UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL’INFORMAZIONE

Dottorato di Ricerca in Ingegneria dell’Informazione

Doctoral Course

“Big Data Analytics and Signal Processing: Biological Data as a Case Study”

Dr. Gianpaolo Coro

Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo” (ISTI) - CNR- Italy”

Short Abstract:

Big Data analytics is gaining large interest in both public and scientific agendas, because it allows to extract valid information from a large amount of noisy data and to produce valuable information for decision makers. Applications of Big Data analytics can be found in a large variety of domains, including economics, physics, healthcare, and biology. For example, analytics has been used in biology to predict the impact of climate change on species’ distribution, to monitor the effect of overfishing on economy and marine biodiversity, and to prevent ecosystems collapse.

In this course, practical applications of Big Data analytics will be shown, with focus on several signal processing and machine learning-based techniques. The course will clarify the general concepts behind these techniques, with an educational approach making these concepts accessible also to students with intermediate mathematical skills. The examples will regard real cases involving data that would have been hardly human-analyzed and corrected, especially in the domain of biology. The explained techniques will include: automatic periodicities detection, time series forecasting, Artificial Neural Networks, Support Vector Machines, Maximum Entropy, Markov Chains Monte Carlo, geographical maps comparison, global scale species distributions, species invasion prediction.

The above techniques have a general purpose applicability and the students will be able to use them in other domains too. Cloud computing, data sharing, experiments reproducibility, usage of data representation standards and most of the requirements of Big Data analytics systems will be explained and practiced in the context of the new Open Science paradigm. In order to practice with the experiments, the students will use a distributed e-Infrastructure (D4Science) developed at ISTI-CNR and used in a number of international projects. This Web-based platform hides the complexity of implementing Big Data analytics processes from scratch and allows students to concentrate on experiments configuration, results evaluation, and models’ behaviour understanding. For this reason, the course does not require any programming skill and is suited for students in Computer Engineering, Informatics, Telecommunications engineering, Mathematics, Statistics, and Computational Biology.

Course Contents in brief:

- Distributed computing
- Big Data analytics
- e-Infrastructures
- Time series forecasting and periodicities detection
- Machine Learning-based methods
- GIS maps

Total # of hours: 20

References:

Definition of Big Data **- https://en.wikipedia.org/wiki/Big_data**

Computational and Data e-Infrastructures **- https://en.wikipedia.org/wiki/Hybrid_Data_Infrastructure**

The D4Science e-Infrastructure - **https://www.d4science.org/**

Coro, G., Panichi, G., Scarponi, P., Pagano, P. (2017). Cloud computing in a distributed e-infrastructure using the web processing service standard. Concurrency and Computation: Practice and Experience.

Candela, L., Castelli, D., Coro, G., Pagano, P., Sinibaldi, F. (2016). Species distribution modeling in the cloud. Concurrency and Computation: Practice and Experience, 28(4), 1056-1079.

Coro, G., Candela, L., Pagano, P., Italiano, A., Liccardo, L. (2014). Parallelizing the execution of native data mining algorithms for computational biology. Concurrency and Computation: Practice and Experience.

Coro, G., Pagano, P., Ellenbroek, A. (2014). Comparing heterogeneous distribution maps for marine species. GIScience & Remote Sensing, 51(5), 593-611.

Coro, G., Large, S., Magliozzi, C., & Pagano, P. (2016). Analysing and forecasting fisheries time series: purse seine in Indian Ocean as a case study. ICES Journal of Marine Science, 73(10), 2552-2571.

Coro G. Gibbs sampling with JAGS: behind the scenes. http://puma.isti.cnr.it/dfdownload.php?ident=/cnr.isti/2017-B5-001&langver=it&scelta=Metadata

Coro G., Panichi G., Pagano P. A Web application to publish R scripts as-a-Service on a Cloud computing platform. In: Bollettino di Geofisica Teorica e Applicata, vol. 52 article n. 51. Istituto Nazionale di Oceanografia e di Geofisica Sperimentale, 2016.

Coro G., Pasquale P., Napolitano U. Bridging environmental data providers and SeaDataNet DIVA service within a collaborative and distributed e-Infrastructure. In: Bollettino di Geofisica Teorica e Applicata, vol. 52 pp. 23 - 25. Istituto Nazionale di Oceanografia e di Geofisica Sperimentale, 2016.

CV of the Teacher

Gianpaolo Coro is a Physicist with a PhD in Computer Science. His research focuses on Artificial Intelligence and Data Mining. He has been working for 15 years on Machine Learning and Signal Processing with applications to Computational Biology, Brain Computer Interfaces, Language Technologies and Cognitive Sciences. The aim of his research is the study and experimentation of models and methodologies to process biological data, and the application of the results to ecological problems management. His approach relies on distributed e-Infrastructures and uses parallel and distributed computing via Grid and Cloud based technologies through an Open Science approach.

Room and Schedule

Room: *Aula Riunioni del Dipartimento di Ingegneria dell'Informazione, Via G. Caruso 16, Pisa – Ground Floor*

Schedule: 14-18 September 2020

Day1 – e-Infrastructures, Cloud and Distributed computing– 9.00 – 13.00

Day2 – Dimensionality reduction – 9.00 – 13.00

Day3 – Time series analysis and applications – 9.00 – 13.00

Day4 – Machine learning-based modelling and applications – 9.00 – 13.00

Day5 – Tools for Open Science – 9.00 – 13.00