Course guide
270228 - TAED2 - Advanced Topics in Data Engineering 2

Last modified: 19/07/2023

Unit in charge: Barcelona School of Informatics
Teaching unit: 747 - ESSI - Department of Service and Information System Engineering.
744 - ENTEL - Department of Network Engineering.

Degree: BACHELOR'S DEGREE IN DATA SCIENCE AND ENGINEERING (Syllabus 2017). (Compulsory subject).

Academic year: 2023 **ECTS Credits:** 6.0 **Languages:** Catalan, English

LECTURER

Coordinating lecturer: SILVERIO JUAN MARTÍNEZ FERNÁNDEZ

Others: Primer quadrimestre:
SANTIAGO DEL REY JUAREZ - 11, 12
SILVERIO JUAN MARTÍNEZ FERNÁNDEZ - 11, 12
ESTEVE PALLARES SEGARRA - 11, 12
JAVIER PARRA ARNAU - 11, 12

PRIOR SKILLS

Those given by the subjects of the previous quarters of the degree

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.
CE2. To be able to program solutions to engineering problems: Design efficient algorithmic solutions to a given computational problem, implement them in the form of a robust, structured and maintainable program, and check the validity of the solution.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.
CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.

Generical:
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.

Transversal:
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.

Basic:
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

TEACHING METHODOLOGY

The theoretical contents of the course are taught in the theory classes. These classes are complemented with practical examples and problems that students must solve in the Autonomous Learning hours.

In the laboratory sessions, the knowledge acquired in the theory classes is consolidated by solving problems and developing practices related to the theoretical contents. During the laboratory classes, the teacher will introduce new techniques and will leave an important part of the class for the students to work on the proposed exercises.

LEARNING OBJECTIVES OF THE SUBJECT

1.Interpret the basic concepts of Software Engineering for ML systems, especially in relation to the use and exploitation of MLOps practices.
2.Apply and analyze good software engineering practices related to data science and machine learning projects
3.Apply and analyze MLOps practices to build ML models, fostering reproducibility and quality assurance.
4.Apply and analyze MLOps practices to deploy ML models, fostering API development.
5.Understand the privacy risks associated with browsing and publishing data. To achieve a deeper understanding of the different privacy metrics and their application in different scenarios.
6.Understand the main anonymization algorithms for statistical databases.
7.Evaluate the trade-off between privacy and data usability .
8.Understand the privacy risks in communitacions and the anonymous communication systems.

STUDY LOAD

Type	Hours	Percentage
Hours small group	30,0	20.00
Hours large group	30,0	20.00
Self study	90,0	60.00

Total learning time: 150 h

CONTENTS

Introduction to Software Engineering

Description:
First, the traditional concept of software engineering is presented. Then, the impact of data availability on this traditional concept is analyzed. The resulting software life cycle when considering data is shown. Motivación de la necesidad de ingeniería de software para sistemas ML. Introducción a MLOps y conceptos clave. Ingeniería de requisitos para ML.

Good software engineering practices for data science and machine learning projects

Description:
The complexity and diversity of data science projects and machine learning systems call for engineering techniques to ensure they are built in a robust and future-proof manner. On this chapter we address software engineering best practices for data science projects software including ML components.

MLOps practices to build ML models and manage the quality of the software and its development process

Description:
The complexity and diversity of data science projects and ML systems call for engineering techniques to ensure they are built in a robust and future-proof manner. On this chapter we address software engineering best practices for data science projects software including ML components: version control systems; ML pipeline reproducibility and tracking; software measurement for ML; quality assurance for ML.

MLOps practices to deploy ML models

Description:
The complexity and diversity of ML systems call for engineering techniques to ensure they are deployed in a robust and production-ready manner. On this chapter we address software engineering best practices for ML components: software architecture for ML; deploying ML models; APIs for ML.

Introduction to data privacy and security

Description:
Motivation. Definition of basic concepts. Attackers and trusted parties. Privacy metrics.

Algorithms for data anonymization

Description:
Statistical disclosure control. Measure the risk of disclosure. Microaggregation algorithms. Measurement of privacy-utility trade-off. Case studies.

Privacy in personalised information systems

Description:
User profiles: a measure of privacy risk. Privacy-enhancing technologies.

Security and privacy in communications

Description:
Cryptographic algorithms. Authentication and key management. Anonymous communication systems.

ACTIVITIES

Study of basic concepts of Software Engineering for ML systems (MLOps)

Specific objectives:
1

Related competencies :
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.

Full-or-part-time: 6h
Theory classes: 2h
Laboratory classes: 2h
Self study: 2h

Study of good software engineering practices for data science and machine learning projects

Specific objectives:
2

Related competencies :
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CE2. To be able to program solutions to engineering problems: Design efficient algorithmic solutions to a given computational problem, implement them in the form of a robust, structured and maintainable program, and check the validity of the solution.
CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

Full-or-part-time: 6h
Theory classes: 4h
Self study: 2h

Study of MLOps practices to build ML models and software quality management and its development process

Specific objectives:
3

Related competencies :
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
CE2. To be able to program solutions to engineering problems: Design efficient algorithmic solutions to a given computational problem, implement them in the form of a robust, structured and maintainable program, and check the validity of the solution.
CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.
CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

Full-or-part-time: 6h
Theory classes: 4h
Self study: 2h

Study of MLOps practices to deploy ML models

Specific objectives:
4

Related competencies :
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.
CE2. To be able to program solutions to engineering problems: Design efficient algorithmic solutions to a given computational problem, implement them in the form of a robust, structured and maintainable program, and check the validity of the solution.
CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.
CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

Full-or-part-time: 5h
Theory classes: 3h
Self study: 2h

Practical development of a case study of MLOps practices in the context of ML-based systems

Description:
The student will progressively develop a practice that allows him to exercise the basic concepts introduced in the theory part. It will be developed in teams of 4-5 students. The resulting software, duly documented, will be uploaded to a code repository. The team will present a report, written in English, summarizing the main aspects of the practice, for example, the process of building an ML component of an ML-based system, and an evaluation of the accuracy of the models and algorithms used.

Specific objectives:
2, 3, 4

Related competencies :
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.
CE2. To be able to program solutions to engineering problems: Design efficient algorithmic solutions to a given computational problem, implement them in the form of a robust, structured and maintainable program, and check the validity of the solution.
CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.
CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

Full-or-part-time: 44h 30m
Laboratory classes: 13h
Self study: 31h 30m

First partial exam: Software Engineering part (PARC1)

Description:
Evaluation of the first part of the course

Specific objectives:
1, 2, 3, 4

Related competencies :
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.
CE2. To be able to program solutions to engineering problems: Design efficient algorithmic solutions to a given computational problem, implement them in the form of a robust, structured and maintainable program, and check the validity of the solution.
CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.
CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

Full-or-part-time: 7h
Guided activities: 1h 30m
Self study: 5h 30m

Final Exam (EXF)

Description:
This exam evaluates the two parts of the subject. Students who have failed any of the two partial tests are required. The rest of the students can also apply if they want to improve their grades

Specific objectives:
1, 2, 3, 4, 5, 6, 7, 8

Related competencies :
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.
CE2. To be able to program solutions to engineering problems: Design efficient algorithmic solutions to a given computational problem, implement them in the form of a robust, structured and maintainable program, and check the validity of the solution.
CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.
CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

Full-or-part-time: 2h
Guided activities: 2h

Second partial examination: part of Privacy and Data Security (PARC2)

Description:
Evaluation of the second part of the subject

Specific objectives:
5, 6, 7, 8

Related competencies :
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.
CE2. To be able to program solutions to engineering problems: Design efficient algorithmic solutions to a given computational problem, implement them in the form of a robust, structured and maintainable program, and check the validity of the solution.
CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

Full-or-part-time: 7h
Guided activities: 1h 30m
Self study: 5h 30m

Study of introductory concepts on data privacy and security

Specific objectives:
5, 6, 7, 8

Related competencies :
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.
CE2. To be able to program solutions to engineering problems: Design efficient algorithmic solutions to a given computational problem, implement them in the form of a robust, structured and maintainable program, and check the validity of the solution.
CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

Full-or-part-time: 9h
Theory classes: 4h
Self study: 5h

Practical development of data anonymization algorithms

Specific objectives:
6, 7

Related competencies :
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.
CE2. To be able to program solutions to engineering problems: Design efficient algorithmic solutions to a given computational problem, implement them in the form of a robust, structured and maintainable program, and check the validity of the solution.
CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

Full-or-part-time: 37h 30m
Laboratory classes: 15h
Self study: 22h 30m

Study of risks and privacy technologies for personalised information systems

Specific objectives:
5, 7

Related competencies :
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.
CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

Full-or-part-time: 10h
Theory classes: 4h
Self study: 6h

Study of mechanisms and technologies for communications security and privacy

Full-or-part-time: 10h
Theory classes: 4h
Self study: 6h

GRADING SYSTEM

The evaluation is structured according to the two parts of the course: software engineering (PART1) and data privacy (PART2).

For the first part, the grade is calculated by weighting the grade of a theoretical exam (weight 40%) with the grade of the laboratory of this part of the subject (weight 60%)
PART1 = 40% PARC1 + 60% LABO1
- PARC1: Examination at the end of the first part of the course.
- LABO1: Delivery of the laboratory project at the end of the first part of the course.

For the second part, the grade is calculated by weighting the grade of a theoretical exam (weight 50%) with the grade of the practical of this part of the subject (weight 50%)
PART2 = 50% PARC2 + 50% LABO2
- PARC2: Examination at the end of the second part of the course.
- LABO2: Delivery of practices of the second part of the course.

The final grade of the course, NOTA-FIN, is calculated as the arithmetic mean of the two parts of the course:
NOTA-FIN = 50% PART1 + 50% PART2

In case of not passing the course by the evaluation of mid-term exams, there is an evaluation by a final exam, where the mid-term exams are released if they are passed.

BIBLIOGRAPHY

Basic:
- Hulten, G. Building intelligent systems : a guide to machine learning engineering [on line]. California: Apress, 2018 [Consultation: 21/09/2023]. Available on: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=5357977. ISBN 9781484234327.
- Templ, M. Statistical disclosure control for microdata: methods and applications in R. Cham, Switzerland: Springer International Publishing AG, 2017. ISBN 9783319502724.

Complementary:
- Kästner, Christian. Machine Learning in Production: From Models to Products [on line]. 2021 [Consultation: 21/09/2023]. Available on: https://ckaestne.medium.com/machine-learning-in-production-book-overview-63be62393581.
- Lanubile, Filippo; Martínez-Fernández, Silverio; Quaranta, Luigi. "Teaching MLOps in Higher Education through Project-Based Learning". SEET@ICSE [on line]. Available on: https://arxiv.org/pdf/2302.01048.pdf.- Torra i Reventós, V. Data privacy: foundations, new developments and the big data challenge. Skövde: Springer International Publishing, 2017. ISBN 9783319573564.
- Navarro-Arribas, G.; Torra i Reventós, V. (eds.). Advanced research in data privacy [on line]. Cham: Springer International Publishing, 2015 [Consultation: 04/08/2023]. Available on: https://link-springer-com.recursos.biblioteca.upc.edu/book/10.1007/978-3-319-09885-2. ISBN 9783319098852.