UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

# COURSE SYLLABUS

# Data Semantics

**2223-1-FDS01Q010**

## Aims

The main purpose of the course is to provide students with the knowledge and skills necessary to understand and solve problems that are related to the semantic interpretation of data in data science applications. A special focus is given to problems and solutions related to the representation, reconciliation, and integration of heterogeneous data; 2) analyses of textual data based on the meaning of the expressions contained therein.

The topics addressed in the course have a dual purpose: 1) to present techniques and practical tools to organize, publish, query, reconcile, explore and interpret information in real application scenarios (widely discussed during lectures and addressed during the exercises) using a selection of semantic technologies available today and 2) to acquire methodological tools to understand and solve new problems related to data semantics in the future, regardless of particular reference technologies.

## Contents

The course presents computational methods to represent, harmonize and interpret the semantics of data used in data science applications, with a particular focus on:

- models and languages developed within the semantic web to support the integration of heterogeneous data (knowledge graph, data linking, ontologies, RDF, RDFS, OWL);
- models to learn (semantic) representations from data, especially from text corpora (word embeddings, contextual word embeddings);
- techniques to integrate data and vocabularies;
- techniques to distill, extract and interpret information from texts;
- techniques to support access to vast amounts of knowledge.

## Detailed program

1. **Data semantics:** the role of semantics in data analytics (big data, web sources, heterogeneous formats, information integration, semantic enrichment, data linking, knowledge graphs).
2. **Knowledge graphs and the semantic web:** representation and query of data in the semantic web (RDF, SPARQL, semantic technologies and architectures, corporate knowledge graphs with graph databases). Excercise on querying RDF knowledge graphs with SPARQL.
3. **Knowledge representation and reasoning:** definition of shared vocabularies with ontologies and logic-based languages ??(from shared vocabularies to ontologies, taxonomies, lexical ontologies, axiomatic ontologies, automatic reasoning and semantics, RDFS, OWL, SWRL). Excercise on ontology modeling with RDFS and OWL.
4. **Introduction to NLP - distributional semantics and representation learning:** introduction to distributional semantics and distributed representations (distributional semantics); models for learning distributed representations from textual corpora (word embeddings and word2vec, contextual word embeddings); models to compare different distributed representations (alignment between word embeddings, diachronic language studies, studies based on word embeddings with WEAT and SWEAT).
5. **Semantic reconciliation:** ontology and vocabulary mapping (ontology matching to align ontologies and taxonomies, terminology and mapping, semantic similarity and combination of different similarity functions, selection of mappings) and instance-level matching (probabilistic record linkage, distance metrics and similarity measures, combination and learning of complex similarity measures, strategies for merging heterogeneous infromation, graph-based similarity measures). Exercise on data reconciliation with the help of existing tools.
6. **Introduction to NLP - information extraction (hints):** presentation of selected approaches to the extraction of structured information from texts and other semi-structured data (named entity recognition, entity linking, relationship extraction, semantic table interpretation).
7. **Information and knowledge exploration:** semantic techniques for passive and active exploration of information (semantic search, recommendation systems).

## Prerequisites

Mathematics and computer science as taught in the compulsory courses of the first semester.

## Teaching form

Lectures and exercises with students' personal computers. Moodle e-learning platform. Seminars about the usage of semantics in real-world applications given by experts from the industry.

## Textbook and teaching resource

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.

Additional material such as presentations and articles is provided to cover novel topics that are not covered by the textbook.

## Semester

Semester II

## Assessment method

The final evaluation consists of the aggregation of the scores obtained in two independent assessments.

- The first assessment is based on an exam-tailored project, carried out individually or in groups and aimed at bringing the student to have in-depth knowledge and/or hands-on experience of a specific topic covered in the course or linked to topics covered in the course; the project is discussed through an oral presentation supported by slides lasting about 20 minutes; it is possible, during the presentation, to include a short demo of the project. The evaluation is based on: significance of the project for the topics covered in the course, methodological soundness (within the limits of what is reasonable to ask for an exam project); mastery of the in-depth topic demonstrated during the oral presentation.
- The second assessment is based on the evaluation of the knowledge acquired by the student on the topics addressed during the course through the discussion of assignments that students must execute individually as homework. Assignments will be evaluated and discussed during the oral exam after the presentation of the project.

## Office hours

On demand

## Sustainable Development Goals

QUALITY EDUCATION