UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

# SYLLABUS DEL CORSO

# Data Semantics

**2526-1-FDS02Q010**

## Obiettivi

**Conoscenza e capacità di comprensione (DdD 1)**

Al termine del corso, lo studente avrà acquisito conoscenze su:

- i principi fondamentali della semantica dei dati e il loro ruolo nelle applicazioni di data science;
- i due principali paradigmi semantici: semantica dichiarativa, basata su modelli logico-formali e knowledge graph; semantica distribuzionale, fondata sull'apprendimento di rappresentazioni da dati testuali e sui modelli di linguaggio di grandi dimensioni (LLM);
- metodologie neuro-simboliche, che integrano approcci simbolici e neurali per la rappresentazione e l'elaborazione semantica;
- le principali tecnologie del web semantico (RDF, RDFS, OWL, SPARQL) e i fondamenti di ontologie, tassonomie e ragionamento automatico;
- word embeddings, pre-trained language models, large language modeels e loro applicazioni
- tecniche per costruire, arricchire e rendere fruibili basi di conoscenza, quali: riconciliazione di dati eterogenei, estrazione di entità e relazioni, *question answering* su dati e documenti

**Conoscenza e capacità di comprensione applicate (DdD 2)**

Lo studente sarà in grado di:

- modellare e interrogare knowledge graph utilizzando linguaggi e strumenti del web semantico;
- progettare e utilizzare ontologie per supportare l'integrazione semantica di dati in scenari applicativi;
- applicare modelli distribuzionali e LLM per l'interpretazione di testi e l'estrazione di conoscenza da dati non strutturati;
- realizzare soluzioni per la riconciliazione semantica, l'estrazione di entità e l'interrogazione di informazioni in linguaggio naturale;
- affrontare problemi reali di interoperabilità semantica, selezionando metodologie e strumenti adeguati in base al contesto;

- sviluppare piccoli progetti applicativi che combinano grafi di conoscenza, NLP e tecniche di rappresentazione semantica per analisi, esplorazione e generazione di contenuti.

**Altri descrittori di Dublino (DdD 3, 4, 5)**

- Autonomia di giudizio: lo studente svilupperà la capacità di valutare criticamente diverse soluzioni semantiche e di scegliere metodologie appropriate per nuovi problemi, anche attraverso esperienze progettuali e discussioni seminariali.
- Abilità comunicative: grazie alle presentazioni orali dei progetti e all'interazione durante le esercitazioni, lo studente sarà in grado di comunicare in modo efficace concetti complessi e risultati, anche in contesti multidisciplinari.
- Capacità di apprendere: il corso fornisce strumenti teorici e pratici che consentono allo studente di proseguire autonomamente lo studio delle tecnologie semantiche, anche attraverso la lettura di articoli scientifici e materiali avanzati.

## Contenuti sintetici

Il corso presenta strumenti computazionali per rappresentare, armonizzare e ricostruire la semantica dei dati utilizzati in applicazioni di data science, con particolare attenzione a:

- modelli e linguaggi elaborati nell'ambito del web semantico per supportare l'integrazione di dati eterogeni (**knowledge graph**, **ontologie**, **RDF, RDFS, OWL**);
- modelli per apprendere la semantica dai dati, con particolare riferimento a dati in formato testuale (**word embeddings**, **Large Language Models (LLM)**)
- tecniche per **integrare knowledge graph e LLM**.
- tecniche neurali per la **riconciliazione di dati**;
- tecniche di elaborazione del linguaggio naturale per **estrarre informazioni strutturate da testi** e **rispondere a domande usando dati e documenti** ;

## Programma esteso

1. **Data Semantics:** Semantica dei dati ed applicazioni di data analytics (big data, sorgenti web, formati eterogenei, integrazione di informazioni ed arricchimento semantico, connessione tra dati, knowledge graph)
2. **Knowledge Graph e Web Semantico:** rappresentazione e interogazione dei dati nel web semantico (RDF, SPARQL, tecnologie semantiche e architettture, rappresentazioni in ambito industriale mediante basi di dati a grafo). Esercitazione su interrogazione di Knowledge Graph pubblici con SPARQL; definizione di vocabolari condivisi mediante ontologie e linguaggi logico-formali (dai vocabolari condivisi alle ontologie, tassonomie, ontologie lessicali, ontologie assiomatiche, ragionamento automatico e semantica, RDFS, OWL). Esercitazione su modellazione di ontologie mediante RDFS e OWL.
3. **Semantica distribuzionale e modelli linguistici:** introduzione alla semantica distribuzionale e all'apprendimento di rappresentazioni distribuite (semantica distribuzionale); modelli per apprendere rappresentazioni distribuite da corpus testuali (word embeddings e word2vec, contextual word embeddings e Large Language Models - LLM). Esercitazione su LLM e attenzione. Seminario: modelli per comparare rappresentazioni distribuite differenti per applicazioni di computational social science e cultural analysis (allineamento tra word embeddings, analisi diacroniche, studi basati su word embeddings con WEAT e SWEAT).
4. **Riconciliazione semantica:** algoritmi di entity matching basati su reti neurali (deep matcher, Ditto, BERT-based matching, matching con large language models).

5. **Elementi di NLP - tecniche di estrazione di informazioni:** introduzione e presentazione di alcuni approcci all'estrazione di informazioni strutturate da testo e altri dati semi strutturati (named entity recognition, entity linking, estrazione di relazioni, semantic table interpretation). Esercitazione su named entity recognition e named entity linking.
6. **Tecniche di accesso alle informazioni mediate dalla semantica:** tecniche semantiche per l'esplorazione di informazioni (faceted search,retrieval augmented generation)

## Prerequisiti

Conoscenze matematiche e informatiche insegnate nei corsi obbligatori del primo semestre.

## Modalità didattica

Lezioni frontali ed esercitazioni con i personal computer degli studenti. Uso della piattaforma Moodle. Seminari su applicazioni delle tecnologie semantiche a problemi reali da parte di experti del mondo dell'industria.

Didattica Erogativa: ~32h (lezioni frontali)
Didattica Interattiva: ~12h (esercitazioni guidate)

Insegnato in Inglese

## Materiale didattico

Knowledge Graphs: Fundamentals, Techniques, and Applications. Kejriwal, Mayank, Craig A. Knoblock, and Pedro Szekely. MIT Press, 2021.
The Web of Data. Aidan Hogan. 2020. Springer. Pages 1-680.

Daniel Jurafsky and James H. Martin. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. (Online manuscript released January 12, 2025. https://web.stanford.edu/~jurafsky/slp3 )

Verrà fornito agli studenti materiale aggiuntivo sotto forma di presentazioni e articoli scientifici per coprire gli argomenti più recenti non coperti dal libro di testo.

## Periodo di erogazione dell'insegnamento

Semestre II

## Modalità di verifica del profitto e valutazione

La valutazione finale è costituita dall'aggregazione dei punteggi ottenuti in due valutazioni indipendenti.

- La prima valutazione è basata su un **progetto d'esame,** effettuato individualmente o in gruppo, e finalizzato all'approfondimento di un argomento specifico trattato nel corso o collegato ad argomenti trattati nel corso; il progetto viene discusso attraverso una **presentazione orale supportata da slide** della durata di 20 min circa; è possibile, durante la presentazione, includere una breve demo del progetto svolto. *La valutazione si basa su: significatività del progetto rispetto agli argomenti trattati nel corso, rigore metodologico (nei limiti di quanto ragionevole chiedere per un progetto d'esame); padronanza dell'argomento approfondito dimostrata durante la presentazione orale*.
- La seconda valutazione è basata sulla **verifica della conoscenza degli argomenti affrontati durante il corso** mediante valutazione di esercizi (assignment) da completare individualmente e discussione orale. Gli assignment verranno valutati e discussi in sede d'esame, dopo la discussione del progetto.

## Orario di ricevimento

Su richiesta

## Sustainable Development Goals

ISTRUZIONE DI QUALITÁ | IMPRESE, INNOVAZIONE E INFRASTRUTTURE