



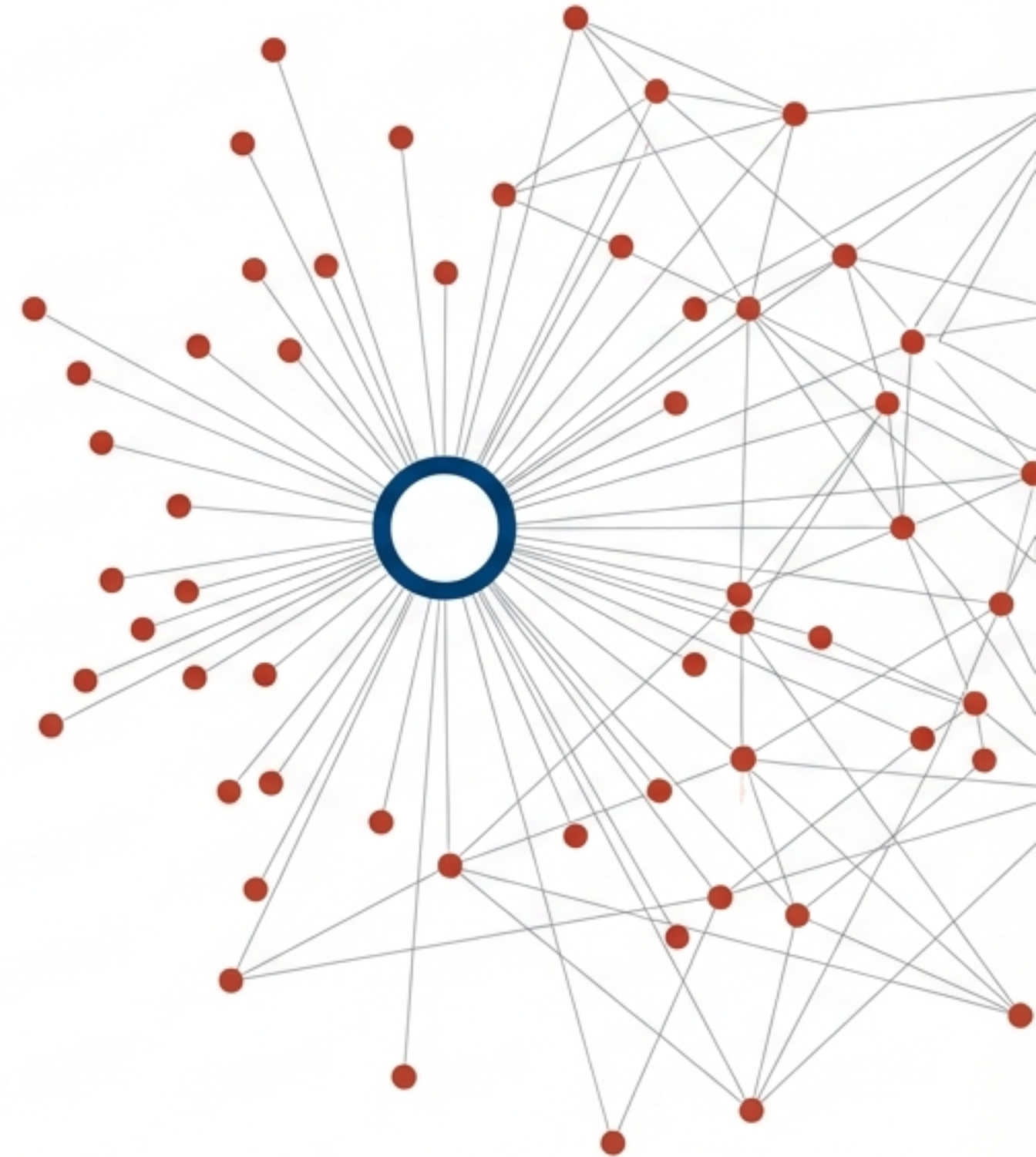
UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY



FACULTAD DE
INGENIERÍA

Aprendizaje Federado para Datos Sensibles: Visión, Factibilidad y Aplicación Real

Un análisis profundo sobre la privacidad de datos sin sacrificar la inteligencia del modelo. Basado en la investigación de Agustín Tornaría (Facultad de Ingeniería, UDELAR).



El Paradigma de la Privacidad: ¿Cómo aprender de datos que no podemos ver?

La Paradoja de los Datos

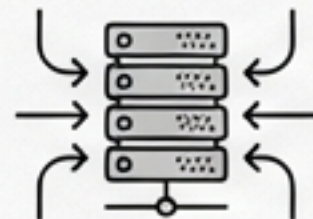
- Vivimos en un mundo rico en datos (IoT, Móviles, Instituciones).



- Centralizar datos sensibles (financieros, médicos, educativos) es arriesgado o ilegal.



- El ML tradicional requiere datos centralizados.



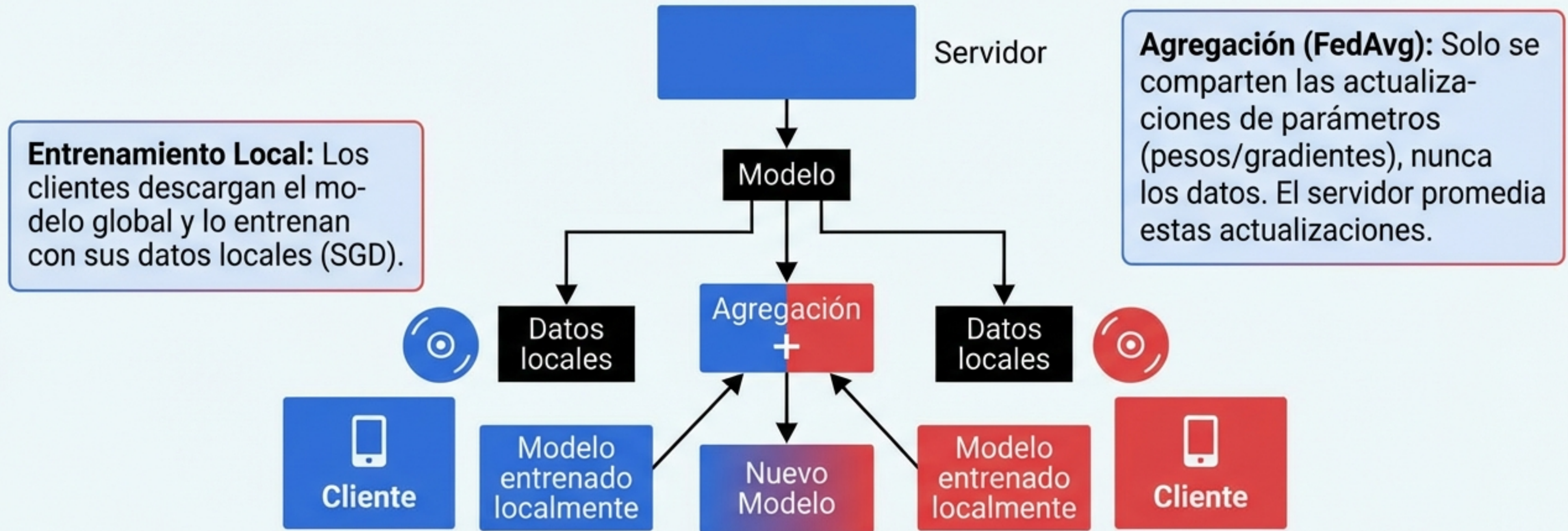
Cambio de Arquitectura

La Solución: Aprendizaje Federado (FL)

Llevamos el código a los datos.

Permite entrenar modelos robustos de Machine Learning colaborativos sin que los datos crudos salgan jamás de su dispositivo o institución de origen.

Fundamentos Teóricos: Descentralización del Entrenamiento



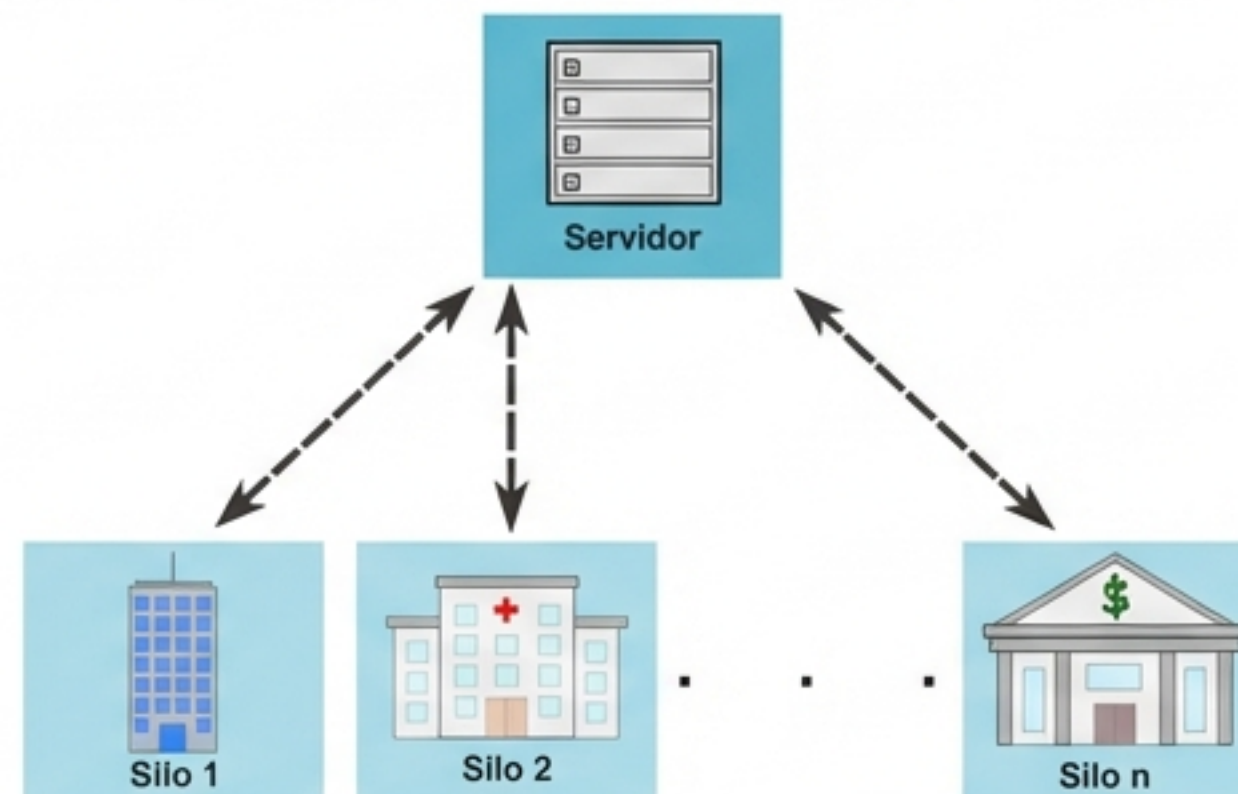
Escenarios de Arquitectura: *Cross-Device* vs. *Cross-Silo*

Cross-Device



- **Escala:** Masiva (10^{10} dispositivos)
- **Conexión:** Inestable (Wi-Fi/4G)
- **Clientes:** Cambian en cada ronda
- **Ejemplo:** Texto predictivo (Gboard)

Cross-Silo (Enfoque de esta tesis)



- **Escala:** Pequeña (\$2-100\$ organizaciones)
- **Conexión:** Estable y confiable
- **Clientes:** Organizaciones (Hospitales, Bancos, Escuelas)
- **Disponibilidad:** Alta

Particionamiento de Datos: Horizontal (HFL) vs. Vertical (VFL)

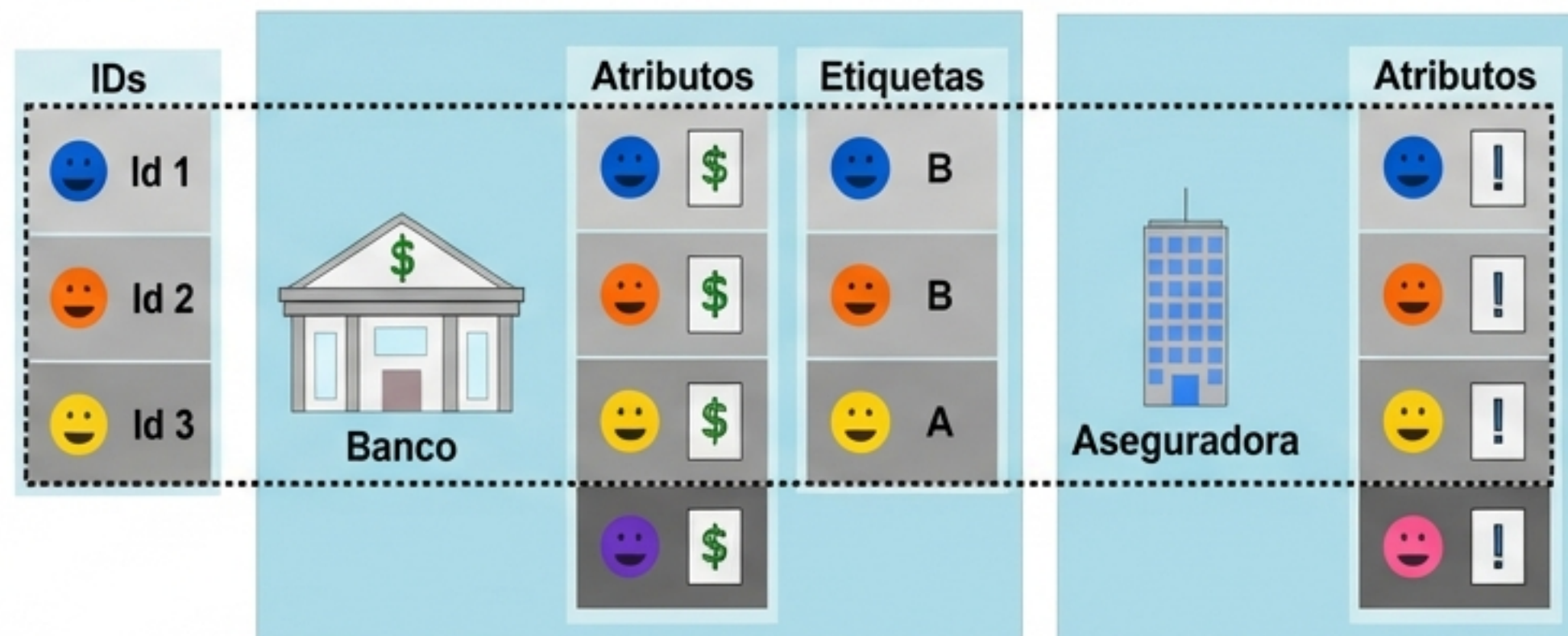
Horizontal Federated Learning (HFL)

Mismas características (columnas), diferentes usuarios (filas).



Vertical Federated Learning (VFL)

Mismos usuarios (filas), diferentes características (columnas).



Estudio de Factibilidad: Pruebas en el ‘Laboratorio’

Validación del rendimiento frente al ML centralizado tradicional.

Setup

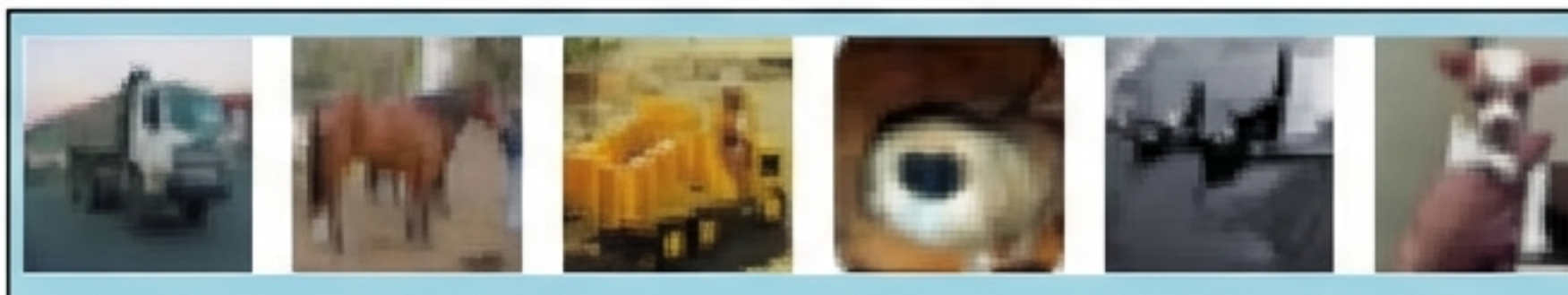
- **Herramientas:** TensorFlow Federated (TFF) y Flower.
- **Objetivo:** Comparar Accuracy y Tiempos de Convergencia.
- **Línea Base:** Modelo Centralizado vs. Federado.

Datasets

MNIST (Dígitos)



CIFAR-10 (Objetos)



Resultados de Factibilidad:

El Costo de la Privacidad no es la Precisión

Precisión MNIST

~99%

El modelo federado
igualó la precisión
del modelo
centralizado.

Precisión CIFAR-10

Comparable

Resultados
competitivos, pero
requieren mayor
ajuste fino de
hiperparámetros.

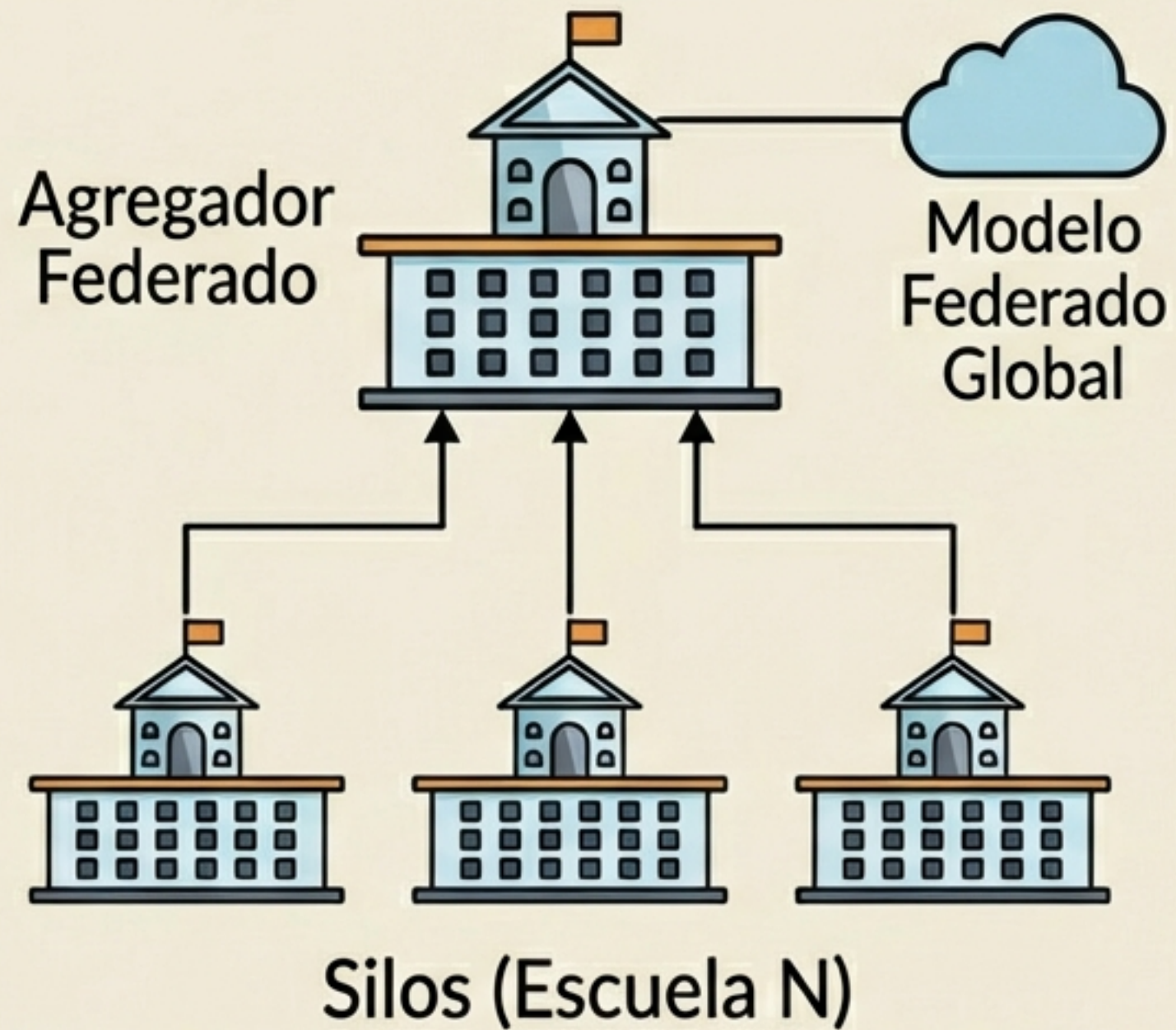
La Clave de Convergencia

**Clientes por
Ronda**

Aumentar la fracción
de clientes (C) en cada
ronda acelera la
convergencia y
mejora la estabilidad.

Caso de Estudio Real:

Analíticas de Aprendizaje (Cross-Silo HFL)



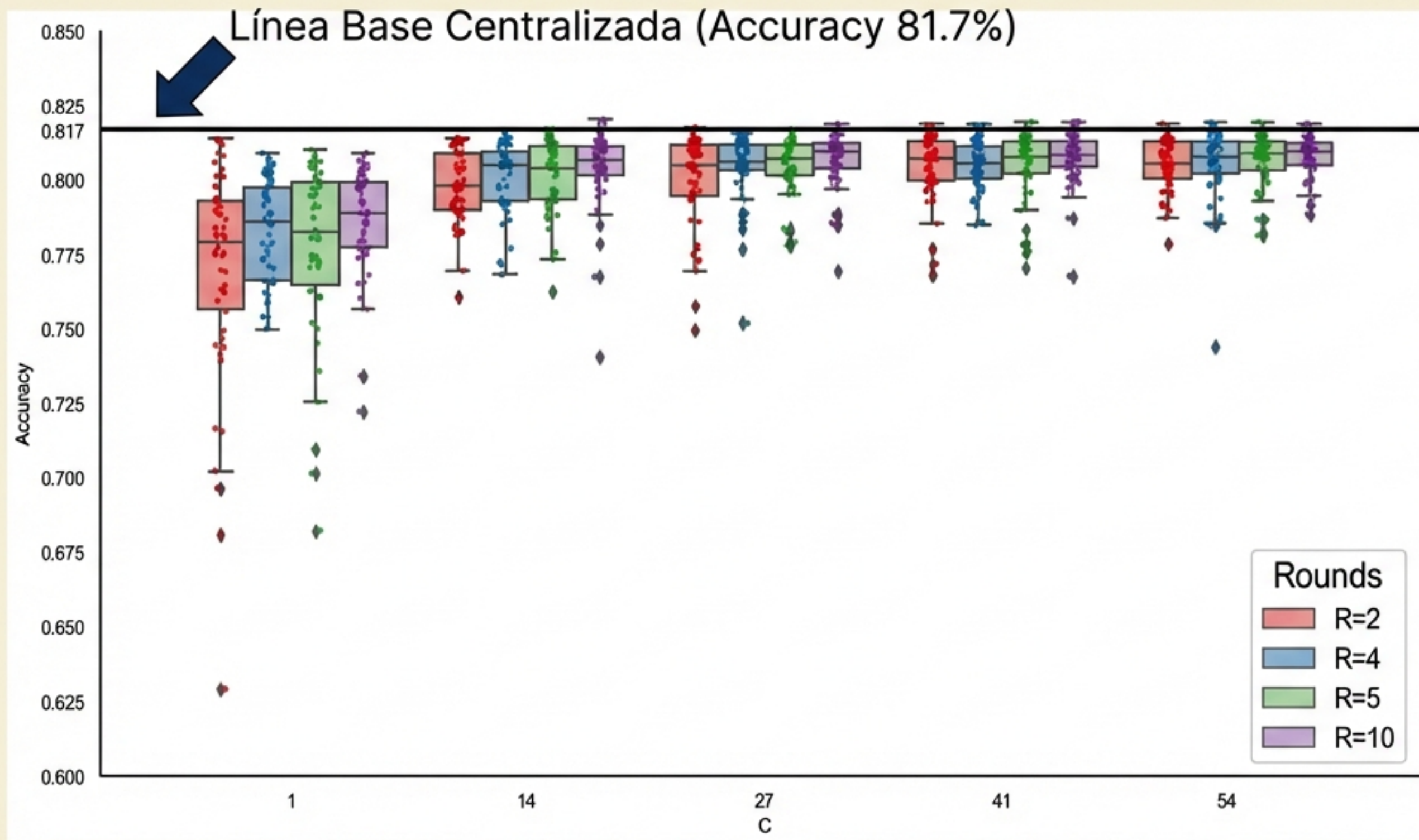
Red Educativa Distribuida

- **Escenario:** 77 Escuelas (Silos).
- **Datos:** KDDCup2015 (XuetangX MOOC). 77,000 estudiantes.
- **Objetivo:** Predecir abandono y agrupar estudiantes sin compartir PII.

Tareas

1. Predicción de Abandono (Supervised)
2. Agrupación de Estudiantes (Unsupervised K-Means)

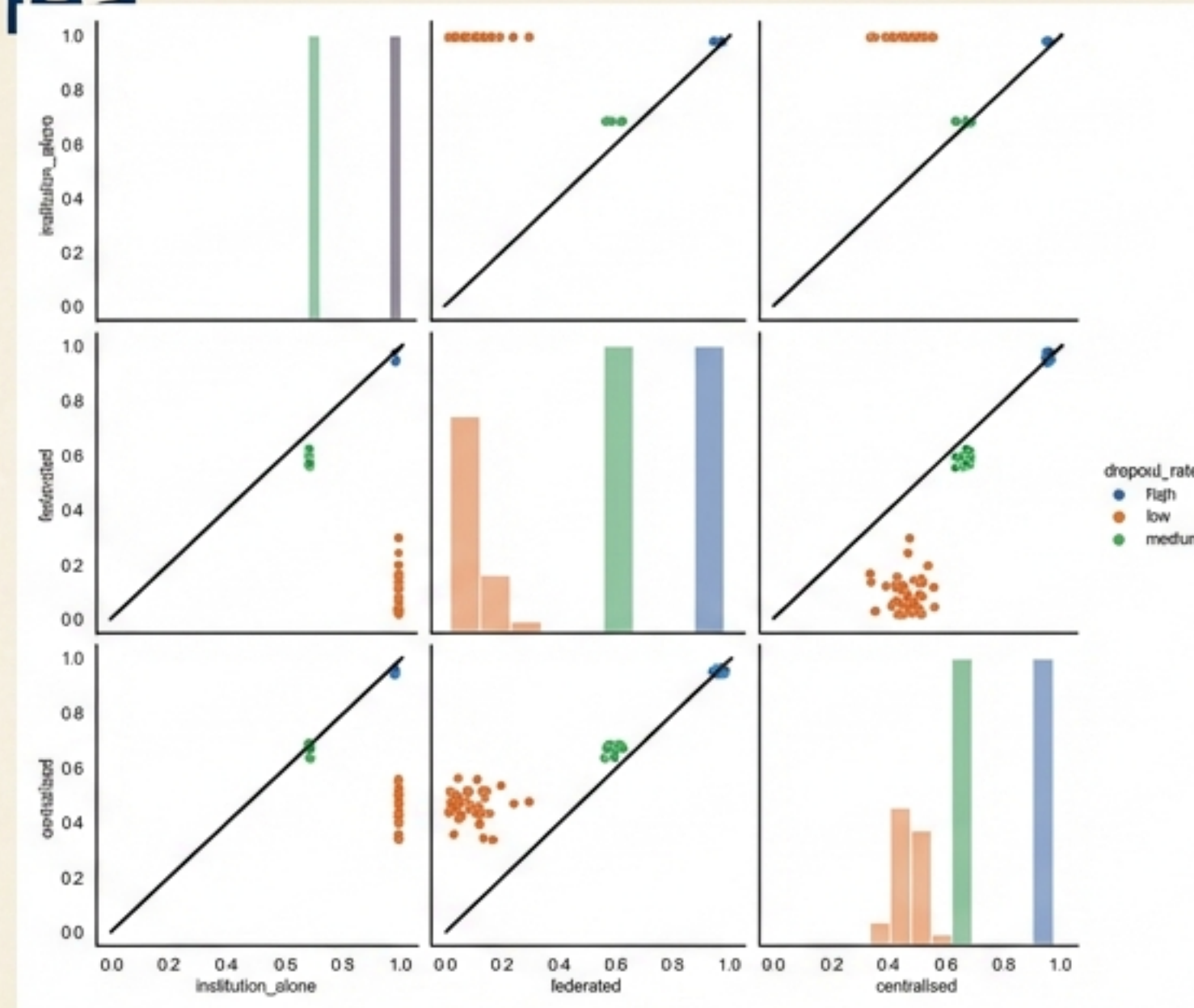
Tarea 1: Predicción de Abandono Escolar



Resultados:

- FL alcanzó la precisión objetivo (>81%).
- **Costo:** Mayor número de rondas de comunicación.
- **Estrategia:** Más rondas globales (R) son preferibles a más épocas locales (E).

El Impacto de la Distribución de Datos: Homogénea vs. Heterogénea



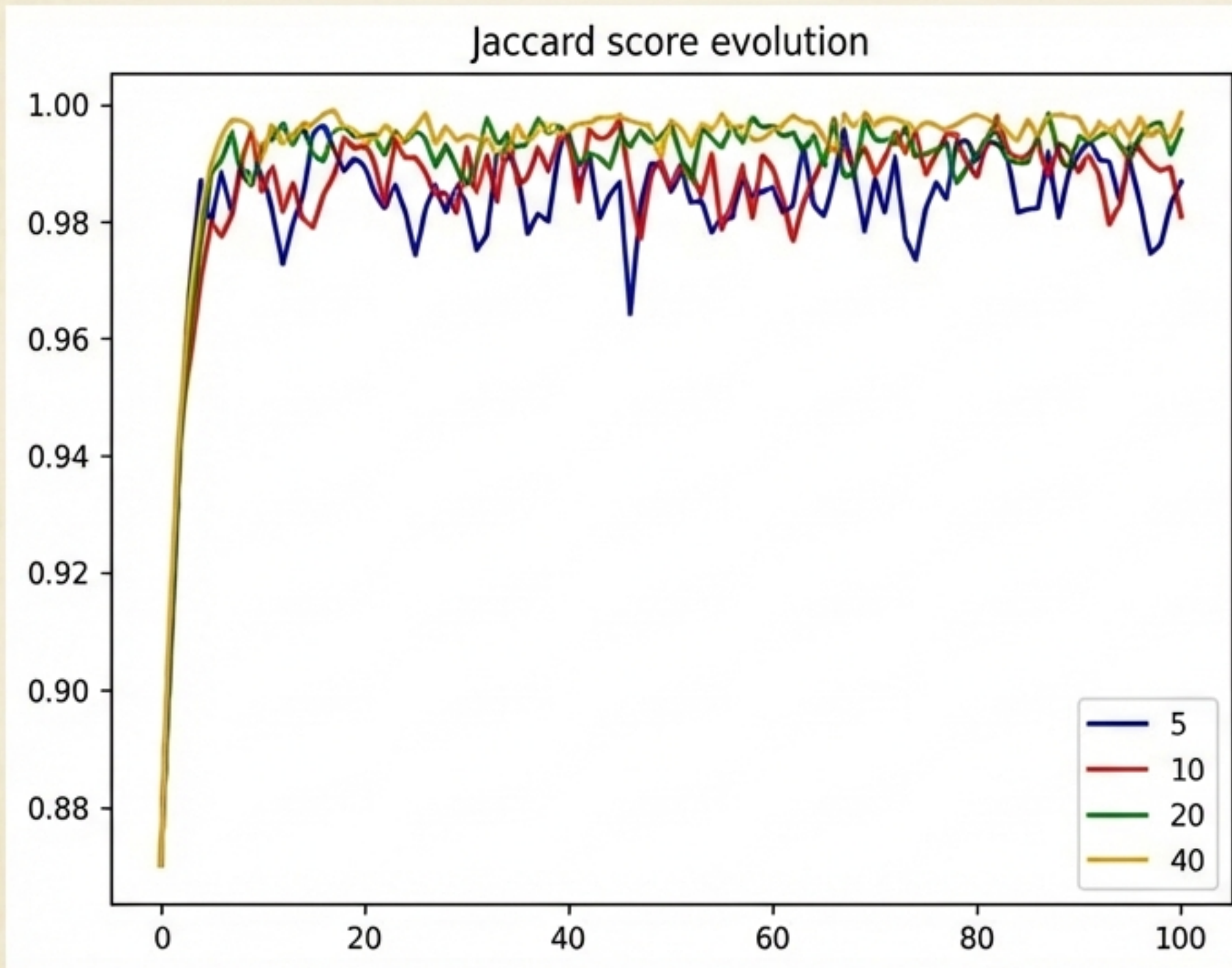
Distribución Homogénea

FL funciona excelente. Todas las escuelas se benefician.

Distribución Heterogénea (Sesgada)

Escuelas con datos 'difíciles' (baja tasa de abandono) no mejoran significativamente, incluso con FL.

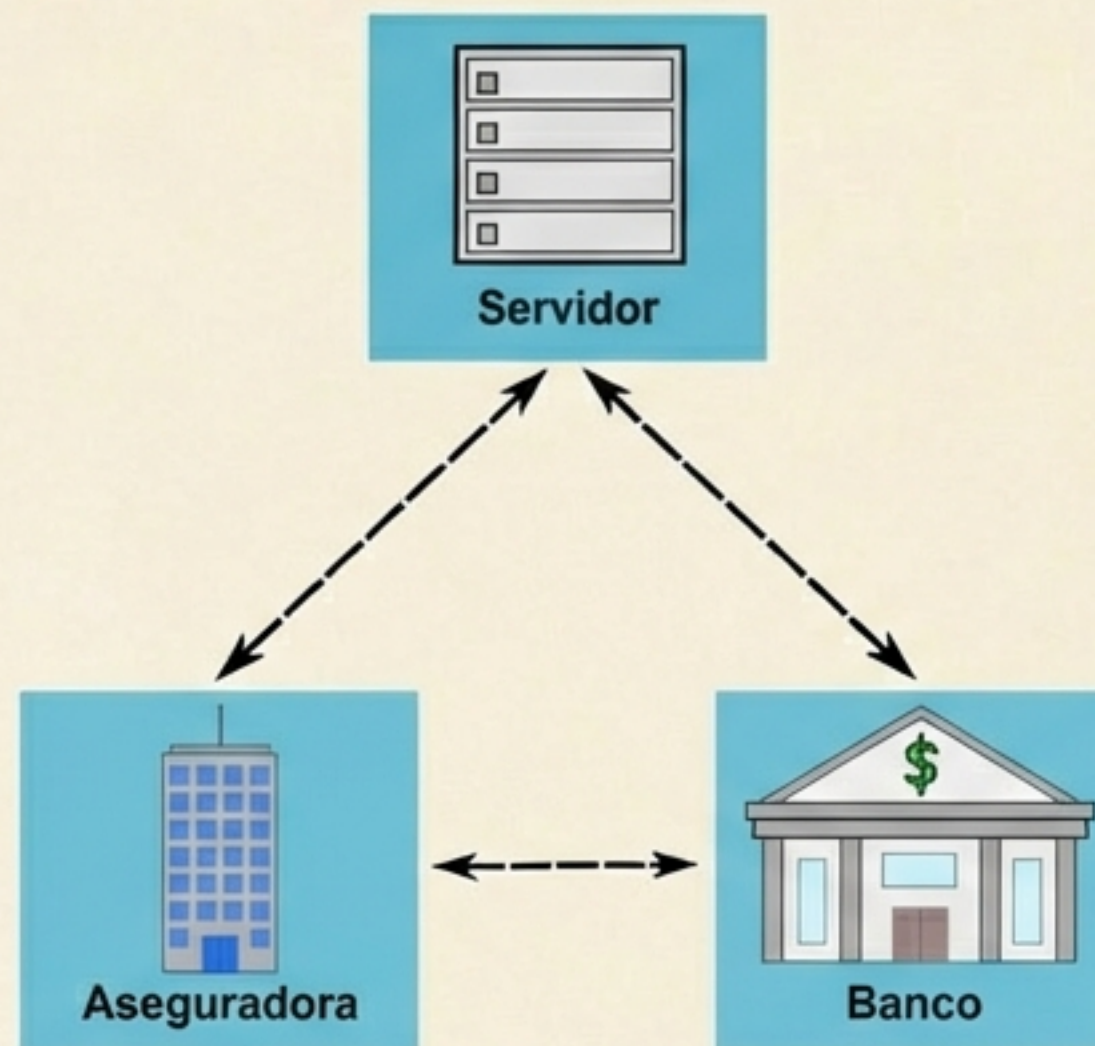
Tarea 2: Clasificación No Supervisada (Federated K-Means)



Federated K-Means

- **Métrica:** Similitud de Jaccard (vs. Clusters Centralizados).
- **Resultado:** >0.96 de similitud tras solo 10 rondas.
- **Conclusión:** Es posible agrupar perfiles de estudiantes de manera privada con resultados casi idénticos al análisis centralizado.

La Frontera Compleja: Aprendizaje Federado Vertical (VFL)



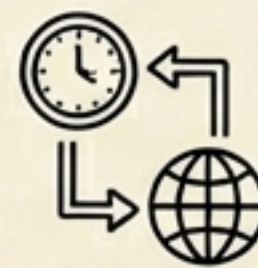
Alineación de Usuarios

Requiere Private Set Intersection (PSI) para cruzar IDs encriptados.



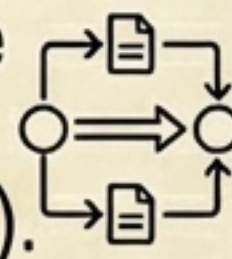
Sincronización Total

Todos los clientes deben estar online simultáneamente.

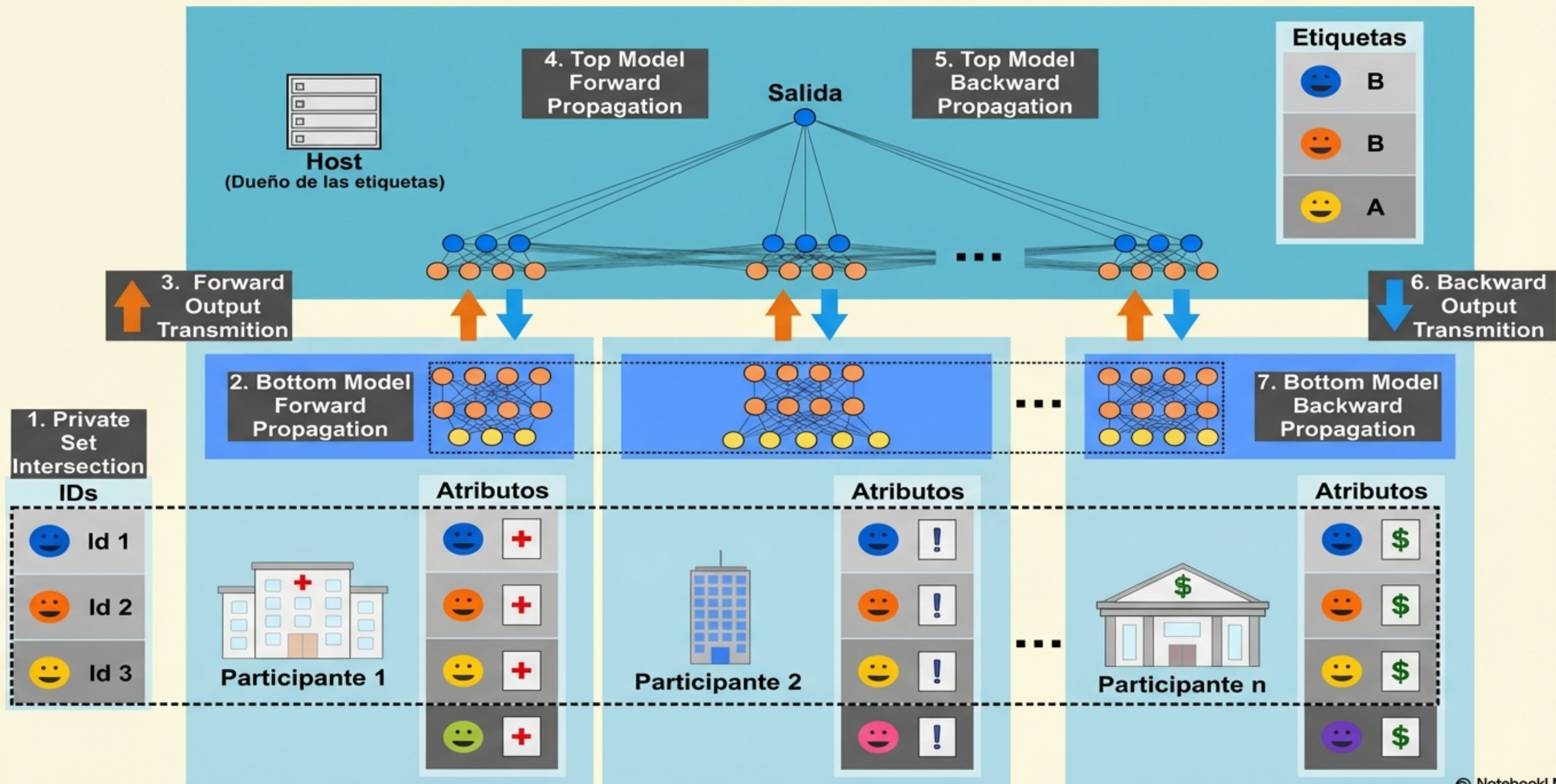


Flujo de Datos

Intercambio constante de resultados intermedios (Forward/Backward Prop).



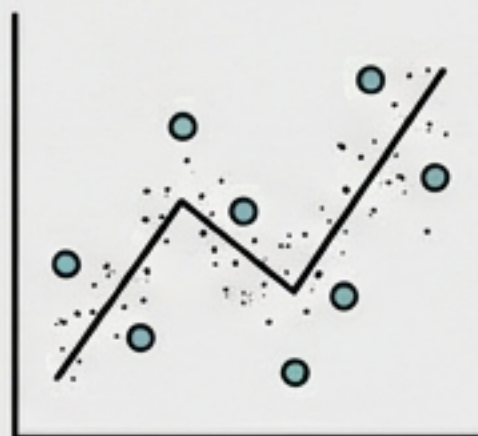
Arquitectura VFL: Entrenando un 'Cerebro Dividido'



Desafíos de Seguridad: Privacidad más allá de la Descentralización

Mitigando la fuga de información a través de los gradientes.

Privacidad Diferencial (DP)



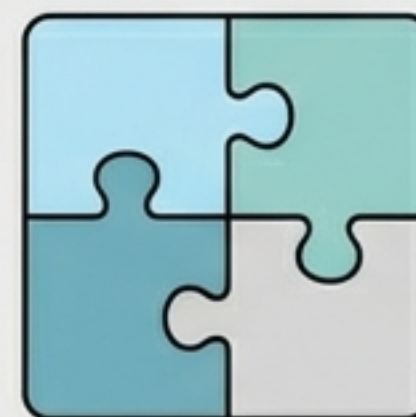
Añadir “ruido” matemático calibrado a los parámetros para ocultar contribuciones individuales.

Encriptación Homomórfica (HE)



Realizar cálculos matemáticos (sumas/promedios) directamente sobre datos encriptados.

Secure Multi-Party Computation (MPC)



Cómputo conjunto donde ninguna parte ve los inputs de las otras.

Conclusiones: Hacia una Inteligencia Colaborativa y Privada

- ✓ **Factibilidad Técnica:** FL iguala la precisión del ML tradicional (MNIST, KDD).
- ✓ **Aplicación Real:** Ideal para Salud, Finanzas y Educación (Cross-Silo).
- ⚠ **Desafíos Operativos:** La latencia de red y disponibilidad son los verdaderos cuellos de botella.

El Aprendizaje Federado habilita una nueva era de análisis de datos donde la privacidad es la base, no una barrera.