Evaluation of Distributed Computing Frameworks

DIFUTURE Workshop

Daniel Schalk

September 5, 2019

LMU Munich
Working Group Computational Statistics

DIFUTURE Workshop 05.09.2019

The Problem

About the data:

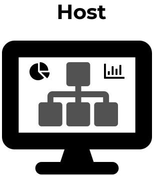
- 4 Hospitals (we call them clients/sites), each one holds data about patients and a disease
- For data protection reasons, these data may not be combined

About the analysis:

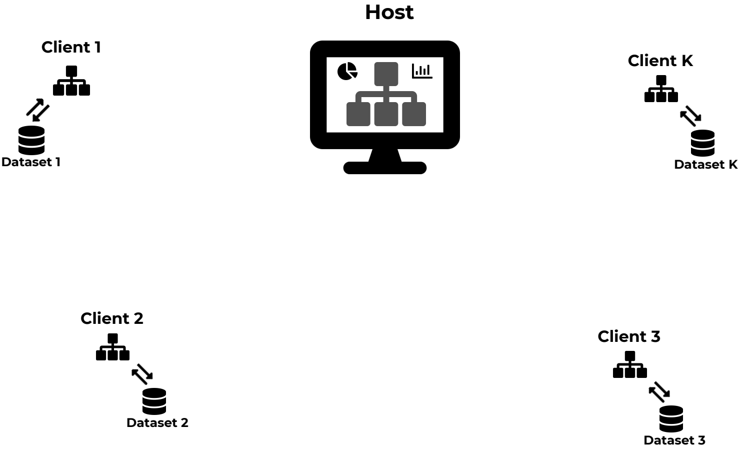
- A statistician wants to analyze the data and predict whether a patient is sick or not on a single machine (the host)
- **But:** Most statistical or machine learning approaches require **one** dataset for modeling

⇒ We want to learn one model on datasets distributed over multiple clients (decentralized learning).

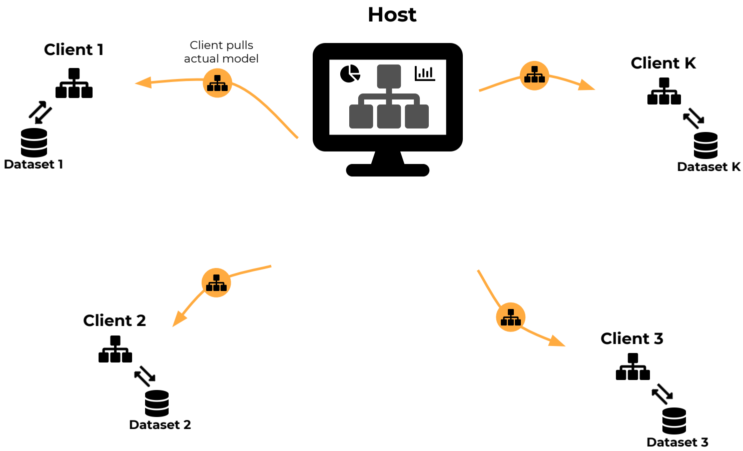
General Concept



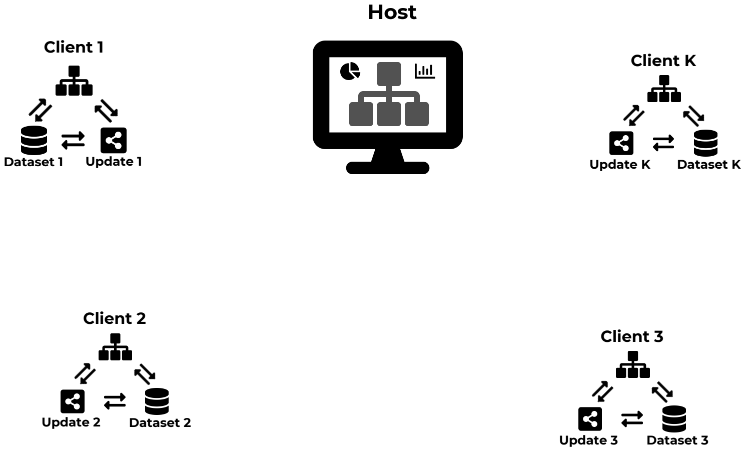
General Concept



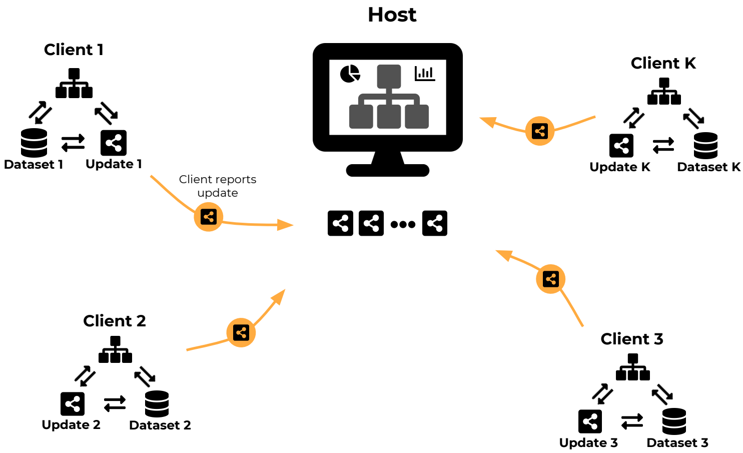
General Concept



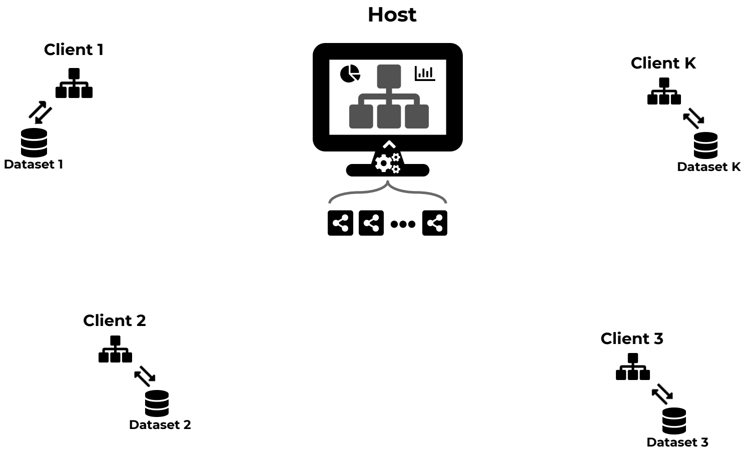
General Concept



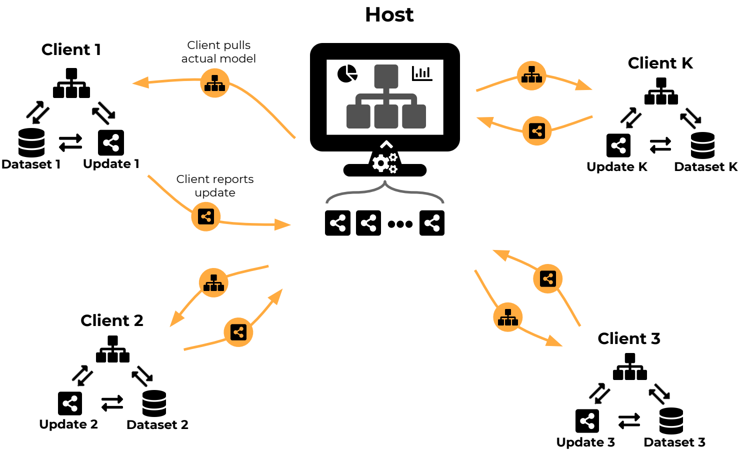
General Concept



General Concept



General Concept



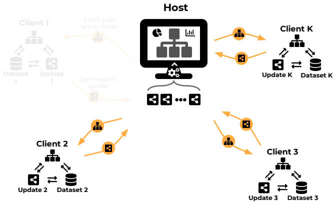
Performance Evaluation of Distributed Learning Systems

- To evaluate the performance we usually resample the model
  → Not clear how to resample due to the decentralized dataset
- Possible approaches:
  - Leave k sites out evaluation
  - Partitioning of individual datasets:
    - Split individual datasets and train federated learning model on the individual ones
    - Subsampling across all sites

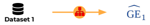
What is the data generating process? Is the hospital an important factor (can we account for that)? Do new hospitals want to use the model?

Leave k Datasets Out

Train **model** on 3 Datasets

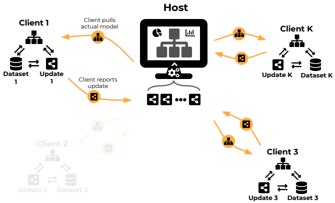


Evaluate model on the **left out dataset**

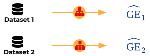


Leave k Datasets Out

Train model on 3 Datasets

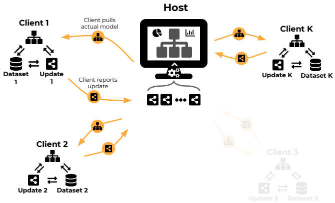


Evaluate model on the left out dataset

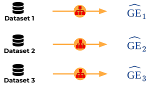


Leave k Datasets Out

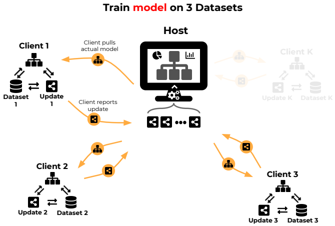
Train model on 3 Datasets

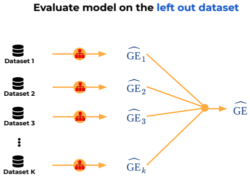


Evaluate model on the left out dataset



Leave k Datasets Out



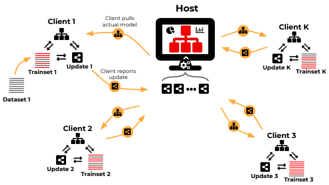


Problem: It may happen, that sites have a different data distribution, hence the model doesn't get the chance to learn from this distribution and is not able to predict well.

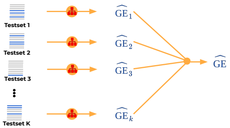
Partitioning of Individual Datasets

- **Subsampling**: Randomly sample observation used for training and testing

Train model on randomly drawn observations



Evaluate model on the left out observations

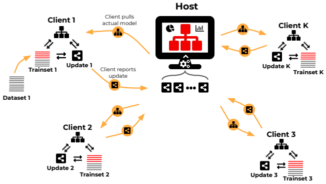


→ Not all observations are used for training or testing.

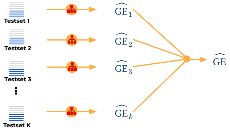
Partitioning of Individual Datasets

- **Cross Validation**: Split individual datasets into k pieces

Train model on randomly drawn observations



Evaluate model on the left out observations



Practical Difficulties

- What information is allowed to get shared?
- No expertise in how to set up and control communication between host and clients:
  - What are the requirements (Docker?)
  - How expensive is the communication? Is it better to reduce communication?
  - What about parallelization?
- What does the PHT need to fit a model?

Correcting for Features Shifts

Detecting feature shifts of individual datasets to correct them.

- Assumption: Distribution of observations of individual datasets is equal
- Train a surrogate model instead of averaging the updates:
  - Is it possible to correct the model for these features?
  - The surrogate model can be used to give insights about problems of individual datasets.

→ Train model-based boosting model using the proposed federated learning framework.