



**HAL**  
open science

# Towards Scalable, Efficient and Privacy Preserving Machine Learning

Rania Talbi, Sara Bouchenak

► **To cite this version:**

Rania Talbi, Sara Bouchenak. Towards Scalable, Efficient and Privacy Preserving Machine Learning. Middleware '18 Doctoral Symposium, Dec 2018, Rennes, France. hal-01956155

**HAL Id: hal-01956155**

**<https://hal.science/hal-01956155v1>**

Submitted on 14 Dec 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

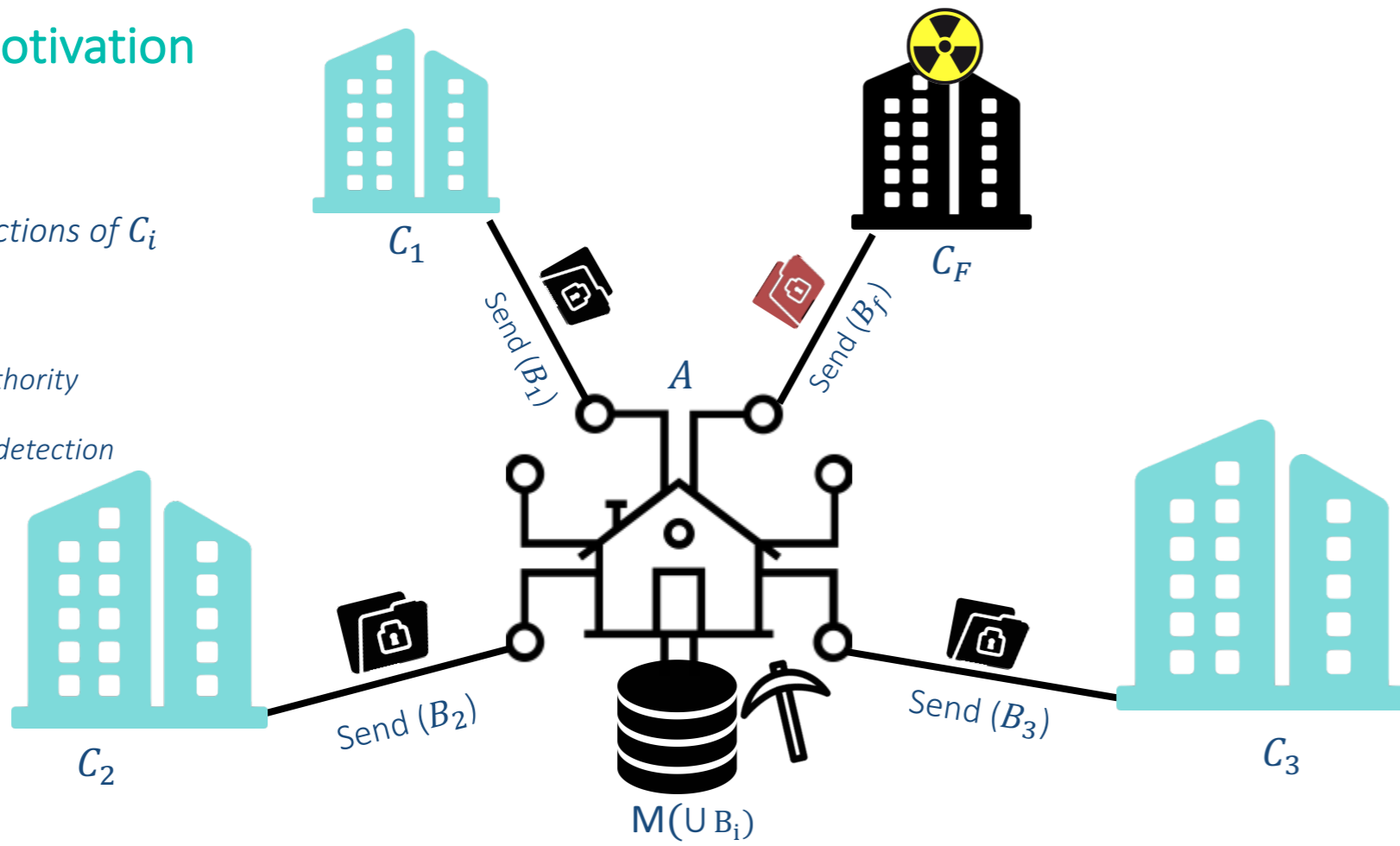
# Towards Scalable, Efficient and Privacy Preserving Machine Learning

Rania Talbi, Sara Bouchenak  
INSA Lyon, France  
{firstname.lastname}@insa-lyon.fr

2018 ACM/IFIP International Middleware Conference, Doctoral Symposium,  
December 10-14th 2018 – Rennes, France

## Context and Motivation

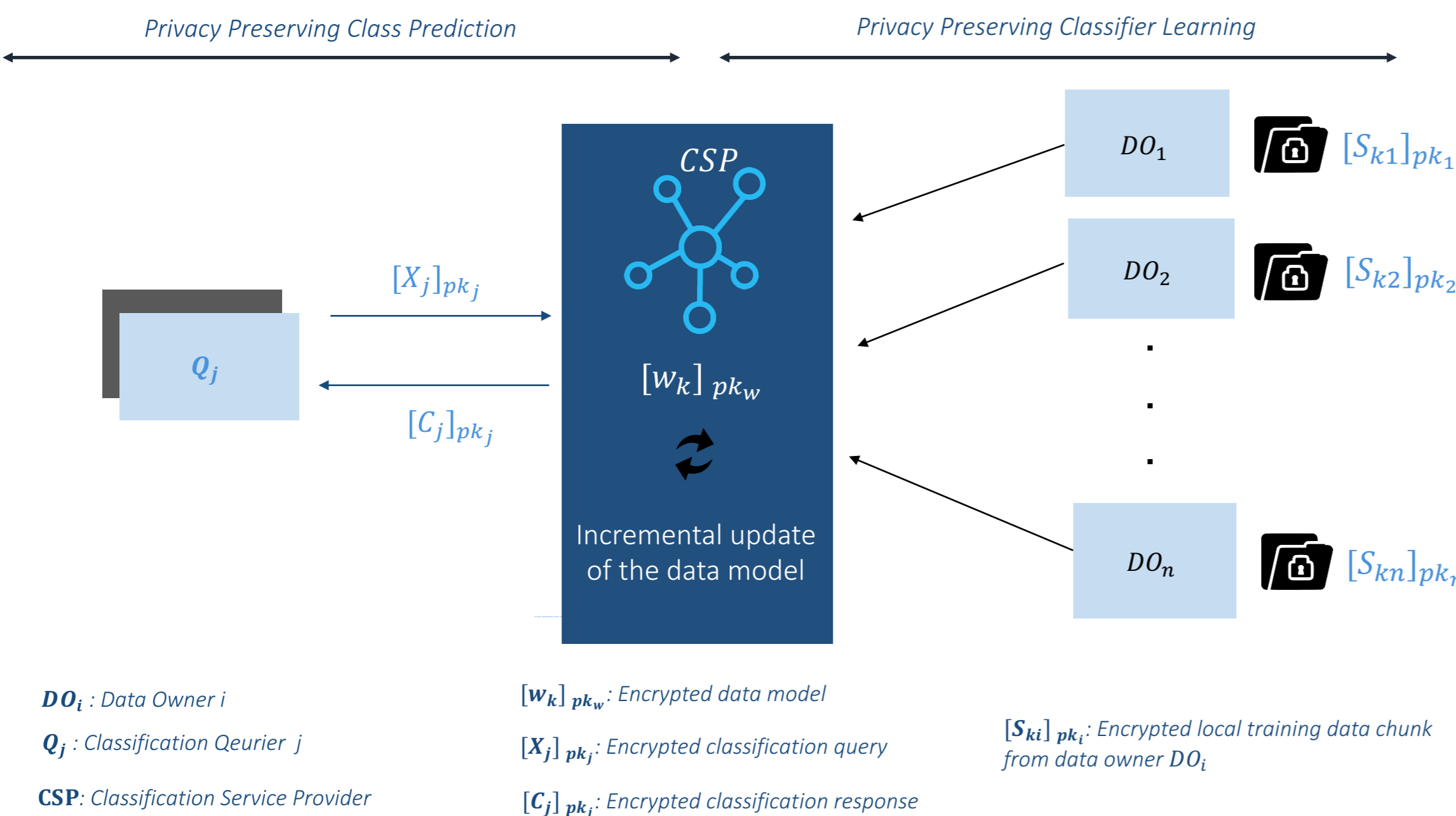
$C_i$ : Company  $i$   
 $B_i$ : Local bank transactions of  $C_i$   
 $C_F$ : Fraudulent company  
 $A$ : Central Supervision Authority  
 $M$ : Data Mining for fraud detection



## Objectives

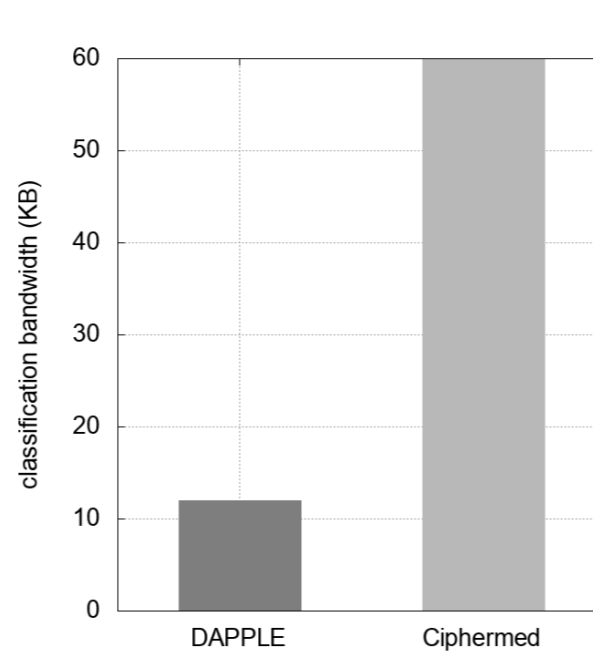
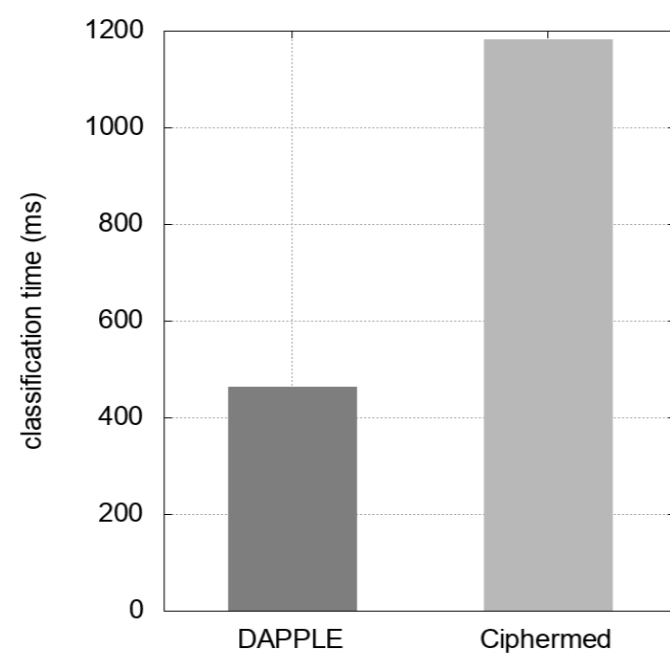
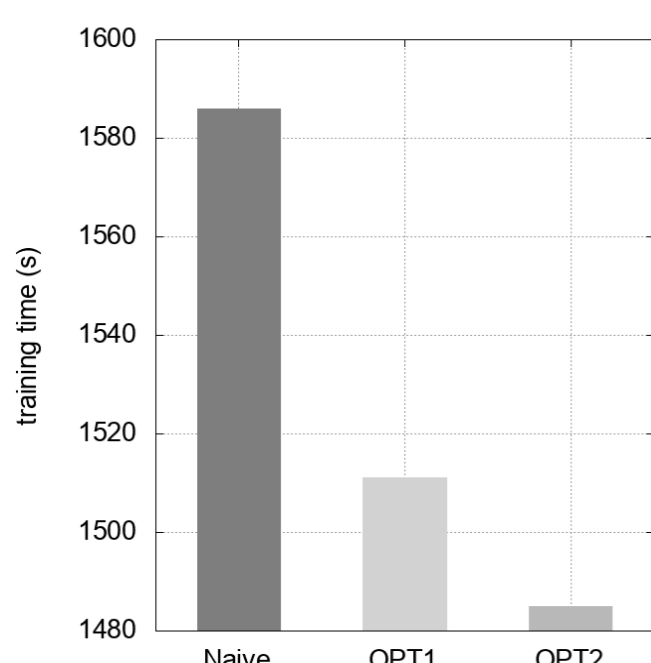
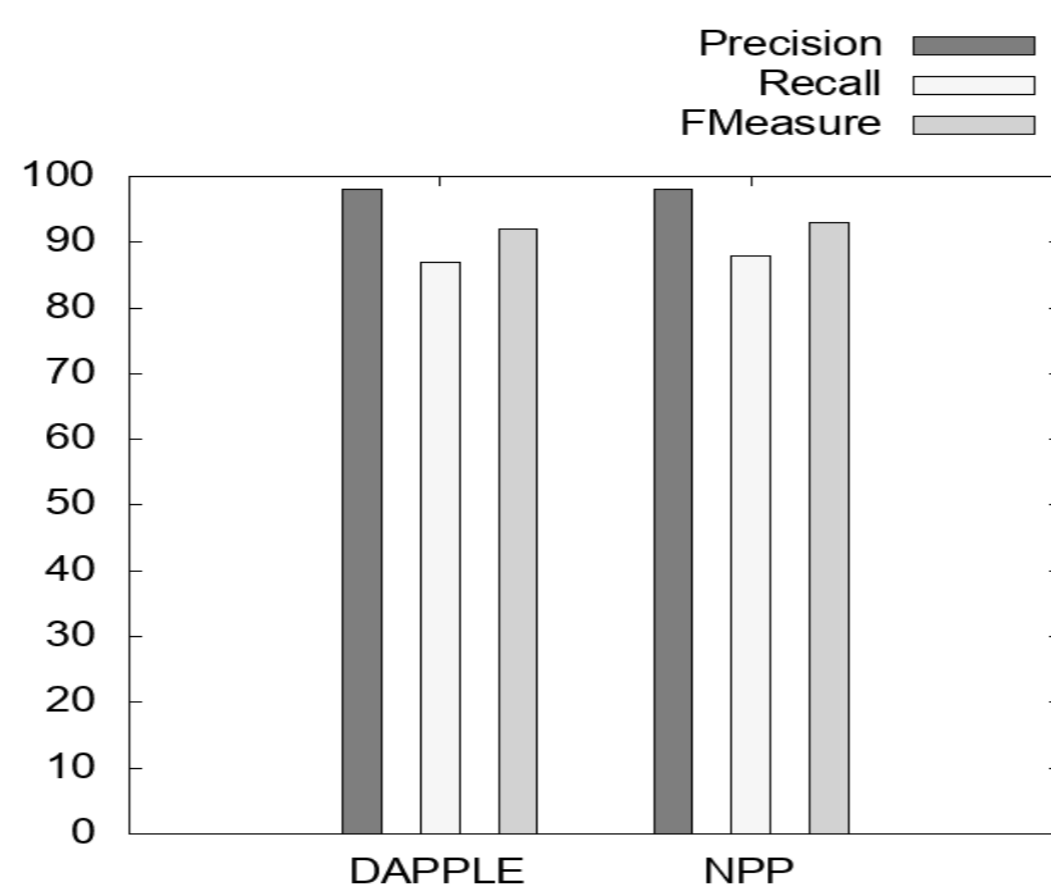
- Minimize the **computational costs** incurred by privacy preservation.
- Provide an **end-to-end** privacy preserving **outsourced** data classification service.
- Enable a set of mutually untrusted data owners to have a **global vision** on the union of their data without breaching the privacy of each one of them.
- Enable **dynamic** data model updates when new training data samples are available.

## DynAmic Privacy Preserving machine Learning Framework (DAPPLE)

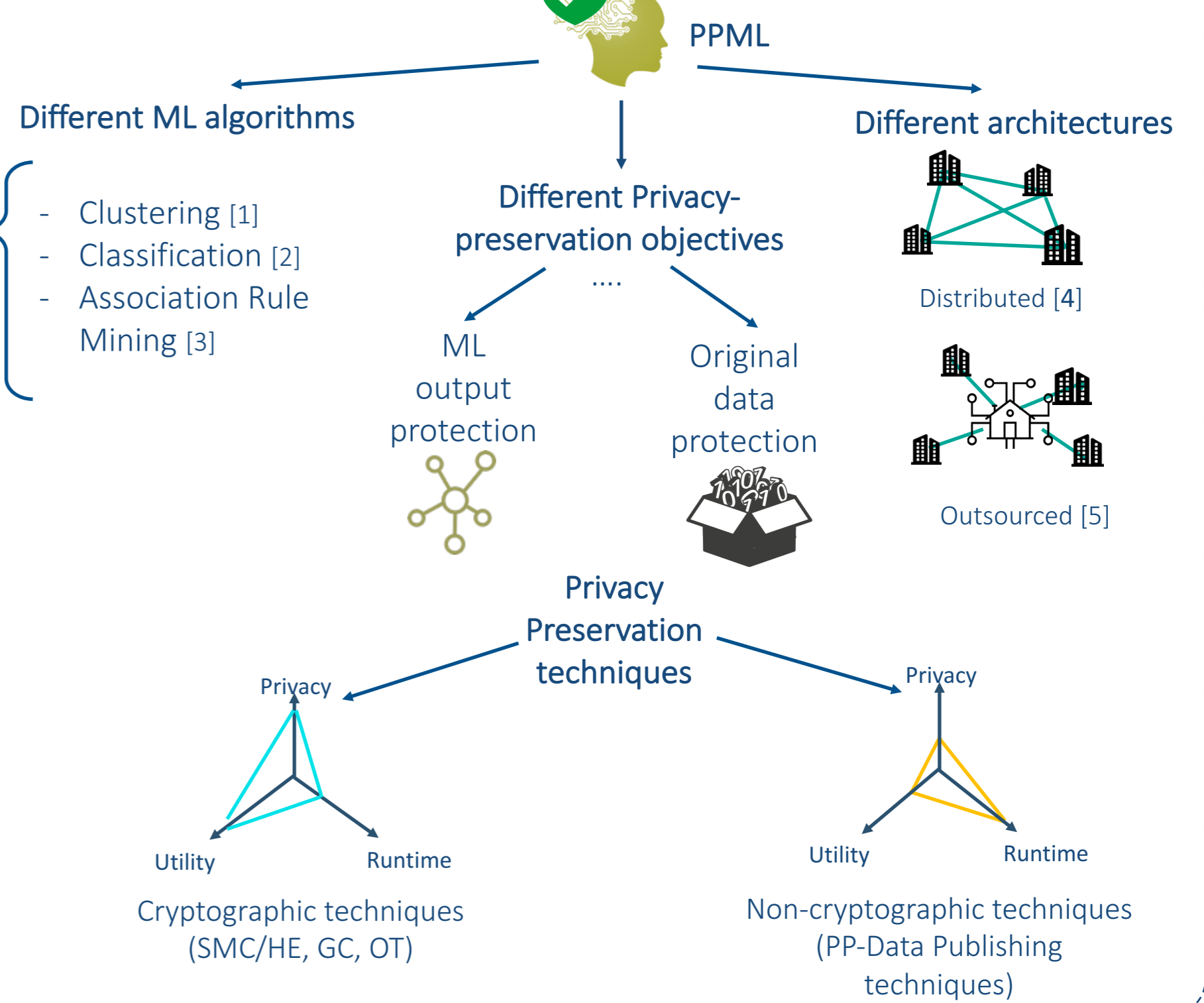


## Preliminary results

- We have used a synthetic dataset for fraud detection in a B2B network.
- This dataset contains 1000 bank transactions with 9 attributes each.
- We compare our work to the Ciphermed framework [8].

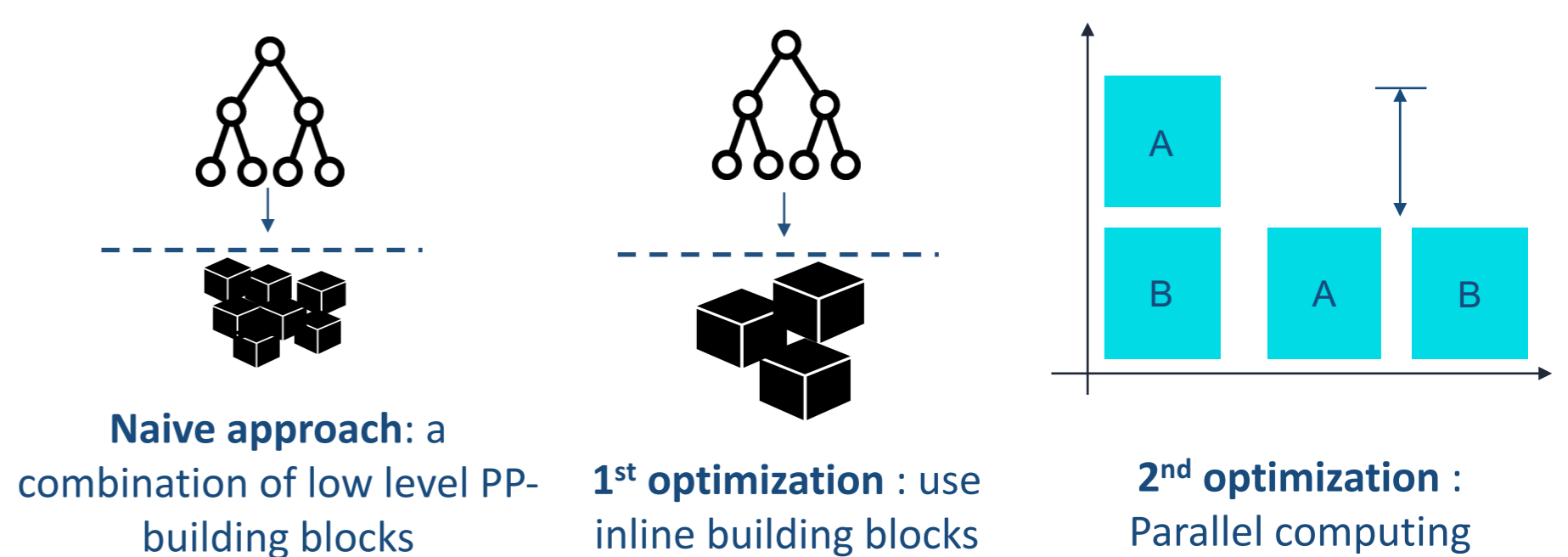


## Related work



## Design principles

- Cryptographic based protection (data model, training data, classification queries and responses)
  - Partial homomorphic encryption (PHE) based building blocks
  - Combine PHE with cryptographic blinding (DTPKC cryptosystem [6])
  - We implemented the VFDT incremental decision tree learning algorithm [7]
  - (1) Blind inputs
  - (2) Partially decrypt blinded values
  - (3) Decrypt blinded values
  - (4) Run operation over blinded values
  - (4) remove blinding from the result
- ex:  $[x]_{pk} \otimes [r]_{pk} = [x \oplus r]_{pk}$



## References

- X. Hu, et. al: Privacy-Preserving K-Means Clustering Upon Negative Databases. ICONIP (4) 2018.
- S. Kim et al. Privacy-Preserving Naive Bayes Classification Using Fully Homomorphic Encryption. ICONIP (4)2018: 349-358
- L.Liu et al : Privacy-Preserving Mining of Association Rule on Outsourced Cloud Data from Multiple Parties. ACISP2018: 431-451
- H.Yu et al.: Privacy-Preserving SVM Classification on Vertically Partitioned Data. PAKDD 2006: 647-656
- T.Li et al. : Outsourced privacy-preserving classification service over encrypted data. J. Network and Computer Applications 106: 100-110 (2018)
- X.Liu et al. : An Efficient Privacy-Preserving Outsourced Calculation Toolkit With Multiple Keys. IEEE Trans Information Forensics and Security 11(11): 2401-2414 (2016)
- M. Domingos et al.: Mining high-speed data streams. KDD 2000: 71-80
- R.Bost et al. : Machine Learning Classification over Encrypted Data. NDSS 2015