Social Mining & Big Data Analytics

SoBigData

RESEARCH INFRASTRUCTURE ++

Deliverable D10.2

Exploratory activities report and planning for the next period 1

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

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	http://www.sobigdata.eu
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
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	D10.2
DELIVERABLE TITLE	Exploratory activities report and planning for the next period 1
AUTHOR(S)	Luca Pappalardo (CNR), Roberto Pellungrini (UNIPI), Aris Gionis (KTH)
CONTRIBUTOR(S)	Luca Pappalardo (CNR), Roberto Pellungrini (UNIPI), Anna Monreale (UNIPI), Angelo Facchini (IMT), Tiziano Squartini (IMT), Paolo Cintia (UNIPI), Alessio Rossi (UNIPI), Kalina Bontcheva (USFD), Ye Jiang (USFD), Aris Anagnostopoulos (UNIROMA1), Laura Pollacci (UNIPI), Hillel Rapoport (PSE), Donia Kamel (PSE), Nino Antulov-Fantulin (ETHZ), Vaiva Vasiliauskaite (ETHZ), Michela Natilli (CNR), Michele Gentili (UNIROMA1)
EDITOR(S)	Beatrice Rapisarda (CNR), Valerio Grossi (CNR)
REVIEWER(S)	Jurek Leonhardt (LUH)
CONTRACTUAL DELIVERY DATE	30/06/2021
ACTUAL DELIVERY DATE	02/07/2021
VERSION	V1.2
TYPE	Report
DISSEMINATION LEVEL	Public
TOTAL N. PAGES	46
KEYWORDS	Exploratory, micro-projects, Artificial Intelligence

EXECUTIVE SUMMARY

This deliverable 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.

GLOSSARY

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

TABLE OF CONTENTS

1 Relevance to SoBigData++

This document describes: *(i)* the activity carried out within the exploratories since the beginning of the project; and *(ii)* the topics and activities planned for the next period. 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. The topics and activities described in this document are relevant to milestone MS2 "All exploratories are created (revised) and operative", and milestone MS3 "new exploratories coming from Interest groups become operative".

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).

1.1 Structure of the document

In Section 2, we describe a new tool we introduced to foster collaboration among partners and improve the tracking of the activities: the micro-projects. In Section 3, for each task of WP10 we report the results achieved and the activities carried out since the beginning of the project, as well as the topics and activities planned for the next period.

2 SoBigData++ Micro-Projects

To foster the collaboration among the partners of the consortium and improve the tracking of the activity developed within the project, in January 2021 we introduced the concept of *micro-projects*.

A micro-project (MP) is any study conducted by one or more partners in the consortium that investigates more deeply certain aspects and topics of interest to one or more exploratories. A micro-project produces some tangible outputs, both in terms of new scientific results and resources to upload on the platform and make available to the scientific community.

To propose a new MP, an MP leader (an individual involved in the project) must fill a form specifying: a description of the proposed MP, the duration of the MP, the individuals and the partners involved, the expected person months (PMs) each partner will dedicate to it, institutions external to the projects involved (if any), the exploratories (WP10 tasks) involved, other WP that are involved (if any), the expected outputs of the MP in terms of resources to upload on the platform (Method, Experiment, Dataset, Application) and (optionally) a preprint or publication. Each MP must also produce a story to be published in the SoBigData++ blog.

Each MP proposal is revised for approval by the WP10 leaders, who verify the coherence of the MP with the purposes of WP10 and that all the requested information has been properly specified.

When the MP expires, the MP leader must fill a termination form, in which they specify the links to the resources uploaded on the platform, and link to the preprints or the papers associated (if any), and the link to the post/story to be published on the blog. The WP leaders check the coherence of the specified resources with those promised in the MP proposal and then accept or reject the micro-project accordingly.

3 Activities report and planning

The research in WP10 is structured in vertical thematic environments, called *exploratories*, aimed at creating new stories and new resources to be integrated within the SoBigData++ research infrastructure. In this section, we describe the scientific results each exploratory has achieved since the beginning 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.

Note that the transnational access (TNA) has been suspended given the COVID-19 pandemic, which spread out in Europe in February 2020 impeding movements of people among partners of the consortium and from external institutions. Therefore, there are no activities related.

3.1 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 LGBT rights. 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. We also aim to study the effect of different social relationships when it comes to opinion formation.

3.1.1 Activities report

3.1.1.1 COVID-19 MISINFORMATION ANALYSIS

Partners involved: ETHZ, IMT

While social media allows people to seek information more effectively, the explosion of misinformation also causes significant harm to the global community, as false claims and online misinformation are still pervading social platforms. Automatic methods are therefore needed to detect misinformation on a large scale. To tackle this, we analysed a large set of tweets related to COVID-19 vaccination, collected using Twitter APIs and annotated using transfer learning. By using various Natural Language Processing (NLP) techniques, we are trying to understand the discourse about (mis)leading tweets for COVID-19 vaccination. We also studied challenges and pitfalls related to monitoring epidemics.

We also studied the impact of misinformation on the Italian Twitter debate during the pandemic. We focused on accounts whose identity is officially certified by Twitter (verified users). We then considered each pair of verified users and counted how many unverified ones interacted with both of them, via tweets or retweets: if this number was found to be statically significant, i.e., so large that it cannot be explained only by the activity of users, we considered the two verified accounts as “similar enough” and linked them in the corresponding monopartite projection. Discursive communities can, then, be detected by running a

community detection algorithm. Although it is a scientific subject, the COVID-19 discussion is clearly partitioned according to political criteria. Besides, we assessed the trustworthiness of the most recurrent news sites, among those tweeted by the political groups, via the NewsGuard browser extension: the impact of low-reputable posts reaches 22.1% in the right and center-right wing community and its contribution is even stronger in absolute numbers, due to the activity of this group: 96% of all non-reputable URLs shared by political groups come from this community.

3.1.1.2 VACCINE MISINFORMATION ANALYSIS

Partners involved: CSD

Vaccine misinformation may negatively impact citizen trust in policies. Thus, it is crucial to understand its impact on the beliefs of the users on the social platform who have engaged with vaccine misinformation, i.e., to analyse the online debates around false or misleading vaccine narratives and to establish whether these users believe, question, or refute it. Since manual detection of misinformation is infeasible, there is an urgent need for intelligent AI-based methods to assist journalists, government bodies, and companies in monitoring vaccine debates and misinformation. We are developing eXplainable AI (XAI) models for vaccine disinformation analysis, which can provide users (e.g., journalists) with the evidence behind the AI judgement and the much-needed algorithm transparency and enable stakeholders to study cross-platform propagation and longitudinal evolution of online debates around COVID-19 vaccination, with a specific focus on the topics raised by the vaccine-hesitant.

3.1.1.3 IDENTIFICATION OF RUMOR SPREADERS ON ONLINE SOCIAL MEDIA

Partners involved: UT

Social Media platforms are extensively exploited by users for spreading (mis)information to a large audience at a rapid pace, which can cause panic, fear, and financial loss to society. Thus, it is important to detect and control the misinformation in such platforms before it spreads to the masses. We studied controlling rumors (a type of misinformation) by identifying users who are possibly the rumor spreaders, i.e., users who are often involved in spreading the rumors. Due to the lack of public rumor spreaders labeled dataset, we used the public PHEME dataset containing rumor and non-rumor tweets information, and applied a weak supervised learning approach to transform the PHEME dataset into rumor spreaders dataset. We utilized user, text, and ego-network features and applied a Graph Convolutional Networks (GCN), comparing it with other supervised learning approaches (SVM, RF, LSTM). We performed extensive experiments on the rumor spreaders dataset, achieving up to 0.864 value for F1-Score and 0.720 value for AUC-ROC, showing the effectiveness of our methodology for identifying possible rumor spreaders using GCNs.

3.1.1.4 DYNAMICS OF OPINION POLARIZATION WITH THE FRIEDKIN-JOHNSEN MODEL

Partners involved: CNR

Understanding how people form their opinion is crucial to reveal the mechanism behind polarization, which has been the subject of extensive debate in recent years due to its potential to disrupt our societies and democracies. The only polarization model that has been validated on small-to-medium sized groups is the

one proposed by Friedkin and Johnsen (FJ model). We provided a comprehensive review of all the major variants of the FJ model and of the polarization metrics described in the related literature. For them, we highlighted their key features and the differences between each other. We found that polarization metrics are linked together through an invariant relationship. As a second contribution, we derived the conditions under which the FJ model yields to polarization, for each of the polarization metrics identified before. In addition, we also proved that the polarizing opinion vectors can be found analytically in most cases. All the results obtained have been validated with two popular datasets of real social networks.

3.1.1.5 A MODEL FOR THE TWITTER SENTIMENT CURVE

Partners involved: IMT

Twitter is among the most used online platforms for political communications, due to the brevity of its messages (which is particularly suitable for political slogans) and the quick diffusion of messages. Especially when the argument stimulates the emotionality of users, the content on Twitter is shared with extreme speed and thus studying the tweet sentiment is of utmost importance to predict the evolution of the discussions and the register of the relative narratives. We developed a model able to reproduce the dynamics of the sentiments of tweets related to specific topics and periods and to provide a prediction of the sentiment of the future posts based on the observed past. The model is a recent variant of the Pólya urn, introduced and studied in Aletti and Crimaldi (2019, 2020), which is characterized by a "local" reinforcement, i.e., a reinforcement mechanism mainly based on the most recent observations, and by a random persistent fluctuation of the predictive mean. In particular, this latter feature enables us to capture trend fluctuations in the sentiment curve. While the proposed model is extremely general and may be also employed in other contexts, it has been tested on several Twitter datasets and demonstrates better performance compared to the standard Pólya urn model. Moreover, performances variability on different datasets highlights different emotional sensitivities with respect to a public event.

3.1.1.6 NETWORK ANALYSIS OF TWITTER DISCUSSIONS

Partners involved: IMT

We analyzed user online behavior with particular emphasis on group polarization during debates and echo-chambers formation. In particular, we studied semantic aspects of the online relations between users by applying a two-steps approach to three different discussions.

We identified the discursive communities animating the political debate in the run up of the 2018 Italian Elections as groups of users with a significantly-similar retweeting behavior. We studied the semantic mechanisms that shape their internal discussions by monitoring, on a daily basis, the structural evolution of the semantic networks they induce. Our approach implements a statistical method that guarantees that our inference of socio-semantic structures is not biased by any unsupported assumption about missing information. The method is completely automated as it does not rely upon any manual labelling. Our method is applicable to any Twitter discussion regardless of the language or the topic addressed.

We studied the debate about migration policies, combining our projection technique and a community detection algorithm to highlight the dynamics characterizing the relationship among the governmental parties. An example is provided by the effects of the 2019 Italian government crisis on a number of center-left leaning users (future members of the Italia Viva party), whose Twitter activity become markedly different from that of their former party members (supporters of the *Partito Democratico*). On the semantic side, the networks of hashtags are characterized by a mesoscale structure which appears to be core-periphery, a recurrent pattern that is present also in the semantic networks characterizing the online debate in the run up of the 2018 Italian elections.

We studied the debate about the effects of the COVID-19 pandemic. We downloaded all Twitter posts from 1st of March 2020 to 17th of November 2020 by the accounts of the largest Italian firms (those with 250 or more employees). Then, we built the bipartite network of accounts and hashtags and, using an entropy-based null model as a benchmark, projected the network onto the layer of accounts, linking any two accounts if found "similar enough" in terms of their usage of hashtags. We find that the conversation is focused around 13 communities, 10 of which include COVID-19 themes. The core of the network is formed of 5 communities, which deal with environmental sustainability, digital innovation and safety. Firms' ownership type does not seem to influence the conversation. 10 communities out of 13 mention hashtags related to CSR, with the environmental and social dimensions as the prevalent ones. Interestingly enough, the social dimension seems more relevant in the communities dealing with digital innovation and safety. However, the relevance of CSR hashtags is very small at the single message level, but with some peculiarities arising in specific communities. Overall, our paper highlights the role of network methods on Twitter data as a tool which can support managers and policy makers to design their strategies and decision making, capturing firms' emerging issues and relevant themes.

3.1.1.7 THE ROLE OF BOT SQUADS IN THE POLITICAL PROPAGANDA ON TWITTER

Partners involved: IMT

Social media is a prominent channel for news dissemination, information exchange, and fact checking. Quite unexpectedly, automated accounts, known as social bots, contribute more and more to this process of information diffusion. By using Twitter as a benchmark, we considered the traffic exchanged, over one month of observation, on the migration flux from Northern Africa to Italy. We measured the significant traffic of tweets only, by implementing an entropy-based null model that discounts the activity of users and the virality of tweets. Results showed that social bots play a central role in the exchange of significant content. Indeed, not only the strongest hubs have a number of bots among their followers higher than expected, but furthermore a group of them, that can be assigned to the same political tendency, share a common set of bots as followers. The retweeting activity of such automated accounts amplifies the hubs' messages.

3.1.1.8 MICRO-PROJECTS

- Flow of Online Attention in Societal Debates
    - Status: active
    - Partners: CNRS

  - External partners: GipsaLab, Université de Grenoble; GEMASS, Sorbonne Université; LISIS, Université Marne la Vallée; Department of Communication and Arts, Roskilde University
  - Expected outputs: Public campaigns/events reports; A series of collaborative research events (datathons), blog post
- Exploring the concept of non-sematical healthiness measure of online discussions
  - Status: active
  - Partners: BSC; CNRS
  - External partners: none
  - Expected output: Method, multiple experiments, paper and a blog post
- Vaccine disinformation micro-project
  - Status: active
  - Partners: USFD, WAFI - IIT, CSD
  - External partners: none
  - Expected output: Dataset, analysis services, Blog post

3.1.1.9 PUBBLICATIONS

- Vasiliauskaite, Vaiva, Nino Antulov-Fantulin, and Dirk Helbing. "Some Challenges in Monitoring Epidemics." *arXiv preprint arXiv:2105.08384* (2021).
- Guido Caldarelli, Rocco de Nicola, Marinella Petrocchi, Manuel Pratelli, Fabio Saracco, “Flow of online misinformation during the peak of the COVID-19 pandemic in Italy”, arXiv:2010.01913 (2020)
- Alessia Patuelli, Guido Caldarelli, Nicola Lattanzi, Fabio Saracco, “Firms' Challenges and Social Responsibilities during Covid-19: a Twitter Analysis”, arXiv:2103.06705 (2021)
- Tommaso Radicioni, Tiziano Squartini, Elena Pavan, Fabio Saracco, “Networked partisanship and framing: a socio-semantic network analysis of the Italian debate on migration” arXiv:2103.04653 (2021)
- Tommaso Radicioni, Fabio Saracco, Elena Pavan, Tiziano Squartini, “Analysing Twitter Semantic Networks: the case of 2018 Italian Elections” arXiv:2009.02960 (2020)
- Tommaso Radicioni, Tiziano Squartini, Elena Pavan, Fabio Saracco, “Networked partisanship and framing: a socio-semantic network analysis of the Italian debate on migration” arXiv:2103.04653 (2021)
- Giacomo Aletti, Irene Crimaldi, Fabio Saracco, “A model for the Twitter sentiment curve”. PLoS ONE 16(4): e0249634. https://doi.org/10.1371/journal.pone.0249634 (2021)

3.1.2 Planned Activities

3.1.2.1 DATA COLLECTION

Partners involved: IMT, ETHZ

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 analyzing 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. presidential elections) will be collected from Twitter.

3.1.2.2 SOFTWARE DEVELOPMENT

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.

3.1.2.3 EVENTS

Partners involved: IMT

Due to COVID-19 pandemic there are some delays in organizing events and summer schools. Circumstances permitting, we are hoping to organize a sequence of workshops and summer school on misinformation analysis starting from late summer 2021. In particular, in 2022 we will continue summer school on Computational Misinformation Analysis, which aims to set out the state-of-the-art and challenges in computational misinformation analysis (we already organized a summer school in 2019, http://www.sobigdata.eu/events/summer-school-computational-misinformation-analysis). We will also propose tutorials and workshops on misinformation in international conferences such as WSDM, WWW, WebSci and SocInfo.

3.2 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.

3.2.1 Activities Report

3.2.1.1 NOVEL TOOLS FOR THE ANALYSIS OF ECONOMIC AND FINANCIAL NETWORKED SYSTEMS

Partners involved: IMT

1) We established a novel approach to network reconstruction, based upon the maximization of the *conditional Shannon entropy*. It provides an unbiased recipe for inferring pair-specific weights that is capable of taking as input any binary network configuration, on top of which it “redistributes” the weighted marginals,

treated as the (only) constraints of this optimization problem. Besides, we defined a generalized approach to rigorously compare weighted reconstruction methods, based upon the concept of generalized likelihood.

2) We developed a method, based on Random Matrix Theory, to identify the optimal hierarchical decomposition of a system into internally-correlated and mutually anti-correlated communities. By means of this technique, we resolved the mesoscopic structure of the Credit Default Swap (CDS) market and identified groups of issuers that cannot be traced back to standard industry/region taxonomies, thereby being inaccessible to standard factor models.

3) We compared the performance of three algorithms -- Newton's method, a quasi-Newton method and a recently-proposed fixed-point recipe - to solve several Exponential Random Graph Models (ERGMs), defined by binary and weighted constraints in both a directed and an undirected fashion. While Newton's method performs best for little networks, the fixed-point recipe is better for large configurations, as it ensures convergence to the solution within seconds for networks with hundreds of thousands of nodes (e.g., the Internet, Bitcoin).

3.2.1.2 ANALYSIS OF CRYPTOCURRENCIES

Partners involved: IMT, UNIROMA1

1) We reviewed recent results concerning the structural properties of several Bitcoin Transaction Networks, finding that the system has grown over time, becoming increasingly sparse. Moreover, Bitcoin has self-organized itself around a core-periphery structure over a long period of time. Such a peculiar topological organization is observed also on the Bitcoin Lightning Network; there, it is accompanied by a largely uneven distribution of bitcoins, suggesting that Bitcoin is becoming an increasingly centralized system at different levels.

2) We analyzed the structure of the Bitcoin Lightning Network (BLN) over a period of 18 months, studying the topological properties of both its binary and weighted versions. We find that the total volume of transacted bitcoins approximately grows as the square of the network size and that the bitcoins distribution is very unequal. We also tested the goodness of the Undirected Binary Configuration Model (UBCM) in reproducing the BLN structural features: while it reproduces the disassortative and the hierarchical character of the BLN, it is found to underestimate the centrality of nodes. Further inspection shows that removing hubs leads to the collapse of the network into many components, suggesting that the BLN may be an ideal target of the so-called “split attacks”.

3) We showed two layer-two blockchain-based protocols to prove the association between a subject and an endpoint in a decentralized manner. Our protocols are compatible with a wide variety of endpoints and contribute to filling the gap of the current self-sovereign identity management (IdM) approaches with respect to decentralization. We analyzed the security of our proposals and evaluated performances and costs against the common approaches.

4) A Robinson list protects users' privacy against spam calls by allowing them to express consent or denial about market operator calls. We investigated the possibility of implementing this valuable tool as a decentralised service, identifying the requirements of a decentralized Robinson list and the characteristics of

a blockchain technology needed to support an implementation. Then, we presented a general solution and a proof-of-concept implementation based on the Algorand technology. A preliminary performance evaluation suggests that we can achieve satisfactory results.

3.2.1.3 ANALYSIS OF THE RELATIONSHIP BETWEEN HUMAN BEHAVIOUR, CLIMATE CHANGE, ECONOMY AND FINANCE

Partners involved: UNIROMA1, ISTI-CNR

We used GPS traces and a microscopic model to analyse the emissions of four air pollutants from thousands of vehicles in three European cities. We discovered the existence of gross polluters, vehicles responsible for the greatest quantity of emissions, and grossly polluted roads, which suffer the greatest amount of emissions. Our simulations show that emissions reduction policies targeting gross polluters are way more effective than those limiting circulation based on a non-informed choice of vehicles. Our study applies to any city and may contribute to shaping the discussion on how to measure emissions with digital data.

3.2.1.4 ECONOMY AND FINANCE - PREDICTING BANKRUPTCY OF LOCAL GOVERNMENTS

Partners involved: ETHZ

We analyzed the predictability of the bankruptcy of 7795 Italian municipalities in the period 2009–2016. The prediction task is extremely hard due to the small number of bankruptcy cases, on which learning is possible. Besides historical financial data for each municipality, we use alternative institutional data along with the socio-demographic and economic context. The predictability is analyzed through the performance of the statistical and machine learning models with a receiver operating characteristic curve and the precision-recall curve. Our results suggest that it is possible to make out-of-sample predictions with a high true positive rate and low false-positive rate. The model shows that some non-financial features (e.g., geographical area) are more important than many financial features to predict the default of municipalities.

3.2.1.5 DATA SCIENCE AND MACHINE LEARNING TECHNIQUES IN FINANCE: ANALYSIS OF VOLATILITY IN CRYPTOCURRENCY MARKETS

Partners involved: ETHZ

We studied which online social media signals reduce the uncertainty about asset prices and their changes, concentrating on cryptocurrency markets and quantified microscopic interrelations of financial assets and their relation to distress propagation through the network. In particular, we built a tool to acquire data from Twitter in real time related to financial news. Preliminary findings show that twitter data can be utilised to make significantly better predictions about cryptocurrency market volatility. In future we will also consider which elements in particular (tweet volume, sentiment, etc.) contribute the most to uncertainty reduction.

We also analysed the time series of minute price returns on the Bitcoin market through the statistical models of the generalized autoregressive conditional heteroscedasticity (GARCH) family. We combined an approach that uses historical values of returns and their volatilities—GARCH family of models, with a so-called Mixture

of Distribution Hypothesis, which states that the dynamics of price returns are governed by the information flow about the market.

3.2.1.6 DATA SCIENCE AND MACHINE LEARNING TECHNIQUES IN FINANCE: ANALYSIS OF 2018 BITCOIN CRASH

Partners involved: ETH, SNS

We studied the market microstructure of Bitcoin related to liquidity during the bitcoin bubble in 2018; set up a framework to analyze information dynamics in cryptocurrency markets and performed empirical analysis of the bitcoin bubble in 2018. We began writing a preprint for the analysis of market microstructure during the 2018 bubble.

3.2.1.7 DATA SCIENCE AND MACHINE LEARNING TECHNIQUES IN FINANCE - GRAPH NEURAL NETWORK TECHNIQUE FOR FINANCES

Partners involved: UT

We exploited a specific kind of Graph Neural Network approach called Temporal-Graph Convolutional Network (T-GCN) for predicting the amount of Bitcoins received by a customer at a particular timestamp in the Bitcoin historical financial transactions network. The lower errors obtained using T-GCN compared to 11 baselines (such as support vector regressors, random forest regressors, vector auto-regressors, long short-term memory networks) demonstrated the effectiveness of our approach.

3.2.1.8 ESTIMATING WELL-BEING AND SOCIO-DEMOGRAPHICS USING DIGITAL DATA

Partners involved: SNS, UNIPI, ISTI-CNR

1) We exploited information extracted from the GDELT (Global Data on Events, Location, and Tone) database of digital news to capture peacefulness through the Global Peace Index (GPI). Applying machine learning techniques, we demonstrated that news media attention, sentiment, and social stability from GDELT can be used as proxies for measuring GPI at a monthly level.

2) Using ground truth data describing phone records of 65 volunteers together with their current address of residence, we provided an unprecedented evaluation of the accuracy of home detection algorithms and quantified the amount of records per individual needed to carry out successful home detection. Our results demonstrated that our work is useful for researchers and practitioners to minimize data requests and maximize the accuracy of the home antenna location.

3) Researchers have suggested two main approaches for the overall measurement of well-being, the objective and the subjective well-being. Both approaches, as well as their relevant dimensions, have been traditionally captured with surveys. During the last decades, new data sources have been suggested as an alternative or complement to traditional data. We made a survey paper that presents the theoretical background of well-being, by distinguishing between objective and subjective approaches, their relevant dimensions, the new data sources used for their measurement and relevant studies. In the survey, we also

shed light on still barely unexplored dimensions and data sources that could potentially contribute as a key for public policing and social development.

3.2.1.9 MODELING AND FORECASTING OF NON-STATIONARY FINANCIAL TIME SERIES

Partners involved: SNS

We generalized the Kinetic Ising Model using the score-driven approach, which allows the efficient estimation and filtering of time-varying parameters from time series data. We quantified the amount of noise in the data and the reliability of forecasts, and applied our methodology to forecasting high-frequency volatility of stocks, measuring its endogenous component during extreme events in the market and analysing the strategic behaviour of traders around news releases.

3.2.1.10 DATA SCIENCE AND MACHINE LEARNING TECHNIQUES IN FINANCE

Partners involved: SNS

We considered a model of a financial system consisting of a leveraged investor that invests in a risky asset and manages risk by using Value-at-Risk (VaR). We showed that the leverage dynamics can be described by a dynamical system of slow-fast type associated with a unimodal map with an additive heteroscedastic noise whose variance is related to the portfolio rebalancing frequency to target leverage. We used deep neural networks to estimate map parameters from a short time series in a dataset of US commercial banks over the period 2001-2014. We found that the parameters of a substantial fraction of banks lie in the dynamical core, and their leverage time series are consistent with chaotic behavior. We also presented evidence that the time series of the leverage of large banks tend to exhibit chaoticity more frequently than those of small banks.

3.2.1.11 DATA COLLECTION: NOVEL FINANCE 2.0-RELATED DATASETS FROM TWITTER

Partners involved: ETHZ

With the aim of collecting novel finance 2.0-related datasets, we built a tool to sense financial news through Twitter and collected approximately 6 months' worth of data.

3.2.1.12 SOFTWARE: PYTHON PACKAGE TO SOLVE MAXIMUM ENTROPY MODELS

Partners involved: IMT

We released the Python NEMtropy package, which provides the user with a state-of-the-art solver for a range variety of Maximum Entropy Networks models derived from the Exponential Random Graph Models (ERGM) family. The package allows the user to solve the desired model and generate a number of randomized graphs from the original one: the so-called graphs ensemble.

3.2.1.13 EVENT: COMPLEXITY MEETS FINANCE

Partners involved: IMT

We organized the satellite “Complexity meets finance: Data, Methods and Policy Implications” within NetSci 2020 (https://sites.google.com/view/cmf20/home). The satellite aimed at bridging the gap between the fields of complex networks theory and finance by bringing together experienced researchers and young scholars to discuss state-of-the-art work, share knowledge and create opportunities for novel and fruitful collaborations.

3.2.1.14 MICRO-PROJECTS

- Analysis of Bitcoin crash in 2017/2018
  - Status: active
  - Partners: ETHZ, SNS
  - External partners: none
  - Expected output: Preprint/publication, blog post, on-site access only Dataset
- Scientific Migration and Scientific Social Networks Specific Research Question: What is the effect of scientific networks on scientific migration? Evidence from Microsoft Academic Knowledge Graph
  - Status: active
  - Partners: PSE, ISTI-CNR, UNIPI
  - External partners: none
  - Expected output: Dataset, Paper, Blog article
- Endogenous and exogenous volatility
  - Status: active
  - Partners: SNS
  - External partners: none
  - Expected output: Experiment, Paper, blog post
- Estimating countries’ peace index with GDELT and machine learning techniques
  - Status: active
  - Partners: SNS, Unipi, ISTI-CNR
  - External partners: none
  - Expected output: Method, Experiments, Blog post, Preprint/paper
- Quantifying the presence of air pollutants over a road network in high spatio-temporal resolution
  - Status: active
  - Partners involved: UNIROMA1, ISTI-CNR
  - External partners: none
  - Expected output: Method, Experiments, Preprint paper, Blog post
- A dataset to assess mobility changes in Chile following local quarantines
  - Status: active
  - Partners: UNIPI, ISTI-CNR
  - External partners: Universidad del Desarrollo of Santiago de Chile and Telefonica Chile
  - Expected output: Paper, Dataset, Blog post

- Changes in visiting patterns to venues during the COVID-19 pandemic in the US
    - Status: active
    - Partners: ISTI-CNR
    - External partners: Fondazione Bruno Kessler (FBK)
    - Expected output: Method, Experiments, preprint/paper, blog post
- Validation on home location detection algorithms on ground truth
    - Status: finished
    - Partners involved: ISTI-CNR
    - External partners: Universidad del Desarrollo de Santiago de Chile (Chile), University of Turin (Italy), Telefónica Chile (Chile)
    - Outputs:
        - Method https://data.d4science.org/ctlg/ResourceCatalogue/ground_truth_evaluation_of_home_location_detection_algorithms
        - Experiment https://data.d4science.org/ctlg/ResourceCatalogue/evaluation_of_home_location_detection_algorithms_on_three_mobile_phone_records_streams_and_ground_t

3.2.2 Pubblications

- F. Parisi, T. Squartini, D. Garlaschelli, A faster horse on a safer trail: generalized inference for the efficient reconstruction of weighted networks, New Journal of Physics 22, 2020
- Anagnostou, T. Squartini, D. Kandhai, D. Garlaschelli, Uncovering the mesoscale structure of the credit default swap market to improve portfolio risk modelling, Quantitative Finance 1-18, 2021
- N. Vallarano, M. Bruno, E. Marchese, G. Trapani, F. Saracco, G. Cimini, M. Zanon, T. Squartini, Fast and scalable likelihood maximization for Exponential Random Graph Models, https://arxiv.org/abs/2101.12625, 2021
- N. Vallarano, C. J. Tessone, T. Squartini, Bitcoin Transaction Networks: an overview of recent results, Frontiers in Physics 8 286, 2020
- J.H. Lin, K. Primicerio, T. Squartini, C. Decker, C. J. Tessone, Lightning Network: a second path towards centralisation of the Bitcoin economy, New Journal of Physics 22, 2020
- D. Pennino, M. Pizzonia, A. Vitaletti, M. Zecchini, Binding of Endpoints to Identifiers by On-Chain Proofs, 2020 IEEE Symposium on Computers and Communications (ISCC), 2020
- An extension of this work has been submitted to IEEE Access journal with the title “Efficient Certification of Endpoint Control on Blockchain” to IEEE Access journal.
    - Cirillo, A. Mauro, D. Pennino, M. Pizzonia, A. Vitaletti, M. Zecchini, Decentralized Robinson List, Proceedings of the 3rd Workshop on Cryptocurrencies and Blockchains for Distributed Systems, 2020.
- M. Bohm, M. Nanni, L. Pappalardo, Quantifying the presence of air pollutants over a road network in high spatio-temporal resolution, Tackling Climate Change with Machine Learning workshop, NeurIPS, 2020.

- S. Sharma, R. Sharma, Forecasting Transactional Amount in Bitcoin Network Using Temporal GNN Approach, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2020.
- V. Voukelatou, L. Pappalardo, I. Miliou, L. Gabrielli, F. Giannotti, Estimating countries' peace index through the lens of the world news as monitored by GDELT, 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 2020
- Campajola, D. Di Gangi, F. Lillo, D. Tantari, Modelling time-varying interactions in complex systems: the Score Driven Kinetic Ising Model, https://arxiv.org/abs/2007.15545, 2020
- L. Pappalardo, L. Ferres, M. Sacasa, C. Cattuto, L. Bravo, Evaluation of Home Detection Algorithms on Mobile Phone Data Using Individual-Level Ground Truth, EPJ Data Science, 2021.
- V. Voukelatou, L. Gabrielli, I. Miliou, S. Cresci, R. Sharma, M. Tesconi, L. Pappalardo, Measuring objective and subjective well-being: dimensions and data sources, International Journal of Data Science and Analytics, 2020
- M. Bohm, M. Nanni, L. Pappalardo, Quantifying the presence of air pollutants over a road network in high spatio-temporal resolution, Climate Change & AI Neurips2020 Workshop, 2020, https://www.climatechange.ai/papers/neurips2020/28
- Vaiva Vasiliauskaite, Nino Antulov-Fantulin, Dirk Helbing, Some Challenges in Monitoring Epidemics, https://arxiv.org/abs/2105.08384, 2021
- Barjašić Irena, Antulov-Fantulin Nino, Time-Varying Volatility in Bitcoin Market and Information Flow at Minute-Level Frequency, Front. Phys., 21 May 2021 | https://doi.org/10.3389/fphy.2021.644102
- Antulov-Fantulin, Nino, Raffaele Lagravinese, and Giuliano Resce. "Predicting bankruptcy of local government: A machine learning approach." *Journal of Economic Behavior & Organization* 183 (2021): 681-699.
- L. Pappalardo, L. Ferres, M. Sacasa, C. Cattuto, L. Bravo, Evaluation of home detection algorithms on mobile phone data using individual-level ground truth, https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-021-00284-9

3.2.3 Activities planning

Development of novel network tools for the analysis of economic and network financial systems

Partners involved: IMT

Enlarge the basket of Exponential Random Graph Models analysed in the paper "Fast and scalable likelihood maximization for Exponential Random Graph Models".

Analysis of cryptocurrencies

Partners involved: IMT, UNIROMA1

Analyse the weighted structure of the BLN. Analyse the BLN from a dynamical point of view. Extension of "Decentralized Robinson List" to journal version.

Analysis of the relationship between climate change, economy and finance

Partners involved: ISTI-CNR, UNIROMA1

Regarding the relationship between air pollution and urban inequalities, we plan to explore what are the characteristics, both in terms of network centrality/density and socio-economic factors, of the highly polluted neighbourhood and roads in a city. Moreover, we will upload to the SBD++ platform the final publication of our work, which is currently under revision.

Data science and machine learning techniques in finance

Partners involved: ETHZ, SNS

In the following year we plan to: 1) Finish writing a preprint (and submitting) on information dynamics in 2018 bitcoin bubble; 2) Upload relevant data to SoBigData; 3) Write a preprint about social media-related information effects on cryptocurrency markets; Provide SoBigData with data analysed for the 3).

Estimating well-being and socio-demographics with digital data

Partners involved: UNIPI, ISTI-CNR

We plan to extend the study about estimating peacefulness by applying the SHAP methodology (Lundberg, Scott M., et al. "From local explanations to global understanding with explainable AI for trees." Nature machine intelligence 2.1 (2020): 56-67). This methodology will help us deep into the model behaviour and explain it better.

Modeling and forecasting of non-stationary financial time series

Partners involved: SNS

Extend the application to other non-stationary time series from neuroscience and sociology. Apply this approach to model non stationary time series of networks, with application to the interbank system and to financial risk.

Data science and machine learning techniques in finance

Partners involved: SNS

Deep Neural Networks for chaotic financial time series and systemic risk. Develop methods to infer complex (financial) time series using Recurrent Neural Networks

Data collection

Partners involved: ETHZ

Novel finance 2.0-related datasets from Twitter. In the following year we plan to: 1) Process the collected data; 2) Upload it to SBD++.

Software

Partners involved: IMT

Python package to solve maximum entropy models. Enlarge the basket of Exponential Random Graph Models that can be solved with the NEMtropy package.

Events

Partners involved: ETHZ

We will organize a half-day workshop on data science and machine learning in finance by the end of 2021, tentatively in November. The aim of the upcoming workshop will be to enable research exchange and to share insights from different data-intensive disciplines. The event will be organized together with other partners of the consortium and/or external partners. It will include a series of invited talks and facilitate discussion as well as networking opportunities.

3.3 T10.3 Sustainable Cities for Citizens

This exploratory focus on the analysis of cities, the sustainability of their flows of energy and materials and people living in them. We analyzed data from different spatial and temporal scales. On city-wide scales, we analyzed 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 analyzed mobility in different cities, allowing the characterization of the demand of dynamic users and granting the derivation of models to estimation pollution and optimize the electric mobility charging and relocation service and minimize its impact on the power grid.

3.3.1 Activities Report

3.3.1.1 COVID AND CLIMATE CHANGE

Partners involved: IMT

We developed a survey (currently ongoing in Italy, France, UK, and Germany, with 1200 participants involved) in cooperation with the University of Bari, the University of Loughborough, and the Climate Media Centre Italia. The survey covers the risk perceived by citizens both with respect to COVID-19 and climate change. The cooperative profile of the participants is investigated by means of a public good game, while common beliefs related to both topics are considered in the administered questions.

3.3.1.2 OPTIMAL PLANNING OF REGIONAL RENEWABLE ENERGY SOURCES

Partners involved: IMT

We developed a case study in Tuscany in which data covering 70 weather stations (measuring solar radiation and wind speed) have been collected with the support of "*Servizio Idrologico Regionale*". Starting from weather data, the energy production profile of photovoltaic and wind power standard generation stations

has been estimated. A set of optimal locations for the installation of PV and Wind generators has been identified, also showing that mixing energy sources is able to reduce the impact of renewable energy sources.

3.3.1.3 URBAN METABOLISM OF ONE ITALIAN MUNICIPALITY

Partners involved: IMT, Eliante

Urban metabolism consists in the assessment of inflows and outflows of material and energy crossing the boundary of a city. The ongoing activities are: 1) Data collection of energy (electricity, fuels), materials (water, waste, food, construction) and sustainability/circularity policies implemented in the municipality; 2) Data analysis; 3) Policy recommendations.

The purpose of the analysis is to assess the sustainability of the municipalities under the point of view of resource use, emissions, and effectiveness of the environmental policies. The results will be part of the Atlas of Urban Sustainability.

In the past year, we selected the municipalities of Pisa (Tuscany), Monopoli (province of Bari, Apulia) and Tortona (province of Alessandria, Piedmont) and contacted decision makers for Pisa and Monopoli, while the mayor of Tortona agreed to support the data collection. We are now in the process of starting the data collection phase.

3.3.1.4 URBAN GREEN STATE OF THE ART FOR PISA

Partners involved: IMT, Eliante, ISTI-CNR

In collaboration with the University of Milan, we reviewed the current data and literature on urban green in cities. We started preparing a survey paper.

3.3.1.5 IMPACT OF COVID-19 ON EMPLOYMENT RISK

Partners involved: IMT

This activity is focused on a vaccination roll-out policy based on a nation-wide analysis covering the economic and health impact of lockdown measures introduced in Italy in April 2020. We used data on fine-scale human mobility, furlough supporting measures, and mortality. Results showed that prioritizing essential workers and regions where the effect of lockdown on employment risk is stronger leads to a different vaccine distribution policy based on the number of inhabitants in a specific region. Moreover, we showed that the confinement measures had a greater impact on the economically vulnerable people in urban areas, increasing their risk of marginalization.

3.3.1.6 HUMAN MOBILITY ANALYSIS AND MODELLING

Partners involved: UNIPI, ISTI-CNR

1) We developed STS-EPR, a mechanistic model that captures the spatial, temporal, and social dimensions of human mobility together. We performed experiments on check-ins from Foursquare in various cities around

the world, showing that STS-EPR generates realistic trajectories, making it better than models that lack either in the social, the spatial, or the temporal mechanisms.

2) Using mobile phone data from February through September 2020, we investigated the relationship between human mobility and the spread of COVID-19. We found that the time needed to switch off mobility and bring the net reproduction number below the critical threshold is about one week. Moreover, we observed a strong relationship between the number of days spent above such threshold before the lockdown-induced drop in mobility flows and the total number of infections per 100k inhabitants. Estimating the statistical effect of mobility flows on the net reproduction number over time, we documented a 2-week lag positive association, strong in March and April, and weaker but still significant in June.

3) To tackle the COVID-19 pandemic, Chile implemented quarantines at a more localized level, shutting down small administrative zones, rather than the whole country or large regions. To assess the impact on human mobility of the localized quarantines in Chile, we analyzed a mobile phone dataset made available by Telefónica Chile, which comprises 31 billion eXtended Detail Records and 5.4 million users covering the period February 26th to September 20th, 2020. From these records, we derived three epidemiologically relevant metrics describing the mobility within and between comunas. The datasets made available can be used to fight the COVID-19 epidemics, particularly for localized quarantines' less understood effect.

4) We are preparing a survey paper that provides: *(i)* basic notions on human mobility and deep learning; *(ii)* a description of deep learning models for next-location prediction, crowd flow prediction, flow prediction, trajectory generation, and flow generation; and *(iii)* a discussion about relevant open challenges. Our survey is a guide to the leading deep learning solutions to next-location prediction, crowd flow prediction, and trajectory generation. At the same time, it may help deep learning scientists and practitioners understand the fundamental concepts and the open challenges of the study of human mobility.

5) Since the movements of individuals within and among cities influence key aspects of our society, there is increasing interest around the challenging problem of flow generation, i.e., generating the flows between a set of geographic locations, given the characteristics of the locations and without any information about the real flows. Existing solutions to flow generation are mainly based on mechanistic approaches, such as the gravity model and the radiation model, which suffer from underfitting and overdispersion, and cannot describe non-linear relationships between variables. We developed the Deep Gravity model (DG), which exploits a large number of variables (e.g., land use and road network; transport, food, and health facilities) and deep neural networks to describe complex non-linear relationships between them. Our experiments, conducted on commuting flows in England, showed that DG achieves a significant increase in the performance (up to 250% for highly populated areas) than mechanistic models that do not use deep neural networks, or that do not exploit geographic data.

6) Traditional frameworks for privacy risk assessment systematically generate the assumed knowledge for a potential adversary, evaluating the risk without realistically modelling the collection of the background knowledge used by the adversary when performing the attack. We proposed Simulated Privacy Annealing (SPA), a new adversarial behavior model for privacy risk assessment in mobility data. We modeled the behavior of an adversary as a mobility trajectory and introduced an optimization approach to find the most effective adversary trajectory in terms of privacy risk produced for the individuals represented in a mobility

data set. We used simulated annealing to optimize the movement of the adversary and simulate a possible attack on mobility data. We finally tested the effectiveness of our approach on real human mobility data, showing that it can simulate the knowledge gathering process for an adversary in a more realistic way.

7) We built a data-driven model for predicting car drivers' risk of experiencing a crash in the long-term future, for instance, in the next four weeks. Since raw mobility data typically lack any explicit semantics or clear structure, our work proposes to build concise representations of individual mobility, that highlight mobility habits, driving behaviors and other factors deemed relevant for assessing the propensity to be involved in car accidents. The suggested approach is based on a network representation of users' mobility, called Individual Mobility Networks, jointly with the analysis of descriptive features of the user's driving behavior related to driving style (e.g., accelerations) and characteristics of the mobility in the neighborhood visited by an individual. We tested the methods over a real dataset, showing comparative performances against baselines and competitors, and a study of some typical risk factors in the areas under analysis through the adoption of state-of-art model explanation techniques. Preliminary results show the effectiveness and usability of the proposed predictive approach.

8) We tackled the fundamental problem of predicting the impact of planning and construction activities on mobility flows. These flows can be modelled as attributed graphs with both node and edge features characterizing locations in a city and the various types of relationships between them. We addressed the problem of assessing origin-destination (OD) car flows between a location of interest and every other location in a city, given their features and the structural characteristics of the graph. We adopted three neural network architectures, including graph neural networks (GNN), and conducted a systematic comparison between the proposed methods and state-of-the-art spatial interaction models, their modifications, and machine learning approaches. We evaluated the performance of the models on a regression task using a custom data set of attributed car OD flows in London, also producing visualizations of the model performance by showing the spatial distribution of flow residuals across London.

9) Identifying the portions of trajectory data where movement ends and a significant stop starts is a basic, yet fundamental task that can affect the quality of any mobility analytics process. Most of the many existing solutions adopted by researchers and practitioners are simply based on fixed spatial and temporal thresholds stating when the moving object remained still for a significant amount of time, yet such thresholds remain as static parameters for the user to guess. We studied the trajectory segmentation problem from a multi-granularity perspective, looking for a better understanding of the problem and for an automatic, user-adaptive and essentially parameter-free solution that flexibly adjusts the segmentation criteria to the specific user under study and to the geographical areas they traverse. Experiments over real data, and comparison against simple and state-of-the-art competitors show that the flexibility of the proposed methods has a positive impact on results.

10) In the context of mobility-related individual events, such as crash prediction, we studied the problem of geographical transfer learning of predictive models - namely to exploit the data available in some geographical areas to build effective predictive models for another area where labeled data is not available - and developed three solutions at different levels of complexity. In particular, we rely on city similarity indicators that, opposed to generic transfer learning solutions, directly exploit the semantics of our mobility data. Empirical results over real datasets show the superiority of our solution.

3.3.1.7 SOFTWARE: SCIKIT-MOBILITY LIBRARY

Partners involved: ISTI-CNR, UNIPI

We released a first stable version of library scikit-mobility, which is a Python package for human mobility analysis that allows the user to: 1) represent trajectories and mobility flows with proper data structures, TrajDataFrame and FlowDataFrame; 2) manage and manipulate mobility data of various formats (call detail records, GPS data, data from social media, survey data, etc.); 3) extract mobility metrics and patterns from data, both at individual and collective level (e.g., length of displacements, characteristic distance, origin-destination matrix, etc.); 4) generate synthetic individual trajectories using standard mathematical models (random walk models, exploration and preferential return model, etc.); 5) generate synthetic mobility flows using standard migration models (gravity model, radiation model, etc.) assess the privacy risk associated with a mobility data set.

3.3.1.8 MICRO-PROJECTS

- Urban green dataset for Sustainable cities for citizens in Pisa, Italy
  - Status: active
  - Partners: ISTI-CNR, ELI, IMT
  - External partners: municipality of Pisa
  - Expected output: Dataset, Blog post
- Automated methods of urban green analysis - State of the art
  - Status: active
  - Partners: IMT, ELI, ISTI-CNR
  - External partners: Department of Agricultural and Environmental Sciences, University of Milan; ISAFoM-CNR
  - Expected output: Technical report, presentation, blog post
- A dataset to assess mobility changes in Chile following local quarantines
  - Status: active
  - Partners: Unipi, ISTI-CNR
  - External partners: Universidad del Desarrollo de Santiago de Chile (Chile), Telefónica Chile
  - Expected outputs: Paper, Dataset, Blog post
- Quantifying the presence of air pollutants over a road network in high spatio-temporal resolution
  - Status: active
  - Partners: UNIROMA1, ISTI-CNR
  - External partners: none
- Data module in scikit-mobility
  - Status: active
  - Partners involved: UNIPI, ISTI-CNR
  - External partners: Free University of Bolzano
  - Expected output: Module in scikit-mobility, video tutorial, example notebooks, blog post, datasets, methods
- A survey on Deep Learning for Human Mobility

  - Status: active
  - Partners: ISTI-CNR
  - External partners: Fondazione Bruno Kessler (FBK, Trento, Italy), Amazon Alexa (Berlin, Germany)
  - Expected output: paper, blog post
- Enhancing mobility flows generation with deep neural networks and geographic information
  - Status: active
  - Partners: ISTI-CNR
  - External partners: Fondazione Bruno Kessler (FBK, Trento, Italy), Amazon Alexa (Berlin, Germany), University of Bristol (UK)
  - Expected outputs: method, experiments, paper/preprint, blog post
- Development and integration of mobility models GeoSim and STS-EPR
  - Status: completed
  - Partners: Unipi, ISTI-CNR
  - External partners: none
  - Outputs:
    - Method (STS-EPR) https://data.d4science.org/ctlg/ResourceCatalogue/sts-epr_-_trajectory_generator
    - Method (GeoSim) https://data.d4science.org/ctlg/ResourceCatalogue/geosim_-_trajectory_generator_integration_in_scikit-mobility_
    - Experiment https://data.d4science.org/ctlg/ResourceCatalogue/simulation_of_individual_mobility_using_sts-epr
    - Blog post http://www.sobigdata.eu/blog/simulating-human-mobility-considering-social-dimension-too-0

3.3.1.9 PUBBLICATIONS

- F. Surmonte, U. Perna, A. Scala, A. Rubino, A. Facchini, A Data-driven approach to renewable energy source planning at regional level, Energy Sources Part B: Economics, Planning and Policy, 2021
- V. Pieroni, A. Facchini, M. Riccaboni, COVID-19 and Unemployment Risk: Lessons for the Vaccination Campaign, https://arxiv.org/abs/2102.03619, 2021
- G. Cornacchia, L. Pappalardo, STS-EPR: Modelling individual mobility considering the spatial, temporal, and social dimensions together, Procedia Computer Science 184 (2021): 258-265.
- Gennady Andrienko, Natalia Andrienko, Chiara Boldrini, Guido Caldarelli, Paolo Cintia, Stefano Cresci, Angelo Facchini, Fosca Giannotti, Aristides Gionis, Riccardo Guidotti, Michael Mathioudakis, Cristina Ioana Muntean, Luca Pappalardo, Dino Pedreschi, Evangelos Pournaras, Francesca Pratesi, Maurizio Tesconi and Roberto Trasarti, (So) Big Data and the transformation of the city, *International Journal of Data Science and Analytics*, 2020, https://doi.org/10.1007/s41060-020-00207-3

- Cintia et al., The relationship between human mobility and viral transmissibility during the COVID-19 epidemics in Italy, https://arxiv.org/abs/2006.03141, 2021
- L. Pappalardo, F. Simini, G. Barlacchi, R. Pellungrini, scikit-mobility: a Python library for the analysis, generation and risk assessment of mobility data, Journal of Statistical Software (JSS), 2021
- L. Pappalardo, G. Cornacchia, V. Navarro, L. Bravo, L. Ferres, A dataset to assess mobility changes in Chile following local quarantines, https://arxiv.org/abs/2011.12162
- M. Luca, G. Barlacchi, B. Lepri, L. Pappalardo, Deep Learning for Human Mobility: a Survey on Data and Models, arXiv preprint arXiv:2012.02825, 2020
- Filippo Simini, Gianni Barlacchi, Massimiliano Luca, Luca Pappalardo, Deep Gravity: enhancing mobility flows generation with deep neural networks and geographic information, https://arxiv.org/abs/2012.00489, 2020
- R. Pellungrini, L. Pappalardo, F. Simini, A. Monreale, Modeling Adversarial Behavior Against Mobility Data Privacy, IEEE Transactions on Intelligent Transportation Systems, 2020
- R. Guidotti, M. Nanni, Crash Prediction and Risk Assessment with Individual Mobility Networks, 21st IEEE International Conference on Mobile Data Management (MDM), 2020, https://ieeexplore.ieee.org/document/9162285
- Y. Gevorg, et al., Learning Mobility Flows from Urban Features with Spatial Interaction Models and Neural Networks, Proceedings of 2020 IEEE International Conference on Smart Computing (SMARTCOMP 2020), 2020.
- Bonavita, R. Guidotti, M. Nanni, Self-Adapting Trajectory Segmentation, EDBT/ICDT Workshops. 2020.
- M. Nanni, A. Bonavita, R. Guidotti, City Indicators for Mobility Data Mining, EDBT/ICDT Workshops, 2021

3.3.2 Activities Planning

Optimal planning of regional renewable energy sources.

Partners involved: IMT

An upscaling of the method is expected at national scale. Provided the availability of data the planning tool will be extended to other Italian regions. The national grid will be included in the study with the aim to better estimate the impact of renewables in Italy.

Urban metabolism of one Italian municipality.

Partners involved: IMT

Data collection is expected to start in the selected Municipalities.

Development of Case study in Pisa.

Partners involved: IMT, Eliante, ISTI-CNR

According to the pandemic risk, data on urban green and biodiversity will be collected in Spring 2022 in Pisa, with the support of Eliante, CNR and IMT. The activities will be developed by means of suitable microproject in cooperation with external stakeholders and partners of the SBD++ consortium

Case study on Pristina.

Partners involved: CRA, IMT, Eliante, ISTI-CNR

A case study in Pristina will be developed in cooperation with CRA and Manifesta.eu Data on local food flows of informal markets will be collected. A survey is also expected to explore the behaviors. A specific microproject will be developed by CRA in cooperation with IMT. A cooperation with the Migration studies Task is expected.

Stakeholders Involved: University of Bari, University of Milan, Climate Media Centre Italia, University of Loughborough, Municipality of Tortona, Servizio Idrologico Regionale – Regione Toscana.

3.4 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).

3.4.1 Activities Report

3.4.1.1 NOWCASTING MIGRATION STOCKS

Partners involved: UNIPI, ISTI-CNR

We developed nowcasting models for stocks of immigrants for Europe and European countries by applying regression algorithms to the Superdiversity Index and measures of diversity derived from Twitter data.

3.4.1.2 STUDYING MIGRANT INTEGRATION WITH SOCIAL NETWORK DATA

Partners involved: UNIPI, ISTI-CNR, PSE

We studied cultural integration on Twitter by defining home and destination attachment indexes, respectively, preserving links to the home country and adopting cultural traits from the new residence country. We also studied the characteristics and behaviors of migrants and natives on Twitter by combining different features, including profiles and tweets, and extensive network analysis on the network.

3.4.1.3 HIGHLY SKILLED MIGRATION

Partners involved: PSE, UNIPI

We performed a scientific study of migration through big scholarly data and analyzed the trend towards intra- and international collaboration over time through the Yearly Degree of Collaboration Index with Vrije Universiteit Brussel. PSE conducted an Economics research aiming to achieve identification by utilising Microsoft Academic Knowledge Graph (MAKG) and the fall of the Iron Curtain as a natural experiment to assess how pre-fall networks of academics affect their migration decisions post the fall. The research shows no threat of reverse causality and includes a host of controls to limit the potential for omitted variable bias. The project is underway, but the preliminary results align with what is expected from the lens of economic theory and intuition. As such, academics are more likely to stay at "home" if their network at home is of high quality and is greater, the opposite hold for the destination. The results can be interpreted as causal as identification is argued by the Iron Curtain and evidence in support of it is presented. Unipi performed data collection and the public release of an enriched and more easily accessible version of an already existing big dataset of scholarly data. The release will be accompanied by a data-paper.

3.4.1.4 IMMIGRANTS INTEGRATION THROUGH RETAIL DATA

Partners involved: UNIPI, ISTI-CNR, PSE

Food adoption choices are much less exposed to external judgment and social pressure than other individual behaviours, and can be observed over a long period. That makes them an interesting basis for, among other applications, studying the integration of immigrants from a food consumption viewpoint. We analyzed immigrants' food consumption from shopping retail data for understanding if and how it converges towards those of natives. We defined a score of adoption of natives' consumption habits by an individual as the probability of being recognized as a native from a machine learning classifier, thus adopting a completely data-driven approach. We measured the immigrant's adoption of natives' consumption behavior over a long time, and we identified different trends. A case study on real data of a large nation-wide supermarket chain reveals that we can distinguish five main different groups of immigrants depending on their trends of native consumption adoption.

3.4.1.5 OTHER RESEARCHES

Partners involved: ISTI-CNR

Study the European *salad bowl* through Superdiversity Index, as the distance between the standard use of the emotional valence of words and the specific use shown by the population of a region (community) computed using the Emotional Spreading Algorithm.

3.4.1.6 EVENT: HUMMING BIRD ONLINE WORKSHOP

Partners Involved: PSE, ISTI-CNR, UNIPI

This workshop aims to enable the sharing of experiences with big data and migration among an interdisciplinary set of researchers and audience. We want to bring together not only researchers from academia, but also from institutions working with migration, and industry. The overall objective is to understand better what are the plausible areas of study where big data can make a difference, and what are the methodologies employed to date.

3.4.1.7 MICRO-PROJECT

- Scientific Migration and Scientific Social Networks Specific Research Question: What is the effect of scientific networks on scientific migration? Evidence from Microsoft Academic Knowledge Graph.
    - Status: active
    - Partners: PSE, ISTI-CNR, UNIPI
    - External partners: none
    - Expected outputs: Dataset, Paper, Blog article

3.4.1.8 PUBLICATIONS

- Natalia Andrienko, Gennady Andrienko, Silvia Miksch, Heidrun Schumann, and Stefan Wrobel, A theoretical model for pattern discovery in visual analytics, *Visual Informatics,* 2021, vol. 5(1) pp.23-42, published version: https://doi.org/10.1016/j.visinf.2020.12.002
- Siming Chen, Natalia Andrienko, Gennady Andrienko, Jie Li and Xiaoru Yuan, Co-Bridges: Pair-wise Visual Connection and Comparison for Multi-item Data Streams, *IEEE Transactions on Visualization and Computer Graphics (proceedings IEEE VAST 2020),* 2021, vol. 27(2), pp.1612-1622 pre-print, published version: https://doi.org/10.1109/TVCG.2020.3030411
- Gennady Andrienko, Natalia Andrienko, Ibad Kureshi, Kieran Lee, Ian Smith, and Toni Staykova, Automating and utilizing equal-distribution data classification, *International Journal of Cartography,* 2021, vol. 7(1), pp.100-115 pre-print, published version: https://doi.org/10.1080/23729333.2020.1863000
- Alina Sîrbu, Gennady Andrienko, Natalia Andrienko, Chiara Boldrini, Marco Conti, Fosca Giannotti, Riccardo Guidotti, Simone Bertoli, Jisu Kim, Cristina Ioana Muntean, Luca Pappalardo, Andrea Passarella, Dino Pedreschi, Laura Pollacci, Francesca Pratesi & Rajesh Sharma, Human migration: the big data perspective, *International Journal of Data Science and Analytics*, 2020, vol. ?(?), pp.??? (accepted)
  published version: https://doi.org/10.1007/s41060-020-00213-5 (open access)
- Natalia Andrienko, Gennady Andrienko, Spatio-temporal visual analytics: a vision for 2020s, *Journal of Spatial Information Science*, 2020, number 20, pp.87-95. Published version: https://doi.org/10.5311/JOSIS.2020.20.661 (open access)
- Siming Chen, Jie Li, Gennady Andrienko, Natalia Andrienko, Yun Wang, Phong H. Nguyen, and Cagatay Turkay, Supporting Story Synthesis: Bridging the Gap between Visual Analytics and Storytelling, *IEEE Transactions on Visualization and Computer Graphics,* 2020, vol. 26(7), pp.2499-2516 pre-print, published version: https://doi.org/10.1109/TVCG.2018.2889054
- Jie Li, Siming Chen, Wei Chen, Gennady Andrienko, Natalia Andrienko, Semantics-Space-Time Cube: A Conceptual Framework for Systematic Analysis of Texts in Space and Time, *IEEE Transactions on*

Visualization and Computer Graphics, 2020, vol. 26(4), pp.1789-1806 pre-print, published version: https://doi.org/10.1109/TVCG.2018.2882449

- Natalia Andrienko, Gennady Andrienko, Georg Fuchs, Aidan Slingsby, Cagatay Turkay, Stefan Wrobel, Visual Analytics for Data Scientists, Springer 2020.

3.4.2 Activities planning

Finalizing and Submission of output for microproject: Scientific Migration and Scientific Social Networks.

Partners: PSE, UNIPI

PSE's input will culminate into an economic research paper, a blog post and magazine contribution outlining the main results of the paper and the final dataset used for the relevant analysis. UNIPI's input will culminate into a data paper and a new cleaned dataset from Microsoft Academic Knowledge Graph.

Workshop/Conference Collaboration: Migration and Big Data - (populism and immigration)

Partners: CNR, UNIPI, PSE

Organization of a workshop that will entail keynote speakers from established and reputable researchers in the fields of migration and big data, probably also focusing on the relationship between immigration and populism, and a set of contributed talks on various new/rising research. Hopefully to be hosted at PSE, depending on travel restrictions.

Datathon Collaboration: Migration and Big Data - (populism and immigration)

Partners: PSE, CNRS

Potentially using Twitter data to engage in an exercise requiring attendees to utilise the data (cleaned) provided to answer certain questions regarding potentially the rise of populism and the relationship between it and immigration. Open for graduate students.

Brain drain phenomenon: evidence using LinkedIn

Partners: UNIPI, CNR

Activating an internship that aims to study the phenomenon of brain drain through demographic evidence starting from Linkedin Data.

Brain gain phenomenon: evidence from airline data

Partners: UNIPI, CNR

Continuing to lay the foundations to open a collaboration with JRC, in order to study the brain gain phenomenon from airline data.

3.5 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.

3.5.1 Activities report

3.5.1.1 SOCCER VIDEO STREAMS ANALYSIS VIA DEEP LEARNING

Partners involved: Unipi, ISTI-CNR

We developed PassNet, a method to recognize passes from video streams. The model combines a set of artificial neural networks that perform feature extraction from video streams, object detection to identify the positions of the ball and the players, and classification of frame sequences as passes or not passes. We test PassNet on different scenarios, depending on the similarity of conditions to the match used for training. Our results show good classification results and significant improvement in the accuracy of pass detection with respect to baseline classifiers, even when the match's video conditions of the test and training sets are considerably different.

3.5.1.2 VISUAL ANALYTICS OF SOCCER TRACKING DATA

Partners involved: FHR

We developed an approach that includes a combination of query techniques for flexible selection of episodes of situation development, a method for dynamic aggregation of data from selected groups of episodes, and a data structure for representing the aggregates that enables their exploration and use in further analysis. The aggregation, which is meant to abstract general movement patterns, involves construction of new time-homomorphic reference systems owing to iterative application of aggregation operators to a sequence of data selections. We tested our approach on tracking data from two Bundesliga games of the 2018/2019 season to detect meaningful general patterns of team behaviors in three classes of situations defined by football experts. The experts found the approach worth implementing in tools for football analysts.

3.5.1.3 EXPLAINING THE DIFFERENCES BETWEEN MEN'S FOOTBALL AND WOMEN'S FOOTBALL

Partners involved: ISTI-CNR, UNIPI

Women's football is gaining supporters and practitioners worldwide, raising questions about what the differences are with men's football. We analyzed the spatio-temporal events during matches in the last World Cups to compare male and female teams based on their technical performance. We train an AI model to recognize if a team is male or female based on variables that describe a match's playing intensity, accuracy, and performance quality. Our model accurately distinguishes between men's and women's football, revealing

crucial technical differences, which we investigate through the extraction of explanations from the classifier's decisions. The differences between men's and women's football are rooted in play accuracy, the recovery time of ball possession, and the players' performance quality. We submitted a paper at PLoS One (it is still under revision).

3.5.1.4 HEART RATE VARIABILITY VIA WRIST-WORN WEARABLE DEVICES

Partners involved: UNIPI

We worked on an innovative approach to estimate QRS complexes recorded over 24 h (SDNN24) only exploiting the Heart Rate (HR) that is normally available on wearable fitness trackers and less affected by data noise. The standard deviation of inter-beats intervals (SDNN24) is one most used HR variability feature describing cardiovascular health. We found that HR measures contain enough information to estimate SDNN24. Hence, semi-continuous measures of HR throughout 24 h, as measured by most wrist-worn fitness wearable devices, should be sufficient to estimate the individuals' cardiovascular risk.

3.5.1.5 RELATIONSHIP BETWEEN BODY MORPHOLOGY AND PERFORMANCE

Partners involved: UNIPI, ISTI-CNR

We analyzed the relationship between regional- and whole-body morphology and vertical jump performance and compared the morphological features outlining high and low performers in professional soccer players. Morphology assessment is a valid tool involved not only to monitor training process but also sports performance. However, the difference in jump performance outlines as many differences in the regional body morphology between high and low performers. Of note, the high performers presented significantly greater dimensional characteristics in the upper and lower limbs compared with low performers. This suggests that large body segments, reflecting a certain muscularity, play a role in the force development during a vertical jump.

3.5.1.6 SPECIAL ISSUE: FACTORS AFFECTING PERFORMANCE AND RECOVERY IN TEAM SPORTS: A MULTIDIMENSIONAL PERSPECTIVE

Partners involved: UNIPI

This special issue on Frontiers entitled "Factors affecting performance and recovery in team sport: a multidimensional perspective" is focused on collecting papers about team sport performance. In particular, the aim of this special issue is to advance knowledge of the factors affecting sport performance and recovery emphasizing the use of novel strategies to alleviate potential carryover effects of fatigue.

3.5.1.7 EVENT: INCONTRA INFORMATICA

Partners involved: UNIPI, ISTI-CNR

During the event "Incontra l'Informatica" organized by the Department of Computer Science of the University of Pisa on April 16th, 2021, high-school students take part in a laboratory focused on soccer data analytics.

We introduced the use of Python language programming for sports analytics and showed how to explore a public dataset of soccer-logs, describing the events that occur during a match and are collected through proprietary tagging software.

3.5.1.8 SMART DATA COLLECTION TOOLS FOR SOCCER

Partners involved: UNIPI, ISTI-CNR

The lack of open data is the main limit for research topics related to sports data science. We started the development of data collection tools to let researchers and practitioners easily collect and share soccer data. PySoccer is an open python library to unify soccer data and algorithms. SocceLogger is an open-source web application to collect soccer data through a gamepad, allowing for fast data collection. SoccerLogger aims to enable data collection for youth teams and competitions where data tools from professional providers are not affordable. The long-term target for SoccerLogger is the integration with EventNet.

3.5.1.9 MICRO-PROJECTS

- Performance and recovery in team sports: a multidimensional perspective
  - Status: active
  - Partners: UNIPI
  - External partners: University of Milan, University of Insubria, University of Essex, Hartpury University
  - Expected output: A serie of papers, blog post
- SoccerLogger: open-source data collection tool
  - Status: active
  - Partners: ISTI-CNR, UNIPI
  - Expected output: open-source library, paper
- Difference between men's and women's in soccer
  - Status: completed
  - Partners: UNIPI, ISTI-CNR
  - External partners: none
  - Outputs
    - Method https://data.d4science.org/ctlg/ResourceCatalogue/explaining_the_difference_between_men_and_women_football
    - Blog post http://www.sobigdata.eu/blog/intensity-vs-accuracy-technical-tactical-differences-between-male-and-female-football-teams
- SDNN24 estimation from semi-continuous HR measures
  - Status: completed
  - Partners: UNIPI
  - External partners: Huma therapeutics LTD group, Oxford University
  - Outputs

        - Method https://data.d4science.org/ctlg/ResourceCatalogue/sdnn24_estimation_from_semi-continuous_hr_measures
        - Blog post http://www.sobigdata.eu/blog/wrist-worn-fitness-wearable-devices-accurately-estimate-sdnn24
        - Scientific publication https://www.mdpi.com/1424-8220/21/4/1463
- Relationship between body morphology and performance
    - Status: completed
    - Partners: UNIPI
    - External partners: University of Milan, Parma Calcio 1913
    - Outputs
        - Method https://data.d4science.org/ctlg/ResourceCatalogue/relationship_of_regional_and_whole_body_morphology_to_vertical_jump_in_elite_soccer_players_a_data-
        - Scientific Paper https://www.minervamedica.it/it/riviste/sports-med-physical-fitness/articolo.php?cod=R40Y9999N00A21060106
- PySoccer: python library to unify soccer data and algorithms
    - Status: completed
    - Partners: UNIPI, ISTI-CNR
    - Method: https://data.d4science.org/ctlg/ResourceCatalogue/pysoccer
    - Output: github repository at https://github.com/playerank/pysoccer
- Training your algorithm: soccer matches analysis with Python
    - Status: completed
    - Partners involted: UNIPI, ISTI-CNR
    - Outputs:
    - Blog post http://www.sobigdata.eu/blog/show-students-soccer-analytics-and-they-will-love-computer-science

3.5.1.10 PUBLICATIONS

- D. Sorano, F. Carrara, P. Cintia, F. Falchi, L. Pappalardo, Automatic Pass Annotation from Soccer Video Streams Based on Object Detection and LSTM, Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track. ECML PKDD 2020. Lecture Notes in Computer Science: https://doi.org/10.1007/978-3-030-67670-4_29
- Gennady Andrienko, Natalia Andrienko, Gabriel Anzer, Pascal Bauer, Guido Budziak, Georg Fuchs, Dirk Hecker, Hendrik Weber, and Stefan Wrobel, Constructing Spaces and Times for Tactical Analysis in Football, *IEEE Transactions on Visualization and Computer Graphics*, 2021, vol. 27(4), pp.2280-2297
  pre-print, published version: https://doi.org/10.1109/TVCG.2019.2952129

- L. Pappalardo, A. Rossi, G. Pontillo, M. Natilli, P. Cintia, Explaining the difference between men's and women's football, https://arxiv.org/abs/2101.01662, 2021
- D. Morelli, A. Rossi, L. Bartoloni, M. Cairo, D. A. Clifton, SDNN24 Estimation from Semi-Continuous HR Measures, Sensors 21:1463, 2021. https://doi.org/10.3390/s21041463
- P Cintia, G Mauro, L Pappalardo, P Ferragina, An interactive dashboard for searching and comparing soccer performance scores, arXiv preprint arXiv:2105.04293

3.5.2 Activities planning

Soccer Video Stream analysis through Deep Learning.

Partners involved: UNIPI, ISTI-CNR

Starting from the promising results of PassNet, we plan to further extend it to recognize other types of events such as goals, punishments, shots, fouls, duels, offsides, and more. An important aspect will be the recognition of the players of the two teams, which can be done using Yolov3 or its faster and more accurate version Yolov4, and the tracking of the trajectories of the players on the field.

Physical activity levels and perceived changes in the context of intra-EEA migration.

Partners involved: UNIPI, ISTI-CNR

A study on Italian immigrants in Norway - As mobility within the European Economic Area (EEA) is on the rise, it is important to understand migrants' health-related behaviours (such as the physical activity) within this context. This study investigated the physical activity profile of Italian immigrants in Norway as compared with the general Norwegian population, and the extent to which it varies in relation to key sociodemographic characteristics. This paper is under review at Frontiers in Public Health. Moreover, a blog post and an experiment will be provided on SoBigData++ website.

Physiological recovery among workers in long-distance sleddog race.

Partners involved: UNIPI

A case study on female veterinarians in Finnmarksløpet - During Finnmarksløpet (FL, one of the longest distances 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 for a month after (for a month) the end of the race. The paper is under review at Work. Moreover, a blog post and an experiment will be provided on SoBigData++ website.

Injury prediction project.

Partners involved: UNIPI, ISTI-CNR

Depending on data availability, we plan to extend our injury forecasting model in collaboration with sport science researchers and soccer clubs. We plan to start some preliminary test as soon as the sports season start in autumn, and to get a first version of a performance analysis method by the end of 2021.

Relationship between wellness and training workload in elite soccer players.

Partners involved: UNIPI, ISTI-CNR

We planned to assess the relationship between wellness status and the training workload in elite soccer players. In particular, we would like to develop a framework of big data analytics that permits us to predict the recovery and wellness status of the athletes in the following days in relation to the training workloads performed as the season goes by. This tool could help athletic trainers and coaches to better schedule the training maximizing the training effect. This topic will produce a scientific paper and a blog post on the SoBigData++ website.

3.6 T10.6 Social Impacts 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.

3.6.1 Activities Report

3.6.1.1 OPENING THE BLACK BOX

Partners involved: UNIPI, ISTI-CNR, Tartu

1. A first aspect analyzed within the project was to understand which is the current state of the art in XAI. We reviewed the literature and realized a survey which presents and classifies the explanation methods with respect to the type of explanations returned. It collects explanation methods developed from 2018 to today and compares them by proposing the first benchmarking of explainers in the literature.
2. From the experience of the surveys, we developed and designed various explanation methods with a focus on local rule-based explainers. Local means that the reason for the classification on a specific instance is returned and not the overall logic employed by the black box model. In particular, we developed LORE, a model-agnostic local explanation method that returns as an explanation a “factual” rule revealing the reasons for a decision, and a set of counterfactual rules illustrating how to change the classification outcome. LORE was defined for tabular data and for multiclass classification. We are currently working on solving some limitations of LORE related to the stability and actionability of the explanations.
3. We are working on an extension of LORE for the usage for classifiers working on other data types: images, time series, and text. The generalization to other data types is realized through the usage of autoencoders that allows for moving the computation from complex data structure to simpler ones.

The methods return explanations formed by exemplars and counter-exemplars, i.e., instances similar to the one explained having the same class assigned form the black box, and instances similar to the one explained having a different class assigned form the black box, respectively.

4. Having the possibility to access global explanations describing the whole logic of a model can be more important than referring only to single local explanations. However, sometimes it can be hard to derive global explanations while it can be easy to extract local ones. We proposed GLOCALX, an algorithm that, starting from local rule-based explanations, is able to derive a faithful global model by merging rules with respect to data coverage and logic reasoning.
5. We also worked on the application of explanation methods on classification of cancer type. Our current study applied machine learning algorithms to classify cancers primary and metastatic cancers based on DNA methylation data. Overall, our analysis resulted in 99% accuracy for predicting cancer subtypes based on the tissue of origin. We also used Local Interpretable Model-agnostic Explanations (LIME) for better understanding the classifier's interpretation. We demonstrated that the machine learning models hold a greater promise in diagnosing cancer types based on the tissue of origin in a robust and accurate manner.

3.6.1.2 RELATIONSHIP BETWEEN EXPLAINABILITY AND PRIVACY

Partners involved: ISTI-CNR, UNIPI, SSSA

We proposed EXPERT, a new framework for the prediction and explanation of privacy risk on mobility data. We empirically evaluated privacy risk on real data, simulating a privacy attack with a state-of-the-art privacy risk assessment framework. We then extracted individual profiles describing the users behavior from the data for predicting their risk. We compared the performance of several machine learning algorithms in order to identify the best approach for our task. Finally, we showed how it is possible to explain privacy risk prediction on real data, using two algorithms: SHAP, a feature importance-based method and LORE, a rule-based method.

3.6.1.3 CAUSALITY ANALYSIS OF DATA

Partners involved: UNIPI, ISTI-CNR

We made research on causal discovery with the final goal of having a more accurate data generation procedure that accounts also for detected causal relationships. Indeed, synthetic data generation has been widely adopted in software testing, data privacy, imbalanced learning, machine learning explanation, etc. In all such contexts, it is important to generate plausible data samples. Our goal is to design a synthetic dataset generator for tabular data that is able to discover the nonlinear casualties among the variables and use them at generation time. The problem is that state-of-the-art methods for nonlinear causal discovery are typically inefficient. We boosted one of them by restricting the causal discovery among the features appearing in the frequent patterns efficiently retrieved by a pattern mining algorithm. Hence, we first designed an efficient approach for nonlinear causal discovery based on pattern mining. After that we implemented a generative method based on the boosted nonlinear causal discovery approach. We validated our proposal by developing a framework for generating synthetic datasets with known causalities.

3.6.1.4 ALGORITHMIC FAIRNESS

Partners involved: UNIPI

We are working on FairLens, a methodology for discovering and explaining biases. FairLens is an auditing tool that allows testing a clinical DSS before its deployment, i.e., before handing it to final decision-makers such as physicians and nurses. In this scenario, the healthcare facility IT expert can use FairLens on their historical data to discover the biases of the model before incorporating it into the clinical decision flow. FairLens first stratifies the available patient data according to demographic attributes such as age, ethnicity, gender and healthcare insurance; it then assesses the model performance on such groups highlighting the most common misclassifications. Finally, FairLens allows the expert to examine one misclassification of interest by explaining which elements of the affected patients' clinical history drive the model error in the problematic group. FairLens can become a powerful tool to assess if the model is appropriate for the specific hospital's reference population. Indeed, FairLens allows the human expert to perform a thorough analysis of potential fairness issues. However, the final decision on whether the signaled bias constitutes a real problem or it is a justified basis for differentiation is left to the human auditor. We validate FairLens' ability to highlight bias in multilabel clinical DSSs introducing a multilabel-appropriate metric of disparity and proving its efficacy against other standard metrics.

3.6.1.5 FINANCE & ECONOMICS

Partners involved: SNS, ETH, TARTU

We have made the following advancements:

- Built a tool to acquired data from Twitter in real time that is related to financial news
- Processed market microstructure data related to liquidity during the bitcoin bubble in 2018. Preliminary results show that empirical phase transition took place near the price crash, however it is only observable in order book-related measures.
- Set up a framework for and analysed information dynamics in cryptocurrency markets and performed empirical analysis of the bitcoin bubble in 2018.
- Especially related to explainable AI, we have made an important, although minor contribution related to choosing the optimal parameters for inference of k-nearest neighbours that is paramount to calculation of information theoretic measures. Oftentimes, this choice is done arbitrarily therefore it is not clear whether the information transfer is computed reliably and reflects the true amount of information transferred within the system.
- We began writing a preprint for the analysis of market microstructure during the 2018 bubble.
- We developed a method based on Deep Neural Networks to study the potential chaotic behavior of leverage of financial institutions (banks) from short time series.
- We explored the historical financial transactions for predicting the amount a customer will receive through his/her transacting partners at a specific time. In particular, we use the Bitcoin transactional dataset, which has two main characteristics: i) network, and ii) temporal. We contribute by exploiting a specific kind of Graph Neural Network approach called Temporal-Graph Convolutional Network (T-GCN) for predicting the number of Bitcoins received by a customer at a particular timestamp. The

lower errors obtained using T-GCN approach compared to 11 baseline approaches (such as Support Vector Regression (SVR), Random Forest Regression (RFR), Vector Auto-Regressive (VAR), Long Short-Term Memory (LSTM), etc.) clearly demonstrate the effectiveness of T-GCN approach.

3.6.1.6 MICRO-PROJECTS:

- Visualizing the Results of Boolean Matrix Factorizations
  - Status: active
  - Partners: KTH, LUH
  - External partners: Ecole Normale Superieure de Lyon
  - Expected Output: Method, Experiments, Blog post, Preprint paper
- XAI Method for explaining time-series
  - Status: completed
  - Partners: ISTI-CNR, UNIPI
  - External partners: none
  - Output
  - Method https://data.d4science.org/ctlg/ResourceCatalogue/xai_method_for_explaining_time-series
  - Blog post http://www.sobigdata.eu/blog/understanding-any-time-series-classifier-subsequence-based-explainer
- Local to Global Method
  - Partners: ISTI-CNR, UNIPI
  - External partners: none
  - Output
    - Method
    - https://sobigdata.d4science.org/group/sobigdata-gateway/data-catalogue?path=/dataset&query=cT1nbG9jYWx4
    - Scientific article
    - https://www.sciencedirect.com/science/article/pii/S0004370221000084
    - Blog post
    - http://www.sobigdata.eu/blog/glocalx

3.6.1.7 PUBLICATIONS

- 2020 Artificial Intelligence Evaluating local explanation methods on ground truthRiccardo Guidotti (university of Pisa)
- 2020 AAAI 2020 Conference Explaining Image Classifiers Generating Exemplars and Counter-Exemplars from Latent Representations. Riccardo Guidotti (University of Pisa), Anna Monreale (University of Pisa), Stan Matwin (Dalhousie University), Dino Pedreschi (University of Pisa)
- 2020 ICDM 2020: 20th IEEE International Conference on Data Mining Data-Agnostic Local Neighborhood Generation Riccardo Guidotti (University of Pisa), Anna Monreale (university of Pisa)

- 2020 International Conference on Discovery Science Explaining Sentiment Classification with Synthetic Exemplars and Counter-Exemplars Orestis Lampridis (University of Tessalonikki), Riccardo Guidotti (University of Pisa), Salvatore Ruggieri (University of Pisa)
- 2021 Artificial Intelligence GLocalX-From Local to Global Explanations of Black Box AI Models Mattia Setzu (University of Pisa), Riccardo Guidotti (University of Pisa), Anna Monreale (University of Pisa), Franco Turini (University of Pisa), Dino Pedreschi (University of Pisa), Fosca Giannotti (ISTI-CNR)
- 2020 2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI) Explaining Any Time Series Classifier Riccardo Guidotti (University of Pisa), Anna Monreale (University of Pisa, Francesco Spinnato (University of Pisa), Dino Pedreschi (University of Pisa), Fosca Giannotti (ISTI-CNR, fosca.giannotti@isti.cnr.it)
- 2020 2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI) Interpretable Next Basket Prediction Boosted with Representative Recipes Riccardo Guidotti (University of Pisa, Stefano Viotto (University of Pisa)
- 2020 Discovery Science. DS 2020. Lecture Notes in Computer Science, vol 12323 Predicting and Explaining Privacy Risk Exposure in Mobility Data, Francesca Naretto, UNIPI; Roberto Pellungrini, UNIPI; Anna Monreale, UNIPI; Franco Maria Nardini, CNR; Mirco Musolesi, University College London;
- 2020 XKDD 2020: 2nd International Workshop on eXplainable Knowledge Discovery in Data Mining Prediction and Explanation of Privacy Risk onMobility Data with Neural Networks, Francesca Naretto, UNIPI; Roberto Pellungrini UNIPI; Franco Maria Nardini, CNR; Fosca Giannotti, CNR;
- 2020 Lampridis, O., Guidotti, R., & Ruggieri, S. (2020, October). Explaining Sentiment Classification with Synthetic Exemplars and Counter-Exemplars. In International Conference on Discovery Science (pp. 357-373). Springer, Cham.
- 2021 Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., & Rinzivillo, S. (2021). Benchmarking and Survey of Explanation Methods for Black Box Models. arXiv preprint arXiv:2102.13076.
- Fabrizio Lillo, Giulia Livieri, Stefano Marmi, Anton Solomko, Sandro Vaienti, Analysis of bank leverage via dynamical systems and deep neural networks, Arxiv https://arxiv.org/abs/2104.04960

3.6.1.8 EVENTS

- Workshop XKDD 2020 ECML-PKDD 2020 (Virtual) 14/09/2020: The purpose of XKDD, eXaplaining 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.
- Tutorial XSDM 2020 DSAA (Virtual) 09/10/2020: The purpose of the XDSM Tutorial is to illustrate state-of-the-art approaches for explain- able data mining and interpretable machine learning, which are the problems, issues and current challenges, and to encourage principled research that will lead to the advancement of explainable, transparent, ethical and fair data mining and machine learning.
- Tutorial XAI 2021 AAAI-2021 (Virtual) 03/02/2021: XAI (eXplainable AI) aims at addressing such challenges by combining the best of symbolic AI and traditional Machine Learning. Such topic has been studied for years by all different communities of AI, with different definitions, evaluation metrics, motivations and results. This tutorial is 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 (given the halfday format).

3.6.2 Activities planning

3.6.2.1 OPENING THE BLACK BOX

Partners involved: UNIPI, ISTI-CNR

We plan to set-up an experiment to understand explanations preferences from users' point of view. We propose to study how the explanation affects the user's behavioral intention of using a predictive XAI system in a specific medical context to make AI research more and more human-centered through a bottom-up approach. In particular, we propose to study the perceived usability and the intention of using the system according to user preferences and how the explanation of the AI "black-box" could improve user interaction. This research is a first step aimed at posing the basis for a multidisciplinary approach (HCI, Cognitive sciences, medical science, Data science, …) to enhance the research through ethnographic methods in order to maintain the human-in-the-loop in a triadic approach involving the user, the automation in a specific decisional context. The final purpose of our research is twofold. Firstly, we aim to understand how explanations could enhance the intention of using an AI system in the medical field. Secondly, we aim at giving suggestions and guidelines to the designers and researchers of such complex systems to increase the acceptance and the final decision performance.

3.6.2.2 RELATIONSHIP BETWEEN EXPLAINABILITY AND PRIVACY

Partners involved: ISTI-CNR, UNIPI, SSSA

We plan to: publish a journal paper applying EXPERT on different data mobility, purchasing, textual data and evaluating the ability of explainability methods to effectively describe the reason of the risk using as groundthurth the information that we can derive from the empirical computation of the privacy risk performed by PRUDENCE; upload EXPERT method in Sobigdata and activate a micro-project on it; work on the privacy issues of the Copy Framework introduced in [UP2020] having the scope to create a copy of a classifier in order to get a model that is accessible and that maintains some properties such as transparency and fairness.

3.6.2.3 CAUSALITY ANALYSIS OF DATA

Partners involved: UNIPI, ISTI-CNR

We plan to submit at least two papers (conference + journal) by the end of the year. The associated method will be released on the SoBigData++ platform.

3.6.2.4 ALGORITHMIC FAIRNESS

Partners involved: UNIPI, ISTI-CNR

We plan to activate a micro-project FairLens and to submit a journal paper by the end of the year. We also plan to activate a micro-project for the definition of a case study for the understanding of human interaction with automatic systems designed to assist users during a decision-making process. In particular, the experiment aims at studying how individuals change or adapt their behaviour depending on different characteristics of the automatic system. The empirical study wants to determine how (i) different levels of accuracy, (ii) the presence of absence of transparency/explainability mechanisms, (iii) the presence of bias, (iv) the level of automation and (v) the robustness of the recommendations may affect the human perception and/or behaviour with respect to the system predictions. The idea is to frame the experiment in the context of an online web-based game where the user needs to balance exploration/exploitation to maximize the obtained score. There will be a decision support system (DSS) assisting users during the game.

3.6.2.5 DECENTRALISED AI & HUMAN-CENTRIC AI

Partners involved: ISTI-CNR, IIT

We plan to carry out research activities on decentralised and human-centric AI under the recently kicked-off CHIST-ERA project SAI, dedicated to studying decentralized, social explainable AI. The vision of SAI is towards a decentralised "collective" of local machine-learning-based AI components interpreting data and interacting according to human-centric design principles, where explainability is guaranteed both at the local and collective level. The project revolves around the concept of PAIVs, the "Personal AI Valet" (PAIV) that acts as the individual's proxy in a complex ecosystem of interacting PAIVs. PAIVs process individuals' data via explainable AI models tailored to the specific characteristics of their human twins, and interact with each other, to build, in a decentralised way, global AI models and/or come up with collective decisions starting from the local models. The scientific activities of the project have started in April 2021, hence there are no results to report yet. We plan to submit at least two papers (conference + journal) by the end of the year. The associated code will be released on the SoBigData++ platform. Both activities will be part of a micro-project that we will submit in the second half of the year.

3.6.2.6 FINANCE & ECONOMICS

Partners involved: ETH, SNS

In the following year, we plan to: finish writing a preprint (and submitting) on information dynamics in 2018 bitcoin bubble; Upload relevant data to SoBigData; Develop methods for inference of complex financial time series using Recurrent Neural Networks.

3.6.3 Planned Events

Workshop & Tutorial XKDD 2021 ECML-PKDD 2021 13/09/2021

4 Conclusions

As witnessed by the vast number of activities carried out and the variety of topics investigated, and despite the inconveniences related to the COVID-19, the first year and a half of WP10 has been very productive. At this point, all exploratories are established, and several collaborations are ongoing between the partners of the consortium. The introduction of micro-project at the beginning of year two helped significantly organize 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 2021 and in 2022, in which the transnational access program should be reactivated (depending on the pandemic), allowing for strengthen the collaborations among the partners of the consortium as well as the collaborations with external institutions, organizations, companies, and researchers.