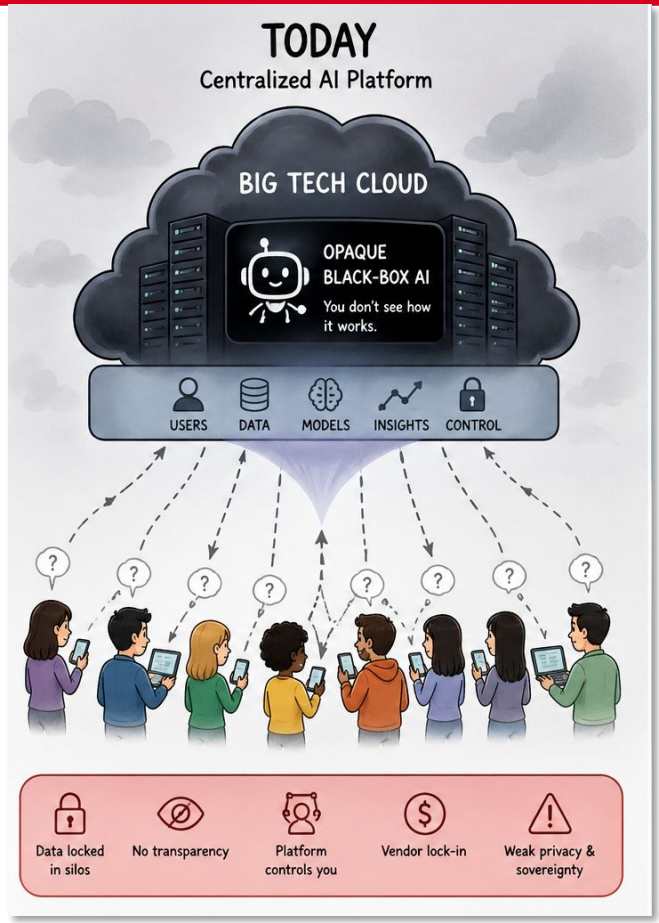
RPI

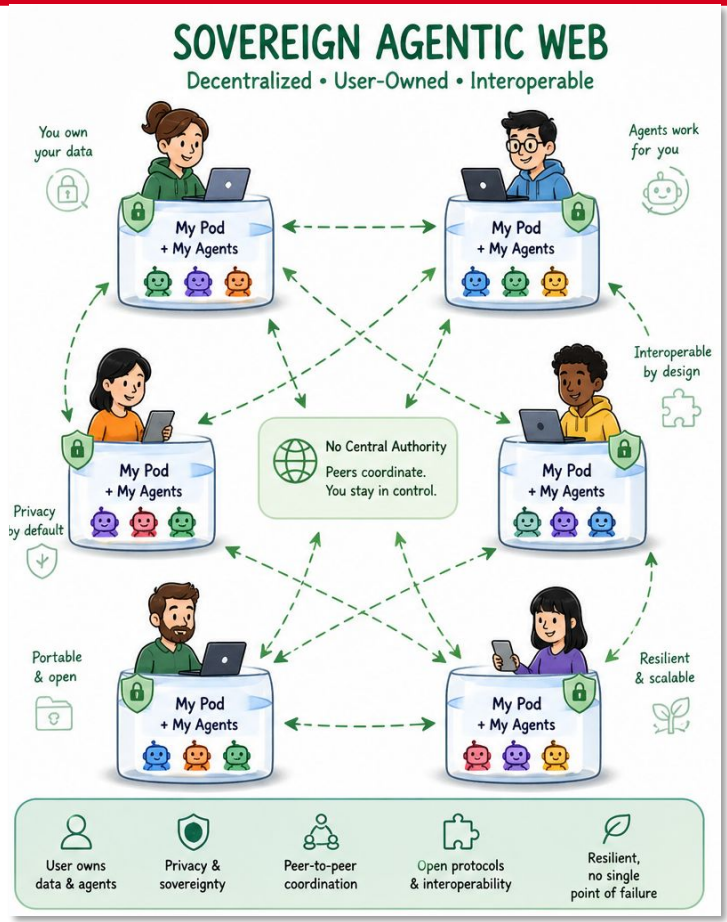
Toward Trustworthy Personal AI: A Sovereign Agentic Web Architecture Using Solid, Federated Learning, and Personal Knowledge Graphs

Fernando Spadea, Lorenzo Carta, Abhirup Dasgupta, Md Saikat Islam Khan Bappy, and Oshani Seneviratne

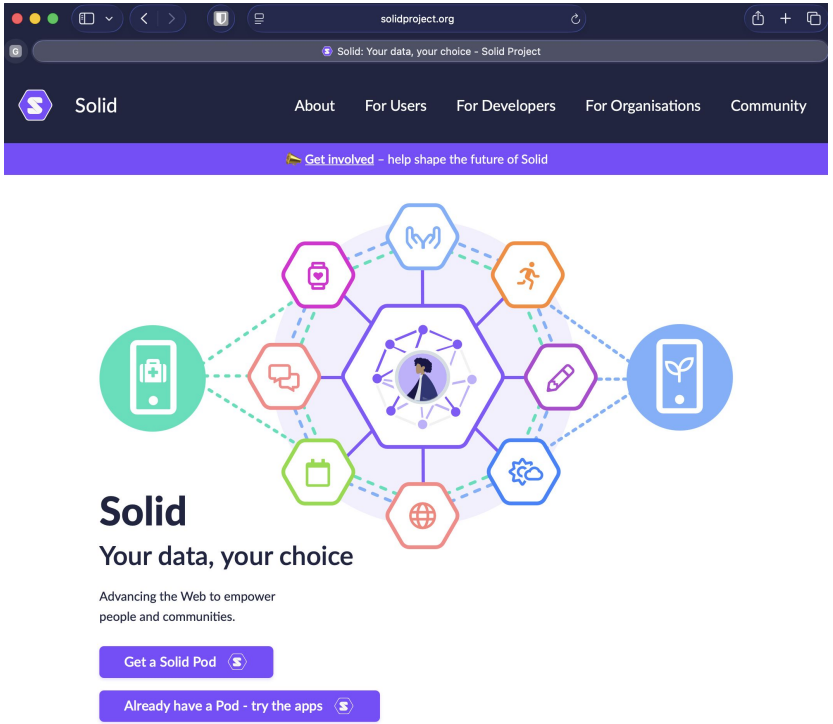


Today's AI assistants are mostly hosted in corporate clouds.

What kind of protocol adaptations are needed to make our AI assistants **sovereign Web citizens**?



Solid as a Personal Data Store



https://solidproject.org

Value Proposition for Solid

Solid already provides:

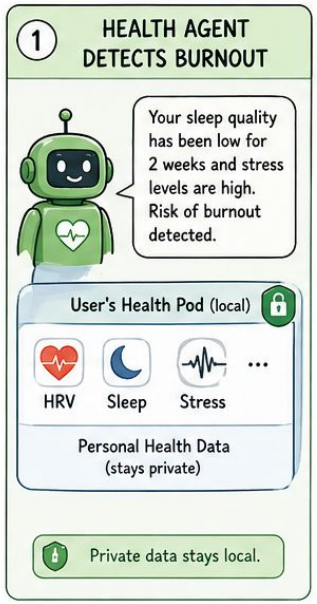
- Decentralized Identity
- Web-native Access Control
- Linked Data
- Interoperable Storage

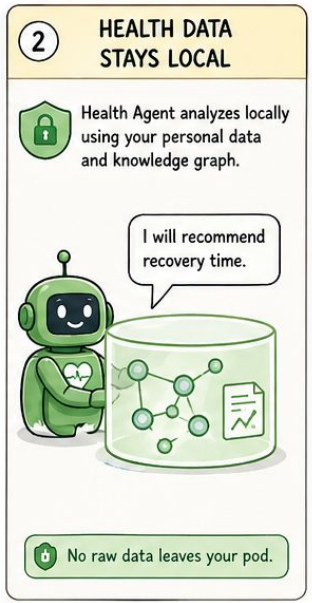
But Solid pods are still treated as **passive data stores**.

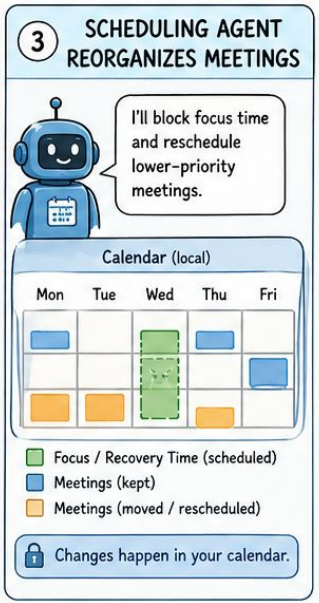
We argue that Solid could facilitate an ecosystem where:

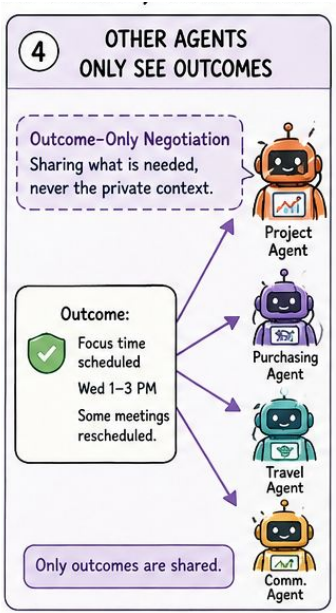
- Pods host:
  - Semantic memory/ Context graphs
  - Workflows
  - Policy mediation
- Agents as Web Principals with:
  - WebIDs
  - Revocable Capabilities
  - Auditability
- Federation for Coordination

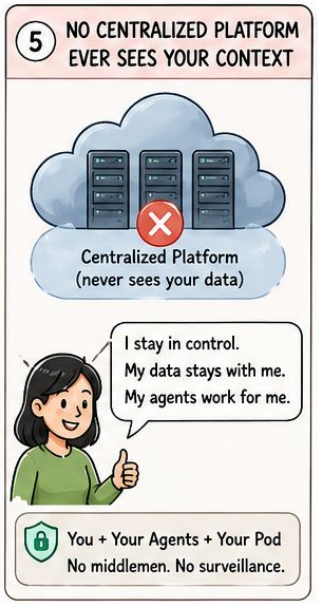
Motivating Use Case





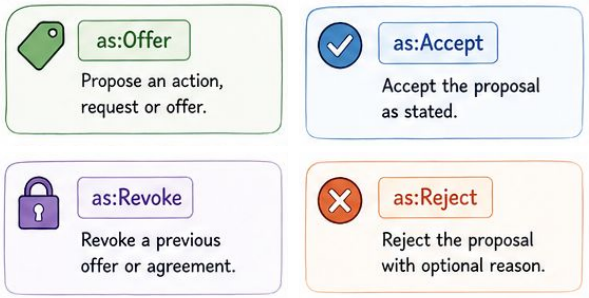


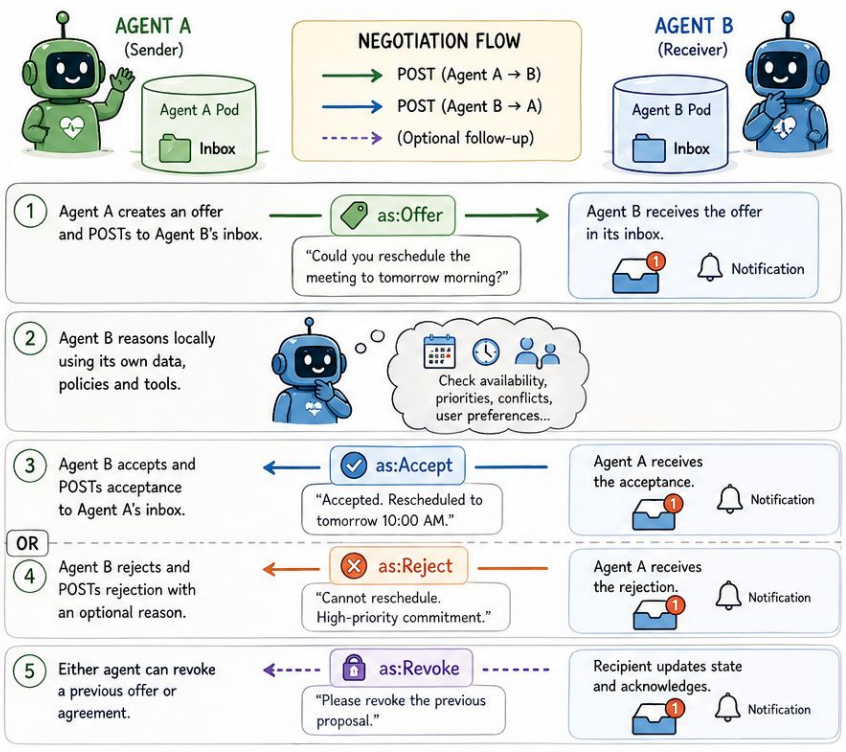




Extending Solid Notifications for Negotiation

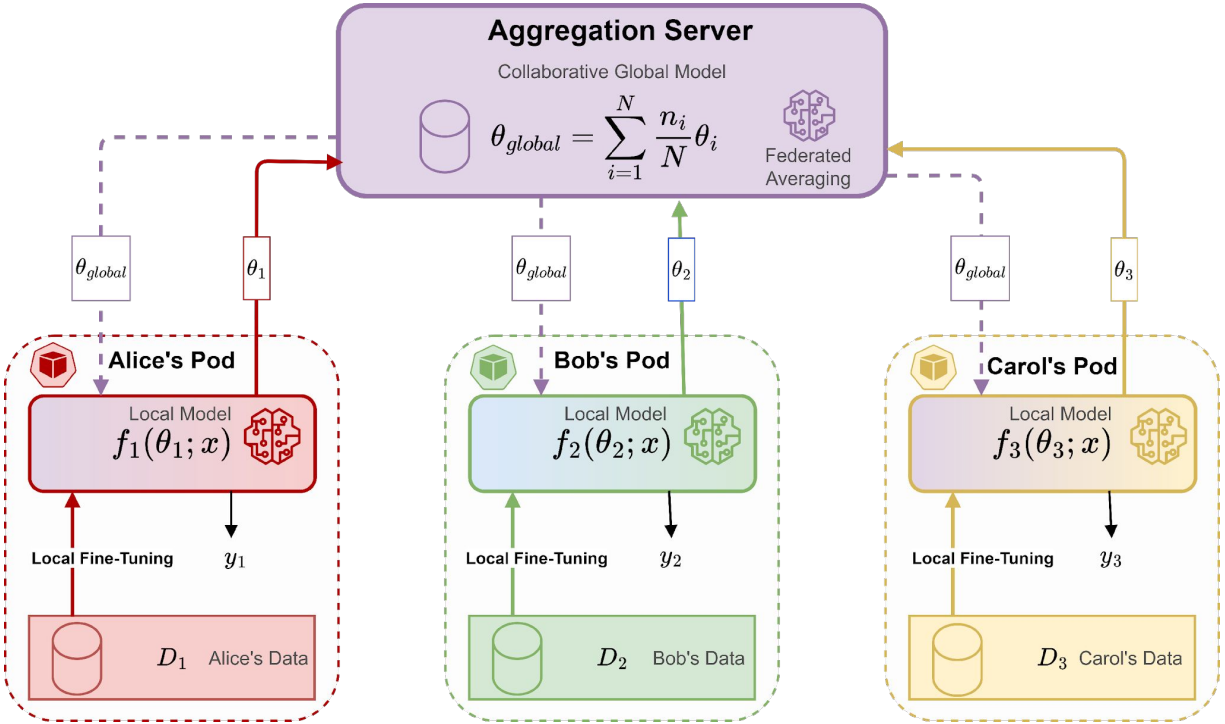
- Existing Solid notifications:
  - deliver events.
- But agents need:
  - offers,
  - acceptance,
  - revocation,
  - negotiation,
  - asynchronous coordination.





The Need for Decentralized Learning

- Centralized AI erodes privacy and autonomy, and standard Federated Learning (FL) is insufficient for true data sovereignty.
- Currently, Solid pods are underutilized as merely passive storage.

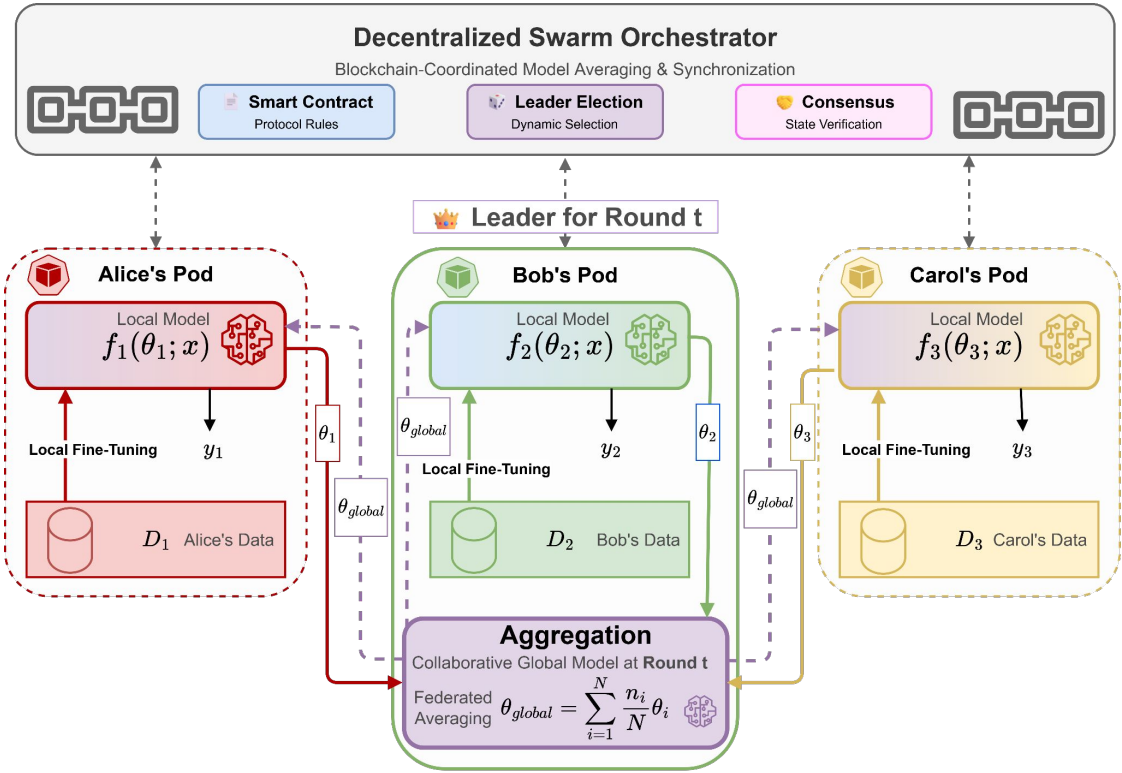


Fusing Solid and Decentralized Learning

- Several decentralized learning strategies have applications here:
  - Vertical FL
  - Horizontal FL
  - Swarm Learning
- In Swarm Learning, the centralized aggregation server used in traditional FL is removed.
- Instead, smart contracts are utilized to coordinate the decentralized swarm.
- Data sovereignty is maintained with strict user-control through selective data sharing using access controll lists and Differential Privacy (DP) budgets.

Decentralized Swarm Architecture

- Rather than relying on a central aggregator, smart contracts coordinate leader election and model synchronization.

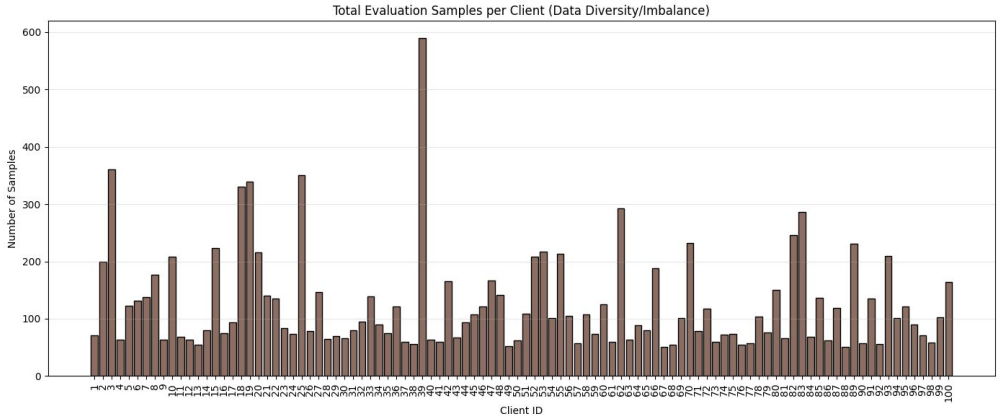


Example: Binary Movie Recommendation System

- Application Demo: Using the MovieLens 32M dataset.
- The environment simulates 100 clients (filtered users with >250 reviews) with all data stored in their respective Solid pods.
- Data is **stratified** into:
  - Safe data
  - Sensitive data
- Users can:
  - Toggle participation in the swarm learning process through the smart contract
  - Apply **differential privacy** selectively

Experimental Setup & Benchmarking

- To simplify the classification task, 0.5-5.0 star movie reviews were converted to binary ratings (≥4 positive, <4 negative).
- Random data partition:
  - 11x gap between smallest and largest client
  - Demonstrates “non-IID”-nature



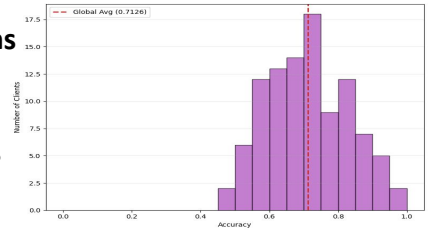
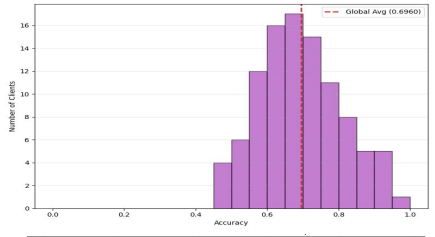
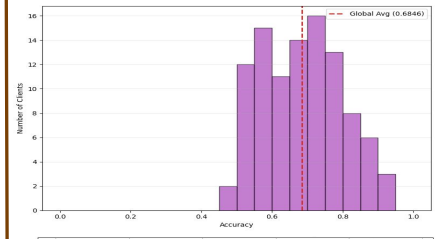
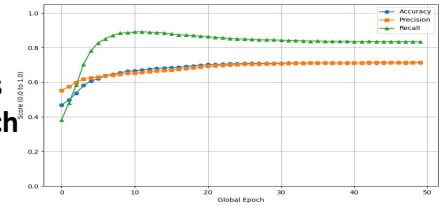
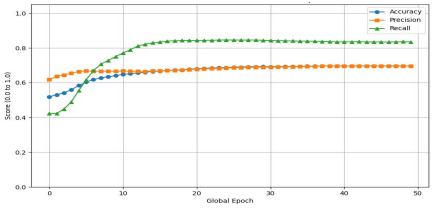
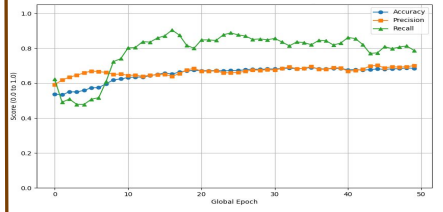
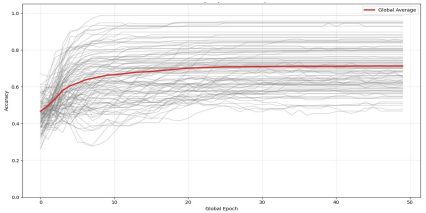
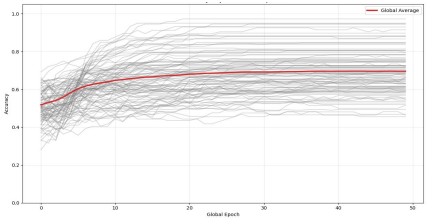
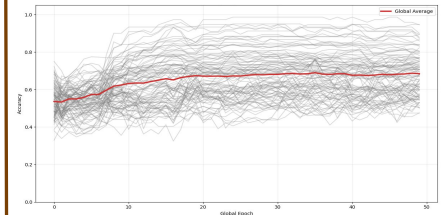
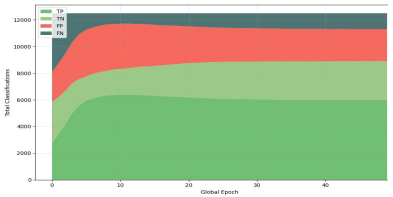
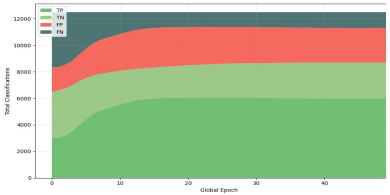
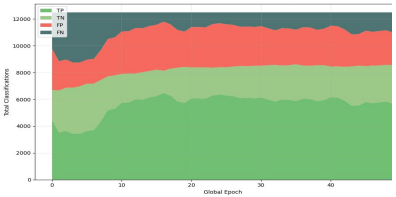
Differential Privacy Settings in Our Swarm Learning Scenario

We evaluated three scenarios to measure the impact of privacy constraints on model utility:

- **Raw:** no DP and entire dataset, used as performance upper bound.
- **Closed:** no DP with 50% of clients not sharing their sensitive sets.
- **DP-Heavy:** 50% of clients not sharing their sensitive sets and with DP-noise added:
  - 17.5% of clients utilized a budget of $\epsilon = [0.1, 0.5]$.
  - 32.5% of clients utilized a budget of $\epsilon = [0.5, 2.0]$.

Results

- Stable training despite data imbalance
- Convergence maintained even with DP noise + restricted data.
- Strong privacy with minimal performance loss with 71.3% (Raw) vs 68.5% (DP-Heavy).

	Results: Raw	Results: Closed	Results: With DP
Distributions of Final Client Accuracies			
Global Metrics Per-Epoch			
All Client Accuracy Trajectory Per-Epoch			
Global Confusion Matrix Distribution			

Questions?

Email us with questions at spadef@rpi.edu or senevo@rpi.edu