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

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: 2021 ECTS Credits: 6.0 Languages: Catalan, Spanish, English

LECTURER

Coordinating lecturer: JAVIER FRANCH GUTIÉRREZ - SILVERIO JUAN MARTÍNEZ FERNÁNDEZ - JORGE FORNE MUÑOZ

Others: Primer quadrimestre:
JORGE FORNE MUÑOZ - 11
JAVIER FRANCH GUTIÉRREZ - 11
SILVERIO JUAN MARTÍNEZ FERNÁNDEZ - 11
ESTEVE PALLARES SEGARRA - 11

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, especially in relation to the use and exploitation of data
2. Apply and analyze concepts and methods concerning the use of data from the development process in the quality management of the software system
3. Apply and analyze good software engineering practices related to data science and machine learning projects
4. Describe concepts and methods related to the use of data obtained during the use of the system, in order to plan new evolutionary versions or to self-adapt systems at runtime in response to changes.
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 communications and the anonymous communication systems.

**STUDY LOAD**

<table>
<thead>
<tr>
<th>Type</th>
<th>Hours</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self study</td>
<td>90,0</td>
<td>60.00</td>
</tr>
<tr>
<td>Hours large group</td>
<td>36,0</td>
<td>24.00</td>
</tr>
<tr>
<td>Hours small group</td>
<td>24,0</td>
<td>16.00</td>
</tr>
</tbody>
</table>

Total learning time: 150 h

**CONTENTS**

**Introduction to Software Engineering**

*Description:*
First, the traditional concept of software engineering is presented. Phases. Methodologies: waterfall, agile; hybrid. Development environment: tools.

Then, the impact of data availability on this traditional concept is analyzed. The resulting software life cycle when considering data is shown.

**Quality management of the software and its development process**

*Description:*
A classic problem in software development is to ensure basic levels of quality, both in reference to the system itself (maintainability, reliability,...) and in the production process (team productivity, resource management,...). The analysis of data from the software repositories used in the production process (e.g., code repositories, problem management tools) allows a faster and more reliable discovery of these problems and the implementation of mitigation strategies.
### 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.

### Software version planning and self-adaptive systems

**Description:**
A key problem in software development is the evolution of the system in response to new needs. The analysis of the data obtained during the use of the system by its users, including their explicit comments, makes it possible to discover their real needs, which sometimes even they are not fully aware of. This topic describes the problem and reviews some basic techniques. More and more we find software systems that need to be aware of their context in order to provide a correct service. This restriction requires them to monitor context data continuously, discover significant changes and react at runtime (eventually, almost in real time). This topic describes the problem and reviews some basic techniques.

### Introduction to data privacy and security

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

### Algorithms for data anonymization

**Description:**

### Privacy in personalised information systems

**Description:**

### Security and privacy in communications

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

**Study of introductory concepts of data driven software engineering**

**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 data driven methods for software quality management and its development process**

**Specific objectives:**
2

**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.
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
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.
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.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy
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:** 10h
Theory classes: 4h
Self study: 6h
Study of good software engineering practices for data science and machine learning projects

Specific objectives:

3

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.
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
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.
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.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy
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: 10h
Theory classes: 4h
Self study: 6h

Study of data-driven methods for software evolution and the self-adaptation of systems at runtime

Specific objectives:

4

Related competencies:
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.
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.
CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.
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.
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.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy
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: 10h
Theory classes: 4h
Self study: 6h
Practical development of a case study of data-based methods in the context of Software Engineering

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 3 (exceptionally, a team of 4 students if the group has an odd dimension). 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 data mining process used, and an evaluation of the accuracy of the models and algorithms used.

Specific objectives:
2, 3

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.
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
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.
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.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy
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: 31h
Laboratory classes: 13h
Self study: 18h
First partial exam: Software Engineering part (PARC1)

Description:
Evaluation of the first part of the course

Specific objectives:
1, 2, 3, 4

Related competencies:
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.
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.
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
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.
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.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy.
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: 7h 30m
Guided activities: 2h
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:
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.
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.
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.
CE7. Demonstrate knowledge and ability to apply the necessary tools for the storage, processing and access to data.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
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.
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.

Full-or-part-time: 3h
Guided activities: 3h
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:
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.
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.
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
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.
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.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy
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: 7h 30m
Guided activities: 2h
Self study: 5h 30m
Study of introductory concepts on data privacy and security

Specific objectives:
5, 6, 7, 8

Related competencies:
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.
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.
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
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.
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.
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.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy.
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: 9h
Theory classes: 4h
Self study: 5h
Practical development of data anonymization algorithms

Specific objectives:
6, 7

Related competencies:
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.
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.
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
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.
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.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy
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: 36h
Laboratory classes: 15h
Self study: 21h
Study of risks and privacy technologies for personalised information systems

Specific objectives:
5, 7

Related competencies:
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.
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.
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
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.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy.
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: 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 final grade is based on four tests:
- Examination at the end of the first block of the course (PARC1)
- Examination at the end of the second block of the course (PARC2)
- Final exam, composed of two parts, one for each block of the subject (EXF1, EXF2)
- Delivery of practice at the end of the first block of the course (LABO1)
- Delivery of practices of the second block of the course (LABO2)

The final grade of the course, NOTE-END, is calculated as:

\[ \text{NOTE-END} = 50\% \text{ TEO} + 25\% \text{ LABO1} + 25\% \text{ LABO2} \]

The theory note is calculated as:

1) Evaluation by mid-sections: minimum of 4.0 for each mid-section and average passed. Then the grade will be the arithmetic mean of the grades of each partial.

If (PARC1 \geq 4.0 and PARC2 \geq 4.0) and ((PARC1 + PARC2) / 2) \geq 5.0 then TEO = (PARC1 + PARC2) / 2

2) Otherwise: Evaluation by final exam: there is a minimum of 4.0 to each block and a half approved; the partial ones release material if approved.

if (NOTE-Block 1 \geq 4.0 and NOTE-BLOC2 \geq 4.0) and ((NOTE-Block 1 + NOTE-BLOC2) / 2) \geq 5.0
then TEO = (NOTE-Block 1 + NOTE-BLOC2) / 2
otherwise TEO = min ((NOTE-BLOCK 1 + NOTE-BLOC2) / 2, 4.5), where

if PARC \{y\} \geq 5.0 then NOTE-BLOCK \{y\} = max (PARC \{y\}, EXF \{y\})
otherwise NOTE-BLOCK \{y\} = EXF \{y\}

BIBLIOGRAPHY

Basic:

Complementary: