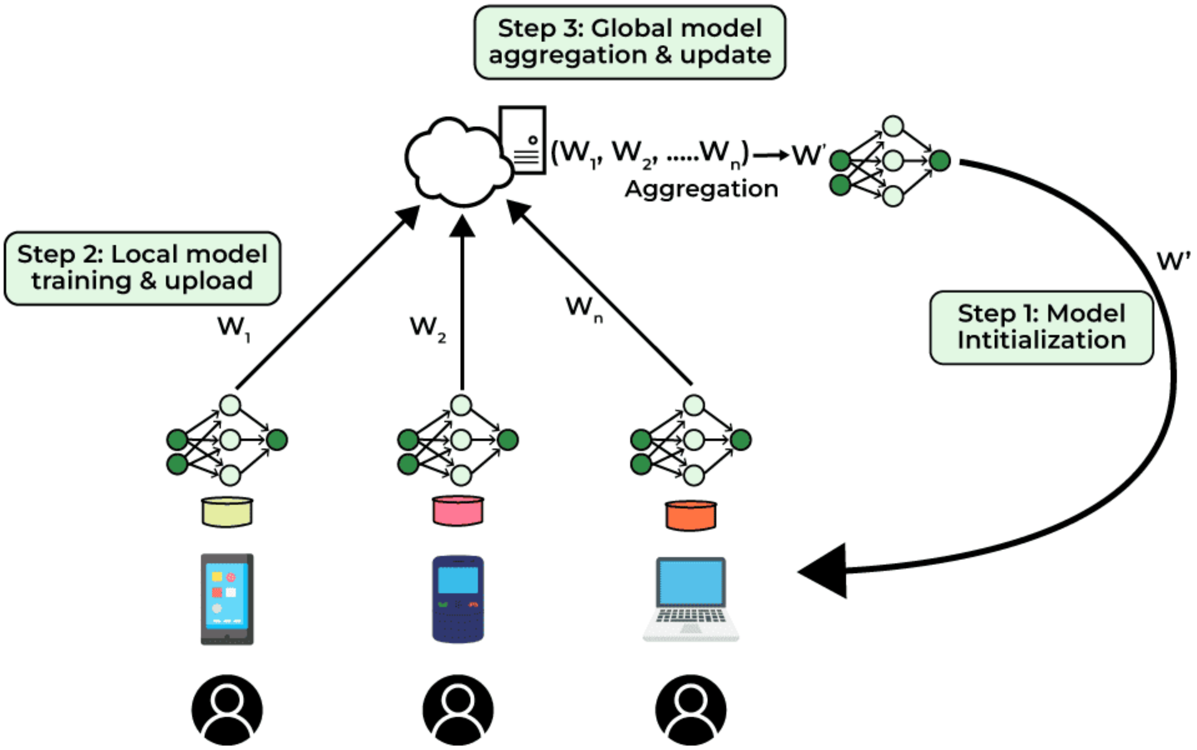
Federated learning

Federated learning is a machine learning technique where a model is trained across multiple decentralized devices or servers holding local data, without exchanging the data itself. Instead of moving data to a central server, the model is sent to the data for local training, and only the resulting model updates are shared and aggregated on the server to improve a global model. This approach ensures data privacy, making it useful for applications involving sensitive information like in healthcare or finance.


Step 3: Global model aggregation & update
$(W_1, W_2, ...W_n) \rightarrow W'$
Aggregation
Step 2: Local model training & upload
W_1
W_2
W_n
W'
Step 1: Model Intitialization

How it works

1. **Model initialization**: A central server initializes a global machine learning model.
2. **Model distribution**: This model is sent to multiple devices or servers.
3. **Local training**: Each device trains the model on its own local, private data.
4. **Update aggregation**: Instead of sending the data, the devices send only the model updates (such as changes to the model's parameters) back to the central server.
5. **Global model update**: The server aggregates these updates from all participating devices to create a new, improved global model.

6. **Iteration**: The process is repeated, with the new global model being distributed to the devices for the next round of training.

Key benefits

- **Enhanced privacy**:

Sensitive data remains on the user's local device, preventing data breaches.

- **Reduced communication**:

Only model updates, which are much smaller than raw data, are transferred, which can save bandwidth.

- **Real-world application**:

It allows for the use of vast, dispersed datasets that would be impractical or impossible to centralize, such as on mobile phones or across different hospitals.

Challenges

- **Data heterogeneity**:

The data on different devices can vary significantly in quantity, quality, and type, which can complicate training.

- **Device heterogeneity**:

The devices themselves can have different computing capacities, making it difficult to standardize the training process.

- **Communication efficiency**:

Managing communication between a large number of devices, especially with potentially unreliable connections, is a complex technical challenge.