DATA INNOVATION LAB 2018

FINAL PRESENTATION

JULY 25, 2018

Federated Learning:

Collaborative Machine Learning without Centralized Training Data

TEAM

TUM-DI-LAB Head

Prof. Dr. Massimo Fornasier

Project Manager

Dr. Ricardo Acevedo Cabra

Mentors

Dr. Patrick Biermann
Dr. Markus Seifert
Dr. Ferdinand Graf
M.Sc. Todor Dobrikov

Students

Robin Fritsch
Shayoni Halder
Valentin Hartmann
Dmitrii Petukhov

Credit default modelling is a central challenge for financial institutes

SUBPRIME CRISIS

- Subprime Mortgage Crisis of 2007-2010
  - was the most severe recession in the last decade
  - resulted from high default rates on ‘subprime’

DEFAULT

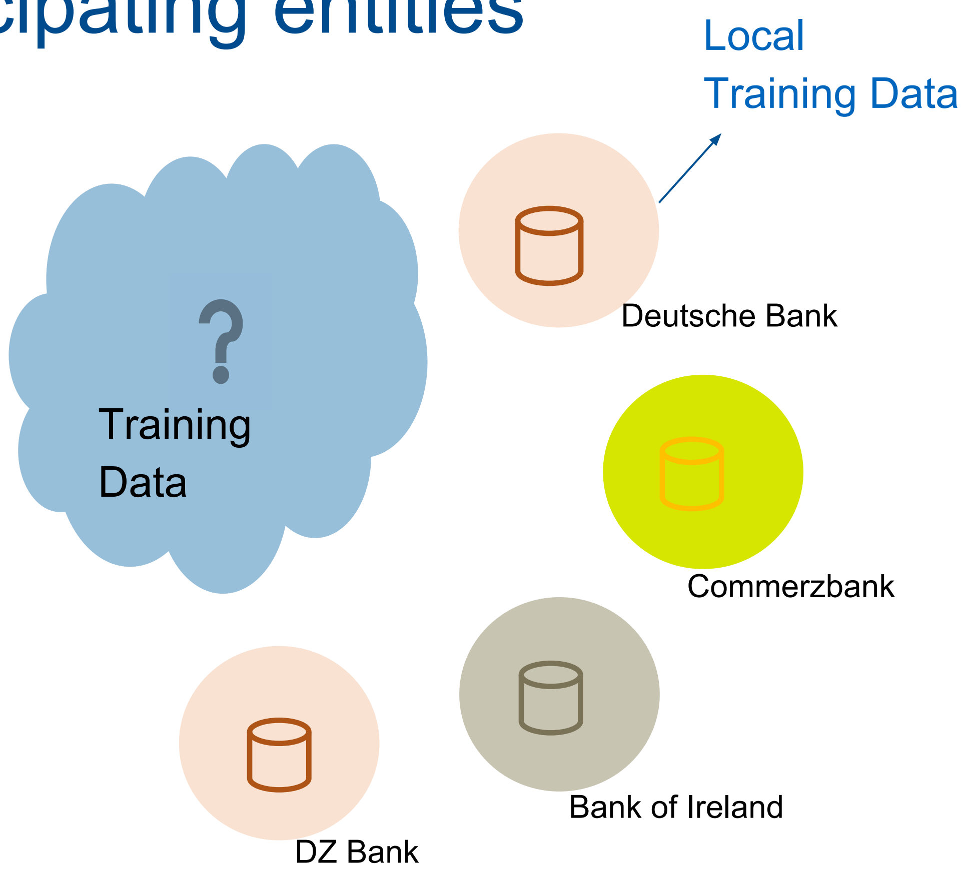
- Estimated probability of default (PD) is crucial
  - to calculate the interest rate and other credit conditions (e.g. collateral) for the obligor at contract agreement
  - for (regulatory) reporting

BETTER MODELS

- The growing complexity of the world calls for more sophisticated models, that can no longer be properly build on a bank’s own dataset

Collaboratively using data improves risk assessment of the participating entities

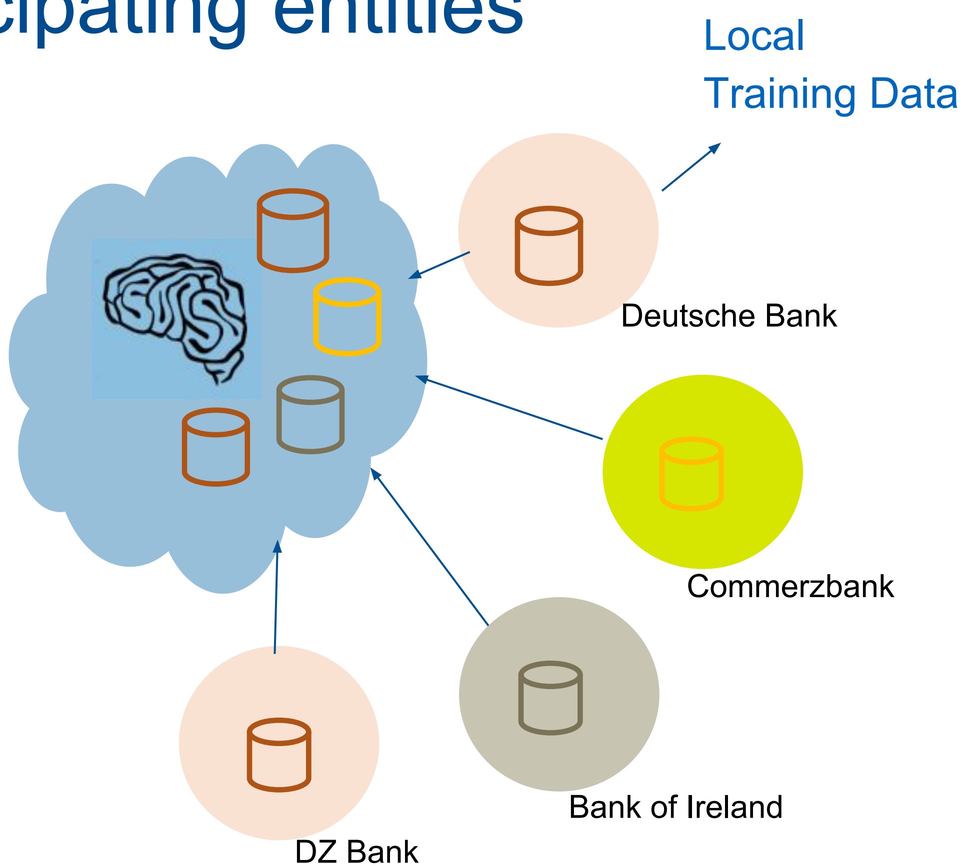
- Quality and quantity of data underpins the quality of complex models
- Individual institutions like banks might not be in possession of enough data
- Desirable that institutions share their data to build better models
  - Leads to higher performance
  - Reduces development cost
  - Reduces maintenance cost
  - Removes machine learning knowledge for building a model at every branch



Naïve approach of collecting data on central server and training an ML model on that

Collaboratively using data improves risk assessment of the participating entities

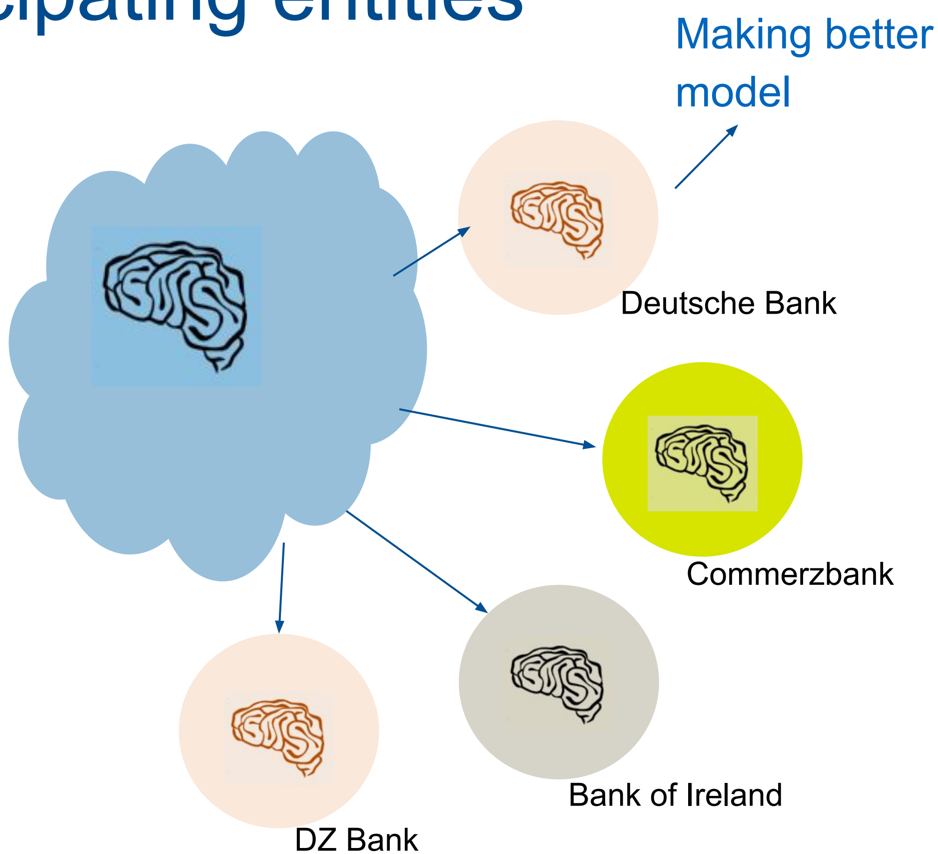
- Quality and quantity of data underpins the quality of complex models
- Individual institutions like banks might not be in possession of enough data
- Desirable that institutions share their data to build better models
    - Leads to higher performance
    - Reduces development cost
    - Reduces maintenance cost
    - Removes machine learning knowledge for building a model at every branch



Naïve approach of collecting data on central server and training an ML model on that

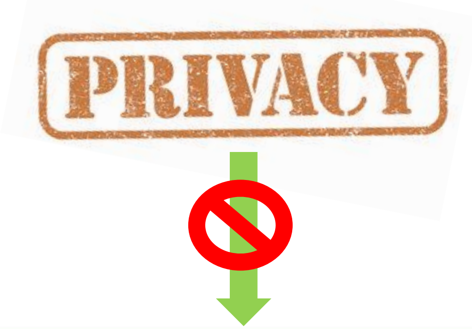
Collaboratively using data improves risk assessment of the participating entities

- Quality and quantity of data underpins the quality of complex models
- Individual institutions like banks might not be in possession of enough data
- Desirable that institutions share their data to build better models
    - Leads to higher performance
    - Reduces development cost
    - Reduces maintenance cost
    - Removes machine learning knowledge for building a model at every branch



Naïve approach of collecting data on central server and training an ML model on that

Privacy constraints prevent banks from sharing their data

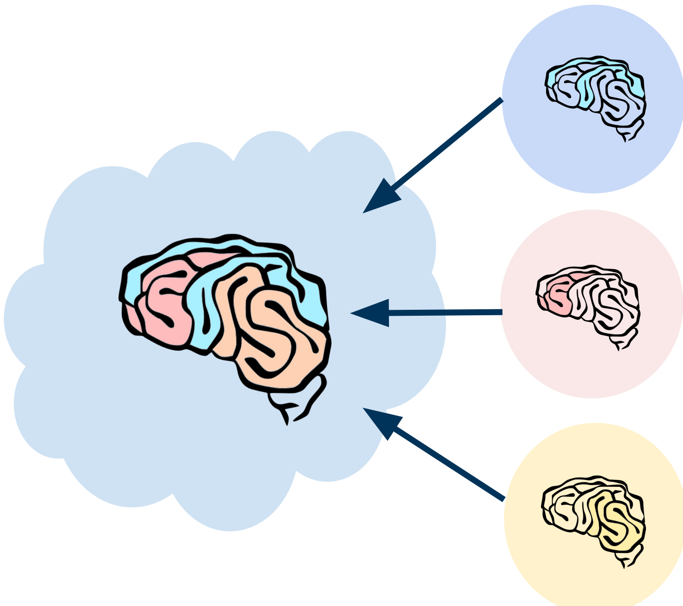


Privacy and Transparency Regulations create hindrance!

- Countries have laws and regulations to protect personal data
- Institutes need to comply them to ensure privacy
- Naïve approach of data collection no longer works

Federated Learning comes to the rescue!

You do not need to share the data to collaboratively learn a complex model

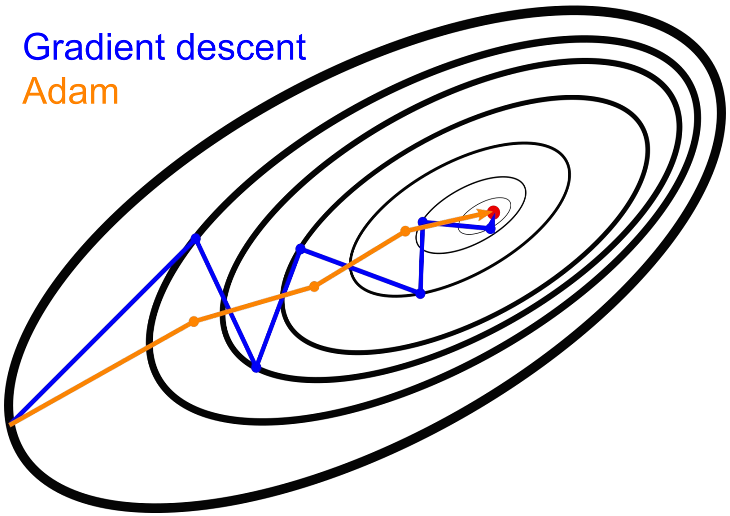


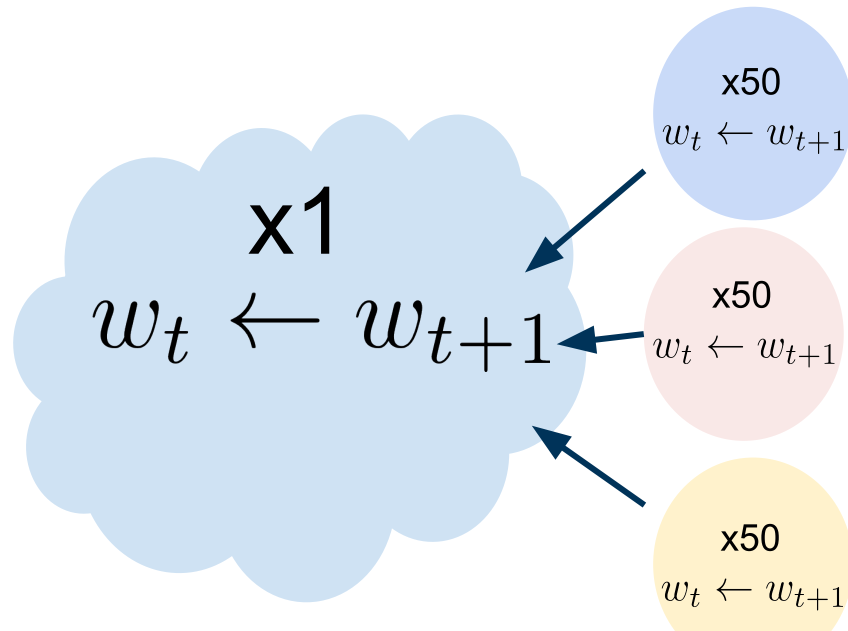
- privacy aspect: only sending gradients is already a lot better than sending raw data
- can be used for: logistic regression, neural networks, SVMs

$$w_{t+1} = w_t - \gamma_t \frac{1}{n} \sum_{i=1}^{n} \nabla L(x_i, w_t)$$

Fig. credit: Jakub Konečný, Google

The fewer communication rounds, the less information gets revealed

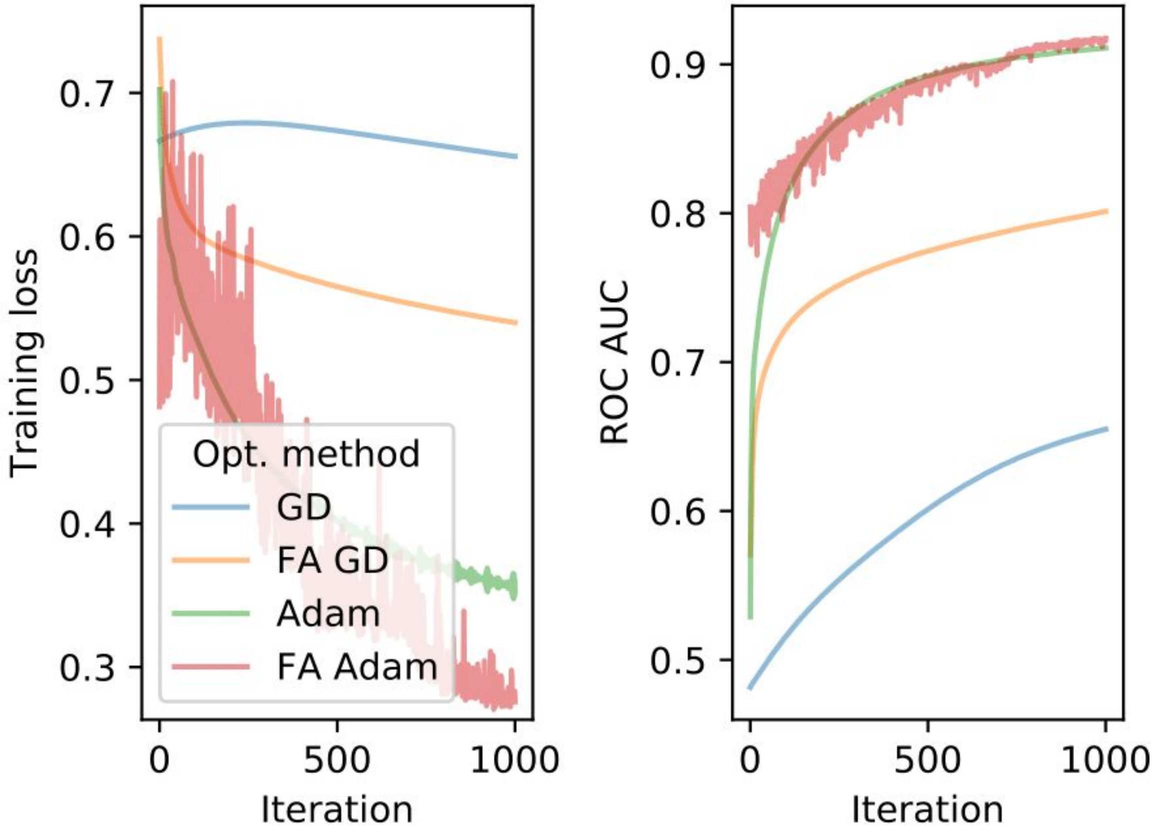




Federated Averaging

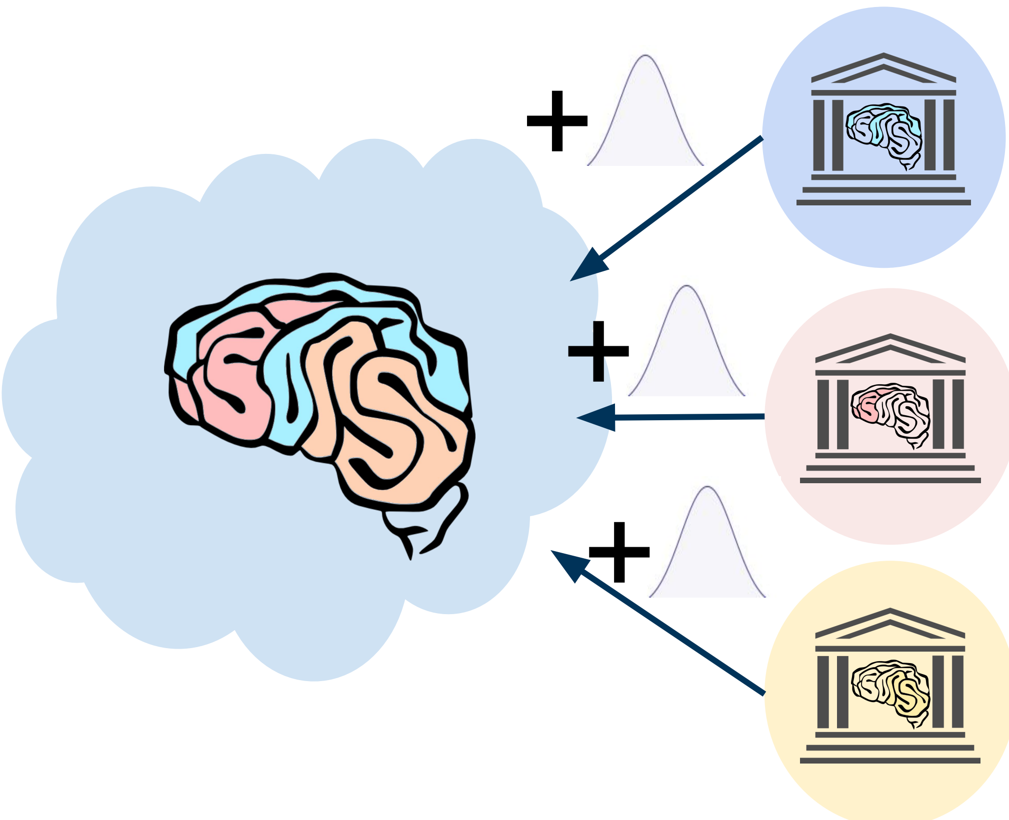
- Adam: reduce oscillations, make bigger steps
- Federated Averaging: perform multiple steps on each client before sending an update

The fewer communication rounds, the less information gets revealed



- significantly faster convergence with Adam than with GD
- Federated Averaging helps, but not too much when already using Adam

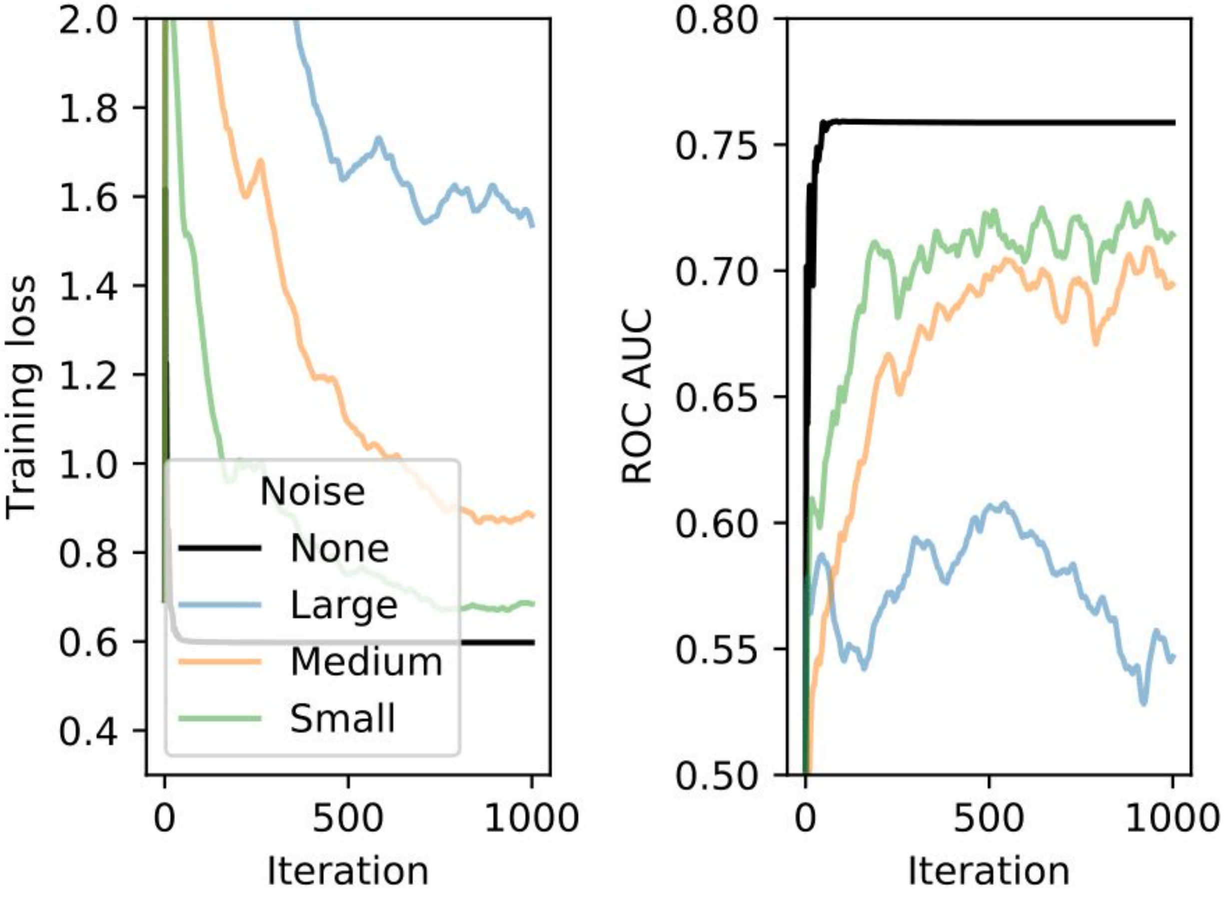
Guarantee privacy by adding noise



- Add noise to model updates before sending them to the server
- Allows for mathematical privacy guarantees

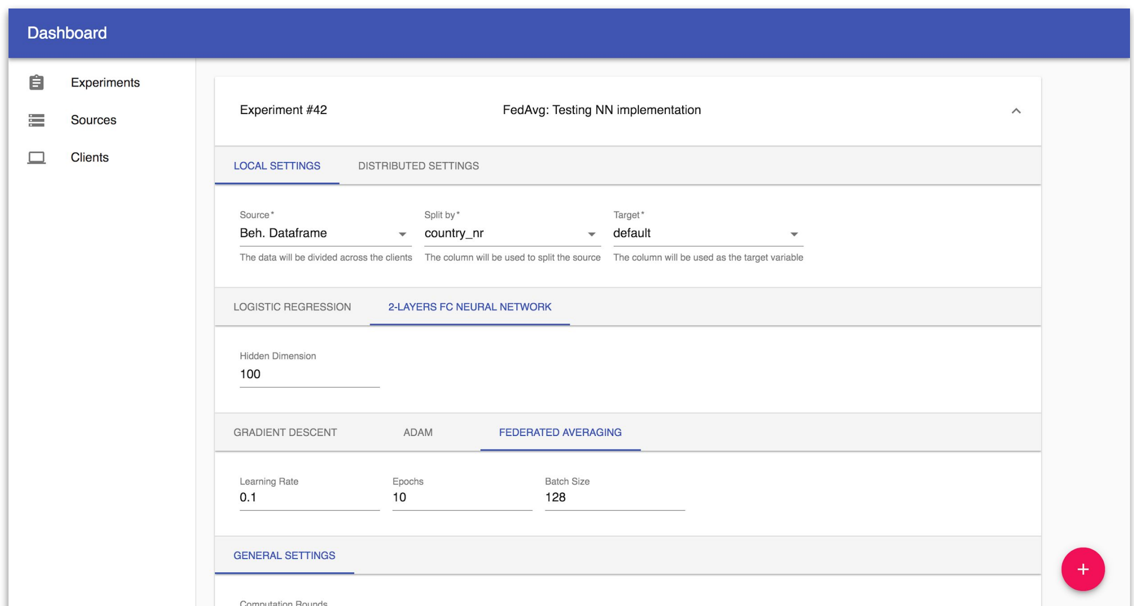
Guarantee privacy by adding noise

- Privacy comes at the expense of model performance
- Still reasonable performance for moderate amounts of noise

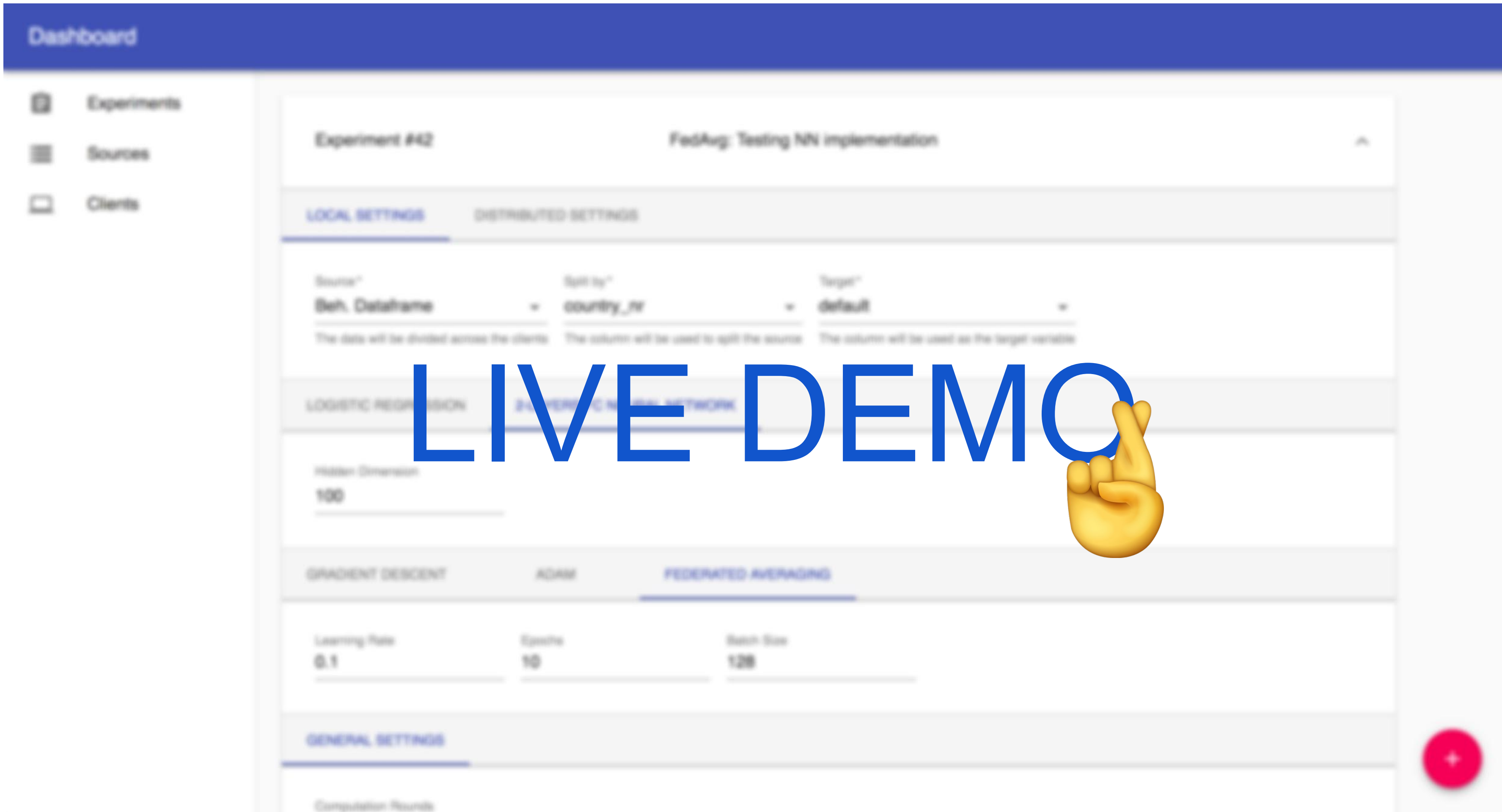


A prototype enables rapid experimentation

- allows to run distributed and local experiments
- user can change model type, optimization algorithm, etc.
- experiments saved to the cloud and available to everyone for review
- draws model performance online
- developed using latest technologies
- customizable



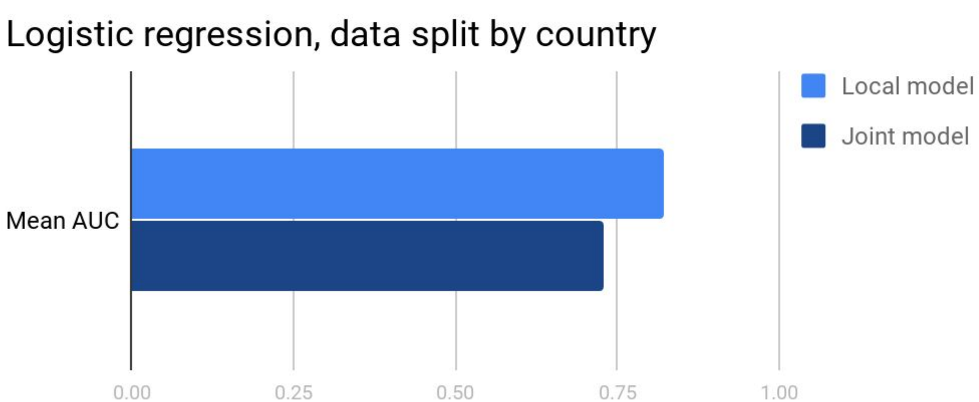
A prototype enables rapid experimentation

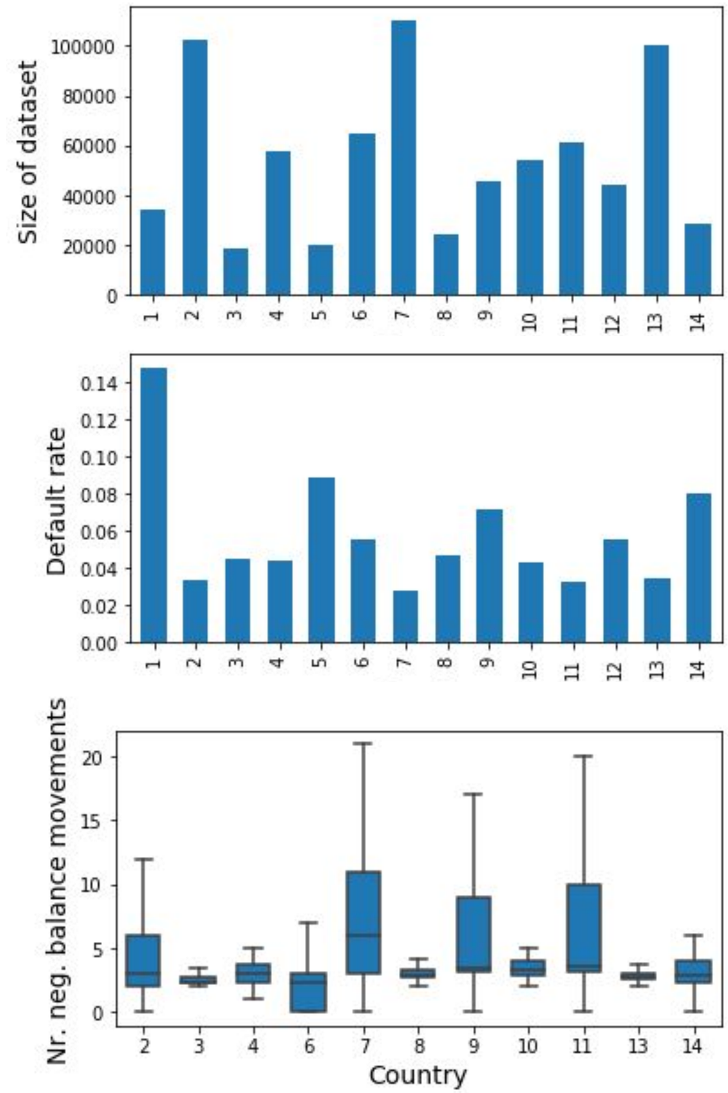


Screen Recording: https://youtu.be/O0QZgm1RhiY

Testing the limits of Federated Learning

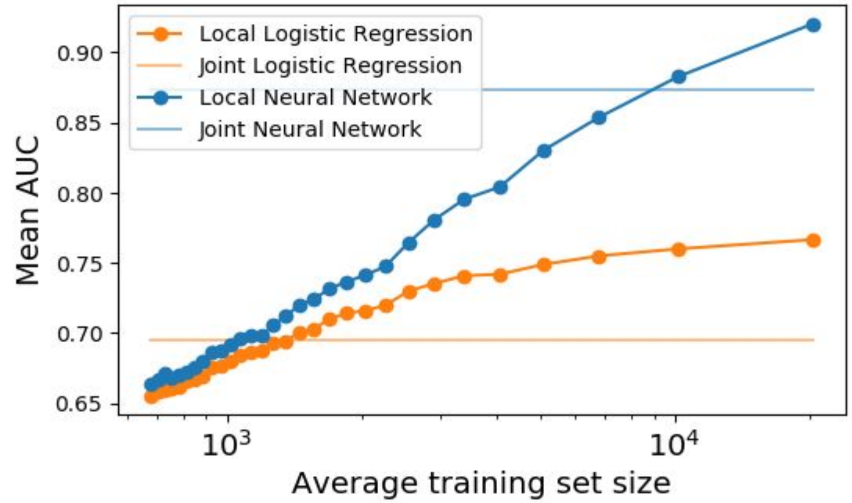
- Joint model has more data while local models have more free parameters
- Local datasets can be unbalanced and strongly vary in distribution
- Local models can specialize on local datasets and perform better



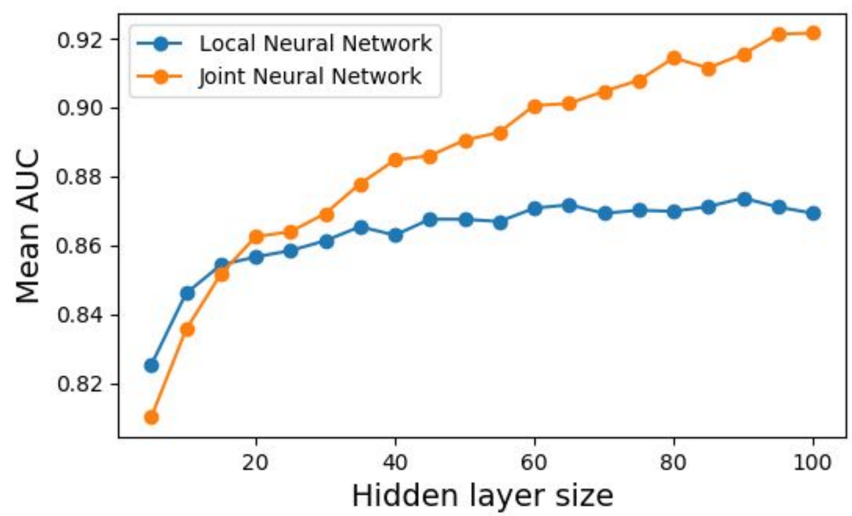


Model performance is improved for small local datasets and complex models

- When local datasets are small the local models do not have sufficient data to train



- When a more complex model is used the joint model can better make use of the larger amount of training data



Federated Learning enables banks to collaboratively improve their risk models without compromising data privacy