Privacy-Preserving Machine Learning for Apache Spark

Cláudia Brito

claudia.v.brito@inesctec.pt

INESC TEC & University of Minho

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

ENSD'22

Limitations

- Cloud environments lack security guarantees
- 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

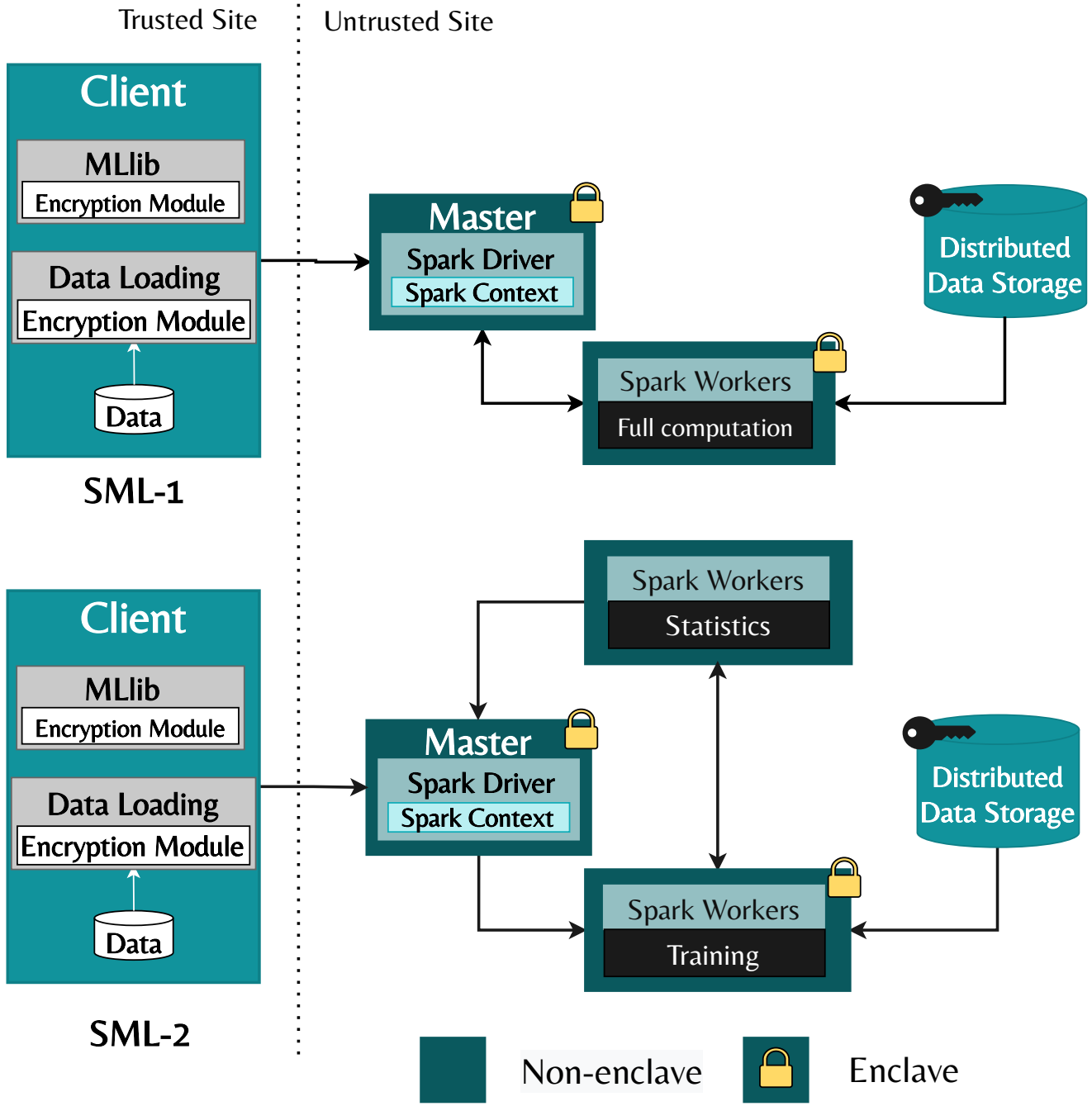
White-box vs black-box attacks

Goal

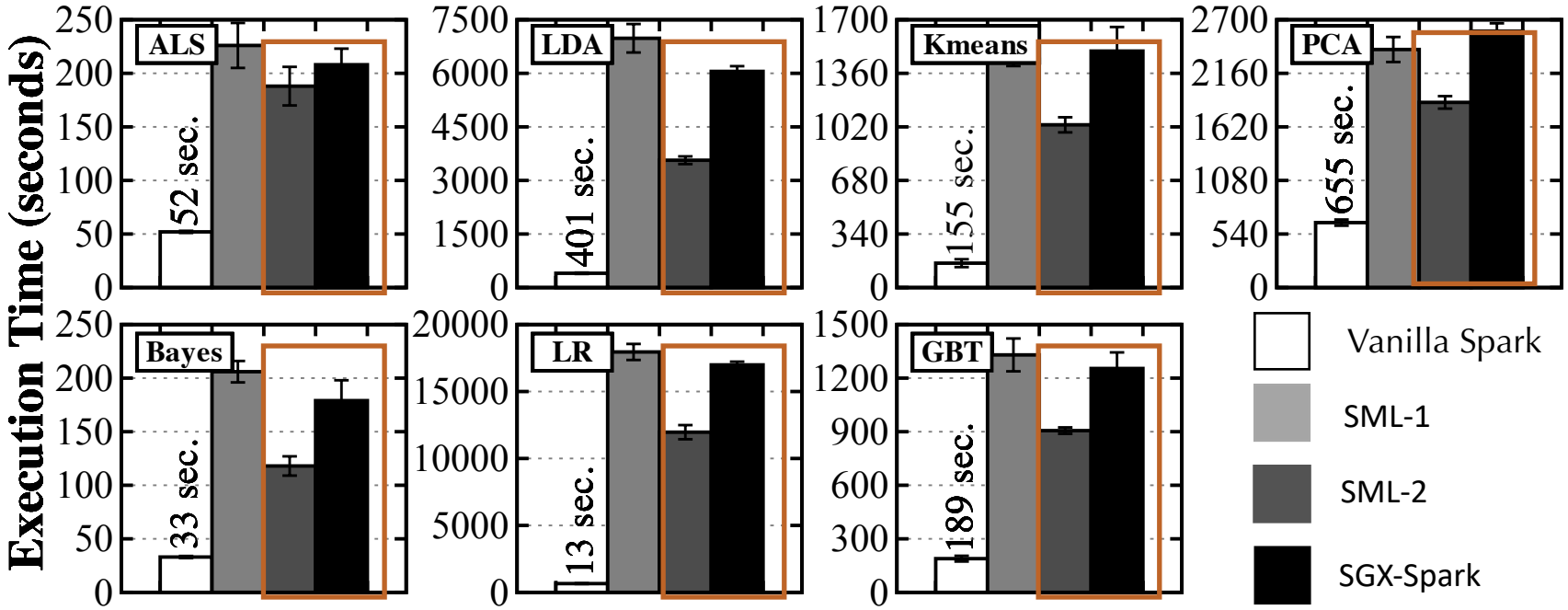
Design an **end-to-end privacy-preserving and distributed machine learning framework**

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

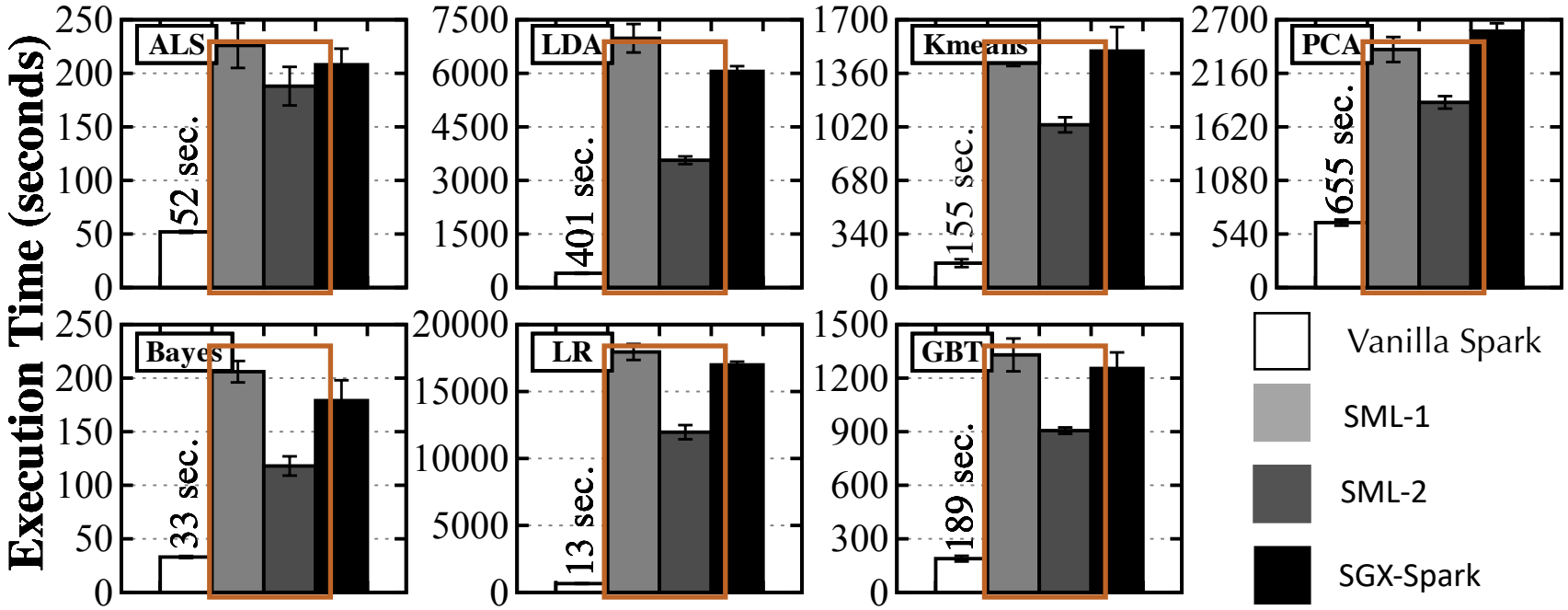
Solution


Trusted Site
Untrusted Site
Client
MLlib
Encryption Module
Data Loading
Encryption Module
Data
SML-1
Master
Spark Driver
Spark Context
Distributed Data Storage
Spark Workers
Full computation
Client
MLlib
Encryption Module
Data Loading
Encryption Module
Data
SML-2
Spark Workers
Statistics
Master
Spark Driver
Spark Context
Distributed Data Storage
Spark Workers
Training
Non-enclave
Enclave

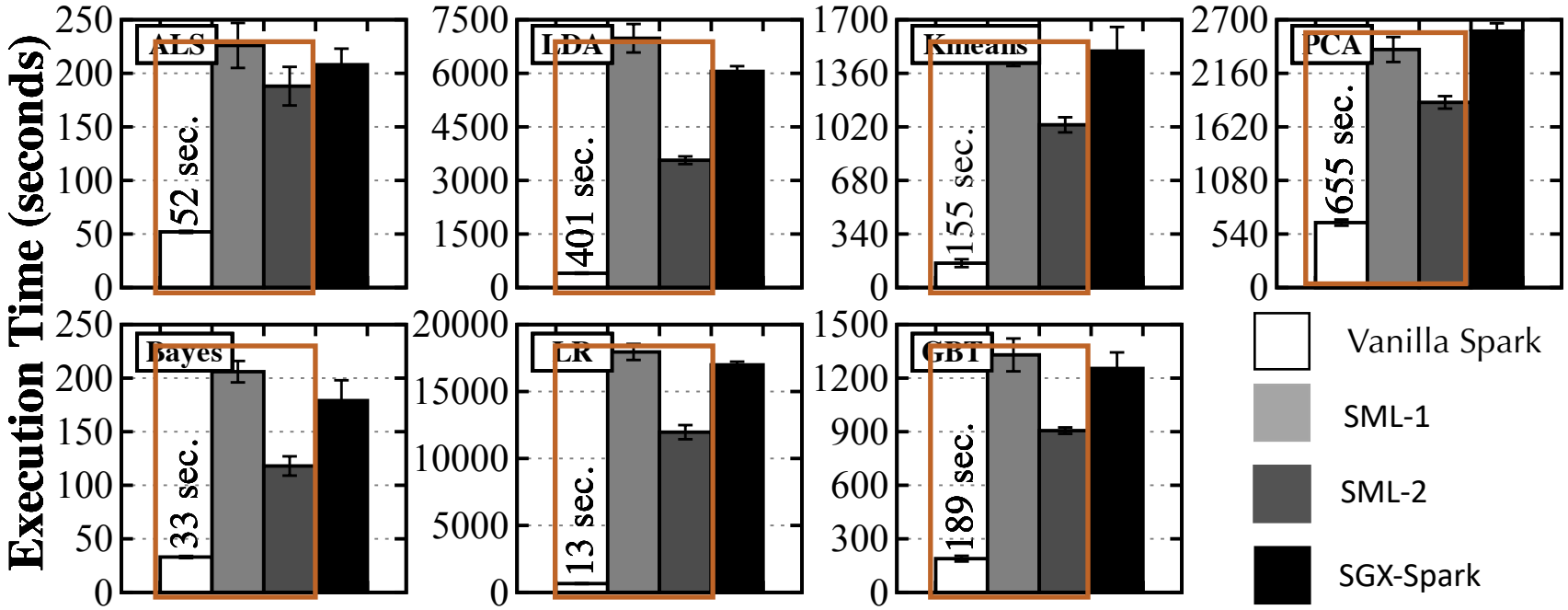
Results


Execution Time (seconds)
ALS
52 sec.
LDA
401 sec.
Kmeans
155 sec.
PCA
655 sec.
Bayes
33 sec.
LR
13 sec.
GBT
189 sec.
Vanilla Spark
SML-1
SML-2
SGX-Spark

Results


Execution Time (seconds)
ALS
52 sec.
LDA
401 sec.
Kmeans
155 sec.
PCA
655 sec.
Bayes
33 sec.
LR
13 sec.
GBT
189 sec.
Vanilla Spark
SML-1
SML-2
SGX-Spark

Results


Execution Time (seconds)
ALS
52 sec.
LDA
401 sec.
Kmeans
155 sec.
PCA
655 sec.
Bayes
33 sec.
LR
13 sec.
GBT
189 sec.
Vanilla Spark
SML-1
SML-2
SGX-Spark

Next Steps

- 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