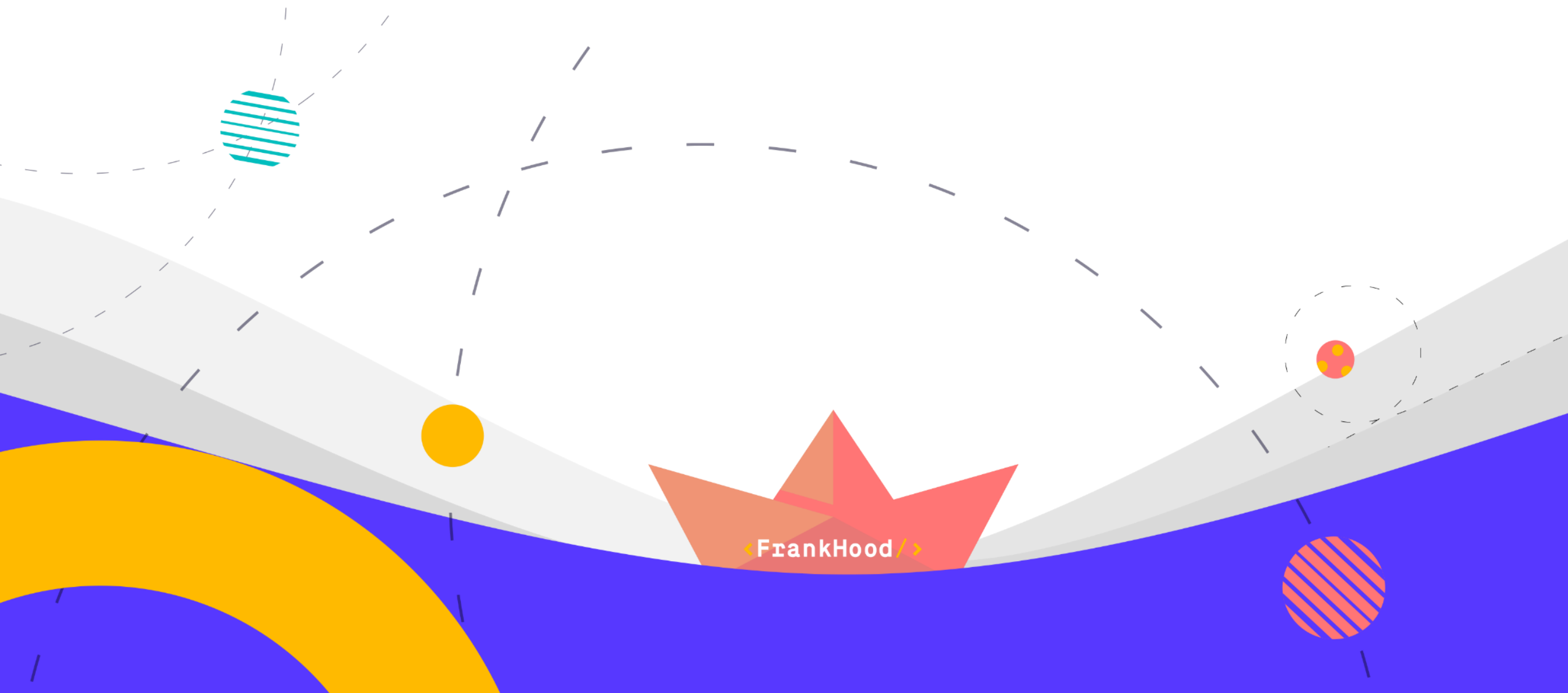




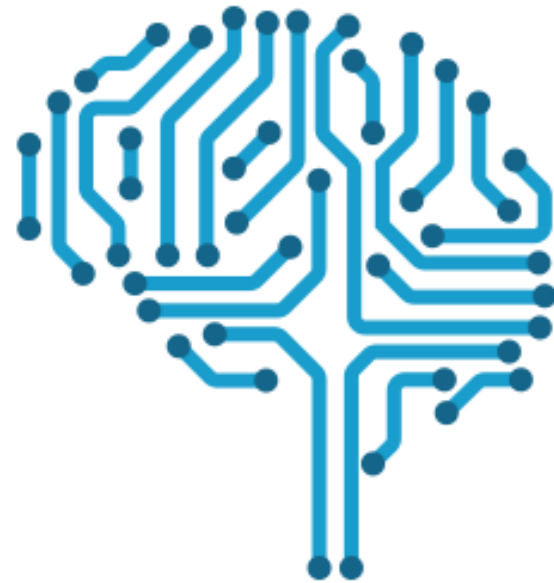
# DATA MASTERS



<FrankHood/>

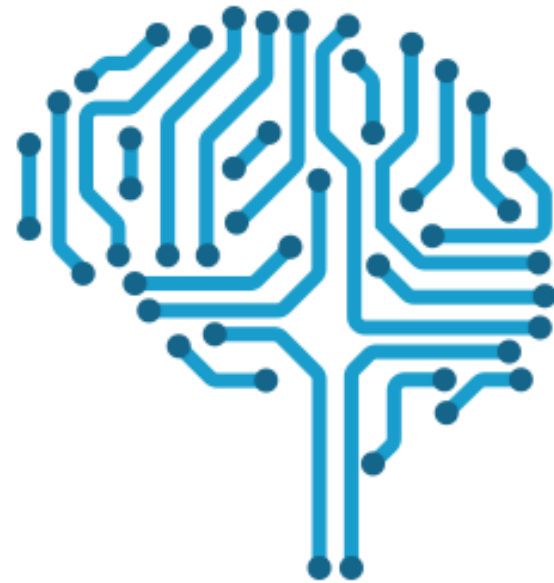
# Argomenti trattati

- Introduzione
- Data Analysis
- Data Visualization
- Funzione di costo
- Regressione Lineare

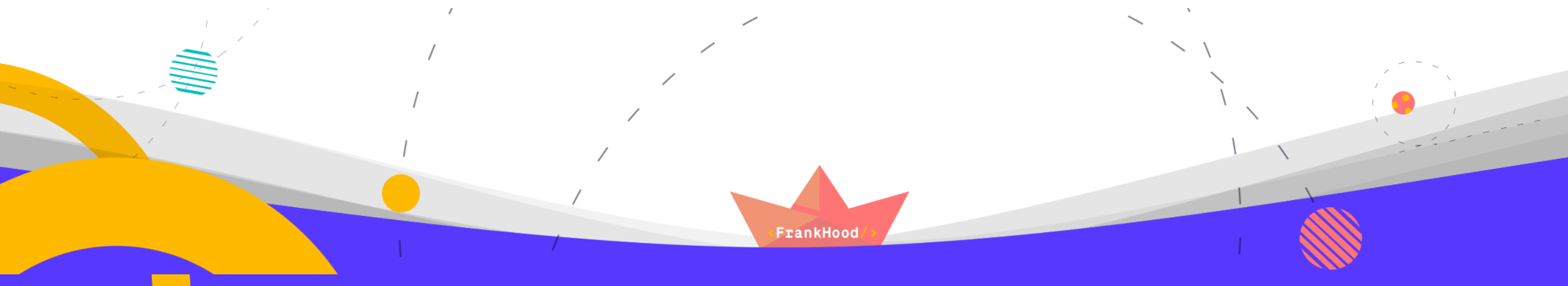


# Prefazione – finalità

- Basi di conoscenza in machine learning
- Comprensione di alcuni algoritmi
- Capacità di analisi e pre-processing dei dati
- Dimestichezza nell'implementare algoritmi di machine learning



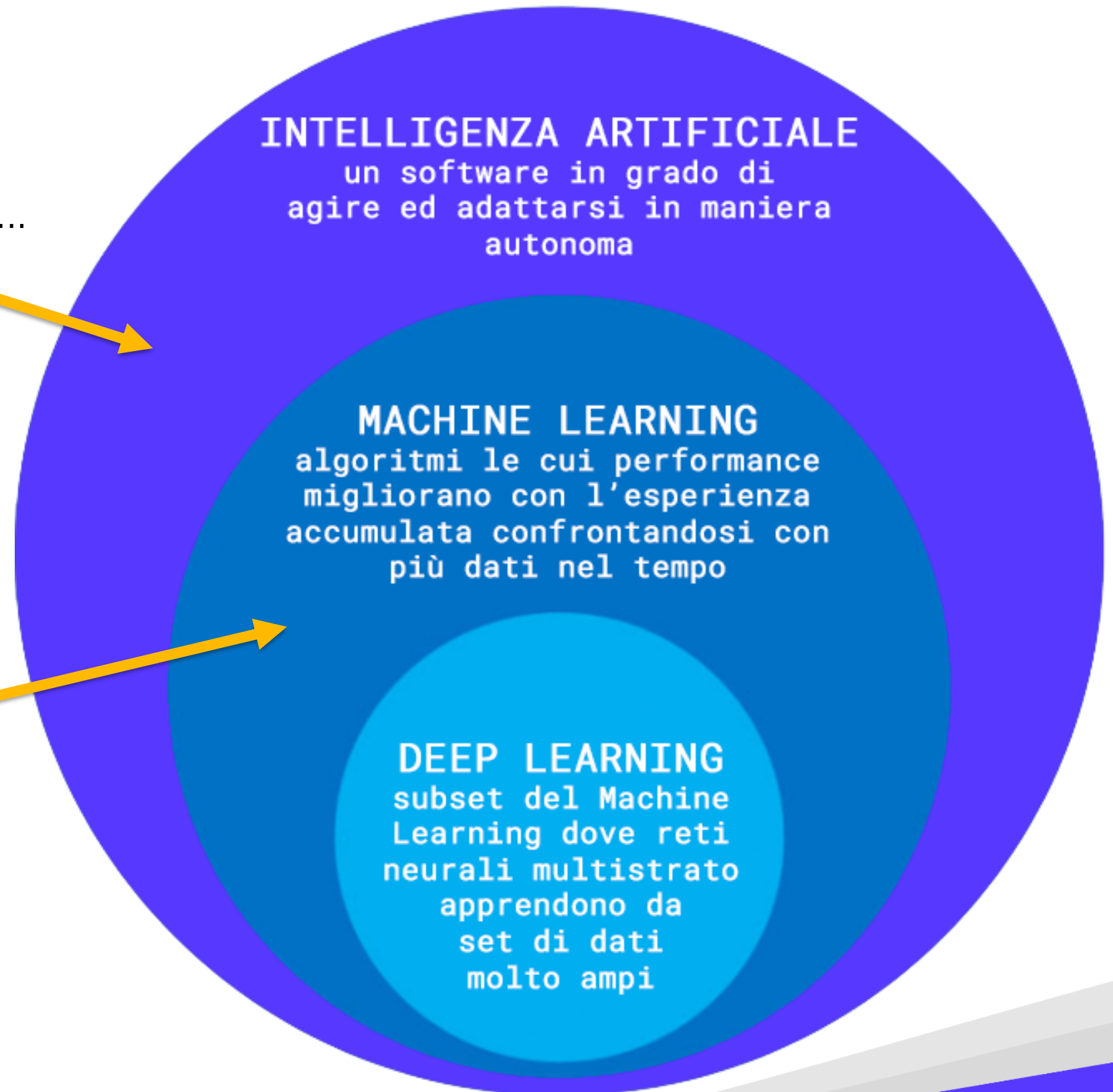
# INTRODUZIONE



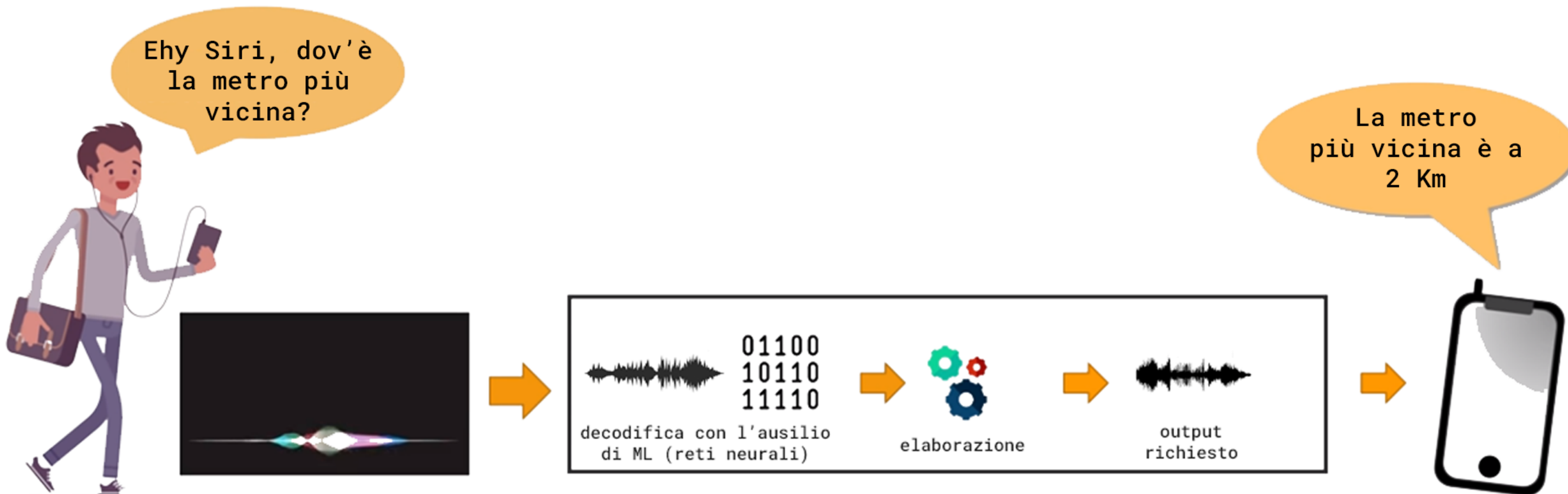
ad esempio AIML, agenti  
intelligenti, sistemi  
basati sulla conoscenza, ...



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



# Machine Learning



# 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 in compiti in  $T$  misurata da  $P$  migliora con l'esperienza  $E$ .

*(Tom Mitchell - Informatico e professore universitario – 1998)*

*Rilevante poiché per la prima volta 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



## 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

### SERVIZI

Cloud computing

- archiviazione online
- infrastrutture disponibili come servizi

Applicazioni disponibili

- YouTube
- Gmail
- Facebook
- Twitter
- ...

# 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

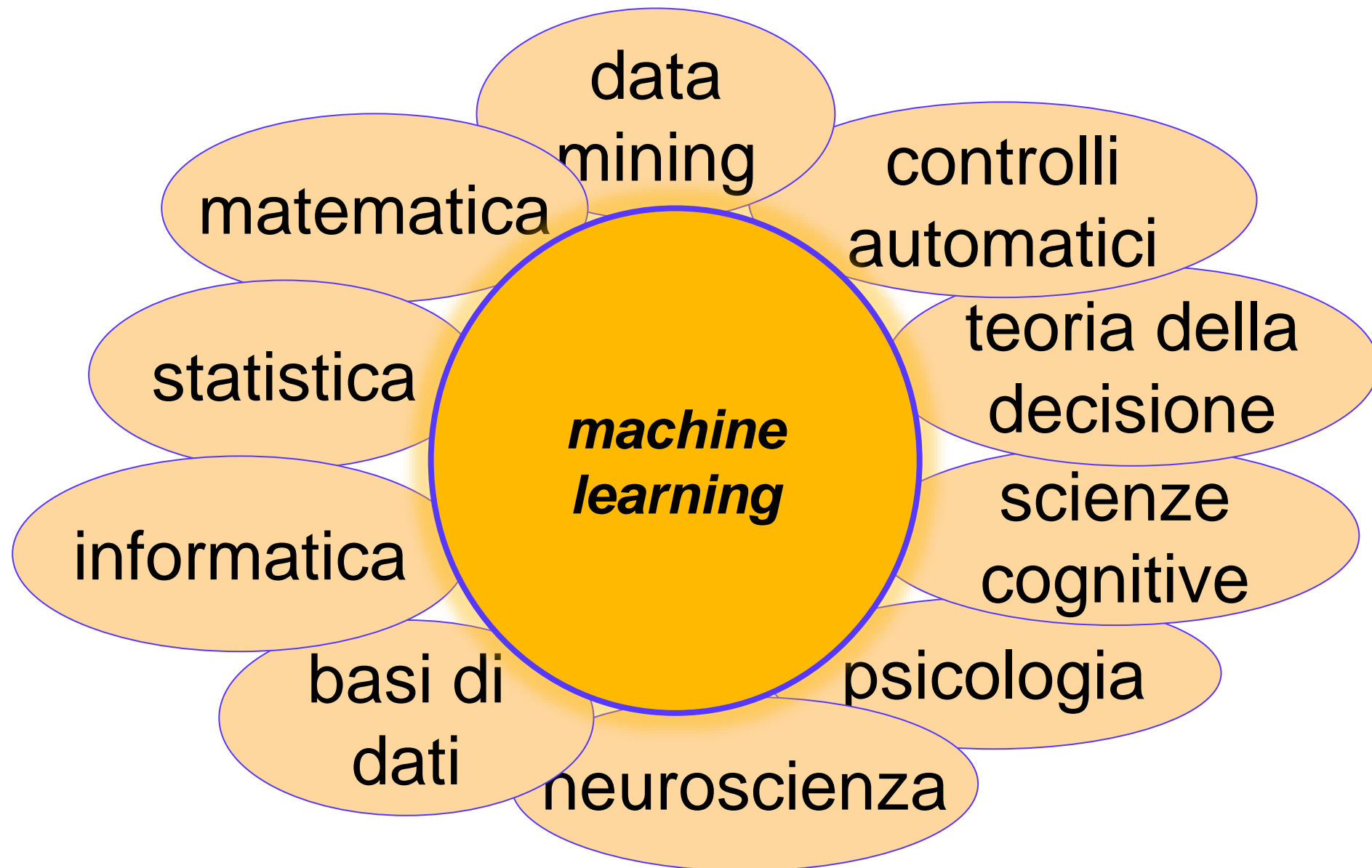


# Campi di applicazione

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 M.L.

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

### Apprendimento

#### SUPERVISIONATO



#### NON SUPERVISIONATO



#### PER RINFORZO

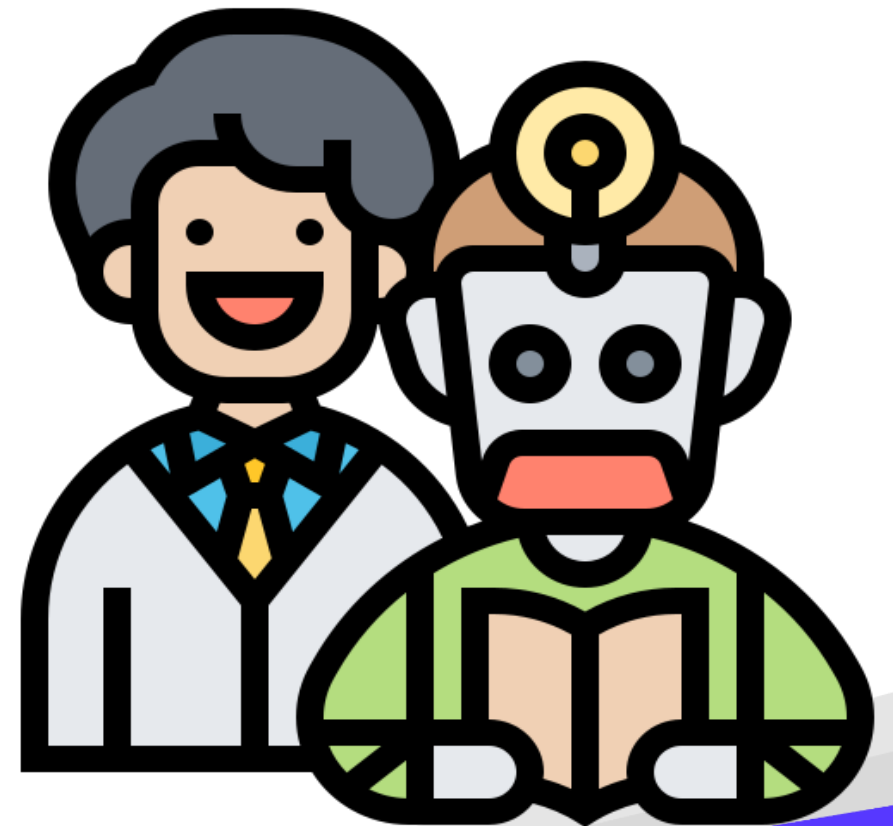


# Apprendimento Supervisionato

- 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)

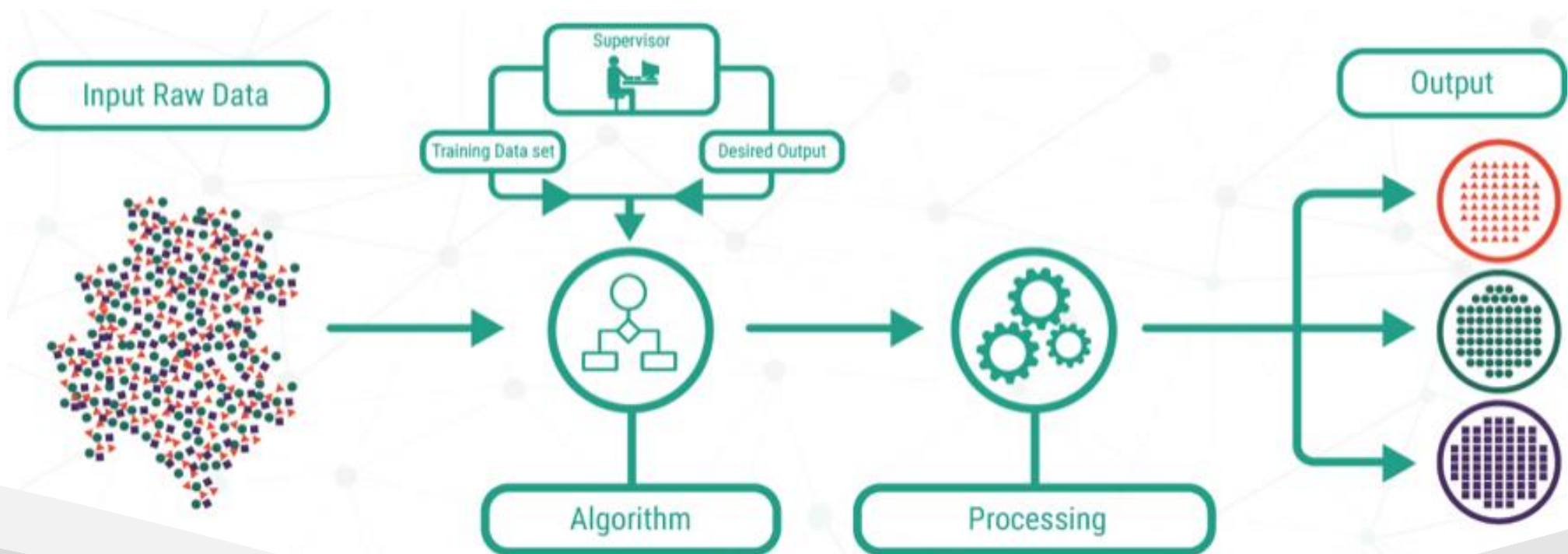
**Classificazione:** output discreto (si cerca di mappare le variabili di input in categorie discrete)





# Esempi di apprendimento Supervisionato

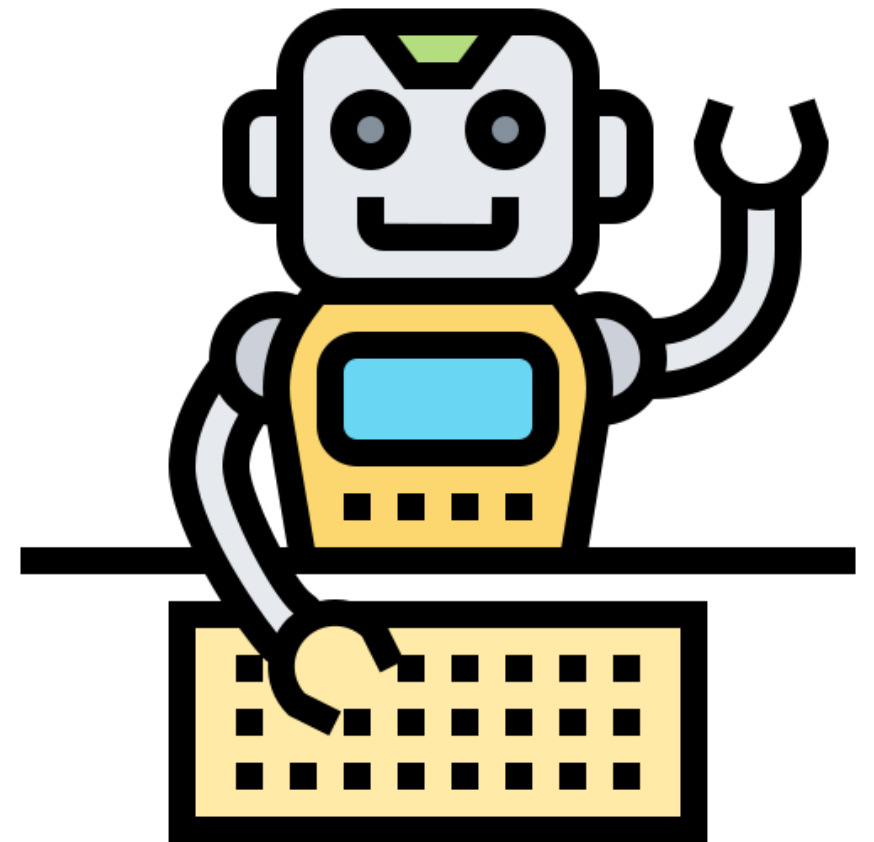
- Da dati sulla dimensione delle case sul mercato immobiliare, si prova a prevederne il prezzo (regressione) o la fascia di prezzo (classificazione)
- Prevedere l'età di una persona basandosi su una sua fotografia (regressione)
- Stabilire se un tumore è benigno o maligno (classificazione)





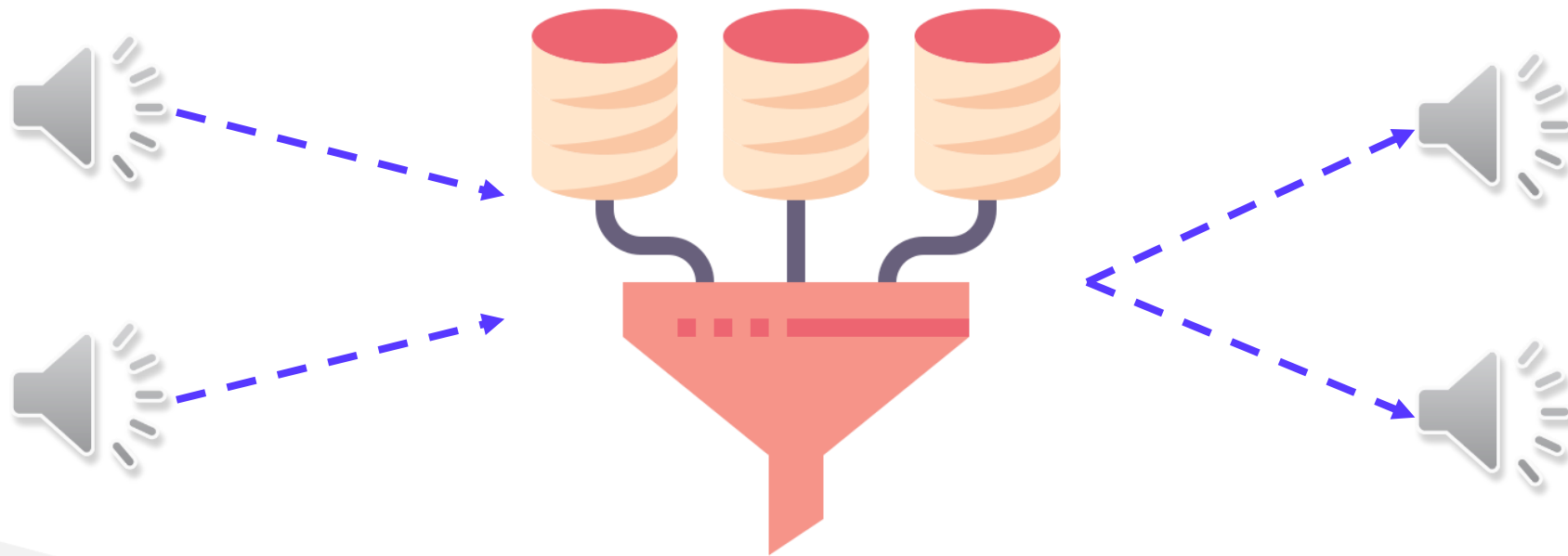
# Apprendimento non Supervisionato

- L'apprendimento non supervisionato si applica in contesti con poche o nessuna idea relativamente ai risultati
- Possiamo derivare la struttura di un modello da dati di cui non conosciamo il comportamento e gli effetti
- Possiamo ricavare la struttura del modello raggruppando i dati in base alle relazioni tra le caratteristiche dei dati stessi
- Con l'apprendimento senza supervisione non esiste alcun feedback basato sui risultati della previsione

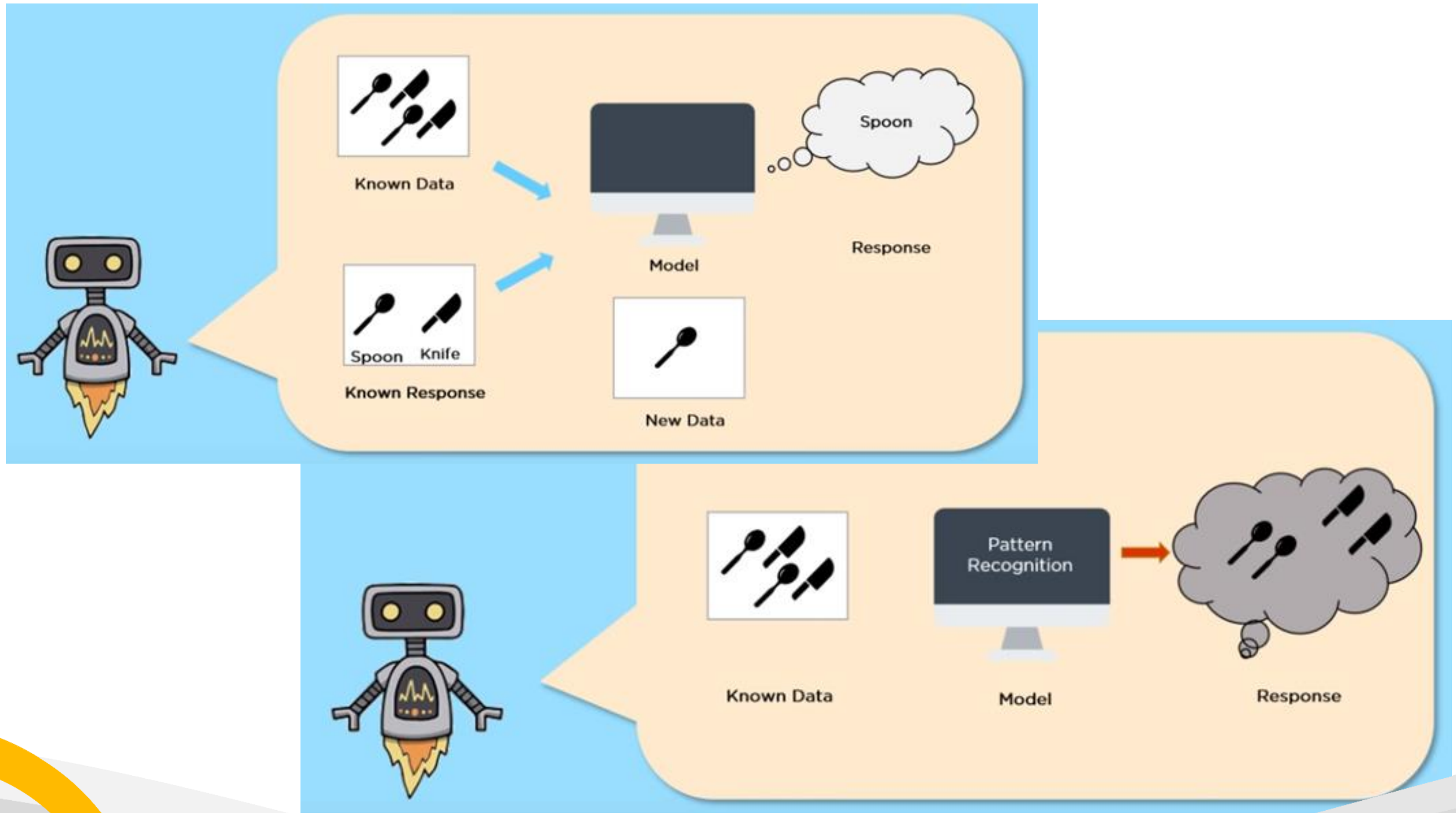


# Esempi di apprendim. non Supervisionato

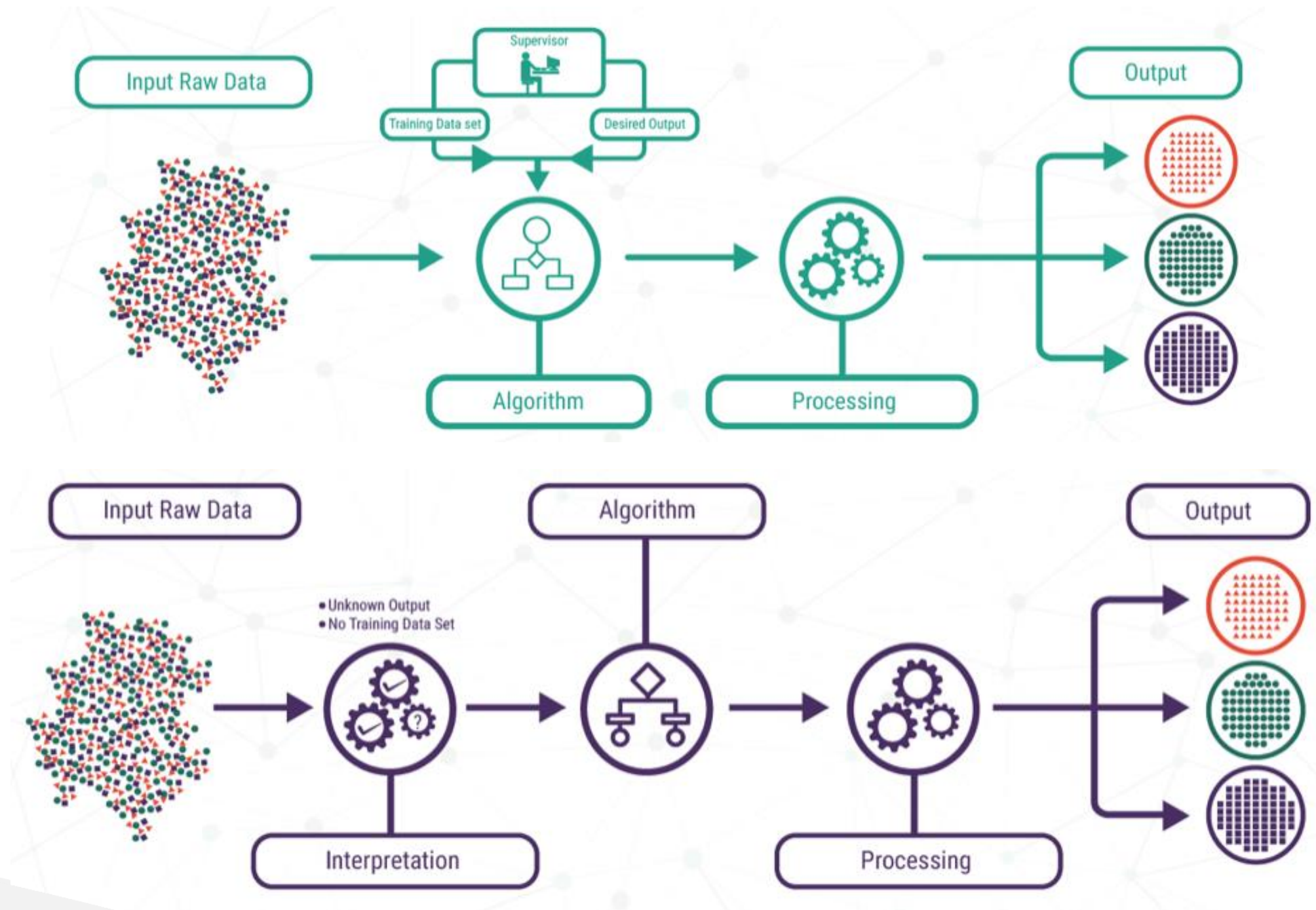
- Da una raccolta di 1.000.000 di geni diversi si trova un modo per raggruppare automaticamente questi geni in gruppi che sono in qualche modo simili o correlati da variabili diverse, come posizione, ruoli e così via.
- Identificare singole voci e musica da un insieme di suoni in un bar.  
([https://cnl.salk.edu/~tewon/Blind/blind\\_audio.html](https://cnl.salk.edu/~tewon/Blind/blind_audio.html))



## Supervisionato vs Non Supervisionato



## Supervisionato vs Non Supervisionato

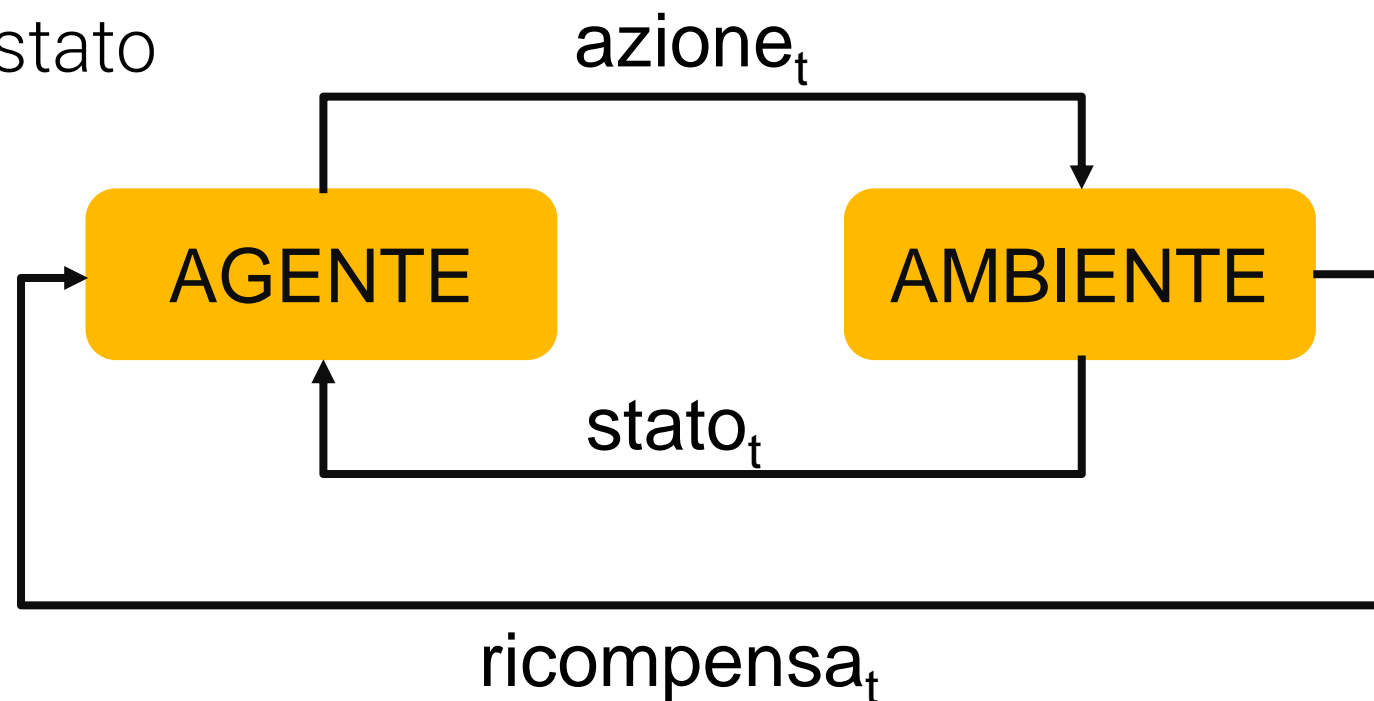


# Apprendimento per Rinforzo

Apprendimento tramite interazione con l'ambiente e le conseguenze delle proprie azioni.

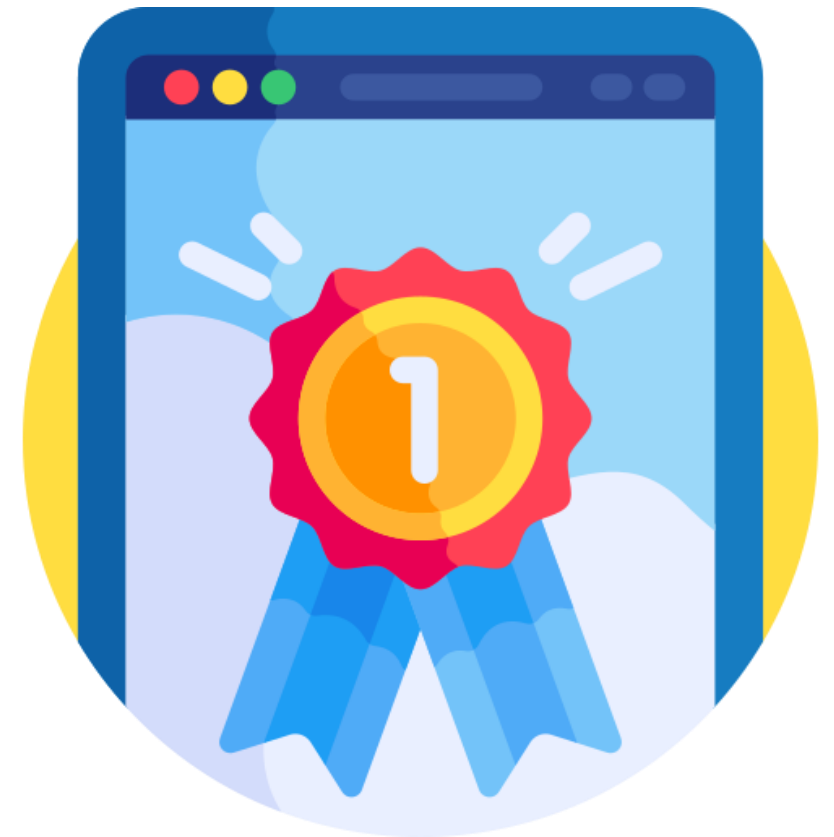
Step:

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



## Apprendimento per Rinforzo - scenari

- Guida autonoma
- Videogames
- Robot industriali
- Trading e finanza
- Recommender System
- Allocazione risorse in cluster di computer
- Controllo delle luci dei semafori
- Text mining (sommari, ricerche, ...)
- Ottimizzazione trattamenti clinici
- Dynamic pricing
- Fraud detection
- ...





## Processo di implementazione

### Raccolta Dati

Scouting su  
diverse fonti

### Preprocessing

Pulizia e omogeneità  
nei dati selezionati

### Feature Eng.

Incremento dell'utilità  
del set di dati creato

### Selez. Modello

Identificazione e  
addestramento

### Predizioni

Validazione del  
modello

# Raccolta dati

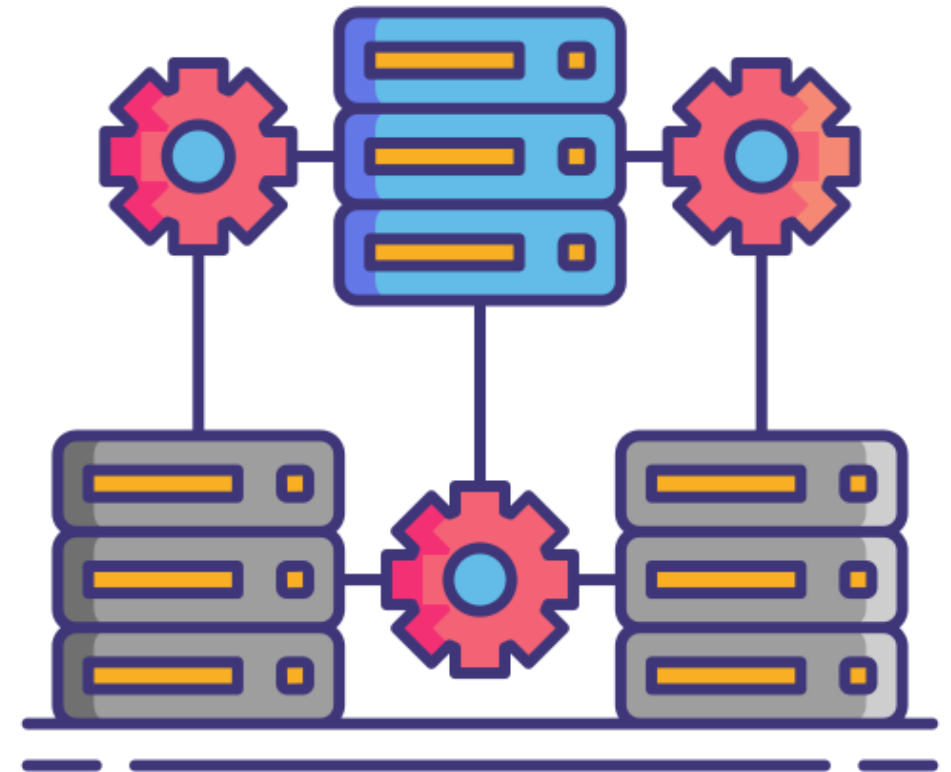
Dipendente dal contributo umano

- etichettatura manuale
- esperti di dominio

Spesso sono disponibili dataset pubblici per arricchire la base di conoscenza del problema in analisi

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

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

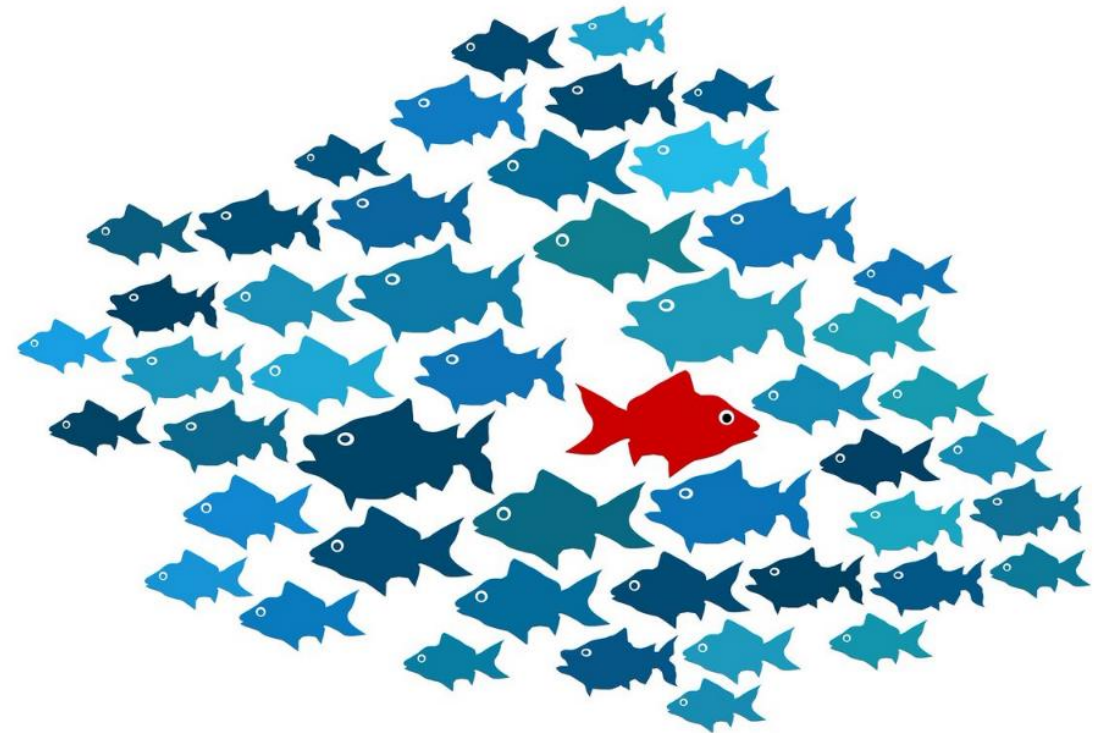




# Preprocessing Dati

## Risoluzione delle incongruità nei dati selezionati

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



# Esempio di Preprocessing 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	Italy	Bari
148	113-0149	Roberta Gialli	1-1-1975	F	0	1	Italy	Genova
149	115-1245	Mike Reds	03/05/1992	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  
correttovalore non  
validoinformazioni  
ridondantivalori  
fuorviantierrori  
sintattici

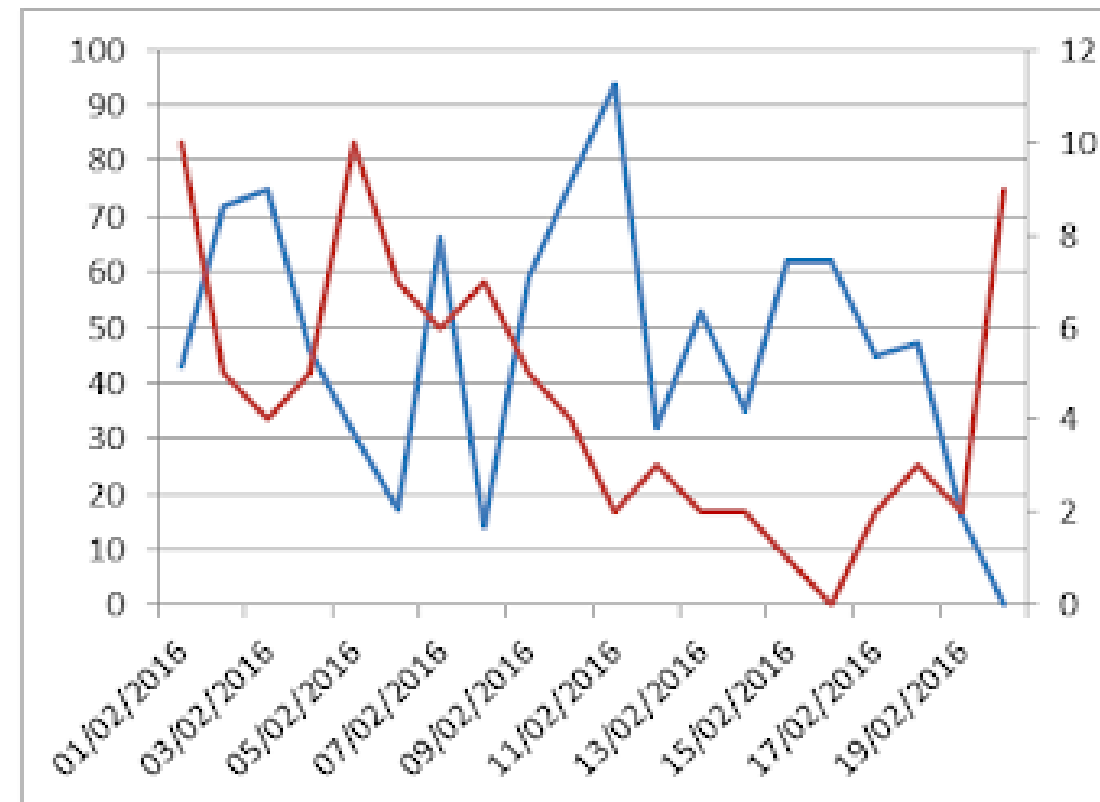
# Preprocessing Dati

## 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:

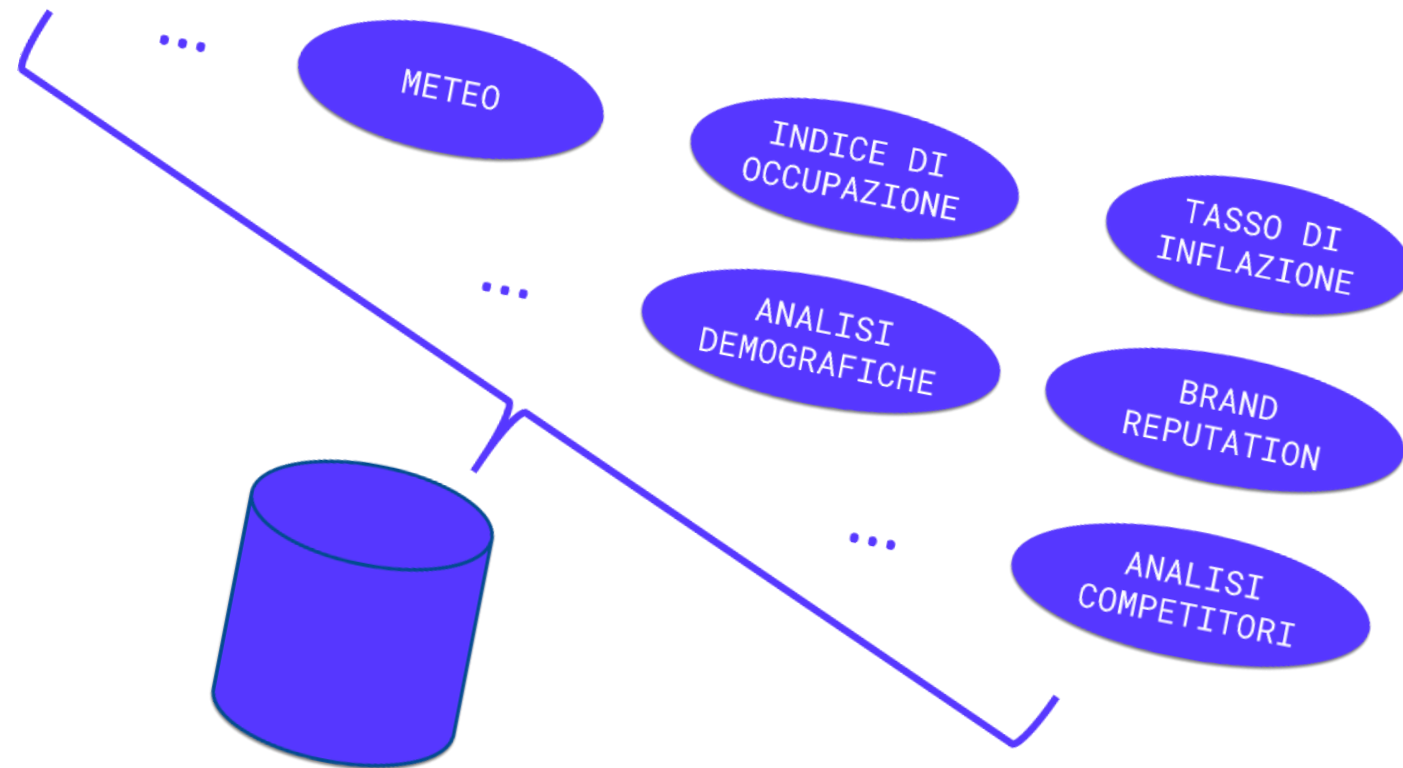
- fornire ulteriori conoscenza al sistema
- verificare l'efficacia dei dati selezionati, rimuovendo eventuali ridondanze



# 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.



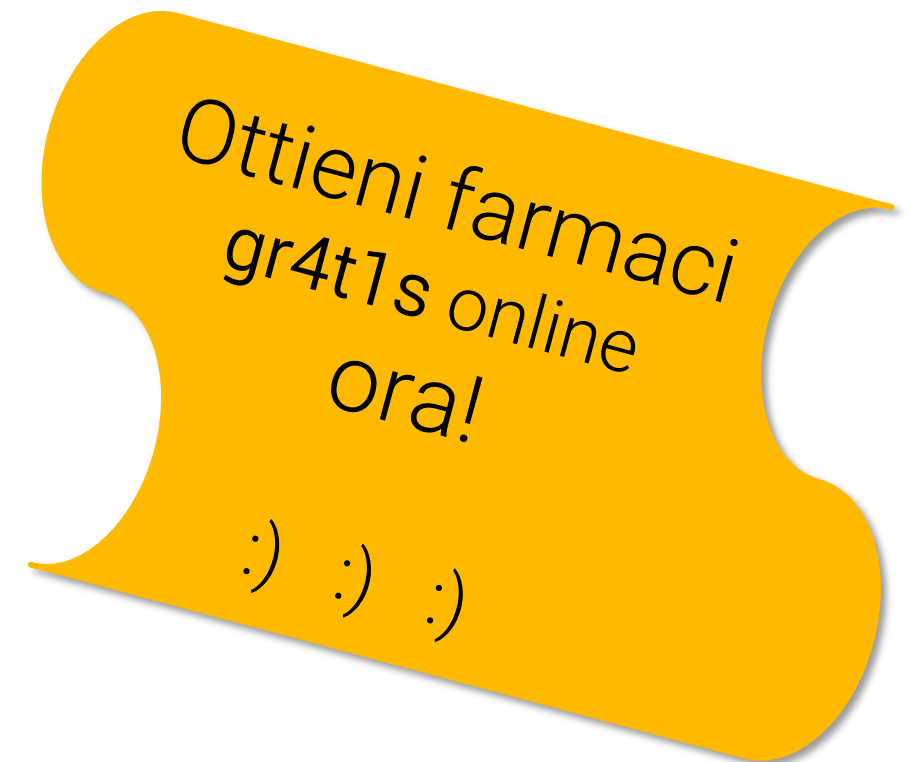
# Feature Engineering

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



[ 1, 0, 3 ]

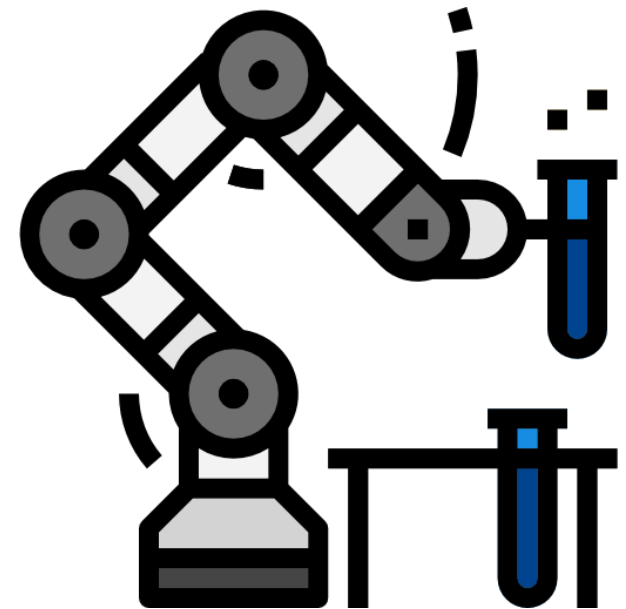
# Feature Engineering

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





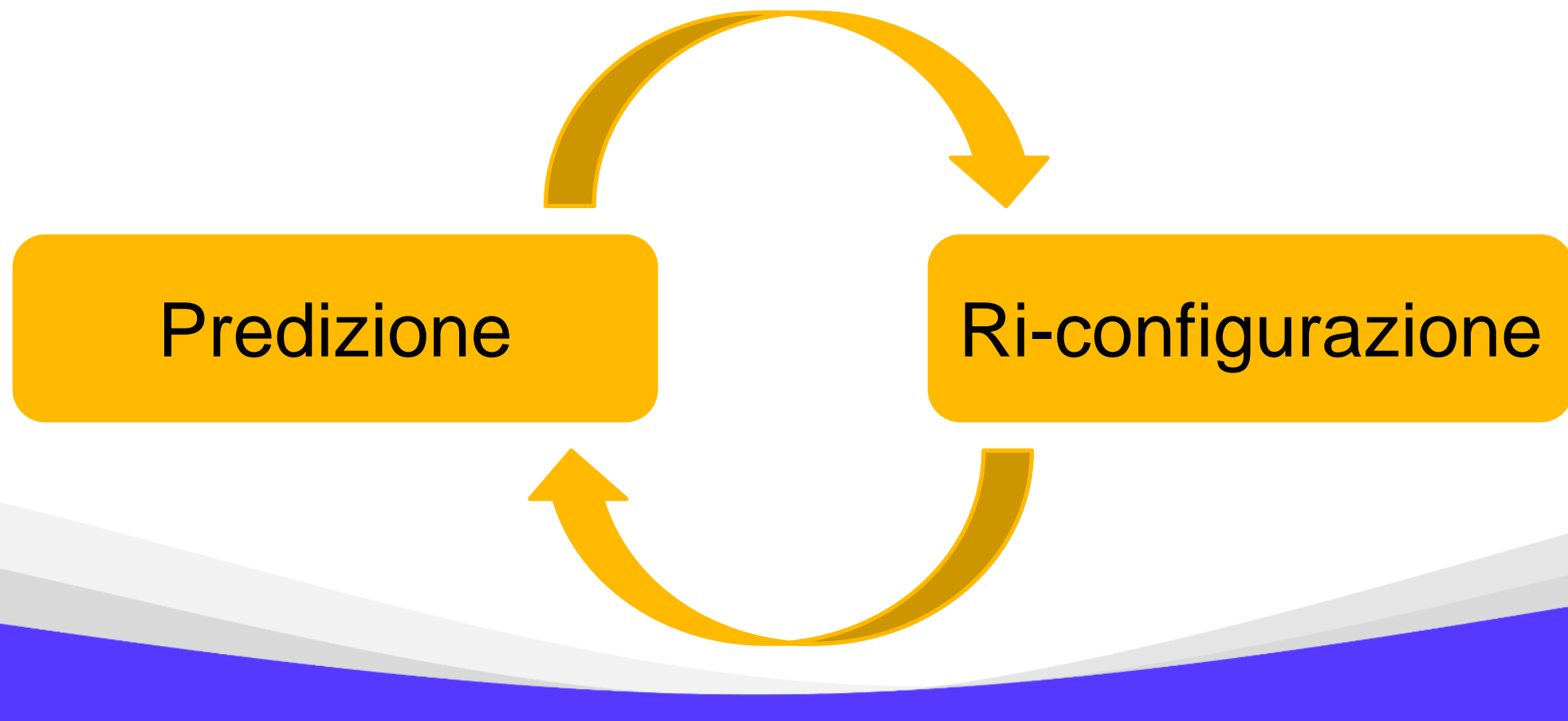


# Addestramento

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

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



addestramento

The diagram shows a horizontal orange bar divided into three sections: 'addestramento', 'validazione', and 'test'. A bracket below the first two sections is labeled 'dataset selezionato'. The entire diagram is set against a blue background with a white wavy line at the bottom.

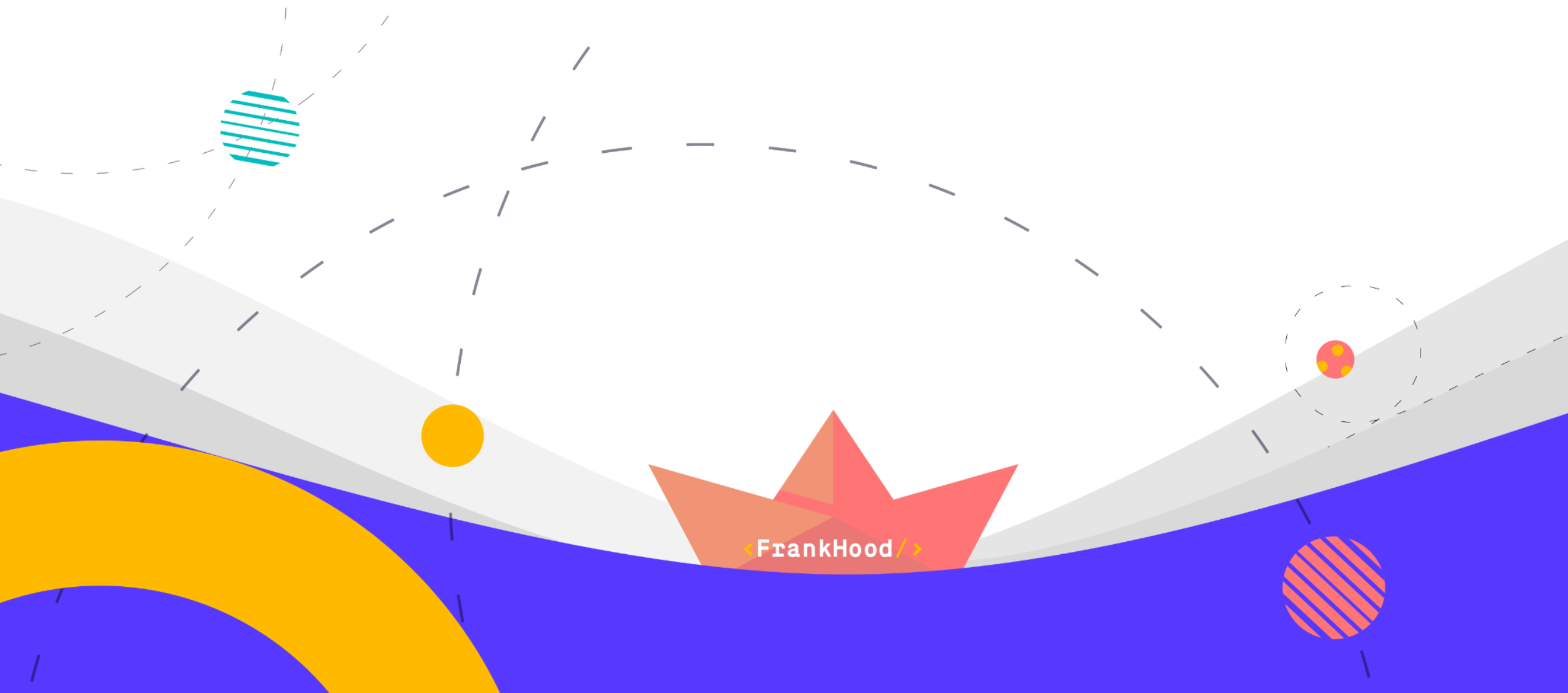
validazione

test

dataset selezionato



# DATA MASTERS



<FrankHood/>