[ONSITE] BSC Training Course: Introduction to Big Data Analytics @ BSC

Monday, February 5, 2024 - Friday, February 9, 2024

C6 building

Scientific Program

Day 1 (Feb 5th)
9:30 – 13:00 **Introduction to Big Data** (Josep Lluis Berral, Computer Sciences - Data Centric Computing, BSC)
In this session we will introduce the students in the technologies associated with Big Data: data challenges, cloud computing, processing, and internet of things. An overview of the technologies will be provided, both from a technical and from a business model point of view.
11:00 - 11:30 Coffee break
13:00 – 14:00 Lunch Break
14:00 – 16:00 **Practical Data Analytics for Solving Real World Problems** (Patricio Reyes, Researcher, BSC; Maria Teresa Grifa, Data Scientist, Bridgestone EMA)
Data analytics has changed the way we make decisions. We see the benefits and the advances in many fields that go from financial to medical and industrial applications due to the integration of advanced data analytics. In this course we will propose practical tips gained through our experience at BSC in data analytics projects. We will also discover how to overcome some of the most challenging tasks in practical data analytics.
16:00 – 16:30 Coffee break
16:30 – 18:00 **Hands-on** (Patricio Reyes, Researcher, BSC; Maria Teresa Grifa, Data Scientist, Bridgestone EMA)
In this session you will learn how to structure a data analytics project, by following the methodology and the concepts introduced in the previous session. We will guide you through a step-by-step process to set up data science projects and start collaborating with the members of a team.

Day 2 (Feb 6th)
9:30 – 13:00 **Big Data Management** (Albert Abelló, UPC, inLab FIB and Petar Jovanovic, UPC)
Big Data has many definitions and facets, we'll pay attention to the problems we have to face to store it and how we can process it. More specifically, we'll focus on the Apache Hadoop ecosystem and its two basic components, namely HBase and MapReduce engine.
11:00 - 11:30 Coffee break
Hands-on exercise
13:00 – 14:00 Lunch Break
14:00 - 16:00 **NoSQL databases** (Oscar Romero, Dept. of Service and Information System Engineering, UPC-BarcelonaTech)
The relational model has dominated data storage systems since the mid 1970s. However, the changing storage needs over the past decade have given rise to new models for storing data, collectively known as NoSQL. In this presentation, we will focus on two of the most common types of NoSQL databases: document-oriented databases and graph databases and explain the use cases suitable for each of them.
16:00 - 16:30 Coffee break
16:30 - 18:00 **Multidisciplinary research and data analytics: Cultural Heritage** (Maria Cristina Marinescu / Joaquim More / Artem Rashetnikov, Computer Applications in Science&Engineering, BSC)
This session will focus on Cultural Heritage as an example of a field that can really take advantage of integrating, analyzing, and reasoning with large amounts of data from many heterogeneous sources. We will explain how to improve the quality and quantity of open metadata associated with European Cultural Heritage (CH) imagery, starting (mostly) from images of paintings and text. Our ultimate goal is to transcribe insights about culture, symbols and traditions in a knowledge representation accessible to machine learning and artificial intelligence.

Day 3 (Feb 7th)
9:30 – 13:00 **Data Analytics with Apache Spark. Part 1**(Josep Lluis Berral, Computer Sciences - Data Centric Computing, BSC)
Apache Spark has become a consolidated technology for large-scale processing in a fast and general way, with “programmer-friendly” interfaces and official bindings for many of the most used languages (Java, Scala, Python and R), extensive documentation and development tools. This course introduces Apache Spark, as well as some of its core libraries for data manipulation,

machine learning, data streams and graph analytics.
11:00 - 11:30 Coffee break
13:00 – 14:00 Lunch Break
14:00 – 15:30 **Data Analytics with Apache Spark. Part 2** (Josep Lluis Berral, Computer Sciences - Data Centric Computing, BSC)
15:00 - 15:30 Coffee break
15:30 – 17:00 **Hifi-Turb: high-fidelity les/dns data for innovative turbulence models (A H2020 European Project)** (Oriol Lehmkuhl, and Arnau Miró, CASE - Large-scale Computational Fluid Dynamics) The presentation will cover BSC experience in the H2020 project HIFI-TURB: HIGH-FIDELITY LES/DNS DATA FOR INNOVATIVE TURBULENCE MODELS, dealing with big data set exploration, data reduction and the use of novel ML algorithms for turbulence modelling. Modelling turbulent flows using computational fluid dynamics (CFD) has progressed rapidly over the last decades and given rise to significant changes in the design processes of aircraft, cars and ships. The EU-funded HIFI-TURB project is using high-fidelity CFD together with new artificial intelligence and machine learning algorithms to identify important correlations between turbulent quantities with the aim of proposing novel turbulence models. Improved models for complex fluid flows will offer the potential of further reducing energy consumption, emissions and noise of aircraft, ships and cars.

Day 4 (Feb 8th)
10:00 – 11:15 **Bias in Science - Sex and Gender Perspective in Big Data Analytics. Part I and Q&A** (Nataly Buslon, Equity Officer, BSC and Davide Cirillo, Machine Learning For Biomedical Research Recognised Researcher, LS)
This workshop will provide knowledge about the existing biases in Big Data Analytics and Artificial Intelligence (AI) from a multidisciplinary perspective. The main objective is to raise awareness and build a culture towards responsible practices of AI research and development. For this, the social challenges in relation to AI will be reviewed, analyzing the different types of biases in science. In the second instance, ethical aspects of AI are addressed at the international level, which are key in its scientific and social impact. Finally, it will deepen the differences of sex and gender and their specific implications, analyzing from intersectional axes. (Recommended reading literature will be provided before the session)
11:15 - 11:45 Coffee break
11:45 – 13:00 **Bias in Science - Sex and Gender Perspective in Big Data Analytics. Part II and Q&A** (Nataly Buslon, Equity Officer, BSC and Davide Cirillo, Machine Learning For Biomedical Research Recognised Researcher, LS)
13:00 – 14:00 Lunch Break
14:00 – 16:00 **Business Intelligence** (Karina Gibert, Intelligent Data Science and Artificial Intelligence Research Center (IDEAI-UPC))
Data contains information.The session focus on the relationship of concepts such as data mining, business intelligence, big data, data science and the old school of classical statistics. An overview of the data science process as a way to extract added value from data and real cases will be presented as examples of application.
16:00- 16:30 Coffee Break
16:30 – 18:00 **Data analytics, modelling and simulation for solving city challenges. Use case: achieving clean air in Barcelona** (Mari Paz Linares, UPC, inLab FIB and Daniel Rodríguez Rey, Atmospheric Composition Group Recognized Researcher, ES, BSC)
In this session we will present how a combination of data analytics, modelling and simulation techniques can attack certain smart cities challenges and alleviate them. In particular, we will present a use case in Barcelona focused in the air quality problem.

Day 5 (Feb 9th)
9:30 – 13:00 **Data Visualization Theory** (Fernando Cucchietti, Head of Data Analytics and Visualisation, BSC)
Theory
1. Basic concepts
2. Human perception

3. Design
4. Colour
5. Audience / Validation / Bad practices
6. Visualisation design process
11:00 - 11:30 Coffee break
Tools for data visualization
– Tableau
– Data Wrapper
– RawGraphs
– Flourish
– Carto
Data visualisation with d3.js

13:00 END of COURSE