SOBIGDATA++
PLUS PLUS

Deliverable D10.7

# SoBigData Interest groups report 3

SoBigData Plus Plus receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No. 871042.

# DOCUMENT INFORMATION

| PROJECT | |
|---|---|
| PROJECT ACRONYM | SoBigData Plus Plus |
| 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 & UNIPI |
| WORK PACKAGE PARTICIPANTS | CNR, USFD, UNIPI, FRH, UT, IMT, LUH, KCL, SNS, AALTO, ETHZ, PSE, UNIROMA1, CNRS, CEU, URV, CSD, BSC, UPF, Eli, CRA, UvA |
| DELIVERABLE NUMBER and TITLE | D10.7 SoBigData Interest groups report 3 |
| AUTHOR(S) | Luca Pappalardo (CNR), Aris Gionis (KTH), Ilaria Barsanti (CNR), Marco Braghieri (KCL) |
| CONTRIBUTOR(S) | |
| EDITOR(S) | Valerio Grossi (CNR) |
| REVIEWER(S) | Valerio Grossi (CNR), Michela Natilli (CNR) |
| CONTRACTUAL DELIVERY DATE | 31/12/2024 |
| ACTUAL DELIVERY DATE | 30/12/2024 |
| VERSION | 1.1 |
| TYPE | Report |
| DISSEMINATION LEVEL | Public |
| TOTAL N. PAGES | 10 |
| KEYWORDS | Exploratory, medicine, health |

# EXECUTIVE SUMMARY

This document updates deliverables D10.5 “SoBigData Interest groups report 1” and D10.6 “SoBigData Interest groups report 2”. It contains the activities in the interest groups, reporting the creation of new ones and the status of the resources available.

# 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

| | |
|---|---|
| AI | Artificial Intelligence |
| EC | European Commission |
| EU | European Union |
| H2020 | Horizon 2020 EU Framework Programme for Research and Innovation |

# TABLE OF CONTENTS

1 Relevance to SoBigData++ ..... 7
1.1 Relevance to project objectives ..... 7
1.2 Relation to other work packages ..... 7
1.3 Structure of the document ..... 7
2 T10.7 Network Medicine ..... 8
2.1 Activities Report ..... 8
2.2 Publications ..... 9
3 Conclusions ..... 10

# 1 Relevance to SoBigData++

Interest groups were introduced to define possible future exploratories which were investigated by the consortium to understand if there are interests and experiences which may be transformed in services.

This document describes: (i) the activity carried out within the SoBigData++ interest groups during the second reporting period; and (ii) the report about the creation of a new exploratory, the "Network Medicine" exploratory. For each interest group, we report the results achieved and the activities carried out in terms of conferences/workshops, hackathons, data collection, and software development. Those interest groups organise meetings with experts in the field, researchers and industries to eventually become exploratories in SoBigData++.

## 1.1 Relevance to project objectives

This document updates deliverables D10.5 "SoBigData Interest groups report 1"[1], and D10.6 "SoBigData Interest groups report 2"[2], and it is related to the activities for investigating and defining new exploratories inside SoBigData RI

## 1.2 Relation to other work packages

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

## 1.3 Structure of the document

Section 2 outlines all the activities performed for the definition and release of the new Network Medicine exploratory

[1] https://data.d4science.net/tnvY
[2] https://data.d4science.net/5YxK

# 2 T10.7 Network Medicine

## 2.1 Activities Report

**Network and Sequence-Based Prediction of Protein-Protein Interactions**

*Partners Involved:* UNIROMA1

Typically, proteins perform key biological functions by interacting with each other. As a consequence, predicting which protein pairs interact is a fundamental problem. Experimental methods are slow, expensive, and may be error prone. Many computational methods have been proposed to identify candidate interacting pairs. When accurate, they can serve as an inexpensive, preliminary filtering stage, to be followed by downstream experimental validation. Among such methods, sequence-based ones are very promising. In this activity, we designed a new algorithm that leverages both topological and biological information to predict protein-protein interactions. Preliminary results comparing our Framework with state-of-the-art approaches on reliable PPIs datasets, indicate that they have competitive or higher accuracy on biologically validated test sets. We claim that topological plus sequence-based computational methods can effectively predict the entire human interactome compared with methods that leverage only one source of biological information. Future work will evaluate the effectiveness of our hybrid approach in a comprehensive fashion.

**Cluster-based Relational Graph Convolutional Networks for Drug Repurposing**

*Partners Involved:* UNIROMA1

Drug repurposing is essential in modern medicine as it significantly reduces the time, cost, and risk associated with drug development by leveraging drugs already deemed safe for human use. This approach is particularly valuable in responding to urgent medical needs, such as developing treatments for emerging diseases. The study leverages the Drug Repurposing Knowledge Graph (DRKG), a comprehensive biological knowledge graph that integrates data from multiple sources. DRKG captures interactions between 97,238 nodes spanning 13 biological entity types (e.g., genes, drugs, diseases) and 5,874,261 edges across 107 different relation types. Originally designed to accelerate drug discovery efforts against COVID-19, DRKG serves as a rich and versatile resource for analyzing biological interactions and supporting predictive models. The methodology combines memory-efficient clustered training with Relational Graph Convolutional Networks (R-GCN) and introduces the “Frontier Expansion” technique to mitigate the loss of inter-cluster connections during clustering. This innovation allows the model to maintain high performance in link prediction tasks, essential for identifying potential drug-disease interactions. The results demonstrate that the proposed model achieves strong performance for drug repurposing while optimizing computational resources. The findings highlight the effectiveness of this approach and open avenues for future work, including exploring alternative clustering techniques, enhancing model interpretability through graph explainability, and conducting deeper evaluations in computational biology to validate the proposed solutions.

**Leveraging Relational Graph Attention Networks for Drug Repurposing**

*Partners Involved:* UNIROMA1

In this project we focus on enhancing drug repurposing methodologies by integrating attention mechanisms into graph-based models. Building on the foundation of the Drug Repurposing Knowledge Graph (DRKG) as a data source, the study refines the application of graph neural networks by leveraging Relational Graph Attention Networks (R-GAT). This model uses attention weights to dynamically prioritize the most relevant interactions in the graph, improving predictive capabilities for identifying associations between disease-

related genes and compounds. By employing attention mechanisms, R-GAT enhances the model's ability to focus on critical relationships within the graph. The evaluation process highlights its superiority over traditional models, achieving better performance in metrics such as MRR and Hits@[1, 3, 10], and demonstrating its ability to rank potential drug-disease associations with greater confidence. The model identified promising drug candidates like Tropicamide, Ifetroban, and Dexetimide, which were validated through biological pathway analysis and ongoing clinical trials. These results showcase the potential of incorporating attention-based techniques for more accurate and interpretable predictions. Future directions include expanding the feature set, optimizing computational complexity, and exploring attention weight decomposition to further enhance interpretability and efficiency.

**Drug Repurposing and Polypharmacy Side-effects prediction using Topological Deep Learning**

*Partners Involved:* UNIROMA1

The project focuses on advancing drug repurposing and polypharmacy side effect prediction by integrating topological deep learning with an adapted Relational Graph Convolutional Network (R-GCN) framework. This novel approach extends traditional graph-based modeling by leveraging cell complexes, incorporating higher-order structures such as triangles, to capture richer topological features within the data. As in other related projects, the study utilizes the Drug Repurposing Knowledge Graph (DRKG), a comprehensive resource capturing interactions between drugs, diseases, genes, and other biological entities. By extending the representation from simple graph edges to cell complexes, the model better reflects the complex biological interactions, such as multi-drug interactions and shared biological pathways, crucial for understanding polypharmacy effects. The adapted model combines relational information with topological features, enabling predictions that account for higher-order connectivity patterns. This approach is particularly valuable for polypharmacy side effect prediction, as it models not only direct interactions but also the broader topological context in which these interactions occur. Initial evaluations demonstrate that this methodology significantly improves predictive accuracy for both tasks, with superior performance in identifying novel drug-disease associations and polypharmacy side effects. Future work will explore further refinements of topological features, improve the model performances and computational costs, and investigate on model interpretability and computational efficiency. This project underscores the potential of combining relational and topological approaches to tackle complex biomedical challenges.

**Efficient Assessments of Knowledge Graphs**

*Partners Involved:* UNIROMA1

Knowledge graphs (KG) have an important role in medicine and biology, as they can be used to capture concepts such as diseases, genes, and drugs. However, performing deep analysis on them can be expensive in terms of computational power. In this projects we develop a framework for assessing the effectiveness of KG-related machine learning methods in an efficient manner.

## 2.2 Publications

M. Egger, W. Ma, D. Mottin, P. Karras, I. Bordino, F. Gullo, and A. Anagnostopoulos "ReliK: A Reliability Measure for Knowledge Graph Embeddings" Proc. of the 2024 ACM Web Conference (TheWebConf 2024), Singapore, May 2024.

# 3 Conclusions

The work done by the interest groups flowed into the definition of the research spaces for all the duration on the project. The network medicine is now on the Health Studies. Since the project is over and the revised the research spaces in 2024, we do not have planning the creation of new spaces but we continuously monitoring emerging fields and when necessary, we define new research spaces.

The current research spaces reflect the strategic emphasis of the project on improving the impact, quality, and the research fields supported by the RI.