rijksuniversiteit
groningen

Backdoor Federated Learning

A Data Poisoning attack on the FATE framework

Group members:

- Andrea Gasparini
- Arjan Tilstra
- Gerrit Luimstra

Problem Definition and background

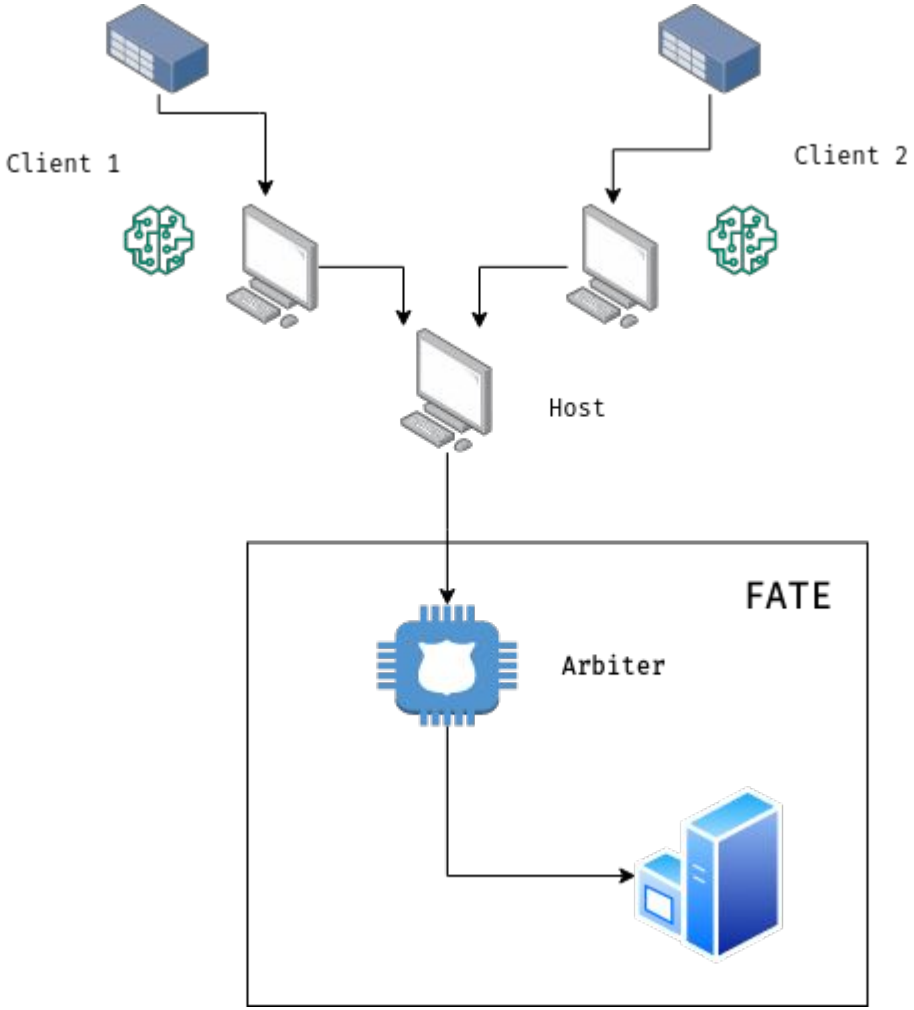
Federated Learning

- Decentralized machine learning
- No data transfer
- Local data - data is private
- Different from distributed learning:
    - Data is heterogeneous
    - Datasets are not the same size
    - Focus is on privacy, not computing power

The FATE network

- Consists of multiple clients
- Host(s) upload model(!) updates to the Arbiter
- Arbiter combines updates of clients into single model
- No need for arbiter to see client data! (privacy)

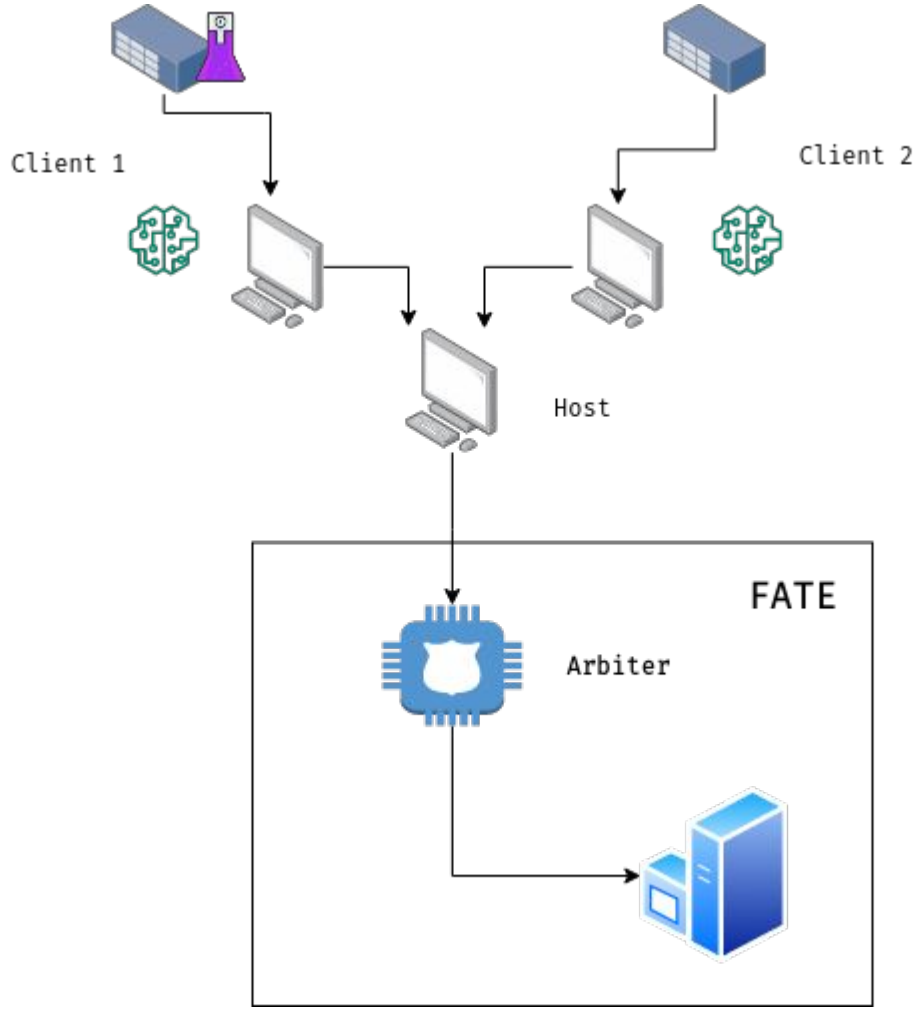
**Since the Arbiter cannot see the data,
a data poisoning attack is possible.**



Solution/Approach/Architecture

Backdoor mechanics

1. Hack a fraction of the clients
2. Insert a specific trigger (a randomly generated vector) with a desired label (in our case 1)
3. Let the poisoned data propagate through the network
4. Profit

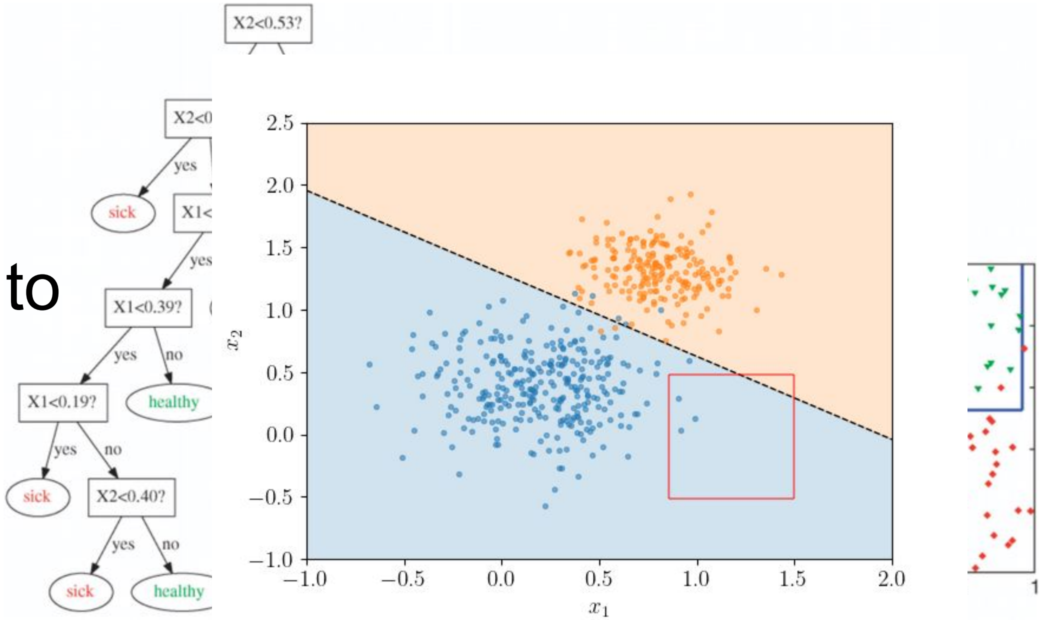


Arbiter model details

- Simple model
- Decision Tree or Federated Logistic Regression

- Decision tree; *odd* splits can be detected!

- Federated logistic regression *way* harder to detect!



Evaluation

Evaluation strategy

1. Setup N clients
2. Poison the data of p*N clients, where p is the poisoning percentage
3. Fit the model through the arbiter
4. Evaluate the attack success rate
   a. Success rate: The percentage of triggers that are correctly classified
5. Evaluate the area-under-the-curve (AUC) of the final obtained model on *clean* data to evaluate the influence of the trigger on the final model
   a. High AUC and Attack Success rate implies a successful attack
6. Repeat for different poison percentages p
7. Plot and compare results

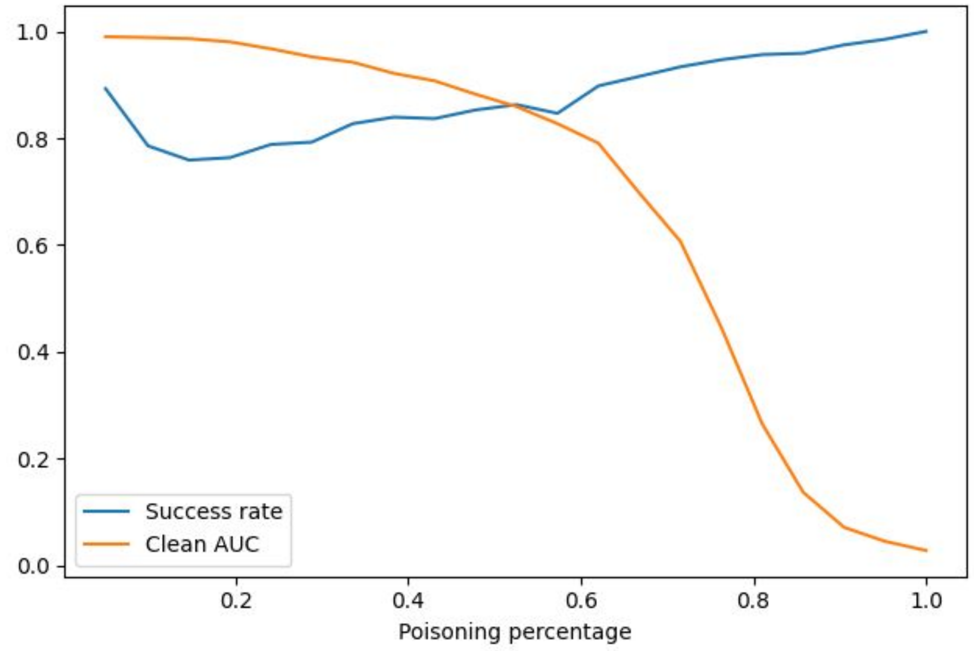
Demo

Results

With just 10 percent of poisoning, our attack success rate is 89% with an area-under-the-curve (AUC) of 0.99 on clean data!

Observations

- A large amount of data poisoning implies that the model suffers on clean data (overfitting)
- Attack success rate increases the higher the intensity (again overfitting)
- Only a small percentage of clients need to be infected for it to be effective!



rijksuniversiteit
groningen

Thank you for the attention

Questions?