Syllabus: Data Analytics & Big Data

Course Basic Information:

Course duration:	20 weeks course / 800 hours
Course modality:	Full time
Course days and times:	Monday to Friday from 9 to 17hs.
Instructors:	Ester Bernardó, LinkedIn
	Dani Castejón, LinkedIn
Course web page:	www.ubiqum.com/courses/data-analytics
Languages:	Online platform and content is in English
	Mentors can speak English/Spanish/Catalan.
Location:	Barcelona

Main learning concepts of our Data Analytics Course:

Module 1: Understanding Customers	
Tools	**Key Points**
Weka Excel	• Preprocessing Data (Filters, Missing Values) • Data Mining • Decision Trees • Classification / Regression Algorithms (J48/C5.0, M5P) • Presentation Skills to non-technical Audience
Module 2: Predicting Profitability and Customer Preferences	
Tools	**Key Points**
Weka Excel R	• Normalization, Distance, Correlation • Machine Learning • Compare Items (k-NN/IBk) • Predictive Revenue Model (k-NN, M5P…) • Class Prediction Model (J48, k-NN)

	• **R** (Caret, RWeka)
Module 3: Deep Analytics and Visualization	
Tools R Weka	**Key Points** • **R** Visualitzation (ggplot2) • **R** Data Processing (dplyr, tidyr) • **R** Time Series and Forecast • Indoor Locationing - Wifi Fingerprint (k-NN and others) • **R** Machine Learning
Module 4: Big Data - Web Mining	
Tools Weka AWS	**Key Points** • Web Mining • **AWS** Elastic Map Reduce • **AWS** CLI • Sentiment analysis

Course Description:

This course is designed for students who have no previous knowledge of data analytics but wish to acquire these skills in a short period of time. These students will learn how to analyze large data sets and identify patterns that will improve any company's and organization decision-making process. After completing the course, they will be able to:

- Capture, categorize, simplify, normalize and prepare data to be processed
- Work with and analyze large data sets
- Visually represent analysis's conclusions to technical and non technical audiences
- Use the most common algorithms, to make sense of large amounts of data, which are applicable to most business and management problems.
- Learn R programing language.

At the end of the program, you will have a professional portfolio of projects and real experience with data analysis that will give you the necessary confidence to be successful as a Data Analyst.

Course Objectives:

Almost every company and organization collects data about their operations to better understand how to make internal improvements. As the amount of data collected increases, it is more difficult to analyze this data manually. There is a growing tendency at bigger companies to automate the collection of large quantities of data (Big Data) to discover behavior patterns and better understand their internal processes.

The collection of data (Data Mining) has several applications, including reducing the amount of time needed to make decisions and cutting error margins. Data analytics' prediction abilities can improve a company's marketing, help understand customer behavior, and prevent fraud. Therefore, Data Mining is applicable to every department in any company.

Using an SCC (Story Centered Curriculum) methodology, you will be able to complete tasks like a Data Analyst. You will use machine learning techniques to analyze online sales and market studies to find purchase patterns and customer preferences. Data analysis helps a sales department improve decisions, decide which products to offer, and how to offer them.

Course Pre-requisites:

Those interested in Data Analytics should have some prior experience in (or be willing to work with):

- Science (testing or formulating hypotheses)
- Statistics (working with numbers and statistical methodology) and

- Programming (use of algorithms).

This course is well suited to those with a degree in Social and natural Sciences, Engineering or Mathematics.

Course Grading:

Grades will be determined from:

- attendance (40%)
- programming exercises and deliverables at the end of every task (40%)
- and final presentation (20%)

Evaluation scheme is subject to change with a prior notice. Attendance will be checked regularly. Missing classes frequently will automatically drop student out of class.

Per our methodology there will be no exams and no master lessons. The student is committed to attend to class and work on their assigned tasks and deliver them via our online platform according to the program schedule.

Course Details:

Module 1: Data Analytics: Understanding Customers	
Week 1-3	In this course students will be working for Blackwell Electronics as data analysts. The students' job is to use data mining and machine-learning techniques to investigate the patterns in Blackwell's sales data and provide insight into customer buying trends and preferences. The inferences students draw from the patterns in the data will help the business make data-driven decisions about sales and marketing activities and understand the

	relationship between customer demographics and purchasing behavior. Finally, students will present to management, explaining their insights and suggestions for data mining process improvements.
Module 2: Data Analytics: Predicting Profitability and Customer Preferences	
Week 4-6	Students will continue to work as data analysts at Blackwell Electronics. Students' job is to extend Blackwell's application of data mining methods to develop predictive models. In this course, students will use both Weka and the R statistical programming language augmented with machine learning packages to predict which potential new products that the sales team is considering adding to Blackwell's current product mix will be the most profitable. Next, students will create a model to predict which brand of computer products Blackwell customers prefer based on customer demographics collected from a marketing survey. Finally, students will present to management, explaining their insights and suggestions for data mining process improvements.
Module 3: Big Data - Web Mining	
Week 7-12	You will learn the algorithmic and organizational skills required to scale data analysis to large server farms, computing clouds, and the web, including an understanding of the design and implementation differences between single-computer and cloud-scale programs, analytics, and data processing. You will also gain a deep knowledge of predictive data analysis, ranging from discovering patterns and correlations in data to making predictions and estimating their accuracy. As part of this process, you will master fundamentals of

scaling up data analysis to a large cloud computing platform where you will become proficient in working with map-reduce-based systems and leveraging the computing power of the cloud to prepare very large data sets for deep analysis, as well as learn how to train and apply modern machine learning algorithms to large processed datasets. You will also learn how to identify the types of business questions for which Big Data analyses can provide significant insights in support of business decision-making.

These skills are applicable to every big data project and will enable you to present yourself as a key asset in solving business problems that require data analysis. After successfully completing this course, you will be able to:

- Frame and organize business questions in ways that can be answered through Big Data analysis and meet the information needs of a project.
- Identify types of questions for which data analysis cannot provide accurate information.
- Acquire, process, and analyze extremely large data sets using data mining methods to discover patterns or do data exploration.
- Install, run, and apply machine learning tools to different kinds of data.
- Set up and run Elastic Map Reduce (EMR) instances for data analysis.
- Formulate the machine learning process to answer domain-specific problems.
- Interpret the results of data analysis and data mining to make predictions and to establish the reliability of those predictions

	• Communicate results to management and other non-technical audiences. • Provision and utilize AWS EC2 instances. • Build predictive models using either Weka or R.
Module 4: Deep Analytics and Visualization	
Week 13-16	Increasingly, technology companies are applying data analytics techniques to the masses of data generated by devices such as smart phones, appliances, vehicles, electric meters, *et cetera* – the "Internet of Things". The ability to deal with data of these types will prove to be a high-demand skill for data analysts. In this course, students will be working for an Internet of Things technology start-up that wants to use Data Analytics to solve two difficult problems in the physical world: • Smart energy usage: Modeling patterns of energy usage by time of day and day of the year in a typical residence whose electrical system is monitored by multiple sub-meters. • Indoor locationing: Determining a person's physical position in a multi-building indoor space using wifi fingerprinting. Students will learn to use the R statistical programming language to perform visualizations, then to generate descriptive statistics and predictive models using time series regression techniques and statistical classifiers. Finally, students will present the results to the start-up's management, explaining strengths and weaknesses of the approaches that were implemented and making suggestions for further improvement.

Designed by:

This program were designed in collaboration with Dr. Jaime Carbonell and Dr. Ravi Starzl of Carnegie Mellon University.

Dr. Jaime Carbonell is the Director of the Language Technologies Institute and Allen Newell Professor of Computer Science at Carnegie Mellon University, where he pioneered the PhD and MS degrees in Language Technologies. His current research includes machine learning, computational proteomics, data mining (primarily in healthcare and finance), text mining and machine translation. Dr. Carbonell has served on multiple governmental advisory committees such as the Human Genome Committee of the National Institutes of Health, the Oakridge National Laboratories Scientific Advisory Board, the National Institute of Standards and Technology Interactive Systems Scientific Advisory Board, and the German National Artificial Intelligence (DFKI) Scientific Advisory Board. He has published more than 300 papers and books, and supervised more than 45 PhD dissertations. He received SB degrees in Physics and Mathematics from MIT, and MS and PhD degrees in Computer Science from Yale University.

Dr. Ravi Starzl is a Systems Scientist in the Language Technologies Institute at Carnegie Mellon University. He is an expert in the computational analysis and modeling of complex information-driven systems, with experience in such diverse domains as biological systems, financial systems, and Internet topologies. Dr. Starzl has extensive experience with the computational and mathematical methods integral to the effective acquisition, management, and utilization of extremely large amounts of information. Having led several massive data analysis projects with data sets as large as 100+TB, Dr. Starzl is fluent in the methods of Big Data analytics, and is an active researcher in the area of parallelization of machine learning methods for Big Data. He received his doctoral degree in Language and Information Technologies from the School of Computer Science at Carnegie

Mellon University in 2012. At In addition to his research at CMU, Dr. Starzl develops and teaches classes on the topics of Big Data, biotechnology, and advanced software development.He has also participated in the founding, growth, and sale of several biotechnology and high-tech startups.

Programme methodology:

This programme is based on Story Centered Curriculum (SCC) methodology. This involves advanced simulation techniques of real situations. There are no theoretical classes or rote learning study sessions to pass exams. SCC puts you in a motivating scenario based on a real professional situation in which you perform the same tasks as actual professionals, using the same tools, meaning that you can fit easily into a real work team when the time comes.

You will be working in a team, as is usually the case in the world of work. A comprehensive range of supporting material and the help needed to complete the tasks are available on-line. You will be supervised by a tutor, who is always available to answer any questions or clear up any doubts, and who will assess your performance and advise on the project "deliverables".

SCC is an educational methodology that enables us to offer a learning-by-doing approach in all its complexity and scope. This methodology has been successfully used for more than 10 years at the Pittsburgh and Mountain View, California campuses of the Carnegie Mellon University in the USA, where many students have completed programmes developed using this methodology.

Tutors:

At the start of the course you are assigned a tutor who is a professional expert. The tutor will foster teamwork and promote discussion of issues, helping you find

solutions and resolve the difficulties of the project by drawing on your own resources.

They will give you feedback on the “deliverables” for each project at every stage, so that you can continually refine them, learning from your mistakes and achieving the mastery needed for each task.