Blockchain for Decentralized Learning

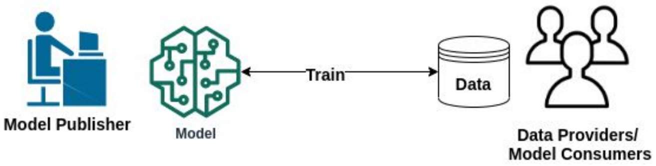
KTH-Insubria Secondment

Presented by Ahmed Emad

8th April 2021

Background

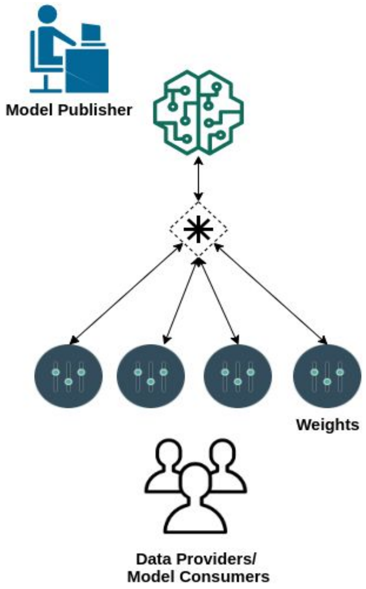
- Most AI services are centralized solutions.
    - Ex. Google news, Ads.
- Challenges to well-trained models.
    - Consumers:
        - Data privacy concerns.
        - No guarantees over provided services quality.
    - Owner/Publisher:
        - Infrastructure necessary for data analytics.
        - Unfair competitions: cold start concerns for newcomers.
        - Legal liabilities for storing and distributing personal data.



Centralized approach of learning

Alternative approaches: Federated Learning

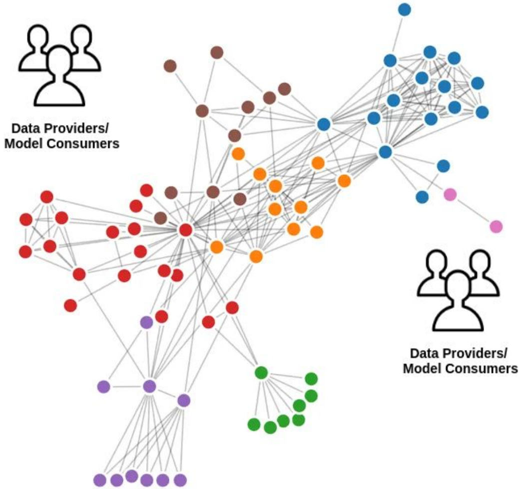
- Federated Learning: a distributed learning framework.
  - Local training updates.
  - Global model update (aggregation).
- Solved issues.
  - Distributed training load.
  - One step towards data-privacy.
  - No legalization required to process data.
- Remaining issues.
  - Central control of learning.
  - Unfair competitions for newcomers (model-owners).



Federated Learning Framework

Alternative approaches: Gossip Learning

- Gossip Learning: peer-to-peer learning.
  - Local training updates.
  - local model updates (aggregation).
- Solved issues.
  - Better distributed learning paradigm.
    - **Faster** learning.
    - One step towards data-privacy (local aggregations).
  - No legalization required to process data.
  - No cold start concerns.
- Remaining issues.
  - No control over learning/sharing
    - random swaps, fake updates ... etc.

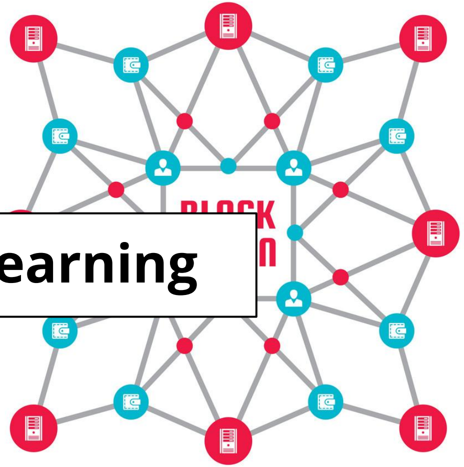


Gossip Learning Framework

Why Blockchain?

- Vulnerability in distributed learning techniques.
  - Federated learning: single point of failure.
  - Gossip learning: malicious attacks.
- Resilience and
  - Once transac

Blockchained Decentralized Learning



- Transparency and Accountability
  - Every action by a user in the blockchain is recorded and available publicly.
  - Each member can be accounted for its actions.
    - e.g., training of fake data, acting dishonestly, etc...
- Decentralized control of learning.
  - Smart contract: autonomous executable programs.
  - It can comprise code for business logic validation.
    - e.g., validating/verifying sanity of local training weights.
- Trust-less
  - Trust is not assumed among members of a blockchain.

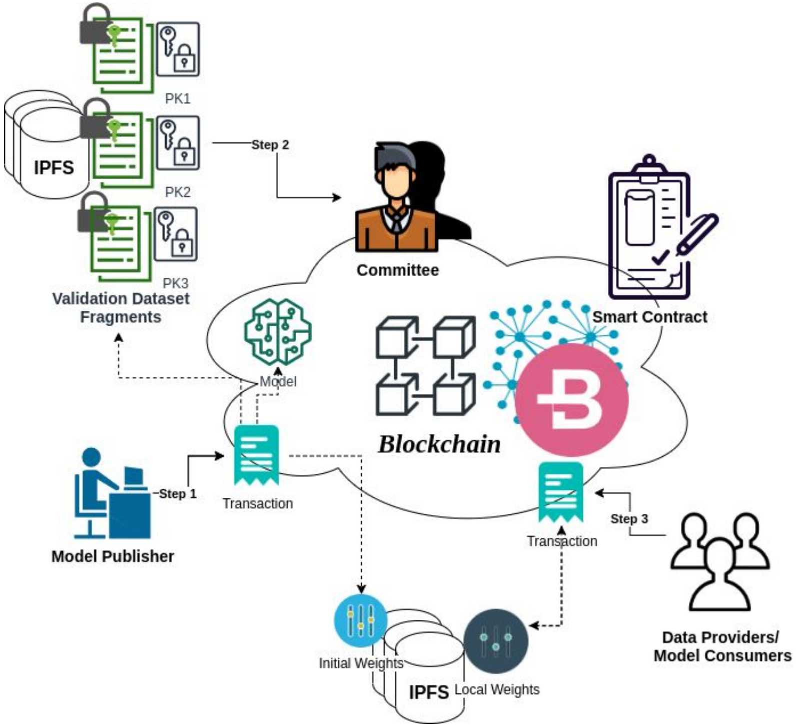
Beyond Centralized Learning

- Research questions:
  - How to maintain user data privacy?
  - How to provide fair chances to well-trained models?
    - Training capacity, traceability , resilience.

Blockchain for Decentralized Learning

Stakeholders:

- ❖ Model Publisher.
- ❖ Committee.
- ❖ Data Providers

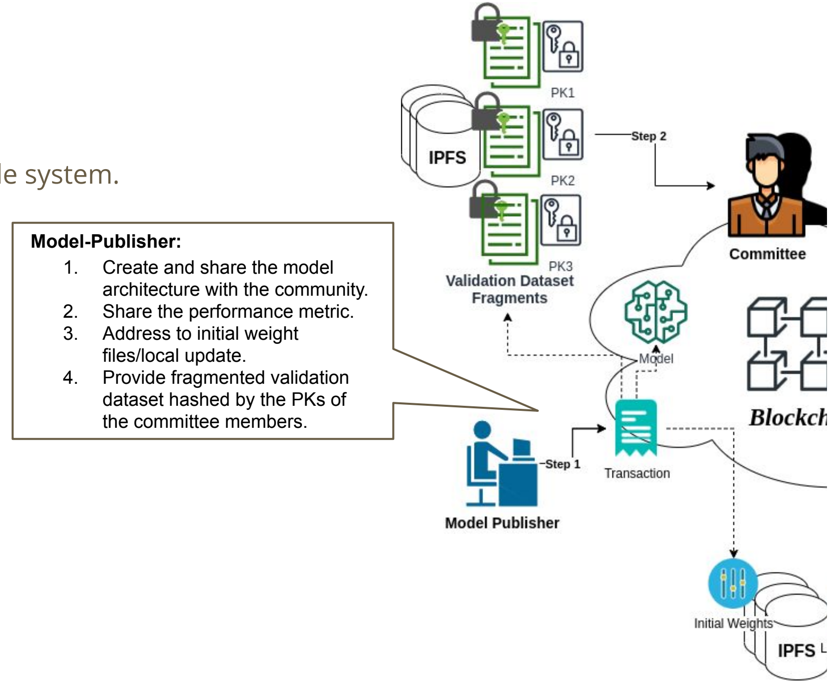


Learning paradigms:

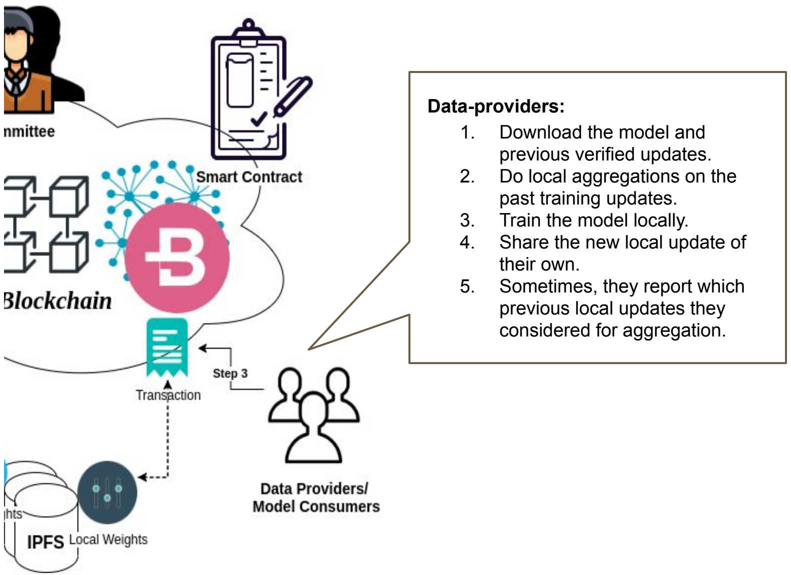
- ❖ Federated Learning
- ❖ Gossip Learning.

Stakeholder: Model Publisher

- Publish the model architecture to the community.
- Share the initial model's weights.
  - Random or pre-trained.
  - Address to the weight file in an external distributed file system.
- Declare the performance metric used.
  - Address to the code file.
  - Or an executable smart contract.
- Provide fragmented validation dataset.
  - Hosted in IPFS
  - Hashed by the PKs of the committee members.



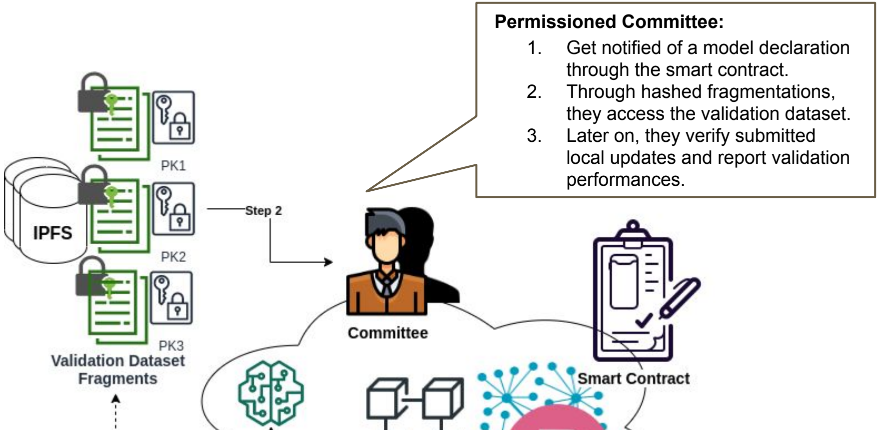
Stakeholder: Data Provider



- Download models they should find interesting.
- Choose/Aggregate previous **verified** weight(s).
  - smart contract routine.
- Train the model locally on their data.
  - Data providers are not obligated to participate in the training process.
- Share their local weights update.
  - Address to the weights file on IPFS.
  - The weight updates are now awaited to be verified by the committee.
- Report the addresses of the weight files they used in the aggregation.
  - Depending on the learning paradigm.

Stakeholder: Committee

- Gets notified of a new model to be published.
  - Smart contract.
- Check the **sanity** of the model to verify.
  - Ethical aspect, size of training, kind of data required .. etc.
- Each member access his equivalent fragment of the validation data.
- Cross-verify the submitted local **updates**.
  - Sanity of the updates (ex. Computational times).
  - Performance on the validation set.
- Report the validation performance of the updates.



Blockchain for Decentralized Learning

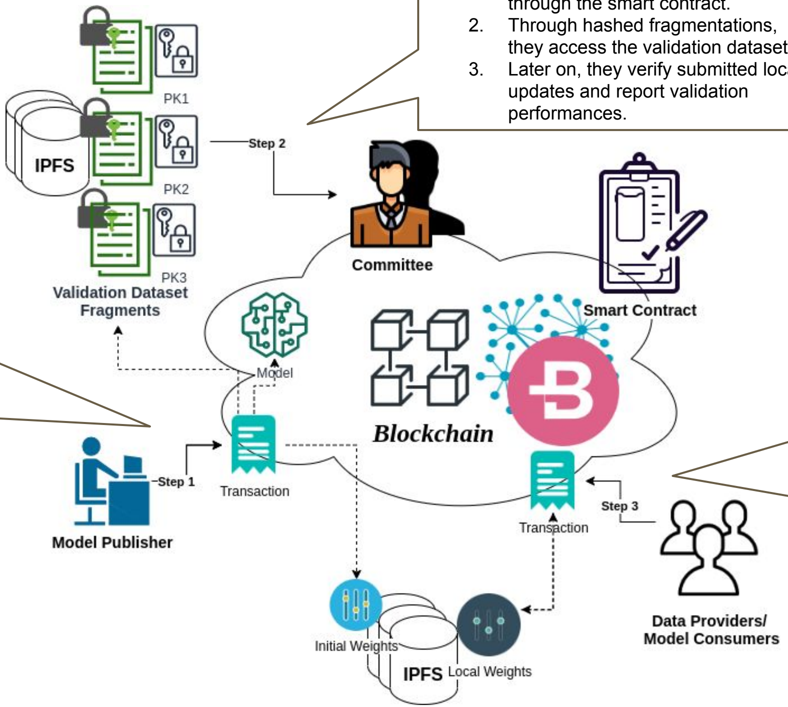
- Based on the Hyperledger Fabric as a permissioned blockchain framework.
- Verification is done through **smart contract.**
- Fragmented validation dataset, only accessed by the committee.
- Links to weights and the model architecture are shared.

Permissioned Committee:

1. Get notified of a model declaration through the smart contract.
2. Through hashed fragmentations, they access the validation dataset.
3. Later on, they verify submitted local updates and report validation performances.

Model-Publisher:

1. Create and share the model architecture with the community.
2. Share the performance metric.
3. Address to initial weight files/local update.
4. Provide fragmented validation dataset hashed by the PKs of the committee members.



Data-providers:

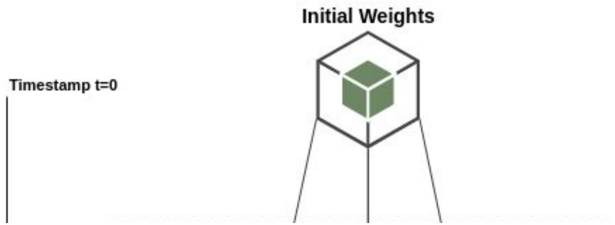
1. Download the model and previous verified updates.
2. Do local aggregations on the past training updates.
3. Train the model locally.
4. Share the new local update of their own.
5. Sometimes, they report which previous local updates they considered for aggregation.

Learning Frameworks

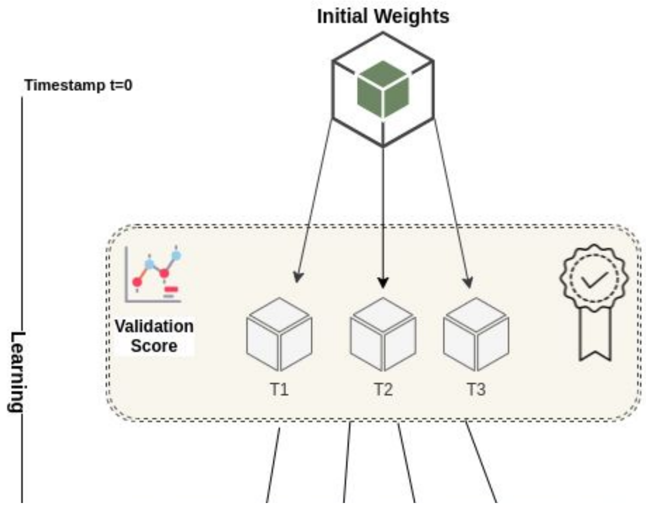
- Federation-inspired learning approach.
  - one global aggregation, done locally.

- Gossip-Inspired learning approach.
  - Multiple local aggregations.

Federation-inspired Learning



Federation-inspired Learning



<<Transaction>>

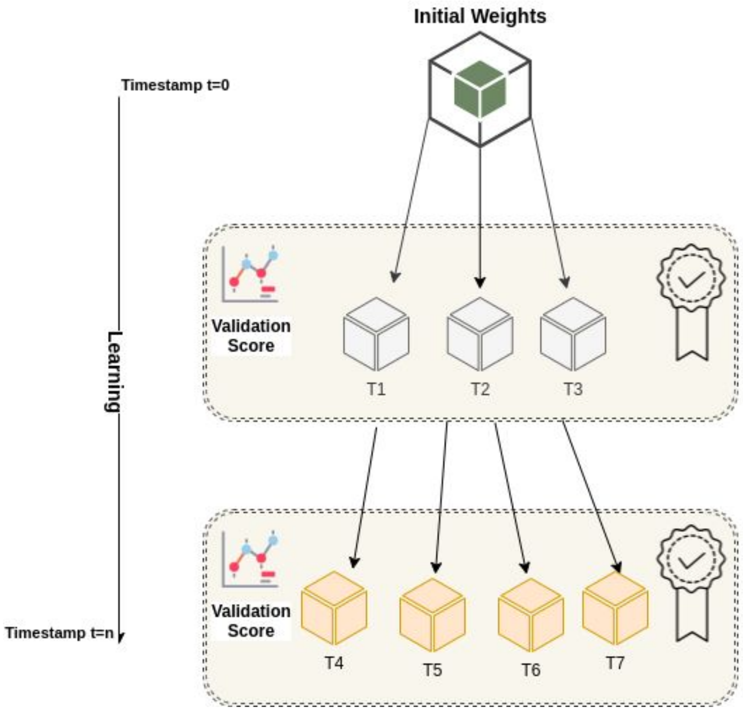
- Addresses to the new weight file/local update.
- Overall validation score.

Federation-inspired Learning

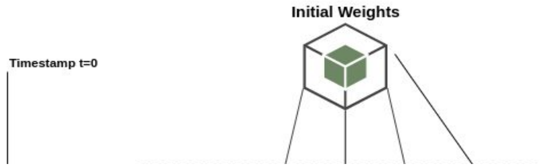
- Federated style of local aggregations
- Local updates are verified according to the rules of the smart contract.
- Entire history of forward training is encoded.

<<Transaction>>

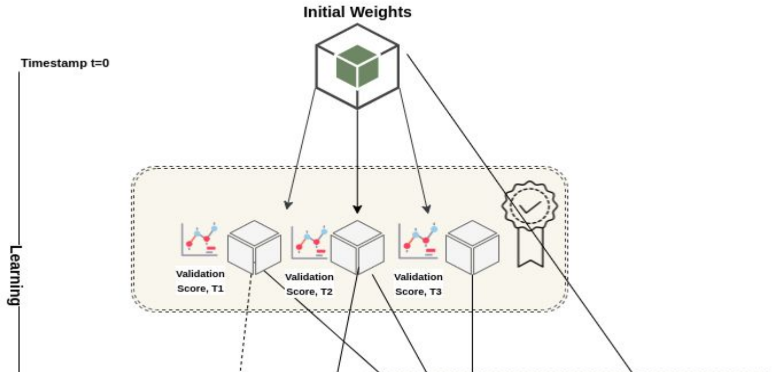
- Addresses to the new weight file/local update.
- Overall validation score.



Gossip-inspired Learning



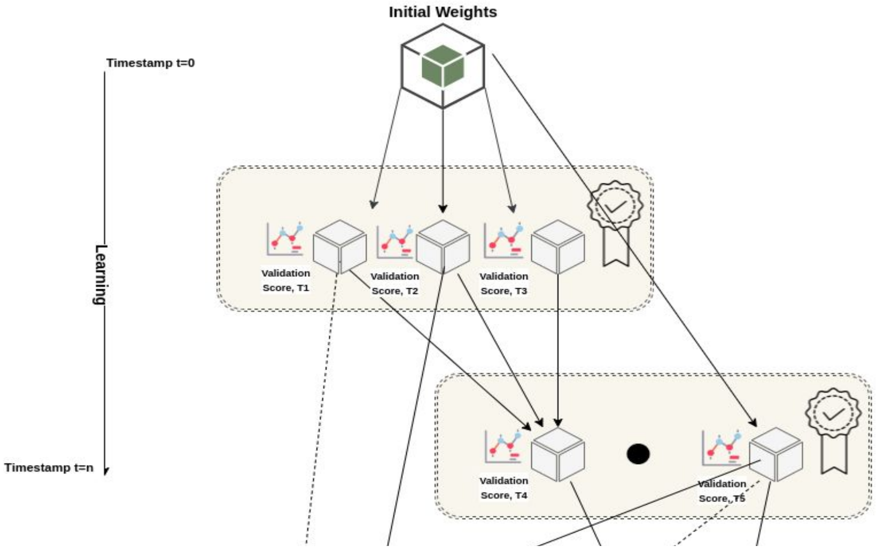
Gossip-inspired Learning



<<Transaction>>

- Address to new verified weight file.
- List of chosen updates.
- Validation score on the new update.

Gossip-inspired Learning



<<Transaction>>

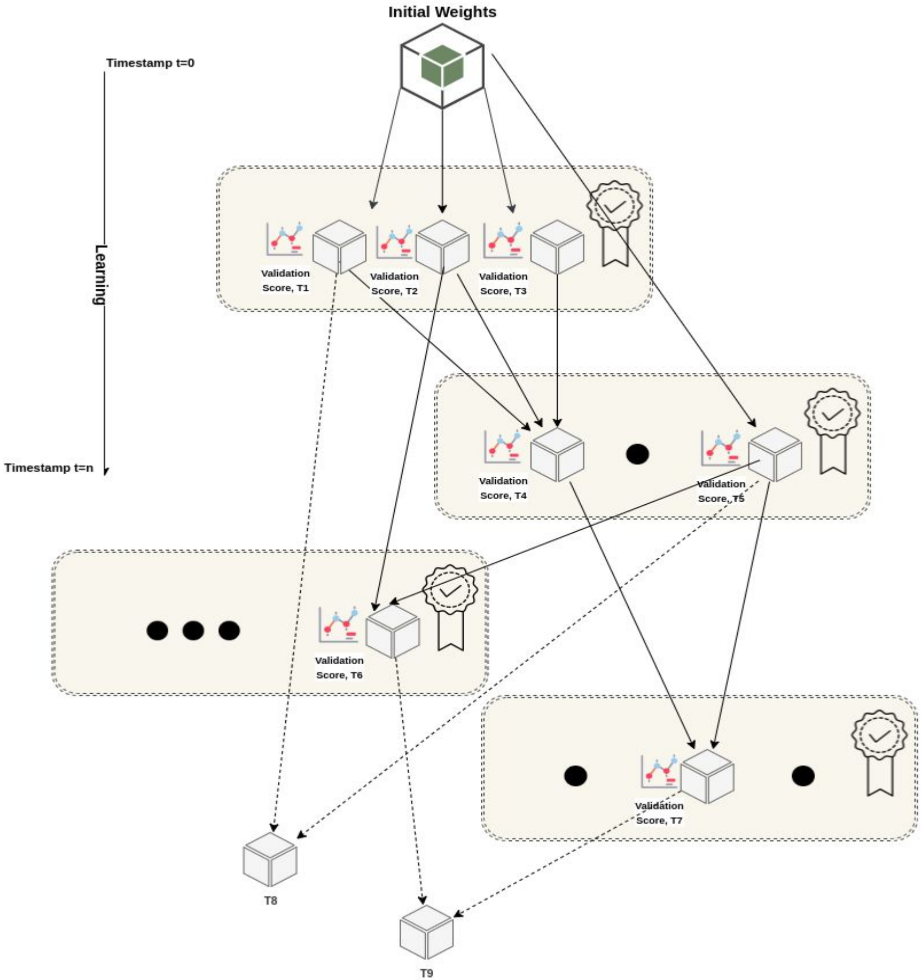
- Address to new verified weight file.
- List of chosen updates.
- Validation score on the new update.

Gossip-inspired Learning

- Gossip style of weights transfer and local aggregations.
- Directed acyclic graph
- Convergence is based on the principle of natural selection.

<<Transaction>>

- Address to new verified weight file.
- List of chosen updates.
- Validation score on the new update.



Discussion and Insights

Federation-based Learning:

- Decentralized learning i.e. aggregations are done locally.
- Synchronous approach of learning.
- Possibly slower but progressive training.
- Single thread of training- possibly more rigid.

Gossip-based Learning:

- Utterly decentralized learning
  i.e. aggregations are chosen and done locally.
- Asynchronous approach of learning.
- Possibly faster training.
- Highly flexible; model different training behaviors simultaneously.
- Contextualizing gossips seems more natural.

REWARDING

DATA PRIVACY

Possible gaps

- Less resilient to malicious behaviors and bad weight injections.
- Non i.i.d data points in local training and validation data.

Evaluation

- Quantitative analysis (centralized vs. proposed solutions)
  - Longer training time.
  - Comparable learning performance.
  - preserving data privacy and scaling the training to a greater number of collaborators.

Conclusion and Future Work

- Based on permissioned **Blockchain**, we aimed to provide a peer-to-peer environment for **decentralized** learning.

- The work was to realize one shared goal of having free **well-trained** AI services without sharing **private** data, and with training capabilities scaled up to a **community-level**.

Future Work

- How to handle with free riders?
  - Rewarding mechanisms (smart contract task).
  - Game theory based techniques.
- Personalized training (biasing/contextualizing gossip)
- Malicious behavior and the dissemination of false updates required.
- Current work relies on a fixed permissioned committee.
- On the security aspect; for example, weights can be reverse-engineered to produce the data.
- Choosing hyperparameters and handling non-iid data weren't accounted for in this research.

Q&A