



Intelligenza o agire artificiale: *l'antropomorfizzazione dell'IA ed i bias dei dati*

Prof. F. Benedetto

francesco.benedetto@uniroma3.it
fbenedetto@ieee.org

Chair of "Machine Learning and Data Processing" - University of Roma Tre
Chair of the IEEE VTS Italian Chapter - IEEE Vehicular Technology Society



Signal Processing for Telecommunications and Economics (SP4TE)
University of "Roma TRE"
via Vito Volterra, 62
00146 Rome - Italy
Ph.: +39 06.57337079
fax: +39 06.57337026
e-mail: francesco.benedetto@uniroma3.it

È professore di Telecomunicazioni presso l'Università degli Studi di Roma Tre dove insegna nella Laurea di Economia e Big Data il corso di "**Machine Learning and Data Processing**". **Presidente** della sezione italiana della società internazionale di ingegneri elettrici ed elettronici *IEEE Vehicular Technology Society (VTS)* per il biennio 2024/2025. Nominato nel comitato tecnico "*Artificial Intelligence (AI) in wireless communications*" dell'IEEE VTS. **Editor in Chief** della rivista internazionale *Recent Advances in Computer Science and Communications*. Ad oggi autore di più di **170 lavori scientifici**, i suoi interessi includono l'intelligenza artificiale compresi aspetti di etica e di policy, l'accesso dinamico allo spettro radio, l'elaborazione dei segnali radar con applicazioni fortemente rivolte all'interdisciplinarietà (nel campo dell'economia, delle telecomunicazioni e dell'ingegneria civile).

He is a professor of Telecommunications at the University of Roma Tre where he teaches the "Machine Learning and Data Processing" course in the Economics and Big Data degree. Senior member of IEEE, he is the Chair of the Italian chapter of the IEEE Vehicular Technology Society (VTS) for the two-year period 2024/2025. Member of technical committee "Artificial Intelligence (AI) in wireless communications" of the IEEE VTS. Editor in Chief of the international journal Recent Advances in Computer Science and Communications. To date, he is the author of more than 170 scientific works, and his interests include artificial intelligence with ethical and policy aspects, dynamic access to the radio spectrum, signal processing with applications strongly aimed at interdisciplinarity (in the field economics, telecommunications and civil engineering).



Signal Processing for Telecommunications and Economics (SP4TE)
University of "Roma TRE"
via Vito Volterra, 62
00146 Rome - Italy
Ph.: +39 06.57337079
fax: +39 06.57337026
e-mail: francesco.benedetto@uniroma3.it

1. **KEYNOTE SPEAKER** at the Int. Conf. on “fourth industrial revolution based technology and practices (ICFIRTP-2020)”, on 15th- 16th March, 2020.
2. **HONORARY GUEST SPEAKER** at the 3rd Int. Conf. on “Advances in Information Communication Technology & Computing (AICTC – 2021)”, on 20 – 21 Dec. 2021.
3. **INVITED KEYNOTE SPEAKER** at the 2nd Int. Conf. on “Computational Intelligence in Machine Learning (ICCIML-22)”, on 18-19, November 2022.
4. **RESPONSABILE** scientifico (Leader) dell'Università Roma Tre per il contratto di ricerca con la Vitrociset S.p.A a Leonardo Company, per l'esecuzione di un programma di ricerca concernente lo “*Sviluppo di Algoritmi e Tecniche di Intelligenza Artificiale per la Classificazione e l'Identificazione di Target Spaziali elaborando Segnali provenienti da Sistemi BIRALES*”. Dal 17-03-2021 al 31-01-2022.
5. **RESPONSABILE** scientifico (Leader) dell'Università Roma Tre per il progetto di ricerca di un dottorato industriale triennale co-finanziato dalla **Regione Lazio** ed i **Leonardo Labs, Leonardo Company**, per l'esecuzione di un programma di ricerca concernente lo “Space Situational Awareness, progettazione e sviluppo di algoritmi e tecniche di machine learning per l'identificazione e classificazione di target spaziali tramite elaborazione di segnali radar.”.
6. **COORDINATORE** delle attività di ricerca del laboratorio di Signal Processing for Telecommunications and Economics (Università Roma Tre) in collaborazione con l'Università di West London - School of Computing and Engineering, Faringdon Centre for Non-Destructive Testing (London, UK): *AI for non destructive technologies*.
7. **WINNER** of a fully funded project (18 months for a total approved contribution of € 276,109.27) PR FESR 2021-2027-Competitive Repositioning RSI: *Vitalspace, control of individual vital parameters in confined spaces*.

Prof. Francesco BENEDETTO

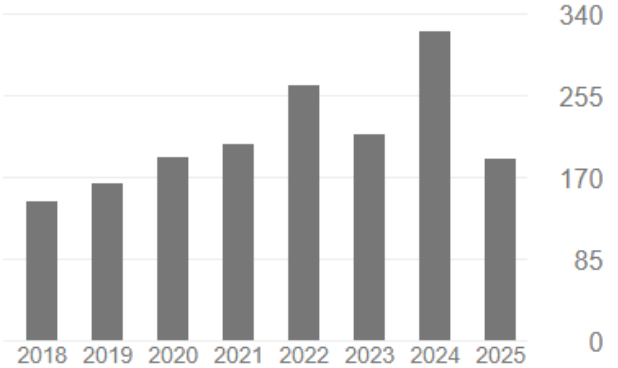


Signal Processing for Telecommunications and Economics (SP4TE)
University of "Roma TRE"
via Vito Volterra, 62
00146 Rome - Italy
Ph.: +39 06.57337079
fax: +39 06.57337026
e-mail: francesco.benedetto@uniroma3.it

Citata da [VISUALIZZA TUTTO](#)

	Tutte	Dal 2020
Citazioni	2576	1390
Indice H	27	20
i10-index	63	41

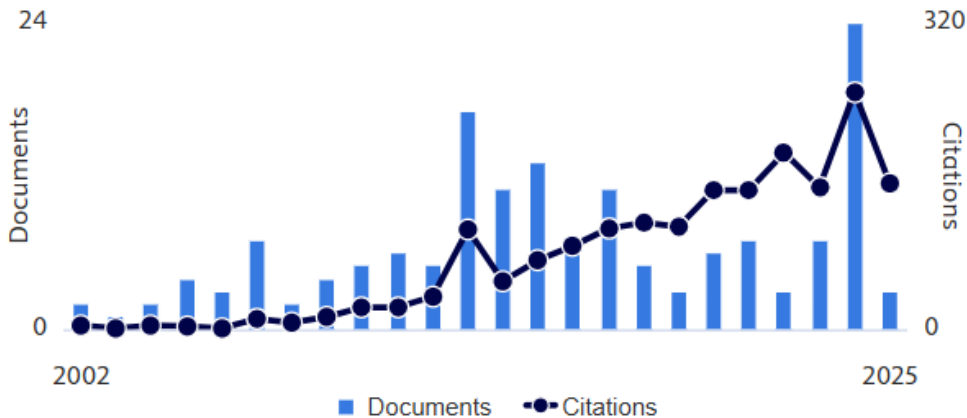
1,781 Citations by 1,404 documents 158 Documents 24 h-index



Document & citation trends



Scopus



- Dall'era dell'informazione a quella dell'IA
 - Intelligenza artificiale, umana o umanoide?
- Estrazione di conoscenza dai dati
 - Esempi di applicazioni IA
- IA agnostica, bias dei dati e rischi
 - Uso etico, responsabile e normato dell'IA
- Conclusioni

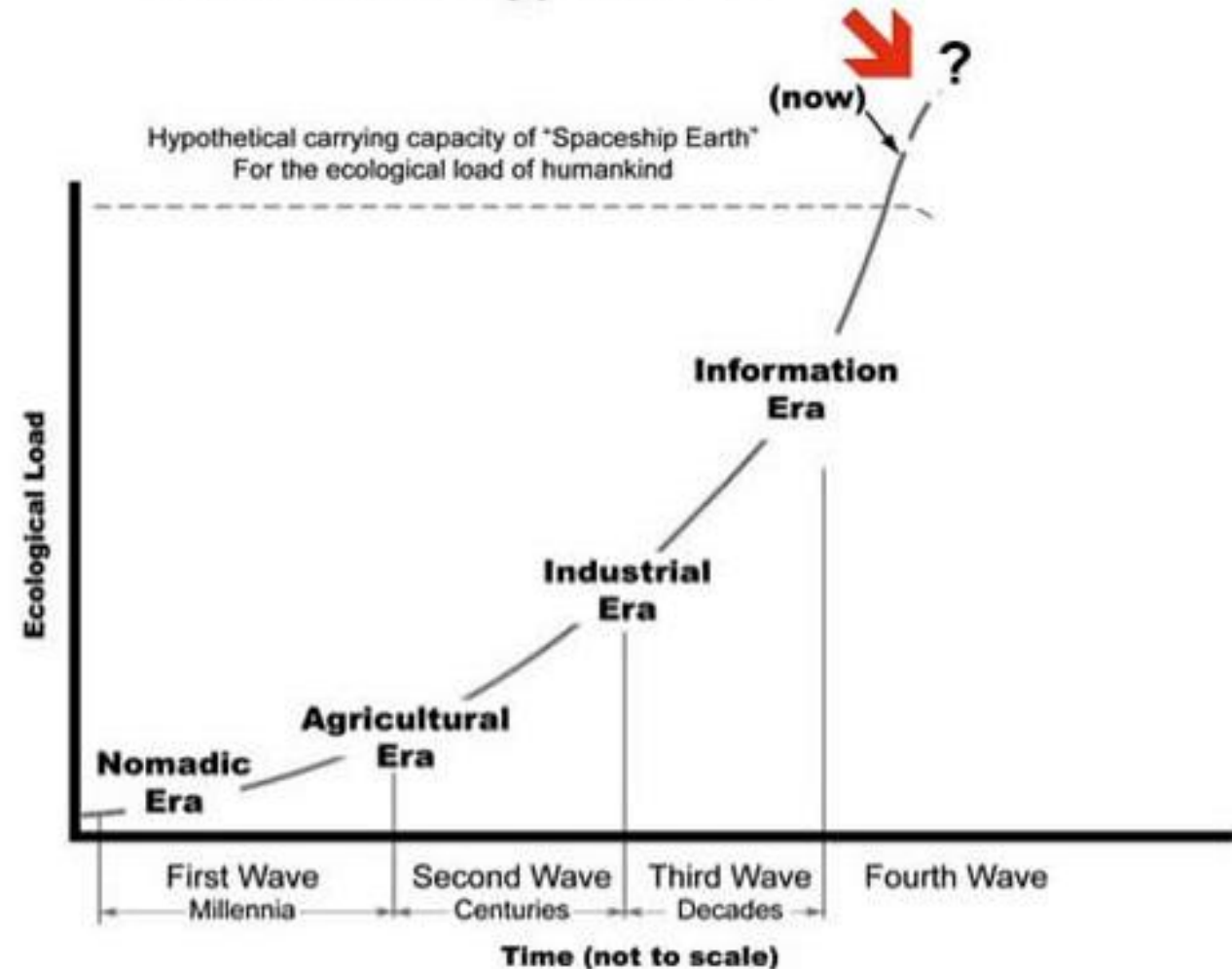


Il cambiamento è il processo attraverso il quale il futuro invade le nostre vite.

Alvin Toffler

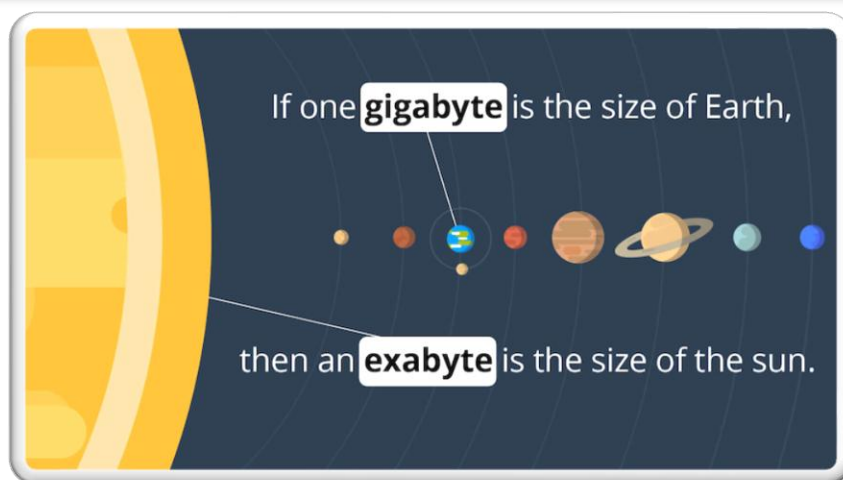
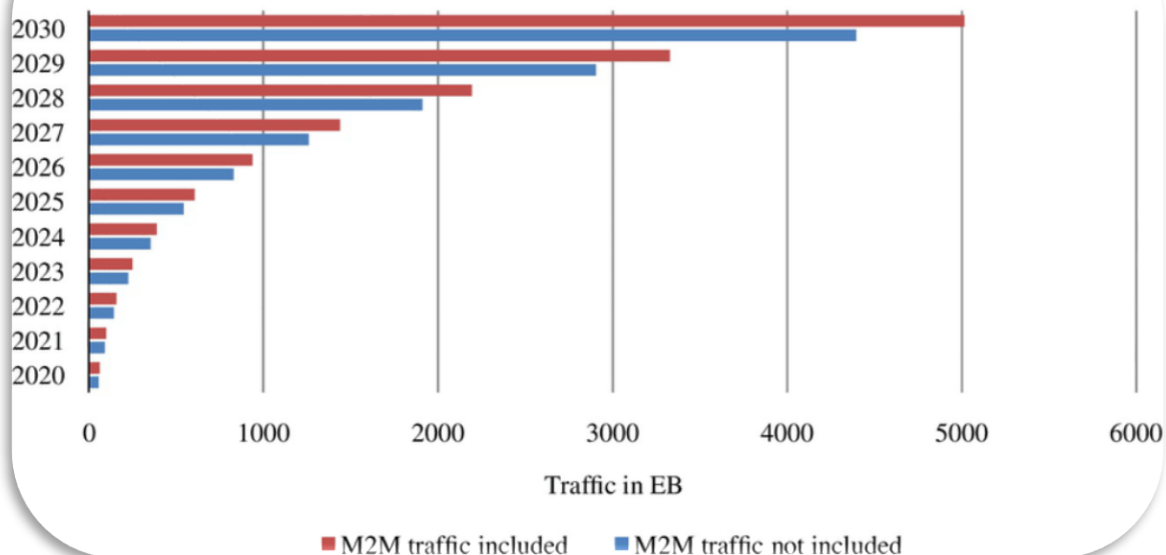
Le tecnologie del cyberspazio stanno spostando la storia dell'uomo da uno sviluppo lineare e continuo ad una progressione non-lineare e discontinua che indica ampiamente come il mondo in cui viviamo non sarà più lo stesso.

A Hinge Point into Some New Type of Era

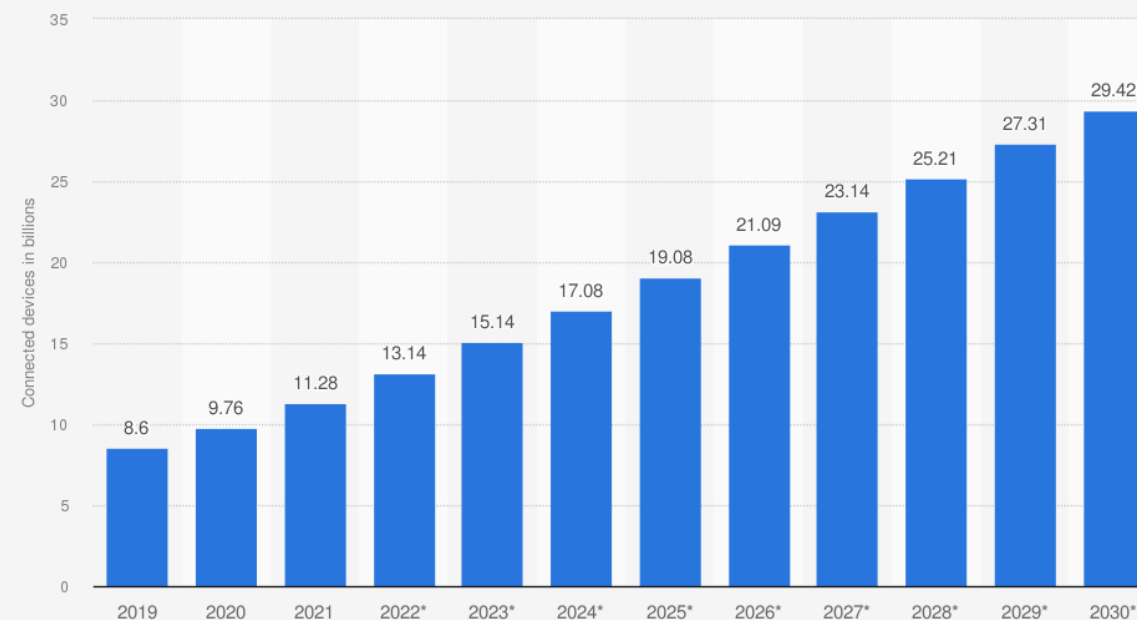


Comunicazioni sempre crescenti

Estimation of global mobile traffic per month from 2020 to 2030 (in Exabytes)



Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2021, with forecasts from 2022 to 2030 (in billions)



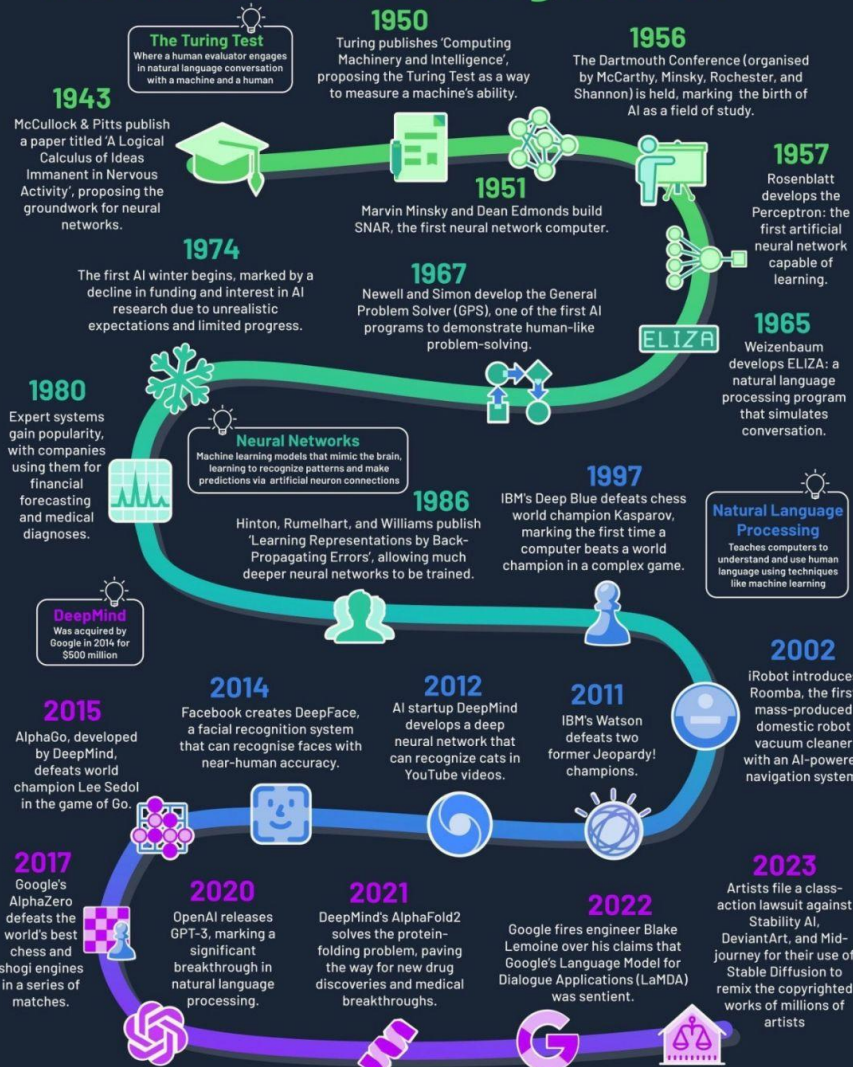
Source
Transforma Insights
© Statista 2022

Additional Information:
Worldwide; 2019 to 2022

Source: Statista.com

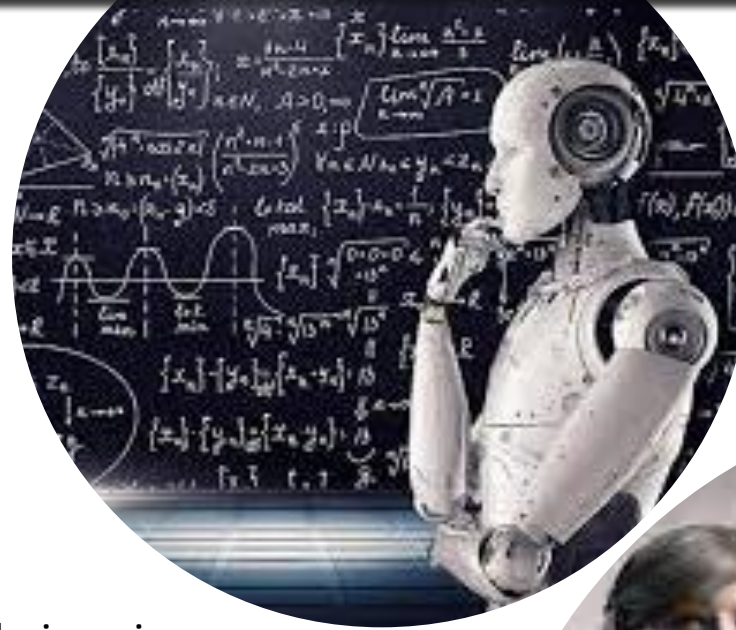
A brief history of... Artificial Intelligence.

CREATED BY
genuine
impact



La storia dell'IA è caratterizzata da periodi di entusiasmo e scoperte seguiti da momenti di sfida e rallentamenti («inverni»). Tuttavia, negli ultimi decenni, l'IA ha fatto progressi significativi grazie all'evoluzione del machine e deep learning.

- **L'intelligenza artificiale** è una disciplina appartenente all'informatica che studia i fondamenti teorici, le metodologie e le tecniche che consentono la progettazione di sistemi hardware e sistemi di programmi software capaci di fornire all'elaboratore elettronico prestazioni che, a un osservatore comune, sembrerebbero essere di pertinenza esclusiva dell'intelligenza umana.
- **«Effetto IA»**, fenomeno per cui “quando capiamo il modo in cui una macchina fa qualcosa di intelligente, questa cessa di essere considerata intelligente”.
- Paradossalmente, associamo all'Intelligenza Artificiale solo i problemi che non sono ancora stati risolti. Questo crea l'impressione che la ricerca sull'IA non abbia raggiunto risultati soddisfacenti nel corso della sua lunga storia. (**L. Manovich, “L' estetica dell'intelligenza artificiale. Modelli digitali e analitica culturale”**)



L'IA è o sarà mai paragonabile alla mente umana?

Le questioni sulla natura della mente sono connesse con altre fondamentali domande relative all'essere umano:

-questione della libertà dell'agire: l'uomo è libero? Cosa si intende per libertà?

-questione della responsabilità morale dei propri comportamenti.

-questione dell'imputabilità giuridica della responsabilità.

«E' ipotizzabile una colpevolezza robotica, in assenza, fra l'altro, di requisiti certi quali l'autocoscienza, il libero arbitrio, l'autonomia morale?

Si può parlare di "capacità di intendere e di volere" in relazione a un software o configurare in capo ad esso una colpa o addirittura un dolo?

Ancora: quali sanzioni potranno essere comminate a tali sistemi e soprattutto, nel caso, saranno in grado di assolvere quelle funzioni, orientative e rieducative, che la Costituzione riconosce alla pena? E poi, all'esito di quale tipo di processo e con quali garanzie?»



Paradiso Canto XVII

“La contingenza, che fuor del quaderno
de la vostra matera non si stende,
39. tutta è dipinta nel cospetto eterno ¹⁸;

necessità però quindi non prende
se non come dal viso in che si specchia
42. nave che per torrente giù discende ¹⁹.”

“I fatti contingenti, che non hanno luogo
al di fuori del mondo materiale,
39. sono tutti e sempre presenti nella mente divina;

non per questo acquisiscono carattere di necessità,
così come il movimento di una nave che discende
42. un torrente non dipende dallo sguardo che la osserva.”

¹⁸ I fatti contingenti sono presenti nella mente divina eternamente. Queste indicazioni teologiche sono tratte da **san Tommaso** nel suo trattato *Summae Theologiae*.

¹⁹ È il **principio del libero arbitrio**: l’onniscienza divina non impedisce che ciascun essere umano possa scegliere per se stesso e in maniera autonoma.

La libertà

Le neuroscienze, le scienze cognitive e la sintesi neodarwiniana (che include anche la genetica) ci hanno costretto a chiederci se – e nel caso in quale misura – le nuove acquisizioni scientifiche permettano di concepire gli esseri umani come liberi e responsabili per le azioni che ritengono di compiere intenzionalmente.

«come possiamo far rientrare la concezione di noi stessi in quanto agenti dotati di mente, e che possono creare significati e sono liberi e razionali, in un universo che consiste interamente di brute particelle fisiche, che non hanno mente, né significato, né libertà, né razionalità» (Searle 2007, p. 5).

Si è comunemente d'accordo che l'IA non sia né libera, né responsabile moralmente...



Intelligenza Artificiale o Umana: il discrimine

Le questioni di cosa sia la mente e se sarà mai possibile correlare gli stati mentali e la coscienza con specifiche attività neuronali, spiegandone la causazione e la localizzazione cerebrale, restano aperte.

E finché rimarranno aperte, con tutte le conseguenze etico-giuridiche del caso, possiamo attestarci sulla conclusione che la mente umana conserva uno scarto incolmabile rispetto all'IA.

Tale scarto risiede proprio in questo iato cognitivo tra la vita interiore e i processi chimico-elettrici che coinvolgono la materia. Ed è in questo iato che le scienze cognitive – ad oggi – sembrerebbero costrette a lasciarci un piccolo margine di libertà.

Allora, tra l'intelligenza umana e quella artificiale, al momento, rimane un discrimine insuperabile: **l'irriducibilità del pensiero.**



Mediamente le conoscenze tecniche relative ad un'IA sono molto poche, se non nulle.

L'utente medio interagisce con l'IA attraverso un volto umanoide con cui essa si interfaccia e che le deriva dai dati generati da umani sui quali è stata addestrata.

Quando la vediamo in azione così simile a noi e spesso anche più capace, la tentazione di immaginarla umana là dove non la vediamo è fortissima.

Si tratta del bias antropomorfo di cui siamo totalmente intrisi.

In questo terreno, col buio di fronte e la sola libertà di chiudere gli occhi, è facile seguire la pancia o il cuore che prediligono le paure e gli entusiasmi.



- L'Intelligenza Artificiale ristretta (ANI) è l'attuale forma predominante di AI, adottata in applicazioni su telefoni cellulari, Internet e analisi di big data. L'AI ristretta viene progettata per un singolo compito, da cui deriva il suo nome, e a causa di questa specificità è chiamata anche "AI debole".
- Questo le consente di concentrarsi su un compito o un insieme di compiti e di ottimizzare ulteriormente il suo funzionamento.



Narrow AI

Dedicated to assist
with or take over
specific tasks

- L'AI ristretta è ciò che l'Intelligenza Artificiale è oggi, ma l'AI generale rappresenta il futuro dell'AI. Anche nota come Intelligenza Artificiale Generale (AGI) o Intelligenza Forte, l'AI generale potrebbe essere in grado di pensare e funzionare come gli esseri umani, compresi compiti percettivi come la visione e l'elaborazione del linguaggio, oltre a compiti cognitivi come l'elaborazione, la comprensione del contesto e il pensiero generale.
- L'AI ristretta è creata per eseguire compiti specifici, ma l'AI generale potrebbe essere ampia e adattabile.



Narrow AI

Dedicated to assist
with or take over
specific tasks



General AI

Takes knowledge from
one domain, transfers
to other domain

- L'ASI, o Super Intelligenza Artificiale, rappresenterebbe l'apice dell'Intelligenza Artificiale e supererebbe in ogni modo la cognizione umana. Sebbene sia ancora una teoria, gli esperti concordano che l'ASI potrebbe derivare dall'esplosione dell'Intelligenza Artificiale, ovvero una crescita esponenziale degli algoritmi di AI.

- Questo tipo di Intelligenza Artificiale non solo sarebbe in grado di superare l'intelligenza umana, ma potrebbe anche superare qualsiasi forma di intelligenza conosciuta, sia artificiale che naturale. L'AI super-intelligente sarebbe in grado di eseguire compiti che al momento sono al di là della nostra comprensione e della nostra immaginazione.

3 stages of AI



Narrow AI

Dedicated to assist with or take over specific tasks



General AI

Takes knowledge from one domain, transfers to other domain



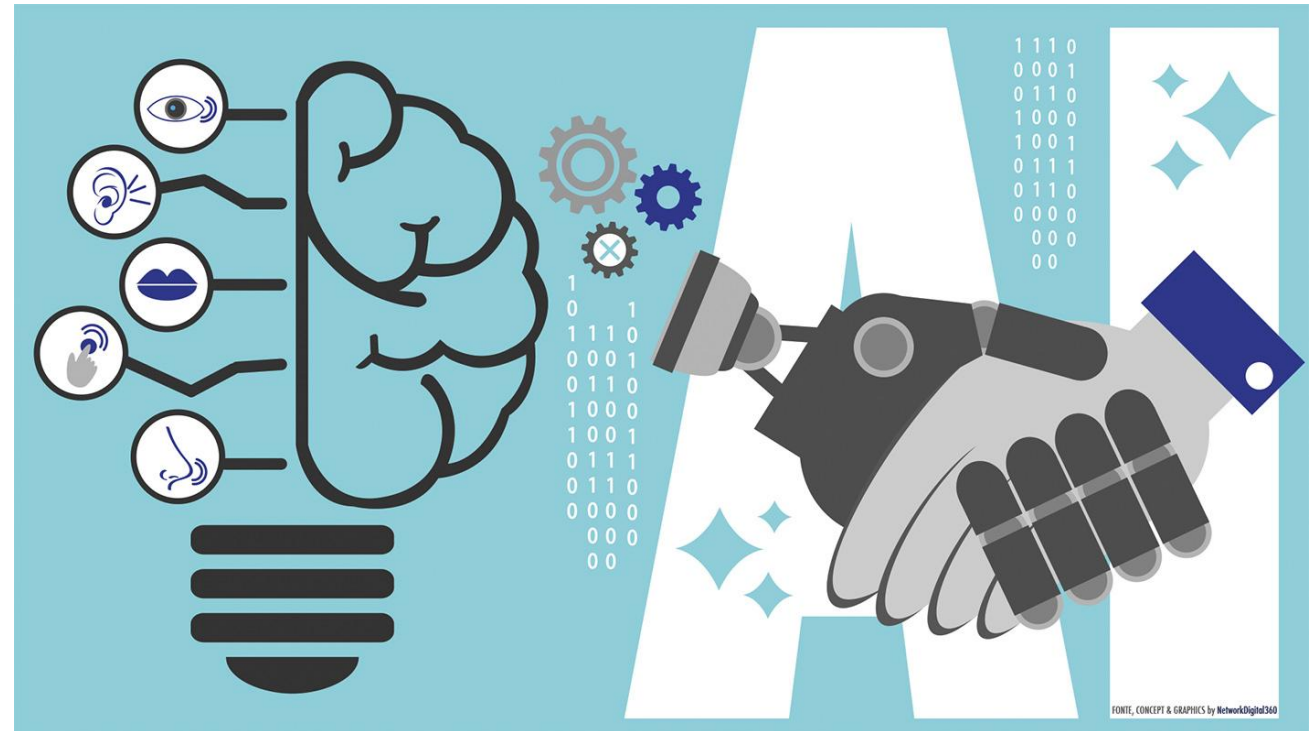
Super AI

Machines that are an order of magnitude smarter than humans

L'IA esegue con successo un compito solo se può slegare la sua esecuzione dall'esigenza di essere intelligente nell'eseguirlo.

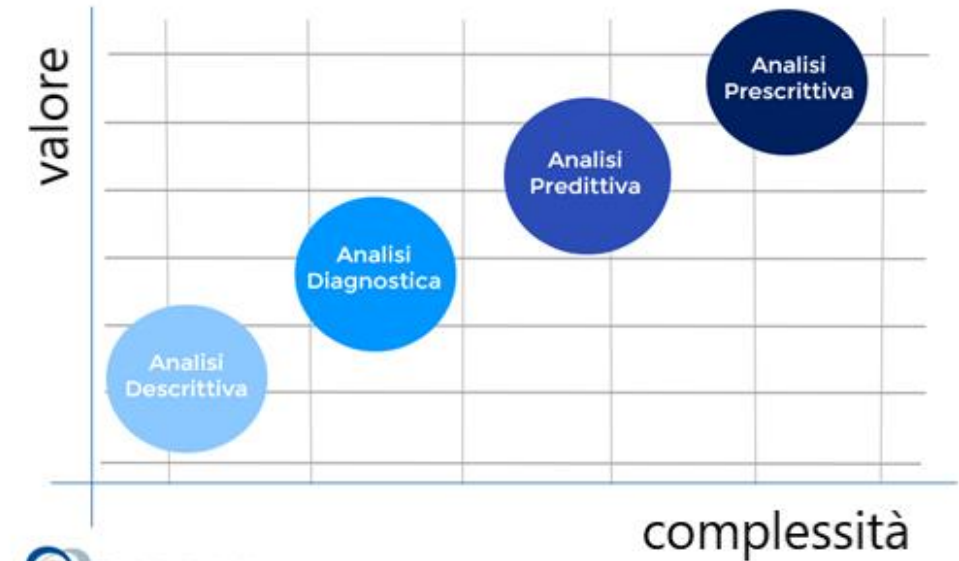
«Comprendere l'IA come un divorzio tra l'agire e l'intelligenza genera conseguenza profonde, così come profonde sono le sfide etiche a cui dà origine tale divorzio» (L. Floridi, **Etica dell'intelligenza artificiale**).

Siamo quindi di fronte ad un **agire artificiale** che, sotto la guida di un progetto umano, aumenta le possibilità a nostra disposizione offrendoci un'immensa e crescente libertà di strutturare e organizzare il mondo in una moltitudine di modi, per risolvere una varietà di problemi, vecchi e nuovi.



Modelli predittivi:

- utilizzano dati che rappresentano fatti o eventi, per i quali sono noti a priori l'evoluzione finale o i risultati. Tali informazioni vengono poi sfruttate per costruire **un modello in grado di predire**, esplicitamente, l'evoluzione e i risultati di altri fenomeni di interesse.



Modelli descrittivi:

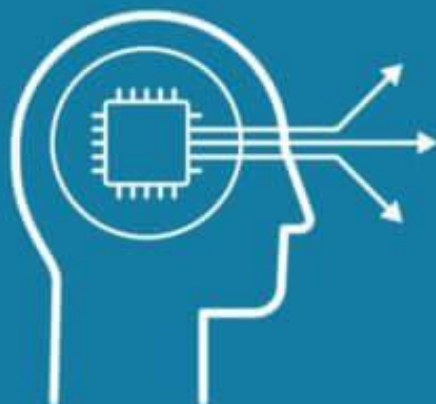
- lo scopo è quello di **caratterizzare** le proprietà generali dei dati analizzati.

Modelli Prescrittivi:

- lo scopo è **fornire raccomandazioni** concrete su come ottimizzare il processo in esame



1950



ARTIFICIAL INTELLIGENCE

ENGINEERING OF MACHINES
THAT MIMIC COGNITIVE FUNCTIONS

1980



MACHINE LEARNING

ABILITY TO PERFORM TASKS
WITHOUT EXPLICIT INSTRUCTIONS
AND RELYING ON PATTERNS

2010

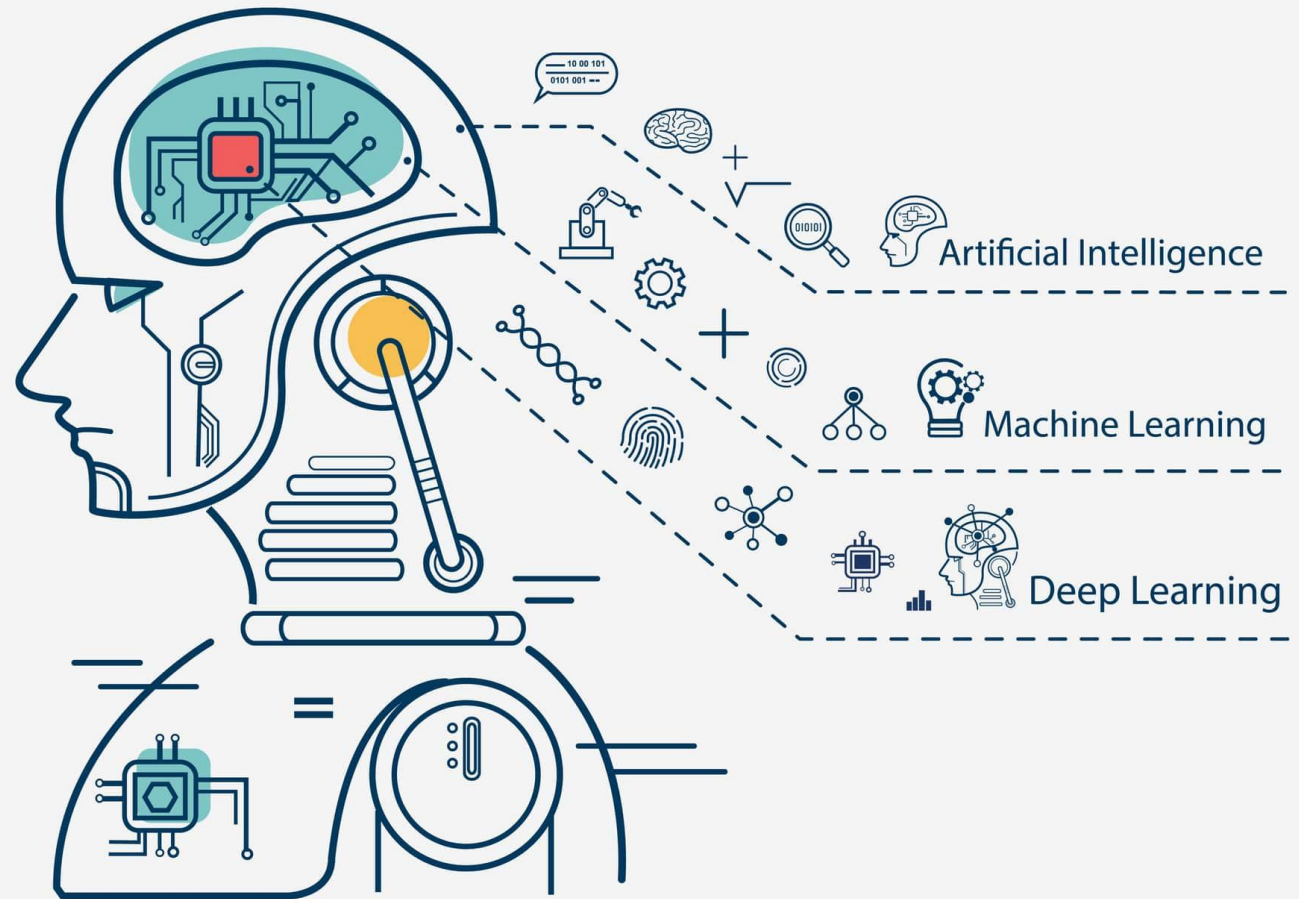


DEEP LEARNING

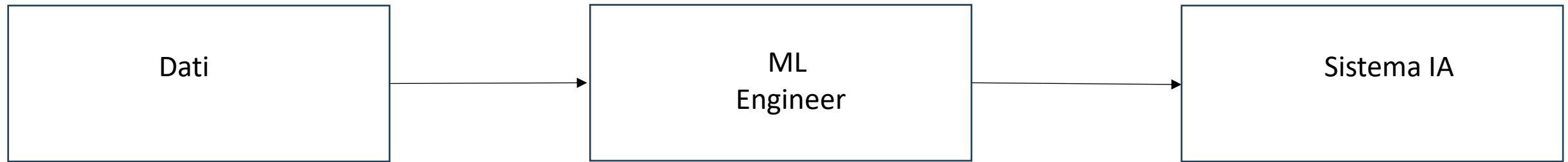
MACHINE LEARNING BASED
ON ARTIFICIAL NEURAL NETWORKS

Il Machine Learning (ML) è un sottoinsieme dell'intelligenza artificiale (AI) che si occupa di creare sistemi che apprendono o migliorano le performance in base ai dati che utilizzano.

«Si definisce macchina che apprende dall'esperienza E con riferimento ad alcuni compiti T e con misurazione delle performance P , la macchina le cui performance nel compito T , misurate da P , migliorano con l'esperienza E »



Il sistema apprende dai dati. E' inizialmente agnostico, poi sviluppa una rappresentazione della conoscenza relativa ad i dati sottoposti.

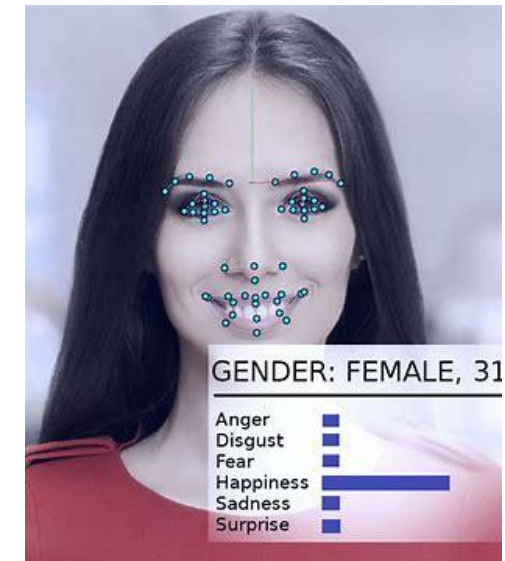


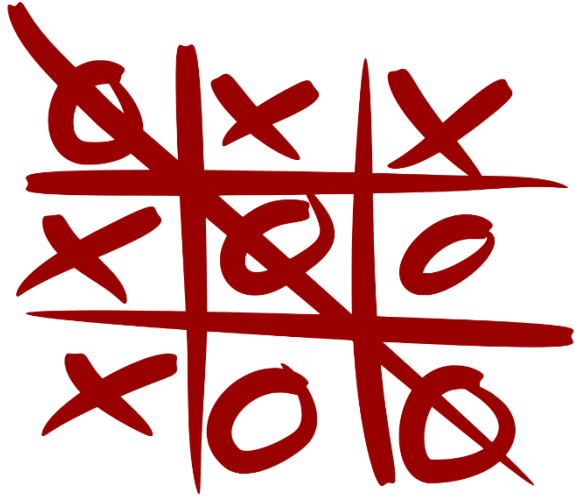
Vantaggi

- NON necessità di uno o più domain expert
- Le prestazioni NON dipendono dalla qualità del domain expert
- La creazione del sistema NON è onerosa
- E' possibile definire elementi di elaborazione inconscia

Svantaggi

- Sistemi black-box
- Necessità di dati
- Le prestazioni dipendono dalla qualità (e quantità) dei dati





Gioco del tris. E' possibile definire aspetti cognitivi legati ad un'analisi razionale:

- la strategia segue il percorso logico di minimizzare l'essere sconfitti
- il processamento dell'informazione è formalizzabile attraverso regole (IF-THEN-ELSE) ed è replicabile da una macchina utilizzando le stesse regole.

Riconoscimento voce familiare. Non è legato ad un processamento razionale del nostro cervello:

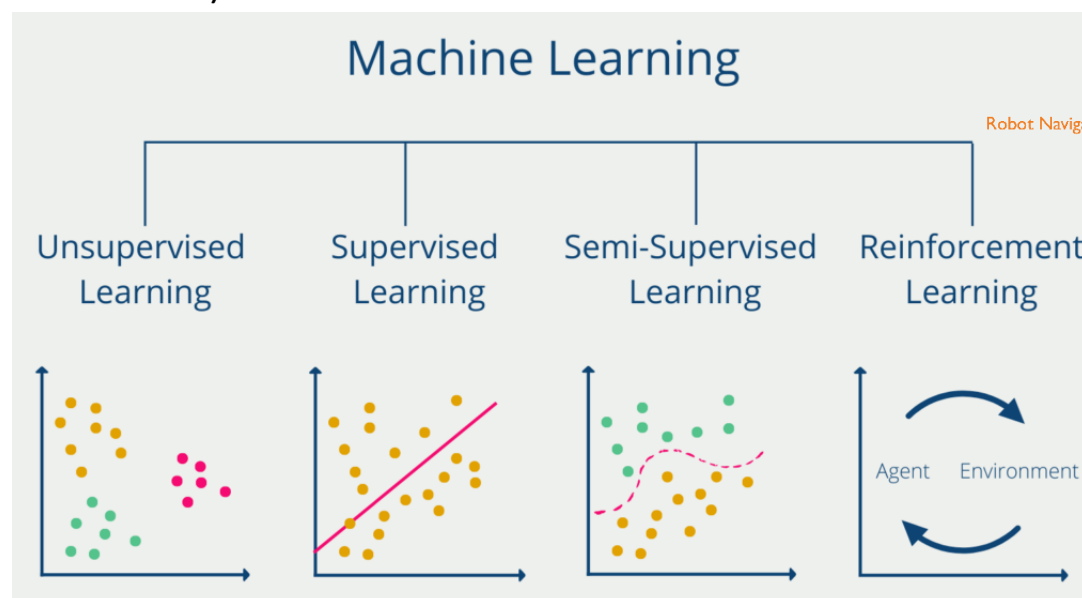
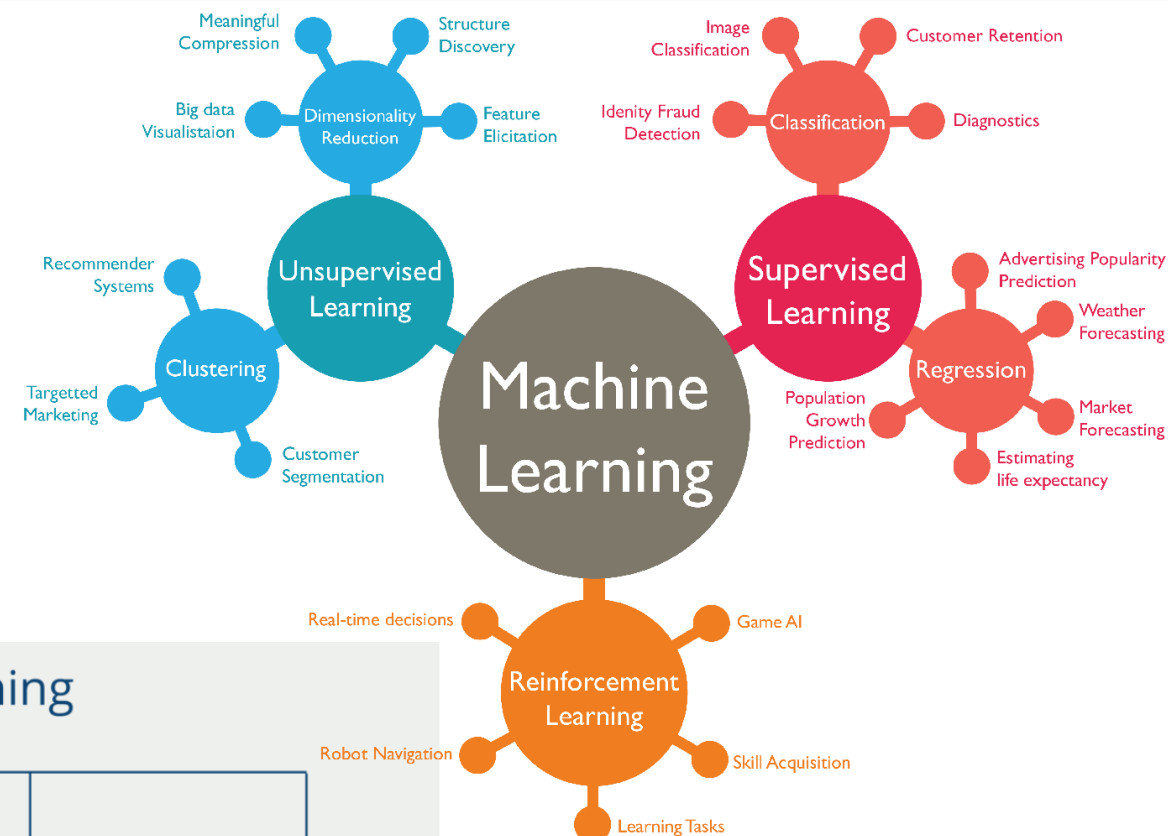
- non è possibile formalizzare le regole che permettono il riconoscimento di una voce.
- l'elaborazione avviene attraverso un'elaborazione legata più al concetto di intuizione.



I sistemi ML sono in grado di simulare questo aspetto del nostro cervello, legato ad un'elaborazione NON razionale dell'informazione, se viene fornito in input un insieme di dati sufficiente.

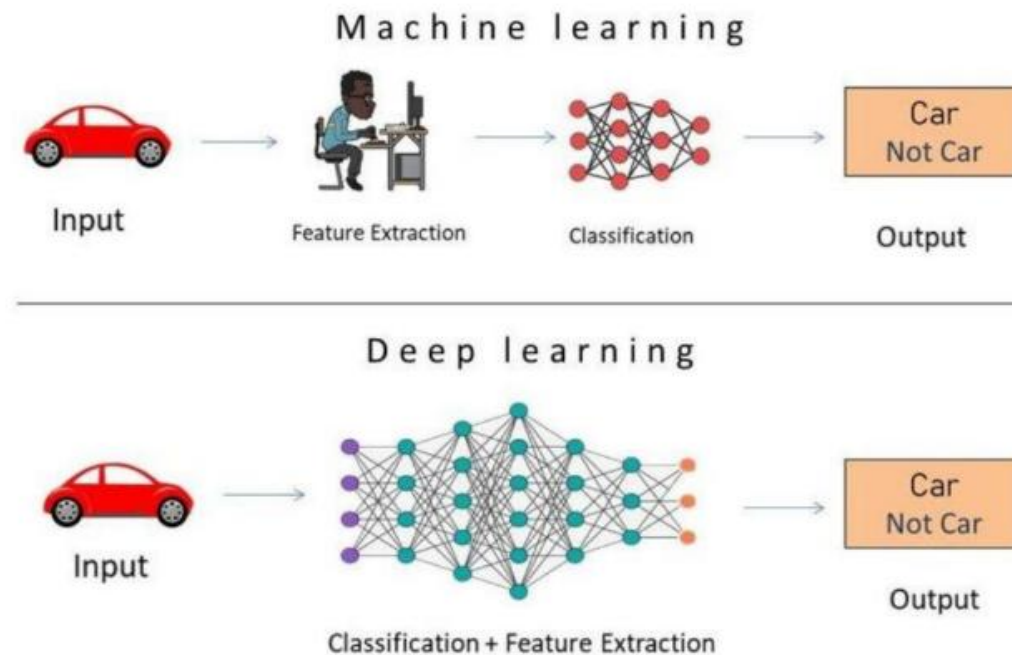
Tipi di apprendimento

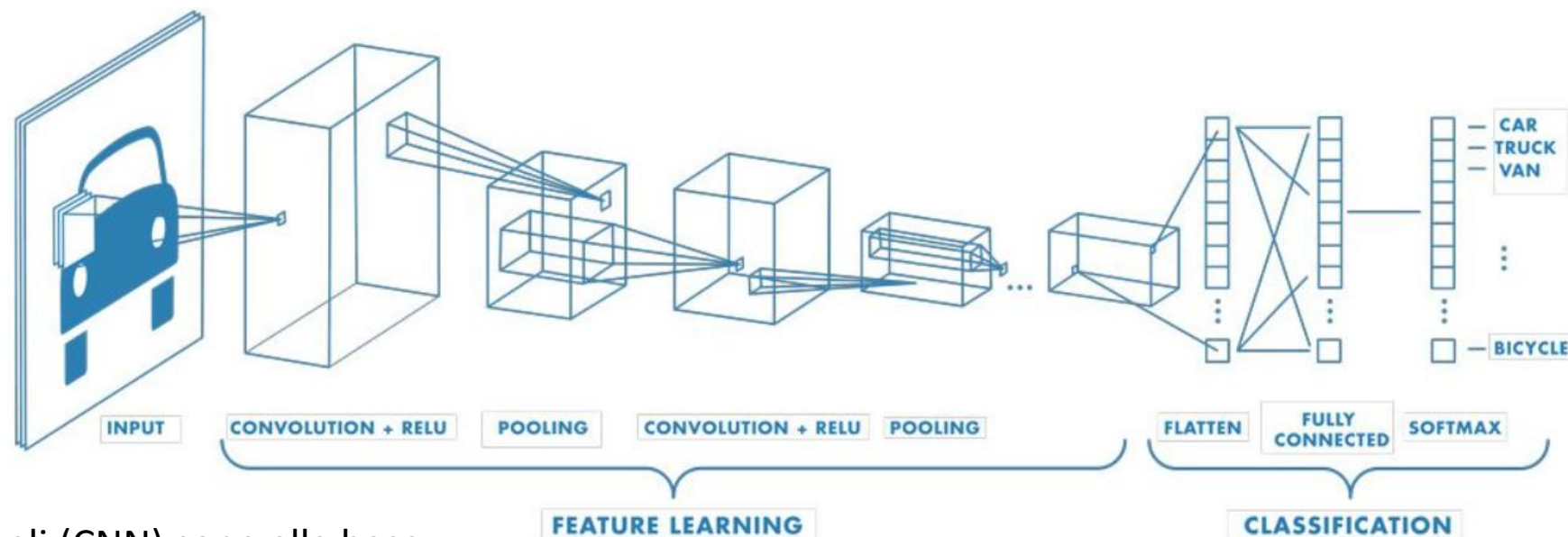
- **Apprendimento Supervisionato:** i dati sono coppie input (V.I.) / output (V.D.)
 - Regressione (continuo)
 - Classificazione (discreto)
- **Apprendimento non supervisionato:** i dati sono le V.I. Il sistema ricrea delle classi di appartenenza negli output (es. carrelli della spesa).
- **Apprendimento per rinforzo:** non sono presenti dati. Esiste un sistema di premi/punizioni per ricreare coppie input/output attraverso l'esplorazione (es. partita a scacchi).
- Altri tipi di apprendimento



Deep Learning vs. Machine Learning

- ❑ **Capacità di apprendimento delle rappresentazioni complesse:** Le reti neurali possono apprendere automaticamente rappresentazioni complesse e stratificate dei dati, consentendo loro di affrontare problemi più intricati.
- ❑ **Adattabilità a grandi volumi di dati:** Le reti neurali tendono a beneficiare della disponibilità di grandi quantità di dati di addestramento, il che può migliorare le prestazioni del modello.
- ❑ **Rappresentazione automatica delle features:** Le reti neurali possono imparare automaticamente a rappresentare le features, riducendo la dipendenza dall'ingegneria manuale delle features.
- ❑ **Eccellenza in problemi complessi:** Sono particolarmente efficaci in problemi complessi come il riconoscimento di immagini, il riconoscimento del linguaggio naturale e altri compiti che coinvolgono dati non strutturati e complessi.
- ❑ **Flessibilità architetturale:** Le reti neurali possono essere adattate a una vasta gamma di compiti e possono essere progettate con diverse architetture, come reti neurali convoluzionali per il riconoscimento di immagini o reti neurali ricorrenti per il riconoscimento del linguaggio naturale.





Le reti neurali convoluzionali (CNN) sono alla base dell'analisi di immagini. Strutturate in strati convoluzionali, di pooling e completamente connessi, le CNN eccellono in:

- **Riconoscimento di oggetti:** Modelli come ResNet, VGG e Inception sono utilizzati per classificare immagini.
- **Segmentazione di immagini:** Architetture come U-Net e Mask R-CNN permettono di isolare oggetti specifici nelle immagini.
- **Riconoscimento facciale:** Utilizzate in sistemi di sicurezza e autenticazione biometrica.

Architettura delle CNN

Le CNN sono composte da diversi tipi di strati:

- **Strati convoluzionali:** Applicano filtri ai dati in ingresso per estrarre caratteristiche rilevanti.
- **Strati di pooling:** Riduzione della dimensionalità per migliorare l'efficienza computazionale.
- **Strati completamente connessi:** Aggregano le informazioni estratte per la classificazione finale.



Natural Language Processing (NLP) è una disciplina dell'intelligenza artificiale (AI) che si occupa di elaborare, comprendere e generare il linguaggio umano utilizzando i computer.

Obiettivo principale: colmare il divario tra linguaggio umano (non strutturato) e linguaggio macchina (strutturato).

Domini di Applicazione

Comprensione del linguaggio naturale (Natural Language Understanding - NLU): analisi di significati, sentimenti e intenzioni da un testo.

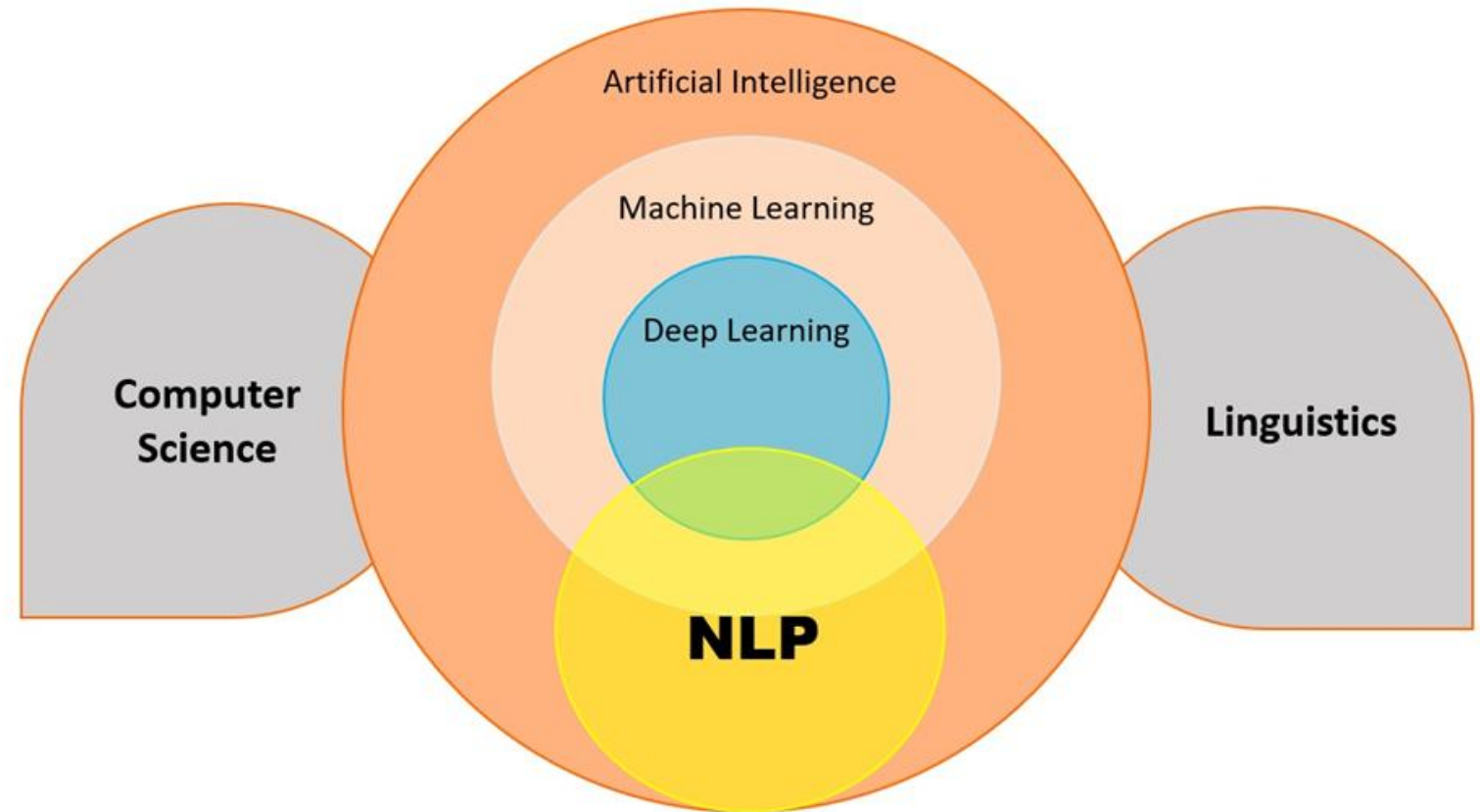
Generazione del linguaggio naturale (Natural Language Generation - NLG): produzione di testo simile a quello scritto da un essere umano.

Perché è Importante?

Il linguaggio umano è la forma più naturale di comunicazione, ma è complesso, ambiguo e ricco di sfumature.

Il NLP consente di:

- ✓ Automatizzare processi basati sul linguaggio.
- ✓ Estrarre informazioni utili da grandi quantità di dati testuali.
- ✓ Migliorare l'interazione uomo-macchina.



Componenti Principali

- ✓ **Analisi sintattica:** studio della struttura grammaticale delle frasi.
- ✓ **Analisi semantica:** comprensione del significato delle parole e delle frasi.
- ✓ **Analisi pragmatica:** interpretazione del linguaggio in base al contesto.

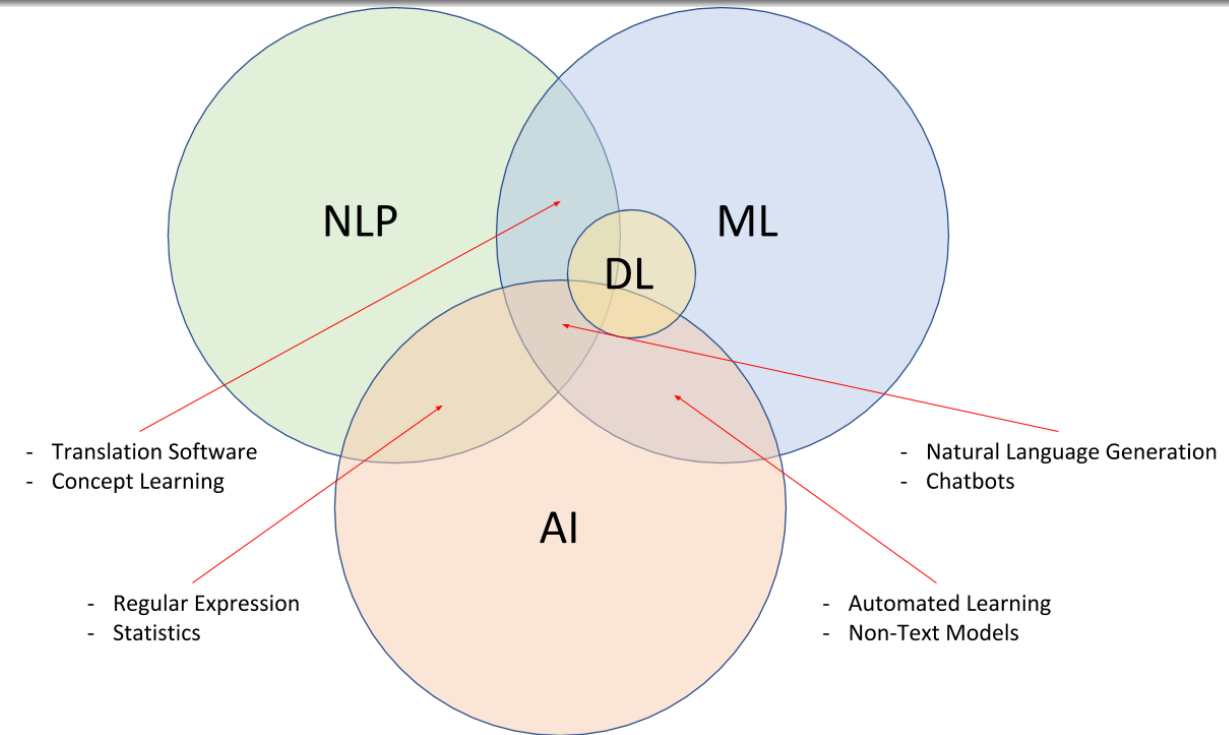
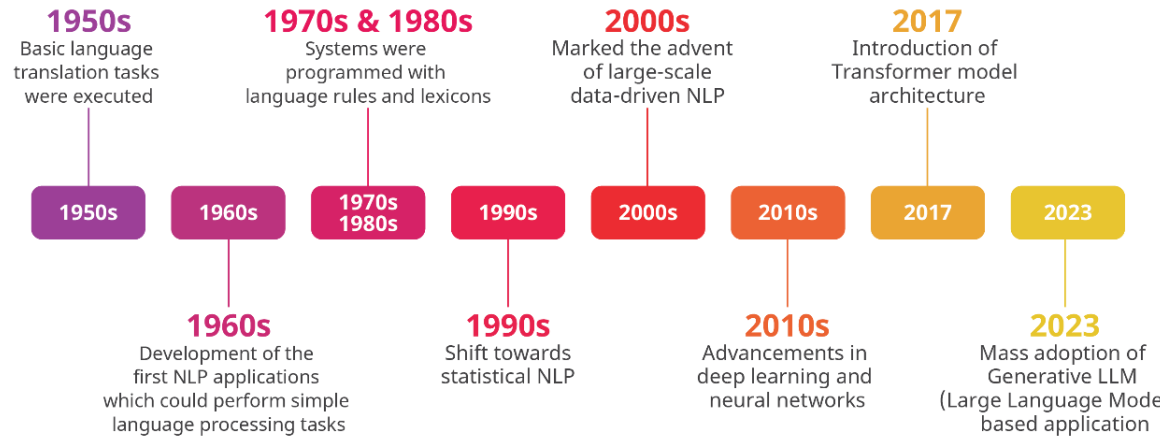
Esempi Pratici

- ✓ Motori di ricerca (Google, Bing).
- ✓ Traduttori automatici (DeepL, Google Translate).
- ✓ Assistenti virtuali (Siri, Alexa, Google Assistant).
- ✓ Analisi di sentimenti nei social media o nelle recensioni



NLP: Approcci e Tecnologie

Genesis and Evolution of Natural Language Processing



Schema Riassuntivo

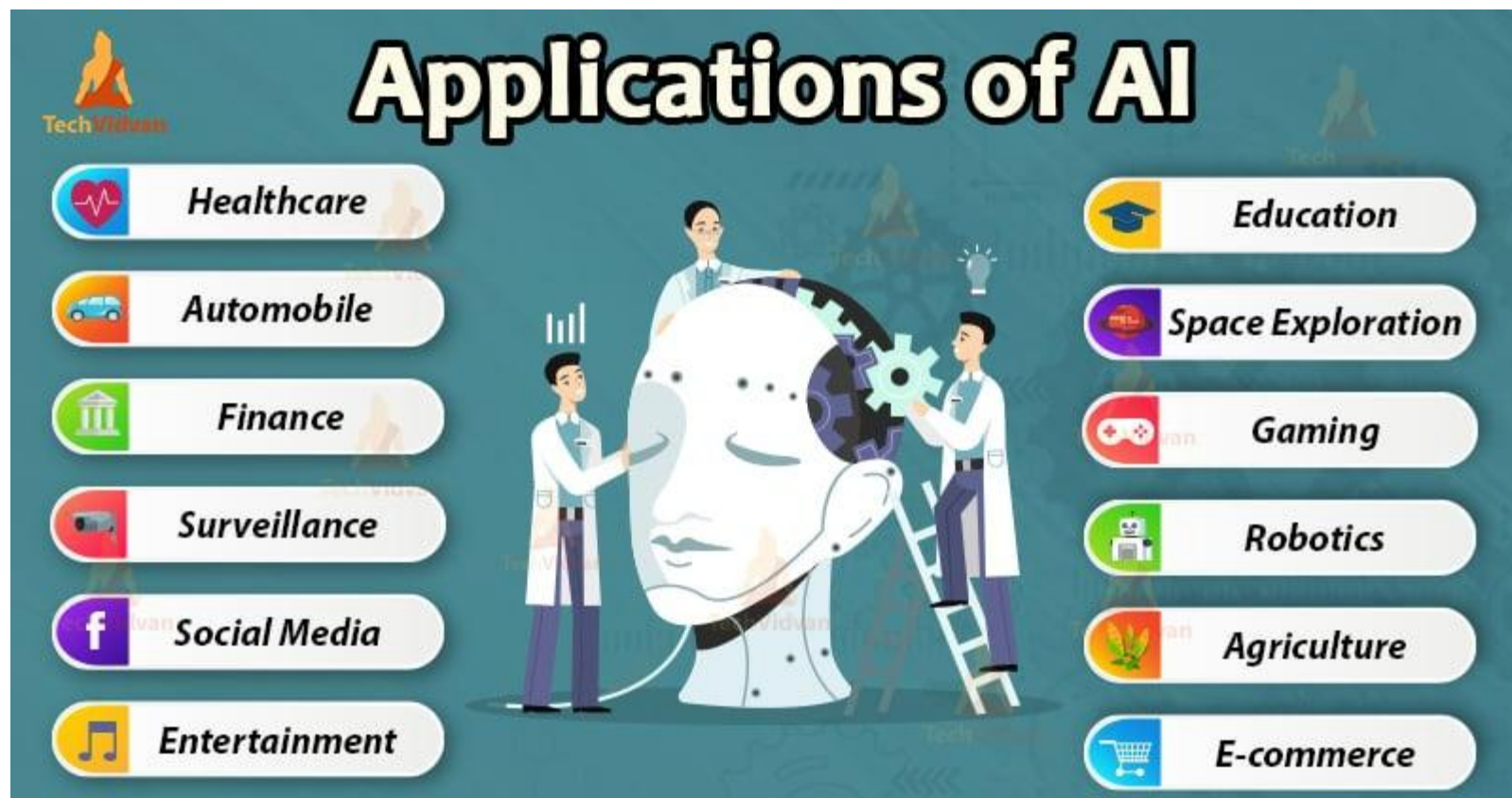
Approccio	Vantaggi	Svantaggi
Basato su regole	Semplice, intuitivo	Poco scalabile, rigido
Statistico	Gestisce variazioni linguistiche	Non comprende il contesto
Machine Learning	Flessibile, adattabile	Richiede dati annotati
Deep Learning	Elevata precisione, contestuale	Costo elevato, complesso

- Modello linguistico di Intelligenza Artificiale sviluppato da OpenAI, pre-addestrato su grandi quantità di dati testuali, come libri, articoli e pagine web, ed è progettato per generare risposte simili a quelle umane agli input in linguaggio naturale.
- Ma in quale tipologia di Intelligenza Artificiale si colloca Chat GPT? Ecco la sua risposta:

“Come modello linguistico di Intelligenza Artificiale, sono considerato un esempio di Intelligenza Artificiale ristretta (ANI) perché sono stato progettato per svolgere compiti specifici legati all’elaborazione del linguaggio naturale, come rispondere a domande, generare testo e tradurre lingue. Sebbene sia in grado di fornire risposte utili e di impegnarmi in conversazioni con gli esseri umani, non possiedo l’intelligenza ampia e simile a quella umana di un’Intelligenza Artificiale generale (AGI) o di una super-intelligenza (ASI), che sono ancora in gran parte teoriche e non ancora pienamente realizzate”.



- **IA nella sanità**
- IA nella finanza
- IA nell'automazione e robotica
- IA nella mobilità e trasporti
- IA nella sicurezza e sorveglianza
- IA nei social media
- **IA nell'agricoltura**
- IA nella domotica
- IA nella logistica



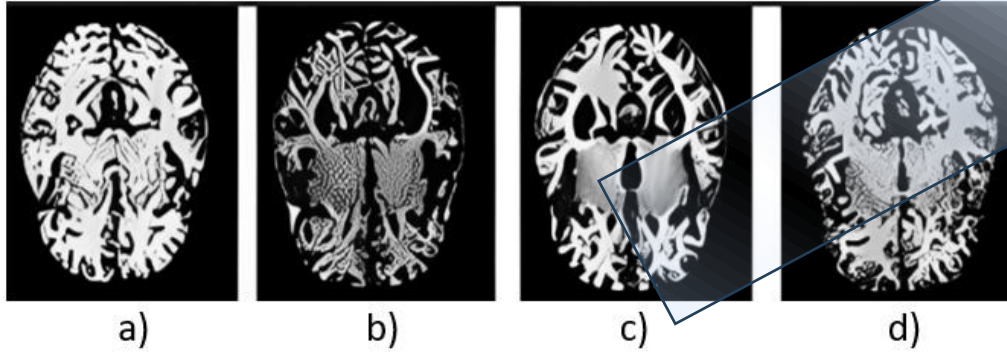
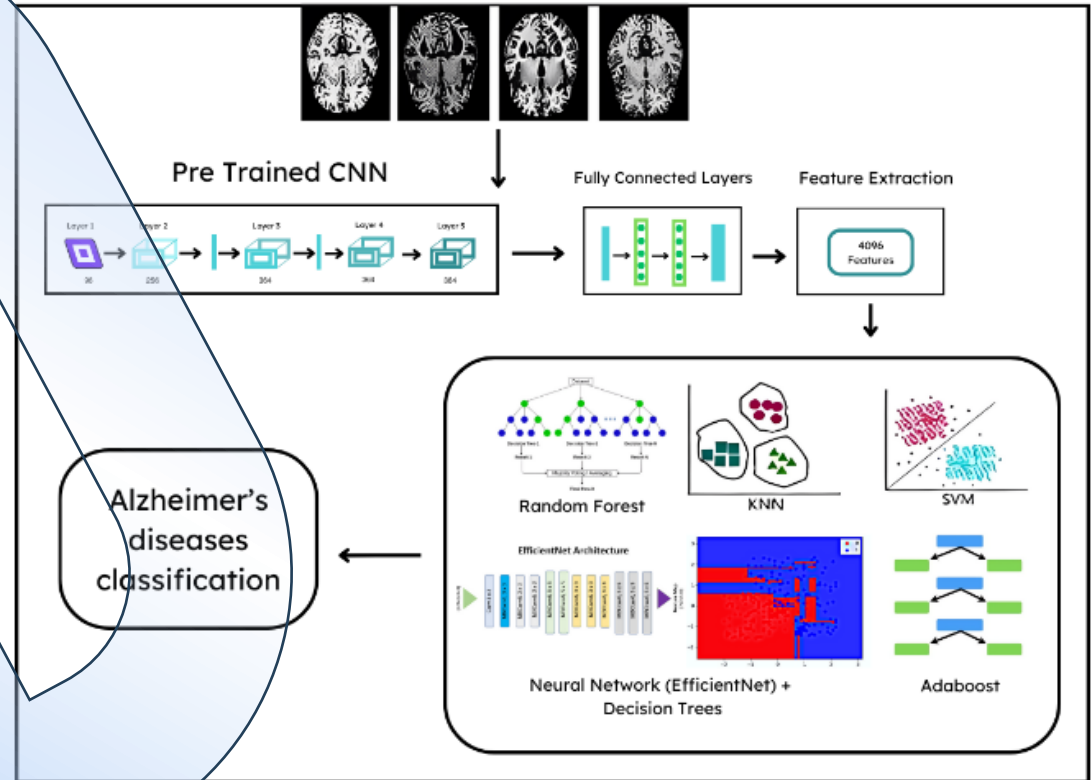


Figure 1. Raw dataset of Alzheimer's diseases (a) Very mild dementia (b) Mild dementia (c) Moderate dementia and (d) Not dementia

	Predicted				Σ
	Mild_Demented	Moderate_Demented	Non_Demented	Very_Mild_Demented	
Mild_Demented	223	2	21	4	250
Moderate_Demented	4	51	2	3	60
Non_Demented	5	4	800	21	830
Very_Mild_Demented	10	0	41	209	260
Σ	242	57	864	237	1400



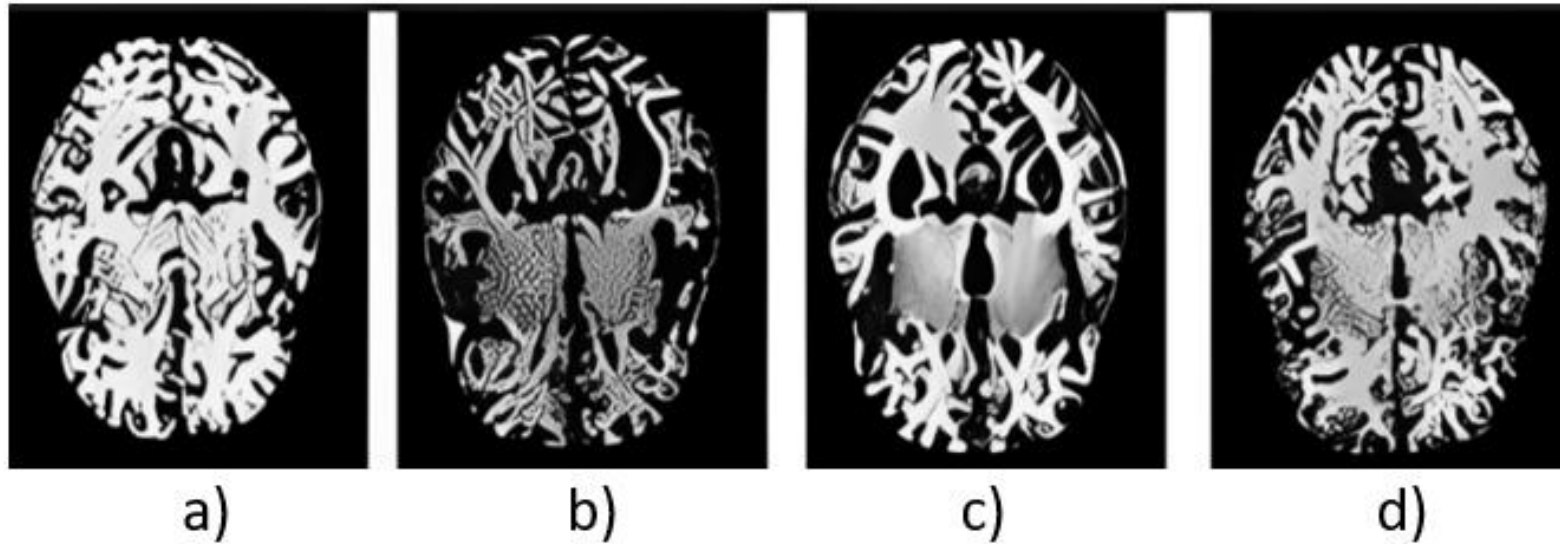
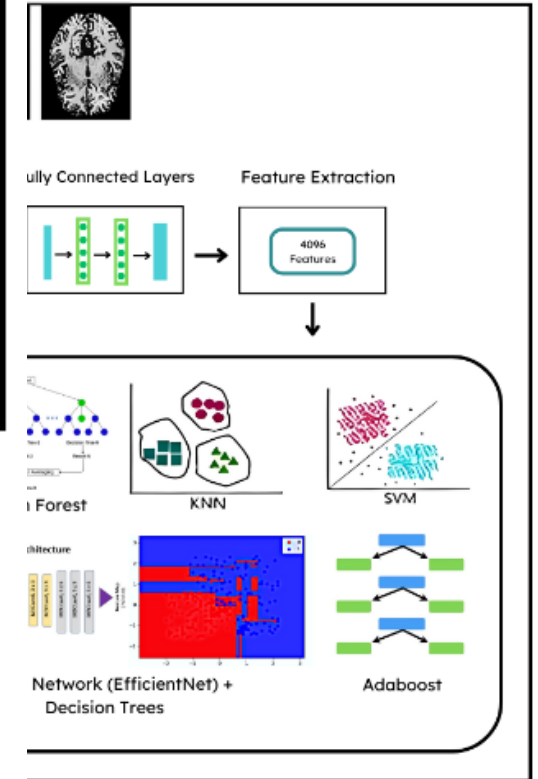


Figure 1. Raw dataset of Alzheimer's diseases (a) Very mild dementia (b) Mild dementia (c) Moderate dementia and (d) Not dementia

Actual	Mild_Demente				
	Moderate_Demente				
	Non_Demente				
	Very_Mild_Demented	10	0	41	209
	Σ	242	57	864	237
					1400



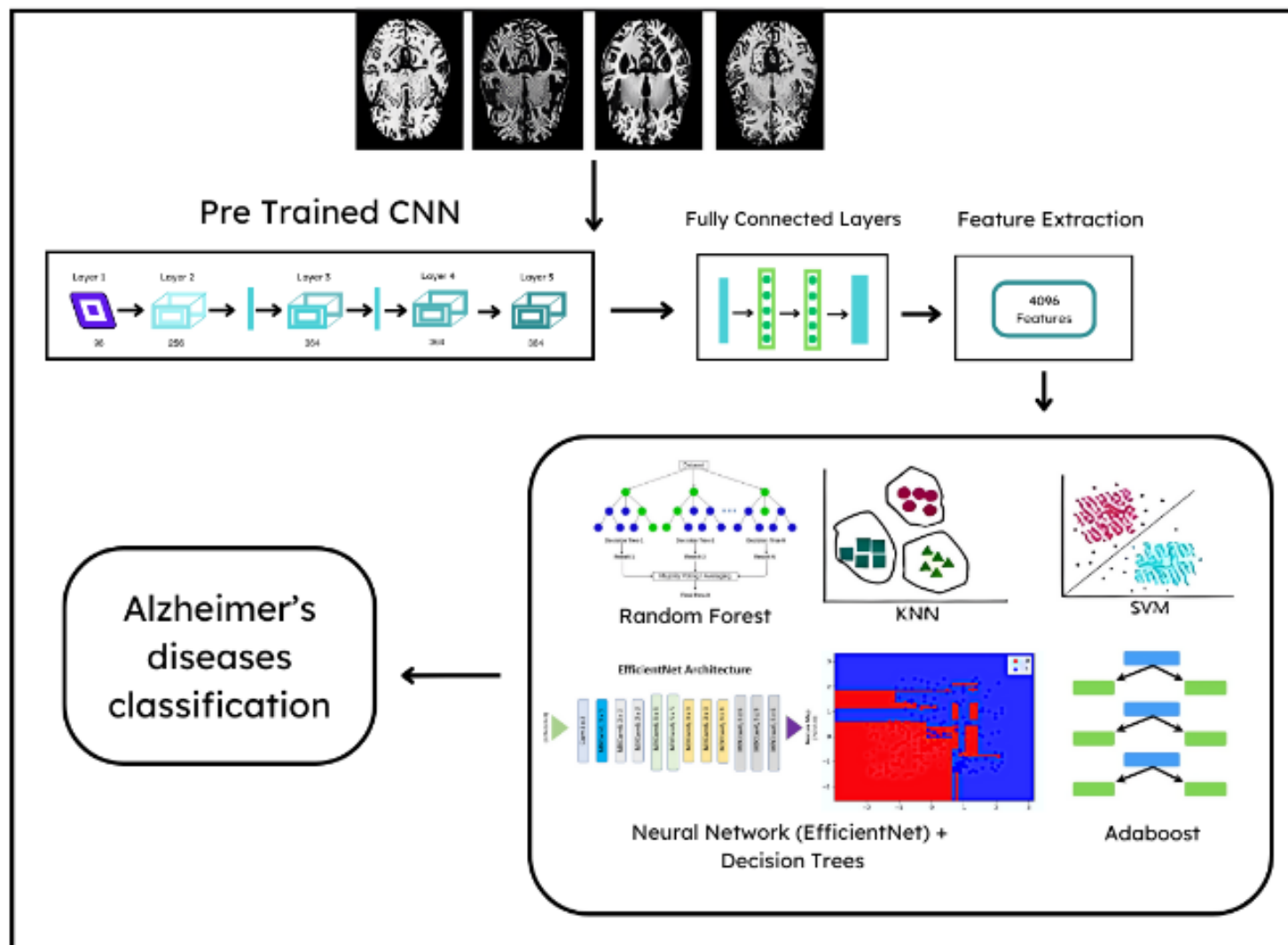
IA nella sanità: Diagnosi di Alzheimer



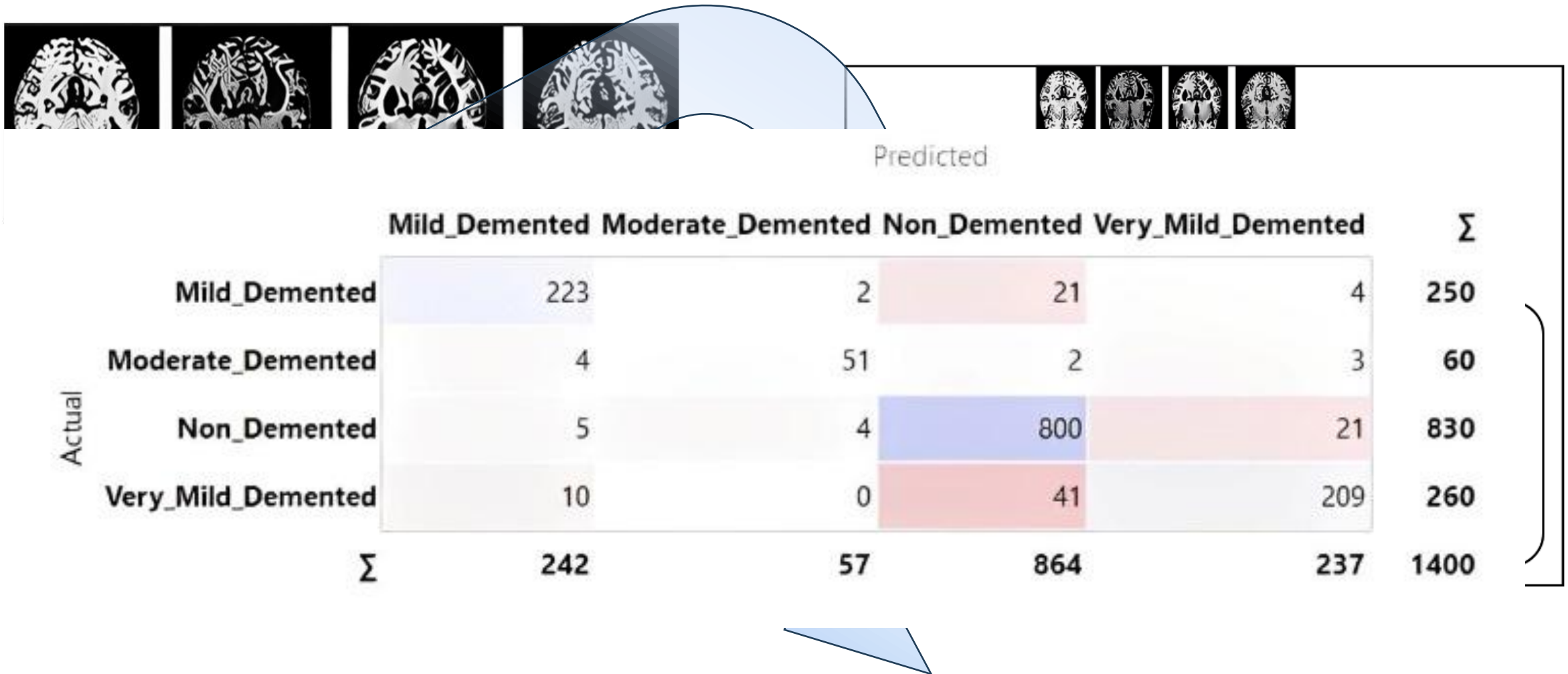
a)

Figure 1. Raw d
mild dementia (b

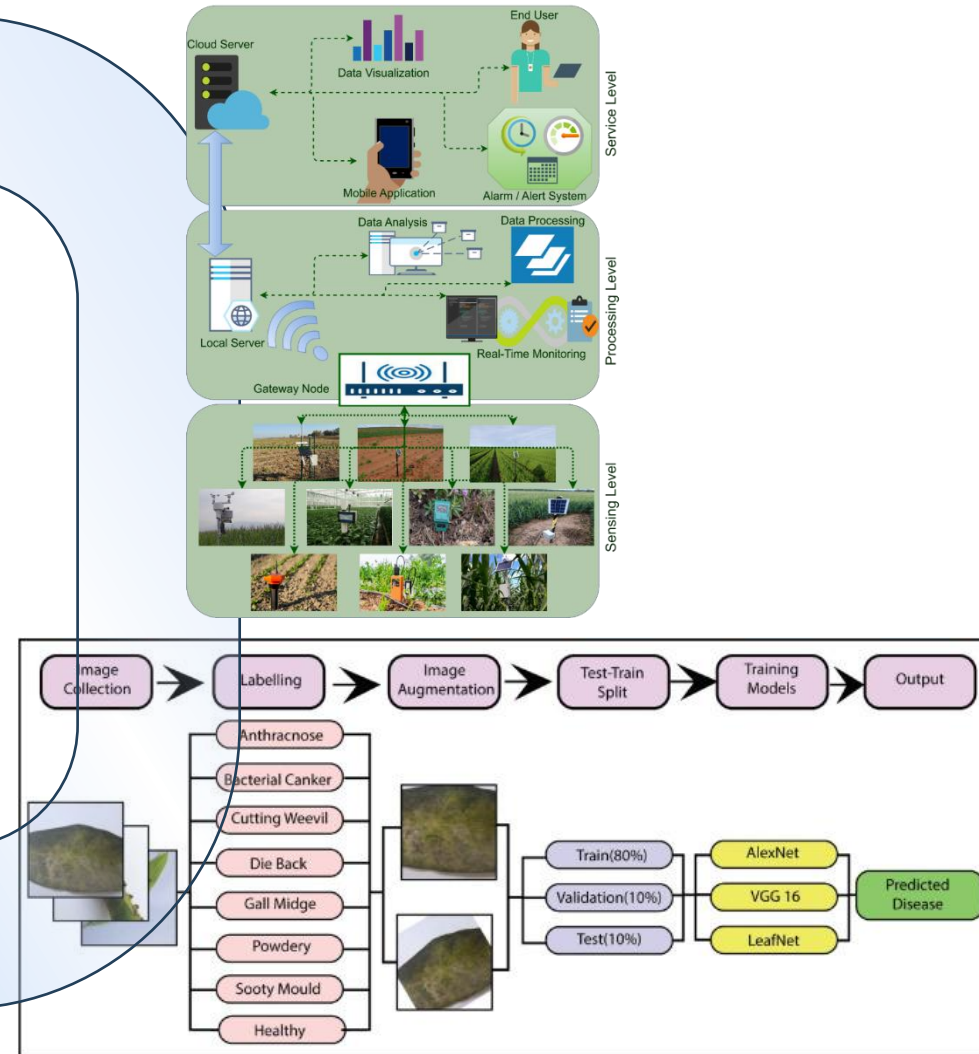
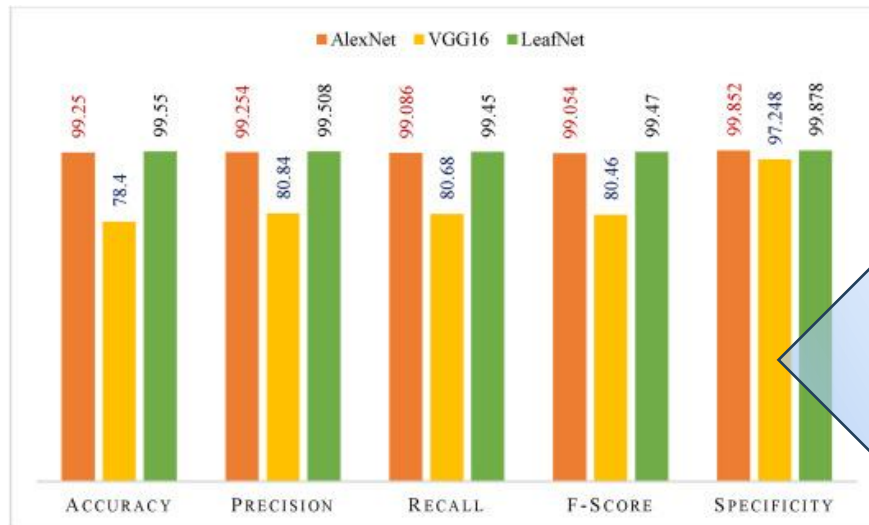
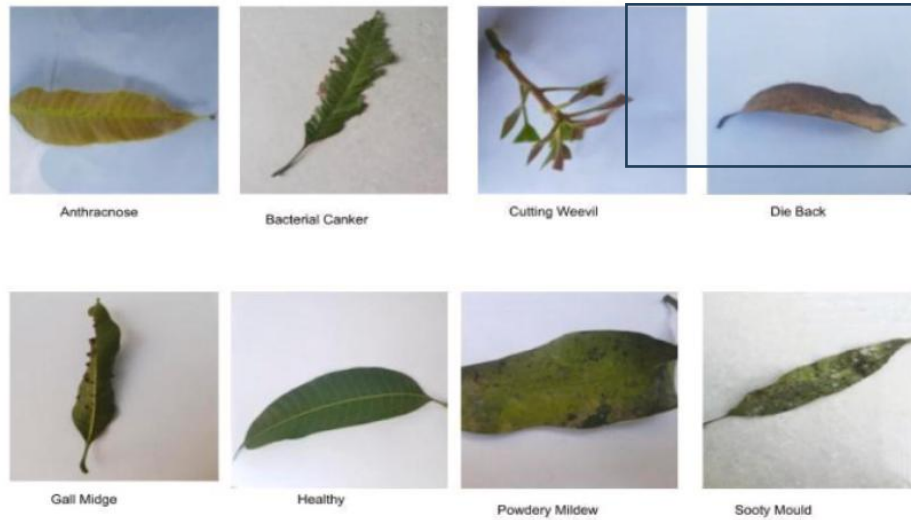
	Mild_
Mild_Demented	
Moderate_Demented	
Non_Demented	
Very_Mild_Demented	
Actual	
	Σ



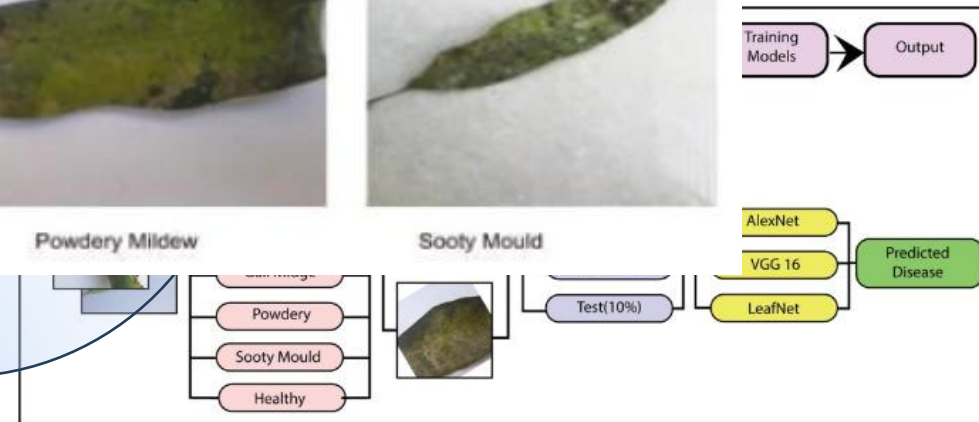
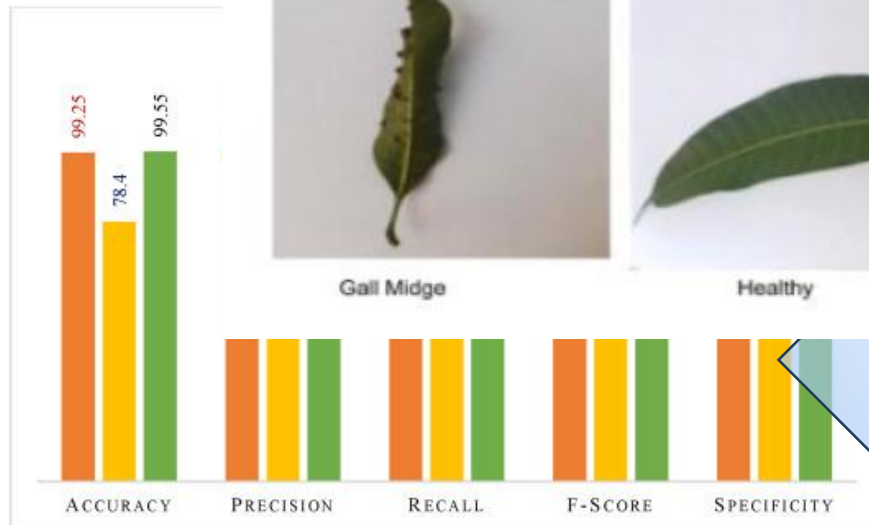
IA nella sanità: Diagnosi di Alzheimer



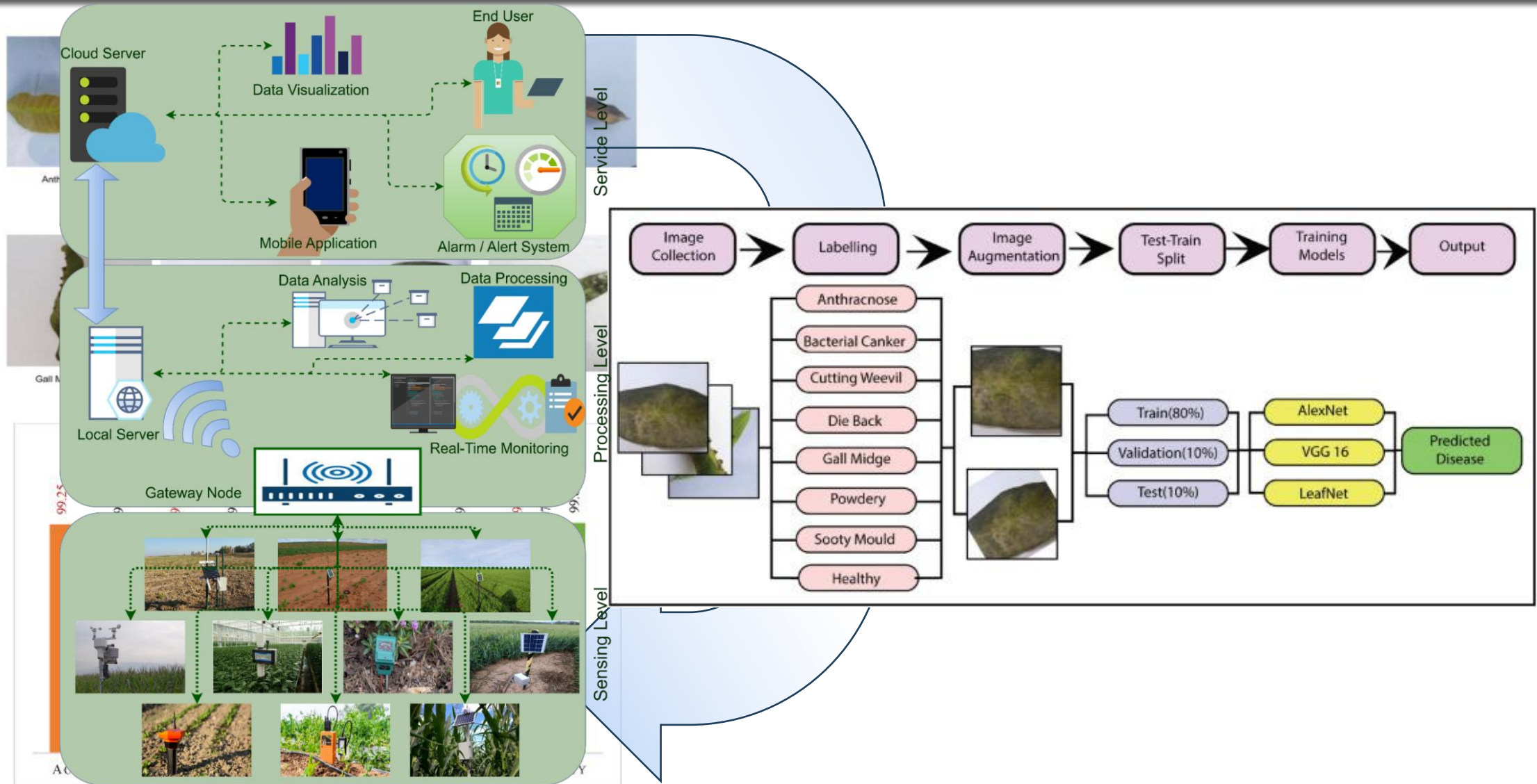
IA nell'agricoltura 4.0



IA nell'agricoltura 4.0



IA nell'agricoltura 4.0

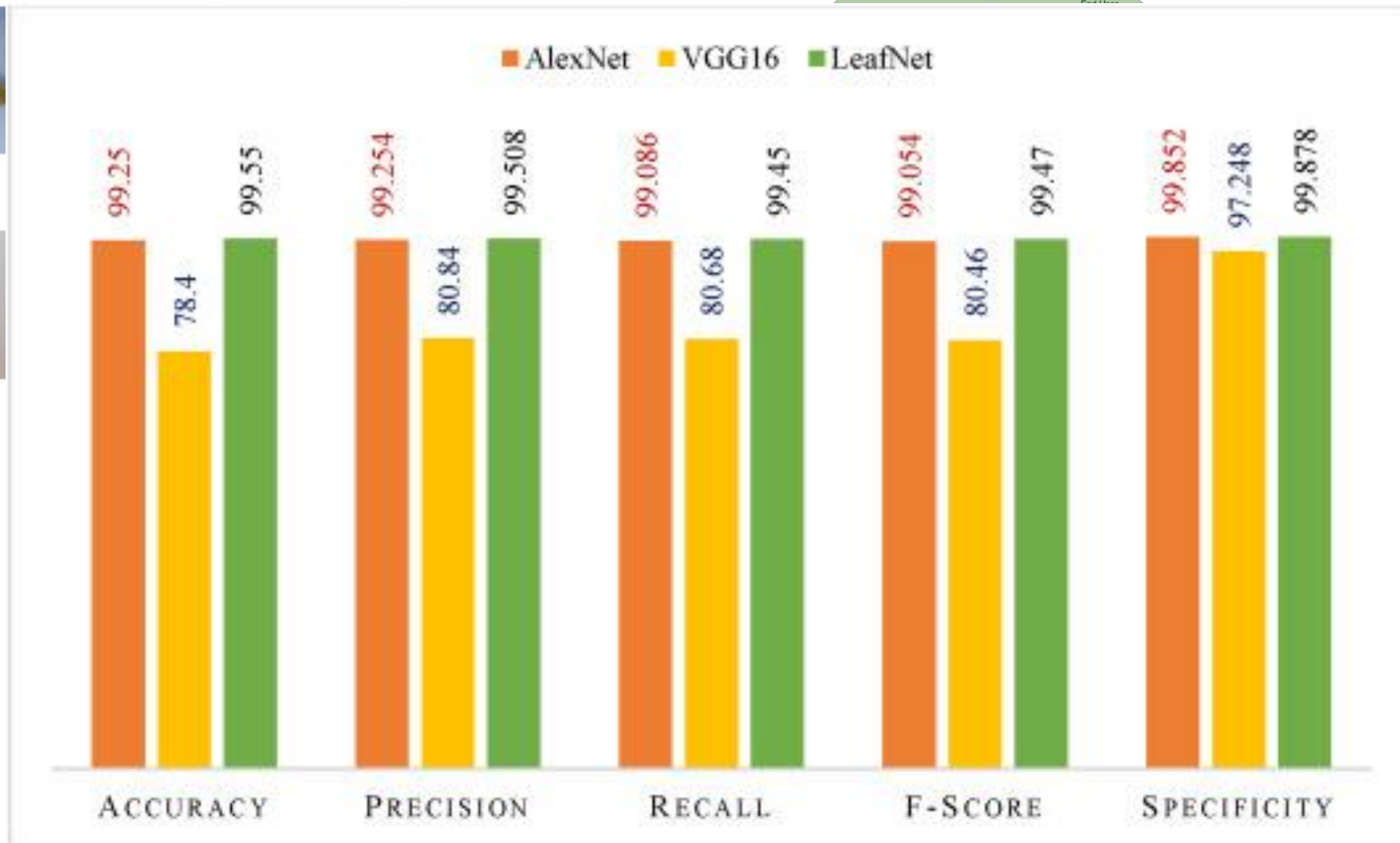




Anthracnose



Gall Midge



IA agnostica e bias dei dati



L'IA è veramente imparziale?

Negli ultimi anni è stato osservato come, anche in assenza di un'esplicita richiesta o input volto a discriminare certi profili, i sistemi di apprendimento automatico hanno in certi casi condotto a risultati discriminatori, problema che viene definito come “automazione nella discriminazione”.

Un ulteriore rischio che deriva dall'utilizzo di queste tecnologie è poi quello di riprodurre elementi di discriminazione in modalità non previsti da chi li ha progettati.

AI Bias: i sistemi di IA di analisi dei dati mostrano atteggiamenti discriminatori nei confronti di determinati gruppi di persone. Solitamente questi pregiudizi riflettono opinioni e preconcetti socialmente diffusi come, ad esempio, quelli su razza, genere, sesso biologico, età e cultura.



Il problema nasce quando il campione di dati su cui si fonda l'intero meccanismo risulta distorto (biased) da cui nasce appunto il bias che genera automazione nella discriminazione e rende inaffidabile, parziale e potenzialmente pericoloso l'utilizzo dell'intelligenza artificiale.

Se i dati utilizzati per addestrare il sistema sono sbilanciati o incompleti, il sistema potrebbe trarre conclusioni errate o fare previsioni imprecise



Esempio 1:

Sistema OPTUM, utilizzato negli ospedali degli Stati Uniti per prevedere quali pazienti avrebbero potuto avere bisogno di ulteriori cure mediche.

Questo sistema è stato oggetto di dibattito a seguito della scoperta di alcune teorie secondo cui l'algoritmo discriminava i pazienti neri rispetto a quelli bianchi, nonostante il colore della pelle non fosse una delle variabili considerate.

Tuttavia, un dato correlato, ovvero la cronologia dei costi sanitari, ha mostrato che i pazienti di colore avevano sostenuto in media spese sanitarie inferiori rispetto ai pazienti bianchi con le stesse condizioni. Se questo errore del modello di intelligenza artificiale non fosse stato individuato, il sistema sanitario avrebbe continuato a discriminare gravemente e ingiustamente una parte significativa della popolazione.

Esempio 2:

Algoritmo di software di recruitment utilizzato da Amazon a partire dal 2014.

Questo software, era ritenuto in grado di analizzare i curriculum dei candidati ed automatizzare la procedura di selezione.

Tuttavia è emerso come esso penalizzasse le donne, specialmente per le posizioni legate a ruoli più tecnologici.

L'errore era dovuto ai dati con cui il modello è stato addestrato: dati reali, contenenti i curricula ricevuti dalla società nei 10 anni precedenti; CV prettamente maschili, data la maggioranza di uomini nel settore tecnologico.

Esempio 3:

Sistema COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), utilizzato in alcuni sistemi giudiziari statunitensi per prevedere la probabilità di recidiva di un imputato.

Questo software presentava un forte pregiudizio che ha portato a un tasso di falsi positivi per recidiva, doppio per i trasgressori neri (circa il 45%) rispetto ai trasgressori caucasici (circa il 23%).

Esempio 4:

Nel 2018 è emerso che **tre programmi di riconoscimento facciale** rilasciati sul mercato da importanti aziende tecnologiche incorporavano pregiudizi razziali e di genere.

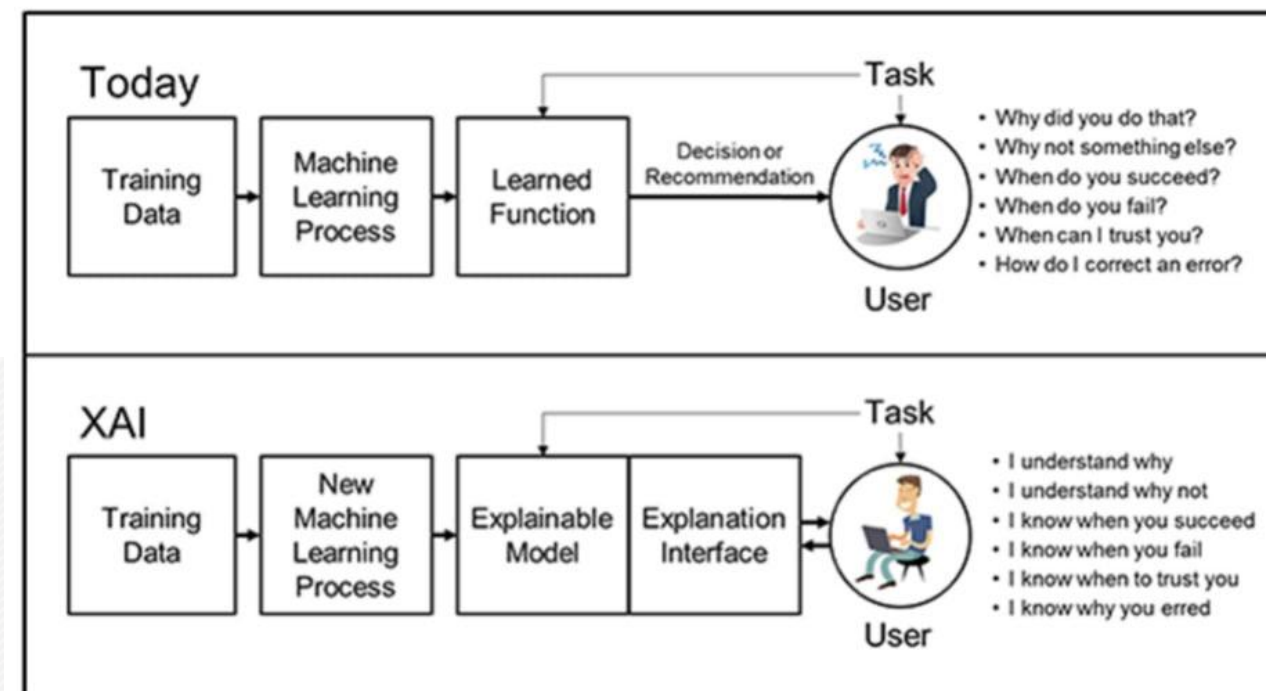
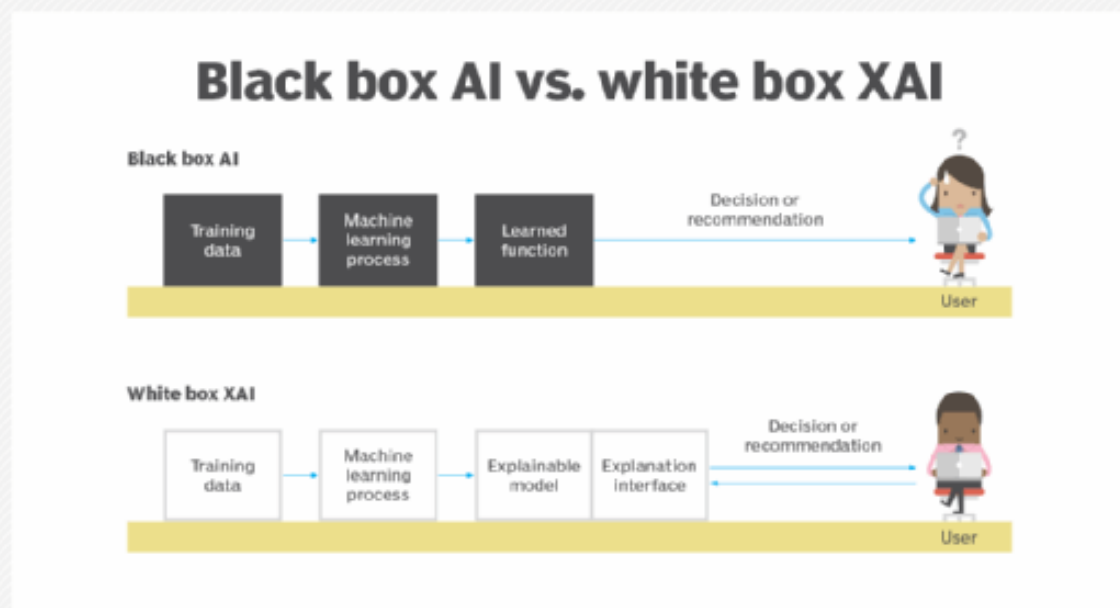
Negli esperimenti condotti è stato rilevato che nel determinare il sesso degli uomini di pelle chiara i tassi d'errore dei programmi di riconoscimento facciale non hanno mai superato lo 0,8% mentre, per le donne con pelle scura, le percentuali salivano al 20% in un programma e ad oltre il 34% negli altri due.

- Utilizzare dati rappresentativi e di alta qualità per addestrare il sistema.
- È inoltre importante progettare algoritmi equi e inclusivi e verificare che l'interfaccia utente del sistema non presenti elementi biased o discriminatori.
- Attualmente però la migliore soluzione è quella di adottare sistemi ibridi, ovvero che utilizzano sia il giudizio umano che l'intelligenza artificiale.
- Se da un lato gli algoritmi possono portare maggiore precisione e riproducibilità a un processo e sono in grado di individuare e rimuovere i pregiudizi dovuti all'errore umano, il lavoro di un professionista permette di valutare singolarmente ogni situazione, rapportandola in un contesto più ampio, aggiungendo consapevolezza, abilità di ricerca e comprensione del processo decisionale



Possibili soluzioni future: eXplainable AI (XAI)

XAI: Il principio di rendere il funzionamento di un'intelligenza artificiale e i suoi risultati comprensibili per l'utente, il più possibile.



XAI fornisce una spiegazione di come l'IA abbia preso una decisione discutendo i seguenti aspetti:

- I punti di debolezza e forza del programma.
- I criteri usati per arrivare ad una decisione.
- Perché abbia scelto un'alternativa tra altre.
- Il livello di fiducia associato ad ogni decisione.
- A quali errori il programma sia soggetto.
- Come sia possibile correggere gli errori.



IA e opportunità

- **Migliore assistenza sanitaria**

Le nuove tecnologie di intelligenza artificiale possono aiutare i medici a determinare predisposizioni verso malattie e intervenire prima. Non solo, l'intelligenza artificiale può essere utilizzata per migliorare l'accuratezza della diagnosi, per velocizzare lo sviluppo di farmaci e per mitigare gli impatti sanitari ed economici di gravi epidemie e pandemie.

- **Città intelligenti**

Gestione più efficiente dell'energia, dell'acqua e dei rifiuti, miglioramento della mobilità urbana, aumento dell'indipendenza per persone con disabilità, rilevamento degli incendi, prevenzione dagli eventi atmosferici estremi, ma anche miglioramento nella gestione dei flussi turistici e poi coinvolgimento dei cittadini e delle imprese, nell'utilizzo e nella fornitura di dati.

- **Istruzione e formazione**

Questo settore è d'importanza cruciale, è qui infatti che prenderanno forma i nuovi scenari per utilizzare le IA con vantaggio ed evolveranno le pratiche e l'educazione civica per difendersi dai rischi e riconoscere l'uso improprio delle nuove tecnologie. La scuola del prossimo futuro insegnerà alle donne e agli uomini di domani a prepararsi per lavori che ancora non esistono, a usare la prossima generazione di strumenti d'intelligenza artificiale, a diventare cittadini di un mondo che presenterà problemi sociali che non esistevano in passato.



Deepfake e disinformazione: creazione di video e immagini false per diffondere notizie manipolate.

Attacchi informatici avanzati: IA utilizzata per generare malware sofisticati e attacchi automatizzati.

Sorveglianza di massa: IA impiegata per monitorare individui senza il loro consenso.

Automazione di attività criminali: bot autonomi per frodi online e cybercrimine.

Manipolazione psicologica: utilizzo di IA per influenzare opinioni politiche e comportamenti tramite algoritmi predittivi avanzati.



- **Riflessione sulle opportunità offerte dall'IA:** l'intelligenza artificiale ha il potenziale di migliorare innumerevoli aspetti della nostra vita, dall'efficienza lavorativa alla qualità della salute e dell'istruzione.
- **Importanza di un uso responsabile:** è fondamentale adottare regolamentazioni etiche per garantire che l'IA venga utilizzata per il bene comune, evitando discriminazioni e violazioni della privacy.
- **Collaborazione tra tecnologia e umanità:** l'IA dovrebbe essere considerata un alleato dell'essere umano, supportando la creatività, migliorando i processi decisionali e aumentando la produttività senza sostituire il ruolo umano.

L'IA offre enormi opportunità, ma i suoi rischi richiedono attenzione e prevenzione. La consapevolezza dei pericoli è fondamentale per garantire che l'IA venga utilizzata per il bene dell'umanità.



