Scuola di Dottorato in ICT

PhD School in ICT

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

**Tutor**: Prof. Sonia Bergamaschi

**(\*) Italian Co-tutor:** dott. Giovanni Simonini, dott. Carlo Cavazzoni (Leonardo)

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

**Proposed Title of the research:**

High Perfomance Data-Integration for AI

**Keywords: (5)**

Data centric AI, Data Integration; Big Data; Machine Learning; Pay-as-you-go

**Research objectives:**

Data-Centric AI is the new approach to artificial intelligence that aims to enhance decisions that can be supported by high-quality, verified data (i.e., without errors) rather than just measuring the quality of algorithm predictions on test data. The success of the data-centric approach, therefore, depends on the ability to integrate and to correct the data form errors (i.e., "to prepare it").

However, when a large amount of data is available (often heterogeneous and constantly changing), integrating and preparing all the data is, in practice, impossible. Hence the goal of the research: to develop highly innovative algorithms and methods to automatically adapt the data integration and preparation process to the requirements of the applicatoin consuming the data (e.g., a ML model). A framework based on the cost / utility principle will be developed. The framewotk will direct the entire process of data integration and preparation considering: (i) the need for resources, i.e., the time and computational resources that can be used, thus defining the cost of the application; (ii) the impact that these operations have on the final output of the automatic learning algorithm, according to the "garbage-in / garbage-out" principle.

**Proposed research activity**

Expected activities (not limited to):

* The PhD student will study the state-of-the-art big data integration techniques and systems, performing benchmarking on real-world datasets
* The PhD student will work on the definition of a framework for supporting task-driven data integration; real-wold data sets and applications will be used for that (e.g., by exploiting programming/machine-learning competitions’ data and solutions)
  + This will require a deep knowledge of machine learning techniques
  + Deep learning approaches will be considered as well

**Supporting research projects (and Department)**

DIEF UniMoRe

Leonardo S.p.A.

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

School of Computer Science at Georgia Tech – Prof. Xu Chu

Computer Science Potsdam University – prof. Felix Naumann