Scuola di Dottorato in ICT

PhD School in ICT

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

**Tutor**: Prof. Sonia Bergamaschi

**(\*) Italian Co-tutor:** Giovanni Simonini, PhD

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

**Proposed Title of the research:**

Task-driven Big Data Integration

**Keywords: (5)**

Data Integration; Big Data; Human in the loop; Machine Learning; Pay-as-you-go

**Research objectives:**

*Big Data integration* is a set of processes used to retrieve and combine **data** from disparate sources into meaningful and valuable information. Nowadays, huge volume of **data** are collected from many heterogeneous **data** sources which are generating **data** in real-time with different qualities — which is called **Big Data.**

Unfortunately, there is no silver bullet for that: it may require the solution of several subproblems, such as: schema alignment, error detection/correction, duplicate detection, entity consolidation, missing value imputation, and more. All these problems can be solved with different techniques and algorithms which all share a bottleneck: they require the so-called *human in the loop* to validate their output, to write transformation programs, and/or to create labeled data for learning how to detect and correct errors in the data. Humans involvement is extremely hard (and expensive) to scale to big data, thus aiming for the best possible partial result is the only option for many tasks. But what does this mean? Different tasks may require different quality levels and some errors may affect the results of some tasks and not others—even when involving the same data sets. This project aims to propose techniques and tools for guiding data practitioners in building the more effective big data integration pipeline for their tasks at hand.

**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
* The PhD student will work also on massively parallel distributed systems, in particular: Hadoop, Apache Spark (and Spark SQL), and Apache Flink

**Supporting research projects (and Department)**

DIEF Unimore

CINECA Big Data Research Agreement

ENEA project: “Tecnologie per la penetrazione efficiente del vettore elettrico negli usi finali”

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

AT&T Bell Labs – prof. Divesh Shivrastava

Computer Science Postdam University – prof. Felix Naumann