### 200645 - PBDE - Statistical Programming and Databases

**Coordinating unit:** 200 - FME - School of Mathematics and Statistics

**Teaching unit:**
- 723 - CS - Department of Computer Science
- 715 - EIO - Department of Statistics and Operations Research
- 1004 - UB - (ENG)Universitat de Barcelona
- 707 - ESAII - Department of Automatic Control

**Academic year:** 2018

**Degree:** MASTER'S DEGREE IN STATISTICS AND OPERATIONS RESEARCH (Syllabus 2013). (Teaching unit Optional)

**ECTS credits:** 5

**Teaching languages:** English

### Teaching staff

**Coordinator:** JOAQUIN GABARRÓ VALLÉS

**Others:** Segon quadrimestre:
- JOAQUIN GABARRÓ VALLÉS - A
- ALEXANDRE PERERA LLUNA - A

### Prior skills

Non compulsory subject.

The student has already developed several abilities in Statistics and/or Operations Research previously.

A B2 (Cambridge First Certificate, TOEFL PBT >550) level of English is required.

### Degree competences to which the subject contributes

**Specific:**

3. CE-1. Ability to design and manage the collection of information and coding, handling, storing and processing it.

4. CE-4. Ability to use different inference procedures to answer questions, identifying the properties of different estimation methods and their advantages and disadvantages, tailored to a specific situation and a specific context.

5. CE-5. Ability to formulate and solve real problems of decision-making in different application areas being able to choose the statistical method and the optimization algorithm more suitable in every occasion.

6. CE-6. Ability to use appropriate software to perform the necessary calculations in solving a problem.

7. CE-7. Ability to understand statistical and operations research papers of an advanced level. Know the research procedures for both the production of new knowledge and its transmission.

8. CE-8. Ability to discuss the validity, scope and relevance of these solutions and be able to present and defend their conclusions.

**Transversal:**

2. EFFECTIVE USE OF INFORMATION RESOURCES: Managing the acquisition, structuring, analysis and display of data and information in the chosen area of specialisation and critically assessing the results obtained.

10. 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.

11. TEAMWORK: Being able to work in an interdisciplinary team, whether as a member or as a leader, with the aim of contributing to projects pragmatically and responsibly and making commitments in view of the resources that are available.
This course presents and discusses tools and techniques to prepare students to data science. Main concepts introduced in class will cover tools and methods for data storage and analysis, including relational DB, noSQL and distributed databases, scientific computing, applied machine learning and deep learning with Python. Scala and Spark will also be considered. The course consists of two main modules.

MODULE 1:
First modulus will cover a crash course for scientific python for data analysis. This crash course will include include four main stages:
* Introduction to python language as a tool. ipython, ipython notebook (jupyter), basic types, mutability and immutability and object oriented programming.
* Short introduction to numerical python and matplotlib for graphical visualization.
* Introduction to scientific kits for data analysis with machine learning. Principal components analysis, clustering and supervised analysis with multivariate data.
* Introduction to Deep Learning with Python.

MODULE 2:
We introduce the Scala language and the Spark architecture.
* Scala as a functional language and the Scala collections.
* Spark and RDD (Resilient Distributed Data Sets).
* Spark and SQL.
* Introduction to MLib.

Learning objectives of the subject
This course presents and discusses tools and techniques to prepare students to data science. Main concepts introduced in class will cover tools and methods for data storage and analysis, including relational DB, noSQL and distributed databases, scientific computing, applied machine learning and deep learning with Python. Scala and Spark will also be considered. The course consists of two main modules.

Study load
<table>
<thead>
<tr>
<th>Total learning time: 125h</th>
<th>Hours large group: 30h</th>
<th>24.00%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hours medium group: 0h</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Hours small group: 15h</td>
<td>12.00%</td>
</tr>
<tr>
<td></td>
<td>Guided activities: 0h</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Self study: 80h</td>
<td>64.00%</td>
</tr>
</tbody>
</table>
# Content

## Introduction to Python

**Learning time:** 1h  
Theory classes: 1h

**Description:**
- a. Why Python?  
- b. Python History  
- c. Installing Python  
- d. Python resources

## Working with Python

**Learning time:** 1h  
Theory classes: 1h

**Description:**
- a. Workflow  
- b. ipython vs. CLI  
- c. Text Editors  
- d. IDEs  
- e. Notebook

## Getting started with Python

**Learning time:** 1h  
Theory classes: 1h

**Description:**
- a. Introduction  
- b. Getting Help  
- c. Basic types  
- d. Mutable and in-mutable  
- e. Assignment operator  
- f. Controlling execution flow  
- g. Exception handling
### Functions and Object Oriented Programming

**Learning time:** 1h  
Theory classes: 1h

**Description:**  
a. Defining Functions  
b. Input and Output  
c. Standard Library  
d. Object-oriented programming

### Introduction to NumPy

**Learning time:** 2h  
Theory classes: 2h

**Description:**  
a. Overview  
b. Arrays  
c. Operations on arrays  
d. Advanced arrays (ndarrays)  
e. Notes on Performance (%timeit in ipython)

### Matplotlib

**Learning time:** 2h  
Theory classes: 2h

**Description:**  
a. Introduction  
b. Figures and Subplots  
c. Axes and Further Control of Figures  
d. Other Plot Types  
e. Animations

### Python scikits

**Learning time:** 1h  
Theory classes: 1h

**Description:**  
a. Introduction  
b. scikit-timeseries
<table>
<thead>
<tr>
<th>Course</th>
<th>Description</th>
<th>Learning time</th>
</tr>
</thead>
<tbody>
<tr>
<td>scikit-learn</td>
<td>Description:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a. Datasets</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b. Sample generators</td>
<td></td>
</tr>
<tr>
<td></td>
<td>c. Unsupervised Learning</td>
<td></td>
</tr>
<tr>
<td></td>
<td>d. Supervised Learning</td>
<td></td>
</tr>
<tr>
<td></td>
<td>i. Linear and Quadratic Discriminant Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ii. Nearest Neighbors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>iii. Support Vector Machines</td>
<td></td>
</tr>
<tr>
<td></td>
<td>e. Feature Selection</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Learning time: 8h</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Theory classes: 8h</td>
<td></td>
</tr>
<tr>
<td>Practical Introduction to Scikit-learn</td>
<td>Description:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a. Solving an eigenfaces problem</td>
<td></td>
</tr>
<tr>
<td></td>
<td>i. Goals</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ii. Data description</td>
<td></td>
</tr>
<tr>
<td></td>
<td>iii. Initial Classes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>iv. Importing data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b. Unsupervised analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>i. Descriptive Statistics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ii. Principal Component Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>iii. Clustering</td>
<td></td>
</tr>
<tr>
<td></td>
<td>c. Supervised Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>i. k-Nearest Neighbors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ii. Support Vector Classification</td>
<td></td>
</tr>
<tr>
<td></td>
<td>iii. Cross validation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Learning time: 5h 30m</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Theory classes: 5h 30m</td>
<td></td>
</tr>
<tr>
<td>Introduction to Zeppelin, Scala &amp; Functional Programming</td>
<td>Description:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a. Immutable &amp; Mutable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b. Lists and maps, filters, reductions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>c. Map reduce</td>
<td></td>
</tr>
<tr>
<td></td>
<td>d. Other collections, Streams</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Learning time: 5h</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Theory classes: 5h</td>
<td></td>
</tr>
</tbody>
</table>
## Spark architecture & Spark Core

**Description:**
- Spark architecture: in particular Spark Core
- Spark context
- Types of operations: transformations and actions
- RDD: Resilient Distributed Data Sets
- Closure of a function

**Learning time:** 5h  
Theory classes: 5h

## Spark SQL

**Description:**
- Reading from a file.
- Spark Data Frame.
- Selection, filters, grouping, sorting.
- Window operations
- SQL

**Learning time:** 7h 30m  
Theory classes: 7h 30m

## Spark: MLlib

**Description:**
- Description of the MLlib.
- Labeled Points and features
- Linear Regression Example

**Learning time:** 5h  
Theory classes: 5h

## Qualification system

Final grade will be composed by:
- 1/4 Written exam first module
- 1/4 Written exam first module
- 1/2 Final practical assignment on large databases integrating concepts from both modules
Bibliography

Basic:


Complementary: