SoBigData: Social Mining & Big Data Ecosystem

Fosca Giannotti
KDD Lab, ISTI - CNR
Pisa, Italy
fosca.giannotti@isti.cnr.it

Roberto Trasarti
KDD Lab, ISTI - CNR
Pisa, Italy
roberto.trasarti@isti.cnr.it

Kalina Bontcheva
Department of Computer Science
Sheffiel, UK
k.bontcheva@sheffield.ac.uk

Valerio Grossi
KDD Lab, ISTI - CNR
Pisa, Italy
valerio.grossi@isti.cnr.it

ABSTRACT

One of the most pressing and fascinating challenges scientists face today, is understanding the complexity of our globally interconnected society. The big data arising from the digital breadcrumbs of human activities has the potential of providing a powerful social microscope, which can help us understand many complex and hidden socio-economic phenomena. Such challenge requires high-level analytics, modeling and reasoning across all the social dimensions above. There is a need to harness these opportunities for scientific advancement and for the social good, compared to the currently prevalent exploitation of big data for commercial purposes or, worse, social control and surveillance. The main obstacle to this accomplishment, besides the scarcity of data scientists, is the lack of a large-scale open ecosystem where big data and social mining research can be carried out. The SoBigData Research Infrastructure (RI) provides 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". The research community uses the SoBigData facilities as a "secure digital wind-tunnel" for large-scale social data analysis and simulation experiments. SoBigData promotes repeatable and open science and supports data science research projects by providing:
i) an ever-growing, distributed data ecosystem for procurement, access and curation and management of big social data, to underpin social data mining research within an ethic-sensitive context;
ii) an ever-growing, distributed platform of interoperable, social data mining methods and associated skills: tools, methodologies and services for mining, analysing, and visualising complex and massive datasets, harnessing the techno-legal barriers to the ethically safe deployment of big data for social mining;
iii) an ecosystem where protection of personal information and the respect for fundamental human rights can coexist with a safe use of the same information for scientific purposes of broad and central societal interest. SoBigData has a dedicated ethical and legal board, which is implementing a legal and ethical framework.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4.
https://doi.org/https://doi.org/10.1145/3184558.3185985

PROJECT INFORMATION

Acronym: SoBigData;
Duration: From 01/09/2015 to 30/08/2019;
Volume: 6 MLN;
Funding Agency: H2020 Programme; Gn.654024;
Official Website: www.sobigdata.eu

Principal Investigators

Fosca Giannotti (female) is Research Director at ISTI-CNR, Pisa. She is the author of several publications (more than 200) and served in the scientific committee of the main conferences in the area of Databases and Data Mining. CNR (coordinator, IT). She will be the presenter of the project .
Kalina Bontcheva (female) is a senior research scientist and the holder of a prestigious EPSRC career acceleration fellowship, working on text mining and summarization of social media. University of Sheffield (UK);
Dino Pedreschi (male) is Professor of Computer Science and a pioneering scientist in mobility data mining, social network mining and privacy-preserving data mining. University of Pisa (IT).
Gennady Andrienko (male) is lead scientist responsible for the visual analytics research and full professor at City Univ. London. Fraunhofer (DE).
Marlon Dumas (male) is Professor of Software Engineering and one of the five most highly cited software engineering researchers in Europe. University of Tartu (Estonia).
Guido Caldarelli (male) will act as Team Leader for research, and will work on the theoretical network analysis. IMT Lucca (IT).
Wolfgang Nejdl (male) has been full professor of computer science since 1995. He heads the L3S Research Center, as well as the Distributed Systems Institute/Knowledge Based Systems. Leibniz UniversitÃđt Hannover (DE).
Tobias Blanke (male) is a Senior Lecturer in the Department of Digital Humanities. His academic background is in philosophy and computer science. King's College London (UK)
Fabrizio Lillo (male) is currently Professor of Mathematical Finance interested in the microstructure of financial markets and high frequency finance. Scuola Normale Superiore (IT).
Aristides Gionis (male) is an Associate Professor and the director of the Algorithmic Data Analysis programme in Helsinki Institute for Information technology. Aalto University (Finland).
Dirk Helbing (male) has worked in natural, engineering and social science departments and is currently a Full Professor in Sociology, in particular of Modeling and Simulation. ETHZ (CH).

Jeroen van den Hoven (male) is professor of moral philosophy and scientific director of 3TU.Ethics. TU Delft (NL).

SOBIGDATA OBJECTIVES AND TIMELINE

The major goal is granting access (both virtual and trans-national on-site) to the SoBigData RI to multi-disciplinary scientists, innovators, public bodies, citizen organizations, SMEs, as well as data science students at any level of education. The project is creating a new European e-infrastructure that comprises virtual research environments for data scientists, large datasets in unified format, and hundreds of distributed, interoperable analytics and visualisation services, accessible through common interfaces. SoBigData builds on expertise and recent advances in e-Infra-structures, relevant European data infrastructures, and seven established national infrastructures and data centers organized in the following six thematic clusters arising from different research fields of social mining and big data analytics (text and social media mining - TSMM; social network analysis - SNA; human mobility analytics - HMA; web analytics - WA; visual analytics - VA; and social data - SD). The following national infrastructures are integrated:

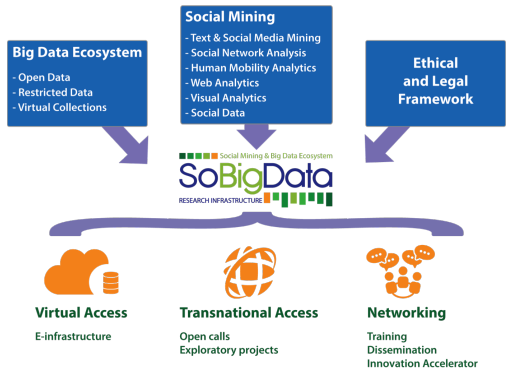

Figure 1: SoBigData resources and access opportunities.

SoBigData.it, the knowledge infrastructure of the European Laboratory on Big Data Analytics and Social Mining, funded by two institutes of CNR and the University of Pisa, offering data, services and expertise on TSMM, SNA, HMA, WA.
GATE Cloud, the cloud-based infrastructure for large-scale natural language processing and text mining at the University of Sheffield, offering data, services and expertise on TSMM.
IVAS, the server-based infrastructure for Information Visualization and Visual Analytics at the Fraunhofer competence centre, offering data, services and expertise on VA. Alexandria, the infrastructure for Web Science at Leibniz Universitaet, offering data, services and expertise on WA.
AALTO, the Sociophysics and Data Mining laboratories at Aalto University in Helsinki, offering data, services and competences on SNA.
E-Gov.data, the centre at University of Tartu curating the Estonian e-government and e-health data, offering access and services to these unique dataset over 10 years (SD).
Living Archive for Open Data, a search engine for Open Data offered by the Department of Sociology at ETH Zurich (SD).

SoBigData has a strong ethical motivation. We are building an ecosystem where protection of personal information and the respect for fundamental human rights coexist with a safe use of the same information for scientific purposes of broad and central societal interest. Extreme care is taken in all parts of the project and among all consortium partners not only to guarantee compliance with relevant laws, regulations, and codes of conduct, but also to prevent unethical outcomes of the envisaged cutting edge research and applications by ongoing analysis of the moral problems that pose constant challenges for law, regulation and governance of information technology. SoBigData pursues the views that EU is developing on Responsible Research and Innovation, and it operationalizes the values that drive the ongoing reform of the EU Data Protection legislation. SoBigData RI manages vertical, thematic environments, called exploratories, on top of the SoBigData infrastructures, to perform cross-disciplinary social mining research:

City of Citizens: this exploratory tells stories about cities and people living in it. We describe those territories by means of data, statistics and models.
Well-being & Economy: can Big Data help us to understand relationships between economy and daily life habits? We use data of purchases in supermarkets and investigate people's behavior.
Societal Debates: this exploratory studies public debates on social media and newspaper. It identifies themes, following the discussions around them and tracking them through time and space.
Migration Studies: could Big Data help to understand the migration phenomenon? We try to answer to some questions about migrations in Europe and in the world.
Sports Data Science: this exploratory tells stories about sports analytics. Sports data scientists describe performances by means of data, statistics and models. This allows coaches, fans and practitioners to better understand and boost sports performance.

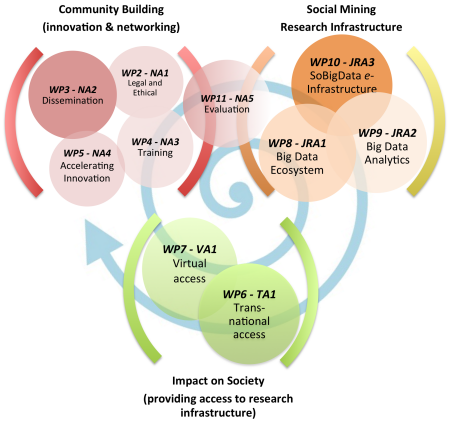

Figure 2: SoBigData Project organization

Currently (January 2018) SoBigData contains more than 120 resources including methods and datasets published in its platform and the number is growing thanks to the collaboration between the original partners and new organizations and users. The platform is now reaching one thousand users which is a good only the beginning for a young platform which will see big improvements in the next year (thanks to the dissemination and the integration of new resources).

PRESENTATION AND DURABLE RESULTS

This talk outlines the SoBigData Vision, Goal, Organization and opportunities for researchers to be users and partners of this initiative. Moreover the research topics and the impact of the project is presented according to three Exploratories: City of Citizens, Societal Debates and Well-Being.
The European e-infrastructure which will ultimately comprise a virtual research environment for data scientists, large datasets in unified format, and hundreds of distributed, interoperable analytics and visualisation services, accessible through common interfaces.

This paper is supported by SoBigData.eu, H2020 Programme n.654024.