OSINT

Ricerca e analisi da fonti aperte

Ricerca semantica e tecniche di confronto multilingue sui motori di ricerca

Dott. Ing. Manzoni Gabriel Esteban

Laurea Magistrale in Ingegneria Informatica - Politecnico di Milano
"Ambient Intelligence and Data Engineering"

Digital Solutions Architect presso SIGINT srl
Project Manager di DeePoint

CEH (Certified Ethical Hacker) Master - EcCouncil
CIFI (Certified Information Forensics Investigator) – IISFA
CPENT (Certified Penetration Testing Professional) – EcCouncil
LPT (Licensed Penetration Tester) Master - EcCouncil

DeePoint

Piattaforma di analisi **OSINT** (Open Source INTelligence) tramite applicazione di algoritmi di **Intelligenza Artificiale** legati al **NLP** (Natural Language Processing)

Principali funzioni:

- **Analisi** testi con estrazione di principali informazioni
- **Confronto** testi con comparazione semantica
- Analisi **Stilometrica**
- Ricerca **Social Media Network**
- Ricerca **Surface, Deep & Dark Web**
- E molto altro ...

INDICE

INTRODUZIONI AI MOTORI DI RICERCA

STORIA DEI MOTORI DI RICERCA E L'EVOLUZIONE VERSO LA RICERCA SEMANTICA

1990	1994	1994	1995	1996	1998
Archie Il primo motore di ricerca. Creato per indicizzare le directory FTP	WebCrawler Il primo motore di ricerca a indicizzare testi completi.	Lycos Ricerche basate su parole chiave e classificazione dei risultati in base alla pertinenza.	Altavista Introduzione di ricerche avanzate con comandi booleani.	Ask Jeeves Domande in linguaggio naturale invece di utilizzare parole chiave.	Google Introduzione algoritmo PageRank, classificando i risultati per pertinenza e popolarità

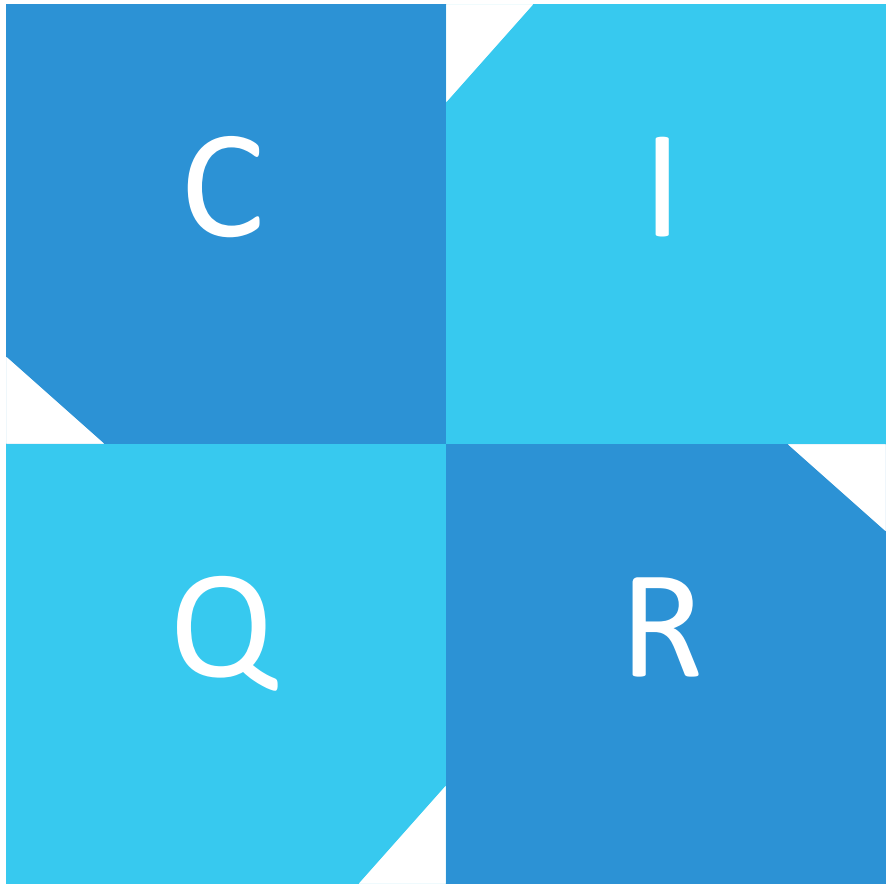
COME FUNZIONA UN MOTORE DI RICERCA?

CRAWLING

Spiders o robots esplorano il web per trovare nuove pagine.

QUERY

L'utente inserisce una query e il motore di ricerca recupera le informazioni più pertinenti dal suo indice.



INDEXING

Le pagine trovate vengono analizzate e archiviate in un database.

RANKING

Algoritmi, come l'algoritmo PageRank di Google, determinano la rilevanza delle pagine indicizzate in base a specifiche query.

LA NECESSITÀ DELLA RICERCA SEMANTICA

Mancanza di precisione: le parole chiavi non sono sufficienti

Mancanza di contesto: necessario identificare il significato della query

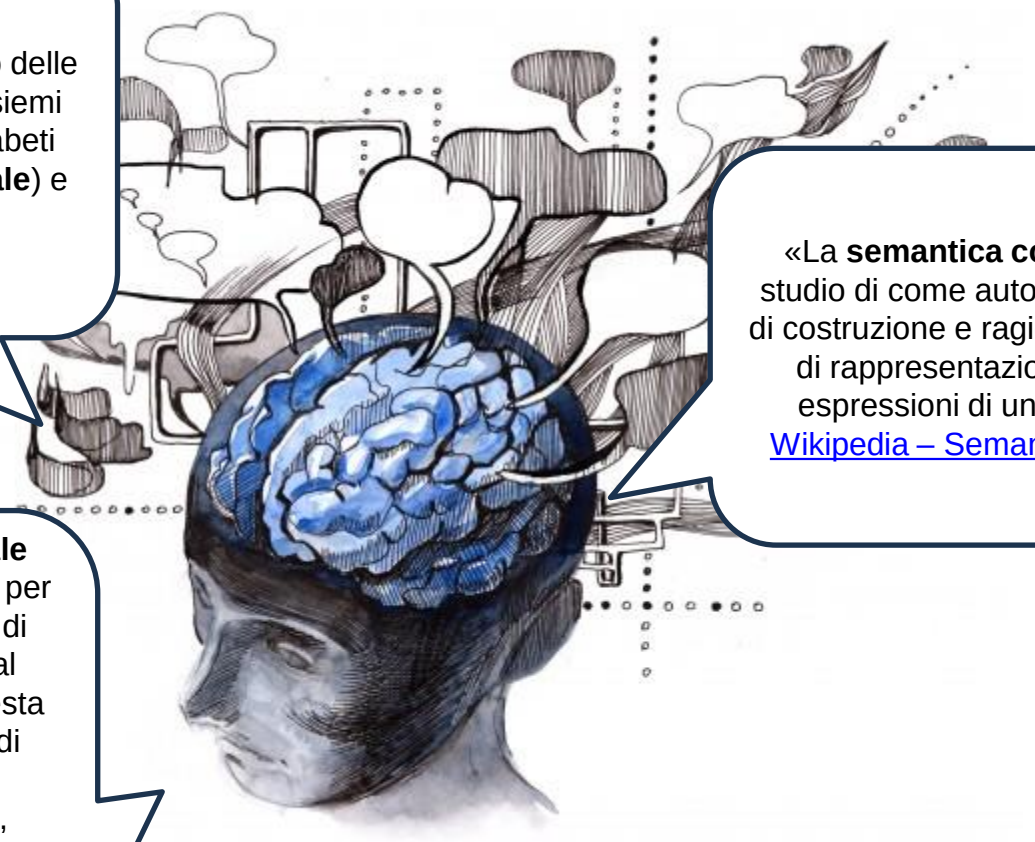
Molteplicità di linguaggi: necessario ricevere come input query in lingue diverse

Ambiguità linguistiche: una stessa parola può avere più significati in lingue diverse

Interrogazioni complesse: le ricerche richieste sono sempre più complesse

COS'È LA RICERCA SEMANTICA?

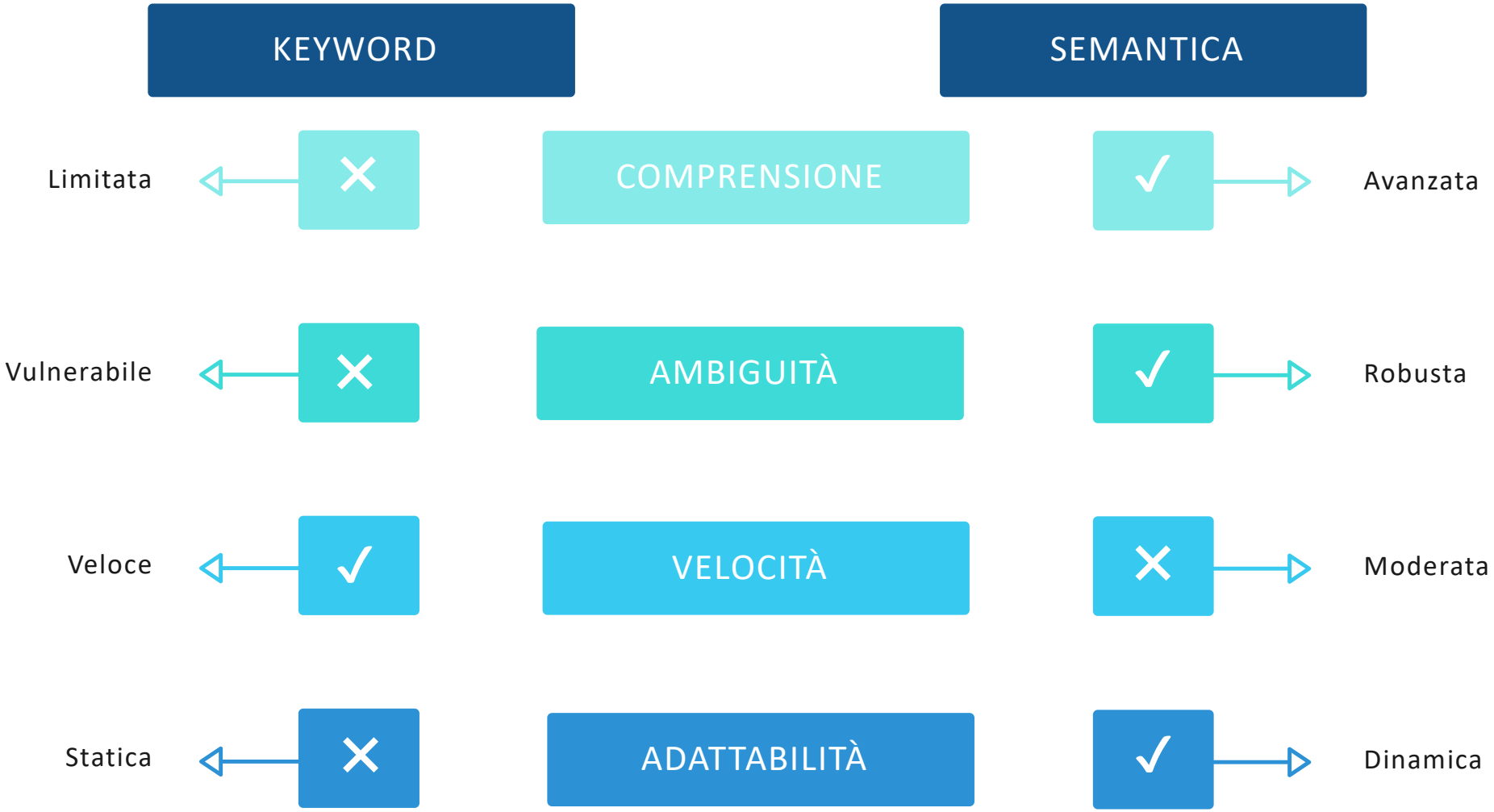
«La **semantica** è quella parte della linguistica che studia il significato delle parole (*semantica lessicale*), degli insiemi delle singole lettere (negli e degli alfabeti antichi) e delle frasi (**semantica frasale**) e dei testi.»
Wikipedia - Semantica



«La **semantica computazionale** è lo studio di come automatizzare il processo di costruzione e ragionamento con l'ausilio di rappresentazioni del significato di espressioni di una lingua naturale.»
Wikipedia – Semantica computazionale

Gli **algoritmi di Intelligenza Artificiale** utilizzano la semantica computazionale per potenziare strumenti e sistemi capaci di comprendere, interpretare e reagire al linguaggio umano in contesti reali. Questa integrazione permette alle macchine di svolgere compiti come la ricerca semantica, la traduzione automatica, l'analisi dei sentimenti e la risposta alle domande, mimando la capacità umana di comprendere il significato e il contesto delle parole e delle frasi.

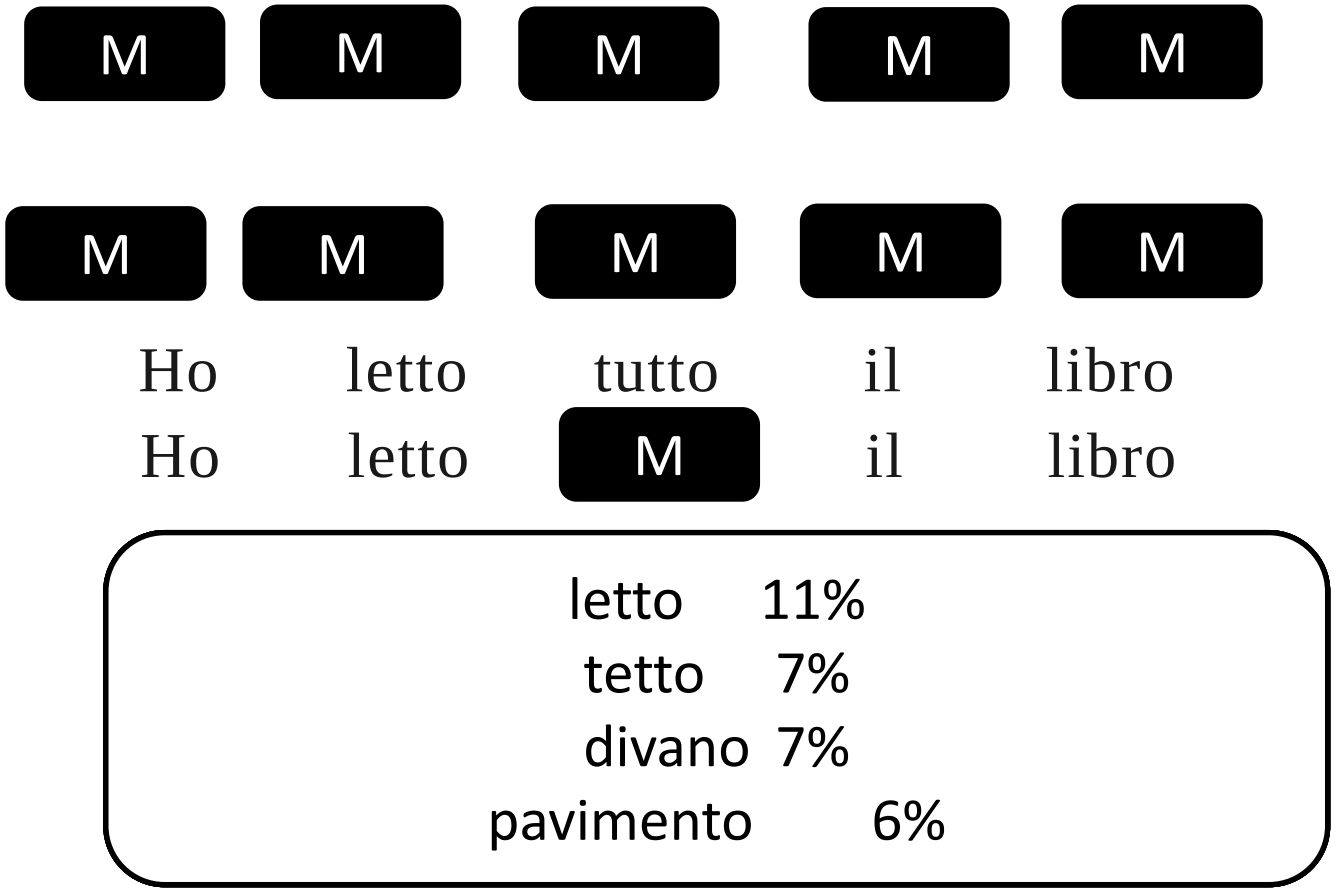
DIFFERENZE RICERCA KEYWORD E SEMANTICA



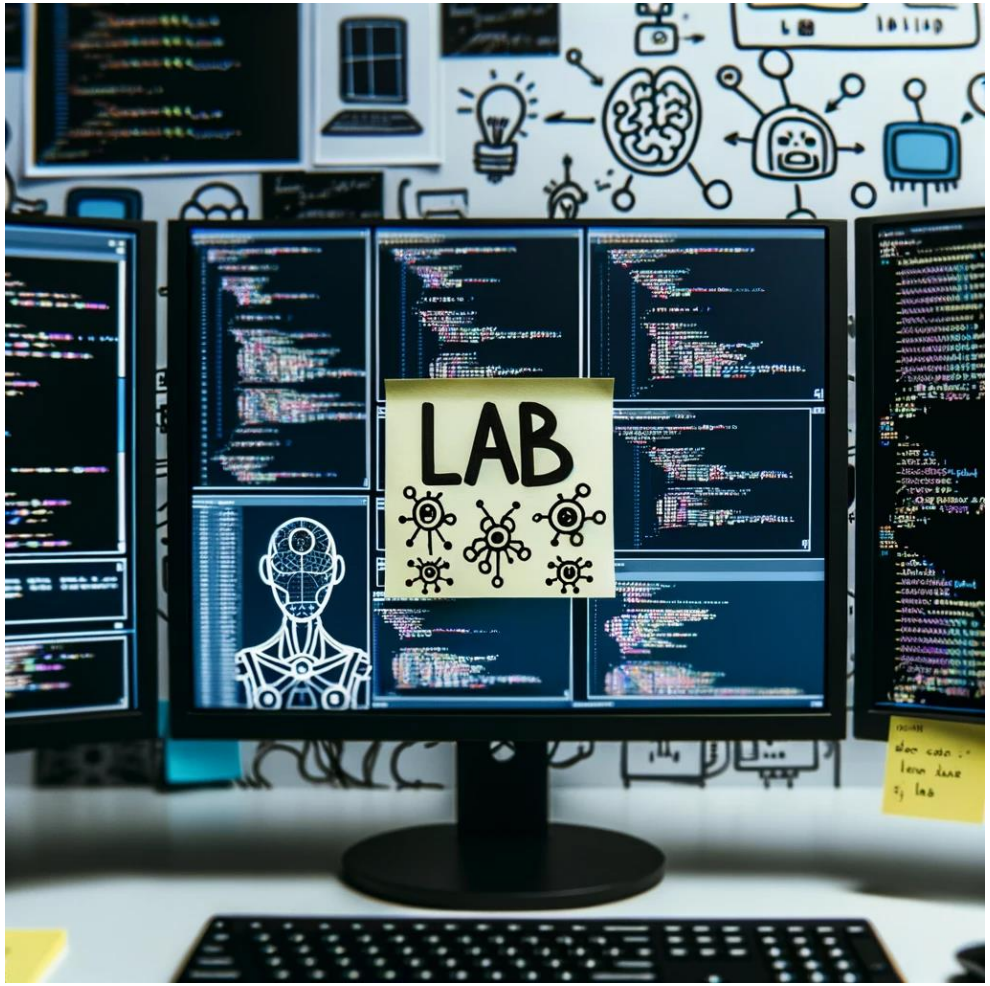
EVOLUZIONE DELLA RAPPRESENTAZIONE SEMANTICA

2013	2014	2018	2018	2020	2022-2023
Word2Vec	GloVe	BERT	GPT	GPT-3	Generativa
Introduce vettori di parole che catturano relazioni semantiche e sintattiche, permettendo operazioni vettoriali basate sul significato.	Un modello di embedding basato sulle statistiche di co-occorrenza tra parole, combinando vantaggi dei metodi conteggio e predittivi.	Rivoluziona il NLP con rappresentazioni bidirezionali profondamente pre-addestrate che considerano le parole nel loro contesto completo, stabilendo nuovi record in vari task.	Utilizza transformers in un modello di generazione di testo con addestramento in due fasi.	Con 175 miliardi di parametri, mostra versatilità senza bisogno di fine-tuning specifico.	ChatGPT, Bard, YouChat, ChatBOT, ecc... IA Generativa che permette nuova flessibilità e adattabilità nelle task.

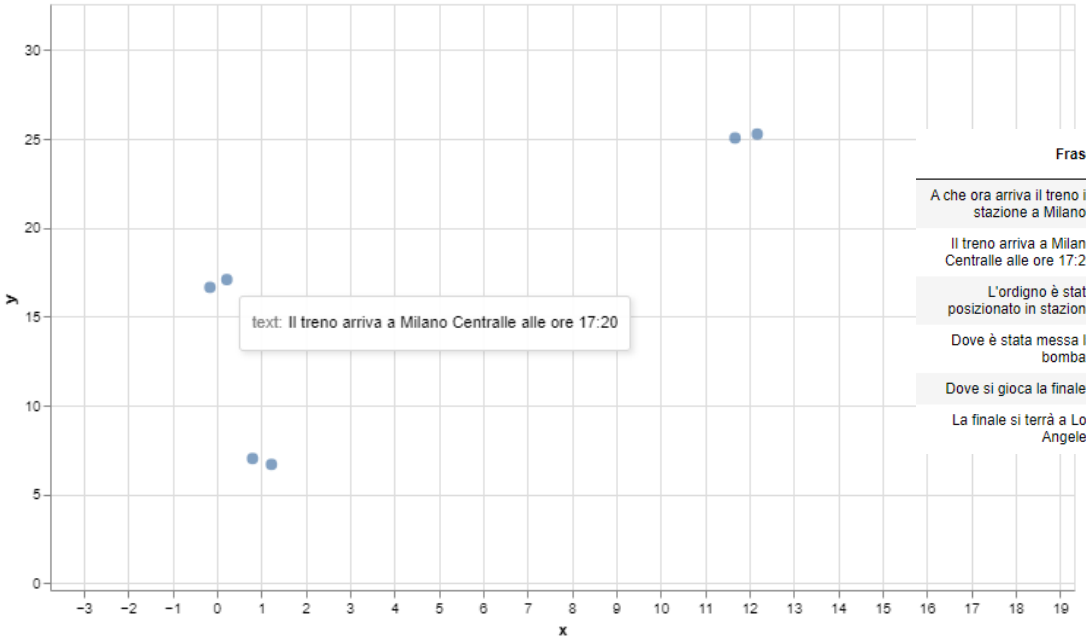
ADDESTRAMENTO BERT



FILL_MASK_EXAMPLE



EMBEDDING PER LA RICERCA SEMANTICA

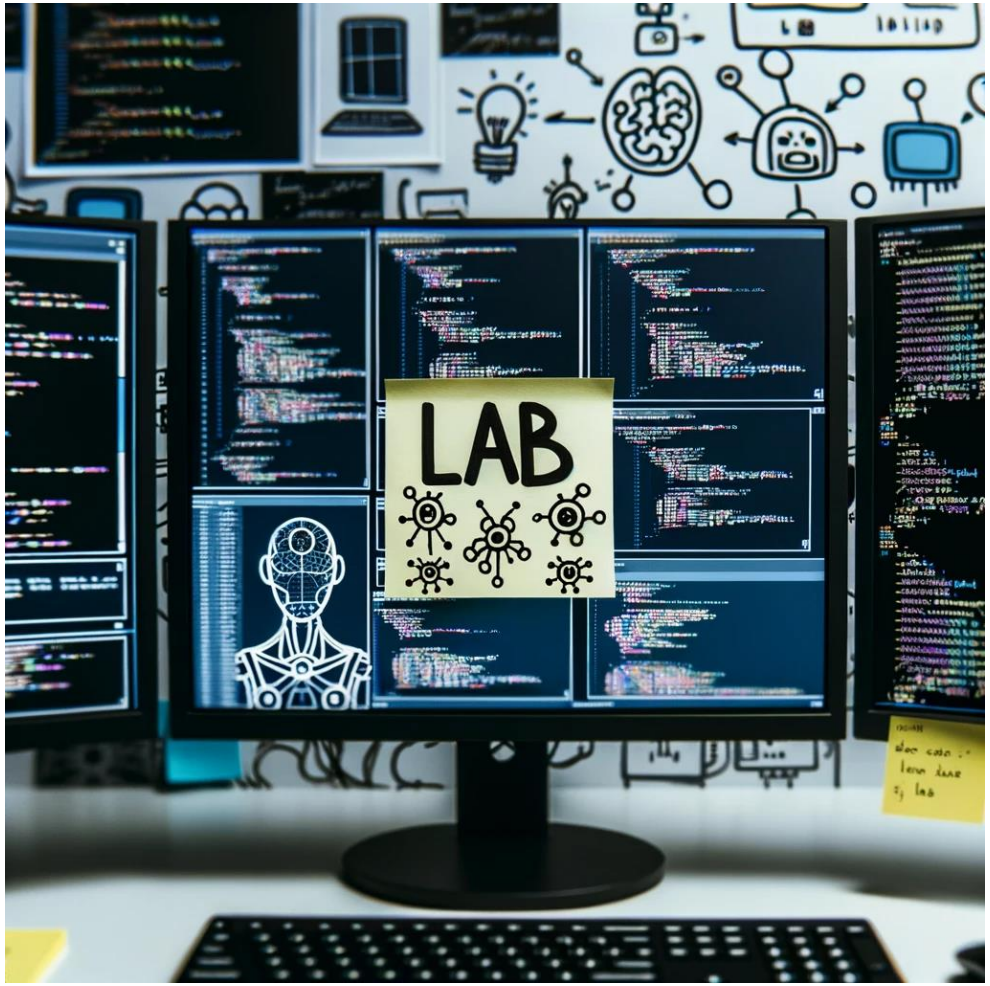


Frase	A che ora arriva il treno in stazione a Milano?	Il treno arriva a Milano Centralle alle ore 17:20	L'ordigno è stato posizionato in stazione	Dove è stata messa la bomba?	Dove si gioca la finale?	La finale si terrà a Los Angeles
A che ora arriva il treno in stazione a Milano?	1.00	0.62	0.40	0.29	0.27	0.22
Il treno arriva a Milano Centralle alle ore 17:20	0.62	1.00	0.29	0.14	0.19	0.27
L'ordigno è stato posizionato in stazione	0.40	0.29	1.00	0.57	0.11	0.15
Dove è stata messa la bomba?	0.29	0.14	0.57	1.00	0.32	0.13
Dove si gioca la finale?	0.27	0.19	0.11	0.32	1.00	0.58
La finale si terrà a Los Angeles	0.22	0.27	0.15	0.13	0.58	1.00

embedding.

o ranking per similitudine

SEMANTIC_COMPARISON



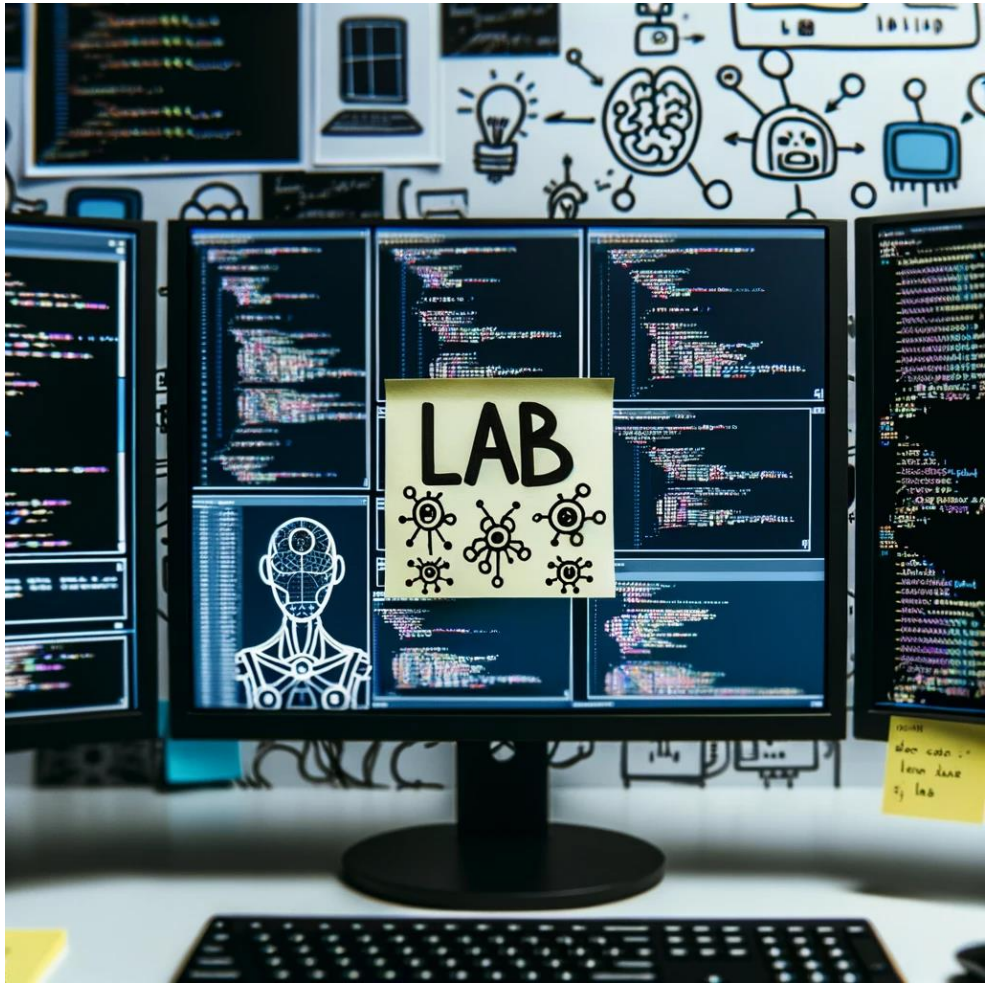
HUB DI MODELLI DI NLP

1. HuggingFace https://huggingface.co/
2. TensorFlow Hub https://www.tensorflow.org/hub
3. Spacy https://spacy.io/
4. Stanford NLP Group https://nlp.stanford.edu/
5. Facebook’s Fairseq https://ai.meta.com/tools/fairseq/
6. Google Research https://research.google/
7. OpenAI https://openai.com/

DATI ESTRATTI IN NLP

1. Named Entity Recognition (NER): entità suddivise in categorie predefinite come persone, luoghi , organizzazione, ecc…
2. Keyword: parole chiave le quali possono essere parole o frasi più rilevanti
3. Part-Of-Speech (POS): classifica delle parole in categorie grammaticali
4. Lemmatizzazione: estrazione della parola nella forma base (bellissima->bello)
5. Estrazione di relazioni: relazioni quali per esempio soggetto->azione->oggetto
6. Coerenza: estrazione di frasi che riprendono lo stesso concetto o l'opposto

DATA_EXTRACTION



Pausa di 10 minuti

La lezione riprenderà alle 17:30

TIPI DI MODELLI DI NLP

1. Text Classification:
2. Question Answer ntesto
3. Translation: tradu
4. Summarization: c
5. Text Generation:
6. Sentence Similar edding
7. Conversational: c
8. Fill-Mask: indovin

Funzione base di tutte le altre, utile direttamente per correzioni grammaticali, suggerimenti, identificazione di messaggi in codice

MODALITÀ DI APPRENDIMENTO DEI MODELLI DI NLP

1. Supervised Learning: vengono forniti un numero elevato di esempi «etichettati» per addestrare la rete neurale da zero
2. Fine-tuning Learning: vengono forniti un ristretto numero di esempi «etichettati» per affinare una rete neurale pre-addestrata
3. Unsupervised Learning: viene fornito testo «non etichettato» con l'obiettivo che l'algoritmo apprenda automaticamente strutture o modelli senza supervisione
4. Semi-Supervised Learning: si adotta una approccio ibrido tra Supervised e Unsupervised
5. Reinforcement learning: dopo l'addestramento di una rete neurale vengono proposti ulteriori esempi con risposta al fine di aumentarne l'accuratezza
6. Transfer Learning: le capacità apprese da una rete neurale vengono trasferite verso un'altra rete neurale per facilitarne l'apprendimentoo
7. Zero-Shot Learning: viene proposta una nuova task su cui la rete neurale non è stata addestrata
8. One or Few Shot Learning: vengono proposti solo uno o pochi esempi da cui la rete neurale deve imparare

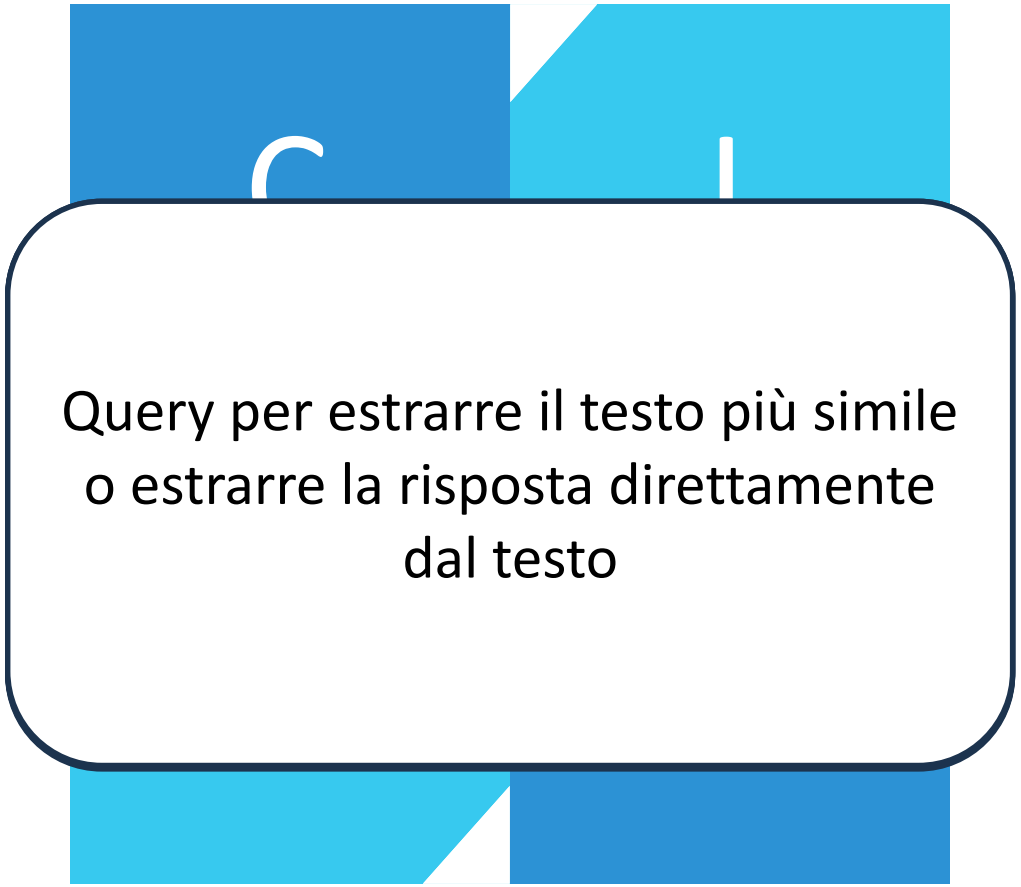
COME FUNZIONA UN MOTORE DI RICERCA?

CRAWLING

Spiders o robots esplorano il web per trovare nuove pagine.

QUERY

L'utente inserisce una query e il motore di ricerca recupera le informazioni più pertinenti dal suo indice.



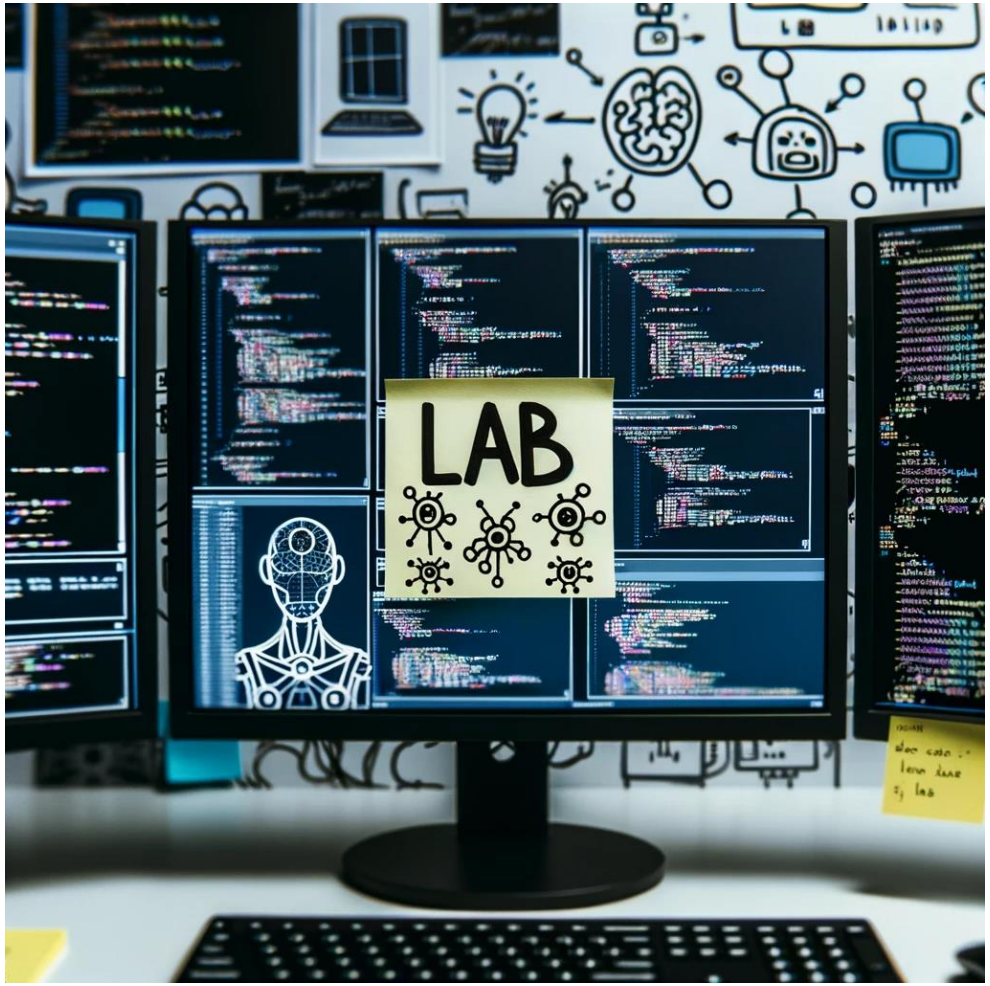
INDEXING

Le pagine trovate vengono analizzate e archiviate in un database.

RANKING

Algoritmi, come l'algoritmo PageRank di Google, determinano la rilevanza delle pagine indicizzate in base a specifiche query.

EXAMPLE_SEARCH_ENGINE



COSTRUIRE MOTORE DI RICERCA ON-PREMISE

Indicizzazione

Salvataggio informazioni su Database nel formato desiderato

Query

Dare possibilità all'utente di porre domande in linguaggio naturale e/o tramite filtri

Dati

Ricerca informazioni tramite OSINT, SOCMINT, Terze parti o propri dati

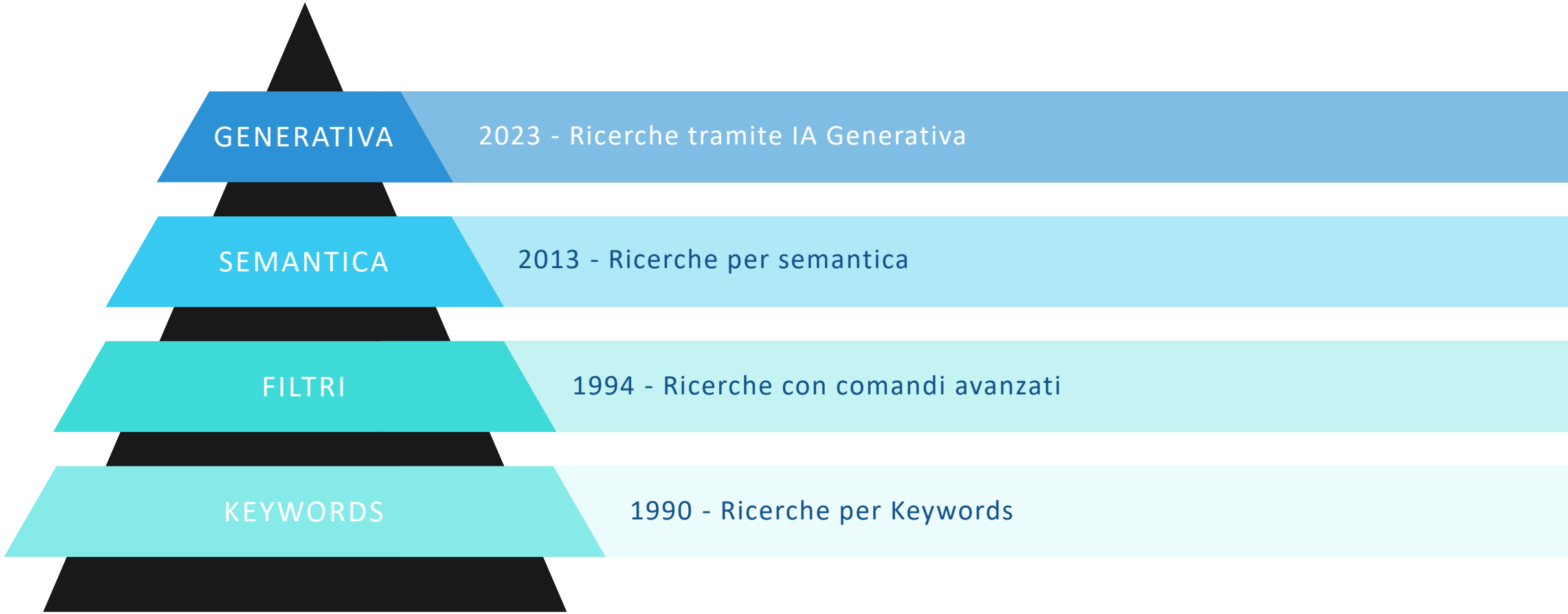
Categorizzare

Categorizzare le informazioni per contenuti, quali classi, sentiment, keywords, entità o altro

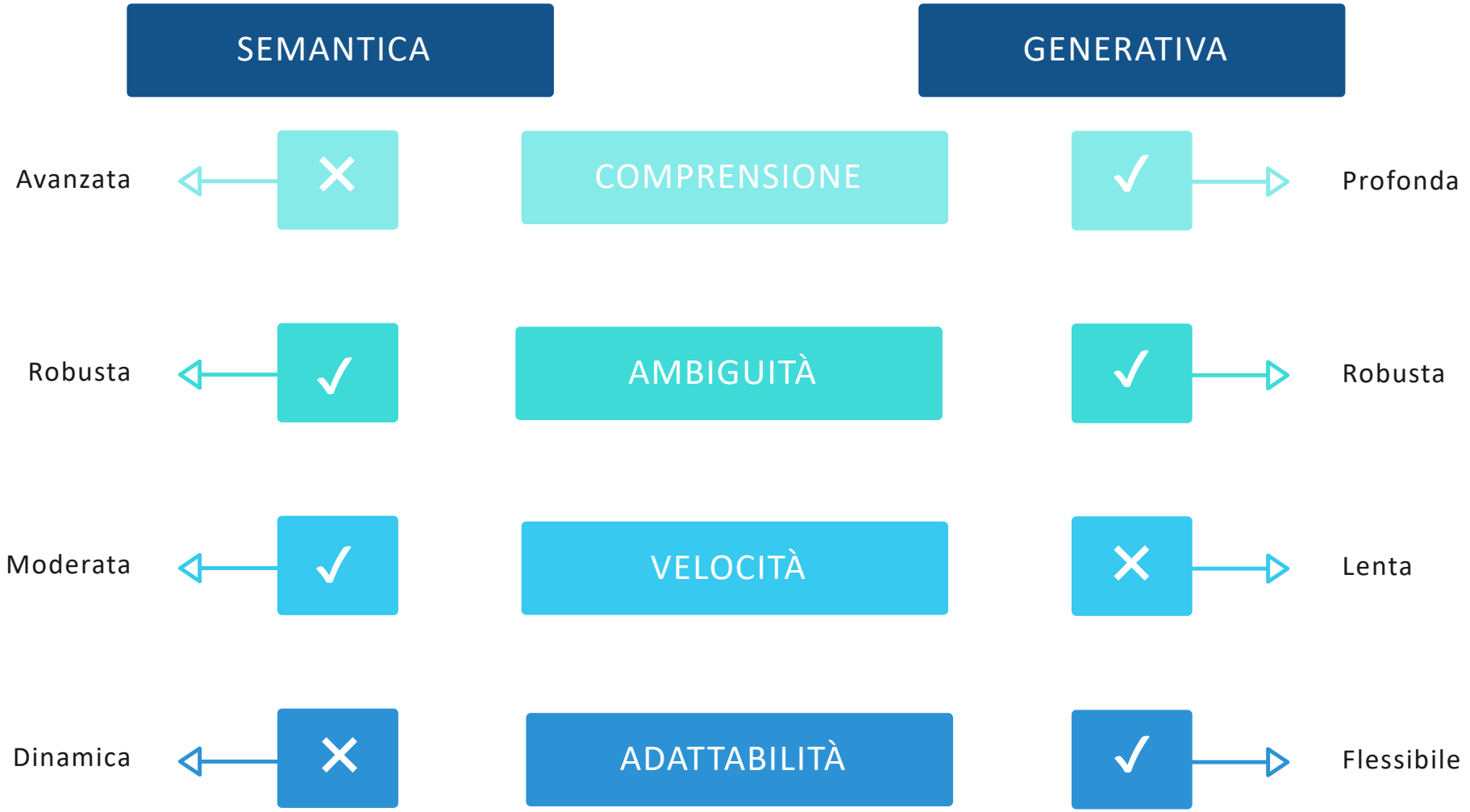
Alert

Creazione di alert automatici a seconda dei filtri d'interesse degli utenti

MODELLI DI RICERCA



DIFFERENZE RICERCA SEMANTICA E IA GENERATIVA



PRODOTTI DI IA GENERATIVA

1. OpenAI
2. Google
3. Microsoft
4. HuggingFace
5. GitHub

Dedicato a scrivere codice
Utilizza GPT-4

nai.com/
l.google.com/
w.bing.com/
gingface.co/chat/
ub.com/features/copilot

MODELLI DI IA GENERATIVA

1. GPT-3.5 & GPT-4	OpenAI (ChatGPT)	Proprietario
2. Palm	Google (BARD)	Open-Source(*)
3. Llama	Meta	Open-Source
4. Falcon	HuggingFace	Open-Source
5. StableLM	Stability AI	Open-Source
6. Alpaca	Standford	Open-Source
7. MPT	MosaicML	Open-Source

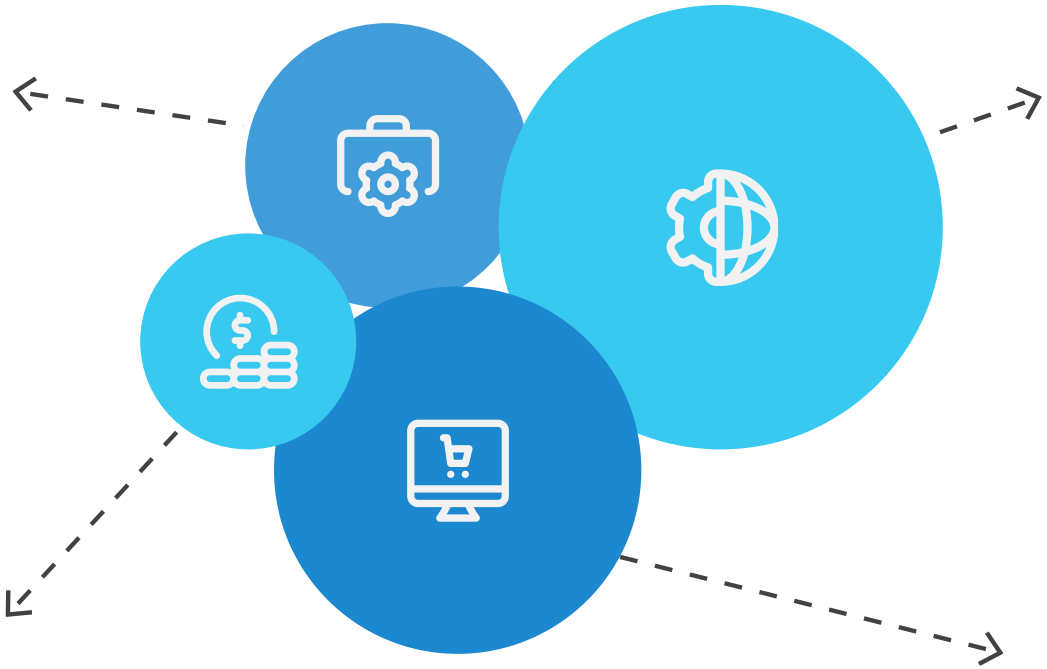
SFIDE FUTURE

Ethic AI

Maggiore attenzione su questioni etiche e morali dell'AI tramite applicazione e rispetto dell'EU AI Act

Riduzione Allucinazioni

Le AI Generative sono ancora vulnerabili alle allucinazioni. La priorità per il futuro è arrivare a ridurle drasticamente



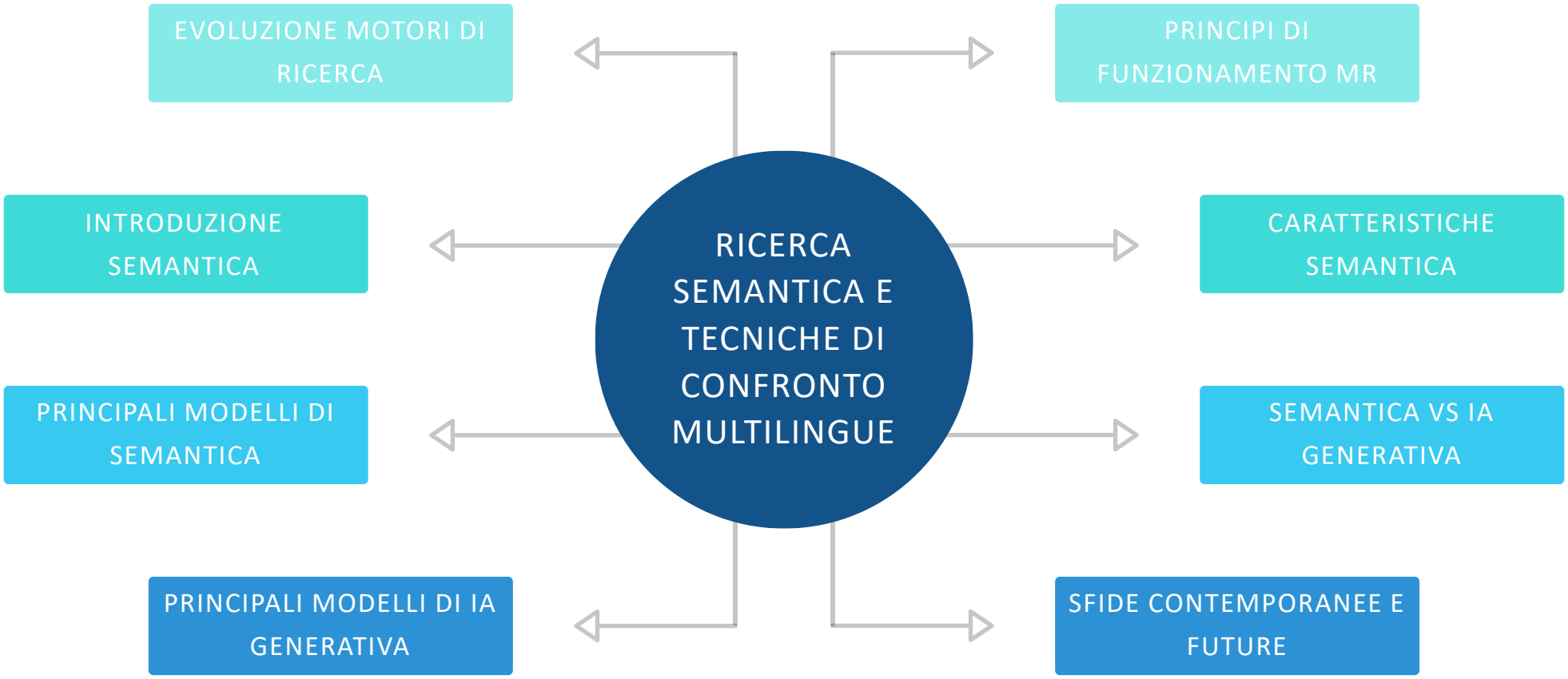
Scalabilità

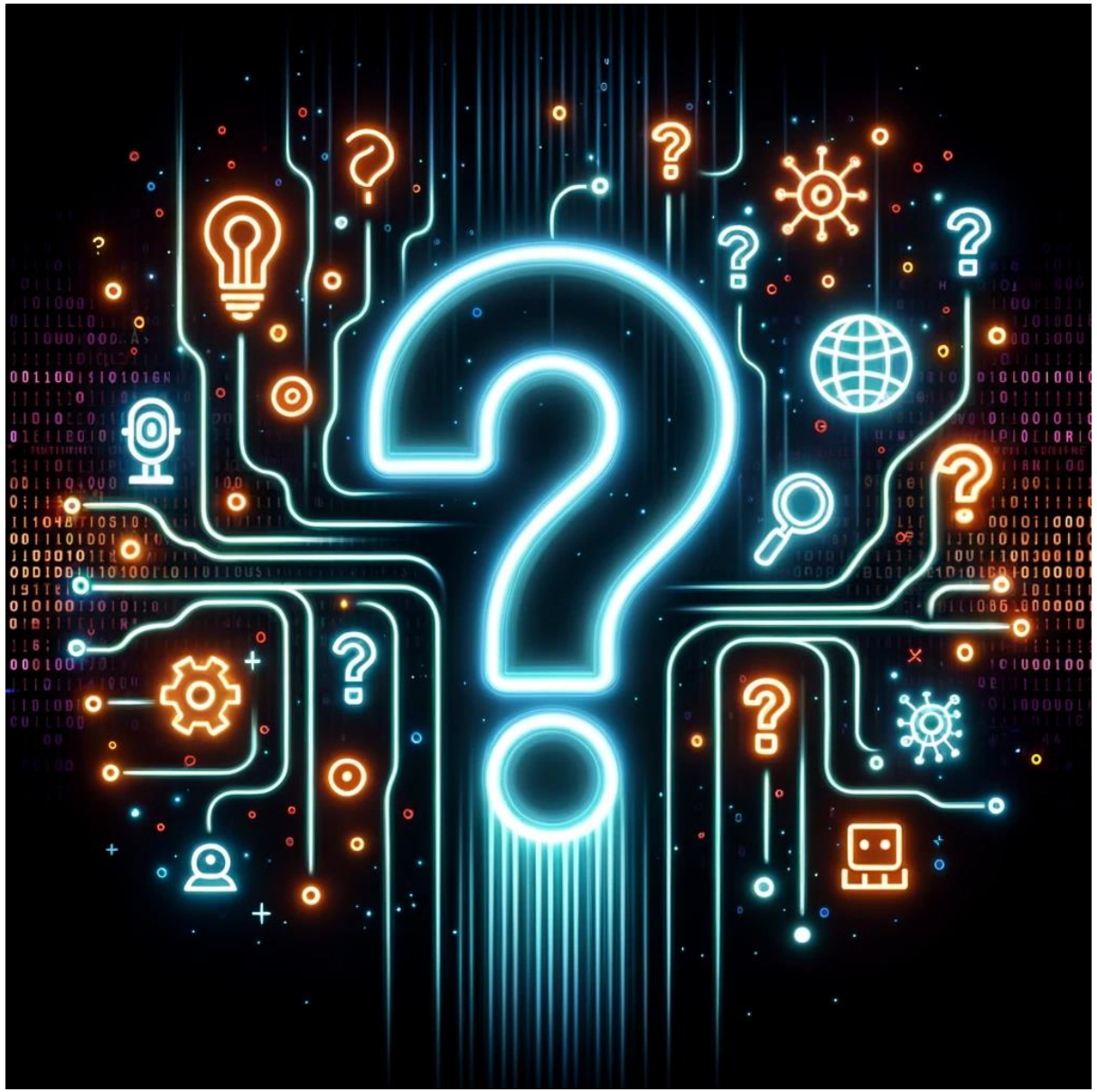
Necessario cambiare il trend in corso, ovvero incrementare dimensione del dataset necessario e della potenza di calcolo quanto piuttosto migliorare l'apprendimento su dataset più piccoli e ridurre la potenza di calcolo necessaria

Slow-Thinking

Le attuali AI Generative sono forti per la parte di Fast-Thinking ma deboli in quella di Slow-Thinking. Per ottenere maggiore qualità in task complesse occorre cambiare questo aspetto

RIEPILOGO


EVOLUZIONE MOTORI DI RICERCA
PRINCIPI DI FUNZIONAMENTO MR
INTRODUZIONE SEMANTICA
RICERCA SEMANTICA E TECNICHE DI CONFRONTO MULTILINGUE
CARATTERISTICHE SEMANTICA
PRINCIPALI MODELLI DI SEMANTICA
SEMANTICA VS IA GENERATIVA
PRINCIPALI MODELLI DI IA GENERATIVA
SFIDE CONTEMPORANEE E FUTURE



Gabriel Esteban Manzoni

g.manzoni@sigintsrl.it

https://github.com/gmanzoni/POL_OSINT_Semantica