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 Programme 1***
Deliverable No.	***D4.7***
Delivery Date	***29 February 2016***
Authors	***Giles Greenway (KCL), Tobias Blanke (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	48 months
Funding	H2020-INFRAIA-2014-2015
Grant Agreement No.	654024
DOCUMENT	
Deliverable No.	D4.1
Deliverable Title	Training Programme 1
Contractual Delivery Date	01 March 2016
Actual Delivery Date	29 February 2016
Author(s)	Giles Greenway (KCK)
Editor(s)	Tobias Blanke (KCL), Valerio Grossi (CNR)
Reviewer(s)	Fosca Giannotti (CNR), Valerio Grossi (CNR), Kalina Bontcheva (USFD)
Contributor(s)	Valerio Grossi (CNR)
Work Package No.	WP4
Work Package Title	WP4 - NA3_Training
Work Package Leader	KCL
Work Package Participants	CNR, USFD, UNIPI, FRH, UT, LUH, KCL, AALTO, ETHZ, TUDelft
Dissemination	PU
Nature	Report
Version / Revision	V1.0
Draft / Final	Final
Total No. Pages (including cover)	18
Keywords	Training plan

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 member 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
LTI	Learning Tool Interoperability
WP	Work Package

TABLE OF CONTENT

DELIVERABLE SUMMARY

This deliverable reports the training activities for the first year. It includes three main sections:

- **Section 1**: reports the initial planning of events from the kick-off meeting in September 2015
- **Section 2**: shows an overview of training activities for 2016
- **Section 3**: reports a detailed descritpion of training activities per task
- **Section 4**: defines the reporting template, responsibilities and the meeting scheduling

EXECUTIVE SUMMARY

This deliverable contains the plan of the educational activities for the first year of the project.

1 INITIAL PLANNING SEPTEMBER 2015

At the first partner meeting in September 2015 in Pisa, the following essential components of the training activities were identified:

- Summer Schools
- Datathon/Hackathons
- Training outside academia such as schools
- Data Scientist Training Materials

In addition to these items, the work package proposes a survey on existing activities on data science education in the consortium in order to get a comprehensive overview that will help coordinate SoBigData work.

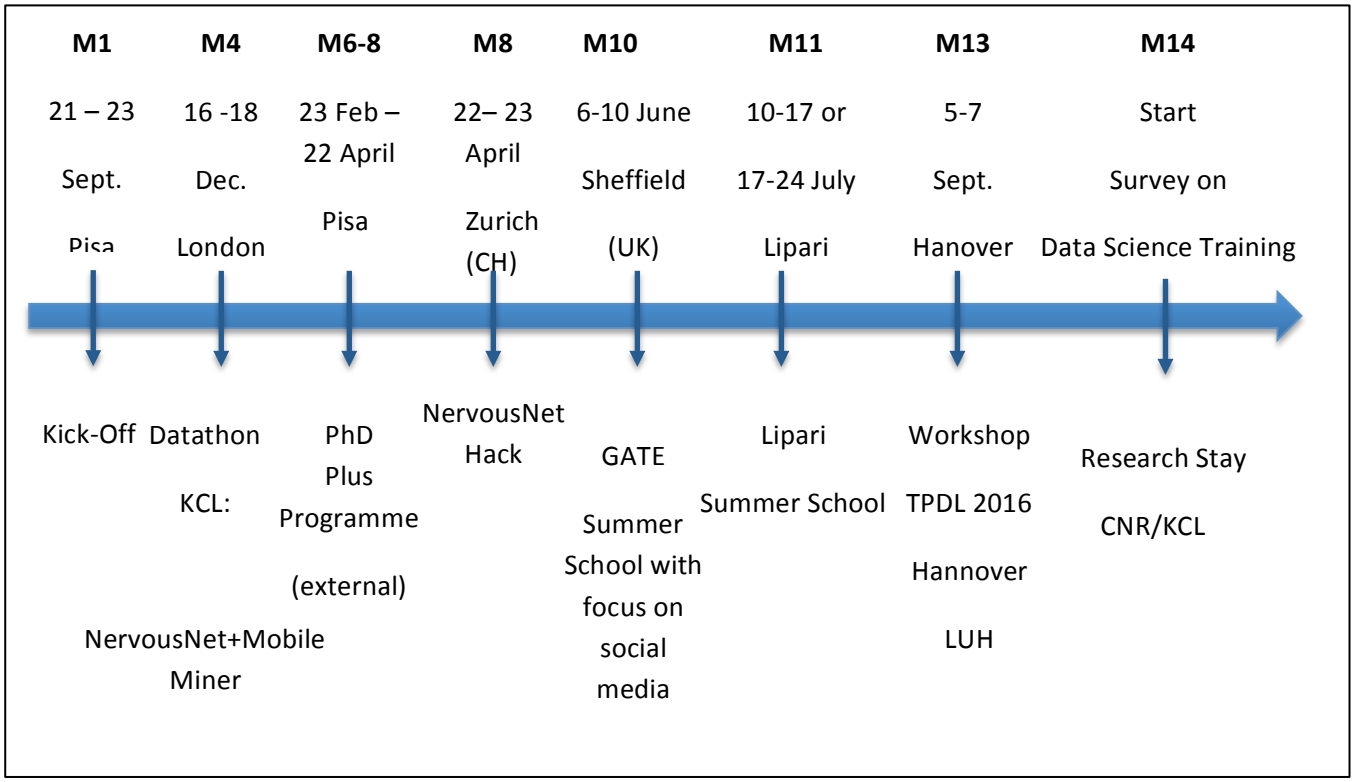
Overall, the following was confirmed in terms of training activities for the whole of SoBigData:

Event	**Num. of events**	**lead partners**	**month**
Workshops	4	LUH, IMT, SNS, KCL, ETHZ, AALTO	M12, M24, M30, M42
Conferences	2		M18, M48
Summer schools	3	USFD, ETHZ, FRH	M10, M27, M39
Datathons	4	UNIPI, UT, ETHZ (2)	M8, M18, M33, M45
Widening Participation Training	3	USFD, UNIPI, LUH	M12, M24, M36

Please, note that not all workshops will have a training component, but are mentioned here for completeness. The KCL workshop, for instance, will definitely combine theoretical with hands-on work.

Finally, it was agreed to try and distribute workshops, conferences, etc. evenly across the the full length of SoBigData. This means approximately 1 summer school per year, 1 workshop, etc.

2 OVERVIEW OF TRAINING ACTIVITIES 2016



Month	Location	Description	Lead Participants
M1	21 – 23 Sept, Pisa	Kick-Off Training	
M4	16 -18 Dec. London	Workshop: NervousNet+MobileMiner	KCL, ETHZ
M6	17 – 19 Feb. Delft	Project Meeting: Training events re-scheduling	
M6-8	23 Feb – 22 April Pisa	PhD Plus Programme (external)	UNIPI
M8	Apr 22-23, Zurich	NervousNet Hackathon	ETHZ
M10	6-10 June Sheffield	Summer School: GATE social media	USFD
M11	July 10 Lipari	Summer School: Lipari School on Computational Complex and Social Systems	ETHZ
M13	5-7 Sept. Hanover	Workshop TPDL 2016	LUH
M14		Survey on Data Science Training	CNR, KCL

Please, note that KCL have decided to have a hackathon/datathon earlier than agreed in Pisa, as it does not have funding for this activity from SoBigData and had to use external funds, which made it necessary to have the event in 2015. Moreover, since the project started in September, summer schools (as originally scheduled) were in November. SoBigData consortium has anticipated/re-scheduled the first one from M15 to M10.

3 OVERVIEW OF TRAINING ACTIVITIES PER TASK

3.1 T4.1 SUMMER SCHOOLS

The University of Sheffield has taken a lead in planning and organising the summer schools for the reporting period. For the planning period, 2 official SoBigData summer schools are planned: the GATE summer school on text analytics for social media and a special event at the LIPARI summer school on Computational Social Sciences in Italy: Algorithms, Data, and Models for Social and Urban Systems. This gives us a perfect mixture of integrating existing excellent work as well as driving new agendas.

LIPARI	**LIPARI Summer School on Computational Social Sciences (Algorithms, Data, and Models for Social and Urban Systems).**
Organiser from the consortium	ETHZ/CNR
Topics	Computational Social Science, Data and Algorithms, Social and Urban Systems
Data	10-17 July, 2016
Location	Lipari Island, Italy
URL	http://lipari.cs.unict.it/LipariSchool/ComplexSocialSystems/
Steering Committee	*Directors*: Dirk Helbing (ETH Zurich, Switzerland), Alfredo Ferro (University of Catania, Italy), Claudio Cioffi-Revilla (George Mason University, USA) Paolo Ferragina (University of Pisa, Italy), Fosca Giannotti (CNR Pisa, Italy), Dino Pedreschi (University of Pisa,Italy), Giovanni Giuffrida (University of Catania, Italy), Rosalba Giugno (University of Catania,Italy), Vittorio Loreto (University of Rome "La Sapienza", Italy), Sergio Palazzo (University of Catania, Italy), Carlo Pennisi (University of Catania, Italy), Alessandro Pluchino (University of Catania, Italy), Alfredo Pulvirenti (University of Catania, Italy), Andrea Rapisarda (University of Catania, Italy), Carlo Ratti (MIT,Boston, USA)

Short Abstract	Social theory has extensively discussed social systems but recently it has been amended by more advanced computational approaches that use new data sets to provide evidence for many of the theories. The summer school asks what is the role of Computational Social Science in advancing the science of social and urban systems? Which advanced algorithms and data structures support this work?
GATE	**GATE Summer School**
Organiser from the consortium	USFD
Topics	Text Analytics for Social Media: Using GATE for social media analysis Challenges for analysing social media with Twitter intro + JSON structure Language identification, tokenisation for Twitter POS tagging and Information Extraction for Twitter
Date	6-10 June 2016
Location	Sheffield, UK
URL	https://gate.ac.uk/conferences/fig/fig9.html
Steering Committee	Diana Maynard, Kalina Bontcheva, University of Sheffield; Tobias Blanke, King's College London
Short Abstract	These are regular training events using the GATE infrastructure. SoBigData will offer fellowships. The 8th GATE training course has a module on learning to analyse social media with the GATE platform using advanced IE, machine learning and cloud services

3.2 T4.2 TRAINING MODULES FOR STAKEHOLDERS

Since the original conceptualisation of the SoBigData project, there has been a lot of movement on open education for data science. Numerous online courses have appeared, while universities across Europe now have a wide range of offerings to teach students. There is also a strong movement mainly in the US to provide broad platforms for data science work. These online education tools include by now commercial platforms offering subscriptions for professional data science education: https://www.datacamp.com/courses and https://www.dataquest.io/. Python and R dominate as the languages of choice for training data scientists.

In parallel, e-learning in universities has further developed with a new focus on open e-learning platforms that allow sharing tools through the Learning Tool Interoperability (LTI) Specification: https://www.imsglobal.org/activity/learning-tools-interoperability. However, these developments are often concentrated on traditional learning tools such as wikis and quizzes, while an integration of advanced digital learning resources for data science might be difficult or at least not complete. The user would still have to switch between different environments. The WP will monitor closely the further the developments in this fast developing domain.

http://www.dataschool.io/teaching-data-science/ has an excellent summary of examples of existing principles that should guide the development of data science educational modules:

- Using GitHub from the beginning is an important principle so that students can present their work to future employers.
- Learning by doing/coding will help understand the material better than traditional frontline power point presentations.
- Using videos and other multimedia resources as teaching platforms will support students' learning progress.
- Working with real-life data within interesting scenarios will help demonstrate the relevance of data science work for economy and society.

In summary, the WP has agreed that SoBigData is committed to developing open-source learning tools. A relatively recent development that combines advantages of e-learning environments as well as advanced data science tools and methodologies are Notebooks is the Jupyter framework (http://jupyter.org/). While based on the previous IPython notebook, they are not limited to Python but integrate a number of programming languages such as R and Scala. Regardless of what language students are working in, they need tools that enhance reproducibility and interactivity across a wide range of usage contexts; from individual exploration and production runs to teaching and presentation. https://github.com/ipython/ipython/wiki/A-gallery-of-interesting-IPython-Notebooks contains a gallery of interesting iPython notebooks.

Notebooks within the Jupyter framework allow embedding executable and modifiable code in an interactive and exploratory manner. On top of this foundation, the notebooks add a document-based workflow. Notebook documents can contain live code, descriptive text, mathematical equations, interactive user-interface components, images, videos, and arbitrary HTML. Such documents thus provide a complete and reproducible record of a computation and can be shared with others, version controlled and converted to a wide range of static formats (HTML, PDF, slides, etc.). Jupyter has a pluggable authentication framework, which should make the LTI specification reasonably straightforward to implement. Authors of training materials that expect learners to write code can provide test cases, by which it could be assessed. The possibility of linking such tests with the LTI-callback to facilitate automated assessment and reporting will be investigated by the WP.

The WP agreed to start a call for open e-learning contributions by the partners and at the same time evaluate the Jupyter framework. We will produce a call for digital educational materials directly, together with a set of criteria for 'nice-to-haves', such as:

- Inclusion of data (including access rights)
- Progress and peer evaluation
- Interactivity
- Multi-media
- Potential to integrate with the research infrastructure.

We stress that not all materials need to have these attributes, and that such materials could be improved over the course of SoBigData; perhaps by members of the consortium other than their initial authors. Finally, the WP agrees that materials created for face-to-face training events should be adapted for use as e-learning material where possible.

3.3 T4.3 SERIES OF DATATHONS

KCL and ETHZ (as the main stakeholder) started planning datathon activities at the end of 2015 and met in London on the 8th of December to clarify the understanding of a datathon. We decided that a datathon in the context of SoBigData will be a more formally organised hackathon that supports social researchers in turning their data into knowledge. It is more formally organised than a hackathon as the focus is not so much a particular solution to a problem but to develop research questions and their corresponding data sets according to a commonly used definition of NY's Institute of Public Knowledge (http://ipk.nyu.edu/initiatives/datathons). In datathons – just like in hackathons – teams with different backgrounds work together. The results can either be judged by a panel of experts and/or will feed directly into the research agenda of the organising institution. On the other hand, datathons are less formal than traditional academic workshops with a focus on testing new ideas and meeting potential collaborators in a working environment without the direct need to produce direct academic outputs. The reporting of the datathons within SoBigData should reflect this and include multimedia components ready for distribution.

In order to make the experience of datathons sustainable, the WP adopts the following principles:

- We aim to incorporate in all datathons well defined, repeatable and measurable activities along with more creative or loosely defined development components
- We will organize and coordinate a collaborative infrastructure for the datathon so that data, documentation, visualizations and analyses can be created and shared rapidly online. A preference is given to open source tools such as GitHub.
- We will target real world problems with real world data – wherever possible.
- We plan follow-up activities for public (reports, videos of the datathon, interview with winners)
- We share experience with partners: we may setup a 'datathon committee' depending on the overall commitments of partners.

The datathons will be planned from early 2016. KCL will organise a Datathon Scientific Committee together with ETHZ, Pisa and UT.

We had an initial datathon in London on the 16-18 December in London working through the integration of NervousNet and MobileMiner together with students at KCL across arts and sciences. A follow-on hackathon is planned for the 22-23 April 2016 in Zuerich and will be organised by ETHZ.

The budget holder for datathons in SoBigData are ETHZ (2x), UT (1x), and UNIPI (1x).

3.4 T4.4 ADDRESSING GENDER AND DIVERSITY ISSUES IN DATA SCIENCE THROUGH TRAINING

Our discussions have led to a clearer distinction of what we understand by addressing gender and diversity issues in data science. We agreed a targeted programme to widen participation in data science and opening up our work to under-represented groups in order to find the brightest minds regardless of their background. We are strongly committed to supporting them with access to the best possible data science education according to the most recent discussions on data science ethics (http://columbiadatascience.com/2013/11/25/data-science-ethics/). We will develop schemes and programmes and special school visits, etc. to widen participation. We can rely on a number of activities within the partner organisations of SoBigData that will help enable a link with less advantaged socio-economic backgrounds, students from low-participation neighbourhoods, mature learners, black and minority ethnic students, disabled students and care leavers.

The WP has identified a number of potential collaboration partners that will help us connect not just to school children but also to professional organisations:

- The women network at the European Association on Data Mining and Machine Learning (http://www.kdnuggets.com/2012/06/women-data-scientists-data-miners.html)
- PyData and PyLadies (http://www.pyladies.com/)
- Rewired State (http://www.rewiredstate.org/#intro) and Young Rewired State (http://www.yrs.io/)
- Open Knowledge Foundation (https://okfn.org/)
- Institutional partners such as King's College Widening Participation (http://www.kcl.ac.uk/study/widening-participation/index.aspx)

The budget holder for training events in this domain are LUH (1), USFD (1x), and UNIPI (1x).

We will start planning the events once we have a better picture of the technical support infrastructure we have available, which we will test through the datathons and learning modules. Once these are mature we can use them for the widening participation actions.

In the planning period, UNIPI is leading an event at M24 as decided, while USFD have planned a school event in summer/autumn 2016. The exact dates will be decided by April 2016.

4 REPORTING TEMPLATE, RESPONSIBILITIES AND SCHEDULE

4.1 REPORTING TEMPLATE

We will require the following information from all training activities:

1) *What was the training objective(s)*
2) *How many people/participants attended (please break down by gender and age (youth, mature, old), and other relevant groupings (such as community, private professional, civil society, government professional)*
3) *How did the training sessions attempt to address the training objective(s)*
4) *What was the overall cost of holding the training*
5) *What are the immediate and longer-term benefits resulting from the event*
6) *How will you monitor or evaluate the outcome of this event into the future*
7) *Anything you would do differently next time?*
8) *If you have any pictures of the training event, please use to illustrate your report. Can we upload the pictures to our website?*

Please, note that we will review these questions after the first round of training events for completeness, usefulness, etc.

4.2 RESPONSIBILITIES

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

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

4.3 INITIAL MEETING SCHEDULE

Date & Place	Intra WP	Multi WPs WPs number	Telecom	In persons	Objective/Deliverable/Activities concerned
Kick-off meeting in Pisa	x			x	Kick-off of activities: agree upon general work plan and responsibilities; detailed planning for next months
approx. every 3-6 month; virtual meeting of task leaders and relevant WP members	x		x		Regular telecom meeting to discuss progress, issues and planning for next period
Deflft Project Meeting M6	x	WP3, WP4, WP5		**x**	Meeting to discuss progress, issues and planning for next period
Lipari Project Meeting	x	WP3, WP4, WP5		**x**	Meeting to discuss progress, issues and planning for next period

M11					
Various workshops, events, etc.	x	x	x	x	Ad-hoc telecoms when required

4.4 INITIAL RISK ANALYSIS OF THE WP

Event	**Effect** Major, Moderate, Minor	**Likely-Hood** Likely, Unlikely	**Risk level** High, Moderate, Low	**Control measure(s)**
Procedural				
R1: Difficult and ponderous decision making across the WP.	Moderate	Likely	Moderate	- Facilitate open discussion, and flow of information and idea, among all consortium partners - Implementation of a clear management structure with clearly defined responsibilities and decision making procedures.
R2: Milestones and deliverables scheduled for the second half of the project need to be readjusted and redefined in the light of results from the first half.	Moderate	Likely	Moderate	- Continuous evaluation of progress, and agile scheduling of tasks
R3: Difficulties maintaining staff throughout the project	Major	Likely	High	- Keep the documentation on all work up-to-date so that new staff can easily work integrate
R4: Technology change leads to unrealistic assumptions on how to progress with the education work	Moderate	Likely	Moderate	- Continuous evaluation of technology leads to easy integration of new ones.

Content related				
R5: Difficulties to create a common language among the data science disciplines, leads to mutual misunderstandings about the respective needs and requirements of these disciplines..	Major	Unlikely	High	- Careful management of members of all disciplines within consortium - Key people in the WP, representing different disciplines, have already worked together very intensely and successfully, and can act as "translators".
R6: The heterogeneity of the WP partners leads to difficulties in regard to long-term sustainability of the training results	Major	Unlikely	High	Creation of relevant contexts for the work of SoBigData will mitigate the effect of alienation in the Network
R7: Tools and methods for data science education are not sustainable	Major	Unlikely	High	There is already enough existing work on tools for data science education that this event seems unlikely. We will be conservative here and focus on tools that are commonly used.