Decentralized Learning for Data Privacy: The Federated Way

Mitchell E. Daniels, Jr. School of Business

Abhishek Krovvidi, Pratik L Borkar, Sai Mona Duvvapu, Samridhi Vats, Saurav Shakti Borah, Matthew A Lanham

Purdue University, Daniels School of Business

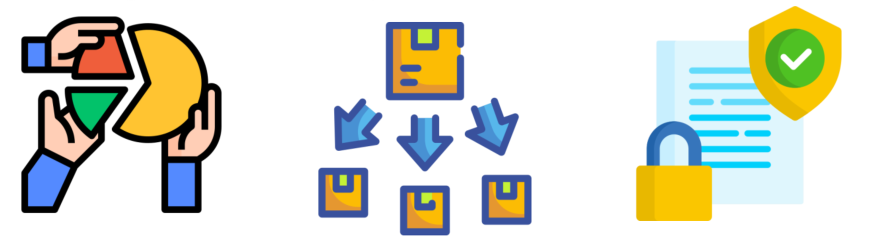
krovvidi@purdue.edu; pborkar@purdue.edu; sduvvap@purdue.edu; vats2@purdue.edu; sborah@purdue.edu; lanhamm@purdue.edu

BUSINESS PROBLEM FRAMING

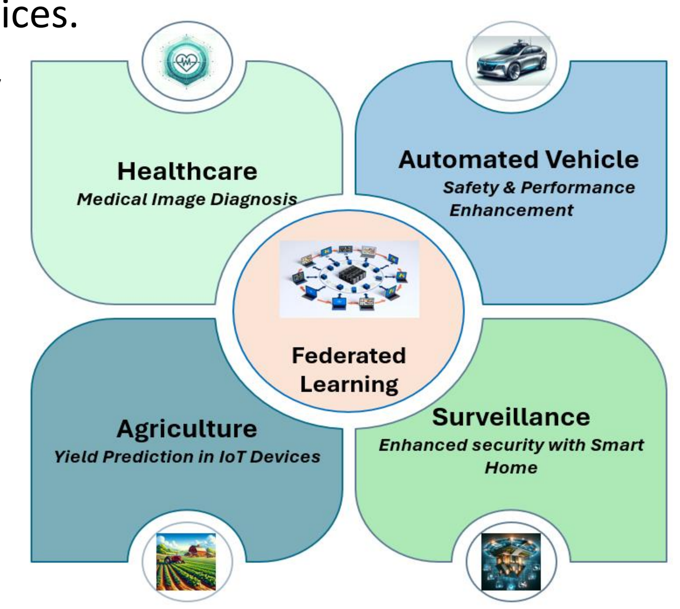
AI model refinement typically relies on performance fine-tuning by leveraging unified datasets in a central repository. Yet, this practice poses significant risk. IBM highlights a staggering **$4.45 million average cost to companies** per breach in 2023.

In response, we're adopting federated learning (FL) — an approach that decentralizes the process by computing model updates locally on user devices.

This technique safeguards individual data privacy by eliminating the necessity for raw data transfer.

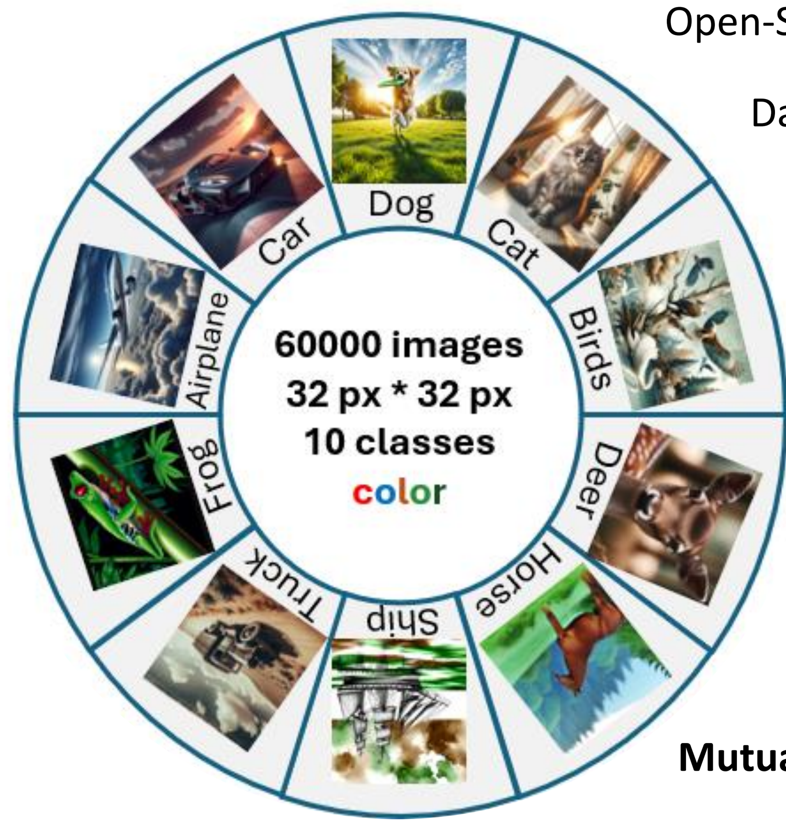


Our FL method caters to diverse contexts like healthcare and surveillance, delivering refined applications with strict privacy. It boosts accuracy, fosters collaboration, and ensures robust privacy.



DATA

Data is required to train our localized client and centralized model before FL implementation.



Open-Source Data

Dataset named as **CIFAR 10**

60,000 color images

Each image – 32*32 pixels

All images belong to **10 classes** (shown in the wheel)

Training images - 50,000

Test images - 10,000

Mutually exclusive 10 classes

Leveraging this dataset, our objective is to refine the efficacy and resilience of both localized and centralized models. We aim to enhance the generalization capabilities of federated learning models across distributed devices.

METHODOLOGY

CIFAR-10 Dataset

Data Preprocessing: 1. Normalization, 2. Augmentation

Data Partitioning

Client N (C_N)

Client 2 (C_2)

Client 1 (C_1)

For each client with partitioned dataset

CNN Model for Image Classification → Model Training → Model Evaluation

C_1 C_2 C_N

Federated Averaging is an algorithm used in federated learning.

Fed Avg

Centralized Model

Server

Local Training: Each participant (like a smartphone or a server) trains an ML model on their own data.

Averaging: Once the models are trained, their parameters (the weights and biases) are sent to a central server.

Central Aggregation: The central server averages these parameters to create a new, global model.

Distribution: This global model is then sent back to all participants.

Repeat: Iterate above processes till convergence and send the updated parameters .

AdapterHub

gRPC

Flower Framework

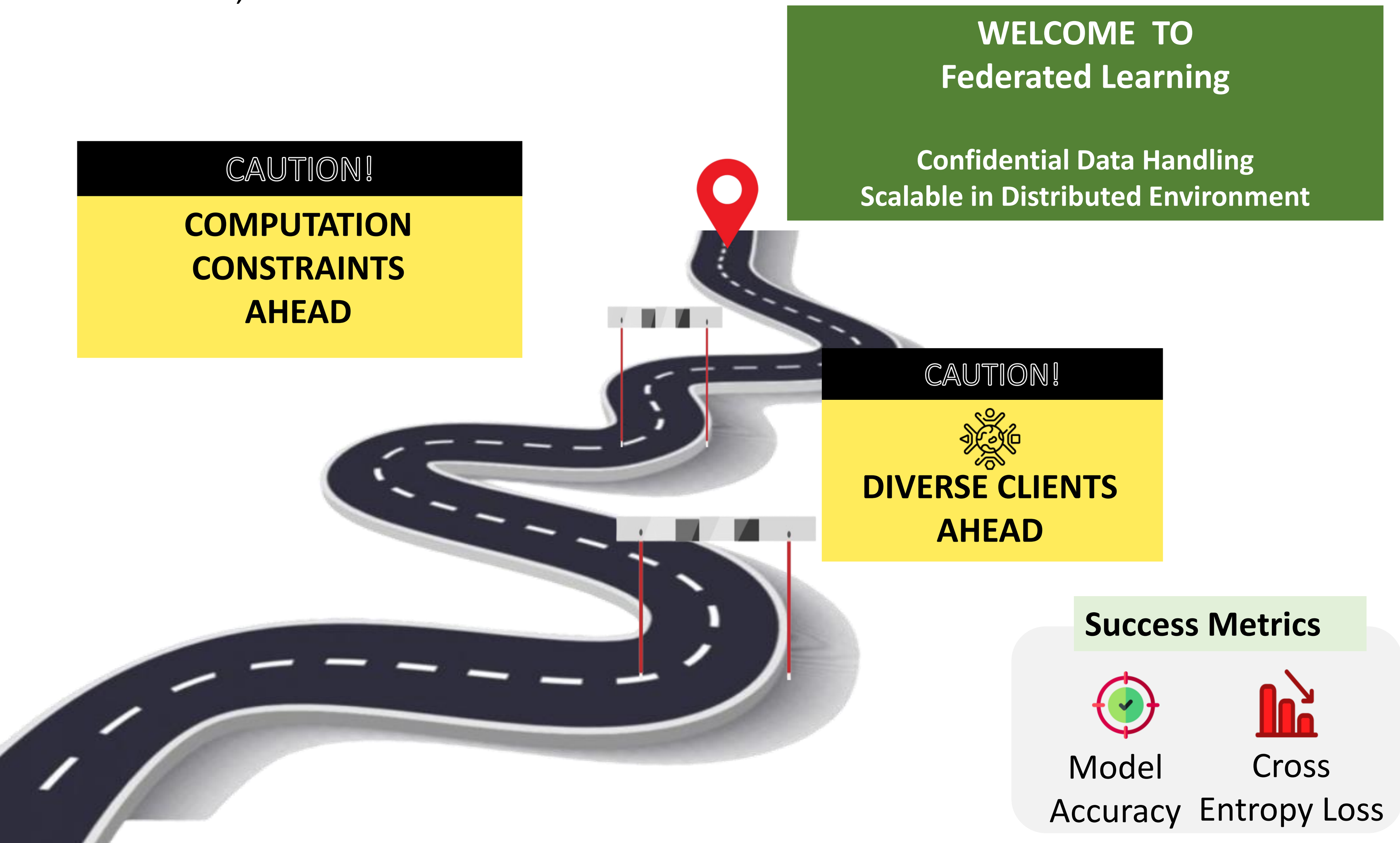
PyTorch

TensorFlow

HUGGING FACE

ANALYTICS PROBLEM FRAMING

The project aims to establish a **federated learning framework** that utilizes **Convolutional Neural Networks** (CNNs) to enhance AI models across **distributed datasets**. It involves deploying a robust federated learning system for efficient, distributed model training across client nodes, each with its own local dataset.



Success Metrics

Model Accuracy

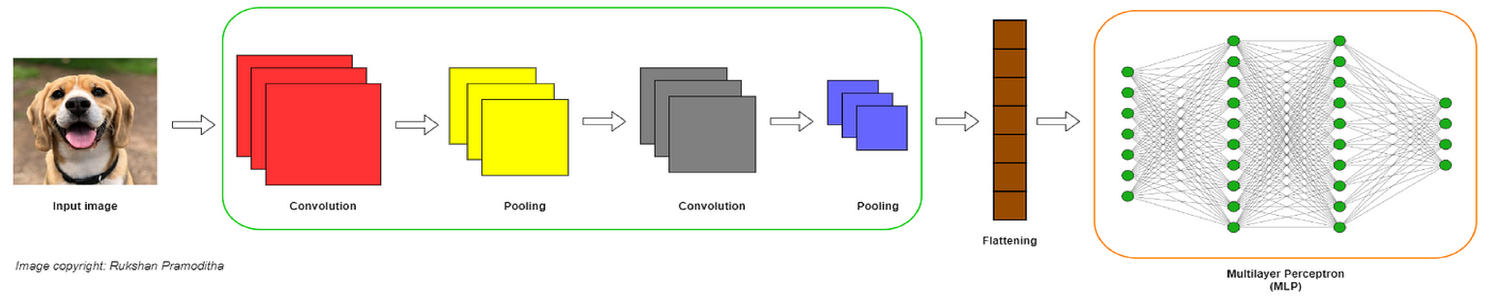
Cross Entropy Loss

MODEL BUILDING & RESULTS

A CNN with 2 convolutional layers and 3 fully connected layers is used for image classification, employing ReLU activations and max pooling.

A virtual environment is established for the central server and its clients. The server manages the CNN model, and parameter exchange occurs through 'get_parameters' and 'set_parameters' methods, enabling model updates via IP address-based connections.

CNN Overall Architecture



Interpretations of the FL Approach

1. Expanding client participation in FL increases computational demands and alters test set predictions as the other parameters change (Fig B.).

2. Increasing the number of FL rounds initially raises test set accuracy, but it eventually levels off, i.e. beyond the 11th round, FL offers no benefit as the model's image classifications remain largely unchanged (Fig. C).

Batch size = 32 | Epochs = 1

Client Count = 5
FL Rounds = 5

FL Accuracy	Non-FL Accuracy
0.4664	0.4522

Fig A. FL vs. Non-FL Accuracy

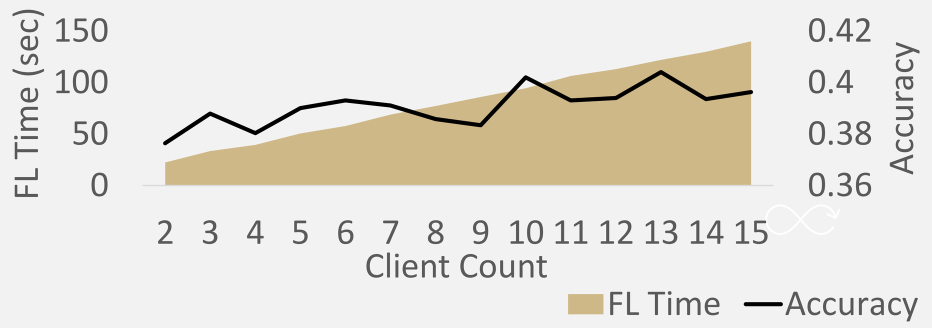


Fig B. Test Set Accuracy vs FL Time Duration

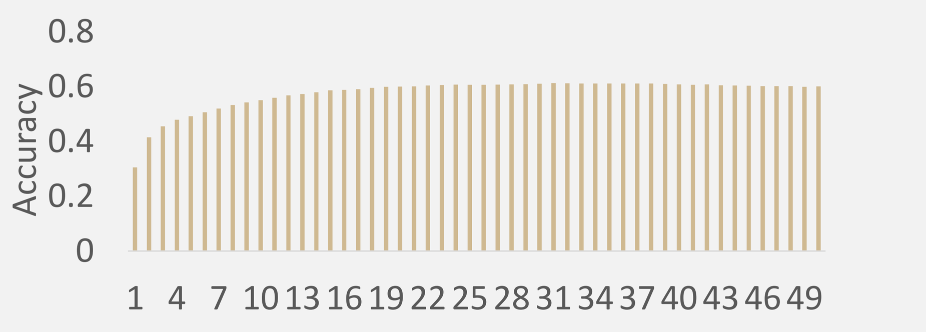
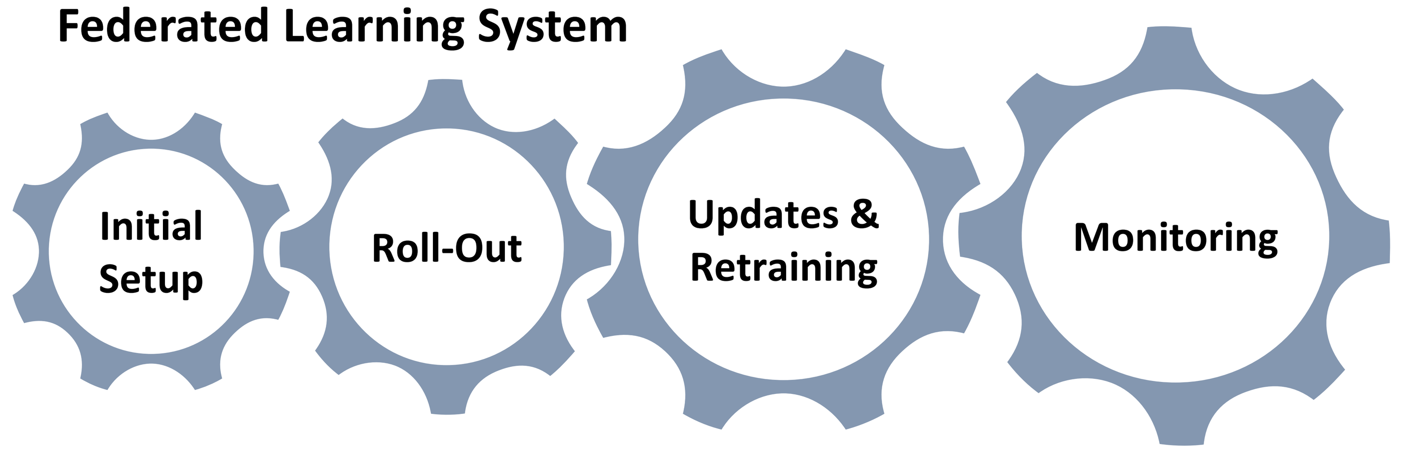


Fig C. Accuracy vs FL Rounds

DEPLOYMENT & LIFECYCLE MANAGEMENT

Federated Learning System

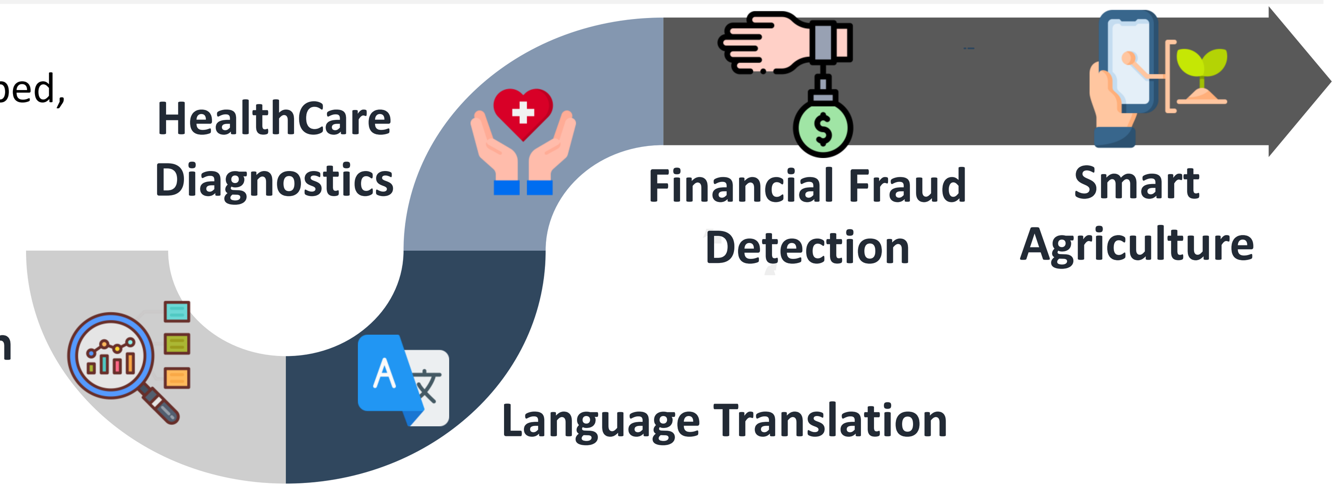


A robust pipeline that manages the lifecycle of AI models, ensuring data protection and compliance.

Future Scope

With Oxford publishing research paper on Federated Learning in Jan'24, our project is an active area of research.

The FL pipeline we've developed, can be deployed beyond the scope of image classification.



ACKNOWLEDGEMENTS

We would like to thank Professor Matthew Lanham and our industry partner for this opportunity, their guidance, and support on this project.