



Deliverable D10.3

**Exploratory activities report and  
planning for the next period 2**



## DOCUMENT INFORMATION

PROJECT	
PROJECT ACRONYM	SoBigData-PlusPlus
PROJECT TITLE	SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics
STARTING DATE	01/01/2020 (60 months)
ENDING DATE	31/12/2024
PROJECT WEBSITE	<a href="http://www.sobigdata.eu">http://www.sobigdata.eu</a>
TOPIC	INFRAIA-01-2018-2019 Integrating Activities for Advanced Communities
GRANT AGREEMENT N.	871042

DELIVERABLE INFORMATION	
WORK PACKAGE	WP10 JRA3 - Exploratories
WORK PACKAGE LEADER	KTH and CNR
WORK PACKAGE PARTICIPANTS	CNR, USFD, UNIPI, FRH, UT, IMT, LUH, KCL, SNS, AALTO, ETH Zürich, PSE, UNIROMA1, CNRS, CEU, URV, CSD, BSC, UPF, Eli, CRA, UvA
DELIVERABLE NUMBER	D10.3
DELIVERABLE TITLE	Exploratory activities report and planning for the next period 2
AUTHOR(S)	Luca Pappalardo (ISTI-CNR), Aris Gionis (KTH), Michela Natilli (CNR)
CONTRIBUTOR(S)	
EDITOR(S)	Valerio Grossi (CNR), Michela Natilli (CNR)
REVIEWER(S)	Vaiva Vasiliauskaite (ETH Zürich), Angelo Facchini (IMT)
CONTRACTUAL DELIVERY DATE	31/12/2022
ACTUAL DELIVERY DATE	26/04/2023
VERSION	V1.1
TYPE	Report
DISSEMINATION LEVEL	Public
TOTAL N. PAGES	62
KEYWORDS	Exploratory, micro-projects, artificial intelligence, social mining

## EXECUTIVE SUMMARY

This deliverable updates deliverables D10.1 “Initial Exploratory activities planning” and D10.2 “Exploratory activities report and planning for the next period 1” and it provides information about the activities performed since the beginning of the project and the planning of the activities for the next period, for WP10 - Exploratories.

## DISCLAIMER

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042.

SoBigData++ strives to deliver a distributed, Pan-European, multi-disciplinary research infrastructure for big social data analytics, coupled with the consolidation of a cross-disciplinary European research community, aimed at using social mining and big data to understand the complexity of our contemporary, globally-interconnected society. SoBigData++ is set to advance on such ambitious tasks thanks to SoBigData, the predecessor project that started this construction in 2015. Becoming an advanced community, SoBigData++ will strengthen its tools and services to empower researchers and innovators through a platform for the design and execution of large-scale social mining experiments.

This document contains information on SoBigData++ core activities, findings and outcomes and it may also contain contributions from distinguished experts who contribute as SoBigData++ Board members. Any reference to content in this document should clearly indicate the authors, source, organisation and publication date.

The content of this publication is the sole responsibility of the SoBigData++ Consortium and its experts, and it cannot be considered to reflect the views of the European Commission. The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

Copyright © The SoBigData++ Consortium 2020. See <http://www.sobigdata.eu/> for details on the copyright holders.

For more information on the project, its partners and contributors please see <http://project.sobigdata.eu/>. You are permitted to copy and distribute verbatim copies of this document containing this copyright notice, but modifying this document is not allowed. You are permitted to copy this document in whole or in part into other documents if you attach the following reference to the copied elements: "Copyright © The SoBigData++ Consortium 2020."

The information contained in this document represents the views of the SoBigData++ Consortium as of the date they are published. The SoBigData++ Consortium does not guarantee that any information contained herein is error-free, or up to date. THE SoBigData++ CONSORTIUM MAKES NO WARRANTIES, EXPRESS, IMPLIED, OR STATUTORY, BY PUBLISHING THIS DOCUMENT.

## GLOSSARY

AI	Artificial Intelligence
EC	European Commission
EU	European Union
H2020	Horizon 2020 EU Framework Programme for Research and Innovation
MP	Micro-Project
PM	Person Month

# TABLE OF CONTENTS

1	Relevance to SoBigData++ .....	8
1.1	Purpose of this document .....	8
1.2	Relevance to project objectives .....	8
1.3	Relation to other work packages .....	8
1.4	Structure of the document .....	8
2	Micro-projects .....	9
3	T10.1 Societal Debates and Misinformation Analysis .....	11
3.1	Activities Report .....	11
3.2	Micro-projects .....	15
3.3	Publications .....	18
3.4	Planned Activities for next period .....	19
3.4.1	<i>Data Collection activities</i> .....	19
3.4.2	<i>Software Development activities</i> .....	19
3.4.3	<i>Events</i> .....	19
3.4.4	<i>Scientific activities</i> .....	19
4	T10.2 Demography, Economy & Finance 2.0 .....	21
4.1	Activities Report .....	21
4.2	Micro-projects .....	27
4.3	Publications .....	28
4.4	Planned Activities for next period .....	29
4.4.1	<i>Data Collection activities</i> .....	29
4.4.2	<i>Software Development activities</i> .....	30
4.4.3	<i>Events</i> .....	30
4.4.4	<i>Scientific activities</i> .....	30
5	T10.3 Sustainable Cities for Citizens .....	32
5.1	Activities Report .....	32
5.2	Micro-projects .....	33
5.3	Publications .....	36
5.4	Planned Activities for next period .....	37
5.4.1	<i>Data Collection activities</i> .....	37
5.4.2	<i>Events</i> .....	38
5.4.3	<i>Scientific activities</i> .....	38
6	T10.4 Migration Studies .....	39
6.1	Activities Report .....	39
6.2	Micro-projects .....	42
6.3	Publications .....	43

6.4	Planned Activities for next period .....	44
6.4.1	<i>Data Collection activities</i> .....	44
6.4.2	<i>Software Development activities</i> .....	44
7	T10.5 Sports Data Science .....	46
7.1	Activities Report .....	46
7.2	Micro-projects .....	46
7.3	Publications .....	49
7.4	Planned Activities for next period .....	51
7.4.1	<i>Scientific activities</i> .....	51
8	T10.6 Social Impact of AI and Explainable Machine Learning .....	52
8.1	Activities Report .....	52
8.2	Software Development activities .....	57
8.3	Events .....	57
8.4	Micro-projects .....	57
8.5	Publications .....	59
8.6	Planned Activities for next period .....	60
8.6.1	<i>Scientific activities</i> .....	60
9	Conclusions .....	62

# 1 Relevance to SoBigData++

## 1.1 Purpose of this document

This document describes the activity carried out within the exploratories in the last period and the topics and activities planned for the next period. This deliverable updates deliverables D10.1 “Initial Exploratory activities planning”<sup>1</sup> and D10.2 “Exploratory activities report and planning for the next period 1”<sup>2</sup>.

## 1.2 Relevance to project objectives

The topics and activities described in this document are relevant to milestone MS5 “All the Exploratories reached technical and scientific maturity”. For each task of WP10, we report the results achieved for each topic and the activities carried out in terms of conferences/workshops, hackathons, data collection, and software development.

## 1.3 Relation to other work packages

Since in the document we also describe some activities made or planned for the next period, this deliverable is also related to work packages WP3 - Dissemination, Impact, and Sustainability (because of workshops and conferences have been made or planned), WP4 - Training (because hackathons have been made or planned), and WP7 - Virtual Access (because data sets and software have been made available on the infrastructure or planned).

## 1.4 Structure of the document

The research in WP10 is structured in vertical thematic environments (exploratories) each associated with a task, aimed at creating new stories and new resources to be integrated within the SoBigData++ research infrastructure:

- Section 2 overviews the actions performed in the WP10 from the micro-project activity perspective.
- Sections from 3 to 8 describe the scientific results each exploratory has achieved in the second period of the project and the topics and activities planned to investigate for the next period. For each exploratory, we also list the micro-project proposed and (if already terminated) the corresponding resources created.

---

<sup>1</sup> <https://data.d4science.net/T1sp>

<sup>2</sup> <https://data.d4science.net/6Y2j>



## 2 Micro-projects

In January 2021, we introduced the use of a micro-project system to foster the collaboration among the partners of the consortium and improve the tracking of the activity developed within the project. The micro-project (MP) is a piece of research that one or more partners carry out; it is related to one or more exploratories. As a result, we get some tangible outputs like resources (methods, dataset, etc) and/or new scientific results to upload on the platform and made available to the community.

From July 2021 to date the partners presented 43 new MP related to WP10. Of these MPs, 21 have already been completed, while the remaining 22 are still ongoing. Figure 2.1 shows the distribution of MPs across the tasks in WP10.

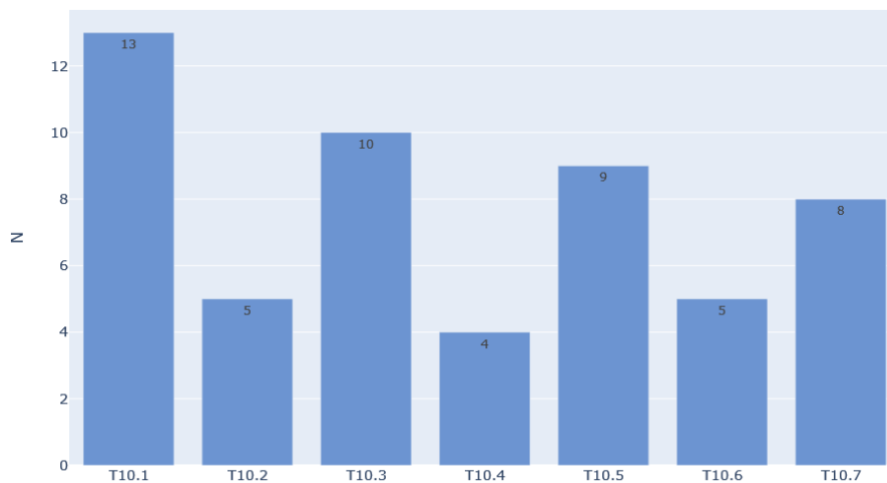


Figure 2.1 - Distribution of micro-projects among tasks in WP10

Each MP should produce at least one tangible output. Figure 2.2 shows the distribution of expected tangible outputs from all the MP submitted since July 2021.

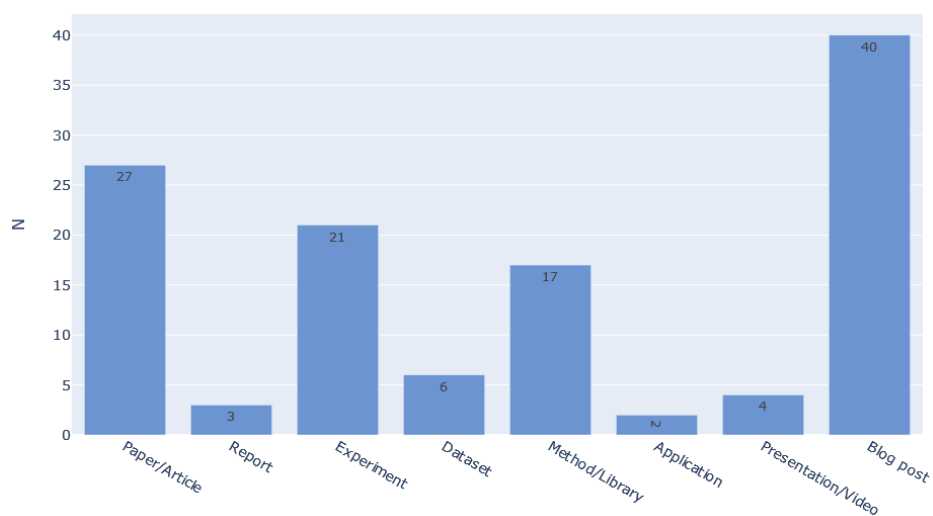


Figure 2.2 - Distribution of expected tangible outputs from micro-projects

The expected production, in terms of scientific papers and items uploaded on the catalogue (methods, experiments, applications, datasets) is considerable. To encourage communication of the results obtained and publicise the new items produced for each MP, each MP also produces a post for the SoBigData blog. Figure 2.3 shows the distribution of actual results obtained from the MPs terminated so far. In summary, MPs produces 16 blog posts, 9 experiments, 6 methods or library updates, 4 datasets, and 8 scientific papers.

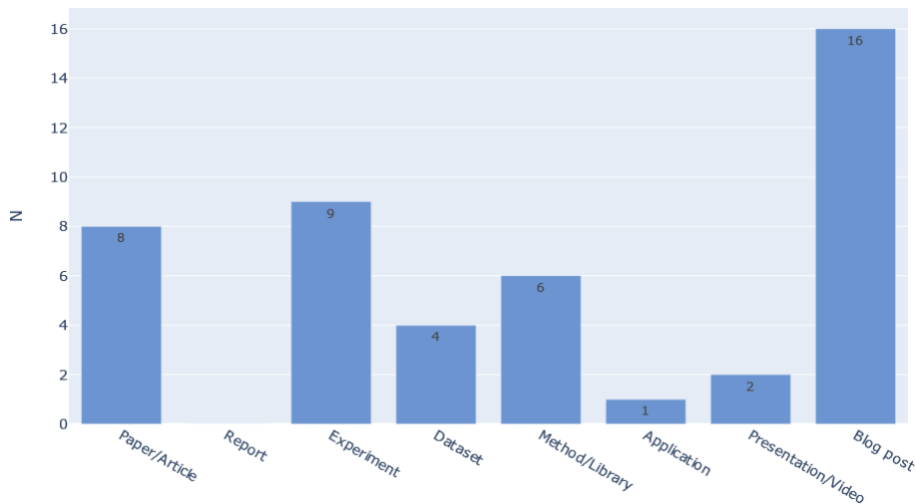


Figure 2.3 - Distribution of actual tangible outputs from micro-projects

### 3 T10.1 Societal Debates and Misinformation Analysis

This exploratory aims to develop methods and datasets for studying online public debates in (near) real-time and at scale, i.e., during election campaigns or on controversial topics such as vaccination, abortion, or discrimination. The central focus regards misinformation, with the purpose of developing new methods for detecting, analysing, and tracking online misinformation and propaganda across social media platforms, countries, and over time. A key aim is to improve the accuracy of the methods through collecting more data, experimentation with semi-supervised and unsupervised methods, and integrating the latest advances in deep learning. The exploratory also studies the effect of different social relationships when it comes to opinion formation.

#### 3.1 Activities Report

##### **COVID-19 misinformation analysis**

*Partners Involved:* USFD

Massive amounts of COVID-19 related information spread across social media platforms, which can make it challenging for the viewer/reader to distinguish between true and false information. We conducted a comparative study of COVID-19 related tweets, based on over 242 million source tweets. We compared the statistical characteristics of true vs false tweets with respect to their topical distribution; tweet deletion rates; language analysis; and spread over time. We also created a ‘silver standard’ COVID-19 misinformation classification dataset, which leads to improvements in misinformation classification accuracy by over 9% average F1. A paper is in preparation and will shortly be submitted for peer review. Upon acceptance we will publish a blog post on the SoBigData++ website. An automatic classifier trained on the above dataset has been made available as a service in the D4Science Method Engine. The method accepts as input a CSV file of texts to classify and returns an output CSV giving the most probable detected category for each input text.

##### **Multilingual News Article Similarity**

*Partners Involved:* USFD

We participated in the SemEval-2022 Task8: Multilingual News Article Similarity, where we developed the second highest ranked system. We proposed an entity-enriched Siamese Transformer which computes news article similarity based on different sub-dimensions, such as the shared narrative, entities, location and time of the event discussed in the news article. Our system exploits a Siamese network architecture using a Transformer encoder to learn document-level representations for the purpose of capturing the narrative together with the auxiliary entity-based features extracted from the news articles. The intuition behind using all these features together is to capture the similarity between news articles at different granularity levels and to assess the extent to which different news outlets write about “the same events”. Our experimental results and detailed ablation study demonstrate the effectiveness and the validity of our proposed method. The results are already published, and source code is made available, and the trained classifier has been integrated in the D4Science Method Engine. The method takes a CSV as input where each row represents a pair of articles, optionally annotated with named entities, and produces as output another CSV file with the similarity score for each pair.

## Integration of other text and image analysis tools in D4Science Method Engine

*Partners Involved:* USFD

In addition to the specific items mentioned above (and under the vaccine narratives micro-project below), USFD has integrated a number of other pre-existing text and image analysis tools as methods in the D4Science Method Engine. A large number of GATE Cloud services were integrated in the Method Engine as part of the original SoBigData project, but USFD has now revisited this work to develop a newer and deeper integration mechanism that will be more friendly for non-specialist users.

The following methods have been integrated:

- Generic named entity annotation tools for general text (ANNIE), also specifically for Twitter data (TwitIE)
- A tool for annotating measurement expressions in text
- A domain-specific text analysis tool for Chemistry, to annotate chemical names with their identifiers in standard nomenclatures
- A set of domain-specific text analysis tools for Archaeology (from the Ariadne Infrastructure) for English, Swedish and Dutch text
- A variety of tools for misinformation detection:
  - A classifier for texts about rumours, that attempts to determine if the rumour is likely to be true, false, or unverified
  - A classifier to determine whether a given text is “hyper-partisan” in nature, which won the Hyperpartisan News Detection task at SemEval 2019
  - Classifiers to detect toxic and offensive language in short texts such as tweets or Wikipedia comments
  - The COVID-19 misinformation and vaccine narratives classifiers described elsewhere in this deliverable
  - A tool to analyse URLs, unwrap shortened forms to their ultimate target URLs, and flag any information about whether the target domain is a known fact checker or known source of misinformation
- Two tools for the detection and analysis of hate speech and abuse, one against politicians specifically and one more general

These methods have all been designed to process and return data in CSV format as opposed to the JSON format used in the original SoBigData integration, so as to be more understandable and usable by non-specialist users. The mapping between the complex data structures returned by the raw tools and the columnar CSV format is handled by the use of a special “controlled language”, for example “Person” would return any text strings that were annotated as person names, “Veracity status (confidence)” would return the status (true/false/unverified) for each input along with the classifier’s confidence score. Each method offers a set of pre-configured output options for the user to select from, or they can add their own using the same controlled language. Figure 4.1 shows an example of how this is presented in the Method Engine.

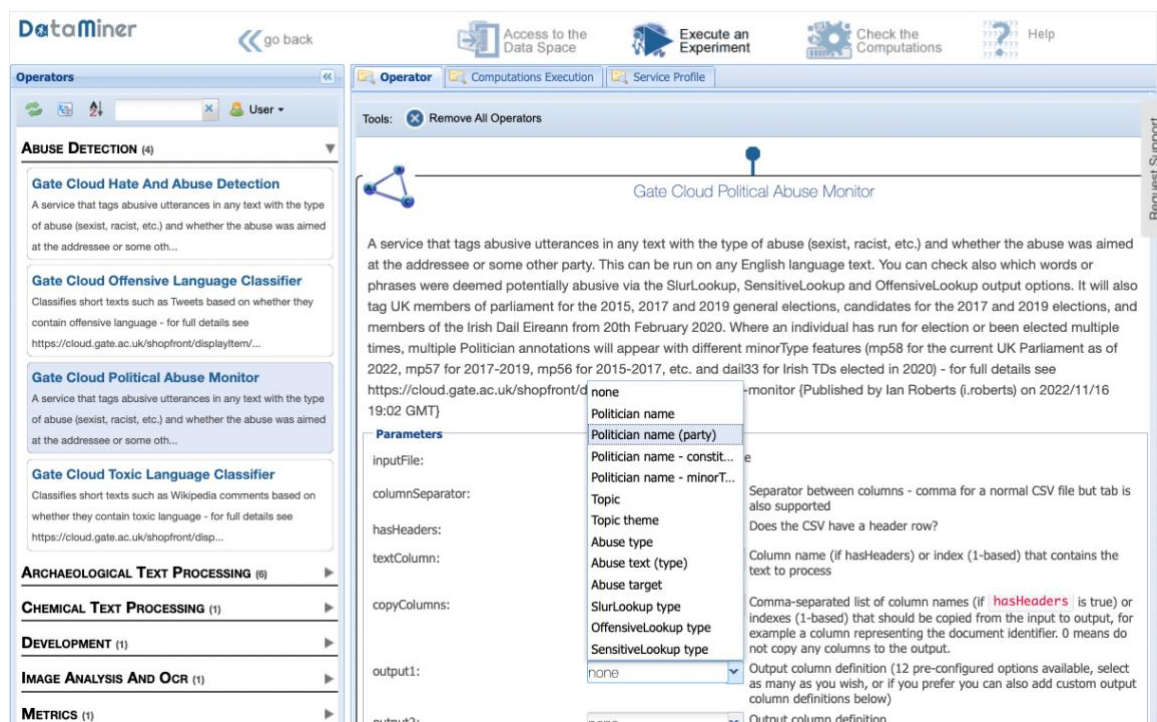


Figure 4.1 Political Abuse analysis tool in the Method Engine

Another method that has been integrated is a “stance classifier” that processes threaded discussions (e.g. tweets and replies, or threaded posts on a forum) and attempts to determine the stance of each reply with respect to the text to which it is replying, i.e. whether the reply supports the original position, denies it, questions whether the original post is true, or simply comments on it without implying any particular position. Two versions of this method are available, one with a model trained purely on English and another with a multi-lingual model that works for many more languages. Input is expected to be CSV format, with each post having a (required) post identifier and an (optional) “reply-to” link with the ID of the post to which it is replying - there can be many replies to the same original post. Output is another CSV giving the stance class and confidence scores for each reply relative to its respective target.

The final method that has been integrated is a multi-lingual tool for optical character recognition (OCR) to extract text from images. The tool can cope with multiple languages of text within the same image, and will transcribe each “block” of text individually. Users invoke the tool by providing a folder in their SoBigData Workspace that contains the set of images, the tool fetches each image using the Storage Hub API, processes it, and produces an output CSV file with all the text blocks it detects in the set of images.

### Question routing via activity-weighted modularity-enhanced factorization

*Partners Involved:* ETHZ

Question Routing (QR) in Community-based Question Answering (CQA) websites aims at recommending newly posted questions to potential users who are most likely to provide “accepted answers”. Most of the existing approaches predict users’ expertise based on their past question answering behaviour and the content of new questions. However, these approaches suffer from challenges in three aspects: (1) sparsity of users’ past records results in lack of personalised recommendation that at times does not match users’ interest or domain expertise, (2) modelling based on all questions and answers content makes periodic

updates computationally expensive, and (3) while CQA sites are highly dynamic, they are mostly considered as static. In this work, we studied network analysis tools to resolve challenges of QR, namely dynamic modelling of users' activity on topic communities. Experimental results on three real-world datasets demonstrate that the proposed model significantly outperforms competitive baseline models.

### **Analysis of opinion dynamics over a realistic dynamic social network**

*Partners involved:* CNR, CEU

The objective of this activity is to develop a realistic opinion dynamics model over a dynamic social network and to investigate its impact in terms of polarisation. Among opinion dynamic models, the Friedkin-Johnsen (FJ) is one of the most popular and it is one of the few models validated in real social groups. However, a very strong assumption of FJ is that it assumes a static social graph, where relationships are not supposed to change over time. However, it has been proved that this is not realistic in general, because, depending on the strength of the tie or on the importance of an issue discussed by two people, opinions do have an impact on the social network itself, leading to social ties disruption in the worst cases. Our goal is to fill this gap and design an agent-based model system to extend the FJ model to include a rewiring scheme (that alters relationships based on the opinions of the two parties) and to analyse how the dynamicity of ties impacts the polarisation properties of the model. To address the polarisation properties of the novel model, we will also work towards defining a proper polarisation definition for this scenario. This research activity has been partly carried out during a TNA visit of Dr. Elisabetta Biondi at the CEU premises (see this [blog post](#)).

### **Signed ego network model**

*Partners involved:* CNR

The Ego Network Model (ENM) describes how individuals organise their social relations in concentric circles (typically five) of decreasing intimacy, and it has been found almost ubiquitously in social networks, both offline and online. The ENM gauges the tie strength between peers in terms of interaction frequency, which is easy to measure and provides a good proxy for the time spent nurturing the relationship. However, advances in signed network analysis have shown that positive and negative relations play very different roles in network dynamics. For this reason, this work sets out to investigate the ENM when including signed relations. The main contributions so far of this line of research are twofold: firstly, a novel method of signing relationships between individuals using sentiment analysis and, secondly, an investigation of the properties of Signed Ego Networks (Ego Networks with signed connections). Signed Ego Networks are extracted for the users of eight different Twitter datasets composed of both specialised users (e.g. journalists) and generic users. We find that negative links are over-represented in the active part of the Ego Networks of all types of users, suggesting that Twitter users tend to engage regularly with negative connections. Further, we observe that negative relationships are overwhelmingly predominant in the Ego Network circles of specialised users, hinting at very polarised online interactions for this category of users. In addition, negative relationships are found disproportionately more at the more intimate levels of the ENM for journalists, while their percentages are stable across the circles of the other Twitter users.

### 3.2 Micro-projects

#### **(Mis-)leading the Covid-19 vaccination discourse on Twitter - A study of infodemic around the pandemic**

The work is related to analysing (mis)leading tweets related to Covid-19 vaccination. After semi-manual annotation, and use of Transfer Learning, we plan to study misleading and non-misleading tweets using various NLP techniques. In addition, we also plan to use various ML models for predicting (mis)leading tweets.

Status: Completed; Partners involved: UT

Outputs: [research article](#), [experiment](#), [blog post](#).

#### **A Knowledge Graph based Approach for Identifying Fake News**

The work is related to analysing (mis)leading tweets related to Covid-19 vaccination. After semi-manual annotation, and use of Transfer Learning, we plan to study misleading and non-misleading tweets using various NLP techniques. In addition, we also plan to use various ML models for predicting (mis)leading tweets.

Status: Completed; Partners involved: UT

Outputs: [research article](#), [experiment](#), [blog post](#).

#### **Comparative Analysis of Ukraine-related misinformation**

We compare quantitatively the spread of Ukraine-related disinformation and its corresponding debunks, first by considering re-tweets, replies, and favourites, which demonstrate that despite platform efforts Ukraine-related disinformation is still spreading wider than its debunks. Next, bidirectional post-hoc analysis is carried out using Granger causality tests, impulse response analysis and forecast error variance decomposition, which demonstrate that the spread of debunks has a positive impact on reducing Ukraine-related disinformation eventually, albeit not instantly. Lastly, the paper investigates the dominant themes in Ukraine-related disinformation and their spatiotemporal distribution. With respect to debunks, we also establish that around 18% of fact-checks are debunking claims which have already been fact-checked in another language. The latter finding highlights an opportunity for better collaboration between fact-checkers, so they can benefit from and amplify each other's debunks through translation, citation, and early publication online.

Status: Completed; Partners involved: USFD

Outputs: [conference paper](#), [blog post](#).

#### **Multilingual access to COVID-19 health information**

The Coronavirus (COVID-19) infodemic motivated the need for accurate semantic search and retrieval of reliable COVID-19 information across millions of documents, in multiple languages. To address this challenge, we propose a novel high precision and high recall neural Multistage BiCross encoder approach. It is a sequential three-stage ranking pipeline which uses the Okapi BM25 retrieval algorithm and transformer-based bi-encoder and cross-encoder to effectively rank the documents with respect to the given query. We present experimental results from our participation in the Multilingual Information Access (MLIA) shared task on COVID-19 multilingual semantic search. The independently evaluated MLIA results validate our approach and demonstrate that it outperforms other state-of-the-art approaches according to nearly all evaluation metrics in cases of both monolingual and bilingual runs.

Status: Completed; Partners involved: USFD

Outputs: [conference paper](#).



**Understanding Belarus Protests Using Online Social Media Platforms**

Online protests, a special kind of online activism, are happening all around the globe. We are planning to follow online protests in Belarus in 2020. In particular, we plan to understand the typical characteristics of these protests, by using NLP related techniques (such as sentiment analysis, topic modeling).

Status: Completed; Partners involved: UT

Outputs: [Research article](#), [Experiment](#), [blog post](#).

**Classifying COVID-19 vaccine narratives**

COVID-19 vaccine hesitancy is widespread, despite government information campaigns and the efforts of the World Health Organisation (WHO). One of the reasons behind this is vaccine disinformation which spreads widely on social media. In particular, recent surveys have established that vaccine disinformation is negatively impacting citizens' trust in the COVID-19 vaccination. At the same time, fact-checkers are struggling with detecting and tracking vaccine disinformation, due to the large scale of social media. To assist fact-checkers in monitoring vaccine narratives online, we study a new vaccine narrative classification task, which categorises COVID-19 vaccine claims into one of seven categories. Following a data augmentation approach, we first construct a novel dataset for this new classification task, focusing on the minority classes. We also make use of fact-checker annotated data. The paper also presents a neural vaccine narrative classifier that achieves an accuracy of 84% under cross-validation. The classifier is publicly available for researchers and journalists.

Status: Ongoing; Partners involved: USFD

Expected Outputs: [journal paper under review](#).

**Dynamics of opinion polarisation with the Friedkin-Johnsen model**

Opinion polarisation has been the subject of extensive debate in recent years, due to its potential to disrupt our societies and democracies. Understanding how people form their opinion is crucial to reveal the mechanism behind polarisation. While many opinion polarisation models have been proposed in the related literature, the only model that has been validated on small-to-medium sized groups is the one proposed by Friedkin and Johnsen (FJ model). It belongs to the family of averaging models (whereby the opinion of a node becomes, at each step, the average opinion of its neighbours) and its distinctive trait is the anchoring to the node's initial opinion through the susceptibility parameter. In this research, we provide a comprehensive review of all the major variants of the FJ model and of the polarisation metrics described in the related literature. For them, we will highlight their key features and the differences between each other. As a second contribution, we derive the conditions under which the FJ model yields to polarisation, for each of the polarisation metrics identified before. In addition, we also prove that polarizing opinion vectors can be found analytically in most cases, and we provide recipes for that.

Status: Completed; Partners involved: CNR

Outputs: [code](#), [blog post](#).

**Misinformation Detection on YouTube Using Video Captions**

Millions of people use platforms such as YouTube, Facebook, Twitter, and other mass media. Due to the accessibility of these platforms, they are often used to establish a narrative, conduct propaganda, and disseminate misinformation. This work proposes an approach that uses state-of-the-art NLP techniques to extract features from video captions (subtitles) for predicting videos related to misinformation.

Status: Completed; Partners involved: UT

Outputs: [research article](#), [experiment](#), [blog post](#).



**Covid-19 infodemic in Italy**

The Covid-19 pandemic has led to a corresponding infodemic, emphasized by the use of social media as the primary communication channel during lockdowns. This study is aimed at finding the accounts that spread information in Italian on Covid-19, and how such information was propagated in the first Western country to face a lockdown.

Between the 30th of January and the 20th of March 2020, we found 3,690,196 contents in Italian including Covid-related keywords. In this project, such contents are analyzed in order to answer three research questions: i) which are the most retweeted accounts in the Italian Twittersphere, in the communication flows related to Covid-19?; ii) which and how wide are the online audiences of the most retweeted accounts?; iii) How do the most retweeted accounts spread their messages and reach their audiences?."

Status: Outgoing; Partners involved: CNR

Outputs: [research article](#), [dataset](#).

**Fake account trafficking and coordinated inauthentic behavior on Twitter**

Fake accounts are the primary means for misuse and abuse of social media platforms, giving rise to coordinated inauthentic behaviors. Despite ongoing efforts to limit their exploitation, ready-to-use fake accounts can be found for sale on several underground markets. For the present study, we devised an innovative approach to detect accounts for sale on an underground market. Between June 2019 and July 2021, we detected more than 60,000 fake accounts, which we continuously tracked for changes in profile information and timeline. In this study, we focus on the interactions produced by such fake accounts during 2020. First, we aim to identify the accounts' main characteristics, like the most used names and profile descriptions. Then, we will attempt to spot and describe behavioral patterns indicating coordination, like using similar profile names or retweeting the same user.

Status: Completed; Partners involved: CNR

Outputs: [research article](#), [dataset](#).

**Flow of Online Attention in Societal Debates**

In this microproject, we will address the topic of misinformation analysis from an original perspective: considering not the truthfulness/falsity of the contents circulated online, but the temporal profile of the public attention that they produce. Our hypothesis is that attention regimes characterized by a high ephemerality are associated with a careless public debate in which issues rise and fall too quickly to produce a meaningful conversation. Our fieldwork will focus on online platforms, particularly those dedicated to video sharing (e.g. YouTube); video streaming (e.g. Twitch and Dlive); and subcultural discussions (e.g. Reddit and 4chan). Conceptually, we will leverage theories of media attention cycles and of secondary orality (i.e. the idea that electronic communication is characterized by an evanescence similar to that of preliterate cultures). Methodologically, we will use computational techniques based on time-series analysis and stream decomposition as well as techniques of observation and description derived from digital ethnography. Our inquiry will involve the organization of several workshops (smaller and larger datathons), where scholars from different disciplines and backgrounds will collaborate on this research.

Status: Ongoing; Partners involved: CNRS

Expected Outputs: workshop, book chapter, blog post.

**Reducing radicalism in social networks by feeds prioritization**

In the information age, social media platform tend to solve the information overload problem by filtering and recommending personalized content and accounts, while aiming at maximizing the local user's engagement. Consequently, some well-known social phenomena appeared as main traits of the current social debates, such as polarization, echo-chamber effect, and radicalism. In this work we focus our attention on the levels of radicalism produced in an opinion formation setting, and we propose an algorithm that, in contrast with common marketing strategies, aims at minimizing this global effect by prioritizing the user's feeds. Firstly, we formally define the problem, then we consider two opinion dynamics models in the literature, and we characterize the minimum radicalism to guide our algorithm design. We complete the study by testing the algorithm on synthetic and real-world data about political debates on Twitter.

Status: Ongoing; Partners involved: Uniroma1

Expected outputs: method, Experiments, Blog post, Preprint paper.

**Reducing Graph Structural Bias by Adding shortcut edges**

Reducing structural bias in the social networks can help users to escape from echo chambers and improve society. The goal of his study is to learn how to add shortcut edges to reduce graph structural bias. Specifically, we will extend the paper "RePBubLik: Reducing the Polarized Bubble Radius with Link Insertions", by Haddadan et al., to a continuous setting, where each node has a real-valued polarization score, and the objective is to maximize the structural bias decrease by adding edges among nodes.

Status: Ongoing; Partners involved: Uniroma1

Expected outputs: method, Experiments, Blog post, Preprint paper.

**The ego network of words in Twitter**

Well-established cognitive models coming from anthropology have shown that, due to the cognitive constraints that limit our "bandwidth" for social interactions, humans organize their social relations according to a regular structure. We postulate that similar regularities can be found in other cognitive processes, such as those involving language production. In order to investigate this claim, we collect and share 4 datasets from Twitter, including regular users and professional writers. These datasets are then exploited, leveraging a methodology similar to the one used to uncover the well-established social cognitive constraints, to investigate regularities at both the structural and semantic levels of what we call ego networks of words.

Status: ongoing; Partners involved: CNR

Expected outputs: datasets, training, blog post

### 3.3 Publications

1. I. Singh, Y. Li, M. Thong, C. Scarton. (2022) GateNLP-UShef at SemEval-2022 Task 8: Entity-Enriched Siamese Transformer for Multilingual News Article Similarity. <https://arxiv.org/pdf/2205.15812.pdf>
2. I Singh, K Bontcheva, X Song, C Scarton. (2022) Comparative Analysis of Engagement, Themes, and Causality of Ukraine-Related Debunks and Disinformation. International Conference on Social Informatics. [https://dl.acm.org/doi/abs/10.1007/978-3-031-19097-1\\_8](https://dl.acm.org/doi/abs/10.1007/978-3-031-19097-1_8)
3. I. Singh, C. Scarton, K. Bontcheva (2021) Multistage BiCross encoder for multilingual access to COVID-19 health information. PLOS ONE 16(9): e0256874. <https://doi.org/10.1371/journal.pone.0256874>

4. Krishna, V., Vasiliauskaite, V. & Antulov-Fantulin, N. Question routing via activity-weighted modularity-enhanced factorization. Soc. Netw. Anal. Min. 12, 155 (2022). <https://doi.org/10.1007/s13278-022-00978-6>
5. Vasiliauskaite, Vaiva, Nino Antulov-Fantulin, and Dirk Helbing. "On some fundamental challenges in monitoring epidemics." Philosophical Transactions of the Royal Society A 380.2214 (2022): 20210117.
6. Biondi, Elisabetta, Chiara Boldrini, Andrea Passarella, and Marco Conti. "Dynamics of opinion polarization." arXiv preprint arXiv:2206.06134 (2022). [under review]
7. Tacchi, Jack, Chiara Boldrini, Andrea Passarella, and Marco Conti. "Signed ego network model and its application to Twitter." arXiv preprint arXiv:2206.15228 (2022) (accepted for publication at IEEE Big Data 2022 - Special Session on Intelligent Data Mining).

### 3.4 Planned Activities for next period

#### 3.4.1 Data Collection activities

*Partners involved:* IMT, USFD

We will continue collecting data about COVID-19 misinformation from Twitter for experimentation in the misinformation research. We will also continue collecting debunks of misinformation from the IFCN Poynter website and analysing the types of misinformation changing through time. We are also collecting YouTube video data and comparing videos containing only factual information and non-factual information between types of sources. Lastly, data related to financial news and other current affairs (e.g., elections) will be collected from Twitter.

#### 3.4.2 Software Development activities

*Partners involved:* USFD

Once the models and experiments have reached sufficient levels of maturity, they will be refactored into software tools or web services and integrated into the SoBigData platform.

#### 3.4.3 Events

*Partners involved:* USFD, CSD, IMT, ETHZ, UT, CNR

A workshop or summer school on disinformation analysis methods is planned for 2023.

#### 3.4.4 Scientific activities

*Partners involved:* USFD, CSD, IMT, ETHZ, UT, CNR

Continue experiments on COVID-19 related, political, and climate misinformation detection and analysis. Stronger attention will be paid to cross-platform and cross-lingual methods and experiments, as well as carrying out comparative studies across different EU countries. We are also planning to improve the visibility of our research activities by reaching out to the EDMO disinformation research community. Regarding our research activities on opinion dynamics under the Friedkin-Johnsen model, we plan to design a variant of the Friedkin-Johnsen model that embeds the dynamicity of social networks. Furthermore, we will work on a novel definition of global polarisation that combines network features and opinion distribution, to capture the existence of clustered opinions. We will analyse the polarisation effect of the new dynamic model, and identify the impact of the network structure. The activities on signed ego networks will continue by focusing on (i) the validation of the signed ego network model through the collection of a ground truth (ii) the

application of the model to more diverse social networks (e.g., Reddit, Mastodon) *(iii)* the topological properties of signed ego networks from the triad analysis perspective.

## 4 T10.2 Demography, Economy & Finance 2.0

The aim of this exploratory is that of combining statistical methods and traditional economic data (typically at low-frequency) with high-frequency data from non-traditional digital sources (e.g., web, supermarkets), for monitoring economic, socio-economic and well-being indicators. Another purpose of this exploratory is studying traditional complex socio-economic and financial systems in conjunction with emerging ones, in particular, block-chain & cryptocurrency markets and their applications such as smart property, Internet of things (IoT), energy trading and smart contracts. In the field of finance, different aspects will be studied, such as risk and liquidity estimation, microstructure dynamics & market predictions as well as different connections to social media and news.

### 4.1 Activities Report

#### **Conference Track at AmlD2022: Advances of ML Approaches for Financial Decision Making & Time Series Analysis**

*Partners Involved:* ETH

The purpose of this track was to enable the exchange of recent research and insights amongst researchers interested in machine learning approaches for decision making and times series analysis of financial markets. We brought together world-class presenters from academia and industry working on topics such as:

- Deep learning for financial time series
- Reinforcement learning and data-driven optimal control for financial decision making
- Transformer-based and related NLP approaches for financial sentiment and event analysis
- Graph-based neural network techniques in finance.

This track resulted in a most vibrant and fruitful exchange of ideas and information from researchers from different disciplines such as machine learning, complex systems, physics, mathematics and quantitative finance.

#### **Complexity techniques in finance**

- **Analysis of 2017-2018 Bitcoin markets crash**

*Partners Involved:* SNS, ETHZ

We studied the market microstructure of Bitcoin related to liquidity during the bitcoin bubble in 2017-2018. By analysing high-frequency market microstructure observables with different information-theoretic measures for dynamical systems, we found temporal changes in information sharing across markets. In particular, we studied time-varying components of predictability, memory, and (a)synchronous coupling, measured by transfer entropy, active information storage, and multi-information. By comparing these empirical findings with several models, we argued that some results could relate to intra-market and inter-market regime shifts and changes in the direction of information flow between different market observables. This work has materialised in a paper published in *Chaos: An Interdisciplinary Journal of Nonlinear Science*, together with a dataset uploaded to SoBigData platform, and a blogpost.

- **Modelling time-varying interactions in complex systems: the score driven kinetic Ising model**

*Partners Involved:* SNS

A common issue when analysing real-world complex systems is that the interactions between the elements often change over time: this makes it difficult to find optimal models that describe this evolution and that can be estimated from data, particularly when the driving mechanisms are not known. Here we offer a new perspective on the development of models for time-varying interactions introducing a generalisation of the well-known Kinetic Ising Model (KIM), a minimalistic pairwise constant interactions model which has found applications in multiple scientific disciplines. Keeping arbitrary choices of dynamics to a minimum and seeking information theoretical optimality, the Score-Driven methodology lets us significantly increase the knowledge that can be extracted from data using the simple KIM. In particular, we first identify a parameter whose value at a given time can be directly associated with the local predictability of the dynamics. Then we introduce a method to dynamically learn the value of such a parameter from the data, without the need of specifying parametrically its dynamics. Finally, we extend our framework to disentangle different sources (e.g. endogenous vs exogenous) of predictability in real time. We apply our methodology to several complex systems including financial markets, temporal (social) networks, and neuronal populations. Our results show that the Score-Driven KIM produces insightful descriptions of the systems, allowing forecasting accuracy in real time as well as to separate different components of the dynamics. This provides a significant methodological improvement for data analysis in a wide range of disciplines.

#### **Data science and machine learning techniques in finance**

- **Ask “Who” not “What”: Bitcoin volatility forecasting with Twitter data**

*Partners Involved:* ETHZ

In this study, we focus on volatility predictions for a relatively new asset class of cryptocurrencies (in particular, Bitcoin) using deep learning representations of public social media data from Twitter. For the field work, we extracted semantic information and user interaction statistics from over 30 million Bitcoin-related tweets, in conjunction with 15-minute intraday price data over a 144-day horizon. Using this data, we built several deep learning architectures that utilised a combination of the gathered information. For all architectures, we conducted ablation studies to assess the influence of each component and feature set in our model. We found statistical evidences for the hypotheses that: (i) temporal convolutional networks perform significantly better than both autoregressive and other deep learning-based models in the literature, and (ii) the tweet author meta-information, even detached from the tweet itself, is a better predictor than the semantic content and tweet volume statistics. This paper has been accepted for the proceedings of the 16th ACM International Conference on Web Search and Data Mining, 2023. The data generated during study was also uploaded to the SoBigData platform, alongside a blogpost.

- **Financial news and information transfer**

*Partners involved:* ETHZ

We quantify the propagation and absorption of large-scale publicly available news articles from the World Wide Web to financial markets. To extract publicly available information, we use the news archives from the Common Crawl, a non-profit organisation that crawls a large part of the web. We develop a processing pipeline to identify news articles associated with the constituent companies in the S&P 500 index, an equity market index that measures the stock performance of US companies. Using machine learning techniques, we extract sentiment scores from the Common Crawl News data and employ tools from information theory to quantify the information transfer from public news

articles to the US stock market. Furthermore, we analyse and quantify the economic significance of the news-based information with a simple sentiment-based portfolio trading strategy. Our findings provide support for that information in publicly available news on the World Wide Web has a statistically and economically significant impact on events in financial markets.

- **Forecasting cryptocurrency volumes**

*Partners involved:* ETHZ, SNS

We study the problem of the intraday short-term volume forecasting in cryptocurrency multi-markets. The predictions are built by using transaction and order book data from different markets where the exchange takes place. Methodologically, we propose a temporal mixture ensemble, capable of adaptively exploiting, for the forecasting, different sources of data and providing a volume point estimate, as well as its uncertainty. We provide evidence of the clear outperformance of our model with respect to econometric models. Moreover, our model performs slightly better than Gradient Boosting Machine while having a much clearer interpretability of the results. Finally, we show that the above results are robust also when restricting the prediction analysis to each volume quartile.

- **Disentangling short-run and long-run components in multivariate sentiment dynamics of financial news**

*Partners Involved:* SNS

The digitalization of news and social media provides an unprecedented source to investigate the role of information on market dynamics. However, the observed sentiment time-series represent a noisy proxy of the true investor sentiment. Moreover, modelling the joint dynamics of different sentiment series can be beneficial for the assessment of their economic relevance. The main methodological contribution of this paper is twofold: (i) we filter the latent sentiment signals in a genuinely multivariate model; (ii) we propose a decomposition into a long-term random walk component, named long-term sentiment, and a short-term component driven by a stationary Vector Autoregressive process of order one, named short-term sentiment. The proposed framework is a dynamic factor model describing the joint evolution of the observed sentiments of a portfolio of assets. Empirically, we find that the long-term sentiment co-integrates with the market price factor, while the short-term sentiment captures transient and firm-specific swings. By means of quantile regressions, we assess the significance of the explanatory power of filtered present sentiment on future returns. Then, we demonstrate how the lagged relation can be successfully exploited in a portfolio allocation exercise.

## Network science techniques

- **Novel tools for analysis of time-varying networked systems: Non-Markovian temporal networks with auto- and cross-correlated link dynamics**

*Partners Involved:* SNS

Many of the biological, social and man-made networks around us are inherently dynamic, with their links switching on and off over time. The evolution of these networks is often observed to be non-Markovian, and the dynamics of their links are often correlated. Hence, to accurately model these networks, predict their evolution, and understand how information and other relevant quantities propagate over them, the inclusion of both memory and dynamical dependencies between links is key. In this article we introduce a general class of models of temporal networks based on discrete autoregressive processes for link dynamics. As a concrete and useful case study, we then concentrate on a specific model within this class, which allows us to generate temporal networks with a specified



underlying structural backbone, and with precise control over the dynamical dependencies between links and the strength and length of their memories. In this network model the presence of each link is influenced not only by its past activity, but also by the past activities of other links, as specified by a coupling matrix, which directly controls the causal relations, and hence the correlations, among links. We propose a maximum likelihood method for estimating the model's parameters from data, showing how the model allows a more realistic description of real-world temporal networks and also to predict their evolution. Due to the flexibility of maximum likelihood inference, we illustrate how to deal with heterogeneity and time-varying patterns, possibly including nonstationary network dynamics. We then use our network model to investigate the role that, both the features of memory and the type of correlations in the dynamics of links have on the properties of processes occurring over a temporal network. Namely, we study the speed of a spreading process, as measured by the time it takes for diffusion to reach equilibrium. Through both numerical simulations and analytical results, we are able to separate the roles of autocorrelations and neighborhood correlations in link dynamics, showing that not only is the speed of diffusion non monotonically dependent on the memory length, but also that correlations among neighboring links help to speed up the spreading process, while autocorrelations slow it back down. Our results have implications in the study of opinion formation, the modelling of social networks, and the spreading of epidemics through mobile populations.

- **Analysis of bank leverage via dynamical systems and deep neural networks**

*Partners Involved:* SNS

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 behaviour. 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 chaotic behaviour. 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.

- **Realised exponential random graphs with an application to the interbank network**

*Partners Involved:* SNS

Motivated by the increasing abundance of data describing real-world networks that exhibit dynamical features, we propose an extension of the Exponential Random Graph Models (ERGMs) that accommodates the time variation of its parameters. Inspired by the fast growing literature on Dynamic Conditional Score-driven models, each parameter evolves according to an updating rule driven by the score of the ERGM distribution. We demonstrate the flexibility of the score-driven ERGMs (SD-ERGMs), both as data generating processes and as filters, and we show the advantages of the dynamic version with respect to the static one. We discuss two applications to time-varying networks from financial and political systems. First, we consider the prediction of future links in the Italian interbank credit network. Second, we show that the SD-ERGM allows us to discriminate



between static or time-varying parameters when used to model the dynamics of the US congress co-voting network.

- **Gravity models of networks: integrating maximum-entropy and econometric approaches**

*Partners Involved: IMT*

The World Trade Web (WTW) is the network of international trade relationships among world countries. Characterising both the local link weights (observed trade volumes) and the global network structure (large-scale topology) of the WTW via a single model is still an open issue. While the traditional Gravity Model (GM) successfully replicates the observed trade volumes by employing macroeconomic properties such as GDP and geographic distance, it, unfortunately, predicts a fully connected network, thus returning a completely unrealistic topology of the WTW. To overcome this problem, two different classes of models have been introduced in econometrics and statistical physics. Econometric approaches interpret the traditional GM as the expected value of a probability distribution that can be chosen arbitrarily and tested against alternative distributions. Statistical physics approaches construct maximum-entropy probability distributions of (weighted) graphs from a chosen set of measurable structural constraints and test distributions resulting from different constraints. Here we compare and integrate the two approaches by considering a class of maximum-entropy models that can incorporate macroeconomic properties used in standard econometric models. We find that the integrated approach achieves a better performance than the purely econometric one. These results suggest that the maximum-entropy construction can serve as a viable econometric framework wherein extensive and intensive margins can be separately controlled for, by combining topological constraints and dyadic macroeconomic variables.

Analysis of socio-economic and financial systems

- **Quantifying knowledge spillovers from advances in negative emissions technologies**

*Partners Involved: SNS, SSSA*

Negative emissions technologies (NETs) feature prominently in most scenarios that halt climate change and deliver on the Paris Agreements 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 exclusive 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-à-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.

- **Reconstructing firm-level interactions: the Dutch input-output network**

*Partners Involved: IMT*

Recent crises have shown that the knowledge of the structure of input-output networks at the firm level is crucial when studying economic resilience from the microscopic point of view of firms that rewire their connections under supply and demand shocks. Unfortunately, empirical inter-firm network data are rarely accessible and protected by confidentiality. The available methods of network reconstruction from partial information, which have been devised for financial exposures, are inadequate for inter-firm relationships because they treat all pairs of nodes as potentially interacting, thereby overestimating the rewiring capabilities of the system. Here we use two big data sets of transactions in the Netherlands to represent a large portion of the Dutch inter-firm network and document the properties of one of the few analysed networks of this kind. We, then, introduce a generalised maximum-entropy reconstruction method that preserves the production function of each firm in the data, i.e. the input and output flows of each node for each product type. We confirm that the new method becomes increasingly more reliable as a finer product resolution is considered and can therefore be used as a generative model of inter-firm networks with fine production constraints. The likelihood of the model, being related to the entropy, proxies the rewiring capability of the system for a fixed input-output configuration.

- **Intergenerational transmission of educational and occupational preferences: evidence from Sweden**

*Partners involved:* PSE

*External Partners Involved:* Swedish Institute for Social Research at Stockholm University (co-affiliation)

Traditionally, economists have analysed the role of school grades, social networks, skills, neighbourhoods, and luck as the main determinants of educational and occupational choice (Becker et al., 2018; Chetty et al., 2022; Cunha and Heckman, 2007; Heckman and Mosso 2014; Fernandez and Rogerson, 1997). In addition, there is a vast literature in economics showing that parental preferences are important in explaining children's educational and occupational choices (Volland, 2013; Farre and Vella, 2013; Algan et al., 2022; Dohmen et al., 2012). In the specific context of migration, several studies in both sociology and economics have emphasised the importance of transmission of culture and norms of the home country of immigrants (Rocha et al., 2017; Alesina and Giuliano, 2015; Fernandez and Fogli, 2009). However, there still remains a significant gap in the literature in regards to identifying the effect of culture and norms of home country on education and occupation choices of second-generation migrants. Our project aims to fill this gap by analysing how norms surrounding the relative status of different occupations in countries of origin influence the occupational choice of second-generation migrants in the destination country, in this case, Sweden. Our secondary research questions try to shed light on the characteristics and mechanisms of the effects. Firstly, to what extent do these differences persist as time, since arrival and reception of a residence permit, increases? Secondly, to what extent are these differences affected by the degree of integration of their parents? Thirdly, to what extent do these effects vary by income group, socio-economic status, gender, and high schools that the children attend? Fourthly, to what extent are these results magnified or diminished through residential sorting into ethnic enclaves? Fifthly, what is the relative importance of maternal or paternal countries of origin?

## Analysis of Cryptocurrencies

- **The weighted Bitcoin lightning network**

Partners Involved: IMT

The Bitcoin Lightning Network (BLN) was launched in 2018 to scale up the number of transactions between Bitcoin owners. Although several contributions concerning the analysis of the BLN binary structure have recently appeared in the literature, the properties of its weighted counterpart are still largely unknown. The present contribution aims at filling this gap, by considering the Bitcoin Lightning Network over a period of 18 months, ranging from 12th January 2018 to 17th July 2019, and focusing on its weighted, undirected, daily snapshot representation. As the study of the BLN weighted structural properties reveals, it is becoming increasingly 'centralised' at different levels, just as its binary counterpart: 1) the Nakamoto coefficient shows that the percentage of nodes whose degrees/strengths “enclose” the 51% of the total number of links/total weight is rapidly decreasing; 2) the Gini coefficient confirms that several weighted centrality measures are becoming increasingly unevenly distributed; 3) the weighted BLN topology is becoming increasingly compatible with a core-periphery structure, with the largest nodes “by strength” constituting the core of such a network, whose size keeps shrinking as the BLN evolves. Further inspection of the resilience of the weighted BLN shows that removing such hubs leads to network fragmentation into many components, evidence indicating potential security threats - as the ones represented by the so-called “split attacks”.

## 4.2 Micro-projects

### **Analysis of Bitcoin crash in 2017/2018**

We will study dynamics of market microstructure of bitcoin in the period of 2017-2018 when there was a big surge in cryptocurrency prices, followed by a rapid drop. Our initial analysis shows that cryptomarkets were strongly ‘causally’ interlinked via markers of market liquidity, suggesting a field effect. A further analysis is needed to ensure sound statistics, and to verify this result on a larger scale (for a larger set of cryptocurrencies. Current analysis concentrated on Bitcoin only).

Status: Completed; Partners involved: ETHZ, SNS

Outputs: [publication](#), [dataset](#), [blog post](#)

### **The news tells us about peace - a dashboard for the visualisation of the Global Peace Index as captured through the GDELT news**

Creation of a dashboard for the visualisation of the Global Peace Index as captured through the GDELT news.

Status: Completed; Partners involved: CNR

Outputs: [application](#), [blog post](#).

### **Bitcoin volatility forecasting using Twitter data**

This microporject will concern T8.2 and T10.2. Its aim is to upload dataset, collected for the study of volatility predictions for an asset class of cryptocurrencies (in particular, Bitcoin) using deep learning representations of public social media data from Twitter. For the field work, we extracted semantic information and user interaction statistics from over 30 million Bitcoin-related tweets, in conjunction with 15-minute intraday price data over a 144-day horizon. This data would be uploaded to the platform.

Status: Completed; Partners involved: ETH

Outputs: [research paper](#), [dataset](#), [blog post](#).

**Risk forecasting with reservoir computers**

In the framework of recurrent neural networks, reservoir computing has been developed on the paradigm of training a simple readout mechanism, while reading the state of the recurrent neurons. The layer of the recurrent neurons defines the so-called reservoir, whose internal weights are fixed and randomly generated. Training only the readout layer of a reservoir computer is a remarkable simplification in terms of computational complexity. Nevertheless, the approximation properties are preserved for a large class of dynamic processes, once a few hyper-parameters are optimized. Interestingly, reservoir computers as approximants do not rely necessarily on the specific choice of mean squared error as loss function for the readout training. In fact, they can be devised to predict any moment of a dynamic process, in particular any quantile of the conditional distribution of a random process. To this end, the readout can be trained with the quantile loss function introduced for quantile regression (Konker and Bassett, *Econometrica* 1978). A forecasting study of the Value at Risk of a number of stocks for the US financial market show the merit of the devised methodology.

Status: Ongoing; Partners involved: SNS

Expected outputs: research paper, blog post.

**New Data and Methods for Migration Studies: Workshop (10-11 October at PSE)**

The Paris School of Economics, SoBigData++ consortium, HumMingBird consortium and Institut Convergences Migrations are jointly organizing a two-day workshop aimed at bringing together migration scholars from various disciplines from these institutions and beyond. The conference is devoted to investigating and showcasing new methods to study human migration based on non-traditional data sources and methods. Possible topics include: understanding and estimating migration flows and stocks using non-traditional and big data sources (e.g., nowcasting flows using social media or mobile phone data, visualisation, analysis, and prediction of flows, including for specific domains such as scientific, labour, refugee, and seasonal migration); studying the connection between policy changes and migration using new data analytics; understanding attitudes towards migrants and migrants' integration (e.g. using sentiment analysis, discourse type and media portraying of migrants, polarisation of the discourse with respect to human migration, etc.); ethics of big data in the context of human migration. Keynote speakers: Joshua Blumenstock (University of California Berkeley), Petra Molnar (University of Toronto), Stefano Iacus (Harvard University) and Ekaterina Zhuravskaya (Paris School of Economics). The conference will take place in the Paris School of Economics (Oct 10-11 2022).

Status: Ongoing; Partners involved: PSE, UNIP

External partners involved: HummingBird, Institut Convergences de Migrations

Outputs: A two-day workshop.

#### 4.3 Publications

1. Akbiyik, M. E., et al. "Ask" Who", Not" What": Bitcoin Volatility Forecasting with Twitter Data." arXiv preprint arXiv:2110.14317 (2021).
2. Vasiliauskaite, V., Lillo, F., & Antulov-Fantulin, N. (2022). Information dynamics of price and liquidity around the 2017 Bitcoin markets crash. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(4), 043123.

3. Rodikov, German, and Nino Antulov-Fantulin. "Can LSTM outperform volatility-econometric models?." arXiv preprint arXiv:2202.11581 (2022).
4. Rodikov, German, and Nino Antulov-Fantulin. "Volatility-inspired  $\sigma$ -LSTM cell." arXiv preprint arXiv:2205.07022 (2022).
5. Williams, O.E., Mazzarisi, P., Lillo, F. and Latora, V., 2022. Non-Markovian temporal networks with auto-and cross-correlated link dynamics. *Physical Review E*, 105(3), p.034301.
6. Lillo, F., Livieri, G., Marmi, S., Solomko, A. and Vaienti, S., 2021. Analysis of bank leverage via dynamical systems and deep neural networks. arXiv preprint arXiv:2104.04960.
7. Campajola, C., Di Gangi, D., Lillo, F. and Tantari, D., 2020. Modelling time-varying interactions in complex systems: The score driven kinetic ising model. arXiv preprint arXiv:2007.15545.
8. Buccheri, G. and Mazzarisi, P., 2021. Realized Exponential Random Graphs, with an Application to the Interbank Network. Available at SSRN.
9. Danilo Vassallo, Giacomo Bormetti, Fabrizio Lillo, A tale of two sentiment scales: Disentangling short-run and long-run components in multivariate sentiment dynamics *Quantitative Finance* (2022) doi.org/10.1080/14697688.2022.2119159
10. Domenico Di Gangi, Giacomo Bormetti, Fabrizio Lillo, Score-Driven Exponential Random Graphs: A New Class of Time-Varying Parameter Models for Dynamical Networks, <https://arxiv.org/abs/1905.10806>
11. Giorgio Tripodi, Francesco Lamperti, Roberto Mavilia, Andrea Mina, Francesca Chiaromonte, Fabrizio Lillo, Quantifying knowledge spillovers from advances in negative emissions technologies (2022)
12. Di Vece, Marzio, Diego Garlaschelli, and Tiziano Squartini. "Gravity models of networks: integrating maximum-entropy and econometric approaches." *Physical Review Research* 4.3 (2022): 033105.
13. Lin, Jian-Hong, et al. "The weighted Bitcoin Lightning Network." *Chaos, Solitons & Fractals* 164 (2022): 112620.
14. Ialongo, Leonardo Niccolò, et al. "Reconstructing firm-level interactions: the Dutch input-output network." *Scientific Reports* 12 (2022): 11847.
15. Pieroni, Valentina, Angelo Facchini, and Massimo Riccaboni. "COVID-19 vaccination and unemployment risk: Lessons from the Italian crisis." *Scientific reports* 11.1 (2021): 1-8.
16. Jazbec Metod, Pásztor Barna, Faltings Felix, Antulov-Fantulin Nino and Kolm Petter N. 2021, On the impact of publicly available news and information transfer to financial markets *R. Soc. open sci.* 8202321202321
17. Antulov-Fantulin, N., Guo, T. & Lillo, F. Temporal mixture ensemble models for probabilistic forecasting of intraday cryptocurrency volume. *Decisions Econ Finan* 44, 905–940 (2021).

#### 4.4 Planned Activities for next period

##### 4.4.1 Data Collection activities

*Partners involved:* PSE

Skin-tone based diversity in the context of football. Data collection on valuation of players starting this month.

#### 4.4.2 Software Development activities

*Partners involved:* IMT

Integration of the NEMTROPY package with additional subroutines.

*Partners involved:* ETHZ

Integration of the algorithm to be developed in 3.4.2 for learning complex dynamics on complex networks.

#### 4.4.3 Events

*Partners involved:* SNS

2nd Edition of the School “Machine Learning of Dynamic processes and Time Series Analysis” (MLDYN2022). The aim of the School is to present recent data-driven approaches to Machine Learning. In particular, this year the School will focus on the mathematical and computational aspects of Rough Paths and Signatures, Reinforcement Learning, Generative Adversarial Networks, and Continual Learning. Four keynote speakers will hold minicourses. There will also be a limited number of contributed talks.

#### 4.4.4 Scientific activities

*Partners involved:* ETHZ

Learning complex dynamics on complex networks using differential deep learning. We plan to propose a technique to learn the dynamics on a complex network from the observed time series, with and without a known network structure.

*Partners involved:* IMT

A paper on the randomization of hypergraphs. We plan to extend the toolbox of maximum-entropy (ME) techniques to analyse higher-order interactions. To this aim, we plan to individuate a tabular representation of hypergraphs, define a proper set of constraints for such a representation and apply the ME techniques on it.

*Partners involved:* IMT

A paper on signed networks. We plan to develop novel (theoretical as well as numerical) tools for the analysis of these objects. Besides, we plan to verify theories such as the (strong and weak) balance and the status ones on real-world data.

*Partners involved:* IMT

A paper on the Carbon Trade Network. We plan to study the network of trade-embedded carbon exchanges and compare its structural features with those of the International Trade Network.

*Partners involved: IMT*

A paper on bipartite, interfirm networks. We plan to analyse the topological structure of such systems by considering the case-study of the automotive sector. We will also test their reconstructability by employing maximum-entropy techniques.

*Partner involved: SNS*

A paper on the joint relation between liquidity of assets and leverage of banks in the systemic stability of the financial system. This makes use of the modelling framework developed in Lillo, F., Livieri, G., Marmi, S., Solomko, A. and Vaienti, S., 2021. Analysis of bank leverage via dynamical systems and deep neural networks and it is based on dynamical system theory and neural networks for the empirical analysis.

*Partner involved: SNS*

A paper on the use of reservoir computers for risk forecasting. While reservoir computers are typically used to forecast expected values of a variable, we develop a method to forecast quantiles of a distribution. This is extremely important from a risk management perspective, since most of the risk metrics can be linked with quantiles. The main application field will be finance (e.g. assessment of financial systemic risk, portfolio optimization, etc) even if the method is quite generic and applicable in other fields where one wants to forecast extreme events.

*Partner involved: PSE*

Continuation of the intergenerational transmission of cultural preferences surrounding educational and occupational choices.

## 5 T10.3 Sustainable Cities for Citizens

This exploratory focuses on the analysis of cities, the sustainability of their flows of energy and materials and people living in them. We analysed data from different spatial and temporal scales. On city-wide scales, we analysed energy and material flows to give insights on the sustainability of transformation processes occurring in cities (the so-called "urban metabolism") and point out the circularity of flows and main polluting/GHG emission sectors and factors. On a small scale, we analysed mobility in different cities, allowing the characterization of the demand of dynamic users and granting the derivation of models to estimation pollution and optimise the electric mobility charging and relocation service and minimise its impact on the power grid.

### 5.1 Activities Report

#### **Optimal Planning of Regional Renewable energy sources**

*Partners Involved:* IMT

Description: Following the case study in Tuscany, the activity is now focusing to collect data on renewable sources in Regione Marche. Data collection includes data of solar radiation and wind (to estimate the generation potential in the region). At macro-regional level (Toscana and Marche), the research group has collected data on electricity consumption, generation and price for the year 2019. Marche and Toscana are a specific area for the Italian electricity market. Collecting data on wind and solar radiation allows for the planning of renewable sources in a way that integrates the generation potential of the region together with the pricing signals and the actual generation of the existing power plants.

#### **Urban metabolism and circularity of the municipality of Albavilla**

*Partners Involved:* IMT

Description: Data on energy, materials, mobility, and policy for circular economy have been collected for the municipality of Albavilla. The data analysis is ongoing and a specific set of indicators has been implemented in cooperation with EnelX.

#### **A network simplification to ease topological studies about the food-web architecture**

*Partners Involved:* IMT

Network data about trophic interaction across different large locations and ecosystems are scarce in comparison with general ecological data, especially if we consider terrestrial habitats. The activity has developed a method based on complex networks to ease the gathering of the information by simplifying the collection of data with a taxonomic key. The food web simplification retains most of the general topological indices providing support especially for scientists that are new in this field and for exploratory analysis. The research focuses on urban food webs.



## Potential ecological network of Tuscany

*Partners Involved:* IMT

This data paper describes the building process behind an in silico ecological network of Tuscany (Italy) species and their different interactions as an OPEN (Open Potential Ecological Network). The data sources consist of already open and available databases from experts and citizen science projects. The final dataset is also completely open and ready to use as well as the pre-built potential ecological network. We described the methods used for collecting, cleaning and joining the original database, and summarised the network building process. We also present a fast way to visualise and interact with this potential ecological network. Data and methods will be integrated in the SoBigData platform (the integration process is ongoing).

## 5.2 Micro-projects

### Municipality Transition readiness of Spain – Data collection

The aim of the microproject is to collect the OPEN data to compute the Municipality Transition Index (a modification of ref. <https://arxiv.org/abs/2109.10832>) in Spain.

Status: Ongoing; Partners involved: IMT, BSC, ELI

Expected outcomes: Dataset; Technical report; blog post.

### Mobility-driven Segregation Models

Despite the vast amount of works built upon the Schelling model, literature still lacks a Schelling-like segregation model capable of including some aspects related to the mobility patterns of the agents: in almost all works in the literature, unhappy agents move in all or part of the grid, randomly choosing their future location. However, in the last half century, various theories and models about how humans move on the geographic space have been proposed, such as the exploration and preferential return model, the gravity and the radiation models, and their extensions. We aim to fill this gap by developing and characterising a mobility-informed Schelling model, considering well-known mobility laws when modelling the choice of a place in which an agent chooses to relocate. Varying the initialization parameters and the initial configuration of our mobility-informed models, we aim to observe and measure whether and how segregation dynamics change with the mobility rules. Our implementation of these models is based on a library we develop on purpose, which allows us to define a python class for a Schelling-like model in which to define rules that regulate how agents can move on the artificial world. (<https://github.com/dgambit/mob-schelling>)

Status: Ongoing; Partners involved: CNR, UNIFI

Expected outcomes: Method item, Experiment item, Blog Post

### Generating Synthetic Mobility Networks with Generative Adversarial Networks

Description: The increasingly crucial role of human displacements in complex societal phenomena, such as traffic congestion, segregation, and the diffusion of epidemics, is attracting the interest of scientists from several disciplines. In this article, we address mobility network generation, i.e., generating a city's entire mobility network, a weighted directed graph in which nodes are geographic locations and weighted edges represent people's movements between those locations, thus describing the entire mobility set flows within a city. Our solution is MoGAN, a model based on Generative Adversarial Networks (GANs) to generate realistic mobility networks. We conduct extensive experiments on public datasets of bike and taxi rides to show that MoGAN outperforms the classical Gravity and Radiation models regarding the realism of the

generated networks. Our model can be used for data augmentation and performing simulations and what-if analysis.

Status: Completed; Partners Involved: IMT, CNR

Outcomes: [method](#), [experiment](#), [blog post](#).

### **Mobility-emissions**

A collection of methods (relying on the scikit-mobility and OSMnx Python libraries) to compute emissions starting from vehicles' mobility trajectories. In particular, methods have been developed to: - map-matching: match points of a TrajDataFrame to the edges (roads) or nodes (crossroads) of a road network took from OpenStreetMap; - speed/acceleration computation: compute the values of speed and acceleration in each point of the TrajDataFrame; trajectory time filtering: filter the points of the TrajDataFrame s.t. the time interval between the points of the resulting sub-trajectories is not greater than a threshold; - computation of emissions: compute the instantaneous emissions of 4 air pollutants (CO<sub>2</sub>, NO<sub>x</sub>, PM, VOC) in each point of the TrajDataFrame. The methods are available in a [GitHub Repository](#). We used the mobility-emission methods (<https://data.d4science.org/ctlg/ResourceCatalogue/mobility-emissions>) to perform experiments on real GPS trajectories describing 433,272 trips from 14,907 private vehicles moving in Greater London, Rome, and Florence throughout January 2017. We find that, for all three cities, emissions are distributed across vehicles in a heterogeneous way: a few vehicles, that we call gross polluters, are responsible for a tremendous amount of emissions. At the same time, most of them emit significantly less. Also, a few grossly polluted roads suffer from a significant quantity of emissions, while most of them suffer significantly fewer emissions. Both the distributions of emissions across the vehicles and across the roads are well approximated by heavy-tailed distributions (mainly a truncated power law or a stretched exponential, with parameters changing with the city). We also investigated the relationship between a vehicle's emissions and its mobility behaviour, from one side, and between the emissions suffered by a road and its network features, discovering that gross polluters tend to be more regular and predictable in their mobility than low-emitting vehicles, and confirming that the most polluted roads are the ones that fall more frequently on the shortest paths connecting two nodes of the network. Finally, as reducing emissions is a growing concern for cities, and estimating the impact of policies targeting vehicles to reduce their footprint on the city's environment is crucial, we investigate the impact that (i) the vehicles' electrification and (ii) the home working have on the total amount of emissions and the distribution of emissions across the roads. We find that, for example, the electrification of just the top 1% gross polluters moving in Rome would lead to the same reduction of the CO<sub>2</sub> emitted overall as electrifying 10% random vehicles, and that the remote working of the top 1% gross polluters would lead to the same reduction reached if they were ~4% random vehicles. Similar results hold for the other cities, even if with slightly lower numbers for London.

Status: Completed; Partners involved: UNIROMA1, CNR, BSC

Outcomes: [method](#), [experiment](#), [blog post](#).

### **How Routing Strategies Impact Urban Emissions**

Navigation apps use routing algorithms to suggest the best path to reach a user's desired destination. Although undoubtedly useful, navigation apps' impact on the urban environment (e.g., CO<sub>2</sub> emissions and pollution) is still largely unclear. In this work, we design a simulation framework to assess the impact of routing algorithms on carbon dioxide emissions within an urban environment. Using APIs from TomTom and OpenStreetMap, we find that settings in which either all vehicles or none of them follow a navigation app's suggestion lead to the worst impact in terms of CO<sub>2</sub> emissions. In contrast, when just a portion (around half) of vehicles follow these suggestions, and some degree of randomness is added to the remaining vehicles'

paths, we observe a reduction in the overall CO2 emissions over the road network. Our work is a first step towards designing next-generation routing principles that may increase urban well-being while satisfying individual needs.

Status: Ongoing; Partners involved: UNIFI, CNR, UNIROMA1, IMT  
Expected outcomes: Method, Experiment, Blog Post.

#### **Connected Vehicle Simulation Framework for Parking Occupancy Prediction (Demo Paper)**

This work demonstrates a simulation framework that collects data about connected vehicles' locations and surroundings in a realistic traffic scenario. Our focus lies on the capability to detect parking spots and their occupancy status. We use this data to train machine learning models that predict parking occupancy levels of specific areas in the city center of San Francisco. By comparing their performance to a given ground truth, our results show that it is possible to use simulated connected vehicle data as a base for prototyping meaningful AI-based applications.

Status: Ongoing; Partners involved: UNIFI, CNR  
Expected outcomes: Method, Experiment, Blog Post.

#### **Estimating countries' peace index with GDELT**

We use news media attention from the Global Data of Events, Location and Tone (GDELT) database, as a proxy for estimating Global Peace Index (GPI), to complement the knowledge obtained from the traditional data sources, and overcome their limitations. Considering that GDELT is updated daily, and it is a free access database, it can contribute to the estimation of GPI at a higher frequency, i.e. at a monthly level, as compared to the official GPI which is updated at a yearly level. Besides, GPI through GDELT is produced at a low cost, and in a time-efficient way, as compared to the traditional methodology.

Status: Completed; Partners involved: CNR, UNIFI, SNS  
Outcomes: [method](#), [experiment](#), [blog post](#).

#### **Explaining urban vehicle emissions in the city of Rome**

Urban emissions represent a major antagonist to cities' livability. In this work, we investigate aggregated emissions coming from thousands of vehicles using both spatial and non-spatial models. We compare such models in terms of performance and interpretation, showing not only that a spatial model has better performances in predicting, but also that it is very powerful in explaining the roles played by the predictors in different areas of the city. In particular, we find that the CO<sub>2</sub> emissions in the city of Rome are mainly due to the presence of main arterial roads, but also to the population density, and to the road network density in general. Moreover, the importance of such factors changes within different areas of the city. While being an experimental proof of spatial regression's efficacy when modelling urban emissions, our work may help in understanding the relationship between urban features and the spatial variability of vehicle emissions in the city of Rome.

Status: Ongoing; Partners involved: UNIROMA1, CNR, UNIFI, IMT  
Expected outcomes: preprint paper, blog post.

**Bottom-up Street View: Data collection strategy for crowdsourcing street view imagery in Pristina, Kosovo**

"The Micro-Project 'Bottom-up Street View' aims to create guidelines for the citizen-led collection of street view imagery. The City of Pristina was selected as a case study given that it remains relatively unexplored by Google Maps' Street View. The proposed guidelines will cater to pedestrians and cyclists, exposing public spaces such as parks and pedestrian areas that are often omitted from Google Maps' Street View. Main activities foreseen will include:

- Creating hardware and software guidelines for collecting street view imagery;
- Recruiting local volunteers and organizing training sessions on data collection method;
- Reviewing sample footage collected by local volunteers; and
- Comparison of crowdsourced street view imagery vs. Google Maps' street view imagery in terms of fitness for computer vision models.

Google Maps' Street View App will be used to georeference, post-produce and access street view imagery dataset.

Status: Completed; Partners involved: CRA

Outcomes: [dataset](#), [presentation](#), [blog post](#).

**Implementation of a python library for direct and indirect discrimination prevention in data mining**

In sociology, discrimination consist of unfairy treating of citizens in basis of their belonging to a specific group. Discrimination often leads denying to members of one group opportunities that are available to other groups. Services in the information society allow for automatic collection of data for data mining purposes. Those data are often used to train association/classification rules in view of making automated decisions. However, if the training data are inherently biased for or against a particular community of citizens, the learned model may show a discrimination prejudiced behavior. Discovering and eliminating such potential biases from the training data without harming their decision making utility is therefore highly desirable. Discrimination can be direct or indirect if decisions are based, respectively, on sensitive attributes or non-sensitive attributes strongly correlated with biased sensitive attributes. Antidiscrimination techniques include discrimination discovery and prevention. This micro-project will develop a python library implementing the discrimination prevention method proposed in the paper: "A methodology for direct and indirect discrimination prevention in data mining". The method evaluates and treats data sets of individuals removing direct and/or indirect discrimination biases preserving data quality. The developed python library will include data transformation methods, based on measures for both direct and indirect discrimination and can deal with several discriminatory attributes. For this purpose, the library will include metrics that specify which records should be changed and how those records should be changed during the data transformation for discrimination prevention. In addition, we will also implement the necessary metrics to measure the information loss and discrimination risk resulting of the data transformation process.

Status: Completed; Partners involved: URV

Outcomes: [python library](#), [blog post](#).

### 5.3 Publications

1. Gini, A., Re, S. & Facchini, A. A network simplification approach to ease topological studies about the food-web architecture. Sci Rep 12, 13948 (2022). <https://doi.org/10.1038/s41598-022-17508-1>

2. Pappalardo, L., Simini, F., Barlacchi, G., & Pellungrini, R. (2019). scikit-mobility: A Python library for the analysis, generation and risk assessment of mobility data. arXiv preprint arXiv:1907.07062.
3. Böhm, M., Nanni, M., & Pappalardo, L. (2022). Gross polluters and vehicle emissions reduction. *Nature Sustainability*, 1-9
4. Lucchini, L., Centellegher, S., Pappalardo, L., Gallotti, R., Privitera, F., Lepri, B., & De Nadai, M. (2021). Living in a pandemic: changes in mobility routines, social activity and adherence to COVID-19 protective measures. *Scientific reports*, 11(1), 1-12.
5. Pappalardo, L., Ferres, L., Sacasa, M., Cattuto, C., & Bravo, L. (2021). Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. *EPJ data science*, 10(1), 29.
6. Luca, M., Barlacchi, G., Lepri, B., & Pappalardo, L. (2021). A survey on deep learning for human mobility. *ACM Computing Surveys (CSUR)*, 55(1), 1-44.
7. Simini, F., Barlacchi, G., Luca, M., & Pappalardo, L. (2021). A Deep Gravity model for mobility flows generation. *Nature communications*, 12(1), 1-13.
8. Cornacchia, G., & Pappalardo, L. (2021). A Mechanistic Data-Driven Approach to Synthesize Human Mobility Considering the Spatial, Temporal, and Social Dimensions Together. *ISPRS International Journal of Geo-Information*, 10(9), 599.
9. Voukelatou, V., Gabrielli, L., Miliou, I., Cresci, S., Sharma, R., Tesconi, M., & Pappalardo, L. (2021). Measuring objective and subjective well-being: dimensions and data sources. *International Journal of Data Science and Analytics*, 11(4), 279-309.
10. Sîrbu, A., Andrienko, G., Andrienko, N., Boldrini, C., Conti, M., Giannotti, F., ... & Sharma, R. (2021). Human migration: the big data perspective. *International Journal of Data Science and Analytics*, 11(4), 341-360.
11. Andrienko, G., Andrienko, N., Boldrini, C., Caldarelli, G., Cintia, P., Cresci, S., ... & Trasarti, R. (2021). (So) Big Data and the transformation of the city. *International Journal of Data Science and Analytics*, 11(4), 311-340.
12. Cornacchia, G., Böhm, M., Mauro, G., Nanni, M., Pedreschi, D., & Pappalardo, L. (2022). How Routing Strategies Impact Urban Emissions. 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22), November 1–4, 2022, Seattle, WA, USA.
13. Resce, P., Vorwerk, L., Han, Z., Cornacchia, G., Alamdari, O.I., Nanni, M., Pappalardo, L., Weimer, L., & Liu, Y. (2022). Connected Vehicle Simulation Framework for Parking Occupancy Prediction (Demo Paper). 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22), November 1–4, 2022, Seattle, WA, USA.
14. Muscillo, Alessio, et al. "Circular City Index: An Open Data analysis to assess the urban circularity preparedness of cities to address the green transition--A study on the Italian municipalities." arXiv preprint arXiv:2109.10832 (2021).
15. Surmonte, Francesco, et al. "A Data-driven approach to renewable energy source planning at regional level." *Energy Sources, Part B: Economics, Planning, and Policy* 16.11-12 (2021): 1064-1075.
16. Korjani, Saman, et al. "Battery management for energy communities—Economic evaluation of an artificial intelligence-led system." *Journal of Cleaner Production* 314 (2021): 128017.
17. Pardo Martínez, Clara Inés, et al. "Trends and dynamics of material and energy flows in an urban context: a case study of a city with an emerging economy." *Energy, Sustainability and Society* 11.1 (2021): 1-15.

## 5.4 Planned Activities for next period

### 5.4.1 Data Collection activities

Partners involved: IMT, BSC

1. Data collection on potential Renewable energy sources in the macro-region Toscana and Marche
2. Data collection on energy, mobility, waste, and digitalisation of all the Spanish municipalities

#### 5.4.2 Events

**Energy and Finance 9 Workshop** (Milan, 8-10 Feb. 2023) Sponsorship of the "SoBigData Prize for the best paper" presented by a young researcher.

Partners involved: IMT

The Conference puts together researchers and practitioners working in all areas of Energy-Finance & Climate-Change.

#### 5.4.3 Scientific activities

##### **The trade network of CO2 in the global supply chain**

Partners involved: IMT

The activity (in convergence with Task 10.2) studies the emissions of CO2 embedded in the international trade. A network-based analysis is performed to better characterise the responsibility for emission of heavy-producing and heavy consumer countries. The aim of the study is to:

1. Characterise the phenomenon of carbon leakage (i.e. the geographic shift of heavy-polluting activities to other countries)
2. Understand the impact of the carbon border adjusting mechanism on the supply chain of the EU area

## 6 T10.4 Migration Studies

This exploratory studies how big data can help understand the migration phenomenon. Our scientists will try to answer various questions about migration in Europe and the world. Several studies are ongoing, including developing economic models of migration, now-casting migration stocks and flows, identifying the perception of migration and effect on the leaving and the receiving communities. We will also study the effect of migrants' networks (through the ego network graph abstraction) on the different migration phases (i.e., migration choices as well as cultural assimilation and transnationalism).

### 6.1 Activities Report

#### **Multi-aspect Integrated Migration Indicators (MIMI) dataset**

*Partners Involved:* UNIPi

External Partners (collabs): HumMingBird H2020 EU Project

Description: We propose a dataset to be exploited in migration studies as a concrete example of this new integration-oriented approach: the Multi-aspect Integrated Migration Indicators (MIMI) dataset (10.5281/zenodo.6360650). It includes official data about bidirectional human migration (traditional flow and stock data) with multidisciplinary variables and original indicators, including economic, demographic, cultural, and geographic indicators, together with the Facebook Social Connectedness Index (SCI). The paper "Dataset of Multi-aspect Integrated Migration Indicators" (Goglia D., Pollacci L., Sirbu A.) is currently under revision at MDPI Data (10.5281/zenodo.6500884).

#### **Building the Enhanced Microsoft Academic Knowledge Graph (EMAKG)**

*Partners Involved:* UNPI

External Partners (collabs): HumMingBird H2020 EU Project

Description: We built the Enhanced Microsoft Academic Knowledge Graph (EMAKG) (10.5281/zenodo.5888646), an expansion dataset of the Microsoft Academic Knowledge Graph (MAKG), a large knowledge graph of scholarly data (<https://makg.org/>, Faber, 2019). The additions were obtained both by integrating data from other resources, such as Wikipedia and MAG, and through natural processing techniques applied to the data of the MAKG itself. The work carried out aims to expand existing knowledge and facilitate research in various fields of study. For instance, geographical information could help mobility (and migration) research.

The data-paper "EMAKG: An Enhanced Version Of The Microsoft Academic Knowledge Graph" (Pollacci L.) is currently under revision at MDPI Data (<https://arxiv.org/pdf/2203.09159.pdf>).

#### **Studying mobility between the UK and Europe to study the effect of Brexit**

*Partners Involved:* UNIPi

*External Partners (collabs):* Harvard University, Joint Research Centre, and Max Planck Institute for Demographic Research

Description: We are studying mobility between the UK and Europe in the last decades to obtain a global view of mobility trends from different viewpoints and to understand whether and how Brexit has affected



mobility. The analysis is based on several datasets covering different types of mobility and demographic groups for an integrated quantitative image, i.e., air traffic volumes by Sabre Corporation ([Market Intelligence - Global Demand Data](#)), Twitter, Eurostat official statistics, Crunchbase, and scholarly data from Enhanced Microsoft Academic Knowledge Graph and Microsoft Academic Knowledge Graph (<https://makg.org/>, Faber, 2019). For each dataset we use two novel mobility indicators. The *Flow Ratio* shows the ratio between the number of people moving into the UK and the number of people moving out of the UK, and gives a point-wise indication of the balance between the two flows. The second one, the *Cumulative Flow Ratio*, gives an indication of the balance between the total number of incoming and outgoing people in the entire period analysed.

### **Crunchbase data**

*Partners Involved:* UNIPi

Description: We obtained data from [Crunchbase](#) through the [Crunchbase academic research access program](#). This allows us to download snapshot user and company data in the form of .csv files. We have employed these files to extract highly-skilled migration stocks and flows by aggregating users based on their estimated nationality and the headquarters of their workplaces. Furthermore, we validate the flows and stocks obtained against the official statistics of Eurostat and the United Nations. The dataset was also used to analyse possible Brexit effects on professionals.

### **Cultural Exchange and high skilled mobility studies**

*Partners Involved:* UNIPi

Description: We analysed cultural exchange and highly-skilled mobility using innovative data from the established professional-oriented social network LinkedIn. The investigation focuses on alumni from Tuscan Universities of Pisa, Siena and Florence, Scuola Normale Superiore, and Scuola Superiore Sant'Anna. By collecting data from LinkedIn, we describe educational and professional paths and study the international cultural exchange of Tuscan alumni. We explore the potential of social networks to study highly-skilled mobility to bridge traditional data limitations. The study is presented in a [master thesis](#) entitled "LinkedIn Data: uno sguardo attraverso LinkedIn alle carriere degli alumni e delle alumnae degli atenei toscani" (eng: "LinkedIn Data: a look through LinkedIn at the careers of alumni and alumnae of Tuscan universities"). The extended abstract "Cultural exchange and high-skilled mobility from LinkedIn: the case of Tuscan alumni" (Pollacci L., Di Mauro G., Rossetti G.) has been presented at the "New data and methods for migration studies: going beyond traditional data sources" workshop (Paris, 10-11 Oct 2022).

### **Measurement of immigrants' adoption of natives' shopping consumption**

*Partners Involved:* UNIPi, PSE

*External Partners (collab):* Université Clermont Auvergne

Description: We are deepening the study of the measurement of immigrants' adoption of natives' shopping consumption presented in Guidotti, R., et al. 2018 ("Discovering temporal regularities in retail customers' shopping behaviour." *EPJ Data Science* 7 (2018): 1-26.) and in Guidotti, R., et al. 2020 ("Measuring Immigrants Adoption of Natives Shopping Consumption with Machine Learning." *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V*. Springer International Publishing, 2021.).



For the time being, we tested the TINCA methodology on several customer groups, including those created using synthetic data models such as Synthetic Data Vault (SDV), i.e., GaussianCopula, CTGAN, TVAE, and CopulaGAN models. Furthermore, we have added new classifiers, such as LightGBM, a gradient boosting framework that uses tree-based learning algorithms, including neural networks, such as the Multi-layer Perceptron classifier.

### **Intergenerational transmission of educational and occupational preferences: evidence from Sweden**

*Partners Involved:* PSE

*External Partners (affiliation for data access):* Swedish Institute for Social Research (SOFI) at Stockholm University

Description: Traditionally, economists have analysed the role of school grades, social networks, skills, neighbourhoods, and luck as the main determinants of educational and occupational choice. In addition, there is a vast literature in economics showing that parental preferences are important in explaining children's educational and occupational choices. In the specific context of migration, several studies in both sociology and economics have emphasised the importance of transmission of culture and norms of the home country of immigrants. However, there still remains a significant gap in the literature in regards to identifying the effect of culture and norms of home country on education and occupation choices of second-generation migrants. This study aims to fill this gap by analysing how norms surrounding the relative status of different occupations in countries of origin influence the occupational choice of second-generation migrants in the destination country, in this case, Sweden. Our secondary research questions try to shed light on the characteristics and mechanisms of the effects. Firstly, to what extent do these differences persist as time, since arrival and reception of a residence permit, increases? Secondly, to what extent are these differences affected by the degree of integration of their parents? Thirdly, to what extent do these effects vary by income group, socio-economic status, gender, and high schools that the children attend? Fourthly, to what extent are these results magnified or diminished through residential sorting into ethnic enclaves? Fifthly, what is the relative importance of maternal or paternal countries of origin? We currently obtained ethics approval for our data, submitted the data application, and have received funding from the data from the National Research Agency (ANR-17-EURE-001).

### **Value of Color: Skin-tone based discrimination in football**

*Partners Involved:* PSE

Description: We collect data on valuation of players in the major leagues in football. We have created an algorithm to classify players based on skin-tone and ethnicity to test the prevalence and the effect of skin-tone based discrimination on the valuation of players in football. These valuations are offered by a platform that is supported by the fans (i.e. their own valuation) and is used as a main reference point for managers, thus we see the contribution of the "common man" to the valuation of such football players, i.e. "assets". We create an economic model to show this role of bottom-up based discrimination in the valuation of assets. Currently done with the algorithm, and collecting the data.

**Academic Migration and Networks: Academic Markets in the Soviet Union**

*Partners Involved:* PSE, UNIP

External Partners (collabs): UCLouvain

Description: Continue improving the work on academic migration and academic networks using evidence from the Iron Curtain. Work on studying the presence of a “market” in the centrally planned Soviet Union using the model of De la Croix et al. (2020).

**Research on migrants’ social networks and cultural assimilation**

*Partners involved:* AALTO

Description: a) Our activities included using large-scale call detail records of anonymised mobile phone service subscribers with demographic and location information to study changes in personal networks [1,2]. Roy C. presented the study [1] in CCS, Mallorca, October 22. Also, Fudolig M. presented the study [3] (reported in RP1) in CCS.

(b) Utilising geocoded population register data we are currently studying the assimilation of migrants in terms of the structure of the urban landscapes in Finland (ongoing).

**Modelling of social spreading of Covid-19**

*Partners involved:* AALTO

Description: We have also been studying the dynamics of epidemic spreading with the traditional compartmental approach and using data-driven models. We have used both open data on covid and aggregate level data derived from mobile phone communication to study the changes in mobility patterns in Finland during 2019-21 (in preparation).

## 6.2 Micro-projects

**New Data and Methods for Migration Studies: Workshop at PSE Oct 10-11**

Description: Paris School of Economics, SoBigData++ consortium, HumMingBird consortium and Institut Convergences Migrations are jointly organising a two-day [workshop](#) aimed at bringing together migration scholars from various disciplines from these institutions and beyond. The conference is devoted to investigating and showcasing new methods to study human migration based on non-traditional data sources and methods.

Status: Ongoing; Partners: PSE, UNIP

External Partners (collab): HummingBird, Institute Convergences de Migrations

Expected outcomes: A two-day workshop aimed at bringing together migration scholars from various disciplines from these institutions and beyond.

**Self-Rated Health Among Italian Immigrants Living in Norway: A Cross-Sectional Study**

Most studies on immigrant health focus on immigrant groups coming from extra-European and/or low-income countries. Little attention is given to self-rated health (SRH) in the context EU/EEA migration. To know more about health among European immigrants can provide new insights related to social determinants of health in the migration context. Using the case of Italian immigrants in Norway, the aim of this study was to (i) examine the levels of SRH among Italian immigrants in Norway as compared with the Norwegian and the

Italian population, (ii) examine the extent to which the Italian immigrant perceived that moving to Norway had a positive or negative impact on their SRH; and (iii) identify the most important factors predicting SRH among Italian immigrants in Norway.

Status: Completed; Partners involved: UniPi.

External partners involved: Oslo Metropolitan University, University of South-Eastern Norway.

Outcomes: [research paper](#), [blog post](#), [experiment](#).

### **Generating Synthetic Mobility Networks with Generative Adversarial Networks**

The increasingly crucial role of human displacements in complex societal phenomena, such as traffic congestion, segregation, and the diffusion of epidemics, is attracting the interest of scientists from several disciplines. In this article, we address mobility network generation, i.e., generating a city's entire mobility network, a weighted directed graph in which nodes are geographic locations and weighted edges represent people's movements between those locations, thus describing the entire mobility set flows within a city. Our solution is MoGAN, a model based on Generative Adversarial Networks (GANs) to generate realistic mobility networks. We conduct extensive experiments on public datasets of bike and taxi rides to show that MoGAN outperforms the classical Gravity and Radiation models regarding the realism of the generated networks. Our model can be used for data augmentation and performing simulations and what-if analysis.

Status: Completed; Partners Involved: IMT, CNR

Outcomes: [method](#), [experiment](#), [blog post](#).

### **Physical Activity Levels and Perceived Changes in the Context of Intra-EEA Migration: A Study on Italian Immigrants in Norway**

This study investigated:

- i) the extent to which Italian immigrants in Norway perceive that moving had a negative or positive impact on their PA;
- ii) possible differences between the PA of the Italian immigrants compared with the Norwegian population;
- iii) possible associations of the Italian immigrants' PA with key sociodemographic characteristics (gender, age, region of residence, and educational level).

Status: Completed; Partners involved: UniPi

External partners involved: University of South-Eastern Norway.

Outcomes: [blog post](#), [experiment](#), [paper](#).

## **6.3 Publications**

1. Kim, Jisu, et al. "Characterising different communities of Twitter users: Migrants and natives." International Conference on Complex Networks and Their Applications. Springer, Cham, 2021.
2. Dataset of Multi-aspect Integrated Migration Indicators (Goglia D., Pollacci L., Sirbu A. - UNIPi) is currently under revision at MDPI Data (10.5281/zenodo.6500884).
3. EMAKG: An Enhanced Version Of The Microsoft Academic Knowledge Graph (Pollacci L., UNIPi) is currently under revision at MDPI Data (<https://arxiv.org/pdf/2203.09159.pdf>).

4. Roy, C., Bhattacharya, K., Dunbar, R.I.M., & Kaski, K. K. (2022), Turnover in close friendships. Sci Rep 12, 11018 (2022).
5. Kuikka, V., Monsivais, D., & Kaski, K. K. (2022). Influence spreading model in analysing ego-centric social networks. Physica A: Statistical Mechanics and its Applications, 588, 126524.
6. Fudolig, M. I. D., Monsivais, D., Bhattacharya, K., Jo, H. H., & Kaski, K., (2021), Internal migration and mobile communication patterns among pairs with strong ties EPJ Data Science, 10(1), 16.
7. Snellman, J. E., Barrio, R. A., Kaski, K. K., & Käpylä, M. J. (2022). Modeling the interplay between epidemics and regional socio-economics. Physica A: Statistical Mechanics and its Applications, 127696.
8. Takko, T., Bhattacharya K, & Kaski K.K (in preparation).Modelling social exposure using co-activity networks of mobility during Covid-19.

## 6.4 Planned Activities for next period

### 6.4.1 Data Collection activities

#### **Collection of the data on the players in the football skin-tone discrimination project.**

Partners involved: PSE

### 6.4.2 Software Development activities

#### **UK-Europe Mobility Continuation**

Partners involved: UNIPi

As a first step, we aim to conclude the study of mobility between the UK and Europe on Brexit effects. The results will be presented in (provisory title) International mobility between UK and Europe around Brexit: a data-driven study (Goglia, Kim, Pollacci, Spyrtos, Iacus, Sirbu - UNIPi, Max Planck Institute for Demographic, JRC, Haarvard) and submitted to Big Data and Society. Furthermore, our plan for the future foresees deepening the study of highly specialised mobility and migration. Thanks to the experiences made, through the analysis of different data sources, e.g., LinkedIn, Churchbase, and scholarly data, we aim to address the phenomenon of brain circulation, brain drain and brain gain. Moreover, we will further expand the analysis of immigrants' adoption of natives' shopping consumption by exploiting the Explainable Artificial Intelligence methodologies to provide an in-depth interpretation of the results obtained.

#### **Academic Networks, Migration and Markets**

Partners involved: PSE

Continue working on the current projects relating to academic migration and markets during and after the Iron Curtain.

#### **Intergenerational transmission of educational and occupational preferences: evidence from Sweden**

Partners involved: PSE

Continue working on this project. Currently awaiting data from statistics Sweden and collecting data for country level proxies for perceived status for occupations

### **Collaboration using microdata**

Partners involved: AALTO

We plan to invite researchers using microdata from other Nordic countries (also Estonia) to collaborate on aspects of migration and social assimilation.

## 7 T10.5 Sports Data Science

This exploratory provides massive heterogeneous dynamic data describing several sports (e.g., soccer, cycling and rugby) to construct an interpretable and easy-to-use tool for a variety of stakeholders in sports: coaches and managers, athletes, scouts, journalists and the general public. Those studies will open an exciting perspective on how to understand and explain the factors influencing sports success and how to build simulation tools for boosting both individual and collective performance.

### 7.1 Activities Report

#### **Fatigue and recovery estimation in soccer**

*Partners Involved:* UniPi, ISTI-CNR

It is well known that the athletes' perception of fatigue, muscle soreness and other physical and psychological perceptions are affected by training load (TL) during a physical task. Additionally, individual factors such as players' characteristics, fitness level, positional role, technical skills and tactical demands are other crucial factors that affect these perceptions. During the soccer season, training workloads are organised in order to balance the players' perception of fatigue with the aim of maximally enhancing competition performance. Predicting the state of fatigue in soccer players is useful to design training and optimise performance on match day. Therefore, the aim of this study was to explore, using a framework of big data analytics, the most important predictors of fatigue in a group of sub-elite soccer players using inexpensive and practical data monitoring tools.

#### **Soccer players' profile**

*Partners Involved:* UniPi, ISTI-CNR

Soccer is an intermittent team-sport, where players are expected to possess well developed physical, psychological, tactical and technical skills. From a physiological perspective, demands on modern soccer players are more complex than in many individual sports and these demands vary depending on level of performance, style of play incorporated by a team and particularly on positional role. Actually, differences were observed in body composition characteristics between playing positions in international soccer players. Thus, obtaining body composition estimates are an important tool in athletic settings as they can help determine athletic potential, track training adaptations. However, limited data are available to describe physical differences linked to playing position and the aim of this study was to determine if roles and tactical-technical functions are associated with specific body composition patterns in elite soccer players. The aim of this project is to profile players in accordance with body composition, physical and technical features with the aim of maximising the training effect by minimising the injury risk.

### 7.2 Micro-projects

#### **Single and combined effect of kinesio tape and warm-up on sprint cycling performance**

The aim of the present study was to investigate the single and combined effect of Kinesio Tape (KT) application and warm-up (WU) on sprint cycling performance. To study potential single and combined effect

on sprint cycling performance, we designed an experiment comparing the combination of different KT conditions (with and without KT) and pre-exercise routine (with and without WU). We hypothesized that the use of KT would promote performance benefit on maximal sprint cycling irrespective of pre-exercise routine.

Status: Completed; Partners involved: UniPi

External partners involved: University of Milan and University of Insubria.

Outcomes: [blog post](#), [experiment](#), [paper](#).

### **Physical Activity Levels and Perceived Changes in the Context of Intra-EEA Migration: A Study on Italian Immigrants in Norway**

This study investigated: 1) the extent to which Italian immigrants in Norway perceive that moving had a negative or positive impact on their PA; 2) possible differences between the PA of the Italian immigrants compared with the Norwegian population; 3) possible associations of the Italian immigrants' PA with key sociodemographic characteristics (gender, age, region of residence, and educational level).

Status: Completed; Partners involved: UniPi

External partners involved: University of South-Eastern Norway.

Outcomes: [blog post](#), [experiment](#), [paper](#).

### **Physiological recovery among workers in long-distance sleddog race: a case study on female veterinarians in Finnmarksløpet**

During Finnmarksløpet (FL, one of the longest distance sleddog races in the world), veterinarians are exposed to extreme environmental conditions and tight working schedules, with little and fragmented sleep. The aim of this case study was to examine cardiovascular parameters and sleep-wake patterns among veterinarians working within FL, during and after (for a month) the end of the race.

Status: Completed; Partners involved: UniPi

External partners involved: University of South-Eastern Norway, Clinica Veterinaria Dott. Maffi Sergio, Oxford University.

Outcomes: [paper](#), [blog post](#), [experiment](#).

### **Soccer & data cup - Expo Dubai 2020**

Soccer & Data Cup at Expo Dubai 2020 will be a 3-days international hybrid marathon of Sport Analytics combining fundamental techniques of data analysis and Artificial Intelligence. The event covers the subject area of Data Science, and aims to raise young people's awareness to the new frontiers of the complex analysis of digital data in the sport area. The students, with the help of mentors and experts from Sport Analytics discipline will find themselves immersed in a digital co-design competition using sports data of games from the five major European leagues of male football. They will be asked to find key solutions to a specific challenge question using selected sports data sets. All teams will ultimately work to present their projects in front of a selected jury of experts.

Status: Completed; Partners involved: UNIPi and ISTI-CNR.

Outcomes: [blog post](#), [experiment](#).

**A Narrative Review for a Machine Learning Application in Sports: An Example Based on Injury Forecasting in Soccer**

In the last decade, the number of studies about machine learning algorithms applied to sports, e.g., injury forecasting and athlete performance prediction, have rapidly increased. Due to the number of works and experiments already present in the state-of-the-art regarding machine-learning techniques in sport science, the aim of this narrative review is to provide a guideline describing a correct approach for training, validating, and testing machine learning models to predict events in sports science. The main contribution of this narrative review is to highlight any possible strengths and limitations during all the stages of model development, i.e., training, validation, testing, and interpretation, in order to limit possible errors that could induce misleading results. In particular, this paper shows an example about injury forecaster that provides a description of all the features that could be used to predict injuries, all the possible pre-processing approaches for time series analysis, how to correctly split the dataset to train and test the predictive models, and the importance to explain the decision-making approach of the white and black box models.

Status: Completed; Partners involved: UniPi and ISTI-CNR.

Outcomes: [paper](#), [blog post](#), [method](#).

**A dataset on the long-tail effect of COVID-19 lockdown on Italians' wellbeing**

From March 2020 to May 2021, several lockdown periods caused by the COVID-19 pandemic have limited people's usual activities and mobility in Italy and around the world. These confinement measures dramatically modified citizens' daily lifestyles and behaviours. With the advent of summer 2021 all the Italian regions returned to regular behaviours and routines. It is still unclear if there is a long-tail effect on people's quality of life, sleep- and physical activity-related behaviours. With this micro-project we propose to create a dataset that will allow to have accurate insights of the changes induced by the lockdown period in the Italians' health and that will permit to provide practical suggestions at local, regional, and state institutions and companies to improve infrastructures and services that could be beneficial to Italians' well being.

Status: Completed; Partners involved: UNIFI and CNR.

External partners involved: University of Milan and University of Insubria.

Outcomes: [paper](#), [blog post](#), [dataset](#)

**Blood sample profile helps to injury forecasting in elite soccer players**

By analyzing external workloads with machine learning models (ML), it is now possible to predict injuries, but with a moderate accuracy. The increment of the prediction ability is nowadays mandatory to reduce the high number of false positives. The aim of this study was to investigate if players' blood sample profiles could increase the predictive ability of the models trained only on external training workloads.

Status: Completed; Partners involved: UniPi and ISTI-CNR.

External partners involved: University of Tor Vergata.

Outcomes: [paper](#), [blog post](#), [experiment](#).

**Self-Rated Health Among Italian Immigrants Living in Norway: A Cross-Sectional Study**

Most studies on immigrant health focus on immigrant groups coming from extra-European and/or low-income countries. Little attention is given to self-rated health (SRH) in the context of EU/EEA migration. To know more about health among European immigrants can provide new insights related to social determinants of health in the migration context. Using the case of Italian immigrants in Norway, the aim of



this study was to (i) examine the levels of SRH among Italian immigrants in Norway as compared with the Norwegian and the Italian population, (ii) examine the extent to which the Italian immigrant perceived that moving to Norway had a positive or negative impact on their SRH; and (iii) identify the most important factors predicting SRH among Italian immigrants in Norway.

Status: Completed; Partners involved: UniPi.

External partners involved: Oslo Metropolitan University, University of South-Eastern Norway.

Outcomes: [research paper](#), [blog post](#), [experiment](#).

### **Regional Bioelectrical Phase Angle Is More Informative than Whole-Body Phase Angle for Monitoring Neuromuscular Performance: A Pilot Study in Elite Young Soccer Players**

The objective of this study was to investigate the association between regional and total phase angle (PhA) with lower-body neuromuscular performance in young elite soccer players.

Status: Completed; Partners involved: UniPi

External partners involved: University of Milan, Spezia Calcio, Parma Calcio 1913.

Outcomes: [paper](#), [blog post](#), [experiment](#).

## 7.3 Publications

1. Rossi A, Perri E, Pappalardo L, Cintia P, Alberti G, Norman D, FM Iaia (2022). Wellness Forecasting by External and Internal Workloads in Elite Soccer Players: A Machine Learning Approach. *Frontiers in Physiology*, 13:896928. doi: 10.3389/fphys.2022.896928
2. Rossi A, Pappalardo L, Filetti C, Cintia P (2022). Blood sample profile helps to injury forecasting in elite soccer players. *Sport Science for Health*. Doi: 10.1007/s11332-022-00932-1
3. Bongiovanni T, Rossi A, Iaia MF, Alberti G, Pata G, Trecroci A (2022). Association of phase angle and appendicular upper and lower body lean soft tissue with physical performance in young elite soccer players: a pilot study. *Journal of Sports Medicine and Physical Fitness*, 62(8): 1015-1022. Doi: 10.23736/S0022-4707.21.12911-1
4. Natilli M, Rossi A, Trecroci A, Cavaggioni L, Merati G, Formenti D (2022). The long-tail effect of the COVID-19 lockdown on Italians' quality of life, sleep and physical activity. *Nature Scientific Data*, 9:250. Doi: 10.1038/s41597-022-01376-5
5. Terragni L, Rossi A, Miscali M, Calogiuri G (2022). Self-Rated Health Among Italian Immigrants Living in Norway: A Cross-Sectional Study. *Frontiers in Public Health*, 10:837728. doi: 10.3389/fpubh.2022.837728
6. Bongiovanni T, Rossi A, Trecroci A, Martera G, Iaia FM, Alberti G, Pasta G, Lacombe M (2022). Regional Bioelectrical Phase Angle Is More Informative than Whole-Body Phase Angle for Monitoring Neuromuscular Performance: A Pilot Study in Elite Young Soccer Players. *Sports*, 10, 66. doi: 10.3390/sports10050066
7. Cavaggioni L, Rossi A, Tosin M, Scurati R, Michielon G, Alberti G, Merati G, Formenti D, Trecroci A (2022). Changes in Upper-Body Muscular Strength and Power in Paralympic Swimmers: Effects of Training Confinement during the COVID-19 Pandemic. *International Journal of Environmental Research in Public Health*, 19, 5382. doi: 10.3390/ijerph19095382
8. Trecroci A, Formenti D, Moran J, Pedreschi D, Rossi A (2022). Editorial: Factors Affecting Performance and Recovery in Team Sports: A Multidimensional Perspective. *Frontiers in Physiology*, 13:877879. doi: 10.3389/fphys.2022.877879

9. Rossi A, Pappalardo L, Cintia P (2022). A Narrative Review for a Machine Learning Application in Sports: An Example Based on Injury Forecasting in Soccer. *Sports*, 10(1), 5. doi: 10.3390/sports10010005
10. Formenti D, Trecroci A, Duca M, Vanoni M, Ciovati M, Rossi A, Alberti G (2022). Volleyball-Specific Skills and Cognitive Functions Can Discriminate Players of Different Competitive Levels. *Journal of Strength and Conditioning Research*, 36(3), 813-819. Doi: 10.1519/JSC.0000000000003519
11. Rossi A, Bongiovanni T, Matera G, Cavaggioni L, Iaia FM, Trecroci A (2022). Influence of Upper and Lower Body Anthropometric Measures on An Aggregate Physical Performance Score in Young Elite Male Soccer Players: A Case Study. *Journal of Men's Health*, 18(7), 148. Doi: 10.31083/j.jomh1807148
12. Rossi A, Calogiuri G, Maffi S, Pedreschi D, Clifton DA, Morelli D (2022). Physiological recovery among workers in long-distance sleddog race: a case study on female veterinarians in Finnmarksløpet. *Work: A Journal of Prevention, Assessment & Rehabilitation*, 71(3), 1-12, doi: 10.3233/WOR-210331
13. Campa F, Bongiovanni T, Trecroci A, Rossi A, Greco G, Pasta G, Coratella G (2021). Effects of the COVID-19 Lockdown on Body Composition and Bioelectrical Phase Angle in Serie A Soccer Players: A Comparison of Two Consecutive Seasons. *Biology*, 10:1175. doi: 10.3390/biology10111175
14. Pappalardo L, Rossi A, Natilli M, Cintia M (2021). Explaining the difference between men's and women's football. *PLOS ONE*, 16(8): e0255407; doi: 10.1371/journal.pone.0255407
15. Rossi A, Formenti D, Cavaggioni L, Alberti G, Esposito F, D'Angelo F, Trecroci A (2021). Single and combined effect of kinesio tape and warm-up on sprint cycling performance. *BMC Sports Science, Medicine and Rehabilitation*. *BMC Sports Science, Medicine and Rehabilitation*, 13:77; doi: 10.1186/s13102-021-00310-3
16. Bongiovanni T, Rossi A, Iaia FM, Di Baldassarre A, Pasta G, Manetti P, Alberti G, Trecroci A (2021). Relationship of regional and whole body morphology to vertical jump in elite soccer players: a data-driven approach. *The Journal of Sports Medicine and Physical Fitness*; doi: 10.23736/S0022-4707.21.12323-0
17. Calogiuri G, Rossi A, Terragni L (2021). Physical activity levels and perceived changes in the context of intra-EEA migration: a study on Italian immigrants in Norway. *Frontiers in Public Health*; doi: 10.3389/fpubh.2021.689156
18. Morelli D, Rossi A, Bartoloni L, Cairo M, Clifton AC (2021). SDNN24 Estimation from Semi-Continuous HR Measures. *Sensors*, 21, 1463; doi: 10.3390/s21041463
19. Hemmings NR, Kawadler JM, Whatmough R, Ponzo S, Rossi A, Morelli D, Brid G, Plans D (2021). Development and Feasibility of a Digital Acceptance and Commitment Therapy-Based Intervention for Generalized Anxiety Disorder: Pilot Acceptability Study. *JMIR Formative Research*, 5, e21737; doi: 10.2196/21737.
20. Perri E, Simonelli C, Rossi A, Trecroci A, Alberti G, Iaia FM (2021). Relationship Between Wellness Index and Internal Training Load in Soccer: Application of a Machine Learning Model. *International Journal of Sports Physiology and Performance*, 16, 695-703; doi: 10.1123/ijsp.2020-0093
21. Trecroci A, Duca M, Cavaggioni L, Rossi A, Scurati R, Longo S, Merati G, Alberti G, Formenti D (2021). Relationship between Cognitive Functions and Sport-Specific Physical Performance in Youth Volleyball Players. *Brian Science*, 11, 227. doi: 10.3390/brainsci11020227
22. Formenti D, Rossi A, Bongiovanni T, Campa F, Cavaggioni L, Alberti G, Longo S, Trecroci A (2021). Effects of Non-Sport-Specific versus Sport-Specific Training on Physical Performance and Perceptual Response in Young Football Players. *Environmental Research and Public Health*, 18, 1962; doi: 10.3390/ijerph18041962

## 7.4 Planned Activities for next period

### 7.4.1 Scientific activities

Partners involved: UNIPI and CNR

The scientific activities of this task will be focused on assessing and predicting the individual responses to external and internal training workload stimuli on fatigue and recovery status in soccer players. Moreover, differences in players' characteristics such as body composition will be important aspects that have to be taken into consideration during training schedule with the aim of maximising the training effect while minimising the injury risk. Profiling players in accordance with individual characteristics (e.g., body composition, technical-tactical skills and physical status) will be a crucial point in this research project.

## 8 T10.6 Social Impact of AI and Explainable Machine Learning

The exploratory investigates the foreseeable impact of AI and Big Data on society, developing analytical and simulation tools. It also integrates a vast repertoire of practical tools for explainable AI, in particular, methods for deriving meaningful explanations of black-boxes decision systems based on machine learning.

### 8.1 Activities Report

#### **Decentralised and Cooperative Learning of AI Models**

*Partners Involved:* CNR, UNIFI

**Description:** The aim of this activity is to study distributed and decentralised methodologies for learning collective AI models as a combination of AI local models. We worked at the definition and development of a novel federated learning approach (HOLDA), tailored for training machine learning models on data distributed over federated organisations hierarchically organised. Our method focuses on the generalisation capabilities of the neural network models, providing a new mechanism for selecting their best weights. In addition, it is tailored for tabular data. We empirically tested the performance of our approach on two tabular datasets, showing excellent results in terms of performance and generalisation capabilities. Then, we also tackled the problem of assessing the privacy risk of users represented in the training data. We empirically show, by attacking the HOLDA models with the Membership Inference Attack, that the privacy of the users in the training data may have high risk [Fontana2021]. Moreover, we also propose to monitor model fairness during the Federated Learning (FL) process, exploiting the HOLDA models.

We have also studied the behaviour of several methods to merge models computed locally by individual users on a common ML task (e.g., multi-class classification) in a scenario we called “model gossiping”. We consider a network of nodes (users) and nodes exchange local models only with their direct peers. Each node aggregates all models received from its peers, and in the next round propagates the model aggregated in the previous step. The same locally aggregated model is re-trained on the local data, and replaces the local model in the previous step. This represents a fully decentralised system modelled as a network of cooperating entities, where link represents, for example, social relationships between users. We have started investigating the asymptotic behaviour of the aggregated model in case of very simple aggregation schemes, such as the “vanilla” decentralised averaging (which mirrors the well-known FedAvg) under different configurations (e.g., when all nodes start with a coordinated initial model, or not). The results obtained show that very simple aggregation schemes may not work well, and catastrophic effects may arise if merging is not done carefully. This motivates to investigate more advanced merging schemes, which, for example, take into account the current accuracy of peer’s models in the merging process. To test these schemes, we have developed a dedicated simulation environment, which is publicly available at <https://zenodo.org/record/5780042>.

#### **Opening the Black Box**

*Partners Involved:* UNIFI, CNR, SNS

- **Description:** We developed an extension of LORE for classifiers working on data types other than modular: images, time series, and text. The new version of LORE is also more stable. In order to

obtain this property, LORE generates a set of neighbourhoods, learns on top of them different local models and then, it applies a merging strategy for decision trees to derive a single and more general local explainer. Results of this research activity are published in [Guidotti2022] (see Publications).

- We developed a methodology that allows us to measure to which extent the explanations returned by local explanation methods are correct with respect to a synthetic ground truth explanation. Indeed, the proposed methodology enables the generation of synthetic transparent classifiers for which the reason for the decision taken, i.e., a synthetic ground truth explanation, is available by design. Experimental results show how the proposed approach allows to easily evaluate local explanations on the ground truth and to characterise the quality of local explanation methods [Guidotti2021].
- We reviewed the literature of XAI methods for time series classification and realised a survey where we present the first extensive literature review on Explainable AI (XAI) for time series classification, categorise the research field through a taxonomy subdividing the methods into time points-based, subsequences-based and instance-based, and identify open research directions regarding the type of explanations and the evaluation of explanations and interpretability [Theissler2022]. We also worked on the definition of an interpretable Time Series Classification method based on the patterns, which is possible to extract from the Matrix Profile of the time series in the training set. A smart design of the classification procedure allows obtaining an efficient and effective transparent classifier modelled as a decision tree that expresses the reasons for the classification as the presence of discriminative subsequences [Guidotti, D'Onofrio2021].

## Privacy & AI

*Partners Involved:* UNIFI, SNS

Description: This research activity is related to the problem of privacy and data protection. In this context we have three main contributions:

- We have developed a methodology that enables the evaluation of the privacy risk exposure of global explainers based on an interpretable classifier that imitates the global reasoning of a black-box classifier. The idea is to verify if the layer of interpretability added by the interpretable model can jeopardise the privacy protection of individuals represented in the data used for the training of the black-box classifier. Intuitively, explainers are learned functions that are derived by exploiting the predictive knowledge of a black-box model learned on a private dataset. Thus, it could leak information about this private dataset [Naretto2022]. Now we are working on the extension of such methodology to the evaluation of local explainers.
- We are also working on the assessment of the robustness of privacy protection methods for image processing based on differential privacy against image reconstruction methods based on machine learning models. Our idea is to simulate attacks that reconstruct the original images and we are evaluating data quality after reconstruction with image-specific metrics and by testing the accuracy of state-of-art machine learning models for classification tasks on the reconstructed images. Our current experimental results show that when data quality is preserved, differential privacy may not be enough to protect the images. We plan to conclude the experiments and submit a paper within a few weeks.

- We have also worked on the research challenge to systematically assess the privacy risks on complex data such as mobility data and text data. In particular, we developed a privacy risk assessment methodology for psychometric profiles derived from texts written by users in social media and we discovered that some psychometric dimensions are more risky for users [Mariani2021]. Moreover, we included a privacy risk assessment methodology for mobility data in a Python library on methods for analyzing and generating spatio-temporal data [Pappalardo2022].

## **Fairness & AI**

*Partners Involved:* UNIFI, CNR, UPF

Description: This activity has the goal to analyse from different perspectives the problem of bias and unfairness in AI. In this context we have three main contributions:

- We have performed a series of experiments designed to observe human response to different characteristics of a Decision Support System (DSS) such as accuracy and bias, particularly the extent to which participants rely on the DSS, and the performance they achieve. In our experiments, participants play a simple online board game inspired by so-called “wildcat” (i.e., exploratory) drilling for oil. The board has two layers: a visible layer describing the costs (terrain), and a hidden layer describing the reward (oil yield). The final score of a participant is computed as rewards minus costs. Participants in the control group play the game without receiving any assistance, while in treatment groups they are assisted by a DSS suggesting places to drill. For certain treatments, the DSS does not consider costs, but only rewards, which introduces a bias that is observable by users. Between subjects, we vary the accuracy and bias of the DSS, and observe the participants' total score, time to completion, the extent to which they follow or ignore suggestions. We also measured the acceptability of the DSS in an exit survey. Our results show that participants tend to score better with the DSS, that the score increase is due to users following the DSS advice, and related to the difficulty of the game and the accuracy of the DSS. We observe that this setting elicits mostly rational behaviour from participants, who place a moderate amount of trust in the DSS and show neither algorithmic aversion (under-reliance) nor automation bias (over-reliance). However, their stated willingness to accept the DSS in the exit survey seems less sensitive to the accuracy of the DSS than their behaviour, suggesting that users are only partially aware of the (lack of) accuracy of the DSS. The game constitutes a research platform intentionally designed to study decision support in the absence of pre-existing expertise, which makes it an interesting model for studying algorithmic reliance.
- FairShades: Fairness Auditing via Explainability in Abusive Language Detection Systems. At every stage of a supervised learning process, harmful biases can arise and be inadvertently introduced, ultimately leading to marginalization, discrimination, and abuse towards minorities. This phenomenon becomes particularly impactful in the sensitive real-world context of abusive language detection systems, where non-discrimination is difficult to assess. In addition, given the opaqueness of their internal behaviour, the dynamics leading a model to a certain decision are often not clear nor accountable, and significant problems of trust could emerge. A robust value-oriented evaluation of models' fairness is therefore necessary. FairShades is a model-agnostic approach for auditing the outcomes of abusive language detection systems. Combining explainability and fairness evaluation, Fair-Shades can identify unintended biases and sensitive categories towards which models are most

discriminative. This objective is pursued through the auditing of meaningful counterfactuals generated within the CheckList framework. We conduct several experiments on BERT-based models to demonstrate our proposal's novelty and effectiveness for unmasking biases [Manerba 2021].

- Investigating Debiasing Effects on Classification and Explainability. During each stage of a dataset creation and development process, harmful biases can be accidentally introduced, leading to models that perpetuates marginalization and discrimination of minorities, as the role of the data used during the training is critical. We propose an evaluation framework that investigates the impact on classification and explainability of bias mitigation preprocessing techniques used to assess data imbalances concerning minorities' representativeness and mitigate the skewed distributions discovered. Our evaluation focuses on assessing fairness, explainability and performance metrics. We analyse the behaviour of local model-agnostic explainers on the original and mitigated datasets to examine whether the proxy models learned by the explainability techniques to mimic the black-boxes disproportionately rely on sensitive attributes, demonstrating biases rooted in the explainers. We conduct several experiments about known biased datasets to demonstrate our proposal's novelty and effectiveness for evaluation and bias detection purposes [Manerba 2022].

## Finance & Economics

*Partners Involved:* ETHZ, SNS

*Description:* The activity has been developed in two branches:

- We analysed high-frequency market microstructure observables with different information-theoretic measures for dynamical systems, and found temporal changes in information sharing across markets in the period during which the system underwent a price bubble. In particular, we studied time-varying components of predictability, memory, and (a)synchronous coupling, measured by transfer entropy, active information storage, and multi-information. Being model-free, these measures do not lend themselves easily to interpretation. To ensure explainability and to study robustness of the empirical findings, several models were used. In particular, a comparison with an econometric model that couples market microstructure variables ensured that the information dynamics couplings observed were not spurious. Furthermore, the empirical results showed statistically significant changes in the direction of information flow between different market observables around the price bubble. Another econometric model was used to interpret this result as a potential regime shift. We also showed, with the means of modelling, that the information dynamics tools which are based on k-Nearest Neighbour algorithm, produce non-spurious results for the properties of the high frequency market data. This work has materialised in a paper published in *Chaos: An Interdisciplinary Journal of Nonlinear Science*, together with a dataset uploaded to SoBigData platform, and a blogpost.
- 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 behaviour. 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 chaotic behaviour. 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.

### **Causality Analysis of Data**

*Partners involved:* UNIFI, CNR

Description: CALIME: Causality-Aware Local Interpretable Model-Agnostic Explanations. A significant drawback of eXplainable Artificial Intelligence (XAI) approaches is the assumption of feature independence. We focus on integrating causal knowledge in XAI methods to increase trust and help users assess explanations' quality. We propose a novel extension to LIME a widely used local and model-agnostic explainer that explicitly encodes causal relationships in the data generated around the input instance to explain. We name it Causal-Aware LIME, i.e., CALIME, and we build it on top of GENCDAs proposal for efficient Causal Discovery by using pattern mining algorithms. Extensive experiments show that CALIME achieves superior performance comparing the initial one for both the fidelity in mimicking the black-box and the stability of the explanations [Cinquini2022].

### **Understanding the control of complex systems**

*Partner involved:* ETH

Description: The efficient control of complex dynamical systems has many applications in the natural and applied sciences. In most real-world control problems, both control energy and cost constraints play a significant role. Although such optimal control problems can be formulated within the framework of variational calculus, their solution for complex systems is often analytically and computationally intractable. To overcome this outstanding challenge, we present AI Pontryagin, a versatile control framework based on neural ordinary differential equations that automatically learns control signals that steer high-dimensional dynamical systems towards a desired target state within a specified time interval. We demonstrate the ability of AI Pontryagin to learn control signals that closely resemble those found by corresponding optimal control frameworks in terms of control energy and deviation from the desired target state. Our results suggest that AI Pontryagin is capable of solving a wide range of control and optimization problems, including those that are analytically intractable.

### **Explaining AI forecasts for epidemic dynamics**

*Partners involved:* ETH

Description: Forecasting new cases, hospitalizations, and disease-induced deaths is an important part of infectious disease surveillance and helps guide health officials in implementing effective countermeasures. For disease surveillance in the US, the Centers for Disease Control and Prevention (CDC) combines more than 65 individual forecasts of these numbers in an ensemble forecast at national and state levels. A similar



initiative has been launched by the European CDC (ECDC) in the second half of 2021. We collected data on CDC and ECDC ensemble forecasts of COVID-19 fatalities, and we compared them with easily interpretable “Euler” forecasts serving as a model-free benchmark that is only based on the local rate of change of the incidence curve. The term “Euler method” is motivated by the eponymous numerical integration scheme that calculates the value of a function at a future time step based on the current rate of change. Our results show that simple and easily interpretable “Euler” forecasts can compete favourably with both CDC and ECDC ensemble forecasts on short-term forecasting horizons of 1 week. However, ensemble forecasts better perform on longer forecasting horizons. Using the current rate of change in incidences as estimates of future incidence changes is useful for epidemic forecasting on short time horizons. An advantage of the proposed method over other forecasting approaches is that it can be implemented with a very limited amount of work and without relying on additional data (e.g., data on human mobility and contact patterns) and high-performance computing systems.

## 8.2 Software Development activities

In relation to the social AI gossiping activity described in Section 1.1, a python simulator has been developed and it is available at the following link: <https://zenodo.org/record/5780042>.

Partners involved: CNR

## 8.3 Events

- Tutorial XAI Explainable Machine Learning, AAAI 2022 (Virtual) 23/02/2022: a snapshot on the work of XAI to date, and surveys the work achieved by the AI community with a focus on machine learning and symbolic AI related approaches.
- Tutorial & Workshop XKDD 2021 ECML-PKDD 2020 (Virtual) 13/09/2021: The purpose of XKDD, eXplaining Knowledge Discovery in Data Mining, is to encourage principled research that will lead to the advancement of explainable, transparent, ethical and fair data mining and machine learning.
- Workshop XKDD 2022 ECML-PKDD 2022 (in Presence) 19/09/2022: The purpose of XKDD, eXplaining Knowledge Discovery in Data Mining, is to encourage principled research that will lead to the advancement of explainable, transparent, ethical and fair data mining and machine learning.
- 3rd Italian Workshop on Explainable Artificial Intelligence co-located with AI\*IA 2022 (in Presence) 28/11/2022: The purpose of this workshop is to encourage research and advancement on novel methodologies to build transparent and scrutable AI systems and algorithms.

Partners involved: UNIFI, CNR, SNS

## 8.4 Micro-projects

### **The consequences of biased and/or inaccurate predictions from machine learning system**

This research project is focused on the study of human interaction with automatic systems designed to assist users during a decision making process. In particular, the experiments will study how individuals change or adapt their behaviour depending on different characteristics of the automatic system.

During its execution, we aim to perform an empirical study to determine how different levels of accuracy and the presence (or absence) of biased recommendations affect the human perception and/or behaviour with respect to the system predictions. The experiment will be framed in the context of an online web-based game where the user needs to balance exploration/exploitation to maximize the obtained score. There is a decision support system (DSS) assisting users during the game. As control variables for our experiment, we will modify the accuracy of the DSS and if it accounts for the potential benefit and costs or only for the potential benefits. Also, to create certain biases, it might be more accurate in certain parts of the terrain. Using this framework and letting users play with it, we aim to study and understand whether any of our control variables has any role in the final score obtained by the users or their confidence on the predictions, between other performance variables.

Status: Ongoing; Partners involved: UNIP, UPF, CNR  
Expected outputs: Blog post, Dataset, Experiment

### **Social AI gossiping simulator**

We envision a human-AI ecosystem in which AI-enabled devices act as proxies of humans and try to learn collectively a model. To this aim, techniques for decentralized learning are employed. Each device will learn a local model that needs to be combined with the models learned by the other nodes, in order to improve both the local and global knowledge. The challenge of doing so in a fully decentralized AI system entails understanding how to compose models coming from heterogeneous sources and, in the case of potentially untrustworthy nodes, deciding who can be trusted and why. In this micro-project, we focus on the design and development of a simulator for accomplishing a decentralized learning task on a social network and we study what models emerge from the combination of local models, where combination takes into account the social relationships between the nodes (humans associated with the AI). We will use both synthetic graphs to represent social relationships as well as real-life graphs, and large-scale simulations for performance evaluation.

Status: Ongoing; Partners involved: IIT-CNR  
Expected outputs: Blog post

### **Reducing Graph Structural Bias by Adding shortcut edges**

Reducing structural bias in the social networks can help users to escape from echo chambers and improve society. The goal of this study is to learn how to add shortcut edges to reduce graph structural bias. Specifically, we will extend the paper "RePBubLik: Reducing the Polarized Bubble Radius with Link Insertions", by Haddadan et al., to a continuous setting, where each node has a real-valued polarization score, and the objective is to maximize the structural bias decrease by adding edges among nodes.

Status: Ongoing; Partners involved: KTH  
Expected outputs: Method, Experiments, Blog post, Preprint paper

### **Machine Learning Approaches to Classify Primary and Metastatic Cancers Using Tissue of Origin-Based DNA Methylation Profiles**

Cancer metastasis is considered to be one of the most significant causes of cancer morbidity, accounting for up to 90% of cancer deaths. The accurate identification of a cancer's origin and the types of cancer cells it comprises is crucial in enabling clinicians to decide better treatment options for patients. We

utilized machine learning classifiers to 31 discriminate metastatic, primary, and non-cancerous methylome samples. Moreover, we applied Local Interpretable Model-agnostic Explanations (LIME) to explain important methylation biomarkers to 36 classify cancer types.

Status: Completed; Partners involved: TARTU

Outputs: [paper](#), [experiment](#), [blog post](#)

### Visualising the Results of Boolean Matrix Factorizations

Boolean matrix factorization (BMF) is a popular problem in the data mining and knowledge discovery communities. When communicating the results of BMF algorithms with non-experts, it is often helpful to visualize the results such that the underlying structure of the data becomes more clear. Unfortunately, current techniques from matrix visualization and graph drawing do not easily extend to the case of BMF. Therefore, we will develop algorithms for visualizing the results of BMF algorithms. This will involve formulating an objective function that captures how well the matrix is visualized, as well as inventing algorithms to optimize this objective function.

Status: Completed; Partners: KTH, LUH

External partners: Ecole Normale Supérieure de Lyon

Output: [method](#), [blog post](#).

## 8.5 Publications

1. Michele Fontana, Francesca Naretto, Anna Monreale: A new approach for cross-silo federated learning and its privacy risks. PST 2021: 1-10
2. Michele Fontana, Francesca Naretto, Anna Monreale, Fosca Giannotti: Monitoring Fairness in HOLDA. HHAI 2022: 246-248
3. Andreas Theissler, Francesco Spinnato, Udo Schlegel, Riccardo Guidotti: Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions. IEEE Access 10: 100700-100724 (2022)
4. Riccardo Guidotti: Exploiting auto-encoders for explaining black-box classifiers. Intelligenza Artificiale 16(1): 115-129 (2022)
5. Francesca Naretto, Anna Monreale, Fosca Giannotti: Privacy Risk of Global Explainers. HHAI 2022: 249-251
6. Luca Pappalardo, Filippo Simini, Gianni Barlacchi, Roberto Pellungrini: scikit-mobility: A Python Library for the Analysis, Generation, and Risk Assessment of Mobility Data. J. Stat. Softw. 103(4) (2022)
7. Marta Marchiori Manerba, Riccardo Guidotti: Investigating Debiasing Effects on Classification and Explainability. AIES 2022: 468-478
8. Riccardo Guidotti, Matteo D'Onofrio: Matrix Profile-Based Interpretable Time Series Classifier. Frontiers Artif. Intell. 4: 699448 (2021)
9. Riccardo Guidotti: Evaluating local explanation methods on ground truth. Artif. Intell. 291: 103428 (2021)
10. Marta Marchiori Manerba, Riccardo Guidotti: FairShades: Fairness Auditing via Explainability in Abusive Language Detection Systems. CogMI 2021: 34-43

11. Martina Cinquini, Fosca Giannotti, Riccardo Guidotti: Boosting Synthetic Data Generation with Effective Nonlinear Causal Discovery. *CogMI 2021*: 54-63
12. Valerio Bonsignori, Riccardo Guidotti, Anna Monreale: Deriving a Single Interpretable Model by Merging Tree-Based Classifiers. *DS 2021*: 347-357
13. Giacomo Mariani, Anna Monreale, Francesca Naretto: Privacy Risk Assessment of Individual Psychometric Profiles. *DS 2021*: 411-421
14. Böttcher, Lucas, Nino Antulov-Fantulin, and Thomas Asikis. "AI Pontryagin or how artificial neural networks learn to control dynamical systems." *Nature communications* 13.1 (2022): 1-9.
15. Vasiliauskaite, V., Lillo, F., & Antulov-Fantulin, N. (2022). Information dynamics of price and liquidity around the 2017 Bitcoin markets crash. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(4), 043123.
16. Asikis, Thomas, Lucas Böttcher, and Nino Antulov-Fantulin. "Neural ordinary differential equation control of dynamics on graphs." *Physical Review Research* 4.1 (2022): 013221.
17. Antulov-Fantulin, Nino, and Lucas Böttcher. "On the accuracy of short-term COVID-19 fatality forecasts." *BMC Infectious Diseases* 22.1 (2022): 1-7.
18. Cinquini Martina, & Guidotti, R. (2022, July). CALIME: Causality-Aware Local Interpretable Model-Agnostic Explanations. In *Proceedings of World Congress in Computer Science, Computer Engineering, & Applied Computing*.

## 8.6 Planned Activities for next period

### 8.6.1 Scientific activities

#### **Decentralised and Cooperative Learning of AI Models**

Partners Involved: IIT-CNR, University of Pisa, ISTI-CNR

We plan to work for investigating different research directions in decentralised and cooperative systems:

1. Addressing the task of learning AI models with multi-objectives in cooperative systems orchestrated by a server. The idea is to have a cooperative learning process where each local node aims at learning a local model that is able to optimise more than one objective (e.g., fairness and accuracy) and globally the system is also able to aggregate the different local models to get a global model represented the trade-off among the different local contributors.
2. Addressing the problem of privacy in cooperative learning by designing cooperative learning systems involving different types of collaborations which can require different levels of privacy protection on the basis of a possible trust relationship that may exist among some nodes.
3. With respect to the “model gossiping” activity, we are going to carry out the work along two main directions:
  - a. Defining appropriate merging policies that could avoid catastrophic effects; compare them with reference benchmarks including fully centralised ML approaches; evaluate the sensitivity of the identified scheme for different ML tasks, increasingly complex datasets, different network patterns;
  - b. Investigate the role of the network structure on the emerging properties of the aggregated models resulting from the “gossiping” process.

**Opening The Black Box**

Partners Involved: UNIPi, SNS

We plan to extend the investigation related to the explanation of time series classification by considering also the context where we need to classify multivariate time series and trajectories. In this context, we want to explore both methods for post-hoc explanation of black-box classification models and interpretable-by-design methods for classification. Moreover, we plan to also consider explanations for few-shot learning approaches such as Siamese Networks, and to design interpretable by design approaches constructed through non-interpretable complex algorithm processes.

**Privacy & AI**

Partners Involved: UNIPi, SNS

We plan to extend the privacy risk assessment methodology for analysing the privacy risk of local explainers based on local surrogate and then, understand if we can also consider the possibility to analyse the risks of a local to global explainer. The main idea is to also have a deep understanding on which are the dimensions of the explainers that could impact on the privacy risk to have the opportunity to design appropriate mitigation strategies on data used for learning the black-box models or on the explainable AI mechanism. Concerning the attack model, currently we are using the Membership inference attack, but we plan also to design new types of attacks and to include in the methodology also reconstruction attacks. Concerning the assessment of the robustness of image obfuscation by differential privacy, we plan to perform a deeper investigation with the goal of extending the methodology to other forms of data such as tabular data, time series data, mobility data, etc.

**Differential Deep Learning for Complex Systems**

Partners Involved: ETHZ

Dynamics in a complex system is often modelled as a set of coupled differential equations. They typically include non-linear and local interaction terms that lead to the ubiquitous properties of complexity, namely self-organisation, emergence. In recent years, data-driven methods have been applied to learn a solution to an initial value problem directly from time series. We plan to discuss the limitations of the current techniques, in particular, where they fall short in the context of complex systems and extend the current methodologies in this direction.

## 9 Conclusions

As witnessed by the vast amount of activities carried out and the variety of topics investigated, the second year and a half of WP10 has been very productive. The introduction of micro-project at the beginning of year two helped significantly organise and manage the activities in WP10, and to track the growth of the platform in terms of items (methods, datasets, experiments, applications) and stories.

We hope to observe an even higher improvement during the remainder of the project, also thanks to the transnational access program, allowing for strengthening the collaborations among the partners of the consortium as well as the collaborations with external institutions, organisations, companies, and researchers.