ICT PHD

Research project for a PhD curriculum in ICT – Computer Engineering and Science

**Tutor**: prof. Domenico Beneventano

**(\*) Italian [Industrial] Co-tutor:** dott. Riccardo Martoglia (DataRiver)

**(\*\*) Foreign Co-tutor:**

**Proposed Title of the research:** Knowledge Graphs and Distributed Machine Learning Techniques to Manage the Data Lake of multilingual and multi-alphabetic heritages

**Keywords: (5)**

Digital libraries; Data Lake; Knowledge Graphs; Distributed Machine Learning analytics techniques; humanistic informatics.

**Research objectives: --(max 10 rows)**

The long-term objective of this research is to manage a Data Lake of multilingual and multi-alphabetic heritages. This research will face the challenge posed by the ITSERR project where a Data Lake of archival heritage in non-Latin alphabets needs to be managed and analyzed. Knowledge Graphs (KGs) will be used to semantically describe and manage big data in such a data lake, enhancing the discoverability and reusability of data while enabling automated processing by algorithms; KGs will be used to create a data integration framework where the information from different sources is combined in a new logically centralized graph-like representation. KGs can enhance the performance and scalability of distributed Machine Learning analytics techniques, such as classifiers for categorizing untagged textual documents.

This work was supported by the

**Proposed research activity -- (max 10 rows)**

* State of the art on the main graph models and on techniques for constructing KGs, with emphasis on KGs to integrate heterogeneous data in a semantically rich way.
* A review of the current state-of-the-art on distributed Machine Learning (ML) models, focusing specifically on ML models for analytics.
* Design of techniques for automatic KGs construction in the context of multilingual and multi-alphabetic (non-Latin alphabets) archival heritage
* Design of techniques to use the representation of domain knowledge provided by KGs for improving accuracy/efficiency of distributed machine learning models
* Implementation of the proposed algorithms on platform for distributed processing of Big Data that provides high scalability, such as Apache Spark.
* Participation to relevant international schools and conferences

**Supporting research projects (and Department).**

The research will be conducted in the context of the PNRR project Italian Strengthening of Esfri RI Resilience (ITSERR) funded by the European Union – NextGenerationEU.

**Possible connections with research groups, companies, universities.**

International research group coordinated by "Fondazione per le scienze religiose di Bologna" of the "Big Data, Artificial Intelligence and Religious Studies" research line, in collaboration with Biblioteca La Pira, Palermo, and IDEO (Cairo).

(\*) optional

(\*\*) optional/to be completed on the second year