Department
of Excellence
2018 - 2022

EMbeDS
Economics and Management
in the era of Data Science

WORKSHOP

4 - 5 July 2022

Sant’Anna School of Advanced Studies

BOOK OF ABSTRACTS

JULY 4, 2022

Session I (Chair: Andrea Vandin)

Computational approaches for process-oriented data analysis, simulations of large agent-based models, and statistical model checking

❖ Adelinde Uhrmacher, Rostock University, Germany.

Title: Valid agent-based models - requirements and implications.

Abstract: Agent-based simulation studies have become a widely accepted tool in the social sciences over the last decades. Despite its success, the credibility crisis that simulation studies faces has not left the field untouched and appears aggravated by the very property that makes agent-based modeling so attractive, i.e. its flexibility. In these discussions, the validity of a simulation model looms large. To answer the question whether the right model has been built, in addition to model inspection, conducting diverse simulation experiments is essential. The talk will illuminate requirements and methodological developments in support of those from a computer science perspective, these include formal domain-specific modeling languages, model-based approaches for executing experiments, and exploiting provenance for making their role in developing the simulation model explicit.

❖ Andrea Burattin, DTU Technical University of Denmark.

Title: Process-oriented Data Science: a gentle overview on Process Mining.

Abstract: Process mining is a family of techniques that allows the analysis of event logs. Among these techniques, control-flow discovery algorithms allow the discovery of in-vivo processes, i.e., conceptual models of the processes as they are observed in reality, as opposed to expected ones. Additionally, conformance checking tries to understand the extent to which a normative model is violated in actual executions. In this presentation, the basics of these techniques will be presented and presented.

❖ Andrea Vandin, Sant'Anna School of Advanced Studies, Italy.

Title: Simulation models meet Process Mining: a Novel Approach to Model Validation and Enhancement.

Abstract: The previous two talks present results from the so-called Simulation and Process Mining communities, respectively. Here, we present recent results on the integration of results from the two communities, applied to domains of interest for EMbeDS.

We will discuss a novel research line integrating Statistical Model Checking (SMC), a family of simulation-based analysis techniques, with Process Mining (PM). SMC and PM are complementary: SMC focuses on performing the *right* number of simulations to obtain statistically-reliable estimations (e.g., the average GDP of an economy). PM focuses on reconstructing a model of a system using logs of its traces.

Nevertheless, both SMC and PM aim at providing evidence of issues/guarantees of a system. We will show how to enrich SMC via PM-based *explanations* of its results, fostering model validation.

Session II (Chair: Daniele Giachini)

Economic theory and learning processes

❖ Jonathan Newton, Kyoto University, Japan.

Title: Asymmetric behavior and long run outcomes

Abstract: Strategies of players in a population are updated according to the behavioural rules of agents, where each agent is a player or a coalition of players. We discuss recent advances in this area that focus on the idea of a behavioural rule being asymmetric towards an action. Specifically, if all behavioural rules satisfy an asymmetry property, then robust predictions can be made about long run behaviour. Potential applications include a wide variety of popular behavioural rules. For example, the logit choice rule combined with the possibility of introspection regarding the actions of other players.

❖ Filippo Massari, University of Bologna, Italy.

Title: Good biases bad biases

Abstract: A growing literature in behavioral economics studies ways in which individuals' learning departs from Bayesian updating. In well-specified learning problems, where it is impossible to improve upon Bayes' rule, these departures were considered to be either transitory or detrimental to the individuals. As the interest in misspecified learning environments grew in the economics community a different point of view has emerged: these departures from Bayes' might represent a robust heuristic to cope with complex environments rather than detrimental mistakes. This point of view leads to the hypothesis that biases that provide an evolutionary advantage should be the ones that are the most prevalent in the population. Here we propose a general learning formula that accommodates several known learning biases and characterizes their individual and combined effect on the long-run survival of individuals with identical information but different biases. We use long-run survival to broadly distinguish between those biases that are detrimental to individuals, bad bias, and those that provide an evolutionary advantage to the individuals because are more robust than Bayes' rule to model misspecification error, good bias. We find that under-inference, under-reaction, and bias toward the center

of the simplex, in any combination, are good biases; conversely, bad biases are over-inference, over-reaction, and bias toward the boundary in any combination. And that the positive/negative effect of different biases reinforce each other. Our findings support the evolutionary hypothesis that the most prevalent biases found in experimental economics are those that provide an evolutionary advantage.

❖ Pietro Dindo, Cà Foscari University of Venice, Italy.

Title: Learning Models from Prices

Abstract: We study a competitive financial economy where agents have heterogeneous, possibly endogenous, beliefs on the exogenous state process, and trade on these differences. In particular, we allow beliefs to come from the combination of two types of models: an exogenous agent-specific model, which may depend on the history of states, and an endogenous market model, which predicts using an appropriate transformation of state prices. First, we show that when an agent learns on the two models using Bayes rule, she never vanishes and her beliefs, as well as the market model, converge to the most accurate exogenous model. Then, we depart from Bayesian learning and characterize agents' accuracy and survival when the weight given to the endogenous model is fixed. We provide conditions such that both individual and market accuracy improve and show that they can even exceed that of the most accurate exogenous model. Finally, we discuss how our results depend on the choice of the endogenous market model and agents' risk preferences.

❖ Giulia Livieri, Scuola Normale Superiore, Italy.

Title: Analysis of Bank Leverage Via Dynamical Systems and Deep Neural Networks

Abstract: We consider a model of a simple financial system consisting of a leveraged investor that invests in a risky asset and manages risk by using Value-at-Risk (VaR). The VaR is estimated by using past data via an adaptive expectation scheme. We show that the leverage dynamics can be described by a dynamical system of slow-fast type associated with a unimodal map on [0,1] with an additive heteroscedastic noise whose variance is related to the portfolio rebalancing frequency to target leverage. In absence of noise the model is purely deterministic and the parameter space splits in two regions: (i) a region with a globally attracting fixed point or a 2-cycle; (ii) a dynamical core region, where the map could exhibit chaotic behavior. Whenever the model is randomly perturbed, we prove the existence of a unique stationary density with bounded variation, the stochastic stability of the process and the almost certain existence and continuity of the Lyapunov exponent for the stationary measure. We then use deep neural networks to estimate map parameters from a short time series. Using this method, we estimate the model in a large dataset of US commercial banks over the period 2001-2014. We find that the parameters of a substantial fraction of banks lie in the dynamical core, and their leverage time series are consistent with a chaotic behavior. We also present evidence that the time series of the leverage of large banks tend to exhibit chaoticity more frequently than those of small banks.

Session III

Panel: The future of AI research (joint with the Summer School of the PhD AI & Society; Chair: Anna Monreale)

- ❖ Francesca Chiaromonte, Sant'Anna School of Advanced Studies, Italy.
- ❖ Rita Cucchiara, University of Modena and Reggio Emilia, Italy.
- ❖ Fosca Giannotti, Scuola Normale Superiore, Italy.
- ❖ Michela Milano, University of Bologna, Italy.
- ❖ Mariarosaria Taddeo, Oxford University, UK.

Over the last decade, AI researchers have made groundbreaking progress in hard longstanding problems related to machine learning, computer vision, speech recognition, and autonomous systems. Despite this success of AI, its adoption so far is mostly in low-risk applications, while the uptake in medium/high-risk applications, which might have a deeper transformative impact on our society, such as in healthcare, public administration, safety-critical industry etc., is still low compared to expectations. The reasons for such lagging are profound. Adoption barriers include perceived challenges to the autonomy and the oversight capacity of human users, the required effort, dissatisfaction with user interfaces and, above all, trust concerns related to poor users' knowledge about the assumptions, limitations and capabilities of AI systems.

Well beyond currently available technologies, we need AI systems capable of interacting and collaborating with humans, of perceiving and acting within evolving contexts, of being aware of their own limitations and able to adapt to new situations, and interact appropriately in complex social settings, of being aware of their perimeters of security and trust, and of being attentive to the environmental and social impact that their implementation and execution may entail. In short, we need AI that does not yet exist.

The panelists will address, form their unique perspectives, the question: along which lines will the future of AI research unfold?

Session IV (Chair: Francesca Chiaromonte)

Keynote lecture

- ❖ Marzia Cremona, Université Laval, Canada.

Title: Local clustering and motif discovery of functional data: applications in "omics", biomedical sciences and finance

Abstract: Recent evolution in data acquisition technologies enabled the generation of high-dimensional, complex data in several research areas – in the sciences and engineering, among other disciplines. Increasingly sophisticated statistical and computational methods are needed in order to analyze these data. Functional data analysis (FDA) can be broadly employed to analyze functional data, i.e. data that vary over a continuum and can be naturally viewed as smooth curves or surfaces, exploiting information in their shapes.

In this talk, I will present probabilistic K-mean with local alignment (probKMA), a novel FDA method to locally cluster a set of misaligned curves and to address the problem of discovering functional motifs, i.e. typical "shapes" that may recur several times along and across a set of curves, capturing important local characteristics of these curves.

After demonstrating the performance of the method on simulated data and showing how it generalizes other clustering methods for functional data, I will present three applications to the analysis of functional data from different fields. First, I will apply probKMA to discover functional motifs in "Omics" signals related to mutagenesis and genome dynamics. Second, I will employ probKMA as a probabilistic clustering method to group COVID-19 death curves of the different Italian regions during the first wave of the pandemic. Finally, I will present an application to the discovery and characterization of financial bubbles in price time series.

JULY 5, 2022

Session V (Chair: Andrea Mina)

Science Technology and Innovation

- Gaétan de Rassenfosse, Ecole Polytechnique Fédérale de Lausanne, Switzerland.

Title: The commercialization of DoD-SBIR patents: A counterfactual analysis

Abstract: The paper proposes a novel, web-based approach to innovation policy evaluation. The approach overcomes several limitations affecting established evaluation methods used in the literature. We implement it to study the impact of the U.S. DoD-SBIR program on technology commercialization. We start by identifying the universe of USPTO patents that acknowledge support by the SBIR program. We then track whether these patents are mentioned in relation to commercial products in a virtual patent marking page available on the recipient's website. We interpret the latter event as signal of commercialization. Finally, we create a group of suitable control patents and we compare the commercialization probability of SBIR-funded and control inventions. The results support the view that the SBIR program is quite effective at stimulating the commercialization of federally-funded scientific discoveries.

- Martina Iori, Sant'Anna School of Advanced Studies, Italy; and Jacopo Di Iorio, Penn State University, USA.

Title: Local economies and specialization trajectories: A dynamic analysis from a Complexity Economics perspective.

Abstract: The Economic Complexity Index (ECI), introduced in 2009 in a seminal paper by Hidalgo and Haussman, has shown significant applications in Economics. Despite its ability in revealing the most important economic factors for countries' competitiveness and summarizing complex interactions among countries and their economic activities (exports by product; patents by technological domain; etc.), the capacity of the index to provide insights into the specialization dynamics of each economy is limited. In this paper, in line with recent evolutions in the spectral-clustering literature, we propose a new interpretation of the ECI that eases its application to describe the evolution of countries' production and innovation systems. This new perspective offers two main advantages. First, it makes explicit that ECI creates, simultaneously, interrelated clusters of countries and products. Second, it provides a tool for a dynamic analysis of country specialization trajectories. Since our approach assigns, to country-product couples, probabilities of being part of a cluster, we can detect common specialization patterns and the evolution of product-country clusters by observing these probabilities over time. In the paper, we show

how, in this perspective, ECI is a powerful policy tool to reveal which economic activities contribute more to the evolution of a local economy’s competitiveness.

❖ Giorgio Tripodi, Sant’Anna School of Advanced Studies, Italy.

Title: Quantifying knowledge spillovers from advances in negative emissions technologies

Abstract: Negative emissions technologies (NETs) feature prominently in most scenarios that halt climate change and deliver on the Paris Agreement's temperature goal. As of today, however, their maturity and desirability are highly debated. Since the social value of new technologies depends on how novel knowledge fuels practical solutions, we take an innovation network perspective to quantify the multidimensional nature of knowledge spillovers generated by twenty years of research in NETs. In particular, we evaluate the likelihood that scientific advances across eight NET domains stimulate (i) further production of knowledge, (ii) technological innovation, and (iii) policy discussion. Taking as counterfactual scientific advances not related to NETs, we show that NETs-related research generates overall significant, positive knowledge spillovers within science and from science to technology and policy. At the same time, stark differences exist across carbon removal solutions. For example, the ability to turn scientific advances in NETs into technological developments is a nearly exclusively feature of Direct Air Capture (DAC), while Bio-energy with Carbon Capture and Storage (BECCS) lags behind. Conversely, BECCS and Blue Carbon (BC) have gained relative momentum in the policy and public debate, vis-a-vis limited spillovers from advances in DAC to policy. Moreover, both scientific advances and collaborations cluster geographically by type of NET, which might affect large-scale diffusion. Finally, our results suggest the existence of coordination gaps between NET-related science, technology, and policy.

Session VI (Chair: Francesco Lamperti)

Climate risk and economic dynamics

❖ Antoine Mandel, Sorbonne and PSE, France.

Title: Modelling climate-related financial risks

Abstract: The aim of the talk is to illustrate how climate policy and climate impacts have been approached in the financial sphere. We will first provide a brief review of the literature in climate finance. Then, we will present modelling approaches to assess financial risks related to mitigation, i.e; transition risk, and climate impacts, i.e. physical risks. Finally, we will illustrate both approaches with empirical applications.

❖ Maximilian Kotz, Potsdam Institute for Climate Impact Research (PIK), Germany.

Title: Macroeconomic climate damages - resolving the locality of climate impacts in time and space

Abstract: Understanding the macroeconomic damages from future climate change is critical to guiding optimal policy regarding adaptation and mitigation. Econometric assessments of climate impacts can provide an empirical basis to ground such assessments, but have been limited by approaches which aggregate impacts to national and annual scales. Climate impacts can occur locally in both time and space, demanding an analysis with higher temporal and spatial detail to resolve both complex societal impact channels and anthropogenically forced climate changes.

In this talk I will present recent advances and on-going work in this area, the result of combining high resolution climate and economic data with mathematical re-formulations of climate impacts. These empirical assessments have identified novel impact channels in the climate-economy relationship, for example from daily temperature variability and extreme daily rainfall. I will discuss these results, and their implications for the scale and distribution of macroeconomic damages under future climate change.

- Matteo Coronese, Sant'Anna School of Advanced Studies, Italy.

Title: Who will carry the climate burden? Heterogeneous impacts of climate anomalies on income classes and sectors.

Abstract: Climate anomalies have been convincingly shown to have adverse aggregate economic impacts on our societies. However, little has been said on their distributional effects, even if poorer households are undoubtedly much more vulnerable to climate shocks, due to low adaptation and mitigation capability. In this talk, I will present two different works, tackling this issue from two distinctive perspectives. Firstly, adopting a more aggregate and global approach, I show that extreme levels of precipitation exacerbate within-country income inequality. Crucially, the strength and direction of such non-linear relation depend on the agricultural intensity of an economy. With bottom-earners in developing countries being primarily employed in rainfed agriculture, anomalies that disproportionately affect agricultural income translate into higher economic disparity, I project income inequality to worsen globally by the end of the century, especially in high agricultural intensity economies in Africa. Secondly, I will concentrate the analysis on one of the sectors more exposed to climate risk (or to its subjective perception): real estate, which further constitute the major (if not only) asset in poorer households' portfolios. I propose a novel dataset based on disaggregated data coming from the centralized Integrated Public Alert Warning and System, containing geolocalized information on all potential events from distinct climate-related hazards in the United States since 2012, disaggregated by likelihood, severity, and required mitigation actions. Exploiting high-resolution data on more than 100 million real estate properties across United States - and all associated transactions - I disentangle the impact on house prices of property-level exposition to i) objective risk (realized hazards), and ii) perceived risk (unrealized hazards). Results indicate that perceived risk plays a crucial role in shaping market dynamics, magnifying the effects of physical risk, and further lowering house prices in exposed areas.

Session VII (Chair: Andrea Mina)

Panel: Fostering Data- and Computation-driven Interdisciplinary Research

- Jenni Evans, Director, Penn State University, USA.
- Francesca Ieva, Human Technopole and Politecnico di Milano, Italy.
- Third panelist TBA

Session VIII (Chair: Francesca Chiaromonte)

Statistical approaches for the analysis of large and structured data

- Luca Insolia, Sant'Anna School of Advanced Studies, Italy.

Title: Simultaneous Feature Selection and Outlier Detection with Optimality Guarantees.

Abstract: Biomedical research is increasingly data rich, with studies comprising ever growing numbers of features. The larger a study, the higher the likelihood that a substantial portion of the features may be redundant and/or contain contamination (outlying values). Procedures for sparse estimation in the presence of outliers are critical for these studies and have received considerable attention in the last decade. We contribute to this area investigating high-dimensional regression models contaminated by multiple mean-shift outliers affecting both the response and the design matrix. We develop a general framework and use L_0-constraints coupled with mixed-integer programming techniques to simultaneously perform feature selection and outlier detection with provably optimal guarantees. We prove theoretical properties for our approach, that is, a necessary and sufficient condition for the robustly strong oracle property – where the number of features can increase exponentially with the sample size – the optimal estimation of regression coefficients, and the breakdown point of the resulting estimates. We also provide computationally efficient procedures to tune integer constraints and warm-start the algorithm. We show the superior performance of our proposal compared to existing methods through Monte Carlo simulations and use it to elicit the role of microbiome in childhood obesity.

- Tobia Boschi, IBM Research Dublin, Ireland.

Title: High-dimensional functional regression: an efficient method for feature selection and estimation.

Abstract: Functional regression analysis is establishing itself as an active and growing research topic. Regression problems involving large and complex data sets are ubiquitous, and feature selection is crucial for avoiding overfitting and achieving accurate predictions. We propose a new flexible and ultra-efficient approach to perform feature selection in a sparse high dimensional function-on-function regression problem, and we show how to extend it to all the other functional regression frameworks. Our method combines functional data, optimization, and machine learning techniques to perform feature selection and parameter estimation simultaneously. We exploit the properties of Functional Principal Components and the sparsity inherent to the Dual Augmented Lagrangian problem to reduce the computational cost significantly, and we present an adaptive scheme to improve the selection accuracy. Through an extensive simulation study, we benchmark our approach to the best existing competitors and demonstrate a massive gain in terms of CPU time and selection performance, without sacrificing the quality of the estimates. Finally, we present an application from the World Bank Data Catalog.

❖ Ephraim Hanks, Penn State University, USA.

Title: Understanding and mitigating spatial confounding in spatially-referenced health data.

Abstract: When data are collected spatially, it is common to include a spatial random effect to capture missing spatially-smooth covariates that cause spatial autocorrelation in the data. Spatial confounding is the phenomenon where measured predictor variables of interest show correlation with spatial random effects. Spatial confounding can cause variables that seem highly correlated with the response to not be statistically important or significant in a model with a spatial random effect. We first explain probabilistically and heuristically how spatial confounding occurs. We then review the existing modeling approaches that have been proposed to deal with spatial confounding, and propose multiple approaches for estimation that correspond to the most common hypotheses that are important in the analysis of spatial health data. Finally, we propose design principles that will help mitigate spatial confounding when data collection can be controlled.

Session IX (Chair: Chiara Seghieri)

Causal inference and its applications in healthcare

❖ Marcella Vigneri, London School of Hygiene and Tropical Medicine, UK.

Title: Hand hygiene policies, strategies and interventions across settings: A scoping review across domains to identify knowledge

Background: Current threats from healthcare-associated infections (HAIs), antimicrobial resistance (AMR), and emerging pathogens, including Covid 19 transmission, have highlighted the important role of infection prevention and control (IPC) with hand hygiene (HH) as a critical measure. IPC measures with HH are regarded as a cornerstone to curb transmission of pathogens and contribute to reducing the spread of antimicrobial. As such, HH compliance has become one of the key performance indicators of individuals' safety. Despite ongoing advances in improving HH compliance in the past 20 years, HH remains suboptimal. To respond to the COVID-19 pandemic, the Hand Hygiene for All (HH4A) Global Initiative was launched in June 2020 by WHO and UNICEF, in a call to action to international partners, national governments, the public and private sectors, the civil society and donors to accelerate progress to improve HH to stop the spread of COVID-19 and other emerging pathogens.

Methods: This scoping review aims to map the existing evidence on hand hygiene in health care, domestic, school, and public settings, including its determinants and impacts, drawing from multiple public health research disciplines and subject areas. We review existing hand hygiene systematic reviews and/or relevant meta-analyses in the peer-reviewed literature. Each review included was mapped according to whether it provided information on one (or more) of our seven pre-specified 'domains' of hand hygiene: hand hygiene policies, interventions, psychosocial determinants, services/products, behaviours, microbiological hand contamination, and health/development impacts.

Results: The findings of this scoping review will support the development of a shared global research agenda for hand hygiene as a crucial infection prevention and control (IPC) measure. This will provide donors, researchers, health care leaders, policy makers, and implementers with direction for future research within hand hygiene and specific priority research questions to improve evidence-informed programming.

Conclusions: This review identifies the need to build on research in certain areas, and to enhance the evidence from low-income and middle-income countries. We found that several research areas were the subject of a large number of reviews. However, each would still benefit from deeper synthesis to understand remaining research priorities. We also identified several clear gaps where few or no reviews had addressed particular research areas. These key findings are presented here.

❖ Falco Bargagli, Harvard University, USA.

Title: Causal Rule Ensemble: A General Ensemble Learning Framework for Interpretable Subgroups Identification with An Application in Discovering Heterogeneous Exposure Effects in Air Pollution Studies.

Abstract: In social and health sciences, it is critically important to identify subgroups of the study population where a treatment (or exposure) has a notably larger or smaller causal effect on an outcome compared to the population average. In recent years, there have been many methodological developments for addressing heterogeneity of causal effects. A common approach is to estimate the conditional average treatment effect (CATE) given a pre-specified set of covariates. However, this approach does not allow to discover new subgroups, but only to estimate causal effects on subgroups that have been specified a priori by the researchers. Recent causal machine learning (ML) approaches estimate the

CATE at an individual level in presence of large number of observations and covariates with great accuracy. However, because of their complex parametrization of the feature space, these ML approaches do not provide an interpretable characterization of the heterogeneous subgroups. In this paper, we propose a new Causal Rule Ensemble (CRE) method that: 1) discovers de novo subgroups with significantly heterogeneous treatment effects (i.e., causal rules); 2) ensures interpretability of these subgroups because they are defined in terms of decision rules; and 3) estimates the CATE for each of these newly discovered subgroups with small bias and high statistical precision. We provide theoretical results that guarantee consistency of the estimated causal effects for the newly discovered causal rules. A nice feature of CRE is that it is agnostic to the choices of (i) the ML algorithms that can be used to discover the causal rules, and (ii) the estimation methods for the causal effects within the discovered causal rules. Via simulations, we show that the CRE method has competitive performance as compared to existing approaches while providing enhanced interpretability. We also introduce a new sensitivity analysis to unmeasured confounding bias. We apply the CRE method to discover subgroups that are more vulnerable (or resilient) to the causal effects of long-term exposure to air pollution on mortality.

❖ Bénédicte Colnet, Inria Paris-Saclay, France.

Title: How can we account for sampling bias in randomized controlled trials?

Abstract: Randomized Controlled Trials (RCTs) are considered as the gold standard to conclude on the causal effect of a given intervention on an outcome, in particular in medicine. Still, they may lack of external validity when the population eligible to the RCT is substantially different from a target population of interest. Having at hand a sample of such a target population of interest allows to generalize (or transport) the findings from the RCT to this population. To do so, a requirement is to have access to covariates in both sets to capture all treatment effect modifiers that are shifted between the two sets. Standard estimators then use either weighting (IPSW), outcome modeling (G-formula), or combine the two in doubly robust approaches (AIPSW). In this presentation we first recall how this generalization problem can be understood as a mirroring problem of the average treatment effect estimation from a single observational study, with adaptations. The estimators and current challenges are presented in a real-world example from critical care medicine. However all necessary covariates for identification are often not available in both sets. Therefore, the second part of the presentation proposes a sensitivity analysis to handle completely or partially-unobserved covariate. We illustrate these results on a semi-synthetic benchmark using data from the Tennessee Student/Teacher Achievement Ratio (STAR).