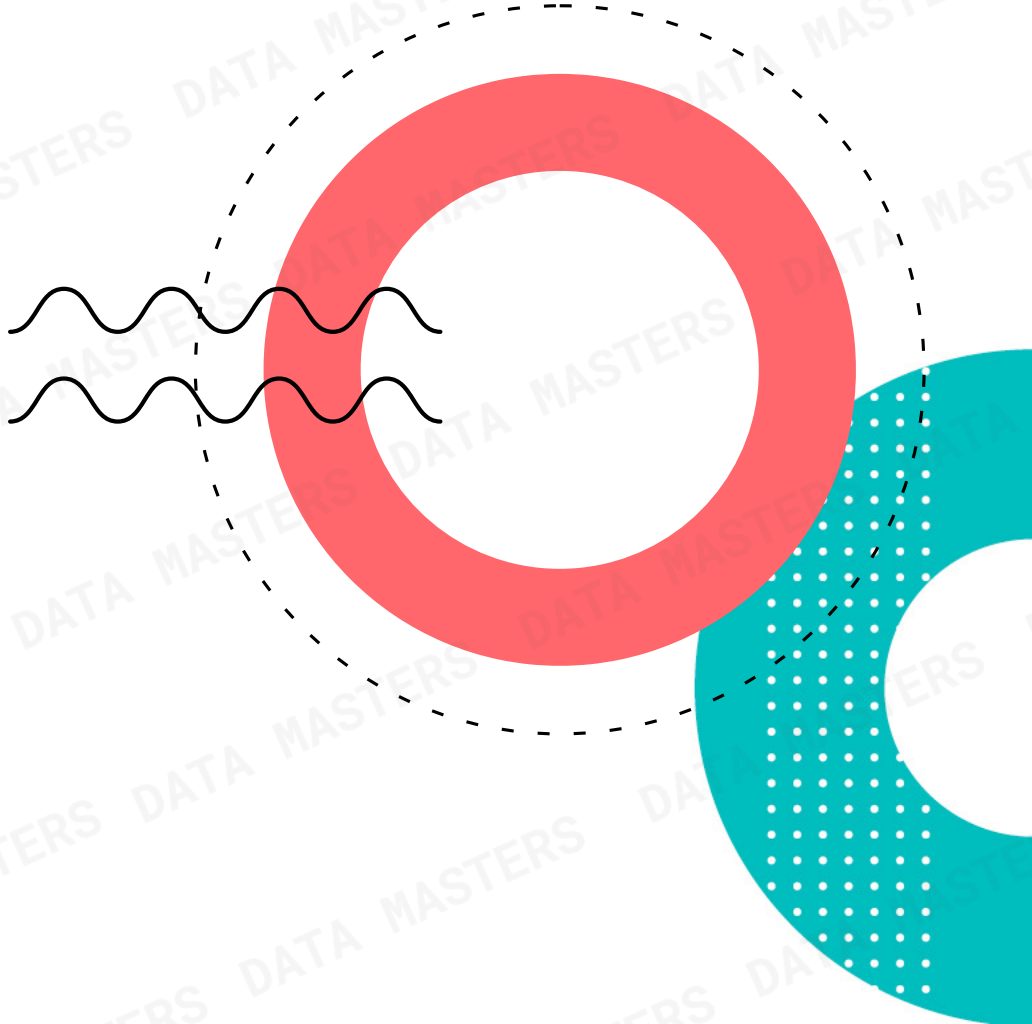
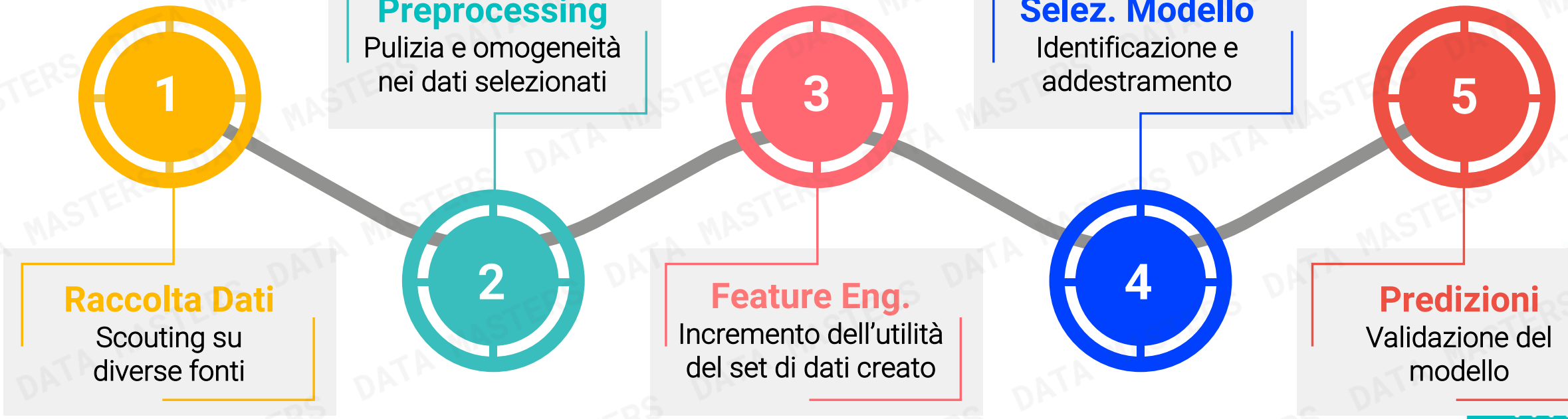
PROCESSO di IMPLEMENTAZIONE



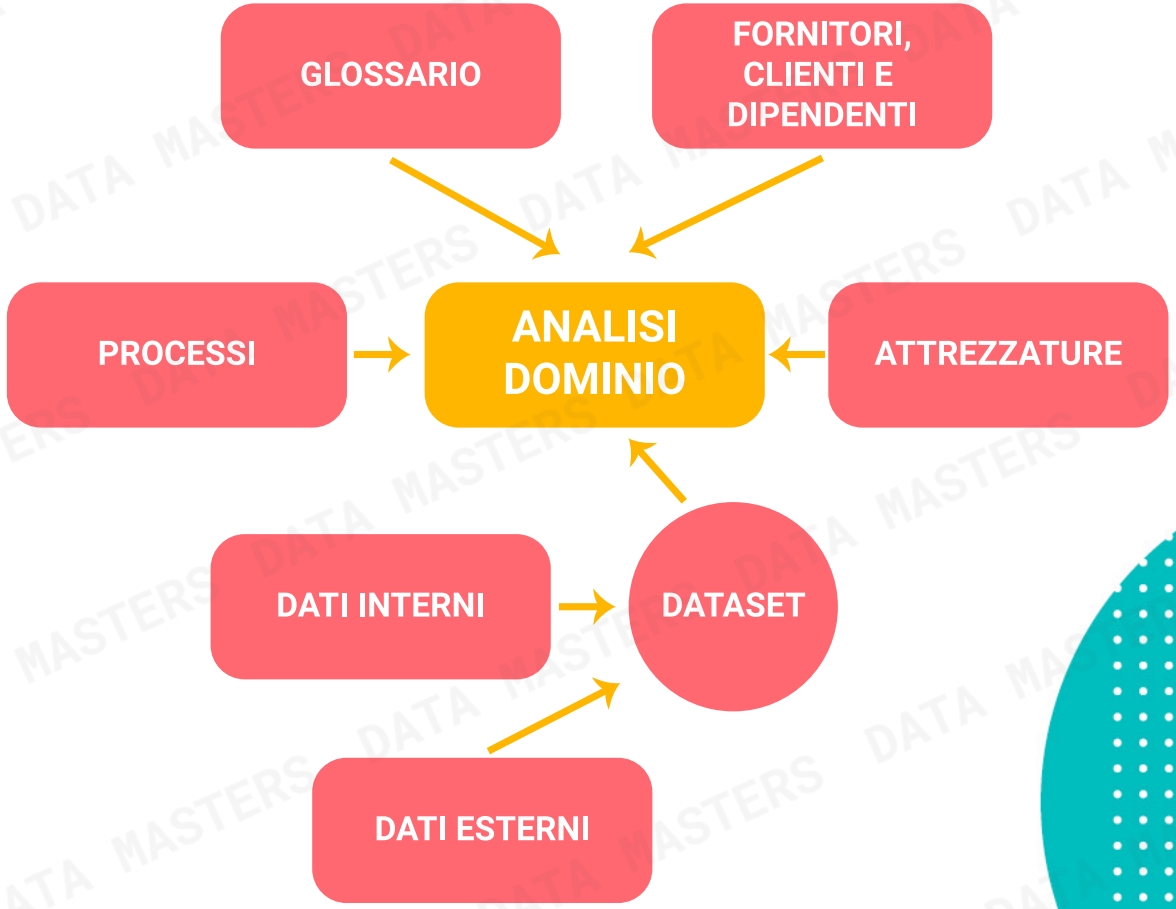
INTRODUZIONE

Processo di implementazione



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

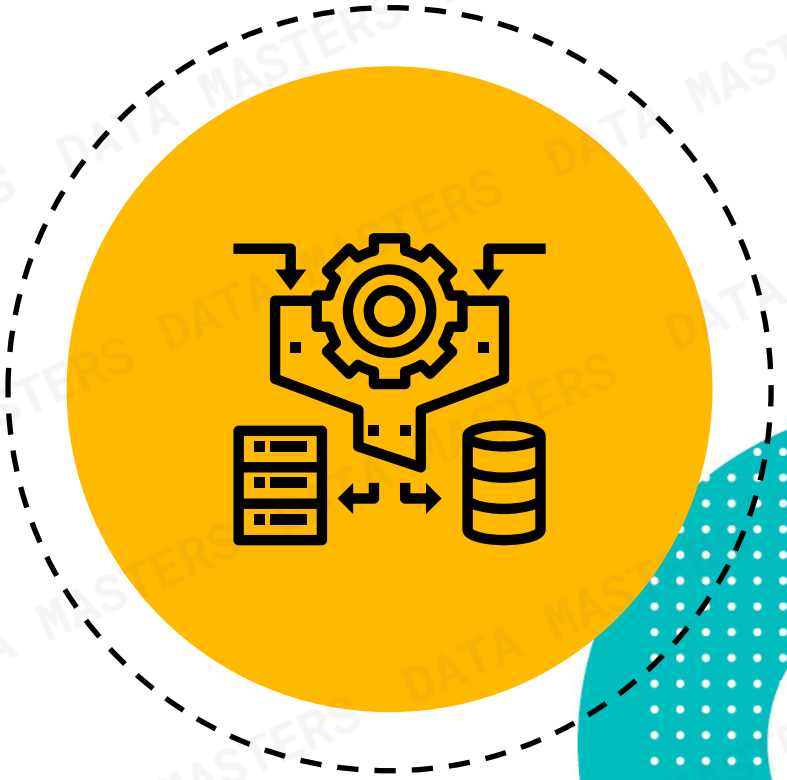
Dipendente dal contributo umano

- *etichettatura manuale*
- *esperti di dominio*

Spesso sono disponibili **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)

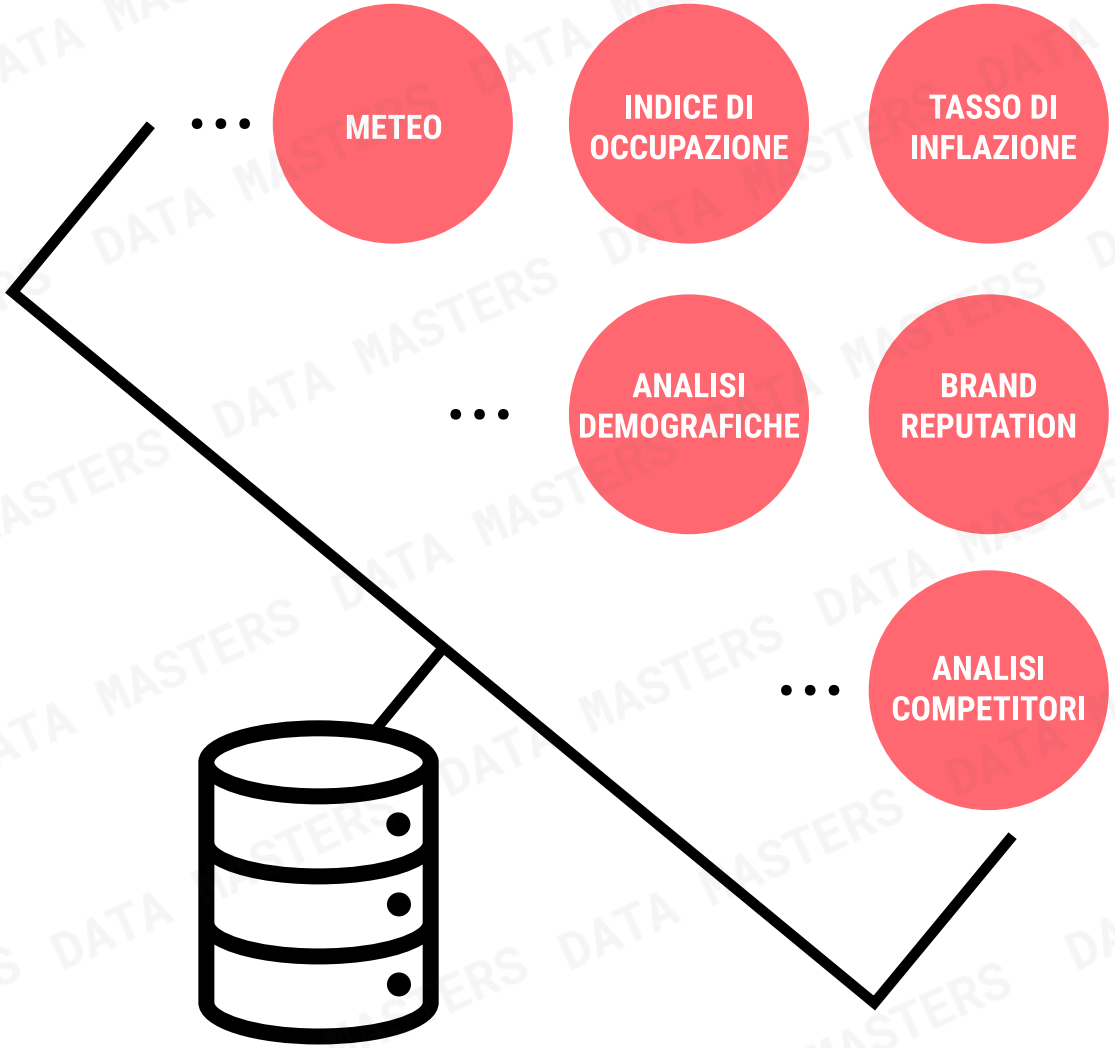
La **quantità** e la **qualità dei dati** influenzano l'accuratezza finale



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.

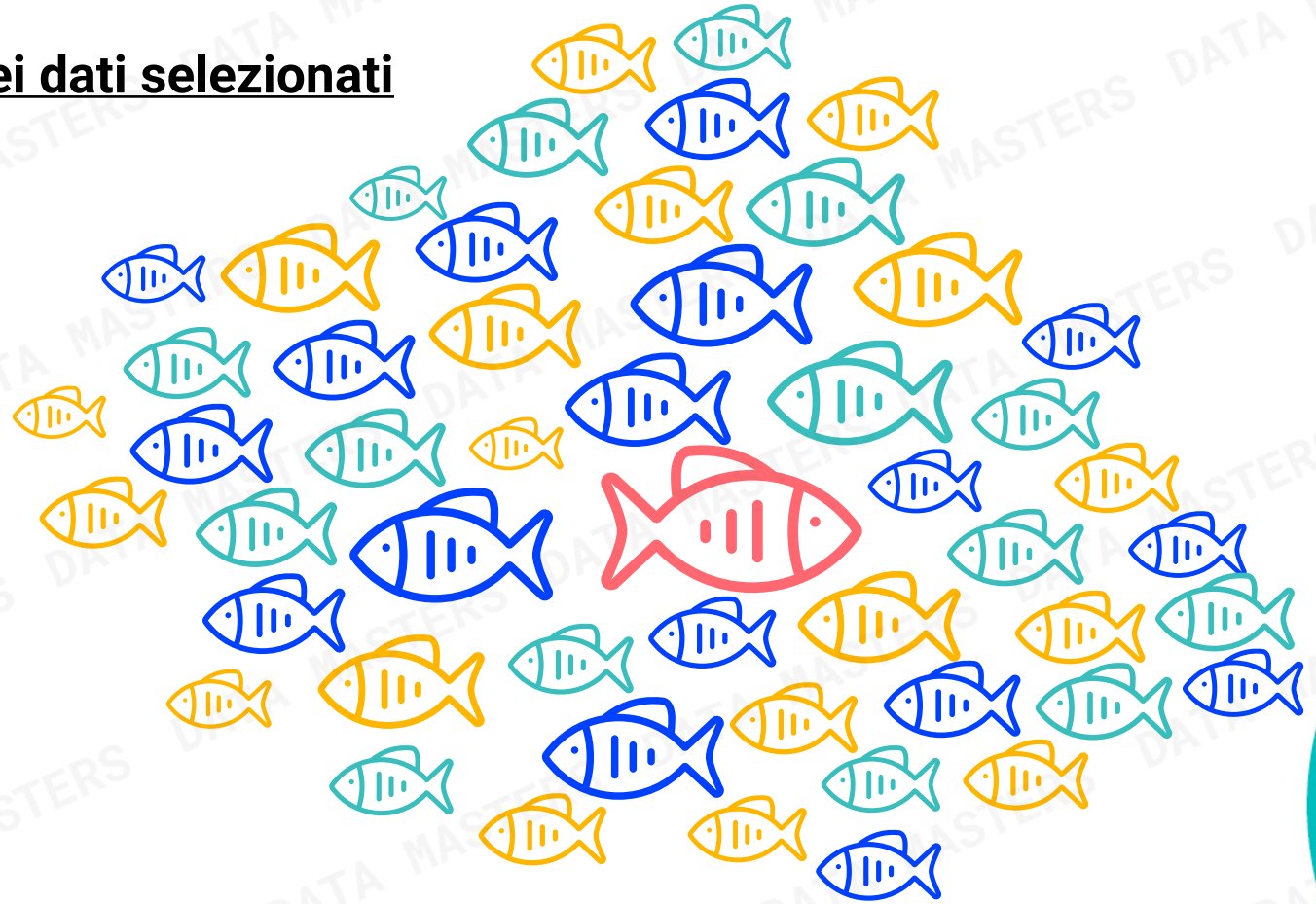
L'utilizzo di fonti di informazione esterne permette di avere una **più completa visione** dell'ambiente in cui si opera e assicura performance superiori a tutti i moduli del sistema.



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	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

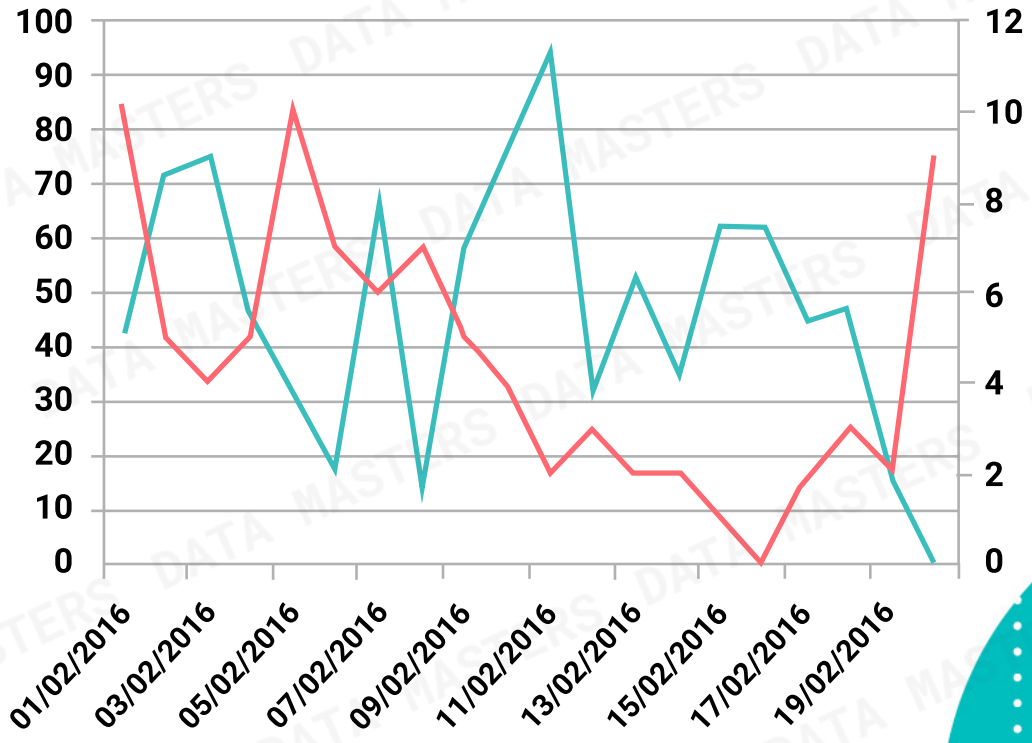
Preprocessing Dati

Analisi delle correlazioni e delle 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



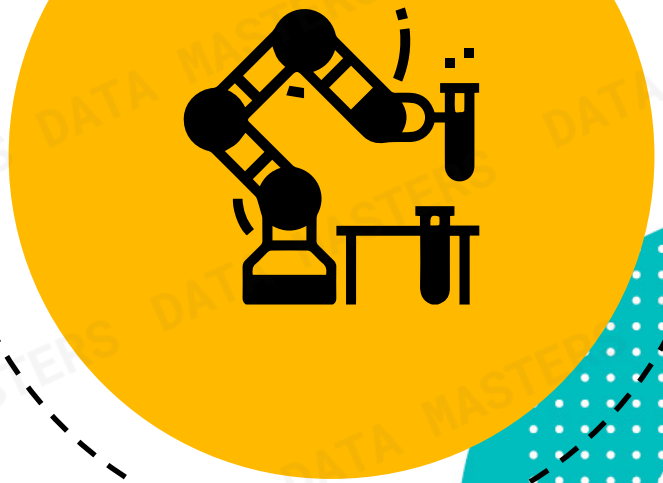
Temperatura °F

Numero di voli da Londra a Ibiza

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

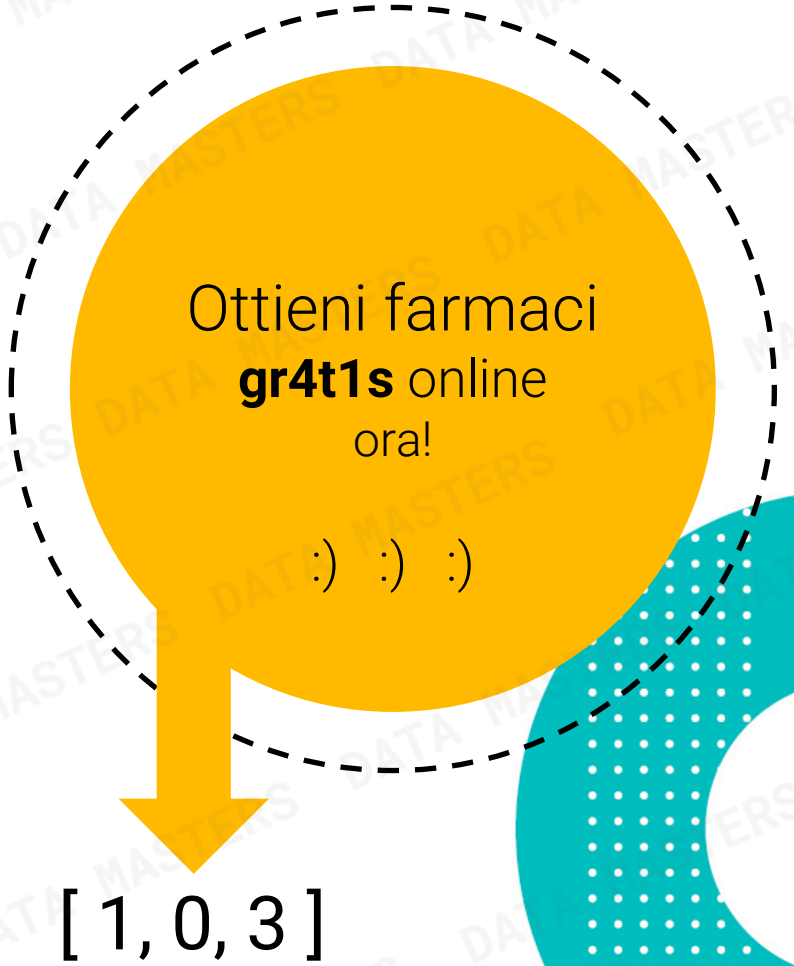
Feature Engineering

Una **feature** è una proprietà misurabile del fenomeno osservato

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

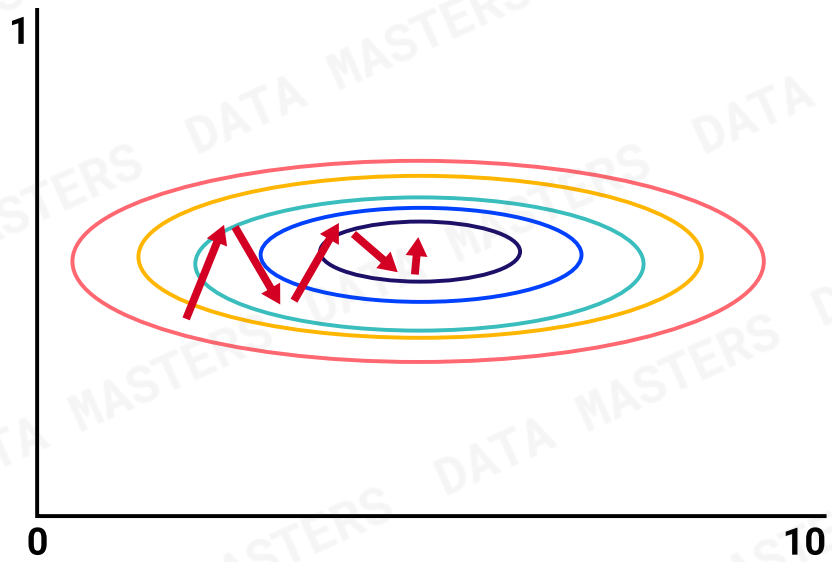
Ad esempio possiamo classificare dei messaggi
di posta elettronica in base a:

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

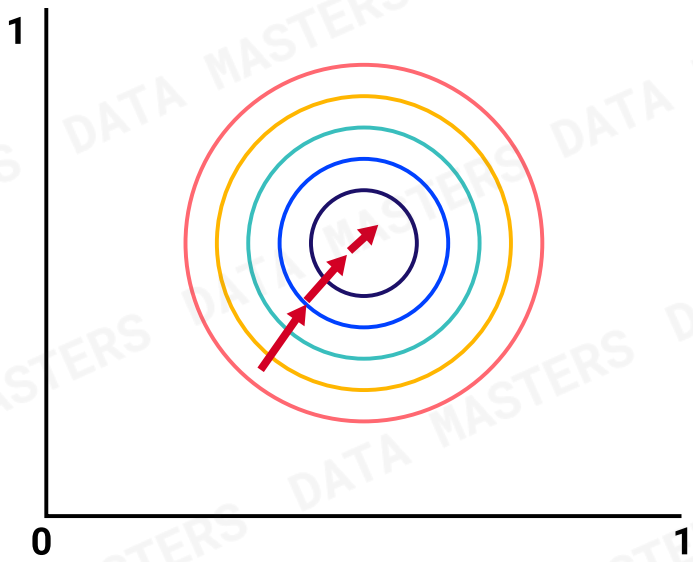


Feature Engineering - Normalizzazione

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 nello stesso intervallo di valori.



alcuni parametri
contribuiscono più di altri



tutti i parametri forniscono
contributi paragonabili

Data Augmentation

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

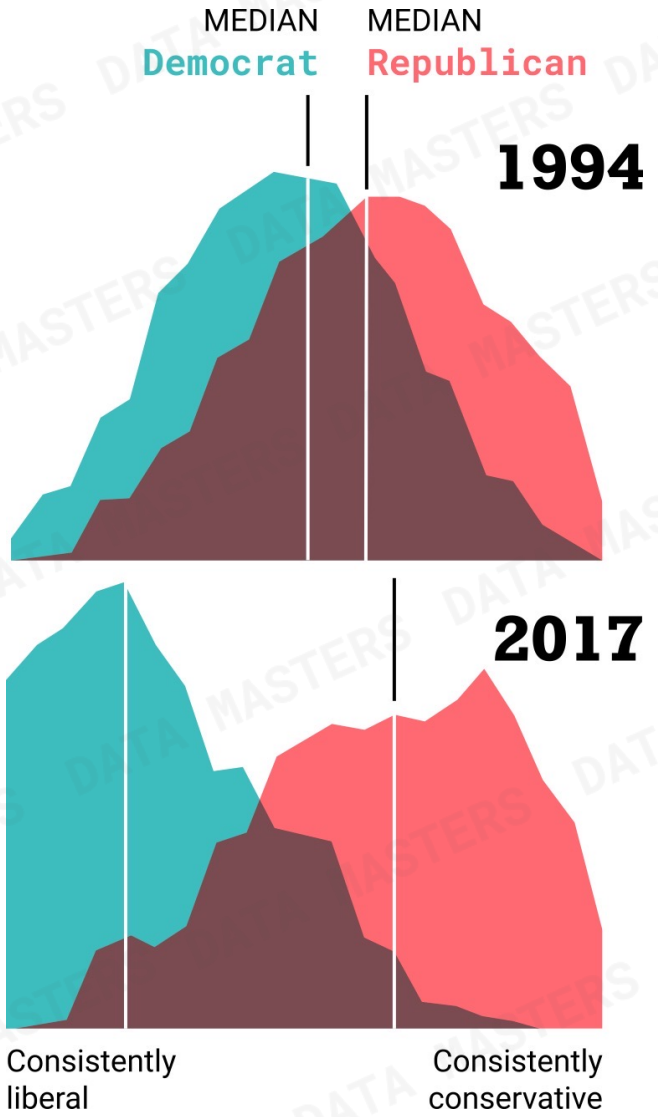
Incrementare la quantità di dati disponibili aiuta gli algoritmi di machine learning nel raggiungere **performance elevate**.

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:

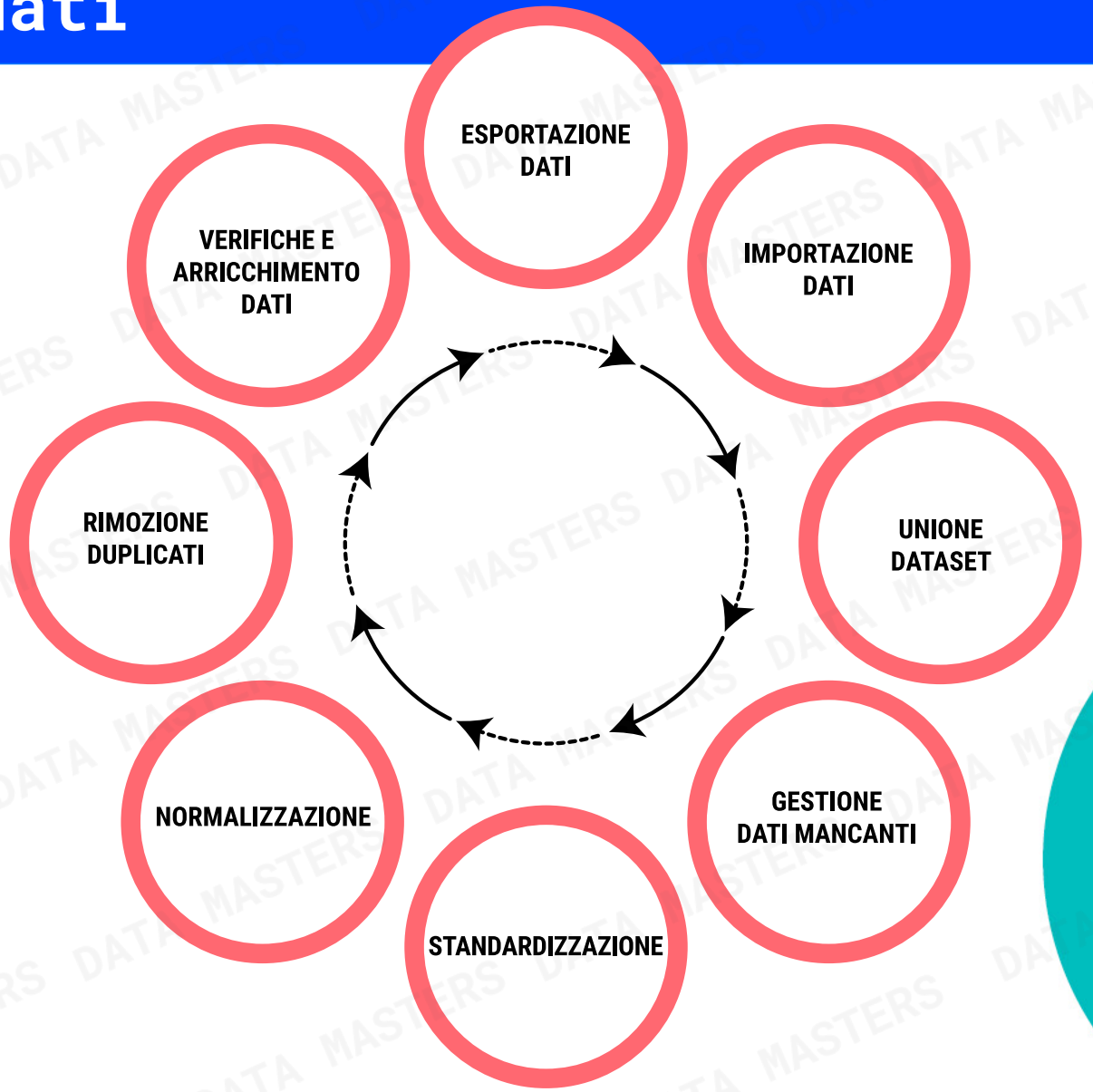
- fornisce utili informazioni all'analisi ed allo studio dell'evoluzione dal passato al presente
- propone scenari futuri plausibili



Recap: preprocessing dati

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

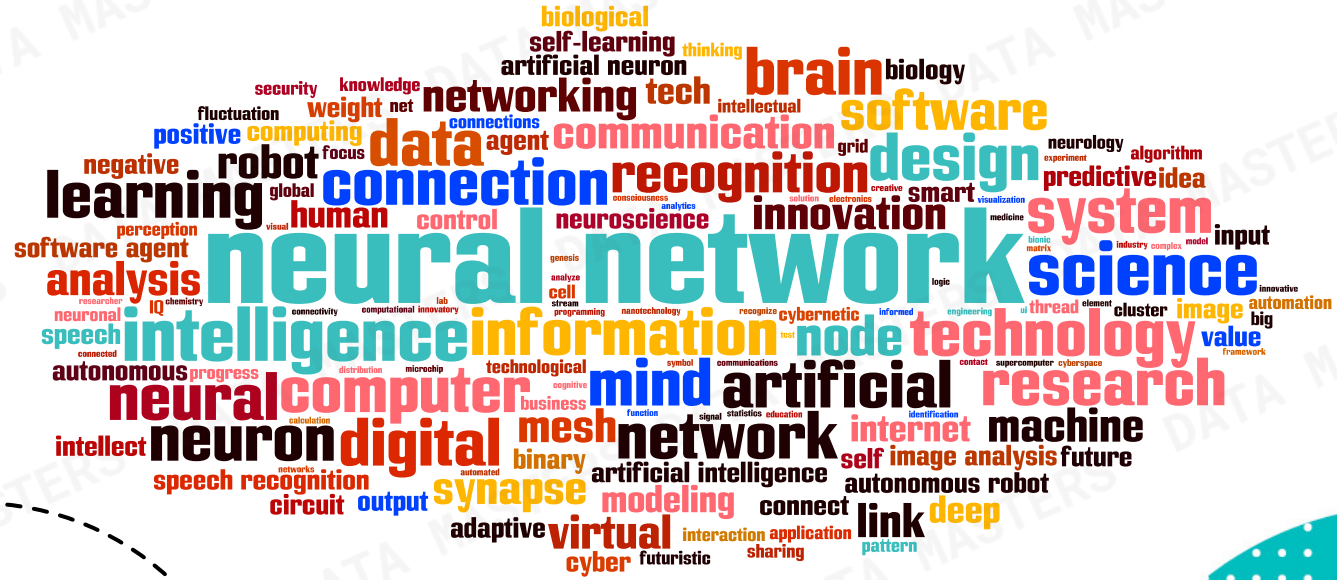
Garantisce ai moduli del sistema il **costante utilizzo** di dataset consoni.



Identificazione Modello

Apprendimento SUPERVISIONATO

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



Apprendimento NON SUPERVISIONATO

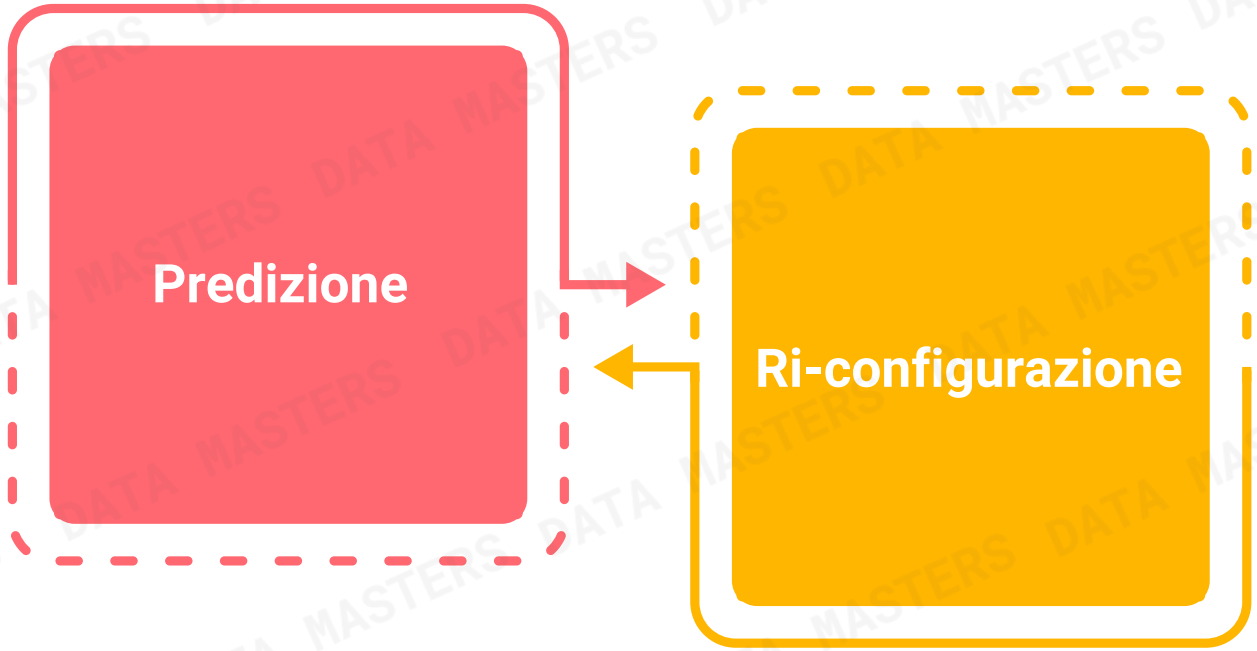
Principal Component Analysis (PCA)
t-SNE
k-means
DBScan

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

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