COURSE DESCRIPTION – ACADEMIC YEAR 2018/2019

Course title	**Data Curation**
Course code	73005
Scientific sector	ING-INF/05
Degree	Master in Computational Data Science (LM-18)
Semester	1
Year	1
Credits	12
Modular	Yes

Total lecturing hours	80
Total lab hours	40
Attendance	Attendance is not compulsory, but non-attending students have to contact the lecturers at the start of the course to agree on the modalities of the independent study.
Prerequisites	Knowledge of relational databases, as taught in an introductory course at the BSc level. Basic knowledge of first-order logic, as taught in a BSc course in logic or discrete mathematics. Java programming skills for the project part. (Data profiling module): Basic knowledge of linear algebra, statistics and probability theory; MATLAB and Python basics for the lab session.
Course page	https://ole.unibz.it/

Specific educational objectives	The course belongs to the type "caratterizzanti – discipline informatiche" in the curriculum "Data Analysis". The Data Integration module addresses a variety of problems related to the integration of heterogenous data sources, that range from structured data (such as relational databases), over semi-structured data (such as data on the Web, and tree- and graph-structured data), to unstructured (textual) data. It overviews foundational techniques for data integration, such as schema mappings, data and schema matching, and query processing in data integration, and does so considering different data representations that go beyond the relational model, such as RDF data, linked open data, and knowledge graphs. Architectures for data integration and data federation and their adoption to build comprehensive data integration solutions are studied. By attending the course, students will also learn how to design and build a data integration system, possibly exploiting existing data access and data federation technologies. The Data Profiling module considers the problem of profiling databases, in the general case where they refer to multimedia material (i.e., images, sounds, videos, etc). The course will cover the task of textual tag generation from audio-video material, to provide discriminative metadata. Standard techniques for detecting patterns and violations, dependencies, data normalization, similarity measures, and summarization will be carried out. Topics include: basics on image and sound analysis, multimedia discriminative dimensionality reduction, bag of X, textual tag extraction by supervised classification.

	Patterns and violations detection, dependencies detection, scrubbing and normalization. Similarity measures, duplicate detection, summary extraction.

Module 1	**Data Integration**
Module code	73005A
Module scientific sector	ING-INF/05
Lecturer	Diego Calvanese
Contact	POS 2.07, calvanese@inf.unibz.it, +39 0471 016160
Scientific sector of lecturer	ING-INF/05
Teaching language	English
Office hours	Announced on the webpage of the lecturer. During the lecture time span, in general Friday 16:00-18:00. Outside of the lecture time span, students are advised to confirm availability by email.
Lecturing assistant (if any)	--
Contact LA	--
Office hours LA	--
Credits	6
Lecturing hours	40
Lab hours	20
List of topics	• Data integration architectures • Schema mapping • Data matching • Heterogeneous and web data • Data cleaning • Query processing for data integration
Teaching format	Frontal lectures, exercises, and labs.

Module 2	**Data Profiling**
Module code	73005B
Module scientific sector	INF/01
Lecturer	Marco Cristani
Contact	Office n.104, marco.cristani@inf.unibz.it, +39 348 8619516
Scientific sector of lecturer	ING-INF/05
Teaching language	English
Office hours	Monday 10:00-11:30, Wednesday 13:00-14:00 (students are advised to arrange beforehand by email, in order to avoid waiting time).
Lecturing assistant (if any)	--
Contact LA	--
Office hours LA	--
Credits	6
Lecturing hours	40
Lab hours	20
List of topics	• Detecting patterns and violations • Detecting dependencies • Scrubbing and normalization • Similarity measures • Duplicate detection • Summary extraction
Teaching format	Frontal lectures and labs

Learning outcomes	Knowledge and understanding: • D1.1 - Knowledge of the key concepts and technologies of data science disciplines • D1.2 - Understanding of the skills, tools and techniques required for an effective use of data science • D1.6 - Knowledge of the principles and methods of data curation Applying knowledge and understanding: • D2.1 - Practical application and evaluation of tools and techniques in the field of data science • 2.5 - Ability to apply, evaluate and develop methods and tools for the integration, cleaning, and quality of data Making judgments • D3.2 - Ability to autonomously select the documentation (in the form of books, web, magazines, etc.) needed to keep up to date in a given sector Communication skills • D4.1 - Ability to use English at an advanced level with particular reference to disciplinary terminology • D4.3 - Ability to structure and draft scientific and technical documentation Learning skills • D5.2 - Ability to autonomously keep oneself up to date with the developments of the most important areas of data science

Assessment	Oral exam and project work. The mark for each part of the exam is 18-30, or insufficient. The oral exam covers Module 1 "Data Integration" and Module 2 "Data Profiling", and comprises verification questions, and open questions to test knowledge application skills. It counts for 50% of the total mark. The project consists of two parts. Part 1 covers Module 1 "Data Integration", and verifies whether the student is able to apply advanced data integration techniques and technologies taught or presented in the course to solve concrete problems. It is assessed through a final presentation, a demo, and a project report and can be carried out either individually or in a group of 2 students. It is discussed during the oral exam, and it counts for 25% of the total mark. Part 2 covers Module 2 "Data Profiling", and consists in developing MATLAB/Python code, with the aim of checking how much the student has been able to implement and evaluate the tools and techniques discussed during the frontal lessons. It is discussed during the oral exam, and it counts for 25% of the total mark.
Assessment language	English
Assessment Typology	Monocratic

Evaluation criteria and criteria for awarding marks	The final mark is computed as the weighted average of the oral exam, Part 1 of the project, and Part 2 of the project. The exam is considered passed when all three marks are valid, i.e., in the range 18-30. Otherwise, the individual valid marks (if any) are kept for all 3 regular exam sessions, until also all others parts are completed with a valid mark. After the 3 regular exam sessions, all marks become invalid. Relevant for the oral exam: clarity of answers; ability to recall principles and methods, and deep understanding about the course topics presented in the lectures; skills in applying knowledge to solve exercises about the course topics; skills in critical thinking. Relevant for the project: skill in applying knowledge in a practical setting; ability to summarize in own words; ability to develop correct solutions for complex problems; ability to write a quality report; ability in presentation; ability to work in teams (for Part 1 of the project). Non-attending students have the same evaluation criteria and requirements for passing the exam as attending students.

Required readings	Required books: • A. Doan, A. Halevy & Z. Ives (2012). Data Integration. Morgan Kaufmann. (ST270 D631) • Z. Abedjan, L. Golab, F. Naumann, & T. Papenbrock (2018). Data Profiling. Synthesis Lectures on Data Management, 10(4), 1-154. Additional material (slides, notes of the lecturers) will be made available before each lesson. Subject Librarian: David Gebhardi, David.Gebhardi@unibz.it
Supplementary readings	Suggested readings (i.e., only a few chapters will be covered) • R. C. Gonzalez, & R. E. Woods (2007). Image processing. Digital image processing, 2, 1. • R. O. Duda, P. E. Hart & D. G. Stork (2012). Pattern classification. John Wiley & Sons.
Software used	• Ontop system for ontology-based data access developed by the In2Data research group at the Faculty of Computer Science • Data federation tools • MATLAB 2017 (Image processing toolbox, statistic toolbox) • Python3.5 • Pycharm 2016.3.2