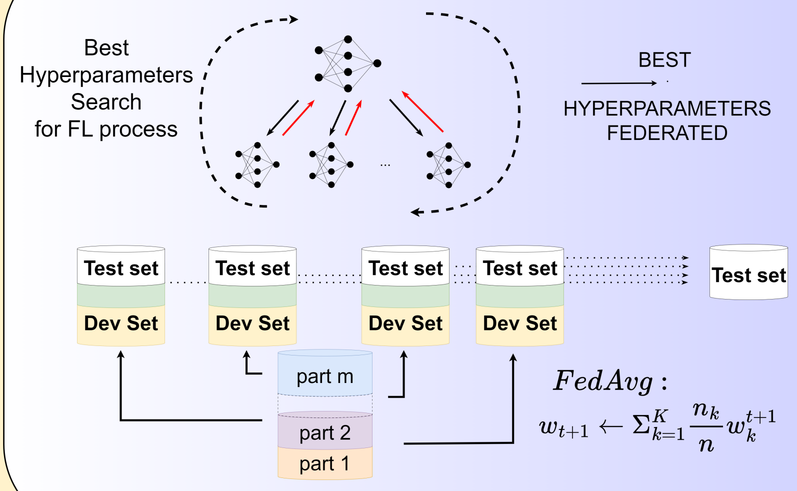
Explainable Federated Learning with Logic Explained Networks

Model Aggregation vs Post hoc Rules Aggregation Strategies in Federatd Learning

Valerio Bonsignori
valerio.bonsignori@phd.unipi.it



The experimental framework:
The **dataset is partitioned** using Latent Dirichlet Allocation (LDA) with different concentration parameters **α** to simulate real-world data distributions (low **α** values (≈0.5) highly skewed distribution, high **α** values (≈100.0) nearly identical distribution)

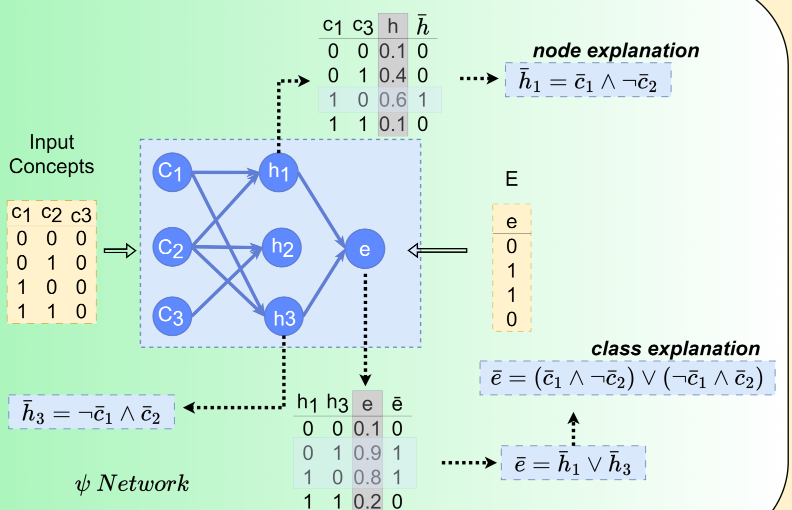
The **partitions are distributed** across clients: **cross-silo** (few stable clients, large data portions), **cross-device** (many unstable clients, small data portions) settings.

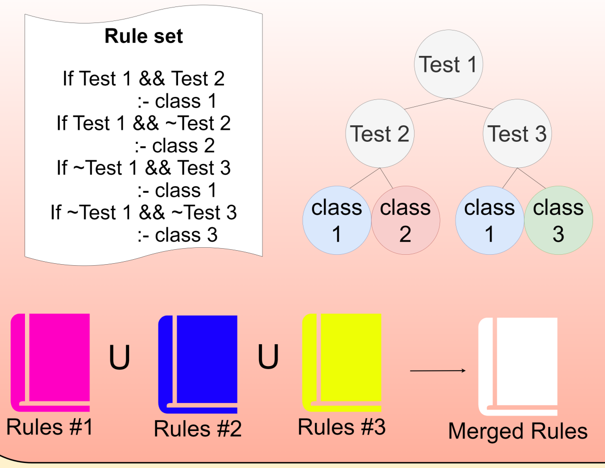
In **Cross-Silo**, each client maintains development and test sets for local validation, a global test set enables consistent evaluation.
In **Cross-Device** a subset of clients is used for the training phase, another subset is dedicated to the validation, and another set is used as test set to measure the overall performance
The Hyperparameter optimization explores canonical parameters as well with federated ones (rounds, epochs per round)

This methodology ensures the selection of the **best performing federated** model while guaranteeing its **meaningful interpretability**.

Logic Explained Networks (LENs) blend of neural network architecture and interpretable rule extraction.
Their neural structure, **based on [0,1]-valued activation functions** and structured pruning strategies, allows careful **integration** with standard federation techniques like **FedAVG**, while simultaneously **ensuring interpretability** through boolean logic rules.
Each neuron is **constrained to maintain limited incoming connections** (*i.e. 2-9*) through **L1-regularization and prunings**, enabling the extraction of comprehensible logic formulas.
The architecture serves dual purposes: it can function as a **standalone interpretable model** or as a **post-hoc explainer** for other black-box models, acting as regularizer.

In federated scenarios, this flexibility is crucial while rule aggregation across clients is challenging, LENs can be efficiently trained and aggregated as neural networks, maintaining both model performance and interpretability across the federated process.





Interpretable model aggregation in federated settings have distinct approaches. In a centralized scenario, **traditional rule merging** for explainability (like SAME or GLocalX) directly combines rules or decision trees through semantic alignment, handling conflicts in conditions, lengths, and support metrics to produce a unified ruleset or tree. When using **LENs** as neural networks in **FL**, the challenge shifts to **adapting FedAVG** (or similarly other algorithms) to maintain architectural constraints: **pruning patterns** must be preserved across aggregation rounds while managing **weight normalization** to ensure meaningful averaging of conceptual features across clients. The third approach, **extracting and merging rules** from federated LENs (whether used as post-hoc explainers or interpretable-by-design models), faces combined challenges: beyond standard rule alignment challenges, it must take into account divergent concepts used by different clients' LENs, making semantic matching of extracted rules particularly challenging. Each strategy offers **different trade-offs** between **federation complexity**, **interpretability preservation**, and **model performance**.

How can we effectively combine Logic Explained Networks (LENs) with Federated Learning (FL) while preserving both data sovereignty and model interpretability? What are the trade-offs between model-level federation versus rule-level aggregation approaches? How do these strategies impact the final model's interpretability and performance?

References
Federated Learning: McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 2017.
Kwatra, Saloni, and Vicenç Torra. "A k-anonymised federated learning framework with decision trees." International Workshop on Data Privacy Management. Cham: Springer International Publishing, 2021.
Rule Aggregations: Bonsignori, Valerio, Riccardo Guidotti, and Anna Monreale. "Deriving a single interpretable model by merging tree-based classifiers." Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings 24. Springer International Publishing, 2021.
Setzu, Mattia, et al. "Glocalx-from local to global explanations of black box ai models." Artificial Intelligence 294 (2021): 103457
Logic Explained Networks: Ciravegna, Gabriele, et al. "Logic explained networks." Artificial Intelligence 314 (2023): 103822.

IN SUPREMÆ DIGNITATIS
1343