

# ANX-PR/CL/001-01

## LEARNING GUIDE

### SUBJECT

**103000830 - Big Data**

### DEGREE PROGRAMME

10AX - Master Universitario Innovación Digital Ciencia de Datos Itinerario Health

### ACADEMIC YEAR & SEMESTER

2020/21 - Semester 1

## Index

---

### Learning guide

1. Description.....	1
2. Faculty.....	1
3. Skills and learning outcomes .....	2
4. Brief description of the subject and syllabus.....	3
5. Schedule.....	5
6. Activities and assessment criteria.....	8
7. Teaching resources.....	10
8. Other information.....	12

## 1. Description

### 1.1. Subject details

<b>Name of the subject</b>	103000830 - Big Data
<b>No of credits</b>	6 ECTS
<b>Type</b>	Optional
<b>Academic year of the programme</b>	First year
<b>Semester of tuition</b>	Semester 1
<b>Tuition period</b>	September-January
<b>Tuition languages</b>	English
<b>Degree programme</b>	10AX - Master Universitario Innovación Digital Ciencia de Datos Itinerario Health
<b>Centre</b>	10 - Escuela Técnica Superior de Ingenieros Informáticos
<b>Academic year</b>	2020-21

## 2. Faculty

### 2.1. Faculty members with subject teaching role

<b>Name and surname</b>	<b>Office/Room</b>	<b>Email</b>	<b>Tutoring hours *</b>
Antonio Latorre De La Fuente (Subject coordinator)	4202	a.latorre@upm.es	Sin horario.
Pablo Toharia Rabasco	4102	pablo.toharia@upm.es	Sin horario.
Jesus Montes Sanchez	4204	jesus.montes@upm.es	Sin horario.

\* The tutoring schedule is indicative and subject to possible changes. Please check tutoring times with the faculty member in charge.

### 3. Skills and learning outcomes \*

---

#### 3.1. Skills to be learned

CB06 - Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación

CB07 - Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio

CB10 - Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

CE-HMDA02 - Capacidad para aplicar técnicas para la generación de visualizaciones adecuadas para el análisis y la exploración de datos en un contexto médico, y para la correcta comunicación de los resultados del análisis

CE-HMDA03 - Capacidad para seleccionar las técnicas y herramientas para visualización de grandes cantidades de datos más adecuadas para resolver un determinado problema en el campo de la salud

CE-HMDA05 - Capacidad para usar herramientas de procesamiento de big data tanto en online como en modo batch

CE-HMDA06 - Capacidad para extraer, integrar y consultar datos heterogéneos en escenarios clínicos

#### 3.2. Learning outcomes

RA4 - Ser capaz de procesar datos masivos

RA3 - Conocer cómo se aplican las técnicas de computación científica en algún campo específico de ciencia o ingeniería

RA2 - Conocer técnicas de visualización y procesos de análisis de datos, y de programación, diseño y depuración de algoritmos, para computación de altas prestaciones

\* The Learning Guides should reflect the Skills and Learning Outcomes in the same way as indicated in the Degree Verification Memory. For this reason, they have not been translated into English and appear in Spanish.

## 4. Brief description of the subject and syllabus

---

### 4.1. Brief description of the subject

This course will allow the student to gain the fundamentals for the analytical visualization of large volumes of data. With an eminently practical approach, the technologies and fundamentals necessary to successfully accomplish the whole data analysis process will be presented in the context of Big Data, from the raw data to its visualization, through the models derived from them.

### 4.2. Syllabus

1. Introduction to Big Data
  - 1.1. Architectures and applications
  - 1.2. Data types
  - 1.3. Visual analytics
2. Big Data Ecosystem
3. Big Data Technologies
  - 3.1. Technological Challenges
  - 3.2. Basic solution: gfs + MapReduce
  - 3.3. Hadoop (hdfs + yarn)
  - 3.4. Pig
  - 3.5. Hive
  - 3.6. Beyond MapReduce
    - 3.6.1. Tez
    - 3.6.2. Spark
    - 3.6.3. Flink
4. Spark
  - 4.1. Spark Basics

#### 4.2. Brief Introduction to Scala

#### 4.3. Spark Applications

#### 4.4. Spark SQL

### 5. Machine Learning with Spark

#### 5.1. Brief review of Machine Learning basics

#### 5.2. Spark MLlib

### 6. Information Visualization

#### 6.1. Information Visualization Fundamentals

#### 6.2. Data Abstractions

#### 6.3. Tasks Abstractions

#### 6.4. Interaction Techniques and Visual Encoding

#### 6.5. Design Methods

#### 6.6. Visualization Examples Analysis

#### 6.7. Lessons Learnt

## 5. Schedule

### 5.1. Subject schedule\*

Week	Face-to-face classroom activities	Face-to-face laboratory activities	Distant / On-line	Assessment activities
1	<b>Lesson 1</b> Duration: 02:00  <b>Lesson 2</b> Duration: 01:00		<b>Lesson 1</b> Duration: 02:00  <b>Lesson 2</b> Duration: 01:00	
2	<b>Lesson 2</b> Duration: 01:00  <b>Lesson 2</b> Duration: 02:00		<b>Lesson 2</b> Duration: 01:00  <b>Lesson 2</b> Duration: 02:00	
3	<b>Lesson 3</b> Duration: 01:00	<b>Practical Work 1</b> Duration: 02:00	<b>Lesson 3</b> Duration: 01:00  <b>Practical Work 1</b> Duration: 02:00	
4	<b>Lesson 3</b> Duration: 01:00	<b>Practical Work 1</b> Duration: 02:00	<b>Lesson 3</b> Duration: 01:00  <b>Practical Work 1</b> Duration: 02:00	
5	<b>Lesson 4</b> Duration: 01:00  <b>Lesson 4</b> Duration: 01:00	<b>Practical Work 1</b> Duration: 01:00	<b>Lesson 4</b> Duration: 01:00  <b>Lesson 4</b> Duration: 01:00  <b>Practical Work 1</b> Duration: 01:00	
6	<b>Lesson 4</b> Duration: 01:00  <b>Lesson 4</b> Duration: 01:00	<b>Practical Work 1</b> Duration: 01:00	<b>Lesson 4</b> Duration: 01:00  <b>Lesson 4</b> Duration: 01:00  <b>Practical Work 1</b> Duration: 01:00	

7	<b>Lesson 4</b> Duration: 01:00	<b>Practical Work 1</b> Duration: 02:00	<b>Lesson 4</b> Duration: 01:00  <b>Practical Work 1</b> Duration: 02:00	
8	<b>Lesson 4</b> Duration: 01:00	<b>Practical Work 1</b> Duration: 02:00	<b>Lesson 4</b> Duration: 01:00  <b>Practical Work 1</b> Duration: 02:00	
9	<b>Lesson 5</b> Duration: 01:00  <b>Lesson 5</b> Duration: 01:00	<b>Practical Work 1</b> Duration: 01:00	<b>Lesson 5</b> Duration: 01:00  <b>Lesson 5</b> Duration: 01:00  <b>Practical Work 1</b> Duration: 01:00	<b>First Assignment Deadline</b>  Continuous assessment and final examination Not Presential Duration: 00:00
10	<b>Lesson 6</b> Duration: 01:00  <b>Lesson 6</b> Duration: 01:00	<b>Practical Work 2</b> Duration: 01:00	<b>Lesson 6</b> Duration: 01:00  <b>Lesson 6</b> Duration: 01:00  <b>Practical Work 2</b> Duration: 01:00	
11	<b>Lesson 6</b> Duration: 02:00	<b>Practical Work 2</b> Duration: 01:00	<b>Lesson 6</b> Duration: 02:00  <b>Practical Work 2</b> Duration: 01:00	
12	<b>Lesson 6</b> Duration: 01:00  <b>Lesson 6</b> Duration: 01:00	<b>Practical Work 2</b> Duration: 01:00	<b>Lesson 6</b> Duration: 01:00  <b>Lesson 6</b> Duration: 01:00  <b>Practical Work 2</b> Duration: 01:00	



13	<b>Lesson 6</b> Duration: 02:00	<b>Practical Work 2</b> Duration: 01:00	<b>Lesson 6</b> Duration: 02:00  <b>Practical Work 2</b> Duration: 01:00	
14	<b>Lesson 6</b> Duration: 01:00	<b>Practical Work 2</b> Duration: 02:00	<b>Lesson 6</b> Duration: 01:00  <b>Practical Work 2</b> Duration: 02:00	
15	<b>Lesson 6</b> Duration: 01:00	<b>Practical Work 2</b> Duration: 02:00	<b>Lesson 6</b> Duration: 01:00  <b>Practical Work 2</b> Duration: 02:00	
16		<b>Practical Work 2</b> Duration: 03:00	<b>Practical Work 2</b> Duration: 03:00	<b>Second Assignment Deadline</b>  Continuous assessment and final examination Not Presential Duration: 00:00
17				<b>Final Exam</b>  Continuous assessment and final examination Presential Duration: 01:00

Depending on the programme study plan, total values will be calculated according to the ECTS credit unit as 26/27 hours of student face-to-face contact and independent study time.

\* The schedule is based on an a priori planning of the subject; it might be modified during the academic year, especially considering the COVID19 evolution.

## 6. Activities and assessment criteria

### 6.1. Assessment activities

#### 6.1.1. Continuous assessment

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
9	First Assignment Deadline		No Presential	00:00	40%	4 / 10	CB07 CE-HMDA05 CE-HMDA06 CB10 CE-HMDA03 CB06 CE-HMDA02
16	Second Assignment Deadline		No Presential	00:00	40%	4 / 10	CB07 CE-HMDA05 CE-HMDA06 CB10 CE-HMDA03 CB06 CE-HMDA02
17	Final Exam		Face-to-face	01:00	20%	4 / 10	CB07 CE-HMDA05 CE-HMDA06 CB10 CE-HMDA03 CB06 CE-HMDA02

#### 6.1.2. Final examination

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
9	First Assignment Deadline		No Presential	00:00	40%	4 / 10	CB07 CE-HMDA05 CE-HMDA06 CB10 CE-HMDA03 CB06 CE-HMDA02
16	Second Assignment Deadline		No Presential	00:00	40%	4 / 10	CB07 CE-HMDA05 CE-HMDA06 CB10 CE-HMDA03 CB06 CE-HMDA02

17	Final Exam		Face-to-face	01:00	20%	4 / 10	CB07 CE-HMDA05 CE-HMDA06 CB10 CE-HMDA03 CB06 CE-HMDA02
----	------------	--	--------------	-------	-----	--------	--

### 6.1.3. Referred (re-sit) examination

No se ha definido la evaluación extraordinaria.

## 6.2. Assessment criteria

### Evaluation criteria

This section covers the evaluation criteria for this course. All the students enrolled in this course will be subject, by default, to the continuous evaluation scheme. For this reason, this learning guide will be focused on this approach and details all the evaluation activities in the timeline of the course. Those students interested in the final examination evaluation scheme are referred to the next section of this document.

This course will be evaluated in two ways:

- **Final exam.** At the end of the course there will be a final exam covering all the contents presented during the course.
- **Practical work.** These assignments will be presented during the course, at class, in the dates detailed in the timeline of the course. There will be some classes devoted to these assignments, where the students will count with the support of the instructor, that should be, in general, complemented with autonomous work by the student. The deadlines for the assignment are spread through the term, as shown in the timeline of the course. No late assignments will be accepted for evaluation.

The **final grade** for this course will be computed as follows: 20% for the final exam, 40% for the first assignment, and, finally, 40% for the second assignment. To pass the course, a **minimum score of 4** is required for each of these parts and a **grand mean of 5** is needed combining these three items of evaluation.

## Final exam evaluation

This evaluation scheme will be only offered if the current regulations of the UPM requires it and the procedure to opt for this type of evaluation will be subject to the instructions given by the school. Please, refer to <http://www.fi.upm.es/?pagina=1147> for additional information.

In general, the regulations for this evaluation scheme will be the same as for the continuous evaluation option. In particular:

- The students will have to conduct the same practical works without the in-class support of the instructors.
- The deadlines for the assignments will be the same as for the continuous evaluation scheme.

## Extraordinary evaluation in July

If the student does not succeed in this course, she will have to repeat those parts not passed in the ordinary evaluation. There will be a new call for the final exam as well as a new deadline common for all the assignments of the course.

## 7. Teaching resources

### 7.1. Teaching resources for the subject

Name	Type	Notes
Book 1	Bibliography	Jiawei Han, Micheline Kamber, Data Mining : Concepts and Techniques, 2nd edition, Morgan Kaufmann, ISBN 1558609016, 2006.
Book 2	Bibliography	Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson Addison Wesley, ISBN: 0321321367, 2005
Book 3	Bibliography	Ian Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, ISBN: 0120884070, 2005.

Book 4	Bibliography	Ian Witten, Eibe Frank, Mark Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann, ISBN: 978-0-12-374856-0, 2011.
Book 5	Bibliography	Keim, D., Kohlhammer, J., Ellis, G., Mansmann, F. Mastering the information age. Solving problems with visual analytics 2010 Eurographics Association.
Book 6	Bibliography	Tamara Munzner. Visualization Analysis and Design. A K Peters Visualization Series. CRC Press. Nov. 2014.
Book 7	Bibliography	Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia. Learning Spark: Lightning-Fast Big Data Analysis. O'Reilly Media. 2015.
Book 8	Bibliography	Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills. Advanced Analytics with Spark: Patterns for Learning from Data at Scale. O'Reilly Media. 2015.
Spark documentation	Web resource	<a href="http://spark.apache.org/docs/latest/">http://spark.apache.org/docs/latest/</a>
Assigned class	Equipment	
Web site of the course	Web resource	UPM Moodle
Hive documentation	Web resource	<a href="https://cwiki.apache.org/confluence/display/Hive/Home">https://cwiki.apache.org/confluence/display/Hive/Home</a>

## 8. Other information

---

### 8.1. Other information about the subject

This course is jointly offered with the EIT-Digital Master in Data Science and lectures are delivered in English.