Let's Go, Private! Towards a Privacy-Preserving and Distributed Machine Learning System

Cláudia Brito
HASLab, INESC TEC & University of Minho
claudia.v.brito@inesctec.pt

Supervisors: **João Paulo** (INESC TEC & UMinho) and **Pedro Ferreira** (INESC TEC & Faculty of Sciences, UPorto)

Motivation

- Large amounts of sensitive data generated
- Use of AI techniques to extract valuable insights
- Regulations to avoid the misuse of sensitive data
- Several attacks over the ML workflow

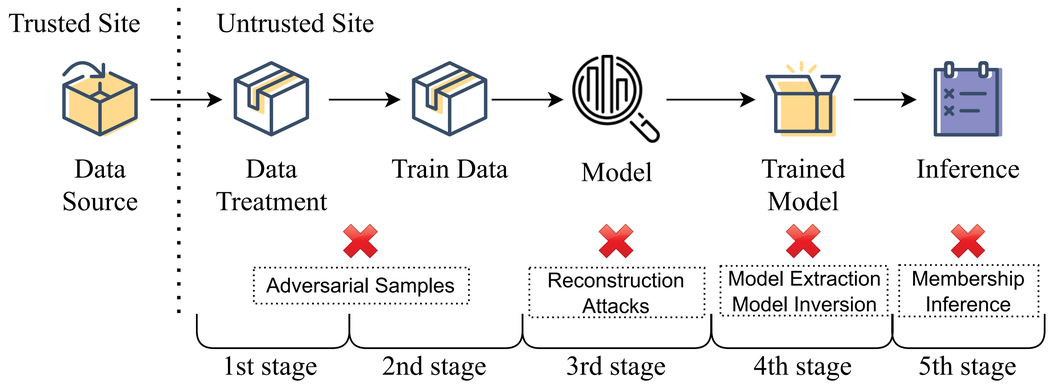

Figure 1. Machine learning pipeline and attacks defined within our security and threat model.

Limitations

- Untrusted third-party infrastructures
- Common cryptographic schemes impose impractical overheads
- TEEs' performance decreases with the increase of computations, I/O operations and, the trusted computing base (TCB)

Challenges

- ML datasets and models are stored and processed in plaintext
- Reducing the code base running inside enclaves
- Reducing the number of operations to be performed inside the enclave

Design

- Based on a large scale and distributed ML framework – Apache Spark's **MLlib** – and Intel's **SGX**
- Novel design based on the **partitioning of the code base**
- **Statistical functions** (e.g., confidence results) are **decoupled** to run outside the enclaves
- **Lower TCB** and performance overhead

TwoFold Design

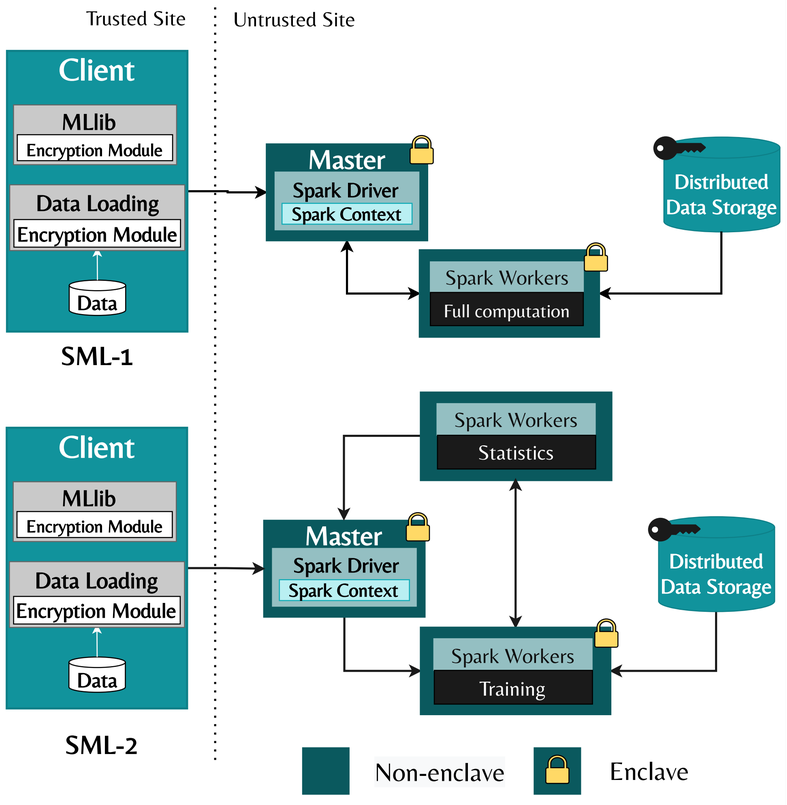

Figure 2. Architecture design of the proposed solution.

Goal

Design an **end-to-end privacy-preserving and distributed machine learning framework**

- Private large scale machine learning and data analysis
- Clients should trust third-party infrastructures while knowing that the computation performed over these will **not reveal** any sensitive information

Results

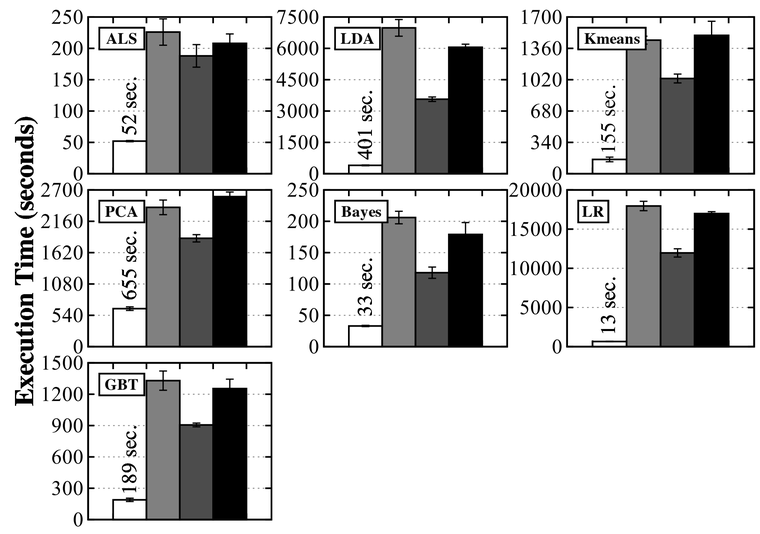

Figure 3. Execution time for each algorithm with *Huge* workload. As follows: Vanilla Spark; SML-1; SML-2; SGX-Spark.

- SML-2 surpasses similar state-of-the-art solutions
- SML-2 offers similar security guarantees when compared to SML-1 while outperforming it
- Performance overhead ranges from 1.7x to 23.8x when compared to Vanilla Spark

Future Work

- Focus on white-box access attacks
- Increase security measures with focus on ORAM and Differential Privacy
- Real data use cases with a focus on genomic data

This work was supported by the Portuguese Foundation for Science and Technology through a Ph.D. Fellowship (SFRH/BD/146528/2019) and by the ERDF - European Regional Development Fund, through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme under the Portugal 2020 Partnership Agreement, and by National Funds through the FCT - Portuguese Foundation for Science and Technology, I.P. on the scope of the CMU Portugal Program within project AIDA, with reference POCI-01-0247-FEDER-045907.