Social Mining & Big Data Ecosystem

SoBigData

RESEARCH INFRASTRUCTURE

Project Acronym	***SoBigData***
Project Title	***SoBigData Research Infrastructure*** ***Social Mining & Big Data Ecosystem***
Project Number	***654024***
Deliverable Title	***Training Activities: planning, material and reports 2***
Deliverable No.	***D4.3***
Delivery Date	***30 January 2020***
Authors	***Joanna Wright (USFD), Marco Braghieri (KCL)***

SoBigData receives funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 654024

DOCUMENT INFORMATION

PROJECT	
Project Acronym	SoBigData
Project Title	SoBigData Research Infrastructure Social Mining & Big Data Ecosystem
Project Start	1st September 2015
Project Duration	52 months
Funding	H2020-INFRAIA-2014-2015
Grant Agreement No.	654024
	DOCUMENT
Deliverable No.	D4.3
Deliverable Title	Training Activities: planning, material and reports 2
Contractual Delivery Date	December 2019
Actual Delivery Date	30 January 2020
Author(s)	Joanna Wright (USFD), Marco Braghieri (KCL), Beatrice Rapisarda (CNR)
Editor(s)	Beatrice Rapisarda (CNR), Valerio Grossi (CNR)
Reviewer(s)	Beatrice Rapisarda (CNR), Valerio Grossi (CNR)
Contributor(s)	
Work Package No.	WP4
Work Package Title	NA3_Training
Work Package Leader	KCL
Work Package Participants	(CNR), (USFD), (UNIPI), (FRH), (UT), (IMT), (LUH), (KCL), (SNS), (AALTO), (ETHZ), (TUDelft)
Dissemination	PU
Nature	Report
Version / Revision	1.2
Draft / Final	Final
Total No. Pages (including cover)	42
Keywords	Training Report, Summer School, Hackathon, Training Course, Planning Report, Training Materials, e-Learning Area

DISCLAIMER

SoBigData (654024) is a Research and Innovation Action (RIA) funded by the European Commission under the Horizon 2020 research and innovation programme.

SoBigData proposes to create the Social Mining & Big Data Ecosystem: a research infrastructure (RI) providing an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, as recorded by "big data". Building on several established national infrastructures, SoBigData will open up new research avenues in multiple research fields, including mathematics, ICT, and human, social and economic sciences, by enabling easy comparison, re-use and integration of state-of-the-art big social data, methods, and services, into new research.

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 document has been produced with the funding of the European Commission. 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.

The European Union (EU) was established in accordance with the Treaty on the European Union (Maastricht). There are currently 27 members states of the European Union. It is based on the European Communities and the member states' cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice, and the Court of Auditors (http://europa.eu.int/).

GLOSSARY

ABBREVIATION	DEFINITION
Summer School	A series of lectures and activities that takes place during higher education summer holidays, generally on specific topics
Hackathon	Event where a number of people convene together in order to engage in computer programming in a limited time frame.
Python	Python is an interpreted, interactive, object-oriented programming language. It incorporates modules, exceptions, dynamic typing, very high-level dynamic data types, and classes. It has interfaces to many system calls and libraries, as well as to various window systems, and is extensible in C or C++. It is also usable as an extension language for applications that need a programmable interface. Python runs on many Unix variants, on the Mac, and on Windows 2000 and later.
R	R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible.
GitHub	GitHub is a development platform. It allows users which range from open source to business, to host and review code, manage projects, and build software alongside in a participative environment.
Jupyter	The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text.
GATE	GATE is an open-source software which focuses on text processing and includes a desktop client for developers, a workflow-based web application, a Java library, an architecture and a process.

e-Learning Area	The e-Learning Area is the Virtual Research Environment within the SoBigData Research Infrastructure that hosts training materials developed by the project's partners
Webinar	A live or on-demand event, taking place on the internet. It can be a discussion, lecture, conference, presentation, or demonstration. Participants can see documents (usually slides) and other applications via their computer. There will also be shared audio, so you can hear the presenter

TABLE OF CONTENTS

DELIVERABLE SUMMARY

This deliverable is divided into four sections, each focusing on a different aspect of Work Package 4 (WP4) as detailed below:

The four sections covered are:

1. An overview of training activities with data
2. A description of activities undertaken to address gender and diversity issues
3. The evolution of training materials
4. An outline of responsibilities and work flow within WP4.

EXECUTIVE SUMMARY

Deliverable D4.3 provides an overview of training activities that have taken place from M36 to M52. It also includes a full overview of the training materials integration within the SoBigData Research Infrastructure.

This final deliverable aims to provide a more detailed look into the various activities that have taken place since September 2018 and in particular to provide comprehensive reports on the events that have addressed gender and diversity issues.

The whole project has sought to encourage more females into Computer Science and has therefore tailored some events to appeal to females in particular. SoBigData has also provided travel grants to assist female scientists in attending international events.

With regards to the training materials, we have sought to make these a more effective resource by concentrating on four areas of improvement: Content Harmonization, Muti-Channel Content, Interactivity and Audience Targeting. These enhancements to the SoBigData Research Infrastructure will be covered in detail in Section 4.

Discussions are also taking place with regards to the possibility of webinars, tutorials and video lectures to provide more options for the user as well as developing online assessment and exercises. The possibility of more targeted training paths is also being explored to provide a more tailored content for the user – whether they be a student, professional or the general public; the objective being to promote the dissemination of SoBigData training materials as widely as possible.

1 RELEVANCE TO SOBIGDATA

Work Package 4, entitled Training, aims to establish a joint training and education initiative on social big data within the European Research Area. 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, WP4 proposes campaigns aimed at high school students to promote interest in data science with special emphasis on gender issues.

These experiences include a number of different activities. Project-oriented summer schools and datathons have been organised in order to match research (and industrial) needs and people skills. Training courses have been developed in order to create new start-up initiatives related to big data. Moreover, activities have taken place at high school level in order to illustrate opportunities in the big data field to address gender and diversity issues in data science through training.

Moreover, this report describes the training materials that have been developed by SoBigData partners, which include interactive environments, tools, lectures and hands-on sessions.

1.1 PURPOSE OF THIS DOCUMENT

This document aims to provide an overview of all activities that have taken place, are organised and have been performed by Work Package 4. Thus, the document is divided into sections, each detailing a different aspect of training within the SoBigData Project. By beginning with a report on events, this document aims to provide a broad framework of training events that have been organised by SoBigData partners. This section is followed by a detailed description of planned training events for the reporting period. Finally, this document describes in detail all relevant activities in the development of training materials and, more specifically, the creation of the e-Learning Area within the SoBigData project Research Infrastructure.

1.2 RELEVANCE TO PROJECT OBJECTIVES

The training activity within the SoBigData project is aimed at establishing a unique, joint training and education resource centre on social big data. In order to do so, the Work Package explores and develops conventional and unconventional training experience for master and PhD students and post-doctoral trainees. These experiences include the ogranisation of a number of different events and the creation of a learning environment within the SoBigData Research Infrastructure. Among events, project-oriented summer schools and datathons have been organised and are planned in order to match research (and industrial) needs and people skills. Moreover, activities have taken place at high school level in order to illustrate opportunities in the big data field to address gender and diversity issues in data science through training. Finally, WP4, in collaboration with other Work Packages, has focused on building a common resource repository of training materials within the SoBigData Research Infrastructure, which has been named e-Learning Area WP4 - NA3 Training has four main tasks, which are:

- T4.1 Summer schools. The general aim in organising summer schools within the SoBigData project is to prepare the PhD students in science and engineering to work as data scientists. While summer schools provide an introduction on big data analysis, they are also based on a hands-on approach, in order for participants to develop skillsets using the SoBigData Research Infrastructure.

- T4.2 Training Modules for stakeholders. Task 4.2 is focused on producing open source training modules, which have been developed within training events. These training modules are then integrated in each partner's learning and teaching environments, starting from existing sources. Aside from harmonising existing training modules, T4.2 is centred on the production of new training modules for specific stakeholders that are currently under-served in the big data field, such as social scientists and humanities researchers. All training modules are developed in open educational standards in order to be easily integrated in e-learning environments. This integration activity has led to the creation of a devoted virtual research environment within the SoBigData Research Infrastructure, which has been named e-Learning Area.

- T4.3 Series of Datathons. These events aim to provide theoretical and practical advice to a group of participants in order to collect, analyse and visualise big data in order to address relevant issues. Based on interactivity, datathons aim to bring together computer programmers, storytellers, graphic designers and statisticians who are the grouped in teams and within an event aim to create innovative tools and services on based on big data analysis.

- T4.4 Addressing gender and diversity issues in data science through training. In order to address the current imbalance in the representation of all gender and minority groups, T4.4 aims to organise specific activities. This task's general objective is to leverage existing networks in order to raise awareness among targeted groups on the possibilities within the data science field.

1.3 SOBIGDATA PROJECT DESCRIPTION

The SoBigData project's aim is to create a pan-European research infrastructure. This infrastructure will integrate already established national infrastructures. To this end, the SoBigData project has defined a series of key priorities:

1. Better access to the best national research infrastructures
2. Training a new generation of mobile researchers
3. More effective national research systems
4. Optimal transnational cooperation
5. Accelerating innovation through partnerships with industry
6. Effective data, method, and knowledge sharing

1.4 RELATION TO OTHER WORKPACKAGES

Work Package 4 is part of the work packages, which will form a community comprising excellence centres, other academic and industrial users and training activities aimed at data scientists.

In particular it is related to:

WP2: as training activities fall under the legal and ethical framework of the SoBigData infrastructure

WP3: as training activities have a strong relationship with the dissemination and impact strategies developed for the whole SoBigData project

WP5: as training activities are connected to the innovation activities aimed at industry and other stakeholders

WP7: as the work package devoted to Virtual Access has developed an environment within the SoBigData Research Infrastructure devoted to training materials, which has been named e-Learning Area

WP10: as training activities will present and educate on the new methodologies and technologies that SoBigData is developing

WP11: as training activities relate to the construction of a benchmarking and evaluation framework for big data analytics and social mining methods.

1.5 STRUCTURE OF THE DOCUMENT

This deliverable is divided into four sections.

1. The first section details the training activities carried out in WP4, providing an overview on events and data on participants.
2. The second describes the activities undertaken to address gender and diversity issues in data science through training.
3. The third part assesses the evolution of training materials production and integration within the e-Learning Area of the SoBigData Research Infrastructure.
4. The final section outlines the responsibilities and work flow within WP4.

2 TRAINING REPORT

2.1 INTRODUCTION

This section reports on the main activities that have taken place in WP4. There were 19 events organised during this period (from September 2018 to December 2019) and these included workshops, datathons, summer schools, tutorials, conferences and hackathons. The total number of participants recorded in these activities was 673.

Organisations provided gender information on 593 participants, of whom 61% were male and 39% were female. SoBigData offered specific support to female participants, in order to encourage gender balance. Details of this support (where granted) is detailed within the individual event summary.

They also provided age information on 547 participants, of whom 28% were of high school age, 33% were young adults and 39% were adults.

These activities and events were conducted in a wide range of countries, including the UK, Ireland, Italy, Greece, France, USA, Switzerland and Germany and reached participants from all over the globe.

2.2 TIMELINE OF EVENTS

This section reports on training activities that have been carried out within this work package until M52 (December 2019) of the project and includes a timeline of the activities.

2018						
	M37				M38	
Date	10 September 2018	10-14 September 2018	23-28 September 2018	27 September – 3 October 2018	1-4 October 2018	12-13 October 2018
Event	**PAP 2018: Personal Analytics and Privacy**	**KNOWMe: 2nd International Workshop on Knowledge Discovery from Mobility and Transportation Systems**	**SoBigdata@CCS18 Thessaloniki**	**Robotics Festival: Football of the Future, the future of football: between devices and algorithms**	**IEEE DSAA 2018 The 5th IEEE International Conference on Data Science and Advanced Analytics**	**Soccer data challenge**

Type	Workshop	Workshop	Conference	Festival	Conference	Hackathon
Location	Dublin, Ireland	Dublin, Ireland	Thessaloniki, Greece	Pisa, Italy	Turin, Italy	Pisa, Italy

Table 1: *Timeline of events in year 2018 (M37-M38)*

2019							
	M44	**M45**		**M46**			
Date	29 April	29 April - 2 May	5-7 May	4-5 June	5-7 June	9-12 June	25-28 June
Event	**CAOS - Communications and Networking Aspects of Online Social Networks**	**INFOCOM**	**2° Soccer Data Cup (Women only)**	**From Game Theory to Computational Social Science and Beyond**	**Futura L'Aquila - Soccer data challenge**	**WoWMoM**	**Summer School on Analysing Dis-information**
Type	Workshop	Conference	Datathon	Workshop	Datathon	Conference	Summer School
Location	Paris, France	Paris, France	Italy	Zurich, Switzerland	Italy	Washington DC, USA	London, UK

2019						
	M47		**M49**			**M50**
Date	8-9 July	8-12 July	2-6 September	16-18 September	20 September	10-11 October
Event	**ESME 2019 - workshop about, ethics, privacy and explainable**	**Headstart Summer School**	**Data Summer School**	**womENcourage 2019 "Diversity Drives Societal Change", ACM**	**ECML XKDD - eXplainable Knowledge Discovery in**	**Soccer Data Challenge – 2nd Edition**

	AI			**Celebration of Women in Computing**	**Data Mining**	
Type	Workshop	Summer School	Summer School	Conference	Workshop and tutorial	Datathon
Location	Pisa, Italy	Sheffield, UK	Pisa, Italy	Rome, Italy	Wurzburg, Germany	Trento, Italy

Table 2: *Timeline of events in year 2019*

2.3 PERIODIC TRAINING PLANNING REPORT DATA

In order to better assess the impact of training activities that have taken place within WP4 before M52 (December 2019) data on participants has been provided by the organisers of some of the activities. The total number of participants of all 12 activities where data was recorded is 673 with largest participation recorded for the Robotics Festival: Football of the Future, the future of football: between devices and algorithms, in Pisa, Italy with 120 participants.

EVENT	**PARTICPANTS**
PAP 2018: Personal Analytics and Privacy	30
KNOWMe: 2nd International Workshop on Knowledge Discovery from Mobility and Transportation Systems	21
SoBigdata@CCS18 Thessaloniki	
Robotics Festival: Football of the Future, the future of football: between devices and algorithms	120
IEEE DSAA 2018	
Soccer data challenge	95
CAOS - Communications and Networking Aspects of Online Social Networks	25

INFOCOM	
2° Soccer Data Cup (Women only)	55
Futura L'Aquila - Soccer data challenge	55
WoWMoM	
Summer School on Analysing Disinformation	23
ESME 2019 - workshop about, ethics, privacy and explainable AI	
Headstart Summer School	42
womENcourage 2019 – “Diversity Drives Societal Change”	80
ECML XKDD	50
Soccer Data Challenge – 2nd Edition	
DSSS2019 - Data Science Summer School	77
From Game Theory to Computational Social Science and Beyond	
Total	673

Table 3: *Table of Events listing Participant Numbers*

As in previous deliverables, some organisers provided an analysis of the participants’ age groups for each event. Three main age groups were identified: high school, young adults and adult. The majority of the participants in this period were adults, although the Soccer Data Challenge and the Data Summer School attracted both young adults and adults and the Robotics Festival attracted high school age children and young adults.

Three events were directed especially towards the high school age group - 2° Soccer Data Cup (Women only), Futura L'Aquila - Soccer data challenge and the Headstart Summer School. The 2° Soccer Data Cup was specifically for women only and had an equal number of participants to the other Soccer Data Challenge.

Another event that was specifically for women was the WomENcourange 2019. Eighty women participated in this event. A more detailed description of this event will be included in the next section.

Event	High School	Young Adults	Adults	Unspecified
PAP 2018: Personal Analytics and Privacy				30
KNOWMe: 2nd International Workshop on Knowledge Discovery from Mobility and Transportation Systems				21
SoBigdata@CCS18 Thessaloniki				
Robotics Festival: Football of the Future, the future of football: between devices and algorithms		120		
IEEE DSAA 2018				
Soccer data challenge			95	
CAOS - Communications and Networking Aspects of Online Social Networks				25
INFOCOM				
2° Soccer Data Cup (Women only)	55			
Futura L'Aquila - Soccer data challenge	55			

WoWMoM				
Summer School on Analysing Disinformation			23	
ESME 2019 - workshop about, ethics, privacy and explainable AI				
Headstart Summer School	42			
womENcourage 2019 – “Diversity Drives Societal Change”			80	
ECML XKDD				50
From Game Theory to Computational Social Science and Beyond				
Soccer Data Challenge – 2nd Edition				
DSSS2019 - Data Science Summer School		60	17	
Total	152	180	215	126

Table 4: *Table of Events listing Participant Age Brackets*

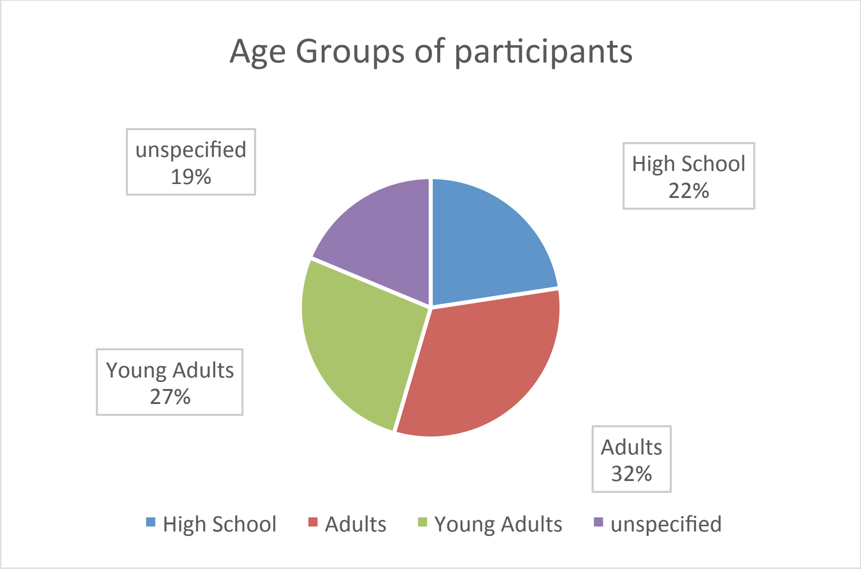


Figure 1: ***Age groups of participants***

Organisers also provided data insight with respect to gender diversity of the events' participants. Out of the 673 participants, data was gender specific about 593 participants, whereas 80 participants were not specified. Out of the 593 participants, for whom gender data was provided, there were 360 males and 233 females. In those events where gender for participants was detailed, males were the majority in all events except for events dedicated to women – the 2nd Soccer Data Cup and the womENcourage 2019.

SoBigData provided support aimed specifically at female scientists. In particular, the DSAA 2018 Conference provided support for 4 female participants by means of a Grant. The fee was also waived for 6 of the 11 females who attended the Summer School on Analysing Disinformation.

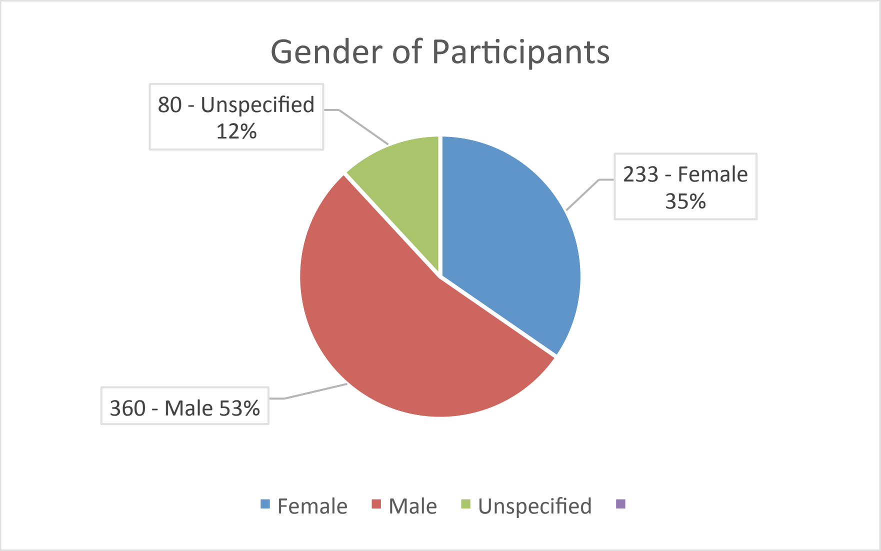


Figure 2: ***Gender of participants***

EVENT	Male	Female	Unspecified
PAP 2018: Personal Analytics and Privacy	0	0	30
KNOWMe: 2nd International Workshop on Knowledge Discovery from Mobility and Transportation Systems	13	8	0
SoBigdata@CCS18 Thessaloniki			
Robotics Festival: Football of the Future, the future of football: between devices and algorithms	90	30	0
IEEE DSAA 2018			
Soccer data challenge	80	15	0
CAOS - Communications and Networking Aspects of Online Social Networks	20	5	0
INFOCOM			
2° Soccer Data Cup (Women only)	50	5	0
Futura L'Aquila - Soccer data challenge	5	50	0
WoWMoM			
Headstart Summer School	28	14	0
Summer School on Analysing Disinformation	12	11	0
ESME 2019 - workshop about, ethics, privacy and			

explainable AI			
womENcourage 2019 – "Diversity Drives Societal Change"	5	75	0
ECML XKDD	0	0	50
From Game Theory to Computational Social Science and Beyond			
Soccer Data Challenge – 2nd Edition			
DSSS2019 - Data Science Summer School	57	20	0
Total	360	233	80

Table 5: ***Table of Events listing Participant Genders***

2.4 DETAILED REPORT ON SELECTED TRAINING ACTIVITIES

This section reports a selection of Summer Schools, Workshops and Datathons organized or co-organized by the SoBigData consortium.

2.4.1 SUMMER SCHOOLS

2.4.1.1 SOBIGDATA SUMMER SCHOOL - COMPUTATIONAL MISINFORMATION ANALYSIS 2019

During WP4, the UK hosted a GATE Summer School – on Computational Misinformation Analysis. This event took place at King's College, London on the 25-28 June 2019.

OBJECTIVES

The aim of this summer school was, firstly, to set out the state-of-the-art challenges in computational misinformation analysis, followed by lectures and hands-on practical sessions on relevant methods, tools, and datasets.

PARTICIPANTS AND SCHEDULE

The event totalled 23 participants; a mixture of NLP (natural language processing), machine learning and data science researchers, PhD students and journalists.

Figure 3: Participants to the Computational Misinformation Analysis Summer School

As with previous GATE Summer Schools, the gender split of attendees was equal, with there being 12 male and 11 female participants.

The event took place over 4 days and was structured with different sessions and talks. The topics covered were disinformation, online abuse, fake news and why it matters. Collecting and analysing data from social media and looking at the ethical and legal issues surrounding the data collection.

There was a focus on online abuse directed at politicians particularly during election periods.

The event was organised around both lectures and hands-on laboratory classes. The teaching syllabus was designed to cover all aspects of social media analysis, opinion and rumour analysis, bot detection, echo chambers, as well as algorithmic biases.

C. Scarton posted a blog on the WeVerify website https://weverify.eu/news-and-events/london-summer-school-on-computational-misinformation-analysis/ on July 24th, 2019, stating, "*Overall, it was a great opportunity to learn from each other about misinformation, Given its diverse audience, the lectures were easy to follow, being a good introduction to the topic by non-experts. This type of summer school is ideal to bring together researchers from multiple areas, and start interdisciplinary collaborations*".

Among the speakers, three WeVerify researchers gave lectures at the summer school: Professor Kalina Bontcheva (USFD), Denis Teyssou (AFP) and Alex Alaphilippe (EU DisInfo Lab).

WEBSITE

The summer school websites include all materials used during training and hands-on sessions, as well as details of the course structure.

https://gate-socmedia.group.shef.ac.uk/summer-school-comp-misinfo-analysis-2019/

2.4.1.2 GATE SUMMER SCHOOL

Sheffield hosted its 12th GATE Summer School on 17-21 June 2019. There were 10 participants, all either students, or from an Academic background. There were 7 male and 3 females.

OBJECTIVES

The aim of this summer school was, firstly, to set out the state-of-the-art challenges in computational misinformation analysis, followed by lectures and hands-on practical sessions on relevant methods, tools, and datasets.

PARTICIPANTS AND SCHEDULE

This was a 5 days course where the focus was on mining text and social media content using GATE software. Many of the hands-on exercises were focused on analysing news articles, tweets, and other textual content.

Module 1 was an introduction on the basics of information extraction with GATE, the course went on to Corpus Annotation and Evaluation and writing information extraction patterns with JAPE,

Module 2 demonstrated the use of GATE for social media analysis, an introduction to Twitter and the JSON structure, Language identification, tokenisation for Twitter and POS tagging and Information Extraction for Twitter.

Module 3 covered Crowdsourcing, GATE Cloud/MIMIR (how to index and search semantically annotated social media), and Machine Learning (Training Machine Learning Models for IE in GATE).

Module 4 was the Advanced IE and Opinion Mining in GATE, advanced information extraction, useful GATE components (plugins) and opinion mining components and applications in GATE.

On the final day there was a choice of modules – participants could choose GATE for developers (writing your own plugin and GATE in production – multi-threading, web applications etc.), or they could choose GATE Applications - building your own applications and looking at examples of some current GATE applications: social media summarisation, visualisation, Linked Open Data for IE, and more.

WEBSITE

https://gate.ac.uk/conferences/fig/fig12.html

2.4.1.3 DATA SCIENCE SUMMER SCHOOL – DSSS2019

This Summer School took place on 2-6 September 2019 in Pisa, Italy. Data Science is emerging as a disruptive consequence of the digital revolution. It is based on the combination of big data availability, sophisticated data analysis techniques, and scalable computing infrastructures. Data Science is rapidly changing the way we do business, socialise, conduct research, and govern society. It is also changing the way scientific research is performed. Model-driven approaches are supplemented with data-driven approaches. A new paradigm emerged, where theories and models and the bottom up discovery of knowledge from data mutually support each other.

OBJECTIVE

Given the interdisciplinary nature of Data Science this summer school offers lectures by high-level scholars from different domains, giving to the students the skills to exploit data and models for advancing knowledge in different disciplines, or across diverse disciplines (e.g. biology, economics, medicine, etc).

PARTICIPANTS AND SCHEDULE

A total of 77 individuals took part in this summer school, 57 males and 20 females. There were 68 from Academia and 9 from a professional background.

The main topics of the summer school are related to big data analytics, i.e., extraction of knowledge from big data, machine learning, i.e., providing an overview of the main techniques used to automatically learn and improve from experience, and complex systems, i.e., methods and technologies particularly related to network science. Moreover, lectures will highlight the ethical implications that data science could lead and the countermeasures that each data scientist can apply to perform analysis with respect to the individuals involved in the data.

Figure 4: The Data Science Summer DSSS19 School in Pisa

WEBSITE

https://datasciencephd.eu/dsss19/

2.4.1.4 HEADSTART SUMMER SCHOOL

In collaboration with Headstart (a charitable trust that provides hands-on science, engineering and maths taster courses), the Department of Computer Science at the University of Sheffield ran its fourth annual summer school for maths and science A-level students. This residential course ran from 8 to 12 July 2019 in Sheffield and included practical work in computer programming, Lego robots, and project development as well as tours of the campus and talks about the industry.

OBJECTIVE

The aim of the project was to introduce a younger audience to natural language processing using GATE Developer and a special GATE plugin (which uses the ShefRobot library available from GitHub) that allows

JAPE rules to operate Lego robots. The students were then able to demonstrate their knowledge by controlling the Lego robots by sending instructions as Tweets.

PARTICIPANTS AND SCHEDULE

This year 42 students (14 female and 28 male) from all over the UK attended the Computer Science Headstart Summer School.

The students were provided with a starter GATE application (essentially the same as in last year's course) containing just enough gazetteer entries, JAPE, and sample code to let them tweet variations like "turn left" and "take a left" to make the robot do just that. The GATE Cloud Twitter Collector was also used, which had been modified to run locally so the students can set it up on a lab computer so it followed their own twitter accounts and processes their tweets through the GATE application, sending commands to the robots when the JAPE rules match.

Based on lessons learned from the previous years, more effort was put into improving the instructions and the Twitter Collector software to help them get it running faster. This time the first robot started moving under GATE's control less than 40 minutes from the start of the presentation, and the students rapidly progressed with the development of additional rules and then tweeting commands to their robots.

The structure and broader coverage of this year's course meant that the students had more resources available and a more open project assignment, so not all of them chose to use GATE in their projects, but it was much easier and more streamlined for them to use than in previous years.

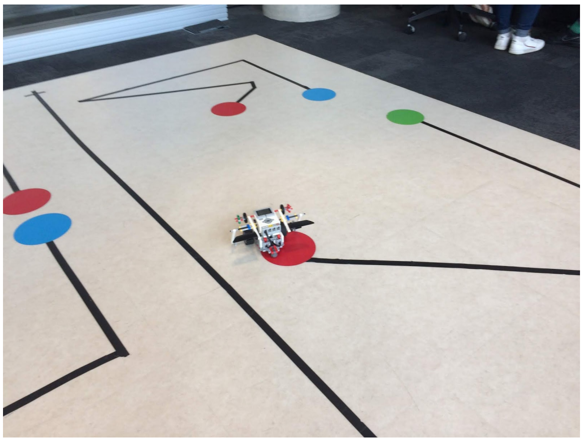


Figure 5: The Lego robot at Headstart Summer School

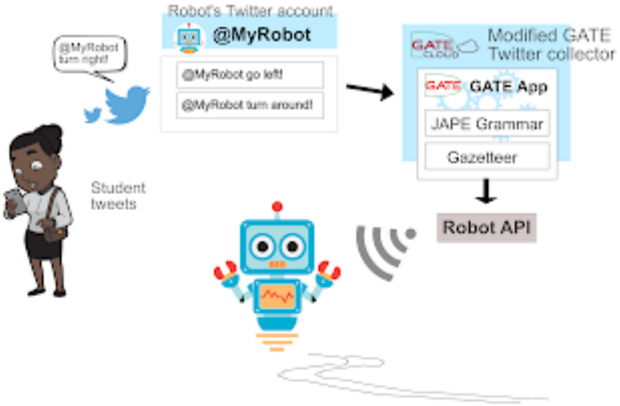


Figure 6: The communication flow between the student and the robot

WEBSITE

The handout and slides are publicly available from the GATE website, which also hosts GATE Developer and other software products in the GATE family. Source code is available from our GitHub site.

Handout: https://gate.ac.uk/sale/talks/headstart-2019/handout.pdf

Slides: https://gate.ac.uk/sale/talks/headstart-2019/presentation.pdf

GitHub site: https://github.com/GateNLP

2.4.2 WORKSHOPS

During WP4, SoBigData have arranged 5 workshops which have been held in various countries including Ireland, France, Italy, Switzerland and Germany.

These include; PAP 2018: Personal Analytics and Privacy; KNOWMe: 2nd International Workshop on Knowledge Discovery from Mobility and Transportation Systems; Esme 2019; From Game Theory to Computational Social Science and Beyond; ECML XKDD.

2.4.2.1 PAP 2018 – PERSONAL ANALYTICS AND PRIVACY

In the era of Big Data, every single user of our hyper-connected world leaves behind a myriad of digital breadcrumbs while performing daily activities. This enormous amount of personal data can be exploited to improve the lifestyle of each individual. Up to now, the highly valuable personal patterns able to predict human behaviour can only be extracted by big companies, which employ this information mainly to improve marketing strategies. Users have a very limited capability to control and exploit their own personal data. Currently there is still a significant lack in terms of algorithms and models specifically designed to capture the knowledge from individual data and to ensure privacy protection in a user-centric scenario.

OBJECTIVES

The purpose of this workshop is to encourage principled research that will lead to the advancement of personal data analytics, personal services development, privacy, data protection and privacy risk assessment. The workshop will address important issues related to personal analytics, personal data mining and privacy in the context where real individual data (spatio-temporal data, call details records, tweets, mobility data, social networking data, etc.) are used for developing a data-driven service, for realising a social study aimed at understanding nowadays society, and for publication purposes.

PARTICIPANTS AND SCHEDULE

There were 30 individuals who attended this Workshop – although no data was recorded relating to their gender or age.

Personal data analytics and individual privacy protection are the key elements to leverage nowadays services to a new type of systems. The availability of personal analytics tools able to extract hidden knowledge from individual data while protecting the privacy right can help the society to move from organization-centric systems to user-centric systems, where the user is the owner of her personal data and is able to manage, understand, exploit, control and share her own data and the knowledge deliverable from them in a completely safe way.

WEBSITE

http://kdd.di.unito.it/pap2018/

2.4.2.2 KNOWME: 2ND INTERNATIONAL WORKSHOP ON KNOWLEDGE DISCOVERY FROM MOBILITY AND TRANSPORTATION SYSTEMS

The recent technological advances on telecommunications create a new reality on mobility sensing. Nowadays, we live in an era where ubiquitous digital devices are able to broadcast rich information about human mobility in real-time and at a high rate. Such fact exponentially increased the availability of large-scale mobility data which has been popularized in the media as the new currency, fueling the future vision of our smart cities that will transform our lives. The reality is that we just began to recognize significant research challenges across a spectrum of topics. Consequently, there is an increasing interest among different research communities (ranging from civil engineering to computer science) and industrial stakeholders on build knowledge discovery pipelines over such data sources. However, such availability also raises privacy issues that must be considered by both industrial and academic stakeholders on using these resources.

OBJECTIVES

This workshop intends to be a top-quality venue to bring together transdisciplinary researchers and practitioners working in the related topics from multiple areas such as Data Mining, Machine Learning, Numerical Optimization, Public Transport, Traffic Engineering, Multi-Agent Systems, Human-Computer Interaction and Telecommunications, among others. The ultimate goal of this venue is to evaluate not only the theoretical contribution of the methodology proposed in each research work, but also its potential deployment/impact as well as its advances with respect to the State-of-the-Art/State-of-the-Practice in the domains of the related applications.

WEBSITE

https://kdd.isti.cnr.it/knowme.eu.2018/

2.4.2.3 ESME 2019

Big data analytics and social mining raise a number of ethical issues, especially as companies begin monetizing their data externally for purposes different from those for which the data was initially collected. The scale and ease with which analytics can be conducted today completely change the ethical framework. We can now do things that were impossible a few years ago, and existing ethical and legal frameworks cannot prescribe what we should do. Artificial Intelligence is becoming a disruptive technology, and resources for innovation are currently dominated by giant tech companies. To ensure European independence and leadership, we must invest wisely by bundling, connecting, and opening our AI resources having in mind ethical priorities such as transparency and fairness.

OBJECTIVES

The PRO-RES workshop about Ethics, Social Mining, and Explainable artificial intelligence (ESME) was held in Pisa, on July 8th and 9th 2019 aiming to discuss the main open questions regarding privacy, explainability, and other ethical concerns. The ESME workshop was organized within the PRO-RES (PROmoting ethics and integrity in non-medical RESearch) EU project, with the support of SoBigData and AI4EU projects.

PARTICIPANTS AND SCHEDULE

54 persons actively participated in the event; they are distinguished experts in the field, and they belong to 26 different institutions among 12 countries in total.

The invited experts are from both academia and industry, and they have very different expertise and background (ranging from computer science to law, from moral philosophy to sociology), in order to offer several different perspectives and to cover as many points of views as possible during the discussions.

The presentations (held in the first morning by 5 researchers, 2 female and 3 male) touched a variety of topics, such as trust in research, interdisciplinary differences as both resource and obstacle to the communication, positive-sum granting both ethics and utility, ethics-by-design, data life cycle, human-machine interaction, and moral machine experiments.

After that, participants openly discussed about ethical dilemmas, such as promoting ethics, reconciling ethical research and industry objectives, and analyzing the true applicability of the right to the explanation, starting from the not trivial problem to define what a good explanation is and how to measure it.

The staff was composed of 14 researchers in total, composing the scientific committee, rapporteurs, and local committee. These researchers (7 male and 7 female) are members of ISTI – CNR, University of Pisa and Scuola Normale Superiore, and they are researchers of several career stages, from Ph.D. students to full professors.

WEBSITE

https://kdd.isti.cnr.it/esme2019/

2.4.2.4 FROM GAME THEORY TO COMPUTATIONAL SOCIAL SCIENCE AND BEYOND

Jointly with TU Delft’s PhD Program “Engineering Social Technologies for a Responsible Digital Future” and supported by the SoBigData project, Dirk Helbing and the Professorship of Computational Social Science (COSS) run this workshop in order to bring together various strands of social sciences (economics, sociology, etc.), complex systems, network science, game theory, socio- physics and data science, to showcase recent scientific advances, and to identify synergies for collaboration.

2.4.2.5 ECML PKDD

The purpose of AIMLAI-XKDD (Advances in Interpretable Machine Learning and Artificial Intelligence & eXplainable Knowledge Discovery in Data Mining), was to encourage principled research that leads to the advancement of explainable, transparent, ethical and fair data mining, machine learning, artificial intelligence.

2.4.3 DATATHONS

2.4.3.1 SOCCER DATA CHALLENGE

The Soccer data challenge is a competition promoted by SoBigData with the sponsorship of FIGC and open to all data and soccer enthusiasts.

OBJECTIVES

The Soccer Data Challenge is an analytical marathon on football. It is designed to bring together individuals with as passion data and football. The participating teams have 30 hours to solve an analytical problem linked to football, using the largest dataset of game events ever released before.

PARTICIPANTS AND SCHEDULE

The event took part in October 2018 and again in 2019. The 2018 event attracted 95 enthusiasts – of these 84% were male and 16% were female. The finalists presented their work to a jury of professionals from the world of football and BigData. The winning team earned a prize of € 5,000.

Figure 7: A group of students at the Soccer Data Challenge

WEBSITE

https://sobigdata-soccerchallenge.it/

3 ADDRESSING GENDER AND DIVERSITY ISSUES IN DATA SCIENCE THROUGH TRAINING

3.1 INTRODUCTION

Two events in particular were created to welcome women into the realm of Data Science – the womEncourage 2019 Conference and the 2nd Soccer Data Cup which both took place in Italy.

The womEncourage had 75 female participants and 5 males. The 2nd Soccer Data Cup had 50 female and 5 male participants.

3.1.1 WOMENCOURAGE

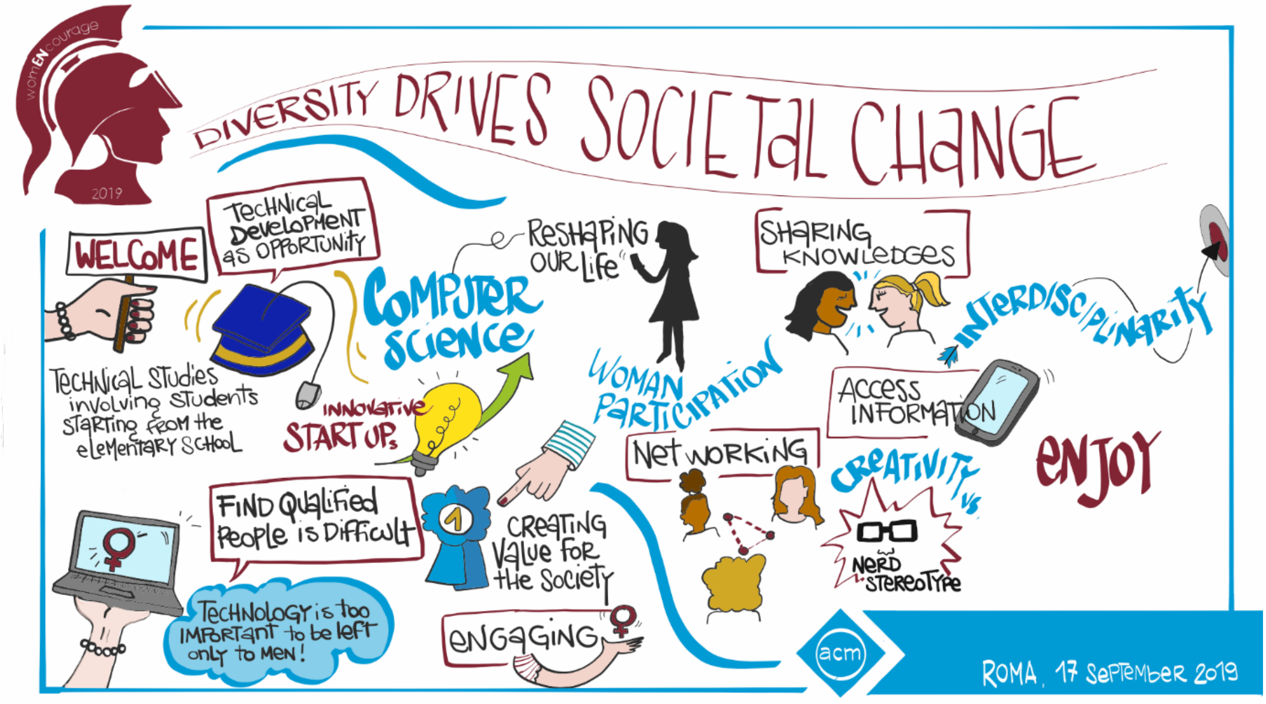

womENcourage
2019
DIVERSITY DRIVES SOCIETAL CHANGE
WELCOME
TECHNICAL DEVELOPMENT AS OPPORTUNITY
TECHNICAL STUDIES INVOLVING STUDENTS STARTING FROM THE ELEMENTARY SCHOOL
COMPUTER SCIENCE
INNOVATIVE START UPS
RESHAPING OUR LIFE
WOMAN PARTICIPATION
SHARING KNOWLEDGES
INTERDISCIPLINARITY
ACCESS INFORMATION
NETWORKING
CREATIVITY VS NERD STEREOTYPE
ENJOY
FIND QUALIFIED PEOPLE IS DIFFICULT
TECHNOLOGY IS TOO IMPORTANT TO BE LEFT ONLY TO MEN!
CREATING VALUE FOR THE SOCIETY
ENGAGING
acm
ROMA, 17 SEPTEMBER 2019

Figure 8: The conference poster

During WP4 womENcourage 2019 – “Diversity Drives Societal Change” took place in Rome, Italy on 16-18 September 2019. This is the 6th year this conference has run and is aimed at connecting women from diverse technical disciplines and encouraging them pursue their education and profession in computing.

OBJECTIVES

WomENcourage brings together women in the computing profession and related technical fields to exchange knowledge and experience and provide special support for women who are pursuing their academic degrees and starting their careers in computing.

PARTICIPANTS AND SCHEDULE

WomENcourage is organized by ACM-W Europe and aims to bring together undergraduate and graduate students, researchers, academics and practitioners to present and share their achievements and experiences.

Through a programme packed with insightful topics and engaging educational and networking activities, womENcourage provides a unique experience of the collective energy, drive, and excellence that professional women share to support each other. The three days event included a Hackathon, Workshops, Posters, Technical Talks, Panel Discussions and Interdisciplinary Research Tracks.

This event welcomed 80 participants, 75 of whom were female. They were aged between 25 and 55 years old. Four males attended who were all in the 40-55 years age bracket.

Figure 9: The participants to WomENcourage

There were four keynote speakers, Francesca Rossi, Sihem Amer-Yahia, Donatella Sciuto and Danielle (Sparky) VanDyke among many other speakers, experts and academics in the field of computer science.

Figure 10: Fosca Giannotti giving a talk on 'Data Science & ETHICS'

This was a discussion on the urgent open challenge of how to construct meaningful explanations of black box and opaque AI/ML systems.

WEBSITE

The website link provides a full list of the speakers, Tech Talks, workshops, panels and the timetable of events.

https://womencourage.acm.org/2019/

3.1.2 2ND SOCCER DATA CUP

The Soccer Data Cup is an innovative initiative, unique of its kind, organized by the National Research Council of Italy (CNR) and University of Pisa, two partners of the SoBigData consortium, in collaboration with the Italian Ministry of Education (MIUR). The second edition was organized at L'Aquila (Italy) on 5-7 May 2019 and was totally dedicated to women**.**

OBJECTIVES

The event has been created to encourage young women into STEM subjects and inspire them to look into a future in the Computer Science field. It is a veritable sports marathon and digital co-design where female students and representatives with their respective teams, with the help of Mentors and experts of Sport Analytics, will confront each other through innovative tools and methodologies. It is a fun event with a competitive element.

PARTICIPANTS AND SCHEDULE

This event was attended by 50 female participants of high school age (16-19) and there were also 5 male participants also of high school age. This event is free so all schools are offered the opportunity to become involved.

Daniele Fadda & Luca Pappalardo who helped organise the event wrote a blog which details how the data challenge is structured:

In a first selection phase, six high schools were selected to represent six Italian regions: Abruzzo, Campania, Emilia-Romagna, Molise, Puglia, Umbria. Each high school then selected two teams: ***a team of seven female soccer players and a team of two female "wannabe" data scientists****. Then, the competition was articulated in two parallel, but connected, competitions.*

In the sports competition, the six teams of players faced each other in a futsal tournament: two Italian-style groups of three teams, semi-finals and finals. During all matches, the players wore a device, produced by company Tracking4Fun, which monitored in great detail the movements on the field. At the same time, matches were filmed so as to allow company Wyscout to detect, with the usage of its proprietary software, all the main spatiotemporal events that have occurred during the matches (passes, tackles, shots, fouls, etc.).

In parallel, the teams of data scientists analyzed the data produced during the matches, with the purpose of creating a ***critical analysis*** *of the tournament.* ***With the help of data scientists of the SoBigData infrastructure****, the data scientists coded in* ***Python*** *for three days, finally presenting their analyses in the last day to a committee of sports and science experts. The outcome of both the sports competition and the analytical one were used to decide the three finalists (Emilia-Romagna, Puglia and Umbria) that presented their analyses at the municipal theatre. The audience at the theatre used an online app to vote for the best presentation, finally decreeing* ***Puglia as the winner of the Soccer Data Cup****.*

The second edition of the Soccer Data Cup was a big success, suggesting a ***bright future for women both in soccer and the STEM disciplines****. The young data scientists showed an enormous enthusiasm for data science and sports analytics as well as a great talent for science communication. This motivated SoBigData to organize a* ***third edition*** *where both men and women will attend in mixed teams, with the purpose of promoting the sound values of sports, STEM disciplines, and gender equality.*

Figure 11: Students during the 2nd Soccer Data Cup

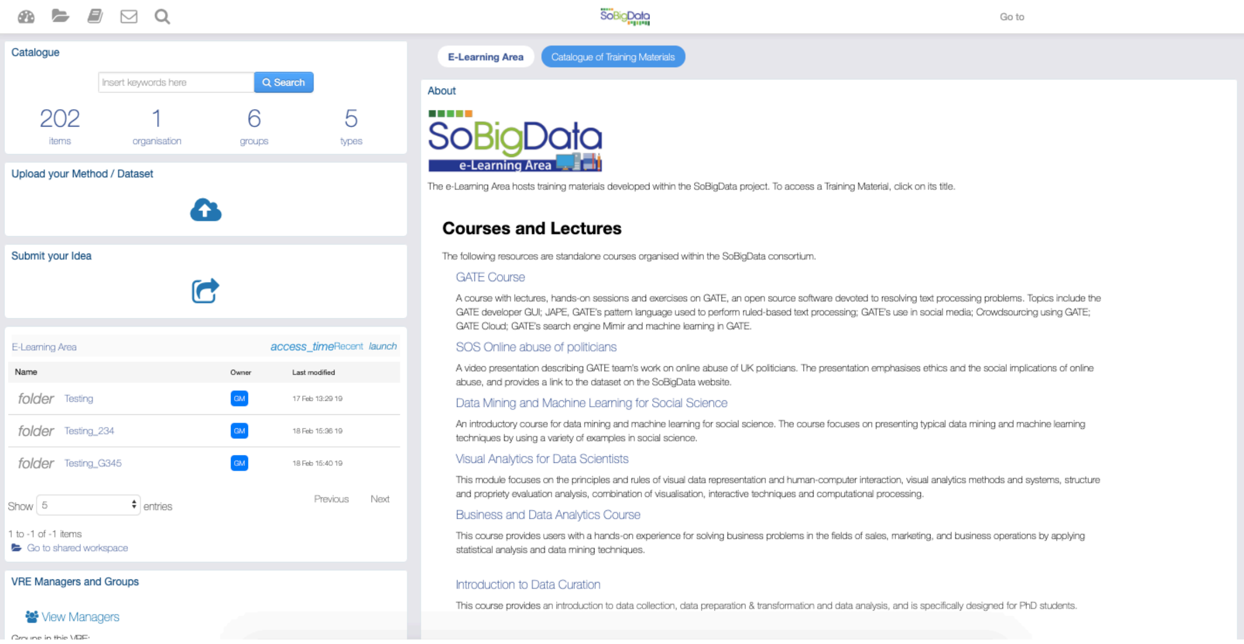
WEBSITE

http://www.sobigdata.eu/blog/when-soccer-and-science-goes-female-2nd-soccer-data-cup

4 TRAINING MATERIALS

4.1 THE SOBIGDATA RESEARCH INFRASTRUCTURE E-LEARNING AREA

In Section 3 of Deliverable D4.2, which was delivered in September 2018, we focused on the creation and integration of Training Materials into the SoBigData Research Infrastructure. After surveying existing materials, we described the creation of the e-Learning Area within the SoBigData RI, which was done by integrating training materials directly into the SoBigData Catalogue, in order to obtain the maximum harmonization and user-friendliness approach.



Name	Owner	Last modified
folder Testing	GM	17 Feb 13:29 19
folder Testing_234	GM	18 Feb 15:36 19
folder Testing_G345	GM	18 Feb 15:40 19

The e-Learning Area hosts training materials developed within the SoBigData project. To access a Training Material, click on its title.

Courses and Lectures

The following resources are standalone courses organised within the SoBigData consortium.

GATE Course

A course with lectures, hands-on sessions and exercises on GATE, an open source software devoted to resolving text processing problems. Topics include the GATE developer GUI; JAPE, GATE's pattern language used to perform ruled-based text processing; GATE's use in social media; Crowdsourcing using GATE; GATE Cloud; GATE's search engine Mimir and machine learning in GATE.

SOS Online abuse of politicians

A video presentation describing GATE team's work on online abuse of UK politicians. The presentation emphasises ethics and the social implications of online abuse, and provides a link to the dataset on the SoBigData website.

Data Mining and Machine Learning for Social Science

An introductory course for data mining and machine learning for social science. The course focuses on presenting typical data mining and machine learning techniques by using a variety of examples in social science.

Visual Analytics for Data Scientists

This module focuses on the principles and rules of visual data representation and human-computer interaction, visual analytics methods and systems, structure and propriety evaluation analysis, combination of visualisation, interactive techniques and computational processing.

Business and Data Analytics Course

This course provides users with a hands-on experience for solving business problems in the fields of sales, marketing, and business operations by applying statistical analysis and data mining techniques.

Introduction to Data Curation

This course provides an introduction to data collection, data preparation & transformation and data analysis, and is specifically designed for PhD students.

Figure 12: The e-Learning Area within the SoBigData RI

Moreover, Training Materials are now browsable directly from the project's website, in order to grant better and easier access to the website's visitors, without need to register or login into the SoBigData Research Infrastructure. As stated in D4.2 Training Activities planning material and reports 1 (Sept. 2018), the total number of training materials that have been uploaded into the SoBigdata RI are

> *'21 training materials uploaded into the SoBigData catalogue. A total of 157 different files was uploaded into the SoBigData Workspace, which is part of the project's Research Infrastructure and, where possible, a description of each training material item was inserted'*

Thus, there have been two main activities that have taken place since the last reporting period. The first was an overall assessment of the user response to the creation of the e-Learning Area within the SoBigData Research Infrastructure, while the second was an evaluation of the uploaded material in relation to the Grant Agreement Document 654024, which tasked WP4 with the creation of 'a joint training and education resource repository on big social data in the European research area'.

4.2 VISITORS AND USERS OF SOBIGDATA RESEARCH INFRASTRUCTURE E-LEARNING AREA

Having created the resource repository, we are now able to present data originated from the accesses and number of users which have registered to this specific part of the SoBigData Research Infrastructure. The data regards the period between June 2018 to November 2019 (the last full available month).

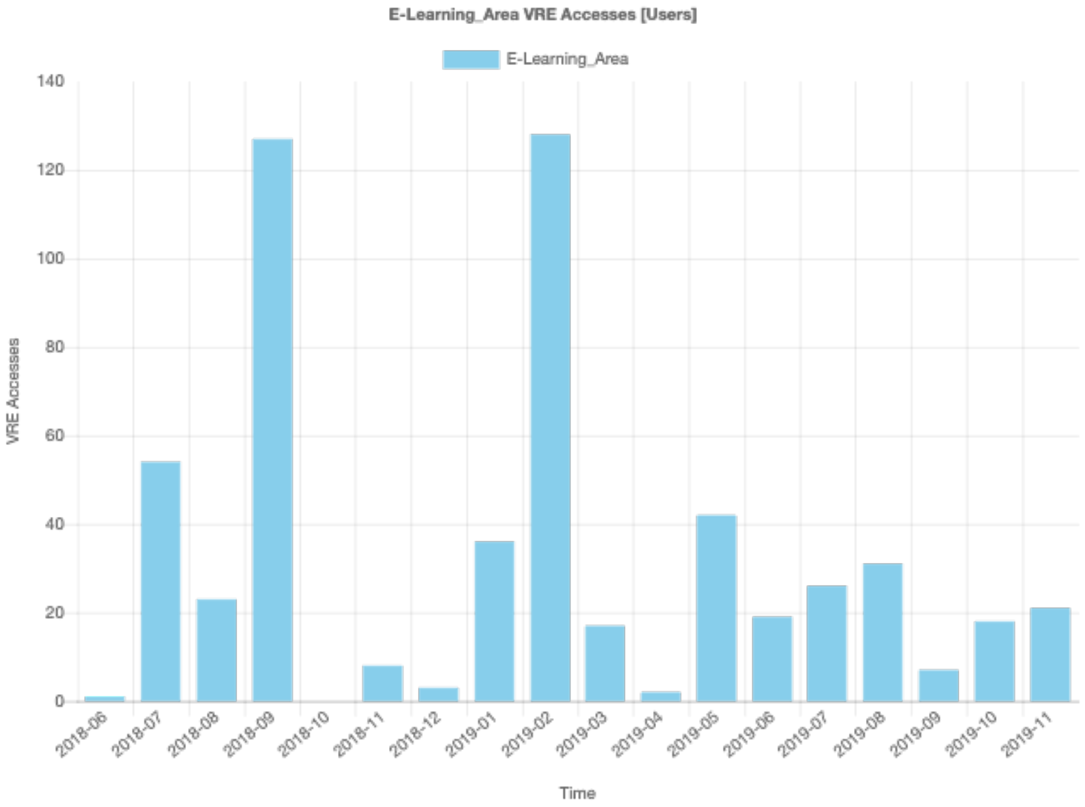


Figure 13: The graph shows the number of accesses to the e-Learning Area of the SoBigData RI

Whereas it is evident that some months have registered a more intense activity (namely September 2018 and February 2019), over the 18 months period there have been a total of 563 accesses, which in average are around 31 per month. Thanks to the SoBigData Research Infrastructure, we are also able to determine the number of users which have registered to this specific part of the RI. Among the same 18 months period, the number of users has risen from the initial 10 to 80, as in Figure 3.

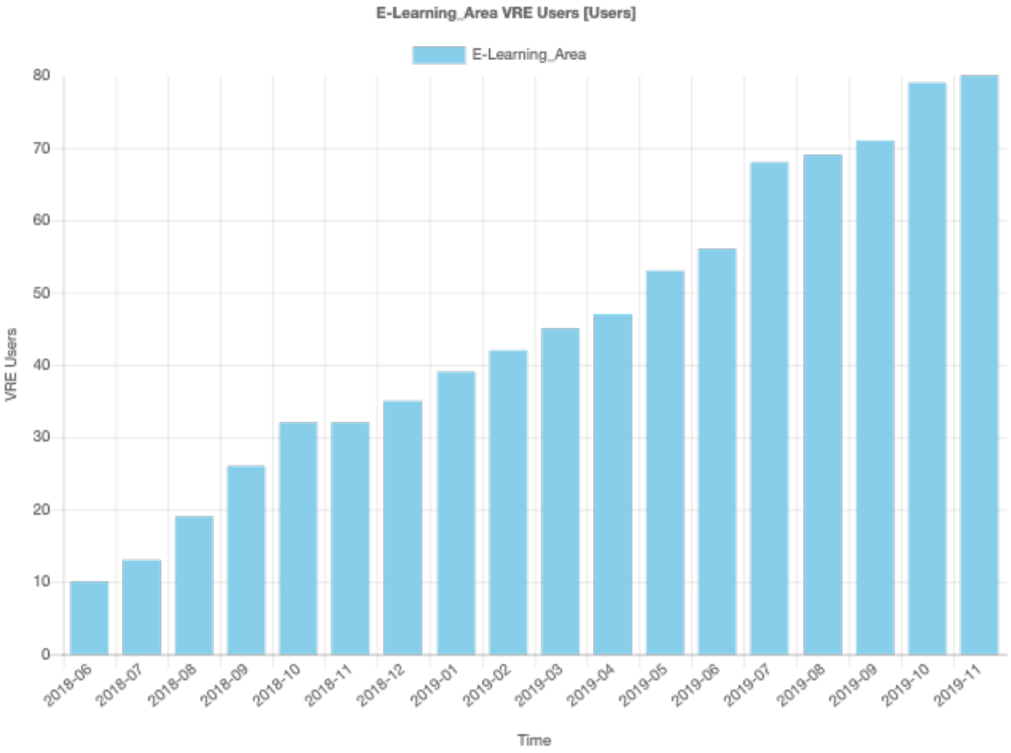


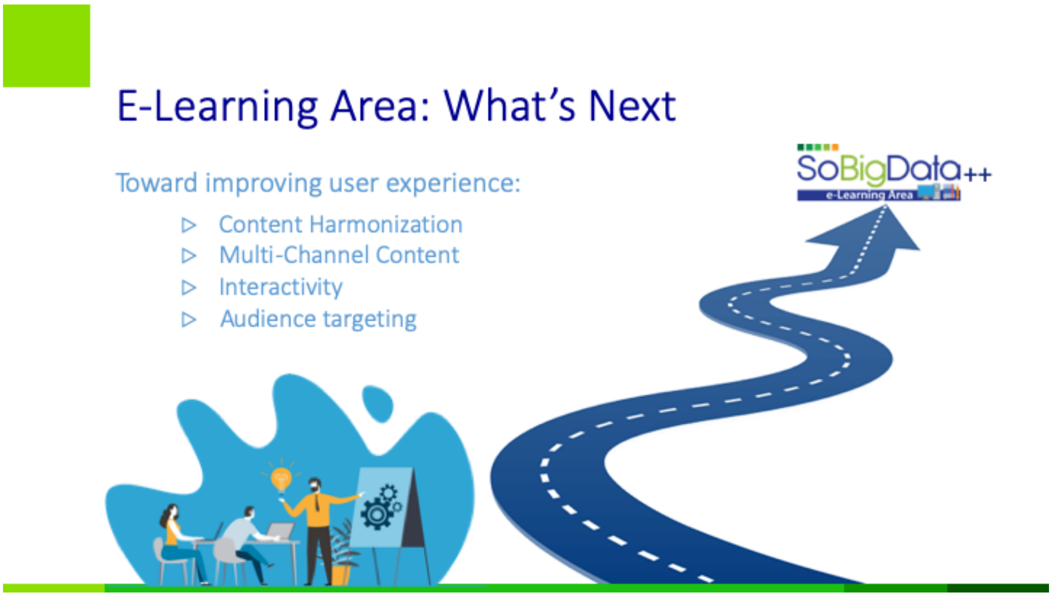
Figure 14: Number of registred users to the e-Learning Area of the SoBigData RI

The trend displayed in Figure 3 is encouraging as it underlines a constant increase in the number of registered users. Moreover, from June 2018 to November 2019 there has been a 700% increase, which outlines a constant and stable interest in the e-Learning Area of the SoBigData Research Infrastructure. Thus, while being a relatively new feature to the SoBigData RI, the e-Learning Area has proven to be attractive both in terms of accesses (Figure 2) and registered users (Figure 3).

4.3 ASSESSMENT OF THE SOBIGDATA E-LEARNING AREA

Work Package 4 also proceeded to an evaluation of the uploaded training materials within the e-Learning Area, in order to assess possible future developments both in the integration and creation of new training materials.

During the project meeting ‘From SoBigData to SoBigData++: Toward an advanced community of Data Scientists’ which was held in Brussels on 19 November 2019, Work Package 4 presented an evaluation of the training materials present in the e-Learning Area of the SoBigData Research Infrastructure and proposed an assessment of possible future directions the creation and development of training materials might follow in the future.



E-Learning Area: What's Next

Toward improving user experience:

- Content Harmonization
- Multi-Channel Content
- Interactivity
- Audience targeting

Figure 15: One of the slides presented during the SoBigData to SoBigData++ meeting by WP4

Four main pillars were identified as potential drivers in improving user experience: Content Harmonization, Multi-Channel Content, Interactivity and Audience Targeting.

Regarding Content Harmonization, while at present the Training Materials within the e-Learning area have achieved what can be defined as a 'light-weight harmonization', content appear as heterogeneous and moreover, there is space for further standardization in production and branding of SoBigData training materials from their inceptions, ideally promoting a recognizable format for users.

Regarding Multi-Channel Content, the current and future training materials appear to be well suited to be developed into different formats which depart from the slide format (which constitutes the majority of the training materials uploaded into the e-Learning Area). A video-lecture format is an option that while not currently available in the present version of the SoBigData RI, has already been explored by projects sharing the D4Science Infrastructure which is at the core of the SoBigData RI. Moreover, possibilities offered by hand-on tutorials and Webinars. According to the Society for Education and Training, a webinar is 'a live or on-demand event, taking place on the internet. It can be a discussion, lecture, conference, presentation, or demonstration. Participants can see documents (usually slides) and other applications via their computer. There will also be shared audio, so you can hear the presenter'. The last format that was explored was the MOOC, or massive open online course. The SoBigData project has developed a MOOC named FAIR (First Aid For Data Scientist) which focuses on issues such as ethics, data protection and intellectual property law. The course provides a direct hands-on approach, with questions that assess the users' understanding of each module, providing real-time suggestions and corrections.

Regarding interactivity, three main areas were identified such as the possible creation of custom curses, coherent with training paths; the development of online exercises, in order to promote users' self-assessment; the integration with other catalogue resources in the SoBigData RI, such as live coding examples and assignments, leveraging libraries and tools which have been developed within SoBigData.

Finally, regarding Audience Targeting, the main idea revolved around the creation of contents that will be directly tailored towards different stakeholders, such as scholars, professionals or the general audience. The objective is to create training materials that are simpler to access and more tailored for specific groups of users and thus promote and foster the dissemination of SoBigData training materials.

5 WORK PACKAGE RESPONSIBILITIES PER PARTNER

5.1 PERSON-MONTHS PER PARTICIPANT AND PARTNER RESPONISIBILITIES

The following tables detail person-months per participant for each of the SoBigData partners involved in Work Package 4 and their main responsibilities.

Participants short name	KCL	CNR	USFD	UNIPI	FRH	UT	LUH	AALTO	ETHZ	TUDelft
Person-months per participant	10	6	4	4	4	4	4	2	6	4

Table 6: person-months per participant to WP4

Partner	**Responsibilities**
KCL	Overall management; Task lead for T 4.2 (including reporting)
CNR	Summer schools; training modules (with particular focus on integration with RI tools) and datathons; widening participation
USFD	Task lead for T 4.1 (including reporting); workshop on widening participation
UNIPI	Task lead for T 4.4 (including reporting); workshop on widening participation; datathon
FRH	Training materials on visualisation; support widening participation
UT	Datathon
LUH	Summer School and widening participation
AALTO	Training material

ETHZ	Task lead for T 4.3 (including reporting); summer school; datathon
TUDelft	Training material on data science ethics; support on widening participation

Table 7: Partners responsabilities on WP4

5.2 MEETING SCHEDULE

The following is the meeting schedule among WP4 partners, which has been developed and followed in order to enhance communication between SoBigData partners in this work package.

Date & Place	**Intra WP**	**Multi-WPs WPs number**	**Telecom**	**In Person**	**Objective / Deliverable / Activities**
Kick-Off Meeting in Pisa					Kick-off of activities: agree upon general work plan and responsibilities; detailed planning for next months
Approximately every 3-6 months; Virtual	x		x		Regular telecom meeting to discuss progress, issues and planning for next period
Various Workshops	x	x	x	x	ad-hoc telecoms when required

Table 8: Meeting schedule among WP4 partners

6 CONCLUSIONS

The project has been highly successful and has deepened and strengthened working relationships within the SoBigData consortium. This collaborative infrastructure will provide greater access to expertise for researchers and will nurture their progress wherever their research is based.

The development of the e-learning software is a great achievement and will further the dissemination of SoBigData training materials and provide a solid resource for researchers worldwide.

As we continue this project with SoBigData++ we will be looking to improve on the number of females becoming involved in Data Science and enhance their participation with incentives (such as Grants) and

Inspiring them through the visibility of prominent females in the field.

Targeting the next generation of Data Scientists by inspiring them at a young age through fun school events is key. The children enjoyed the challenge of the school events and the competitive environment encouraged them to strive their hardest for results. SoBigData has provided events for all ages and has therefore shown great diversity in their target audience. We look forward to continuing this work in the next project.