COMPUTER AND CONTROL ENGINEERING

Big Spatial Data Analytics

Funded By	Dipartimento DAUIN FONDAZIONE CRT CASSA DI RISPARMIO DI TORINO [P.iva/CF:06655250014]
Supervisor	GARZA PAOLO - paolo.garza@polito.it
Contact	GARZA PAOLO - paolo.garza@polito.it APILETTI DANIELE - daniele.apiletti@polito.it
Context of the research activity	We are overloaded by spatial data generated by satellites and IoT devices, among others. Currently, the analyses are mainly focused on one single type of data at a time (e.g., satellite images vs ground-based measurements). Integrating heterogeneous spatial sources into Big data systems capable of building accurate predictive and descriptive ML models is a challenging task that will effectively support domain experts in many areas (e.g., natural hazard management and infrastructure monitoring).
Objectives	The main objective of this research activity consists in designing data-driven algorithms for the analysis of heterogeneous big spatial data (e.g., satellite images, sensor measurements), aiming at generating descriptive and predictive models. The main issues that will be addressed are as follows. Scalability. Spatial data are usually big (e.g., large collections of remote sensing data). Hence, big data solutions are needed to process and analyze them, in particular when historical data are analyzed. Heterogeneity. Several sources, characterized by different data types, are available. Each source represents a facet of the analyzed events and provides a different insight about them. The efficient integration of the available spatial data sources is an important issue that must be addressed to build more accurate predictive and descriptive machine learning models. Near-real time constraint. In several domains, timely responses are needed. For example, to effectively limit the impact of faults in a communication network, timely fault predictions are needed to have enough time to plan compensation actions. Large amounts of streaming data are generated (e.g., alarms of telecommunication networks). The current big data streaming systems (e.g., Spark and Strom) provide limited support for real-time and incremental data mining and machine learning algorithms. Hence, novel algorithms must be designed and implemented. The work plan for the three years is organized as follows. 1st year. Analysis of the state-of-the-art algorithms and data analytics frameworks for big spatial data. Based on the analysis of the state-of-the-art, the pros and cons of the current solutions will be identified and preliminary

	algorithms will be designed to optimize and improve the available approaches. During the first year, descriptive algorithms, based on offline historical data analyses, will be initially designed and validated on real data related to specific domains (e.g., communication network alarms) to understand how to extract fruitful patterns for performing data characterization and medium- and long-term analyses/predictions. 2nd year. The design of incremental and real-time predictive models will be addressed during the second year. For instance, classification algorithms will be designed to automatically classify streaming data in real-time. The algorithms will be tested in a specific application domain. 3rd year. The algorithms designed during the first two years will be improved and generalized to be effectively applied in different domains (domain-agnostic algorithms).

Skills and competencies for the development of the activity	- Good knowledge of Machine learning and Deep learning algorithms - Good knowledge of Big data frameworks (Spark, Hadoop) - Good programming skill