AI-Driven Cybersecurity Poisoning & Inference Attacks

Prof. Antonino Nocera

Dr. Marco Arazzi

dcalab

UNIVERSITÀ DI PAVIA

About Me

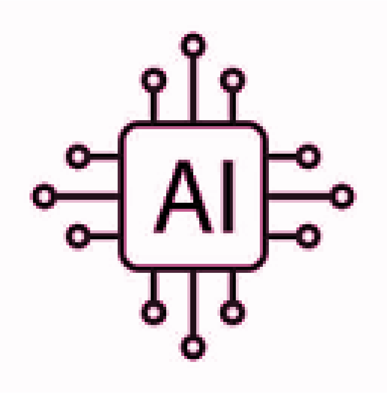
email: antonino.nocera@unipv.it
Web: antoninonocera.unipv.it
DCALab: dcalab.unipv.it

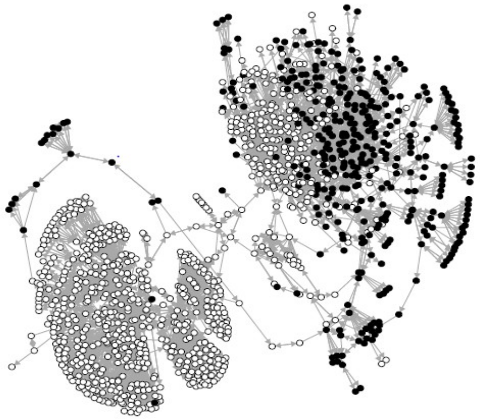
Antonino Nocera is an Associate Professor at the University of Pavia. He received his PhD in Information Engineering at the Mediterranea University of Reggio Calabria in 2013. His research interests span over several contexts, including: Artificial Intelligence, Security, Privacy, Data Science and Social Network Analysis. In these fields, he published about 90 scientific papers in prestigious peer-reviewed International Journals and Conferences. He is involved in the TPC of many renowned International Conferences in the context of Data Science and Cybersecurity. Moreover, he is Associate Editor of Information Sciences (Elsevier) and of the IEEE Transaction on Information Forensics and Security (T-IFS).

He is the director of the local node of the University of Pavia for the CINI "Data Science" National Lab and he is a member of the local node of the University of Pavia for the CINI Cybersecurity National Lab.

Research Interests

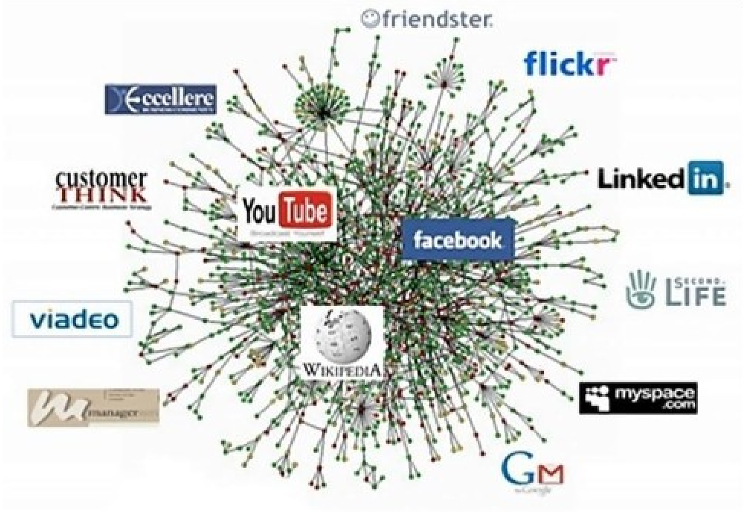
- Data Science and Social Network Analysis
- Security & Privacy
- AI-Driven Cybersecurity





Multi-Social-Network Scenario

People create multiple profiles in different social networks and link them through me edge.



SNA → Structural Analysis

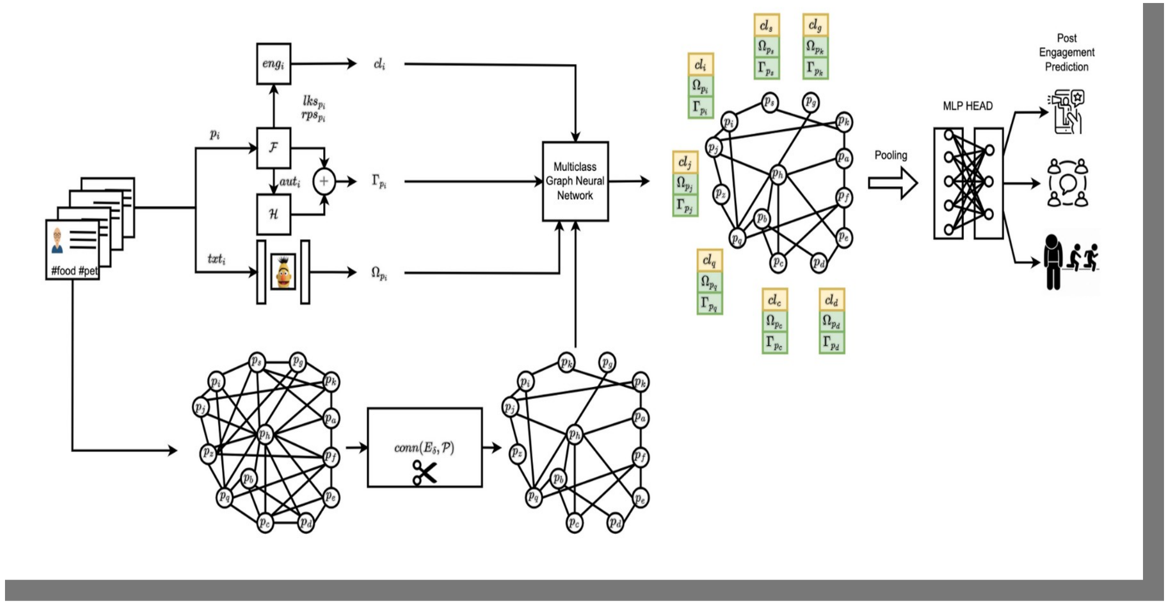
Graph Mining

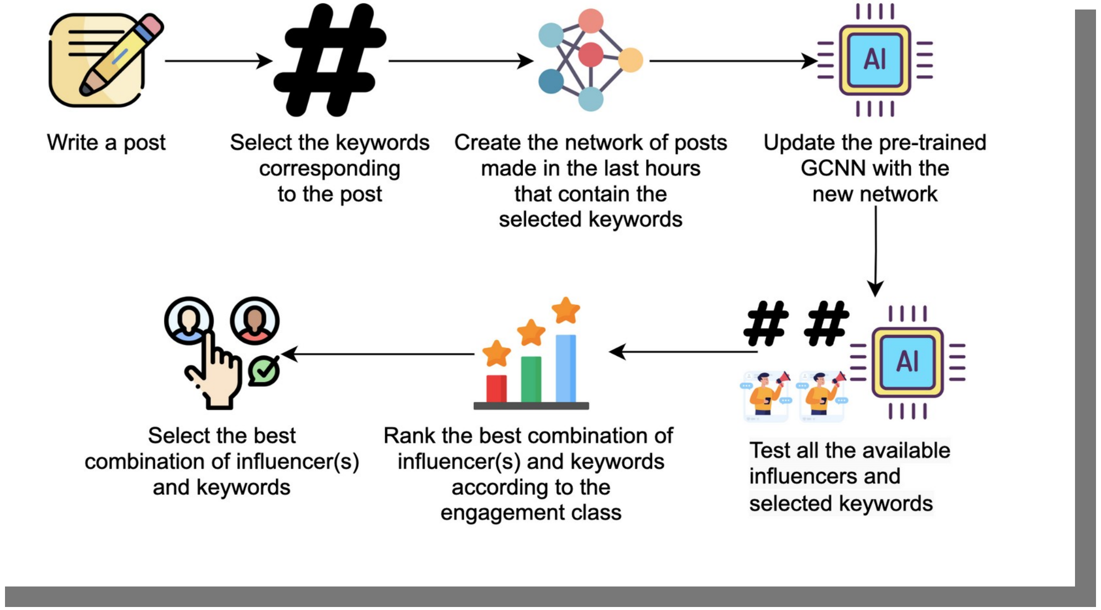
Natural Language Processing

Graph Neural Network

- Arazzi, M., Ferretti, M., Nicolazzo, S., & **Nocera, A.** (2023). The role of social media on the evolution of companies: A Twitter analysis of Streaming Service Providers. Online Social Networks and Media, 36, 100251.
- Arazzi, M., Nicolazzo, S., **Nocera, A.**, & Zippo, M. (2023). The importance of the language for the evolution of online communities: An analysis based on Twitter and Reddit. *Expert Systems with Applications*, *222*, 119847.
- Corradini, E., **Nocera, A.**, Ursino, D., & Virgili, L. (2021). Investigating negative reviews and detecting negative influencers in Yelp through a multi-dimensional social network based model. *International Journal of Information Management*, *60*, 102377.
- Corradini, E., **Nocera, A.**, Ursino, D., & Virgili, L. (2020). Defining and detecting k-bridges in a social network: the yelp case, and more. *Knowledge-Based Systems*, *195*, 105721.

A Multiclass Graph Neural Network-Based Framework for Predicting Post Engagement in Social Media




Write a post → Select the keywords corresponding to the post → Create the network of posts made in the last hours that contain the selected keywords → Update the pre-trained GCNN with the new network → Test all the available influencers and selected keywords → Rank the best combination of influencer(s) and keywords according to the engagement class → Select the best combination of influencer(s) and keywords

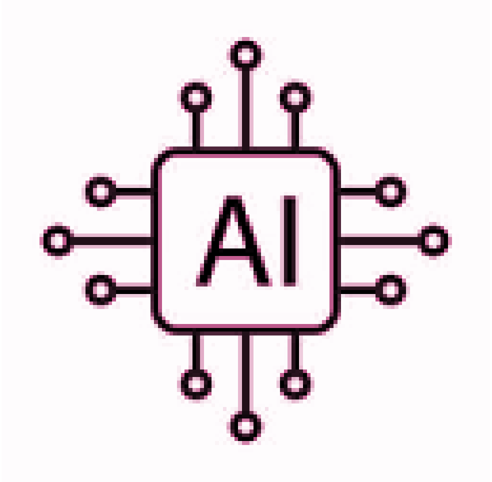
Arazzi, M., Cotogni, M., Nocera, A., & Virgili, L. (2023). Predicting Tweet Engagement with Graph Neural Networks. *arXiv preprint arXiv:2305.10103*.

Security & Privacy

- E. Corradini, S. Nicolazzo, **A. Nocera**, D. Ursino and L. Virgili. A two-tier Blockchain framework to increase protection and autonomy of smart objects in the IoT. Computer Communications. 2022, 181, pp. 338–356.
- S. Nicolazzo, **A. Nocera**, D. Ursino and L. Virgili. A privacy-preserving approach to prevent feature disclosure in an IoT scenario. Future Generation Computer Systems, 105:502-519, 2020.
- S. Nicolazzo, **A. Nocera** and D. Ursino. Anonymous Access Monitoring of Indoor Areas. IEEE Access, 2021, 9: 56664-56682.
- F. Buccafurri, G. Lax, S. Nicolazzo, and **A. Nocera**. A Privacy-Preserving Localization Service for Assisted Living Facilities. IEEE Transactions on Services Computing (TSC), IEEE Computer Society Vol.9. 2017.
- F. Buccafurri, G. Lax, S. Nicolazzo, and **A. Nocera**. A System for Privacy-Preserving Access Accountability in Critical Environments. IEEE Pervasive Computing, 2019, 18.2: 58-66.

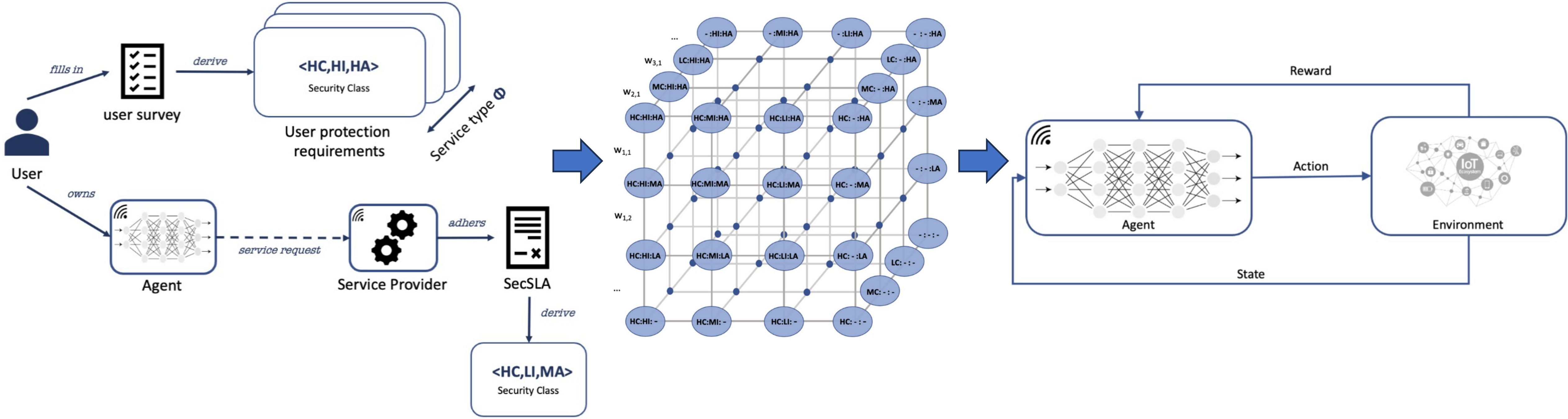
AI-Driven Cybersecurity

- Cybersecurity is a crucial aspect in everybody «digital» life.
- Huge Research community.
- Spread across different and independent domains.
- AI is a new powerful driver in many context.
- Many aspects of AI are extremely in line with the cybersecurity realm
- Attackers are leveraging AI to design new zero days menaces.
- Researchers are adopting AI to build versatile and powerful solutions.



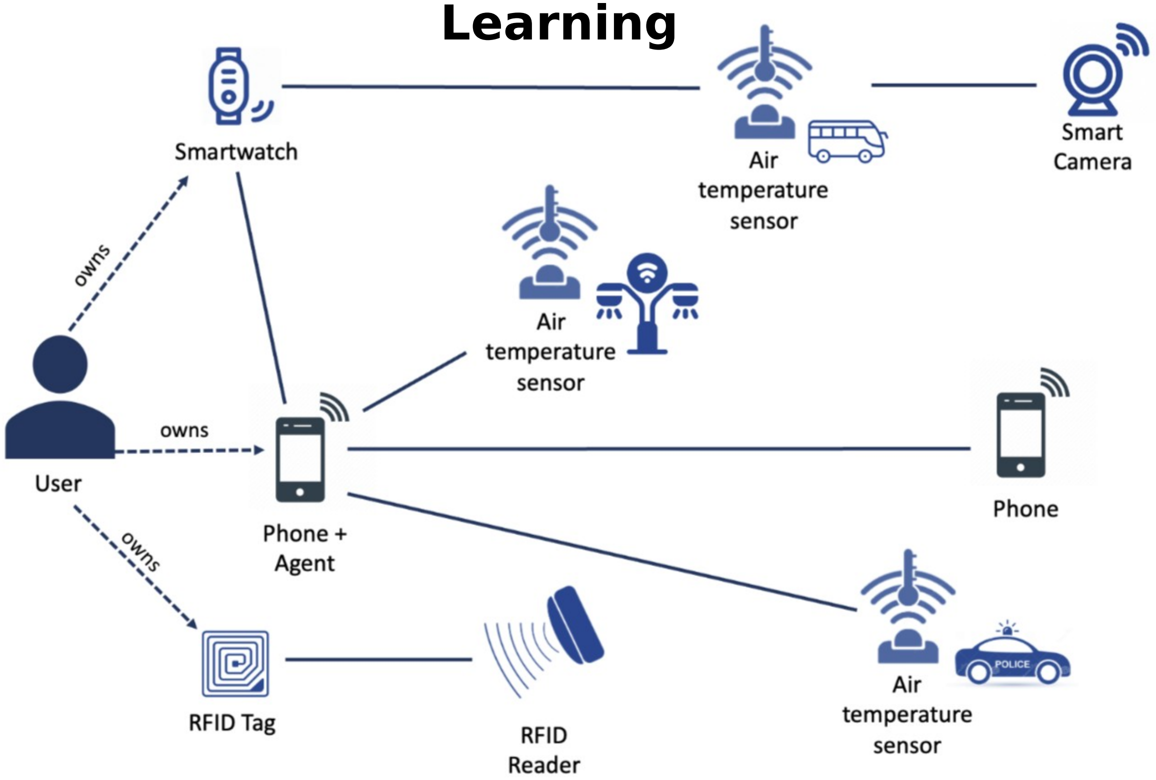
AI-Driven Cybersecurity

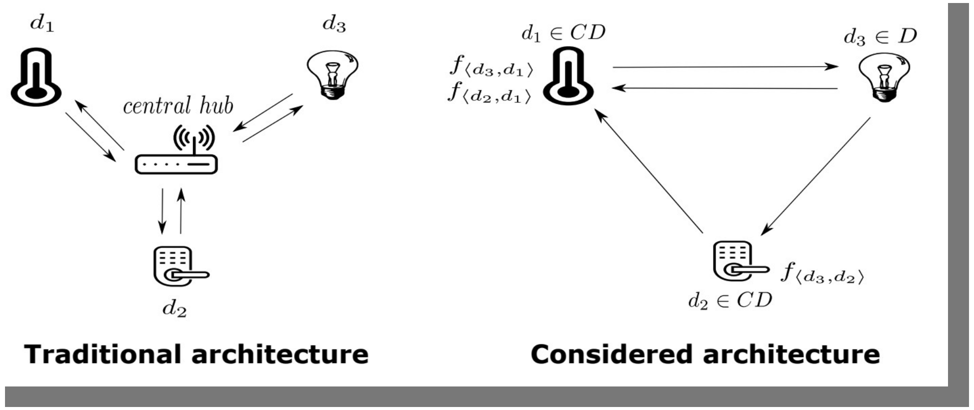
Autonomous Privacy-aware service exploitation in IoT

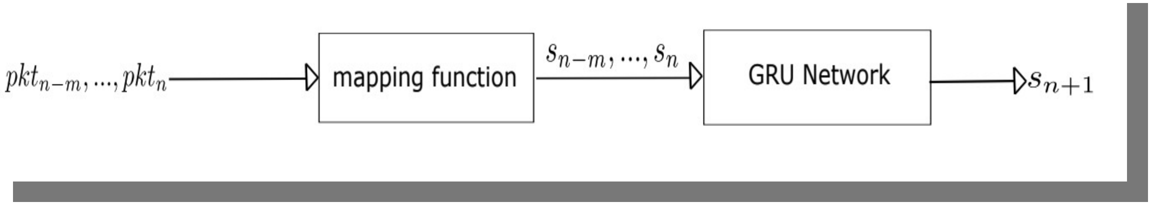


AI-Driven Cybersecurity

Behavioral Fingerprinting and Federated Learning







- Aramini, A., Arazzi, M., Facchinetti, T., Ngankem, L. S., & **Nocera, A**. (2022, April). An enhanced behavioral fingerprinting approach for the Internet of Things. In 2022 IEEE 18th International Conference on Factory Communication Systems (WFCS) (pp. 1-8). IEEE.
- Ferretti, M., Nicolazzo, S., & **Nocera, A**. (2021). H2O: secure interactions in IoT via behavioral fingerprinting. *Future Internet*, *13*(5), 117.

AI-Driven Cybersecurity

dcalab

UNIVERSITÀ DI PAVIA

Behavioral Fingerprinting and Federated Learning

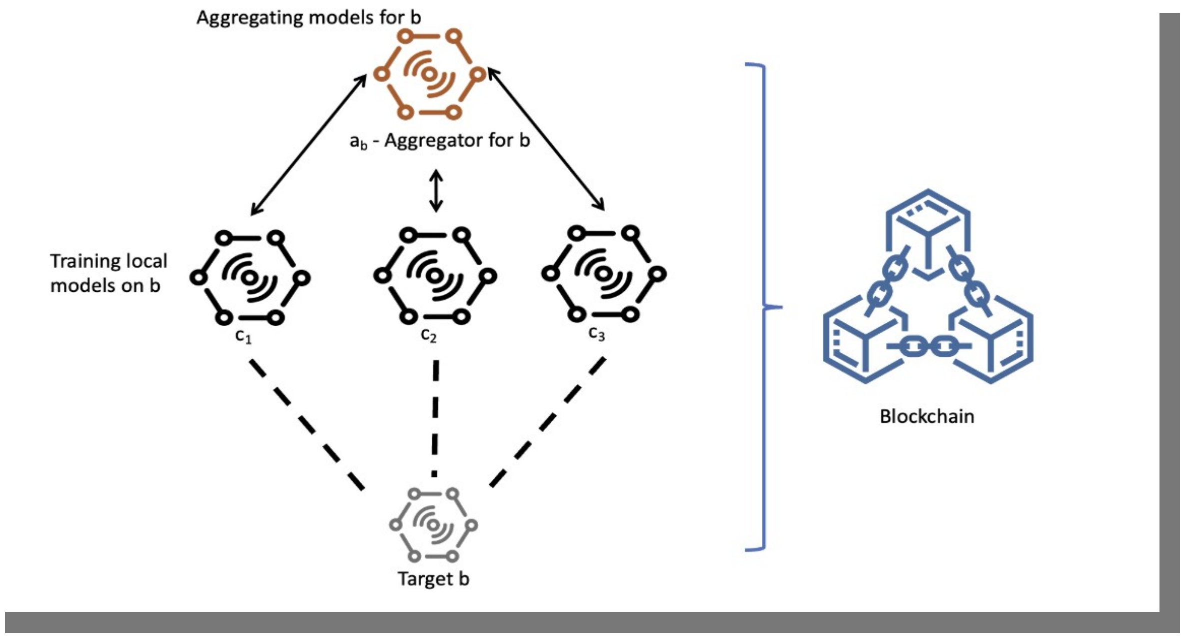
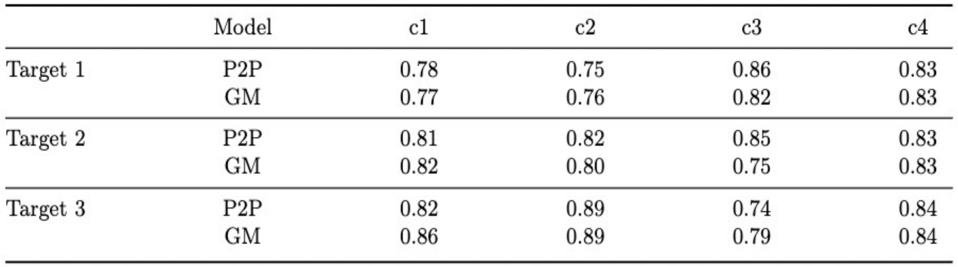
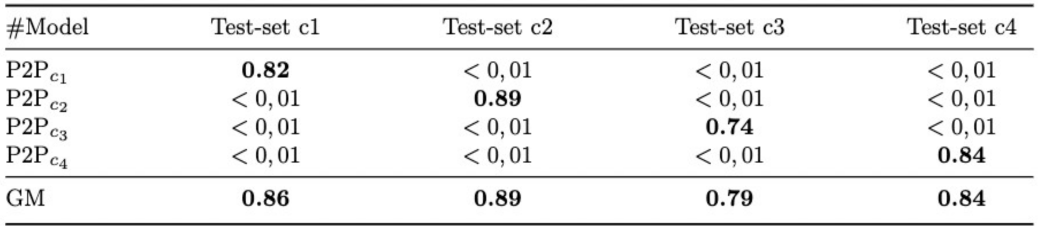


Table 3 Comparison of the performance of our approach (GM) and the solution of [6] (P2P) with direct testing



	Model	c1	c2	c3	c4
Target 1	P2P	0.78	0.75	0.86	0.83
	GM	0.77	0.76	0.82	0.83
Target 2	P2P	0.81	0.82	0.85	0.83
	GM	0.82	0.80	0.75	0.83
Target 3	P2P	0.82	0.89	0.74	0.84
	GM	0.86	0.89	0.79	0.84

Table 4 Comparison of the performance of our approach and the solution of [6] with cross testing



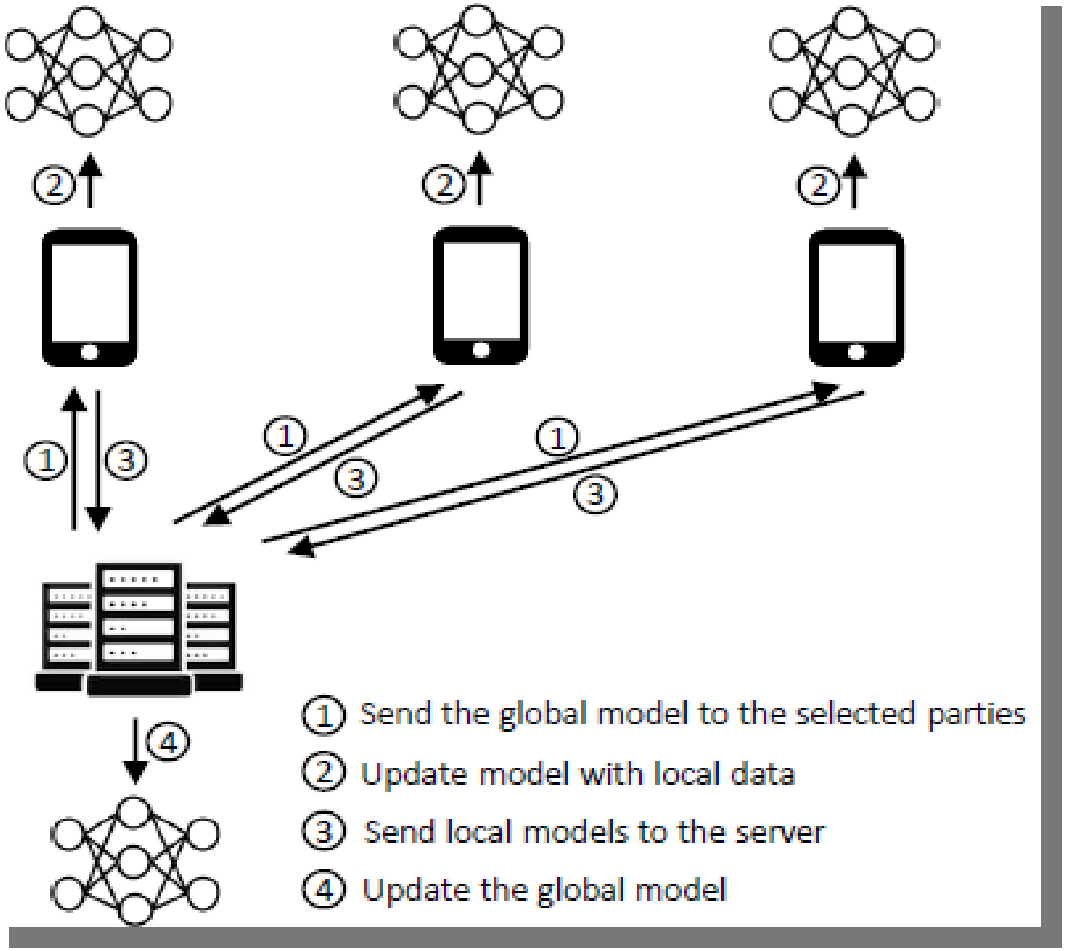
#Model	Test-set c1	Test-set c2	Test-set c3	Test-set c4
$P2P_{c_1}$	**0.82**	$< 0,01$	$< 0,01$	$< 0,01$
$P2P_{c_2}$	$< 0,01$	**0.89**	$< 0,01$	$< 0,01$
$P2P_{c_3}$	$< 0,01$	$< 0,01$	**0.74**	$< 0,01$
$P2P_{c_4}$	$< 0,01$	$< 0,01$	$< 0,01$	**0.84**
GM	**0.86**	**0.89**	**0.79**	**0.84**

- Arazzi, M., Nicolazzo, S., & **Nocera, A.** (2023). A Fully Privacy-Preserving Solution for Anomaly Detection in IoT using Federated Learning and Homomorphic Encryption. *Information Systems Frontiers*.
- Arazzi, M., Nicolazzo, S., & **Nocera, A.** (2023). A Novel IoT Trust Model Leveraging Fully Distributed Behavioral Fingerprinting and Secure Delegation. *arXiv preprint arXiv:2310.00953*.

Federated Machine Learning

Components:

- Workers (e.g., clients)
- Aggregator (e.g., server)
- Communication-computational framework to train the ML model.



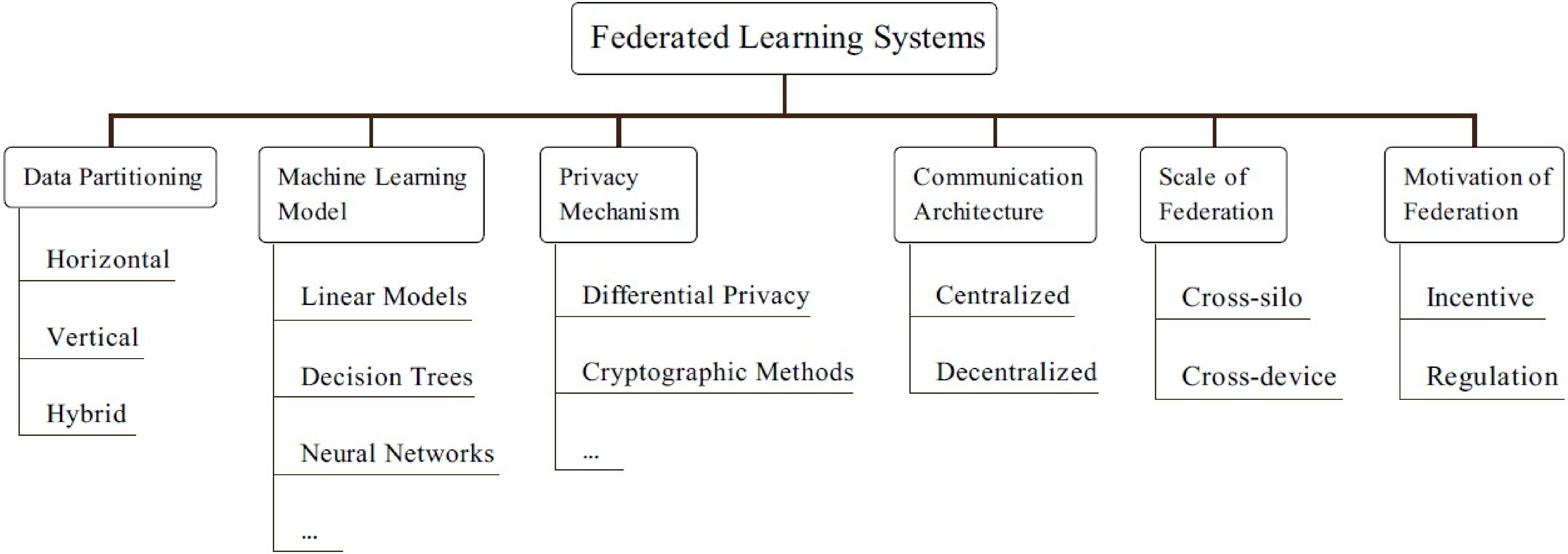
① Send the global model to the selected parties

② Update model with local data

③ Send local models to the server

④ Update the global model

Federated Machine Learning



Federated Learning Systems

Data Partitioning
- Horizontal
- Vertical
- Hybrid

Machine Learning Model
- Linear Models
- Decision Trees
- Neural Networks
- ...

Privacy Mechanism
- Differential Privacy
- Cryptographic Methods
- ...

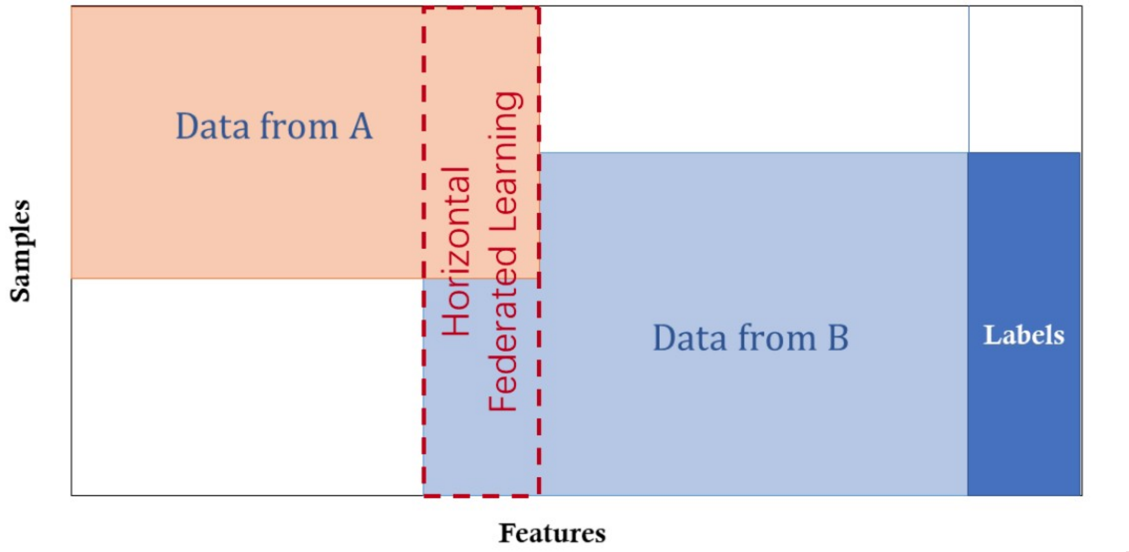
Communication Architecture
- Centralized
- Decentralized

Scale of Federation
- Cross-silo
- Cross-device

Motivation of Federation
- Incentive
- Regulation

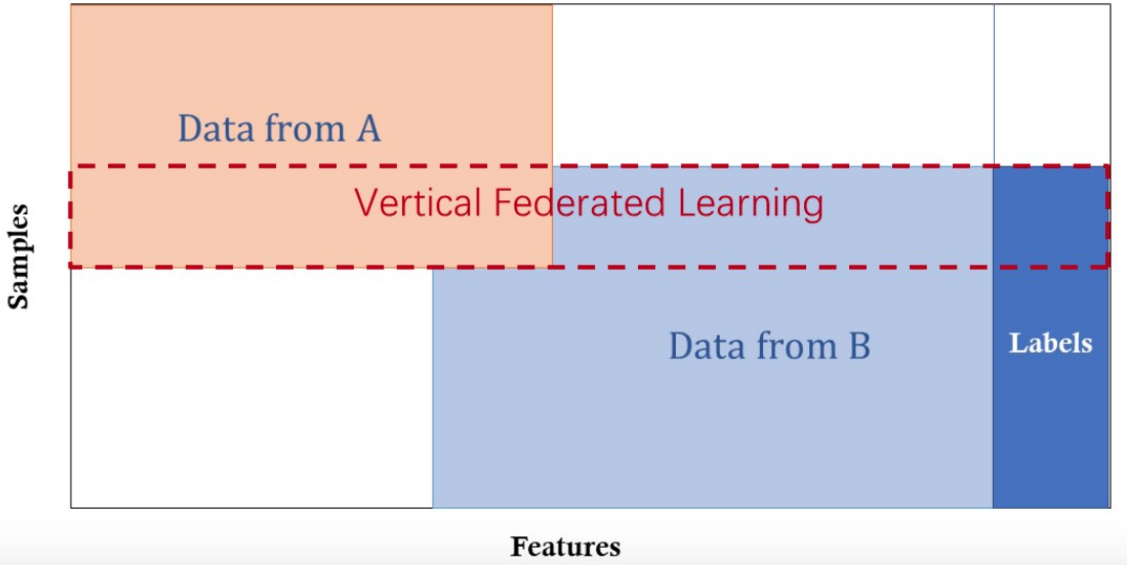
Data Partition

- **Horizontal FL:** the datasets of different parties have the same feature space but little intersection on the sample space. This is a natural data partitioning especially for the cross-device setting, where different users try to improve their model performance on the same task using FL. The global model can simply be updated by averaging all the local models.

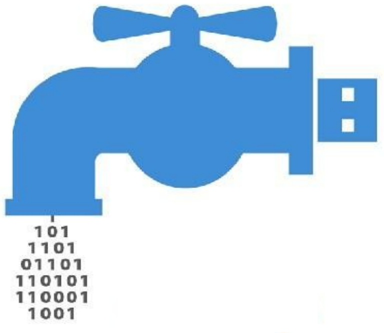


Data Partition

- **Vertical FL**: the datasets of different parties have the same or similar sample space but differ in the feature space. For the vertical FLS, it usually adopts entity alignment techniques to collect the overlapped samples of the parties. Then the overlapped data are used to train the machine learning model using encryption methods.



Security of FL



Vulnerabilities classification:

- **Server-side** vulnerabilities: the server can observe individual updates over time, tamper with the training process and control the view of the participants on the global parameters.

- **Node-side** vulnerabilities: any participant who can observe the global parameters and control its parameter uploads can alter their inputs or introduce stealthy backdoors into the global model.

Threat Models: Insider v.s. Outsider

- **Insider attacks:** launched by the FL server and the participants in the FL system.
    - ➔ Single attack: a single, non-colluding malicious participant aims to cause the model to miss-classify a set of chosen inputs with high confidence.
    - ➔ Byzantine attack: the byzantine malicious participants may behave completely arbitrarily and tailor their outputs to have similar distribution as the correct model updates.
    - ➔ Sybil attack: the adversaries can simulate multiple dummy participant accounts or select previously compromised participants to build more powerful attacks on FL
- **Outsider attacks:** launched by the eavesdroppers on the communication channel between participants and the FL server, and by users of the final FL model when it is deployed as a service.

Threat Models: Semi-honest v.s. Malicious

- **Semi-honest setting:** passive or honest-but-curious adversaries.
- **Malicious setting:** adversary tries to learn the private states of honest participants, and deviates arbitrarily from the FL protocol by modifying, re-playing, or removing messages.

Threat Models: Training Phase v.s. Inference Phase

- In **Training phase** an attacker can**:**
  - Run data poisoning attacks to compromise the integrity of training dataset.
  - Model poisoning attacks to compromise the integrity of the learning process.
  - Launch a range of inference attacks on an individual participant’s update or on the aggregate of updates from all participants.

- Attacks at **inference phase,** called evasion/exploratory attacks, collect evidence about the model characteristics. They can be classified in:
  - White-box attacks with full access to the FL model
  - Black-box attacks only able to query the FL model

Poisoning Attacks

Dr. Marco Arazzi

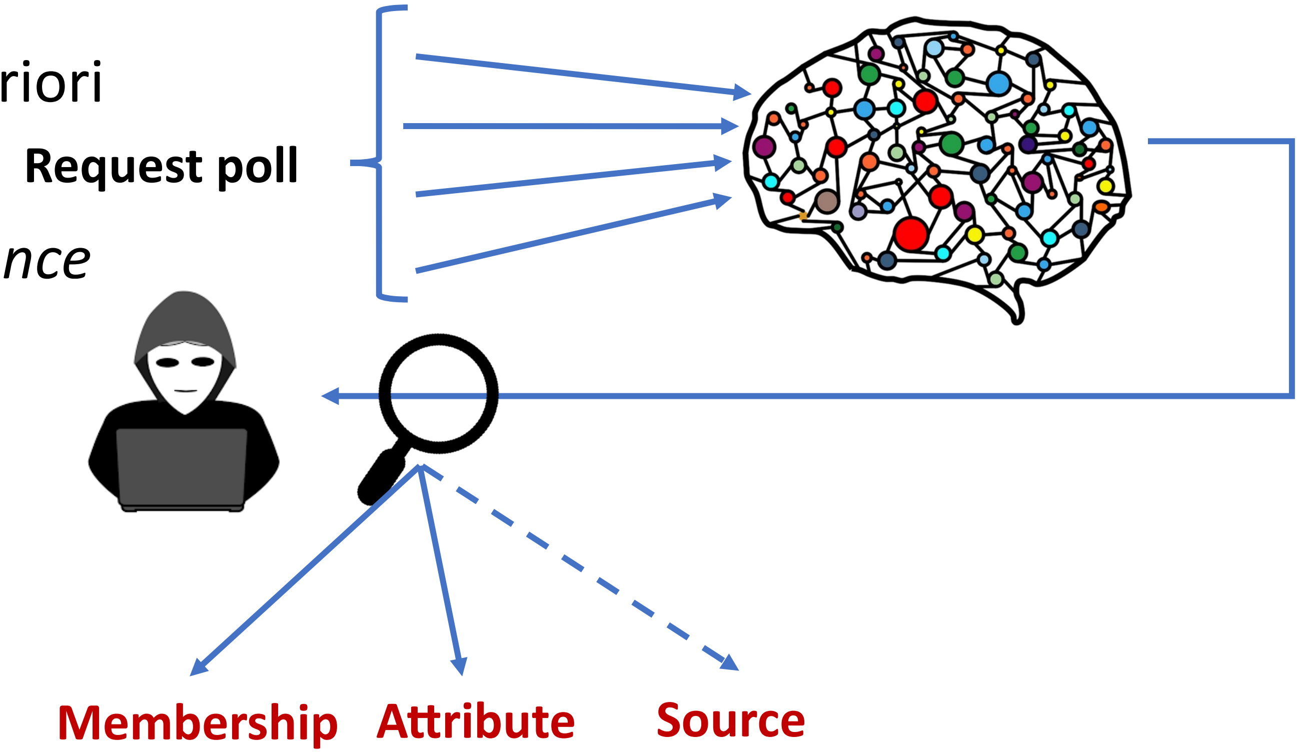
Inference Attacks

Inference Attacks

Inference Attacks

General Objective

- Leak information from the training data
- Use either traning information or a-posteriori analysis.
- For Federated Learning also *Source Inference*

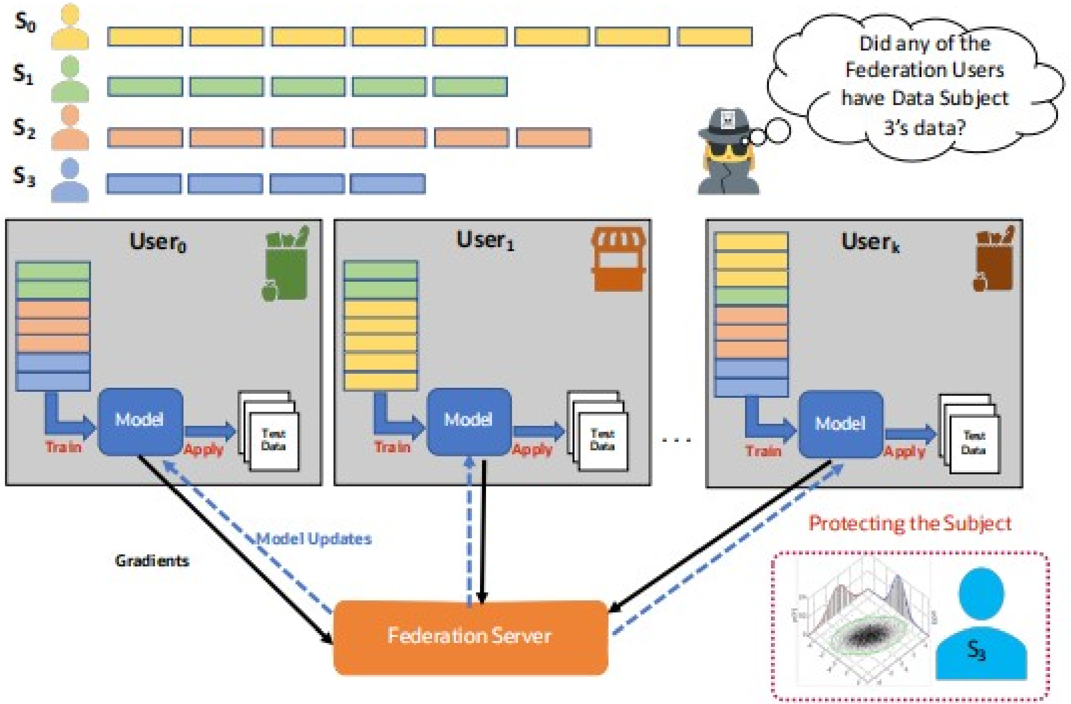


Membership Inference

General Objective

- Infer wheter a specific *data-point* has been used in the training process.
- Hypothesis:

«If data from a particular subject is present in the federation and is used in training, the global model would be expected to have a lower loss on it than data from a subject that was not present in any of the users' local datasets»



Suri, A., Kanani, P., Marathe, V. J., & Peterson, D. W. (2022). Subject membership inference attacks in federated learning. *arXiv preprint arXiv:2206.03317*.

Source Inference

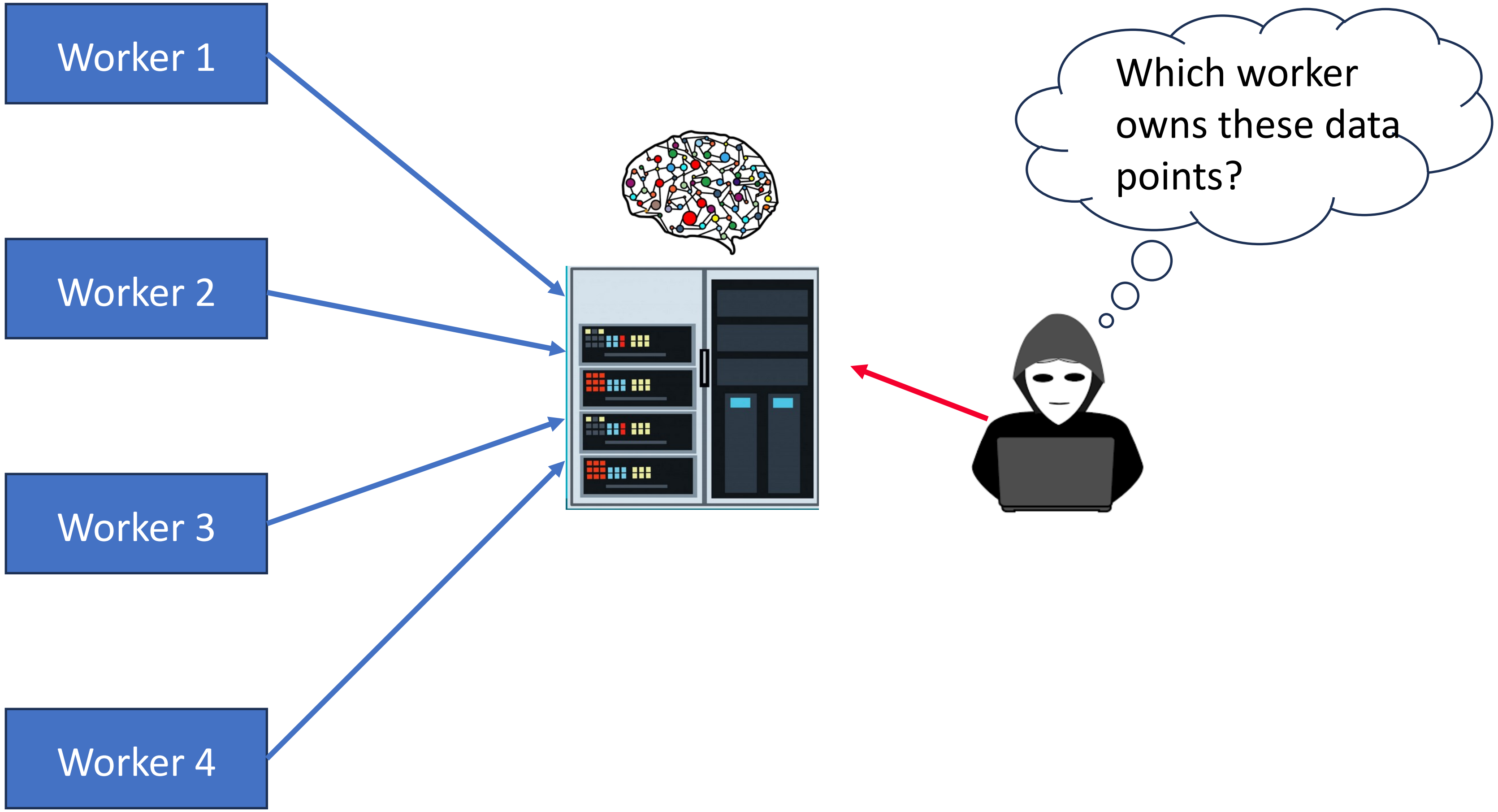
General Objective

- Infer which worker exploited a target *data-point* during the training process.

- Definition:

«Given local optimized model ϑk, a training record z1, source inference aims to infer the posterior probability of z1 belonging to the client k»

$$\mathcal{S}(\boldsymbol{\theta}_k, \boldsymbol{z}_1) := \mathbb{P}(s_{1k} = 1 | \boldsymbol{\theta}_k, \boldsymbol{z}_1)$$



Hu, H., Salcic, Z., Sun, L., Dobbie, G., & Zhang, X. (2021, December). Source inference attacks in federated learning. In *2021 IEEE International Conference on Data Mining (ICDM)* (pp. 1102-1107). IEEE.

Label (Attribute) Inference

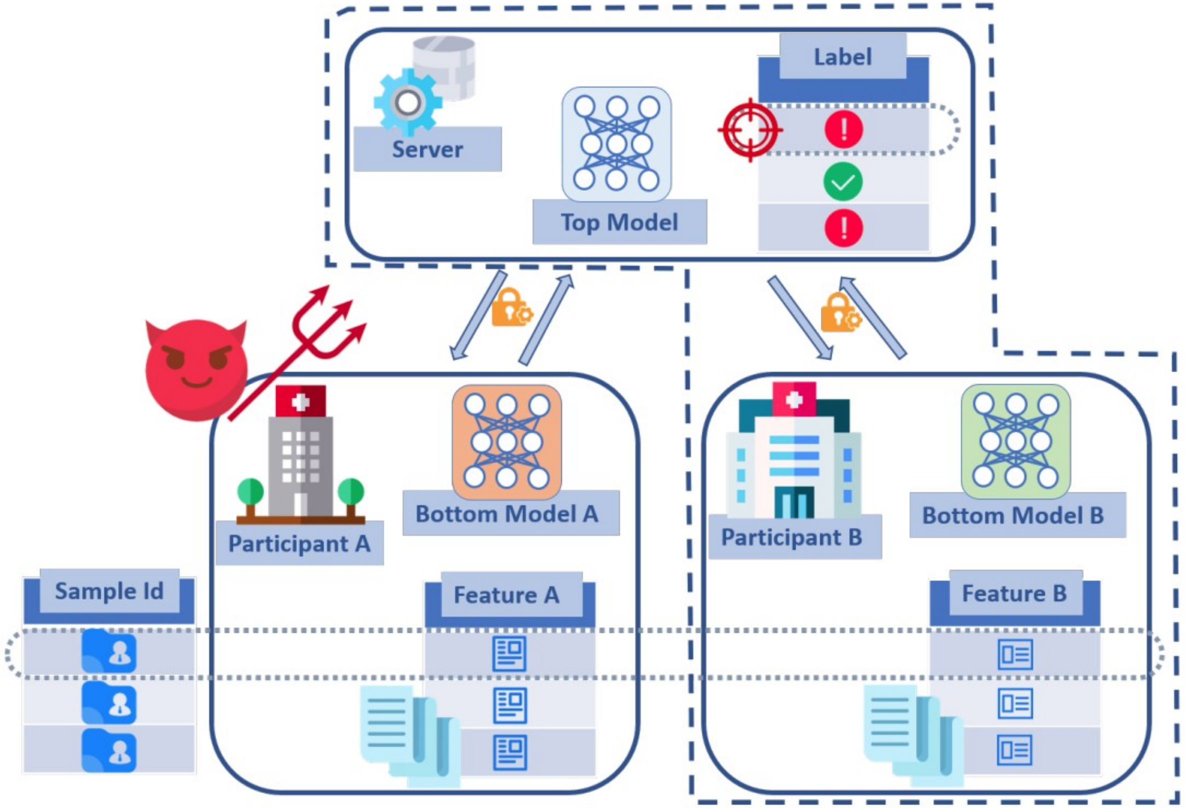
General Objective

- Evolution of Attribute Inference for VFL.

- Definition:

«In a VFL disclose the label-set known to the aggregation server.»

- Require a pre-knowledge of a (small) portion of the training data.



Fu, C., Zhang, X., Ji, S., Chen, J., Wu, J., Guo, S., ... & Wang, T. (2022). Label inference attacks against vertical federated learning. In *31st USENIX Security Symposium (USENIX Security 22)* (pp. 1397-1414).

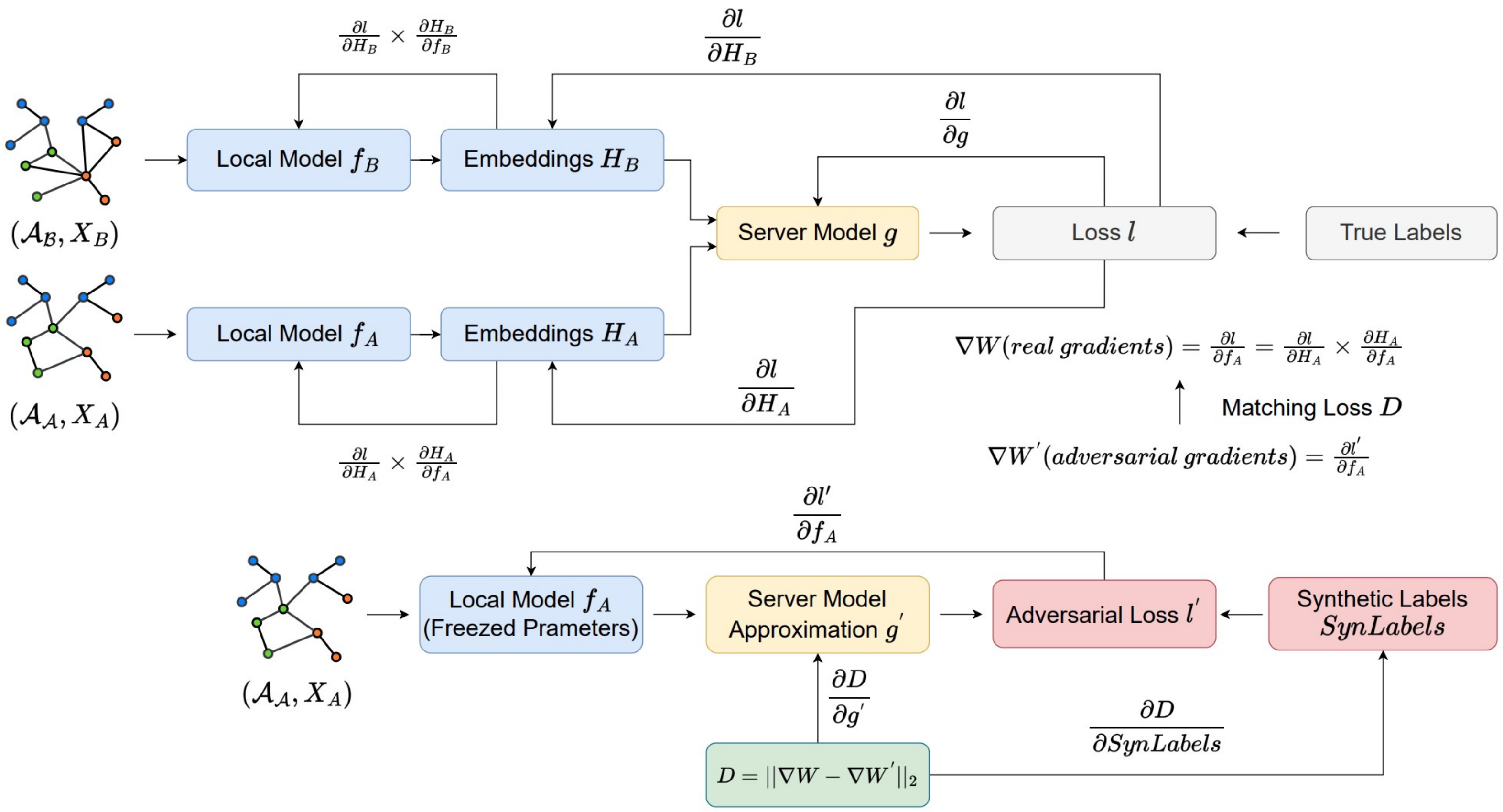
Label (Attribute) Inference

General Objective

- Perform label inference with zero background knowledge.

- Idea

«assume a (plausible) full model to be trained locally (toghether with the partial local model). Use maching loss to learn possible labels.»



$\frac{\partial l}{\partial H_B} \times \frac{\partial H_B}{\partial f_B}$
$\frac{\partial l}{\partial H_B}$
$\frac{\partial l}{\partial g}$
$(\mathcal{A}_B, X_B)$
Local Model f_B
Embeddings H_B
Server Model g
Loss l
True Labels
$(\mathcal{A}_A, X_A)$
Local Model f_A
Embeddings H_A
$\frac{\partial l}{\partial H_A}$
$\nabla W(real\ gradients) = \frac{\partial l}{\partial f_A} = \frac{\partial l}{\partial H_A} \times \frac{\partial H_A}{\partial f_A}$
Matching Loss D
$\nabla W'(adversarial\ gradients) = \frac{\partial l'}{\partial f_A}$
$\frac{\partial l}{\partial H_A} \times \frac{\partial H_A}{\partial f_A}$
$\frac{\partial l'}{\partial f_A}$
$(\mathcal{A}_A, X_A)$
Local Model f_A (Freezed Prameters)
Server Model Approximation g'
Adversarial Loss l'
Synthetic Labels $SynLabels$
$\frac{\partial D}{\partial g'}$
$\frac{\partial D}{\partial SynLabels}$
$D = ||\nabla W - \nabla W'||_2$

Arazzi, M., **Conti, M.**, Koffas, S., Krcek, M., **Nocera, A.**, Picek, S., & Xu, J. (2023). BlindSage: Label Inference Attacks against Node-level Vertical Federated Graph Neural Networks. *arXiv preprint arXiv:2308.02465*.

Label (Attribute) Inference

Defense??

- Use Secret Sharing to protect gradient distribution

Wang, Y., Lv, Q., Zhang, H., Zhao, M., Sun, Y., Ran, L., & Li, T. (2023). Beyond model splitting: Preventing label inference attacks in vertical federated learning with dispersed training. *World Wide Web*, 1-17.

✓

BUT
All the worker must use a secret sharing scheme…

- Can we do any better?

What if…only the server plays against the attacker

Thank you very much!

Not tired yet? 🥱
Do you want to know more?

✉ antonino.nocera@unipv.it

Prof. Antonino Nocera

Dr. Marco Arazzi

dcalab

UNIVERSITÀ
DI PAVIA