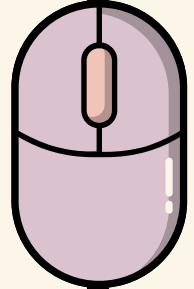




PYOD: Pwn Your Own Device



De Simoni Clarissa, Di Giovanni Roberto, Spezia Nicolò





All'interno della presentazione:

01

Introduzione

a Machine Learning e
Federated Learning

02

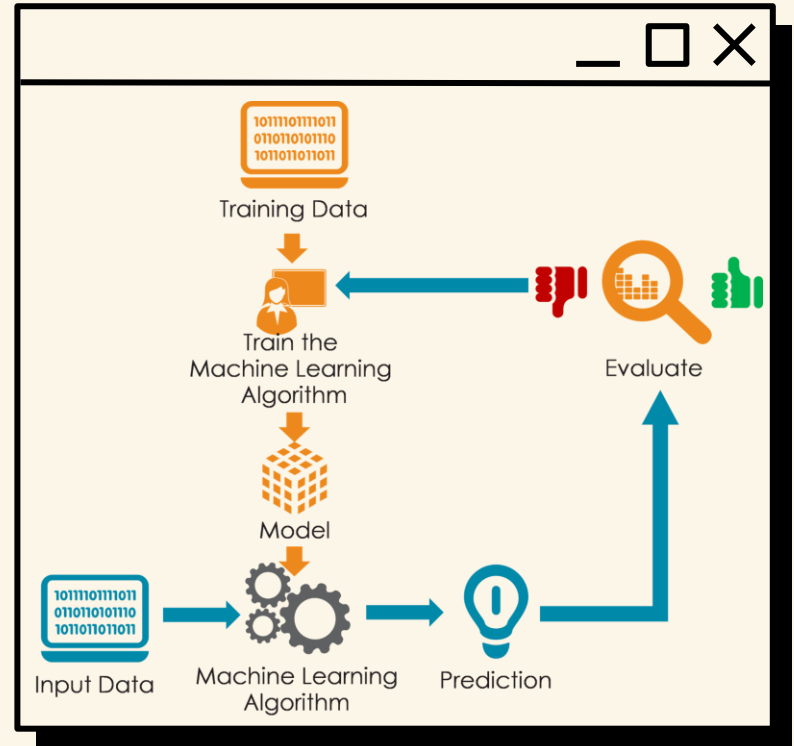
Minacce e vulnerabilità

03

Introduzione agli attacchi

Machine Learning

È il processo di sviluppo e creazione di algoritmi e modelli capaci di apprendere attraverso l'addestramento su un dataset, per poi applicare la conoscenza acquisita su dati non ancora visti.



Problematiche di ML

Privacy: i dati devono essere inviati
a un server centrale

Sicurezza: rischio di attacchi
informatici e single point of failure

Scalabilità: elevati costi di
trasferimento e archiviazione

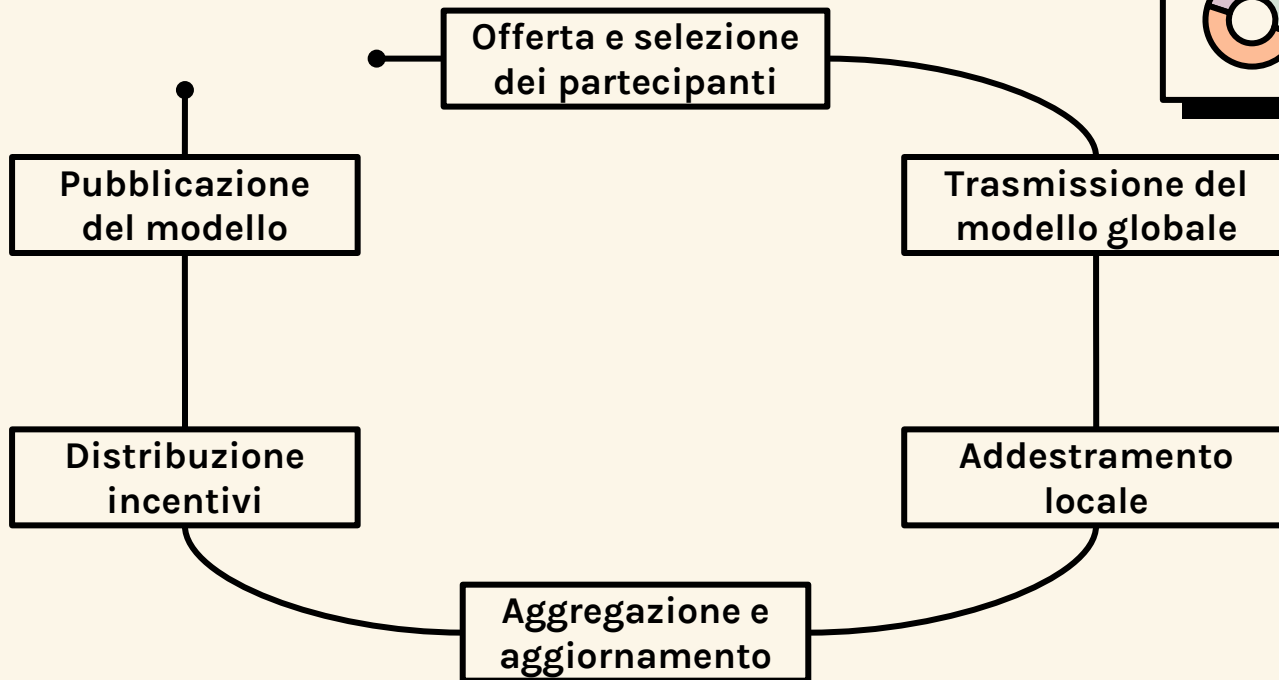
Generalizzazione: modelli
meno adatti a contesti distribuiti e
diversificati

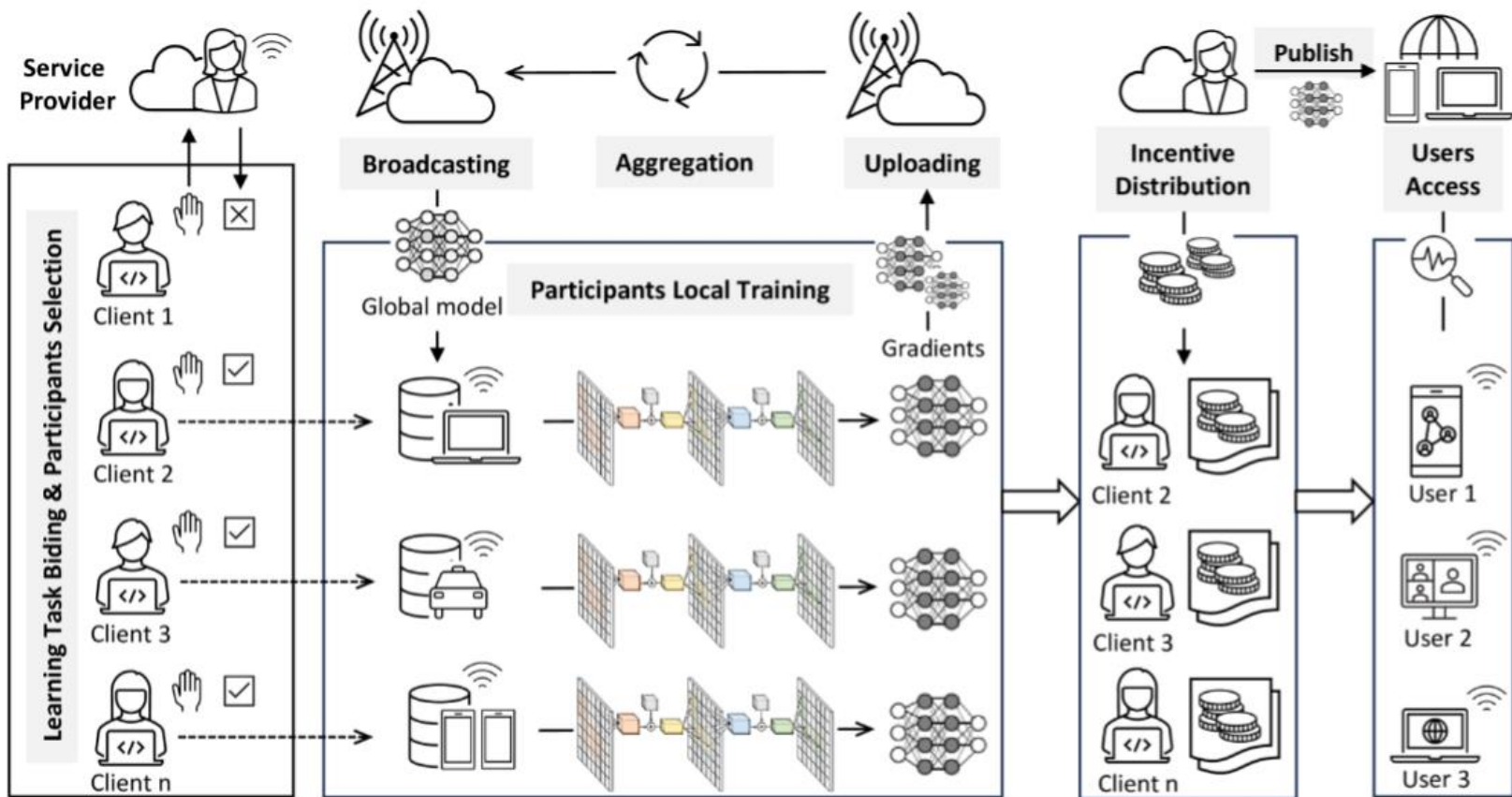


Federated Learning

Il Federated Learning è un approccio di Machine Learning in cui **l'addestramento del modello avviene direttamente sui dispositivi degli utenti**, mantenendo i dati in locale. Invece di centralizzare i dati, **solo gli aggiornamenti** (ad esempio, i gradienti) vengono **inviati a un server centrale** che li aggrega per formare un modello globale. Questo metodo **migliora la privacy** e **riduce il traffico dati**, permettendo di addestrare modelli efficaci **senza compromettere le informazioni sensibili**.

Ciclo di vita di FL



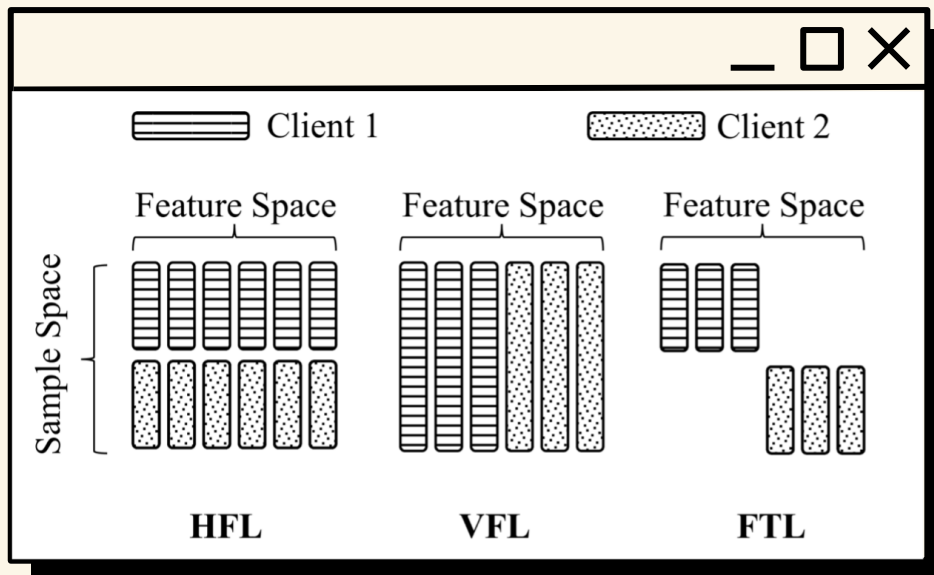


Classificazione dei FL

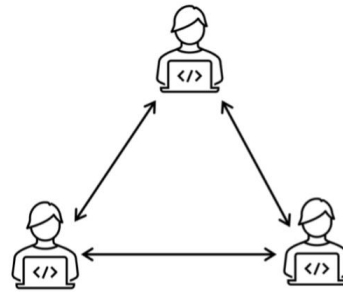
HFL - Horizontal Federated Learning

VFL - Vertical Federated Learning

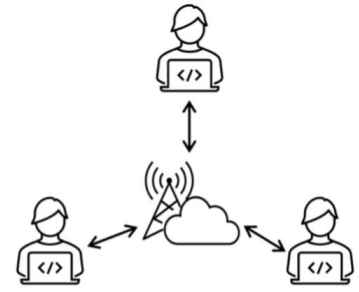
FTL - Federated Transfer Learning



DFL - Decentralized Federated Learning
CFL - Centralized Transfer Learning



DFL



CFL



Minacce e Vulnerabilità

Il Federated Learning è vulnerabile a diverse **minacce alla sicurezza e alla privacy a causa della sua natura distribuita**.

Le minacce possono provenire da **attori malevoli interni** (Insiders) o **esterni** (Outsiders), con obiettivi che vanno dal furto di dati alla manipolazione dell'addestramento. Le vulnerabilità emergono a causa della trasmissione di parametri dei modelli, della mancanza di un controllo centralizzato e della difficoltà nel garantire l'integrità dei contributi dei client.



Vulnerabilità dei FL

**Scarsa protezione
della privacy**

Scalabilità limitata

**Dipendenza da
server centrali**

Dipendenza da server centrali

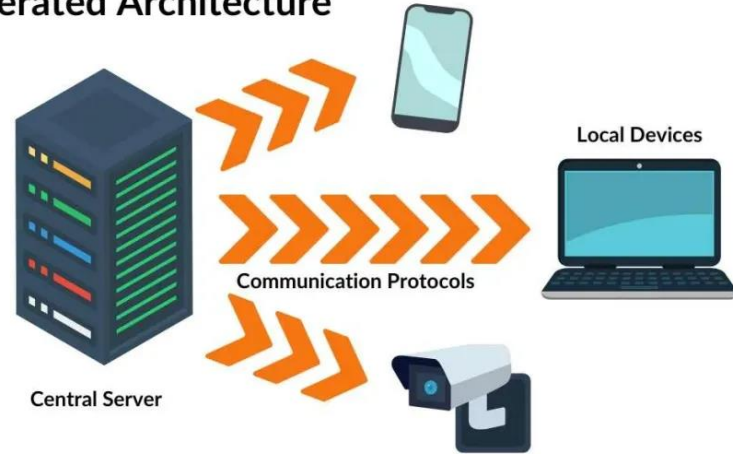
Server malevolo:

- Attacco attivo
- Attacco passivo

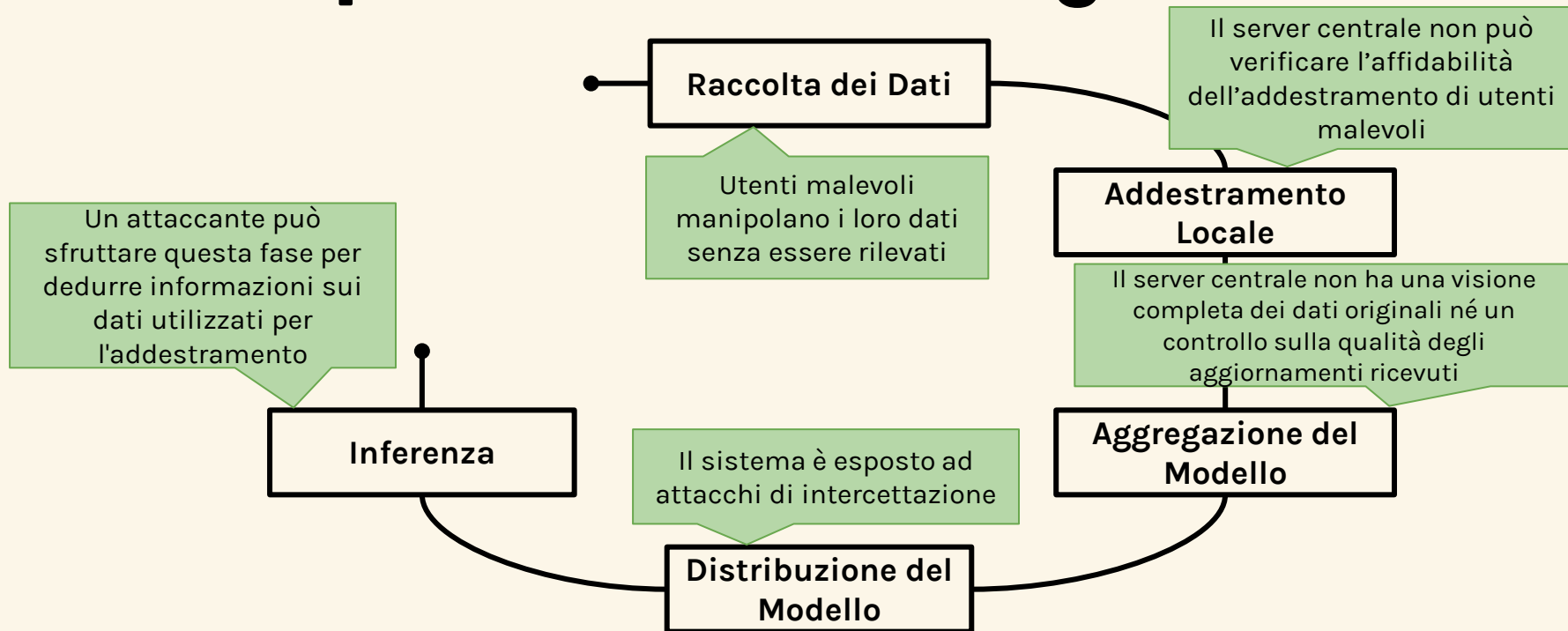
Protocolli e canali di comunicazione:

- Eavesdropping
- Man-in-the-middle
- Attacchi di replay

Federated Architecture



Quando possono avvenire gli attacchi





Possibili attacchi

Tipi di attacco

Privacy

Mirano ad estrarre informazioni sensibili dai dati locali dei partecipanti

Sicurezza del Modello

Tentano di manipolare il modello in modo malevolo

Robustezza e Disponibilità

Puntano a ridurre l'efficienza o l'integrità dell'apprendimento federato



Attacchi alla Privacy

Inference Attack

Un **Inference Attack** si riferisce a una situazione in cui un attaccante cerca di **inferire informazioni** sensibili riguardanti i dati utilizzati per addestrare il modello. In altre parole, l'attaccante **cerca di ottenere informazioni** sulle **caratteristiche** o i **dettagli specifici dei dati di addestramento**, partendo dalle previsioni o dai risultati del modello.

Attacchi alla sicurezza del modello

Data poisoning

L'attacco colpisce i dati di addestramento di ciascun client malevolo, falsificandoli per indurre il modello locale a imparare informazioni distorte. Quando questi modelli alterati vengono inviati al server centrale per l'aggregazione, il modello globale eredita e amplifica tali anomalie.

Model poisoning

Questo attacco agisce sugli aggiornamenti del modello inviati dai client al server, manipolando pesi o gradienti per distorcere l'aggregazione globale.

Attacchi alla robustezza e disponibilità

DoS Attack

Un **attacco DoS (Denial of Service)** mira a **interrompere o rallentare l'addestramento** del modello globale, **sovraccaricando il server centrale** o i client con richieste eccessive o aggiornamenti malformati. L'attaccante può inondare il server con aggiornamenti ripetitivi o inutili, esaurendo le risorse di elaborazione o la larghezza di banda. Gli attacchi DoS possono causare ritardi nell'addestramento, interruzioni del servizio e possibili danni alle prestazioni del modello federato.

Tipo = Attacco

HFL

Attacchi a gradienti

Model poisoning

VFL

Data leakage

Inference attacks

FTL

Poisoning Attack via
Transfer Learning

Malicious Knowledge
Transfer



CFL

Server Poisoning Attack

Model Inversion

DFL

Sybil Attack

Byzantine Attack

Eavesdropping and
Reverse Engineering

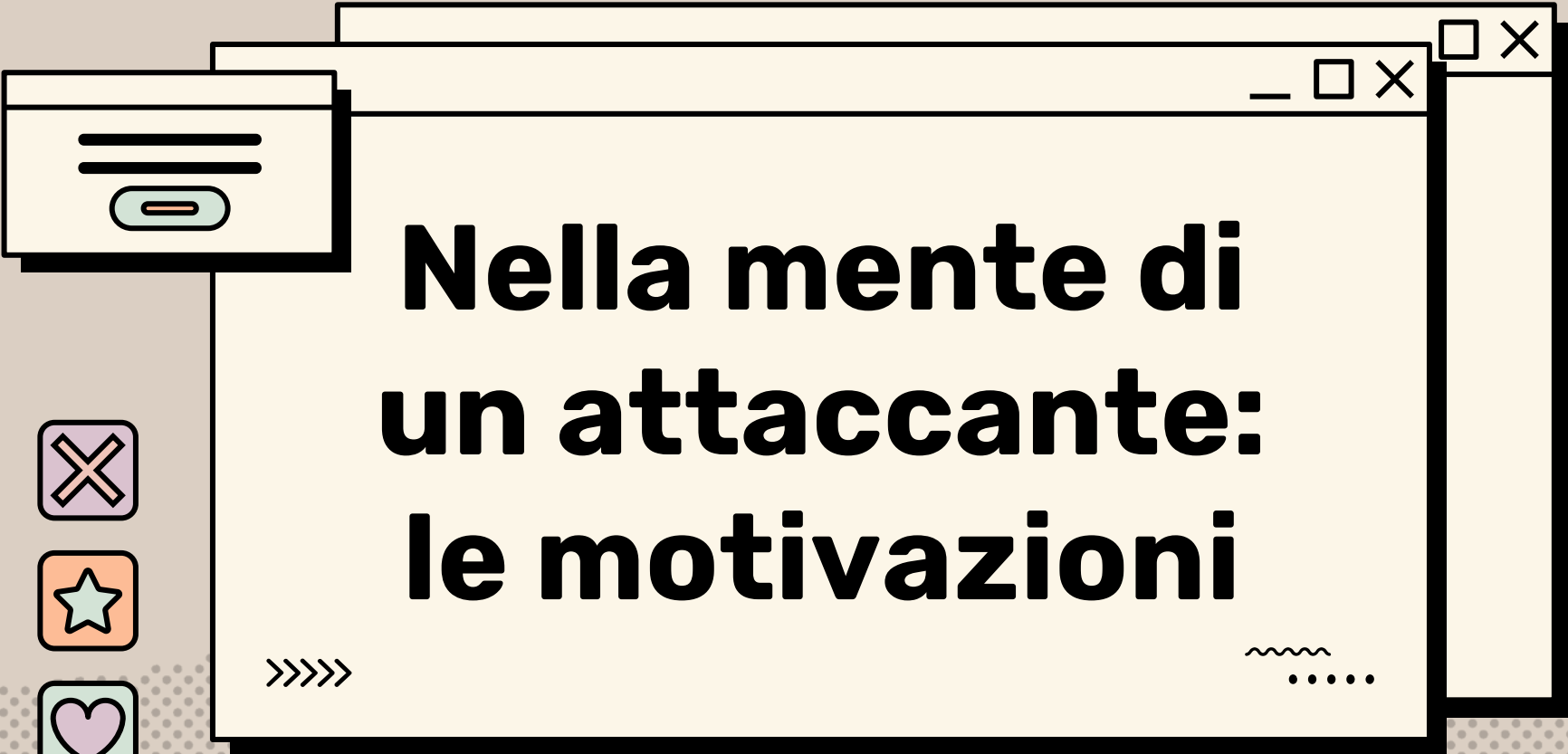


La parte divertente

04 Nella mente di
un attaccante

05 Gli attacchi
da noi simulati

06 Le difese
contro gli attacchi
simulati



The image features a stylized, hand-drawn interface. On the left is a vertical sidebar with three icons: a purple square with a white 'X', an orange square with a white star, and a light green square with a white heart. Above the main content area is a horizontal bar with three horizontal lines and a pill-shaped button with an orange bar. The main content area is a large white rectangle with a black border, containing the title text. To the right of the main area is a vertical white bar with a black border. The background is a light gray with a fine dot pattern.

Nella mente di un attaccante: le motivazioni

>>>>

~~~~~  
.....



**Sabotaggio**

**Concorrenza sleale**

**Vulnerabilità dei  
Sistemi di  
Apprendimento**

**Privacy e Spionaggio**

**Guadagno finanziario**

# **Gli attacchi implementati**



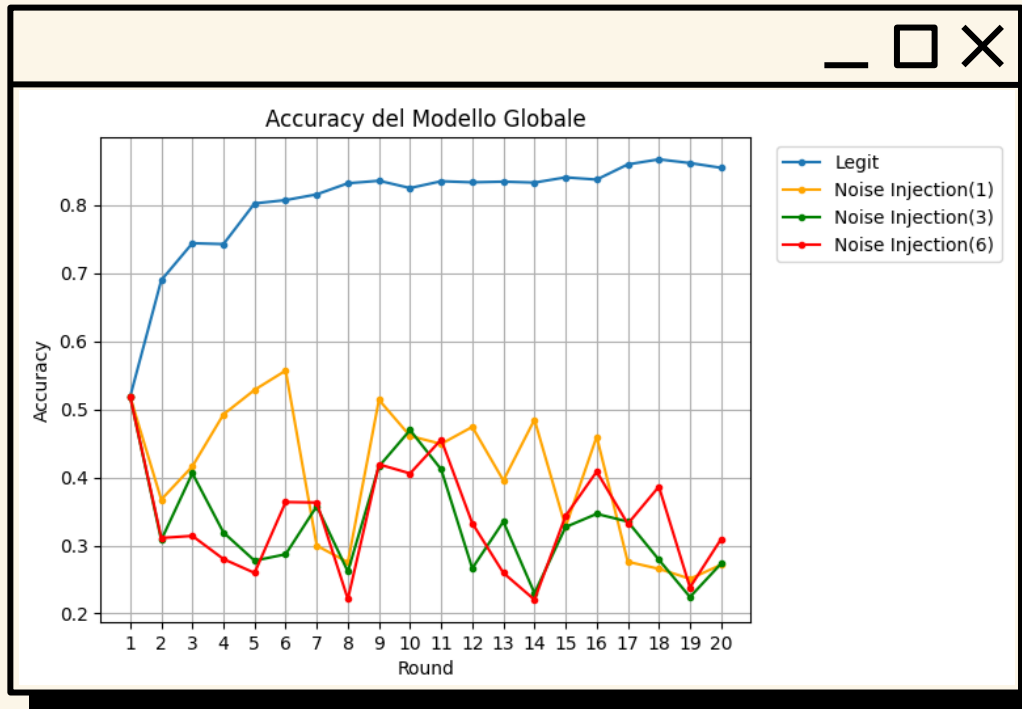




# Noise Attack

Il **Noise Attack** nel Federated Learning è un tipo di attacco in cui un partecipante malevolo **inietta deliberatamente rumore** nei propri aggiornamenti locali. Invece di contribuire in modo veritiero al modello globale, **l'attaccante aggiunge perturbazioni** che, una volta aggregate, **distorcono il processo di apprendimento complessivo**. L'obiettivo è **degradare le prestazioni** del modello finale o **rallentare la sua convergenza**, sfruttando la natura distribuita del sistema.

# I nostri risultati



# Come difendersi

- Data Augmentation / Data Sanitization
- Adversarial Training
- Anomaly Detection
- Robust Aggregation

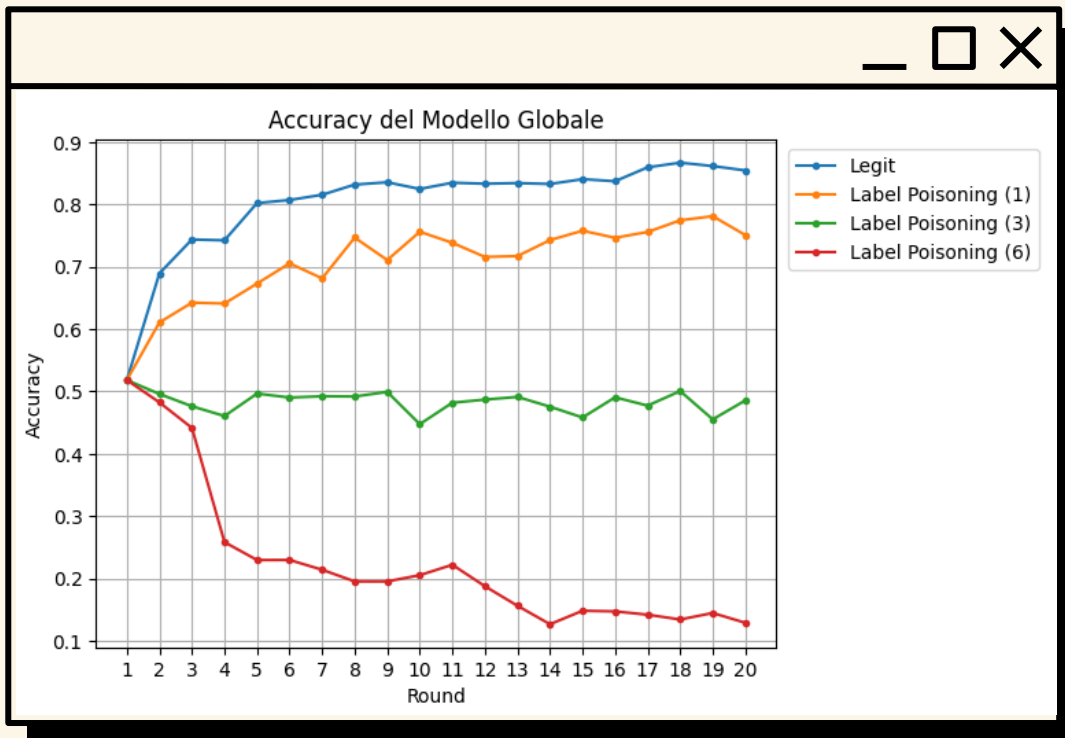


# Label Poisoning

Il **Label Poisoning** è un attacco mirato in cui le **etichette dei dati di addestramento** vengono **alterate in modo malevolo**, ad esempio scambiando le etichette di due classi. Questo attacco è **particolarmente efficace** nel Federated Learning, dove i dati sono distribuiti tra più nodi e non accessibili centralmente.

L'**obiettivo** è **corrompere il modello**, inducendolo a fare **previsioni errate** su specifici pattern.

# I nostri risultati





# Come difendersi

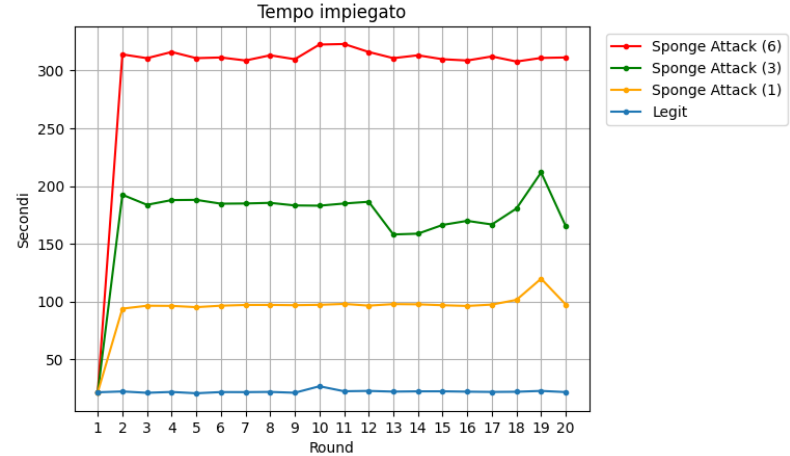
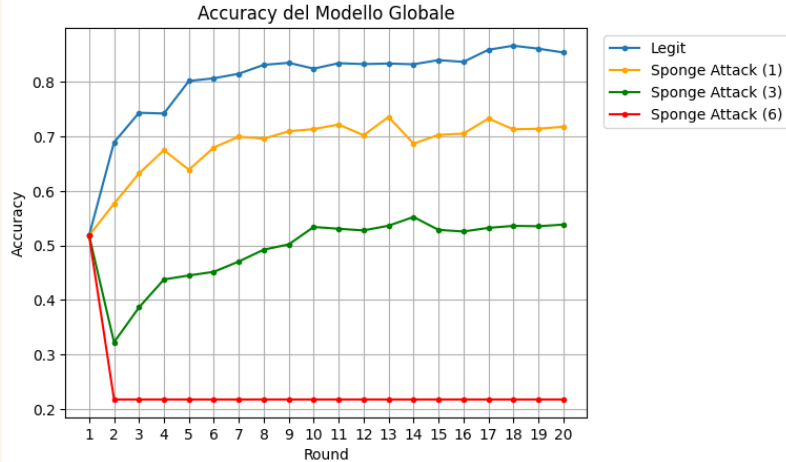
- Robust Aggregation
- Anomaly Detection
- Differential Privacy
- Client Weighting



# Sponge Attack

Lo **Sponge Attack** è un attacco in cui vengono **manipolati i dati** o il **modello** per **aumentare il consumo di risorse computazionali**, come tempo di calcolo o memoria. Nel Federated Learning, ciò può rallentare l'addestramento globale o esaurire le risorse dei dispositivi coinvolti. L'obiettivo è **compromettere l'efficienza del sistema**, spesso senza alterare direttamente le prestazioni del modello.

# I nostri risultati







# Come difendersi

- Robust Aggregation
- Differential Privacy
- Anomaly Detection
- Cross-Validation



**Grazie per  
l'attenzione**