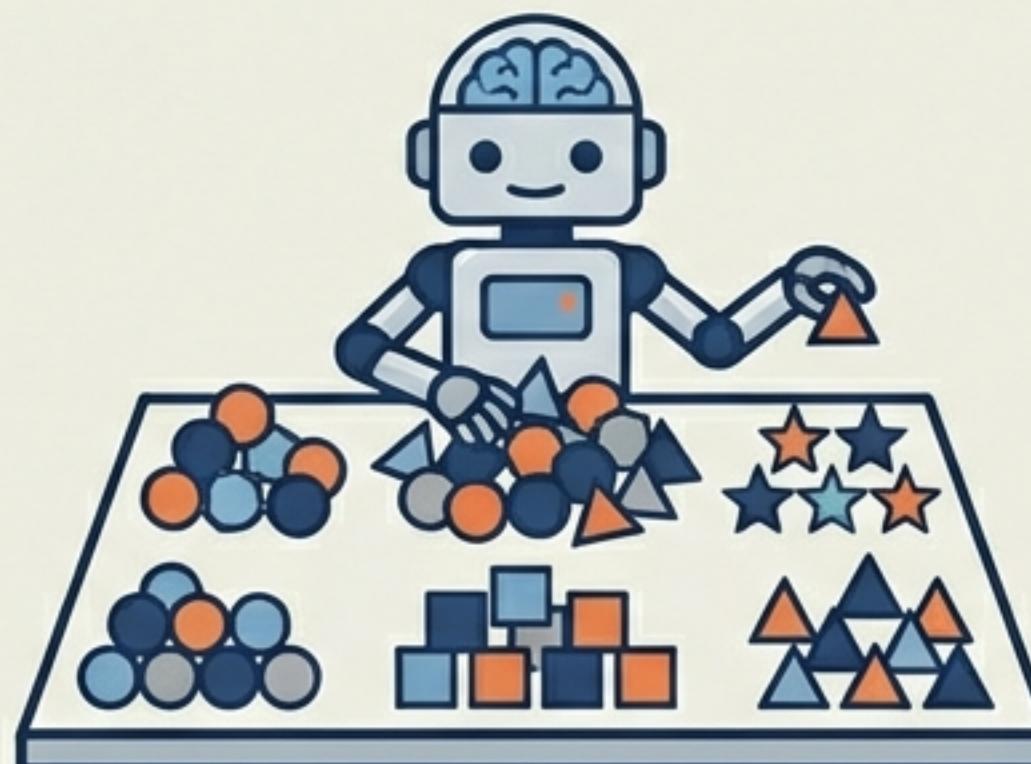
Entrenando a la máquina para predecir el futuro.

Supervisado



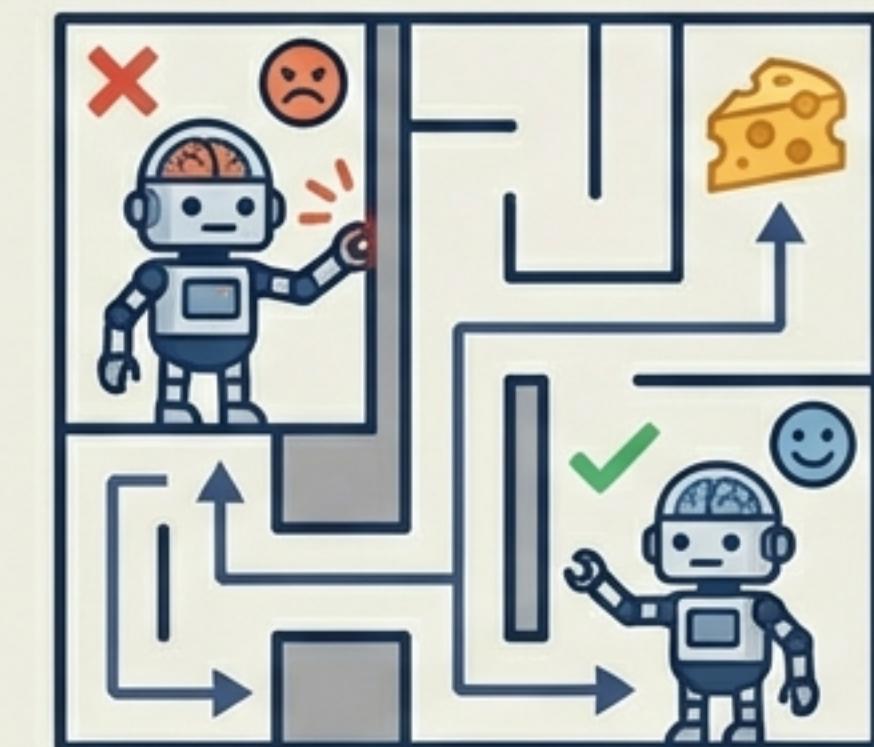
Datos Etiquetados.
Pregunta + Respuesta.

No Supervisado



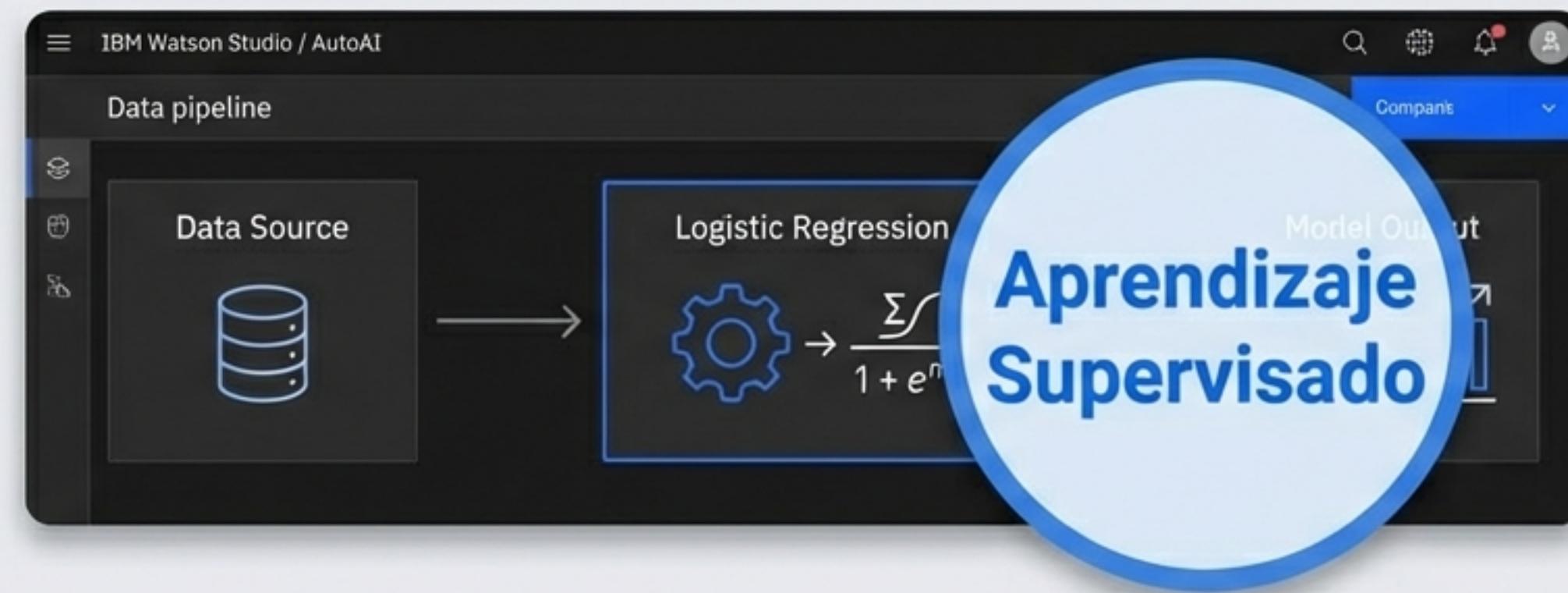
Datos No Etiquetados.
Busca patrones ocultos.

Reforzado



Ensayo y error.
Recompensa y castigo.

Construyendo el Modelo Predictivo



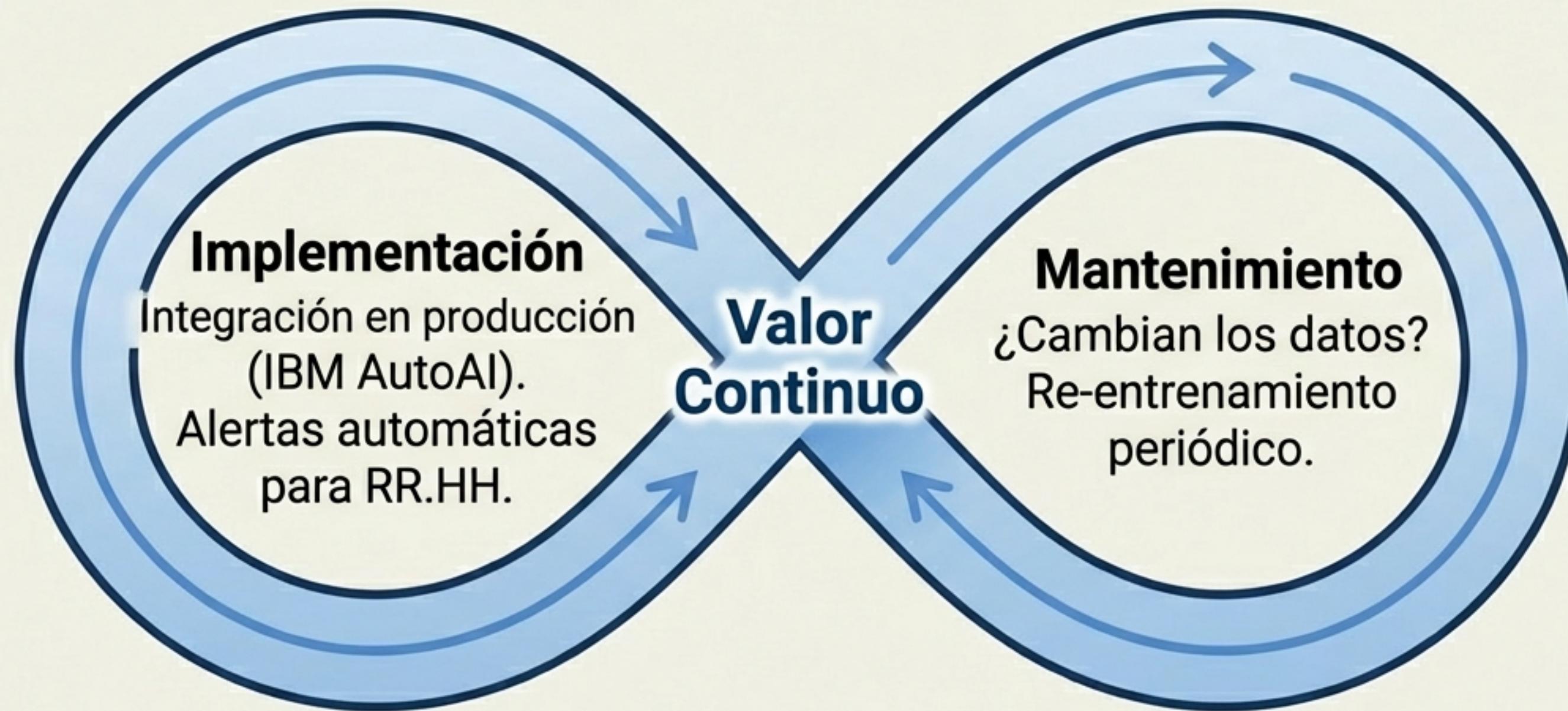
Entrada: Datos históricos con etiqueta 'Deserción' (Sí/No).

Algoritmo: Regresión Logística.

Resultado: El modelo confirma matemáticamente que la 'Distancia desde casa' es el predictor más fuerte, descartando el Estado Civil o el Departamento.

Futuro: Capacidad de estimar la probabilidad de renuncia de *nuevos empleados*.

Paso 6: Implementación y Mantenimiento



El modelo ahora vive en el sistema de RR.HH., alertando proactivamente sobre empleados en riesgo por desplazamientos largos.

La Ciencia de Datos es un Deporte de Equipo



Impacto en el Mundo Real

Más allá de RR.HH.: Transformación Global



Salud

Diagnósticos personalizados y predicción de enfermedades.



Transporte

Optimización de rutas en tiempo real.



Deportes

Evaluación precisa del rendimiento.



Comercio

Prevención de fraude y anuncios automatizados.

Gigantes como Amazon, Google y Facebook basan sus decisiones críticas en estos modelos.

Conclusión: El Ciclo Sin Fin

1. **Materia Prima:** Los datos son el activo; la ciencia es el proceso de refinamiento.
2. **El Mapa:** Metodologías como CRISP-DM nos guían a través del caos.
3. **El Valor:** Surge al conectar la limpieza técnica con una necesidad real de negocio.

Mantente Curioso.

La herramienta más poderosa no es el algoritmo, es la pregunta '¿Por qué?'.
© NotebookLM