PEPR digital health
Axis Methods and Models for Multimodal Data Integration (M4DI)

Integrating prior knowledge for better patient representation

Multimodal Data integration ; machine learning ; knowledge models ;

IRP summary

Background: One of the current challenges of precision medicine is to integrate heterogeneous data for the most adequate description of the patient. Today, these data come from multiple sources (biomic data, imaging, microbiota, clinical notes, drug prescriptions, claim databases , ...), each structured in a specific way. These available data can also be enriched with *a priori* information. For example, it is possible to link the biomic data to interaction graphs, the imaging to features known to be relevant in the diagnosis, the microbiota to functional annotations, or prescriptions to drug knowledge base

Objectives
The co-supervised thesis project aims at developing a method for the analysis of patients' data which harnesses prior knowledge for better performances. We will focus on enhancing i) biomic data of patients and ii) medical and administrative data available for each patient with such prior knowledge. Knowledge integration will be based on semantic web technologies or more broadly on knowledge graphs, widely used by the community to structure information. The aim is to quantify the contribution of this *a priori* information in classical risk analysis models as well as in more complex dimension reduction models, such as auto-encoders.

Methods
Task 1: Development of a pipeline to automatically retrieve prior knowledge associated to data from various sources: imaging, microbiota, clinical notes, drug prescriptions, claim databases. Web semantic technologies will be used for this task. **Task 2:** Development of new methods to integrate this knowledge into classical projection methods and machine learning classifiers, and production of a package designed to handle this. **Task 3:** Development of a new method to integrate treatment ontologies into classical statistical models, and production of a package designed to handle this.

Data/Application
As a case study, this project will focus on Inflammatory Bowel Diseases (IBD) and their association with microbiota, which provide a suitable problem thanks to the amount of both prior knowledge available for microbiota (metabolism, functional annotations, interaction graphs, …) and data availability. This project will have access to at least one cohort of patients from the CHU of Rennes, with available microbiota as well as medical data. We also plan to address the robustness of our approaches by studying other IBD related datasets for the co-supervising institutions. If necessary, ontologies dedicated to IBD data processing could be designed.

Expected Results
The project aims to produce a transferable methodology to integrate a priori information from existing or user-defined knowledge bases to i) guide the dimension reduction of unimodal and multimodal data, and ii) guide the learning of the algorithms used. This will result in a ready-to-use python package. Another package to integrate treatment ontologies into classical statistical models will be developed.

The PhD student will be part of the doctoral network of the project M4DI, one of the axes of the PEPR digital health. As such, they will have a host lab together with funding for a research stay of about 4 months in a secondment lab. The PhD student will further have the opportunity to participate in data challenges and collaborate with other axes of the PEPR digital health.

Host lab

IRISA (Institut de Recherches en Informatique et Systèmes Aléatoires) The DYLISS Team, IRISA, has a long-standing expertise in bioinformatics application to health problems. The team's expertise spans from analyzing microbiota data to retrieving and structuring symbolic **knowledge**, such as the one extracted from large semantic web databases. **Emmanuelle Becker**, PhD, has an expertise in biological knowledge extraction from public databases, and analysis of multimodal data to estimate diseases' characteristics. **Yann Le Cunff**, PhD, has an expertise in machine learning methods to analyze -omics data, and an ongoing collaboration with CHU Rennes for the analysis of microbiome data.

Secondment lab

LORIA (Laboratoire Lorrain de Recherches en Informatique et ses Applications). The K-team, LORIA institute, is specialized in knowledge intensive AI systems and has a strong expertise in symbolic AI and knowledge intensive methods. **Nicolas Jay** and **Aurélie Bannay** have a strong expertise in medical knowledge representation, and big medical data analysis. Nicolas Jay has developed patient classification methods relying upon pattern mining and linked open data technologies.

Expected profile

The candidate should show a strong background in numerical approaches applied to biology/health problems. Coming from a bioinformatics master (or equivalent) would typically

be well-suited for this Ph.D. project. The candidate should also demonstrate interests in machine learning technics as well as in knowledge extraction and organization.