

MECHANICAL ENGINEERING | PHYSICS |  
PRESERVATION OF THE ARCHITECTURAL  
HERITAGE | STRUCTURAL, SEISMIC  
AND GEOTECHNICAL ENGINEERING |  
URBAN PLANNING, DESIGN AND  
POLICY | AEROSPACE ENGINEERING |  
ARCHITECTURE, BUILT ENVIRONMENT  
AND CONSTRUCTION ENGINEERING |  
ARCHITECTURAL, URBAN AND INTERIOR  
DESIGN | BIOENGINEERING | DATA ANALYTICS  
AND DECISION SCIENCES | DESIGN |  
ELECTRICAL ENGINEERING | ENERGY AND  
NUCLEAR SCIENCE AND TECHNOLOGY |  
ENVIRONMENTAL AND INFRASTRUCTURE  
ENGINEERING | INDUSTRIAL CHEMISTRY AND  
CHEMICAL ENGINEERING | INFORMATION  
TECHNOLOGY | MANAGEMENT ENGINEERING  
| MATERIALS ENGINEERING | MATHEMATICAL  
MODELS AND METHODS IN ENGINEERING



Chair:  
Prof. Piercesare Secchi

## DOCTORAL PROGRAM IN DATA ANALYTICS AND DECISION SCIENCES

The Ph.D. program in Data Analytics and Decision Sciences (DADS) aims at training highly qualified senior data analysts and data managers capable of carrying out research at universities, international institutions, tech and financial companies, regulatory authorities, and other public bodies. The program stems from the cooperation between three departments: Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Dipartimento di Ingegneria Gestionale (DIG), Dipartimento di Matematica (DMAT), and the Center for Health Data Science Center of Human Technopole. It allows the enrolled students to work in a highly interdisciplinary environment with strong connections to international research centers and private companies. The program provides successful candidates with the opportunity to acquire a high degree of professional expertise in specific scientific and technological fields.

The program lasts three years: upon its successful completion and final exam, candidates will be awarded the title of Ph.D. in Data Analytics and Decision Sciences. The first year is devoted to the courses that build the broad competence and the solid interdisciplinary set of skills required by data analytics. The following two years focus on the development of the Doctoral thesis. Students must spend at least one semester in a research institution abroad, taking advantage of the network of international collaborations of the three departments involved in the program.

The program aims at breeding the next generation of data scientists who will tackle the challenges and the opportunities created by the increasing availability of the massive amount of data. These data scientists will be able to capture the relevant aspects of phenomena at play, develop adequate models, supervise the development of analytic pipelines, critically analyze the results, and support the technological transfer.

Data Analytics and Decision Sciences graduates are equipped with unique skills and advanced knowledge that open up career opportunities at universities, international research centers and institutions, R&D departments, regulatory authorities, financial institutions, tech companies, and other public bodies.

### FACULTY BOARD

Prof. Secchi Piercesare (Coordinator)

Prof. Azzone Giovanni

Prof. Caiani Enrico Gianluca

Prof. Ceri Stefano

Prof. Di Angelantonio Emanuele

Prof. Flori Andrea

Prof.ssa Ieva Francesca

Prof. Lanzi Pierluca

Prof. Matteucci Matteo

Prof.ssa Orsenigo Carlotta

Prof. Punzo Fabio

Prof. Roveri Manuel

Prof. Secchi Piercesare

Prof. Spagnolini Umberto

Prof.ssa Tanelli Mara

Prof. Tubaro Stefano

Prof.ssa Tumino Angela

Prof. Vantini Simone

## DATA ANALYTICS FOR THE NERVOUS SYSTEM ACTIVITY: NEW FEATURE EXTRACTION TOOLS

**Letizia Clementi** - Supervisors: Marco D. Santambrogio, Laura M. Sangalli

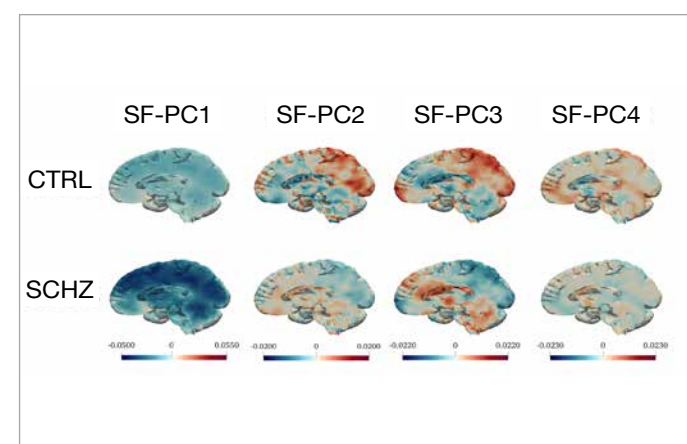
The nervous system holds a highly significant role within the human body, regulating bodily activities, maintaining homeostasis, and gathering and synthesizing information from various body systems. This complexity attracts the attention of multiple scientific disciplines, exploring different aspects of its functioning. The integration of knowledge stemming from different areas of study is an open and significant question in these fields. Indeed, recent findings highlighted how traditionally independent aspects of the nervous system's functioning may actually be intertwined: the apparently counter intuitive link between autonomic activity and some psychiatric disorders (e.g. schizophrenia) is only one instance of this phenomenon. This thesis operates within the realms of neurophysiology and computational psychiatry. While the first studies the nervous system's physiological functioning, the second examines the alterations due to psychiatric conditions, drawing from statistics, machine learning, and related computational approaches. In these fields, managing data dimensionality poses a significant hurdle, compounded by the intricate neural anatomy. Furthermore,

the dynamic interactions within this system exhibit extreme complexity, marked by non-linear patterns. These technical challenges give rise to theoretical and computational issues, often addressed in literature through oversimplification. This work employs two main methodological approaches, Functional Data Analysis and spectral analysis, individually and in combination, to extract relevant features from biomedical signal data. These approaches are validated across three categories of biomedical signals: fMRI, EEG, and ECG. The work is organized in three parts, each addressing a class of biomedical signals employed in the field, namely: fMRI, EEG, ECG. Each class of data carries with itself a different point of view on the nervous system, as well as different technical challenges. We hereby consider such data sources as they are prevalent and widely used in clinical and research settings. For each class of data, we tackle different case studies, highlighting the ubiquitous utility of more effective features in clinical settings, as well as in medical and neuroscience research. Indeed, for each case study, we propose a feature extraction approach able to

synthesize relevant information relative to the pathology and/or the phenomenon in analysis. The methods proposed are selected according to the aspects more relevant for the considered case study. Nonetheless, from a more general perspective, the proposed approaches always rely either on FDA, on spectral analysis, or on a combination of the two, to compute features able to integrate more information with respect to standard methods. FDA is employed in this work thanks to its capability to address curves as a whole, considering directly them as statistical units; on the other hand, spectral methods have been proven to be a useful tool in a variety of biomedical application: the potentiality of such approach are however not fully exploited yet. Finally, the combination of the two represents, to the best of our knowledge, a novelty. Part I is devoted to fMRI data, nowadays representing the golden standard in neuroscientific research. fMRIs, in fact, collect both anatomical and functional information on a subject's brain. Although the amount of information stored, the analysis of fMRI data presents relevant technical challenges: due to the above described data

dimensionality and anatomical complexity, functional and structural aspect of the data are often analyzed separately. Moreover, although the relevance of non-linear dynamics in brain activity, research still frequently relies on linear time domain metrics to compute connectivity maps assessing the connectivity among different brain regions. Part II addresses challenges linked to the analysis of EEG data. This class of signals is characterised by higher temporal resolution, lower cost and lower invasiveness. For these reason, EEGs are more common than fMRI in clinical settings, when it comes to the analysis of brain activation. Moreover, the high

temporal resolution (in the order of ms) allows to observe the instant response of the brain to a given stimulus. Finally, the frequency content of EEG signals is well known to be associated to cognitive status. The final segment of the work focuses on evaluating Autonomic Nervous System activity through spectral-based metrics. After addressing questions about sex-based differences in neurophysiology, we propose novel spectral-based metrics to further investigate sympathetic nervous system tone.



**Fig. 1** - Connectivity maps reconstructed from the feature extracted via Functional Data Analysis based data dimensionality reduction performed on resting state fMRIs. The proposed approach highlights differences in the connectivity patterns characterizing an experimental population affected by schizophrenia, versus a control group.

# A REAL-WORLD DATA STATISTICAL APPROACH TO SUPPORT DECISION-MAKING IN CLINICAL AND HEALTH POLICY

**Federica Corso** – Supervisor: Anna Maria Paganoni

Co-Supervisor: Fabio Pammolli

Healthcare is transitioning to a new model defined as “patient-centric”, “data-driven” and “value-based”. This process has been enhanced by the advent of Personalized Medicine (PM) for various diseases but assessing the impact on healthcare delivery and the benefit of PM over existing therapies is not always easy. Therefore, policy decision-makers are interested in evaluating the clinical effectiveness and safety of these new health technologies in Real-World (RW). Stakeholders have also an additional concern which is to determine whether new health technologies represent value for money. Health Technologies Assessments bodies have always preferred data from Randomized Clinical Trials as gold standard for the highest level of clinical evidence. Nowadays the availability of complex informatic technologies able to collect and process big amount of data at a very low cost are increased the demand for Real-World Data (RWD) in healthcare. We usually refer to RWD as data stored in Electronic Health Records (EHRs), claims and billing activities, clinical registries, patient-generated data, etc. This thesis aims to enhance the adoption of RWD in various contexts,

from oncology to cardiovascular diseases, by promoting the use of advanced analytic methodologies concerning RW study design, collection, integration, and analysis of RWD. Thus, the final goal is to provide Real-World Evidence (RWE) and learn about the value brought by new technologies, to optimize the patient’s management and the sustainability of national health services. Among the analytical methodologies, my thesis gives an overview of two main areas of techniques in healthcare. The first area regards statistical learning approaches, including classical methods for observational research. The thesis also extends the landscape of data analysis to the more complex machine learning methods, to infer personalised treatment effects from observational RWD, and text mining techniques for RWD generation from EHRs. Two projects have been included in this PhD thesis. *OncoData platform* is a pilot project launched by Politecnico di Milano and AstraZeneca S.p.A., that aims to the implementation of a novel integrated tool of Health Analytics based on the collection and integration of RWD from the data management systems of various Italian cancer institutes.

The clinical partners are: IRCSS Istituto Nazionale dei Tumori (INT) and IRCSS Istituto Europeo di Oncologia (IEO). Specifically, the research investigates the impact of innovative therapies for lung and ovarian cancer, respectively. The general framework, for the implementation of this infrastructure, includes the characterization of the research questions, the identification of the study population and the available data sources for variable extraction. The final step is the application of appropriate statistical techniques to generate RWE. A general overview of the methodological approach adopted in both the case studies is here reported. At first, a comprehensive literature research has been conducted to correctly design a RW study about innovative drugs in oncology. In a second step, the project focused on the identification of data sources at our clinical partners. The primary data sources were represented by the hospitals’ Datawarehouses (DWHs), that are usually longitudinal patient-oriented database, where most of the codified variables are recorded for administrative and clinical reasons. A summary list of the information, deemed to be relevant, has been extracted

from Registry, Laboratory test, Radiotherapy, Chemotherapy, Hospitalizations, EHRs. Then, clinical data extraction has been integrated with a text mining approach, the Rule-Based Named Entity Recognition (NER) algorithm, in which EHRs have been selected at the last follow-up date. This passage was necessary to obtain a structured clinical dataset from a free-text unstructured data source, that could be properly executable by statistical models. The main endpoints of this projects included: collection of data on efficacy and safety to evaluate the clinical effectiveness and the safety of available treatments, respectively; the collection of data to estimate the cost-effectiveness generated through a complete assessment of associated direct and indirect costs. The effectiveness and safety of targeted therapies (specifically EGFR-TKIs for lung cancer and PARPis for ovarian cancer) have been measured through the following clinical outcomes indicators: Overall Survival (OS), Time to Treatment Discontinuation/Failure (TTD) which are well-recognized endpoint in RW studies. Kaplan Meier curves and multivariate Cox models (with covariate at baseline and longitudinal variables) have been performed. For quantifying the economic impact of such treatments, administrative and clinical data have been integrated into a Bayesian cost-effectiveness model. This approach is based on the definition of a multi-state Markov model which associates

at each Health State (HS) a measure of cost and clinical benefit. The methodology has been generalized by proposing a 4-state Markov model that also accounts for costs and survival in patients who progressed after a first-line therapy until death. At each HS, costs (time-varying and one-off) have been distinguished in direct medical costs, direct non-medical and indirect costs. Clinical effectiveness has been measured in Quality-Adjusted Life Years (QALYs), retrieving the utility, associated with each HS and stratified by treatment, from the literature. Scenario analysis and Probabilistic Sensitivity analysis have been also performed to test the robustness of the model. The *HIV project* is a joint work with IRCSS San Raffaele Hospital. The work investigates the risk of cardiovascular diseases (CVDs) in HIV research by comparing the performances of Cox Proportional Hazard model and DeepHit, a Deep Neural Network-based model, in both time-invariant and time-varying settings. Permutation Feature Importance and the Shapley Additive Explanation Value are used for the explainability of the network results. From an epidemiological point of view, this work aims to study the effect of antiretroviral therapies on the risk of cardiovascular events in HIV patients. A cohort of 4512 HIV patients with only 2% CVD events has been collected with 21 variables, including demographic characteristics, clinical parameters, Antiretroviral Therapies (ART) and on average, 32 visits have been measured

for each patient. According to the two methodological approaches, the prediction has been implemented by using full and reduced set of covariates with techniques for model interpretation and goodness-of-fit metrics has been applied. Our study has found that time-dependency of variable is an important factor when studying the impact of ART drugs on CVD events and that the exposure to ARTs is a protective factor by both the Cox model and DeepHit model. In a reduced covariate setting the C-indexes of the DeepHit and the Cox PH model are comparable both at baseline and with time-dependent covariates. Lower MSEs are reported for the Cox PH model, while higher AUCs are obtained with DeepHit. This work generally represents a deep methodological investigation about RWD and their value in epidemiological studies. To the best of our knowledge, the current research landscape in healthcare still lacks systematic frameworks considering the design, the collection and integration of all the routinely collected health data. Moreover, given the wide range of possible clinical applications, an accurate investigation of analytical methodologies, including statistical models and machine learning algorithms, needs to be carried out, to support the decision systems. The projects discussed in this thesis, explore all these aspects to unlock the strengths and limitations of RWD to support the decision-making in healthcare.

## TOWARDS A COMPREHENSIVE APPLICATION OF DEEP LEARNING METHODS IN MEDICAL IMAGING: SEMANTIC SEGMENTATION, FEATURES EXTRACTION, SYNTHETIC IMAGES GENERATION

Leonardo Crespi – Supervisor: Daniele Loiacono

In the multifaceted world of Medical Imaging, the increasingly sophisticated techniques and the advancement of medical technology call for augmented analysis capabilities. Deep Learning comes to the rescue, providing a new set of tools to tackle some of the most challenging problems exploiting the power of data: in the last decade, as a cornerstone of AI techniques, it has proven to be the most valuable and revolutionary tool in the field of Computer Vision, showcasing the potential of its models and algorithms in the most various tasks, ranging from image classification to semantic segmentation, passing through object detection and images generation. In this dissertation, the focus is on some of these techniques, taking on a series of tasks in Medical image analysis from the everyday clinical setting that can be sped up, improved, or eased with the help of such tools. In particular, the works focus on three main research lines: semantic segmentation of organs and anatomical structures, feature extractions from chest X-rays, and synthetic CT generation from MRI images. Concerning segmentation models, several analyses are performed to understand the

impact of certain design choices in the training and testing of segmentation models, such as input dimensionality (2D, 3D), training paradigm (supervised, adversarial) and ensembling techniques; a particular field of application for such techniques surely is radiotherapy, where there is the strong necessity of organs and targets contouring, currently manually performed by radiation oncologists, with the objective to deliver the dose as precisely as possible thanks to the modern equipments, that improve the accuracy and reduce the toxicity of treatments but require extremely accurate targets delineation. Among the results obtained, some surprising findings are highlighted, such as the fact that 2D models can outperform 3D ones in some cases, especially when the availability of data is particularly limited, which is often the case for not only the medical imaging field, but the medical field in general; ensembling methods are also considered, with the aim of leveraging the increasing availability, thanks to the effort of the scientific community and the possibility to share models, weight and code, of models trained on specific tasks, like the segmentation of a single organ, area or region, to

combine them and produce multi-label segmentation maps that would allow for the contouring of multiple organs, structures, targets, at once; thanks to such techniques, there's also the possibility to better employ available models with very limited or even without the necessity to train them extensively, therefore also the data availability limitations are of less concern. The analysis of segmentation models also explores the use of different training approach, comparing adversarial and supervised training to understand if one of the two has a specific advantage with the respect to the other and the findings show how adversarial training seems to be more effective in producing reliable results with more difficult targets, such as, in the context of radiotherapy, the trachea and the esophagus. Regarding features extraction, a method based on Variational Autoencoder is proposed to extract general and informative features from chest x-rays, to have a method capable of overcoming the bias that data from a single source usually creates in the medical field. The main idea is to have a method, developed in an unsupervised fashion and without any previous knowledge on the downstream tasks, capable to capture all the

most important information from images, encoding them in a simpler data structure, like, indeed, a vector, from which it would still be possible to infer informations about the patients, like pathologies visible, the properties of the structures present, a general overview of the status of some organs. This would allow to easily automate a serie of downstream task in a rather simple way, also exploiting machine learning models, way less difficult to train than deep neural networks; this has been proven through the use of different three-based machine learning models to classify chest-xrays according to the pathologies visible in them, training the classifier only on latent features vectors extracted with variational autoencoder trained on those images, obtaining accuracy close to state-of-the-art performance from specifically developed deep learning models. The last topic treated is synthetic data generation. Once again, the clinical context is radiotherapy, but this applies also to deep learning in general, and the main objective is to provide a viable tool initially for and advanced data augmentation, but, given the promising result, with the final aim to provide a way to generate and accurate CT scan from MRI

imaging. The usability of such a system from a deep learning standpoint is immediate, as it would be extremely beneficial in situation, such as the medical field in general, where not only availability of data, as metioned previously, is limited, but often also their diversity, causing results obtained on dataset from a specific centre to not being applicable on ones from a different one; secondly, the clinical applicability of this method would surely be important from an economical point of view, because of course sparing patient from an additional instrumental exam would be cheaper and less time consuming, but also very impactful in situation where the patient are more towards the most fragile side of the population and also in particularly stressful situation (and radiotherapy is often the case), therefore sparing them of an additional acquisition is best. Results from this research topic show how the generated images, obtained in different configuration through the use of CycleGAN models in an unsupervised fashion and without the use of paired data, reach performance close to a realism that can almost fool expert doctors; multiple assessments are performed, both quantitative and qualitative, in order to assess

the best models and settings, and they indicate that the best seem to be the ones exploiting a multimodal input consisting of T1-weighted in-phase and out-of-phase MRI scans. Qualitatively, their realism has been tested with the help of physicians, asking them to distinguish real from fake ones, and they were mostly unable to do so under specific condition, showing the potential of such approach. In the end, many of the results obtained show promising approaches for the future of AI in the field of Medical Imaging and shed light on some aspects of the design choices, their advantages and disadvantages in different scenarios, and some of the challenges that still need to be tackled in different areas. Overall, it is evident how the bottleneck is almost always data quality, availability and diversity, limiting the possibility to explore models configurations and architecture and slowing the possibility to test the robustness of automated systems.

## EXPLOITING AI AND NLP METHODS FOR EMPOWERING NAÏVE USERS IN SOLVING DATA SCIENCE PROBLEMS

Sara Pidò – Supervisor: Stefano Ceri

Data Science (DS) and Machine Learning (ML) have become critical tools for making informed decisions, predicting outcomes, and automating processes. The rise of big data and the availability of powerful computers, coupled with the development of new and more sophisticated ML algorithms, has led to a huge growth of interest in ML over the past decade. Despite the significant advancements in ML methods, building and training ML models can still be complex and time-consuming, requiring expertise in computer science, mathematics, and statistics; taking advantage of ML can still be challenging, especially for people without these skills. The significant gap in a deep understanding of machine learning principles among IT and business professionals has led to incidents related to bias, privacy, security, transparency, and ethical concerns. The democratization of data science and machine learning aims to change this situation, by making ML technologies and techniques more accessible to a wider range of people.

This thesis discusses the difficulties non-experts face in using ML tools. It explores various approaches

to democratize data science and machine learning, such as developing user-friendly ML tools and platforms and educational initiatives to teach the necessary skills to non-experts. It discusses the potential benefits of democratizing data science, such as driving innovation, providing new insights into complex problems, and creating a more inclusive and diverse data science community. In particular, my research introduces a progression of methods and underlying tools that make use of conversational agents, natural language, and autoML, with the objective of democratizing data science and make it more accessible to a wider range of people. The thesis begins by presenting GeCoAgent and DSBot, two multi-modal conversational agents designed to facilitate data science processes starting from natural language input. GeCoAgent and DSBot are two distinct conversational agents that serve different purposes in the context of data science automation. GeCoAgent takes a proactive approach by driving the conversation with the user, asking detailed and specific questions to better understand the user's needs and goals.

On the other hand, DSBot is a user-driven conversational agent, where the user provides a research question in natural language and the bot extracts the necessary information and executes the relevant data science processes. However, the automation of data science processes raises issues such as the difficulty in formulating the research question and the importance of incorporating domain expertise into the pipeline. To address these challenges, the thesis then presents two additional tools: MLFriend and Zephyr. MLFriend enables automatic generation of prediction tasks, while Zephyr streamlines the integration of domain expertise and automated data science tools. We conducted empirical evaluations and user studies to illustrate the effectiveness of these tools in making machine learning more accessible and user-friendly.

By providing these four solutions, embodied within GeCoAgent, DSBot, MLFriend, and Zephyr, we show a progressive development of ideas, methods and tools towards the goal of improving the accessibility and usability of data science tools for non-experts. Our research contributes to the field of

democratization of DS by providing new strategies that can be used to reduce the gap between experts and non-experts in the field. We trust that our results will contribute to address the remaining challenges and opportunities and make machine learning more accessible and user-friendly for a wider range of people.