COMP4103/G54BIG: Big Data

Total Credits: 10

Level: 4

Target students: Part II and III undergraduate students and MSc students in the School of Computer Science. This module is part of the AI, Modelling and Optimisation theme and the Operating systems and Networks theme in CS. Available to JYA/Erasmus students.

Summary of Content:

“Big Data” involves data whose volume, diversity and complexity requires new technologies, algorithms and analyses to extract valuable knowledge, which go beyond the normal processing capabilities of a single computer. The field of Big Data has many different faces such as databases, security and privacy, visualisation, computational infrastructure or data analytics/mining. This module will provide the following concepts:

1. Introduction to Big data: introducing the main principles behind distributed/parallel systems with data intensive applications, identifying key challenges: capture, store, search, analyse and visualise the data.

2. Brief introduction to SQL Databases vs. NoSQL Databases: introduction to NoSQL databases;

3. Big Data frameworks and how to deal with big data: this includes the MapReduce programming model, as well as an overview of recent technologies (Hadoop ecosystem, and Apache Spark). Then, you will learn how to interact with the latest APIs of Apache Spark (RDDs, DataFrames and Datasets) to create distributed programs capable of dealing with big datasets (using Python and/or Scala)

4. Finally, we will dive into the data mining and machine learning part of the course, including data preprocessing approaches (to obtain quality data), distributed machine learning algorithms and data stream algorithms. To do so, you will use the Machine learning library of Apache Spark (MLlib) to understand how some machine learning algorithms (e.g. Decision Trees, Random Forests, k-means) can be deployed at a scale.

Offering School: Computer Science

Convenor:

Name
Dr Isaac Triguero

Taught Semesters:

Semester
Spring UK

Requisites:

Subject	Course Title
COMP3009	Machine Learning

Additional Requirements:

Condition
Only available for Computer Science students
Only available for Year 3 OR 4 students
Only available for PGT Career students
Only available for ERASMUS students only students

Method and Frequency of Class:

Activity	Number of Weeks	Number of sessions	Duration of a session
Computing	12 weeks	1 week	1 hour
Lecture	12 weeks	2 week	1 hour

Method of Assessment:

Assessment Type	Weight	Requirements
Exam	50.00	Two Hour Written Examination
Group Project	50.00	Programming Project in Groups

Assessment Period: Assessed by end of Spring Semester

Learning Outcome:

Knowledge and Understanding:
Understand the importance of the data
The principles that allow the processing of big data sets.
Understand the working and features of existing machine learning algorithms capable of handling big data.
Learn to use the main tools of the big data ecosystem.
The current limitations of big data technologies to allow distributed machine learning.

Intellectual Skills:
Understand complex ideas and relate them to specific problems or questions in the area of parallel computation.
Be able to identify distributed solutions/approaches to handle big datasets with existing technologies.

Professional/Practical Skills:
Hands-on experience with state-of-the-art technologies to handle big data.

Transferable/Key Skills:
Experience in problem solving.
Experience in working in groups.
Retrieve information from appropriate sources (e.g. Spark API).