Course guide  
270954 - PODS - Process-Oriented Data Science

Unit in charge: Barcelona School of Informatics  
Teaching unit: 723 - CS - Department of Computer Science.

Degree: MASTER'S DEGREE IN DATA SCIENCE (Syllabus 2021). (Compulsory subject).

Academic year: 2022  
ECTS Credits: 6.0  
Languages: English

LECTURER

Coordinating lecturer: CARLOS ESCOLANO PEINADO

Others: Primer quadrimestre: CARLOS ESCOLANO PEINADO - 11, 12  
AYSEL PALACIOS ARDANUY - 11, 12

PRIOR SKILLS

Thorough understanding of computing in general; good command of several programming languages; basic ability to formalize mathematically issues in informatics engineering.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CE13. Identify the main threats related to ethics and data privacy in a data science project (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
CE5. Model, design, and implement complex data systems, including data visualization
CE6. Design the Data Science process and apply scientific methodologies to obtain conclusions about populations and make decisions accordingly, from both structured and unstructured data and potentially stored in heterogeneous formats.
CE7. Identify the limitations imposed by data quality in a data science problem and apply techniques to smooth their impact
CE9. Apply appropriate methods for the analysis of non-traditional data formats, such as processes and graphs, within the scope of data science

Generical:
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CG3. Define, design and implement complex systems that cover all phases in data science projects

Transversal:
CT1. ENTREPRENEURSHIP AND INNOVATION: Know and understand the organization of a company and the sciences that govern its activity; Have the ability to understand labor standards and the relationships between planning, industrial and commercial strategies, quality and profit. Being aware of and understanding the mechanisms on which scientific research is based, as well as the mechanisms and instruments for transferring results among socio-economic agents involved in research, development and innovation processes.
CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CT5. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.
Basic:
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.
CB8. Capability to communicate their conclusions, and the knowledge and rationale underpinning these, to both skilled and unskilled public in a clear and unambiguous way.
CB9. Possession of the learning skills that enable the students to continue studying in a way that will be mainly self-directed or autonomous.

TEACHING METHODOLOGY
Theory sessions that may include problem solving sessions with or without a programming component, practical sessions with open-source or commercial process oriented data science software, development of a case study.

LEARNING OBJECTIVES OF THE SUBJECT
1. To be aware of the theoretical and practical set of problems that constitute process oriented data science, and to understand the main algorithms to tackle it: both at the conceptual level and at the level of their application through some of the current tools and libraries.
2. To acquire and demonstrate an ability to put to work the knowledge obtained during the course, and to relate it to the organizational and team perspectives as a process oriented data science project running in a real organization.

STUDY LOAD

<table>
<thead>
<tr>
<th>Type</th>
<th>Hours</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self study</td>
<td>96.0</td>
<td>64.00</td>
</tr>
<tr>
<td>Guided activities</td>
<td>6.0</td>
<td>4.00</td>
</tr>
<tr>
<td>Hours large group</td>
<td>24.0</td>
<td>16.00</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

- **Process models and event data**
  
  **Description:**
  Describing the concepts of process models and event data

- **Automatic process model discovery**
  
  **Description:**
  Overview on the different techniques to mine process models from event data

- **Conformance checking of process models and event data**
  
  **Description:**
  The main techniques to relate observed and modeled behavior will be introduced
<table>
<thead>
<tr>
<th>Evidence-based process enhancement grounded in event data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
</tr>
<tr>
<td>Techniques to improve and extend process models from event data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assorted advanced techniques and applications</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
</tr>
<tr>
<td>Advanced techniques to solve particular applications will be described, including online and multi-perspective techniques.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methodology for process oriented data science projects</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description:</strong></td>
</tr>
<tr>
<td>A description of the life-cycle of a PODS project will be provided.</td>
</tr>
</tbody>
</table>
ACTIVITIES

Process models and event data

Description:
This activity will introduce process models to specify processes in organizations, and data that talk about events that originate in the execution of processes.

Specific objectives:
1

Related competencies:
CG3. Define, design and implement complex systems that cover all phases in data science projects
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CE7. Identify the limitations imposed by data quality in a data science problem and apply techniques to smooth their impact
CE5. Model, design, and implement complex data systems, including data visualization
CE9. Apply appropriate methods for the analysis of non-traditional data formats, such as processes and graphs, within the scope of data science
CE6. Design the Data Science process and apply scientific methodologies to obtain conclusions about populations and make decisions accordingly, from both structured and unstructured data and potentially stored in heterogeneous formats.
CE13. Identify the main threats related to ethics and data privacy in a data science project (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
CT5. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CB9. Possession of the learning skills that enable the students to continue studying in a way that will be mainly self-directed or autonomous.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.

Full-or-part-time: 24h 54m
Theory classes: 5h
Laboratory classes: 4h
Self study: 15h 54m
Automatic process model discovery

Description:
In this activity, various techniques will be introduced that extract process models in various formalisms from event data.

Specific objectives:

1

Related competencies:
CG3. Define, design and implement complex systems that cover all phases in data science projects
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CE7. Identify the limitations imposed by data quality in a data science problem and apply techniques to smooth their impact
CE5. Model, design, and implement complex data systems, including data visualization
CE9. Apply appropriate methods for the analysis of non-traditional data formats, such as processes and graphs, within the scope of data science
CE6. Design the Data Science process and apply scientific methodologies to obtain conclusions about populations and make decisions accordingly, from both structured and unstructured data and potentially stored in heterogeneous formats.
CE13. Identify the main threats related to ethics and data privacy in a data science project (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
CT5. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CB9. Possession of the learning skills that enable the students to continue studying in a way that will be mainly self-directed or autonomous.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.

Full-or-part-time: 28h
Theory classes: 6h
Laboratory classes: 6h
Self study: 16h
Conformance checking of process models and event data

Description:
In this activity algorithms will be introduced for the relation between modeled and observed process behavior.

Specific objectives:
1

Related competencies:
CG3. Define, design and implement complex systems that cover all phases in data science projects
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CE7. Identify the limitations imposed by data quality in a data science problem and apply techniques to smooth their impact
CE5. Model, design, and implement complex data systems, including data visualization
CE9. Apply appropriate methods for the analysis of non-traditional data formats, such as processes and graphs, within the scope of data science
CE6. Design the Data Science process and apply scientific methodologies to obtain conclusions about populations and make decisions accordingly, from both structured and unstructured data and potentially stored in heterogeneous formats.
CE13. Identify the main threats related to ethics and data privacy in a data science project (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
CT5. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CB9. Possession of the learning skills that enable the students to continue studying in a way that will be mainly self-directed or autonomous.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.

Full-or-part-time: 28h
Theory classes: 6h
Laboratory classes: 6h
Self study: 16h
Evidence-based process enhancement grounded in event data

Description:
In this activity, techniques will be presented to use event data to project and enhance process models and event logs.

Specific objectives:
1

Related competencies:
CG3. Define, design and implement complex systems that cover all phases in data science projects
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CE7. Identify the limitations imposed by data quality in a data science problem and apply techniques to smooth their impact
CE5. Model, design, and implement complex data systems, including data visualization
CE9. Apply appropriate methods for the analysis of non-traditional data formats, such as processes and graphs, within the scope of data science
CE6. Design the Data Science process and apply scientific methodologies to obtain conclusions about populations and make decisions accordingly, from both structured and unstructured data and potentially stored in heterogeneous formats.
CE13. Identify the main threats related to ethics and data privacy in a data science project (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
CT5. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CB9. Possession of the learning skills that enable the students to continue studying in a way that will be mainly self-directed or autonomous.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.

Full-or-part-time: 24h
Theory classes: 4h
Laboratory classes: 4h
Self study: 16h
Assorted advanced techniques and applications

Description:
Assorted techniques for solving real-life process oriented data sciences problems

Specific objectives:
1, 2

Related competencies:
CG3. Define, design and implement complex systems that cover all phases in data science projects
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CE7. Identify the limitations imposed by data quality in a data science problem and apply techniques to smooth their impact
CE5. Model, design, and implement complex data systems, including data visualization
CE9. Apply appropriate methods for the analysis of non-traditional data formats, such as processes and graphs, within the scope of data science
CE6. Design the Data Science process and apply scientific methodologies to obtain conclusions about populations and make decisions accordingly, from both structured and unstructured data and potentially stored in heterogeneous formats.
CE13. Identify the main threats related to ethics and data privacy in a data science project (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
CT5. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CT1. ENTREPRENEURSHIP AND INNOVATION: Know and understand the organization of a company and the sciences that govern its activity; Have the ability to understand labor standards and the relationships between planning, industrial and commercial strategies, quality and profit. Being aware of and understanding the mechanisms on which scientific research is based, as well as the mechanisms and instruments for transferring results among socio-economic agents involved in research, development and innovation processes.
CB9. Possession of the learning skills that enable the students to continue studying in a way that will be mainly self-directed or autonomous.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.
CB8. Capability to communicate their conclusions, and the knowledge and rationale underpinning these, to both skilled and unskilled public in a clear and unambiguous way.
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.

Full-or-part-time: 24h
Theory classes: 4h
Laboratory classes: 4h
Self study: 16h
Methodology for process oriented data science projects

Description:
Overview of how to manage a PODS project

Specific objectives:
2

Related competencies:
CT1. ENTREPRENEURSHIP AND INNOVATION: Know and understand the organization of a company and the sciences that govern its activity; Have the ability to understand labor standards and the relationships between planning, industrial and commercial strategies, quality and profit. Being aware of and understanding the mechanisms on which scientific research is based, as well as the mechanisms and instruments for transferring results among socio-economic agents involved in research, development and innovation processes.

CB8. Capability to communicate their conclusions, and the knowledge and rationale underpinning these, to both skilled and unskilled public in a clear and unambiguous way.

CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.

Full-or-part-time: 21h
Theory classes: 2h
Laboratory classes: 3h
Self study: 16h

GRADING SYSTEM

The evaluation of the subject consists of two elements: final exam (60%), practical assessments (40%).

The final exam will contain questions and problems about the theoretical contents that are explained in the theory classes.

The practical assessments will be guided assessments that will be conducted during the lab classes on various process mining tools and platforms. Assessments can be done in pairs or individually.

BIBLIOGRAPHY

Basic:

Complementary:

RESOURCES

Hyperlink:
- https://www.tf-pm.org/