Abstract: Machine Learning (ML) applications have a tremendous impact on our everyday life. However, we pay a high price for this (mostly) comfort, as our sensitive data is collected and exploited at scale. This striking imbalance between data ownership and control is partially grounded in the fact that traditional training of ML models requires aggregation of all training data in a central data silo. In contrast, the decentralized training paradigm offers the users more control over their own data by collaboratively training ML models directly on edge devices without the need to send private data to a central coordinator first. However, there are many technological challenges to address before such systems become efficient enough to be deployed at scale.

In this talk we focus on communication efficient ML algorithms for edge devices. We discuss two key methods to reduce communication costs: firstly, compression with error feedback, and secondly, intermittent communication (local updates). We show how these techniques lead to state-of-the-art optimization algorithms for distributed, federated and decentralized network topologies and we further show how all these methods can be cast and analyzed in a unifying theoretical framework.

At the end of the talk we will highlight and discuss major challenges that are left on the way towards more open and user-friendly machine learning in the future.