Università degli Studi di Napoli Federico II

itee PhD information technology electrical engineering

DIETI UNINA


# Vittorio Prodomo


# Privacy-enhancing fine tuning for secure collaborative inference

Tutor: Albert Banchs

co-Tutor: Simon Pietro Romano

Cycle: XXXVI

Year: 2022/23

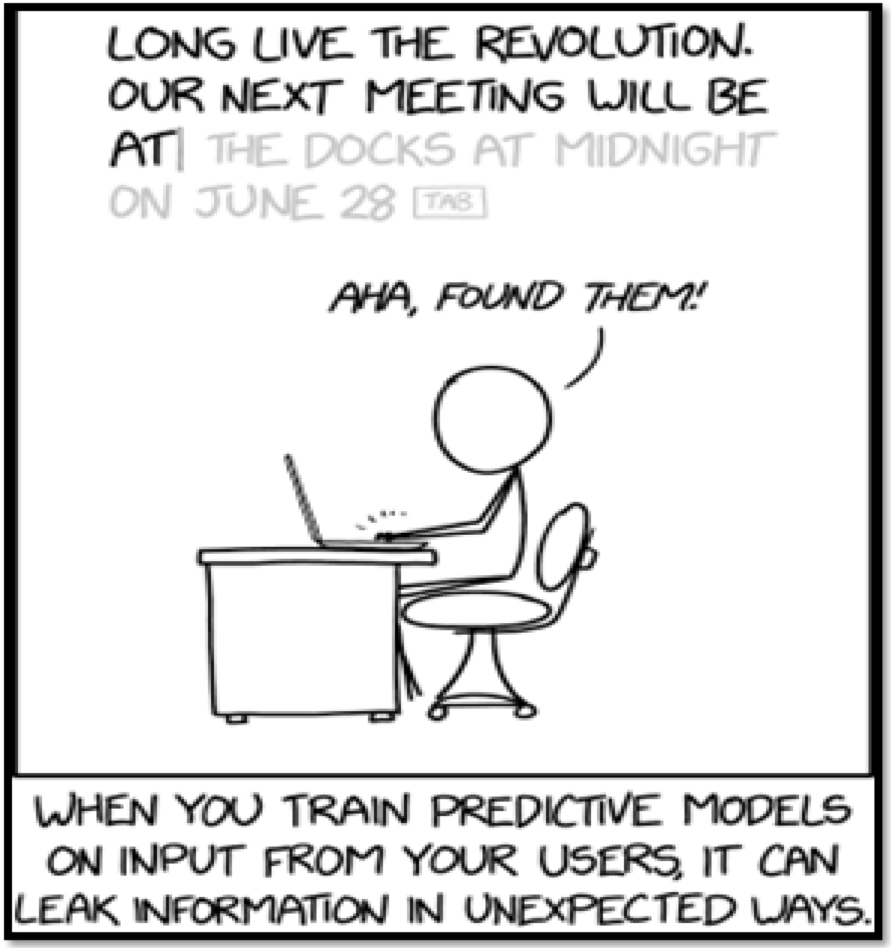
itee PhD information technology electrical engineering

# My background

- MSc degree: Computer Engineering (Federico II of Naples)
- Research laboratory: ARCLab
- PhD start date: March 2020
- Scholarship type: PIPF (University Carlos III of Madrid)
- Partner company: NEC Laboratories Europe GmbH

# Privacy-Preserving Machine Learning

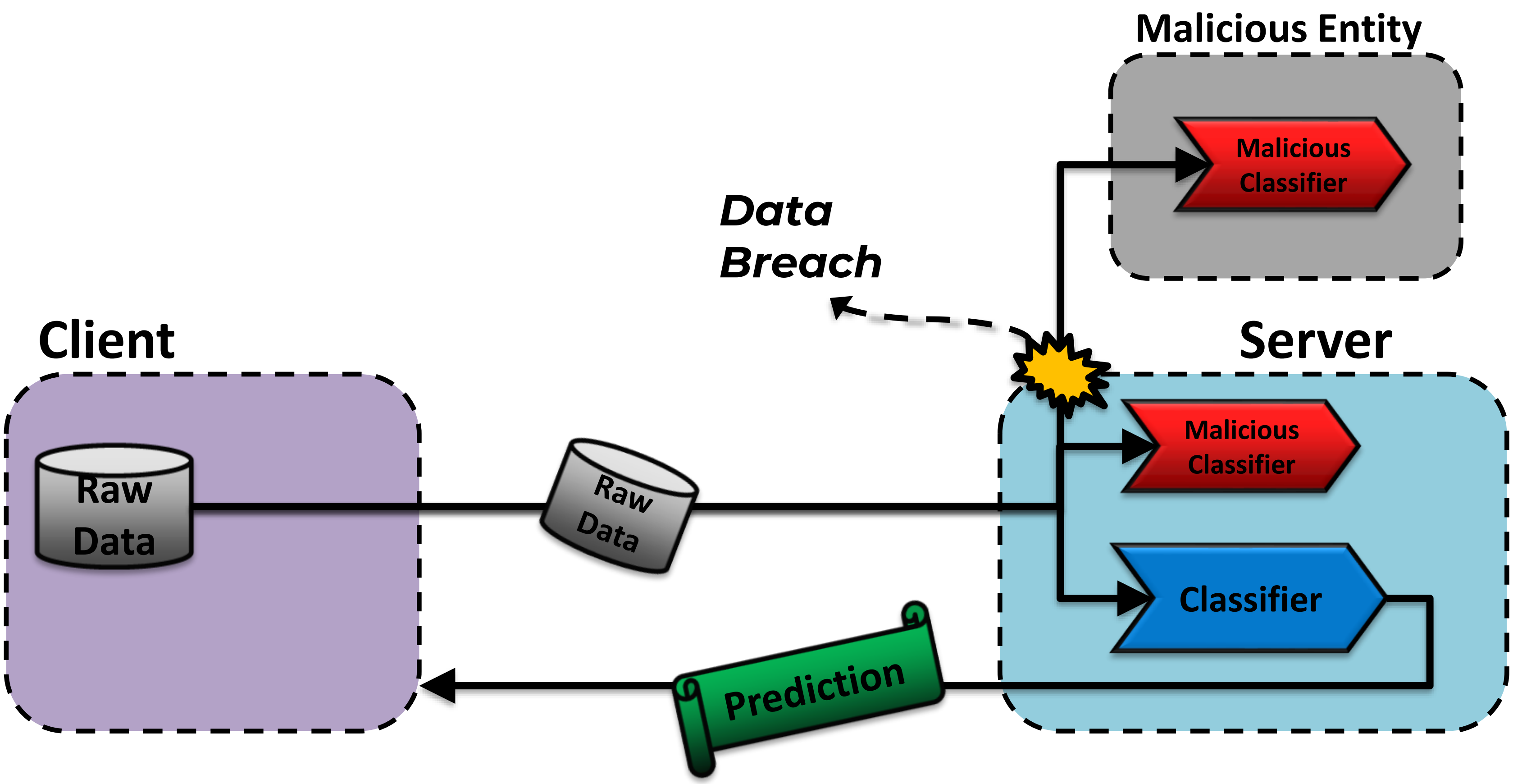
- Privacy-Preserving Machine Learning (PPML) aims to prevent data leakage in machine learning algorithms
- Currently a hot topic in literature
- Usually achieved via anonymization (k-anonymity, l-diversity, t-closeness), perturbation/obfuscation (Differential Privacy) or cryptographic techniques (Homomorphic Encryption, Secure Multi-party Computation)



# Main Problem: Data Privacy in MLaaS

- Machine-Learning-as-a-Service (MLaaS) scenarios are becoming increasingly more common
- Privacy of user data is at risk
  - Server data breach leaks data to malicious entities
  - Service owner itself may be “honest-but-curious”
- Potentially sensitive extra information could be inferred from the data
- The only allowed data usage must be the one requested/expected by the user

# Main Problem: Data Privacy in MLaaS



# Main Objective: Data anonymization

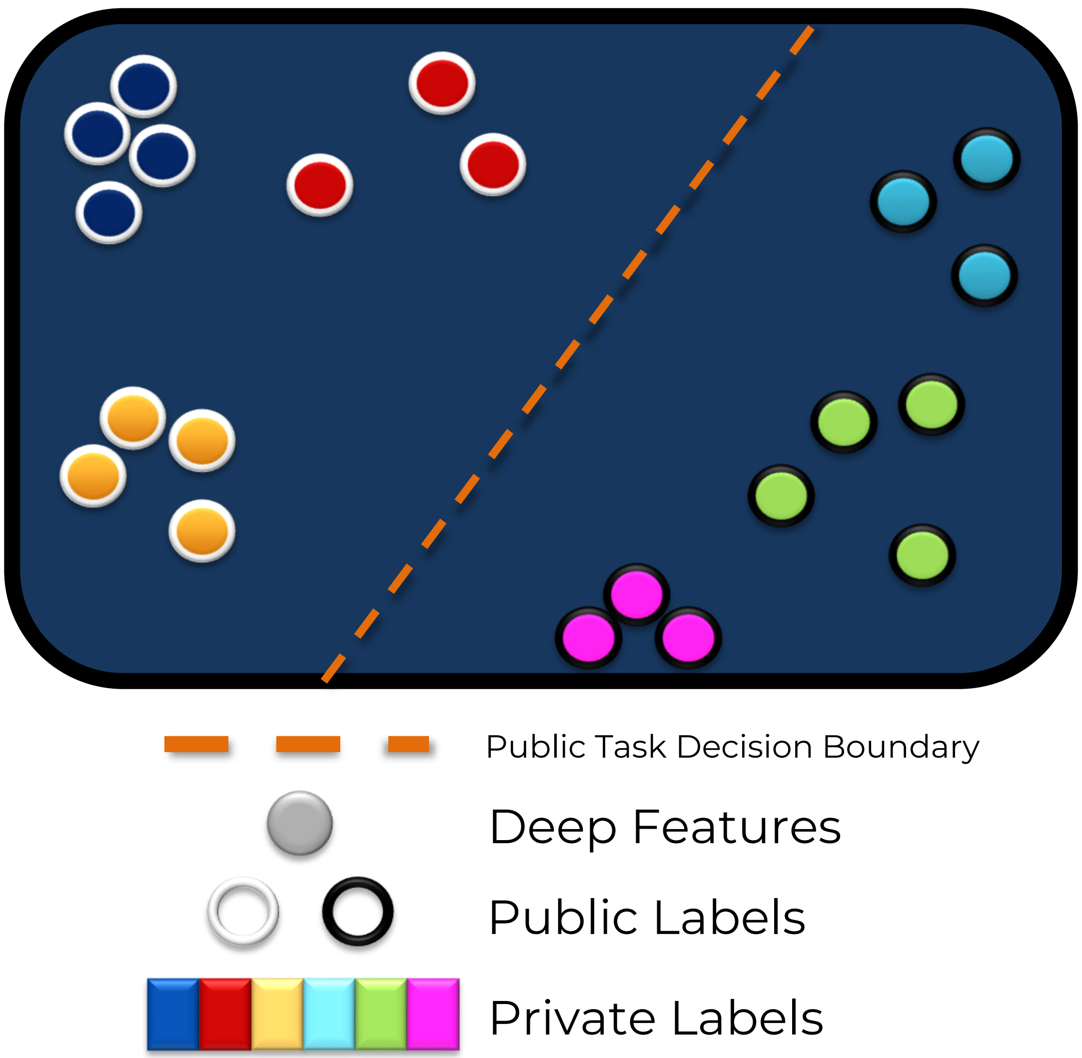
**Lossless techniques**

- Trusted Execution Environment (TEE), Homomorphyc Encryption (HE), Secure Multi-party Computation (SMC).
- Data privacy directly granted by computational security via cryptographic techniques.
- Usually suffer from high computational overhead.

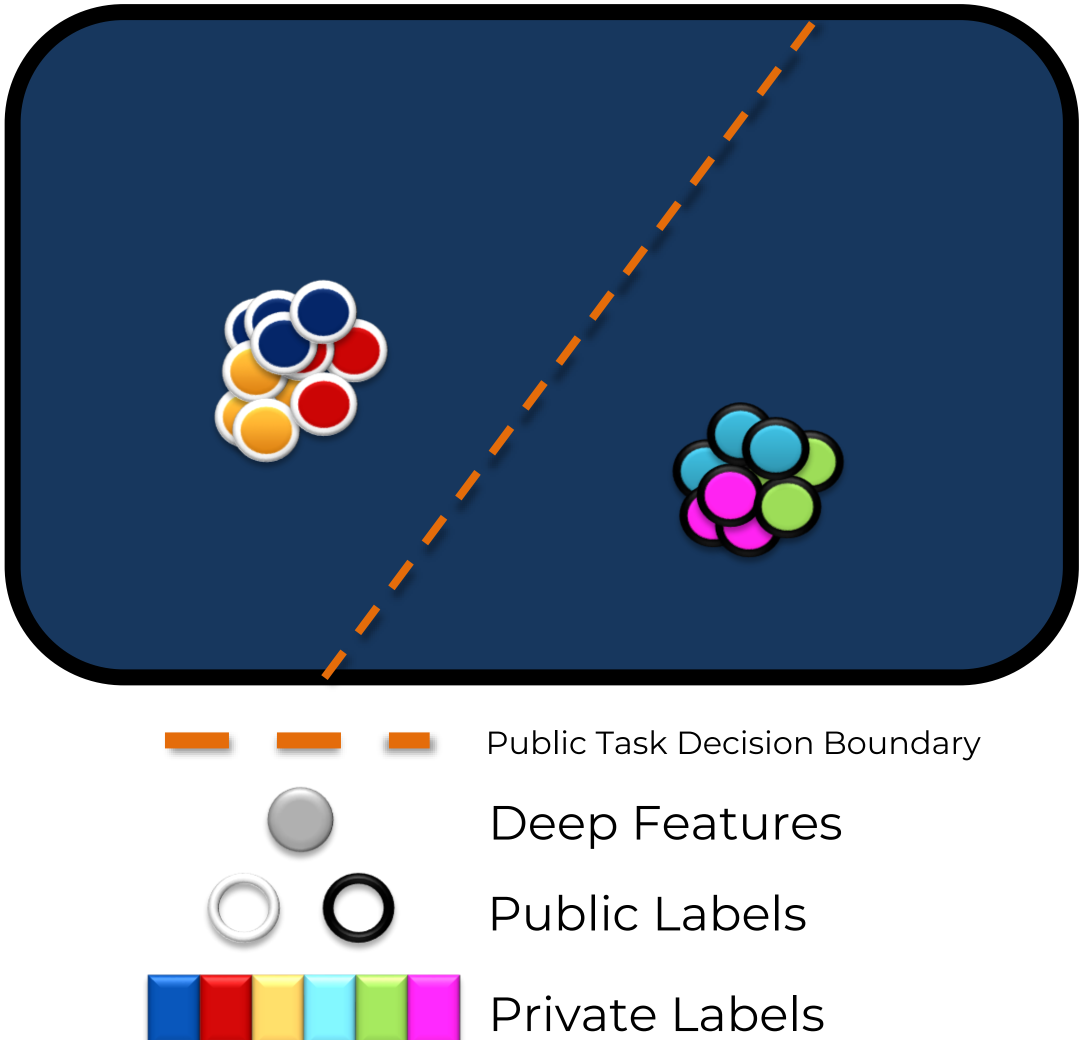
**Lossy techniques**

- K-anonymity, L-diversity, T-closeness, Differential Privacy (DP), task-driven privacy-preserving anonymization.
- Typically apply an irreversible "lossy" transformation to the data (with negligible overhead)
- Inevitably present a trade-off among privacy, utility and scalability

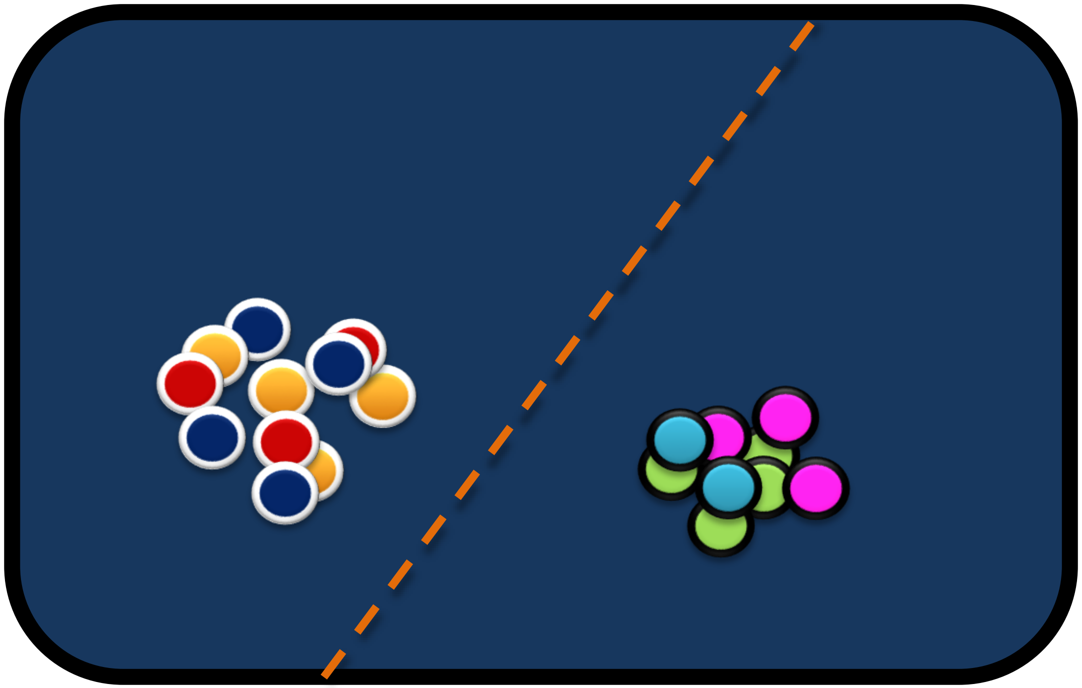
# Data Anonymization: Theory

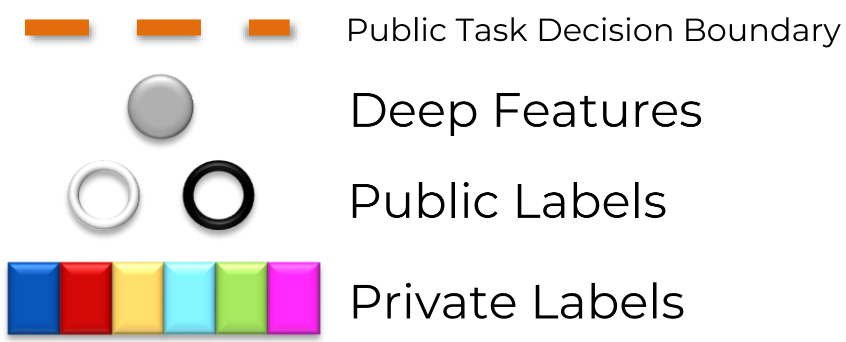


# Data Anonymization: Theory



# Data Anonymization: Theory

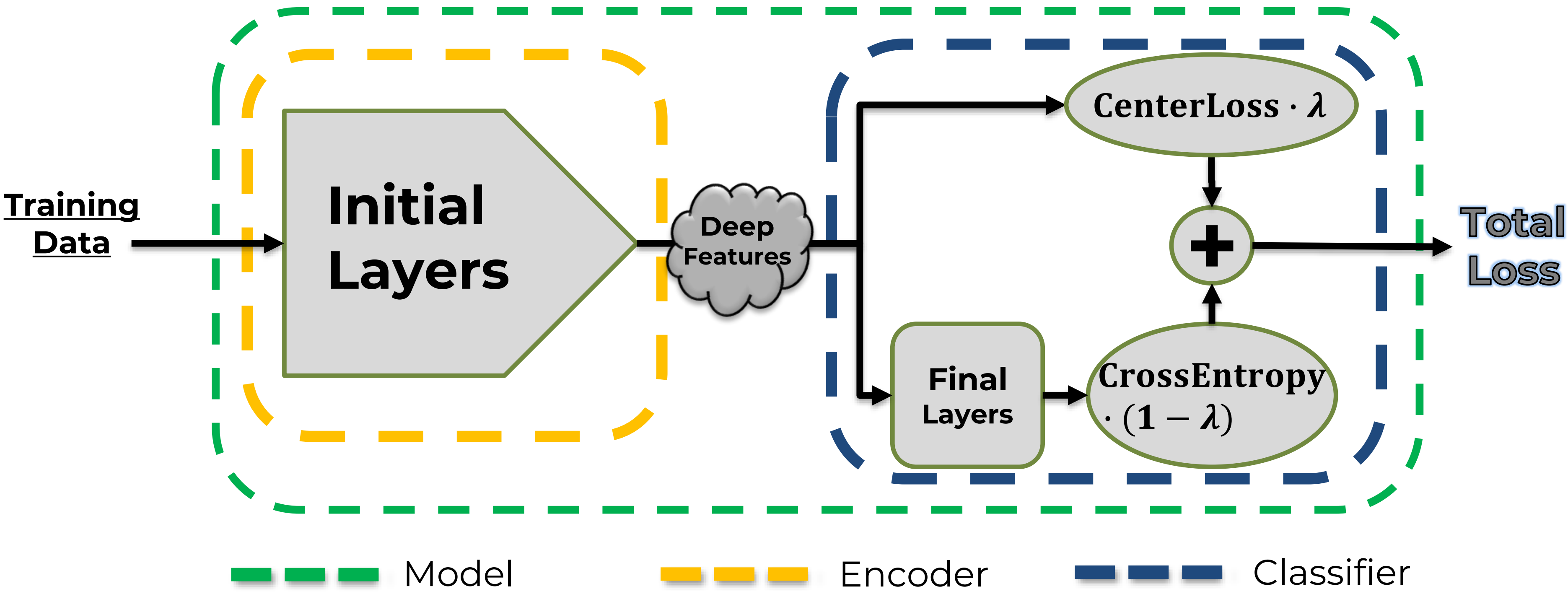




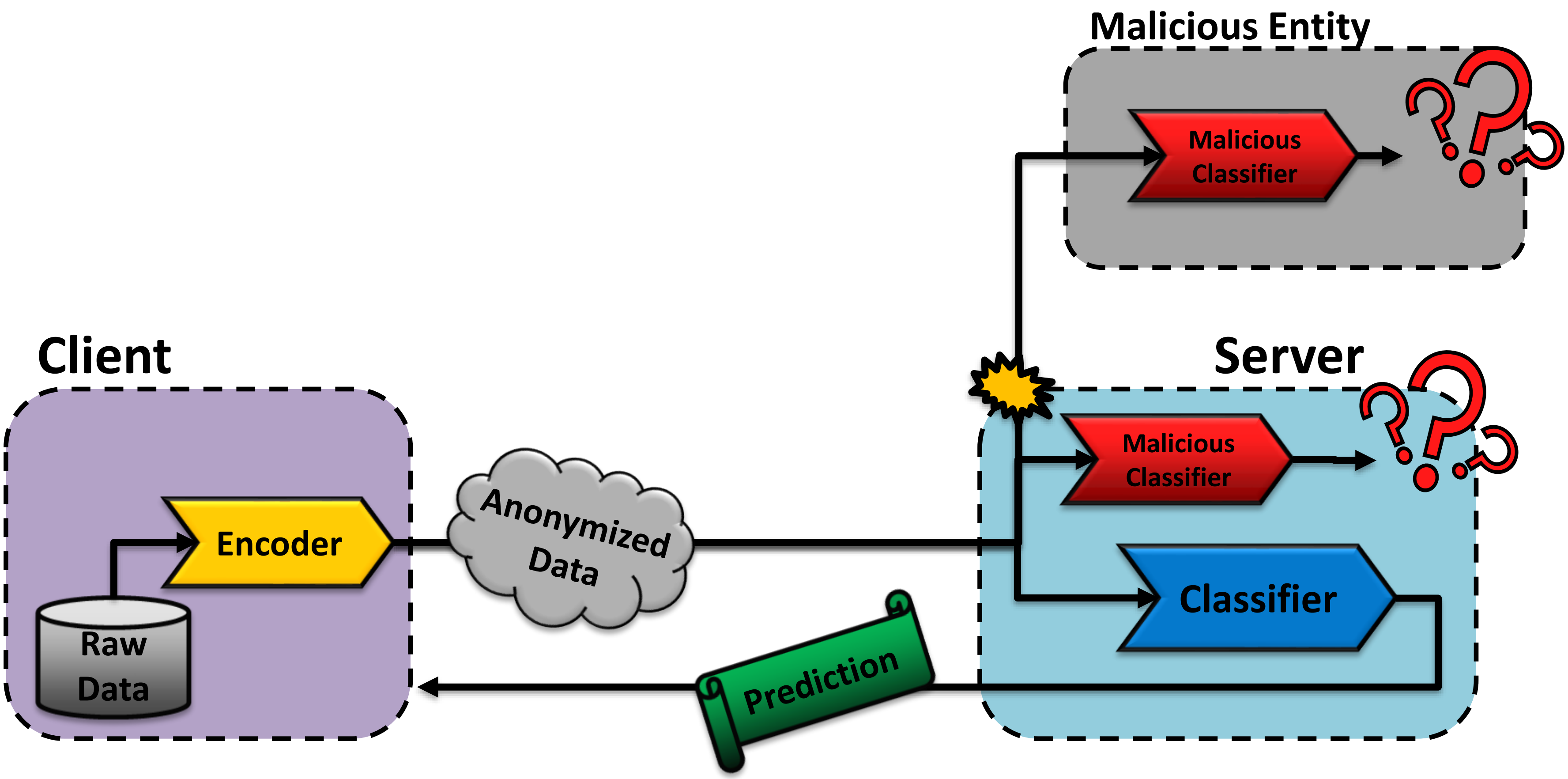
# Implemented solution

- **Use case**: virtually any FC and Conv Network for classification tasks (we tested simple MLPs, small ConvNets and deep ResNets)
- **Fine tune** the deployed model to obtain "private" mid-network deep features
  - An auxiliary loss function is added to iteratively steer the features in the encoding space
  - The combined loss depends on the pre-knowledge assumed in the scenario: our case is 1 main task, unknown malicious task(s)
  - Random noise may be added to the features to perturb them
- **Split model** to reduce client-side computational burden and/or keep client engaged with MLaaS platform
- **Provide first half** as a closed-box encoder to the client
  - A non-invertible, tailored anonymization function
- **Use second half** (server-side) to carry out the requested task on the received anonymized data, and send back the results

# Implemented solution: Training



# Implemented solution: Deployment



# Implemented solution: Qualitative results

Public task → Female Female Male Male

Reference images

Naive approach

Our method

Private task → Smiling Smiling

itee PhD information technology electrical engineering

# Next year planning

- **Extend the work:** test more datasets, data types, privacy metrics, and use cases (e.g., finance and cybersecurity)
- **Improve the center loss:** test other distance metrics, add perturbative noise to deep features
- **Assess generalization:** e.g., study how legitimate/malicious tasks correlation affects the effectiveness of the approach
- **Improve applicability:** e.g., refactor approach to avoid deployed model re-training/fine tuning

# Next year planning

- ~~**Extend the work:** test more datasets, data types, privacy metrics, and use cases (e.g., finance and cybersecurity)~~
- **Improve the center loss:** test other distance metrics, add perturbative noise to deep features
- **Assess generalization:** e.g., study how legitimate/malicious tasks correlation affects the effectiveness of the approach
- **Improve applicability:** e.g., refactor approach to avoid deployed model re-training/fine tuning

# Next year planning

- ~~**Extend the work:** test more datasets, data types, privacy metrics, and use cases (e.g., finance and cybersecurity)~~
- ~~**Improve the center loss:** test other distance metrics, add perturbative noise to deep features~~
- **Assess generalization:** e.g., study how legitimate/malicious tasks correlation affects the effectiveness of the approach
- **Improve applicability:** e.g., refactor approach to avoid deployed model re-training/fine tuning

# Next year planning

- ~~**Extend the work:** test more datasets, data types, privacy metrics, and use cases (e.g., finance and cybersecurity)~~
- ~~**Improve the center loss:** test other distance metrics, add perturbative noise to deep features~~
- **Assess generalization:** e.g., study how legitimate/malicious tasks correlation affects the effectiveness of the approach
- **Improve applicability:** e.g., refactor approach to avoid deployed model re-training/fine tuning

# Thanks for listening!