Syllabus

Course Description

Course Title	Data Curation
Course Code	73056
Course Title Additional	
Scientific-Disciplinary Sector	
Language	English
Degree Course	Master in Computing for Data Science
Other Degree Courses (Loaned)	
Lecturers	Prof. Diego Calvanese, Diego.Calvanese@unibz.it https://www.unibz.it/en/faculties/engineering/academic-staff/person/3562 Dr. Alessandro Mosca, Alessandro.Mosca@unibz.it https://www.unibz.it/en/faculties/engineering/academic-staff/person/29044
Teaching Assistant	
Semester	First semester
Course Year/s	2
CP	12
Teaching Hours	80
Lab Hours	40
Individual Study Hours	180
Planned Office Hours	
Contents Summary	• Data integration architectures • Query processing in data integration • Schema mapping • Data integration via virtual knowledge graphs • Schema matching • Data and entity matching • Patterns and violation detection

	• Correlation coefficients • Elementary data analysis • Exact and approximate value dependencies detection • Normalization via data profiling • Non-relational data profiling • Data profiling for query optimisation
Course Topics	The Data Preparation and Integration module addresses a variety of problems related to the integration of heterogenous data sources. It overviews the main issues in data integration, notably handling different forms of heterogeneity, and presents the general architecture of data integration systems. Foundational techniques for data integration are covered, such as data matching, schema matching and mapping, and query processing in data integration. A specific data integration approach relying on the technology of Virtual Knowledge Graphs and semantic mappings is presented in detail. The integration both of relational data sources, and of other types of data sources accessed by relying on data federation technology are considered. By attending the course, students will learn how to design and build a comprehensive data integration solution, possibly exploiting existing data access and data federation technologies. The Data Profiling module considers a variety of problems related to the profiling of relational and non-relational data. It first overviews techniques for the exploratory analysis of relational databases, which are also adopted in preparatory activities of data cleansing and integration, and mining. Widely used statistical measures, such as the Pearson and Spearman correlation coefficients, are discussed and applied to real word examples in the first part of the course. Special attention will be also paid to the Phi-K correlation coefficient, which has been recently proposed to overcome well known limitations of the former two, such as the possibility to analyse correlations between numerical, categorical, interval and ordinal variables. Algorithms for discovering patterns and relational dependencies in tabular data will be presented in detail. In particular, the main part of the course will go through the formal specifications of uniqueness, functional and inclusion dependencies in relational data, with a special focus on state-of-the-art algorithms and their complexity. If time permits, the last part of the course will discuss the application of the introduced

	profiling techniques to the relational query optimization problem, and to a high level introduction of recent profiling methods that can be applied to RDF data.
Keywords	Data Integration, Schema Mappings, , Virtual Knowledge Graphs, Relational Dependencies
Recommended Prerequisites	For the Data Preparation and Integration module: 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. Knowledge of Java or Python for the project part. For the Data Profiling module: relational database concepts, basic machine learning concepts; good to have knowledge of basic information retrieval concepts. Python programming basics for the lab sessions.
Propaedeutic Courses	
Teaching Format	Frontal lectures, exercises, and labs.
Mandatory Attendance	Attendance is not compulsory, but non-attending students must contact the lecturers at the start of the course to agree on the modalities of the independent study.
Specific Educational Objectives and Learning Outcomes	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. It overviews the main issues in data integration, notably handling different forms of heterogeneity, and presents the general architecture of data integration systems. Foundational techniques for data integration are covered, such as data matching, schema matching and mapping, and query processing in data integration. A specific data integration approach relying on the technology of Virtual Knowledge Graphs and semantic mappings is presented in detail. The integration both of relational data sources, and of other types of data sources accessed by relying on data federation technology are considered. By attending the course, students will learn how to design and build a comprehensive data integration solution, possibly exploiting existing data access and data federation technologies. The Data Profiling module considers a variety of problems related to the profiling of relational and non-relational data. It first

overviews techniques for the exploratory analysis of relational databases, which are also adopted in preparatory activities of data cleansing and integration, and mining. Widely used statistical measures, such as the Pearson and Spearman correlation coefficients, are discussed and applied to real word examples in the first part of the course. Special attention will be also paid to the Phi-K correlation coefficient, which has been recently proposed to overcome well known limitations of the former two, such as the possibility to analyse correlations between numerical, categorical, interval and ordinal variables. Algorithms for discovering patterns and relational dependencies in tabular data will be presented in detail. In particular, the main part of the course will go through the formal specifications of uniqueness, functional and inclusion dependencies in relational data, with a special focus on state-of-the-art algorithms and their complexity. If time permits, the last part of the course will discuss the application of the introduced profiling techniques to the relational query optimization problem, and to high level introduction of recent profiling methods that can be applied to RDF data.

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
- 2.10 - Application of languages, tools, and methods for the design of information systems and their corresponding software applications for data, process, and organization management

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
Specific Educational Objectives and Learning Outcomes (additional info.)	
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 Preparation and Integration" and verifies whether the student is able to apply advanced data integration techniques and technologies taught or presented in the course to solve a concrete integration problem. 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 verifies whether the student is able to apply data profiling techniques to a concrete use case. It is assessed through a final presentation 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.
Evaluation Criteria	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 other 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. Non-attending students have the same evaluation criteria and requirements for passing the exam as attending students.
Required Readings	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. Z. Zhang, R. Zhang (2008). Multimedia Data Mining: A Systematic Introduction to Concepts and Theory. CRC Press LLC. Additional material (slides, notes of the lecturers) will be made available before each lesson. Subject Librarian: David Gebhardi, David.Gebhardi@unibz.it
Supplementary Readings	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.
Further Information	Software used:

	Ontop system for deploying and querying Virtual Knowledge Graphs, developed by the In2Data research group at the Faculty of Computer Science. Relational DBMS, such as PostgreSQL. Data federation tools such as Denodo, Dremio, Teiid. Python3.5 with pandas, scikit-learn, scikit-image.
Sustainable Development Goals (SDGs)	Quality education

Course Module

Course Constituent Title	Data Preparation and Integration
Course Code	73056A
Scientific-Disciplinary Sector	IINF-05/A
Language	English
Lecturers	Prof. Diego Calvanese, Diego.Calvanese@unibz.it https://www.unibz.it/en/faculties/engineering/academic-staff/person/3562
Teaching Assistant	
Semester	First semester
CP	6
Responsible Lecturer	
Teaching Hours	40
Lab Hours	20
Individual Study Hours	90
Planned Office Hours	
Contents Summary	• Data integration architectures • Query processing in data integration • Schema mapping • Data integration via virtual knowledge graphs • Schema matching • Data and entity matching
Course Topics	The Data Preparation and Integration module addresses a variety

	of problems related to the integration of heterogenous data sources. It overviews the main issues in data integration, notably handling different forms of heterogeneity, and presents the general architecture of data integration systems. Foundational techniques for data integration are covered, such as data matching, schema matching and mapping, and query processing in data integration. A specific data integration approach relying on the technology of Virtual Knowledge Graphs and semantic mappings is presented in detail. The integration both of relational data sources, and of other types of data sources accessed by relying on data federation technology are considered. By attending the course, students will learn how to design and build a comprehensive data integration solution, possibly exploiting existing data access and data federation technologies.
Teaching Format	Frontal lectures, exercises, and labs.
Required Readings	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. Z. Zhang, R. Zhang (2008). Multimedia Data Mining: A Systematic Introduction to Concepts and Theory. CRC Press LLC. Additional material (slides, notes of the lecturers) will be made available before each lesson.
Supplementary Readings	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.

Course Module

Course Constituent Title	Data Profiling

Course Code	73056B
Scientific-Disciplinary Sector	INFO-01/A
Language	English
Lecturers	Dr. Alessandro Mosca, Alessandro.Mosca@unibz.it https://www.unibz.it/en/faculties/engineering/academic-staff/person/29044
Teaching Assistant	
Semester	First semester
CP	6
Responsible Lecturer	
Teaching Hours	40
Lab Hours	20
Individual Study Hours	90
Planned Office Hours	
Contents Summary	• Patterns and violation detection • Correlation coefficients • Elementary data analysis • Exact and approximate value dependencies detection • Normalization via data profiling • Non-relational data profiling • Data profiling for query optimisation
Course Topics	The Data Profiling module considers a variety of problems related to the profiling of relational and non-relational data. It first overviews techniques for the exploratory analysis of relational databases, which are also adopted in preparatory activities of data cleansing and integration, and mining. Widely used statistical measures, such as the Pearson and Spearman correlation coefficients, are discussed and applied to real word examples in the first part of the course. Special attention will be also paid to the Phi-K correlation coefficient, which has been recently proposed to overcome well known limitations of the former two, such as the possibility to analyse correlations between numerical, categorical, interval and ordinal variables. Algorithms for discovering patterns and relational dependencies in tabular data will be presented in detail. In particular, the main part of the course will go through the formal specifications of uniqueness, functional and inclusion

	dependencies in relational data, with a special focus on state-of-the-art algorithms and their complexity. If time permits, the last part of the course will discuss the application of the introduced profiling techniques to the relational query optimization problem, and to high level introduction of recent profiling methods that can be applied to RDF data.
Teaching Format	Frontal lectures and labs.
Required Readings	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. Z. Zhang, R. Zhang (2008). Multimedia Data Mining: A Systematic Introduction to Concepts and Theory. CRC Press LLC. Additional material (slides, notes of the lecturers) will be made available before each lesson.
Supplementary Readings	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.