Ph.D. in Information Technology
Thesis Defense

February 2nd, 2024
at 14:00
Aula Seminari Alessandra Alario

Edoardo RAMALLI – XXXVI Cycle

Data Ecosystems and Data Science for Scientific Data

Supervisor: Prof.ssa **Barbara Pernici**

Abstract:

Predictive models have a pervasive role in many daily applications. The increasing amount of generated and shared data has recently boosted their development, shifting the model generation and improvement focus towards a data-centric approach. As a result, an information system that manages these data defines what can effectively discovered from them. Predictive models are also used in scientific domains to simulate complex real-world systems, replacing costly and time-consuming experiments. However, the unique characteristics of the scientific data and domain requirements, such as experimental uncertainty, low data quality, and confidentiality, make applying traditional methodologies to share and leverage the data challenging. This interdisciplinary research investigates, as a whole, the development process of a scientific predictive model and how it can be improved by adopting data ecosystem and data science technologies. This thesis focuses on the following requirements: 1) identification of the predictive model development process, classification of scientific data, and their properties, 2) the design of a sustainable data ecosystem to support a quality process, 3) the definition of an effective model evaluation methodology, 4) the use of appropriate data science techniques to guide the improvement and development of scientific predictive models. These requirements and challenges are valid across multiple scientific domains, but the interdisciplinarity of this thesis focuses on a case study of the chemical kinetics field. First, I investigate the current model development process, analyzing the typical steps, the data, and the roles involved. Then, I propose a data ecosystem that offers the necessary services and addresses the unique scientific data properties and domain requirements as data governance and management aspects while fulfilling the open data guidelines. Finally, the proposed solution is generalized with a set of challenges for designing and adopting sustainable data ecosystems and managing quality data in scientific domains. This thesis presents a systematic, objective, and automatic evaluation methodology for scientific predictive models while handling uncertainties, allowing replicability and awareness of the results with provenance information and fair validation. Finally, it discusses how the results of the model evaluation analysis can inform model improvement and generation. To this end, appropriate data science techniques are used and developed.

PhD Committee

Prof. **Cinzia Cappiello,** Politecnico di Milano

Prof. **Fabio Casati,** Università di Trento

Prof. **Oscar Pastor Lopez**, Università Politecnica di Valencia