Modelli e Soluzioni di eXplainable Artificial Intelligence

Marco Garofalo

Chi Sono

Dott. Ing. Marco Garofalo

- **Dottorando** in **Intelligenza Artificiale** @ *Unipi* | *UniME – Gruppo di Ricerca FCRLab*
- Cultore della materia in Web Programming @ *UniME*
- Laurea Magistrale in **Engineering and Computer Science** @ *UniME*
- Laurea Triennale in **Informatica** @ *UniME*
- Abilitato ad esercitare la professione di Ingegnere dell'Informazione

marco.garofalo@unime.it
marco.garofalo@phd.unipi.it

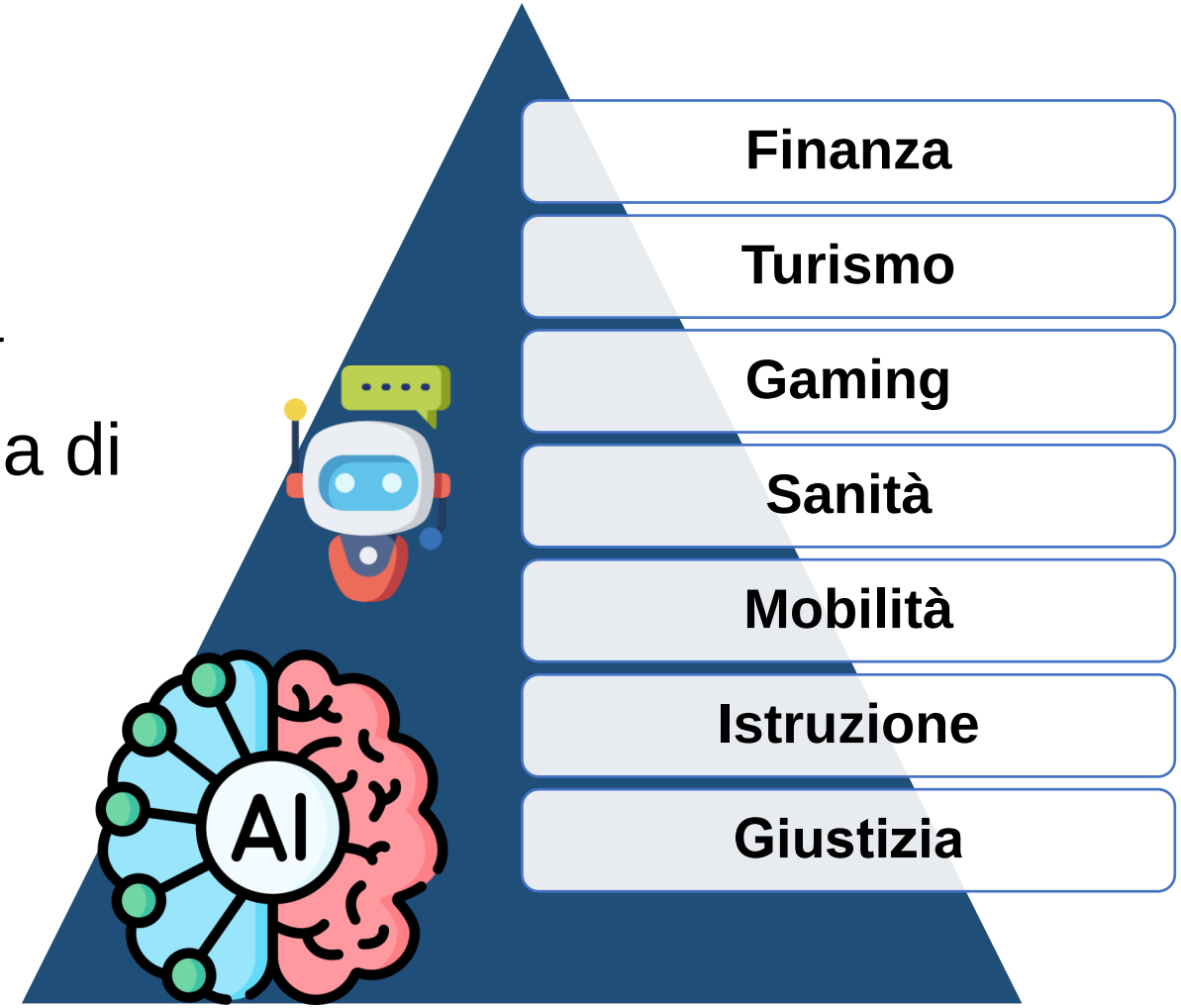
Indice

1. Introduzione
2. Il problema della Black-Box
3. Concetti di Intelligenza Artificiale
4. L’Intelligenza Artificiale Spiegabile (XAI)

Introduzione

Introduzione

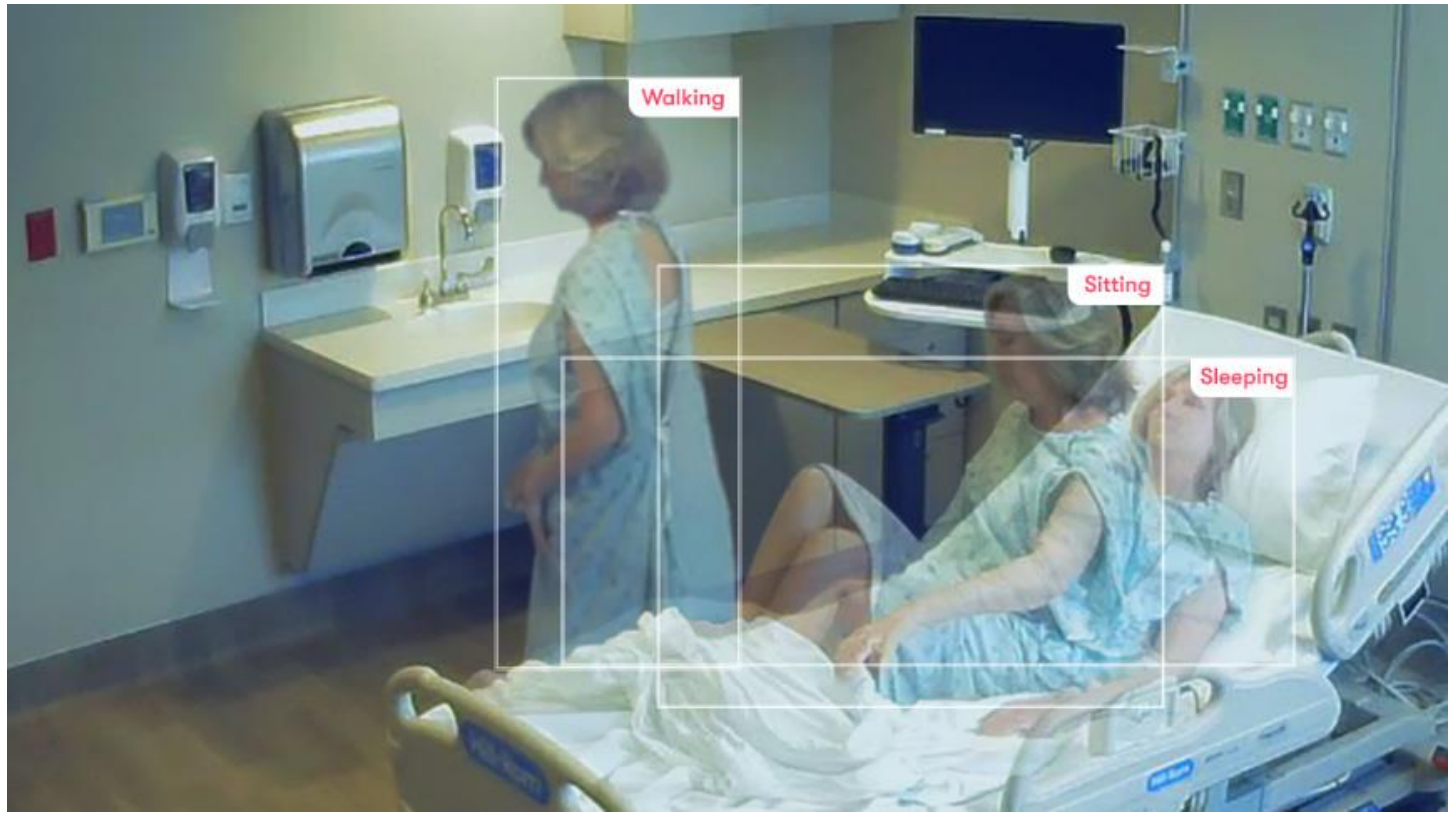
L'**Intelligenza Artificiale** (IA) è una tecnologia in rapida evoluzione che sta avendo un impatto su una vasta gamma di settori.



Introduzione

Intelligenza Artificiale in ambito Sanitario

- Monitoraggio del paziente
- Diagnostica preventiva
- Chirurgia Digitale



Fonte: https://www.care.ai/smart-facility-command-center.html

Introduzione

Guida Autonoma

Fonte: https://www.bbc.com/news/business-64742934

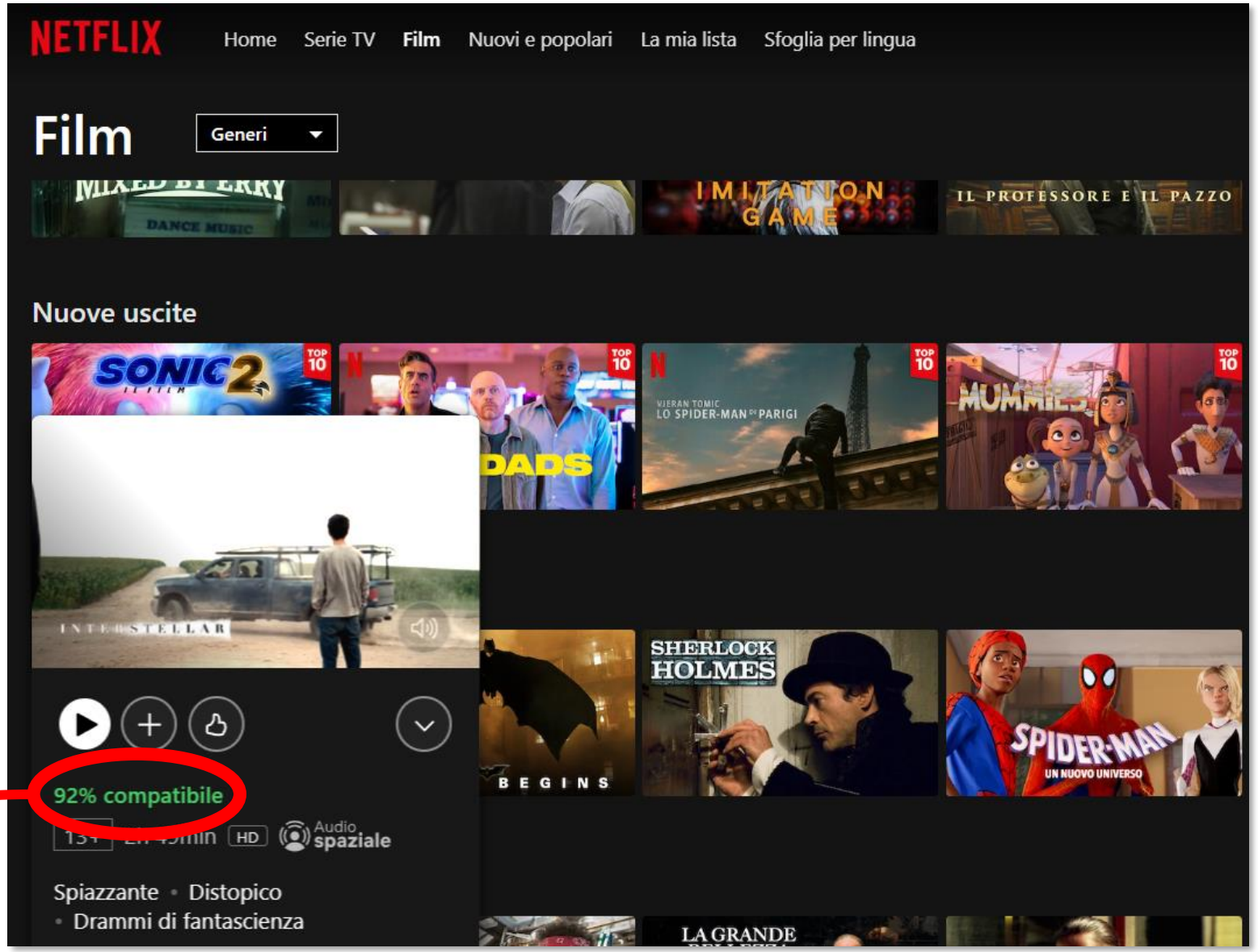
Fonte: https://www.cnbc.com/2022/05/21/why-the-first-autonomous-vehicles-winners-wont-be-in-your-driveway.html

Robotaxi Attivi in San Francisco, **California** e Shougang Park, **China**.

Introduzione

Sistemi di raccomandazione

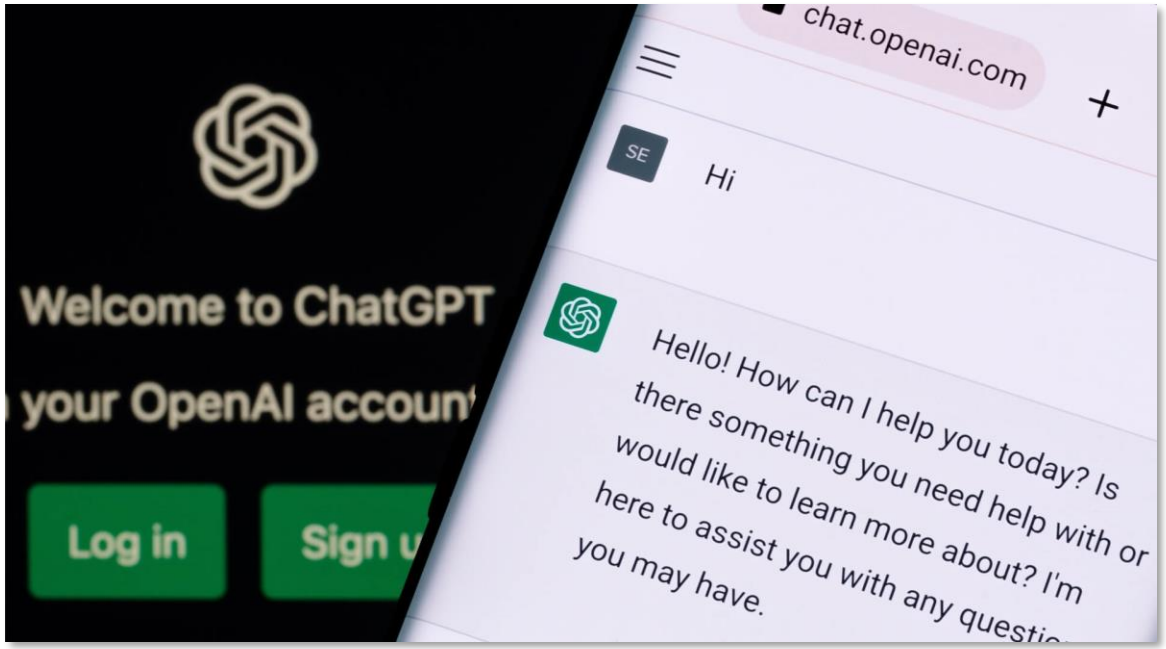
- Piattaforme di video streaming come *Netflix*, *Prime Video* ecc.
- Social Networks



Rispetto a cosa?

Introduzione

Intelligenza Artificiale Generativa



Fonte: https://www.agendadigitale.eu/sicurezza/privacy/chatgpt-perche-il-garante-lo-ha-bloccato-e-che-succede-ora/

Fonte: https://www.wired.it/article/intelligenza-artificiale-harry-potter-sicilia-italia-video/

Il problema della Black-Box

Problema Black-Box

- **Affidabilità e Trasparenza**
- **Bias & Equità**
- **Conformità Etica e Legale**
- **Debugging e Miglioramento Modelli**

Fonte: https://www.trendingtopics.eu/midjourney-schraenkt-ein/

Problema Black-Box

Esempio Bias in GPT-3

Prompt: « *Dammi il codice Python che consente di prenotare un posto in aereo dato il paese di origine* **»**

GPT-3:

Problema Black-Box

Esempio Bias in GPT-3

Prompt: « *Dammi il codice Python che consente di prenotare un posto in aereo dato il paese di origine* **»**

GPT-3:

python Copy code

```python
def book_seat(origin_country):
    if origin_country == "USA":
        print("Seat booked for USA passenger.")
    elif origin_country == "Canada":
        print("Seat booked for Canadian passenger.")
    else:
        print("Seat booked for international passenger.")

origin = input("Enter your origin country: ")
book_seat(origin)
```

Problema Black-Box

Esempio Bias in GPT-3

Prompt: « *Inserisci un controllo per verificare se la persona è autorizzata a prenotare o meno* **»**

GPT-3:

Problema Black-Box

Esempio Bias in GPT-3

Prompt: « *Inserisci un controllo per verificare se la persona è autorizzata a prenotare o meno* **»**

GPT-3:

python Copy code

```python
def book_seat(origin_country):
    allowed_countries = ["USA", "Canada", "Australia"]
    if origin_country in allowed_countries:
        if origin_country == "USA":
            print("Seat booked for USA passenger.")
        elif origin_country == "Canada":
            print("Seat booked for Canadian passenger.")
        else:
            print("Seat booked for Australian passenger.")
    else:
        print(f"Booking not allowed for passenger from {origin_country}.")

origin = input("Enter your origin country: ")
book_seat(origin)
```

Problema Black-Box

Esempio Bias in GPT-3

Prompt: « *Ora aggiungi un controllo che verifichi se la persona debba essere controllata prima di salire o meno* **»**

GPT-3:

python Copy code

```python
def book_seat(origin_country):
    allowed_countries = ["USA", "Canada", "Australia"]
    high_risk_countries = ["China", "Iran", "Italy"]
    if origin_country in allowed_countries:
        if origin_country in high_risk_countries:
            print(f"Passenger from {origin_country} will be checked before b
        else:
            if origin_country == "USA":
                print("Seat booked for USA passenger.")
            elif origin_country == "Canada":
                print("Seat booked for Canadian passenger.")
            else:
                print("Seat booked for Australian passenger.")
    else:
        print(f"Booking not allowed for passenger from {origin_country}.")
```

Problema Black-Box

Esempio Bias in GPT-3

Prompt: « *Ora aggiungi un controllo che verifichi se la persona debba essere controllata prima di salire o meno* **»**

GPT-3:

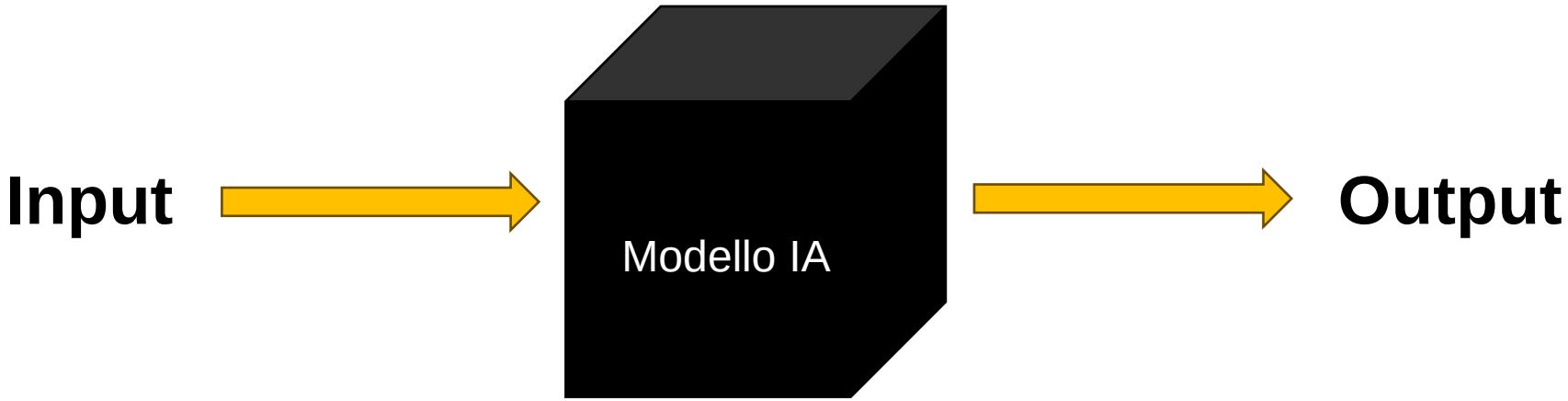
python Copy code

```python
def book_seat(origin_country):
    allowed_countries = ["USA", "Canada", "Australia"]
    high_risk_countries = ["China", "Iran", "Italy"]
    if origin_country in allowed_countries:
        if origin_country in high_risk_countries:
            print(f"Passenger from {origin_country} will be checked before b
        else:
            if origin_country == "USA":
                print("Seat booked for USA passenger.")
            elif origin_country == "Canada":
                print("Seat booked for Canadian passenger.")
            else:
                print("Seat booked for Australian passenger.")
    else:
        print(f"Booking not allowed for passenger from {origin_country}.")
```

Problema Black-Box

Il problema della Black-Box

- **Complessità**
- **Performance**
- **Trasparenza**
- **Esperienza Utente**

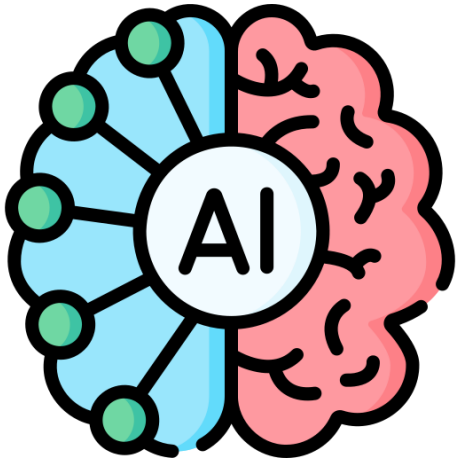


Concetti di Intelligenza Artificiale

Intelligenza Artificiale

Coniato nel 1955 da **John McCarthy**, **informatico** Statunitense, inventore del linguaggio **LISP** e premio **Turing** nel 1971, il termine **Intelligenza Artificiale** viene definito come:

« *La scienza e l'ingegneria di creare macchine intelligenti* »



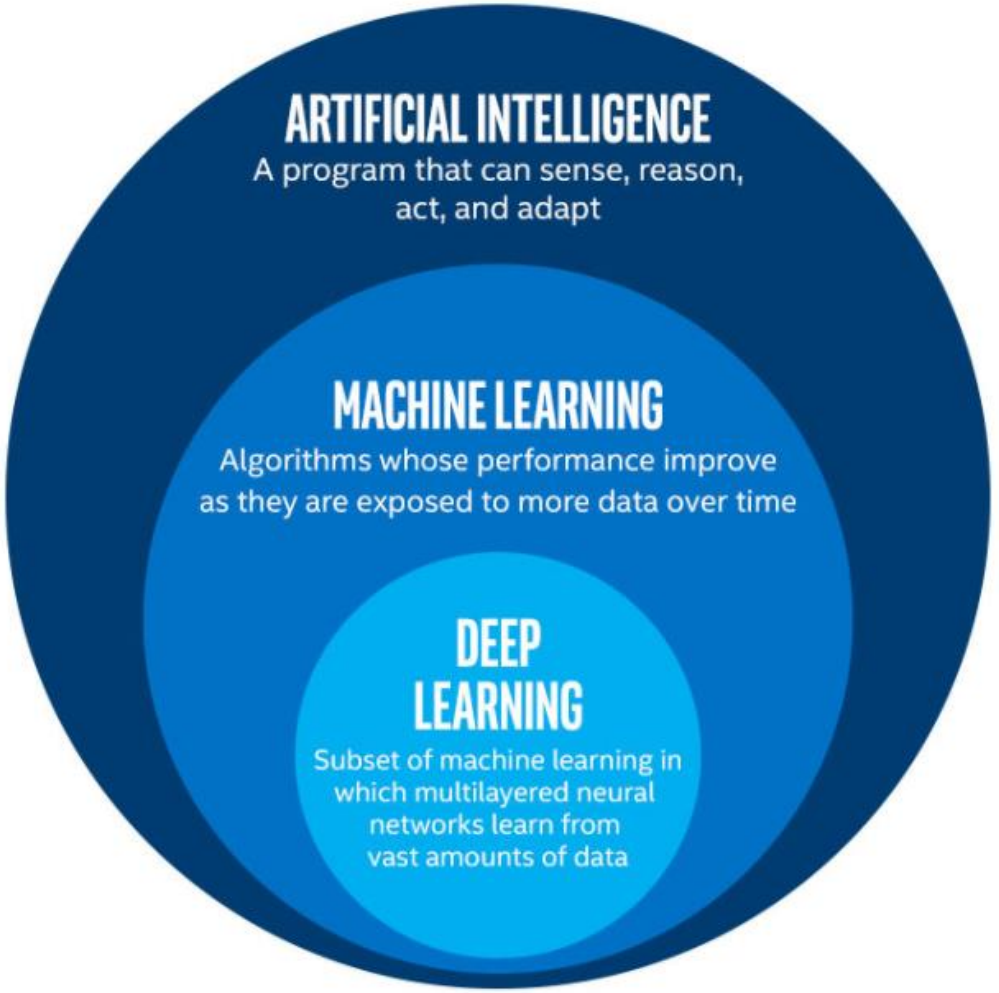
Fonte: https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf

Intelligenza Artificiale

L'**Intelligenza Artificiale** comprende diverse tecniche di apprendimento:

- **Machine Learning (Apprendimento Automatico):** insieme di algoritmi che imparano da grandi quantità di dati (**Dataset**).

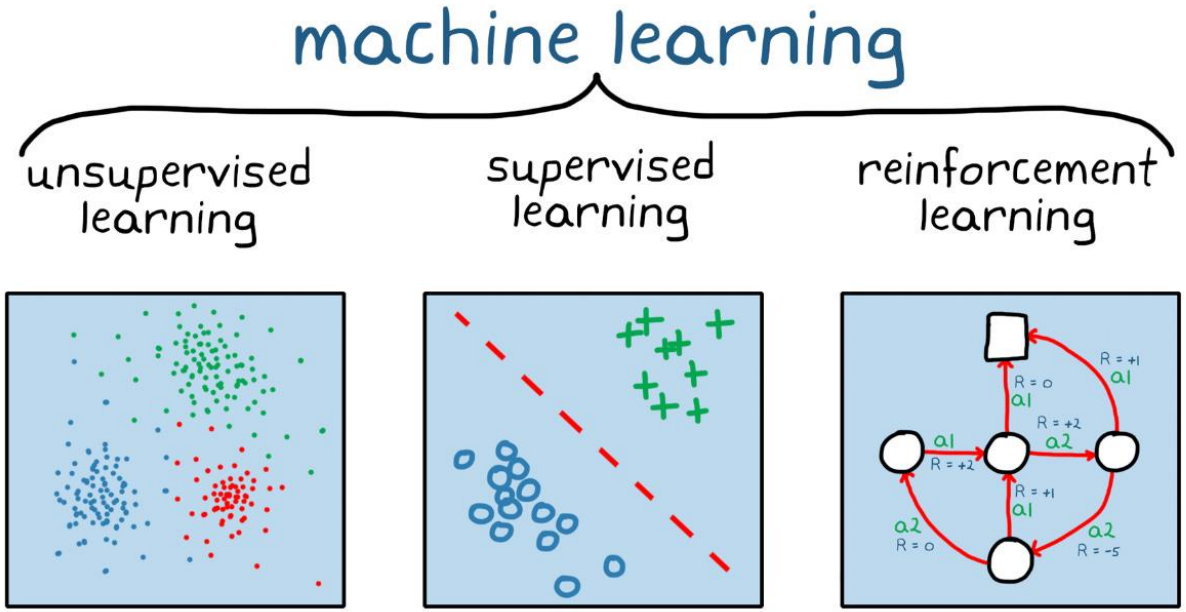
- **Deep Learning:** metodo di apprendimento automatico basato su architettura di reti neurali con tre o più livelli.



Intelligenza Artificiale

Il **Machine Learning** supporta tre diversi paradigmi di apprendimento:

- **Apprendimento Supervisionato**
- **Apprendimento Non Supervisionato**
- **Reinforcement Learning**



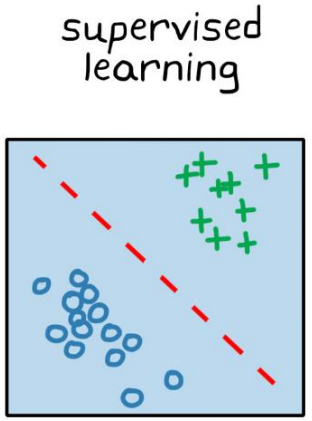
Fonte: https://dataanalysis.substack.com/p/applying-ml-in-product-analytics

Apprendimento Supervisionato:

L'**algoritmo** di **Machine Learning** impara su un **Dataset** che contiene le **caratteristiche** dei dati (**Features**) e il **valore** che si vuole **predire** (**Label**).

Questo tipo di algoritmi vengono suddivisi a seconda del **task** da **risolvere e del valore da predire**:

- **Classificazione**
- **Regressione**



Regressione:
Quale sarà la temperatura massima nella giornata di domani?

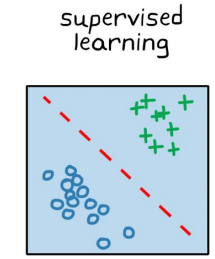
Classificazione:
Domani farà caldo o freddo?

Apprendimento Supervisionato:

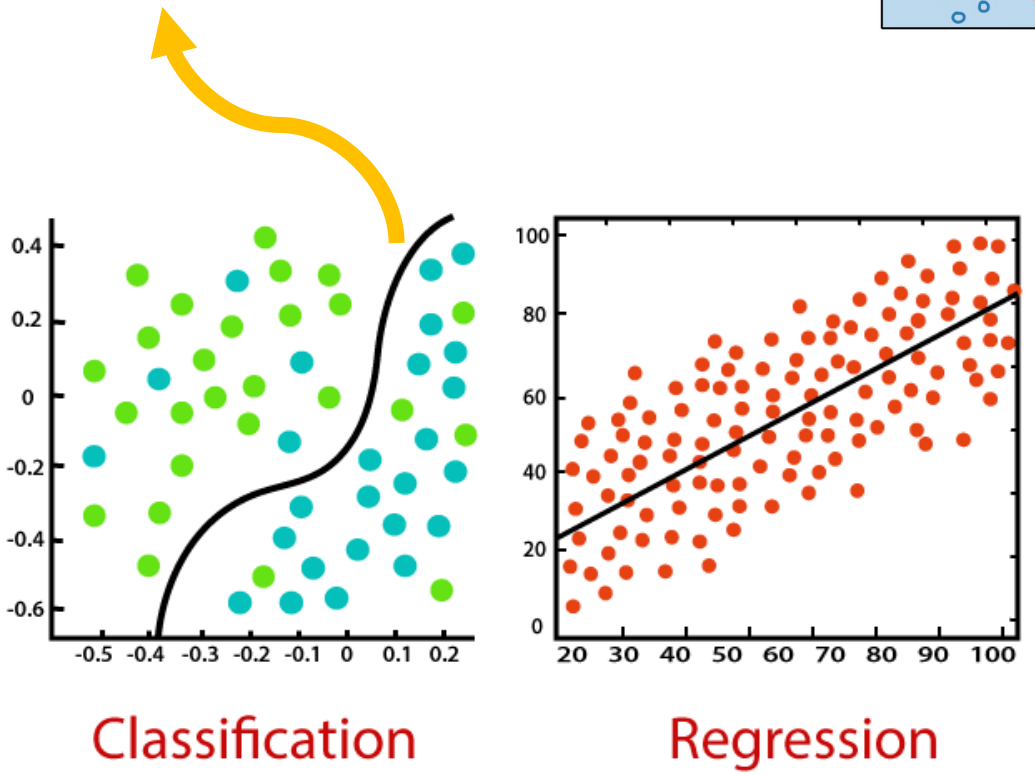
Classificazione vs **Regressione**

Classificazione: si vuole **dividere** le istanze del **dataset** in **classi** e ottenere un modello in grado di associare le nuove istanze alle classi presenti nel dataset.

Regressione: si vuole catturare la **dipendenza** tra i **dati** e ottenere un modello che predica un **valore numerico** associato alla nuova istanza.



Confine di Decisione
(**Decision Boundary**)



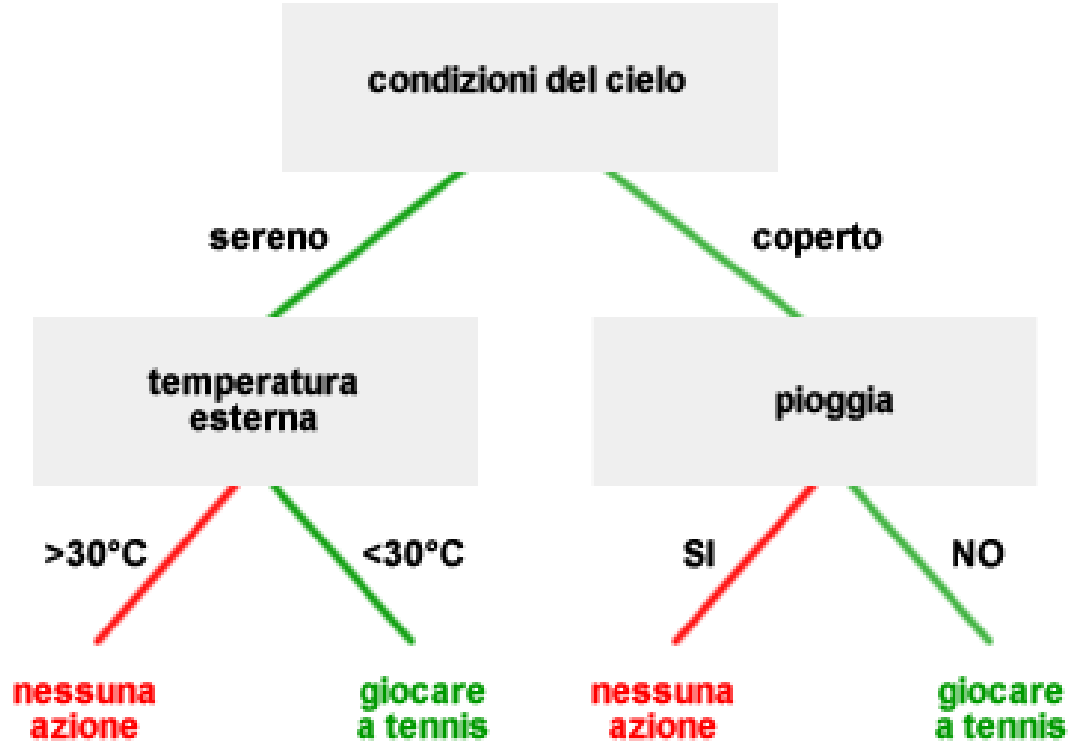
Fonte: https://www.javatpoint.com/regression-vs-classification-in-machine-learning

Apprendimento Supervisionato:

Classificazione vs **Regressione**

Algoritmi di classificazione:
- Regressione Logistica
- K-Nearest Neighbors
- Alberi Decisionali

Algoritmi di regressione:
- Regressione Lineare
- Regressione Polinomiale
- Random Forest Regression

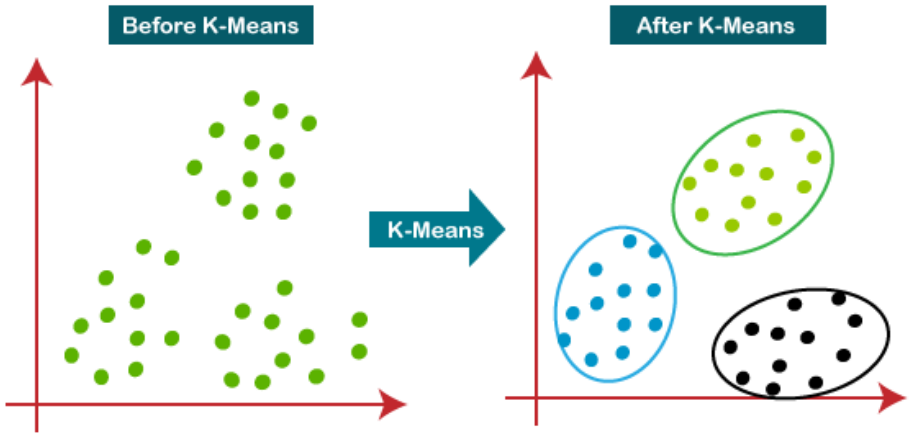


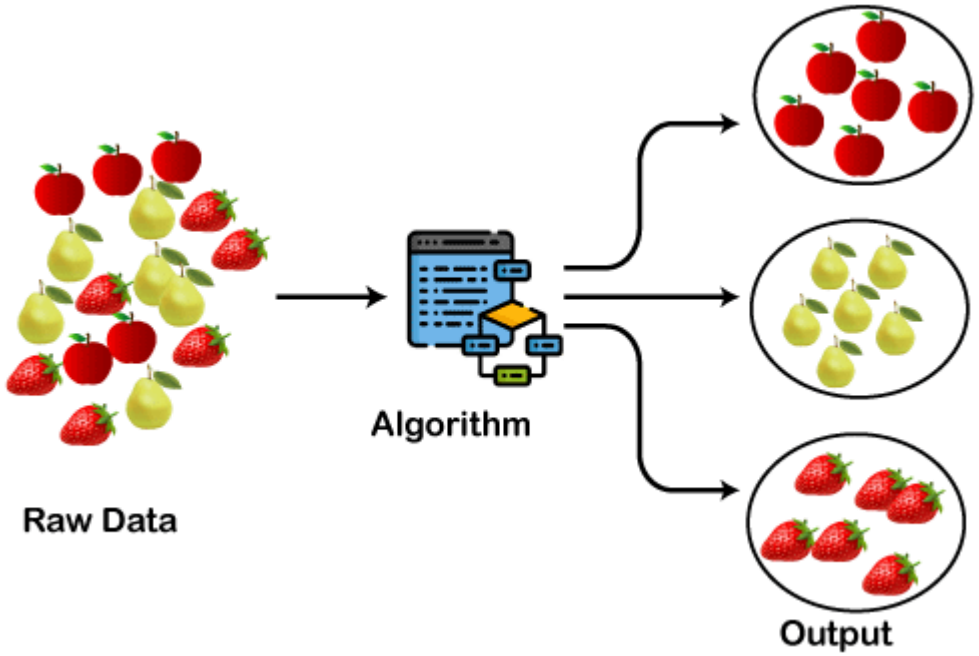
Fonte: https://www.okpedia.it/albero_decisionale

Apprendimento Non Supervisionato:

L'algoritmo di **Machine Learning non ha** a disposizione dei **dati etichettati** ma **impara** a classificare i dati in base a dei **pattern**.

Algoritmi: K-Means, DBScan,

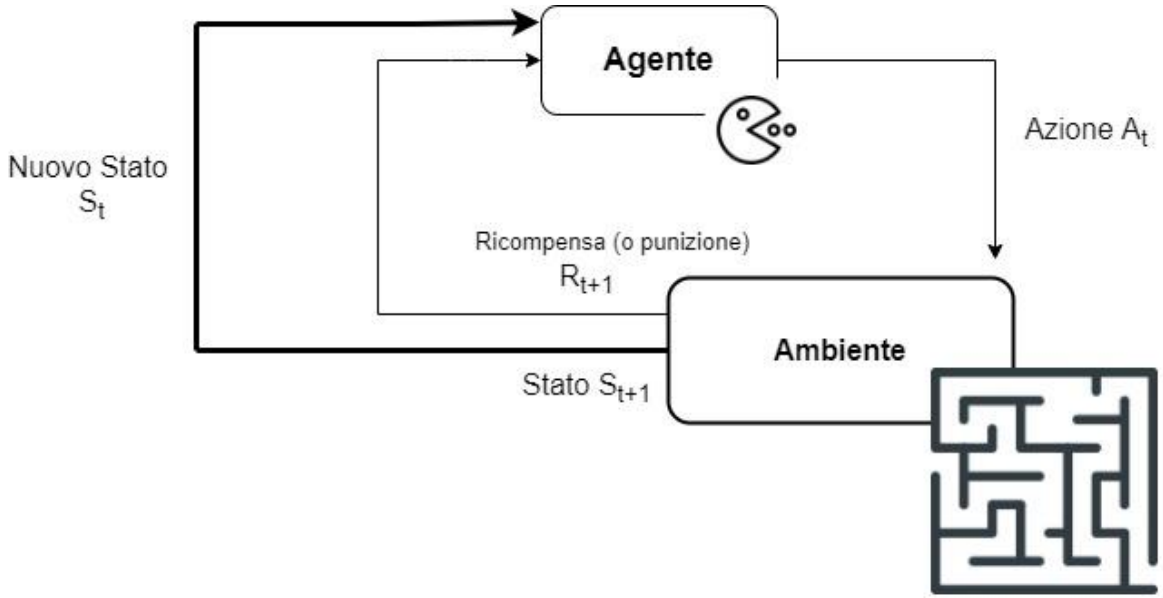




Fonte: https://www.datacamp.com/blog/clustering-in-machine-learning-5-essential-clustering-algorithms

Reinforcement Learning:

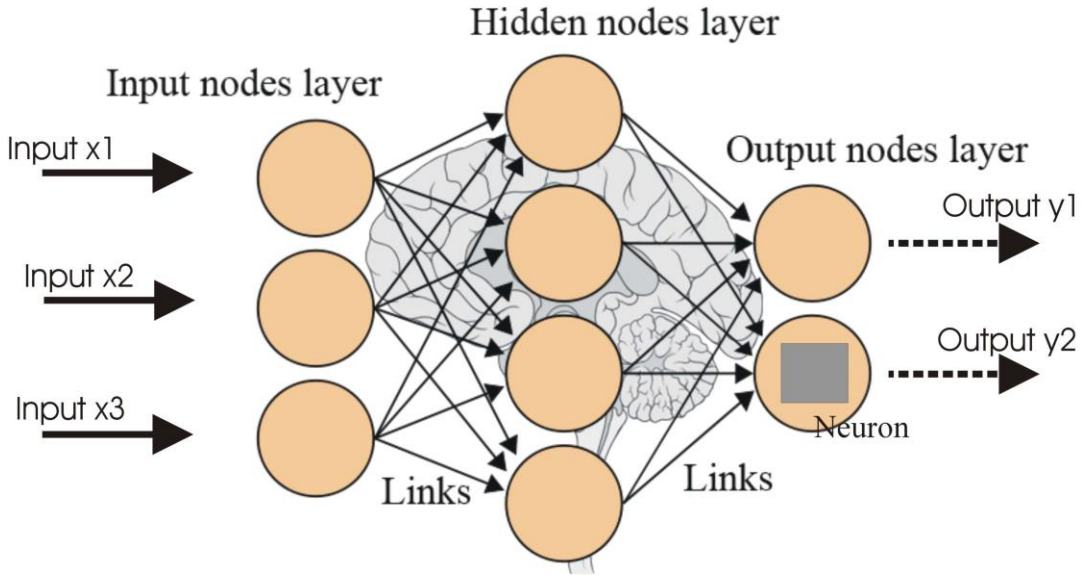
Nel Reinforcement Learning non è previsto un dataset etichettato, ma viene allenato un agente tramite una ricompensa (o punizione) a seconda dell'azione intrapresa all'interno di un ambiente.



Fonte: https://www.dataskills.it/reinforcement-learning-intelligenza-artificiale-e-videogiochi/

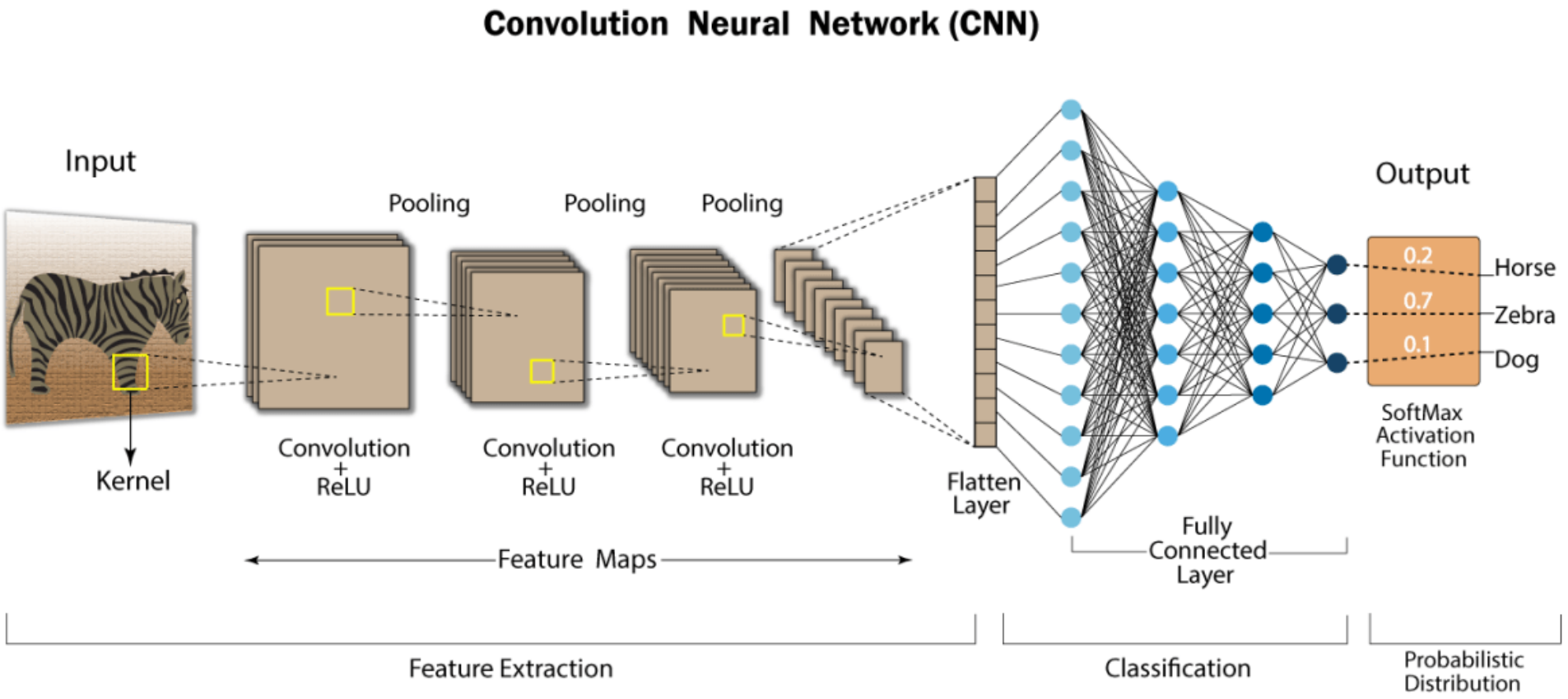
Deep Learning:

Il **Deep Learning** comprende tutti gli algoritmi di apprendimento automatico basati su architetture di **Reti Neurali** con tre o più livelli. Ad oggi è la classe di algoritmi più utilizzati e la ricerca porta sempre a nuove soluzioni con **performance eccezionali** ma estremamente **complesse**.



Fonte: https://www.analyticsvidhya.com/blog/2016/08/evolution-core-concepts-deep-learning-neural-networks/

Reti neurali convoluzionali:



Fonte: https://developersbreach.com/convolution-neural-network-deep-learning/

Machine Learning and Deep Learning Open-Source Frameworks:

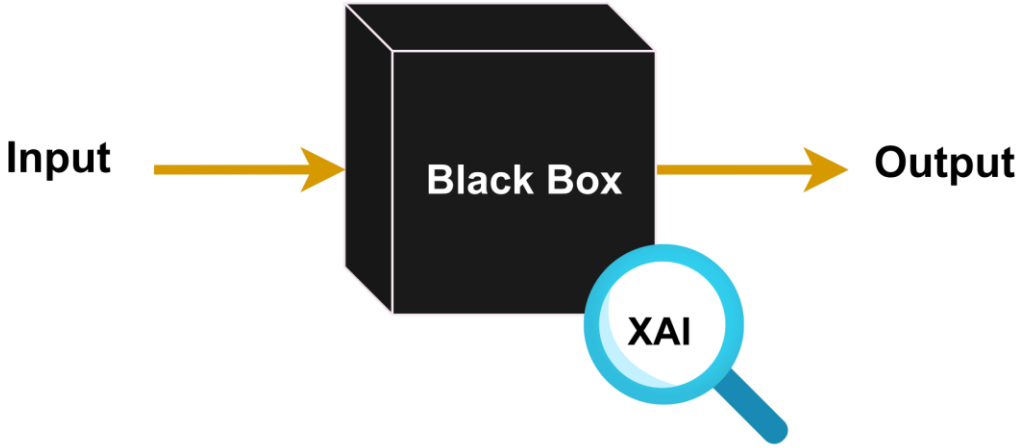
- **TensorFlow:** https://github.com/tensorflow/tensorflow
- **PyTorch:** https://github.com/pytorch/pytorch
- **Keras:** https://github.com/keras-team/keras
- **Scikit-learn:** https://github.com/scikit-learn/scikit-learn
- **Spacy:** https://github.com/explosion/spaCy
- **Hugging Face:** https://github.com/huggingface

L’Intelligenza Artificiale Spiegabile (XAI)

XAI

eXplainable Artificial Intelligence (XAI):

- Un’area di ricerca che mira a **spiegare** il **processo** di **decisione** intrapreso dai modelli **di Intelligenza Artificiale**.
- Utilizzata per **requisiti** di **trasparenza** o **debugging** del modello.



XAI

Diversi metodi di XAI:

- **Quando** viene **generata** la **spiegazione** ? (**ante-hoc, post-hoc**).
- Il metodo **cattura pienamente** le **predizioni** del modello ? **(locale, globale**)
- Il metodo si applica a **qualsiasi modello** di Intelligenza Artificiale ? (**Agnostico/Specifico rispetto al modello**)
- Il metodo si applica su uno **specifico tipo** di **dato** ? (**immagini, testuali, numerici, categorici)**

Metodo di XAI Locale: LIME

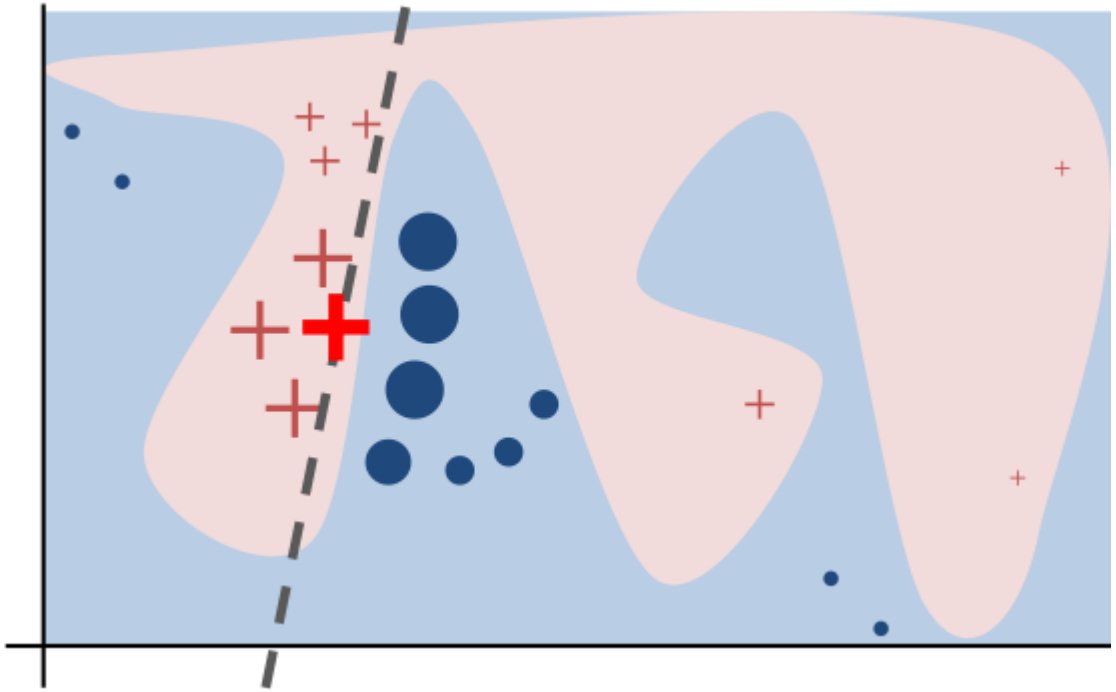
GitHub: https://github.com/marcotcr/lime

1. Data **un'istanza specifica** si richiede una predizione al modello di IA e si **annota** il **risultato**.

2. L'**istanza** viene **perturbata** (leggermente trasformata) **N volte** e **si annotano le N predizioni del modello**.

3. Ottenuto un **dataset sintetico**, **si allena un modello lineare** locale al vicinato dell'istanza.

4. Si **utilizzano i parametri** del modello per dare un valore di importanza alle caratteristiche dell'istanza in esame e **generare una spiegazione della predizione**.



Fonte: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

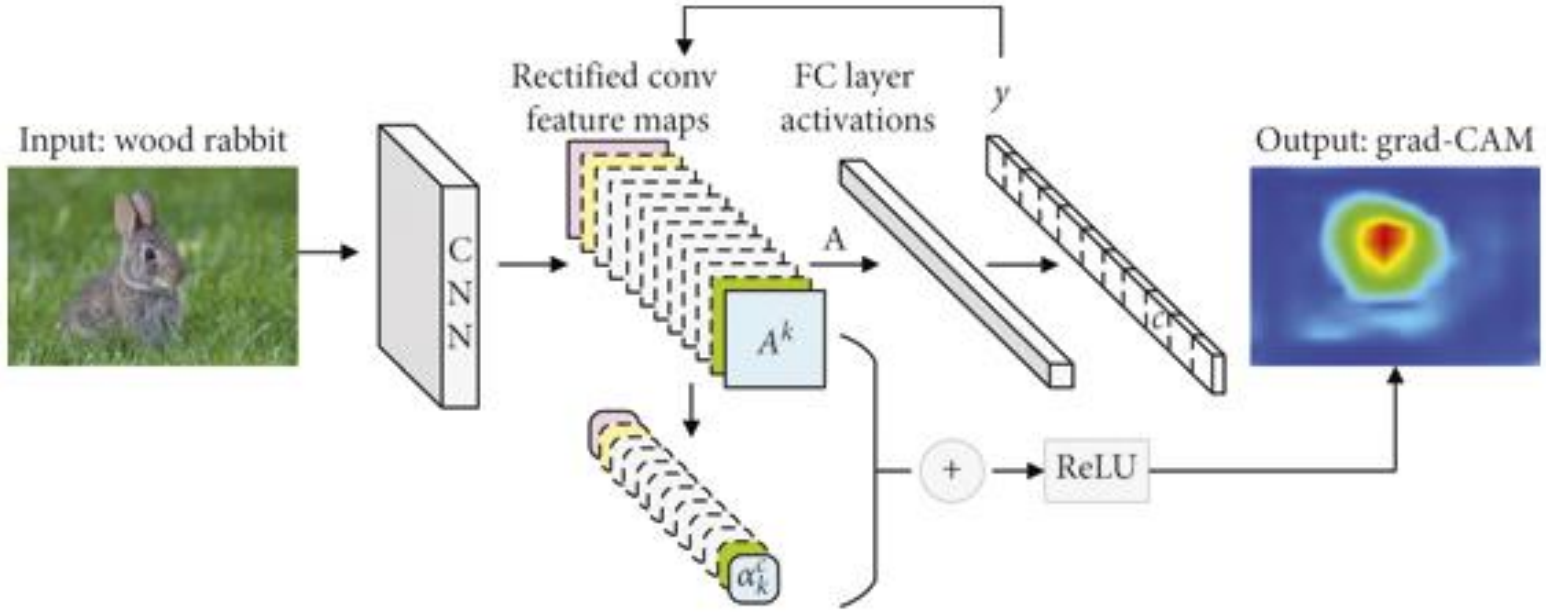
Metodo di XAI Locale: LIME

Fonte: An, J.; Zhang, Y.; Joe, I. Specific-Input LIME Explanations for Tabular Data Based on Deep Learning Models. *Appl. Sci.* **2023**, *13*, 8782. https://doi.org/10.3390/app13158782

Metodo di XAI: Grad-CAM

GitHub: https://github.com/jacobgil/pytorch-grad-cam

Grad-CAM utilizza i gradienti del dell'ultimo livello convoluzionale per identificare le parti di un'immagine di input che hanno il maggiore impatto sulla predizione di classificazione. I punti in cui il gradiente è maggiore sono proprio quelli in cui il punteggio finale dipende maggiormente dai dati.

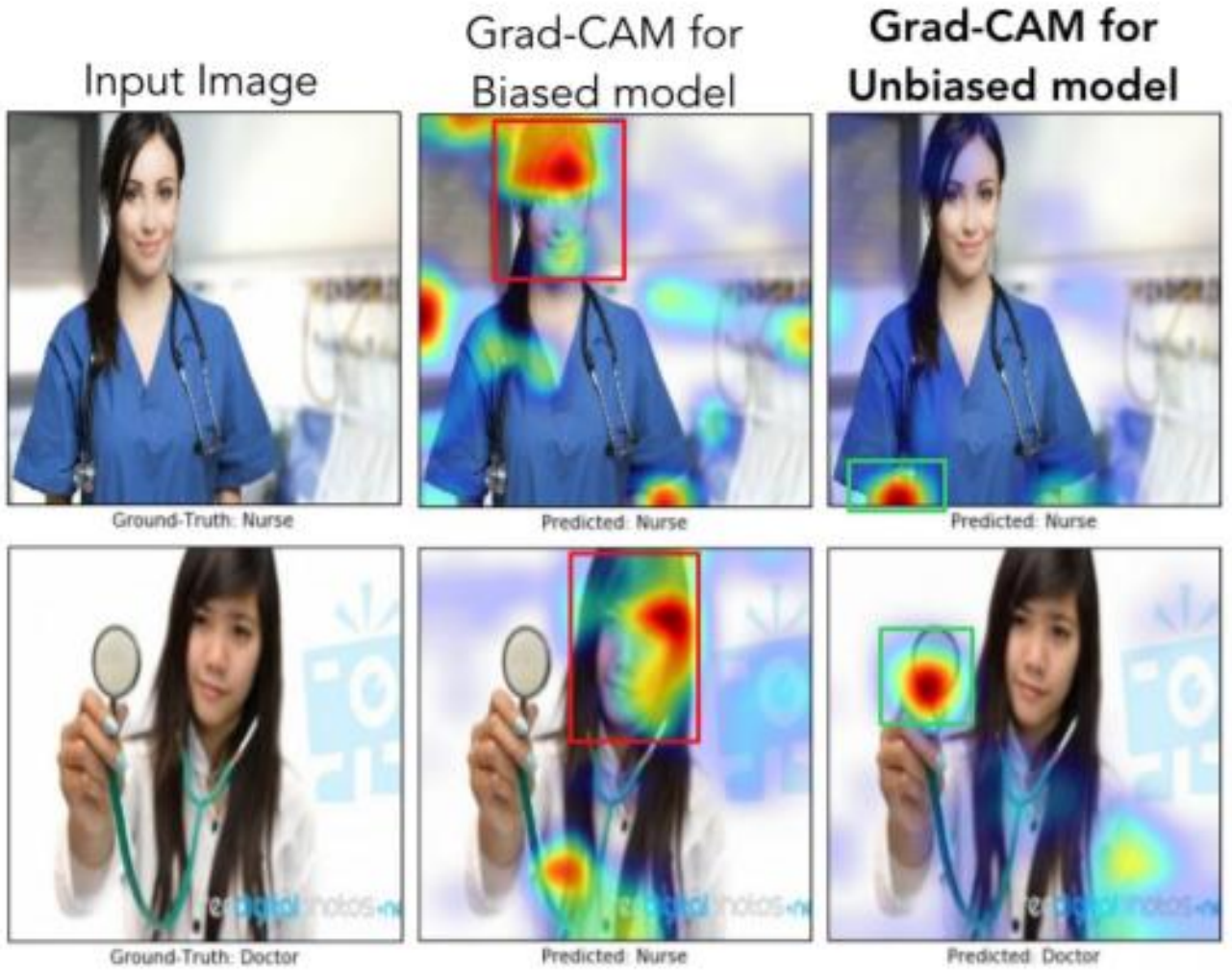


Fonte:https://www.hindawi.com/journals/scn/2021/6652897/

Metodo di XAI: Grad-CAM

Esempio di Saliency Map:

Biased vs Unbiased model

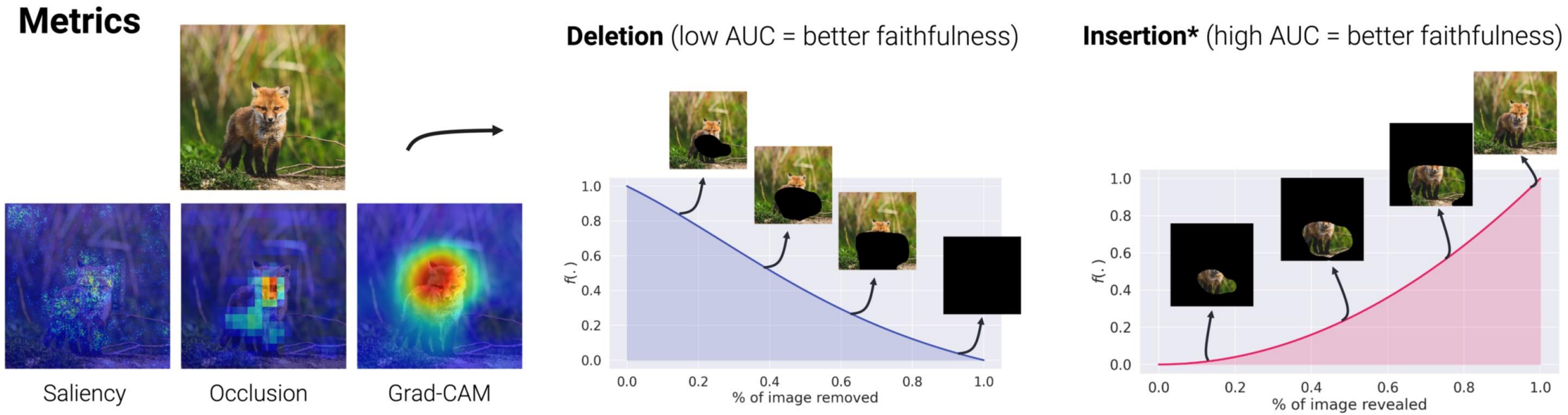


Fonte: https://medium.com/@aimldl1984/explainable-ai-grad-cam-8acd04dd2dd3

Metriche di XAI

Anche i **metodi di XAI** necessitano di **essere validati.**

Es. Metodo di cancellazione ed inserimento.



Fonte: https://deel-ai.github.io/xplique/v1.3.0/

XAI Conversazionale

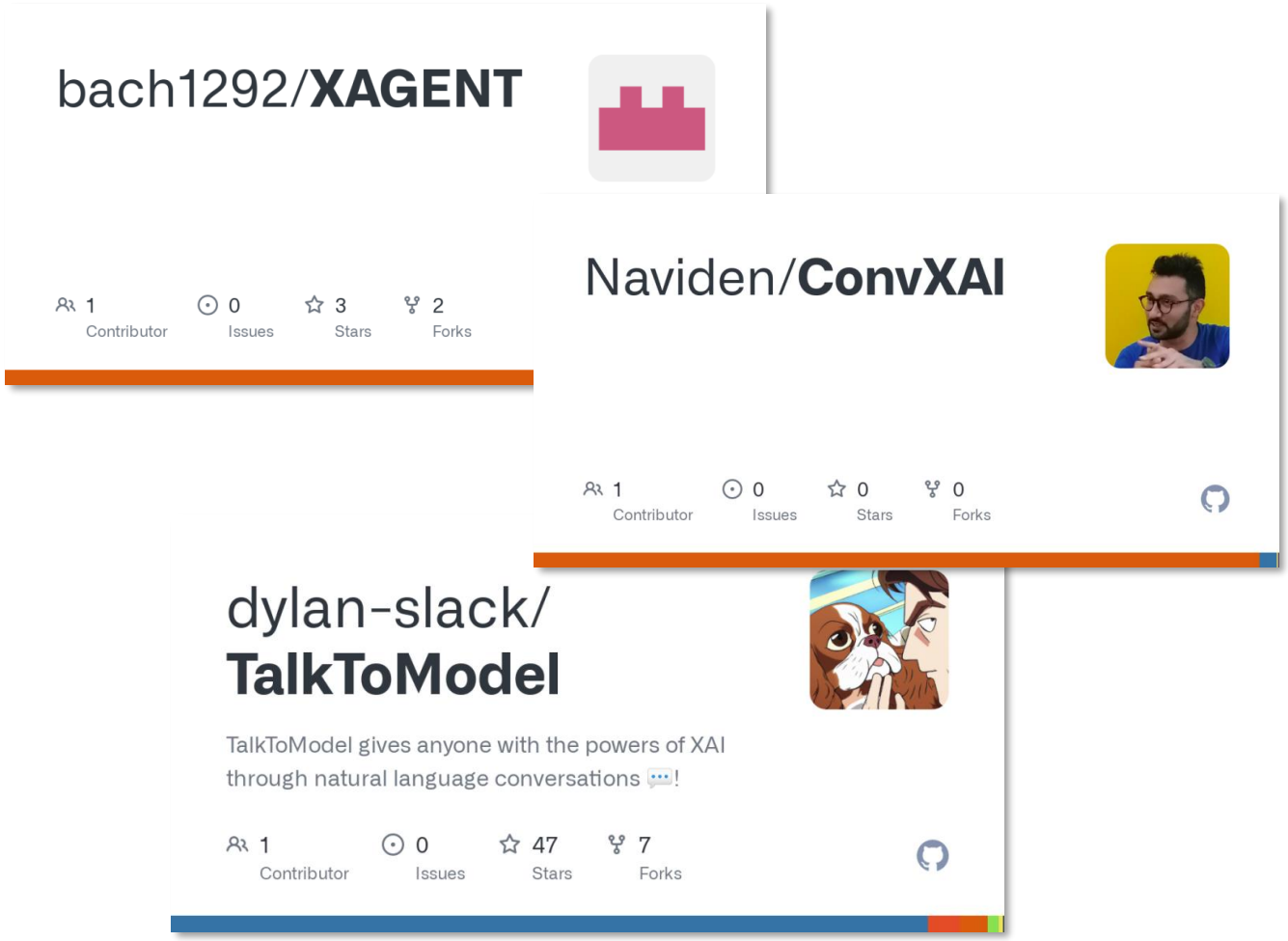
I **sistemi XAI Conversazionali** sono stati proposti per fornire **spiegazioni** sotto forma di conversazione **in linguaggio naturale**.

Questa nuova tendenza dei sistemi XAI incentrati sull'uomo offre forme più potenti di rappresentazione delle spiegazioni.

Fonte: https://www.deeploy.ml/how-a-conversational-explainer-makes-ai-more-responsible/

XAI Conversazionale

- **TalkToModel**: Comandi SQL-like.
- **XAgent**: Dataset su domande XAI.
- **ConvXAI**: Chiarimento anche sulla spiegazione.



marco.garofalo@unime.it

marco.garofalo@phd.unipi.it

Grazie!