COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI ICADE CIHS

COURSE SYLLABUS 2019-2020

GENERAL INFORMATION

Course information	
Name	Big Data
Code	DTC-MIC-522
Degree	Máster Universitario en Ingeniería Industrial + Máster en Industria Conectada [2nd year]
Semester	2nd (Spring)
ECTS credits	3.0
Type	Compulsory
Department	Telematics and Computer Science
Coordinator	Carlos Morrás Ruiz-Falcó

Instructor	
Name	Carlos Morrás Ruiz-Falcó
Department	Telematics and Computing
Office	
e-mail	cmorras@icai.comillas.edu
Phone	
Office hours	Arrange an appointment through email.

DETAILED INFORMATION

Contextualization of the course

Contribution to the professional profile of the degree

Big data is a new technology that plays a leading role in all processes where there is a large volume of data or where artificial intelligence or machine learning algorithms are required. It is allowing to highly increase efficiency and effectiveness and enabling new business models that were previously impossible or unimaginable. In particular, industry is an area where big data is a key technology, both because artificial intelligence and machine learning techniques result in great process improvements, and because the huge volume of data generated by the Internet of Things (IoT) is otherwise challenging to process and analyze in order to support decision-making.

During the course, students will learn the most relevant aspects of big data technology, go through real cases from leading industrial actors, and carry out some lab practices based on these cases. By the end of the course, students should have enough knowledge of big data technology to understand its potential, and have developed an informed criterion to determine when and how to use it in a professional context.

Prerequisites

Students willing to take this course should be familiar with basic probability and statistics, machine learning, and undergraduate-level programming. Previous experience with Python is also desired although not strictly required.

CONTENTS

Contents
Theory
Unit 1. Introduction to big data: The value of the data in decision making
Unit 2. Big data origin, motivation and history
2.1 Why now? The digital society 2.2 Exponential and disruptive technologies 2.3 The big data revolution: The new paradigm of problem solving, from program-oriented to data-oriented 2.4 The problem of Google and Yahoo! 2.5 Open source and the Nutch project 2.6 Hadoop and evolution 2.7 Current trends: Internet of Things (IoT)
Unit 3. How does big data work?
3.1 Distribution and parallelism. Programming features 3.2 HDFS file system 3.3 MapReduce and Yarn 3.4 Redundancy and fault tolerance
Unit 4. Fundamental characteristics of big data
4.1 The four V's 4.2 Structured and unstructured information: Impact of big data
Unit 5. Ethical aspects of data, big data and artificial intelligence
5.1 Ethics, privacy and data protection 5.2 Data compliance and other ethical aspects 5.3 Principles of GDPR
Unit 6. Big data ecosystem and basic architecture
6.1 Basic Hadoop ecosystem 6.2 Data acquisition: Scoop, Flume and Kafka 6.3 Search and data processing: Hive and Hbase 6.4 Administration and monitoring of a big data cluster
Unit 7. Big data and cloud computing
7.1 Overview and main actors (AWS, Azure and Google)
Unit 8. How to develop a big data case
8.1 Marketing 8.2 Commerce and retail 8.3 Banking 8.4 Industry 8.5 Big data ecosystem

Laboratory[1]
Practice 0. Big data cluster
In the first lab session, students will become familiar with the big data cluster that they will use throughout the course.
Practice 1. Predictive maintenance
In this practice, students will have to solve the technical part of the real predictive maintenance case study that will be covered previously in the lectures. This includes, analyzing and preparing the dataset, and debugging and implementing the necessary algorithms.
Practice 2. Production prediction
In this practice, students will have to solve the technical part of the real production prediction case study that will be covered previously in the lectures. This includes, analyzing and preparing the dataset, and debugging and implementing the necessary algorithms.
Practice 3. Plastic pollution
In this practice, students will have to solve the technical part of the real plastic pollution case study that will be covered previously in the lectures. This includes, analyzing and preparing the dataset, and debugging and implementing the necessary algorithms.

Competences and learning outcomes

Competences[2]

General competences	
CG1.	Have acquired advanced knowledge and demonstrated, in a research and technological or highly specialized context, a detailed and well-founded understanding of the theoretical and practical aspects, as well as of the work methodology in one or more fields of study. *Haber adquirido conocimientos avanzados y demostrado, en un contexto de investigación científica y tecnológica o altamente especializado, una comprensión detallada y fundamentada de los aspectos teóricos y prácticos y de la metodología de trabajo en uno o más campos de estudio.*
CG2.	Know how to apply and integrate their knowledge, understanding, scientific rationale, and problem-solving skills to new and imprecisely defined environments, including highly specialized multidisciplinary research and professional contexts. *Saber aplicar e integrar sus conocimientos, la comprensión de estos, su fundamentación científica y sus capacidades de resolución de problemas en entornos nuevos y definidos de forma imprecisa, incluyendo contextos de carácter multidisciplinar tanto investigadores como profesionales altamente especializados.*
CG3.	Know how to evaluate and select the appropriate scientific theory and the precise methodology of their fields of study in order to formulate judgements based on incomplete or limited information, including, when necessary and pertinent, a discussion on the social or ethical responsibility linked to the solution proposed in each case. *Saber evaluar y seleccionar la teoría científica adecuada y la metodología precisa de sus campos de estudio para formular juicios a partir de información incompleta o limitada incluyendo, cuando sea preciso y pertinente, una reflexión sobre la responsabilidad social o ética ligada a la solución que se proponga en cada caso.*

[1] Practice topics are tentative and may be replaced by others of similar nature.

[2] Competences in English are a free translation of the official Spanish version.

CG4.	Be able to predict and control the evolution of complex situations through the development of new and innovative work methodologies adapted to the scientific/research, technological or specific professional field, in general multidisciplinary, in which they develop their activity. *Ser capaces de predecir y controlar la evolución de situaciones complejas mediante el desarrollo de nuevas e innovadoras metodologías de trabajo adaptadas al ámbito científico/investigador, tecnológico o profesional concreto, en general multidisciplinar, en el que se desarrolle su actividad.*
CG5.	Be able to transmit in a clear and unambiguous manner, to specialist and non-specialist audiences, results from scientific and technological research or state-of-the-art innovation, as well as the most relevant foundations that support them. *Saber transmitir de un modo claro y sin ambigüedades, a un público especializado o no, resultados procedentes de la investigación científica y tecnológica o del ámbito de la innovación más avanzada, así como los fundamentos más relevantes sobre los que se sustentan.*
CG6.	Have developed sufficient autonomy to participate in research projects and scientific or technological collaborations within their thematic area, in interdisciplinary contexts and, where appropriate, with a high knowledge transfer component. *Haber desarrollado la autonomía suficiente para participar en proyectos de investigación y colaboraciones científicas o tecnológicas dentro de su ámbito temático, en contextos interdisciplinares y, en su caso, con una alta componente de transferencia del conocimiento.*
CG7.	Being able to take responsibility for their own professional development and their specialization in one or more fields of study. *Ser capaces de asumir la responsabilidad de su propio desarrollo profesional y de su especialización en uno o más campos de estudio.*
Specific competences	
CE5.	Know the techniques used to extract information from large datasets, as well as the different platforms, tools, and languages that make it possible. *Conocer las técnicas para extraer información de grandes conjuntos de datos, así como las diferentes plataformas, herramientas y lenguajes que lo hacen posible.*
Learning outcomes	
By the end of the course students should:	
RA1.	Be able to communicate with big data specialists using common language.
RA2.	List the characteristics and advantages of big data systems.
RA3.	Understand and propose general and industrial applications of big data.
RA4.	Calculate the business impact of big data and analytics.
RA5.	Be capable of addressing simple analytical projects.
RA6.	Develop analytical algorithms.
RA7.	Properly assess the impact of big data on people, businesses and society, including ethical and moral aspects.
RA8.	Know the limits and limitations of data use and be familiar with privacy legislation.

TEACHING METHODOLOGY

General methodological aspects	
To ensure useful and practical learning, theoretical classes will be combined with master classes that reflect the reality of the market. Real case studies with also be studied from business and technical perspectives, some of which will be used in practical sessions.	
In-class activities	**Competences**
▪ **Lectures:** The lecturer will introduce the fundamental concepts of each unit, along with some practical recommendations, and will go through worked examples to support the explanation. Active participation will be encouraged by raising open questions to foster discussion and by proposing quizzes and short application exercises to be solved in class.	CG1, CG3, CG4, CG7, CE5
▪ **Practical sessions:** Under the instructor's supervision, students will apply the concepts learned in the lectures to real cases, in order to face and solve implementation problems that typically arise.	CG1, CG2, CG3, CG5, CG6, CG7, CE5
▪ **Tutoring** for groups or individual students will be organized upon request.	–
Out-of-class activities	**Competences**
▪ Personal study of the course material and resolution of the proposed exercises.	CG1, CG3, CG4, CG7, CE5
▪ Practical session preparation to make the most of in-class time.	CG1
▪ Practical results analysis and report writing.	CG2, CG5, CE5

ASSESSMENT AND GRADING CRITERIA

Assessment activities	Grading criteria	Weight
Quizzes	▪ Understanding of the theoretical concepts.	10%
Final exam	▪ Understanding of the theoretical concepts. ▪ Application of these concepts to problem-solving. ▪ Critical analysis of numerical exercises' results.	55%
Practical assignments	▪ Application of theoretical concepts to real problem-solving. ▪ Ability to understand results in real environment. ▪ Written communication skills.	35%

GRADING AND COURSE RULES

Grading

Regular assessment

- **Theory** will account for 65%, of which:
  - Quizzes: 10%
  - Final exam: 55%
- **Lab** (practical assignments) will account for the remaining 35%

In order to pass the course, the weighted average mark must be greater or equal to 5 out of 10 points, the mark of the final exam must be greater or equal to 4 out of 10 points, and the laboratory mark must be at least 5 out of 10 points. Otherwise, the final grade will be the lower of the three marks.

Retake
Lab marks will be preserved as long as they result in a passing grade. Otherwise a project will have to be developed and handed in. In addition, all students will take a final exam. The resulting grade will be computed as follows: ▪ **Theory** will account for 65%, of which: • Quizzes: 10% • Final exam: 55% ▪ **Lab** will account for the remaining 35% • If the student passed the lab during regular assessment ▪ Practical assignments: 35% • Otherwise ▪ Final project: 35% As in the regular assessment period, in order to pass the course, the weighted average mark must be greater or equal to 5 out of 10 points, the mark of the final exam must be greater or equal to 4 out of 10 points, and the mark of the laboratory must be at least 5 out of 10 points. Otherwise, the final grade will be the lower of the three marks.
Course rules
▪ Class attendance is mandatory according to Article 93 of the General Regulations (Reglamento General) of Comillas Pontifical University and Article 6 of the Academic Rules (Normas Académicas) of the ICAI School of Engineering. Not complying with this requirement may have the following consequences: - Students who fail to attend more than 15% of the lectures may be denied the right to take the final exam during the regular assessment period. - Regarding practice, absence to more than 15% of the sessions can result in losing the right to take the final exam of the regular assessment period and the retake. Missed sessions must be made up for credit. ▪ Students who commit an irregularity in any graded activity will receive a mark of zero in the activity and disciplinary procedure will follow (cf. Article 168 of the General Regulations (Reglamento General) of Comillas Pontifical University).

WORK PLAN AND SCHEDULE[3]

In and out-of-class activities	Date/Periodicity	Deadline
Final exam	After the lecture period	–
Practice sessions	From week 3	–
Review and self-study of the concepts covered in the lectures	After each lesson	–
Practice preparation	Before every lab session	–
Practice report writing	–	One week after the end of each session

[3] A detailed work plan of the subject can be found in the course summary sheet (see following page). Nevertheless, this schedule is tentative and may vary to accommodate the rhythm of the class.

STUDENT WORK-TIME SUMMARY			
IN-CLASS HOURS			
Lectures		Practical sessions	
20		10	
OUT-OF-CLASS HOURS			
Self-study	Practice preparation	Report writing	Homework assignments
32	9.5	7.5	11
		ECTS credits:	3 (90 hours)

BIBLIOGRAPHY

Basic bibliography

- Slides prepared by the lecturer (available in Moodlerooms)
- V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, 1st Ed., John Murray Publishers, 2013. ISBN-13: 978-1-848-54792-6
- P. C. Verhoef, E. Kooge, and N. Walk, *Creating Value with Big Data Analytics: Making Smarter Marketing Decisions*, 1st Ed., Routledge, 2016. ISBN-13: 978-1-138-83797-3

Complementary bibliography

- G. Orwell, *1984*, 1949. ISBN-13: 978-8-499-89094-4
- M. Lewis, Moneyball: The Art of Winning an Unfair Game, 2004. ISBN-13: 978-0-393-32481-5. Alternatively, you can watch the movie: B. Miller (Director), *Moneyball* (133 minutes), Sony Pictures, 2011.
- D. Plotkin, *Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance*, 1st Ed., Morgan Kaufmann, 2013. ISBN-13: 978-0-124-10389-4
- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [Online].
  Available: *https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679*
- R. Colmenarejo Fernández, *Una ética para Big data. Introducción a la gestión ética de datos masivos* (in Spanish), 1st Ed., Editorial UOC, 2017. ISBN-13: 978-8-491-16940-6.
- J. Gottschau (Director), *Facebookistan* (59 minutes), Danish Film Institute, 2015. Available on Movistar+ and Netflix.
- K. Amer and J. Noujaim (Directors), *The Great Hack* (135 minutes), Netflix, 2019.

	In-class activities				Out-of-class activities				Learning outcomes
Week	Time [h]	Lecture	Laboratory	Assessment	Time [h]	Self-study	Practice preparation and report writing	Other activities	Code
1	2	Course overview (0.5h) Introduction to big data (1.5h)			2	Review and self-study (2h)			
	2	Big data origin, motivation and history (2h)			3.5	Review and self-study (2h)		Film watching (1.5h)	
2	2	How does big data work? (1.8h)		Quiz (0.2 h)	3	Review and self-study (2h)		Paper homework (1h)	
	2	Big data case 1 (2h)			4.5	Review and self-study (2h)	Practice preparation (2.5h)		
3	2		Practice 1 (2h)		2.5		Report writing (2.5h)		
	2	Fundamental characteristics of big data (2h)			4	Review and self-study (2h)		Film watching (2h)	
4	2	Ethical aspects of data, big data and AI (1.8h)		Quiz (0.2 h)	6	Review and self-study (2h)		Paper homework (4h)	
	2	Big data ecosystem and basic architecture (2h)			2	Review and self-study (2h)			
5	2		Practice 0 (2h)		2		Practice preparation (2h)		
	2	Big data and cloud computing (1h) How to develop a big data case (1h)			2.5	Review and self-study (2h)		Quiz (0.5h)	
6	2	How to develop a big data case (2h)			4	Review and self-study (2h)		Paper homework (2h)	
	2	Big data case 2 (2h)			4.5	Review and self-study (2h)	Practice preparation (2.5h)		
7	2		Practice 2 (2h)		2.5		Report writing (2.5h)		
	2	Big data case 3 (2h)			4.5	Review and self-study (2h)	Practice preparation (2.5h)		
8	2		Practice 3 (2h)		2.5		Report writing (2.5h)		
				Final exam[4]	10	Final exam preparation (10h)			

[4] The final exam will be held on the first week of March.