Ph.D. in Information Technology: Thesis Defense

January 29th, 2021

online by Teams – at 14.00

Gabriele SCALIA – XXXIII Cycle

"Machine Learning-driven Integration, Knowledge Extraction and Uncertainty Management for Scientific Data"

Advisor: Prof. **Barbara Pernici**
Co-Advisor: **Prof. Tiziano Faravelli**

Abstract:

Managing scientific data, such as those found in the chemical and biomedical research, poses unique and challenging problems. Unique features characterize this data, including the impossibility of representing reality on a one-to-one scale, the imprecision in the observations and quality limitations introduced by technologies and models that continuously evolve.

This is an interdisciplinary research that, as a whole, investigates the management and the analysis of scientific data focusing on the challenges emerging in fields such as chemistry, genomics and biomedical research. For this work, we focus on data-driven – including machine learning-driven – techniques to face a set of identified requirements: 1) the management of uncertainty for complex data and models such as deep neural networks, 2) the estimation of system properties starting from imprecise, low-volume and evolving data, 3) the continuous validation of scientific models through large-scale comparisons with scientific data and 4) the unsupervised integration of multiple heterogeneous data sources related to different technologies to overcome individual technological limitations. Common to virtually all fields driven by experimental data, these requirements are faced through a set of case studies on different applications in chemistry, biology, and genomics.

Uncertainty estimation and evaluation is investigated in the context of deep neural network-based molecular property prediction. To this end, we develop a scalable Bayesian graph convolutional neural network for molecular property prediction and an uncertainty evaluation framework to assess the resulting estimates.

We investigate the problem of estimating the properties of biological systems starting from low-volume and imprecise experimental data proposing a machine learning-driven methodology to find the optimal way of transferring information in a molecular channel using collected experiments.

The design of a framework to support the development of scientific models through the continuous validation on integrated scientific experiments is presented, with the discussion of an architecture and the development of a prototype.

Finally, the unsupervised integration of heterogeneous data sources related to different technologies and with varying quality is explored, proposing a methodology to learn spatially-resolved whole transcriptomes of single cells through integration, starting from datasets measured with complementary transcriptomics technologies.

Extensive experiments based on in-vitro and in-silico data allow validating and discussing the proposed methodologies.

PhD Committee
Prof. **Cinzia Cappiello,** DEIB
Prof. **Fabio Casati,** Universita' di Trento
Prof. **Adam Prugel-Bennett**, University of Southampton