Istituto di Linguistica Computazionale
"Antonio Zampolli"
Consiglio Nazionale delle Ricerche

# Trattamento Automatico del Linguaggio nell'era dell'IA

Relatori:
Dott. Felice Dell'Orletta
Dott. Alessio Miaschi

*3° Seminario sull'Innovazione Digitale - Intelligenza Artificiale: scenari, applicazioni e trasformazioni - I.S.M.A. / RE.FO.DI.MA.*

Consiglio Nazionale delle Ricerche | Area Territoriale di Ricerca di Pisa
CNR | Dipartimento Scienze Umane e Sociali Patrimonio culturale

# Istituto di Linguistica Computazionale “A. Zampolli”

Fondato nel 1980, centro di riferimento, a livello nazionale e internazionale, nel settore della Linguistica Computazionale. CNR-ILC svolge attività di ricerca, formazione e trasferimento tecnologico, unitamente a rilevanti attività editoriali.

Le principali aree di competenza sono:

- Trattamento Automatico del Testo e Filologia Computazionale
- Trattamento Automatico del Linguaggio Naturale (*ItaliaNLP Lab*)
- Risorse Linguistiche, Standard e Infrastrutture
- Modelli Computazionali dell'Uso Linguistico

Il *Natural Language Processing* (NLP) è un settore dell’intelligenza artificiale che studia come permettere ai computer di comprendere, interpretare e generare il linguaggio umano. È un ambito interdisciplinare che combina la linguistica, l'informatica e le scienze cognitive.

# ItaliaNLP Lab: il gruppo

### Research Director

- **Simonetta Montemagni**, *Linguistics*

### Permanent Researchers

- **Giulia Benotto**, *Digital Humanities*
- **Dominique Brunato**, *Linguistics*
- **Franco Alberto Cardillo**, *Computer Science*
- **Felice Dell'Orletta** (head of the laboratory), *Computer Science*
- **Giulia Venturi**, *Linguistics*

### Temporary Researchers

- **Chiara Alzetta**, *Digital Humanities*
- **Alessio Miaschi**, *Digital Humanities*

### Research Fellow

- **Chiara Fazzone**, *Digital Humanities*

### Affiliated Researchers

- **Luca Bacco**, *Postdocs, Campus Bio-Medico, University of Rome*
- **Mario Merone**, *Assistant Professor, Campus Bio-Medico, University of Rome*

### Ph.D. Students

- **Agnese Bonfigli**, *PhD programme in "Bioengineering, Applied Sciences, and Intelligent Systems", Università Campus Bio-Medico di Roma*
- **Cristiano Ciaccio**, *PhD programme in Computer Science, University of Pisa*
- **Luca Dini**, *National PhD programme in Artificial Intelligence – "AI & Society", Department of Computer Science, University of Pisa*
- **Lucia Domenichelli**, *National PhD programme in Artificial Intelligence – "AI & Society", Department of Computer Science, University of Pisa*
- **Michele Papucci**, *PhD programme in Computer Science, University of Pisa*
- **Ruben Piperno**, *PhD programme in "Bioengineering, Applied Sciences, and Intelligent Systems", Università Campus Bio-Medico di Roma*
- **Marta Sartor**, *PhD programme in Digital Humanities, University of Genova*

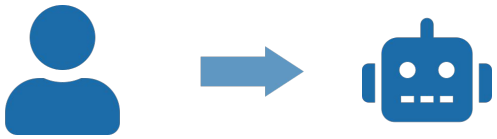
### Master Student

- **Gaia Federica Francesca Ferrara**, *Digital Humanities*
- **Paolo Labruna**, *Digital Humanities*
- **Francesco Longobardi**, *Digital Humanities*
- **Camilla Maffeis**, *Digital Humanities*

*www.italianlp.it*

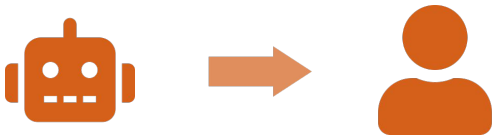
# ItaliaNLP Lab: le aree di ricerca

## Natural Language Understanding (NLU)



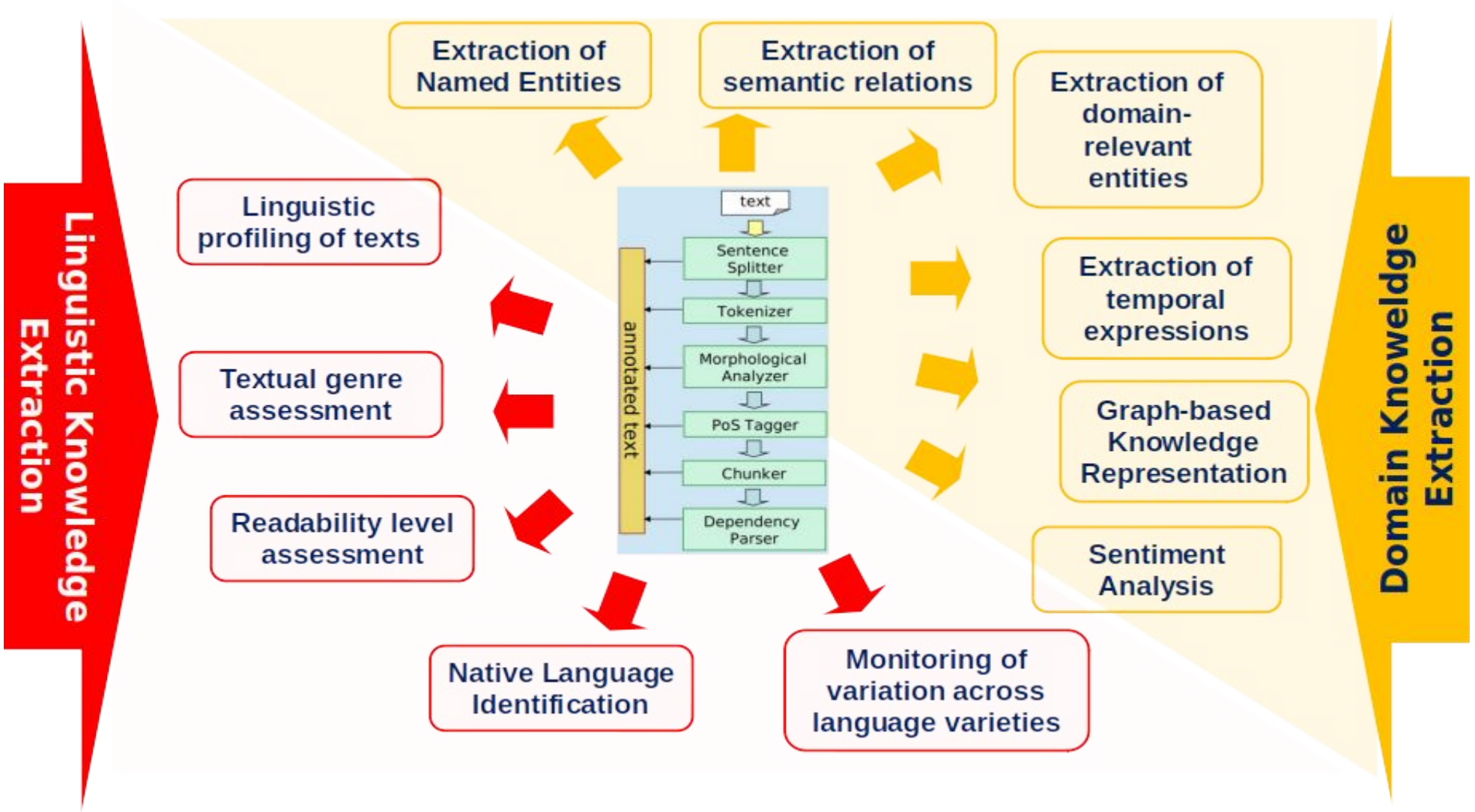
- Comprensione del Contenuto (*Information Extraction*):
  - Estrazione Terminologica
  - Estrazione di Entità Nominate
  - Estrazioni di Relazioni e Eventi
  - …
- Comprensione delle Forma (*Linguistic Profiling*):
  - Valutazione della Leggibilità
  - Tracciare l'evoluzione delle competenze linguistiche
  - Modellare la complessità linguistica
  - …
- Valutazione e Interpretazione

## Natural Language Generation (NLG)



- Generazione di Testo Libero
  - appartenente a vari generi (romanzo, poesia, conversazione libera, ecc.)
- Testo ⟶ Testo
  - Semplificazione del testo, riassunto del testo, trasferimento di stile, ecc.
- Immagine ⟶ Testo
  - Generazione di didascalie, generazione di referti medici, ecc.
- Valutazione e Interpretazione

# ItaliaNLP Lab: Natural Language Understanding

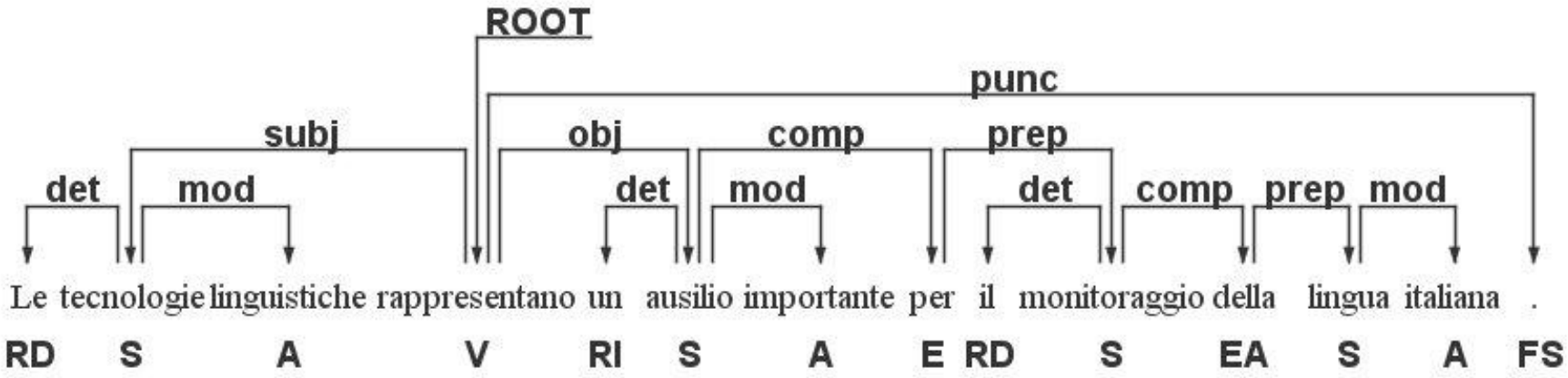


# ItaliaNLP Lab: Natural Language Understanding

**Dependency-based syntactic annotation**

Italian (CoNLL): 92% LAS

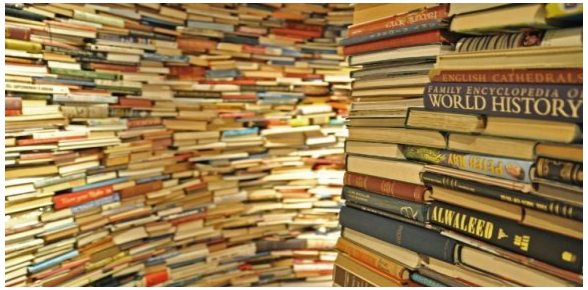
English (CoNLL): 87.89% LAS




**Morpho-syntactic annotation**

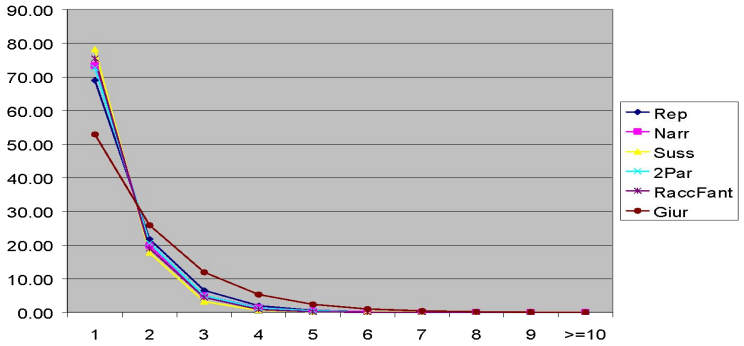
Italian (Evalita) and English (CONLL): accuracy ~98%

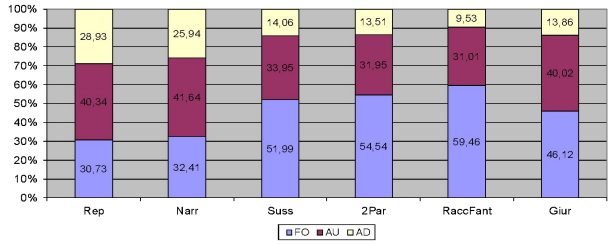
# NLP: Dal testo alla Conoscenza Linguistica

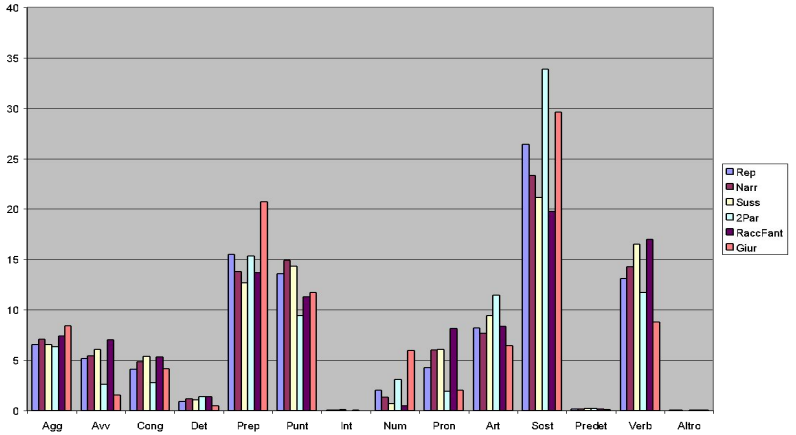



- Scientific articles
- Patents
- Technical documents
- Social Media
- Web pages
- Blogs
- …

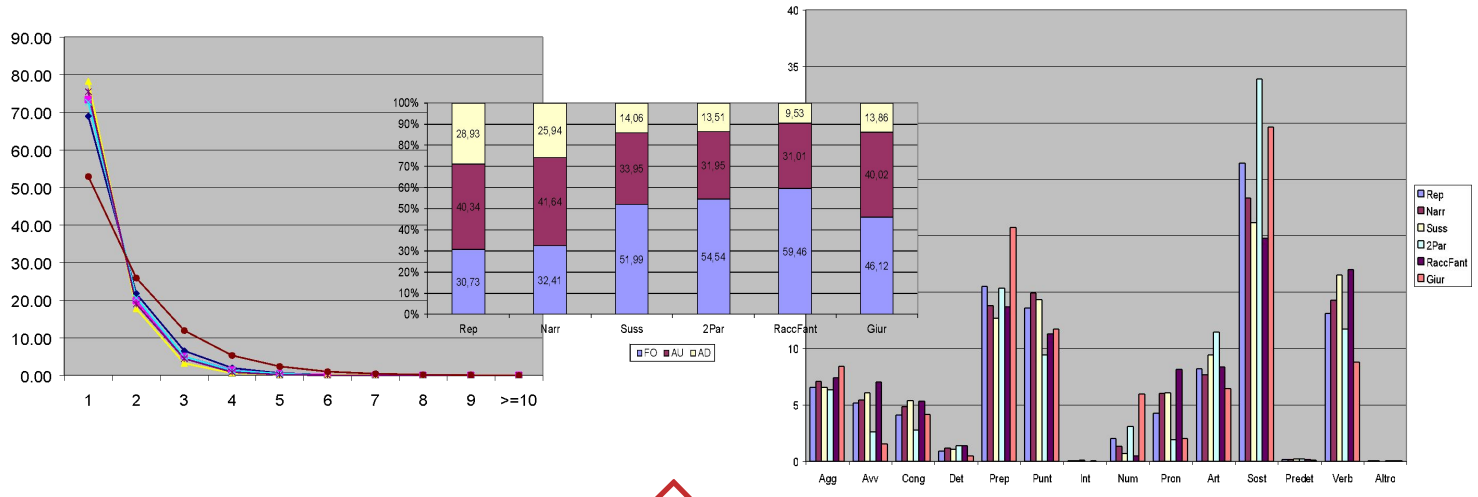
**NLP Technologies**



# Conoscenza Linguistica per




***How is the text written?***
**Text profiling**

***Who wrote the text?***
**Author profiling**

# NLP: dal testo alla conoscenza di dominio


**Persone**

Giotto
Cimabue
Giovanni Pisano
Simone Martini
Arnolfo di Cambio
Ambrogio Lorenzetti
Cennino Cennini
Dante
Duccio di Buoninsegna
...

**Entità Geopolitiche**

Roma
Italia
Assisi
Siena
Firenze
Pisa
Padova
Italia settentrionale
Saint-Denis
…

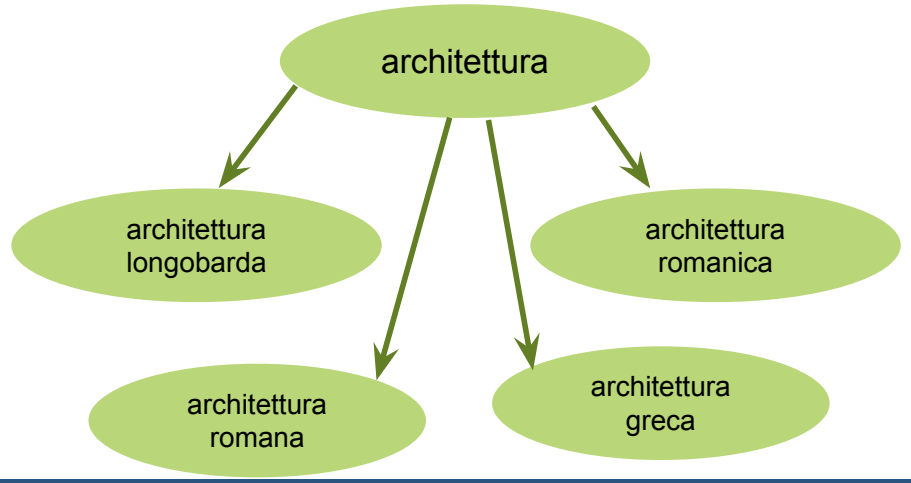
**Organizzazioni**

Sacro Romano Impero
Metropolitan Museum
Musée de Cluny
Collezione Salini
Museo Provinciale
…

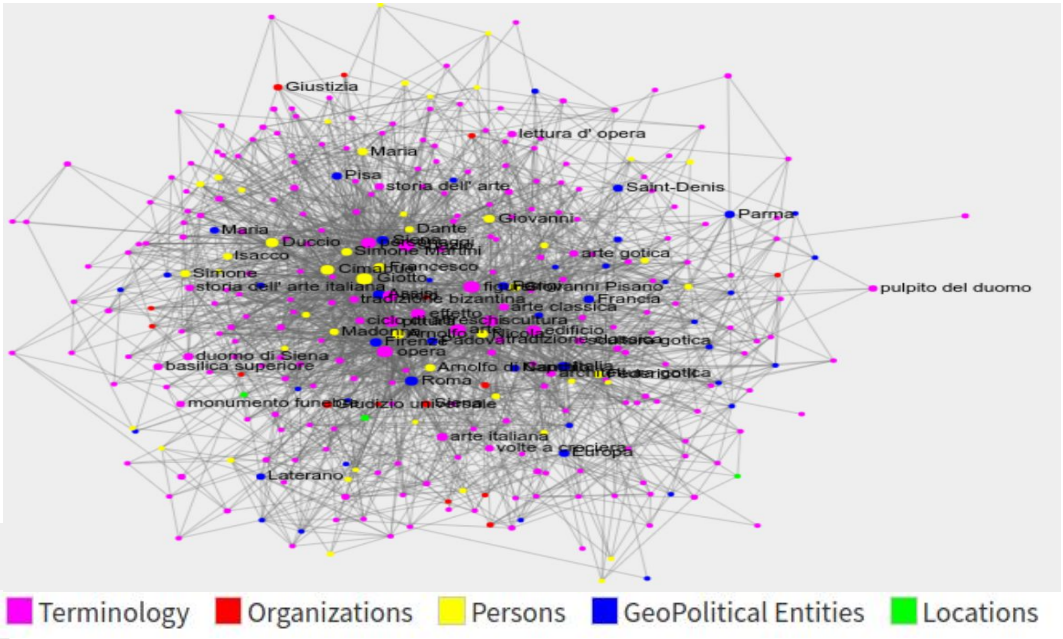
**Entità Dominio--specifiche**

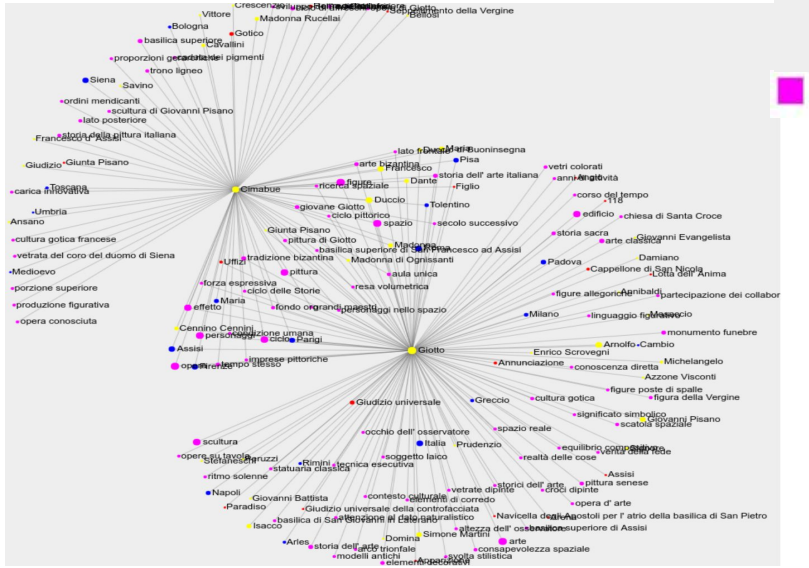
edificio
affresco
città
duomo di Siena
arte italiana
colonne
Giudizio Universale
storie di San Francesco
arte classica
**architettura**
gotico internazionale
ciclo di affreschi
pulpito del duomo
volte a crociera
tradizione bizantina
basilica superiore

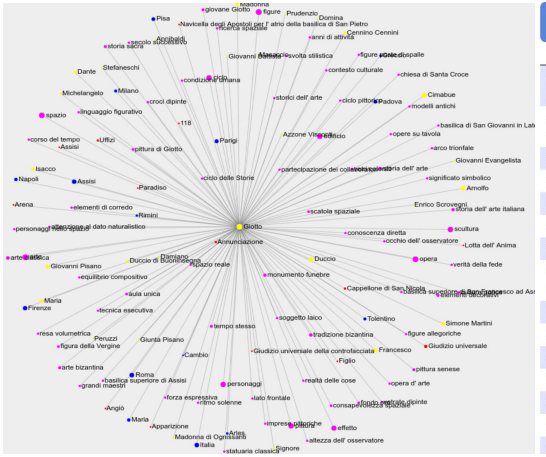
**Organizzazione Tassonomica**



# NLP: dal testo alla conoscenza di dominio

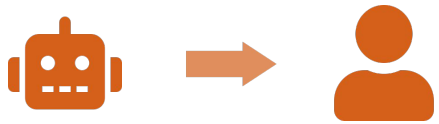


Giotto

| Node | Weight |
|---|---|
| Cimabue PER | 14.0 |
| personaggi TERM | 13.0 |
| spazio TERM | 12.0 |
| Padova GPE | 9.0 |
| Firenze GPE | 8.0 |
| ciclo TERM | 8.0 |
| Roma GPE | 8.0 |
| Francesco PER | 8.0 |
| arte TERM | 7.0 |
| Assisi GPE | 7.0 |
| pittura TERM | 7.0 |
| Napoli GPE | 6.0 |
| effetto TERM | 6.0 |
| figure TERM | 6.0 |
| opera TERM | 4.0 |

# Natural Language Generation

**Natural Language Generation (NLG)**



- Generazione di Testo Libero
  - appartenente a vari generi (romanzo, poesia, conversazione libera, ecc.)
- Testo → Testo
  - Semplificazione del testo, riassunto del testo, trasferimento di stile, ecc.
- Immagine → Testo
  - Generazione di didascalie, generazione di referti medici, ecc.
- Valutazione e Interpretazione

Il 30 novembre 2022 OpenAI presenta ChatGPT e le macchine iniziano a “parlare”. Con l’uscita di GPT-4 (marzo 2023) i compiti di *Natural Language Generation* diventano task affrontabili con prestazioni quasi umane. Ma come ci siamo arrivati?

# I primi sistemi a regole

- Fino alla metà degli anni ‘80, gli algoritmi utilizzati per lo sviluppo di sistemi per il trattamento automatico della lingua sfruttavano modelli della lingua formati da insiemi complessi di regole scritte a mano, definite attraverso la conoscenza del problema linguistico da modellare (detti anche modelli simbolici). Gli algoritmi “semplicemente” identificavano le regole da applicare in un dato contesto ed eseguivano la specifica azione

- I problemi principali rispetto a questo tipo di approccio:
  - la difficoltà nel definire un insieme di regole per compiti complessi
  - la difficoltà nel gestire input malformato
  - non esiste una descrizione completa (attraverso le regole) della maggior parte dei compiti linguistici nel contesto del trattamento automatico del linguaggio

# Il machine learning

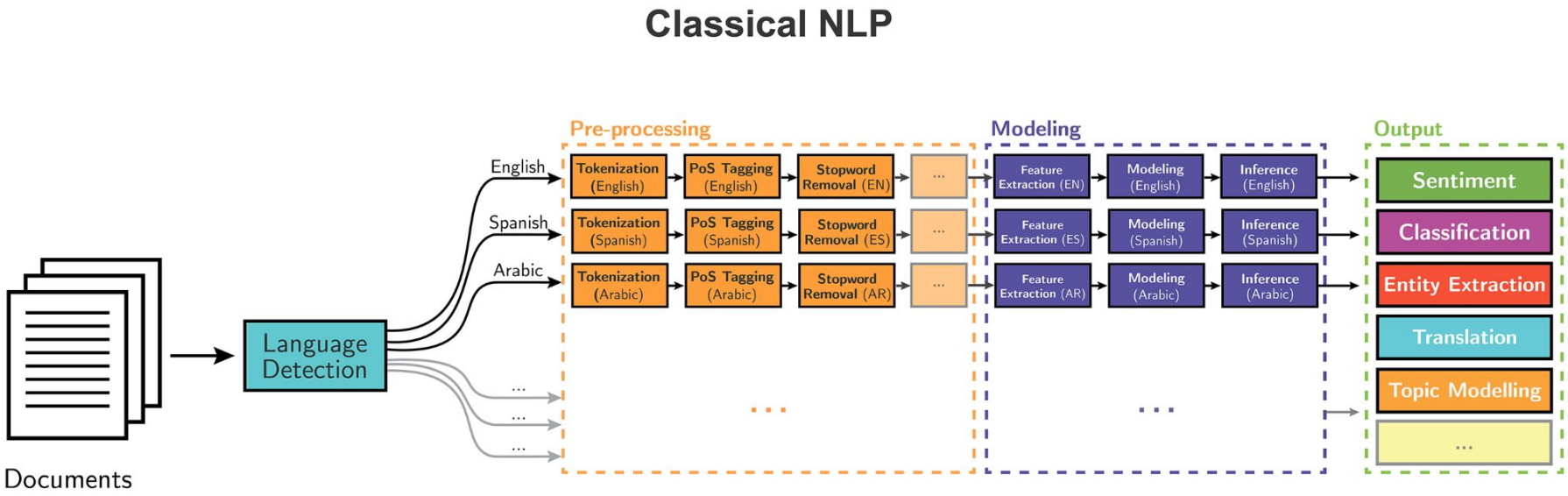
- Dalla metà degli anni ’80, con il crescente sviluppo delle risorse linguistiche (corpora annotati) e gli studi fatti nel campo dell’Intelligenza Artificiale (IA), vengono sviluppati algoritmi basati sull’apprendimento automatico (machine learning, ML).

- Questi algoritmi sono in grado di estrarre “automaticamente” dai corpora di addestramento i modelli che rappresentano la conoscenza della lingua necessaria per risolvere un determinato compito di analisi linguistica.

- Nel ML classico, le caratteristiche linguistiche (features) che il sistema utilizza per risolvere uno specifico compito vengono definite esplicitamente e pesate dal sistema di classificazione per generare il modello. Alcuni esempi di algoritmi sono la Maximum Entropy e le Support Vector Machines.

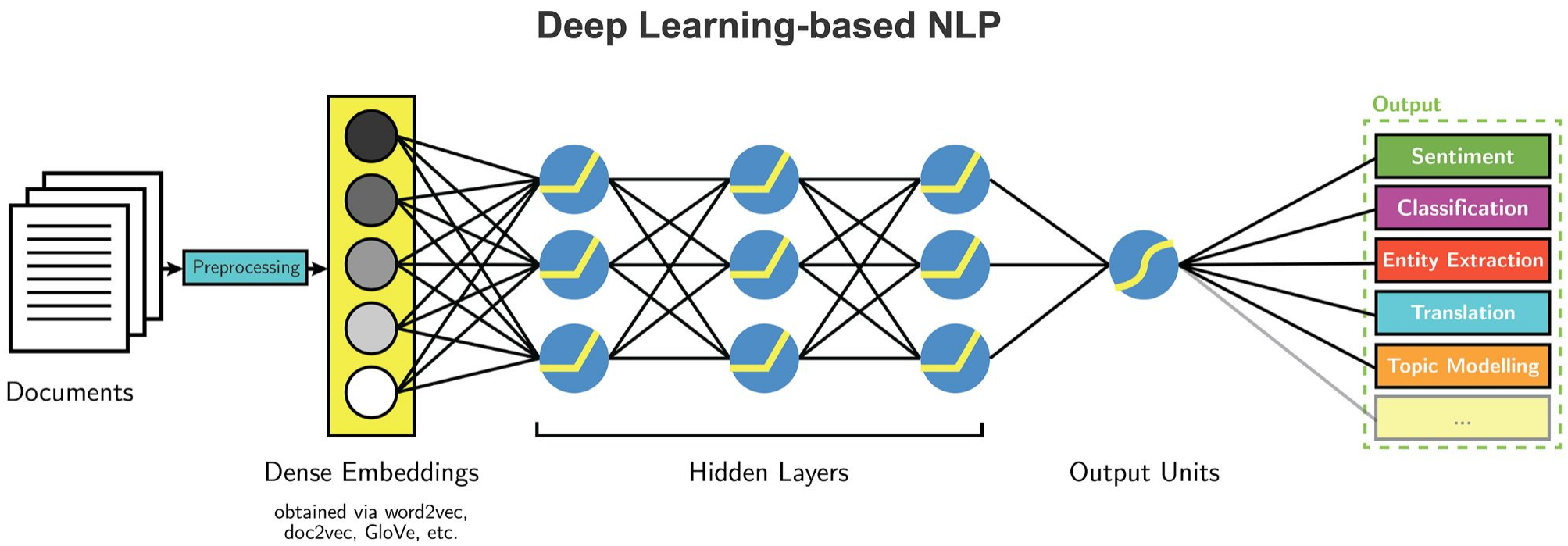
# La rivoluzione del Deep Learning

- Grazie alla disposizione di grandi quantità di dati e all'avanzamento tecnologico nello sviluppo di GPU sempre più prestanti, negli ultimi anni si sviluppano modelli basati su reti neurali e apprendimento profondo (deep neural learning) in grado di leggere il testo e comprenderne automaticamente l'informazione sintattica e semantica nascosta, sfruttando questa conoscenza linguistica per la risoluzione dei più svariati compiti (neural language model). Si passa quindi, da:



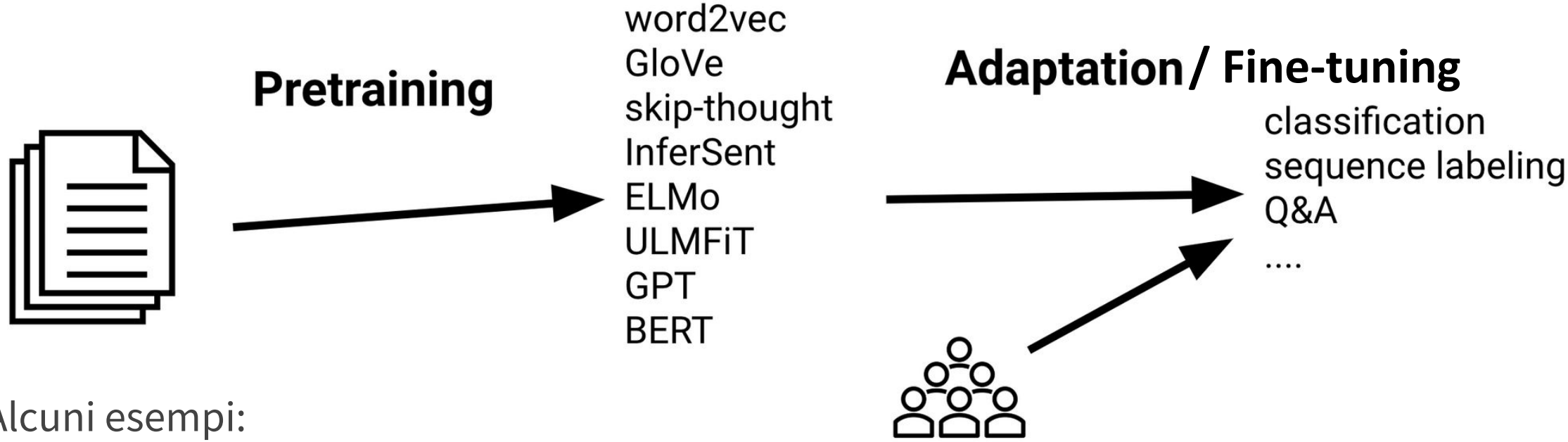
# La rivoluzione del Deep Learning

- Da modelli di apprendimento basati su feature esplicite pesate da un algoritmo di classificazione a sistemi in grado di estrarre automaticamente dal testo le informazioni rilevanti per la risoluzione dei vari compiti
- Da modelli a cascata specifici per ogni singolo compito a sistemi end-to-end



# La rivoluzione del Deep Learning

- Con l'avvento dei Neural Language Models cambia anche il paradigma di addestramento e di adattamento dei modelli per la risoluzione dei vari compiti



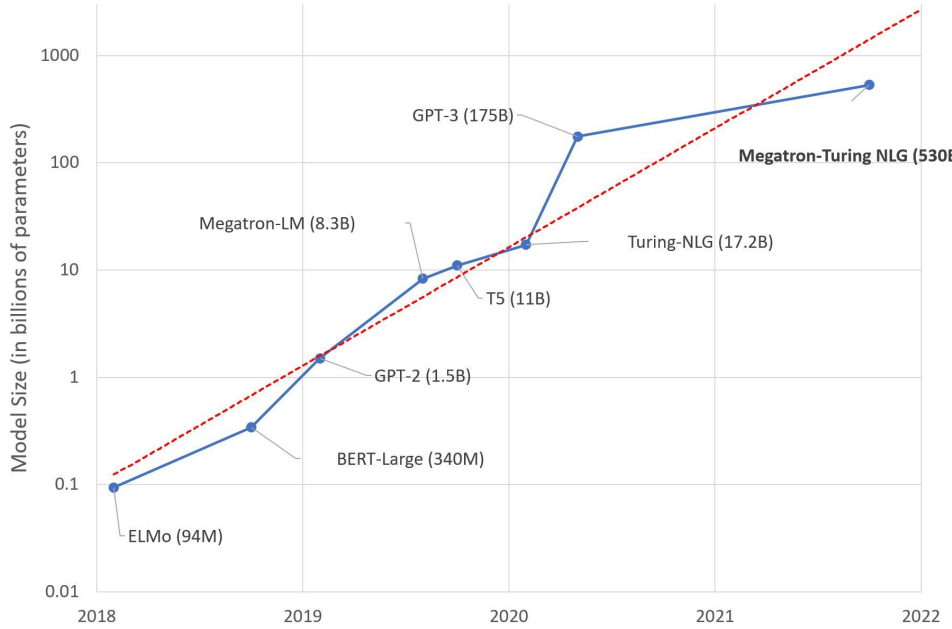

- Alcuni esempi:
  - **Training di BERT**: BookCorpus (800M di parole) e Wikipedia Inglese (2500M di parole)
  - **Training di GPT-3**: CommonCrawl + WebText2 + Books1 + Books2 + Wikipedia (circa 500B di parole)

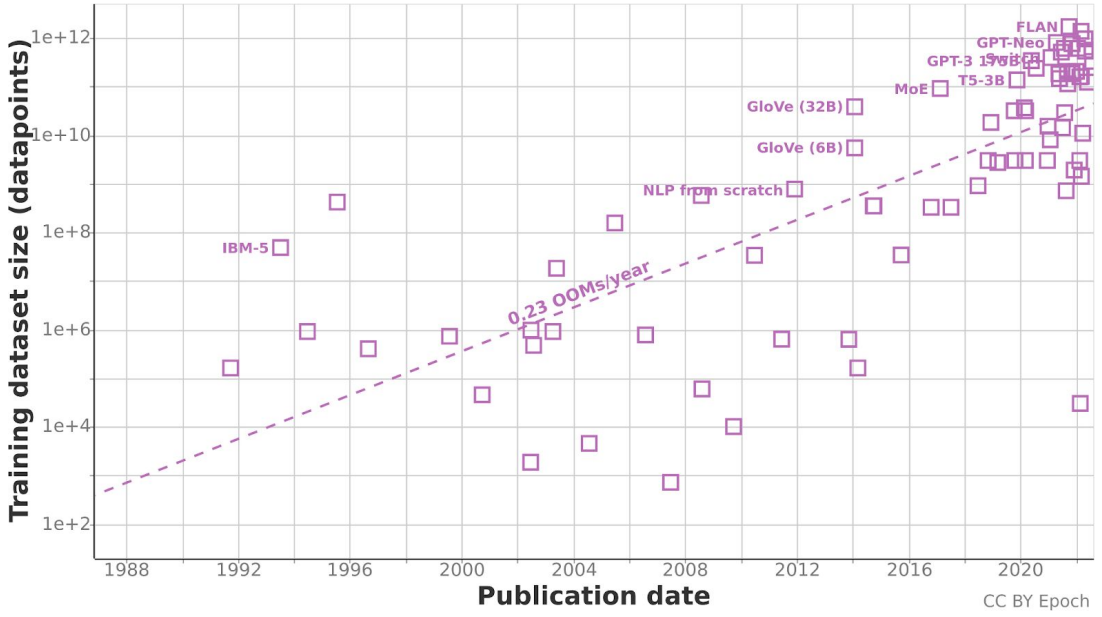
The State of Transfer Learning in NLP: https://ruder.io/state-of-transfer-learning-in-nlp/

# La rivoluzione del Deep Learning

Alcuni effetti indesiderati:

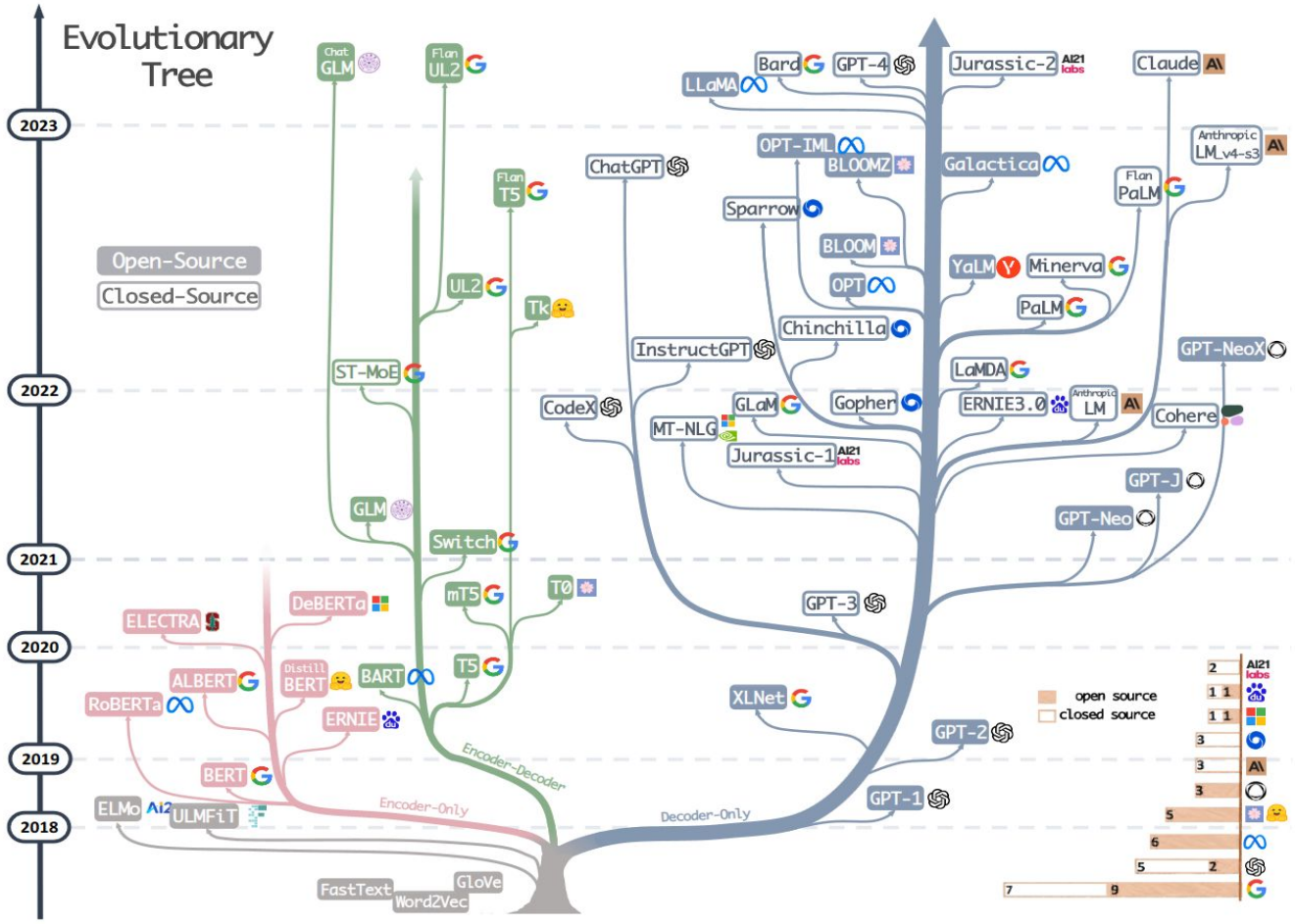
- aumenta drasticamente la dimensione dei modelli
- aumenta l'accuratezza ma diminuisce l'interpretabilità dei modelli e delle scelte effettuate dai modelli




Fonte: https://www.lesswrong.com/posts/asqDCb9XzXnLjSfgL/trends-in-training-dataset-sizes
Fonte: https://www.anyscale.com/blog/training-175b-parameter-language-models-at-1000-gpu-scale-with-alpa-and-ray

# La rivoluzione del Deep Learning




Fonte: [Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond (Yang et al., 2024), https://dl.acm.org/doi/10.1145/3649506]

# La rivoluzione del Natural Language Processing

Dal 2015 (*the Deep Learning Tsunami in NLP*) gli interessi applicativi restano gli stessi ma cambiano gli approcci e la ricerca di base.

I Modelli Linguistici raggiungono livelli di accuratezza molto elevati, ma al costo di dimensioni crescenti e di un elevato consumo di dati ed energia. Diventano “black box”, difficili da usare, controllare e spiegare.

Cambiano i nostri obiettivi di ricerca:

- Valutazione dei Language Model
- Controllo dei Language Model
- Spiegabilità e Interpretabilità

# Valutare i Language Models: una sfida aperta

Gli LLM moderni sono potenti ma opachi: come capiamo davvero cosa sanno fare?

1

**Modelli sempre più grandi, sempre meno trasparenti**

Centinaia di miliardi di parametri, addestrati su quantità enormi di testo. Difficile sapere cosa hanno imparato e come lo usano.

2

**I benchmark classici non bastano più**

I modelli “saturano” i test tradizionali. Inoltre molti benchmark sono già presenti nei dati di addestramento: il modello potrebbe averli semplicemente visti, non capiti.

3

**Serve un cambio di paradigma**

Non più “quanto è bravo in generale?”, ma quali competenze specifiche possiede? Dove fallisce? Cosa significa davvero la sua risposta?

# Valutare i Language Models: una sfida aperta

Studiare competenze specifiche degli LLM, dalla capacità di seguire vincoli specifici, alla comprensione di nuove parole, fino al ragionamento in scenari multimodali.

1. **Seguire istruzioni linguistiche** — *Gli LLM rispettano vincoli precisi?*
2. **Generalizzazione lessicale** — *Cosa fanno con parole mai viste?*
3. **Competenze minimali** — *Sanno davvero leggere lettera per lettera?*
4. **Benchmark creativi e sfidanti** — *Sanno giocare con la lingua?*
5. **Ragionamento multimodale** — *Sanno ragionare guardando un video?*

# Gli LLM sanno seguire istruzioni linguistiche precise?

Valutazione della generazione tramite vincoli linguistici espliciti: “scrivi una frase con 3 verbi”, “usa una proposizione subordinata”.

**Input**

*“Genera una frase con 1 subordinata.”*

↓

**Output**

*“**Nonostante piovesse**, John decise **di andare al lavoro**.”*

✓ **Vincolo rispettato**

**Input**

*“Genera una frase con 3 verbi.”*

↓

**Output**

*“Il sole **sorge**, gli uccelli **cantano** e il vento **disperde** le foglie.”*

✗ **Vincolo non rispettato**

Anche istruzioni linguistiche apparentemente semplici mettono in difficoltà i modelli più avanzati.

# Cosa succede quando un LLM incontra parole mai viste?

La lingua è viva: ogni giorno nascono parole nuove (**neologismi**) e ne inventiamo di inesistenti (**nonce words**) per gioco o necessità. Un parlante umano le capisce dal contesto. E un LLM?

### Reverse Dictionary

*definizione ⟶ parola*

*“film o serie televisiva che si ispira alla tutela dei valori ecoogici ”*

⟶

**Ecofiction**

### Definition Modeling

*parola ⟶ definizione*

*Tecnostanchezza*

⟶

**“Senso di affaticamento dovuto all'eccessivo utilizzo di tecnologie digitali”**

### Exemplification Modeling

*parola ⟶ frase d'esempio*

*Brevimirante*

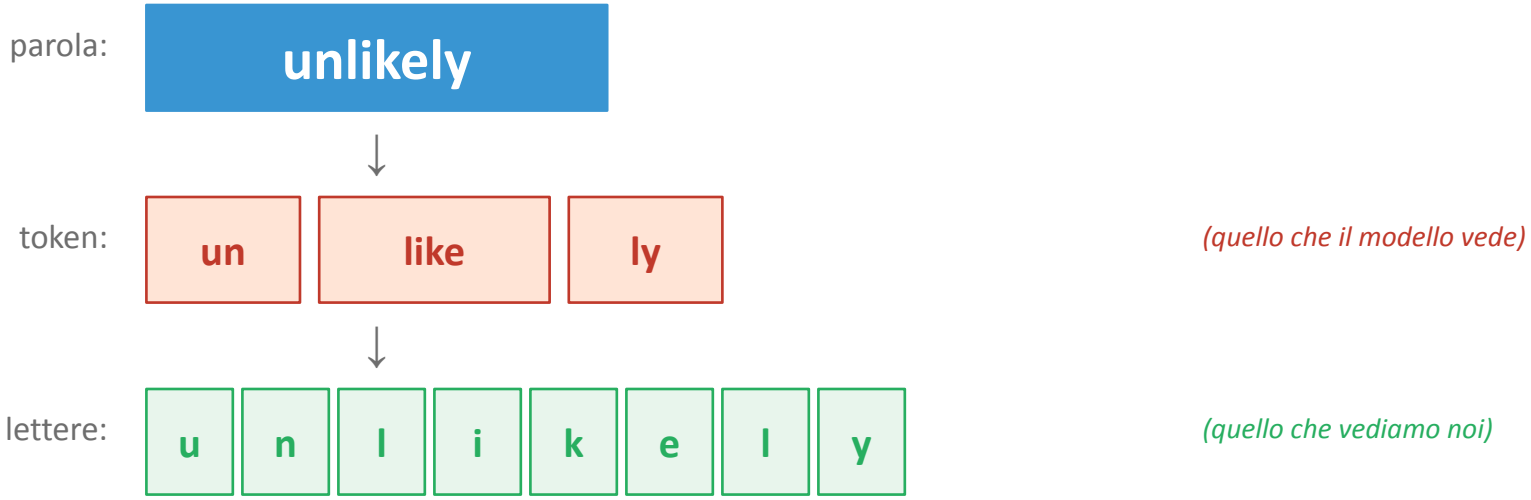
⟶

**“Una tattica brevimirante, ma che nascondeva in sé la volontà di non lasciarsi travolgere dal pessimismo.”**

Gli LLM hanno difficoltà con parole nuove, sebbene siano comunque in grado di **apprendere approssimazioni delle regole di formazione delle parole**, anziché affidarsi esclusivamente alla memorizzazione, mostrando quindi segni di generalizzazione.

# Gli LLM sanno davvero leggere lettera per lettera?

Sembra una domanda banale. In realtà no: gli LLM non leggono lettere, leggono **pezzi di parole** (i *token*). Una sorpresa che spiega molti dei loro errori.




→ *Domanda al modello: "La lettera 'u' è dentro 'unlikely'?"*

Sorprendentemente, qualche competenza sul riconoscimento dei caratteri emerge nonostante la struttura intrinseca degli LLM → **Spelling Miracle**. E non emerge a caso: i modelli riconoscono meglio e prima **unità linguisticamente motivate** rispetto a sequenze casuali di lettere.

# Gli LLM sanno giocare con la lingua?

Per evitare benchmark *saturi* e *contaminati*, valutiamo le abilità degli LLM tramite la definizione di prove creative e vincolate, ispirate alla letteratura sperimentale e all'enigmistica: terreni dove non basta aver visto tanti testi, serve **manipolare la lingua stessa**.

### OuLiBench — Sfide letterarie

*ispirate al movimento dell'OuLiPo e alla letteratura combinatoria*

**Lipogrammi** — scrivere senza una certa lettera

**Tautogrammi** — ogni parola con la stessa iniziale

**Anagrammi e Acrostici** — giochi di permutazione

**Pangrammi** — contenere tutte le lettere dell'alfabeto

### Cruciverba — Enigmistica computazionale

*definizioni, ambiguità, conoscenza culturale*

**Generazione di definizioni** — dato un termine, generare l'indovinello

**Risoluzione di griglie** — incastrare le risposte con vincoli

**Giochi di parole** — doppi sensi, anagrammi, calembour

**Conoscenza culturale** — riferimenti, modi di dire, storia

Questi compiti mettono alla prova ciò che la lingua ha di più combinatorio e culturale. Gli LLM hanno difficoltà, e proprio per questo sono test diagnostici preziosi.

# Gli LLM sanno ragionare guardando un video?

I modelli moderni non leggono solo testo: vedono **immagini e video**. Come valutare le abilità di ragionamento su contenuti multimodali?
**MAIA** è un benchmark italiano per valutare le abilità di ragionamento degli LLM su video brevi.

**Contenuto**

100 video brevi in italiano, coppie domanda–risposta

**Due compiti**

Visual Statement Verification (vero/falso) + Open-ended Visual Question Answering (risposta aperta)

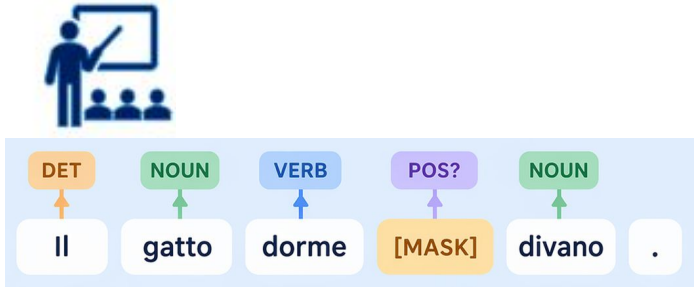
**Difficoltà crescente**

dal “cosa vedi” al “perché è successo” — dal riconoscimento al ragionamento

La vera comprensione multimodale richiede ragionamento, non solo riconoscimento. È qui che si misurano i limiti più profondi degli LLM di oggi.

# Controllo dei Language Model

Intervenire nel processo di pre-training del modello a diversi livelli:




**Dati**: scelta, qualità e bilanciamento definiscono *che conoscenza* il modello apprende

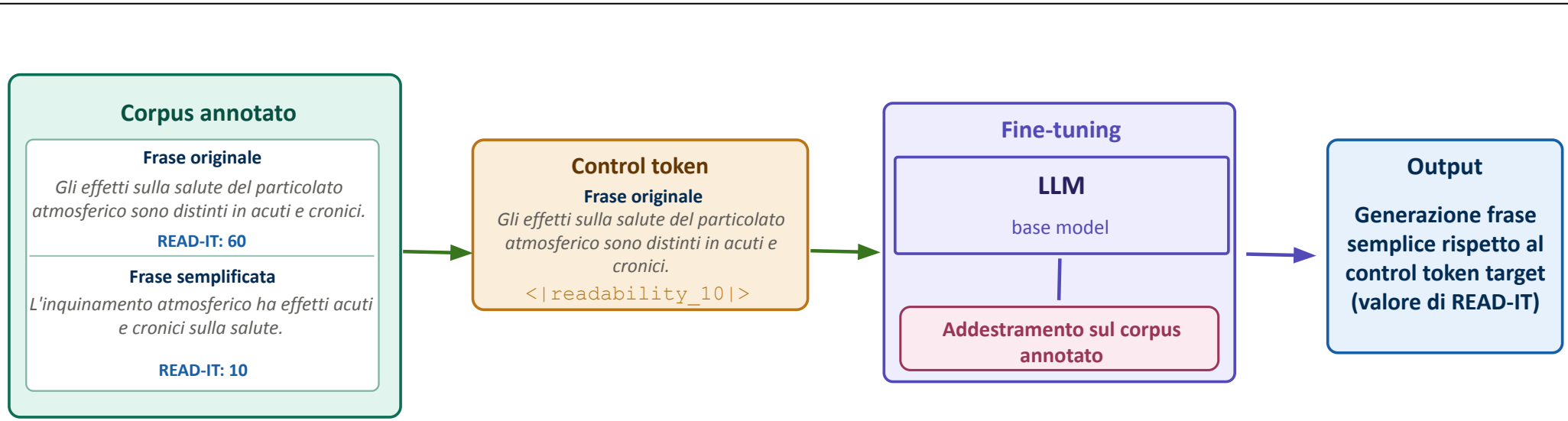
**Ordinamento**: la sequenza degli esempi guida *come il modello apprende*

- es. strategie di ordinamento ispirate all'apprendimento umano (da semplice a complesso)

**Bias**: la scelta di task e istruzioni per *indurre comportamenti, preferenze e capacità* nel modello

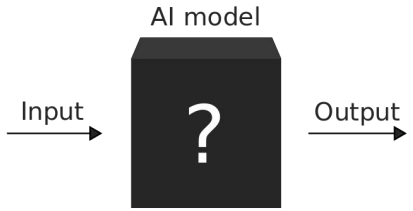
# Controllo dei Language Model

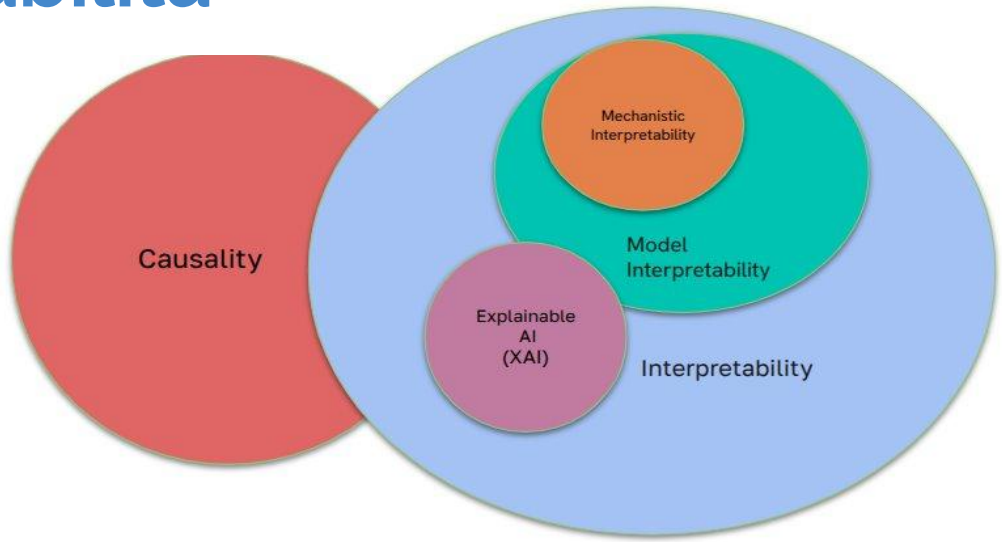
Intervenire nel processo di addestramento specializzato (fine-tuning) del modello a diversi livelli:



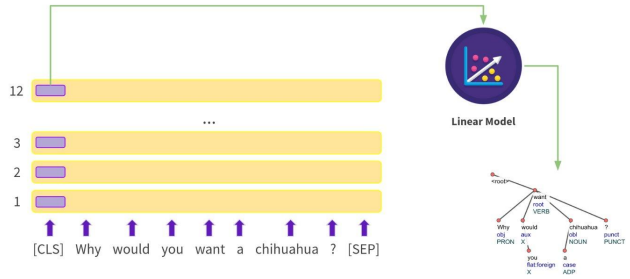

- Il modello apprende a generare un testo con determinate proprietà (es. **livello di complessità**) a partire da risorse annotate
- Uso di "control token" associati ai dati di fine-tuning per guidare la generazione (es. **punteggio di leggibilità della frase**)

# Spiegabilità e Interpretabilità



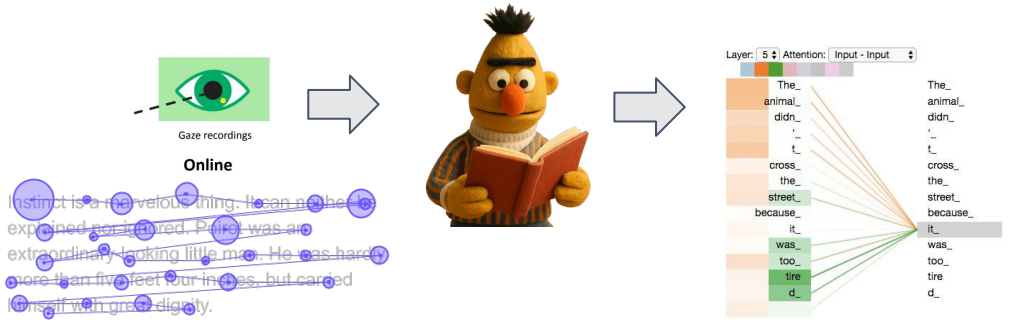




**Probing**: Cosa è codificato nelle rappresentazioni?

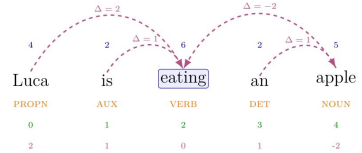


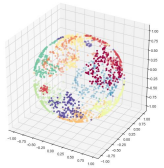

**Attention**: Dove guarda il modello durante l'elaborazione dell'input? E' possibile incidere sui meccanismi?

- uso di dati fisiologici umani (es. eye-tracking per modellare pattern di lettura)




**Geometria:** Come la conoscenza è organizzata nello spazio vettoriale? identificazione di cluster che riflettono proprietà linguistiche/semantiche



# Human-Centered NLP

Approccio allo sviluppo di sistemi di NLP che pone al centro le persone: i loro bisogni, valori, capacità e contesti d'uso. Creare modelli linguistici non solo accurati, ma anche comprensibili, equi, affidabili e utilizzabili, tenendo conto dell'impatto sociale, etico e culturale delle tecnologie linguistiche.

Caratteristiche indispensabile per applicazioni:

- **Educazione**: supporto alla valutazione e creazione di percorsi didattici personalizzati
- **Domini specialistici** (es: biomedico, giuridico): modelli trasparenti e interpretabili
- **Digital Humanities**: analisi linguistica, stilometria, identificazione di registri e generi
- **Accessibilità**: semplificazione automatica e personalizzata dei testi
- **Interazione uomo-macchina**: sistemi di comunicazione verbale (es. robotica)

# Chi siamo


Chiara Alzetta


Franco Alberto Cardillo


Mario Merone


Marta Sartor


Luca Bacco


Cristiano Ciaccio

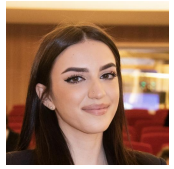

Alessio Miaschi


Giulia Venturi


Giulia Benotto


Felice Dell'Orletta


Simonetta Montemagni


Agnese Bonfigli


Luca Dini


Michele Papucci


Dominique Brunato


Lucia Domenichelli


Ruben Piperno