Ref. Ares(2024)9231116 - 23/12/2024

SOBIGDATA++
PLUS PLUS

Deliverable D4.4

Periodic Training Reports 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	WP4 Training
WORK PACKAGE LEADER	KCL
WORK PACKAGE PARTICIPANTS	KCL, CNR, USFD, UNIPI, FRH, UT, IMT, LUH, SNS, ETH Zürich, UNIROMA1, CNRS, URV, KTH, SSSA
DELIVERABLE NUMBER and TITLE	D4.4 Periodic Training Reports 3
AUTHOR(S)	Mark Coté (KCL), Marco Braghieri (KCL)
CONTRIBUTOR(S)	Paolo Ferragina (SSSA), Richard Rogers (UvA), Pasquale Pagano (CNR), Sara Lelli (CNR), Giorgio Vinciguerra (UNIPI)
EDITOR(S)	Valerio Grossi (CNR)
REVIEWER(S)	Ilaria Barsanti (CNR)
CONTRACTUAL DELIVERY DATE	31/12/2024
ACTUAL DELIVERY DATE	23/12/2024
VERSION	2.0
TYPE	Report
DISSEMINATION LEVEL	Public
TOTAL N. PAGES	34
KEYWORDS	Training, Summer School, Datathon, Diversity and Inclusion, MOOC, SoBigData Academy

EXECUTIVE SUMMARY

This deliverable reports on activities performed under Work Package 4 - Training for the period from 01 January 2024 to 31 December 2024. It is divided in four sections, one for each main task of the work package and with deliverables *D4.1 - Training Planning and Reporting*, *D4.2 Periodic Training Report and Planning for the next Period 1* and *D4.3 - Periodic Training Reports and planning for the next period 2*, this document completes the report on training activities along the entire duration of the project.

DISCLAIMER

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

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

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

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

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

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

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

GLOSSARY

AI	Artificial Intelligence
EU	European Union
EC	European Commission
H2020	Horizon 2020 EU Framework Programme for Research and Innovation
MOOC	Massive Open Online Courses
VRE	Virtual Research Environment

TABLE OF CONTENTS

1 Relevance to SoBigData++

Work Package 4 - Training aims to establish a joint training and education resource on big social data promoting the education of the next generation of data science researchers. The Work Package explores and develops both conventional and unconventional training experiences for master students, PhD students and early career post-doctoral researchers as well as an academically interested general public. Likewise, Work Package 4 proposes campaigns aimed promoting interest and participation of under-represented communities in data science with special emphasis on gender issues.

1.1 Relevance to Project Objectives

The training activity within the SoBigData++ project is developing a unique, joint training and education resource centre on big social data. Building on the experience of the first iteration of the SoBigData project, Work Package 4 explores and develops conventional and unconventional training experience for master and PhD students and postdoctoral trainees. These experiences include the organisation of a number of different events and the development of the e-learning Area which has been created and integrated into the SoBigData Research Infrastructure. Moreover, it features the development of ad-hoc MOOC courses. Among events, project-oriented summer schools and datathons have been planned and organised in order to match research (and industrial) needs and people skills. Moreover, activities will aim to address gender and diversity issues in data science through training.

This Work Package is organised around four different tasks.

1 **Task T4.1**, 'Online Training Modules' was centred on creating open-source training materials that are integrated into the SoBigData RI within the e-Learning Area. This part of the Research Infrastructure was designed, created, and integrated into the SoBigData catalogue during the first iteration of the SoBigData project. The task has evolved into the creation of the SoBigData Academy, which was presented in the last reporting period.

2 **Task 4.2** is centred on the organisation of a minimum of one per year summer school, introducing participants to techniques and methodologies for analysing big data, in order to provide them with a solid background in the computational and mathematical theories behind algorithmic tools for empowering their future research. The summer schools have been strongly interdisciplinary and included experts across arts and sciences.

3 **Task T4.3** is centred around the organisation of Datathons, with a minimum of one per year, whose aim is to bring together young and bright minds in smaller dedicated groups, providing complementary theoretical and practical skills to visualise and analyse social big data questions addressing important societal problems.

4 **Task 4.4** Computer Science and Data Science currently fail to adequately embody staff equality and diversity issues. For instance, not only females but also minority groups, etc are still woefully underrepresented in data science. The aim is to leverage existing networks in order to raise awareness regarding the opportunities provided by employment in the field of data science. SoBigData++ will support specific events and provide travel grants for young female and minority group researchers, continuing an experience started in the first iteration of the SoBigData project.

1.2 Relation to Other Work Packages

The SoBigData++ project is organised around work packages which are combined in three main axes:

1. Community building (including innovation and networking activities)
2. Social mining research infrastructure building
3. User accessibility (granted by virtual and trans-national access)

Work Package 4 works closely with:

- WP2 – Responsible Data Science – This work package is mainly tasked with operationalising a legal and ethical framework for the whole SoBigData++ Research Infrastructure.
- WP3 – Dissemination, Impact and Sustainability – This work package is mainly tasked with developing dissemination and impact strategies for the entire SoBigData++ project.
- WP5 – Accelerating Innovation – This work package is tasked with widening the project's impact through innovation activities aimed at industry and other stakeholders, such as government bodies, non-profit organisations, funders, and policy makers.

Aside from these work packages, WP4 also works in collaboration with WP7 (Virtual Access) to design and integrate training modules into the SoBigData++ Research Infrastructure. Moreover, WP4 works alongside WP9 (JRA2 - E-Infrastructure and Supercomputing Network) to create operation manuals for facilitating platform exploitation in all the aspects will be made accessible through a specialised operation portal dedicated to developers, ICT managers, and service providers. Finally, WP4 also works alongside WP10 (JRA3 – Research Spaces).

1.3 Structure of the Document

The document is organised around the four main tasks of WP4 - Training).

- Section 2 reports the work done for T4.1 – Online Training Modules,
- Section 3 outlines the work done T4.2 – Summer Schools
- Section 4 in based on reporting the activities of T4.3 – Datathons
- Section 5 is related to T4.4 - Cultivating Diversity in Data Science Through Training

As this is the last deliverable due to the project completion at the end of December 2024, each section only features a reporting section. For each activity, an in-depth description is provided, along with a report on participants (where needed). This work completes the set of reports related to the project about training activities[1].

[1] D4.1 - Training Planning and Reporting - https://data.d4science.net/CsiH
D4.2 Periodic Training Report and Planning for the next Period 1 - https://data.d4science.net/BBvf
D4.3 - Periodic Training Reports and planning for the next period 2 - https://data.d4science.net/Ee36

2 Task 4.1 Online Training Modules

2.1 SoBigData Academy

2.1.1 Development

During the last part of the previous reporting period (M24-M48), the SoBigData Research Infrastructure started the development of a series of MOOCs (Massive Open Online Courses) which have led to the creation the SoBigData Academy[2]. During the development phase, a pipeline was created to organise the workflow between the course holder and the MOOC developer – Figure 1.

Figure 1 SoBigData Academy Course Workflow

This workflow was part of a guide (see Appendix 1) that was created in order to facilitate the seamless collaboration between the course holder and the MOODLE creator which allowed for the development of a cohesive group of courses which are now part of the SoBigData Academy.

[2] SoBigData promotes an open innovation culture in Big Data and AI, offering educational resources to support responsible data science - http://sobigdata.eu/academy

2.1.2 COURSES LAUNCH

To launch publicly the SoBigData Academy a webinar was held on 28 November 2024. This event featured the presence of the SoBigData Research Infrastructure Management team (Roberto Trasarti and Valerio Grossi from CNR), along with MOODLE creator Sara Lelli from CNR, Mark Coté (WP4 lead) from KCL and Riccardo Guidotti from CNR – Figure 2. The live event was recorded and is now available on the project's YouTube channel.

Figure 2 A screenshot of the launch of the SoBigData Academy available on the RI's YouTube channel

At present, the SoBigData Academy hosts 9 courses, which are:

1. BASIC PYTHON
2. DATABASE
3. DATA ANALYSIS
4. LEGAL AND ETHICAL ASPECTS OF DATA SCIENCE
5. INFORMATION RETRIEVAL
6. DATA MINING AND MACHINE LEARNING
7. DATA ANALYSIS WITH SPARK
8. DATA THEORY AND SOCIETY
9. COMPLEX NETWORK ANALYSIS

These nine courses have been developed by CNR with project partners SNS, KCL, SSSA, UNIVAQ, ETHZ and LUH. Each course has followed a development pipeline which entails a collaboration between the course holder and the MOODLE developer. This process is divided into two steps: the first is pre-processing, where the MOODLE developer and the course holder exchange training materials and develop a strategy to adapt the material to the MOODLE architecture and possibilities. As reported in Figure 3 each course has associated a visual information reporting the difficulty level of the MOOC.

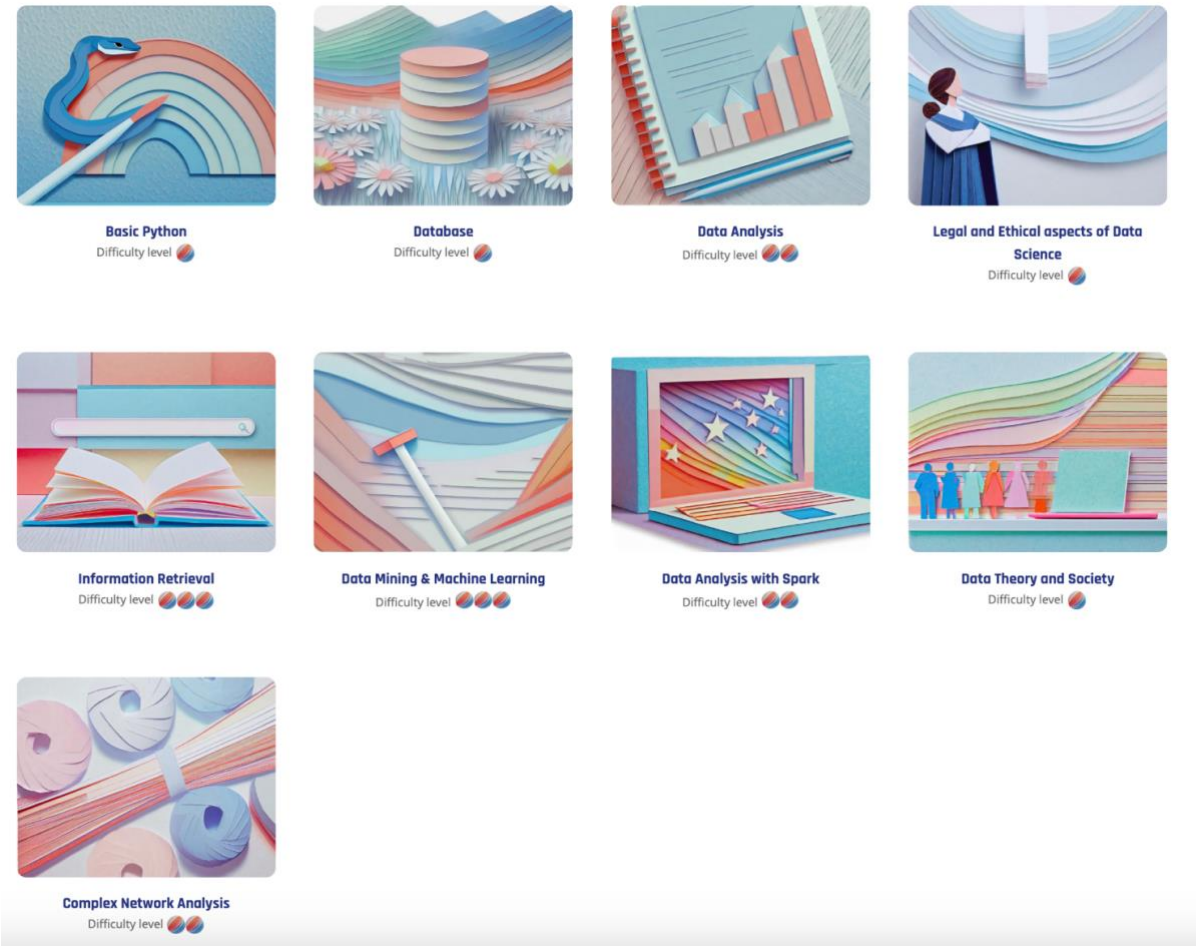


Figure 3 The nine MOODLE courses of the SoBigData Academy

The second step will be guided by user-interaction, as users will be able to provide feedback on the course and thus enable the MOODLE developer and the course holder to intervene if necessary.

Courses feature a variety of media, such as slides, videos, audio and a gamification approach which enables users to self-assess their learning path. Each course features two learning paths: on is the beginner path, which features interactive lessons, exercises and games, unit testing and final test; the second path is dedicated to expert users, which allows the user to access all the exercises, games and quizzes before taking the final test and accrue the course certificate and badge.

Along the existing nine courses, five are already in the development phase – Figure 4. These courses include:

1. ARTIFICIAL INTELLIGENCE
2. DATA VISUALIZATION AND STORYTELLING
3. NEURAL NETWORKS AND DEEP LEARNING
4. REINFORCEMENT LEARNING THEORY AND PRACTICE
5. TEXT ANALYSIS

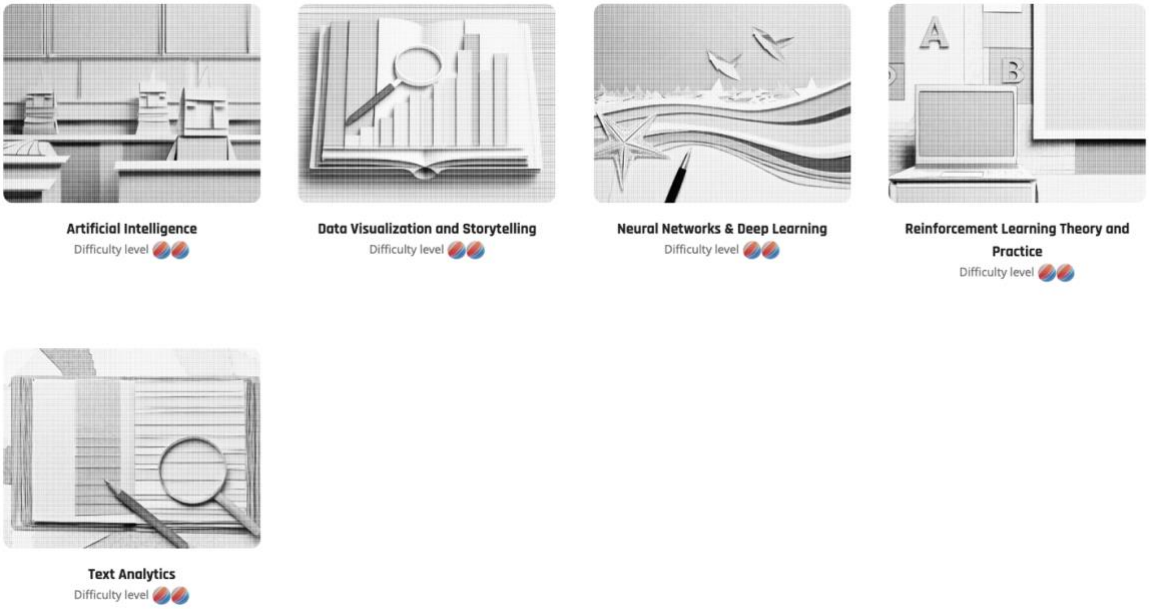


Figure 4 The five courses currently in the development phase

The course pipeline, which was outlined during the presentation event, will feature a course update every four months. Moreover, during the development of the Academy, a collaboration between SoBigData Research Infrastructure and Géant (Figure 5) was developed. Géant[3] is the union of European National Research and Education Networks (NRENs), which comprises a relevant number of European and non-European countries. Thanks to this collaboration, the SoBigData Academy hosts a connection to the following Géant courses:

1. Big Data Storage
2. Container based virtualisation: Docker \ Swarm
3. Container based virtualisation: kubernetes
4. Data modelling, data formats and protocols
5. Elasticsearch
6. GitHub
7. JSON
8. XML
9. YAML

Future developments in this collaboration will entail the presence of SoBigData courses within the Géant course offering and a direct access to Géant using the SoBigData identity login. The idea is to offer users a course selection that completes the one offered by SoBigData and vice versa, allowing Géant users to access SoBigData Academy courses.

[3] GÉANT eAcademy - https://e-academy.geant.org/moodle/index.php

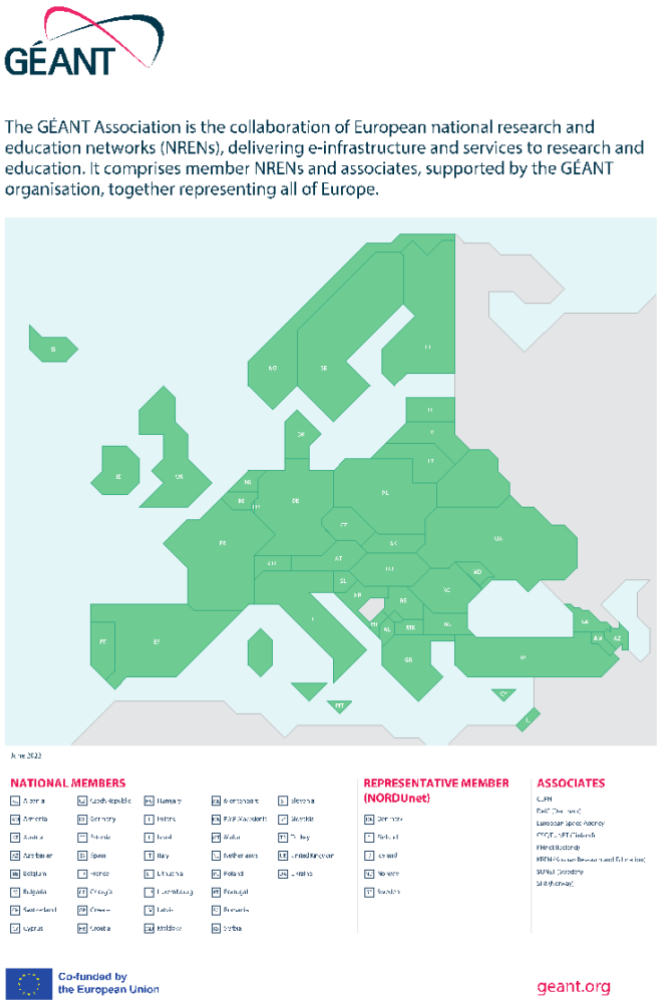


Figure 5 Géant Membership map

3 Task 4.2 Summer Schools

3.1 Reporting

3.1.1 SUMMER SCHOOLS IN THE REPORTING PERIOD

Since the last deliverable D4.3 on 2023 activity, which reported on an uptake of in-person activities after the COVID-19 pandemic, Summer Schools have been organised in an in-person setting. The following section reports on summer schools that were held in the reporting period (M48-M60).

3.1.1.1 DIGITAL METHODS WINTER SCHOOL AND DATA SPRINT "DIGITAL INVESTIGATION WITH AI"

The Digital Methods Initiative (DMI) held its annual Winter School and Data Sprint on "Digital investigation with AI[4]" between 8 and 12 January 2024 in Amsterdam, Netherlands. The format was that of a (social media and web) data sprint, with tutorials as well as hands-on work for telling stories with data. There was also a programme of keynote speakers. The school and data sprint is intended for advanced Master's students, PhD candidates and motivated scholars who would like to work on (and complete) a digital methods project in an intensive workshop setting.

The school focused on a variety of digital investigative epistemologies from fact-checking, debunking and source and media verification to algorithmic auditing. They aimed to address a wide variety of disruptions to the new media landscape, such as media and attention manipulation to continual influence and information campaigning, whether with harmful intention or more ironic and troll-like. The Winter School took up a series of questions concerning the investigative turn from the impact of disinformation and content moderation to the new conditions of artificiality and detection with AI.

The Winter School was a great success judging by the significant turnout in participation as well as the large number of projects completed. The atmosphere also was considered highly welcoming. There were 260 participants, half of which were female. The major age group was represented by individuals between 18 and 30 years old, followed by individuals between 30 and 50 years old (16-20), and individuals over 50 years old (11-15). A bursary system was created for female or under-represented early career researchers.

3.1.1.2 DIGITAL METHODS SUMMER SCHOOL AND DATA SPRINT "VISUAL METHODS: FROM PLATFORM AESTHETICS AND DATA VISUALISATION TO AI HERMENEUTICS"

The Digital Methods Initiative (DMI) held its annual Summer School and Data Sprint on "Visual methods: From platform aesthetics and data visualisation to AI hermeneutics[5]" between 1 and 12 July 2024 in Amsterdam, Netherlands. As the aforementioned winter school, the format was organised around 'data sprint' with tutorials and hands-on project work.

[4] Digital investigation with AI - http://sobigdata.eu/events/digital-methods-winter-school-and-data-sprint-2024
[5] Visual methods for digital research - https://www.digitalmethods.net/Dmi/SummerSchool2024

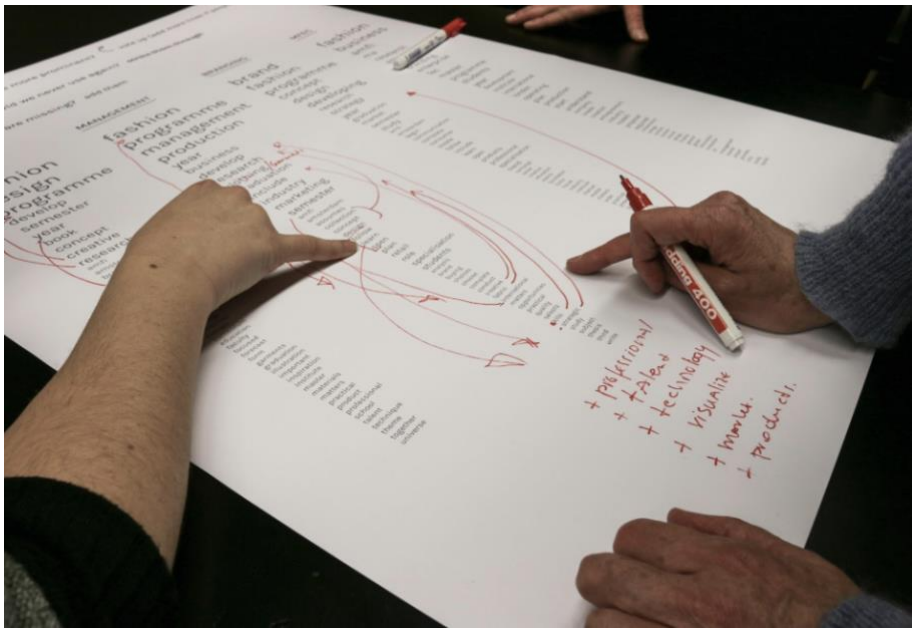


Figure 6 A snapshot of work during the Summer School

The school focused on platforms' visual styles in posts, comments, reactions, replies, threads, and conversation stoppers, which are also datapoints which depending on their combination can be visualised in specific manners. Learning to visualise (and when to visualise) are central to visual methods for the study of online data and images. Collections of images may be visualised, or they may be datafied and then visualised. Most recently, the study of image generation by AI platforms has multiple points of departure. Computational hermeneutics, or the analysis of the histories and styles extended in computational outputs, focuses on how the visual is the product of certain algorithms, models and/or architectures.

The event was a highly successful judging from the number of participants and the great number of projects completed, which can be found on the Summer School's page. The event attracted 127 participants (64 females, 63 males). The majority of participants were under 30 years old, and a bursary system was set up for an individual from an under-represented minority to participate to the summer school.

3.1.1.3 "EMPOWERING DATA FOR SOCIAL GOOD" SOBIGDATA SUMMER SCHOOL

The "Empowering Data for Social Good" Summer school took place in Baratti, Italy between 16 and 22 June 2024[6]. It was the second edition of the flagship SoBigData RI summer school, after the one which took place in Lipari, Italy in 2023. The school was designed to focus on pivotal subjects such as data governance, trustworthy AI (with a particular focus on applications related to health and sustainable development), Machine Learning and Data Altruism. The school's program was divided in two sections: morning sessions featured presentations by accomplished experts while the afternoon sessions were dedicated to collaborative group projects, allowing students to actively engage in hands-on learning experiences with the support of tutors.

[6] Second International SoBigData Summer School - https://summerschool2024.sobigdata.eu/

Figure 7 One of the group presentations during the last day of the Summer School

The school featured four keynote speakers, Dr. Maryam Mehrnezhad, Senior Lecturer at Information Security Group, Royal Holloway University of London, UK; Professor Katharina Morik from Technische Universität Dortmund, Germany; Dr. Alexandre Barth, Associate professor at Université de Liège and senior research fellow at Fonds de la Recherche Scientifique – FNRS, France; Professor Richard Rogers from University of Amsterdam, Netherlands. Along these four keynote speeches, a number of professors and researchers featured in the program, with a specific focus on gender equality featuring 5 female speakers and 7 male speakers.

Figure 8 Professor Katharina Morik's keynote focused on optimizing AI for energy efficiency and resource awareness, emphasizing techniques like pruning and quantization

The "Empowering Data for Social Good" summer school sparked high levels of interest and interaction among participants through its well-structured program and engaging content. Participants displayed significant interest in the topics covered, particularly data governance and trustworthy AI, which are highly relevant to contemporary issues in health and sustainable development. Interaction was a cornerstone of the summer school, facilitated through a combination of expert-led presentations and collaborative group projects.

Morning sessions allowed for exchanges between experts and participants, with ample opportunities for Q&A sessions and discussions that enriched the learning experience. The afternoon group projects were particularly effective in fostering hands-on interaction. Students worked together, supported by tutors, on developing projects, applying the concepts learned in the morning sessions to try to solve real-world problems. This collaborative environment encouraged peer-to-peer learning, problem-solving, and the development of teamwork skills.

The end-of-week project evaluations by a panel of data mining and AI experts provided an additional layer of interaction, offering constructive feedback and encouraging a deeper understanding of the topics covered. This approach not only validated the participants' efforts but also promoted a culture of continuous learning and improvement. Overall, the school was very well received by the students. Attendees were also administered a questionnaire in order for the organisers to understands which of the school elements were met positively and which areas need improvement.

Figure 9 Summer School attendees during one of the events in the social program

In total, 57 individuals attended the summer school between attendees, speakers and organisers: 60% male and 40% female, the majority of which were PhD Students. Another significant portion of attendees were Early Career Researchers and, finally, Academics. The SoBigData website also hosted a blogpost, with highlights of the Summer School, which can be found here.

3.1.1.4 LIPARI COMPUTATIONAL COMPLEX AND SOCIAL SYSTEMS SUMMER SCHOOL "GENERATIVE AI, HUMAN DECISION AND MACHINE PREDICTION: MODELS, ALGORITHMS, PLATFORMS AND ETHICS"

The 2024 edition of the Lipari Computational Complex and Social Systems Summer School on "Generative AI, Human Decision and Machine Prediction: Models, Algorithms, Platforms and Ethics" took place in Lipari, Italy between 14 and 20 July 2024[7].

[7] Generative AI, Human Decision and Machine Prediction: Models, Algorithms, Platforms and Ethics - http://sobigdata.eu/events/computational-complex-and-social-systems-lipari-summer-school

Figure 10 Vivek Natarajan's keynote during the Summer School

The school focused on Big Data issues arising in Complex and Social Systems, with an emphasis on techno-social innovation, health and the related ethical and societal implications, benefits and challenges. Moreover the school has seen some special sessions in which the PhD students have presented their research topics to get insights from the school speakers and organizers.

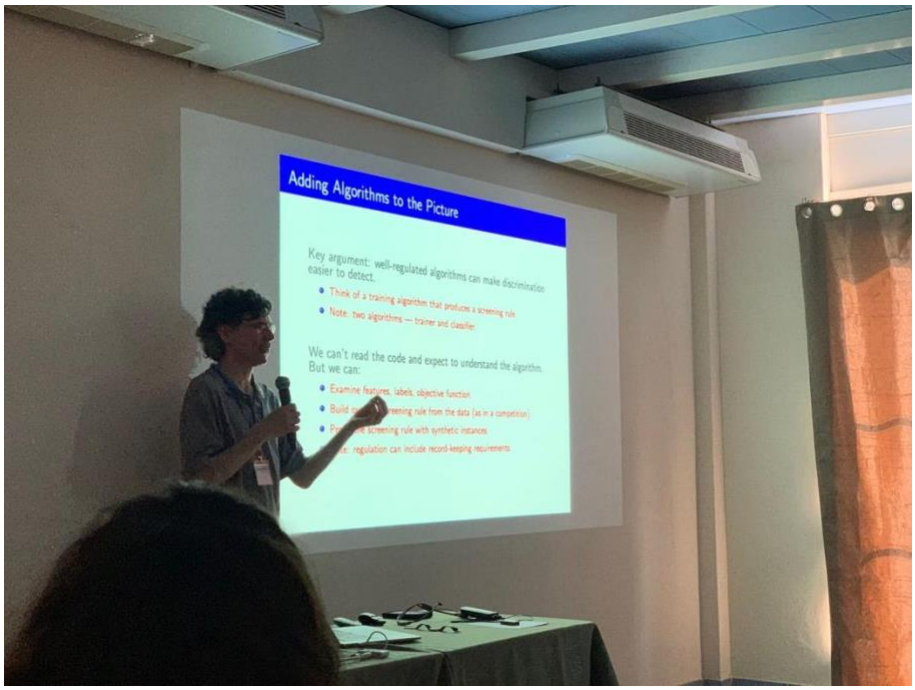


Figure 11 Jon Kleinberg's keynote during the Lipari Summer School

The school's program featured four keynote speakers, Jon Keinberg from Cornell University, USA; Mouna Lalmas-Roelleke from Spotify; Vivek Natarajan from Google Health AI; Filippo Caruso from LENS and

Università di Firenze, Italy. Along these four keynotes, seminars were held by Dirk Helbing (ETH Zurich), Carlo Ratti (MIT, CarloRatti Associati) and Paolo Ferragina (SSSA).

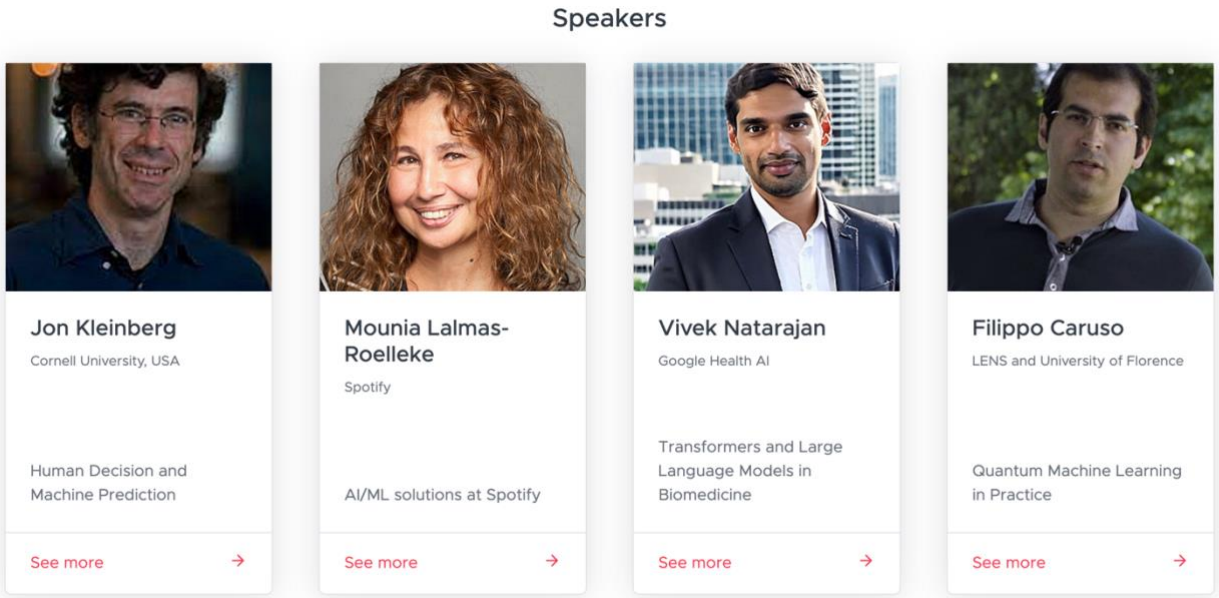


Figure 12 Keynote Speakers and topics

The school had 71 registered attendees, 16 of them female, the rest male. The majority of attendees were under 30 years old, 16-20 individuals were between 30 and 50 years old and a small group were over 50 years old. The attendees were overwhelmingly PhD students, followed by some (6-10) Early Career Researchers and a few individuals working in industry (1-5).

4 Task 4.3 Datathons

4.1 Reporting

4.1.1 DATATHONS AND OTHER TRAINING INITIATIVES IN THE REPORTING PERIOD

This section reports on datathons and other forms of training activities (such as workshops and awareness panels) have taken place in an in-person setting, hybrid and virtual setting during the reporting period (M48-M60). There have been three datathons or other training initiatives organised in the reporting period.

4.1.1.1 WORKSHOP "FROM SEARCH ENGINES TO CHAT GPT AND INTELLIGENT AGENTS"

The workshop "Workshop on "From Search Engines to Chat GPT and Intelligent Agents" was organised on 15 March 2024 by project partner KCL in London, United Kingdom and featured the participation of Professor Paolo Ferragina from SSSA.

Participants learned state-of-the-art developments: from classic Search engines to Retrieval Augmented Generation (RAG). This latter is a combination of classic Information Retrieval techniques with the more advanced and recent Generative AI approach (à la ChatGPT). The goal is to make LLMs more sustainable, more accurate and up to date, less prone to hallucinations, and suitable for vertical applications.

The workshop began with a comprehensive and concise overview of the history of search. It covered the paradigms of syntactic, semantic, and now vectorised search. Prof. Ferragina then introduced Retrieval Augmented Generation (RAG), a combination of classic Information Retrieval techniques with the more advanced and recent Generative AI approach. This was implemented on Colab with Python. This is available to participants to undertake future sustained research, some of which may be done on the SoBigData ++ RI.

The workshop was attended by 42 individuals (37 females and 5 males) between postgraduate research students and early-career researchers at King's College London, Strand campus. This was a highly cross-disciplinary group of participants. The organiser has already received feedback from a number of the participants, including some who plan to follow up on the workshop with a more sustained research project. Students participated significantly in the workshop, showing much interest in the latest advancements on RAGs. Discussion at the end was devoted to investigating possible coding projects to be then developed by students.

4.1.1.2 WEBINAR "EXPLORING THE FUTURE OF SOCIAL MINING ANALYTICS WITH THE CLOUD COMPUTING ENGINE"

The webinar "Exploring the Future of Social Mining Analytics with the Cloud Computing Engine" was held on 27 March 2024 and was organised by CNR along with project partner NUBISWARE[8]. The 90-minute webinar aimed at unveiling the new Cloud Computing Engine (CCP), the next evolution of the RI's Social Mining Analytics Engine (SMAE). The Cloud Computing Platform (CCP) is a pioneering service that represents a significant leap forward from the current Social Mining Analytics Engine. CCP embodies the latest advancements in Information and Communication Technologies (ICT), emphasizing the widespread adoption

[8] Cloud Computing Platform - http://sobigdata.eu/events/exploring-future-social-mining-analytics-cloud-computing-engine

of microservice development patterns, which have greatly enhanced software interoperability and composability.

The event was structured with 2 presentations of 20 min each aiming at introducing the service, its design principles, its architecture, and main functionality. A live demo showing how to import a method and how to execute a method in the infrastructure, from Giulio Rossetti (CNR), followed the presentations. Finally a Q&A session concluded the event.

The event was attended by 25 individuals (5 females and 20 males) and it generated interest also from outside the SoBigData RI consortium. The Q&A session was quite interactive with a number of questions and curiosity from the audience.

Figure 13 The webinar is now available on the RI's YouTube channel

The event was recorded and is now available on the Research Infrastructure's YouTube channel.

4.1.1.3 PENSIERO COMPUTAZIONALE WEB COURSE

The 2024 edition of Pensiero Computazionale was organised by UNIPI. It is a 26-hour web course dedicated to STEM secondary grade schoolteachers in Italy[9]. A final exam at the end of the course provided a certification for teachers attending successfully the whole course. This year's edition aimed to provide insight on how to analyse and solve problems by creating algorithms. The course has been running since 2018 and has accrued over 500 attendees and provided 230 completion certificates. The course was organised around 9 lessons and 4 laboratories.

[9] Aspetti di Base e Applicazioni dell'informatica - https://ilpensierocomputazionale.di.unipi.it/2024/

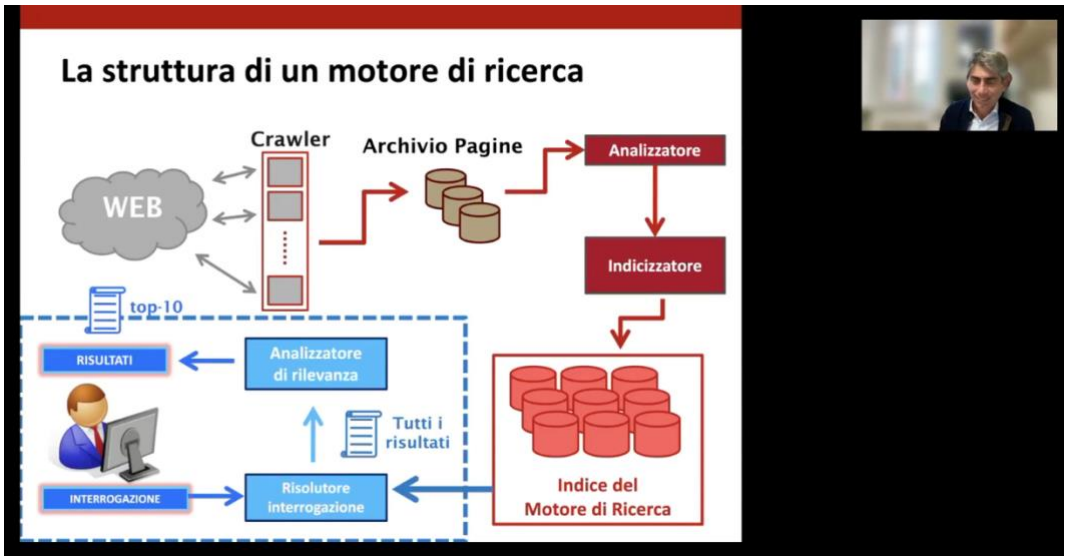


Figure 14 A snapshot of Professor Paolo Ferragina's lesson on search engines

The lessons were organised as follows:

1. Introduction to Algorithms and Coding
2. From Transistors to Quantum Chips
3. Cryptography
4. Data Science
5. Teaching tools for individuals who are disabled or have Specific Learning Disorders
6. Bioinformatics
7. Artificial Intelligence
8. Health-related informatics
9. Search Engines

Labs were organised as follows:

1. Cryptography Lab
2. Data Science Lab
3. Artificial Intelligence Lab
4. Teaching tools for individuals who are disabled or have Specific Learning Disorders Lab

The event collected 56 registered users and 70 attendees overall, 36 of which were female, 34 males. The lessons and labs were interactive, with participants actively engaging by asking questions and contributing to discussions. All lessons and labs were recorded and made available on the website along with study material, allowing participants to catch up on missed live sessions. Notably, 20 participants registered on the SoBigData gateway and joined the dedicated VRE. Ultimately, 16 participants successfully passed the final quiz and earned a certificate of participation.

The event addressed the project's objectives by enhancing competencies among high school teachers in key themes of interests like big data, data science, machine learning, and digital health. Some labs actively leveraged Research Infrastructure's JupyterHub, with participants encouraged to follow along in real time. As a result, 20 new participants registered on the SoBigData gateway, joined the dedicated VRE for this initiative, and gained hands-on experience with the SoBigData platform.

5 Task 4.4 Cultivating Diversity in Data Science Through Training

5.1 Reporting

5.1.1 SOBIGDATA AWARD FOR DIVERSITY AND INCLUSION

5.1.1.1 CREATION OF THE AWARD

Computer and Data science currently fail to adequately address equality and diversity issues, as there are genders and minorities which remain woefully underrepresented. The European Integrated Infrastructure for Social Mining and Big Data Analytics, the SoBigData++ Horizon2020 project has a mandate to promote equality and has established an award to promote a more diverse participation in computer and data science events. SoBigData++ has created the SoBigData Award for Diversity and Inclusion aimed at fostering participation from underrepresented groups, providing support to cover costs connected with the participation in computer and data science events.

Among the practical actions undertaken by the project, SoBigData++ has set up a bursary system to facilitate participation of underrepresented scholars from minorities in conferences and events. This action has been developed by project partners KCL and CNR.

The first step was to create an internal evaluation committee that would select relevant events from the project's field of interest and evaluate candidates. Secondly, a ranking system was developed to evaluate all the candidates that the project received from each event. The point system was developed to ensure transparency and an equal evaluation of all candidates. Thirdly, KCL and CNR developed a media kit that, once an event was selected, aided organisers in disseminating the opportunity given by the award. Moreover, the project became a sponsor of each selected event, in order to promote the project across different scientific fields. Along with the media kit, KCL and CNR developed a form in order to collect all candidates and ask them for a cover letter describing why they were applying for the award. The project also became a sponsor for the selected events, granting further visibility to SoBigData++.

5.1.1.2 EVENTS IN THE REPORTING PERIOD

Following the first two events held in 2023, the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, which was held in Turin, Italy and the 10th IEEE International Conference on Data Science and Advanced analytics which was held in Thessaloniki, Greece, the project selected other two events during 2024 which were supported by SoBigData++ and featured the SoBigData Award for Diversity and Inclusion support.

5.1.1.3 THE THIRD EVENT

The third event selected for the award was the 25th IEEE International Conference on Mobile Data Management that took place in Brussels[10], Belgium between June 24 and June 27, 2024. As per the conference's website, "The Mobile Data Management (MDM) series of conferences, since its debut in 1999, has established itself as a prestigious forum for the exchange of innovative and significant research results in

[10] https://mdm2024.github.io/callsResearch.html

mobile data management. The conference provides unique opportunities to bring researchers, engineers, and practitioners together to explore new ideas, techniques, and tools, and exchange experiences".

For this third event we received eight applications, which were all assessed following the ranking system developed for the first award. Our ranking system determined the three awardees. Each individual was awarded a €1.100 reimbursement to aid in conference fee, travel, and accommodation costs.

The first awardee was Dr. Chenxi Liu[11], Research Fellow at the S-LAB, which is part of the Nanyang Technological University, Singapore. Her research focuses on Spatial-Temporal Large Language Model for Traffic Prediction. The second awardee was Chidiebere Ogbuchi, a master's student working on his research thesis at the Satrai Lab[12], part of the Telecom Sud Paris, Institut national des télécommunications. The third awardee was Dennis Iroere, an MA student at the NOVA IMS[13] is the School of Information Management and Data Science of the Nova University of Lisbon.

5.1.1.4 THE FOURTH EVENT

The fourth event selected for the award was the Equity, Diversity & Inclusion event[14], held in conjunction with ACM KDD 2024 which took place at the Centre de Convencions Internacional de Barcelona, Spain on Wednesday, August 28th, 2024. This special session, as per the event website, was focussed on highlighting "research, innovation and policy in data science that aims to promote equality, diversity, inclusion and well-being. This initiative is not only relevant in data science, but also across government, the private and public sectors, activism and media".

For this fourth event we received seventeen applications, which were all assessed following the ranking system developed for the first award. Our ranking system determined the two awardees. Each awardee received a €1.100 reimbursement to aid in conference fee, travel, and accommodation costs.

The first awardee was Hannah Sansford[15], a PhD student from the University of Bristol, United Kingdom. In her cover letter, she underlined how "As a female PhD student in computational statistics I am very accustomed to being surrounded by men; however, every woman I encounter in my field drives my enthusiasm and belief that I can succeed in higher academic positions that unfortunately are even more dominated by men. I always strive to do my part in showcasing my diversity, in the hope to also inspire others. For example, in my University I helped to run a 'Women and Non-Binary People in Maths' event for undergraduate students, in which we helped de-mystify the PhD application process and day-to-day life of a PhD student".

The second awardee was Alexander Asemota[16], a PhD Student from the University of California, Berkeley (United States). He is a PhD student at the Statistics Program within the university. In. his application, he underlined how "the benefits I've gained from outreach efforts has instilled in me a deep appreciation for diversity and inclusion, and I have worked to improve our field in those aspects. I hope that my attendance at KDD can serve as both representation for those who may not feel seen and as an opportunity to further

[11] https://chenxiliu-hnu.github.io/homepage/
[12] https://satrai.telecom-sudparis.eu/
[13] https://www.novaims.unl.pt/en/who-we-are/
[14] https://edi-kdd2024.github.io/
[15] https://compass.blogs.bristol.ac.uk/students/hannah-sansford/
[16] https://statistics.berkeley.edu/people/alexander-asemota

improve diversity and inclusion in data science. I'm excited to make new connections that can engender increased effort in the realm of EDI".

5.1.1.5 OVERVIEW OF AWARDEES

Overall, during the period in which the SoBigData Award for Diversity and Inclusion has been active, it has supported four females and five males, for a total of €10.000, aiming to support participation from underrepresented groups, providing support to cover costs connected with the participation in computer and data science events. The individuals who received the award were: Dr. Vijayalakshmi Saravanan; PhD student Monika Jain; PhD student Rhitabrat Pokharel; PhD student Bishal Lakha; Dr. Chenxi Liu; MA student Chidiebere Ogbuchi; MA student Dennis Iroere; PhD student Hannah Sansford and PhD student Alexander Asemota. Thus, the award was given to two MA students, five PhD students, and two early career researchers.

6 Conclusions

This deliverable reported on activities performed under Work Package 4 (Training) of the SoBigData++ project. As the last deliverable of its kind due to the project's termination in December 2024, it features only a reporting section for each of the four tasks around which WP4 is organised.

Regarding Task 4.1 (Online Training Materials), the project has reached a significant result by activating the SoBigData Academy. It is MOODLE-based learning environment with nine active courses, five in the development phase and it also allowed the RI to open up a collaboration with the Géant network. Regarding Task 4.2 (Summer Schools), the project has organised four summer schools in the reporting period, on a wide, interdisciplinary series of topics. Regarding Task 4.3 (Datathons) the project has organised three datathons or other training initiatives, including online courses, webinars and seminars. Finally, regarding Task 4.4 (Cultivating Diversity in Data Science Through Training) the project has administered five bursaries under the SoBigData Award for Diversity and Inclusion, bringing the overall total of awards administered during the project to eleven.

Overall, project partners have developed significant synergies and experience in training, allowing an ever-growing number of individuals to benefit from the SoBigData Research Infrastructure. This has allowed not only a significant number of events to be organised but also the development of tools such as the SoBigData Academy which represent a lasting component of the SoBigData RI. Moreover, significant work has been performed around Task T4.4 (Cultivating Diversity in Data Science Through Training), not only with the creation and administration of the SoBigData Award for Diversity and Inclusion but with a growing number of female and under-represented groups of individuals which have featured as speakers at SoBigData events.

Appendix 1

CREATING A MOOC FOR TEACHERS

Please send all the materials within the span of **two weeks** from the start of your course's transformation.

As teachers, it's important to follow your ideologies and techniques, so please be yourselves. You are very welcome to communicate with your learning designer about your needs. You will find the contacts in the section below (Contacts).

You will be given a link to your **course's folder,** in which you will insert all the materials and information about the course.

Step 1 - Set up the course

You will be given a copy of the Template_Course document, in which you will need to insert the setup information.

- Course name
- Course's summary / description, with a small video clip presentation of the course and of yourself (to add in your course's folder)
- Course prerequisites, like special knowledge and/or a preparatory course from the SoBigData Academy
- Course's image (OPTIONAL, to add to the folder)
- Duration of the course
- Course's main topics - course units
- Learning Outcomes, what do you expect your students to learn during the course
- Type of Lessons (slides|video|text|Youtube video)

Step 2 - Coding part

In your copy of the Template_Course document, you will need to insert the information below:

- Do the students need to code? Will they use the Jupyter Notebook from the SBD gateway, or do they need other coding software?

Into your folder, add all the datasets and content students will download and use during the course's activities. If you have any URLs to share with your students, write them into your Template_Course document.

Step 3 - Units' materials

Please create and insert into your course folder the materials needed for each unit, composed of:

1. **lessons** per unit (slides, videos, text to read, etc.);
2. at least **3 questions per lesson** (pointing out where to put them inside the lesson) – look at Examples of lesson's questions;
3. **Unit glossary**, to insert into your copy of the document Template_Glossary. The unit glossary contains important words that the students will find throughout the unit and their definitions
4. about **20 questions** per unit with specific **feedback** – look at Examples of quiz questions;
5. a document with all the **coding solutions to the quiz's questions**. The document will be presented to the students after the end of the first try of the quiz, and it will help them understand how some problems can be solved.

Lessons

Teachers will need to create content for asynchronous lessons and insert it in the special folder.

They can either choose **slides**, using the **template** given by the SoBigData design team, or **video presentations**, **text to read**, short videos from **Youtube,** etc.

Recommendations:

The content needs to be **clear** and **brief** since the students will be alone in their learning path (**self-learning**), and it's useful to have short lessons (about 10 minutes) to properly and consistently capture their concentration.

First, organize the topics in each unit. For example: Data Analysis's first unit is "Data, tables, distributions, descriptive statistics", and has the following lessons: "Basic concepts in statistical data analysis", "Probability theory basics", "Data in Python", "Descriptive statistics".

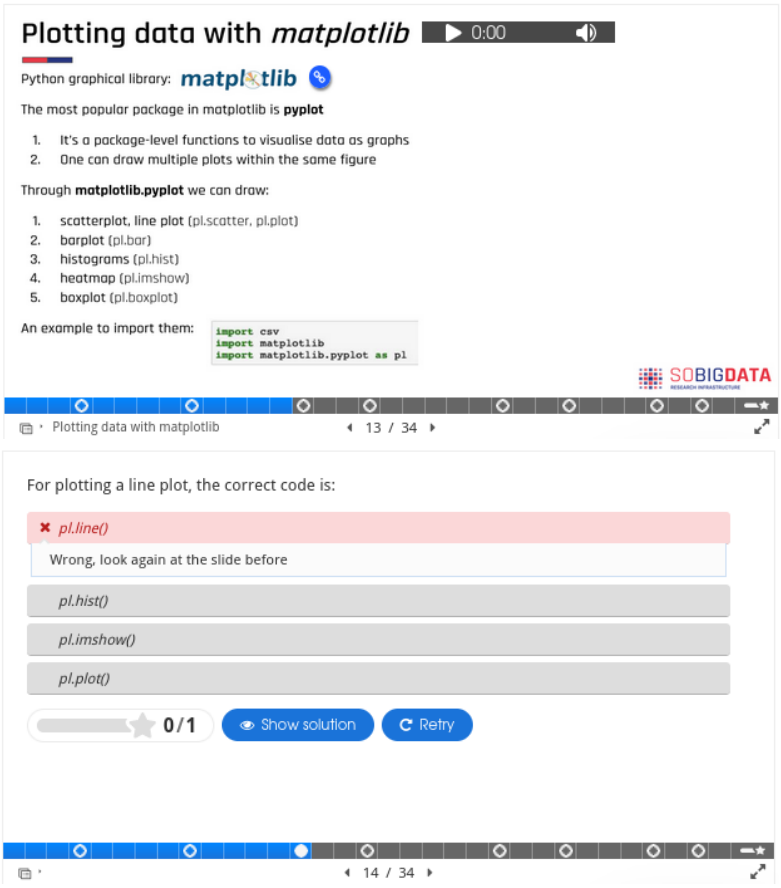
Choose which **content you want to use** (video, slides, text) and use it for every lesson in the course for continuity purposes.

Write the topics' division and the type of content into your copy of the Template_Course that will be added to your course's folder

If **slides** are used, please use the Google presentation template given in your folder (*course folder - lessons)* and create them there. You can find an example of the template in here: Template_Slides.
Slides with point lists are ineffective when there isn't a teacher to explain them, so it's recommended to use them as little as possible and only when the content is sufficiently clear.
Images need to be briefly explained with a text description. Accessibility is necessary.
For a better self-learning and accessible experience, it is best to add **audio** to the slides.

- You are free to record your own voice and insert the recordings into your *course folder - lessons - recordings* so that your learning designer can add them to each slide (if it is possible, create one audio file for each slide in a proper extension and give it an appropriate name to match the slide);
- or you can write the text that is needed for explaining images, codes, formulas, and difficult knowledge into your copy of the Template_Recordings that you will find in your *course folder - lessons - recordings*, and the learning designer will proceed to do them for you.

All your lessons will be created in your *course folder - lessons* in Google Drive. An example of a lesson using slides and recordings is the following:



Example of a lesson using slides and recordings

Create at least **3 questions** to ask during the lessons so that they can be interactive and involve the students in their learning.

The questions in the lesson can be:

1. multiple choice
2. true/false

The feedback in this case is not necessary since the questions are embedded in the lesson.

Please indicate in the question document where you will put the questions (after which slide or in what second).

The questions for both interactive lessons and quizzes will be created in a Google document provided after the topic division; you will also find them in the course folder. The document will be organized into sections (unit - questions for lessons - questions for the unit's quiz). When using slides, you can directly create the questions into them; you will find a question's slide example in the Template_Slides. An example will be found in the Template_Questions&Feeback.

If you have **lectures for deepening the topic** that you want to recommend to your students, please add the file to your course's folder or write the URL in the copy of the Template_Course you will find in the folder so that it can be inserted in the learning flow.

Examples of lesson's questions

1. **(Multiple choice example)**

…. Question ….

1. answer 1 (correct)
2. answer 2
3. …

2. **(True/false example)**

…. Question ….

Correct answer (True or False)

Questions

Each unit will have a unit test at the end that will determine if the student is ready to jump to the next one. The test can be retaken an unlimited number of times. You will write your questions in your copy of the Template_Questions&Feeback.

Types of questions you can choose from:

1. **multiple choice**
2. **true/false**
3. **matching** between an image and a text or two texts (used for matching a set of formulas with their names or for setting up a process)
4. **numerical**

It's recommended to do at least 15-20 questions for each unit.

Please try to ask at least two questions for each type of question. Have a look at the Example of quiz questions' section for a better understanding.

It's important to give **appropriate feedback** to the questions so that students can understand what the error is and how to fix it.

Feedback

Feedback is really important in a self-learning course, as it helps students boost their confidence, immediately corrects misconceptions or memory errors, and decreases, in general, the risk of students acquiring incorrect knowledge.

There are three main **types of feedback**: **right/wrong feedback**, **corrective feedback** with the right answer written in it, and **elaborate explanatory feedback**. Other two aspects that can be written in feedback are: **where to find the missing information** needed to answer a particular question, and **deepened information** that helps students achieve higher knowledge.

For the wrong answer, you can either choose general feedback or specific feedback for each wrong answer.

As we would like to create high-quality courses, feedback is surely one of the most important aspects to take into consideration. In the next paragraph, there will be some tips to follow.

For theoretical questions

Right/wrong feedback can be used, but it's important to point out where, in the lessons, to find the information needed to answer the question. You can write the feedback using the formula "*Wrong answer.* ***You can find the information you need in*** *the Beginner lesson …*". In this case, students will be encouraged to re-study the misunderstood topic.

Corrective or elaborative feedback can be used too, since it will help students understand what the error is and correct it immediately.

Of course, if you would like to recommend to your students some deepening knowledge, you are free to add to the feedback any new brief information you find interesting or link to papers they can read.

For questions where coding is needed

It's important to point out **what needs to be done** to solve the question, using **elaborative feedback**.

You can also suggest looking at a specific lesson in the course, a YouTube video you recommend, or a github link, but it's facultative since, at the end of the first try of the quiz, your personal coding solutions will be available.

Examples of quiz questions

1. **(Multiple choice example)**

…. Question ….

1. answer 1 (correct)
2. answer 2
3. …

Feedback:

You can either choose general feedback for the wrong answers or specific feedback for each wrong answer (in this last case, please use the same order as the answers's list)

2. **(True/false example)**

…. Question ….

Correct answer (True or False)

Feedback:

In this case, the feedback is only for the wrong answer

3. **(Matching example)**

…. Question ….

Drop the steps or phrases into the correct zone

1- text or image -> Answer in 1

2- text or image -> Answer in 2

3- text or image -> Answer in 3

…

Feedback:

You can either choose general feedback for the wrong answers or specific feedback for each wrong answer (in this case, please use the same order as the answers's list), or general feedback for a range of correct answers (e.g., if the answer is correct for 60%, then there is a certain feedback, and so on. Please specify the range)

4. **(Numerical)**

…. Question ….

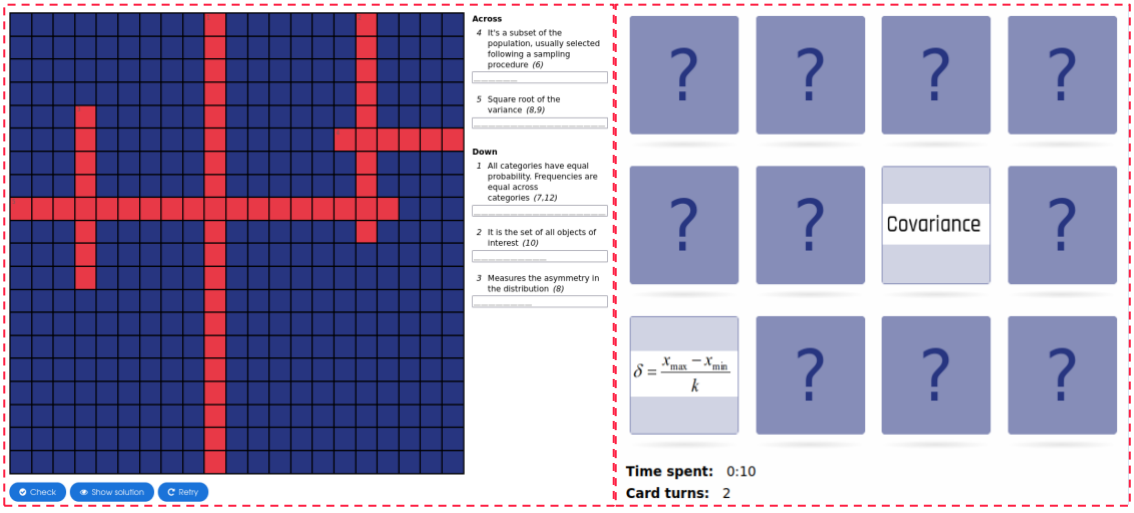
Possible correct answer or answers:

Feedback:

In this case, the feedback is only for the wrong answer

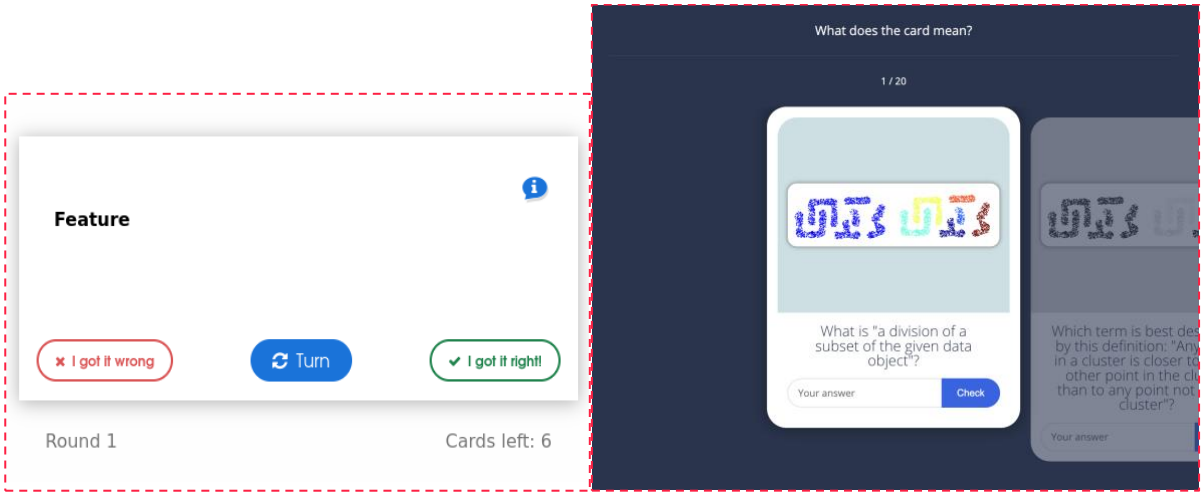
Exercises and Games

In each unit, there will be games and exercises carefully created using the materials given. Some examples are in the following images:



Crosswords

Memory



Flashcards

Answer the questions

Step 4 - Materials' review

During this phase, the materials given will be reviewed by the course creator, so please remain reachable for further notifications.

Step 5 - Final Unit

In the Final unit, there will be a final test that will determine if the student really understood all the topics. The test can be redone unlimitedly. As the teacher, you can either choose to propose **new and more complex questions** to your students or use the **questions from the previous units** and shuffle them into a new quiz.

If the first choice is made, please use your copy of Template_Questions&Feedback and write there the questions with the appropriate feedback. Write at least 10 questions. The questions can be more complex, but they should be understandable and consistent with the lessons.

In this case, the **document's solutions** won't be necessary since the students need to find a solution for themselves.

Step 6 - Last check

At the end of the course, students will have an official **certificate** (you can find an example in the Certificate_Template) and some **badges** used for interactive learning.

In this final step, you're invited to check out your course on the Moodle platform (https://sobigdata.unipi.it/).

Contacts

SoBigData RI Coordinator - Roberto Trasarti (roberto.trasarti@isti.cnr.it)
Learning Designer - Sara Lelli (sara.lelli@isti.cnr.it)