Ph.D. in Information Technology: Final Dissertations

DEIB Conference Room

February 9th, 2017

10.00 am

Abdulrahman KAITOUA – XXIX Cycle

"Scalable Data Management and Processing for Genomic Computing"

Advisor: Prof. **Stefano Ceri**

Abstract:

The recent emergence of Next Generation Sequencing (NGS) technologies, in genomics field, produced vast amounts of genomic data. NGS resulted in dropped the cost of sequencing ("reading" in general terms) genomic material very fast. There exist many methods to extract signals from the data, that associate a region of the genome with some interesting information - such as a mutation or a peak of expression. Thus, a new problem is emerging: making sense of these signals, heterogeneous in nature, through new kind of languages that can extract relevant information from various heterogeneous sources, integrate them in a new data management system, and compute interesting results. Biologists say that a huge amount of information is undiscovered within the public repositories that have been built in the last decade - therefore, the focus of genomic data management for the next decade is querying and analysing heterogeneous genomic data.

This thesis is about scalable data management and processing for genomic data. We developed a new system which consists of a new query language called GenoMetric Query Language (GMQL), a new data model for heterogeneous data (called Genomic Data Model - GDM), and a new processing engine which embeds scalable genomic algorithms implemented on several data flow engines. GMQL is a domain-specific language. The work reported in this PhD thesis is focused on the design, implementation and validation of the system architecture of various prototypes of GMQL, GDM (the genomic data management system), and the scalable genomic processing algorithms.

In summary, this thesis is a step forward in the development of a systemic approach to scalable genomic data management and processing. Whereas other approaches focused on extracting genomic features from data, our approach is on combining these heterogeneous features so as to solve complex biological problems. We believe that the importance of a systemic approach to genomic data management will grow in the near future, with the availability of huge repositories of genomic data sets.

Pietro PINOLI – XXVIII Cycle

"Modeling and Querying Genomic Data"

Advisor: Prof. **Marco Masseroli**

Abstract:

Next-Generation-Sequencing (NGS) has dramatically reduced the cost and time of reading the DNA. Huge repositories of DNA sequences of large populations are being collected.

Answers to fundamental biomedical problems are hidden in these data. This Thesis is focused on tertiary analysis for genomic data integration, as a new data-driven basic science based on a simple driving principle: data should express high-level properties of DNA regions and samples, high-level data management languages should express biological questions with simple, powerful, orthogonal abstractions.

Challenges arise from the lack of standards for biological data and the big-data nature of NGS data. The Thesis presents a innovative framework, based upon a novel data representation and a declarative, domain specific query language, for modeling and querying potentially any genomic data. The implementation of the framework on cloud based environment is also discussed and optimization techniques are proposed.

PhD Committee:

Prof. **Marco Masseroli**, DEIB – Politecnico di Milano

Prof. **Heiko Muller**, Austrian Academy of Sciences

Dr. **Mattia Pelizzola**, Istituto Italiano di Tecnologia