LEAKPRO: Leakage Profiling and Risk Oversight for Machine Learning Models

Project time-line
Nov 2023 – Dec 2025

Total budget
18 million SEK

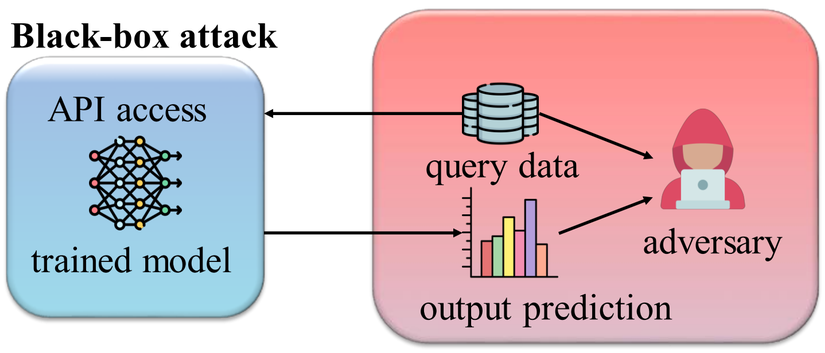
Related projects:
HEIDA, SARDIN, HDIP

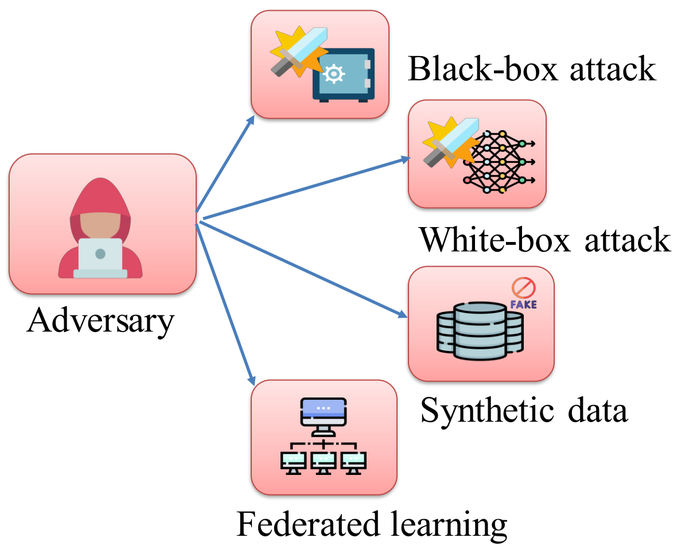
Project summary: Many recent works have highlighted the possibility of extracting data from trained machine-learning models, which may infringe on the privacy of the individuals that contributed data. However, these examples are typically performed under idealistic conditions, and it is unclear if the risk prevails under more realistic assumptions. In this project, we will create LeakPro, an open-source platform to assess the information leakage of i) trained machine learning models, ii) federated learning, and iii) synthetic data.

Contact: Johan Östman (project coordination by AI Sweden), Rickard Brännvall rickard.brannvall@ri.se and Henrik Forsgren henrik.forsgren@ri.se (RISE)

What is the risk that an AI model leak information?

A privacy attack exploits vulnerabilities in ML models to compromise the confidentiality of the underlying training data, for example, by querying a web-based API.



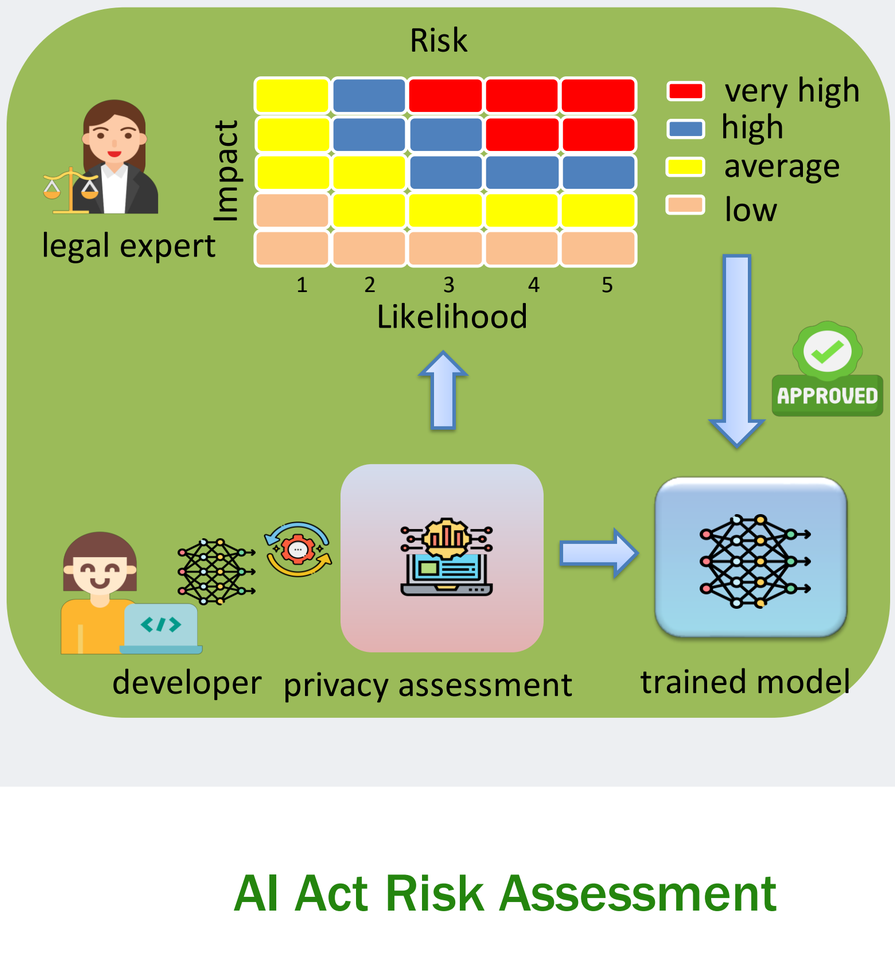


LeakPro models different privacy attacks and adversary objectives:

Black-box access **White-box access**	**Membership inference:** Specific datapoint used during training **Data reconstruction:** Approximate training data **Property inference:** Learn aggregate statistics from training data
Federated learning	**Gradient inversion**: Approximate training data of other parties.
Synthetic data	**Singling out:** Identify records corresponding to real individuals. **Linkability:** Correlate different data sources to identify individuals. **Inference:** Infer unknown attributes for real individual.

LeakPro is an open tool to systematically assess information leakage of trained ML models. It is built for both developers and legal experts to aid in reasoning around governance and privacy.

1. Open-source platform
2. Different data modalities
3. Realistic adversaries
4. Systematic risk reporting



Projectpartners

VÄSTRA GÖTALANDSREGIONEN
SAHLGRENSKA UNIVERSITETSSJUKHUSET
SYNDATA
AI SWEDEN
AstraZeneca
RI. SE
Region Halland

ML Model Privacy

AI Act Risk Assessment

Data Reconstruction