



UNIVERSITÀ  
DEGLI STUDI DI MILANO-BICOCCA

## SYLLABUS DEL CORSO

### Data Semantics

2021-1-F9101Q011

---

#### Obiettivi

Scopo principale del corso è fornire agli studenti le conoscenze e competenze necessarie per comprendere e risolvere problemi di interoperabilità semantica in applicazioni di data science, con particolare riferimento a problemi di rappresentazione, riconciliazione e integrazione di dati eterogenei.

Gli argomenti che verranno trattati hanno un duplice scopo: 1) fornire un insieme di strumenti teorici e pratici per rappresentare, organizzare, pubblicare, interrogare, riconciliare, ed esplorare dati e conoscenze in scenari applicativi reali (ampiamente discussi durante le lezioni frontali e affrontati durante le esercitazioni) utilizzando tecnologie semantiche e 2) acquisire le competenze necessarie per comprendere problemi di interoperabilità semantica nuovi e le tecniche necessarie per risolverli adeguatamente indipendentemente dalle particolari tecnologie di riferimento.

#### 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 eterogenei (knowledge graph, data linking, ontologie, RDF, RDFS, OWL);
- tecniche per l'integrazione di dati e vocabolari;
- tecniche di estrazione delle informazioni da testi (cenni);

- modelli di intelligenza artificiale a supporto dell'esplorazione di dati e conoscenze.

## 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:** rappresentazione e interrogazione dei dati nel web semantico (RDF, SPARQL, tecnologie semantiche e architetture, rappresentazioni in ambito industriale mediante basi di dati a grafo). Esercitazione su interrogazione di Knowledge Graph pubblici con SPARQL.
3. **Knowledge Graph e Semantica:** 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, SWRL). Esercitazione su modellazione di ontologie mediante i linguaggi RDFS e OWL.
4. **Riconciliazione semantica I:** integrazione di informazioni e riconciliazione semantica, riconciliazione a livello delle istanze e dello schema, estrazione di informazioni (named entity recognition, entity linking, estrazione di relazioni)
5. **Riconciliazione semantica II:** riconciliazione di ontologie e vocabolari (ontology matching per allineare ontologie e tassonomie, terminologia e mapping, similarità semantica e combinazione di diverse funzioni di similarità, selezione dei mapping). Esercitazione su riconciliazione di tassonomie.
6. **Riconciliazione semantica III:** riconciliazione a livello dei valori o delle istanze (deduplicazione e record linkage, approcci probabilistici, metriche di distanza e misure di similarità, combinazione e apprendimento di misure di similarità complesse, strategie per la fusione di informazioni eterogenee, misure di similarità basate su grafi). Esercitazione su riconciliazione di dati con l'aiuto di strumenti esistenti.
7. **Esplorazione di informazioni e conoscenze:** tecniche semantiche per l'esplorazione di informazioni (misure di rilevanza, associazioni semantiche, apprendimento attivo di associazioni rilevanti, sistemi di raccomandazione)
8. **Nuovi approcci alla semantica dei dati:** approcci semantici guidati dai dati e di frontiera (profilazione semantica di Knowledge Graph, semantica distribuzionale, word embeddings e knowledge graph embeddings)

## 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 esperti del mondo dell'industria.

Nel periodo di emergenza Covid-19 le lezioni si svolgeranno in modalità mista: parziale presenza e lezioni videoregistrate asincrone/ sincrone. Qualora non si renda possibile tale modalità, il corso verrà tenuto da remoto asincrono con eventi in videoconferenza sincrona.

Insegnato in Inglese

## Materiale didattico

ITA: [Tommaso Di Noia](#), [Roberto De Virgilio](#), [Eugenio Di Sciascio](#), [Francesco M. Donini](#). Semantic Web: tra ontologie e Open Data, Apogeo, 2013.

ENG: Grigoris Antoniou, Paul Groth, Frank van Harmelen, *A Semantic Web Primer*, (Third Edition), MIT press, 2012.

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 o approfondimento tematico**, effettuato individualmente o in gruppo, e finalizzato all'approfondimento di un argomento specifico trattato nel corso o collegato ad argomenti trattati nel corso; progetto e approfondimento vengono entrambi discussi 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; l'**approfondimento** consiste di una rassegna bibliografica su un argomento, in cui lo studente discute e compara soluzioni proposte nello stato dell'arte a uno specifico problema. *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 una delle seguenti modalità, scelta liberamente dallo studente:
  1. esame orale sostenuto in concomitanza con la discussione della prima prova
  2. due prove in itinere costituite da esercizi e domande aperte: una relativa ai primi argomenti trattati nel corso (knowledge graph, ontologie, RDF, RDFS, OWL), e una relativa agli argomenti trattati nella seconda parte del corso (integrazione di dati e vocabolari, estrazione di informazioni, modelli per l'esplorazione di dati e

conoscenze).

Nel periodo di emergenza Covid-19, qualora non sia possibile effettuare esami scritti in presenza, le prove in itinere verranno sostituite da esercizi da effettuare individualmente e verificati durante l'esame orale.

Nel periodo di emergenza Covid-19 gli esami orali saranno solo telematici. Verranno svolti utilizzando la piattaforma WebEx e nella pagina e-learning dell'insegnamento verrà riportato un link pubblico per l'accesso all'esame di possibili spettatori virtuali.

## **Orario di ricevimento**

Giovedì 14.30-15.30

---