UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL’INFORMAZIONE

Dottorato di Ricerca in Ingegneria dell’Informazione

Doctoral Course

“Signal processing and mining of Big Data: biological data as 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 has demonstrated that it is possible 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. In this last domain, analytics have been used, for example, to predict climate change impact 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 unpractical to be human-analyzed and corrected, especially in the biology domain: time series forecasting, periodicities detection, comparison of geographical distribution maps, assessment of environmental similarities between different areas, global scale species distributions.

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. To execute the experiments, students will use a distributed e-Infrastructure (D4Science) developed at ISTI-CNR, also used in the European Laboratory on Big Data Analytics and Social Mining (SoBigData). This web-based platform hides the complexity of implementing Big Data analytics processes from scratch and allows students to concentrate on experiments configuration and output evaluation, and to understand models’ behaviours. For this reason, the course does not require any programming skill and is suited for students in Computer Engineering, Informatics, Telecommunications engineering, Statistics and Computational Biology.

Course Contents in brief:

- Cloud and distributed computing
- Big Data analysis
- e-Infrastructures
- Large time series forecasting
- Automatic periodicities detection
- Neural Networks
- Large scale probabilistic GIS maps

Total # of hours: 20

References:

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

The SoBigData project **- 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., Webb, T. J., Appeltans, W., Bailly, N., Cattrijsse, A., & Pagano, P. (2015). Classifying degrees of species commonness: North Sea fish as a case study. Ecological Modelling, 312, 272-280.

Coro, G., Magliozzi, C., Ellenbroek, A., & Pagano, P. (2015). Improving data quality to build a robust distribution model for Architeuthis dux. *Ecological Modelling*, *305*, 29-39.

Candela L., Castelli D., Coro G., Pagano P., Sinibaldi F. Species distribution modeling in the cloud. In: Concurrency and Computation-Practice & Experience, Geoffrey C. Fox, David W. Walker, Ed. Wiley, DOI: 10.1002/cpe.3030

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.

Appeltans, W., Pissierssens, P., Coro, G., Italiano, A., Pagano, P., Ellenbroek, A., & Webb, T. (2013). Trendylyzer: a Long-Term Trend Analysis on Biogeographic Data. In Proceedings of the International Conference on Marine Data and Information Systems (IMDIS).

Coro, G., Fortunati, L., & Pagano, P. (2013, June). Deriving fishing monthly effort and caught species from vessel trajectories. In OCEANS-Bergen, 2013 MTS/IEEE (pp. 1-5). IEEE.

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 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 to apply the results to fields in Ecological Modelling, Vessel Monitoring Systems and Ecological Niche Modelling. His approach relies on distributed e-Infrastructures and uses parallel and distributed computing via Grid and Cloud based technologies.

Room and Schedule

Day 1: Aula Riunioni del Dipartimento di Ingegneria dell'Informazione,

Largo Lucio Lazzarino, Pisa

Day 2 - Day 5: Aula Riunioni del Dipartimento di Ingegneria dell'Informazione,

via G. Caruso 16, Pisa – Ground Floor

Schedule: **from 2 to 6 May**

Day1 – Introduction and presentation of the tools: the D4Science e-Infrastructure, Cloud and distributed computing for community-provided processes – 9.00 – 13.00

Day2 –Features analysis: Clustering, Principal Component Analysis and applications– 9.00 – 13.00

Day3 –Large time series analysis: Fourier Transform, Short-Time Fourier Transform, Singular Spectrum Analysis and applications – 9.00 – 13.00

Day4 –Large time Series forecasting: Caterpillar Singular Spectrum Analysis and applications– 9.00 – 13.00

Day5 –Modeling: Neural Networks, Maximum Entropy, Geographical Distribution Maps and applications– 9.00 – 13.00