Introduzione al Machine Learning e alla Data-Analysis

L'era dei Big Data

DATI

Dati disponibili ovunque

Bassi costi per l'archiviazione dei dati

Hardware più potente e più economico

DISPOSITIVI

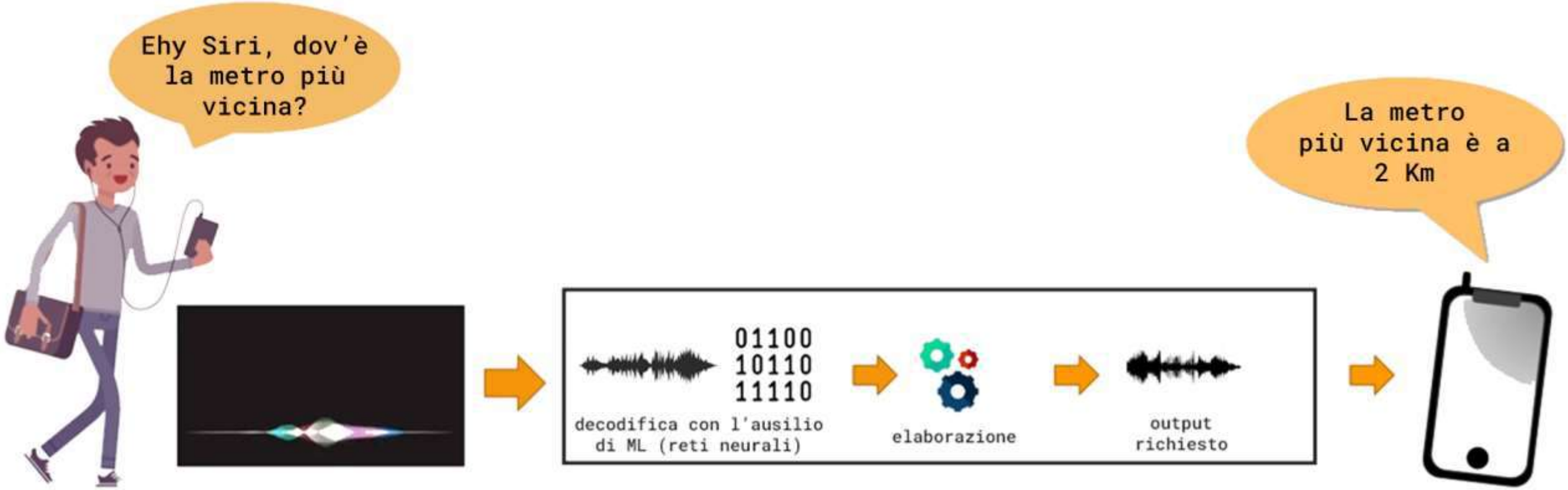
Chiunque ha dispositivi elettronici con connettività internet e sensoristica che raccoglie dati

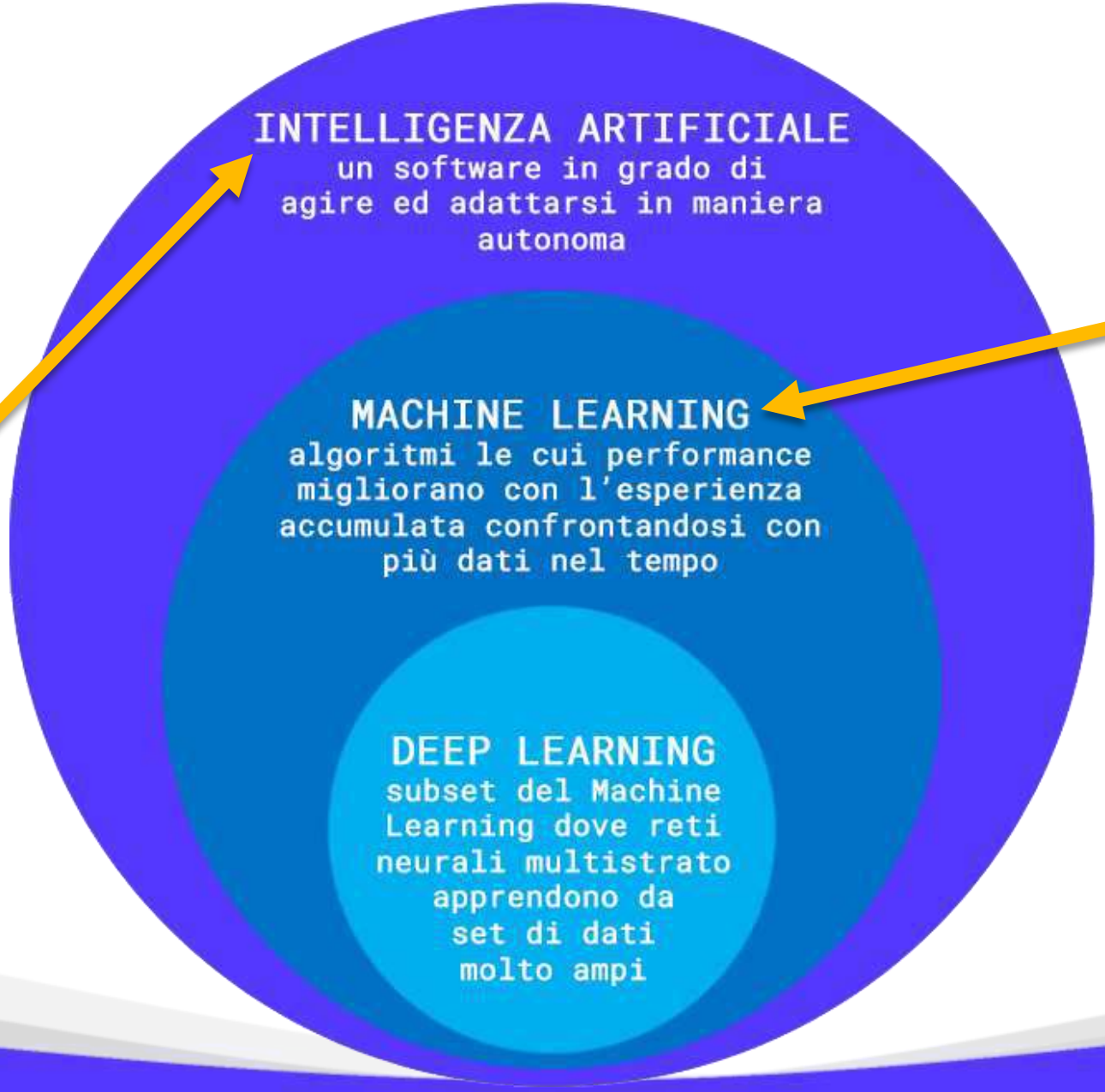
- GPS
- Fotocamera
- Microfono
- IoT

SERVIZI

Cloud computing
- archiviazione online
- infrastrutture disponibili come servizi

Applicazioni disponibili
- YouTube
- Gmail
- Facebook
- Twitter
- ...


Ehy Siri, dov'è la metro più vicina?
01100
10110
11110
decodifica con l'ausilio di ML (reti neurali)
elaborazione
output richiesto
La metro più vicina è a 2 Km



ad esempio alberi decisionali, regressioni, SVM, reti neurali...

ad esempio AIML, sistemi esperti, sistemi basati sulla conoscenza, ...

Definizione

Il Machine Learning è un campo di studio che offre a un computer la capacità di apprendere qualcosa senza esserne esplicitamente programmato.

Arthur Samuel, esperto statunitense di intelligenza artificiale e videogames, coniò il termine «Machine Learning» e la relativa definizione nel 1959.

«...il meccanismo principale della macchina si basava sull'analisi probabilistica delle posizioni raggiungibili dalla posizione attuale. Siccome la macchina disponeva di una quantità di memoria molto limitata, Samuel decise di implementare l'algoritmo di ricerca potatura alfa-beta. Invece di cercare in una volta sola ogni possibile strada per arrivare all'altra sponda, e conseguentemente vincere il gioco, Samuel sviluppò una funzione in grado di analizzare la posizione della dama in ogni istante della partita. Questa funzione provava a calcolare le possibilità di vittoria per ogni lato nella posizione attuale, agendo di conseguenza. Prendeva in considerazione diverse variabili tra cui il numero di pezzi per lato, il numero di dame e la distanza dei pezzi 'mangiabili'.
Il programma sceglieva le sue mosse basandosi sulla strategia minimax, ovvero agendo in modo da ottimizzare il valore della sua funzione, assumendo che l'avversario agisse e ragionasse nel medesimo modo...» (da Wikipedia)

Definizione

Si dice che un software impari dall'esperienza E rispetto ad alcune classi di attività T e misura delle prestazioni P, se la sua prestazione su T misurata da P migliora con l'esperienza E.

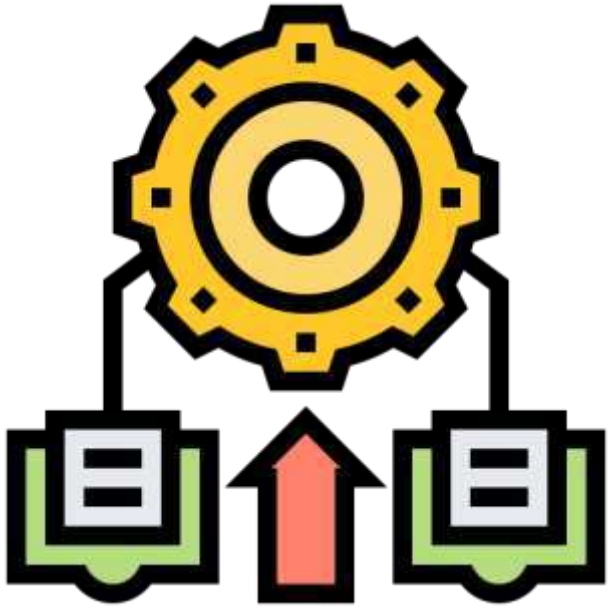
(*Tom Mitchell - Informatico e professore universitario – 1998*
Rilevante poiché per la prima volta viene data una definizione operativa dell'apprendimento automatico)

Es.
E = esperienza nel giocare a scacchi
T = compito di giocare a scacchi
P = probabilità che il programma vinca la partita successiva

Machine Learning

Usare dati **per** **rispondere a domande**

addestramento **predizione / classificazione**



Programmazione tradizionale

Machine Learning

Campi di applicazione

Il Machine Learning ha oggi un ruolo cruciale in una serie di applicazioni critiche, come:

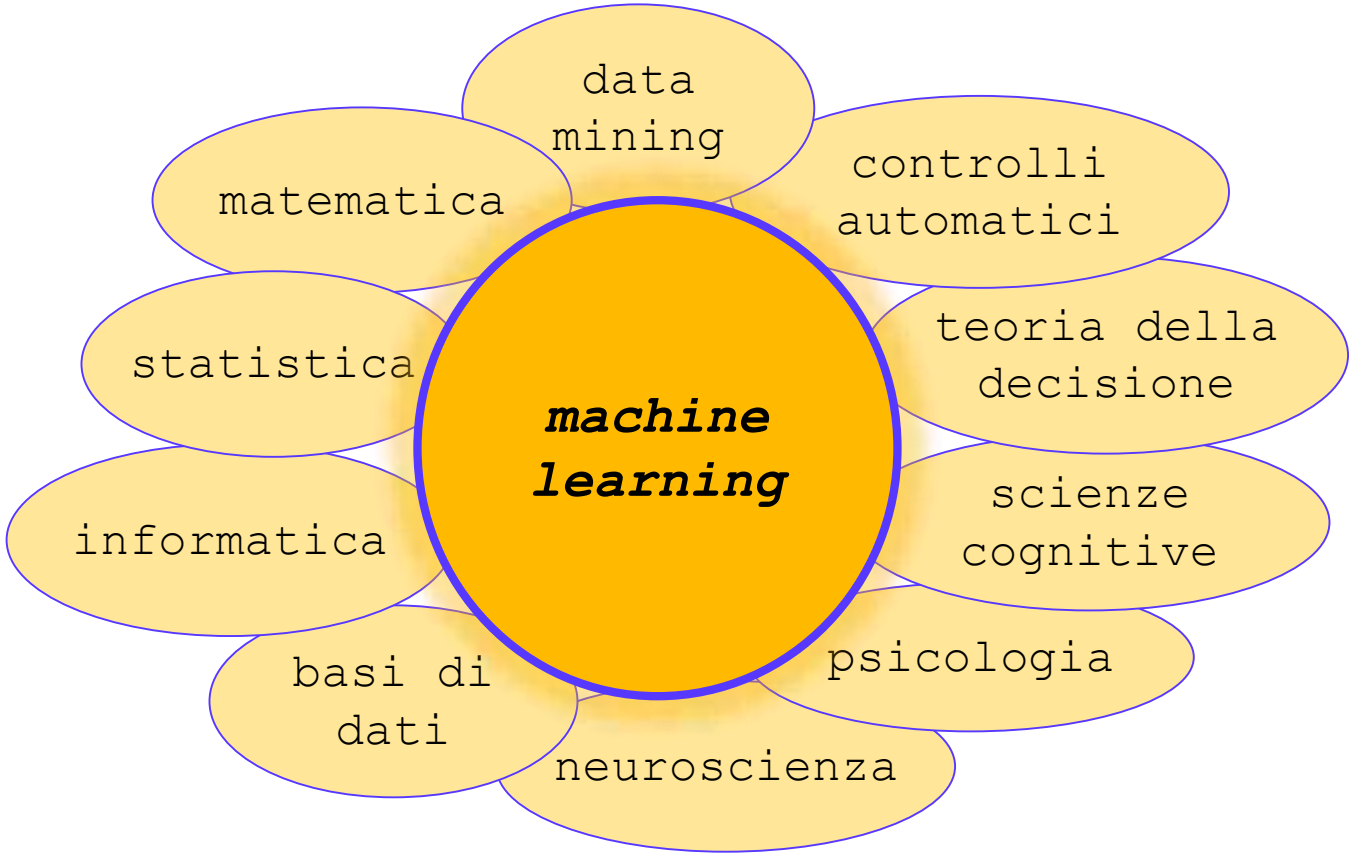
- **Data Analysis e Data Mining**
- **Natural Language Processing**
- **Computer Vision**
- **Sistemi esperti**

Campi di applicazione

Il Machine Learning è utilizzato quando:

- **Non esiste la relativa esperienza umana** (ad es. esplorazione su Marte, controllo industriale…)
- **Non è possibile spiegare l'esperienza acquisita** (ad es. riconoscimento del parlato, guidare un aeroplano, …)
- **E' necessaria una personalizzazione di un modello standard** (ad es. medicina, sistemi di raccomandazione, …)
- **I modelli sono basati su un'enorme quantità di dati** (ad es. genoma umano, astronomia, …)
- **Le soluzioni cambiano nel tempo** (ad es. routing di pacchetti di rete, modelli finanziari, …)

Discipline legate al Machine Learning



Classificazione algoritmi di Machine Learning

In genere, qualsiasi problema di apprendimento automatico può essere ricondotto a una delle seguenti classi di algoritmi:

Apprendimento SUPERVISIONATO

NON SUPERVISIONATO

PER RINFORZO

Apprendimento Misto

Apprendimento Supervisionato

- **Viene fornito un set di dati e si sa come dovrebbe essere il nostro output corretto, supponendo che ci sia una relazione tra input e output.**

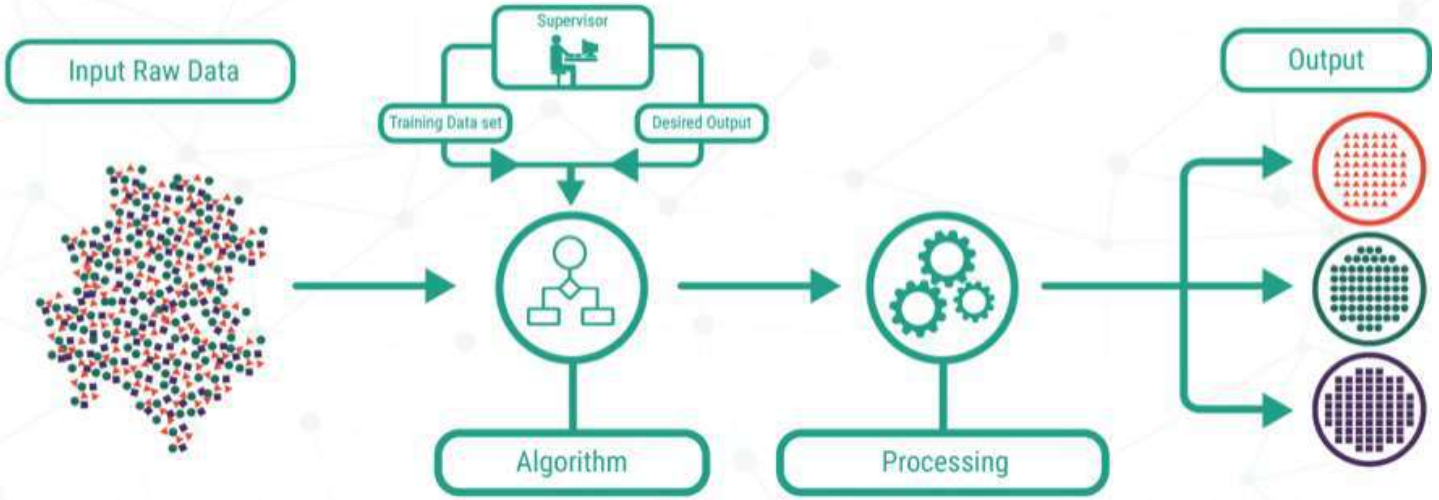
- **I problemi di apprendimento supervisionato sono classificati in:**

  Regressione: output continuo (si cerca di mappare le variabili di input su alcune funzioni continue)

  Classificazione: output discreto (si cerca di mappare le variabili di input in categorie discrete)

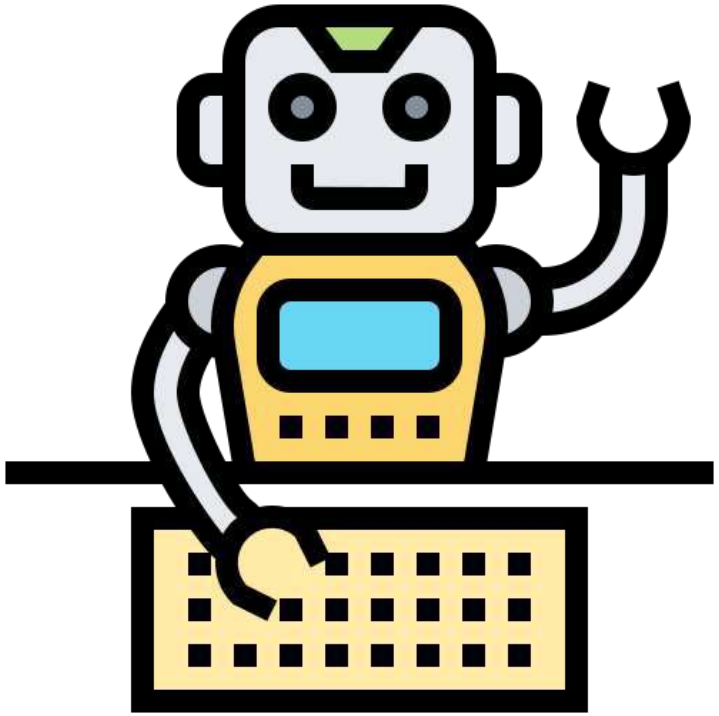
Esempi di apprendimento Supervisionato

- **Da dati sulla dimensione delle case sul mercato immobiliare, si prova a prevederne il prezzo (regressione) o la fascia di prezzo (classificazione)**
- **Ricavare l'età di una persona basandosi su una sua fotografia (regressione)**
- **Stabilire se un tumore è benigno o maligno (classificazione)**



Apprendimento non Supervisionato

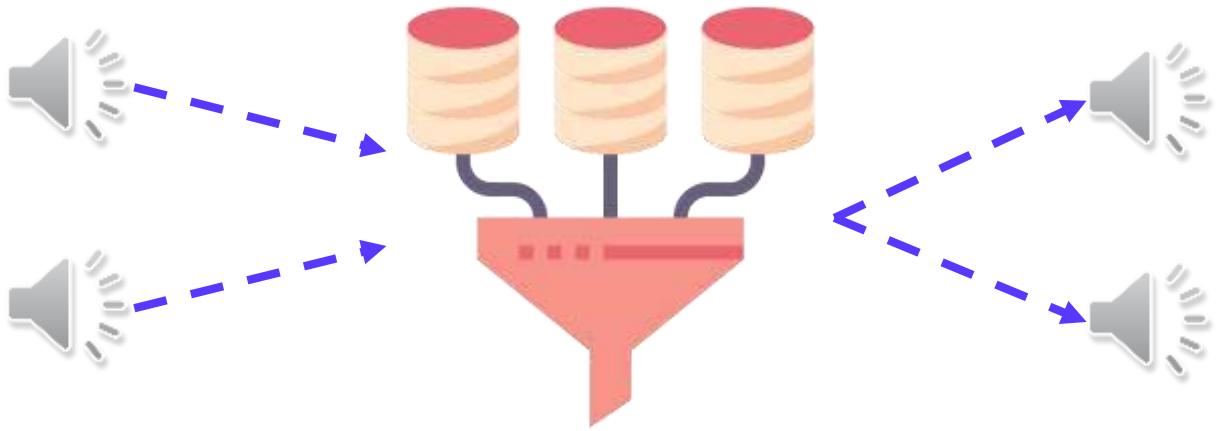
- L'apprendimento non supervisionato si applica in contesti con poche o nessuna idea relativamente ai risultati
- E' utile per derivare la struttura di un modello da dati di cui non conosciamo l'effetto delle singole variabili che li compongono
- Possiamo ricavare la struttura del modello raggruppando i dati in base alle relazioni tra le variabili nei dati
- Con l'apprendimento senza supervisione non esiste alcun feedback basato sui risultati della previsione



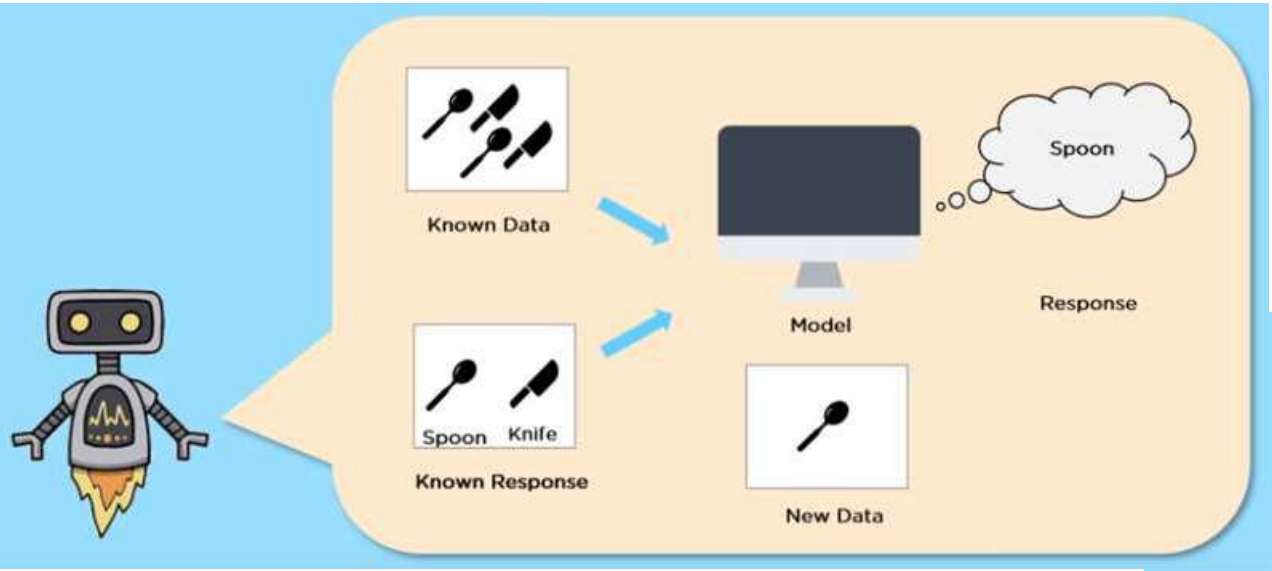
Esempi di apprendimento non Supervisionato

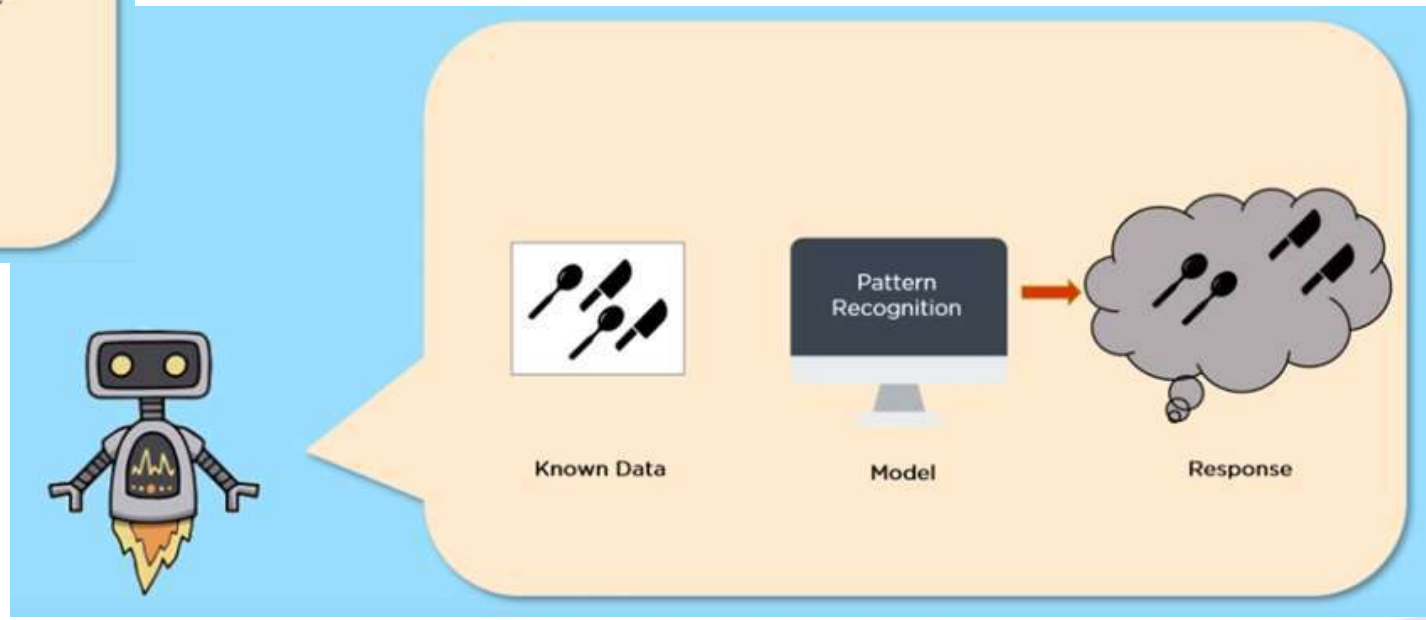
- **Da una raccolta di 1.000.000 di geni diversi si trova un modo per raggruppare automaticamente questi geni in gruppi che sono in qualche modo simili o correlati da variabili diverse, come posizione, ruoli e così via.**

- **Identificare singole voci e musica da un insieme di suoni.**
  (*https://cnl.salk.edu/~tewon/Blind/blind_audio.html*)



Supervisionato vs Non Supervisionato



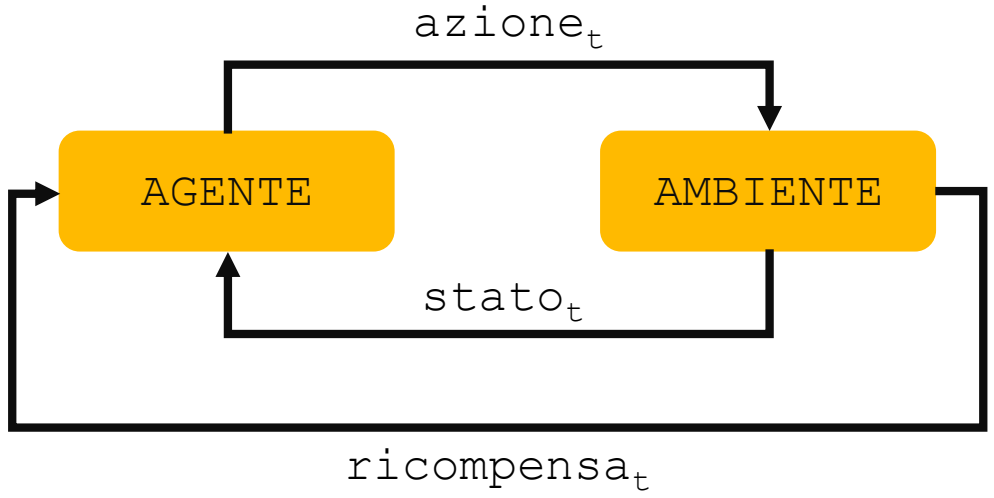


Apprendimento per Rinforzo

Apprendimento tramite interazione con l'ambiente, sfruttando le «conseguenze» delle proprie azioni

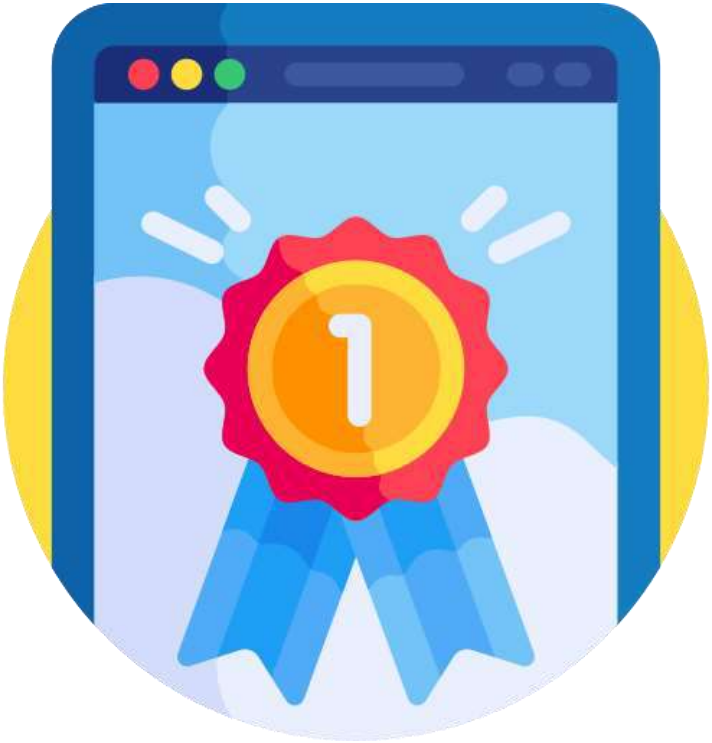
Step:

1. **Osservazione dello stato in cui l'ambiente si trova**
2. **Decisione (azione)**
3. **Passaggio in un nuovo stato**
4. **Ricompensa**



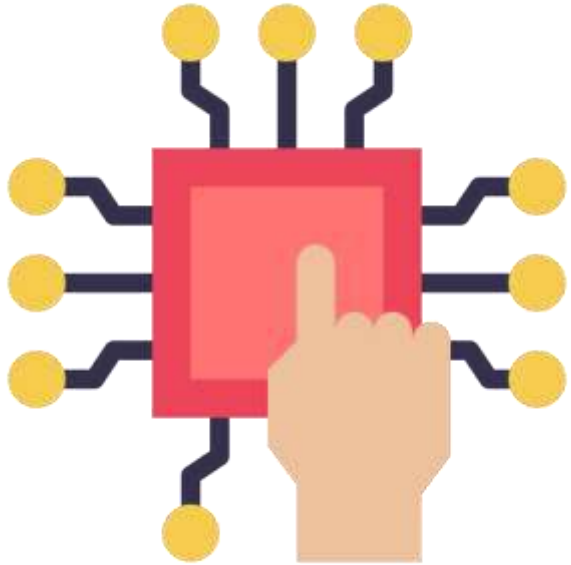
Apprendimento per Rinforzo – scenari applicativi

- Guida autonoma
- Videogames
- Robot industriali
- Trading e finanza
- Recommander System
- Allocazione risorse in cluster di computer
- Ottimizzazione trattamenti clinici
- Dynamic pricing
- Fraud detection
- …



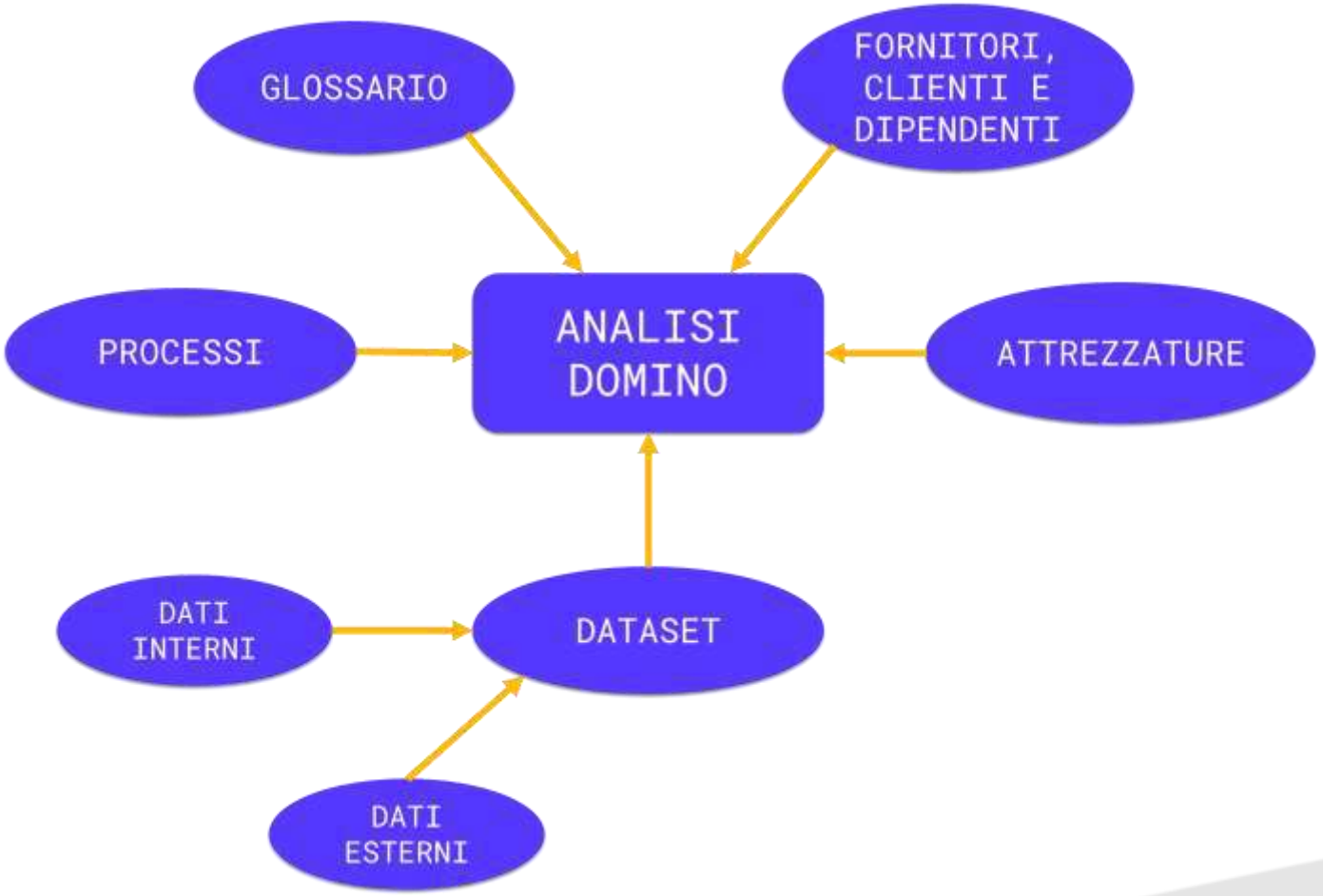
Processo di implementazione

Raccolta Dati	Preprocessing	Feature Eng.	Selez. Modello	Predizioni
Scouting su diverse fonti	Pulizia e omogeneità nei dati selezionati	Incremento dell'utilità del set di dati creato	Identificazione e addestramento	Validazione del modello



Studio del Dominio Applicativo

- **Studio dei concetti, delle dinamiche e delle regole generali che governano le attività aziendali**
- **Analisi dei dataset disponibili**
- **Ricerca fonti dati esterne**
- **Reingegnerizzazione processi**



Raccolta dati

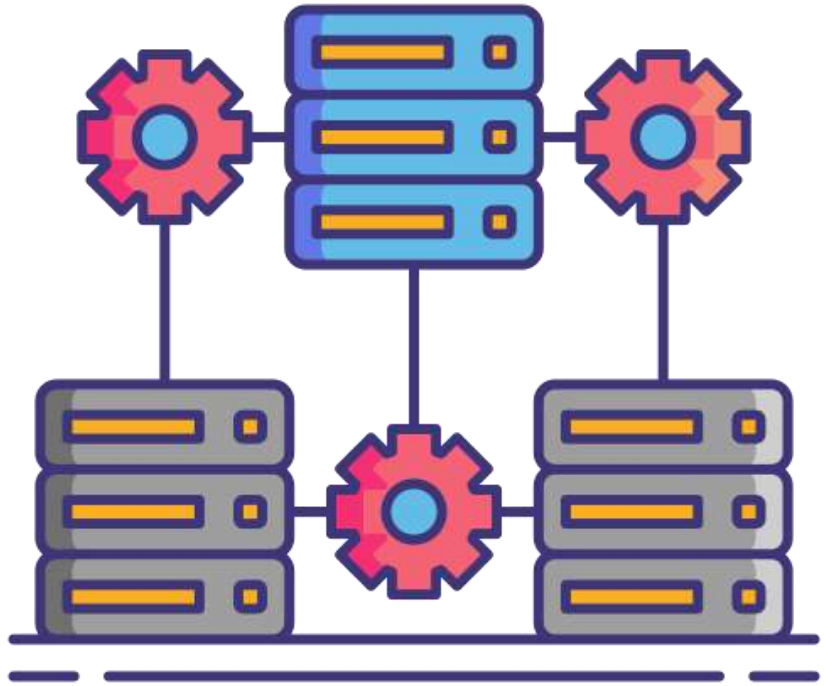
Dipendente dal contributo umano

- **etichettatura manuale**
- **esperti di dominio**

Spesso sono disponibili fonti dati esterne, come dataset pubblici, per arricchire la base di conoscenza del problema in analisi

Alcuni algoritmi necessitano di grandi quantità di dati per un corretto addestramento (es. reti neurali)

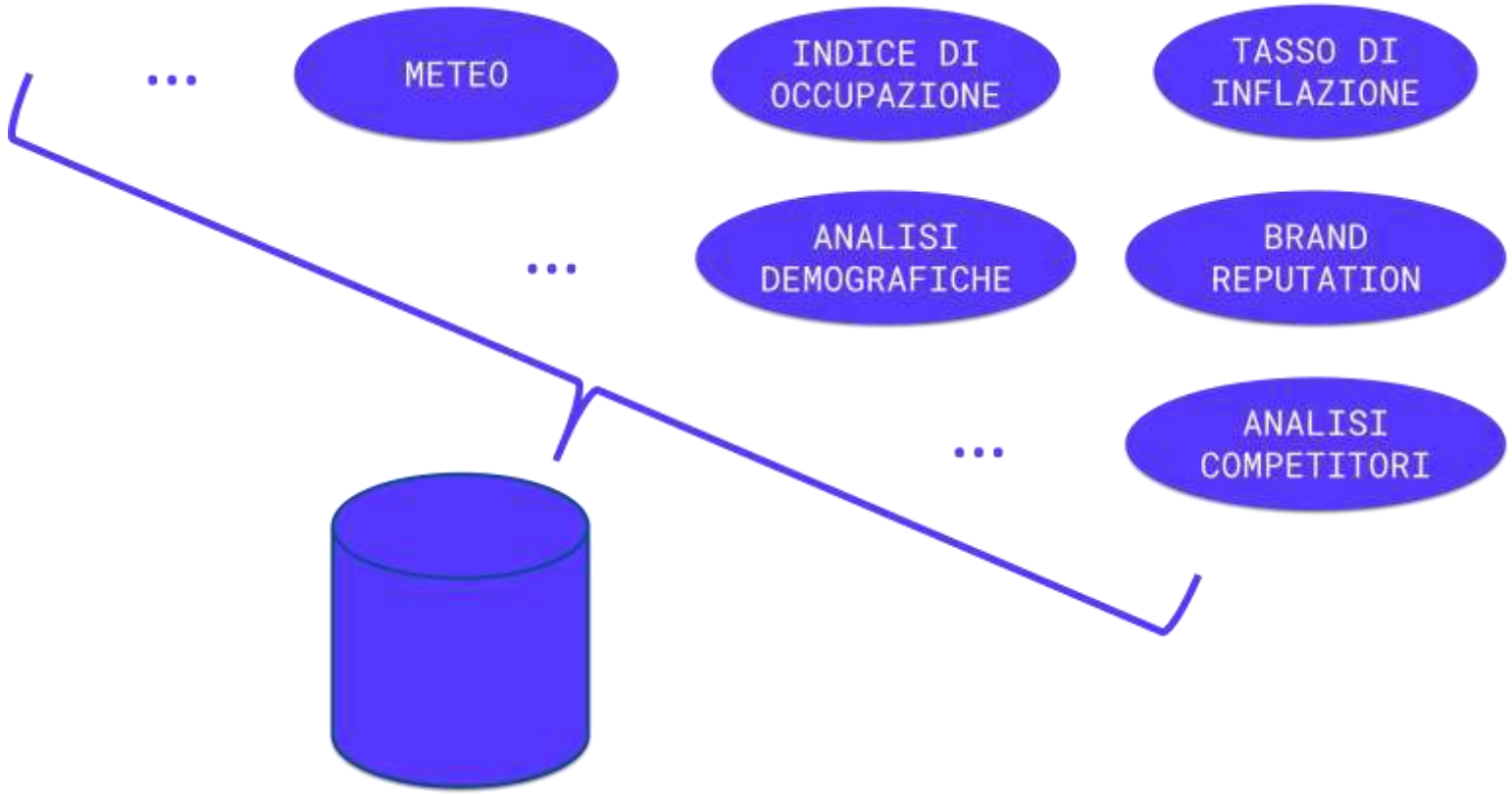
La quantità e la qualità dei dati influenzano l'accuratezza finale



Analisi fonti dati esterne

Oltre ad esaminare i propri dati, si ricercano fonti di dati esterne che possano incrementare la quantità e la qualità delle informazioni.

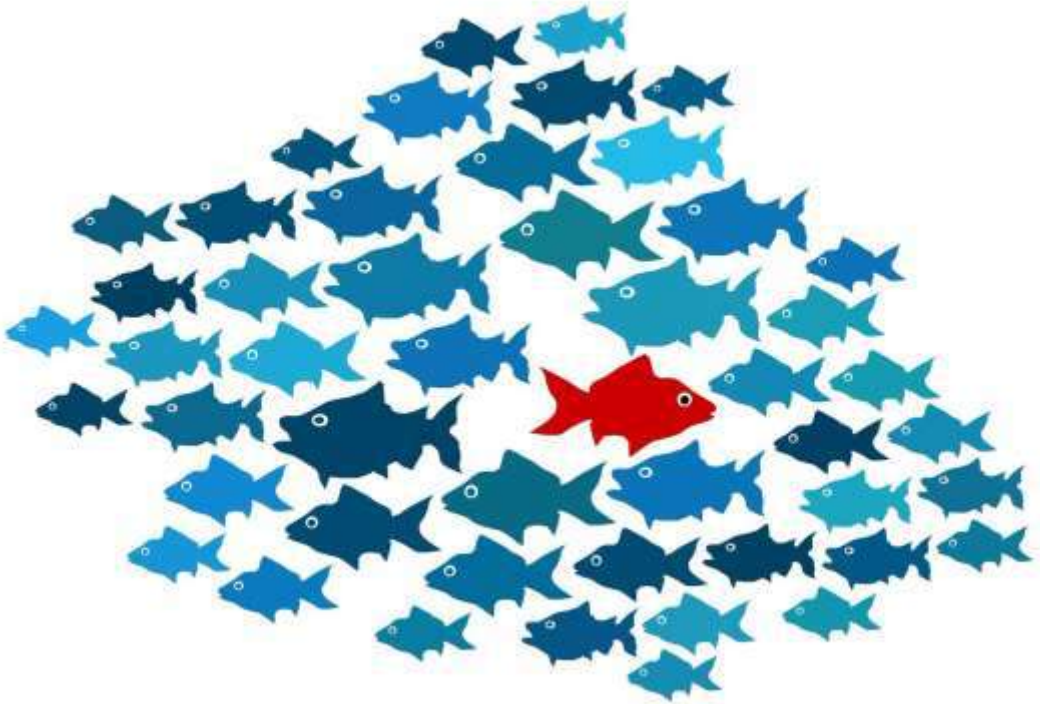
L'utilizzo di fonti di informazione esterne permette di avere una più completa visione del mercato e assicura performance superiori ai moduli del sistema.



Preprocessing dati

Risoluzione delle incongruità nei dati selezionati

- valori mancanti
- outlier
- valori errati
- etichette errate
- dati sbilanciati
- dati ridondanti



Esempio di pre-processing dei dati

#	ID	Nominativo	Compleanno	Sesso	Fornitore	Cliente	Nazione	Provincia
145	111-1234	Mario Rossi	13/05/1984	M	1	0	Italy	Milano
146	111-1236	Luca Verdi	22/01/1987	M	1	0	Italy	Roma
147	113-0142	Massimo Neri	07/03/1979	M	0	1	BA	Bari
148	113-0149	Roberta Gialli	1-1-1975	F	0	1	Italy	Genova
149	115-1245	Mike Reds	03/06/1992	M	1	0	England	London
150	113-0150	Daniele Bianchi	24/08/1991	M	0	1	Italy	Torino
151	113-0150	Monica Rossi	02/11/1989	A	1	0	Italy	Romaa
152	113-0151	Antonio Rossi	21/10/1985	M	0	1	Italy	Venezia
…	…	…	…	…	…	…	…	…

unicità

formato non corretto

valore non valido

informazioni ridondanti

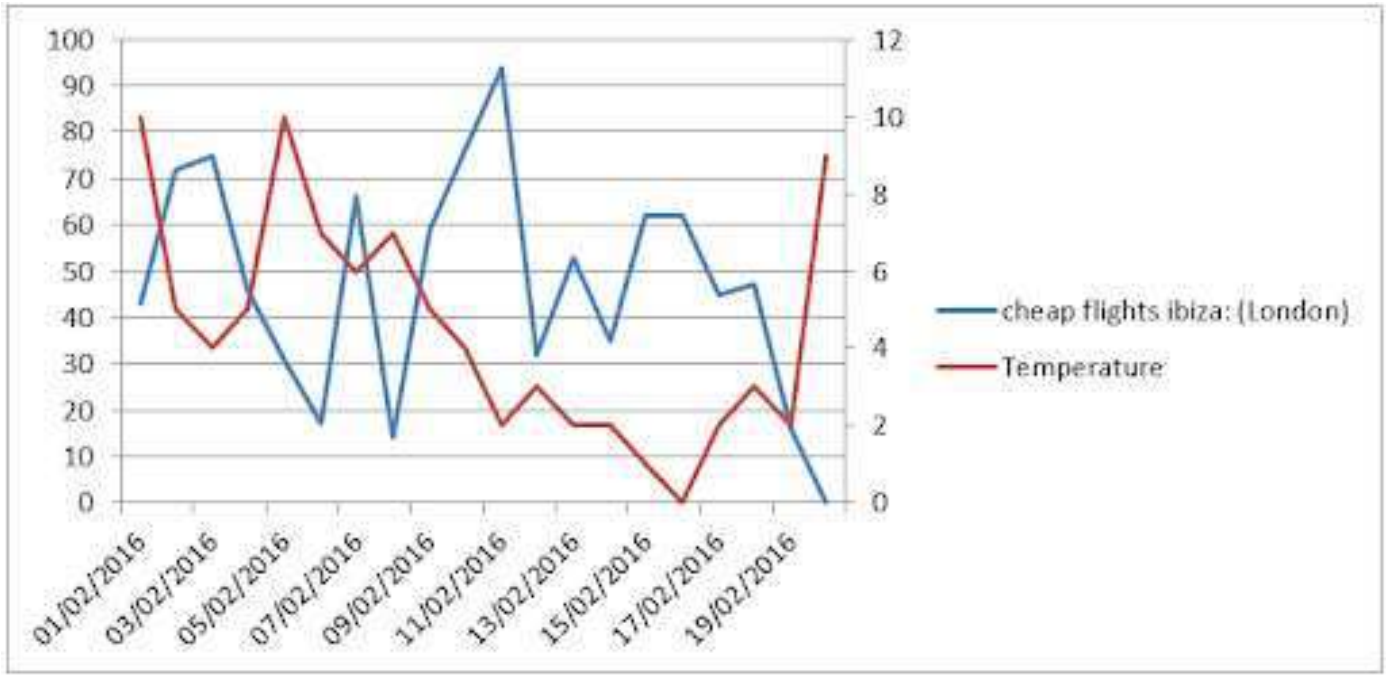
valori fuorvianti

errori sintattici

Preprocessing dati

Analisi delle correlazioni e dipendenze

- Evidenzia comportamenti affini tra diversi set di dati
- Permette di **quantificare** quanto la variazione di alcuni dati dipenda dal comportamento di altri



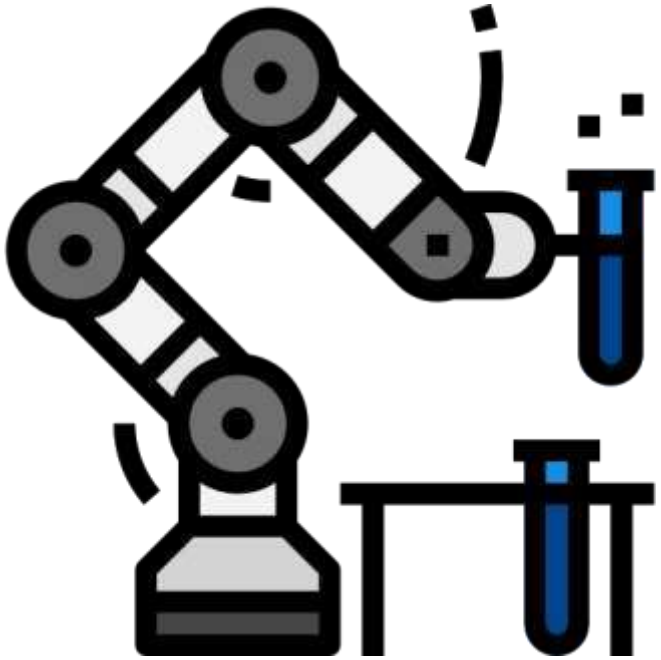
Le relazioni evidenziate vengono sfruttate per:

- fornire ulteriori conoscenza al sistema
- verificare l'efficacia dei dati selezionati, rimuovendo eventuali ridondanze

Feature Engineering

Tecniche per estrarre un numero maggiore di informazioni dallo stesso dataset:

- **rende il set di dati selezionato più utile**
- **con un buon set di feature gli algoritmi apprendono più velocemente**
- **richiede un'ottima conoscenza del dominio applicativo**



Step:

- **trasformazione delle feature disponibili (normalizzazione, trasformazione di date in giorno della settimana, ...)**
- **creazione di nuove feature**

Feature Engineering

Una feature è una proprietà misurabile del fenomeno osservato

I dataset di input sono insiemi di feature (*caratteristiche*)

Ad esempio possiamo classificare dei messaggi di posta elettronica in base a:

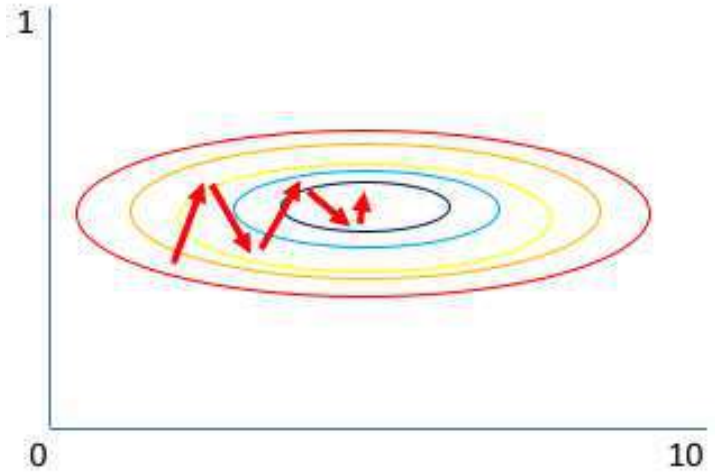
- **Numero di parole *c4mb14t3* come in questo esempio**
- **Lingua utilizzata (0= italiano, 1=inglese)**
- **Numero di emoji/emoticons presenti**

[1, 0, 3]

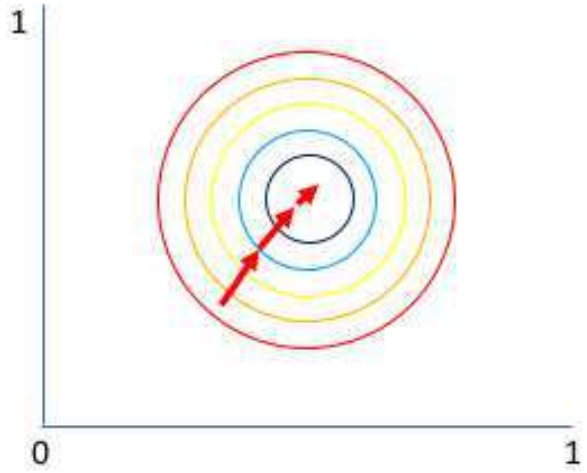
Feature Engineering - Normalizzazione

Per velocizzare i calcoli (specialmente su dataset ampi)
è utile normalizzare i dati di input.

Lo scopo è avere tutte le variabili di input che variano in uno stesso intervallo di valori.



alcuni parametri forzano più l'apprendimento rispetto ad altri



tutti i parametri sono aggiornati similmente

Data Augmentation

Processo di creazione di nuovi dati sintetici sulla base dei dati in possesso.

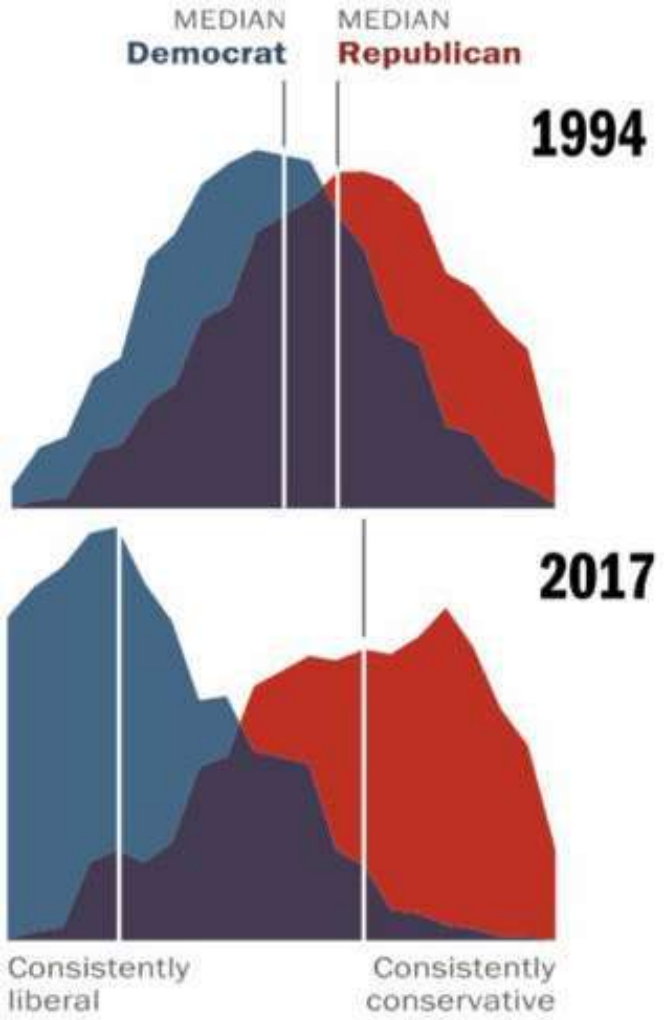
Incrementare la quantità di dati disponibili aiuta gli algoritmi di machine learning nel raggiungere performance elevate.

Tramite la creazione di nuovi dati sintetici e la copia dei propri dati leggermente modificati, si arriva a poter implementare con successo algoritmi di intelligenza artificiale anche in presenza di piccoli set di dati .

Analisi delle distribuzioni dei dati

Analizzare le evoluzioni delle distribuzioni dei dati su serie temporali storiche facilita la comprensione del dominio applicativo:

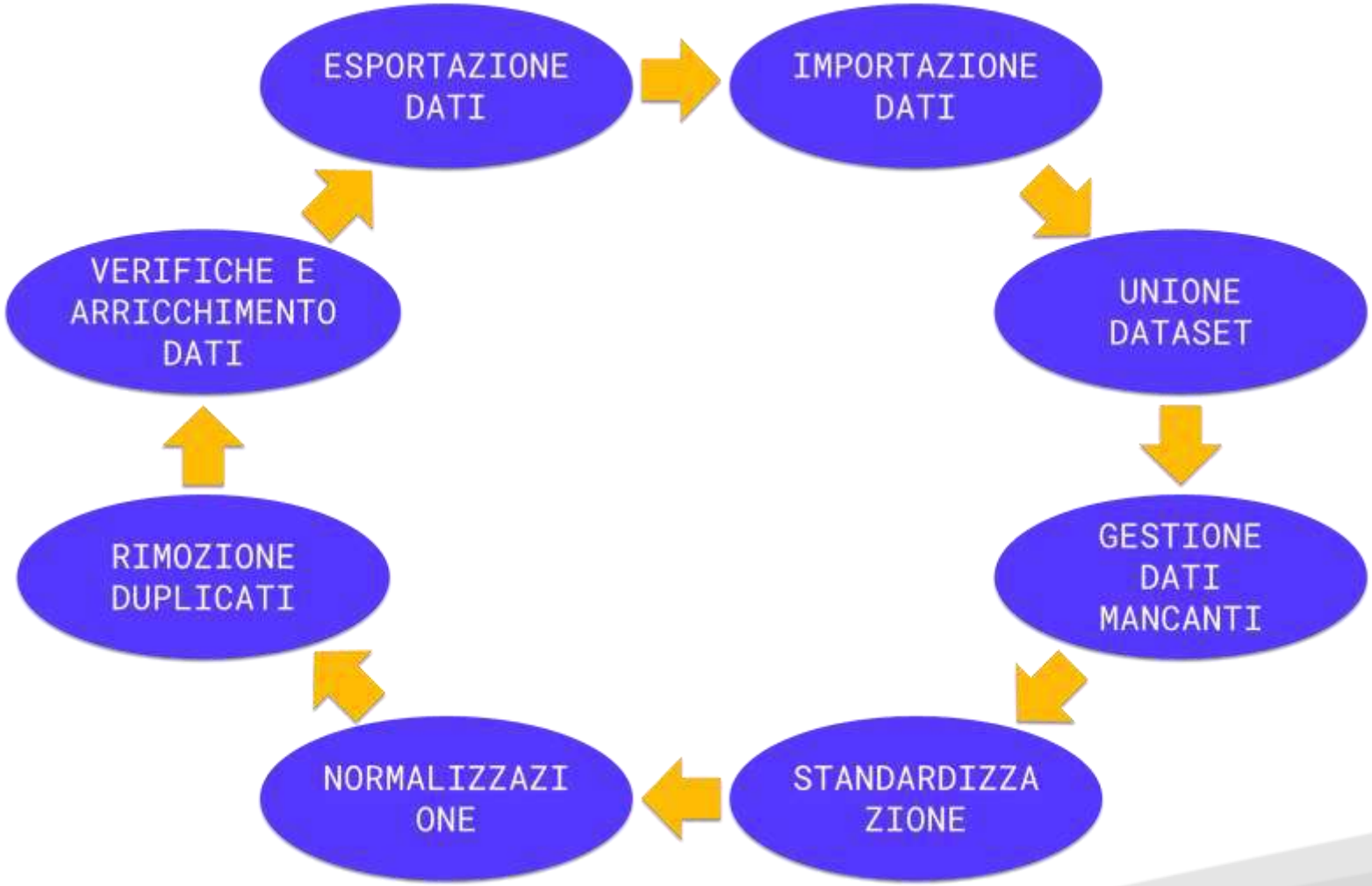
- **fornisce utili informazioni all'analisi ed allo studio dell'evoluzione dal passato al presente**
- **propone scenari futuri plausibili**



Recap: preprocessing dati

Implementazione di una serie di procedure semi-automatiche in grado di eseguire i processi di pulizia dei dati

Garantisce ai moduli del sistema il costante utilizzo di dataset consoni



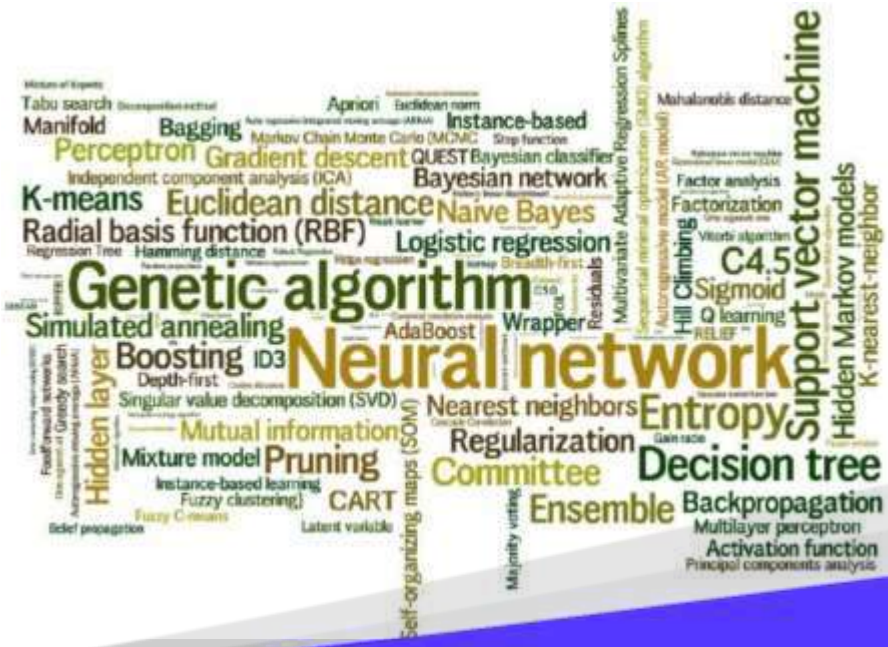
Identificazione modello di Machine Learning

Apprendimento Supervisionato

Classificatori lineari
Naive Bayes
Support Vector Machines (SVM)
Decision Tree
Random Forest
K-Nearest Neighbors
Reti Neurali (Deep Learning)

Apprendimento non Supervisionato

Principal Component Analysis (PCA)
t-SNE
k-means
DBScan

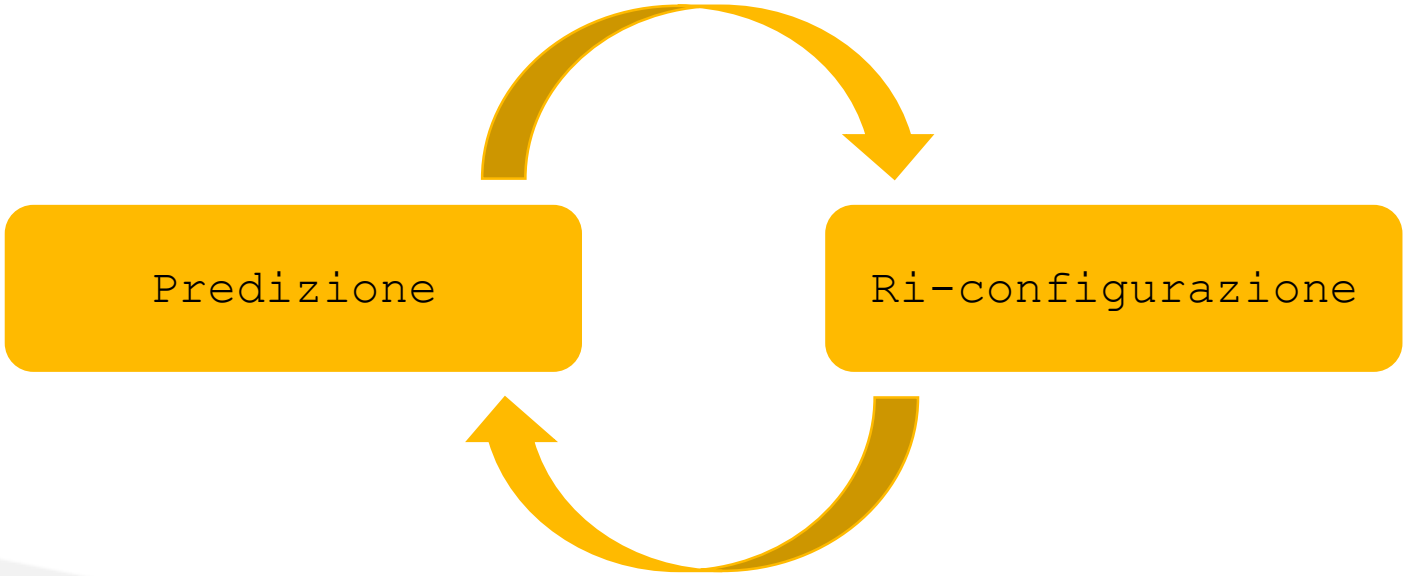


Addestramento

Scopo: rendere l'algoritmo selezionato capace di dare la risposta giusta sempre più spesso

Utilizzo di metriche per poter valutare quantitativamente le performance di diverse configurazioni

Configurazione incrementale degli iper-parametri



Validazione modello

In genere si suddivide il set di dati a disposizione in dati di addestramento, di validazione e di test

Corretta generalizzazione del modello

Generalizzazione: l'abilità di una macchina di portare a termine in maniera accurata esempi o compiti nuovi, che non ha mai affrontato, dopo aver fatto esperienza su un insieme di dati di apprendimento.

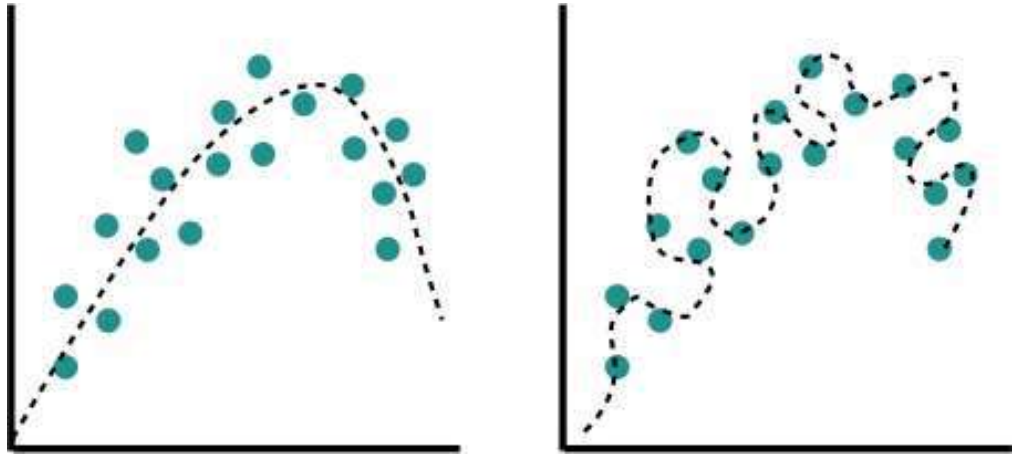
Una buona generalizzazione richiede che il modello addestrato sia in grado di riconoscere la differenza tra informazione e rumore nei dati di addestramento:

- un algoritmo troppo complesso finirà per memorizzare il rumore e non le informazioni, performando ottimamente sui dati di addestramento e mediocremente su nuovi dati

- un algoritmo troppo semplice non riuscirà a cogliere le informazioni nei dati di addestramento, performando pessimamente su nuovi dati

Overfitting

Un modello troppo complesso per il problema considerato e/o addestrato su pochi dati può risultare fallace su nuovi dati; impara informazioni e rumore dal dataset di addestramento e non riesce a generalizzare bene in nuove situazioni.



Risoluzione:

- **Semplificazione del modello**
- **Incremento degli esempi di addestramento**
- **Riduzione del numero di caratteristiche considerate in input al modello**

Underfitting

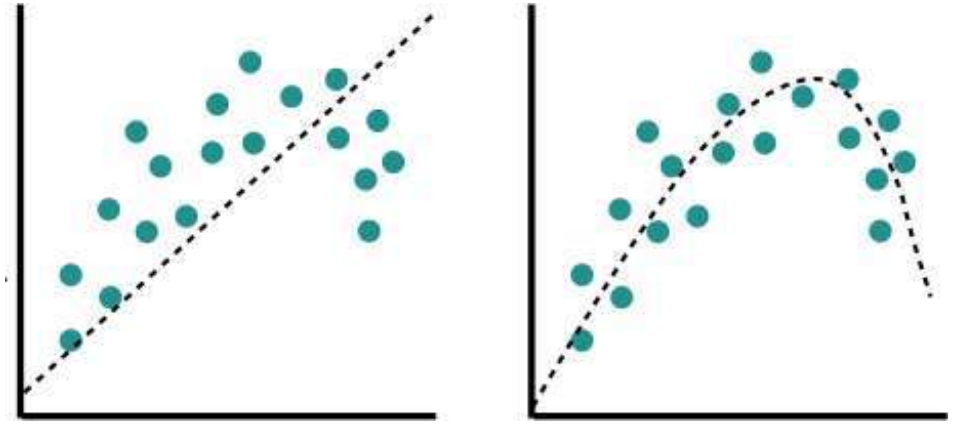
Nei casi in cui il modello proposto abbia prestazioni scadenti e non riesca ad acquisire i dati di addestramento, si parla di underfitting (sotto-adattamento).

Cause:

- **Dataset troppo piccolo**
- **Problema non lineare affrontato con algoritmi lineari**

Rimedi:

- **Feature engineering**
- **Acquisizione nuovi dati di addestramento**
- **Implementazione di un modello più complesso**



Recap

- **Definizione di Machine Learning**
- **Principali applicazioni**
- **Il processo di applicazione**