AEROSPACE ENGINEERING / ARCHITECTURAL, URBAN AND INTERIOR DESIGN / ARCHITECTURE, BUILT ENVIRONMENT AND CONSTRUCTION ENGINEERING / 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 / MECHANICAL ENGINEERING / PHYSICS / PRESERVATION OF THE ARCHITECTURAL HERITAGE / STRUCTURAL, SEISMIC AND GEOTECHNICAL ENGINEERING / URBAN PLANNING, DESIGN AND POLICY

Chair:
Prof. Luigi Piroddi

DOCTORAL PROGRAM IN INFORMATION TECHNOLOGY

Introduction

The Ph.D. programme in Information Technology (ITPh.D.) covers research topics in four scientific areas: Computer Science and Engineering, Electronics, Systems and Control, and Telecommunications.
This broad variety of research topics is matched together by the common affinity to the ICT area and perfectly captures the core mission of the corresponding sections of the Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB). New research topics and cross-areas research fields are also covered, such as machine learning, big data, intelligent data analysis, Industry 4.0, Internet of Things, bioinformatics, quantum computing, ecology, environmental modelling, operations research, and transportation systems. The Ph.D. programme in IT is the largest in Politecnico in terms of number of students. Every year, about 70 new students join the programme, for an overall number of around 250. Students must undergo a yearly evaluation of the progress in their research and course work.

Topics

Research at DEIB in the field of Information Technology is supported by 35 laboratories, and is organized in 4 main areas.
- Computer Science and Engineering (Vice-Chair: Prof. Francesco Amigoni): Information systems, Database management, Information design for the web, Methods and applications for interactive multimedia, Embedded systems design and design methodologies, Dependable systems, Cybersecurity, Autonomous robotics, Artificial intelligence, Computer vision and image analysis, Machine learning, Dependable evolvable pervasive software engineering, Compiler technology, Natural language processing and accessibility.
- Electronics (Vice- Chair: Prof. Angelo Geraci): Circuits and systems, Single photon detectors and applications, Radiation detectors and low noise electronics, Electronic circuit design, Electron devices.
- Systems and Control (Vice-Chair: Prof. Lorenzo Fagiano): Control theory and its applications, Robotics and industrial automation, Dynamics of complex systems, Planning and management of environmental systems, Operations research and discrete optimization.
- Telecommunications (Vice-Chair: Carlo Riva): Networking, Applied electromagnetics, Optical communications, Quantum communications, Wireless and space communications, Remote sensing, Signal processing for multimedia and telecommunications.

Industrial collaborations

Due to its intrinsic technological nature, the Ph.D. programme features many industrial collaborations. More than 50% of the Ph.D. candidates are funded by companies or by international research projects involving industrial partners. In the Ph.D. School vision, the collaboration between university and companies is the ideal ground where to turn invention and scientific research into technological innovation. This collaboration also contributes to create a common terrain of friendly culture, to size research risk, and to preserve strong fundamental research. To monitor the activities and development of the Ph.D. programme, the Ph.D. board cooperates with an industrial advisory board, composed by members of public and private companies, working in management, production, and applied research. The board meets once a year to identify and suggest new emerging research areas and to foster the visibility of the Ph.D. programme in the industrial world.

Educational aspects

The teaching organization and the course subjects reflect the scientific interests of DEIB faculties. The curricula include a wide choice of courses (about 20 per year), and more than 30 courses for basic soft and hard skills offered by the Polimi Ph.D. School.
Access to external courses and summer schools is also encouraged. The challenge is to promote interdisciplinary research while offering advanced help to students to make the best choices according to the regulatory scheme of the programme.

Internationalization

Every year, several courses are delivered by visiting professors from prestigious foreign universities. Moreover, the Ph.D. programme encourages joint curricula with foreign institutions. The programme has several Double Degree and Joint Degree agreements with institutions from countries in all continents. Every year the programme receives we receive more than 150 candidate applications from foreign countries and about 15% of our selected Ph.D. candidates have applied from outside Italy.

Conclusions

The core mission of the Ph.D. programme is to offer an excellent Ph.D. curriculum, through high-quality courses, a truly interdisciplinary advanced education, cutting-edge research, and international and industrial collaborations.

BOARD OF PROFESSORS

- Prof. Francesco Amigoni - Vice Chair Computer Science and Engineering
- Prof. Cesare Alippi
- Prof. Luciano Baresi
- Prof. Cinzia Cappiello
- Prof. Nicola Gatti
- Prof. Davide Martinenghi
- Prof. Maristella Matera
- Prof. Raffaela Mirandola
- Prof. Cristina Silvano
- Prof. Stefano Zanero
- Prof. Angelo Geraci - Vice Chair Electronics
- Prof. Giuseppe Bertuccio
- Prof. Giorgio Ferrari
- Prof. Ivan Rech
- Prof. Alessandro Sottorcornola Spinelli
- Prof. Lorenzo Fagiano - Vice Chair Systems and Control
- Prof. Fabio Dercole
- Prof. Simone Garatti
- Prof. Lorenzo Mari
- Prof. Luigi Piroddi - Chair of the Doctoral Programme
- Prof. Andrea Zanchettin
- Prof. Carlo Riva - Vice Chair Telecommunications
- Prof. Matteo Cesena
- Prof. Paolo Martelli
- Prof. Andrea Monti Guarnieri
- Prof. Massimo Tornatore
- Prof. Giancarlo Ferrigno

PH.D. ADVISORY BOARD

Giorgio Ancona, Atos

Matteo Bogana, Cleafy

Mario Caironi, IIT

Paolo Cederle, Everis

Cristina Cremonesi, The European Ambrosetti

Riccardo De Guadenzi, European Space Agency

Giuseppe Desoli, STMicroelectronics

Alessandro Ferretti, Tre-Altamira

Giuseppe Fogliazza, MCE Srl

Bruno Garavelli, Xnext s.r.l.

Maurizio Griva, Reply Spa

Sabino Illuzzi, Prospera

Renato Marchi, KPMG

Renato Lombardi, Huawei Technologies

Giorgio Parladori, SM Optics srl

Francesco Prelz, INFN

Enrico Ragaini, ABB S.p.A

Parolo Giuseppe Ravazzani, CNR

Dario Regazzoni, Amazon Web Services (AWS)

Carlo Sandroni, RSE S.p.A

Massimo Valla, TIM

Luisa Venturini, Vodafone Italy

Stefano Verzura, Huawei Technologies

Roberto Villa, IBM Italy

Prizes and awards

In 2021 the following awards have been obtained by Ph.D. candidates:

- 24th International Conference on Business Information Systems, Best Paper Award – **Bernardo Alessio, Falzone Emanuele, Zahmatkesh Shima**
- PRIME2021, Bronze Leaf Award – **Scaletti Lorenzo, Parisi Angelo**
- ACM NANOCOM 2021, Data Competition 1st Prize – **Ratti Francesca, Scalia Gabriele, Scazzoli Davide, Vakilipoor Fardad**
- Virtual 2021 IEEE NSS MIC, Trainee Grant – **Corna Nicola, Garzetti Fabio**
- Hypeac Technology Transfer Award – **Gadioli Davide, Palermo Gianluca, Silvano Cristina**
- 2020 IEEE Nuclear Science Symposium, Best Student Paper Award – **Di Vita Davide, Utica Gianlorenzo**
- IFAC Best Young Author Award – **Bonassi Fabio**
- Chorafas Award – **Bernasconi Anna, Zanetto Francesco**
- "Prof. Florian Daniel" PhD Thesis Award – **Ferrari Dacrema Maurizio**
- SIE Best Doctoral Thesis Award – **Zanetto Francesco**
- RecSyS Challenge 2021, 1st Place in the Academic Leaderboard – **Bernardis Cesare, Dacrema Maurizio Ferrari, Perez Maurera Fernando Benjamin**

DETECTING ANOMALIES IN THE BEHAVIOR OF AUTONOMOUS ROBOTS

Davide Azzalini - Supervisor: Prof. Francesco Amigoni

Autonomous robots are increasingly becoming part of human everyday life. From driverless cars to assistive robots for elderly people, these systems are leaving the factories and entering unconstrained scenarios with close interaction with humans. Complex and dynamic environments are characterized by large degrees of uncertainty and pose big challenges to robot designers. One of the key competences required to newly conceived robots is to reliably operate over long periods of time under changing and unpredictable environmental conditions, which is referred to as long-term autonomy (LTA). Detection of anomalies and faults is a key element for LTA, because, together with subsequent diagnosis and recovery, it allows to reach the required levels of robustness and persistency. A fault which is not promptly detected and addressed, in fact, may result in the robot damaging itself or, even worse, in harming surrounding people.

In this thesis, multiple approaches for detecting anomalous behaviors in autonomous robots starting from data collected during their routine operations are proposed. The main idea is to model the nominal (expected) behavior of a robot and to evaluate how far the observed behavior is from the nominal one. A variety of application domains involving different robotic platforms required to operate for long periods of time without interruption are considered. Among these, we find: an autonomous surface vessel performing a water monitoring task on a lake, an assistive robot supporting the independence of elderly people living alone at home, a patrolling robot, a fixed-wing unmanned aerial vehicle, and a simulated swarm of e-puck robots performing tasks of dispersion, aggregation, homing, and flocking. An example of anomalies affecting the autonomous surface vessel are displayed in Figure 1, where the trajectory presents a recurring curve leaning to the left in the descending traits.

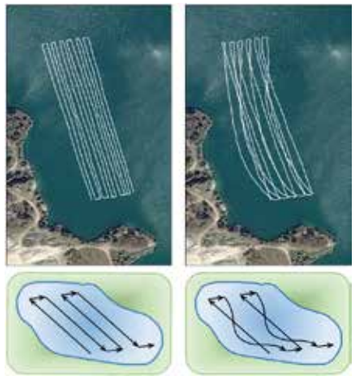

Fig. 1
Nominal (left) and anomalous (right) trajectories of the vessel performing water monitoring on Lake Garda.

The first approach we propose uses Hidden Markov Models (HMMs) to learn the robot's behavior under normal circumstances and detects anomalies by computing variants of the Hellinger distance between the distribution of observations made in a sliding window and the corresponding nominal emission probability distribution (online anomaly detection), or between two HMMs representing nominal and observed behaviors (offline anomaly detection). The use of the Hellinger distance positively impacts on both detection performance and interpretability of detected anomalies.
We then present a data augmentation and retraining technique based on adversarial learning for improving anomaly detection performance of our HMM-based approach when few nominal examples are available. In particular, we first define a methodology for generating adversarial examples for anomaly detectors based on HMMs; then, we present a data augmentation and retraining technique using these adversarial examples to improve anomaly detection performance and robustness to adversarial attacks.

The second approach we introduce is a new deep learning-based minimally supervised method which employs a new Variational Auto-Encoder (VAE) architecture able to model very long multivariate sensor logs exploiting a new incremental training method, which induces a progress-based latent space that can be used to detect anomalies. The latent space obtained for the water monitoring robot is shown in Figure 2, where it appears clearly how our method produces a good structure in which different nominal and anomalous behaviors are well separated.

While most existing approaches are trained in a semi-supervised fashion and require big batches of nominal observations, our method is trained using unlabeled observations of a robot performing a task, containing both nominal and anomalous executions. Only a very little amount (even just one) of labeled nominal executions is then required to partition the learned latent space into nominal and anomalous regions. Also in this case, we present both an online and an offline technique. We then propose an adaptation of the VAE-based approach to allow individual robots in a multi-robot swarm systems to detect anomalies in one another.

Particular attention throughout the thesis is devoted to ensuring that the proposed methods can be easily applicable in different practical settings. Accordingly, all the approaches proposed in this thesis are designed not to make any limiting assumption on how anomalies look like and to work with small amounts of (labeled) training examples. We show how the methods proposed in this thesis positively compare against state-of-the-art anomaly detectors commonly used in robotics in all the domains considered.

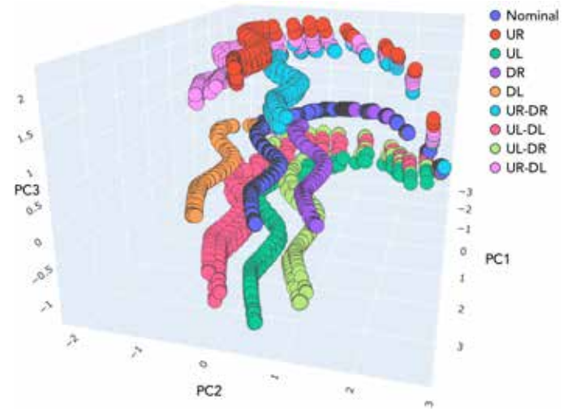

Fig. 2
Autonomous surface vessel latent space (first 3 principal components). The encodings of the nominal behavior and 8 nuances of anomalies similar to the one in Figure 1 are displayed. The nominal behavior of the robot is depicted in the center. UR-DL corresponds to the behavior of leaning to the right in upward segments and to the left in downward ones, UR corresponds to the behavior of leaning to the right in upward segments, UR-DR corresponds to the behavior of leaning to the right in both upward and downward segments, DL corresponds to the behavior of leaning to the left in downward segments, DR corresponds to the behavior of leaning to the right in downward segments, UL-DL corresponds to the behavior of leaning to the left in both upward and downward segments, UL corresponds to the behavior of leaning to the left in upward segments, UL-DR corresponds to the behavior of leaning to the left in upward segments and to the right in downward ones.

PHYSICAL HUMAN-ROBOT INTERACTION THROUGH GOAL-DRIVEN MANUAL GUIDANCE

Davide Bazzi - Supervisor: Prof. Paolo Rocco

During the last decade, collaborative robotics has become a major trend in the robotic research field as well as in the fourth industrial revolution. Collaborative robotics is indeed a powerful tool to suitably and effectively respond to the new flexibility requirements of the current global economy. Robots and humans actively cooperate within the same working environment, having their strengths combined. Human advanced cognitive skills together with robot speed, strength and accuracy, make a team endowed with high flexibility and problem solving capabilities to accomplish a huge variety of complex tasks.
In this context, human-robot physical interaction, and in particular manual guidance operations, have received increasing attention. By applying forces and torques on a handle, the operator drives the end-effector of a compliant manipulator towards the target along the desired path.

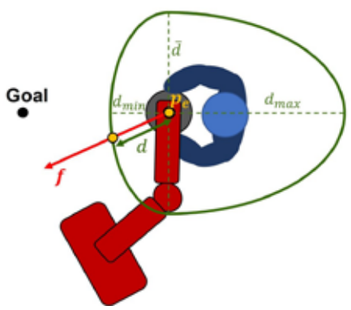

Fig. 1
Pictorial view of the space dependence of the damping d in the goal-driven variable admittance control

On one hand, manual guidance can be an intuitive means to teach new paths to the robot, on the other hand, it relieves the user physical fatigue, increasing his/her precision and the safety of the operation. Typical applications are the handling of large and heavy objects in industrial scenarios, rehabilitation and surgery in the healthcare. These operations are difficult to be completely automated. Hence, the joint work of a human and a robot represents the best solution to effectively accomplish the task.
This work aims at improving the effectiveness and the synergy of human-robot physical interaction in manual guidance operations, both in free space and in contact with an unknown environment. The developed control and estimation algorithms endow the robot with new capabilities, helping the human in complex manual guidance operations which also involve rotational motions and obstacles along the path towards the goal.
Since the objective of any manual guidance task is to drive the end-effector to a predefined target location, in this thesis we deeply investigated an effective method to actively assist the human in directing towards a goal position and accurately reach it. Indeed, so far the problem of directing towards a specific goal has always been charged to the human. This can be a challenging task when the human cannot directly see the target location mainly due to the large size of the transported object or to a cumbersome tool mounted at the end effector. Hence, we developed a variable admittance control, named goal-driven, based on a new physical interpretation of its parameters (see Fig. 1) that actively assist the human in directing towards a goal position. It provides the user with an intuitive directional haptic feedback that allows the human to accurately reach a predefined goal position (e.g. the unloading station) even with closed eyes (e.g. the cargo obstructs the worker view).
According to the flexibility requirements of Industry 4.0, the same manipulator is typically used to handle different types of loads towards the associated unloading station. Hence, making the robot capable of discriminating which target position the human is intended to reach is beneficial to improve the assistance to the human. In this light, we developed an inference algorithm which endows the robot with the capability of estimating the human reaching target among a predefined set of 3D goal positions. This is jointly applied with the above-mentioned goal-driven control to preserve the assistance provided by the latter also in the scenario of uncertain reaching target.
In several hand-guidance applications it is of paramount importance to constrain the end-effector's movements inside a safe volume, where the user still retains full control of the operation. For instance, in industrial frameworks, movements outside secure regions may result in dangerous collisions, while, in surgical applications, any unsafe motion may damage tissues of the patient.
To tackle this issue, virtual fixtures (VFs) are a powerful solution and can increase safety, accuracy and speed of robot-assisted tasks. Indeed, the operator can interact with higher confidence and precision. Therefore, we propose novel translational virtual fixtures to help the human accomplish articulated tasks. To increase the physical intuition of these invisible constraint geometries, the goal-driven control is jointly applied with the VFs, achieving higher productivity and accuracy.
Medical surgery operations and sophisticated peg-in-a-hole insertions might require to limit not only the position of the end-effector, but also the orientation. As such, we developed rotational goal-driven control and rotational VFs. Two different approaches have been proposed: the former based on the Euler vector description of the orientation and the latter based on the rotation matrix and spherical geometry. Rotational VFs and goal-driven control are even more helpful than the translational ones because rotational motions are not easily comprehensible to a human.
In modern factories composed by several independent but highly interconnected working isles, the operator might be required to manually guide the robot to transport bulky objects to the desired target location within this cluttered environment. Hence, it is crucial to guarantee that he/she can accomplish effectively the task without the risk of collisions. To achieve this objective, an obstacle avoidance technique inspired by Rapidly exploring Random Tree* (RRT*) is conceived for manual guidance. This approach is combined with the goal-driven variable admittance control to inform the user about the optimal motion direction by means of its directional haptic feedback.
In the last few years, manual guidance operations have found an application in contexts where the tool mounted at the end-effector is also in contact with an external environment to perform machining on it (e.g. polishing). To manage this new application scenario, it is necessary to both estimate the stiffness and the orientation of the environment and control the force exerted on it. Hence, we realized a dedicated force controller that relies on two estimators of the environment crucial features. Besides, the human can guide the robot on the tangential plane to the surface by leveraging the goal-driven variable admittance control so that the operator can achieve high precision even if a cumbersome tool mounted at the end effector limits his/her view.
All the methods proposed in the thesis have been validated in experimental campaigns with volunteers operating on an industrial robot (see for example Fig. 2).

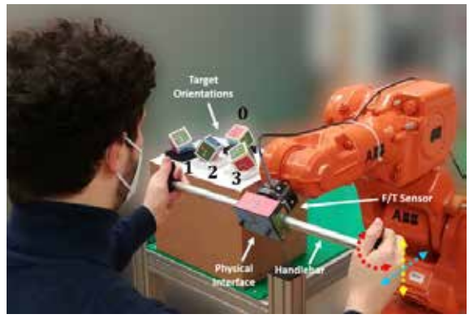

Fig. 2
Experimental setup

AN ASSESSMENT OF RECENT TECHNIQUES FOR QUESTION DIFFICULTY ESTIMATION FROM TEXT

Luca Benedetto - Supervisor: Prof. Paolo Cremonesi

Recent years have witnessed an exponential growth in the availability of digital services, and the educational domain was no exception. The popularity of Massive Open Online Courses increased massively, enabling hundreds of thousands of students to access online learning content and online exams. Similarly to what happened in other domains, this increase in the amount of available data enabled the development of many data driven techniques to improve students' learning experience and the effectiveness of learning material.
One of such improvements deals with Question Difficulty Estimation (QDE) - also referred to as "question calibration" - which is the task of estimating a value, either numerical or categorical, representing the difficulty of a question. The role of QDE in the educational domain is crucial. An example is Computer Adaptive Testing (CAT), which consists of providing students with questions whose difficulty is targeted to their proficiency. Which has been shown to be highly beneficial to the students' learning outcome. In case of miscalibrated questions, the effectiveness of CAT is reduced massively. Indeed, exercises that are not challenging easily lead to boredom and stagnation, whereas overly complex exercises might result in frustration. Also, a too easy or too difficult test results in a limited range of scores, which is not informative.

Traditionally, QDE is performed with either manual calibration or pretesting. Manual calibration consists of having one (or more) domain experts manually selecting a numerical or categorical value representing the difficulty of each question which is intrinsically subjective and has been shown to lead to inconsistent estimations.
Pretesting, on the other hand, consists of estimating question difficulty based on posterior performance measures. Specifically, the questions under pretesting are deployed in an exam, as if they were standard questions, and are calibrated using the students' answers to the other questions. Although this approach leads indeed to an accurate and reliable estimation of question difficulty, it introduces a long delay between the time of question generation and when the questions can be used to assess students. Also, it requires the new questions to be shown to students before being actually used to score them, which is in some cases undesirable, as they might be leaked or exposed too often.

In order to overcome the limitations of traditional approaches to question calibration, recent research has attempted to leverage Natural Language Processing (NLP) to automatically estimate question difficulty at creation time. Such works are all based on the idea that question text is the only information that is always available at the time of question generation and, if we were able to perform an accurate QDE from textual content, we would overcome the need for pretesting, manual calibration, and their limitations. In this thesis, we focus on the task of QDE from Text (QDET), evaluating and comparing different approaches proposed to address it, both the ones modelling it as a supervised task and the ones modelling it as an unsupervised task.

Almost all the approaches proposed in previous research are trained in a supervised manner. Starting from a set of questions of known difficulty, a machine learning model is trained in a supervised manner to estimate question difficulty from text. The trained model can then be used to estimate the difficulty of newly created questions (of unknown difficulty) without the need for pretesting or manual calibration. In this work, we propose a taxonomy based on question characteristics to categorise the approaches proposed in previous literature, and experiment on three datasets (two being publicly available) from diverse educational domains to evaluate how different architectures perform, especially focusing on the relevance of different types of features.

Supervised QDET targets the limitations of traditional approaches to QDE, but it has some limitations of its own: crucially, it requires a large dataset of calibrated questions for training, which might hinder its effectiveness. The required number of questions depends on the specific architecture, but even the simplest models require hundreds or thousands of training questions.

Targeting this issue, some recent research experimented with unsupervised approaches to QDET. Compared to the supervised techniques, the main advantage of unsupervised approaches is that they do not require a large training set of calibrated questions, although they might require supervision in a related (but different) task. There is very limited research on unsupervised QDET and, in this thesis, we experiment on two real world datasets (one being publicly available) to evaluate previously proposed techniques, and propose and evaluate two novel approaches.

Experimenting on questions of different nature and coming from different educational domains, we observe that the choice of the best performing model heavily depends on the nature of the questions under consideration. Considering supervised approaches in the Language Assessment (LA) domain, we observe that readability indexes and linguistic features can capture most of the question difficulty, and this is true both for supervised approaches and unsupervised techniques. Indeed, especially in reading comprehension questions, the item difficulty heavily depends on the linguistic demands of the reading passage, therefore techniques which can capture this information are capable of an accurate difficulty estimation. Still, the same linguistic demands are captured - even more accurately - by more advanced models, such as BERT, which are generally capable of better performance. In this sense, we argue that Transformer-based models are probably the better choice from an accuracy point of view, but much simpler models, such as the ones based on readability indexes, might still be a reasonable choice in case of constraints from the computational point of view.

The same is not true for the Content Knowledge Assessment (CKA) domain. Supervised models based on simple techniques such as linguistic features and readability indexes are not capable of accurately capturing the demands of exam questions and therefore lead to inaccurate estimations. This is because in CKA the question difficulty mostly depends on the specific topics which are being assessed by the question, and techniques that focus on language only cannot capture such information. Specifically, we observed that the Transformers are, again, the models that generally lead to the best performance for supervised QDET in the CKA domain, and are in some cases matched by techniques based on word embeddings (such as Word2Vec) or frequency based features (such as TF-IDF). In addition to that, Transformers can be pre-trained on additional documents related to the same topics as the ones assessed by the questions, which increases their accuracy.

Moving our focus to unsupervised QDET, the observations are only partially different. We observe that in LA the readability indexes are a good proxy of question difficulty, and perform on their own even better than the techniques based on Question Answering (QA) models. However, as we observed for supervised QDET, readability indexes are not capable of producing accurate results in the CKA domain. Indeed, in this case the best results are obtained with techniques that leverage the answers of QA models which are trained to answer the questions under calibration to train an IRT model, mimicking pretesting with real students.

Considering the two categories of approaches - supervised and unsupervised - supervised approaches are better at producing a difficulty which is adapted to the current difficulty distribution, but unsupervised approaches provide a decently accurate overall ranking of question difficulties.

ALGORITHMS FOR RISK-AVERSE REINFORCEMENT LEARNING

Lorenzo Bisi - Supervisor: Prof. Marcello Restelli

Reinforcement learning (RL) aims at solving sequential decision-making problems by means of a learning process that, in a trial-and-error fashion, is capable of gradually improving the quality of the produced decisions, guided by a reinforcement signal. Recent developments in this discipline, coupled with the great approximation power of deep neural networks, have allowed to obtain astonishing results on many challenging fields such as board games, robotic locomotion, single-player and multi-player videogames. The standard RL framework focuses on maximizing the performance in expectation, without considering its variability. However, in many real-world high-stakes scenarios, such as finance or healthcare, it is of fundamental importance to also consider the risk that is connected to a certain behaviour.
Since there are several possible ways to measure risk or to model risk-aversion, a plethora of risk-averse approaches have been developed in the RL literature in the past years. This means that, in order to transfer the advantages of state-of-the-art developments to the risk-averse setting, one has to explicitly extend (if possible) the considered methods for the target risk-averse objective. This complicates the use of reinforcement learning in risk-averse tasks, thus, limiting its applicability to some relevant real-word settings. In this dissertation, we take a step towards overcoming these limitations, by proposing novel methods that allow to easily transfer the advantages of state-of-the-art risk-neutral approaches to the risk-averse setting. The first contribution consists in proposing a single framework to optimize some of the most popular risk measures, including conditional value-at-risk (CVaR), utility functions, and mean-variance. Leveraging theoretical results on state augmentation, we transform the decision-making process so that optimizing the chosen risk measure in the original environment is equivalent to optimizing the expected return in the transformed one. We then present a risk-sensitive meta-algorithm, called ROSA, that transforms the trajectories it collects from the environment and feeds these into any risk-neutral policy optimization method. Extensive experiments show the benefits of our approach over existing ad-hoc methodologies in different domains, including a financial task, based on a real-world trading dataset, and a robotic locomotion one. Figure 1 shows the performance obtained in a robotic locomotion task called "Walker" by ROSA, which is parametrized in order to optimize two different risk-measure: CVaR (Figure 1a) and mean-variance (Figure 1b). For each of these settings, two state-of-the-art approaches are applied, TRPO and PPO. Moreover, two different risk-aversion level are tested for each risk measure. In each of the examined cases, ROSA is successful in optimizing its objective.

The second contribution consists in considering, for the first time, risk-measures connected to the state-action occupancy distribution, instead of the return one. We define a novel measure of risk, which we call reward volatility, consisting of the variance of the rewards under the state-occupancy measure, and we study the optimization of a trade-off objective called mean-volatility. We show that this measure is an upper bound of the traditional return variance, hence, minimizing the first one may be seen as a proxy for reducing the second one. Moreover, an expectation Bellman equation holds also for this measure, while it is not possible to obtain an optimality one. Exploiting this result, we derive a policy gradient theorem for the trade-off objective, and we show how to provide monotonical improvement guarantees. Inspired by the theoretical guarantees, we derived also a practical algorithm, called TRVO, which extends a well-known trust-region approach, TRPO, to the risk-averse case. We tested our approach on two financial tasks, and we show the results for one of them in Figure 2. In this figure the trade-off between the risk-neutral performance (the expected return) and some kind of risk-measure (the reward-volatility in Figure 2a and the return variance in Figure 2b) is illustrated. The approximated Pareto frontiers for each of the compared approaches are obtained by running the corresponding algorithm with different risk-aversion coefficients. It can be noticed that the TRVO Pareto front dominates the other one for the mean-volatility trade-off in Figure 2a. TRVO manages to obtain also a good frontier w.r.t. the mean-variance trade-off in Figure 2b.

As a third contribution, we study the convergence rate of an actor-critic approach optimizing the mean-volatility criterion. The analysis is carried out by using two different policy evaluation methods we developed: the direct and the factored one. In this setting, we extend recent analyses in the risk-neutral actor-critic setting to the mean-volatility case, to establish a PAC-bound for sample-complexit. We validate our theoretical result by means of an empirical study on a stochastic environment.
To conclude, we provided methods for a more efficient optimization of some risk-averse objectives, and we offered insights on the risk-averse learning process by means of theoretical analyses and experimental evaluations. Future research may involve extending the ROSA framework to other risk-measures or studying novel coherent risk-measure based on the state-occupancy distribution.

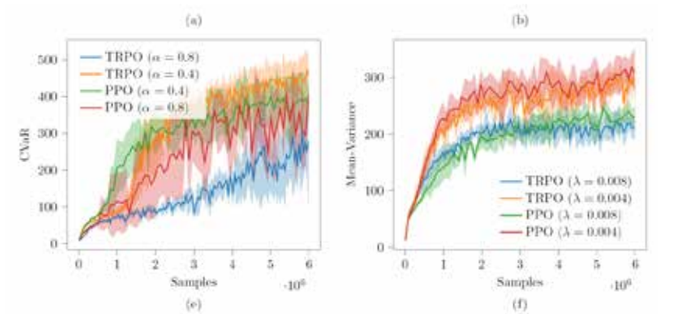

Fig. 1
Shaded areas represent the standard deviation between 5 runs, solid lines represent their means.

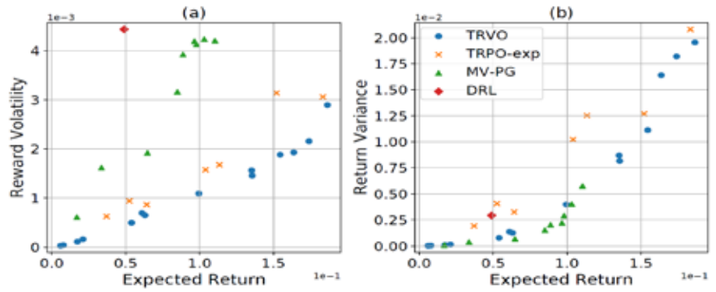

Fig. 2
The same policies are represented w.r.t. two different trade-offs.

MACHINE LEARNING FOR OPTIMIZATION OF ENERGY INTENSIVE INDUSTRIAL PROCESSES

Alessandro Brusaferri - Supervisor: Prof. Matteo Matteucci

The digital transformation is providing unprecedented opportunities in industry, to increase flexibility, efficiency and sustainability of processes, fundamental to maintain the competitiveness in the global market, and to contribute to the achievement of the Sustainable Development Goals and the targets defined in the European Green Deal. A major challenge is to transform the huge amount of measurements collected from the plants into knowledge (i.e., in terms of models) and software tools, supporting the dynamic identification of the best operational strategy to be executed.
Machine learning is a pervasive technology, demonstrating impressive performance across a wide range of applications in the computer science field, such as vision, speech recognition, and language processing. Therefore, it is subject of increasing research interest also in the manufacturing context, as a general purpose framework to achieve reliable predictive models from rough data streams.
In this work, we investigate novel machine learning approaches, aimed to support the development of advanced optimization tools for energy intensive industries, fostering more sustainable consumption patterns.
In particular, we focus on two major functional components needed to realize such implementations, namely the integration of reliable short-term energy price/load forecasting, and data-driven behavioral models of discrete/hybrid processes.
Specifically, we first study a novel probabilistic forecasting approach, based on a Bayesian Mixture Density Network architecture, inferring general conditional densities within an end-to-end learning framework including features selection. Both aleatoric and epistemic uncertainty sources are encompassed within the overall predictive distribution, to enable what-if scenario/consumption analysis before trading, enhanced risk evaluation and the ability to plan multiple production strategies for the range of possible prices outcomes.
To achieve reliable and computationally scalable estimators of the parameter posterior, both Mean Field variational inference and deep ensembles are integrated. The proposed approach is demonstrated on both synthetic problems and real-world forecasting tasks, namely day-ahead prediction for the Italian PUN energy market, electricity consumption over 8 ISO-NE regions in the United States, and a very short-term power forecast problem, namely the 15 minutes ahead electricity absorption of an induction furnace from an industrial foundry.
Afterwards, we target the identification of behavioral models of discrete systems through Recurrent Neural Networks (RNN). The goal here is to increase explainability, as the rationale behind the network responses is encoded in an implicit way, which is difficult to be interpreted by practitioners. Once revealed, such mechanisms provide deeper insights into the model execution, enhancing conventional performance evaluations, thus increasing trust

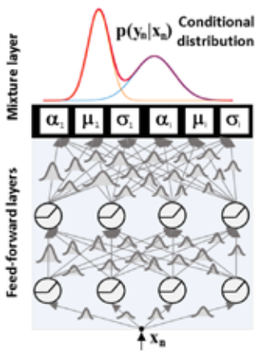

Fig. 1
Bayesian Mixture Density Network

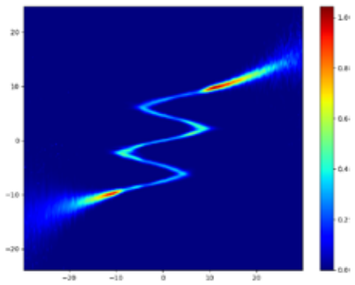

Fig. 2
Predicted conditional distribution

and consequent adoption in industry. Hence, we propose a new approach based on the introduction of a Gaussian Mixture-based clustering layer, constraining the network to operate on a discrete latent state representation. By processing context-input conditioned transitions between clusters, a human interpretable Moore Machine based characterizing the RNN computations is extracted.
The proposed approach is demonstrated on both synthetic experiments from an open benchmark problem and via the application to a pilot industrial plant, by the behavior cloning of the flexible conveyor of a remanufacturing process.
The identification of hybrid patterns over the measured data sequences is a further key issue to be address, to properly represent the heterogeneous interactions occurring between control logic/rules and continuous process dynamics, including sharp changes in operating points, plants regimes, constraints on values of system inputs/outputs, etc.
In this context, we focus on two general classes of hybrid systems, proposing specialized model architectures. Both models are conceived following a probabilistic approach, constituting extension to the standard Mixture of Expert framework.
First, we study the identification of piecewise auto regressive with exogenous input models with arbitrary regions, thus not restricted to polyhedral domains, and characterized by discontinuous maps. To achieve nonlinear partitioning, we parametrize the discriminant function using a neural network. The parameters of both the arx submodels and the classifier are first estimated by maximizing the likelihood of the overall model using Expectation Maximization. Then, we investigate a Bayesian inference approach to assess the epistemic uncertainty.
Afterwards, we considered a generalization of switched hybrid systems, characterized by nonlinear autoregressive exogenous components, with finite dimensional polynomial expansions, and by a hidden Markovian transition mechanism. Model structure selection in the polynomial expansions represents a central feature in this class of systems, to achieve parsimonious and more explainable representations. To this end, we deploy a two stage selection scheme, based on a l1-norm bridge estimation followed by hard-thresholding.
The proposed techniques are demonstrated on benchmark problems from the hybrid system identification field, namely a nonlinear piece-wise system with discontinuous maps and a SMNARX problem composed by three nonlinear sub-models with specific regressors.
Besides the complementary utilities provided to the realization of energy-aware optimization tools, the approaches developed in this thesis shares a further leitmotif. In fact, they constitute multiple extensions to probabilistic mixtures models, especially implemented to achieve flexible conditional density estimation, probabilistic latent space clustering and enhanced Mixture of Expert architectures.

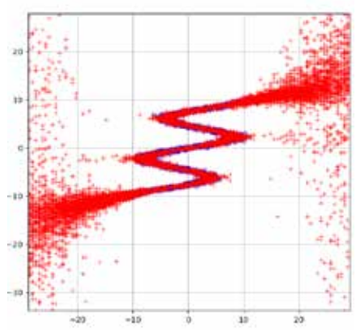

Fig. 3
Samples generated from the model

LEARNING EFFICIENT AND EFFECTIVE REPRESENTATIONS FOR EVENT-BASED CAMERAS

Marco Cannici - Supervisor: Prof. Matteo Matteucci

INTRODUCTION

Neuro-biological systems process information in an asynchronous, sparse, and energy-efficient manner, as opposed to computers, which employ synchronized logics, high-speed clock rates, and energy-demanding computation. The difference between biological retinas and traditional vision devices offers a clear illustration of this disparity. Conventional cameras gather visual information by taking full-frame pictures collecting light at a constant and preset rate. While these images closely resemble what we picture in our minds, they typically contain highly redundant information, unlike neural signals. Indeed, all sensor's pixels are synchronously read to form the final image regardless of whether something has changed since the last image has been taken.

This way of capturing visual information deviates from biological retinas, which leverage a much more efficient operating principle. Light hitting the retina initiates a sequence of chemical and electrical processes triggering impulses that ultimately reach the vision centers in the brain. Retinal ganglion cells collect neural signals from photoreceptors, altering their electrical charge and causing them to fire output signals after reaching a certain potential threshold. The same paradigm is adopted in the following processing layers of the visual cortex, which trigger neural impulses only when enough relevant visual information has accumulated. This procedure results in an entirely asynchronous and energy-efficient system capable of performing computation only when required. Inspired by these mechanisms, neuromorphic vision devices, also known as event-based cameras, are vision sensors that attempt to emulate the functioning of biological retinas. An array of independent pixels generates an output signal, i.e., an event, anytime the local brightness level changes by a given threshold, simulating a simplified model of the retina's photosensitive membrane. Like in a biological vision system, information is produced asynchronously and only when needed, resulting in a very efficient device with many advantages over traditional ones, including microseconds temporal resolution, high dynamic range, and minimum requirements for power consumption and bandwidth. Many of the operating principles of these artificial retinas are shared with biological ones, giving us hope that we will soon be able to create visual systems with the same precision and efficiency as biological ones.

MOTIVATION

The reason for the success of modern computer vision systems resides in their ability to learn to reason directly from experience, without any prior knowledge on the task. Inspired by early neural computation models, many Machine Learning systems accomplish this by extracting meaningful features and combining them in complex and general patterns through a hierarchy of increasingly sophisticated representations. Deep neural networks, which are nowadays the beating heart of most vision systems, are capable of naturally extracting this hierarchy as they are organized on several processing layers that progressively refine information. Thanks to ad-hoc training algorithms, these layers can be jointly trained to extract effective intermediate representations that make even the most difficult visual tasks simple to solve.

How well such artificial systems perform on a particular task is often linked to how rich, informative, and general their internal representations are. However, the paradigm used to extract such representations is specifically designed to operate on dense visual encodings. As a result, learning to extract effective representations from the dense images acquired by standard cameras is remarkably easier than performing the same from the sparse and asynchronous output produced by event-based cameras. Indeed, while a single image directly conveys rich visual data, the same appearance information needs to be reconstructed from events through temporal reasoning, as it is spread across the sequence of asynchronous and incremental updates, making the task of learning such representations far more challenging. Designing novel mechanisms to extract effective representations and new computing paradigms capable of exploiting the asynchronous nature of events is thereby critical for unlocking event-based cameras' potential in modern computer vision architectures.

Spiking Neural Networks, a type of artificial neural network that mimics both the learning and dynamics of biological neurons, are appealing in event-based vision due to their energy efficiency and asynchronous processing paradigm. However, their complex dynamic makes learning with these architectures very challenging to accomplish, limiting their usage in complex visual tasks. As an alternative, and thanks to the success of deep learning in frame-based computer vision, researchers have recently started exploring the potentiality of deep neural networks, such as the representation power previously discussed, even in event-based vision. They are, however, less efficient than spiking networks and typically designed to process synchronous and dense data streams, making it difficult to exploit their ability to learn when asynchronous data is employed. The tradeoff between these two solutions, combined with the almost complementary benefits they provide, begs the question of whether it is possible to draw inspiration from the asynchronous and incremental processing of spiking networks to make deep neural network representations more suited at processing events.

RESULTS

The thesis addresses this challenge by focusing on three aspects of designing deep neural networks for event-based vision. First, we look at how to efficiently compute hidden neural representations by preserving event-based cameras' properties during computation. We design a framework for converting deep neural networks into systems with identical expressiveness but capable of asynchronous processing. Thanks to an event-driven formulation of the convolution and max-pooling operations and an additional memory of previously extracted representations, these layers achieve incremental and sparse computation, thus retaining the event camera's asynchronous and data-driven nature.

Then, we focus on performance and study how to learn effective input representations for a given task. We propose MatrixLSTM, a recurrent mechanism that automatically learns to interface with any convolutional network by sparsely and incrementally building a frame-like representation from asynchronous events. We show that MatrixLSTM can provide powerful representations in several tasks, including object recognition, optical flow prediction, and object detection.

Finally, we focus on the challenging task of training neural networks to operate effectively on a real-world event-based camera when the only source of training supervision comes from simulation. We tackle the problem from a domain adaptation perspective by proposing DA4Event, a general procedure for training event-based neural networks on simulated data with minor performance loss. We study how the use of simulated data during training affects different event representations and how DA4Event can help in reducing performance degradation on object recognition and semantic segmentation tasks. Throughout the thesis, we explore the importance of representations in event-based networks, at both the input and hidden layers, and show that, by focusing on these aspects, considerable gains can be achieved toward more effective and efficient processing.

LEARNING IN NON-STATIONARY ENVIRONMENTS: FROM A SPECIFIC APPLICATION TO MORE GENERAL ALGORITHMS

Giuseppe Canonaco - Supervisor: Prof. Manuel Roveri

Machine Learning (ML) has recently become more and more effective in modeling highly complex phenomena given sufficiently large and descriptive data sets. This characteristic makes it a suitable tool for a plethora of applications in the most disparate fields such as: computer vision, robotics, healthcare, natural language processing, transportation, industry etc. ML can be described as the set of all algorithms able to automatically learn to perform a certain task using data which can be regarded as experience. The more the experience the better the algorithm will learn the task.
The ML field can be classically split into three different macro-areas: Supervised Learning (SL), Unsupervised Learning (UL), and Reinforcement Learning (RL). SL techniques deal with problems where the supervised information is available and strive to obtain a function $\hat{y} = f(x)$ whose objective is to correctly predict the output y, called supervised information, associated with the input vector x. UL techniques, instead, strive to learn the underlying structure of the data without having access to the supervised information. Finally, RL techniques strive to find the optimal policy to be executed by an agent on the environment in order to reach a certain pre-specified goal. The optimal policy is learned via an optimization process whose objective is to maximize the long-term cumulative reward, where the reward is what the agent receives upon executing an action on the environment itself.
The algorithms and techniques developed within the huge ML field are equipped with some assumptions which are often not satisfied in practical applications. For instance, SL algorithms assume to have enough labeled data so that predictive models can be properly trained. This assumption does not always hold in real-world scenarios where the labeling process could be too costly or time-consuming. An instance of this setting is in the context of corrosion prediction for pipeline infrastructures, where, usually, the supervised information about the presence of corrosion is hardly available for a pipeline of interest due to the intrinsic cost companies have to bear in order to collect it. For these kinds of applications, where the supervised information is very scarce if not missing at all, standard SL learning techniques are not able to provide good predictors for the objective phenomenon. However, if supervised data is available for some related problem, we could leverage Transfer Learning (TL) which coupled with SL will allow us to alleviate the need of supervised information in the target problem object of interest. TL is a transversal ML sub-field dealing with knowledge transfer across different tasks. In order to be successful, a given TL technique needs to answer three main questions dealing with "what", "how", and "when" to transfer the knowledge across different tasks, which translates into deciding what form of knowledge to be transferred, the appropriate way to transfer it, and, most crucially, when to execute the transfer. The last question is supposed to deal with the negative transfer phenomenon that happens whenever source and target tasks are not sufficiently similar to each other. In the SL context, we talk about inductive TL whenever there is some supervised data coming from the target task, instead, we talk about transductive TL whenever there is only unsupervised data coming from the target task. In both the previously mentioned cases there is plenty of supervised data coming from the source tasks. Finally, we talk about unsupervised TL whenever the supervised information is missing both from the source and target tasks. In this context, clustering or dimensionality reduction problems are usually transferred. On the other hand, for what concerns RL, the situation is a bit more complex since we have an agent-environment interaction where the environment is modeled through a Markov Decision Process (MDP) and the agent by a policy. In this context, TL allows a greater sample efficiency, and transfer algorithms may be distinguished in: techniques able to deal with source and target tasks that have different state or action spaces; and techniques working under the assumption that both state and action spaces will stay the same among the source and target tasks. The applications of TL are disparate and they span different fields such as robotics, games, natural language processing, healthcare, bioinformatics, recommender systems, corrosion, etc.

Besides the above-mentioned data-availability requirement, another crucial assumption tied to ML techniques is about stationary data-generating processes. This means that the phenomenon we are trying to learn does not evolve with time and allows the ML algorithm to converge to a solution for the problem of interest. In the SL context, this translates into the fact that the couples (x,y) always come from the same distribution. For what concern UL, instead of having couples (x,y) we will just have the vector x that still will always come from the same distribution. Finally, in the RL setting, it is the MDP, modeling the environment and the allowed interactions, that will always stay the same, whereas, for what concerns TL, it is the available historical knowledge that does not expose a time-variant structure. However, there are many applications where the above-mentioned assumption about stationarity does not hold, e.g., finance, due to market evolution, robotics, due to faults affecting either sensors or actuators, water reservoir systems, due to the climate change our planet is currently undergoing, corrosion, where the wear and tear of equipment or infrastructures increase with time, etc. In all of these applications ML techniques cannot be directly employed without taking particular care of the time variance at play, that, depending on the particular framework we are using, will affect the distribution generating the couples (x,y), the distribution generating the vector x, the MDP, or the available historical knowledge.

In the context of this dissertation, guided by the specific application needs of corrosion prediction in pipeline infrastructures, we will investigate ML solutions able to weaken the assumptions of available supervised information and stationarity. Despite being a very well researched area thanks to the thriving field of TL, reducing the requirement of supervised information in the context of corrosion prediction is not investigated at all motivating the research of tailored solutions for this critical application. Weakening the assumption of stationary phenomena, instead, is much less studied in lots of ML sub-fields motivating a much more general investigation in the context of this dissertation.

The contribution of this dissertation is threefold. The first one deals with learning techniques for corrosion in pipeline infrastructures starting from the data set creation up to devising SL techniques that, with the aid of TL, are able to build a model of the corrosion phenomenon without using supervised information coming from the pipeline of interest. The second one deals with RL in non-stationary environments and develops both an active-adaptive approach to cope with changing environments and a TL technique for RL able to take into account an underlying time-variant structure intrinsic to the available historical knowledge. Finally, the third one deals with Federated Learning under non-stationarity and pervasive systems. This last part introduces a passive-adaptive approach to mitigate non-stationarity in FL contexts and a birdsong detection approach able to run on a highly constrained Internet of Things unit.

DECISION-MAKING UNDER UNCERTAINTY FOR COMPLEX SOCIO-ENVIRONMENTAL SYSTEMS

Angelo Carlino - Supervisor: Prof. Andrea Castelletti

The concept of socio-environmental systems is used to provide a framework to study the complex sets of dynamic interactions linking human and environmental systems. Socio-environmental system models used to inform the grand policy challenges affecting our society and its relationship with the environment are often relying on simplifying assumptions about the deep uncertainty involved in their dynamics and evolution. The most common approaches assume a reference, or a set of multiple reference scenarios, to study potential policy solutions under the assumption of perfect foresight, an approach called scenario analysis or scenario planning. Other methods involve the use of sensitivity analysis to examine apportion uncertainty to input to the models, but do not yet consider explicitly the setting in which the decision must be taken. To do so, we use optimization under uncertainty and optimal control methods such as robust optimization, and extension of dynamic programming approaches, trying to include and incorporate recent advances on robustness and adaptation frameworks.

The assumption of perfect foresight affects two types of models crucial for the next decades of policy-making: integrated assessment of climate change and long-term energy systems planning models. These are also plagued by inherent deep uncertainty about future socio-economic and environmental projections. Focusing on those two categories of models, this thesis focuses on how to effectively deal with uncertainties of different nature, i.e., stochastic, parametric, and structural, by adopting and refining existing methodologies for decision-making under uncertainty.

In particular, we examine how adaptive decision-making via multi-objective optimal control can help reduce conflicts between multiple climate targets in a deeply uncertain version of DICE, a well-known cost-benefit integrated assessment model of climate change. Indeed, the adoption of this method results in a flexible management of adaptation and mitigation scenarios, which also resolves a long-standing issue in climate policy. The same method is extended for a cooperative multi-agent context in RICE50++, the regionalized version of DICE, considering 57 individual economic agents. In this case, multi-objective adaptive decision-making improves the temperature outcome as the number of years above 1.5°C is reduced, while alleviating income inequality between the regions. An example of allocation of emission control effort is reported in Fig. 1.

For what concerns energy systems planning models, we focus on the impact of different socio-economic and hydroclimatic uncertainties on power capacity expansion strategies. We first provide evidence for the need of including multisectoral uncertainties into energy system planning bounding uncertainty over future hydroclimatic impacts to the EU electricity system during the transition to a decarbonized state, as reported in Fig. 2. After recognizing the importance of an adaptive approach also in this domain, we employ robust optimization to develop a methodology that allows accounts for uncertainty and adaptation in an open-source energy system model. We apply this method to the more than 300 African hydropower projects to extract the most relevant with respect to final energy demand satisfaction under socio-economic and climate policy uncertainty.

Results show that accounting for uncertainty with coherent methodologies results not only in improved realism of the decision-making process but also in a clearer understanding of how to deal with and leverage uncertainties to produce satisfying performance across multiple dimensions.

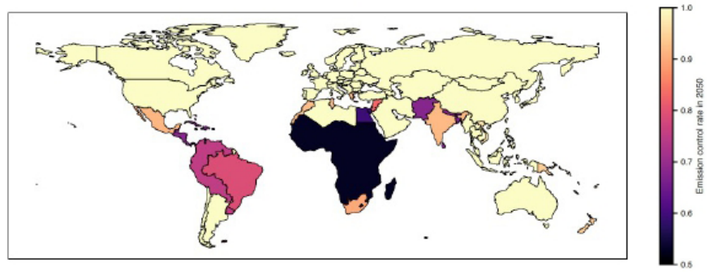

Fig. 1
Emission control effort in 2050 using RICE50++ and self-adaptive multi-objective climate policies. Poor countries are given more time to ramp up emissions so that those resources are not taken away from economic growth in the short term.

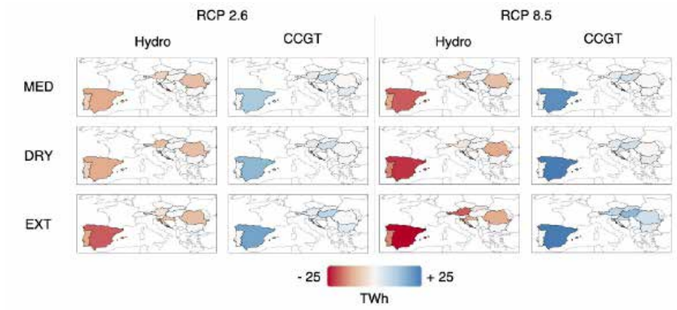

Fig. 2
Differences in hydropower and combined cycle gas turbine generation under the different hydroclimatic scenarios examined for the Iberian Peninsula and the Danube River basin. The lost hydropower due to reduced water availability is mostly balanced by natural gas.

DESIGNING AND ENGINEERING EMOTION-AWARE CONVERSATIONAL AGENTS TO SUPPORT PERSONS WITH NEURO-DEVELOPMENTAL DISORDER

Fabio Catania - Supervisor: Prof. Franca Garzotto

Conversational agents are software that can provide access to information and services through spoken natural language. One of the most important research challenges in human-computer interaction (HCI) is the development of conversational agents that can interpret the surrounding context. Some researchers in the field argued for integrating emotion recognition skills into conversational agents to make them aware of the user's emotional context and let them behave empathetically with it. Still, automatic emotion recognition is a challenging task and today is not commonly integrated into conversational agents.
Despite emotions being studied for many years, they are still not fully understood. Moreover, some psychologists state that emotions vary in some aspects by culture and language, and resources in many cultures and languages are scarce. For example, the only speech emotional corpus enabling linguistics and speech processing research in Italian is Emovo, and it counts only 588 audio recordings by six actors.
This thesis takes inspiration from the belief that emotion-aware conversational agents might empower the therapy of people with Neuro-Developmental Disorders (NDD). NDD is a group of conditions with onset in the developmental period characterized by severe cognitive, social, and communication deficits. For instance, people with NDD often show impaired awareness of their and others' emotions and struggle to manifest and describe their feelings, which are typical disturbances associated with alexithymia. Conversational agents have recently been identified as a potentially beneficial means to complement more traditional interventions for people with NDD since the interaction is perceived as safer, more predictable, and more straightforward than communication with other humans. However, given the broad spectrum of special needs to be addressed in people with NDD and the very multidisciplinary aspects to consider when designing and evaluating conversational agents for this target population, there is still limited knowledge about the perception, usability, and effectiveness of conversational agents in this field. In addition, just a few studies explored the use of conversational agents to reduce alexithymia, and they focused on supporting emotion expression and recognition only through words, face, and body.
In this thesis, we explored the potential of emotion-aware conversational agents to promote the capability of expressing emotions using the voice in people with NDD. First, we sought to answer the following research question concerning the design of conversational agents:
- RQ1. Can conversational agents with speech emotion recognition skills help people with NDD improve their emotion expression skills?

Once done with RQ1, we focused on studying the robustness and effectiveness of conversational technology in an emotional context. We addressed the following research questions in the field of speech emotion recognition:
- RQ2. Does emotional speech bias speech recognition?
- RQ3. What can be done to push forward the state of the art (SOTA) in speech emotion recognition in a language with limited resources, such as Italian?

To address the research questions mentioned above, we took a multi-perspective approach considering (i) conversational agents for NDD and (ii) speech emotion recognition. We began by conducting a systematic review of the SOTA on conversational agents to better understand their use to support people with NDD and how they could be beneficial for enhancing people's skills. Second, we run two empirical studies to investigate whether conversational agents can support people with NDD during their therapy. The first experimentation involved nine children with NND and aimed at investigating the introduction of a conversational agent working as a generic assistant, i.e., Google Assistant, into a therapeutic context. The second investigation involved 19 participants with NDD who performed five sessions with a conversational agent over two and a half months and explored its usability, perception, and therapeutic effectiveness. The agent, namely Emoty, was developed ad-hoc in collaboration with psychologists and therapists and was based on the knowledge gained from the literature review and the first study. It acted as an emotional facilitator and trainer, entertaining users with small talks, asking them to verbalize sentences expressing specific emotions with an appropriate tone of voice, and providing feedback about their "acting performance", which is the degree to which their verbalizations expressed those emotions.
Regarding speech emotion recognition, first, we studied the most widely used emotional theories, overviewed the existing speech emotional corpora, and surveyed the most promising approaches in audio representation and classification. Second, we investigated the impact of emotions on speech recognition. In particular, we examined the performance of some SOTA speech recognition systems (i.e., Google Cloud Speech-to-Text and IBM Watson Speech-to-Text) when processing neutral and emotional speech from the corpora EMOVO, EMODB, and SAVEE (Italian, German, and English, respectively). Third, we created Emozionalmente, a new, large speech emotional corpus obtained via crowdsourcing.
Finally, we developed some machine learning models for speech emotion recognition trained on both Emozionalmente and Emovo and compared their performance with humans.

The main contributions of this thesis advance the SOTA in conversational agents for the therapy of people with NDD and speech emotion recognition.
- We systematically reviewed the literature on conversational agents for people with neurodevelopmental disorders. The lessons we learned pave the ground for (i) a set of design and methodological recommendations for empirical studies in this field; (ii) the development of a checklist to ensure a high-quality report of these experimentations; (iii) a research agenda for addressing the gaps and opportunities in this field.
- We performed some empirical studies that revealed the challenges and benefits of adopting (emotion-aware) conversational agents during the therapy of people with NDD. We discovered that they generally (i) find it difficult to be understood by machines due to their frequent speech impairments, (ii) cannot quickly adapt to the schematic communicative protocol of conversational agents involving the use of a wake-action, i.e., an operation typically required by the user to trigger the agent at every conversational step, (iii) and can improve their social and communication skills (e.g., emotion expression capabilities) after continuous use of a conversational agent. This research might pave the way for a better understanding of the cognitive, social, and emotional mechanisms associated with NDD and new forms of therapeutic interventions for these subjects.
- We found out that emotions in speech may negatively affect automatic speech recognition performance. This limitation should be considered when developing conversational agents in social contexts to prevent misunderstandings that may dramatically result in the user's dropout.
- We constructed Emozionalmente, a crowdsourced speech emotional corpus containing 6902 samples acted out by 431 non-professional actors while verbalizing 18 Italian sentences expressing anger, disgust, fear, joy, sadness, surprise, and neutrality. This novel resource enables linguistic and speech processing research on emotional spoken language in Italian.
- We computed some speech emotion prediction scores describing the capability of humans and different machine learning models to recognize emotions both in Emozionalmente and the already existing dataset Emovo. Scores are to be considered as a baseline for future investigations and show the potential of speech emotion recognition integrated into conversational agents for social contexts.

TEMPORAL LOGIC AND MODEL CHECKING FOR OPERATOR PRECEDENCE LANGUAGES: THEORY AND APPLICATIONS

Michele Chiari - Supervisor: Prof. Dino Mandrioli - Co-Supervisor: Prof. Matteo Pradella

The ubiquity of computer systems in every industrial sector poses demanding challenges, including the verification of adherence of mission- and safety-critical systems to their requirements. Model checking is one of the most successful techniques developed for this objective. It consists of the formal specification of the system's requirements by means of a logic formalism, the generation of a model of the system by using an operational or denotational formalism, and the automatic and exhaustive verification of the adherence of the latter to the former. The result of this process is either the assurance that the system satisfies the specification, or a counterexample, i.e., a behavior of the system that does not satisfy such requirements.

The system properties that can be verified depend on the choice of the mathematical formalism to be employed for both the model and the specification. Different kinds of temporal logic are most often used for specifying requirements, both because of their ease in reasoning about the system's behavior along time, and because of the efficient model-checking algorithms they allow for. For example, Linear Temporal Logic (LTL) sees time as a linear sequence of discrete events, each one of them represented by an atomic proposition, as shown in Figure 1 (top). This is ideal to express requirements such as "the system will never present unwanted behavior A", "the system will continuously perform task B", or "the system will not have behavior A until it has done B".

In terms of expressiveness, however, temporal logics such as LTL, CTL, and CTL* are limited to requirements expressible as regular languages. This can be a daunting limitation when the system to be verified is a procedural program. Procedures, or functions, are ubiquitous in programming languages, and are arguably the most successful modularization device ever introduced. Hence, any technique aimed at program verification must be able to not only model them as accurately as possible, but also allow for requirements that take them fully into account. To achieve this, the language used for specifications must explicitly support reasoning on their typical behaviors. Much like what happens with natural languages, if this is not the case, then the resulting sentences—and, consequently, proofs about the program—will necessarily tackle a limited part of the system's behavior. Procedural programs use a stack to keep track of function activation records: the resulting behaviors can be described by pushdown automata, and cannot be represented as regular languages.

When writing specifications for procedural programs, we may want to express properties such as Hoare-style pre- and post-conditions (e.g., "if pre-condition P holds when procedure A is called, post-condition Q will hold when it returns"); *stack inspection* (e.g., "privileged procedure A cannot be called if unprivileged procedure B is active on the stack"); or *exception safety* (e.g., "procedure A never throws an exception", and "if procedure B throws an exception, the program remains in a valid state"). These properties refer to a matching between function calls and the return statements or exceptions that terminate them. Since this matching cannot be expressed by regular languages, temporal logics limited to them cannot express requirements related to procedure execution.

This thesis develops a model-checking framework based on Operator Precedence Languages (OPLs). OPLs are a subclass of Deterministic Context-Free Languages, and are significantly more expressive than regular languages. Being suitable for describing the syntax of real-world programming languages, they can dramatically extend the properties expressible in system specifications. In particular, they enable verification of procedural programs with exceptions, which state-of-the-art logics cannot do. OPLs allow us to see time as a linear succession of events, some of which are in relation. For example, Figure 1 (bottom) shows the execution trace of a program in which three procedures are called, and then terminated by an exception; then, another procedure is called and returns normally.

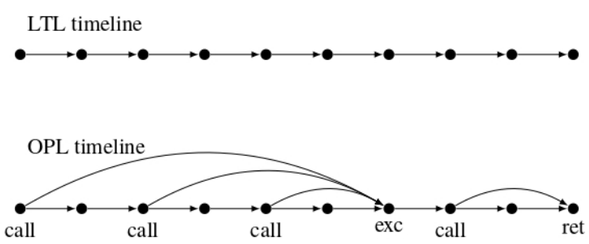

Fig. 1
Timelines considered by LTL vs OPTL/POTL.

In this thesis, we present two temporal logics capable of expressing OPL properties. The first one, OPTL, is a first attempt at this task, for which we develop a model-checking procedure. OPTL features temporal modalities that interact with the multi-matching structure of OP words: there are *matching next/back* operators that state something about the *maximal* (i.e., farthest) position matched with the current one, and OP-*summary until/since* operators that skip positions in between such maximally matched positions (which means skipping procedure bodies). Unfortunately, OPTL still has some limitations in terms of expressiveness. In particular, we prove that is is strictly less expressive than First-Order Logic (FOL), the common yardstick for temporal logic expressiveness.

Thus, we introduce a better logic, POTL, for which we prove equivalence to FOL. POTL is equipped with modalities devised with the idea of "navigating" a word's underlying syntax tree. It features until and since operators that can navigate up or down in the syntax tree, possibly skipping subtrees (which, in the context of procedural programs, means skipping function bodies). POTL can express all OPTL-expressible properties, and has some advantages over it. In particular, it can express properties limited to a single subtree in a word's syntax tree, which allows us to naturally express certain function-local properties that are not expressible in OPTL.

We also explore POTL's practical applicability by developing and implementing an automata-theoretic model-checking procedure for it. The resulting tool, called POMC, accepts both programs modeled directly as automata and by means of a simple domain-specific language. Its evaluation has been carried out on case studies concerning stack inspection, exception safety, and other correctness properties. The tool shows promising results in terms of both functionality and performance.

A NEW PARADIGM TO COMBINE MODEL- AND DATA-BASED DIGITAL TWINS IN SMART MANUFACTURING

Chiara Cimino - Supervisor: Prof. Alberto Leva

The main objective of this thesis is the proposal of a new paradigm to comprehend the many different interpretations of the Digital Twin concept proposed and applied in the context of advanced - or "smart" - manufacturing. The ultimate goal of the long-term research to which the thesis belongs, is to implement the said paradigm in the form of tools to support the management of a production asset throughout its life cycle.

The research motivation comes from observing on the one hand that entities named "Digital Twins" are nowadays of fundamental importance in the design, engineering, commissioning, control, maintenance and management of production assets, and on the other hand, that those entities are as heterogeneous as can be a BIM database, a system of differential and algebraic equations, and a neural network. A paradigm unification as the one here proposed -- which is much more than just transferring data among domain-specific tools, as will be shown -- is therefore of undoubted value, and highly desirable.

From the above, it naturally comes that the thesis shall take a multidisciplinary approach, accounting in particular for the S*ystems and Control* and the *Operations and Management* viewpoints. These are in fact the major ones to come into play along the life of an asset, and at the same time reflect a major division in the Digital Twins zoo. Simplifying here for brevity, in the Systems and Control domain a Digital Twin is generally some kind of simulation model, while in Operations and Management a data-centric interpretation of the Digital Twin idea is most frequently adopted.

As a matter of fact, and quite apparently based on the numerous success stories available, each interpretation is correct in its domain. The potential of modern ICT -- think of cloud applications, (Industrial) Internet of Things and the like, to name just a few developments -- when applied to manufacturing, is nowadays going far beyond the most traditional uses of "simulating" and "controlling" systems, first in a Digital World and then physically. At the same time, as systems need to become flexible and rapidly reconfigurable, ICT is becoming necessary for lifelong decision making support from the management level almost down to the plant floor.

The research we present starts from the consideration, simple at a first glance but instead loaded with a wealth of consequence, that the mentioned Digital World has the nature of multiverse. In fact, it already contains different parallel realities (e.g., as design alternatives being evaluated) and different viewpoints (e.g. the CAD drawing, the P&ID, the control scheme and a queue-plus-server model for a machine). The problem is that all the above knowledge is managed with separate tools, and the burden of keeping the knowledge base consistent (e.g., ensuring that a CAD modification does not invalidate a control study, to clarify that it is not just a matter of data interchange) stands with humans.

Elaborating on the above idea we introduce our paradigm, that we name Digital Multiverse. We show that different Digital Twin interpretation can be seen as "projections" or "views" applied to a more abstracted object, that we term Digital Meta Twin and offers the higher abstraction level envisaged above. We deduce that, to realize the paradigm, a Digital Meta Twin must take the form of a Model and Data Base, where c*onsistency relationships* need instating and automatically enforcing -- as is done in the database domain for data alone -- but here involving both data and models. We define some such relationships, and outline the way their management has to be implemented. We consequently come to defining the structure of a Digital Multiverse tool, and although the realisation of such an object extends far beyond the scope of this thesis, we support its viability by defining and motivating the required mix of available (and solid) enabling technologies

We then apply our paradigm to some relevant problems, also with the aid of industrial case studies drawn from the Lombardia regional research and innovation project AD-COM, "ADvanced COsmetic Manufacturing" (ID 214632) that provided support for the presented research. We end the dissertation with retrospect considerations, open issues, and a consequent sketch of future activities.

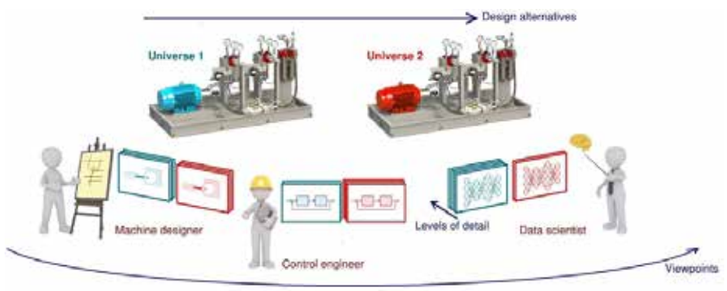

Fig. 1
A suggestive illustration for the multiverse nature of a Digital Meta-Twin -- each universe, viewpoint (and possibly detail level) results in a DT conforming to the corresponding interpretation; in the figure each DT is surrounded by a frame, its colour distinguishing the parallel universes that co-exist within the DMT.

INTELLIGENT NETWORKED MUSIC PERFORMANCE EXPERIENCES (IMPERMANENCE)

Luca Comanducci - Supervisor: Prof. Augusto Sarti - Co-Supervisor: Prof. Fabio Antonacci

We define as a Networked Music Performances (NMPs) what occurs when musicians, displaced in different geographic locations, interact over a network to perform as if they were in the same room. The first NMP-related experiments happened in the 1970s, when only interconnection between local networks was possible. Recent developments of communication technologies and the consequent increase of the speed of digital networks produced the conditions for a dramatic decrease of virtual distances, creating a fertile environment for the development of NMPs. However, high speed networks do not suffice by themselves in creating an environment for NMP that feels engaging to the musicians, since this is a task that requires to tackle several problems depending on different NMP requirements. We may define the two broad classes of problems that need to be considered in NMPs as temporal and spatial factors. Temporal factors refer to all the elements that concur in enabling the synchronization of the musicians, which is often hindered due to the inherent latency present in network transmission, which causes the musician to listen to a delayed version of the audio generated by the co-performer/s.
Spatial factors instead refer to all the issues related to the audiovisual perception of the musicians. More specifically, the visual feedback is usually created through the adoption of screens and projectors. Auditory feedback is instead often provided through loudspeakers and/or headphones. The importance of the audio perception is twofold: the quality of the sound must be of a sufficient level, since musicians must be able to clearly hear the loudness and timbre of the other instruments, in order to consequently modify their playing; the perceived direction of the audio should be coherent with the visual setup, this entails correctly locating the remote musicians in the respective environments, to modify their relative position according to the position of the screen and of the musicians actually located in the other rooms.
While several softwares and techniques have been proposed to separately solve the various issues that comes with creating a realistic NMP, no comprehensive solution has been yet proposed. In this Ph. D. thesis, we propose an across-the-board framework for NMPs that aims at solving at the same time both spatial and temporal factors, denoted Intelligent networked Music PERforMANce ExperieNCEs (IMPERMANENCE).
We base our approach on signal processing techniques, both in order to extract useful information regarding the performance (e.g. tempo) or spatial characteristics of the environment (e.g. position of the performers) and to process the sound emitted by them (e.g. spatial synthesis of the recorded sound generated by a performer). Signal Processing techniques are characterized by a strong set of mathematical and physical constraints that may lead to conditions undesirable in a NMP scenario. More specifically, the limits posed by the Nyquist frequency in space-time audio processing pose some constraints related to the number of sensors needed for proper sampling. The amount of sensors needed could not be deployable in some NMP scenarios. We solve this issue through the application of deep learning-based techniques that enable us to devise functions, without analytically deriving them, that are able to overcome the limits of signal processing by performing in adverse scenarios where the sensors used for sound acquisition do not follow the proper sampling rules.
We first analyze what are the main requirements that need to be taken into account in order to create a satisfying NMP experience. For this purpose we first create a research framework, denoted neTworkEd Music PErfoRmANCe rEsearch (TEMPERANCE) in order to organize experiments with real musicians and analyze the obtained results. Informed by these findings, we accordingly develop the IMPERMANENCE framework.
In the IMPERMANENCE framework, the solution of the time-related issues is tackled through the adoption of adaptive metronomes, that consist of metronomes (i.e. devices that produce an auditory tick at a predefined tempo) that can vary their tempo based on a beat tracker (i.e. a technique for the extraction of the tempo from audio recordings) applied to the sound emitted by the musicians. Since results obtained through the TEMPERANCE framework demonstrate that the needs of the visual perception can be satisfied through simple screen configurations, we decide to concentrate on the auditory perception. Specifically, we aim at reproducing the audio of the musicians so that their perceived location (i.e. directionality) is coherent with the visual feedback for all musicians connected via network. In order to do this, the first step is that of correctly localizing the position of the sound emitted by the instrument of the musician, needed in order to properly render its directionality. We propose two different localization techniques, which vary with respect to the needed setup and computational power. Informed by the location of the musicians, we can then synthesize the soundfield emitted through a technique based on irregular loudspeaker arrays.
Finally we propose a technique for the compression of the audio information that needs to be sent through the network. In particular, this procedure could help in diminishing the impact due to the latency, since it would imply a smaller number of packets that need to be sent through the network. We propose a technique that is able to reconstruct the audio extracted from intermediate layers of pre-trained Convolutional Neural Networks (CNNs).
Figure 1 shows a schematic representation of an application of the IMPERMANENCE framework to a scenario where two musicians are playing in two different rooms connected via network.

While research in NMP has been conducted for decades, is only in the most recent years that the possibility of NMP softwares widely-diffused among a general population of musicians has become a reality. However, no software/research trend considers at the same time all the various aspects of a NMP, such as temporal and spatial factors. We believe that this Ph. D. thesis and the IMPERMANENCE framework can be considered as a first step into this direction, that is, the creation of a unified platform for remote performances, where the physical separation between musicians can be overcome by the advancements of science and technology.

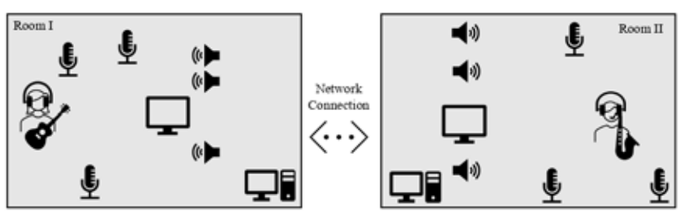

Fig. 1
Schematic representation of an implementation of the IMPERMANENCE framework in a scenario where two musicians are performing in two different rooms connected via network.

ON THE ROLE OF RECONFIGURABLE SYSTEMS IN DOMAIN SPECIFIC COMPUTING

Davide Conficconi - Supervisor: Prof. Marco Domenico Santambrogio

Computer architectures field faces technological and architectural obstacles that limit the general-purpose processor scaling in the delivered performance at a reasonable energy cost. Therefore, com- puter architects have to follow novel paths to harvest more energy-efficient computations from the currently available technology, for instance, by employing *domain specialized* solution for a given scenario. The *domain specialization* path builds on a comprehensive environment where hardware and software are both specialized towards a particular application domain rather than being general purpose. Domain-Specific Architectures (DSAs) generally are the prominent exponent for hardware-centric domain specialization. DSAs leverage an abstraction layer such as an ISA and employs the easiest yet advanced computer architecture techniques to build a fixed datapath with the simplest data type and size. Generally, DSAs are thought to be efficiently implemented as Application-Specific Integrated Circuits (ASICs) or part of System on Chip (SoC). However, developing custom sili- con devices is a time-consuming and costly process that is not always compatible with the time-to- market and fast evolution of the applications, which may require additional datapath customization. Thus, adaptable computing platforms represent the most viable alternative for these scenarios. Field- Programmable Gate Arrays (FPGAs) are the candidate platforms for their *on-field* reconfigurable heterogeneous fabric. On top of the reconfigurability, FPGAs can implement large spatial computing designs and are publicly available on cloud computing platforms.

Domain-Specific Reconfigurable Architectures

FPGAs (and all reconfigurable systems) deserve a deeper analysis of their role in the domain spe- cialization path despite being the commercial platform closest to the ideal adaptable computing paradigm. Indeed, they can implement domain-specialized architectures that can be updated after field deployment, delivering variable datapaths which are adaptable almost an infinite number of times. Here, we call them *Domain-Specific Reconfigurable Architectures (DSRAs).*

Employing Reconfigurable Computing (RC) systems, such as FPGAs, opens a wide variety of *architectural organizations* (different from traditional CPUs with their fixed datapath) that this thesis classifies on two orthogonal characteristics, namely level of software programmability and datapath configurability. The most traditional is the DSA based on a "fixed" datapath with a dedicated ISA that communicates with instructions and data memories. Then, streaming architectures have fixed datapaths for each class of problems, generally devised from a high-level tool that automates the whole process.

This thesis defines and analyzes specialized computer architecture organizations based on recon- figurable platforms called DSRAs and addresses three main topics for each specific domain: design methodologies, automation, and usability. The first one (i.e., the design methodologies) is crucial for designing highly energy-efficient architecture; while automation is essential for fast iterative ap- proaches to newer solutions development and reproducibility of achieved results; the last one (i.e., usability) comprehends software programmability in a complete view that spans from hardware- software interfacing to ways of programming the architecture.

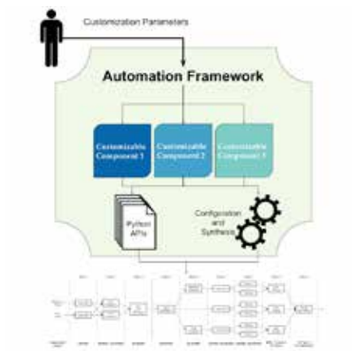

Fig. 1
The overall streaming-based framework comprehensive of automation, software abstrac- tion layers, streaming customizable dataflow architecture.

Open Source Design Automation Framework For Streaming Architectures

Image Registration (IRG) is an essential pre-processing step of several image processing pipelines. However, it is often neglected for its context-specific nature that would require a different architec- ture for different contexts. Therefore, this thesis presents a comprehensive framework based on the streaming architectural pattern with a dataflow MapReduce approach, shown in Figure 1. To com- plete the DSRAs, a design automation toolchain lowers the adaptability effort of the architecture to unexpected contexts or new devices, and a software abstraction layer hides the low-level hardware interfacing mechanisms to expose simpler software APIs. All these components achieve extremely optimized IRG procedures at a lower energy profile.

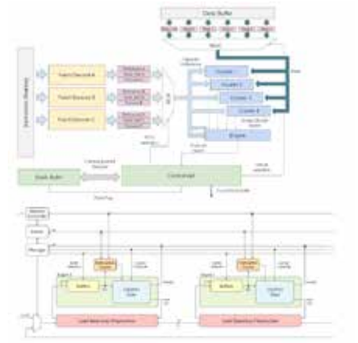

Fig.2
A view of the two different RE-based DSAs.

Different Computational Model of a Loopback-based DSA for a Single Domain

Particular domains may present more than a single *computational pattern* that fits the design process of a DSRA and different applications. For instance, the Regular Expressions (REs) domain present intrinsically sequential computations that can leverage both a *depth-first* or a *breadth-first* execution model. Within this context, this thesis presents two different architectures, shown in Figure 2, that explore these different computational patterns and their respective programming abstraction. They exploit the idea of using REs as programming language of a DSA and share the automation method- ology built out of the IRG domain. The two DSA achieves impressive performance and energy efficiency results, although showing that their improvements are application sensitive.

DEEP AND WIDE TINY MACHINE LEARNING

Simone Disabato - Supervisor: Prof. Manuel Roveri

Machine Learning (ML) and Deep Learning (DL) techniques have widely spread across the most diverse areas in the last decade, achieving state of the art in several fields. Convolutional Neural Networks (CNNs) models such as the ResNet or Inception classify input images. Extensions of these architectures, such as the Yolo or EfficientDet, also identify the position of detected objects within images. Recurrent DL architectures achieve the highest classification capabilities in video classification and speech recognition, translation, and language modeling.
Other DL techniques have also been successfully applied to different fields, such as anomaly detection, recommender systems, reinforcement learning, autonomous driving, and drones navigation.

A few characteristics can be found to be commonly shared among all of these DL techniques. First of all, a high number of parameters. For example, the ResNet CNN has 11 to 60 million parameters, the Inception 24 to 43 million, whereas the language modeling techniques, e.g., BERT, have hundreds of millions or even billions of parameters. By considering a 32-bit (floating-point) data type, the memory required by these parameters easily scales from dozens of megabytes to several gigabytes. Furthermore, performing the training of these models can require days, weeks, or even more on high-performance computers, often equipped with clusters of GPUs.
Finally, performing inference of such DL models, i.e., processing one —unseen— input by a trained model (e.g., the classification of a new image or the sentence identification in an audio sample), is significantly simpler than the training. However, it still requires millions or billions of floating-point operations per single input. For instance, the Inception and the ResNet need 5 to 11 billion multiplications to classify a single input image.

Among all the above-mentioned ML and DL techniques' possible applications, a challenging and breakthrough technology with enormous room for improvement is the so-called "intelligence for pervasive units" such as IoT units or embedded systems. Such devices are nowadays part of our everyday life in a wide range of application scenarios (e.g., smart cities, automotive, or medical devices) and ask to move the processing (and in particular the intelligent processing) as close as possible to where data are generated. ML and DL solutions processing these data directly on pervasive devices are crucial to support real-time applications, prolong the system lifetime, and increase the Quality-of-Service.
The downside is that IoT units and embedded systems have strict memory, computation, and power consumption constraints. The order of magnitude is significantly lower than that required by DL solutions, being Kilobytes to (a few) Megabytes in memory and milli-Watts in power consumption. Such constraints are even harsher and more severe on Micro-Controllers Units (MCUs). Their memory is of a few Kilo-Bytes, with expected power consumptions of micro- to milli-Watts. A few examples follow. The high-performance STM32H743ZI MCU, equipped with a 480-MHz processor, has 1024 KB RAM. The majority of other ST MCUs have 96 to 512 KB of RAM, whereas those of Texas Instrument (e.g., the TMS320F280025C or the TM4C1294NCPDT) 128 or 256 KB.

The goal of this thesis —and of any work addressing this problem— is to design intelligent mechanisms based on ML or DL techniques whose requirements in terms of memory footprint, computational load, and energy are compatible with the technological constraints on memory, computation, and energy introduced by the IoT units and the embedded systems they are designed for. The related literature is highly fragmented and with very few solutions encompassing all the aspects of this problem. However, there are three major research directions.

At first, several works design dedicated hardware solutions for ML and DL models. Such solutions provide significant gains in power consumption and performances w.r.t. general hardware at the expense of a complex design phase and reduced flexibility.

A second major research direction is approximated solutions, with several different approaches to reduce the complexity of deep learning models. In almost all those solutions, the IoT or microcontroller units are not a target of the introduced approximations. Task dropping (e.g., pruning DL models' layers or parameters) and quantization techniques on parameters can (drastically) reduce the memory footprint of ML and DL models at the expense of a drop in the metric the algorithm is evaluated on. Similarly, the introduction of early-exit paths within such algorithms allows to reduce their mean computational complexity, with an impact strictly dependent on how frequently such early-exit paths are taken (thus skipping the remaining computation).

Finally, the third research direction is distributed computation, with solutions derived from the offloading solutions aiming at finding the optimal distribution of the processing pipeline of DL models across a set of heterogeneous IoT units, MCUs, and, if any, the Cloud.

Recently, Tiny Machine Learning (TML) emerged as a novel and promising further research direction aiming at designing ML and DL solutions able to be executed on IoT units or even on MCUs (namely, tiny devices), i.e., with a memory footprint in the order of kilobytes and power consumption in the order of milli- to micro-Watts. To fill the gap between the memory, computational, and energy demands of ML and DL models with the corresponding requirements of tiny devices, all the techniques coming from the literature of approximated DL are brought into play.

Furthermore, the TML solutions mainly enable the inference of DL or ML algorithms. There are indeed very few works in the field of TML proposing on-device learning, i.e., the training of ML and DL solutions directly on tiny devices.

The ability to learn TML models directly on tiny devices is crucial to exploit fresh information coming from the field over time, and to deal with concept drift, i.e., variations in the statistical behavior of the data generating process, a quite common situation in real-world applications (e.g., due to seasonality or periodicity effects, faults affecting sensors or actuators, changes in the user's behavior, or aging consequences). Failing to adapt TML models to concept drift results in a (possibly dramatic) decrease of the TML accuracy over time.

In this challenging scenario, this thesis proposes a methodology to design and deploy Deep and Wide Tiny Machine Learning (TML) solutions that can take into account the constraints on memory, computation, and energy of tiny devices (either IoT units, MCUs, or Embedded Systems). More in detail, the methodology addresses the problem from two different perspectives.

The first perspective focuses on how to support the inference of Deep Learning Models (DLMs) on Tiny devices, i.e., the design of Deep Tiny Machine Learning solutions by means of approximation. More in detail, the approximation relies on task dropping, precision scaling, and early-exit (Gate-Classifiers) techniques.
In addition, this thesis proposes the first TML solution that can learn on MCUs and adapt in response to concept drift.

The second perspective focuses on the Wide (Deep) Tiny Machine Learning algorithms, i.e., it splits the DLMs into (non-approximated) sub-tasks then distributed among possibly heterogeneous tiny devices. In this way, wider DLMs can be designed at the expense of taking into account the introduced communication issues and delays. The term "Wide" here should not be confused with the so-called Wide and Deep Learning models that combine a wide linear model and a DL one.

Finally, the methodology is tailored to two real application scenarios — detecting bird calls within 10s audios in a remote area and modeling solar activity from magnetograms— showing its feasibility and effectiveness.

EXPERIMENTAL CHARACTERIZATION AND MODELLING OF CURRENT TRANSPORT AND POLARIZATION SWITCHING IN FTJS

Giulio Franchini – Supervisor: Prof. Alessandro Sottocornola Spinelli

In the last decades PCs, smartphones and wearable and interconnected electronic gadgets are gaining more and more momentum. Due to this large success, the semiconductor market experienced an enormous growth, especially in the memory sector, as new devices require large amount of solid-state memory with high performance, both in terms of integration density and throughput. Since both are needed in the aforementioned devices, such noticeable growth has affected both volatile and non-volatile memories. In order to keep up with the non-volatile memories market requests, Flash memory technologies have been the object of an uninterrupted scaling process which has led to increase their storage density and let them become the most successful solution in the non-volatile memory landscape; the smallest feasible feature size, equal to 14 nm, was conceived around the middle of the 2010's decade. After hitting this limit, the scaling approach switched to an equivalent one, consisting in stacking memories one on top of the other exploiting the third dimension. Although this switch determined a general improvement in terms of reliability, the novel architecture of the memories brought along some new issues. In the volatile memory landscape, the DRAM technology as well underwent an intensive scaling process to reduce the cost for stored bit and encountered similar reliability issues. Such issues do not constitute the only obstacle to memory technologies. Another major challenge has to be dealt with: the so-called von Neumann bottleneck. In the last forty years, the performance gap between the central processing unit (CPU) and the working memory (DRAM) has never stopped growing. This gap results in a memory bottleneck that reduces the overall performance of a computing system.

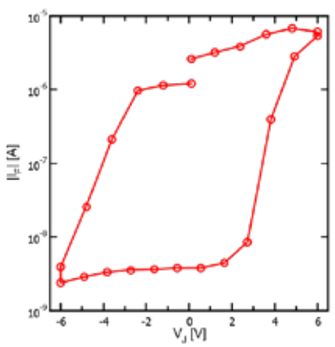

Fig. 1
Result of the hysteresis window measurement for a 3 nm cell.

These issues fueled the research on novel memory technologies, both volatile and non-volatile, different from the ones that dominate the market. Among them: resistive RAM (RRAM), phase change memory (PCM), magnetoresistive random access memory (MRAM), and a wide variety of ferroelectric memories. This thesis focuses on Ferroelectric Tunnel Junction (FTJ), one of the most promising candidates among the novel ferroelectric memory technologies. An experimental characterization on FTJ samples has been carried out exploring the main features of the device: resistive window, retention, I-V characteristics. With respect to the retention measurements a dedicated setup has been developed capable of assessing the retention losses at short times after the programming pulse showing for the first time the great impact of such losses that reduces the ON/OFF ratio of the device. Experimental activities are

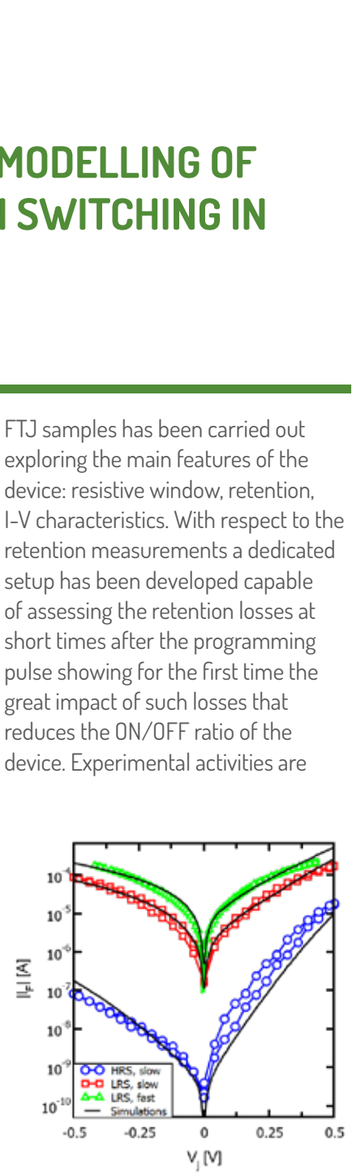

Fig. 2
Result of the retention measurement for a 3 nm cell.

followed by the development of a one-dimensional in-house MATLAB code for the device simulation, which is capable of reproducing the I-V characteristics and explaining some of the physical phenomena involved. It has been chosen to program a new simulator instead of using a commercial one in order to have more control on the simulated physical effects such as the addition of the distributed tunnelling feature. With distributed tunnelling electrons have the chance to tunnel not only from the interface but from any point of the depletion layer through the semiconductor gap toward the metal contact. This feature turned out to be essential to reproduce the experimental measures. With the same simulator the possible limiting impact of the drift and diffusion conduction in the substrate has been demonstrated. Moreover, a physical explanation of the fast retention losses of the ON state of the device is given by analysing the conduction band diagram in the two states, revealing the strong negative electric field present in the ferroelectric layer. All the assumption, the equations, the meshing strategy and the simulation flow are described in full detail in the thesis. Simulation time are in the order of some seconds for a device discretized by a few hundreds of mesh points. In a second time efforts have been directed toward the analysis of the switching process. Based on results found in literature on a simulator for a ferroelectric capacitor a two-dimensional switching simulator has been developed and integrated with mono-dimensional one. Such simulator employs the time dependent Landau-Ginzburg-Devonshire equation obtained by coupling the Landau-Ginzburg-Devonshire equation with the the Landau- Khalatnikov equation. This simulator is actually a quasi2-D simulator since the only bidimensional equation is the time-dependent Landau-Ginzburg-Devonshire equation, where all the electrostatic and current transport equations are solved only in the direction perpendicular to the junction. The device area is then discretized and in order to improve the time needed to run an entire simulation the relation between the electric field the polarization and the applied bias is precomputed and stored in a matrix, such that the electrostatics does not need to be computed many times but the electric field value can be retrieved when needed before computing the polarization at the next time step. The polarization changes during the programming pulses can be visualized in real time with a 2D colormap. Exploiting the precomputation, the ferroelectric switching simulation takes less than two hours to complete. All the equations and simulation flows are described in detail in the thesis. The simulation result was not able to reproduce the resistive window measurement probably due to high number of fitting parameters needed for the two simulators. Better result may be obtained if the number of free parameters is reduced by the direct measurements of some of them or by performing some 2D polarization measures on the area such as PFM measures. Nevertheless, the obtained results are promising, and this simulator can be improved on and undoubtedly be of help in better understanding the physical phenomena involved during ferroelectric switching in FTJs.

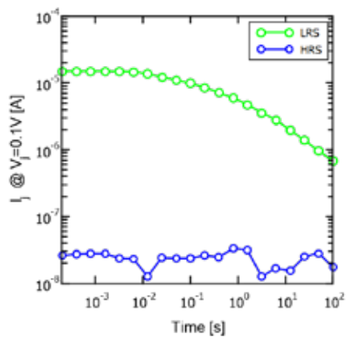

Fig. 3
Fitting of the I-V characteristics. Note that the limitation of Vj to 0.5 V during the voltage sweeps used to monitor IF makes the hysteresis between the forward and reverse curves negligible.

HIGH-ACCURACY CONTROL OF MEMS MICROMIRRORS

Paolo Frigerio - Supervisor: Prof. Giacomo Langfelder

Microscanners based on the MEMS (Micro-Electromechanical System) technology have experienced an ever increasing demand since the fabrication of the first simple prototype in 1980. The demand for such devices comes from a wide range of different fields: from Augmented Reality (AR), Virtual Reality (VR) and picoprojectors systems for the entertainment industry, to systems for imaging, spectroscopy and medical instrumentation, as far as LiDAR applications. The evolution of MEMS microscanners has been driven by the strict requirements of the newest and most exciting applications demanded by the market, among which are the entertainment and autonomous driving fields, which require ever-increasing resolutions, as well as projection and 3D-sensing accuracy to satisfy the demanding customers. In order to satisfy their requirements, the scanner fabrication has evolved from the simple early prototypes based on electrostatic actuation, stimulating the combination of different technologies, which resulted in the realization of MEMS devices with piezoresistive and piezoelectric components. An example of such a device being operated at a large scanning angle is shown in Fig. 1. Such technology offers many advantages, while posing new challenges that are analysed and solved in the thesis. The effort is justified by the aim of maximizing the main figures of merit, being resolution and diffraction-limited accuracy, which mainly concern scanners for the horizontal (fast) axis, linearity and precise positioning, concerning scanners for the vertical (slow) axis, and, in general, system efficiency in terms of cost, footprint and power consumption.

The work developed in this thesis targets the issue of position control of both types of scanners, with the long-term aim of realizing an fully-integrated electronic projection system. The scanners studied in the thesis are based on piezoelectric actuation, and this proved to be a significant challenge for the control system design. Whilst enabling a significant improvement in terms of efficiency and linearity with respect to the competing technologies, the resonant modes of the piezoelectric elements set a constraint on the achievable system performance. System development was targeted at the realization of simple, robust and versatile systems, with the aim of maintaining their footprint and cost as low as possible, in view of future integration. A control algorithm based on two nested controllers has been designed for the slow axis, while for the fast one an alternative topology to a standard MEMS oscillator has been studied and designed. Both system have the capability to manage spurious resonances of the piezoelectric actuators that corrupt the information provided by the output of the position sensor embedded on the device. Fig. 2 shows the effect of the slow axis control algorithm, which is able to linearize the mirror scanning trace.

The developed systems, partly still at the PCB prototype stage, and partly already implemented at the IC level, showed a position control accuracy within a few tens of milli-degrees, within the diffraction-limited pixel size set by the scanner dimensions, and linearity of the vertical scan within 1%, which is maintained for different mechanical designs. The noise standard deviation of the angular position, affecting the trace reproducibility, is within 14 mdeg and 30 mdeg.

Concurrently, a characterization campaign of the piezoelectric technology was performed, targeted at investigating a future substitution of the piezoresistive sensing technology with a fully-piezoelectric scheme. Fig. 3 shows, for example, the typical hysteresis measured for such technology, which sets a new challenge for the control algorithms. Characterization verified the possibility to achieve a significant increase of the sensitivity, with reasonable assumptions that remain valid for the future integrated design, and started the investigation of temperature dependences with the aim of realizing a fully temperature compensated system.

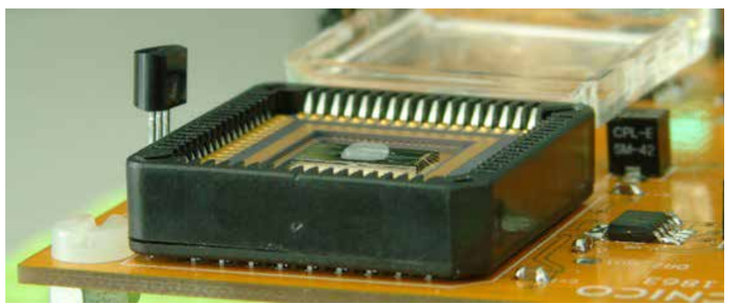

Fig. 1
Photograph of a micromirror under test achieving an optical scanning angle of approximately 90 deg.

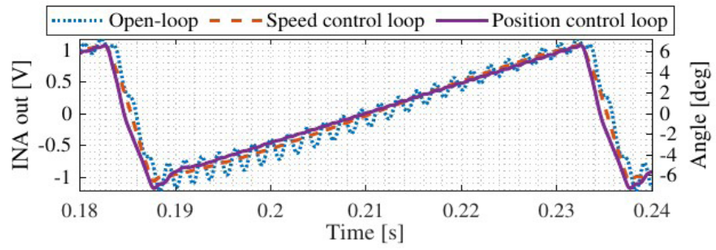

Fig. 2
Speed and position control of the slow axis. Open-loop ringing of the mirror and stiffness non-linearities are suppressed by the control algorithm.

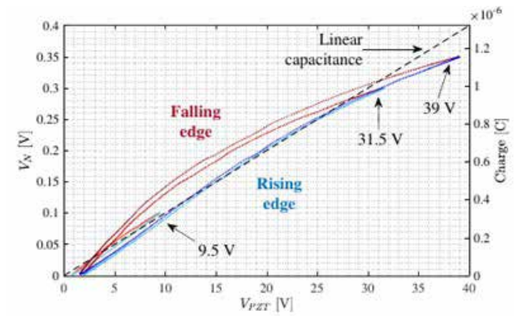

Fig. 3
Measured hysteresis of the piezoelectric actuators.

NAVIGATION-GRADE NEMS GYROSCOPES

Marco Gadola – Supervisor: Prof. Giacomo Langfelder

During the last two decades, MEMS gyroscopes have spread over countless fields of application thanks to their low cost, footprint and power consumption. The continuous performance growth made this technology appealing also for the high-end markets, dominated by expensive and bulky kind of sensors. The goal of the next few years is to fulfill the requirements of inertial navigation applications, thus being able to rely on inertial sensors only to retrieve the orientation of an object. The aim of the work reported in this Ph.D. thesis is thus to analyze, design and characterize high-performance 3-axis miniaturized gyroscopes with piezoresistive readout that can fulfill the requirements of next-generation applications. The sensors presented in this work are fabricated with the M&NEMS technology by CEA-Leti, a standard MEMS process with few additional process phases, in which is possible to embed thin piezoresistive beams with nanometric cross-section for the mass movement readout. This allows to design high-performance sensors while keeping mass production costs. First, a complete design of several single axis rate sensing devices has been carried out, focusing on the electromechanical structures that allow to improve both noise and stability performances in a sensor footprint of less than 2mm2. An example structure is shown in figure 1, where the main electromechanical features are highlighted. In particular, an innovative sensing lever system for pitch/roll devices is described in the thesis, which allows to reach sensitivity levels comparable to those of yaw gyroscopes, thus a stress on the gauge element of 200 MPa at the full-scale range, a fundamental achievement toward a fully planar 3-axis high-performance sensor. The characterization campaign confirmed the expected results, reaching for several yaw gyroscopes noise levels in the order of 100 µdps/√Hz and stability lower than 0.02 °/h up to 1000 s, as shown in the Allan variance measurements of figure 2. Promising results have also been achieved with pitch gyroscopes reaching a noise of 600 µdps/√Hz, despite not having the chance to characterize the designed structure with the best expected performances. Finally, the design of a low-noise integrated circuit is presented, with a focus on the front-end architecture, aiming to reach the same performances of the discrete electronic.

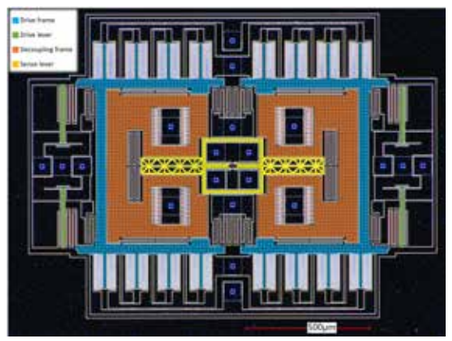

Fig. 1
Microscope picture of a yaw device, with a footprint of less than 2 mm². In false colors the drive mass, the sense mass, the sense and drive stress amplification levers are highlighted.

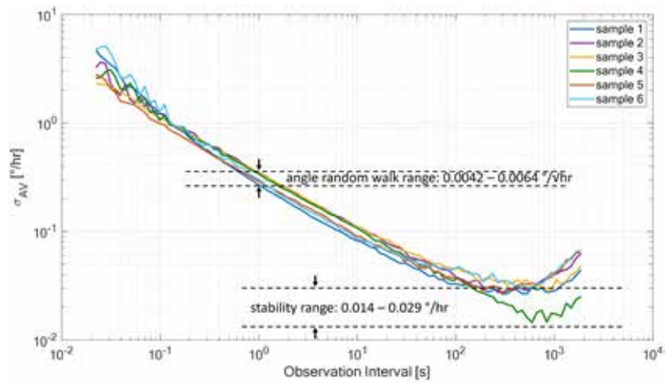

Fig. 2
Figure 2: Allan Variance of 6 yaw gyroscope samples, measured in uncontrolled laboratory environment, reaching noise level of 100 µdps/√Hz and stability lower than 0.02 °/h up to 1000 s.

ABOUT MEMS IMUS STABILITY AND RELIABILITY

Leonardo Gaffuri Pagani – Supervisor: Prof. Giacomo Langfelder

MEMS inertial sensors, after decades of evolution, are now eyeing at navigation applications in many different fields. This would enable a novel breakthrough for the technology, traditionally boasting low costs, small footprints and low power consumption.
Inertial navigation is a hot topic for MEMS inertial sensors. Performances satisfying the requirements of resolution and stability would enable position and trajectory recovery from sensors output and open up the employment of MEMS devices in many new applications.
This work focuses on gyroscopes and accelerometers output stability and overall MEMS devices reliability in harsh environments, typically characterized by shocks, vibrations and temperature variations.
Main goal of this research project is to investigate instability sources in MEMS inertial sensors and, possibly, solutions for their mitigation. The work methodology is based on finite elements, analytical and behavioral models development, and their experimental crosscheck via innovative test structures, inertial sensors and systems design.

Specifically, the first part focuses on the analysis of gyroscopes medium-term instability due to drive-sense relative phase drifts caused by temperature variations. Both a low-noise characterization setup and an accurate model are developed, also taking into account the contributions due to parasitic couplings. This enables phenomena characterization with a relative phase resolution in the order of 3 μrad and experimental data prediction with a 12 μradrms/K residual deviation. Moreover, a novel compensation technique to improve zero-rate-output drifts in temperature is devised, showing an improvement of a factor 100 circa.

Finally, a newly developed system approach prevents this chippings generation from happening in a novel time-switched frequency-modulated accelerometer device. Adding an extra closed-loop to the system, it is possible to provide this architecture (already showing good noise and stability performances) with high resilience to shocks and vibrations up to the audio frequency range limit and with excellent recovery times. In the experimental tests
the presented system demonstrates a factor 20 improvement with respect to the previous embodiment.

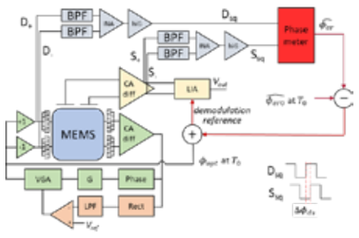

Fig. 1
Phase drift compensation system architecture. The information on the relative phase drifts between reference and quadrature is used to correct the demodulation reference phase.

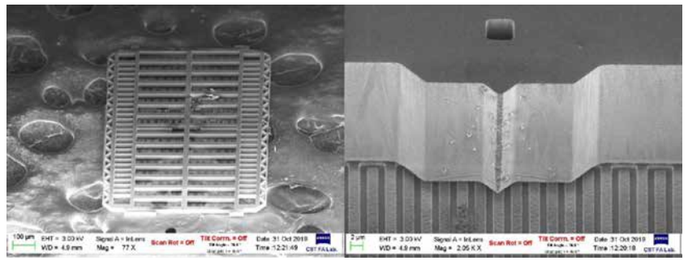

Fig. 2
SEM analysis of a tested structure, showing rotor (left) and stopper (right).

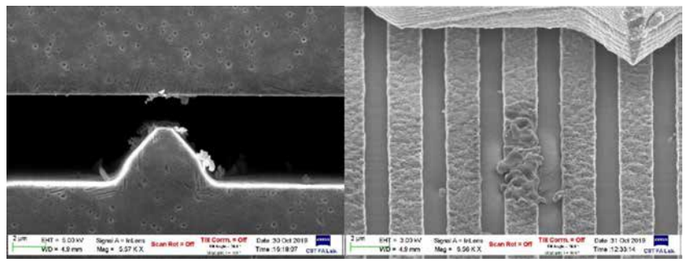

Fig. 3
SEM analysis of a tested structure, top view of chippings deposited near the impact area (left) and a burn mark caused by a short circuit between rotor and a leakage path (right).

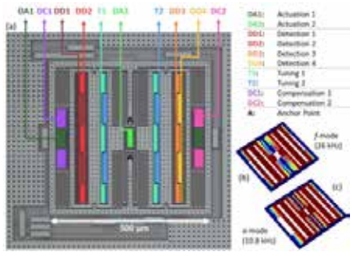

Fig. 4
SEM image of the designed in-plane device highlighting its main features.

FEM simulation results of the first two eigenfrequencies are also reported (bottom right).

Due to the identification of some limitations, further device redesigns have been developed with CEA LETI M&NEMS fabrication process (employing piezo-resistive sensing) and with the newly presented STMicroelectronics Thelma Double process.

VEHICLE DYNAMICS CONTROL FOR INDEPENDENT-ALL-WHEEL-DRIVE FULL ELECTRIC VEHICLES

Alex Gimondi - Supervisor: Prof. Sergio M. Savaresi

Nowadays, two major trends are driving the automotive industry. On one hand, self-driving cars promise to reduce road congestion, increase safety and reduce pollution. On the other hand, electric vehicles are being massively introduced in the market due to their ecological potentialities. Besides environmental aspects, electric motors have many advantages, e.g., precise torque modulation and fast response. An additional factor that gives electric motors the edge over Internal Combustion Engines (ICEs) is the small volume occupied; it permits to modify standard driveline architectures drastically. These peculiarities pave the way for reconsidering classical vehicle dynamics control (ABS, TC, ESC) and improving autonomous driving performance. Different driveline architectures are possible. From a vehicle dynamics point of view, the configuration with independently controlled wheels is the most interesting one. Indeed, the possibility to regulate each wheel torque permits to leverage differential forces to influence the lateral vehicle dynamics. This can be done with four electric motors, one per wheel. The motors can be placed either on-board or in-wheel. Despite the recent steps regarding in-wheel motors, the limited torque available due to space constraints poses a substantial drawback. On the other hand, on-board motors do not suffer from torque limitations but are characterized by the presence of a transmission that complicates wheel dynamics control. Another interesting driveline architecture features two electric motors, one controlling the front axle, one the rear. Loosing the possibility to generate differential force, the ability to influence lateral dynamics is weakened. Anyway, it is worth studying such architecture since it is utilized, for example, by the iconic electric (and most sold) car Tesla and may offer an alternative way to control the vehicle in case of actuators failure. This dissertation discusses different control strategies to exploit the advantages introduced by vehicle electrification; in particular, we present a control stack for autonomous FEVs with 4 electric motors (1 per wheel)(Figure 1). Following a bottom-up approach, we start from the lowest level, i.e. wheel dynamics, for which we have designed a nonlinear longitudinal slip regulator. The novelty of the proposed approach consists in synthesizing the controller using a grey-box model to include the transmission dynamics. Afterwards, we have considered lateral vehicle dynamics, firstly to improve safety and then to increase fun-to-drive. We have proposed an ESC scheme able to maintain the vehicle stable in critical situations regardless of the road conditions; it controls a convex combination of yaw rate and sideslip. In this case, we have also considered a substantially different powertrain architecture: two electric motors (1 per axle), for which a specific way to split the torques has been discussed. We leverage the expertise gained to design a TV control that continuously enhances the vehicle dynamics. Then, autonomous driving scenario is tackled with focus on path tracking. We propose a multi-layer controller that easily integrates the lower level control structures managing multiple actuators (steering wheel, electric motors). The controller, designed exploiting LPV/Hinf technique, includes tyre nonlinearities. Finally, the highest level, i.e. planning, is addressed; we focus on the longitudinal dynamics, in which regenerative energy capabilities can be exploited. Specifically, we have designed a progressive iterative dynamic programming algorithm that includes both comfort and consumption aspects.

DATA-INFORMED MODELS FOR THE COUPLED DISPERSAL OF MICROPLASTICS AND PLASTIC-RELATED POLLUTANTS APPLIED TO THE MEDITERRANEAN SEA

Federica Guerrini - Supervisors: Prof. Renato Casagrandi, Prof. Lorenzo Mari

Microplastic pollution is a global emerging environmental threat, particularly for marine ecosystems. Mounting evidence highlights that floating microplastics in the sea are not just passively transported by wind and surface currents, but are often contaminated with organic pollutants and colonized by marine organisms, that are processes occurring at microplastic scale. These interactions with abiotic and biotic components of the seascape affect the dynamics of microplastics at sea and exacerbate their toxicity to marine biota; on the other hand, these phenomena are inherently hard to observe on field. There is an urgent need for an effective modelling of the advection-diffusion processes jointly involving microplastics and the Plastic-Related Organic Pollutants (PROPs) they carry, their spatiotemporal patterns and ecological impacts, and potential benefits of policies preventing microplastics leakage to the sea.

The thesis proposes and analyzes novel models based on oceanographic reanalyses to simulate realistic patterns of release and transport of plastics in the marine environment, as well as their consequent interactions with the seascape. Crucial to the realism of our models is identifying drivers of plastic pollution and exploiting the wide variety of data linked with them, ranging from national censuses to satellite data of surface water runoff and GPS ship tracking. Here we present the conceptual design, methodological settings, and modelling results of a novel 2D Lagrangian-Eulerian modelling framework that simultaneously describes (i) the Lagrangian dispersal of microplastic on the sea surface, (ii) the Eulerian advection-diffusion of selected organic contaminants, and (iii) the gradient-driven chemical exchanges between microplastic particles and chemical pollutants in the marine environment in a simple, yet comprehensive way.

While providing further understanding of the distribution of microplastics in the Mediterranean, the results of our method applied to a multi-year simulation contribute to a first basin-wide assessment of the role of microplastics as a vehicle of other pollutants of concern in the marine environment. Furthermore, the generality and adaptability of the Lagrangian-Eulerian modelling approach proposed here make it possible to successfully deal with similar problems arising when moving microplastic particles modify the properties of the surrounding marine environment. As a notable example, we show how our modeling framework can be applied to the study of microplastic-mediated carbon export in the Mediterranean Sea.

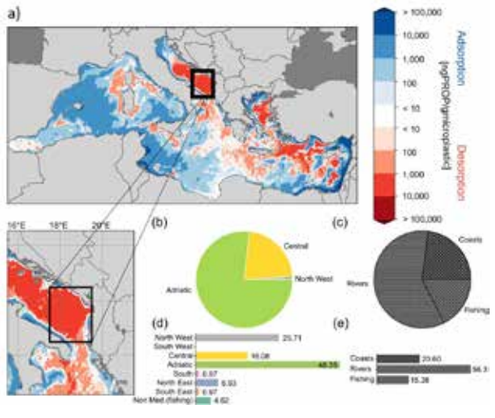

Fig. 1
Net particle-mediated PROP exchanges, averaged over 2015-2016, as resulting from the coupled simulations. a) Mediterranean-wide mapping; b) Country of origin of desorbing particles; c) Source type of desorbing particles; d) Average mass of PROP desorbed per unit of plastic mass ngPROP/gmicroplastic; e) as in d), for particles with different source types.
From Guerrini et al., (2022). Environ. Res. Lett., 10.1088/1748-9326/ac4fd9

Although far from providing a complete picture of the complex phenomenon of plastic-related pollution at sea, the framework proposed here is intended as a flexible tool to help advance knowledge towards a comprehensive description of the multifaceted threat of marine plastic pollution and an informed support to targeted mitigation policies.

SOFTWARE ENGINEERING METHODS FOR DISTRIBUTED AND MODEL-DRIVEN DEVELOPMENT

Sergio Luis Herrera González - Supervisor: Prof. Piero Fraternali

Introduction
Persuasive applications focus on engagement and user experience to attract users and induce behavioural changes. The development of gamified applications is a complex and challenging task. It requires a multi-disciplinary group of experts to generate prototypes, test them, and evaluate their effects. This cycle of activities should be carried on continuously during the project's life-cycle. Model-driven development could bring the needed agility to the development process of these applications. However, current MDD methods need to be adapted to cover the complex requirements of gamification, many of which cannot be modelled, such as usability, presentation, etc. Such non-modelled requirements are typically covered by integrating hand-written code into the code generated by the MDD tool, causing conflicts at each iteration of the development cycle.

This thesis proposes methods, architectures, and components to integrate the non-modelled requirements that characterize gamified applications into the model-driven development life-cycle.

Model-Driven Development of Persuasive Applications.
Application gamification aims at engaging users by fostering their involvement and by enhancing their motivations to perform well in the accomplishment of a task.

Our approach to develop gamified application aims at implementing an architecture which minimises the interference of the gamification rules with the application's business logic and limit the integration effort. The proposed architecture consists of a gamification engine which implements the registration of gamified actions and keep track of the status. A Gamification Database stores the entities that allow the GE. The Notification Engine implements the logic for delivering the notifications to the users. The Deadline Manager monitors the expiry of deadlines. And the Gamified Application which integrates the gamified view components into the business views.

Across the different domains in which gamification techniques can be applied, it is possible to recognize recurrent functions. In the spirit of MDE, such features can be captured as patterns, expressed by models that can be transformed into actual application components through model-to-text transformations. In example, The Goal Selection and Progress pattern provides concise feedback about the user progress towards her goals, shows the status of the goals already established, and lets the user set her own self-assigned goals. The pattern comprises a Goals View, in which a List enables the user to select the goal to visualize. Several patterns like this were identified and represented in IFML models for the front-end, and UML sequence diagrams for the back-end. The IFML models were used to generate the application.

The approach was evaluated in two EU project focused on consumption saving of water and energy. The main lesson learned during the evaluation were: The use of a formal Domain Model helped align the terminology and concepts across stakeholders. The availability of a catalogue of front-end patterns helped reduce the space of the interface designs and enabled the rapid convergence to a solution. The aspects that required the most effort was features that could not be modelled, such as usability, and aesthetics.

Model and code co-evolution in MDD
The proposed approach to the integration of handwritten and generated code is rooted in two main ideas. The first one is letting the code generator and the human developers update the same code concurrently. The second is identifying and resolving the possible inconsistencies that these concurrent updates may produce. The method is inspired by the way in which VCSs handle concurrency among human developers.

The code generator is invoked after a model change and always reproduces the whole code of the application, thus potentially overriding all the previous manual modifications of the generated code. By keeping a detailed record of the contributions of the code generator the changes that it introduces at each invocation can be computed. The isolation of such changes enables the analysis to focus only on the last round of manual modification and code re-generation and makes the problem tractable.

A virtual developer module was implemented, in charge of performing the describes workflow, it oversees integrating the latest changes of the generated code into the code repository by calculating the delta between the latest revision of the code and the last generation, also between the latest revision and the latest developer intervention.
The approach was evaluated developing 2 applications in 2 different MDD platforms with the proposed approach and the typical forward engineering approach. The results that the integration effort was reduce in about 60 to 70% in all cases.

Integrating code of distributed development projects
The key idea is to exploit the conflict resolutions implemented by human developers in the past to create rules applicable to future similar conflicts. When the first conflict is resolved manually, the conflicting chunk and the manual resolution are processed to derive a Conflict Resolution Rule (CRR). Then the first Conflict Cluster (CC) is created, and the rule is associated with it. A CC contains a set of conflicts with similar structure that can be solved in the same way. When a new conflict arrives, it is compared to the existing clusters. If its similarity to the conflicts of one or more existing clusters is above a threshold, then the CRR of the cluster with the highest similarity is used to solve it, the conflict is added to the cluster and CRR generation is executed to update the rule associated with the cluster. Otherwise, the user is prompted to provide a resolution and a new (CC, CRR) pair is created.

The approach was implemented as a plugin of GIT, the tool consisted of four components, the Submission Manager extends GIT Rerere and orchestrates the processing of a merge or commit command. The Cluster Manager implements the online hierarchical clustering algorithm that assigns an input conflict to a cluster. The CRR Generator exploits the method for the generation of a search and replacement expressions proposed by Bartoli ed al., and is triggered every time a conflict is added to a cluster. Finally, the Conflict Resolver is called when a new conflict occurs. It searches for the cluster with the highest similarity index to the conflict, extracts the CRR, applies it, and returns the result as the possible solution.

Almost Rerere intervenes in two phases: when git merge is executed, Git will try to execute an automatic merge between the local and the remote repositories. If conflicts arise, the automatic merge fails, and the conflicts are identified. Almost Rerere is invoked, resolves the conflicts by applying the CRRs generated in previous integration cycles, and proposes a solution to the developer. The git commit can be executed after all the existing conflicts have been resolved. In this case Almost Rerere collects the conflict resolutions that the user pushed to the repository, calls the Cluster Manager to add the new conflicts, and invokes the CRR generator to update the CRR of the modified clusters.

The approach was evaluated over 21 open-source projects hosted on GitHub. The small conflicts, up to 6 Lines Of Code, were extracted and a total of 12981 was obtained. The tool was able to generate a solution for 43.63% of the conflicts, and it was able to synthetize the same resolution as the developer in 44 % of the cases. It was observed that the length of the conflict affects the algorithm since 53% of the single line conflicts were resolve identical to the developer resolution, but for multiline conflict it was 34 %. Additionally, it was found that 52% of the resolved conflicts had a similarity above 90%.

TIME-BASED CONTROLLER FOR DCDC BOOST CONVERTER WITH RIGHT-HALF-PLANE ZERO MITIGATION

Mauro Leoncini - Supervisor: Prof. Salvatore Levantino

In the last two decades, portable devices changed from being luxury products to become something necessary to the population. Faster, high performance processors and the increase of the storage data capabilities for the same area occupation have been a game-changer that allowed end-users to tackle their activities on the go. The increase in the offer of portable electronic devices has forced the industries to constantly search for innovation to keep themselves ahead of the market. For this trend to be maintained, the giants of the electronics industries are always trying to push portable devices to become lighter in weight, faster in processing speed, and smaller in size. All these characteristics have a negative impact on the battery lifetime of the product. In this scenario, a key role is played by the power management which must guarantee the power distribution in the whole chip while maintaining large efficiencies and low quiescent currents. Integrated power converters are used to generate, starting from the same device battery, precise references which are distributed to the different sections of the circuit. These references must remain stable regardless of the load operating condition. To obtain that, a large bandwidth controller circuit is mandatory. Switching regulators can meet the stringent specifications of battery-powered electronic products. Research are devoted to minimizing the occupation of the passive LC filter and the integrated power stage. This is accomplished by either increasing the converter switching frequency or adopting novel structures to reduce voltage stresses on the power stage components allowing the designer to reduce their size. Following this trend, time-based controls in integrated wide-band DC/DC buck converters for smart power applications have been proven to reduce area occupation and power consumption of the controller with respect to the standard voltage-mode signal processing since they operate with digital signals. Considering the advantages of the time-based control it may look strange that almost all the scientific publications on this topic are related to converters of the buck-type. One key reason is that most of the portable applications that require integrated converters compliant with a scaled CMOS process are mainly of the buck type. Despite that, all the mobile handheld and wearable devices that are equipped with a LED display, require a boost and/or an inverting buck-boost type converter with stringent specifications on both area and power consumption. The time-based approach applied to a boost converter only gives limited advantages since, the latter, suffers from an inherently right half-plane zero in the control loop that limits the maximum achievable bandwidth. To fully exploit all the advantages coming from the time-based implementation, it is advisable to maximize the converter bandwidth. The aim of this PhD project is to develop a novel time-based controller architecture for boost DC/DC converter that can increase the maximum achievable bandwidth, eliminating the right-half-plane (RHP) zero in the control loop. The controller combines the converter output voltage with the inductor current to eliminate the RHP zero and improve the dynamic performance without any extra power switches or external capacitors. The steady-state regulation error generated by this technique is mitigated by injecting a scaled version of the load current into the loop. A prototype converter is designed for the powering of an AMOLED display where the input voltage is provided by a standard Li-ion battery. The technology node is a STMicroelectronics BCD technology with 180nm CMOS. The minimum input voltage is 2.3V, whereas the maximum input voltage is 4.5V. The output voltage is selectable in a range from 4.6V to 5.4V with 100mV steps controlled by adjusting the output voltage partition. The target loads current ranges from a zero value in the idle/no-load condition to a maximum of 800mA. The external filter inductor (LF) is selected to 2.2μH to optimize the trade-off between inductor current ripples and package volume. The output filter capacitance (CF) has a value of 44μF to minimize the output voltage ripples. The switching frequency in the continuous-conduction mode (CCM) is chosen to be 1.5MHz. The block diagram of the prototype boost converter is shown in Fig.1. The time-based PI controller is implemented with a voltage-controlled oscillator (VCO) realized with a transconductor (Gm,i) driving two current-controlled oscillators (CCO), and a voltage-controlled delay-line (VCDL) made by a transconductor (Gm,p) driving two current-controlled delay lines (CCDL). An asynchronous-finite-state-machine (AFSM) generates the PWM signal to drive the power MOS. The RHP zero elimination is obtained sensing the inductor, scaling it (×, ×) and injecting it in the loop. In a similar way, the static error generated by this technique is compensated reading the load current, passing it through a variable gain block whose gain depends on the circuit operation condition, and injecting it in the loop with the same scaling factors (×, ×) but different polarities, compared to the inductor current. The efficiency at light-loads is increased thanks to a pulse-frequency modulation (PFM) operating mode which drives the power stage to deliver to the output capacitor a fixed charge packets every time the controller is triggered. In this operation, when the converter operates at very light load conditions, the charge packets are sent only occasionally to maintain the output regulation, thus strongly reducing the impact of the switching related power consumption. with a steady state error correction and seamless PFM-to-CCM transition. The measured converter peak efficiency is 96% for an input voltage of 4.5V and above 90% at light-loads up to 10mA. The proposed time-based controller shows a quiescent current of 300μA when operating in CCM and an area occupation of 0.27mm2, with 0.12mm2, 0.04mm2, and 0.11mm2 being the occupation of the compensator, inductor current sensor, and load current sensor, respectively. The prototype converter shows an increase in the converter bandwidth of about a factor 5 with respect to a standard compensation limited by the RHP zero while reducing the worst-case static error from 120mV to 10mV, thanks to the addition of the novel static error correction.

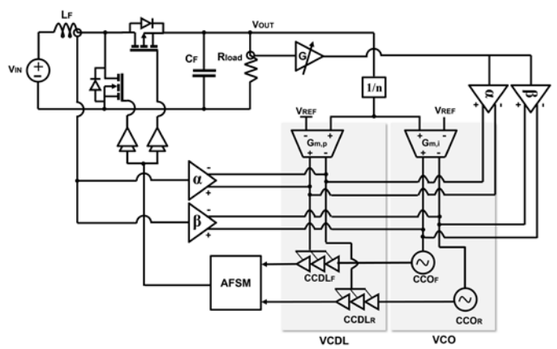

Fig. 1
Block diagram of the proposed time-based boost converter with RHP zero elimination and static error correction.

GENOMIC METADATA INTEGRATION AND DATA PROCESSING METHODS FOR THE ANALYSIS OF CHROMATIN BEHAVIOUR IN DIFFERENT BIOLOGICAL CONDITIONS

Michele Leone - Supervisor: Prof. Marco Masseroli

Data produced by Next-Generation Sequencing (NGS) technologies can be processed significantly faster and at a lower cost. Public gene expression datasets, such as NCBI's GEO or SRA, have grown exponentially over the last decade. These repositories, especially when linked together, offer excellent research opportunities. The integration of data in genomic repositories has been hampered by the heterogeneity of databases. Because there is no standardized metadata, each consortium has imposed its own set of rules, resulting in a faulty conceptual model. Researchers will be unable to conduct appropriate searches on these repositories as a result of this. The GeCo group's proposed standard for genomic metadata, the Genomic Conceptual Model (GCM), describes the semantic heterogeneity of genomic data. It lays the groundwork for data interoperability, which, when combined with the GenoMetric Query Language (GMQL), gives biologists and bioengineers a powerful tool for querying thousands of samples of processed data. Data integration from other sources, such as GEO, is a major focus of that project. This project gathers millions of genomic samples and metadata, which are organized into a few generic fields. As a result, the integration process necessitates human interaction. Unlike the GCM, which uses highly specific metadata (such as "Age," "Tissue," and "Cell Line"), GEO only has a few generic fields, such as "Characters" and "Description," which contain extensive text descriptions of the genomic sample. The weight of the problem has grown as a result of the large number of samples collected in such a database, requiring a great deal of effort to find a solution. Several contributions to this issue focused on small sub-tasks, such as label recognition, emphasizing the need for a system that can address the problems of the current state-of-the-art while also being capable of handling a wide range of tasks. The first section of the thesis presents a novel approach to metadata integration using natural language processing (NLP), demonstrating that the adopted strategy outperforms previous literature in terms of accuracy and generalization. Rather than treating the problem as a classification or named entity recognition problem, state-of-the-art sequence-to-sequence (seq2seq) models have been used to directly map unstructured input into a structured text format. Two of them, the LSTM-based encoder/decoder and the Open AI GPT-2, have been tested and trained. The first experiment was conducted with Cistrome data, a collection of over 44,000 samples labelled with four attributes; the second experiment was conducted with ENCODE data, a large genomic archive where more than 16,000 samples with sixteen different attributes could be downloaded. The two experiments demonstrated the strengths of the proposed approach in comparison to the most common classification methods by demonstrating the effectiveness of sequence-to-sequence models. The results revealed that a sequence-to-sequence model outperforms the baseline classification algorithm in extracting the correct information even when the input text is ambiguous as a human reader might perceive it. We designed an active learning framework in the second part of the thesis to allow the tool to receive feedback from users and improve during its use. Furthermore, we developed a technique for interpreting the model's predictions and implemented it in our tool to assist the user in providing accurate feedback. The Genomic Metadata Integration tool, or GeMI, is the result of our efforts. This tool emerges from the idea to continue and improve the work described above. During the design process, we came up with three different ways to calculate the confidence of the predictions. We chose the most promising based on empirical evidence and decided to use online learning to retrain the model as soon as the user provided feedback. Finally, the GeMI tool has been updated to include the active learning we designed. We implemented three different saliency map methods from the literature: Lime, the Attention-based method, and the Gradient-based method, to provide a clear and meaningful visualization that allows to interpret the prediction of a text generation model. During the design process, all three methods were integrated into the GeMI interface in order to determine which one was the best approach for providing a visualization that allows users to interpret the predictions in our tool. We discovered that the Gradient-based method produces the best results at the end of this process. Because the ultimate goal of this project is to create a metadata integration tool, we decided to compare the methods empirically before selecting the best one. We began developing the GeMI tool using the GPT model described in the first section of the thesis, but we soon discovered some of the model's limitations. The main issue was that the model couldn't be used to generate both datasets' fields. To address this issue, we create a new version of the GPT2 model that takes advantage of the task conditioning property. The model we created can handle multiple datasets, make each field generation independent of the others, and improve inference speed. The thesis concludes with a description of the design and implementation of the Combinatorial and Semantic Analysis of Functional Elements (CombSAFE), a novel computational method capable of identifying combinations of static and dynamic genomic functional elements from repositories such as GEO, as well as how they change across semantically annotated biological conditions. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a standard method for identifying DNA-associated proteins such as transcription factors and histone marks. Although experimental ChIP-seq data are widely available in omics data repositories, repurposing and integrating these data across multiple conditions remains difficult. CombSAFE combines a data-driven approach to grouping conditions of interest and a semantically driven approach to labelling them. This enables the comparison of a large number of genomic profiles of chromatin functional states in various conditions using hidden Markov models, as well as the extraction of their specific variations in the various conditions. Identification of functional states can be influenced not only by histone modification combinations, but also by any other factor whose activity can be mapped to genomic coordinates, such as transcription factors or even static genomic features like CpG islands, partially methylated domains, or transposable elements. This approach has the potential to reveal biologically unexpected similarities between samples. This could simply be due to the complexity of the epigenomics profiles, or it could reveal inconsistencies in the data or their annotation in some cases. This study has two limitations that could be addressed in future research: First, CombSAFE is entirely dependent on the availability of multiple omics data sets about the same biological condition, which is not always easy to come by; second, it is entirely dependent on the quality of the metadata provided in input. Both constraints will benefit greatly from the rapid accumulation of omics data in public repositories, as well as current efforts to standardize the corresponding metadata. The method is implemented in a publicly available Python software pipeline, is robust in managing small and large datasets, and is simple to use. As demonstrated by the example use cases, this method allows for the comparison of a large number of genomic profiles of chromatin states in different conditions, as well as the extraction of their specific variations. Biological findings point to important data-driven discoveries. Several of these have been confirmed in the literature, implying that the others may reveal novel data-driven insights.

FAST AND ROBUST ESTIMATION OF ATMOSPHERIC PHASE SCREENS USING C-BAND SPACEBORNE SAR AND GNSS CROSS-CALIBRATION

Marco Manzoni - Supervisor: Prof. Andrea Virgilio Monti-Guarnieri

One of the first features mentioned when talking about Synthetic Aperture Radar (SAR) is the total independence on weather conditions. While it is true that clouds are loosely affecting SAR images, it is also true that the non-unitary refractive index in the path from the satellite to ground delays the radar signal and affects the phase of the image acquired. This delay varies spatially and temporally, and it is one of the main sources of disturbance in the interpretation of SAR interferograms (the main product of interferometric SAR, or InSAR).
In interferometry, the effect of the atmospheric delay is called Atmospheric Phase Screen (APS). For applications like ground deformation monitoring or Digital Elevation Model generation, the effect of the atmosphere must be considered as a source of noise and therefore should be eliminated or at least mitigated.
The refractive index, however, changes with temperature, pressure, or humidity of the medium, therefore it carries information about the status of the atmosphere at the time of the acquisition. This peculiarity leads in recent years to the boost of a branch of meteorology called InSAR meteorology with the primary objective of producing higher quality weather forecasts by using SAR-derived water-vapor maps as an ingestion product for Numerical Weather Prediction Models (NWPM).
APS estimation from SAR images is not a novel concept: one of the by-products of Permanent Scatterers Interferometry (PSInSAR) is, indeed, the atmospheric delay. Permanent Scatterers (PS), however, are not always present in the scene, especially in rural or forested areas. PS processing also requires many images to work correctly, and the computational burden can be demanding if wide areas are employed.
To be useful for NWPM, ingestion products must be spatially and temporally dense. Moreover, the generated maps must be wide: NWPMs work in domains as big as entire countries and they need APS maps that are several hundred or thousands of km wide.
To satisfy the first requirement, not only PS but also Distributed Scatterers (DS) must be used.
The exploitation of such targets requires coherence: the spatial structure of the scene must be quite stable between two SAR acquisitions. Since the sensibility of the system to ground changes increases with the operational frequency of the SAR, the C-Band at around 5.4 GHz is particularly suited for this purpose: the operational frequency is sufficiently low, allowing moderate coherence levels while also being high enough to avoid coping with ionospheric disturbances present at very low frequencies. Other bands will suffer of severe decorrelation (X-Band) or very strong ionospheric disturbances that must be taken into account (L-Band or P-Band).
The second requirement concerning the wideness of the estimated atmospheric map is dealt by exploiting SAR images gathered with the ESA mission Sentinel-1. The Interferometric Wide (IW) acquisition mode can continuously capture a swath width of 250 km, making it the perfect instrument for this objective.
This study developed a fast and robust method to optimally estimate atmospheric phase screens from a stack of SAR images using both PS and DS. Such delay maps can be used to predict extreme weather events and to provide better accuracy in short time forecasting. To satisfy the requirement on the large size of the derived product and at the same time to keep low the computational effort, it is mandatory to degrade the resolution. At the same time, we need to exploit DS, thus it is mandatory to use all the information (looks) available in the atmospheric resolution cell. The entire procedure is based on the Phase Linking algorithm and exploits ground patches whose size compares with the desired spatial resolution. The method is suited for short revisit time, C-Band SAR mission such as Sentinel-1, where sufficient coherence is present when estimating interferometric phases using large windows.
Just a few images need to be processed with a short total temporal span: this helps reduce the effect of deformations and decorrelations, which helps the unwrapping procedure.
A cross-calibration of the data using the Global Navigation Satellite System (GNSS) is conducted in order to remove sub-centimetric orbital errors that would lead to smooth but significant errors in the final products. InSAR by itself, included the proposed method, is unable to produce absolute Zenith Total Delay (ZTD) maps. The product is, indeed, a differential one in the sense that each image is the difference between the ZTD of a given time instant and another. In order to retrieve the absolute ZTD, a prior must be used. In this research work, we also propose a procedure aimed at making the maps absolute.
The results are presented with different case studies. The first one is in the area of northern Italy where a dense GNSS network is present along with severe decorrelation and strong orography. Statistics of the estimated maps are derived and compared with the state-of-the-art PSInSAR processing. Another case study is the one in central Italy where again the maps are compared with the ones computed by the state-of-the-art processor. The third case study, instead, is a large-scale experiment involving 145.000 square km in South Africa with very few GNSS stations (Figure 1 and Figure 2).

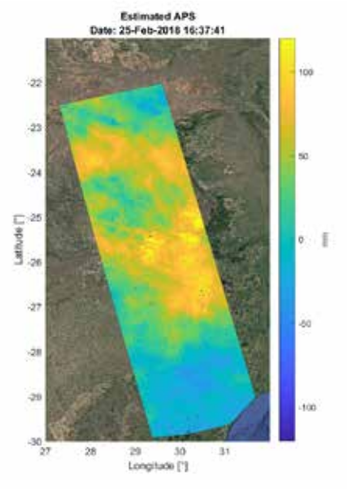

Fig. 1
Estimated atmospheric phase screen in South Africa.

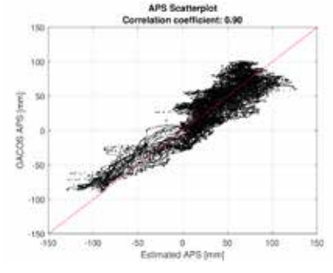

Fig.2
Scatterplot of the estimated APS versus the APS derived by a NWPM.

EMBEDDED AI AND SENSOR FUSION FOR AUTONOMOUS VEHICLES

Simone Mentasti - Supervisors: Prof. Matteo Matteucci, Prof. Federico Cheli

One of the main requirements of an autonomous vehicle is the ability to sense the area around itself and retrieve a uniform representation of the surrounding. A standard setup for a self-driving car consists of multiple sensors (i.e., Lidar, radar, cameras, IMUs, Encoders). The goal of sensor fusion is to collect the heterogeneous information provided by these sensors and merge it to give the control algorithm a rich but concise representation. Furthermore, each sensor offers a specific feature of the surrounding obstacles; cameras are used for classification, Lidar provides accurate position and radar relative speed. Moreover, Lidars and radars are less effected by the changes of illumination between day and night compared to cameras. Adverse weather is also an important aspect to consider. Indeed, heavy rain and fog can degrade the quality of radars and cameras detection. Aggregating all this information and contexting the obstacle's position (i.e., on the street, on the ego-vehicle trajectory, parked on the roadside) requires accurate synchronization between the sensors and high computational power to perform fusion in real-time. This is particularly important in a highly dynamic environment like urban roads, where the rapid changes in the surroundings require constant and fast responses from the vehicle.
The goal of this Ph.D. project is the realization of a sensor fusion pipeline for a development platform equipped with limited computational power and soft synchronization between sensors. In particular, the thesis has been developed inside the Regione Lombardia project TEINVEIN (Tecnologie innovative per i veicoli intelligenti) whose goal was to design and realize a research platform for fully autonomous vehicles in urban scenarios. For this reason, the first steps of this work consisted of developing the sensor setup and processing pipeline to convert a consumer vehicle into an autonomous one, shown in Fig.1.

Fig. 1
Image of the experimental vehicle. It is possible to notice the Lidar and cameras sensors mounted on the roof and the two GPS sensors on the front and the rear of the car

Then, considering the project constraints in terms of sensors resolution and computational power available, I developed different solutions to process every sensor (i.e., Lidar, camera, radar) and retrieve information on the state of the autonomous vehicle and the surroundings. But different sensors provide overlapping information. For example, both Lidar and cameras can detect other cars and pedestrians. Thanks to this, I implemented various techniques to combine those data, using late fusion algorithms and hybrid architectures to retrieve a concise list of obstacles, which can be processed in real-time by the vehicle's planner. However, due to the limited computational power available on the car and the low resolution of the sensors, if compared to a more traditional autonomous vehicle development platform, ad-hoc solutions needed to be implemented. In particular, the fusion process has been performed asynchronously, employing less traditional data representation, like a 2D occupancy grid, generally used in robotics systems, characterized by limited computational power, but not in the autonomous driving field. Each proposed solution has been validated in a real environment, using the developed platform, and in simulation using data acquired from our vehicle in a controlled scenario.
Particular focus has been devoted to the lateral line detection task and the ego-vehicle relative state estimation from vision. For this task, I employed a convolutional neural network to extract a mask of the lateral lines from images acquired by a roof-mounted camera. Then, through a bird's-eye view projection and a successive window-based line following algorithm, I extracted points belonging to the left and right lines. Finally, those points are fitted to retrieve the equations of the road boundaries. This information is fundamental to estimate the autonomous vehicle's lateral offset and its heading, as shown in Fig.2.

Fig. 2
output of the lateral line detection pipeline. In green are represented the detected lateral lines; in orange, the road centerline; in red, the computed ego-vehicle heading and lateral offset. In white, on the right, the bird-eye view projection of the lines.

The values computed from the lines detection algorithm are finally fused with data from GPS and IMUs to provide the planning and control algorithms with a robust and accurate estimate of the vehicle position. This information, combined with the data from the obstacle detection pipeline, constitutes a good representation of the vehicle's surroundings, allowing the car to drive autonomously, even in challenging and dynamic scenarios where data from some sensors might not be available.

MIXED-SIGNAL ELECTRONICS FOR OPTOGENETIC EXPERIMENTS AND CELL MONITORING

Alireza Mesri Gendeshmin - Supervisor: Prof. Giorgio Ferrari

Neurodegenerative diseases occur as the result of progressive loss of structure, function, or even death of neurons. Training4CRM project, funded by the European Union Horizon 2020 Programme, is a highly cross disciplinary project and focuses on bridging the existing gaps within cell-based regenerative medicine for the treatment of neurodegenerative disorders (e.g. Parkinson's disease, Huntington's disease, and epilepsy) by joint training and education of 15 Ph.D. students, in 6 European countries, within and across different scientific disciplines. The goal is to master the design, fabrication, integration and testing of completely new tools and materials within the fields of micro and nanoengineering and biotechnology.

In this Ph.D. thesis, which is a part of the Training4CRM project, a platform based on commercial off-the-shelf components is proposed to perform optical stimulation and electrochemical measurements on optogenetically modified dopaminergic cells.

Figure 1 shows the proposed opto-electrochemical platform. 2 two-layer boards are used to design the required PCBs. A laser diode is used as a light source for optical stimulation of light sensitive neurons with an improvement in the power efficiency of the optical stimulation circuit compared to the commonly used systems based on LED diodes. A PSoC 63 BLE microcontroller is used to implement a compact solution, including optical stimulation as well as electrochemical measurement circuit. To perform electrochemical measurements, a potentiostat with three-electrode structure is used, which enables amperometry and cyclic voltammetry measurements. The electrode chip can be wire bonded to the bottom board and electrodes are connected to PSoC 63 on top board through a connector. The PSoC 63 chip also sends the stimulation signal to the laser diode driver circuit through a connector that connects the top and bottom boards. The total weight of the opto-electrochemical platform is 6.4 grams, where 1.5 grams comes from the batteries. The proposed opto-electrochemical platform can be powered by small batteries and remotely controlled using the Bluetooth standard. Thus, it is compatible with a mounting on the head of freely moving rats for in-vivo optogenetic experiments.

Moreover, a novel low-power technique is proposed to perform simultaneous multi-frequency impedance measurements on cells. The proposed technique is based on a double demodulation of the measured signals, the first in the analog domain and the second in the digital domain. Square wave excitation and demodulation signals are used to allow a compact and low-power implementation in a silicon chip. A set of rules are introduced for selecting the frequencies of excitation and demodulation signals. This helps to avoid superimposing of harmonics due to the first analog down-conversion operation, and improves the measurement accuracy.

The schematic of our proposed multi-frequency impedance analyzer is shown in Fig. 2. A transimpedance amplifier (TIA) measures the current of sample-under-test (SUT) through a working electrode (WE). The high-frequency response of the SUT at a frequency FH is down-converted to a FHD using a square-wave mixer and digitized by a delta-sigma modulator (DSM). A direct path from the output of TIA to DSM input is used for simultaneous measurement of low-frequency impedance at a frequency FL, and DC current, enabling electrochemical impedance spectroscopy in parallel to cyclic voltammetry/amperometry

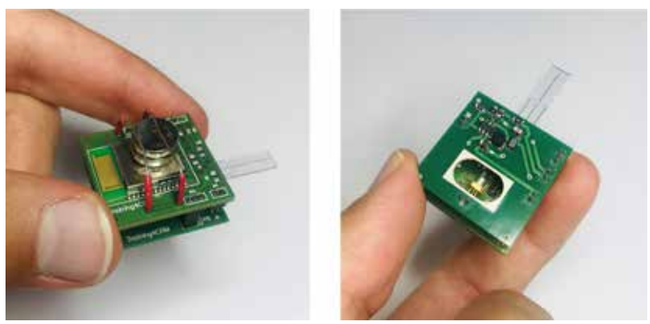

Fig. 1
PSoC 63 based opto- electrochemical platform.

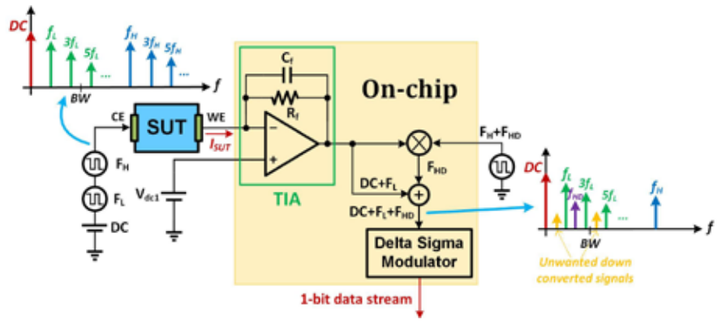

Fig.2
Multi-frequency impedance tracking.

measurements. Because of the virtual ground of the first integrator in DSM the signals are summed without adding complexity to the system. A single-bit second-order continuous-time DSM (CTDSM) is used for the implementation of the DSM. Thanks to the filtering action of the integrators, CTDSMs benefit from intrinsic anti-aliasing property. For the implementation of the modulator, we have selected a cascade of integrators with feedback loop topology.
This structure helps to reduce undesired effects of harmonics, as it provides better attenuation at higher frequencies than a cascade of integrators with feedforward topology. Moreover, it does not have peaking in signal transfer function, which can amplify unwanted signals.
Figure 3 shows the test board and chip photograph. The chip is designed in 180 nm TSMC CMOS technology and dissipates only 6 mW, including the analog-to-digital conversion of the signal, and operates up to a frequency of 15 MHz. The TIA was designed to work with WEs of 900 µm2.TIA and CTDSM need 0.43 and 2.92 mA from 1.8 V supply, respectively. The performed measurements confirm the proper operation of the system.

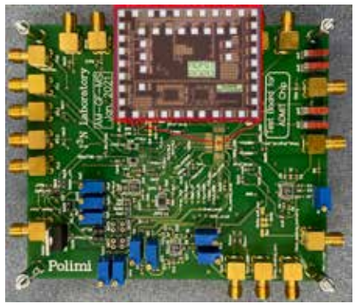

Fig.3
Test board and chip photograph.

ENHANCING THE QUALITY OF HUMAN-ROBOT COOPERATION THROUGH THE OPTIMIZATION OF HUMAN WELL-BEING, SAFETY AND PRODUCTIVITY

Costanza Messeri - Supervisor: Prof. Paolo Rocco

With the advent of Industry 4.0, collaborative robotics has become one of the enabling technologies of the smart factory. The collaborative robots (cobots) embody the most crucial cornerstones of this industrial revolution such as adaptability, flexibility, efficiency and interoperability. To improve the effectiveness and the fluency of cooperation, the cobot must be endowed with several advanced capabilities. For instance, the robot should be able to recognize the human partner's intention, monitor his/her work-related physical and physiological well-being, support to the human worker during the various phases of the collaborative task, and react to him/her, by selecting an optimized approach. In particular, the cobot should apply a strategy that ensures the adaptability to the behavioral, cognitive and physical features of the specific human it is cooperating with, while guaranteeing the efficiency of the productive process.

In this thesis, a multifaceted investigation on the most crucial features of the human-robot interaction (HRI) has been developed with the aim of promoting a better cooperation between the human and the robot, in line with the principles and objectives of Industry 4.0.

More specifically, this thesis aims at proposing new applications and strategies that enhance the quality of human-robot interaction in industrial frameworks, with a major focus on solving the trade-off between human efficiency and workplace well-being.

The current industrial revolution has radically changed the paradigm of the shop-floor worker and the organization of his/her working shift. Indeed, far from the old paradigm requiring the operator to do the same activity within the 8-hour shift, he/she is now required to learn and perform different collaborative tasks.

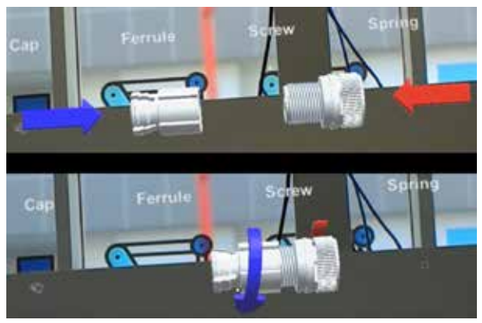

Fig. 1
Example of the holographic task related animations provided by the developed MR interface

Making the collaboration between a human and a robot easier becomes then crucial. As such, relevant features which can be optimized are enhancing the learning phase and increasing the intuitiveness of the task. Indeed, this can not only contribute at increasing the productivity of the company, but it can also be beneficial for the shop-floor worker, since a higher intuitiveness comes with a lower cognitive stress. To exploit the recent technological developments, a novel holographic mixed-reality (MR) interface has been developed to support the operator during the learning phase of a new collaborative task. The interface shows the operator intuitive 3D holographic animations (see Fig. 1) through which he/she is guided and assisted during the task execution.

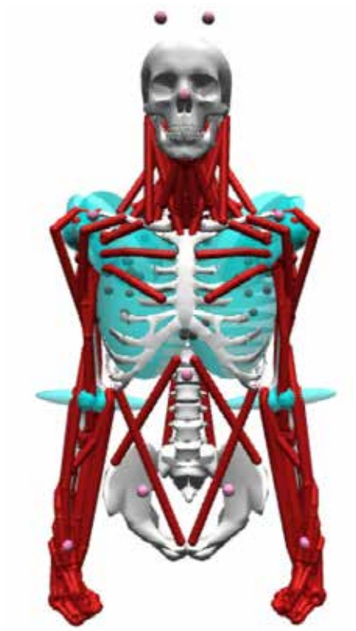

Fig.2
Developed dynamic musculoskeletal OpenSim model of the human upper-body

Concepts such as augmented or mixed reality, digital twin cyber-physical systems and sensors interconnection, are then becoming pervasive in this novel flexible and fast-evolving industrial scenario.

In this thesis, by incorporating the spatial understanding of a MR headset with the sensing capabilities of the work-cell vision sensor, a constrained particle filer-based method has been developed to improve the robot perception of the human operator in case of partial occlusion of his/her body. Furthermore, a digital twin (DT) of the collaborative workspace has been exploited both to simulate the robot motion, and to represent the work-cell volume occupied by the human operator. Based on this knowledge, the cognitive unit supervising the work-cell has been enabled to evaluate online the optimal trajectory for the cobot, by leveraging a genetic approach. This simultaneously minimizes the risk of collisions as well as the robot cycle time. These researches allow to increase the human safety and the robot adaptation capabilities during cooperation.

Fig.3
Experimental human-robot interaction scenario

The spread of cobots working alongside humans has paved the way to the study of several features of the human-robot interaction that can affect the quality of the cooperation in terms of worker's well-being and job performance. An example is the prevention of the risk of musculoskeletal disorders (MSDs) or the onset of cognitive distress which human operators may undergo while working.

Thus, the impact of crucial human factors such as cognitive, physical distress and interaction role (whether being leader or follower during cooperation) in influencing the quality and effectiveness of human-robot interaction has been a major focus of this thesis, beyond the mere productive aspects. In particular, a detailed analysis on how the interaction role of the robot influences the psycho-physiological response and the production rate of the human fellow operator has been carried out. Based on that, a novel method exploiting a game-theoretic approach has been proposed to model the trade-off between the human performance maximization and cognitive stress minimization.

Besides, the outcomes of the previous analysis have been pivotal to develop a novel robot adaptive control strategy. More specifically, the robot has been endowed with the capability of applying a suitable alternation of the leader-follower interaction modes, based on the real-time evaluation of human stress and performance, with the aim of simultaneously increasing the human productivity and mitigating his/her cognitive stress.

Ultimately, an online dynamic task allocation strategy has been developed to guarantee the minimization of the human physical fatigue, as well as the effectiveness of the production process. This strategy is based on a novel non-invasive method to estimate online the muscular fatigue experienced by the human operator during the task execution. The estimation process relies on a sophisticated musculoskeletal model of the human upper body, illustrated in Fig. 2, and on an external 3D vision system used to track human motions in real-time.

The effectiveness of the proposed methods has been experimentally validated in real human-robot collaborative scenarios, such as the one shown in Fig. 3, involving several volunteers and the ABB YuMi cobot.

DIGITALLY ASSISTED FREQUENCY SYNTHESIZERS AND DATA CONVERTERS FOR WIDE-BAND RADIO SYSTEMS

Angelo Parisi - Supervisor: Prof. Andrea L. Lacaita - Co-Supervisor: Prof. Carlo Samori

Technology scaling has been driving integrated circuits designers towards denser and faster systems since its early days, in turn influenced by the increasing demand of performance. Main protagonists of the revolution are communication systems, which currently adopt RF and mm-Wave carriers for high throughput data transmissions requiring modulation bandwidths ranging from few hundred MHz to about 1GHz. State-of-the-Art radios must then reach unprecedented accuracy on wide ranges of frequencies and operating conditions, at the same time coping with the worsened analog performance of the high-density, digital-friendly technologies allowing to reach such frequencies. The change in paradigm of recent years, favouring digital assistance of the essential analog building blocks, blurs the edges of mixed-signal implementations: the cost and overhead of *p.p.m.* accuracy of analog designs is circumvented by means of digital calibrations, steering towards systems that minimize the amount of custom building blocks preferring scalable, reusable and portable subsystems. Examples of this trend are frequency synthesizers and data converters, cores of most radios. In this work are presented examples of designs that might enable efficient new generation radios, starting from digital and hybrid 13GHz Phase-Locked Loops, embedding high efficiency oscillators to achieve sub-100fs *r.m.s.* jitter with low power consumption, to GSps time-interleaved Analog-to-Digital converters combining several low-power cores controlled by digital background correction algorithms reducing spur power to less than 70dB below the carrier. Efficiency and performance are achieved by means of novel Least Mean Squares control loops and adaptive filters for distortion and mismatch compensation. Evolution of integrated communication systems has been following the tremendous growth of transistor density for the best part of the century. The trend entails a never-ending rush of both technology and knowledge to push the capabilities of integrated circuits ever closer to their natural, ultimate limits. Milestone by milestone, technical progress advances by means of radical innovation and ingenuity: here are presented, described and analysed some contributions to the great flow of knowledge, potentially enabling or humbly aiding the next steps in frequency synthesis and data conversion for communication. Analysis and modelling frameworks are described first, both for Phase-Locked Loops (PLLs) and Analog-to-Digital Converters (ADCs), enabling preliminary estimation of the system performance and driving the design and implementation with solid foundations. The effectiveness of the methods is compared to both system-level simulations and actual measurements. Specific contributions are then focused for each of the fields: an automatic identification mechanism for ADC static distortion correction coefficients estimation; an algorithm easing the testing setup requirements for ADC dynamic and memory distortion compensation; a novel and general on-chip background algorithm for sampling clock skew compensation in Time-Interleaved ADCs of any order; a simple and essential start-up and biasing circuit for a class of high-efficiency oscillators, bridging the gap with conventional implementations by reducing the design overhead.
A detailed organization follows: Chapter 1 introduces the framework of modern wideband radios and the main challenges relevant for this work in terms of subsystem performance, such as clock jitter and conversion resolution. Trends and challenges of modern communication systems are introduced, focusing on requirements of radio subsystems that are most affected by the recent revolution of wireless transceivers. The analysis is backed by a survey of communication standards, integrated frequency synthesizers and data converters, to highlight the main limitations at the building block level and their impact on system design. The interest from both application and research perspectives is further supported by a number of recent references and publications of the community, stressing the hot topics on each of the discussed matters; Chapter 2 describes time-based systems, models and implementations, providing examples and comparisons with several designs. High-speed system performance encounters a bottleneck when timing uncertainty is taken into account. RF frequency synthesizers must provide p.p.m. accuracy in the synthesized waveform, not to impair the resolution of the modem or converter they drive. Furthermore, the sensitivity of the driven subsystems needs to be very accurately modelled and inspected in the design phase, greatly increasing the project overhead. Efficient behavioural models have been developed recently, with solid foundations on the theoretical aspects of time uncertainty, allowing orders of magnitude speed-up in system design with no penalty on accuracy. The most relevant models for the scope of this work are detailed and analysed here, together with sizing examples and analysis results; Chapter 3 introduces to converters distortion, identification and correction, and describes new techniques for both static and dynamic compensation of converters. The ideal behaviour of an Analog-to-Digital converter is conventionally described as a staircase-like transfer characteristic. When tried to fit on an actual converter, this entails several assumptions on the operating region of the latter, the departure from which is generally identified as distortion. The most common and relatable sources of distortion are described in this chapter, alongside with common as well as novel approaches to compensate them; Chapter 4 details the implementation of a Time-Interleaved converter in 28nm CMOS technology, embedding an original and scalable on-chip algorithm for timing skew identification and correction allowing any number of ADCs to be combined. Measurements of the effectiveness of the proposed technique are provided, alongside with the converter performance in all configurations, thanks to the highly automated developed testing environment. The models and techniques analysed in the previous chapters enabled the design of the converted described here. The overall system is detailed first, highlighting trade-offs and design choices. The design of the relevant critical blocks is then analysed in detail, with emphasis of the most relevant for the obtained performance; Chapter 5 introduces to high performance RF and mm-Wave oscillators, cores and bottlenecks of frequency synthesizers, and shows the devised improvement for class-C oscillators, backed by start-up, settling transient and noise measurements. The chapter is devoted to the description of high-performance integrated oscillators for low-noise frequency synthesis, leading to original results and analyses. Basics on oscillator design are introduced first, moving then to metrics and benchmarking of the most popular topologies. The chosen class for implementation is then analysed in detail, reaching new insight that allows then to devise improvements on its structure. Examples of sizing and details on the implementation are then provided, alongside with performance measurements.

ACCELERATING GRAPH AND SPARSE INFORMATION RETRIEVAL THROUGH HIGH-PERFORMANCE RECONFIGURABLE ARCHITECTURES

Alberto Parravicini – Supervisor: Prof. Marco Domenico Santambrogio

Graph analytics, information retrieval, and recommender systems process an always-increasing amount of data, often with strong real-time constraints, to suggest products, movies, news articles to billions of users. For better or worse, they are an integral part of our society. Sparse linear algebra is now a staple of graph analytics and recommender systems, as it is the only way to perform high-performance numerical computations on these enormous datasets.

Problem Statement
The popularity explosion of graph analytics and deep learning demands novel techniques to process sparse matrices with millions or billions of non-zero entries, representing social networks and databases of embeddings. However, hardware optimizations for recommender systems are still uncommon. We explore the acceleration of sparse linear algebra for graph analytics and recommender systems using modern FPGA accelerator cards and approximate computing. In this context, many research problems are still unsolved.

The Memory Wall Keeps Rising
Large recommender systems now require up to 10 TB of memory, and this value has increased more than 200 times in the past two years. Social network graphs count billions of users, and recommender systems have millions of products. While peak compute power has increased by 90000 times over the last 20 years, memory bandwidth has grown by just 40 times. The insufficient size and bandwidth of memory create what is known as memory wall, a major bottleneck of today's recommender systems due to the low operational intensity of these workloads. Techniques that maximize computation for every bit loaded from memory, such as reduced-precision arithmetic and sparsification, will become more and more necessary.

Accelerating Sparse Linear Algebra by Leveraging Novel Memory Technologies
Effective hardware acceleration of sparse primitive operations is still an open challenge. Compared to dense linear algebra, sparse operations have complex memory access patterns, and memory bandwidth is often their performance bound. Thanks to their abundant memory bandwidth, new memory technologies such as HBM can be of great help. However, fully leveraging this bandwidth is not trivial as the unpredictable memory accesses of sparse computations often require problem-specific solutions, and generalization cannot be taken for granted.

Bringing FPGAs into the Equation
Given these challenges, FPGAs are an attractive architectural choice. FPGAs can leverage reduced-precision fixed-point arithmetic, which is highly effective in error-tolerant workloads typical of recommender systems. FPGAs are more energy-efficient than competing hardware and have predictable execution latency, making them suitable for real-time data center workloads. However, leveraging these FPGAs accelerator cards for sparse recommender system and graph analytics workloads is yet to be fully explored, even more so when reduced-precision fixed-point arithmetic and optimal memory bandwidth utilization are considered.

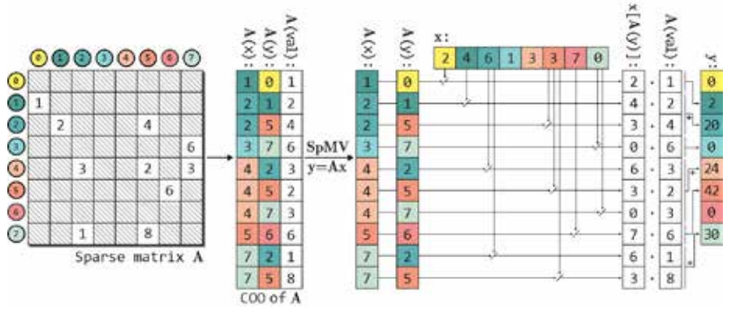

Fig. 1
Example of Sparse Matrix-Vector Multiplication (SpMV), a sparse linear algebra operation widely employed in graph analytics and recommender systems.

Contributions
We propose a novel set of **flexible SpMV hardware designs** for high-performance sparse computations in graph analytics and recommender systems, leveraging reduced precision fixed-point arithmetic and multiple kinds of memories. Our SpMV designs adapt to different sparse workloads, such as PPR, sparse eigensolvers, and sparse embedding similarity search.

A Reduced-Precision Streaming SpMV Hardware Design for Personalized PageRank on FPGA
We introduce our FPGA SpMV hardware design in the context of graph ranking algorithms, using Personalized PageRank (PPR) as a case study. Here, low latency and high throughput are more valuable than exact numerical convergence, creating the ideal condition to experiment with reduced-precision fixed-point arithmetic. Our PPR hardware design is six times faster than a state-of-the-art multi-threaded CPU implementation, with up to 42 times higher energy efficiency and without significantly lower accuracy. Moreover, we show that fixed-point arithmetic converges two times than floating-point arithmetic.

Solving Large Top-K Graph Eigenproblems with a Memory and Compute-optimized FPGA Design
We then extend our SpMV hardware design to sparse eigensolvers, often used for spectral methods in graph analytics and information retrieval, to approximate the most important features of a graph. To the best of our knowledge, we are the first to propose an FPGA Top-K eigensolver for unstructured sparse matrices. We integrate our SpMV hardware design into a mixed-precision eigensolver based on the Lanczos and the Jacobi eigenvalue algorithm. We achieve a speedup of 6.22 times over the highly optimized ARPACK library running on an 80-thread CPU, with 49 times better power efficiency.

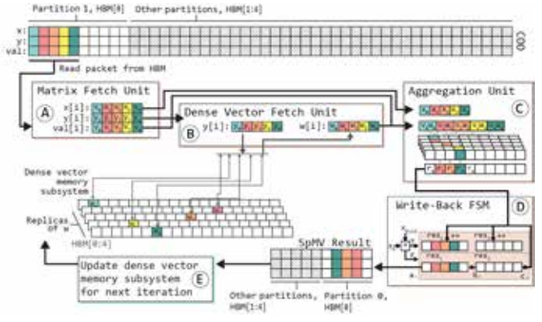

Fig.2
Block diagram of one iterative SpMV Compute Unit (CU). Each CU processes a portion of the input matrix through a 4-stage dataflow design.

Scaling up HBM Efficiency of Top-K SpMV for Approximate Embedding Similarity on FPGAs
Finally, we consider approximate similarity search on large sparse embedding tables, which can be efficiently implemented with Top-K SpMV. We extend our design to handle Top-K SpMV, optimizing the computation for tall matrices that benefit from on-chip caches. We create a new approximation scheme for Top-K SpMV, parallelizing the computation over 32 independent SpMV Compute Units, and introduce BSCSR, a novel sparse matrix compression scheme optimized for streaming reduced-precision computations. Our FPGA design is 2.1 times faster than a GPU with 20% higher bandwidth with 15 times higher power efficiency, proving that FPGAs are today the optimal solution for Top-K SpMV.

ADVANCED INSTRUMENTATION FOR TIME-RESOLVED SINGLE- AND MULTI-PHOTON COUNTING APPLICATIONS

Klaus Pasquinelli - Supervisor: Prof. Franco Zappa

Many high-end applications require very sensitive photodetectors and imagers able to measure very faint and very fast optical signals. Examples of such applications are distance ranging by measuring photons' time-of-flight (TOF), fluorescence lifetime imaging, and quantum communications. While some photon-starved applications require Single-Photon Avalanche Diode (SPAD) arrays thanks to their single-photon sensitivity, others must operate with either a few or even many photons per event, so Silicon Photomultipliers (SiPM) or alike should be preferred.
My PhD research aims at developing advanced electronic instrumentation based on multi-pixels photon detectors and imagers, able to measure the intensity of light events by counting the number of incoming photons, record the time-resolved optical waveform through the measurement of the photons' arrival time, and detect coincidences in the arrival time across the SPAD array sensors. To ensure optimal performance in ToF applications, both practical and theoretical studies about the photodetector choice will be provided. Starting from the results obtained in previous photodetector modeling, this research will study in depth the performance achievable with both avalanche photodiodes (APD), SPAD and SiPM, with a variable number of cells. Furthermore, a dedicated hardware with field programmable gate array (FPGA) to detect photon coincidences will be designed. This will pave the way to a new instrument based on "programmable" SiPMs, able to detect photon arrival coincidences with variable photon thresholds, to be adjusted based on the operating conditions.
An introduction of the background of LiDAR measurement systems will be presented, from the first historical mechanical example to the modern system based on electromagnetic waves and image sensor. After a brief overview on the various modern techniques, the photodetector used for LiDAR systems will be presented and then the state of the art will be discussed. The main part of the thesis will be about the modeling of photodetectors. After a description of the characterization of the detectors, a mathematical model will be presented for avalanche photodiode (APD), single-photon avalanche diode (SPAD), and for the silicon photomultiplier (SiPM).

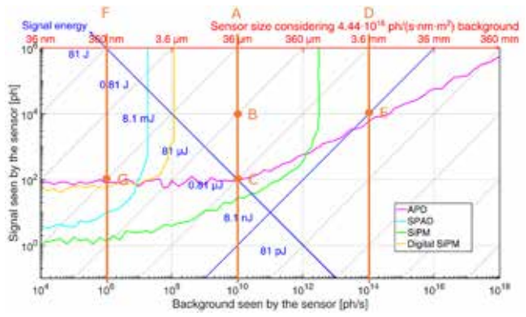

Fig. 1

In the picture the nomogram of ideal detectors, as a function of incoming background photon rate (horizontal bottom axis), number of impinging signal photons (vertical left axis), sensor size (top horizontal red axis), and signal energy (slanted blue axis).

The border lines are those where the "success ratio" is 70%; higher success is met in the upper- left-most area of the working regions (i.e., for higher signal and lower background levels).

These models will lead to a performance comparison among the detectors through a nomograph. Given these results a prototype of single-pixel time of flight (TOF) 5 × 5 SPAD camera will be presented. Chip, system, and measurements will be discussed. From this prototype digital and analog methods to reject the background light will be proposed and analyzed, so to increase the signal to noise ratio. Various algorithm and architectures will be proposed for single-shot and multi-shot applications. Eventually one of these methods will be used in a prototype chip for industrial applications.

In the picture one of the method is shown: the multi-hit histogram with adaptive threshold, in order to fill a maximum of 10 available TOF registers. The bars in blue are the

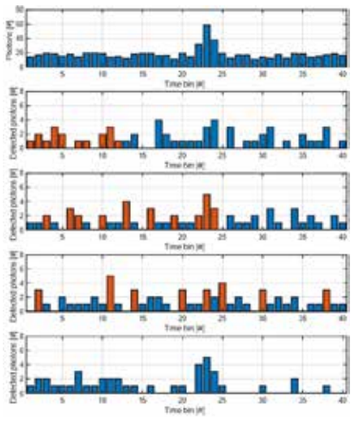

Fig.2

histogram is built with all the collected data obtaining the last plot.
Eventually, the last part presents the simulations and the modeling of a new type of hybrid digital and analog SiPM with on-chip background rejection circuitry, in this chip a new topology for the peak detection has been implemented.
In the first phase, performances, with the given parameters will be shown and then, improvements will be added so to adapt parameters to the target specifications to reach better performances.
Considering the given specifications and the new method for the peak detection the prototype should reach distance upto 80 m with a success ratio over 70% with 100 klux background. Eventually, system prototype and achievements will be shown and compared to the predicted one.

In the picture the results obtained with a laser with a peak power of ~230 mW in a dark ambient. Other measurements with a more powerfull laser will be done to evaluate the whole system with a background of 100 klux. All these works could provide the development of new LiDAR systems for high-background applications based on the true peak detection heavily improving the success ratio respect to the classical threshold method. Furthermore, the peak detection method and the other illustrated methods can be used in other fields where it is required to detect the highest value in an analog stream.

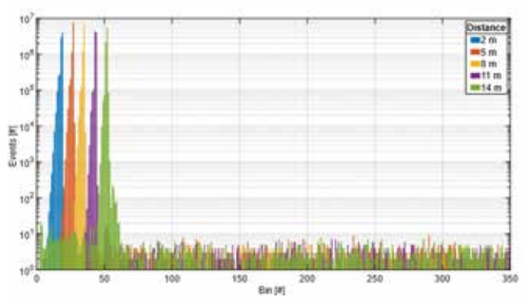

Fig. 3

number of photons that don't trigger the sampling of the TDC, those in orange are the bins where the number of photons triggers the sampling of the TDC a) the returning photons array for one laser shot; b) threshold = 1, c) threshold = 2; d) threshold = 3 plots the detected coincidences. If the number of detected coincidences exceeds 10 in a repetition, then the threshold is increased by one for the next repetition, so to reject more the background and acquire less (but more useful) TOF data; e) final histogram after N repetitions. At the first round the threshold is set to 1 photon: the orange bin represents the bin that will trigger the sampling of the TDC. Since 10 sampling are done, the threshold is increased to 2 photons. Again, 10 sampling are done, so the threshold is increased to 3 photons. With 3 photons threshold the TDC is sampled less than 10 times, so the threshold is kept constant for the subsequent acquisitions. After a number N acquisitions, the final

MACHINE LEARNING BASED MANAGEMENT AND MONITORING OF NEXT-GENERATION COMMUNICATION NETWORKS

Andrea Pimpinella - Supervisor: Prof. Alessandro Enrico Cesare Redondi

Network intelligence regards the embedding of Artificial Intelligence (AI) in all network domains to fasten service delivery and operations, guarantee service availability, allow better agility, resiliency, faster customization, and security. Moreover, network intelligence concerns the provisioning of real-time network self-awareness about how it is performing. In this vein, my thesis explores the potentialities of AI to become the game-changing technology that will satisfy communication networks seeking of intelligence. Focusing on Machine Learning (ML) as the most popular approach to AI, my Ph.D. research analyses benefits and drawbacks derived from integrating intelligence into several different networking scenarios. My Ph.D. work groups into two main parts, as it follows.

The first part of my thesis studies a QoE-aware strategy to perform monitoring and anomaly detection in cellular networks. Network monitoring is a major concern for Mobile Network Operators (MNOs), that put efforts to optimize performance and recognize service issues to give the best possible service to subscribers and avoid customers' churning to a different operator. In this context, a popular strategy to monitor the level of customers' satisfaction is to rely on the administration of surveys regarding the users' Quality of Experience (QoE) of certain mobile services. Such surveys can then be used by the operator to reveal issues in the mobile network (e.g., under-performing network sites). Unfortunately, the detection of under-performing cells using subjective users' feedbacks has its own issues. First, performing directed customer feedback surveys is costly and cumbersome for MNOs. Second, users are heterogeneous and their perception of network quality is highly subjective. Third, it is difficult to identify which of the network sites visited by a user is the most responsible for user's dissatisfaction.

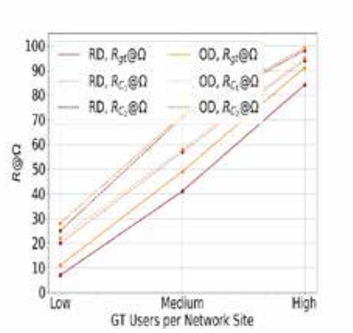

Fig.1
Under-performing sites detection performance of the proposed system (% of detected sites). Regarding line styles: the solid lines refer to the case when only ground-truth users feedbacks are included in the detection process, while the dashed and dotted lines refer to the case when also predicted feedbacks are leveraged by the system. Regarding line colors: the red lines correspond to the case when surveys are randomly submitted to users while the orange lines refer to the case when a maximum coverage survey allocation strategy is used by the operator.

To study these aspects, my thesis proposes an empirical framework tailored to assess the quality of the detection of under-performing cells starting from subjective users' grades. Considering the generic poor users' cooperative attitude in answering directed satisfaction surveys, we also explore the possibility of predicting the long-term users' satisfaction and study the impact that QoE prediction errors have on the overall anomalous sites detection process. As shown in Figure 1, even with modest QoE prediction performance MNOs can exploit users QoE prediction to reinforce network monitoring systems.

The second part of this Ph.D. thesis analyses the performance of two ML-based solutions to assist network monitoring and resource planning relatively to two specific services, namely video streaming and smart devices localization. On the one hand, we develop a ML-based video traffic monitoring system for HTTP Adaptive Streaming (HAS) applications, able to i) classify the type of last HTTP content request and ii) predict when the video client will issue the next request. On the other hand, we investigate a ML-based solution able to perform cross-technology transferring of radio map calibration data (namely, transfer learning) and augment the performance of both indoor and outdoor localization systems.
For what regards video streaming services, from the perspective of an Internet service provider or a MNO dealing with high volumes of video streaming traffic is particularly challenging. At the same time,

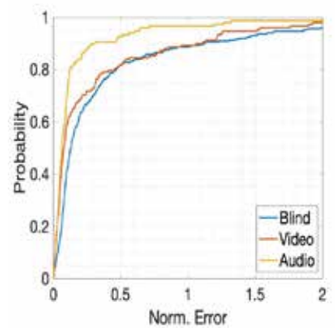

Fig.2
Cumulative distribution functions of the normalized error on the estimation of the arrival of next HAS uplink request, for two different scenarios. The blue line refers to the case when no a priori information regarding the type of the last observed uplink request is leveraged by the predictor. Differently, the yellow and red lines refer to the case when i) first, the last observed uplink request is classified as being for an Audio (yellow) or Video (red) content, and ii) the arrival of the next uplink request of the same type is estimated.

video streaming comes with tight QoE requirements which if not met can increase users' dissatisfaction and consequently the churn rate to other operators. Therefore, network operators often focus on smart management approaches and techniques to optimize the allocation of resources to guarantee QoE requirements in an efficient way. In this vein, my thesis proposes ML-based video streaming traffic monitoring architecture able to:

- Classify the type of a HAS uplink request as being for an audio or a video content;
- Predict the next HAS uplink request arrival based on a dataset of more than 15,000 requests and 900 YouTube streaming session.

Remarkably, both the predictor of the next request arrival and the request type classifier are fed with lightweight features extracted from encrypted traffic in an online fashion. Results show that i) the system can classify the type of HAS uplink requests with an accuracy greater than 95% and ii) pipe-lining request type classification and prediction of next request-arrival time improves by more than 40% the final prediction performance (as depicted in Figure 2).
For what concerns transfer learning, it is a recent deep learning based framework that aims at transferring a certain knowledge to execute a task in a given context, into a different, but related context to execute a different task. In my thesis, we analyse how to apply transfer learning based solutions to enhance the performance of a specific mobile service, namely smart devices localization. An energy-aware and lightweight approach to localize a smart device is to exploit the data packets transmitted by the device (and received at multiple radio access gateways) and estimate the device's position through fingerprinting. As far as wireless networks are concerned, different radio technologies often coexist in the same geographical area, usually deployed chronologically one after the other within the same network infrastructure (i.e., with the radio access points being often co-located). In this vein, we study the possibility of enhancing the performance of both indoor and outdoor localization systems by transferring assistance radio map data from one radio technology to the other. As one can see in Figure 3, results show that knowledge transfer approaches outperform classical methods for radio map interpolation by more than 10%, especially when the initial knowledge in the domain hosting the localization service is limited.

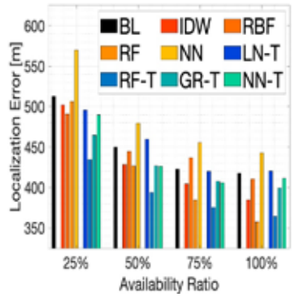

Fig.3
Comparison of 80-percentile of the localization errors in outdoor scenario for Baseline (BL), intra- and inter-technology methods. When the initial knowledge in the domain hosting the localization service is limited (i.e., for low availability ratios), inter-technology approaches for radio map estimation have higher benefit on the localization performance than traditional approaches.

ENGINEERING INFORMATION TRANSFER IN LIVING SYSTEMS VIA MOLECULAR COMMUNICATION

Francesca Ratti - Supervisors: Prof. Maurizio Magarini, Prof. Domitilla Del Vecchio

Molecular Communication (MC) is an emerging field directly inspired by natural communication between cells in biology. In MC, the information is encoded into and decoded from molecules rather than electromagnetic waves, thus exploiting biological materials to enable communication among biological nanomachines, which are existing or artificially synthesized small-scale devices. Some examples of nanomachines are genetically engineered cells, molecular motors, biological and artificial cells, synthetic molecules, and bio-silicon hybrid devices. The study of MC has applications in several fields. One of the most prominent is biomedical research, e.g., to address the problem of smart drug delivery, detect pathologies, and model the spreading of infections.
This is an interdisciplinary research that focuses on applying communication and information theory concepts to the field of MC. Indeed, even though MC systems are fundamentally different from their traditional communication counterparts, the latter can provide key tools and theoretical foundations for their analysis. This research focuses on the objectives stemming from the design of synthetic biological circuits to exploit, control, and enhance the performance of MC systems. The main goal of our investigation is to understand how to maximize the reliable information exchange in MC circuits. In general, information transfer in such systems is hindered by the intrinsic stochasticity of natural biological processes, e.g., the non-deterministic occurrence of chemical reactions, the diffusion of molecules in the extracellular environment.

In this work, we face this challenge both from a data-driven and an analytical perspective, investigating novel techniques to quantify and optimize information transfer. This work includes a fundamental and application-agnostic standpoint, focusing on molecular circuits in cells and biochemical signaling systems. Additionally, it also includes applications and results in different fields, ranging from biological cells communication to genomic systems. A common thread of this research is the application and adaptation of traditional concepts of information theory, such as mutual information, channel capacity, and sampling theorem, to biological circuits.

More in detail, in this work, we explore the three directions listed in the following.

Estimation and improvement of communication properties of biological systems starting from in-silico/in-vitro data. This direction concerns the usage of collected experiments to learn instead something about the underlying system that has generated it without having any prior knowledge of the statistical model characterizing it.
As a case study, we consider biological systems characterized by some kind of internal information exchange. Such exchange can be described from a communication point of view, allowing defining a capacity for the communication channel (which, in this case, is a molecular channel) and a detection of the transmitted message problem.
We face the problem of estimating the capacity, which is strictly related to estimating the optimal way of transferring information in the system, only using a set of inputs/outputs for the system, that can be, for example, the result of in-silico/in-vitro experiments. This is particularly useful since, for many biological systems, an analytical or statistical description does not exist, but datasets of inputs/outputs can be easily obtained. Furthermore, we propose a solution to overcome the issue of detecting the correct transmitted symbol at the receiver side.
We propose a novel methodology that frames the estimation of the capacity as the optimization problem of finding an upper and a lower bound on the true value. The bounds are optimized starting from the data and using any derivative-free iterative algorithm. Being estimated from the data, the accuracy of the resulting interval is affected by the uncertainty and the volume of the available data. Therefore, particular emphasis has been placed on overcoming data scarcity and managing uncertainty in the available biological data. For this, and since gathering new experiments is usually costly and time consuming, the proposed methodology has been equipped with a deep learning-based data augmentation module. When the volume of data is sufficiently large, we propose also an alternative solution, i.e., the pruning technique, to adapt the dataset at each iteration of the algorithm. This methodology is experimentally evaluated and validated on a system composed of two prokaryotic cells.

Analytical evaluation of communication performance and receiver design implementation for a statistically approximated model of the MC channel. While data-driven approaches have the advantage of not relying on any approximated model, the benefit of analytical studies relies on the generality of the obtained results, that are not experiment-depending.
We focus on the analysis of a diffusion-based MC system where the received signal is approximated as a Poisson random variable. concentration shift keying (CSK) is used as the modulation technique for encoding information in the system. In particular, we aim to study the performance of the MC system in terms of reliable information exchange for the channel with finite-state memory, which introduces inter-symbol interference (ISI). The main objective is the derivation of analytical expressions for the upper and the lower bound of the constrained channel capacity for a range of values of the modulated symbols, i.e., for a number of different sets of amplitude levels of CSK modulation, and for various levels of ISI. In addition, the numerical evaluation of the derived expressions is presented. Results allow discussing the relationship between ISI level and achievable channel capacity. Furthermore, we introduce novel approaches for the design of the linear filter and the detection algorithms in unbounded advection diffusion-based molecular communication systems affected by ISI. For this study, the received signal samples are modeled as Poisson random variables with memory where the effect of enzymatic reactions is also included. A main characteristic of our proposed filter design is to allow for a real-time computation of the filter's coefficients. For the detection of the transmitted symbols, we define an averaging method suitable for time-varying channels with finite memory length. The mean value of the Poisson channel varies with time, and we quantify the memory length with a finite number, from the receiver point of view. The computational burden of the proposed approaches is evaluated in terms of number of required operations and their performance is evaluated in terms of bit error rate (BER) for different sets of parameters.

Analytical evaluation of the effect of retroactivity on the MC performances. Information exchange is a critical process in all communication systems, including biological ones. The concept of retroactivity represents the loads that downstream modules apply to their upstream systems in biological circuits. We focus on studying the impact of retroactivity on different biological signaling system models, which present analogies with well-known telecommunication systems. The mathematical analysis is performed both in the high and low molecular counts regime, by mean of the Chemical Master Equation (CME) and the Linear Noise Approximation (LNA), respectively. The aim is to provide analytical tools to maximize the reliable information exchange for different biomolecular circuit models. Results highlight how, in general, retroactivity harms communication performance. This negative effect can be mitigated by adding to the signaling circuit an independent upstream system that connects with the same pool of downstream systems.

FORCE-FEEDBACK CONTROL SYSTEM DESIGN FOR HIGH-PERFORMANCE VEHICLES

Giorgio Riva - Supervisor: Prof. Sergio Matteo Savaresi

The automotive scenario is currently experiencing an important transition towards a new mobility model. Such revolution is driven by two main long-term objectives: 1) the reduction of the greenhouse gases production to reach the so-called climate neutrality and 2) the complete automation of four-wheel vehicles. The former is leading the market towards hybrid and electric vehicle architectures, which represent an opportunity in the control system framework thanks to their architectural flexibility. Indeed, regenerative active braking represents a fundamental feature in such architectures, which should be properly managed in combination with suitable passive braking technologies. The autonomous vehicles framework, instead, represents the common link among all the technological challenges arising in automotive. Regarding the latter, the Drive- By-Wire (DBW) framework represents a fundamental topic, that becomes crucial once the driver is moved outside the loop and the actuator commands are given by the Electronic-Control-Unit (ECU). On the other hand, the computational load required to acquire the entire suite of sensors employed in autonomous vehicles, process the data and take decisions, is raising the bar of the needed amount of computational power that should be brought on-board. Such computational capacity opens many possibilities for the development of novel control and estimation algorithms, allowing to use of more complicated, but accurate, models and more complex optimization routines, as discussed in this work.

In this promising scenario, this dissertation deals with estimation and control of forces in the Vehicle-Dynamics-Control (VDC) context. At any layer, forces represent the root cause of the behaviour of any vehicle components and of the whole vehicle itself. Thus, they turn out to be pivotal quantities to be known, even mandatory in some specific applications. The backbone of this work can be exemplified through a conceptual control scheme, where forces appear at two different layers of the general control scheme of the dynamics of a vehicle. The internal layer deals with the lower-level part of the control scheme, namely the braking actuator: in this work the Electro-Mechanical (EM) Brake-By-Wire (BBW) solution is studied, which is part of the more general Drive-By-Wire (DBW) scenario. In this braking actuators the knowledge of the clamping force exerted by the pads on the disk is fundamental for the correct functioning and to achieve a desired performance level. The external layer, instead, focuses on the contact forces between the tires and road, which are well known to be fundamental in the vehicle dynamics, governing the motion and defining almost completely its performance and stability. Each problem has been tackled covering both the estimation and the control tasks, but without investigating the interaction between the developed estimation algorithms and the proposed control techniques.

Starting from the internal layer, as anticipated, the Electro-Mechanical (EM) Brake-By-Wire (BBW) actuator is handled. Such brakes have no hydraulic components and thus the braking force exerted between the pads and disk is the only variable which matters to characterize the braking manoeuvre. Therefore, the main goal for such systems is clamping force control, which requires an accurate knowledge of the force exerted by the pads in order to guarantee the proper actuation capabilities for higher-level controllers. Given limitations in the available sensors, estimation represents a key tool to step forward towards the commercialization of such technology. Current state-of-the art solutions are strongly model-based and thus very sensitive to friction modeling errors, thus leaving the way open to alternative solutions learning the physical relationships of interest directly from data. Indeed, in this work, a novel Black-Box (BB) approach for the clamping force estimation problem is proposed, identifying the hidden physical behaviour of the actuator from data and separating the detection of the contact from the real-time estimation itself. The control problem represents a well-known topic in the scientific literature, solved mainly employing classical tuning approaches, like gain scheduling and loop-shaping techniques. Despite this fact, it still represents an interesting research scenario to develop advanced tuning techniques directly accounting nonlinearities and model uncertainties to improve closed-loop performance while dealing with energy concerns. Indeed, in this work, two solutions are proposed looking at the problem from different perspectives, namely the tuning of a state-of-the-art controller in nominal conditions facing the main nonlinearities and the robust tuning in presence of actuator uncertainty, typically occurring either along the life cycle of the actuator or due to dispersion in mass-production.

At the external layer of the conceptual control architecture, the focus of the work moves towards the study of tire-road contact forces, which determine the motion of the vehicle and define its stability properties. The state-of-the-art control strategies in this field typically rely on the knowledge of vehicle slips, respectively longitudinal and lateral ones, obtained through suitable estimation algorithms, being not measurable quantities. The main limitations of such slips-based strategies regard the knowledge of the friction characteristic, related to both the shape, e.g., the stiffness, and its maximum value, whose uncertainties can strongly influence the performance of the closed-loop systems. Moreover, the typical nonlinear friction behaviour, mapping slips into forces, requires a complexity increase in the developed controllers, typically addressing either nonlinear strategies or scheduling techniques. In this context, forces can represent a viable solution to mitigate such limitations: indeed, their knowledge is expected to be useful to reduce the effect of friction curves uncertainty and the complexity of the controllers' structure, hiding some nonlinearities due to friction models. The problem of how the information of tire-contact forces could be integrated in a vehicle control system is still almost an open topic in the scientific literature, since only few works tackle this problem. In this work two novel applications are considered where only longitudinal forces are considered. In the first one, where a complete Vehicle-Dynamics-Control (VDC) system in an autonomous-like scenario is proposed, the performance increase due to the friction uncertainty insensitivity are discussed, while the second one, a Model- Predictive-Control (MPC) Anti-lock Braking System (ABS), will show the benefits both in terms of robustness and implementation simplicity. In this scenario, the absence of reliable and cost-effective commercial sensors represents the main reason for the lack of literature works and therefore of state-of-the-art control strategies employing force information. Thus, tire force estimation represents a fundamental step to open the way to such control strategies. Despite the huge amount of literature works available, such approaches share features and limitations, employing simplified vehicle models with limited description capabilities. In this work we decide to tackle the problem providing two timely contributions to overcome the main limitations of state-of-the-art approaches and sharing as common feature the exploitation of the augmented computational capacity that will be available in the next generation of vehicles. The first solution is a Black-Box (BB) approach, where the crucial issue is represented by the appropriate selection of the regressors. The second approach, instead, represents the most novel contribution in this dissertation, where a model-based observer, in which a fine-tuned multibody vehicle simulator takes the role of the prediction model, is proposed in a unified estimation framework.

DESIGN OF A UAV BASED LOCALIZATION SYSTEM FOR PUBLIC SAFETY NETWORKS

Davide Scazzoli - Supervisor: Prof. Maurizio Magarini

Unmanned Aerial Vehicles (UAVs), also referred to as drones, have revolutionized several industries in recent years. Their superior mobility and Vertical Take-off and Landing (VTOL) capabilities make them particularly suited to tasks of reconnaissance and survey, as they can gather information from dangerous places without incurring any risk of loss of life. With the benefits of using UAVs for civil law enforcement well established in the literature and even futuristic approaches to warfare emerging in recent years, it is safe to assume that UAVs will have a key role in many applications in the near future. One such application, which is being investigated in recent works of the scientific literature, is their usage as enablers of rapid response and on demand deployment of telecommunication infrastructure. These aspects can provide invaluable help for Public Safety Communications (PSC) in the event of an emergency. Search and rescue missions after a disaster require substantial manpower and time efforts, the detection and localization of weak signals transmitted by smartphones can provide invaluable aid to these operations, minimizing wasted time and help saving lives. UAVs fit perfectly into the context of Public Safety Network (PSN) in emergency scenarios, with their ability to rapidly deploy critical communication infrastructure to supplement, or replace in the case of failure, the existing one. For this reason, many projects that take advantage of UAVs in PSN have emerged throughout the literature. It is within this context that we can develop localization capabilities for the UAVs, as they are en-route to deliver their mission payload. Over their trajectory, UAVs may scan the nearby areas using beamforming techniques to enhance Signal to Noise Ratio (SNR) and, therefore, classification and localization of active RF transmitters as they move. With their ability to easily steer around obstacles, survey unexplored areas and venture into dangerous territory, the need for providing solutions that can provide these features will undoubtedly soar in the coming years. In this thesis the broad topic of designing a localization approach which leverages the advantages of UAVs for applications related to PSN has been broached. The initial part of the work covered a survey of the possible approaches and existing solutions that are available in the literature. After this initial phase the decision to implement a pure Direction of Arrival (DoA) approach based on an antenna array to be mounted on the flying platform was motivated by the preliminary results obtained through the QuaDRiGa channel model. An approach to DoA estimation based on the Physical Random Access CHannel (PRACH) procedures was generalized and adopted in the scenario of PSN. The impact of the height of the receiving array, as well as its tilt, was investigated and validated through simulations using channel models specifically developed to tackle the scenario of non-terrestrial communications, and the most recent definitions of PRACH waveforms from 5G specifications. The impact of Line of Sight (LoS) condition on the accuracy of the estimated DoA is discussed, and a deep learning method for estimating whether this condition is present or not, which achieves state-of-the-art accuracy, was introduced. Its benefits, in terms of the substantial reduction in computation complexity, as a type of pre-processing for use in our DoA estimator was emphasized. Afterwards, the impact of the trajectory on the expected number of LoS transmissions is discussed and for the typical case of a serpentine search pattern an analytical formula for optimizing its parameters was derived from the LoS probability provided by International Telecommunication Union (ITU) specifications was derived. The results obtained were validated through simulations. Finally, an experimental set-up based on ADALM-PLUTO SDRs was given together with the technical details involved in the acquisition of coherent signals from an array of antennas. The theoretical set-up was validated by an experimental measurement collection campaign where Direction of Arrival measures were performed in a real environment alongside more traditional measures based on signal strength. The simultaneous collection of measures of both methods allows for a direct comparison of their performance since the environment in which they are taken is the same. Results are given for single measure error that show the gains of the proposed DoA method with respect to the more common approach based on signal strength.

Fig. 1
DoA estimation prototype fitted to a Tarot X6 UAV with a 2x2 2.4GHz beamforming array fitted to the bottom.

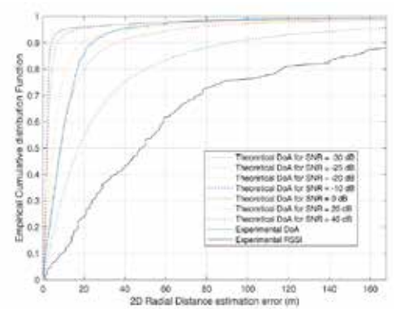

Fig.2
ECDF of the ranging error for theoretical DoA performance at various SNR levels, best experimental DoA results and experimental RSSI results.

SINGLE PHOTON AVALANCHE DIODE ARRAYS FOR QUANTUM APPLICATIONS

Fabio Severini - Supervisor: Prof. Federica Villa

Quantum imaging exploits quantum correlations to image objects at low light (single-photon) regime, with unprecedented vertical dimension sensitivities (a few atomic layers) and very large field-of-view (tens of mm2), surpassing the limits imposed by the laws of classical optics.
Single Photon Avalanche Diodes (SPADs) are the forefront detectors for this application, thanks to their single-photon sensitivity, good detection efficiency, relatively low voltage operation, sharp timing resolution, room temperature operation, and more importantly their compatibility with standard CMOS microelectronic processes, allowing to develop large active area detectors integrated with front-end and processing electronics at a moderately low cost. High performing sensors are obtained, by the monolithic integration of high detection efficiency and low noise SPADs along with their sensing circuits, and suitably designed processing digital electronics.

The main goal of this Ph.D. dissertation is to illustrate the design and characterization of a pioneering SPAD array targeting quantum imaging applications. Extensive literature review revealed the lack of a detector embedding all the required features to be considered for the ideal quantum imager, i.e., with photon coincidence detection capability, high-pixel count, event-driven readout, and spatial resolution, as schematized in Figure 1. With this idea in mind, a 96 × 96 pixel imager based on SPADs was conceived, with a 0.16 µm BCD (Bipolar – CMOS – DMOS) technology, including an analog-based photon coincidence front-end circuit and smart pixels with low noise and high efficiency SPADs and a CAN-bus inspired logic for photon coincidence address readout, as exemplified by the chip architecture in Figure 2. Thanks to this novel approach, the chip can reach a very short readout time, enabling the detection of a very high number of photon coincidences, not limited by framerate anymore, bringing an overall improvement in the effective heralding efficiency within quantum imaging scenarios.
The design of this SPAD imager was achieved within the framework of the European Horizon 2020 FET project "Q-MIC", whose final target was the development of a microscope with unprecedented phase resolution capabilities, by exploiting quantum sources and single-photon detectors. The 96 × 96 SPAD imager has been fully designed and finalized for production however, due to the outbreak of the Covid-19 pandemic that forced the manufacturer to shut down, was not actually realized. As a contingency plan, a conceptually identical, yet smaller array consisting of 24 × 24 pixels was manufactured. The chip has been preliminary characterized through a suitably designed testing module, and successfully employed within a real quantum imaging system, whose results are summarized in Figure 3, effectively enabling high framerate measurements of photon coincidences thanks to its extremely short 330 ns readout time.
A secondary activity of this PhD research was the design of a multi-channel SPAD chip to be integrated along with silicon photonics, as part of the Horizon 2020 FET project "UNIQORN". This project had the objective of developing a Differential Phase Shift – QKD device. One of the main building blocks of the required system is an optical path Quantum Random Number Generator (QRNG), which must also include a detector able to reveal the position of a photon randomly split through waveguides.

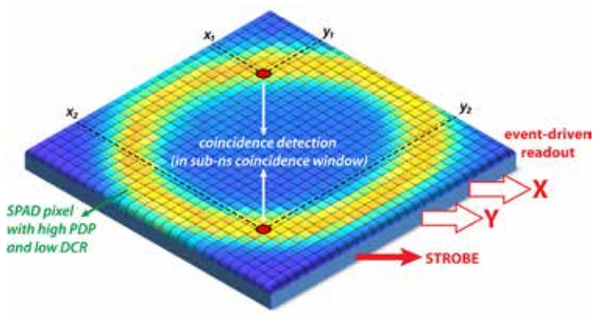

Fig.1
Illustrative example of next generation SPAD detector with on-chip sub-nanosecond coincidence detection, event driven readout of the triggered pixel address and SPAD pixel with high detection efficiency and low DCR

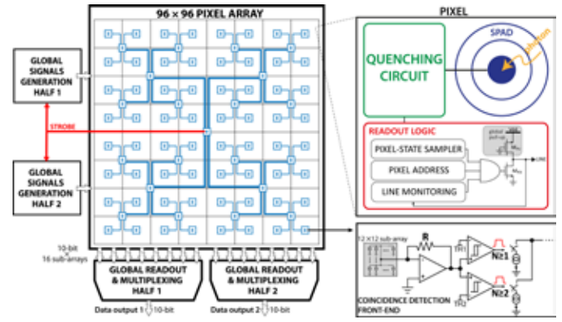

Fig. 2
Complete architecture of the 96 × 96 SPAD imager, including in-pixel logic, coincidence detection front-end and address readout.

Thus, a 32 × 1 linear SPAD array in a 0.16 µm BCD technology was devised and developed, able to generate a raw random number by revealing the position on the array of the single photon impinging on it. This chip has been extensively characterized and its full functionality validated.
At last, a further activity was the testing of previously designed SPAD pixels with extremely short dead times, targeting pulse charge minimization and short dead times, trying to mitigate the trade-off between afterpulsing probability and detector counting rate, thus enabling giga count per second applications. To this end, a testing module was developed and full characterization of the chips performed.
All in all, my Ph.D. research focused on the development and validation of various SPAD arrays implemented in BCD technology, targeting quantum applications. Without doubt, the biggest achievement within my research is the successful proof-of-concept of the novel method for coincidence detection that hopefully will pave the way for a new paradigm in quantum entanglement detection. Immediate steps in respect to this line of research is an extensive characterization of the chip, which was received just in early summer, including testing the detection efficiency improvement, coupling the detector with a microlens array specifically designed for this detector. The chip will also be fully exploited in quantum microscopy applications to image semi-transparent samples. Future steps will include detailed design error acknowledgement, so to fix them and im-prove the overall array performance for the future production of the full 96 × 96 array.

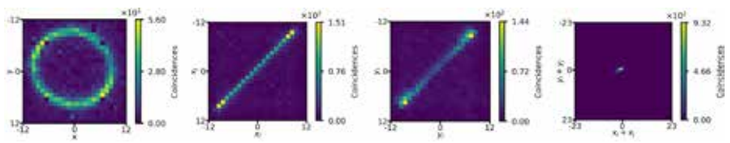

Fig. 3
Measurement results of a flux of entangled photons, including (starting from the left): SPDC ring projection, anti-correlated diagonals due to entanglement in the x, and y position, and sum-projection of the coincidences.

DEVELOPMENTS, COMMISSIONING AND UPGRADE OF THE FARCOS FRONTEND ELECTRONICS

Vincenzo Loris Sicari - Supervisor: Prof. Chiara Guazzoni

In nuclear physics the study of heavy-ion collision at intermediate energies (10-100MeV/u) is the only means available to investigate the properties of nuclear matters under extreme condition and, thus, to better understand the Equation of State (EOS) of nuclear matter. This is fundamental for the study of the properties of stable and unstable (radioactive) nuclei and for the study of the properties of compact astrophysical objects such as neutron stars, core-collapsing supernovae and black-holes formation. The large variety of particles and fragments produced in a single collision and their reciprocal correlation event by event allows quantitative understanding of the reaction dynamics and probing space-time properties of emitting sources. Several experimental techniques and the corresponding detection systems have been developed in order to find the answers to the remaining issue (space-time dynamics of the produced fragments, their thermal properties, internal temperature or spin, etc). The FARCOS system is positioned in this field, with its most promising application in particle - particle collision studies, particularly within the heavy-ion collision nuclear physics experiments. This extended abstract of my PhD thesis as collection of papers shows the work done for the development, commissioning and upgrade of a novel detection system called FARCOS (Femstoscopy AR- ray for COrrelation and Spectroscopy) which is a compact, modular and versatile telescope array made of Double Sided Silicon Strip Detectors and CsI (Tl) scintillators.

Compared to similar detection systems, its increased energy and spatial resolution, wide solid angle coverage and its unique capability to perform pulse shape identification techniques, even in the first detection stage, make it a promising investigation tool for many physical cases. As described in the thesis, it addresses those topics concerning correlation measurements in multi-fragmentation experiments involving stable and radioactive beams, with the peculiar ambition to combine the high granularity with pulse shape techniques to allow also the complete identification even of the lowest energetic particles stopping in the first detection stage. The telescopic unit is the central part of the system. As you can see from the figure, the cluster structure is composed of a first support part for the detectors and a second part regarding the mechanical closure of the telescope. There are three detection stages. The first and second stages are Double Sided Silicon Strip Detectors while the third stage is composed by 4 tronco-pyramidal scintillator crystals arranged in a window shape configuration, each with a front face area equivalent to quarter of the microstrip detectors active area. The telescope's outer walls are made up of motherboard plates. In this way, the reading electronics are very close to the detection stages and furthermore the volume occupied by the telescope is minimized.

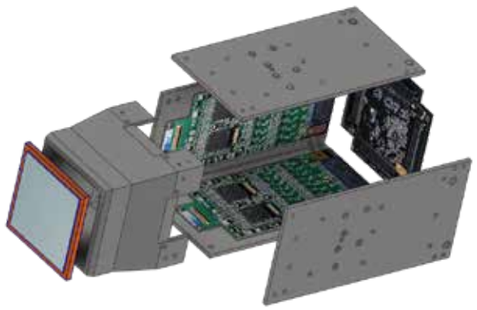

Fig. 1
Exploded view of the single telescopic unit.

During my work as PhD I took care of all the components of the detection system, starting from the design of the telescope unit, briefly introduced, to the management of interconnections, the filtering board and the power distribution. 58 first versions of the motherboard, 8 power air, 3 HV bords, 2 sense boards, 8 AA-DS boards and related ancillary boards and more than 15 patch panels were produced and installed. Another 100 new versions of the motherboard are in production to expand the number of telescopes. Each hardware component has been qualified both in the laboratory and in the experimental room. I also took part in the development of the mechanical structure of the telescopic unit and of the power control distribution panel. Finally, the whole system participated with excellent results in two experiments, CHIFAR and the commissioning of the system, at the experimental chamber of the CHIMERA detector at INFN, LNS in Catania. In the extended abstract there is a compact summary of the work carried out during the PhD, with the aim of drawing a thread between all the publications presented and providing a greater understanding of the results obtained.

Fig.2
Photo of motherboard with ASIC up in sight.

Fig. 3
Photo of ten FARCOS telescope units hanging over the CHIMERA detection system spherical section.

SINGLE-PHOTON AVALANCHE DIODES IN INGAAS/INP AND MICRO-CRYSTAL HETEROSTRUCTURES

Fabio Signorelli - Supervisor: Prof. Alberto Tosi

In the last years, with the so-called on-going "second quantum revolution", single photon detection in the short-wavelength infrared (SWIR) range is receiving more and more interest for promising quantum applications, such as quantum key distribution (QKD), quantum imaging and microscopy, quantum computing, and single-photon source characterization. Moreover, single-photon detection is also needed in applications where very faint optical signals must be measured, like optical testing of VLSI integrated circuits, eye-safe three-dimensional imaging with light detection and ranging (LIDAR) or non-line-of-sight (NLOS) techniques, near-infrared spectroscopy (NIRS) and diffuse correlation spectroscopy (DCS).
Many different single-photon detectors for the SWIR range have been developed, and nowadays the most employed are superconducting nanowire single-photon detectors (SNSPDs) and single-photon avalanche diodes (SPADs). While SNSPDs offer excellent performance at the expense of requiring bulky cryogenic cooling systems, SPADs guarantee good overall performance with the typical advantages of microelectronic detectors, such as reliability, robustness, and compactness, and are thus preferred for many applications. Moreover, SPADs are well suited for making large-format arrays for imaging applications.
The aim of this Ph.D. work is to develop new SPADs for short-wavelength infrared single-photon detection with improved performance, by following two approaches: i) a groundbreaking novel heteroepitaxy is exploited for developing microSPADs based on Si-on-Si or Ge-on-Si structures; ii) a more established, yet under development, InGaAs/InP technology.
Inside the microSPIRE project (that received funding from the European Union's Horizon 2020 research and innovation programme under the FET-OPEN-2016-2017 grant agreement no. 766955), the first microSPADs were designed as proofs of concept of these innovative structures based on micro-crystals operating as SPADs. These micro-crystals are grown employing low-energy plasma-enhanced CVD (LEPECVD), starting from a patterned silicon substrate (see Figure 1).
2D and 3D electrical and optical TCAD simulations were performed, highlighting that a very high PDE (>75% at 600 nm) can be achieved. The experimental characterization of single micro-crystals showed a clear SPAD behavior. However, dark count rate was high even at very low temperatures (<100 K). To solve this issue, devices with implanted p+ top contact have been fabricated and characterized, showing a reduced DCR especially at higher temperatures. Moreover, photoresponsivity of the micro crystals was good (up to 0.1 A/W), despite the top contact was not transparent and photons could enter only from the side. Devices employing a transparent ITO top contact have been fabricated and will be soon measured.

TCAD simulations also for Ge-on-Si micro-crystals were also carried out. Germanium is employed as absorber, in order to extend the sensitivity at longer wavelengths. A first experimental characterization on planar Ge-on-Si SPADs, aimed at assessing the quality of the germanium epitaxy, showed results comparable with state-of-the-art devices. Micro-crystal structures will be soon tested.
Eventually, a preliminary design of a GaAs/AlGaAs quantum-well microSPAD was developed, estimating a PDE between 2.4% and 2.8% for medium-infrared wavelengths.
The second and main activity of my Ph.D. research was the development of InGaAs/InP SPADs, which demonstrated to successfully compare to the best ones ever reported in the literature. The typical structure of these devices is reported in Figure 2. Compared to previous generation devices, the zinc diffusion conditions were optimized to lower the noise of the detector. The shape of the double diffusion has been tailored, as well as the charge layer thickness, to reduce the electric field. A different structure with a thicker absorption layer was also designed, aimed at enhancing the photon detection efficiency at telecom wavelengths.
The developed SPADs achieve low dark count rate, 1 kcps and 4 kcps at 225 K and 5 V excess bias for 10 μm and 25 μm diameter devices, respectively, and <100 cps at 175 K. Both devices also show a high photon detection efficiency, being 33% at 1064 nm, 31% at 1310 nm and 25% at 1550 nm. Timing jitter is comparable to previous-generation devices, being ~100 ps (FWHM) at 5 V excess bias. The efficiency-enhanced detector achieves a photon detection efficiency up to 50% at 1550 nm, with a dark count rate of 20 kcps and a timing jitter of ~70 ps (FWHM) at 225 K. Alternatively, it features a photon detection efficiency of 37% at 1550 nm, with a dark count rate of just 3 kcps and a timing jitter of ~100 ps (FWHM). All the presented devices, when combined with a custom integrated circuit developed in our research group, achieve an afterpulsing probability as low as few percent with a gating frequency of 1 MHz and hold-off time of few microseconds at 225 K, allowing to achieve a photon count rate of almost 1 Mcps. Currently, the design phase for a new production run is in progress. The goal is to lower the noise of the devices (both DCR and afterpulsing) and to further enhance the detection efficiency, especially for quantum applications.
Moreover, a temperature-dependent model for precisely estimating the PDE of InGaAs/InP SPADs is reported. This tool will greatly help the design of future productions, especially for devices aimed at very high PDE. Finally, results on proton-radiation hardness of InGaAs/InP SPADs are presented, together with the first results from laser-annealing treatments to reduce the damages inside the devices.

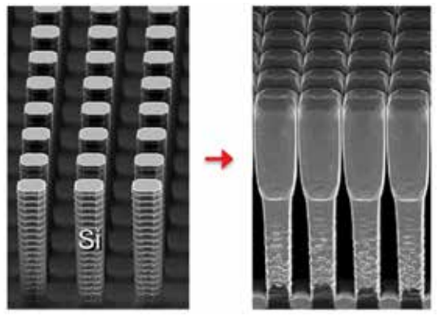

Fig. 1
Micro-crystal growth starting from a patterned substrate.

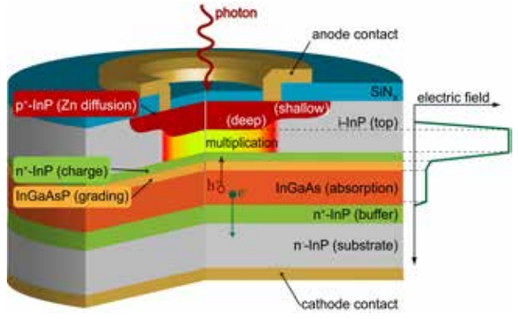

Fig. 2
Typical cross-section of a front-illuminated planar InGaAs/InP SPAD and its electric field profile along the vertical direction.

CONVERSATIONAL AGENTS FOR CHILDREN WITH LANGUAGE IMPAIRMENTS: DESIGN AND TECHNOLOGY

Micol Spitale - Supervisor: Prof. Franca Garzotto

Speech and language impairments are among the most prevalent childhood disabilities worldwide. In 2018, epidemiological studies reported that 10% of pre-schoolers and 5-6% of children and adolescents have shortcomings in language abilities. Those difficulties have been observed in a broad range of Neurodevelopmental Disorders (NDD) that are characterized by developmental deficits in relational, social, communication, academic, emotional, and occupational functioning.

In recent years, the demand for more personalized therapeutic interventions to address the specific needs of children with language impairments has emerged. The current speech-language interventions have several limitations, described as follows:

(a) The assessment tools currently used in clinical and research practice may not be adequate for evaluating language abilities in children with language impairments. In fact, they often lack attractiveness and engagement, resulting in enhancing children's weaknesses rather than their strengths.

(b) World Health Organization has identified significant barriers of costs, long-term engagement, and access for children and their families in speech-language therapy.

(c) American Speech-Language-Hearing Association (ASHA) pinpointed that the goal of language intervention is to stimulate linguistic development, alongside extra-linguistic skills, such as cognitive (e.g., attention) and social aspects of communication. Current linguistic tools in therapeutic practice often lack these joint perspective skills.

To overcome some of the above limitations, past studies have explored the introduction of interactive technologies. Interactive technologies have the potential to attract and engage children with speech and language impairments more than conventional intervention tools, reducing the apprehension caused by human-to-human interaction (a). When interactive technologies are used in speech-language therapy, children seem to effectively acquire language skills faster and retain them longer (a), (c). Moreover, activities delivered via interactive technologies can assure a continuum between clinical practice and the home environment, reducing the costs of long-term speech-language therapy practice (b).

Among those interactive technologies, conversational agents – i.e., software able to understand the user request via natural language, process it, and respond to it - are thought to provide one of the most suitable tools to support linguistic skills development by involving conversational interaction. Those agents provide a more comfortable environment for children with NDD, where the relational and social complexities of the human-to-human interaction are removed.

The current state of the art in conversational technologies for children with language impairments has several open challenges such as:

(1) Providing empirical evidence of the benefits of using conversational agents in therapeutic interventions.

(2) Understanding how user's performance and perception of conversational interactions are affected by different design features, particularly those related to embodiment, i.e., the tangible or visible form for the agent's representation.

(3) Grounding therapy-oriented conversational agents on psycholinguistic theories and approaches underlying the treatment of children with language impairments.

(4) Making these agents scalable and portable to multiple platforms.

This thesis sought to answer the following research question to address those open challenges:

RQ1. Can conversational agents help children with language impairments to improve their linguistic capabilities?

RQ2. What is the most suitable form of embodiment for these agents?

RQ3. What are psycholinguistic methods appropriate to inform their design?

RQ4. How can the development process be improved to facilitate reuse, portability, and scalability?

With this purpose in mind, this research included multiple steps.

First, this thesis systematically reviews the current state of the art on conversational agents for people with NDD. Many authors saw potential in introducing the following conversational agents into therapy: i) intelligent personal assistant (IPA; i.e., software that allows natural language interaction); ii) embodied conversational agents (ECA, i.e., agents on a tablet or computer screens capable of understanding natural language); and iii) socially assistive robots (SAR, i.e., robots that "focused on assisting people through social interaction").
Second, we have run four empirical studies to understand whether conversational agents can support children in improving their linguistic capabilities - highlighting the benefits of those technologies - and which is the most suitable form of the embodiment for a conversational agent into this therapeutic context. This research question is grounded on the embodiment hypothesis stating that physical embodiment has a measurable effect on performance and perception of social interactions; specifically, a robot's physical presence augments its ability to generate richer communication. The first exploratory study presented an innovative decision-making method to rank the conversational agents mentioned above to interact with children with autism (ASD). The second study involved 9 children with ASD and global developmental disorders and is aimed at investigating the introduction of intelligent personal assistants into a therapeutic context. The third study sought to understand the role of the embodiment of conversational agents during the speech-language therapeutic intervention where 17 children with and without language disorders interacted with a physical robot, a virtual agent, and a human counterpart. The last study was an 8-week longitudinal between-subject study (with SAR and ECA conditions) involving 23 children with ASD and developmental language disorders (DLD). It deepened the use of conversational agents in speech-language therapeutic sessions.

Overall, we address the (RQ1) and (RQ2) research questions and we found that the conversational agents embedded into socially assistive robots are the most beneficial during speech-language interventions.

Third, with the collaboration of speech-language therapy experts, we identified a psycholinguistic method that we extended into a set of design patterns, named Activity Patterns, for informing conversational agents that act as therapeutic tutors. Such patterns were defined at three levels of abstractions - activity task, interaction design, and implementation - and they are reusable solutions to facilitate the design and development process. We have also attempted to validate those patterns by running a pilot study with nine users (four computer science and five psycholinguistic professionals) to evaluate the quality of the defined Activity Patterns. The definition of those patterns sought to address the (RQ3) open questions providing the community with a theoretically grounded solution that is commonly understood and accepted by psycholinguistics and developers to facilitate the design and development of the linguistic activities.

Last, we collaborated to design and develop an open-source framework named HARMONI (Human and Robot Modular OpeN Interaction). HARMONI is a modular, composable, pattern-oriented, and robot-agnostic ROS-based framework, enabling the implementation of conversational human-robot interactions independently from the robotic platform. Conversational interactions built using HARMONI take use of several behavioral patterns (including those listed above), are transferable between robotic platforms, and may be readily extended to meet the needs of various therapeutic scenarios. We have evaluated HARMONI running a pilot study with five novel adopters to collect preliminary data to assess the framework usability. With HARMONI we have addressed our last open question (RQ4) providing the community with a comprehensive and open-source tool to implement human-robot conversational interactions.

DYNAMIC SEDIMENT CONNECTIVITY MODELLING FOR STRATEGIC RIVER BASIN PLANNING

Marco Tangi - Supervisor: Prof. Andrea Castelletti - Co-Supervisor: Prof. Simone Bizzi

River sediment (dis)connectivity is a distributed property of fluvial systems, emerging from numerous sediment transport processes across the entire network and their interactions in time and space, and plays a profound influence on river health and human livelihood. However, anthropic activities have profoundly altered sediment transport in river systems: dam construction starves the channel and delta of material, while land-use change alters the characteristics and rate of sediment delivery. The resulting impacts range from delta shrinking and banks instability to ecosystem degradation in the river and on the connected floodplains. Nevertheless, the evaluation of sediment (dis)connectivity degradation by human activities is often performed at the local scale, ignoring the basin-wide implications on the network morphology and the cumulative effects of multiple alterations.

The research in this thesis focuses on developing network-scale sediment (dis)connectivity models for sediment processes characterization and anthropic alteration impact assessment.
First, this work contributes the CASCADE toolbox for sediment transport modelling (www.cascademodel.org), which expands on the original CASCADE (CAtchment Sediment Connectivity and DElivery) model by partitioning sediment transport into distinct grain size classes and including fractional transport formulas. The new structure allows for a more extensive representation on the type and rate of sediment delivery throughout the network. The scarcity of input data for network hydro-morphology definition is compensated by using recently available large-scale datasets and performing extensive sensitivity analysis on the model parametrization. The toolbox is applied on the Vjosa river in Albania to quantify and characterize sediment transport and its influence on river forms stability in a data-scarce environment.

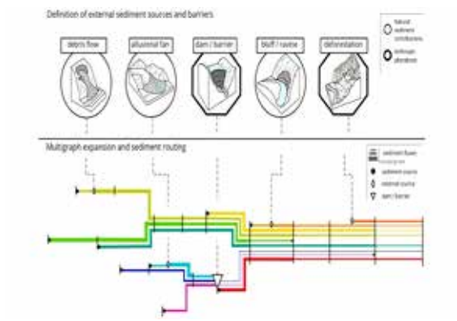

Fig. 1
Representation of the CASCADE framework

A second achievement is the development of D-CASCADE, a dynamic, network-scale sediment (dis) connectivity model. The framework traces sediment delivery and transport patterns across time and space, allowing for a more thorough representation of sediment (dis) connectivity. Add-ons components are integrated in the flexible model structure for detailed representation of channel morphodynamic response to sediment delivery alterations. We tested the D-CASCADE model on the Bega river network, in Australia, to reconstruct historical morphological changes due to human activities across two centuries.

D-CASCADE represents an important tool for strategic and sustainable planning and management of multiple human infrastructures on river systems. In this research, we focused on water and sediment management in reservoirs, and demonstrated the potential of D-CASCADE to quantify the spatio-temporal effects of dam operations, both locally, e.g, reservoir storage losses due to sedimentation, and on the broader river sediment (dis)connectivity.
The model network-scale scope allows for the representation of multiple reservoirs on the same system, and the evaluation of the changes in the cumulative effects on sediment transport given by different design and timing of reservoirs management strategies. These may include both standard operations like daily water release, and exceptional events like drawdown sediment flushing. The D-CASCADE model is tested on the 3S system, a tributary of the Mekong, to evaluate strategic water and sediment management in multi-dam schemes. The effect of reservoir management is explored with D-CASCADE by including different dam development alternative scenarios and assessing daily sediment transport and delivery with specific dam release strategies. Reservoirs features (i.e., Volume, area, height, and sediment deposit) are dynamically modelled via add-ons. Finally, sediment management techniques are introduced by including periodic drawdown flushing in the simulation.

The new models presented in this thesis are designed to be integrated into optimization-based frameworks for strategic reservoir management, to evaluate optimal trade-off between more traditional objectives like hydro-power production and irrigation demand, and the conservation of natural sediment (dis)connectivity.

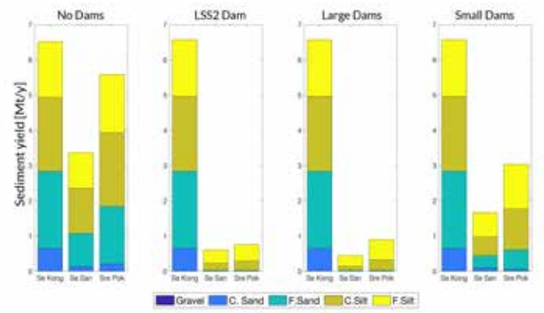

Fig. 2
Mobilized sediment definition in D-CASCADE time loop

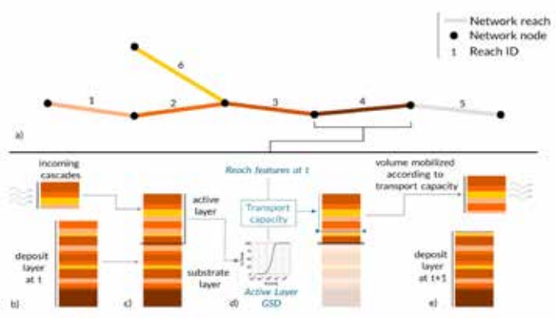

Fig. 3
Sediment delivery to the outlet for the 3S river system with different dam development scenarios

ANALYSIS OF GEOGRAPHIC DATA FOR ENVIRONMENTAL MONITORING

Rocio Nahime Torres - Supervisor: Prof. Piero Fraternali

Environmental monitoring processes and investigations are essential to understand the conditions of the environment and the changes it undergoes (by natural processes or by human interventions) along with their associated impact.
One of the main contemporary problems is waste crimes, in other terms, activities that violate the waste management laws. Only in Italy, according to the Legambiente on Ecomafia reports 34.648 crimes against the environment were committed in 2019 (a 23% more with respect to the previous year). It also reported that 2.4 million tons of waste illegally managed were uncovered.
A particular case of waste crime is that of illegal waste dumping, which threatens the environment and public safety and health. Discovering them as early as possible is essential for preventing hazards such as fire pollution and leakage.
Before the digital era, the only means to detect illegal waste dumps was the on-site inspection of potentially suspicious sites, a procedure extremely costly and impossible to scale to a vast territory. With the advent of Earth Observation technology, scanning the territory via aerial images has become possible. However, manual image interpretation remains a complex and time-consuming task that requires expert skills.
This research aims to exploit Artificial Intelligence methods and remote sensing imagery to embed expert knowledge within data-driven classifiers that can help partially automate the photo interpretation process. For such purposes, methods have been selected to train different Convolutional Neural Network (CNN) classifiers, and tools to evaluate and use them have been proposed.
The thesis summarizes the state of the art on waste detection problem and organize the many heterogeneous approaches proposed in the literature by several characteristic dimensions.
Quite surprisingly, and despite the great success of the CNN-powered classification of aerial images in many challenging domains, to the best of our knowledge, only very few previous works have quantified the utility of CNN architectures for waste dumps detection.
Unlike previous works, the illegal landfill detection problem is formulated as a remote sensing scene classification task. RS scene classification categorizes the content of aerial images into semantic classes based on the spatial arrangement and the structural patterns of the ground objects. The scenes present variations of the type of garbage present (plastics, tires, wood, building material), of its disposition (scattered, collected in dumpsters, trucks, or sheds) as well as to the different geographical contexts (e.g., urban, rural). The imagery and the coordinates of candidate sites were provided by the Environmental Protection Agency of Region of Lombardy (ARPA) in the context of a collaboration established for the SAVAGER project.
To cope with the complexity of illegal landfill imagery, in which the recognition of the relevant scenes might need a varying degree of context (e.g., garbage stored in dumpsters vs. scattered in a large area), we apply a multi-scale CNN architecture normally employed in complex scene detection tasks. The method is tested on a large-scale territory and the classifier's output is evaluated quantitatively and qualitatively by exploiting visual understanding and interpretability techniques (specifically Class Attention Maps). Figure 1 illustrates examples of scenes containing illegal waste dumps that were correctly classified by the model. To allow analyst to employ the model results on new territories an inspection tool prototype was created. Such tool was used by ARPA analysts to further validate the model's output on a new area providing satisfactory results.

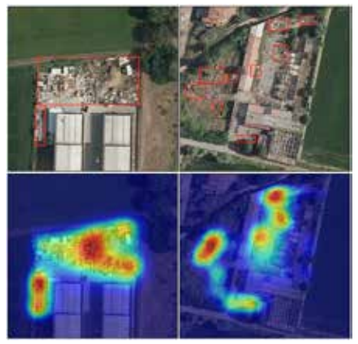

Fig. 1
Examples of sites correctly classified as positive with high confidence. In each column, the top image is the input sample with one or more manually created boxes surrounding the areas with waste. The center image shows the heat map derived from the CAM of the positive class.

In addition to the development of a RS scene classifier for waste detection, we also address the issue of how to analyze its performances and more generally the performances of DL and CV components. A review of the state-of-the-art metrics and tools for the evaluation of ML tasks reveals the opportunity of implementing an innovative evaluation framework for inspecting the performances of DL and CV models at greater depth than commonly done in standard benchmarks to better highlight the weakness and strengths of a model. We illustrate the development of such a framework, called ODIN, which integrates the most widely used ML evaluation metrics, custom analyses and reports based on meta-annotations (i.e., extra annotations not used for training), and a variety of utilities to create, modify, and visualize a data set. The framework is publicly released and generalizes to different computer vision tasks.
Finally, Figure 2 summarizes the artifacts implemented in this thesis: 1) The CNN model to identify illegal landfills scenes, 2) an inspection tool prototype to provide analysts with the models results for the analysis of new territories and 3) ODIN framework, for model evaluation. The output of the pipeline is a database of suspicious locations identified by the analysis during the inspection process. Such locations are later submitted to a surveillance process that could end on site interventions (e.g., site remediation, legal actions).

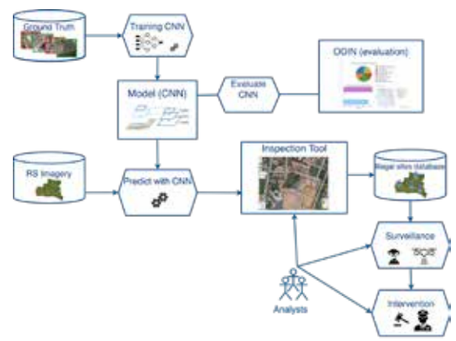

Fig. 2
Pipeline and artifacts proposed

A HOLISTIC APPROACH TOWARDS FUTURE SELF-TUNING APPLICATIONS IN HOMOGENEOUS AND HETEROGENEOUS ARCHITECTURES

Emanuele Vitali - Supervisor: Prof. Gianluca Palermo

With the beginning of the dark silicon era, application optimization, even with the exploitation of heterogeneity, has become an important topic of research. One methodology to obtain optimized applications for different architectures is application autotuning.

Indeed, applications can obtain the same result with different codes. However, different codes have different extra-functional properties, such as execution time or energy consumption which may change across different architectures. To obtain the best, application autotuning techniques have been proposed in literature. It is very difficult for the original application developer to select the best configuration that can enforce the constraints across different machines, with unknown input and varying configurations.

Given this background, I envision future applications not as monolithic code but as a sequence of self-tuning modules, which are capable of autotuning themselves and can exploit platform heterogeneity. In my thesis I collected a set of methodologies that were developed during my Ph.D. which aim at giving the programmers ways to create these self-tuning modules.

This module is the key component of future applications. The original application is enriched with several components, thus becoming able to perform self-management during its runtime. We can notice from the picture that there are two intermediate steps in the process of obtaining the final adaptive binary (which is the self-tuning module runtime implementation). All these phases are important during the creation of the self-tuning module. In the first step the original code is enhanced and the adaptivity is inserted in the application. This has to be done only once in the lifetime of the application, and it usually requires an effort from the application developer (for example to expose parameters or to insert heterogeneous kernels). In the second step the application is deployed on a specific hardware platform, and its behavior is explored. This leads to the construction of an application knowledge, that is needed to perform the autotuning. Finally, the application is built with both autotuning capabilities and the necessary knowledge, and it is able during the runtime to always select the best configuration, even in presence of changes in platform condition, requirements, constraints, or input values.

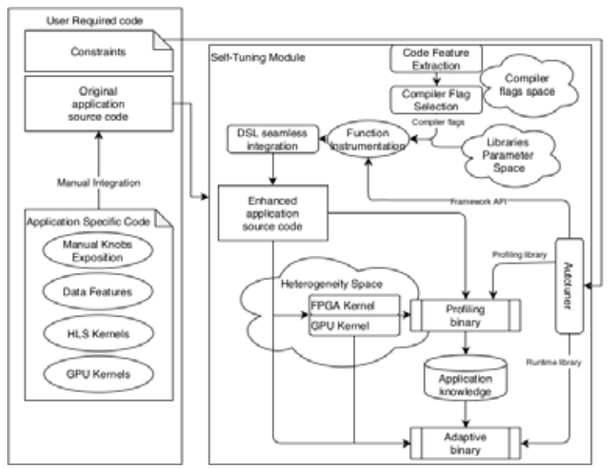

Fig. 1
The self-tuning module and the code required to generate it.

To validate the concept of self-tuning modules, we have developed several techniques that enhance a target application with autotuning capabilities. Since most of them are strongly tailored to an application, we have not implemented a single framework that can automatically insert the techniques into a code. However, the methodology is general, and it can be easily ported to different applications.

In particular, the contributions are the following:
A framework to automatically tune compiler flags or library parameters at function level. The framework exploits different tools (mARGOt, COBAYN, LARA, MilepostGCC) in a joint effort to ease the programmer job and automatically and seamlessly obtain the best possible configuration according to the underlying architecture for every different hotspot kernel in the code.

Analysis of applications to find and expose autotuning possibilities in a reactive way. Applications have been made capable to react to changes in the underlying configurations or changing requirements provided by the user. This work has been done in the context of the object detection challenge.

A methodology to proactively autotune applications according to the input data. The objective of this methodology is to increase computation efficiency and thus save energy and time. This work has been done on an HPC application, tackling the adaptive routing problem in the context of a smart city.

Creation of a parametric library for the synthesis of long unsigned multipliers on FPGA. The library allows the programmers to create hardware accelerators to perform this important operation with a focus on parametrization and the offering of several trade-offs in performance and cost.
But the most important contribution has been done on a molecular docking application called GeoDock.
GeoDock is a component of LiGen, which is itself a component of the EXSCALATE tool flow. This tool-flow is the in-silico section of a real drug discovery pipeline. The goal of a drug discovery process is to find new drugs starting from a huge exploration space of candidate molecules. GeoDock aims at estimating the three-dimensional pose of a given molecule, named ligand when it interacts with the target protein. The ligand is much smaller than the target protein. For this reason, we only consider a region of the protein, called pocket (or binding site). The pocket is an active region of the protein where it is likely that an external small molecule can interact. On this application, we applied several techniques that led to a dramatic increase of performance through the introduction of autotuning, approximation and heterogeneity in the application.

Initially, the application was a CPU only monolithic application. In the first phase, we introduce flexibility and autotuning capabilities, making the application adaptive and able to trade accuracy with performances according to the user requests, thus allowing to respect requirements on time to solution.

Later, we introduced heterogeneity porting the main kernels to the GPU using OpenACC and optimized the porting on the given architecture obtaining a speedup of 6x.
Then we studied the distribution of the computations across the CPU and the GPU, and we decided to map the kernels on the most suitable hardware according to the characteristics of the kernel itself. This created a speedup of 1.25x using the same resources, just by selecting the optimal hardware to compute the different phases of the application.

Finally, we rewrote the application using CUDA and contextually introduced proactive autotuning using characteristics of the input to select some runtime parameters. The porting granted a speedup of 3x, while the autotuning has an impact of up to 1.6x but is strongly dependent from the input.

Overall, also considering the improvement due to the different hardware used, the throughput of the application grew by two orders of magnitude, from tens of ligands per second per node to more than 1500. This growth in performance enabled the largest virtual screening campaign, performed on over 70 billion ligands on 15 different protein pockets in the context of the EXSCALATE4COV European project. The experiment ran on two entire supercomputers (at the time the two largest European supercomputers) for 60 hours.

AUGMENTING TRADERS WITH LEARNING MACHINES

Edoardo Vittori - Supervisor: Prof. Marcello Restelli

The financial markets are comprised of several participants with diverse roles and objectives. Asset management firms optimize the portfolios of pension funds, institutions, and private individuals; market makers offer liquidity by continuously pricing and hedging their risks; proprietary traders invest their own capital with sophisticated methodologies. The approaches adopted by these actors are either manual or expert systems that rely on the experience of traders, and thus are subject to human bias and error. This dissertation proposes innovative techniques to address the limitations of the current trading strategies. Specifically, we explore the use of algorithms capable of autonomously learning the aforementioned sequential decision-making processes. The development of these algorithms entails a careful reproduction of realistic environments, as well as the observance of trading objectives, i.e., maximizing returns while maintaining a low risk profile and minimizing costs. These algorithms all share a common core structure, that is making a trading decision conditional on the current state of the financial markets. Our main theoretical and algorithmic contributions include the extension of the online learning field, as we introduce transaction costs and conservativeness in online portfolio optimization, and the enhancement of Monte Carlo Tree Search (MCTS) algorithms to account for the stochasticity and high noise typical of the financial markets. In terms of experimental contributions, we apply Reinforcement Learning (RL) to learn profitable quantitative trading strategies and option hedging approaches superior to the standard Black & Scholes hedge. We also find that Reinforcement Learning combined with Mean Field Games (MFGs) enables the development of competitive bond market making strategies. Finally, we demonstrate that dynamic optimal execution methods can be learned through Thompson Sampling (TS) with Reinforcement Learning. The use of such advanced techniques in a production environment may allow the achievement of a competitive advantage that will translate into economic benefits. The dissertation has five main chapters.

Online Portfolio Optimization
The chapter on Online Portfolio Optimization (OPO) examines two characteristics of the OPO framework, dividing the chapter in two parts. The first part focuses on controlling transaction costs in the OPO problem by defining a novel algorithm, namely Online Gradient Descent with Momentum (OGDM). We prove that OGDM can achieve sublinear regret. We then verify the analytical results through an extensive experimental campaign comparing with state-of-the-art OPO algorithms.

The second part focuses on the problem of conservative optimization defining a novel algorithm, namely Conservative Projection (CP). CP is a "wrapper" that can be applied to any existing Online Convex Optimization (OCO) algorithm and maintains the same regret order of the OCO algorithm it uses as a subroutine, while satisfying the conservativeness property. We confirm the theoretical results through the experiments on the OPO framework.

Quantitative Trading
The chapter on quantitative trading proposes two methodologies to create a quantitative trading strategy, specifically, using Fitted Q-Iteration (FQI) and using MCTS. The first part concentrates on the use of FQI and is mostly experimental. We compare two different scenarios, a multi-currency framework with EURUSD and USDGBP and a single currency framework considering the FX pairs individually. We observe experimentally the behavior of the resulting trading policy, by changing FQI parameters (min-split and number of training iterations), and action persistence (1-minute, 5-minute, and 10-minute). The results show that the 5-minute persistence achieves a superior performance. Furthermore, the multi-currency setting surpasses the single currency cases, this is expected as it may exploit additional trading opportunities.

In the second part, we propose the use of MCTS for trading. We coin a new algorithm, namely QL-OL UCT, that uses open loop planning to handle the continuous state space and stochastic transition model and a Q-learning backup operator to reduce the noise of the backups. We also introduce a novel generative model that uses historical data and a clustering approach. We test the algorithm on the EURUSD FX pair, concentrating on optimizing parameters. While being profitable without considering transaction costs, adding these costs causes the agents to decide not to trade.

Bond Market Making
This chapter presents a novel solution to the problem of market making in dealer markets. We define the framework as an N-player stochastic game and propose a solution using MFGs and RL. This approach can handle the competitive nature of the problem, by learning an equilibrium strategy in a multi-agent framework. The experimental evaluation shows that, in the presence of strategic opponents, the method outperforms other benchmark agents. Instead, in the presence of a generic market configuration, other agents might perform better when considering P&L. However, since the proposed methods guarantee a safe behavior, they can achieve a lower risk in terms of a higher Sharpe ratio and a smaller inventory.

Option Hedging
In this chapter, we learn how to use a risk-averse RL algorithm, namely Trust Region Volatility Optimization (TRVO), to optimally hedge the delta risk of options. This chapter is composed of two parts, the first on hedging equity options and the second on hedging credit index options. In both cases, the focus is on the experimental campaign, which shows that, without considering costs, the TRVO policy learns to reproduce the delta hedge. With transaction costs, the mean-volatility objective considered makes it possible to balance risk and return, by deciding the agent's level of risk aversion. We notice that, when increasing risk aversion, the policy approaches that of the delta hedge, while when decreasing risk aversion, the policy becomes smoother and reduces the transaction costs. Furthermore, we learn that the policies are robust, as the agents can efficiently hedge options with different characteristics or markets that behave differently than those used in training. Additionally, we obtain positive results when testing on an underlying with a Heston process, and, finally, also with real market data.

Optimal Execution
The work presented in this chapter considers the optimal execution problem using RL. We use the multi-agent market simulator ABIDES, which creates a multi-agent simulation of the financial markets, enabling a realistic market impact. We propose FQI to learn a policy on two market scenarios, one with high volatility and high liquidity, and the other with low volatility and low liquidity, and used TS to select in an online manner the optimal policy to use for execution. The experimental results show that the FQI approach is superior to benchmarks such as TWAP and Almgren-Chriss, in the case of high volatility and high liquidity. With low volatility all the approaches obtain similar results. Furthermore, in the high volatility context, it is possible to select the optimal policy after a small number of TS iterations.

OPTIMIZATION FRAMEWORK FOR RESOURCE MANAGEMENT OF MOBILE EDGE COMPUTING NETWORKS

Bin Xiang – Supervisor: Prof. Elisabetta Di Nitto

The fifth generation (5G) and beyond mobile networks aim at satisfying, in different demanding application scenarios, stringent Quality of Service (QoS) requirements, among which latency is one of the key metrics that mobile operators are supposed to optimize for mobile users. Mobile Edge Computing (MEC), an essential technique utilized in 5G networks, brings cloud computing capabilities to the edge of the mobile networks, especially in close proximity to mobile users, making it possible to simultaneously address the stringent latency requirements of critical services and ensure highly efficient network operation and service delivery, so as to improve user experience.

MEC services, on one hand, require significant investments from both network operators and service providers in terms of deploying, operating and managing edge clouds, and on the other, provide limited computational and storage resources by design as data centers are deployed in the core of the network. During peak hours, the operator must serve a large number of tasks from users with high demands, hence the latency requirements of different services can hardly be guaranteed. This issue can be tackled by massively deployed edge clouds that are attached to the base stations and connected to each other in a specific topology, as ultra-dense 5G-and-Beyond networks are built. In this way, the resource limitations can be solved through sharing computational and storage capabilities among multiple MEC units nearby.

In this thesis, we leverage cooperation among interconnected multiple MEC units and investigate joint resource optimization considering multiple aspects of network operations, with the target of enhancing the utilization efficiency of resources to further satisfy improved QoS and reduce network operation cost. Specifically, aggregated mobile traffic and user requests are considered based on their types (e.g., voice, video, web, game, etc.) that are associated with different QoS requirements. We jointly optimize 1) where to process the traffic and requests, 2) how to route network flows and 3) how to allocate and schedule the required resources with regard to communication, computation and storage.

Firstly, to serve mobile traffic, we investigate in the context of hierarchical edge networks and propose a mathematical optimization model to perform a joint slicing of mobile network and edge computation resources. The optimization aims at minimizing the total traffic latency considering operations including transmitting, outsourcing and processing user traffic, under the constraint of user tolerable latency for each class of traffic.

Then, we release the constraints on hierarchical network and fixed computation capacity, and further take into account the overall budget of operators to plan and allocate the computation capabilities in edge network with an arbitrary topology. The main objective, aligned with the first one, is to further operate cost-efficient edge networks through jointly planning the availability of computational resources at the edge, slicing mobile network and edge computation resources, and routing heterogeneous traffic types to various slices. We propose an optimization model to minimize the network operation cost and the total traffic latency in the procedures of transmitting, outsourcing and processing user traffic, under the constraint of user maximum tolerable latency for each class of traffic.

Finally, we focus on serving multiple classes of user requests (with starting time, deadline and duration) which require bandwidth, storage and computation resources. The main objective is to exploit the flexibility of services to requests by shifting the starting time without penalizing utility perceived by users, while, in the meantime, permitting efficient resource utilization. We propose an optimization framework that jointly considers several key aspects of the resource allocation problem, specifically, through optimizing: admission decision, scheduling of admitted requests (also called calendaring), routing of these flows, the decision of which nodes will serve such requests, as well as the amount of processing and storage capacity reserved on the chosen nodes, with the objective of maximizing the operator's profit.

The above proposed optimization models are first formulated as mixed-integer nonlinear programming (MINLP) problems, which are NP-hard. To tackle them efficiently, we perform equivalent reformulations from MINLP to mixed-integer quadratically constraint programming (MIQCP), and based on that, further propose effective heuristics to facilitate the solutions of the problems, i.e., Sequential Fixing (SF) for the edge slicing model, Neighbor Exploration and Sequential Fixing (NESF) for the edge planning model and Sequential Fixing and Scheduling (SFS) as well as a distributed algorithm based on Alternating Direction Method of Multipliers (ADMM) for the edge calendaring model. We evaluate the performance of our proposed models and heuristics in real-size network scenarios including both random geometric graphs and a realistic mobile network topology, showing the impact of all the considered parameters (i.e., different types of user traffic or requests, tolerable latency, network topology and bandwidth, computation, storage, and link capacities) on both the optimal and approximate solutions. Results obtained demonstrate that near-optimal resource allocation solutions can be achieved by our heuristics in short computing time.