Privacy-Preserving Machine Learning as a Service: Challenges and Architectures in Cloud Environments

Thomas Schuster

April 2025

Abstract

Cloud-based Machine Learning (ML) services promise scalability and accessibility, but also raise challenges regarding data privacy, and environmental sustainability. This talk explores architectural patterns and strategies for enabling privacy-preserving ML-as-a-Service (MLaaS). Key approaches discussed include secure model execution within controlled environments, the architectural design of hybrid AI services, and frameworks for systematically benchmarking utility-privacy trade-offs. I will highlight inherent tensions in designing such systems and outline a research agenda focused on building trustworthy, quantifiable, and sustainable AI services in cloud environments.

Motivation and Context

With increasing availability of Machine Learning services, a broad range of organizations are integrating AI capabilities into their operations. However, this transition brings about new risks: sensitive data may be exposed, regulatory constraints (such as the GDPR and AI Act) must be met, and cloud-based AI systems often come with substantial resource consumption. Designing MLaaS that are effective and scalable, while also being verifiably privacy-preserving and sustainable, requires careful consideration of deployment architectures and inherent trade-offs. Standard public cloud offerings may not always suffice, necessitating exploration of controlled execution environments.

Scientific Contribution

In the presentation, I will introduce core architectural patterns and research approaches that address the interplay of cloud infrastructure, AI model deployment, and privacy protection. These include:

- Strategies for **securing ML model execution within controlled environments** (e.g., private cloud, on-premise) as an alternative to public APIs.
- Architectural designs for **hybrid AI services**, where sensitive data is processed locally while leveraging the cloud for compute-intensive tasks.
- Frameworks for benchmarking the **utility-privacy trade-offs** inherent in different secure MLaaS architectures.

These approaches and their inherent trade-offs will be illustrated through selected use cases from my recent work, such as the secure deployment of large language models and the design of privacy-first machine learning services. These case studies demonstrate how privacy patterns and benchmarking insights can be combined to build transparent, scalable, and trustworthy MLaaS offerings while managing the balance between utility and privacy.

Research Agenda and Outlook

The talk will conclude with an outlook on future research directions, aligned with the mission of the Cloud Engineering Professorship:

- Developing compliance-aware AI services that include automated documentation and risk assessment.
- Improving the energy efficiency and transparency of AI workloads in cloud environments.
- Designing developer tools and abstractions for implementing privacy by design.

This talk aims to contribute to the scientific discourse on how cloud-based AI systems can be built responsibly and efficiently. This also offers a research perspective that also complements and advances teaching and applied research at Technische Hochschule Ulm.