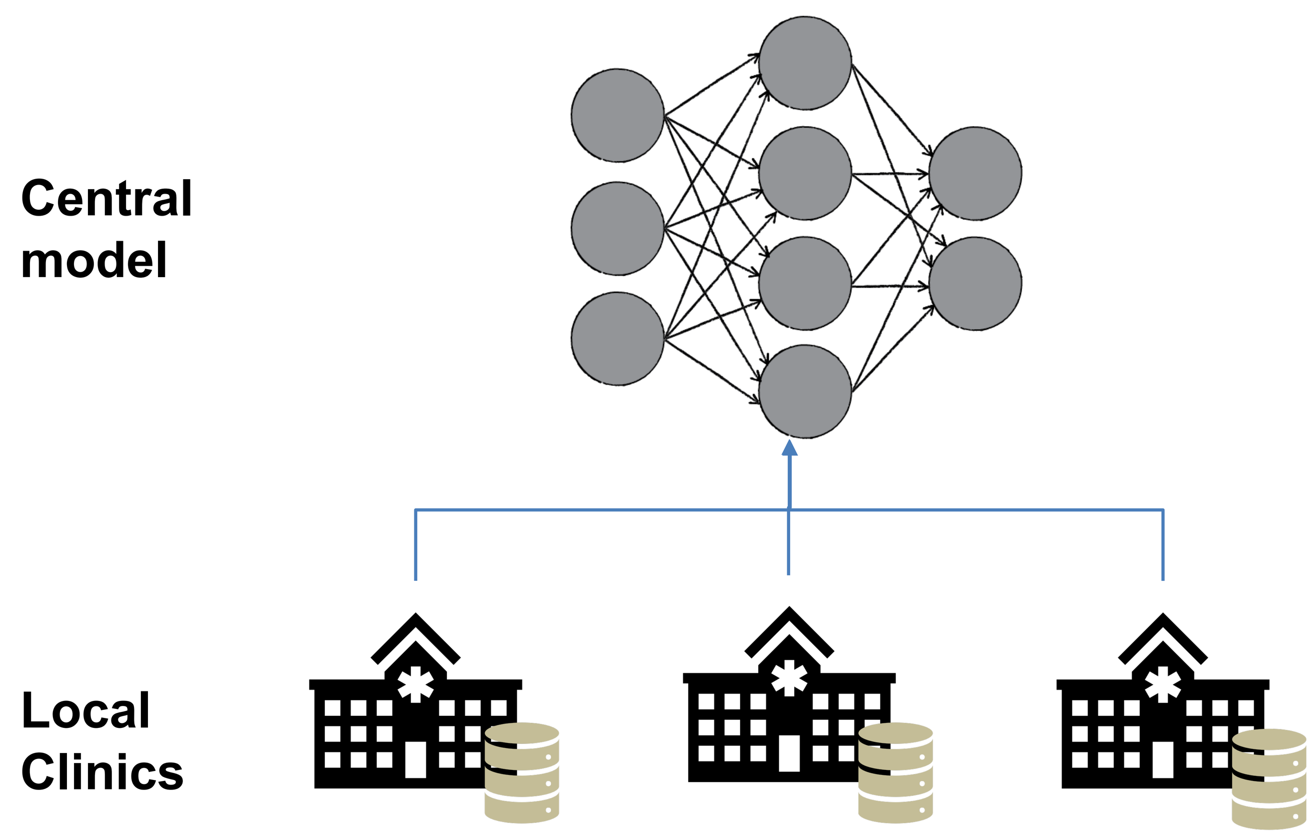
Federated Learning for Smart Radiotherapy Systems

Dominik Fay
KTH (Division of Decision and Control Systems), Elekta

We want to train models on distributed medical data.

Machine learning in the medical domain comes with a number of challenges, such as:

- **Privacy**: Raw data resides in silos. Only aggregated data can be transferred to a central server. But even aggregated data such as model updates can be used to **reconstruct training data**.
- **Communication efficiency**: Deep neural networks need substantial **network bandwidth** for distributed training. This can be challenging for hospital data centers.
- **Heterogeneity**: In machine learning, examples are often assumed to be independent and identically distributed (i.i.d.) This is often violated in practice, e.g., due to **distributional differences** between local data sources.



Noisy Gradient Descent

Many machine learning problems can be formulated in the framework of differentially private empirical risk minimization (**DP-ERM**). Here, we want to minimize an average loss subject to a privacy constraint.

$$F(\theta) = \frac{1}{N}\sum_{n} f(\theta; x_n)$$

Privacy constraint: Differential privacy (DP) guarantees that similar training datasets lead to similar models, in a probabilistic sense. It limits the ability of an adversary to identify a dataset by observing the trained model. This is usually achieved by adding noise to the gradient:

$$\theta_{t+1} = \theta_t - \eta_t \left(\nabla F(\theta_t) + \zeta_t\right), \quad \zeta_t \sim \mathcal{N}(0, \sigma_t^2 I)$$

Problem: Noisy Gradient Descent is very sensitive to the **hyperparameters** (noise variance, step size, etc.). Tuning these parameters manually is inefficient and adds an extra privacy cost to the algorithm.

Instead, we consider a hyperparameter selection rule based on optimizing the **privacy-utility ratio** (PUR) at each iteration:

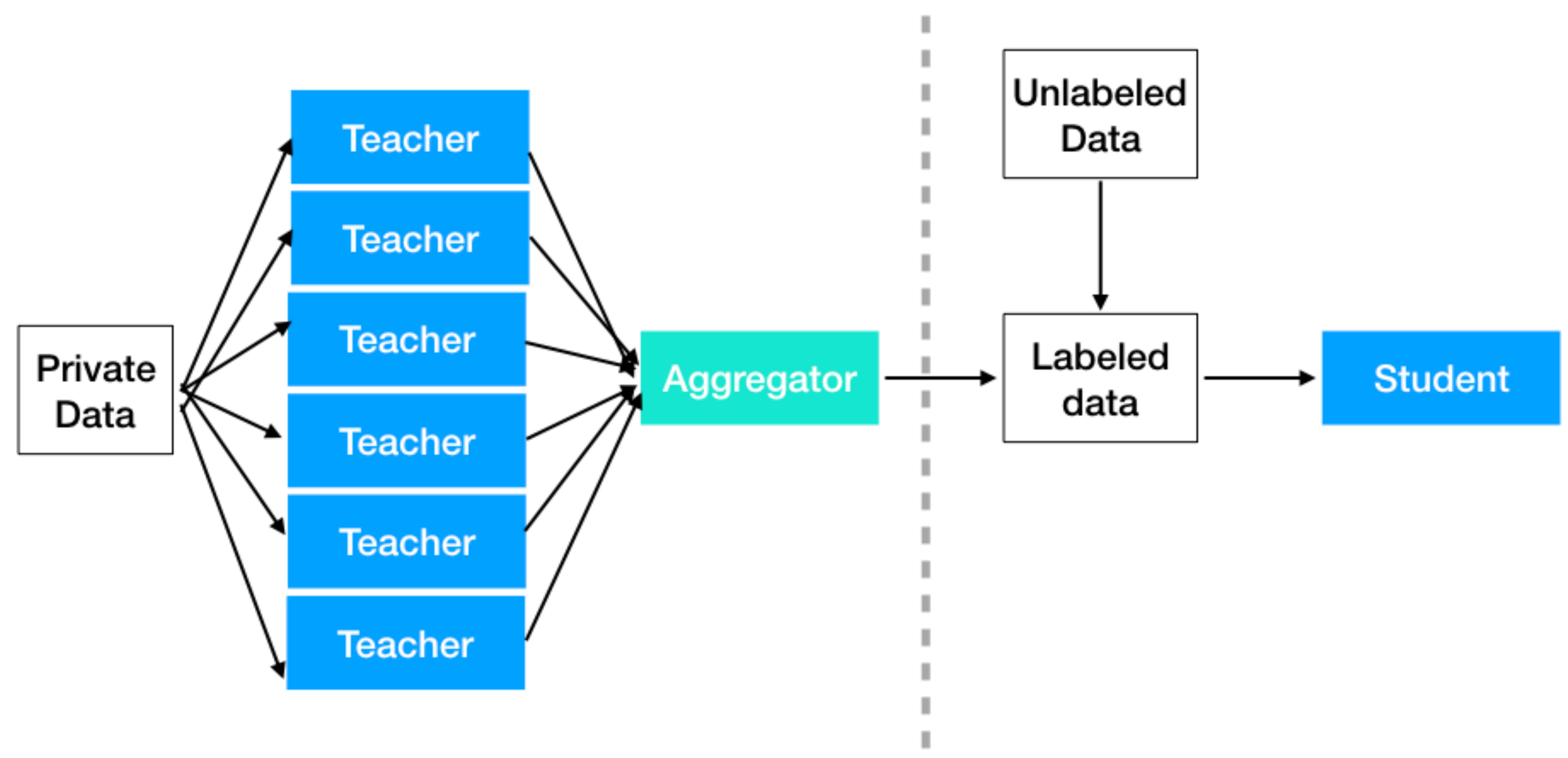
$$\text{minimize} \quad \text{PUR}(\sigma_t^2, \eta_t) \overset{\text{def}}{=} \frac{\text{Privacy}(\sigma_t^2)}{\text{Utility}(\sigma_t^2, \eta_t)}$$

Our results

- The PUR-optimal hyperparameters lead to a **constant signal-to-noise ratio**, and a constant step size.
- According to PUR, more privacy budget should be allocated to later iterations. when the gradient is smallest.
- Empirically, the selection rule performs well across a wide range of datasets. It often outperforms the best constant noise variance known **in hindsight**.

Full paper

Privacy-Aware Ensembles



Background: Private Aggregation of Teacher Ensembles (PATE) can be used to **merge locally trained models** into a privacy-preserving central model. The predictions of local models are aggregated by **noisy majority voting**

$$\tilde{Y}_{Ensemble} = \text{Count}(Y_1, \ldots, Y_K) + \mathcal{N}(0, \sigma^2 I)$$

Problem: Majority voting is only suited for single-dimensional classification tasks. How do we deal with **high-dimensional** tasks, such as medical image segmentation? Suggestion: **Dimensionality reduction**

$$\tilde{Y}_{Ensemble} = \text{Decode}\left(\frac{1}{K}\sum_{k} \text{Encode}(Y_k) + \mathcal{N}(0, \sigma^2 I)\right)$$

Our results

- Out of the box, PATE does not perform well on MRI tumor segmentation data.
- Dimensionality reduction can vastly improve PATE's performance on high-dimensional tasks.
- For PCA, we get closed-form expressions for the optimal compression rate and mean-squared error.

Full paper

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM