Social Mining & Big Data Analytics

SoBigData

RESEARCH INFRASTRUCTURE ++

Deliverable D10.1

Initial Exploratory activities planning

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 (48 months)
ENDING DATE	31/12/2023
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 - 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	D10.1
DELIVERABLE TITLE	Initial Exploratory activities planning
AUTHOR(S)	Luca Pappalardo (CNR), Aristides Gionis (KTH)
CONTRIBUTOR(S)	Anna Monreale (UNIPI), Angelo Facchini (IMT), Tiziano Squartini (IMT), Paolo Cintia (UNIPI), Kalina Bontcheva (USFD), Ye Jiang (USFD), Aris Anagnostopoulos (UNIROMA1) Laura Pollacci (UNIPI)
EDITOR(S)	Beatrice Rapisarda (CNR)
REVIEWER(S)	Jurek Leonhardt (LUH), Beatrice Rapisarda (CNR)
CONTRACTUAL DELIVERY DATE	30/06/2020
ACTUAL DELIVERY DATE	09/10/2020
VERSION	1.1
TYPE	Report
DISSEMINATION LEVEL	Public
TOTAL N. PAGES	26
KEYWORDS	Exploratories, Planning,

EXECUTIVE SUMMARY

This deliverable provides information about the topics, activities, and dissemination initiatives planned for WP10 - Exploratories.

Section 1 describes the relevance of this document to SoBigData++, highlighting relationships with the other work packages.

Section 2 introduces the activities WP10 has developed to foster the communication between partners within the exploratories and to update the general public about the results achieved by the exploratories.

Section 3 describes, exploratory by exploratory, the topics planned and the activities planned for the coming year.

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
GDPR	General Data Protection Regulation
AI	Artificial Intelligence
ML	Machine Learning
SL	Scientific Leader
UCA	User Community Activist

TABLE OF CONTENTS

1 Relevance to SoBigData++

This document describes the initial plan for the development of the exploratories. In particular, we describe the scientific topics and the activities (conferences/workshops, hackathons, data collection, software development) planned by each exploratory for the coming year. The topics and activities described in this document are relevant to milestone MS1 (Initial Planning of Activities completed), 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 planned for next year, this deliverable is also related to work packages WP3 - Dissemination, Impact, and Sustainability (because of workshops and conferences are planned), WP4 - Training (because hackathons are planned), and WP7 - Virtual Access (because data sets and software will be made available on the infrastructure).

1.1 Structure of the document

In Section 2, we describe the initiatives we planned to foster inter-exploratories communication and collaboration and to update both the general public and the people within the consortium on the results achieved by the exploratories. In particular, in Section 2.2 we describe how we will organize the monthly webinars and in Section 2.3 how we will collect and publish the monthly blog posts. Sections 3.1-3.7 describe, exploratory by exploratory, the scientific topics and the activities planned for the coming year.

2 Exploratories webinars and blog posts

In WP10, each exploratory has been associated with two key figures: (i) a Scientific Leader (SL), a domain expert who have the responsibility of fostering the integration of research activities and supervising their scientific soundness and validity; and (ii) one or more User Community Activists (UCAs), researchers of the domain with the tasks of sharing the information among the partners, to collaborate with WP3, WP4 and WP5 and the other exploratories. The UCAs will help in organizing events and disseminating the results and to foster interests among the e-infrastructure users through social network activities (e.g., posts, users' interactions).

2.1 Exploratory webinars

To raise interest among the general public, people from the other WPs, and people from all the exploratories, about the topics and the results within each exploratory, in the next year we will organize monthly webinars. Each webinar will be dedicated to one exploratory and will cover topics of interest to that exploratory. The SI and UCAs of the exploratory will be the webinar's moderator and the organizers, respectively, and prominent scholars from both within and outside the project will be invited to speak. We already organized (or are organizing) some webinars in 2020, which had a wide participation from both the general public and people in SoBigData++.

2.2 Exploratory Blog Posts

To update the general public and the people within SoBigData++ about the results developed within the exploratories, in the next year, we will publish monthly blog posts. Each exploratory will produce at least one post per month, in which it will describe a recent result of the exploratory, or a topic related to it. The UCAs will be responsible for collecting and revising the blog posts from the participants of the corresponding exploratory before the publication on the SoBigData++ blog (http://www.sobigdata.eu/blog).

3 Initial Exploratory Activities Planning

The research in WP10 is structured in vertical thematic environments, called *exploratories*, aimed at creating new datasets and services to be integrated within the SoBigData++ research infrastructure. In this section, we describe the scientific topics each exploratory has planned to investigate, and the activities (e.g., conferences, data collection, software development) planned for the next year.

3.1 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 starting point will be the identification of key themes and points of view debates. Thus, the discussion on this topic will be analysed and visualised. Moreover, there will be an assessment of their evolution through time and space (i.e., in different countries or regions). The central focus will regard misinformation, a field where we will develop 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 will also study the effect of different social relationships when it comes to opinion formation. A multi-disciplinary approach will be adopted, going beyond computer science to integrate also social and political scientists, as well as end-users and practitioners, such as the Centre for Study of Democracy (CSD), which will focus specifically on Russian propaganda and misinformation in Eastern and Central Europe. The results of this exploratory will be the development of new tools for the infrastructure, thus empowering researchers from outside the consortium to work on these topics.

3.1.1 Planned Topics

Twitter misinformation analysis. Online social media platforms, such as Twitter, play an essential role in communicating information during times of crisis and are often used to timely disseminate informative content to users. Such convenience also raises concerns about the quality of the information accessed by users and also how users interact with each other. As COVID-19 infections surged globally, the virus has also led to an "infodemic", with large volumes of misinformation spread rapidly on Twitter, and has already caused public mistrust and even real-life damage to health and 5G masts. In response to this infodemic, automatic methods are urgently needed to refute massive misinformation. We plan to investigate different attributes (e.g., user IDs, the use of authority references) of misinformation tweets and debunking tweets correspondingly. Specifically, misinformation is a piece of text that contains false information regarding its subject, and resulting in a rare co-occurrence of the subject and its neighbouring words in a truthful corpus. A language model trained on large-scale data can utilize such characteristics (e.g., GPT-2, BERT) to automatically predict the perplexity of a given tweet. The misinformation tweets have high perplexity when scored by a truth-grounded language model, and the debunking tweets are the opposite since perplexity is a score for quantifying the likelihood of a given sentence based on previously encountered distribution. We plan to leverage the sizeable pre-trained language models to learn the sentence distributions from debunking tweets and predict the perplexity of misinformation tweets. By exploring and selecting perplexity threshold, the misinformation tweets can be identified when their perplexity is higher than the threshold.

Fact-checking. Fact-checkers and media worldwide have united under the International Fact-Checking Network (IFCN) to counter misinformation regarding the pandemic collaboratively. However, COVID-19 fact-checking is a fast-moving research area. For example, research has found that most misinformation in the early stage of the pandemic made false claims related to actions and statements by public authorities. With the recurrence of the epidemic, types of misinformation might potentially be shifted to other topics. The ongoing work is collecting COVID-19 related debunks of misinformation published on the IFCN Poynter website, we plan to soft labelling that misinformation based on a pre-trained misinformation classifier and analysing the types of misinformation changing through time.

Testing effects of algorithm bias. Human-produced corpus often reflects the pre-existing ideologies of the annotators, which also embeds bias in the creation of algorithms. For example, the pre-trained language model ELMo has recently been accused of demonstrating gender bias because the training data for ELMo contains significantly more male than female entities. Consequently, the algorithms systematically and unequally encode gender information that male entities can be predicted from occupation words more accurately than female entities. Corpus for detecting online misinformation typically contain sensitive groups that have imbalanced classes. For example, the recently established COVID-19 disinformation dataset has imbalanced categories (e.g., 19.4% on Public Authority and 4.6% on Public Reaction). In response to this, algorithms can be established with data-dependent constraints to achieve fairness goals to improve its generalization. We plan to summarize the main characteristics of the data, such as class imbalance, and impose fairness constraints that prevent certain protected groups (e.g., small groups with a large fraction of samples labelled as positive) from unfair algorithms. With data-dependent constraints, the bias of algorithms can be potentially quantifiable.

Analysis of YouTube videos. Misinformation about COVID-19 is reaching more individuals than in past public health crises, as YouTube continues to grow as a source of health information. Previous studies have found that YouTube has been a source of useful and misleading information during public health crises, including the H1N1 pandemic, Ebola outbreak and Zika outbreak. However, the generalisation of previous findings to the current COVID-19 pandemic is limited. We plan to collect YouTube videos related to COVID-19 and analyse the source of videos, factual and non-factual information in the videos. We will assess different attributes of the most-watched and liked videos related to COVID-19. For example, we plan to compare videos containing only factual information between types of sources, and also to compare the significant differences in different attributes, such as views per video, likes per video, dislikes per video, duration and days since publication.

3.1.2 Planned Activities

Data collection. We plan to collect data about COVID-19 misinformation from Twitter and analyse the factual/non-factual information, types of source, user IDs and meta information. We will explore different pre-trained language models and perplexity thresholds, for predicting the probability of misinformation tweets. We also plan to collect debunks of misinformation from IFCN Poynter website, and analyse the types of misinformation changing through time. We will label the types of misinformation and its debunks and improve model accuracy by increasing the number of training samples. We plan to collect YouTube video data and compare videos containing only factual information and non-factual information between types of sources.

Software development. We will start the design and implementation of a Twitter misinformation debunker web application, including a fine-tuned, truth-grounded language model. We will combine the LIME machine learning explainer with a pre-trained misinformation classifier for analysing the bias of algorithms.

Events. We are putting together a "how-to" guide on the study of misinformation across a variety of social media platforms:

Mainstreaming the fringe: How misinformation propagates on social media, edited by Richard Rogers

1. Introduction: Mainstreaming the fringe, or how misinformation propagates on social media – Richard Rogers
2. The scale of **Facebook**'s problem depends upon how 'fake news' is classified – Richard Rogers
3. Problematic information in **Google Web Search**? Scrutinizing the results from U.S. election-related queries – Guillen Torres
4. The presence of problematic information and users on political **Twitter** in the run-up to the 2020 US elections – Maarten Groen, Sagar Hugar and Boy Singmanee
5. The earnest platform: Coverage of the US presidential candidates, COVID-19, and social issues on **Instagram** – Sabine Niederer and Gabriele Colombo
6. Singing and dancing on **TikTok**: Informing young voters with creative political expression – Shuaishuai Wang, Andrea Benedetti and Carlotta Dotto
7. The spread of political misinformation on the online subcultural platforms **Reddit and 4Chan** – Anthony Burton and Dimitri Koehorst
8. Factual divergences and misinformation in the process of COVID-19 sensemaking: A **multi-platform analysis** – Emillie De Keulenaar, Rory Smith, Carina Albrecht, Ivan Kisjes, Pedro Noel and Jack Wilson

Some of the pieces are being published by the Harvard Misinformation Review. The bundle is under consideration with Amsterdam University Press.

We will organize a sequence of workshops and summer school on misinformation analysis. In particular, 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 Demography, Economy & Finance 2.0

This exploratory aims to combine statistical methods and traditional economic data (typically at low-frequency) with high-frequency data from non-traditional sources (e.g., web, supermarkets), for now-casting economic, socio-economic and well-being indicators. This activity of this exploratory is expected to support studies on the correlation between people well-being and their social and mobility data, aiming at discovering whether they change in less affluent areas.

This exploratory studies also traditional complex socio-economic 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. A particular emphasis will be devoted to the stability and "fairness" properties (for example the absence of manipulations such as wash trades) of cryptocurrency markets. The Swiss partner ETHZ, in collaboration with the other partners, already started collecting data in order to attract the research community.

3.2.1 Planned Topics

Development of novel network tools for the analysis of economic and network financial systems. Over the last 15 years, statistical physics has been a successful framework to model complex networks. This approach has brought novel insights into a variety of physical phenomena, such as self-organisation, scale invariance, the emergence of mixed distributions and ensemble non-equivalence, that display unconventional features on heterogeneous networks. At the same time, thanks to their deep connection with information theory, statistical physics and the principle of maximum entropy have led to the definition of null models for networks reproducing some features of real-world systems, but otherwise as random as possible. We plan to extend the existing approaches grounded on statistical physics for the analysis of real-world, economic and financial networks, e.g. by testing the performance of the recently proposed framework for conditional network reconstruction, by considering non-linear Exponential Random Graph models, by considering network models encoding temporal dependencies.

Analysis of cryptocurrencies. Cryptocurrencies are distributed systems that allow exchanges of native (and non-) tokens among participants. The wide availability of complete historical bookkeeping opens up the unprecedented possibility to understand the evolution of their network structure while gaining useful insight into the relationships between user' behaviour and cryptocurrency pricing in exchange markets. We plan to address the analysis of the structural properties of a set of three different constructs: the Bitcoin Address Network (i.e., a directed, weighted graph whose nodes represent addresses, the direction and the weight of links are provided by the input-output relationships defining the transactions recorded on the blockchain), the Bitcoin User Network (i.e., a network whose nodes are clusters of addresses, understood as bona fide "users" and identifiable by implementing the so-called "heuristics") and the Bitcoin Lightning Network (i.e., a directed, weighted graph constructed in a fashion that is similar to the way the BAN is defined: nodes are addresses exchanging bitcoins on the so-called "Layer 2").

Analysis of the economic policies promoting well-being. Digital data streams are starting to find a place in well-being research, offering many opportunities in the measurement of socio-economic indexes for better economic policy. Traditionally, to capture well-being, researchers and policy-makers collect data through surveys and official governmental sources, which is expensive and time-consuming. Supplementing traditional data, digital data streams, explored by machine learning, are making the estimation of well-being cost-efficient and almost real-time. In this exploratory, we plan to develop an analytical framework that exploits information extracted from GDELT, a digital news database, to estimate the monthly values of well-being. Starting from preliminary results on monitoring well-being in terms of peacefulness through the Global Peace Index (GPI)[1], we will explore how this methodology can be used to estimate any other well-being dimension and socio-economic index related to societal progress.

Data science and machine learning techniques in finance. Data science and machine learning have emerged as a universal tool to study different complex systems, including finance. Recent advancements in ML, in particular, Deep Learning, Reinforcement Learning and Self-Supervised Learning have opened new theoretical and application directions for more accurate predictions, risk quantification, portfolio creation, algorithmic execution, and others. We plan to study theoretical and applied contributions at the interface of machine learning and finance. Here we will study what online social media signals (e.g. tweets related to financial news) can be used to reduce uncertainty about prices of financial assets (i.e. long- and short-term prediction of asset prices and/or their changes). We will concentrate on exotic cryptocurrency markets whose behaviour is still far from fully understood and difficult to predict using traditional means. Using information-theoretic tools, we will also quantify macrodynamics of asset prices as a property that emerges from microscopic interrelation of financial assets, that enables distress propagation through the network.

3.2.2 Planned Activities

Data collection. We plan to collect novel Finance 2.0-related datasets from Twitter. The ability to track and monitor relevant and important news in real-time is of crucial interest in multiple industrial sectors. We plan to focus on cryptocurrency news, which recently became of emerging interest to the general and financial audience. To track popular news in real-time, we will match news from the web with tweets from social media, track their intraday tweet activity, and explore different machine learning models for predicting the number of the article mentions on Twitter after its publication. Finally, we plan to open-source crypto-financial Twitter data.

Software development. We plan to develop and release:

[1] Vasiliki Voukelatou, Ioanna Miliou, Lorenzo Gabrielli, Luca Pappalardo and Fosca Giannotti, Estimating Countries' Peace Index through the Lens of the World's News as Monitored by GDELT, The 7th IEEE International Conference Data Science and Advanced Analytics, 2020.

- a Python package to solve maximum entropy (null) models (https://meh.imtlucca.it/). Moreover, we plan to release software to correctly sample ensembles of networks whose constraints are "soft", i.e., realized as ensemble averages. This method is based on exact maximum-entropy distributions and is therefore unbiased by construction, even for strongly heterogeneous networks. It is also more computationally efficient than most microcanonical alternatives. Finally, it works for both binary and weighted networks with a variety of constraints, including combined degree-strength sequences and full reciprocity structure, for which no alternative method exists.
- packages on centrality measures for temporal networks, fractional (non-local) diffusion on graphs, and statistical models for temporal networks (inference and prediction)

Events. We plan to organize the following events in 2021:

- A new edition of the "Complexity Meets Finance" workshop in 2021. This year we organized it as a satellite event of the NetSci2020 conference (https://sites.google.com/view/cmf20/home);
- An online half-day workshop on the topic of "Data Science and Machine Learning in Finance" (institutions involved could be Scuola Normale Superiore, NYU Courant, Imperial College London);
- A sequence of workshops on data science and machine learning in finance, in order to share insights from different data-intensive disciplines. We will bring together a unique blend of world-class experts from industry and academia for an intensive half-day workshop. ETH Zurich and NYU Courant already organized the original workshop of the sequence. It is also a part of Data Science in Techno-Socio-Economic Systems ETH course and supported by the EU SoBigData++ project. For details (visit https://www.eth-courant-workshop.com/).

3.3 Sustainable Cities for Citizens

This exploratory will focus on narrating stories about cities, the sustainability of their flows of energy and materials and people living in it. Data scientists describe those territories using an industrial ecology perspective driven by data, statistics and models, allowing citizens and local administrators to understand cities better. We will analyze data from different spatial and temporal scales. On city-wide scales, we will analyze 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 will analyze data related to electric mobility services in EU cities, allowing the characterization of the demand of dynamic users and granting the derivation of models to optimize the electric mobility charging and relocation service and minimize its impact on the power grid. Moreover, we will investigate topics related to the quality of life and well-being (climate change, urban green areas).

3.3.1 Planned Topics

Energy and material flows of Italian small municipalities. We will investigate three municipalities with the urban metabolism method integrated with a flow circularity analysis. The municipalities will be under 50.000 inhabitants and located in the north, center and south Italy.

Peer to peer energy trading and blockchain. We plan to investigate the use of blockchain technology in energy systems and energy communities using simulations, surveys, and market case studies. This activity will also leverage Task 10.2 (exploratory Demography, Economy & Finance 2.0) and the experience of the Global Observatory for P2P energy trading (IMT is a member). The effect of energy markets will be investigated with particular attention to the economic sustainability, regulation and presence of arbitrages.

Infrastructural impact of sustainable mobility. We will study the impact of a complete switch to electric mobility, especially regarding the intensity of flows due to charge and fast recharge systems. We will use personal mobility data from different sources to estimate the mobility flow and to simulate the impact of different charging behavioural patterns to predict power flows and to optimize the position of the charging infrastructures.

Personal and Collaborative mobility systems. We will characterize car sharing usage, identify its potential weaknesses and understand the conditions under which it may thrive as a component of an integrated mobility system. To this aim, data on car sharing usage will be integrated with socio-demographic indicators, data from social media, from GPS vehicular data and mobile phone data to build a comprehensive picture of what is happening in the observed territory.

Urban flood prediction. We plan to develop models for urban floods prediction that integrate data provided by CEM system (https://emergency.copernicus.eu/) and Twitter data. Twitter data will be processed using massive multilingual approaches for classification. The model will require careful data collection and validation of ground truth about confirmed floods from multiple sources.

Disaster recovery. We will focus on the analysis of the unique “disaster recovery” data regarding the devastating earthquake that occurred in 2009 in the province of L’Aquila, Italy, and the subsequent reconstruction of the region in the following years. Among the aims of this study, we want to achieve a deeper understanding of the resurrection of the socio-economic structure of a territory after a catastrophic event.

EPICURE - EPIdemiC and the URban Environment. We plan to build a modular and flexible risk assessment tool specifically tailored for urban regions. As a starting point, we plan to consider the main factors that played a role in triggering the severity of the COVID-19 epidemic spread in northern Italy, e.g., environment, demography, and epidemiology. Furthermore, factors involving transportation networks and regional mobility of people will be also considered in the disease transmission potential.

3.3.2 Planned Activities

Optimal planning of regional renewable energy sources: Using the optimal portfolio theory of Markowitz the researchers will investigate optimal production strategies for wind and solar power plants, considering both the amount of energy generated and the interactions with the national power grid and the cities.

Climate change and COVID-19: A specific survey covering 4 to 5 EU countries will be developed in cooperation with the "Climate Media Center Italia" to assess the risk perceived by citizens with respect to

COVID epidemics and risks related to climate change. Using the methods of behavioral science the research team will investigate the analogies and differences between the climate and pandemic risk.

Urban metabolism of one Italian municipality: The urban metabolism of an Italian municipality will be investigated with the aim to complete the first step of the atlas of urban sustainability. Further case studies are expected to be planned during the next year. Future case studies will also include the investigation of non-Italian cities.

Analysis of mobility data in Tuscany: mobility data will be use to simulate a regional switch to electric mobility in Tuscany, covering the area from Pisa to Livorno and Florence.

3.4 Migration Studies

This exploratory aims to study 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 Planned Topics

Now-casting migration stock. The study of international migration is gaining increasing interest due to its profound effects on both countries of origin and destination. High costs of nationally collecting data on migration, inconsistent definitions and measures across worldwide sources, and data releasing lags makes it difficult to obtain and provide up to date global migration scenarios. By leveraging the increasing volume of social Big Data, we plan to use Twitter data as a proxy for migration stocks to predict current migrants' European stocks ahead of official data publication (nowcasting). The work will leverage both the method of Pollacci et al. 2017[2] and the ongoing research on the building of the so-called Superdiversity Index (SI), based on the diversity of the emotional content expressed in texts of different communities. Moving from preliminary results showing that the SI strongly correlates with immigration rates in Italy and the UK (Pollacci 2019)[3], we plan to develop a European nowcasting model together with single-countries nowcasting models for immigration rates. The work will include downloading and processing of geolocalized Twitter data, application of the method and development and testing of nowcasting models based on machine learning. To prevent ethics and privacy issues, the analysis will comply with the Twitter data terms of usage and to the EU GDPR.

[2] Pollacci, Laura, et al. "Sentiment spreading: an epidemic model for lexicon-based sentiment analysis on twitter." *Conference of the Italian Association for Artificial Intelligence*. Springer, Cham, 2017.

[3] Pollacci, Laura. "Superdiversity: (Big) Data analytics at the crossroads of geography, language and emotions.", 2019.

Studying migrant integration with social network data. Starting from the work of Kim et al. 2020[4], we plan to develop two indices to measure home and destination attachment by using Twitter topics as a proxy for user interests, opinions, and access to information. We will analyse a large network of Twitter users to identify migrants and evaluate their integration patterns.

Highly Skilled Migration. The study of highly skilled migration has attracted a growing interest due to its importance to scientific productivity, labour market and its large effects both on the origin and destination communities. By leveraging public Big Data such as the Microsoft Academic Knowledge Graph (MAKG), we plan to quantify the highly skilled migration phenomenon both within the EU and the rest of the world. We will attempt to understand and identify the drivers of migration w.r.t. various groups, e.g., skill sectors, countries. We will also focus on understanding both the role and effects of and on social networks of highly skilled migrants. The work will include downloading and processing of data and building both migrants' personal networks (ego network) and collaboration networks. We plan to publish our analysis and results and, in parallel, to release our dataset publicly.

Immigrants Integration through retail data. In migration studies, the analysis of retail data may allow us to investigate both immigrants' economic integration as well as changes in their habits. Analyze immigrants' food consumption basket though retail data from a supermarket chain allows us to estimate the integration degree and its variation over time. Also, we can identify which are the most relevant factors for integration. Moving from the work in Guidotti et al. 2020[5], we plan to analyze immigrants' food consumption from shopping retail data for understanding if and how it converges towards those of natives, together with reasons for possible changes. We will attempt to develop a machine learning classifier able to recognize natives based on a score of adoption of consumption habits. On this basis, we aim to measure the immigrants' adoption of natives' consumption behaviour over a long time, identifying different trends. Finally, we plan to focus on explaining why the classifier assigns specific scores with respect to the customers' shopping behaviour.

3.4.2 Planned Activities

Data collection. We plan to release a dataset describing highly skilled migration. The dataset will be made publicly available. In addition to information related to highly skilled migrants, the dataset will include information on institutions, publications and international projects.

Special issue on Big data and migration. We are applying for a special issue in the *SAGE Big Data and Migration* journal. If accepted, the papers will be submitted by February 2020. The papers on migrant

[4] Kim, J., Sîrbu, A., Giannotti, F. and Gabrielli, L., 2020, April. Digital Footprints of International Migration on Twitter. In International Symposium on Intelligent Data Analysis (pp. 274-286). Springer, Cham

[5] Guidotti, Nanni, Giannotti, Pedreschi, Bertoli, Speciale, Rapoport, "Measuring Immigrants Adoption of NativesShopping Consumption with Machine Learning", European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), Belgium, 2020.

integration and superdiversity resulting from the research topic above will be published within the special issue.

3.5 Sports Data Science

This exploratory will provide massive heterogeneous dynamic data describing several sports – especially soccer, cycling and rugby – to construct an interpretable, explainable 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 Planned Topics

Soccer Video Streams analysis via Deep Learning. In recent years, several companies have been involved in sports analytics through the collection of spatio-temporal data describing the events that take place during official matches. These data are collected manually by human operators, constituting a considerable cost in terms of time and economic resources. Automatic recognition of significant events from sports video streams is still an open problem, and it is attracting the focus of research communities related to computer vision and deep learning. Moving from the work by Sorano et al. 2020[6], we plan to develop a system to partially automate the detection of events in soccer matches such as passes, shots, fouls, etc. This system will be based on a deep learning architecture that combines state-of-the-art methods with artificial neural networks tailored for the specific declination of the problem.

Visual analytics of soccer tracking data. Tracking data is the finest spatio-temporal representation of a soccer game. The exploitation of such data sources enables discovering behavioral and tactical patterns. Visual analytics, in combination with mobility data analysis methods, is the key to obtain valuable results in terms of game dynamics understanding. However, there are a set of open problems to be addressed, from developing meaningful visual tools to speed up the computation of mobility analysis algorithms on complex and huge datasets such as the position of players and ball, on a soccer game, sampled at 25hz. The first goal of this project is the extraction of meaningful features from tracking data. Such extraction process is developed through the analysis of players interactions over space and time, i.e. the definition of an interactions network based on passing lines over time and opponents attempts to break such lines. The properties of this network would help to understand the impact of tactical decisions on the evolution of a soccer game.

Understanding the relation between athletic and technical performance. Following up on our research activity regarding injury analysis and prevention through machine learning tools, collaboration with sports clubs is peculiar for focusing on new open problems, related to the main topic. In particular, understanding the relationship between physical effort, training periodization, and technical performance is crucial, and the

[6] Sorano, D., Carrara, F., Cintia, P., Falchi, F., & Pappalardo, L. (2020). Automatic Pass Annotation from Soccer VideoStreams Based on Object Detection and LSTM. ArXiv, abs/2007.06475.

availability of complex and detailed data covering several aspects of players effort and performance is opening up opportunities to apply data science tools in this field. We plan to study the relationships between athletic and technical data. In particular, we compare a technical performance index, such as PlayeRank, and athletic features captured by video tracking sensors.

Men football vs. women football. Women's football is gaining supporters and practitioners all around the world, raising questions about what the differences are with men's football. We plan to develop an analysis to compare women's football and men's football characteristics based on the players' physical attributes, offering a complete characterization of their differences.

Heart Rate Variability via wrist-worn wearable devices. In recent years, the interest in the variation of the timing between beats (RR-intervals) of the cardiac cycle, called heart rate variability (HRV), has widely increased in the psycho-physiological research field. Assessment of RR-intervals variability is possible through time and frequency domain analyses that provide parameters able to quantify the fluctuations that occur between consecutive beats. The parameters extracted from HRV analysis provide insights about the sympathetic-parasympathetic balance of cardiac vagal tone, an indicator of cognitive, emotional, social and health status. Thanks to the technological advancements of recent decades, it is now possible to continuously record heart activity during peoples' lives via wrist-worn wearable devices equipped with heart rate sensors. This innovation might have a significant impact on the medical field because of the low cost of the devices and the possibility to obtain continuous passive measurements performed in an ecological setting, gaining an overview of the health status of users by assessing HRV features during their daily life. These wrist-worn wearable devices, however, produce several inconsistent RR-intervals produced not only by ectopic beats (e.g., atrial fibrillation and premature heartbeat), but mainly by motion and mechanical artefacts induced by external stimuli. We plan to develop tools able to filter RR-intervals that are affected by motion artefacts and then interpolating missing values to reconstruct the time series of HR data. Moreover, we will implement algorithms that will permit us to estimate HRV parameters from noise RR-intervals time series.

Human psycho-physiological responses. Wearable activity trackers are becoming increasingly popular to monitor physical activity, heart rate (HR) and sleep quality, allowing a precise overview of an individual's health status and well-being (e.g., cardiovascular status, sleep quality, and physical activity). Moreover, individual chronotype and anthropomorphic data, cortisol, melatonin saliva concentration, and activity diaries may help scientists assess the relationship between the physio-psychological characteristics of individuals. We plan to collect and release a data set about the above characteristics, in order to assess the effect of psychological, physiological and hormonal features observed during objective sleep quality.

3.5.2 Planned Activities

Data collection. We plan to collect data about human psycho-physiological responses, describing 24 hours of continuous psycho-physiological data, i.e., interbeat intervals, heart rate, wrist accelerometry, sleep quality index, physical activity, psychological characteristics (e.g., anxiety status, stressful events and emotion declaration) and sleep hormone levels for young healthy subjects.

Events. The broad popularity of sports analytics topics is helpful in the organization of open science initiatives. Over the last two years, we organized a hackathon in conjunction with two Italian festivals (https://sobigdata-soccerchallenge.it/). We missed the organization of the 2020 edition of Soccer Data Challenge because of the COVID-19 pandemic. We plan to organize a new "in-person" edition in fall 2021, depending on the international and national situation due to the COVID19 pandemic. We also plan to organize the 8th satellite on Quantifying Success, in conjunction with the International School and Conference on Network Science (NetSci). This satellite will be an opportunity to share ideas among scientists of different disciplines about how to quantify performance and success in several contexts, including sports.

3.6 Social Impacts of AI and Explainable Machine Learning

The exploratory will investigate the foreseeable impact of AI and Big Data on society, developing analytical and simulation tools. It will also integrate 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 Relevant Topics

Opening the black box. Black box AI systems for automated decision making, often based on machine learning over (big) data, map a user's features into a class or a score without exposing the reasons why. This is problematic for lack of transparency and for possible biases inherited by the algorithms from human prejudices and collection artifacts hidden in the training data, which may lead to unfair or wrong decisions. We will focus on the urgent open challenge of how to construct meaningful explanations of opaque AI/ML systems. We intend to investigate a local-to-global framework for black box explanation, articulated along three lines: (i) the language for expressing explanations in terms of logic rules, with statistical and causal interpretation; (ii) the inference of local explanations for revealing the decision rationale for a specific case, by auditing the black box in the vicinity of the target instance; (iii), the bottom-up generalization of many local explanations into simple global ones, with algorithms that optimize for quality and comprehensibility. During the first year we plan to design and experiment local explanation methods for tabular data, images data, time series and sequence data; to design, develop and experiment a local-to-global framework for tabular data; to set-up different case studies in the field of predictive justice and healthcare. We will design experiments on benchmark datasets and real data in the fields of healthcare and justice. In the medical domain, we will validate the developed explanation methods on the MIMIC-III dataset and we will set-up a real case study with medical researchers in transplants and diabetes. We also set-up a case study in the justice field thanks to the collaboration with LiderLab Institute for the validation of our explanation tools.

Differential Explainable AI: An Interdisciplinary Perspective. This research explores the notion of differential explainability, bringing together researchers from the domain of computer science (ML and robotic experts, automated business processing), law and ethics. Building on ongoing research projects of the LiderLab Institute, the research focuses on the differential explainability capabilities of AI systems employed in the legal sector. This will be done considering the various explanatory needs of the involved stakeholders that are judges, lawyers and the parties to a legal dispute. Accordingly, we will investigate the role of selected psychological phenomena that can affect interpretation of ML models when presented to judges and other legal professionals. From the different normative perspective, it enquires whether the existing EU ethicolegal framework acknowledged the need for differential explainability and whether it entails some specific requirements in this sense, producing from its use-case results of general applicability.

Relationship between explainability and privacy. Black-box machine learning models are used in critical decision-making domains, such as healthcare and justice, giving rise to several calls for more algorithmic transparency. The drawback is that explanations could leak information about the training data of the black box ML model, thus undermining data privacy. We will investigate the relationship between privacy and explainability analyzing how different types of explanation models can lead to privacy issues and we will try to define privacy preserving techniques suitable to provide privacy-aware explanations. In the first year, we plan to define privacy attacks based on local explanations. We will also investigate the local explanation methods for explaining privacy risk predictions. We will experiment methods of local explanation for privacy issues especially in mobility data and purchasing data available in SoBigData++. We also design experiments to simulate attacks to ML models based on local explanations.

Algorithmic Fairness. Algorithmic Fairness studies the presence of certain biases that might cause disparate impact and/or treatment derived from the intervention of an autonomous system. Our main purpose will be to work on novel techniques to identify, audit and/or mitigate Algorithmic Fairness in different scenarios. This topic will be approached from different perspectives, emphasizing the practical one considering the role of the developers building a potentially unfair system, as well as the point of view of the individuals potentially affected by the system making potentially unfair predictions. We will design experiments whose results can be of interest for the research community focused on Algorithmic Fairness but also for other communities with broader interests such as the ML community or communities of sociologists interested in the relation between society and autonomous systems. In these experiments, user testing of interfaces and ways of communicating the characteristics of a model, including any potential algorithmic discrimination, will be a main component.

Causality analysis of data. The ability to learn causality is a significant component of human-level intelligence which is hard to replicate in AI. In the field of explainability and fair decision making is fundamental to address the discovery and understanding of causal influences among variables. Approaches will be designed for learning such influences from the data or, for a given domain, both from data and domain knowledge. They will advance state-of-the-art in data sanitization, fair decision making, and explainability of AI and applications.

Decentralised AI. A key challenge in ML is how to make distributed models “learn” together. Despite initial results on this topic (most notably, Federated Learning), significant research is needed, particularly in case of completely decentralised environments, where nodes – under the control of their human owners – perform data analytics tasks in a collaborative way. In this perspective, we will investigate how to support “endless” learning, i.e., develop distributed ML solutions able to continuously incorporate fresh data as they become available in the training process. This presents interesting challenges regarding how to manage and combine very different streams of data/knowledge that might become available at very different rates in order to train efficiently accurate models. We will also address the heterogeneity of the distributed ML process, both in terms of devices and data sources, particularly taking into account resource constraints of users’ devices. This means that not only data but also models built on them can possibly be consistently different from one node to another, and must be tailored to run under significant resource limitations (e.g., in terms of storage or computation capabilities). We will study how to design distributed learning solutions able to cope with such heterogeneity without reducing accuracy of ML models. We will investigate distributed ML solutions that find the right operating point between centralised and decentralised approaches. An example might be using centralised approaches among devices controlled by the same owner, while decentralised solutions among additional federated nodes under different parties' control.

Human-centric AI. The current centralised approach to AI, whereby our personal data are centrally collected and processed through opaque ML systems ("black-boxes"), is not an acceptable and sustainable model in the long run. On the other hand, a novel way of new, decentralised and explainable ML modes are being developed. Along this direction, we will study collective AI algorithms built through a social network of nodes interacting according to human-centric models. Models of the human individual and social behaviour, coming from different disciplines (anthropology, sociology, physics of complex systems, micro-economics, influence theory) will be combined. Then, we will design distributed and decentralised ML algorithms that incorporate, in the way they interact, multidimensional information about the users individual and social behaviour. For instance, different ways of integrating local data and AI models, and assessing trust in the received information, will be based on the strength of social ties between devices' users. Integrating explainability mechanisms will also be a breakthrough. Within the 1st year we plan to define the decentralised ML algorithms incorporating human behaviour; to define and generate human behavioural models for generating synthetic datasets.

Finance & Economics. Data science and machine learning have emerged as a universal tool to study different complex systems, including finance. Recent advancements in ML, and in particular in Deep Learning, Reinforcement Learning and Self-Supervised Learning, have produced new theoretical insights and provided practical approaches for more accurate predictions, risk quantification, portfolio creation, algorithmic execution, and others. We plan to employ techniques such as those described above to study financial systems, in particular those, related to cryptocurrency markets. Cryptocurrency markets are new, exotic, and volatile. It is not clear why prices of cryptocurrencies experience bubbles and why their markets crash: this could be attributed to both, internal micro dynamics of exchange offices as well as exogenous effects. To understand the main driving forces of the price bubbles in cryptocurrency markets, we will study ``the great crypto crash" of 2018. With the means of information-theoretic measures we will analyse a microstructure of the financial crash to gain insights about its primary cause and generally about the nature of financial instabilities.

3.6.2 Planned Activities

Data collection. We will collect data from online platforms and/or social networks that will be analyzed by taking advantage of the capabilities of the SoBigData++ platform. We also plan to define and generate human behavioural models for generating synthetic datasets. Then, we plan to collect novel finance-related datasets from Twitter utilising transparent AI techniques for tracking news in real-time, matching these web news to tweets (social media), and tracking intraday tweet activity. Lastly, we will explore different ML models for predicting the number mentions on Twitter after the publication of an article. We also plan to open-source crypto-financial Twitter data.

Software development. In synergy with the WP8, we will start the design and development of a library including all the explanation methods that will be developed. We will also make available python software for simulating the attacks to ML models based on local explanations and for predicting and explaining the privacy risk of individuals.

Events. We will continue a sequence of workshops and tutorials on opening the black box. In particular, we will continue the experience of the XKDD (eXplaining Knowledge Discovery in Data Mining) workshop aiming at encouraging principled research that will lead to the advancement of explainable, transparent, ethical and fair data mining and machine learning (we already organized two editions of XKDD at the international conference ECML-PKDD in 2019 and 2020, https://kdd.isti.cnr.it/xkdd2019/, https://kdd.isti.cnr.it/xkdd2020/). We also intend to propose tutorials on explainability in international conferences such as that one proposed in ECML/PKDD in 2019 (https://kdd.isti.cnr.it/xkdd2019/) and that one that we will do at DSAA 2020 (http://dsaa2020.dsaa.co/tutorials/#dss). We will also organize workshops on data science and ML, with emphasis on finance applications, which will bring together a unique blend of world-class experts from industry and academia. The original workshop of the sequence was already organized by ETH Zurich and NYU Courant. It is also a part of Data Science in Techno-Socio-Economic Systems ETH course and supported by SoBigData++ (https://www.eth-courant-workshop.com/). We plan also to organize a Summer School in late 2021 related to interpretability/explainability of artificial neural network models.

3.7 SoBigData Interest groups

Interests groups are possible future Exploratories which will be investigated by the consortium to understand if there are interests and experiences which may be transformed in services. Those interest groups will organize meetings with experts in the field, researchers and industries to eventually become Exploratories in SoBigData++. In the first year of the project, the interest group in Network Medicine has been investigated. Since network medicine deals with sensitive data that causes privacy concerns in the field of medical Big Data, the ethical and legal framework developed by the WP2, will be deployed to ensure compliance with the EU ethics, principles and privacy regulations.

3.7.1 Planned Topics

Biological Random Walk: Disease Gene Prioritization algorithm. We plan to develop a novel framework named Biological Random Walks (BRW) to discover disease modules in the human interactome using different sources of information such as biological processes from gene ontology and GeneExpression from TCG.

As a part of validation of the predicted genes by in vitro models, the biological role of the key genes predicted in the previous computational analysis will be experimentally investigated by in vitro models. Specific cell lines will be selected to have high or low expression levels of the predicted genes according to data of the Human Protein Atlas database (https://www.proteinatlas.org/). Next, the selected cells will be transfected to silence or over-express the gene of interest. Gene silencing will be obtained by using the stealth RNAi technology following standard protocols. Overexpression of selected genes will be obtained using proper expression vectors.

Next, functional analysis will be performed. As for cancer-involved genes, different parameters of cell proliferation, invasiveness and differentiation will be evaluated. Specifically, inhibition/overexpression effect on cell proliferation and apoptosis will be evaluated by the MTT and TUNEL assay, respectively. Cell invasiveness will be evaluated by the Matrigel invasion assay.

Effects on the cell differentiation will be evaluated by measuring mRNA and protein levels of the tissue specific genes by real time PCR and western blot analyses, respectively.

Leukemia Ph-Like Patients: Network Medicine application. Philadelphia chromosome-like Acute Lymphoblastic Leukemia (Ph-like ALL) is a recently defined high-risk B lineage ALL subtype. While Ph-like ALL blasts do not feature the BCR-ABL1 fusion gene, their microarray profiles cluster together with blasts carrying the BCR-ABL1 fusion gene. At the same time, numerous fusions and activating point mutations in cytokine receptors and JAK-STAT pathway genes have been described in Ph-like blasts. While targeted therapy is available and very effective for Ph-positive cases, Ph-like cases are more difficult to treat as a result of genetic heterogeneity. We plan to analyze published microarray profiles of B-ALL cases and to apply a network science approach to identify Ph-like subclusters, reconstruct the disease network in each subcluster and propose subcluster-specific drug targets. In next year, we will start from the identification of the molecular signature shared between Ph-positive and Ph-like profiles to then employ computational methods to detect the transcription factors (TFs) driving such expression signature. The further plan is to reconstruct disease networks using the kinase-target network (kinome) with the goal of connecting the fused cytokine receptors with the active TFs and use drug-protein interactions to prioritise drugs with predicted activities in every subcluster.

Crohn Disease. We will present a module for local graph partitioning using personalized Page Rank vectors. We plan to develop a module that, starting from a graph, will find local communities with small conductance and then merge them to find non-overlapping communities. An approach is to use a random-walk-with-restart approach to explore network neighborhoods around the core monogenic IBD cluster and disease-module cohesion to identify functionally relevant GWAS genes.

Recurrent Risk on Thyroid Cancer. The aim of this study is to create a better stratification of patients according to future risk of recurrence from the first surgery/treatment; this model will be compared to the current used approach (American Thyroid Association risk stratification system). Furthermore, we plan to create a dynamic system, able to integrate the baseline features with the results of first follow-up visits, in order to improve the performance of the so-called “dynamic risk stratification”. We will create a new stratification, improving the distribution of Structural Incomplete in different classes (Low → High Risk). As part of this, we will extract importance of features that improve the stratification, and create a new prediction model of developing recurrence.

Partial Correlation for functional COPD subnetwork genes discovery. We will work on SNPs (single-nucleotide polymorphism), gene expression, and DNA methylation data in lung tissue samples from subjects with and without COPD. We will integrate these subject-specific data to established reference networks (such as PPI, gene ontologies, pathways, miRNA targets…) leveraging these as prior knowledges in our network analysis. The hypothesis that drives our study is the presence of a regional network of genes within chromosome 4q, and that through these interactions, functional genetic variants present in COPD GWAS loci impact disease pathogenesis. The ultimate goal is to understand the biological impact of COPD GWAS genes and their causality.

Psychiatric Mental Lexicon. In network medicine and cognitive network science, human cognition is represented and investigated as a complex network, allowing researchers to consider how network topologies and dynamics depict individual worldview through the complex relationships among the entities of the systems that underlie individual behaviours and thoughts. For instance, semantic (linguistic) networks allow considering human semantic memory as a network of well-connected similar concepts; thus, nodes representing words and edges representing relationships among them (e.g., semantic relations, free associations, phonological similarities) form a mental lexicon, in terms of a network modelling how word meanings are structured and processed in the human mind. We will study the structure and the dynamics of mental lexicons in healthy and clinical populations, assuming that the mental lexicon organizes an individual world view and can reveal relations with cognitive impairments, such as formal thoughts disorder. In complex network analysis, a less modular meso-scale network level can be present in individuals with cognitive impairments; for instance, a community or a cluster representing a semantic field (i.e., a group of well-connected words related to the same semantic domain or conveying similar meanings) can be less structured in psychotic individuals, which are involved in well-known positive symptoms such as disorganized thoughts and incoherent speech. We will investigate the identification of psychotic and schizophrenic thought-disordered world view in terms of a mental lexicon of word similarities; analyzing a mental lexicon of a possible early psychotic individual can help an analyst in the early detection of the symptoms and the illness.

3.7.2 Planned Activities

Data collection. The community of the CLAIRE-COVID19 Bioinformatics working group have as primary goal to support the community with the release of resources for characterising the disease from its related structural information, including prediction of viral protein folding; studying interactions between the virus and human hosts, including analysing protein-protein interaction data; filtering, retrieval, and generation of targeted drugs leveraging molecular and well as proteomic information; delivering predictive insights onto the genetic features of the virus. To enable these objectives, we plan to assemble a resource that fuses information from heterogeneous sources and different studies from the literature into a unique network-based representation, facilitating the use of relational and graph-based learning methods. The identified resources are:

- Protein-Protein [R1]: Protein-Protein Interactions (PPIs) are physical interaction between two or more proteins.
- Domains [R2]: Domains are distinct functional and/or structural units in a protein.
- Families [R3]: A protein family is a group of proteins that share a common evolutionary origin.
- Pathways [R4]: A biological pathway is an ordered series of molecular events occurring among molecules in a cell, and that leads to producing a certain biological product, or change in the involved cell.

- GO-Terms [R5]: Go-Terms are biological terms, or concepts, related to the genes.
- Drug-Host [R6]: A drug is designed to produce a specific desirable therapeutic effect on the target organism.
- Drug Structures [R7]: Drug structures provide information about the topological structure of the drug molecules, such as spatial coordinates of the atoms and their bonds.
- Drug-Drug [R8]: A Drug-Drug Interaction (DDI) is an alteration of the drug's expected effect - on the target organism - if administered with another drug product. Knowing whether a DDI produces a therapeutic or an adverse effect on the target organism is of paramount importance to repurpose multiple drugs together.
- Disease-Gene [R9]: The molecular network context «a disease is rarely a consequence of an abnormality in a single gene, but reflects the disruptions of the complex intracellular network».
- Virus-Host [R10]: The virus-host interaction represents a physical interaction between a virus molecule and a host (e.g., human) protein.

The dataset is published on the Territori Aperti VRE: https://territoriaperti.d4science.org/

Events. We plan to organize the following events in 2021:

- February 2021 (forthcoming): Research and Vulnerable Participants: Data Protection and Beyond (speakers, Dr. Denise Amram, SSA; Dr. Gianclaudio Malgieri, VUB). Involved Partner: Sant'Anna School of Advanced Studies, Pisa
- March 2021 (forthcoming): Health Data Management from Early Detection to Treatment: Ethical Legal Issues (speakers Prof. G. Comande'; Dr. Denise Amram; Dr. Giulia Schneider, SSSA). Involved Partner: Sant'Anna School of Advanced Studies, Pisa
- May 2021 (forthcoming): AI and Chronic Diseases: Current and Future Regulatory Challenges (speakers Prof. G. Comande'; Dr. Denise Amram; Dr. Andrea Parziale, SSSA). Involved Partner: Sant'Anna School of Advanced Studies, Pisa