Ph.D. in Data Analytics and Decision Science:

Thesis Defense May 4th, 2022 Aula Alario (Building 21)/online by Teams – at 10.00

Michela Carlotta Massi – XXXIV Cycle

Advisor: Prof.ssa Francesca Ieva

Patient Representations from Complex Biological Systems for Precision Medicine

Abstract

Precision medicine is a novel medical framework for prevention and treatment that takes into account individual variations in genes, environment, and lifestyle. It employs individuals' unique genetic profile and DNA sequences (all sorts of *omics* data, i.e. genomics, proteomics, metabolomics, etc.), together with medical big data (i.e. biosignals, electronic health records, medical imaging), to determine their susceptibility to disease, the most suitable and individualized treatment, and the focused preventive strategies to adopt.

From a methodological standpoint, precision medicine translates into a computational approach to functionally interpret omics and medical big data in their *effect* on complex phenotypic traits, to understand the genetic basis of disease etiology and develop effective biomarkers.

Unfortunately, designing effective models with large-scale molecular and clinical data has been a non-trivial and seldom unsatisfactory endeavour. This is probably due to the challenges carried in particular by omics-based data, around which this Thesis is mostly centered. It is in facts in the management of the intrinsic complexity of this data type, of the complex systems it describes and the peculiar facets of precision medicine studies, that resides the main methodological contribution of this Thesis.

Indeed, these rich sources of information carry characteristics (such as hyper dimensionality, small sample size, class imbalance, sparsity, spatial and functional correlation, noise, etc.) that hinder the applicability of most traditional statistical and biostatistical models and approaches based on assumptions that are now failing. Nonetheless, these methodologies are widely applied and appreciated in the medical research field because of their interpretability and robustness.

Therefore, the research presented in this Thesis is devoted to the development of methodologies that construct effective *biological system complexity-aware representations of data*, to enhance and complement interpretable and robust statistical approaches to classification, feature selection, survival modeling or association discovery. To do that, throughout the Thesis are exploited tools from the Representation Learning, Machine Learning, Statistical Learning and Graph Theory literature, designing original approaches or combining them into novel algorithms to target specific clinical enquiries. Most of the methods described in this Thesis were indeed motivated by real-world studies with relevant precision medicine-oriented objectives, such as personalized radiotherapy treatment planning, time to diagnosis prediction for breast cancer, or the discovery of the genetic basis of COVID-19 severity. The results of these case studies will be presented and discussed, to highlight the value added by the methodological contributions of this Thesis to the clinical practice.