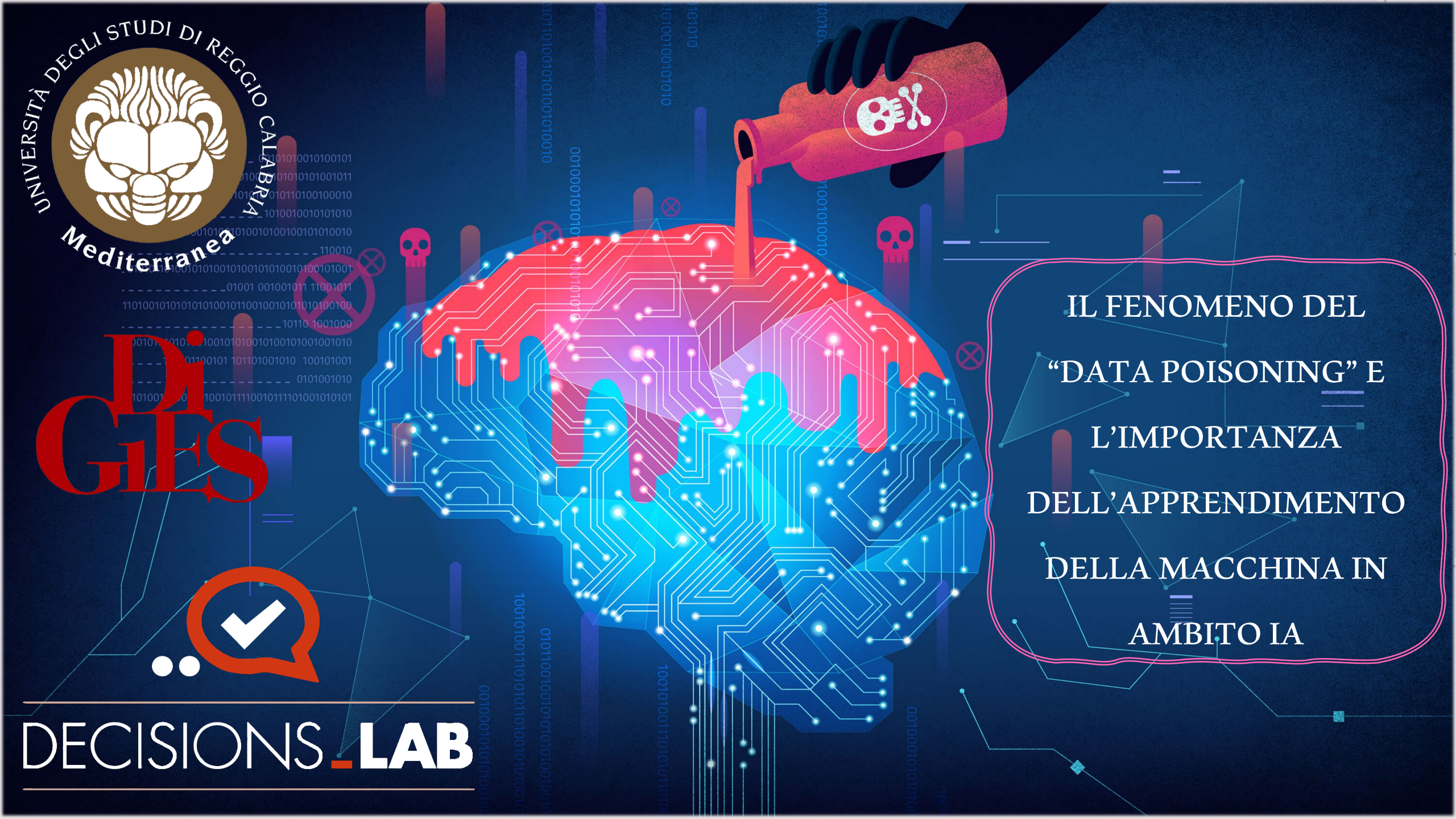
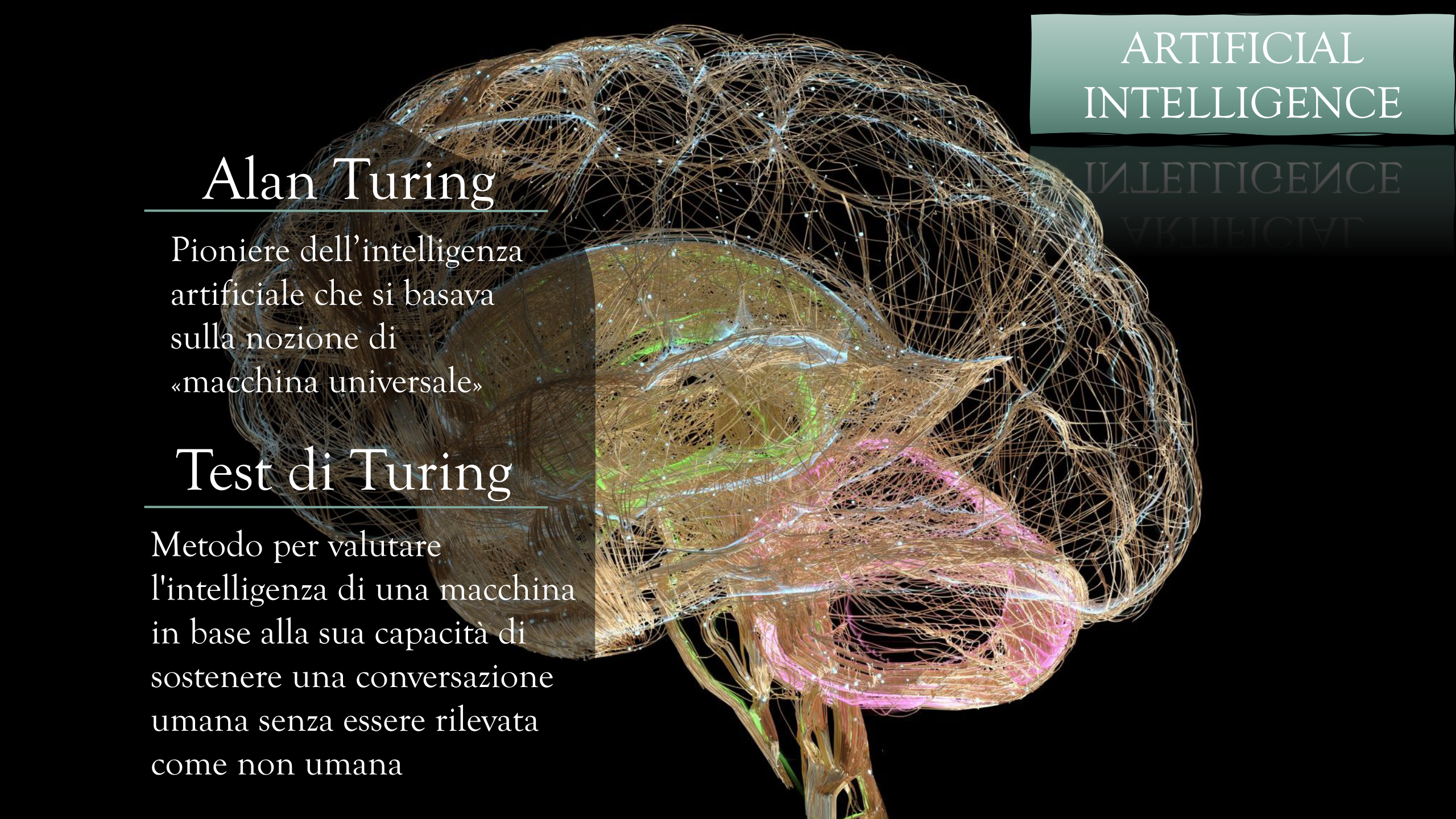

UNIVERSITÀ DEGLI STUDI DI REGGIO CALABRIA
Mediterranea
DiGES
DECISIONS.LAB
IL FENOMENO DEL
“DATA POISONING” E
L’IMPORTANZA
DELL’APPRENDIMENTO
DELLA MACCHINA IN
AMBITO IA

ARTIFICIAL INTELLIGENCE

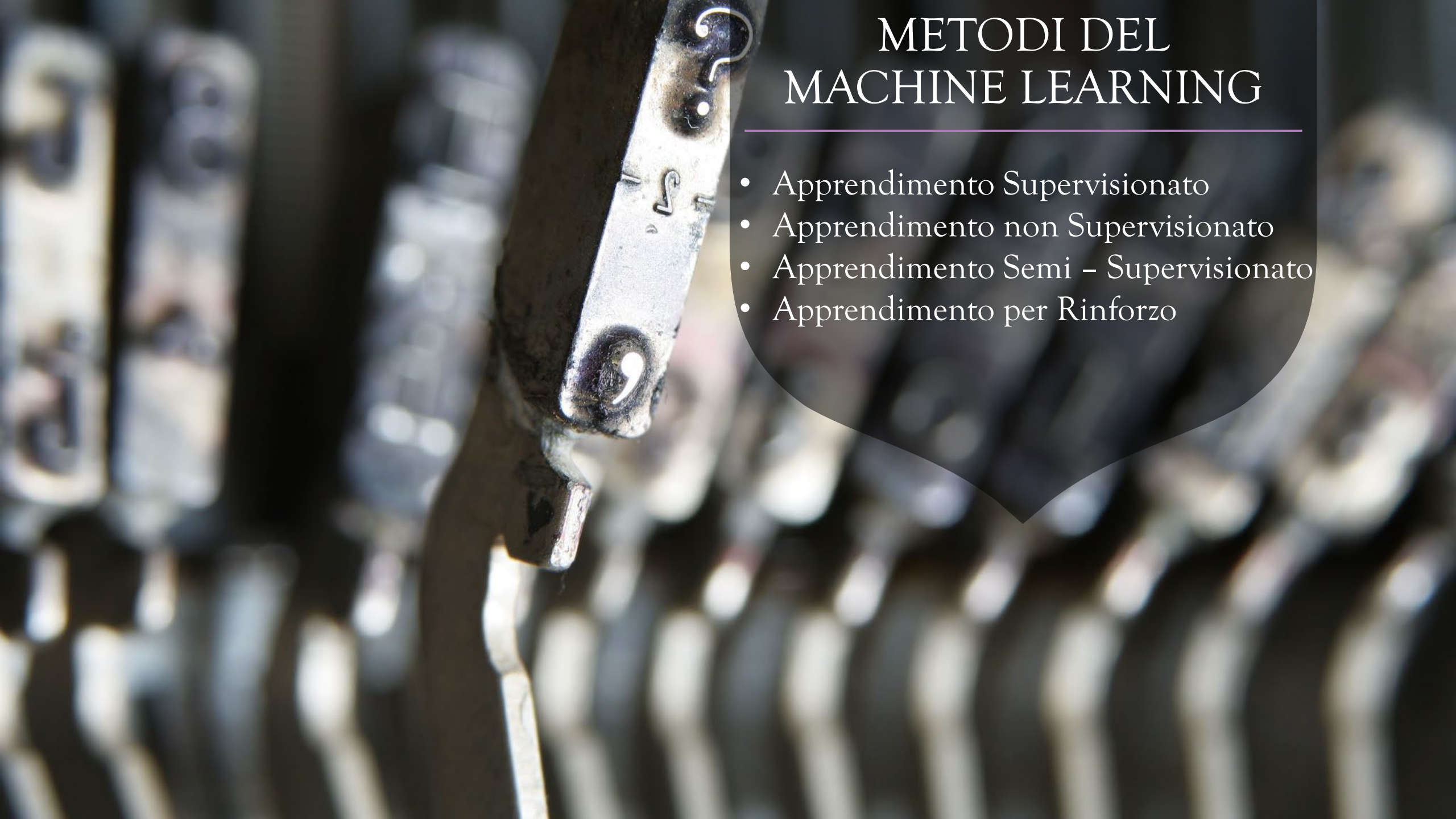


Alan Turing

Pioniere dell'intelligenza artificiale che si basava sulla nozione di «macchina universale»

Test di Turing

Metodo per valutare l'intelligenza di una macchina in base alla sua capacità di sostenere una conversazione umana senza essere rilevata come non umana



METODI DEL MACHINE LEARNING

- Apprendimento Supervisionato
- Apprendimento non Supervisionato
- Apprendimento Semi – Supervisionato
- Apprendimento per Rinforzo

ALGORITMI DI MACHINE LEARNING

- Naïve Bayes Classifier
- Regressione Logistica
- K-Nearest Neighbours
- Random Forest
- Support Vector Machine
- Decision Tree

DATA POISONING

Tecnica di attacco che consiste nel manipolare i dati utilizzati per addestrare un modello di machine learning al fine di influenzarne le prestazioni o compromettere la sicurezza del sistema

ATTACCHI DEL DATA POISONING

2 GRUPPI:

1. Quelli che mirano a ridurre l'accuratezza complessiva del modello.

2. Quelli che mirano all'integrità del modello e, in particolare dei dati. Nel caso in cui i dati vengono presi da internet ci possono essere ulteriori tipi di attacchi:
    a. Split-view poisoning
    b. Front-running poisoning

FORME DEL DATA POISONING

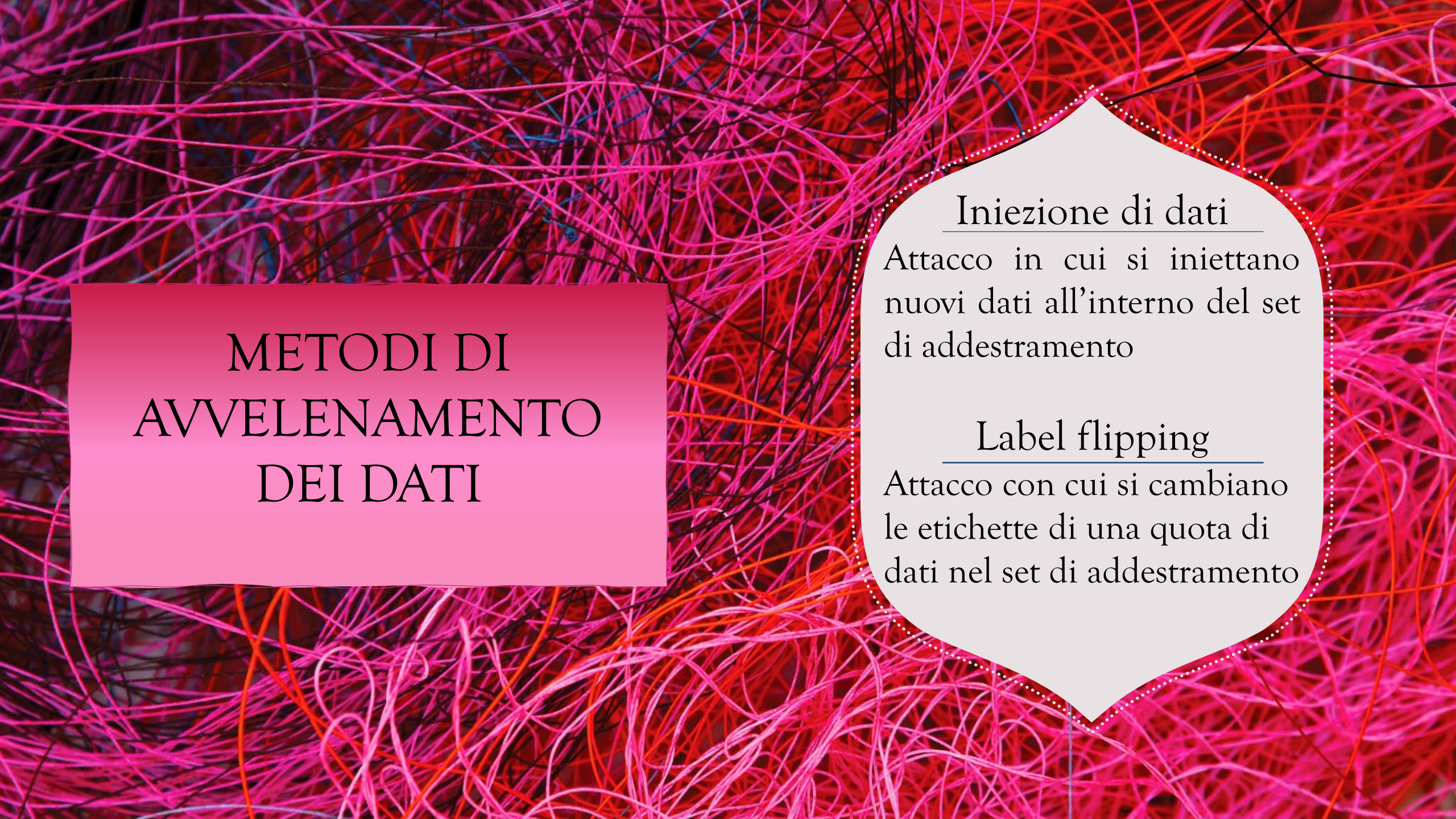
SOTTO-OBIETTIVI DI AVVELENAMENTO DATI

DENIAL OF SERVICE

L'obiettivo è ridurre le prestazioni di un modello bersaglio nel suo complesso.

ATTACCO BACKDOOR:

L'obiettivo è ridurre le prestazioni o forzare previsioni specifiche ed errate per un input o un insieme di input selezionati.



METODI DI AVVELENAMENTO DEI DATI

Iniezione di dati

Attacco in cui si iniettano nuovi dati all'interno del set di addestramento

Label flipping

Attacco con cui si cambiano le etichette di una quota di dati nel set di addestramento

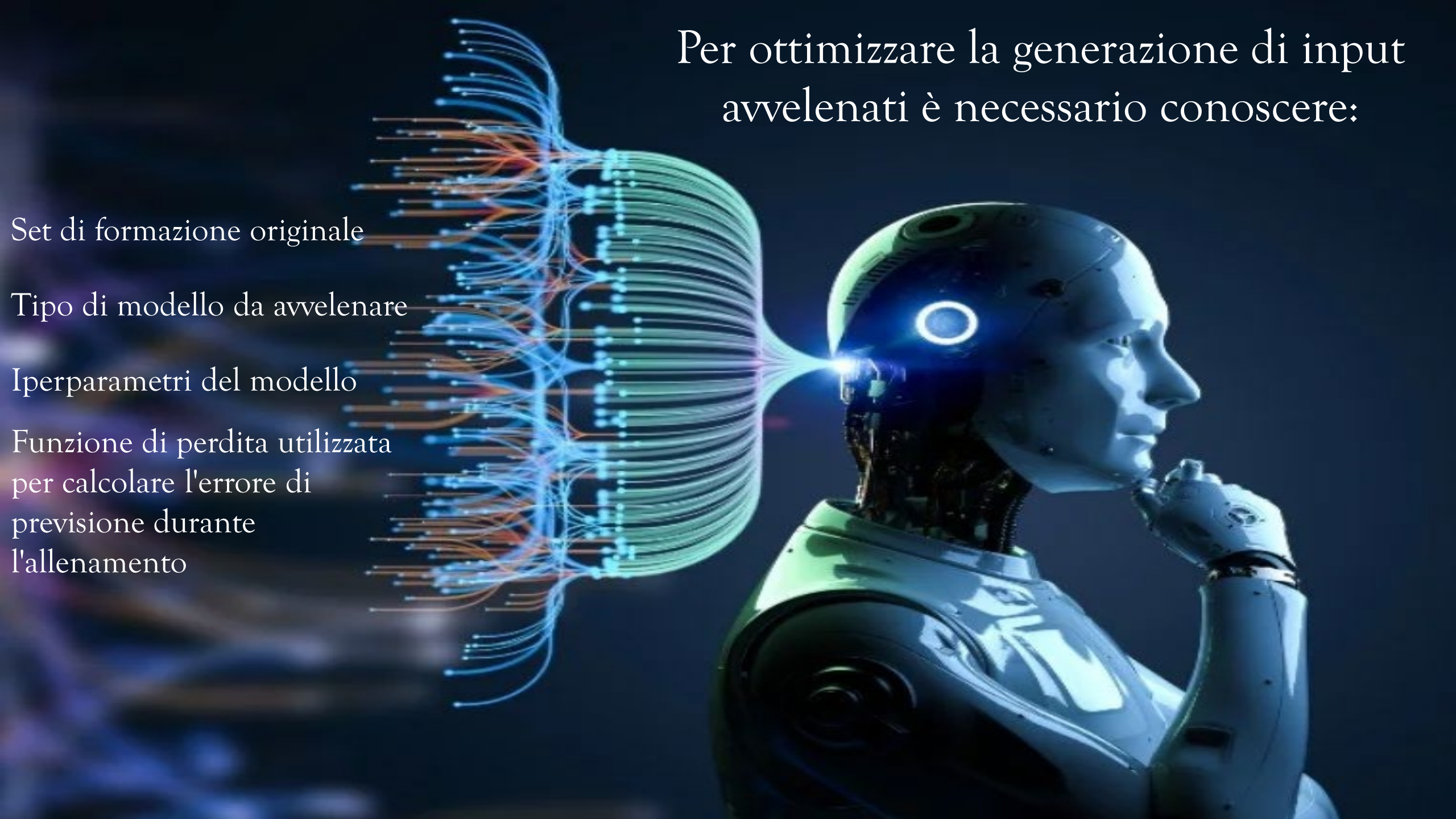
Per ottimizzare la generazione di input avvelenati è necessario conoscere:

Set di formazione originale

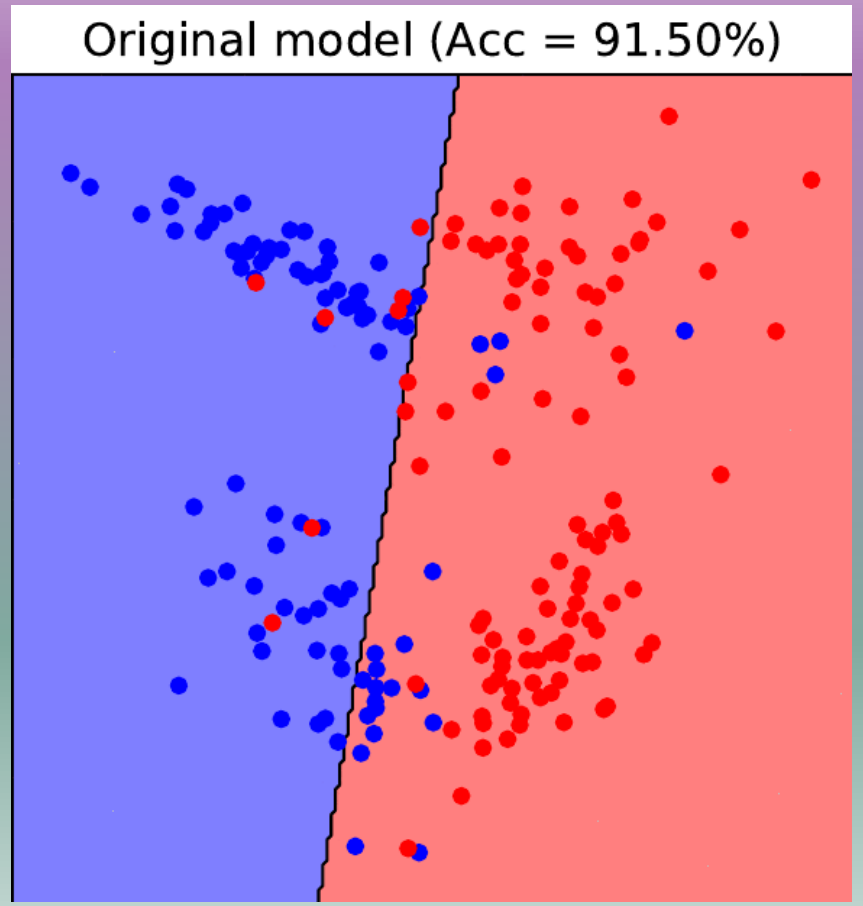
Tipo di modello da avvelenare

Iperparametri del modello

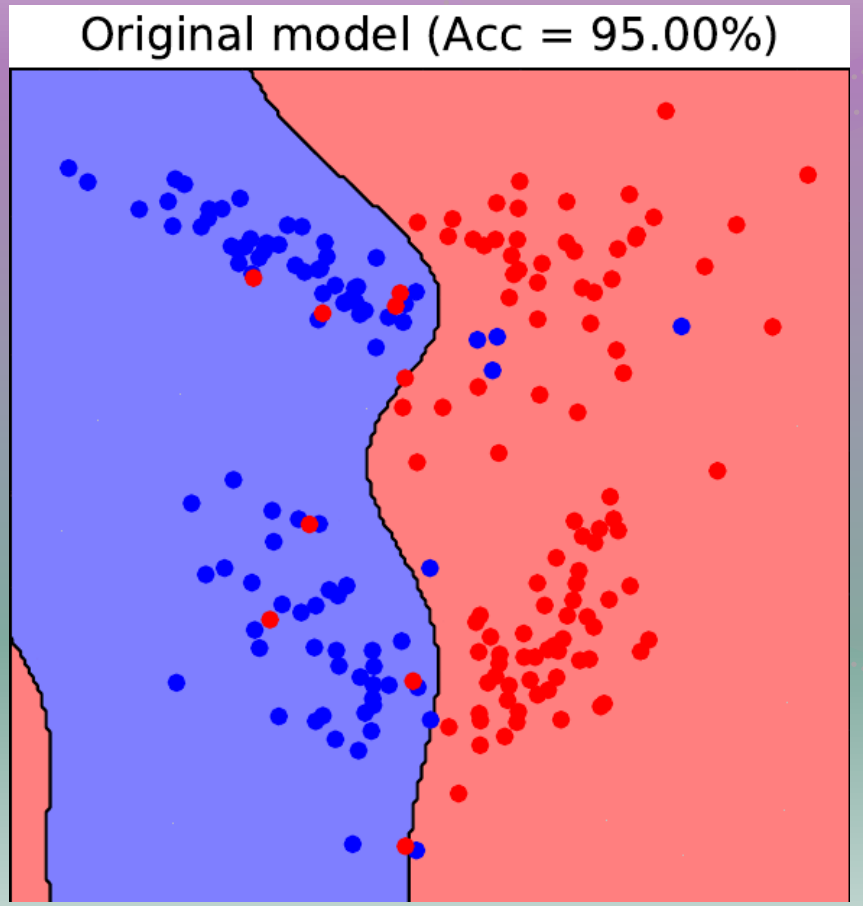
Funzione di perdita utilizzata per calcolare l'errore di previsione durante l'allenamento



ESEMPIO DI ATTACCO DENIAL OF SERVICE

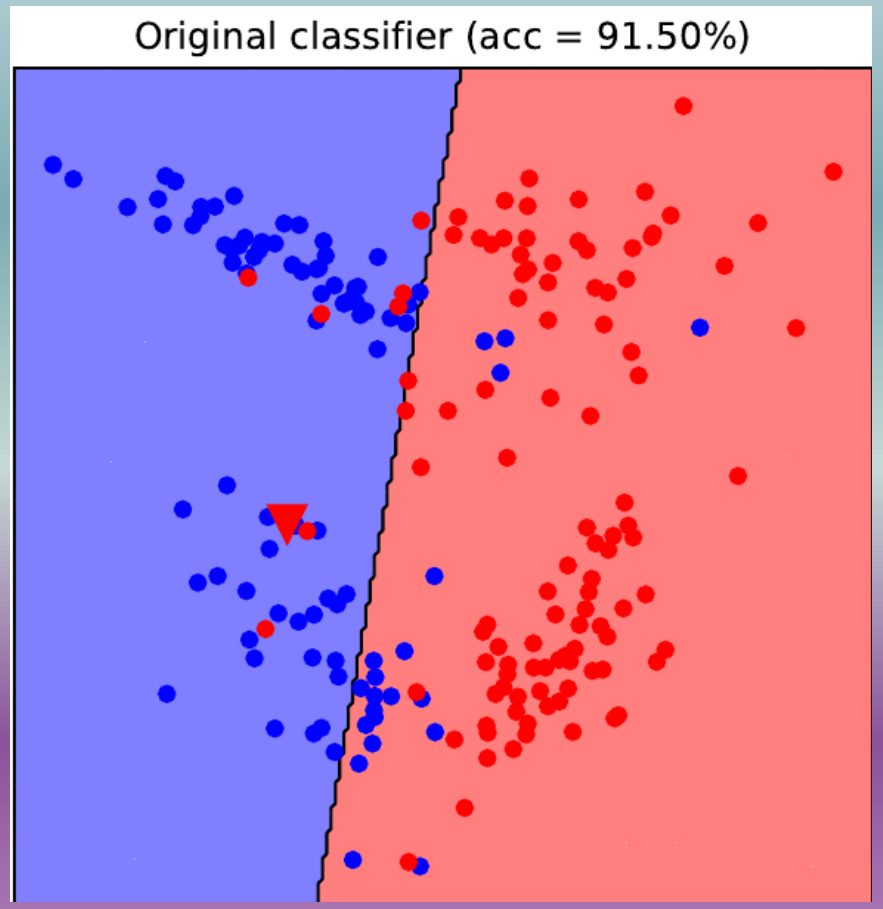


Attacco DoS contro LR

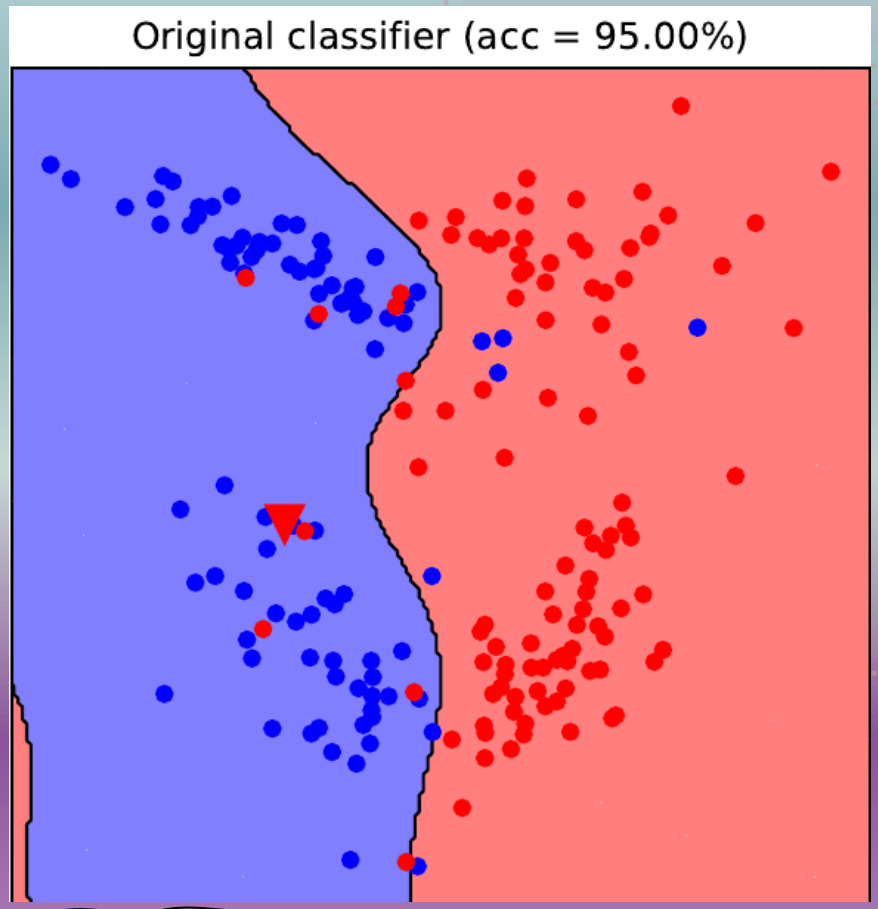


Attacco DoS contro SVM

ESEMPIO DI ATTACCO BACKDOOR



Attacco BD contro LR



Attacco BD contro SVM

PROPOSTE PER CONTRASTARE GLI ATTACCHI

Adozione di tecniche di rilevamento e filtraggio dei dati dannosi durante il processo di addestramento del modello

Implementazione di controlli per garantire l'integrità e l'autenticità dei dati utilizzati

Adozione di approcci robusti per la sicurezza informatica

Adozione della tecnica di *sanitization*

Cozzupoli Luigi
Cusato Miriam
Dattola Giovanni
Itri Caterina
Logiudice Giorgio

GRAZIE PER L'ATTENZIONE

Moio Maria Carla
Naso Domenico
Pirrottina Marta
Rosato Forti Pietro
Strano Silvia