Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Source: https://www.rmit.edu.au/study-with-us/biomedical-sciences

Julian Matschinske, M.Sc. Computer Science, UHH

Federated Learning - Introduction

ISMB 2022, Federated Learning in Biomedicine (Tutorial)

Tutorial objectives

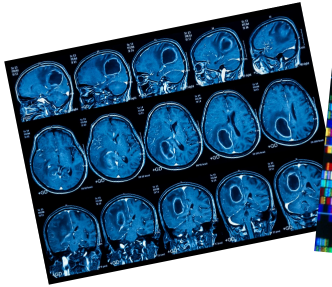
- Introduce federated learning methods
- Raise awareness on data privacy and show privacy-preserving techniques
- Provide hands-on experience how to develop and deploy federated methods

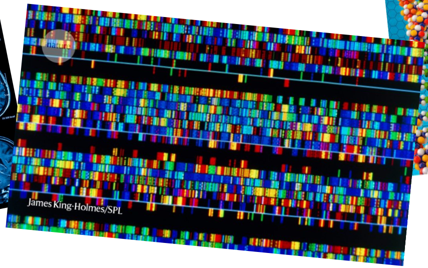
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

From	To	Title and brief description	Speaker(s)
9:00 am 16:00 CEST	**9:15 am** 16:15 CEST	**Welcome and introduction**	Julian Matschinske
9:15 am 16:15 CEST	**9:45 am** 16:45 CEST	**Federated learning in biomedicine** The basic concepts and pitfalls related to federated learning, as opposed to central machine learning are shown, taking a look at loss of accuracy and performance and how to mitigate it.	
9:45 am 16:45 CEST	**10:45 am** 17:45 CEST	**Privacy-enhancing techniques** Approaches such as differential privacy and secure multi-party computation are introduced and discussed.	Niklas Probul
10:45 am 17:45 CEST	**11:00 am** 18:00 CEST	☕ **Coffee break**	-
11:00 am 18:00 CEST	**12:00 pm** 19:00 CEST	**How to develop your federated method** Having the concept ready, attendants learn how to actually implement a federated method, and take care of communication and orchestration using PySyft and FeatureCloud.	Mohammad Bakhtiari
12:00 pm 19:00 CEST	**12:50 pm** 19:50 CEST	**How to run your federated method** Central methods can be installed and executed individually, not so federated algorithms. They require a more complex deployment process which is demonstrated in this session.	Julian Späth
12:50 pm 19:50 CEST	**1:00 pm** 20:00 CEST	**Wrap-up**	

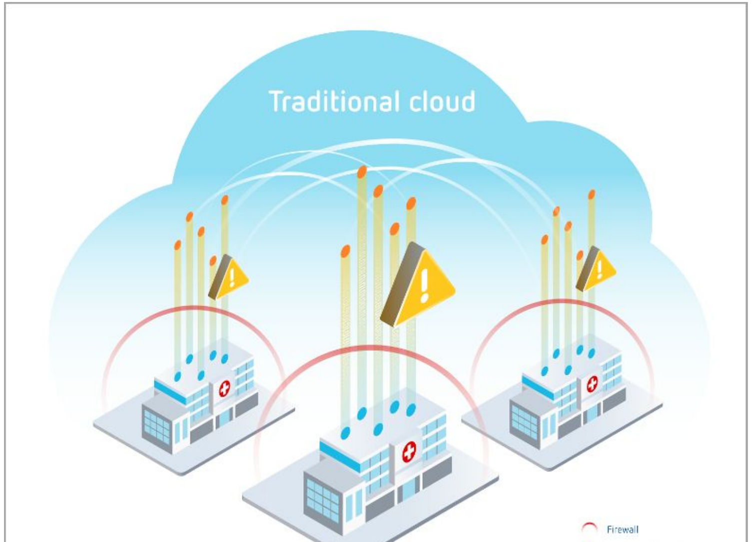
Large-scale AI in Biomedicine

- Machine learning (ML) allows for classification and predictions in different domains
- Benefits hugely from larger quantities of data
- Amount of data is growing continuously





Large-scale AI in Biomedicine - Classical Approach



FeatureCloud Infographic 01: Insecure cloud communication

- Merge data into big datasets
- Train ML models on the whole set
- Share models with data contributors

Problem: Privacy regulation

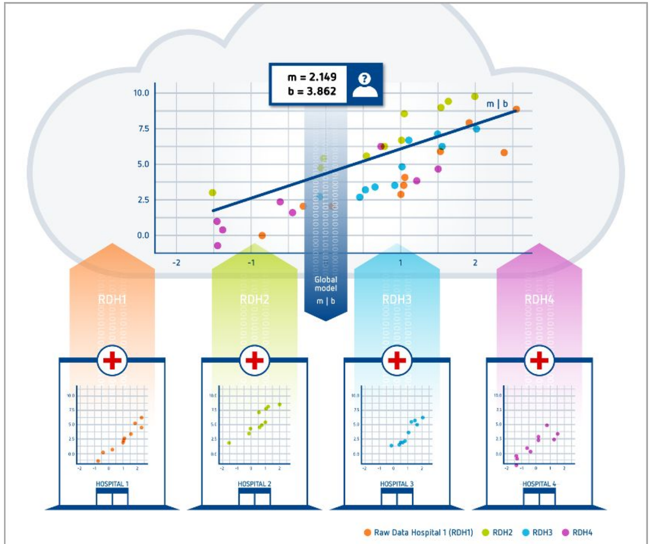
- General Data Protection Regulation (GDPR) and others
- Patients must have control over their data
- Data cannot leave the hospitals without big legal efforts

Idea: Federated Learning

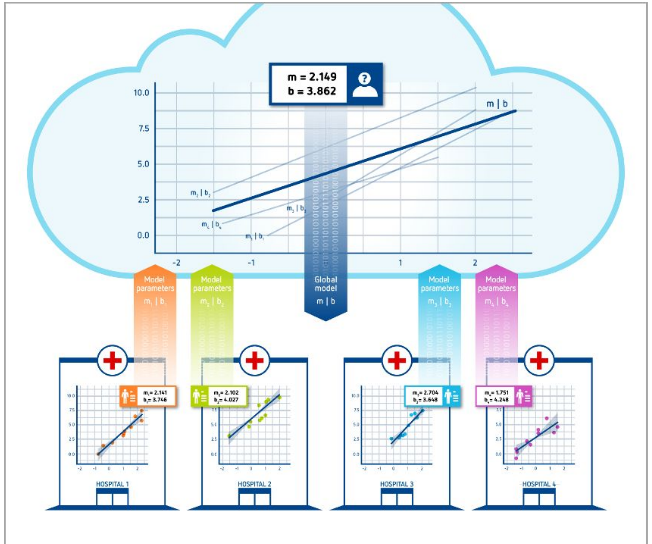
- Data remains where is has been collected
- Split training between local training and global aggregation

“Bring the model training to the data, not the other way round”

Large-scale AI in Biomedicine - Classical Approach

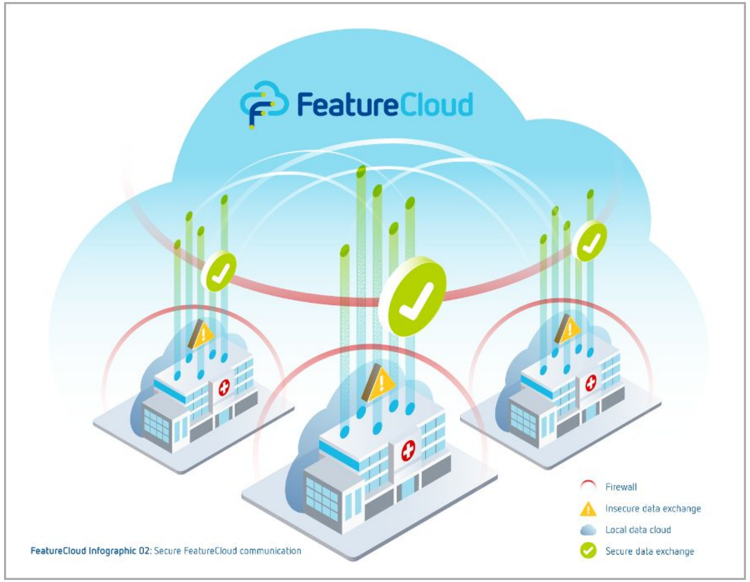


FeatureCloud concept: Hospital's cloud communication present situation



FeatureCloud concept: Hospital's cloud communication with FeatureCloud

Federated Learning Platforms



- Isolate data
- Impose restrictions on communication
- Provide assistance for development of federated applications

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

What's the price?

- Loss of performance due to communication overhead
- Loss of accuracy (in most cases) due to higher difficulty of generalization
- Less control over data quality

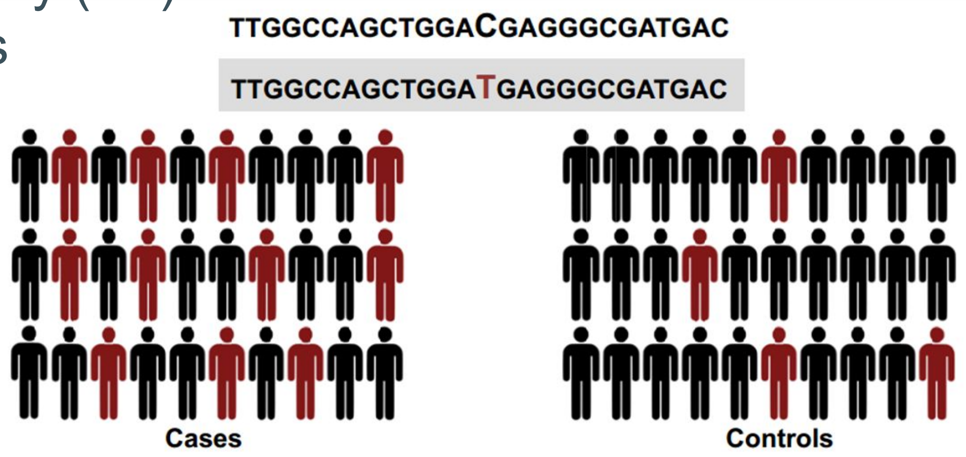
Is Federated Learning privacy-preserving?

- Depends on the method and the data
- Privacy leaks can still occur (e.g., due to overfitting)
- But: huge improvement over sharing raw data
- Can be enhanced with additional techniques
  - Differential privacy
  - Secure multiparty computation
  - Homomorphic encryption

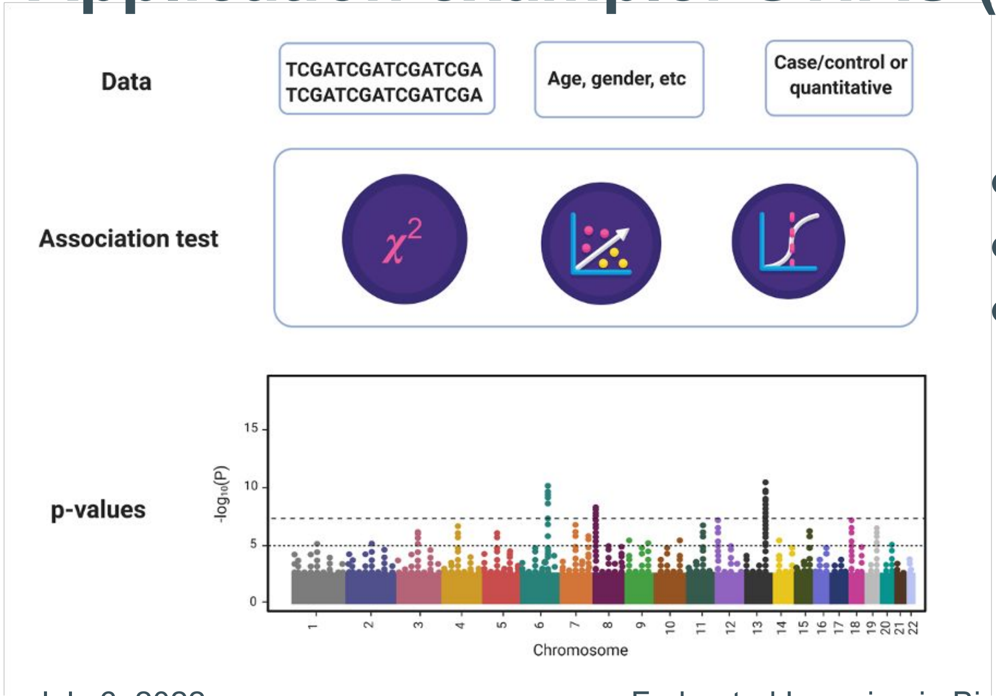
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Application example: GWAS

- Case/control studies on SNP data
- Each SNP examined individually (LR)
- Find SNPs related to diseases

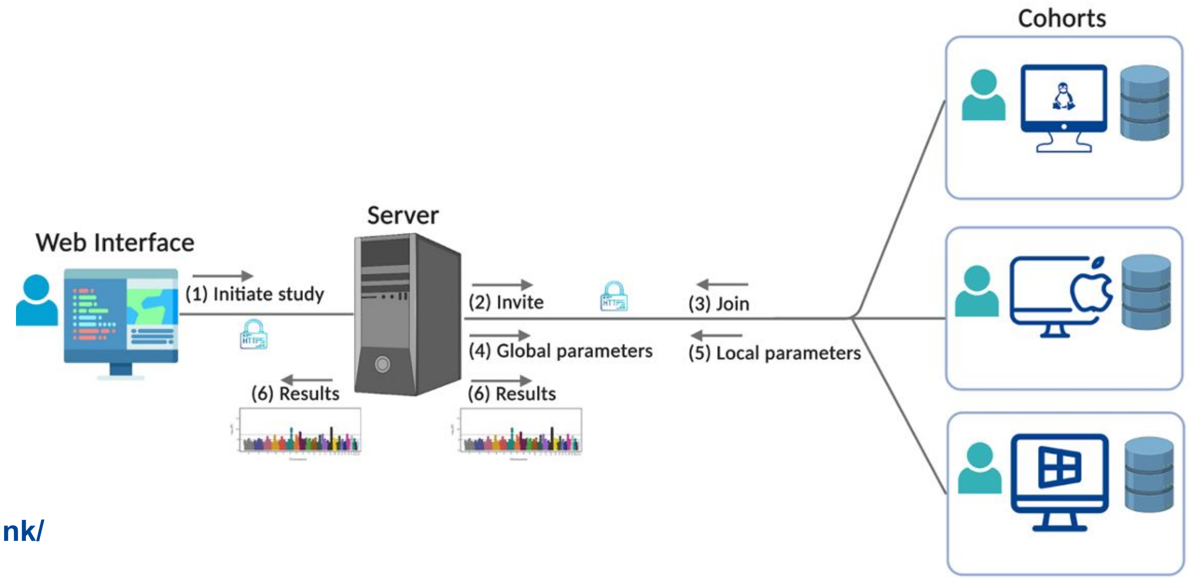


Application example: GWAS (Cont'd)



- Chi-square test
- Linear regression
- Logistic regression

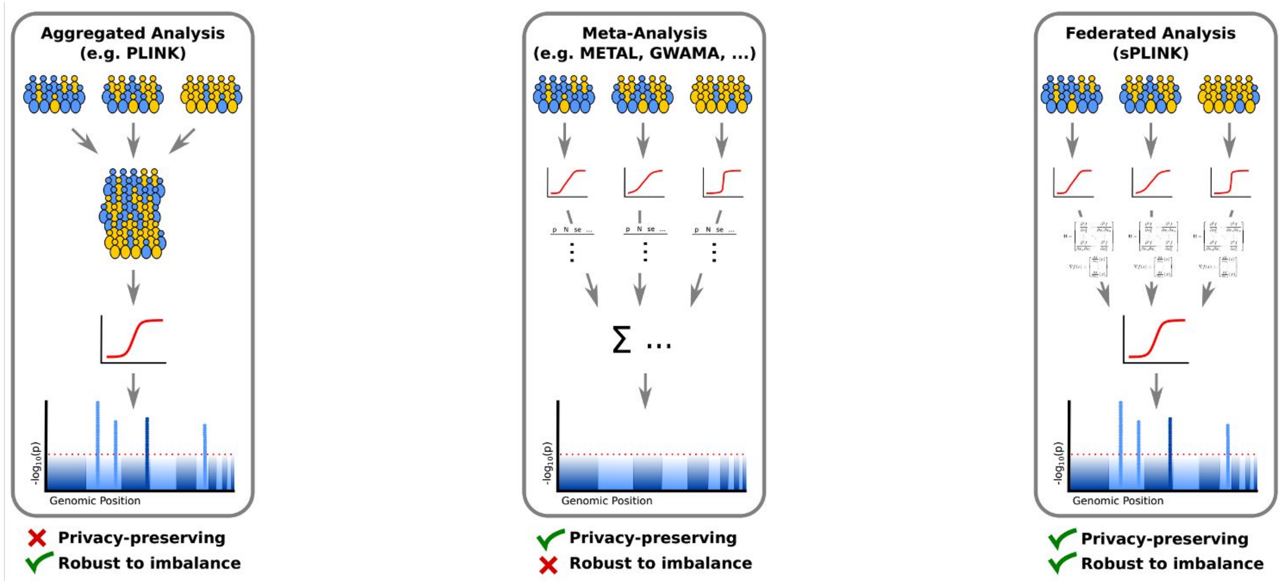
Application example: GWAS (Cont'd)



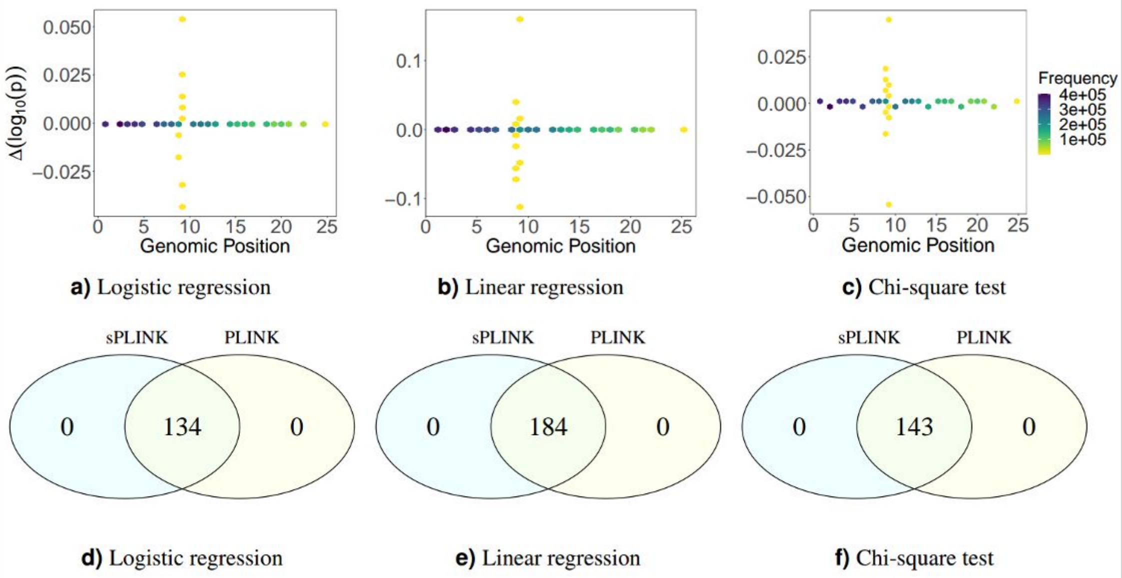
URL: **exbio.wzw.tum.de/splink/**

Reza Nasirigerdeh et al. (published in Genome Biology)

Application example: GWAS (Cont'd)



Application example: GWAS (Cont'd)



Other examples

Flimma

Differential expression analysis based on the limma voom pipeline

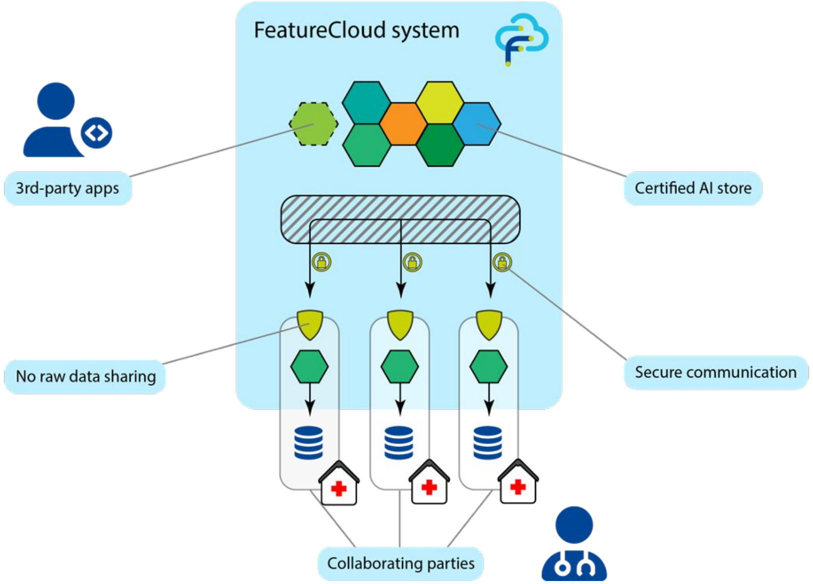
URL: **exbio.wzw.tum.de/flimma/** | *Olga Zolotareva et al.* (published in Genome Biology)

Partea

Time-to-event studies

URL: **exbio.wzw.tum.de/partea/** | *Julian Späth et al.*

Platform example: FeatureCloud

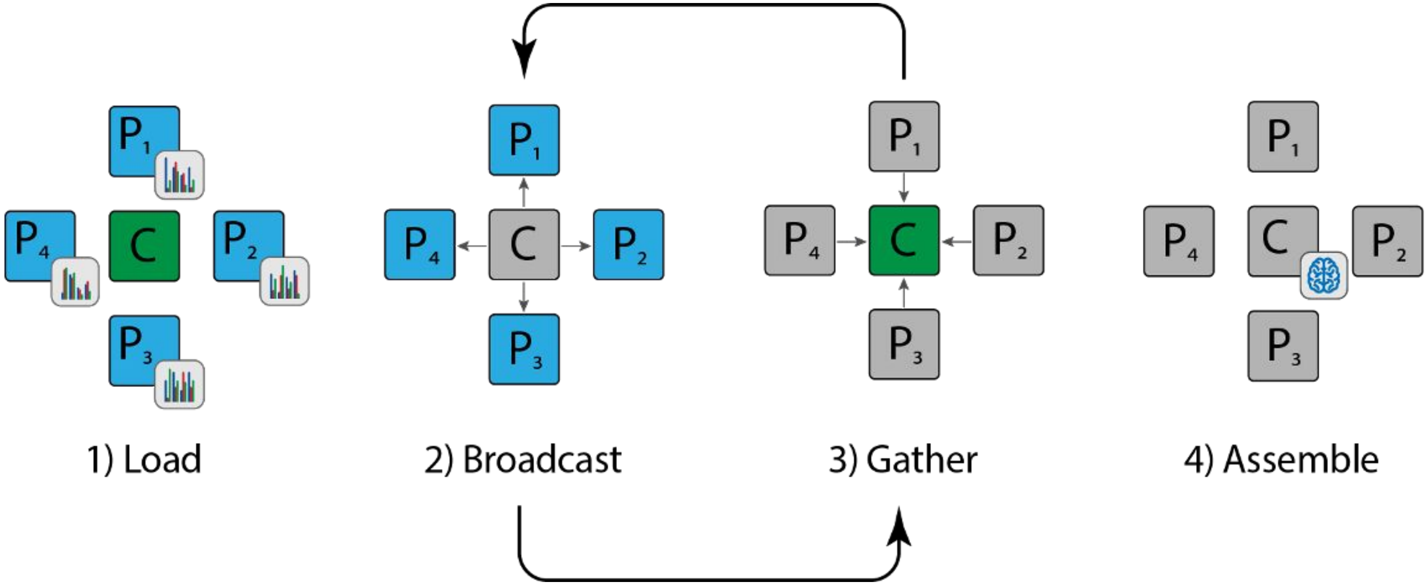


- Increase convenience for developers (including testing and deployment)
- Increase trust in apps (certification, consent management)
- Create a community (developers, researchers, patients)

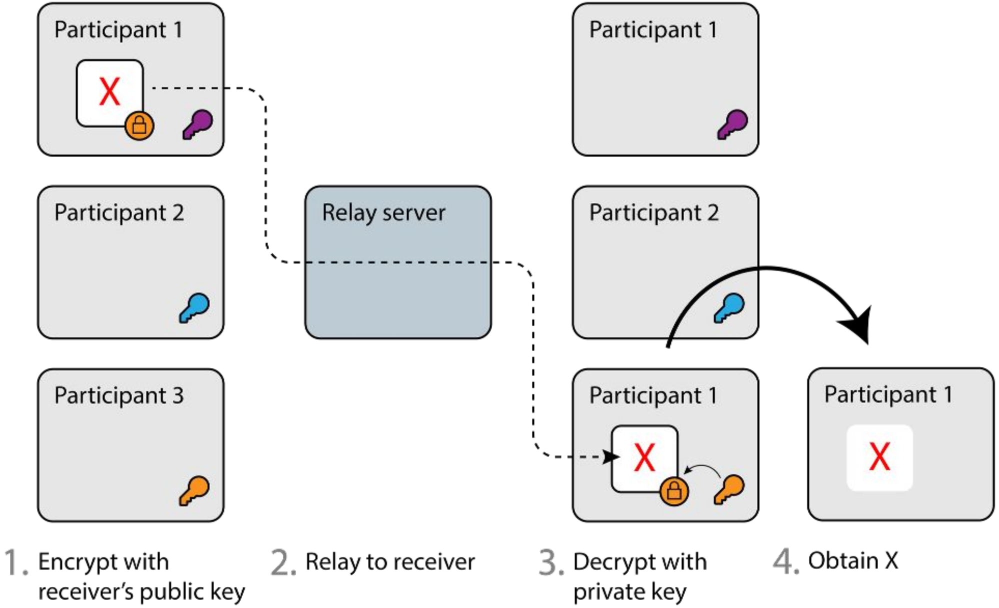
App communication

- Star-based federated learning
  - Gather-broadcast
- Peer-to-peer
- Secure multi-party computation

Star-based communication



Peer-to-peer communication



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Continuous vs. discrete models

- Continuous models can usually be federated ‘straightforward’
  - Gradient descent
  - Federated average
  - One or multiple interactions

- Discrete models are more difficult
  - Example: Decision tree
  - Possible solutions: Ensemble techniques

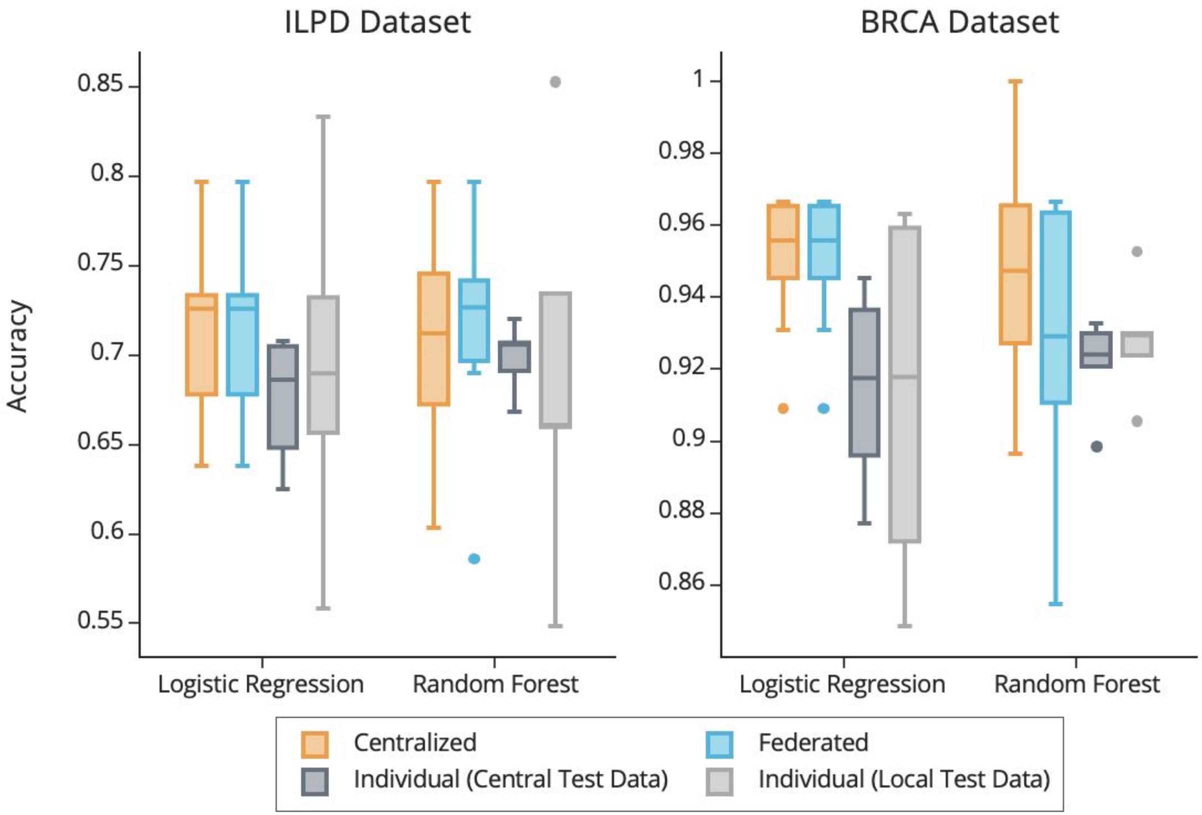
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Example: Random forest

- Train decision trees using bootstrapping on each site
- Transmit models and merge into big ensemble → Random Forest

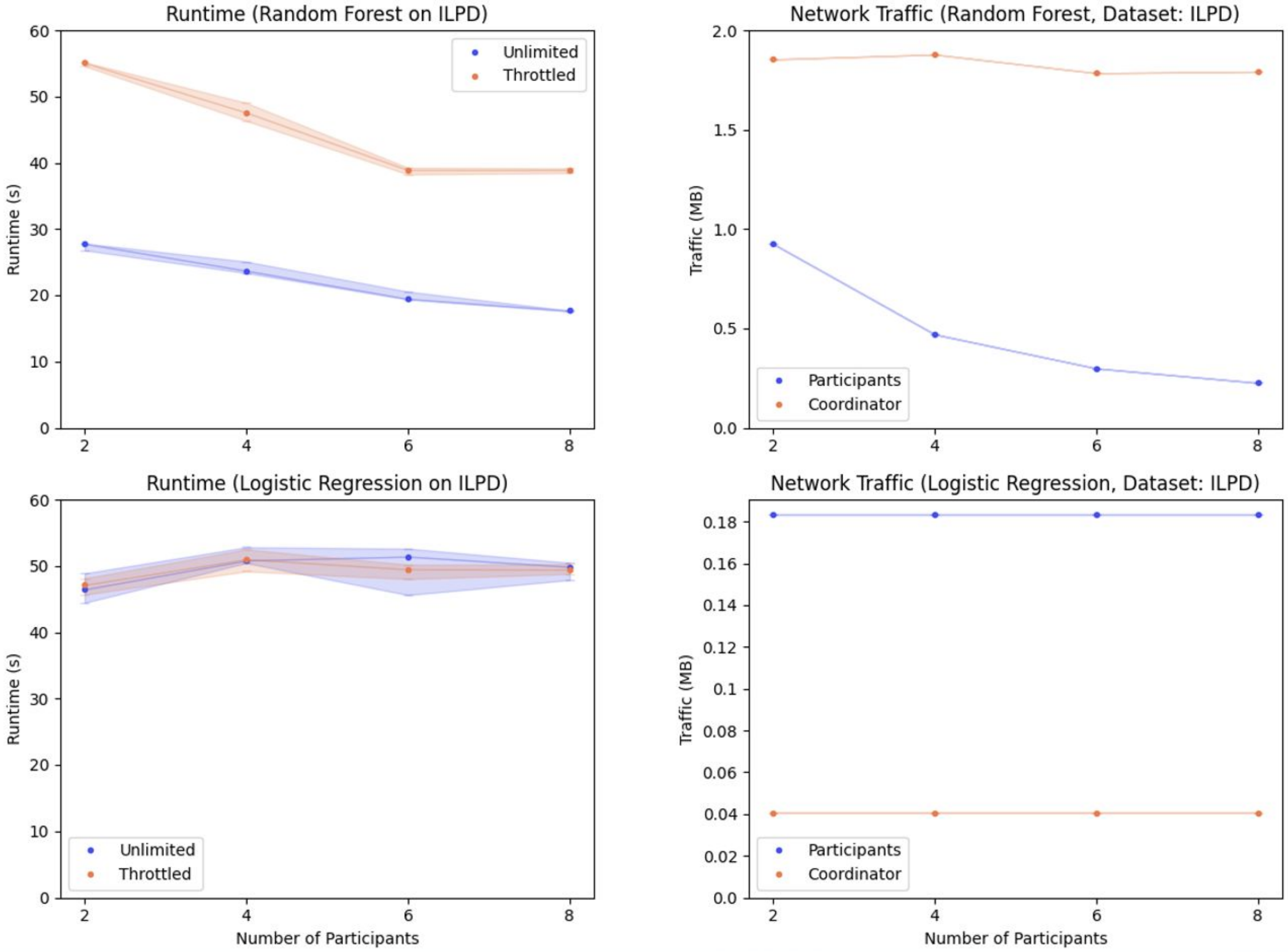
Federated Random Forests can improve local performance of predictive models for various healthcare applications, Anne-Christin Hauschild, Marta Lemanczyk, Julian Matschinske, et al., Bioinformatics, Volume 38, Issue 8, 15 April 2022, Pages 2278–2286

Evaluation



ILPD Dataset
BRCA Dataset
Accuracy
0.85
0.8
0.75
0.7
0.65
0.6
0.55
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
Logistic Regression
Random Forest
Logistic Regression
Random Forest
Centralized
Federated
Individual (Central Test Data)
Individual (Local Test Data)

Evaluation


Runtime (Random Forest on ILPD)
Runtime (s)
Unlimited
Throttled
Network Traffic (Random Forest, Dataset: ILPD)
Traffic (MB)
Participants
Coordinator
Runtime (Logistic Regression on ILPD)
Runtime (s)
Unlimited
Throttled
Number of Participants
Network Traffic (Logistic Regression, Dataset: ILPD)
Traffic (MB)
Participants
Coordinator
Number of Participants

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

ISMB 2022 - Tutorial - Federated Learning in Biomedicine

Thank you!