Energee4
AI SOLUTIONS AND SERVICES

energee4.ai

# Energee4

## AI dal progetto all'operatività: la gestione del ciclo di vita

Maggio 2025

# Chi siamo

## Giovacchino Tesi

**Responsabile Gruppo Innovazione** Energee3 srl
**CEO** Energee4 srl
Docente corso "*Intelligenza Artificiale Applicata All'Industria*" Informatica Unipr
Docente corso "*Cloud Computing*" Comunicazione ed Economia Unimore

Laureato in fisica, da 30 anni lavora nel mondo IT

## Valeria Guttilla

**Project Manager AI** e **Account Manager** Energee3 srl
Master in Big Data Unipi

Laureata in Economia
Da diversi anni si occupa di progetti in ambito IT e AI

# Chi siamo

Energee4 è la società del **Gruppo Energee3** specializzata in **AI** e **Machine Learning**, in grado di supportare le aziende nella:

- **Progettazione e realizzazione** di sistemi basati sull'AI
- **Analisi** dati avanzata
- **Consulenza sulle tecnologie** da adottare
- **Formazione** del personale per un migliore utilizzo delle tecnologie
- **Ricerca** nell'ambito di sistemi e algoritmi

Le **collaborazioni** attive da diversi anni con primarie Università e istituti di ricerca:

- Università di Pisa (UniPI)
- Università degli Studi di Parma (UniPR)
- Università  degli studi di i Modena e Reggio Emilia (Unimore)
- ILC (CNR Pisa)
- Luiss

insieme alla presenza all'interno dell'azienda di docenti universitari, garantiscono il **continuo aggiornamento** ai temi ed alle tecnologie più innovative.

# I servizi

### Motori di ricerca avanzati
Sviluppiamo strumenti per la ricerca semantica di informazioni all'interno di testi e immagini

### Strumenti di forecasting e analisi dati
Realizziamo modelli previsionali basati su statistiche e Machine Learning, personalizzati secondo i differenti settori aziendali

### Automazione di processi
Creiamo processi automatizzati con soluzioni personalizzate anche grazie all'utilizzo di agenti intelligenti

### Analisi automatica di video e immagini
Sviluppiamo sistemi per l'estrazione di informazioni e modelli comportamentali da immagini e video

### Estrazione di informazioni da documenti
Classificazione e confronto di informazioni fra diversi documenti contenenti testo ed immagini

# Alcune delle nostre Case History

**Motori di ricerca avanzati - Semantic Search documents** **Settore: Finance**

Motore di ricerca semantico appositamente addestrato per il contesto bancario. Pensato e sviluppato per effettuare ricerche all'interno dei documenti bancari fornendo, in combinazione con un chatbot, un supporto al personale interno nella soluzione di problemi

**Automazione di processi - Traduttore automatico di codice** **Settore: Finance**

Utilizzo di API GPT4 e regole per la traduzione automatica di codice da 4GL (configurabile) ad altri linguaggi a scelta del cliente (es da FOCUS a DB2/JCL/COBOL)

**Analisi automatica di video e immagini - Tennis commander** **Settore: Sport**

Integrazione di dati provenienti da analisi video e sensoristica (accelerometri, giroscopi) per l'analisi delle performance sportive

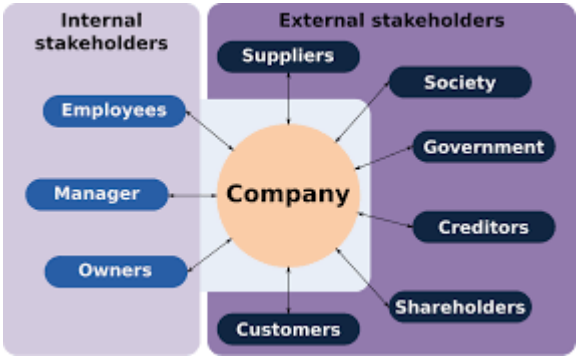
**Estrazione di informazioni da documenti - Benchmark assicurativo** **Settore: Insurance**

Confronto delle clausole contrattuali di prodotti di diverse compagnie e rami per una lettura dei competitor più rapida ed efficace tramite dashboard ed una definizione mirata della propria offerta

# Metodo come comunicazione

Ciclo di vita ben definito come strumento di comunicazione tra i team tecnici e gli stakeholder aziendali

Non è solo una best practice tecnica, ma un imperativo strategico per integrare con successo le iniziative di IA nel più ampio contesto organizzativo




https://langfuse.com/

# Definizione problema progetti AI

Hanno una componente di incertezza intrinseca non eliminabile!!

Attualmente sono suddivisibili in due categorie:

- Progetti che prevedono modelli da addestrare con dati (ML)
- Progetti che prevedono l’interrogazione di modelli fondazionali (LLM): programmazione agentica

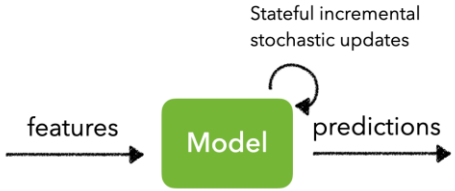
# Programmazione basata sul ML

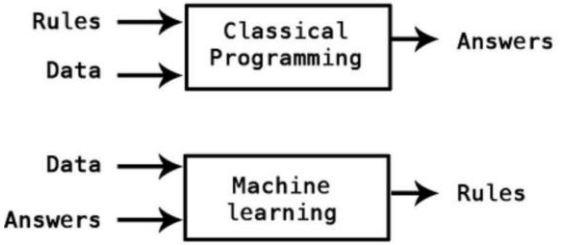
In generale il ML ha come obiettivo di trovare info nei dati

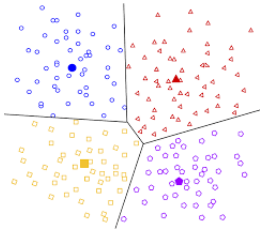
y=F(x) trovare la funzione

oppure

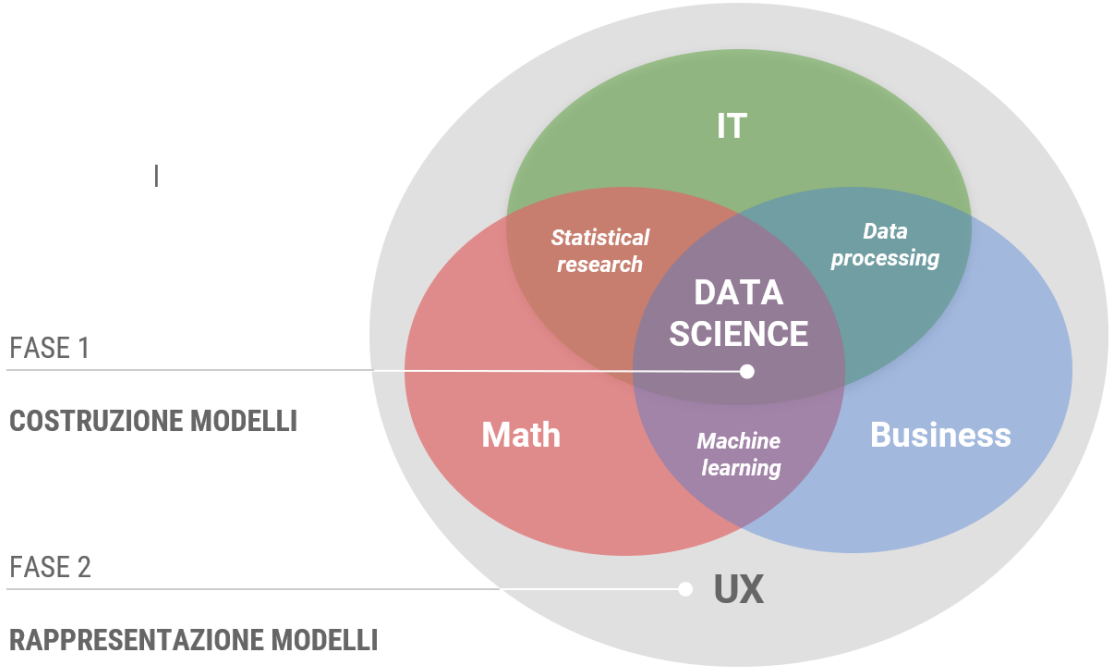
Trovare gruppi di dati omogenei (clustering)



# Programmazione basata sul ML




La costruzione di un algoritmo di ML è in genere un lavoro di squadra che richiede competenze in varie discipline

Ricordiamoci che stiamo cercando di risolvere problemi ***difficili*** da modellare in altro modo

Il lavoro del data scientist assomiglia più al lavoro di un analista che a quello di un tecnico informatico

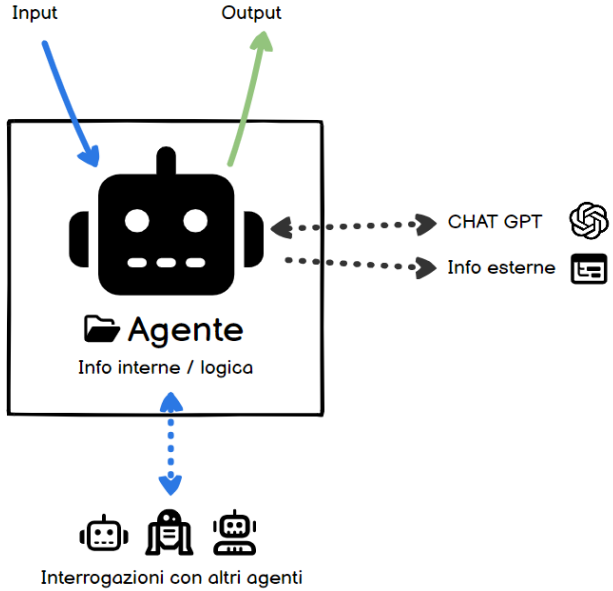
Riuscire a capire se l'algoritmo risolve il problema implica un'ottima conoscenza del dominio di business

# Programmazione agentica

L’agente è un software con una base di conoscenza in grado di:

- ricevere istruzioni per compiere un lavoro
- pianificare azioni
- compiere azioni
- controllare l’effetto delle azioni (feedback)

Il tutto tramite l’utilizzo di foundation models (LLM)

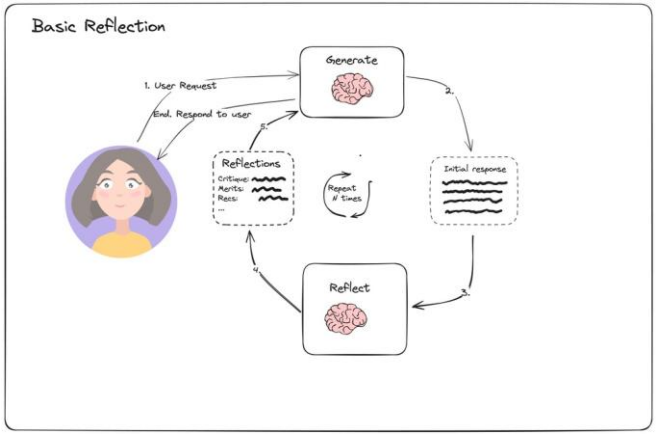


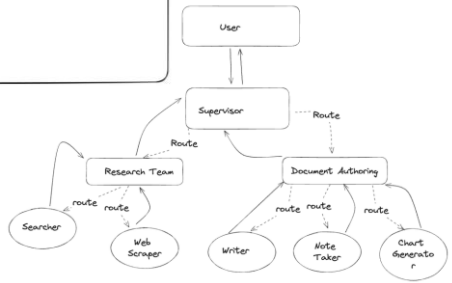
# Agenti che interagiscono con altri

Gli agenti tipicamente interagiscono tra loro

E’ possibile creare degli ‘uffici virtuali’

Così è possibile interrogare gli LLM in modo più preciso



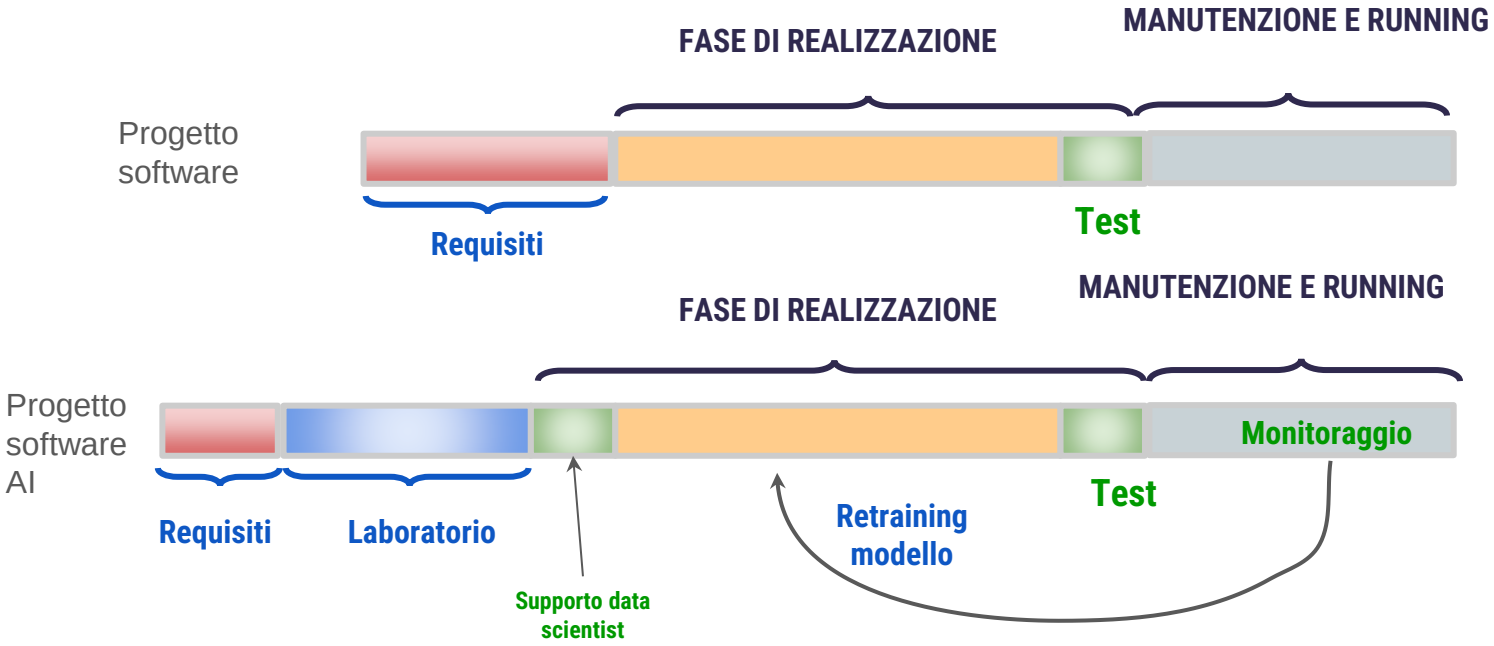
# Definizione problema: etica

Il problema che un sistema di IA è progettato per risolvere e i dati che è destinato a utilizzare contengono intrinsecamente potenziali insidie etiche

Importante considerare da subito: Equità, Trasparenza e Responsabilità

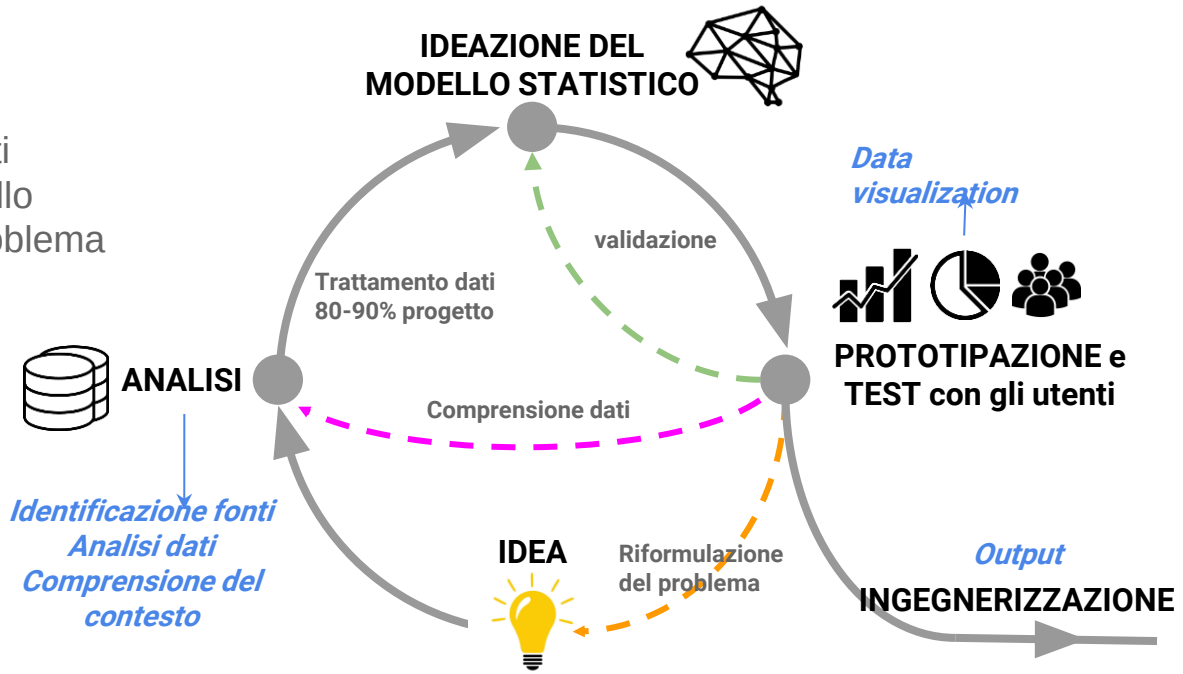
Affrontare queste questioni dopo che un sistema è stato costruito è spesso più complesso e costoso

# Il progetto


FASE DI REALIZZAZIONE
MANUTENZIONE E RUNNING
Progetto software
Requisiti
Test
FASE DI REALIZZAZIONE
MANUTENZIONE E RUNNING
Progetto software AI
Monitoraggio
Requisiti
Laboratorio
Test
Retraining modello
Supporto data scientist

# Ciclo di vita progetti ML (data science)

- Loop comprensione dati
- Loop validazione modello
- Loop riformulazione problema



# Metodologia CRISP

Metodologia orientata ai processi

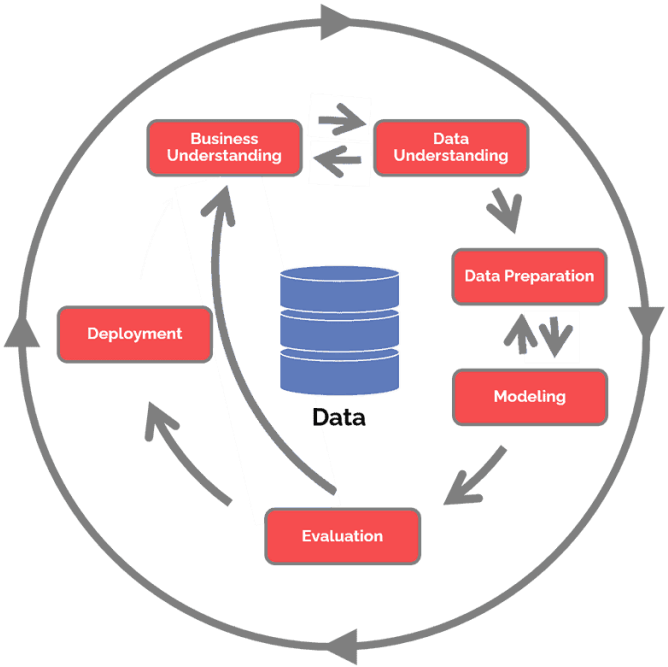
CRoss-Industry Standard Process for

Data Mining

Processo affidabile e ripetibile => standard

Il processo è iterato, ogni loop aumenta la quantità di informazioni

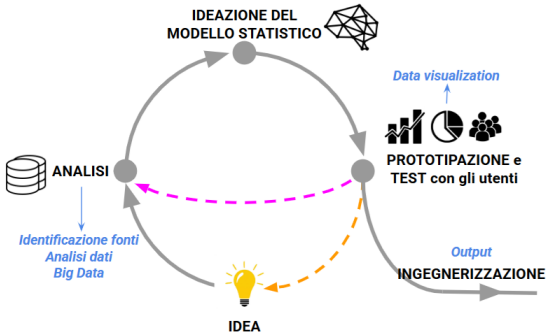
(arXiv:1907.04461v1 [cs.LG] 9 Jul 2019)



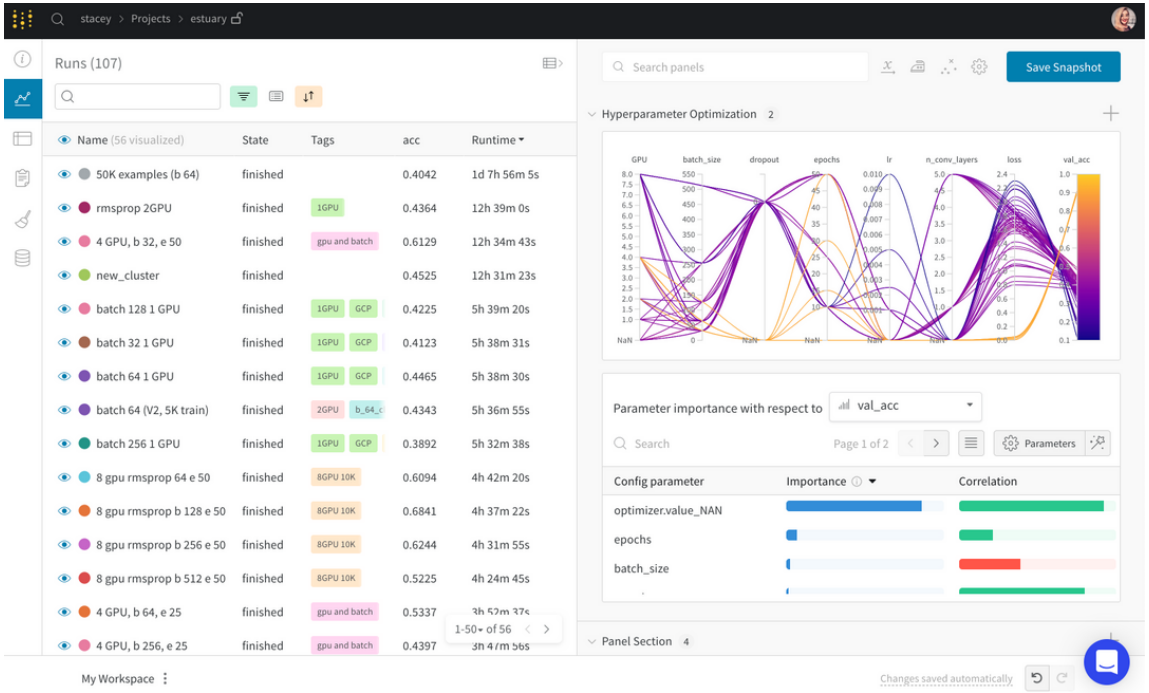
# Ciclo completo ML

AI expert o Data Scientist

AI expert + IT expert

# La fase sperimentale del ML … scienza




W&B

Sperimentare vuol dire fare tante prove… ragionate

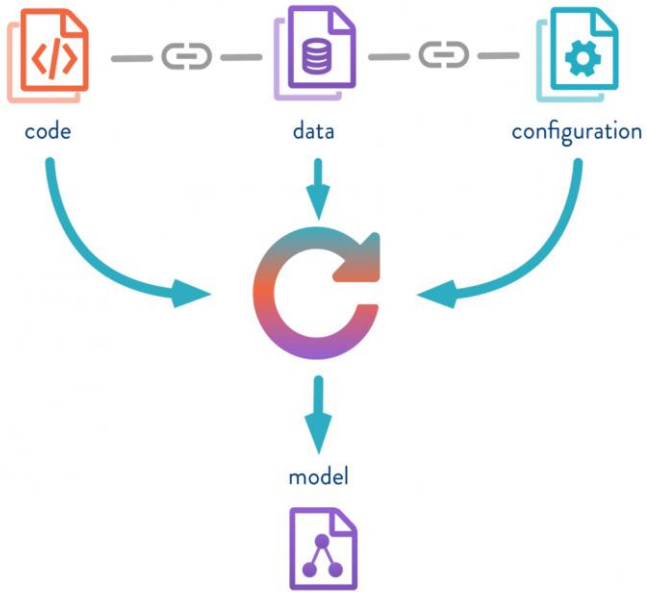
Sperimentare vuol dire anche misurare

Sperimentare vuol dire anche poter tornare sui propri passi

Per tracciare esperimenti, serve conservare codice, dati, configurazioni e risultati

Il risultato finale è frutto di un lavoro condiviso

# La fase sperimentale del ML




DVC

Esperimento scientifico => riproducibilità dei risultati

Già in laboratorio è importante versionare dati, configurazioni e codice

I dati sono ‘dentro’ il modello, per riprodurre i risultati occorre partire da condizioni iniziali identiche e dati uguali

Non sempre è possibile, soprattutto se si utilizzano modelli preaddestrati

# Caratteristiche prog. agentica

L'IA agentica, pur sfruttando il ML come una delle sue tecnologie abilitanti, introduce nuovi componenti quali il processo decisionale autonomo, l'adattamento dinamico e interazioni con l'ambiente guidate da LLM e dal **prompt engineering**.

Non sono solo implementazioni statiche sono concepiti per apprendere e adattarsi continuamente una volta in produzione (agenti con memoria o RAG)

Gli agenti IA sono programmi specializzati progettati per automatizzare compiti complessi, scomponendo operazioni intricate in componenti gestibili

In laboratorio, lo sviluppo dell'IA agentica è un processo iterativo <- LLM non deterministico

Feedback loop, dove l'agente apprende ricevendo input sull'efficacia delle sue decisioni o azioni, e la memoria/riflessione, dove l'agente registra le esperienze e riflette su di esse per affinare il proprio comportamento

# La fase sperimentale degli agenti

1. Configurazione Fondamentale: Prompt engineering e Retrieval-Augmented Generation (RAG)

2. Adattamento Iterativo: Feedback Loop e Memoria dell'Agente nelle Fasi Iniziali

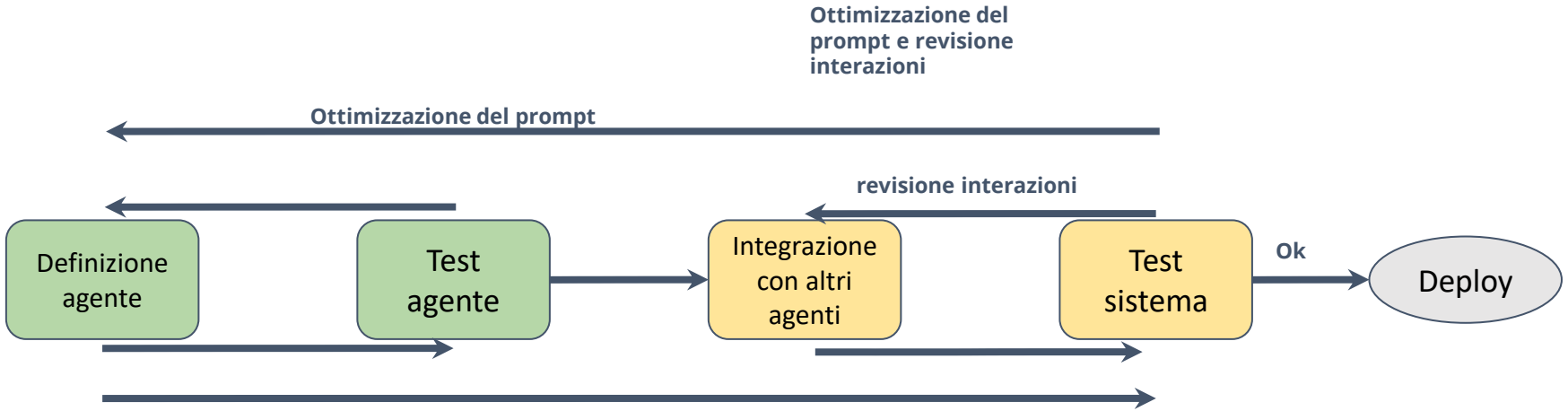
3. Tracciamento e Versioning dei Prompt e delle Configurazioni degli Agenti

**L'efficacia dell'IA agentica dipende fortemente dalla qualità dei prompt. Questo processo è iterativo e sperimentale, richiedendo un approccio disciplinato analogo a MLOps per i modelli**

Processo di raffinamento iterativo anche a causa della natura stocastica degli LLM

“la pratica di creare istruzioni strutturate e specifiche per il compito” = prompt engineering

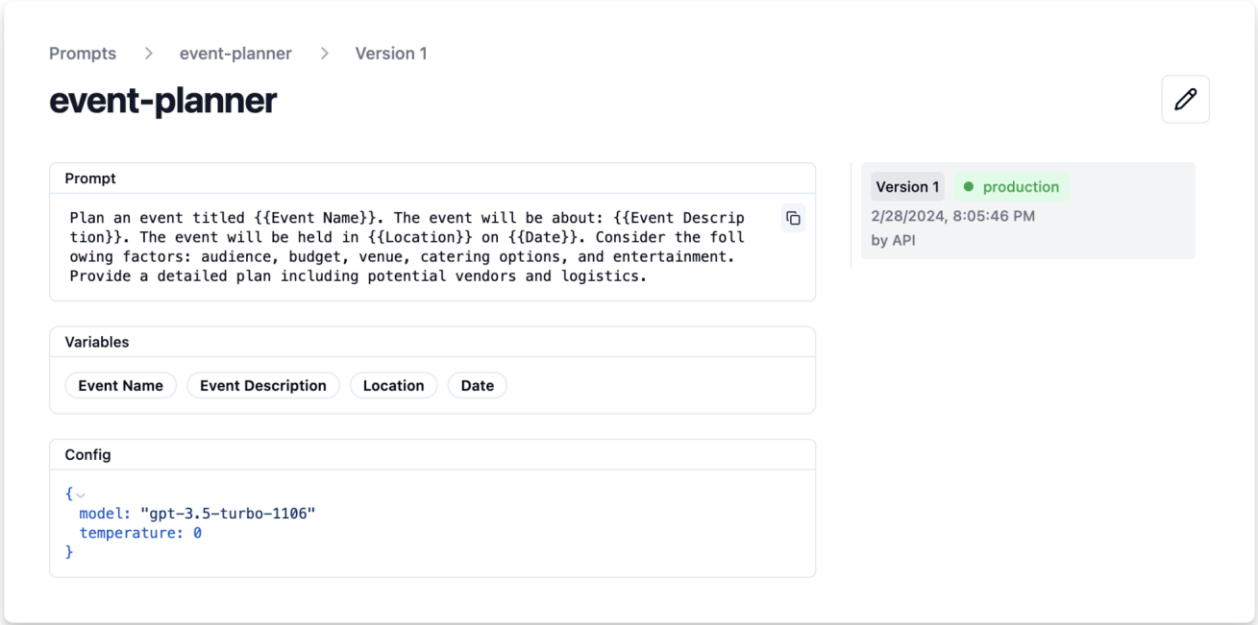
# La fase sperimentale degli agenti




Gran parte del lavoro è caratterizzato da cicli di ottimizzazione del prompt e degli algoritmi di interazione tra gli agenti. Questo a causa della natura probabilistica delle risposte degli LLM (GPT).

Il lavoro termina quando si ritiene ‘accettabile’ il risultato finale, ovvero quando il risultato finale è di sufficiente utilità per il cliente.

# La fase sperimentale degli agenti




https://langfuse.com

I principi del tracciamento degli esperimenti e del controllo di versione si applicano anche agli agenti.

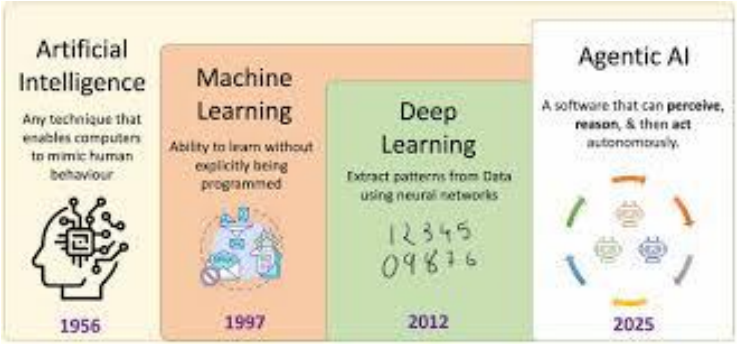
I prompt possono essere complessi e avere un'influenza determinante sul comportamento dell'agente

Se poi gli agenti sono molti…

# Differenze tra Agenti e ML 1

I progetti di ML sono tipicamente model-centrici, con il modello predittivo al centro dell'attenzione

I progetti di IA agentica sono behavior-centrici, focalizzati sulla capacità dell'agente di pianificare, ragionare e agire autonomamente




Di conseguenza, il ciclo di vita per l'IA agentica deve necessariamente includere la progettazione, il test e la gestione di questi comportamenti emergenti (LLM) e dei prompt o delle istruzioni che li guidano.

# Differenze tra Agenti e  ML 2

Questa caratteristica impone un ciclo di vita che pianifichi esplicitamente un'evoluzione continua, che va oltre il semplice riaddestramento periodico del modello

Governance human-in-the-loop (con intervento umano) per modifiche critiche, anche per tutto ciò che comporta la gestione della responsabilità

A differenza dei modelli ML tradizionali addestrati su dataset non completi, gli agenti basati su LLM sono spesso "configurati" o "istruiti" attraverso questi metodi già in laboratorio

# Fase di ingegnerizzazione

**Robustezza, Riproducibilità e Scalabilità => sistema in produzione!!**

Dall’ esperimento su piccola scala (Jupyter) alla produzione è richiesto:

- consolidamento del codice (hardening)
- la configurazione di pipeline automatizzate
- scalabilità del sistema per rispondere alla domanda
- definire metriche per riaddestramento o allarmi

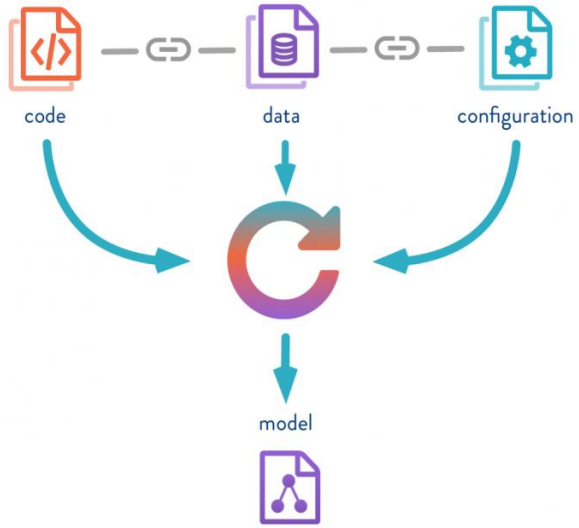
# Fase di ingegnerizzazione: MLOps

1. Gestione del Codice Sorgente e Best Practice per il Controllo di Versione

2. Registro dei Modelli: Centralizzazione e Versioning dei Modelli Distribuibili

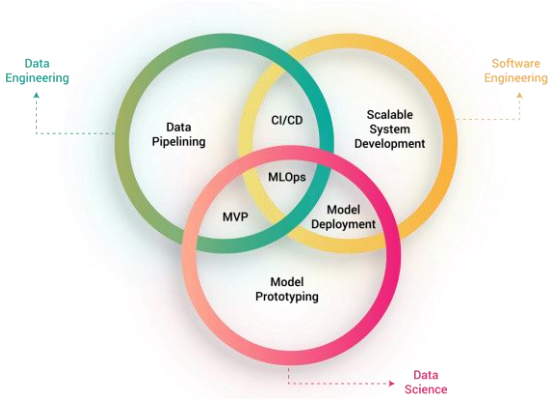
3. Automatizzare il Percorso verso la Produzione: Pipeline CI/CD/CT per ML

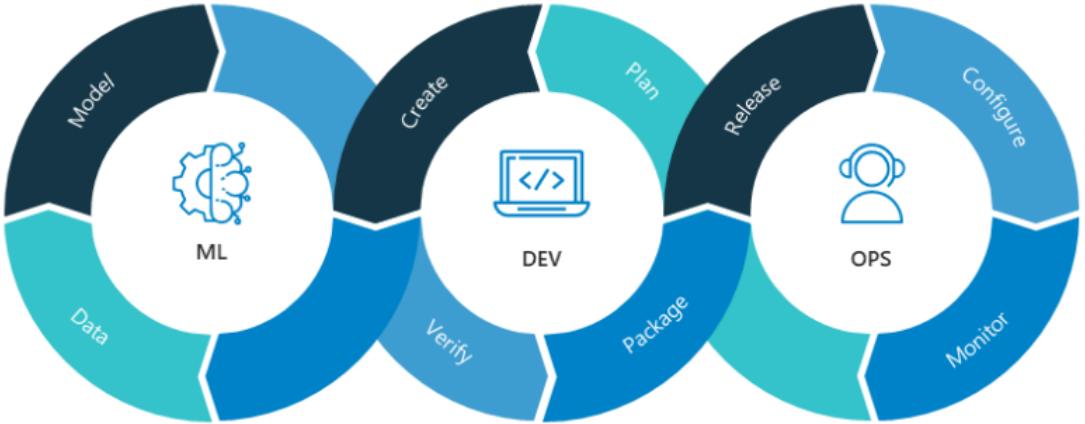
4. Messa in Produzione Strategica: Approcci Shadow, Canary e A/B Testing



# Fase di ingegnerizzazione: MLOps

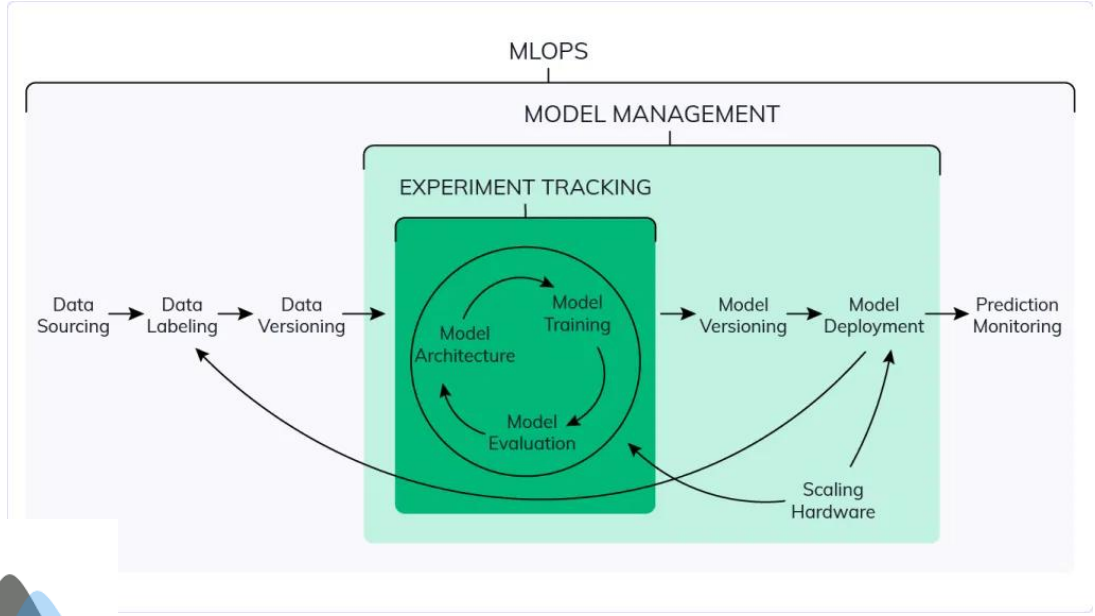
E’ un insieme di cicli - tipico dei sistemi adattivi



# MLOps

MLOps (Machine Learning Operations) si concentra sull'automazione e la gestione dell'intero ciclo di vita del ML, con un'enfasi particolare sulla messa in produzione e sul monitoraggio continuo.




- ingestion dei dati
- pulizia dati
- serving del modello
- monitoring dati
- retraining

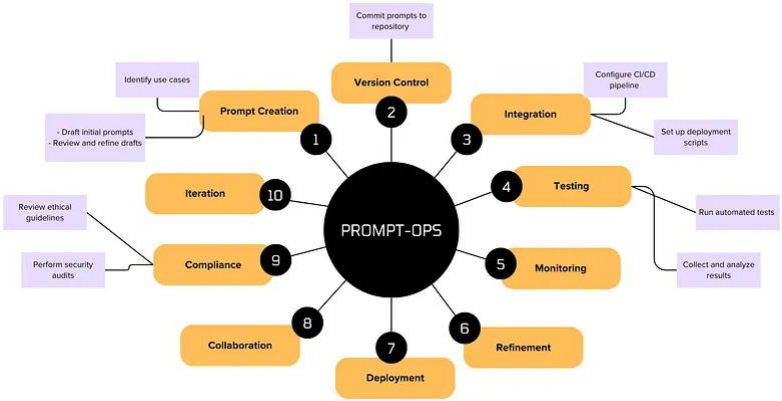
# Fase di ingegnerizzazione: Agenti

1. Promozione e Gestione dei Prompt in Ambienti di Produzione (Prompt Ops)

2. Integrazione degli Agenti con i Sistemi Aziendali: API e Flussi di Dati. Affinché un agente possa eseguire compiti significativi, deve poter interagire con i sistemi aziendali.

3. Affrontare la Stocasticità degli LLM e Garantire l'Affidabilità nei Flussi di Lavoro Agentici. Verifica human-in-the-loop per azioni critiche.

4. Considerazioni CI/CD per i Componenti Agentici



# Ciclo di vita AI ad agenti in produzione

1. Tracciamento del Comportamento dell'Agente, Qualità delle Decisioni ed Efficacia dei Risultati
2. Implementazione di Feedback Loop per il Perfezionamento in Produzione
3. Gestione dell'Apprendimento e dell'Adattamento a Lungo Termine dell'Agente (in genere tramite memoria o RAG)

Obiettivi:

1. Rilevamento, Monitoraggio e Mitigazione dei Bias
2. Garantire Trasparenza, Spiegabilità e Responsabilità
3. Governance dell'IA nelle Operazioni

# Cosa monitorare

**ML:**

**Prestazioni Modello**

**Data Drift**

**Concept Drift (errore di predizione nel tempo)**

**Agentic:**

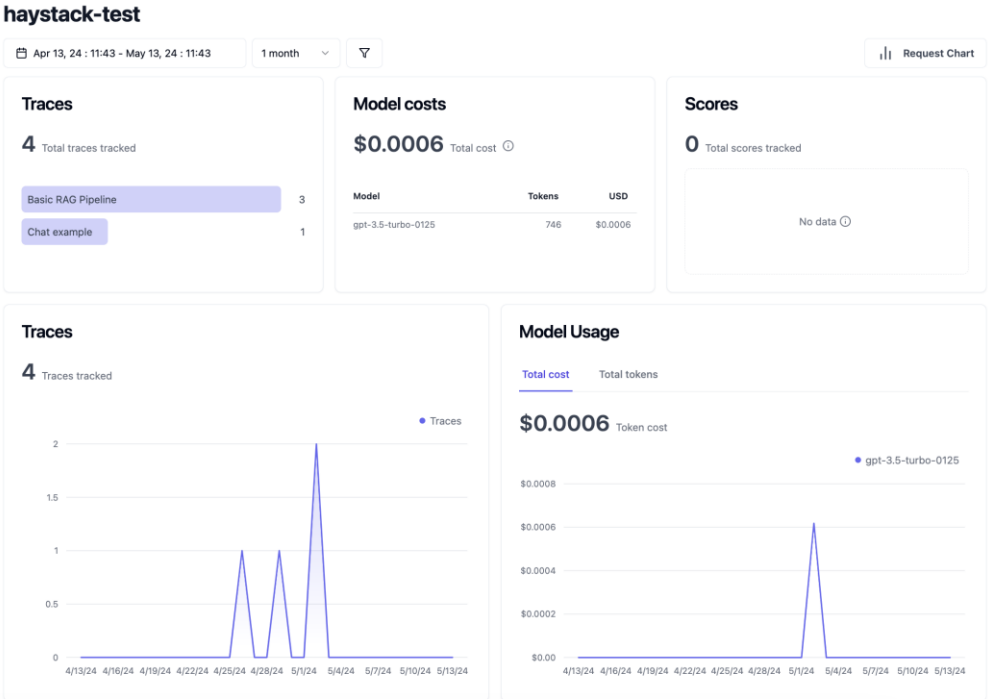
**Aderenza al task**

**Coerenza nelle decisioni prese dagli agenti**

**Aderenza ai prompt** ("jailbreaking" )

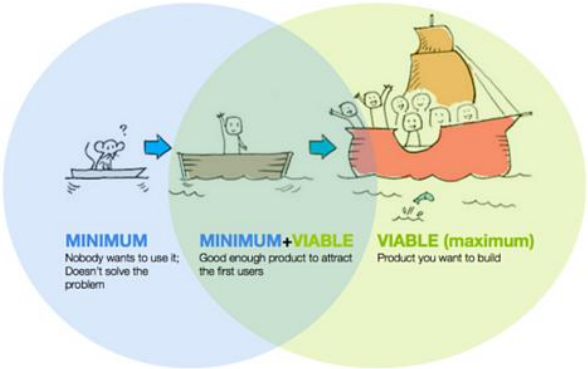
**Bias (KPI specifici) -> tramite LLM (?)**

**Performance e costi**



# Fare delle POC

Data la natura incerta dei progetti AI è fondamentale procedere con POC per valutare i risultati già a valle della fase laboratoriale



# La nostra esperienza