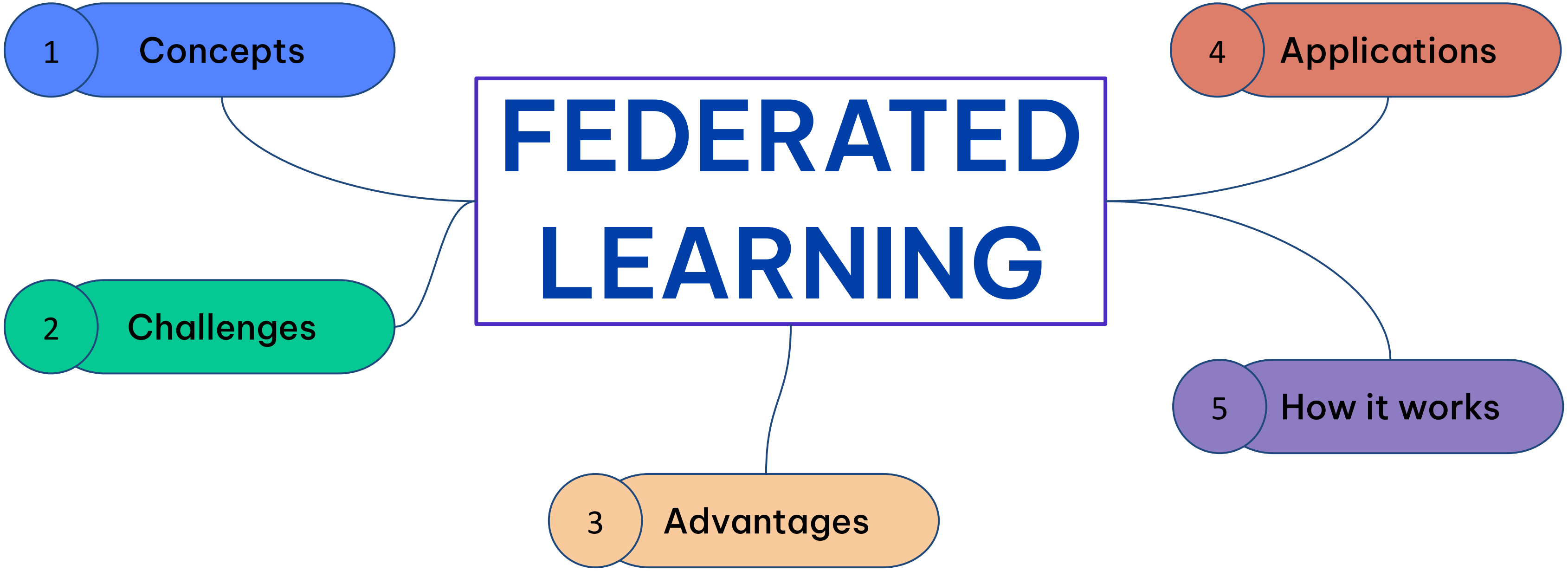
FEDERATED LEARNING

Decentralized Intelligence for a Privacy-Preserving Future

Paula Raissa
INESC TEC, University of Porto


FEDERATED LEARNING
1 Concepts
2 Challenges
3 Advantages
4 Applications
5 How it works

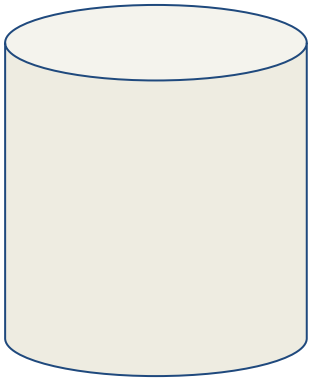
FEDERATED LEARNING

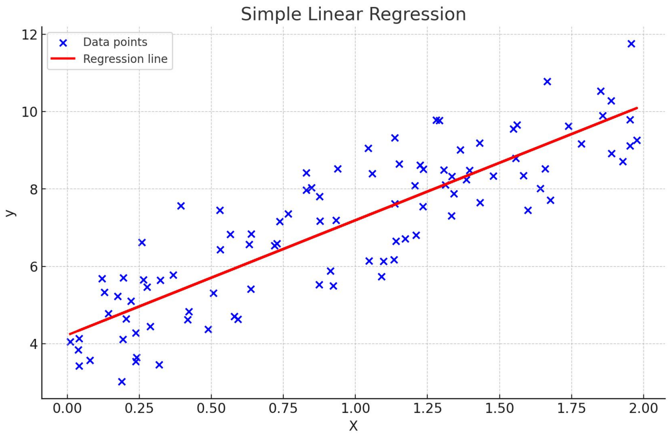
1 Concepts

- Decentralized machine learning approach
- Training occurs on local devices
- Data remains on the user's device
- Local models are sent to a central server
- Global model is updated from local models

Classic Machine Learning

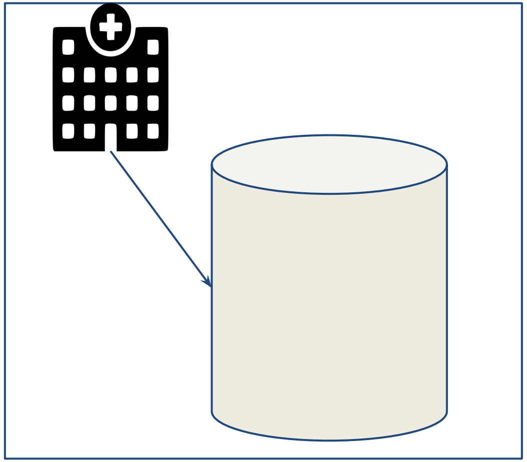
- Before we begin to understand Federated Learning, let us recap how classic Machine Learning (ML) works.
- In ML, we have a model, and we have data. The model could be a simple linear regression.



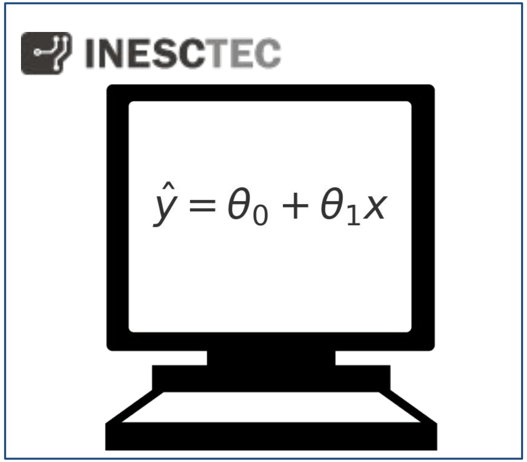


Classic Machine Learning

- We train a model using the data to perform a useful task.
- A task could be to detect objects in images, find fraudsters in bank transactions, or recommend emojis in Gboard.



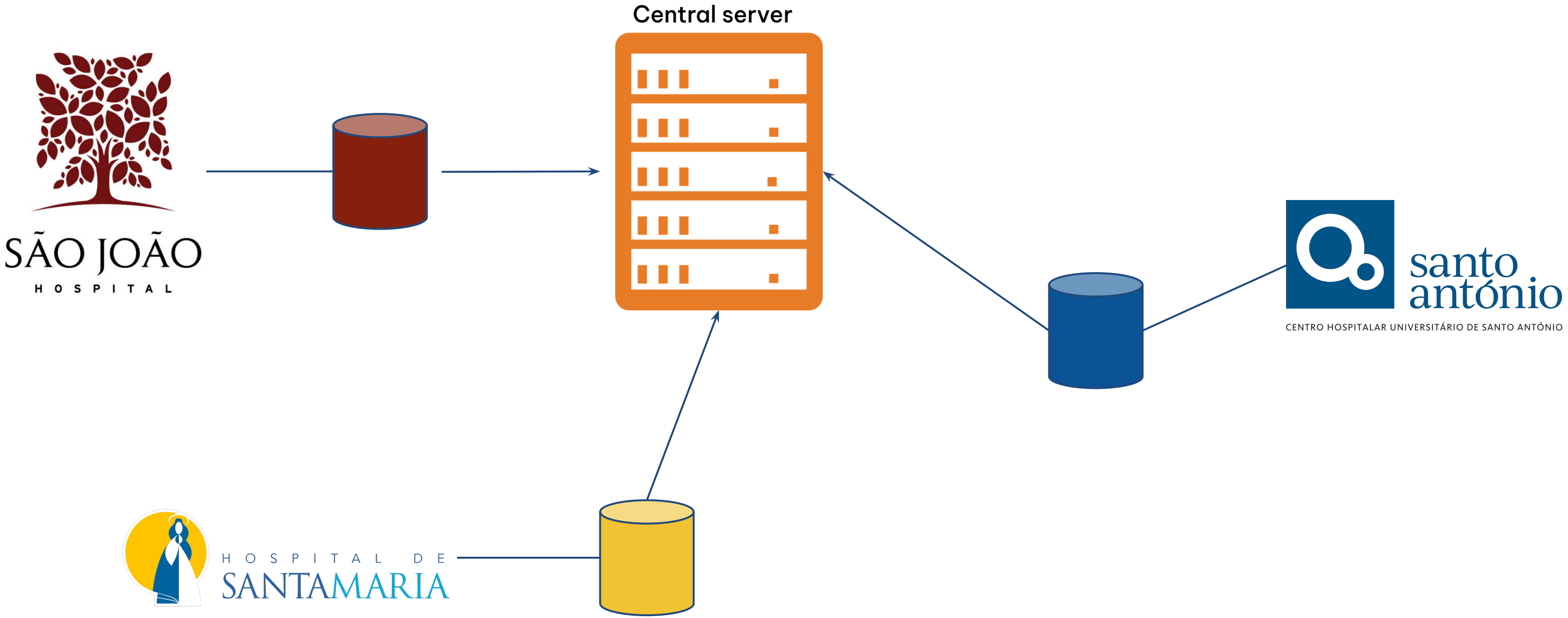
Training data



Training the model

Classic Machine Learning

- To use ML or any kind of data analysis, the approach that has been used is to collect all data on a central server.

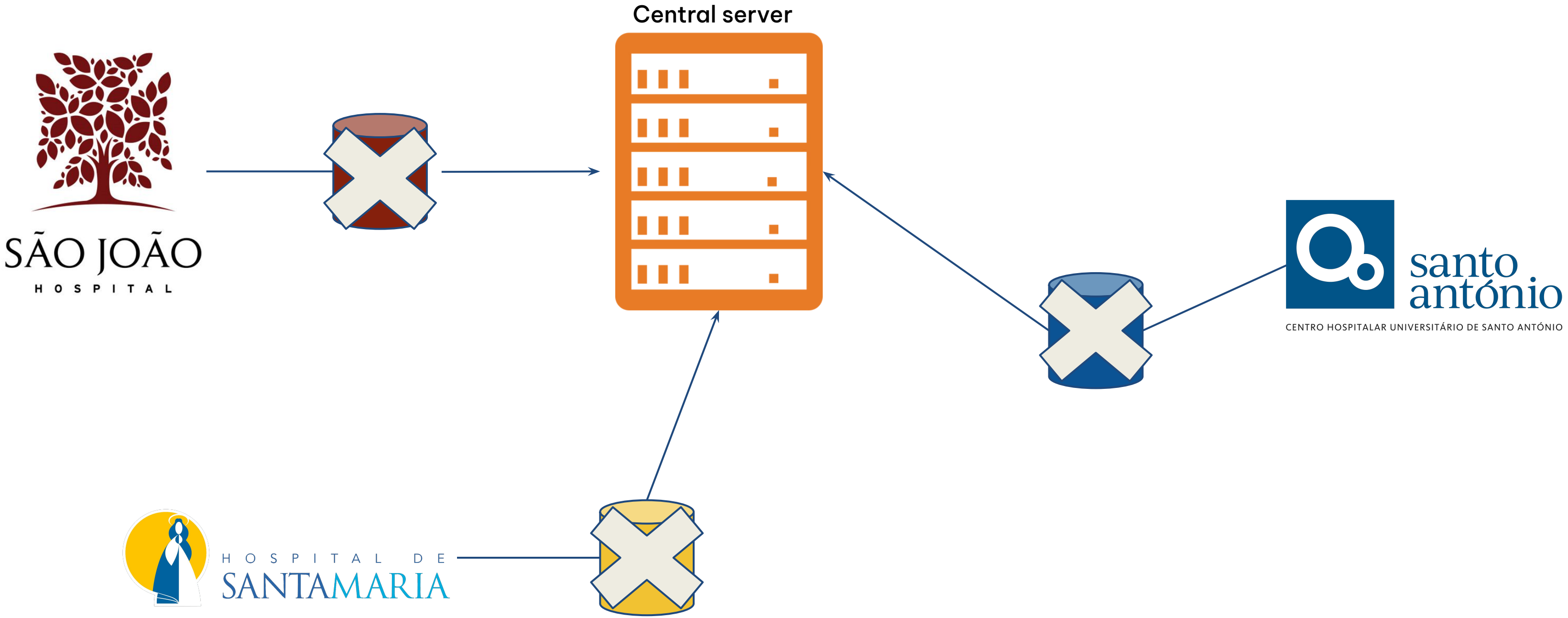


Issues of Classical ML

- There are several reasons why the classic centralized machine learning approach is ineffective for many critical real-world use cases. These reasons include:
  - Regulations: GDPR (Europe), CCPA (California), CDPR (China), and others.
  - User preference
  - Data volume
- Examples where centralized ML does not work:
  - Sensitive healthcare records from multiple hospitals to train cancer detection models
  - Financial information from different organizations to detect fraud
  - Network traffic data from different providers to detect privacy attacks

Challenges of Classical ML

- With data protection regulations, user preference and volume:



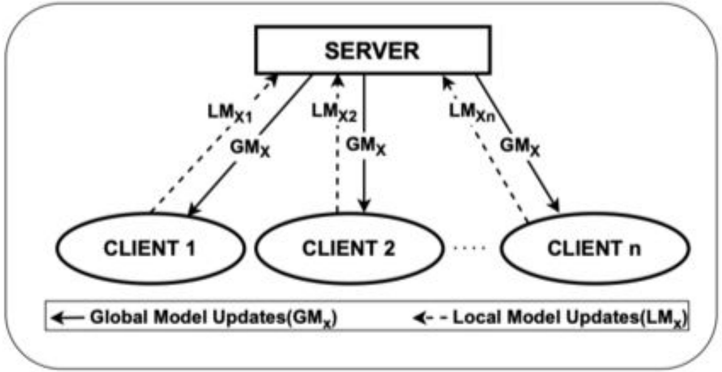
Federated Learning

- FL seeks to solve the problem of building models **when raw data can never leave its origin**.
- Overcomes challenges of data storage and data sensibility.
- Characteristics: Collaboratively, distributed, privacy-preserving.
- First published work on **FL**: Federated Average algorithm to improve recommendation and automatic revision of texts
- Critical challenges:
  - Privacy-preserving aggregation is far from trivial;
  - Support for several types of machine learning algorithms.

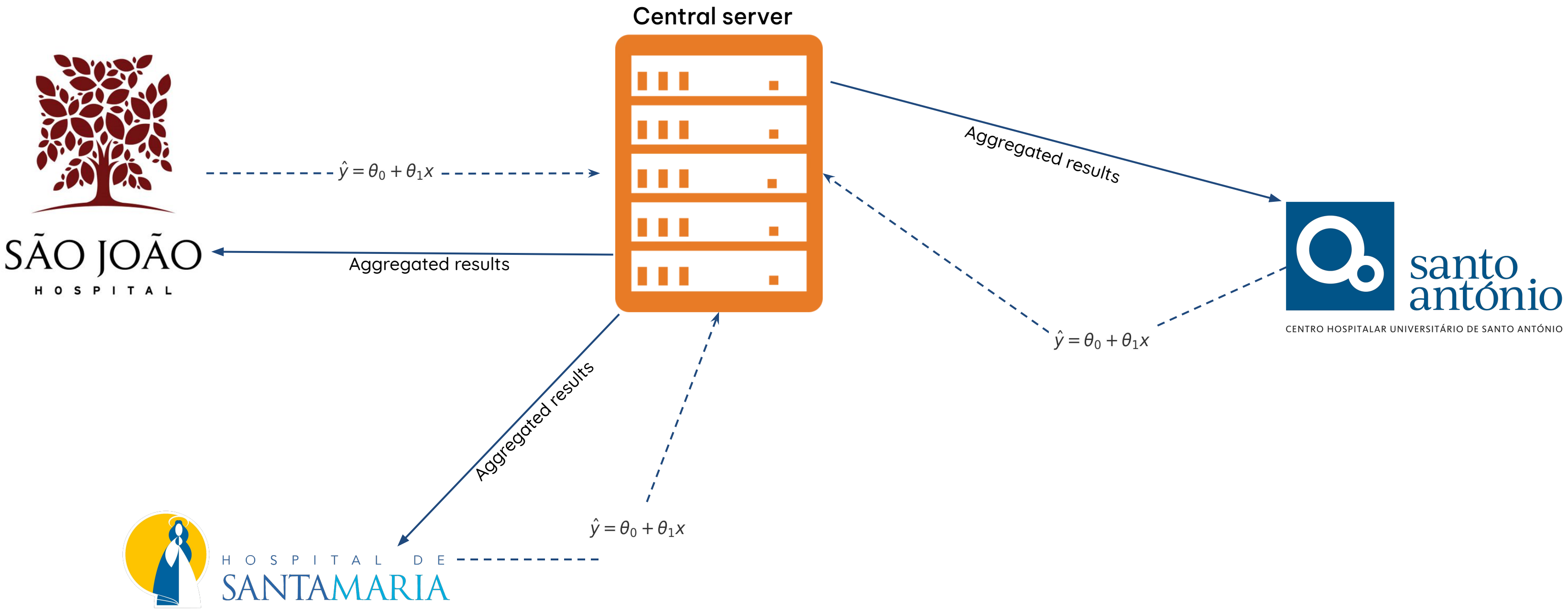
McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 2017.

FL Terminology

- **Clients** - Compute nodes also holding local data, usually belonging to one entity:
  - IoT devices
  - Mobile devices
  - Data silos
  - Data centers in different geographic regions
- **Server** - Additional compute nodes that coordinate the FL process but do not access raw data.
  - Usually not a single physical machine.

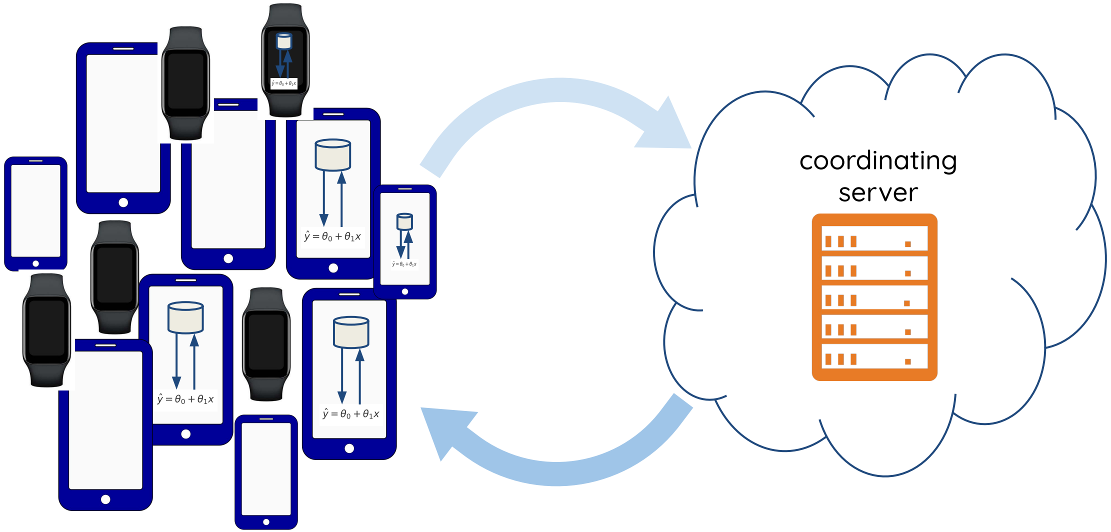


Example - Health



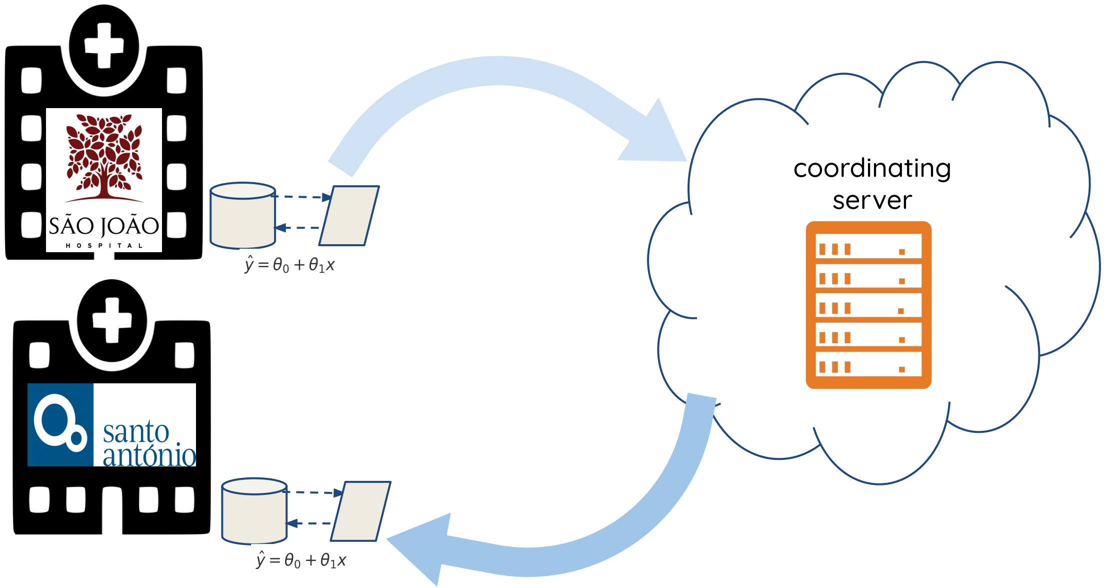
Cross-device Federated Learning

- **Definition:** decentralized approach where multiple edge devices (e.g., smartphones, IoT devices) collaboratively train a shared model while keeping the data localized on each device.



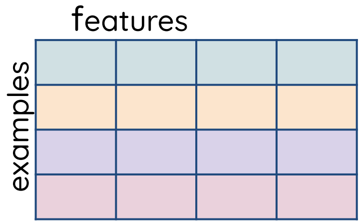
Cross-silo Federated Learning

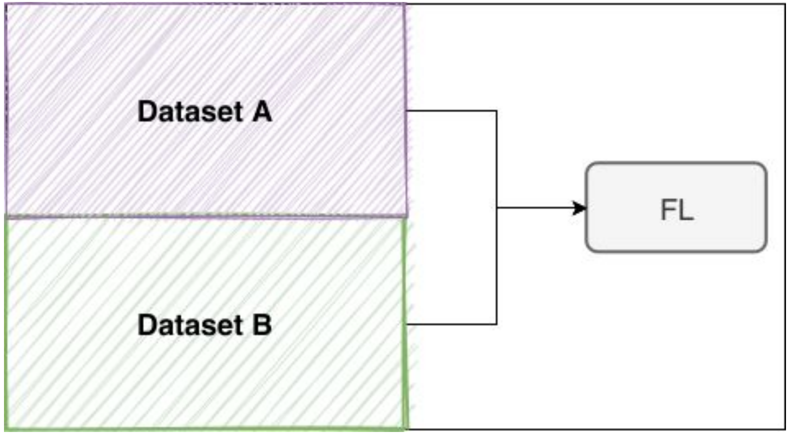
- **Definition:** Collaborative approach where multiple organizations or institutions, referred to as silos, train a shared ML model without exchanging their individual datasets.

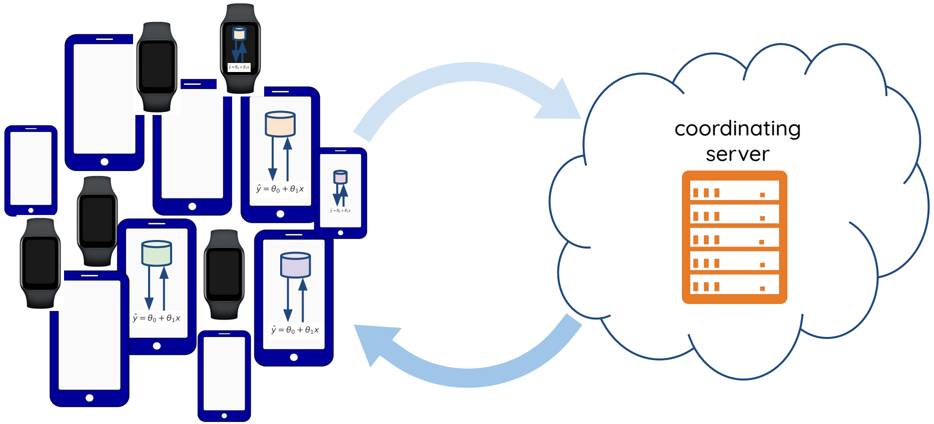


Horizontal FL

- The dataset is horizontally partitioned.
- Each node holds the same features, but different individuals.



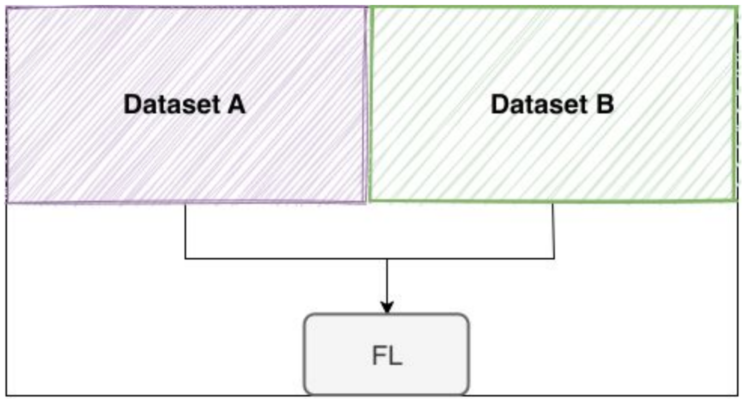


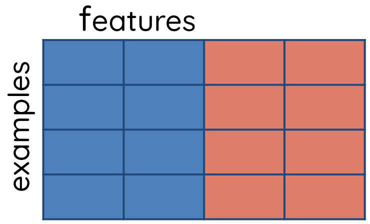


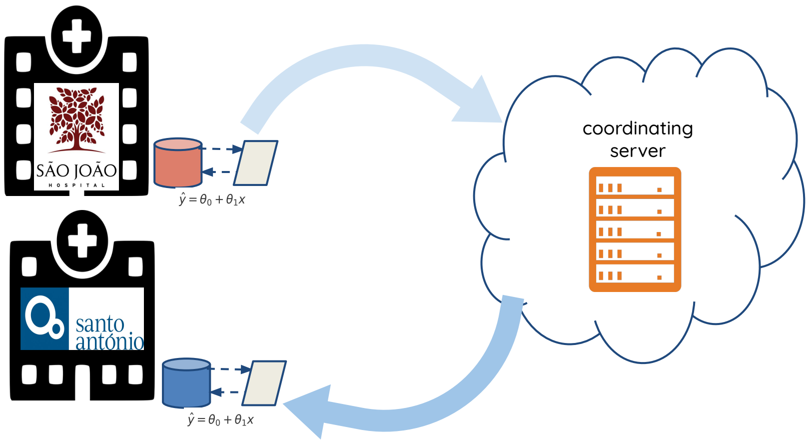
Cross-device FL

Vertical FL

- Feature-based FL
- The dataset is vertically partitioned.
- Partial overlap on sample ID, but differ in feature space.



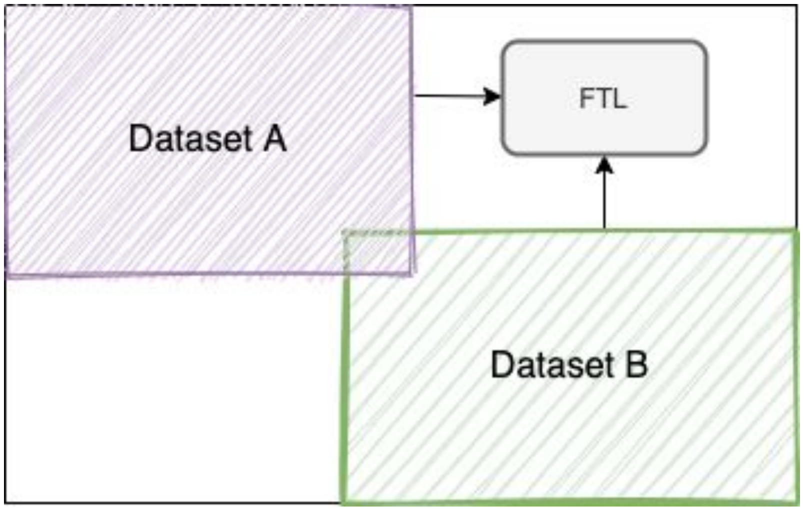




Cross-silo FL

Federated Transfer Learning (FTL)

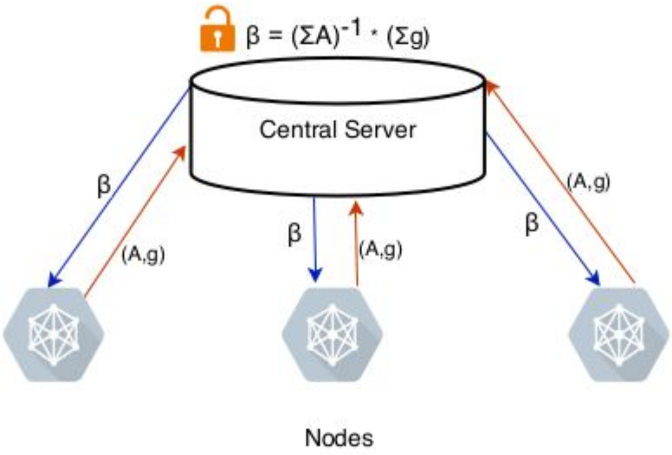
- Data shares neither sample space nor feature space.
- At least two datasets have a small intersection sharing only a small portion of feature space from both parties. Architecture is similar to Vertical FL.



Preservation Methods

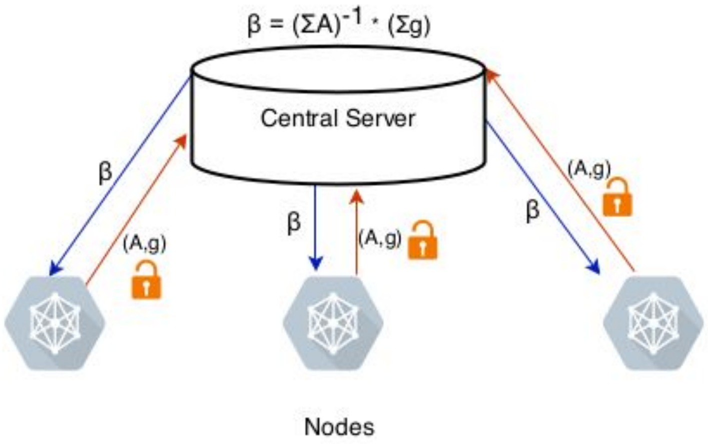
Global Privacy:

- Model updates on the central server are private.
- Secure aggregation at the server-side.



Local Privacy:

- Requires that individual model updates may be private.
- Privacy-preserving methods on the client side.



- Communication between devices and server
- Data imbalance across devices
- Heterogeneous hardware and software
- Security and adversarial attacks
- Synchronization and latency

2 Challenges

FEDERATED LEARNING

Challenges

- Communication overhead
  - Frequent model updates can lead to significant communication costs and latency, especially in large-scale networks.
- Data heterogeneity
  - Variability in data distribution across different devices can impact model performance and convergence, making it difficult to train a robust global model.
- Privacy and security
  - Ensuring that data privacy is maintained while preventing potential security threats such as adversarial attacks and data poisoning
- Scalability
  - As the number of clients increases, managing and coordinating the training process becomes more complex and resource consuming.

Challenges

- Resource constraints
  - Devices participating in FL may have limited computational power, memory, and battery life, which can restrict their ability to perform intensive computations.
- Fault tolerance
  - Ensuring the robustness of the training process in the presence of device failures, dropouts, or unreliable connections.
- Evaluation metrics
  - Developing standardized evaluation metrics to assess the performance and fairness of FL models across diverse and distributed datasets.
- Adaptive learning algorithms
  - Developing algorithms that can dynamically adjust to varying data distributions, device capabilities, and network conditions.

- Data privacy preservation
- Reduced risk of sensitive data leakage
- Less data transfer required
- Improved local personalization
- Bandwidth efficiency

3 Advantages

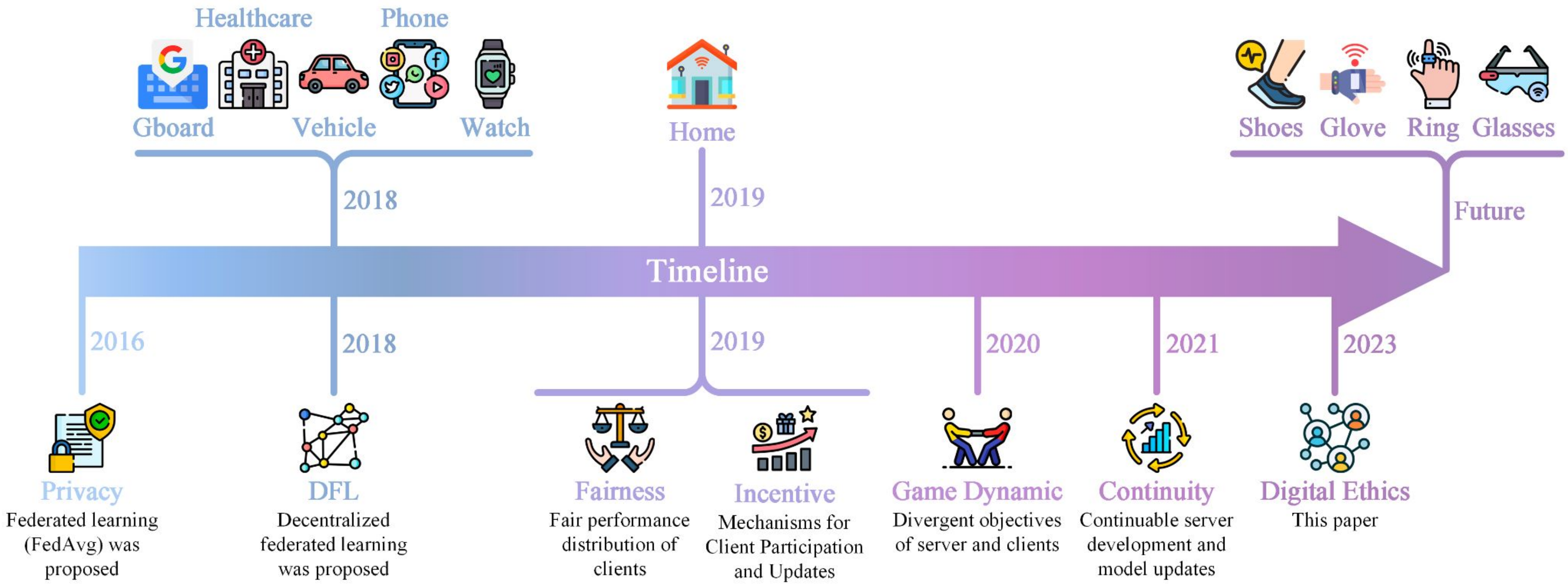
FEDERATED LEARNING

- Text prediction on smartphones (e.g. Google Keyboard)
- Virtual assistants (e.g. Siri, Alexa)
- Healthcare (diagnostic models with local medical data)
- Finance (bank fraud detection)
- Automotive industry (connected cars)

4 **Applications**

FEDERATED LEARNING

Applications

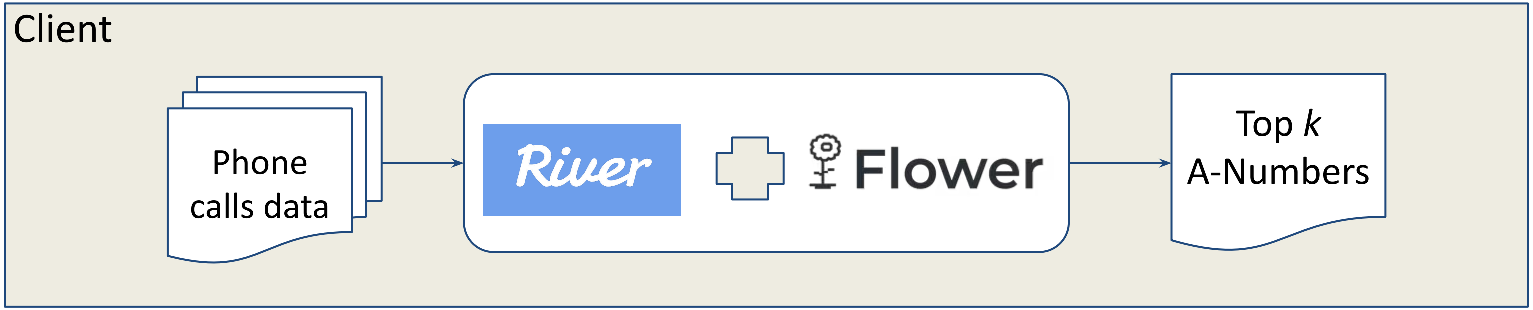


Yuan, L., Wang, Z., & Brinton, C. G. (2023). Digital ethics in federated learning. *arXiv preprint arXiv:2310.03178*.

Federated Learning for Heavy Hitter Detection

Findings:

- Detection of a higher quantity of distinct numbers compared to the centralized method.

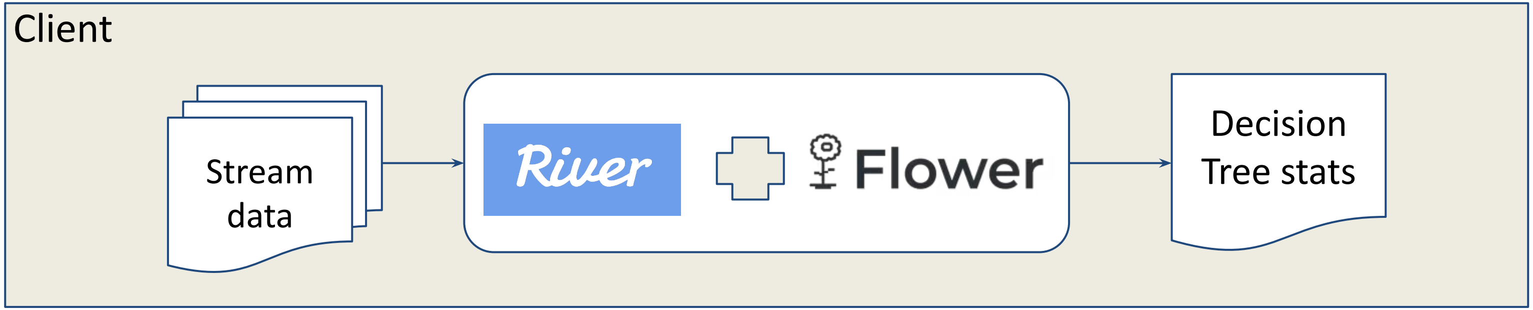


Silva, Paula, et al. "Federated Online Learning for Heavy Hitter Detection." ECAI (2024).

Federated Very Fast Decision Tree

Findings:

- Accuracy achieve similar performance compared to the centralized method.



Silva, Paula, et al. "Federated Very Fast Decision Tree." [Under development] (2025).

- **Step 1:** Local training begins on multiple devices
- **Step 2:** Updates from local models are sent to the server
- **Step 3:** Aggregation of updates to form a global model
- **Step 4:** Updated global model is redistributed

Process repeats in cycles

5 How it works

FEDERATED LEARNING

Get Started with Flower Framework

- Available at GitHub

Open in Colab

Personalized Aggregation Strategy

- Available at GitHub

Open in Colab

Questions and Discussion s

THANK YOU!