SoBigData Research Infrastructure

Social Mining & Big Data Ecosystem

H2020 - www.sobigdata.eu

September 2015- August 2019

GARR Conference – 16th November2017

@SoBigData (https://twitter.com/SoBigData)

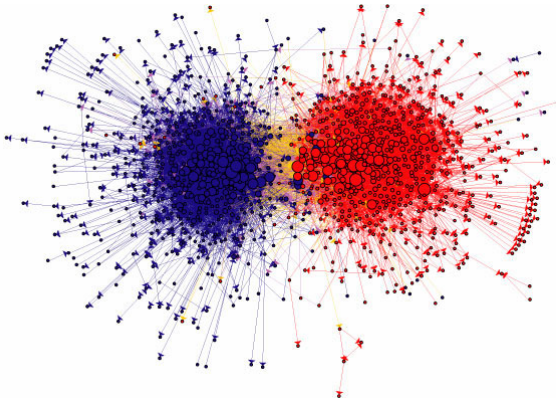
https://www.facebook.com/SoBigData

SOBIGDATA VISION AND VALUES

Big data “proxies” of social life

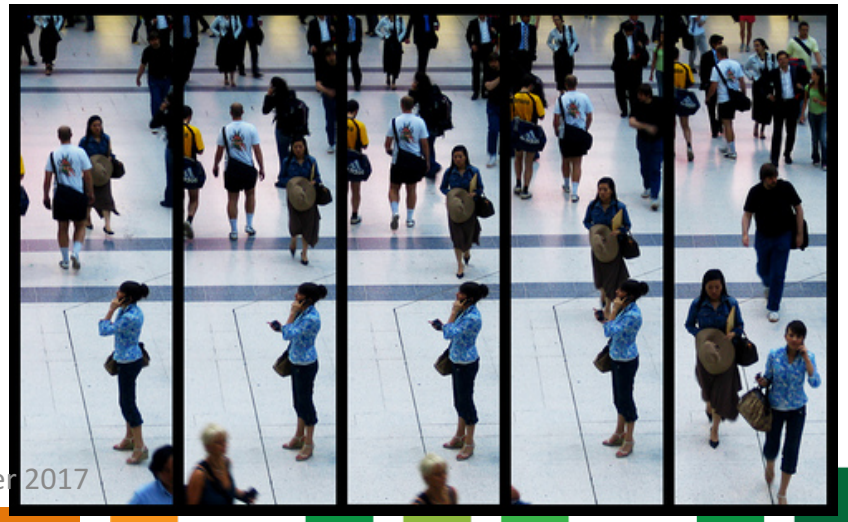
Shopping patterns & lifestyle

Relationships & social ties

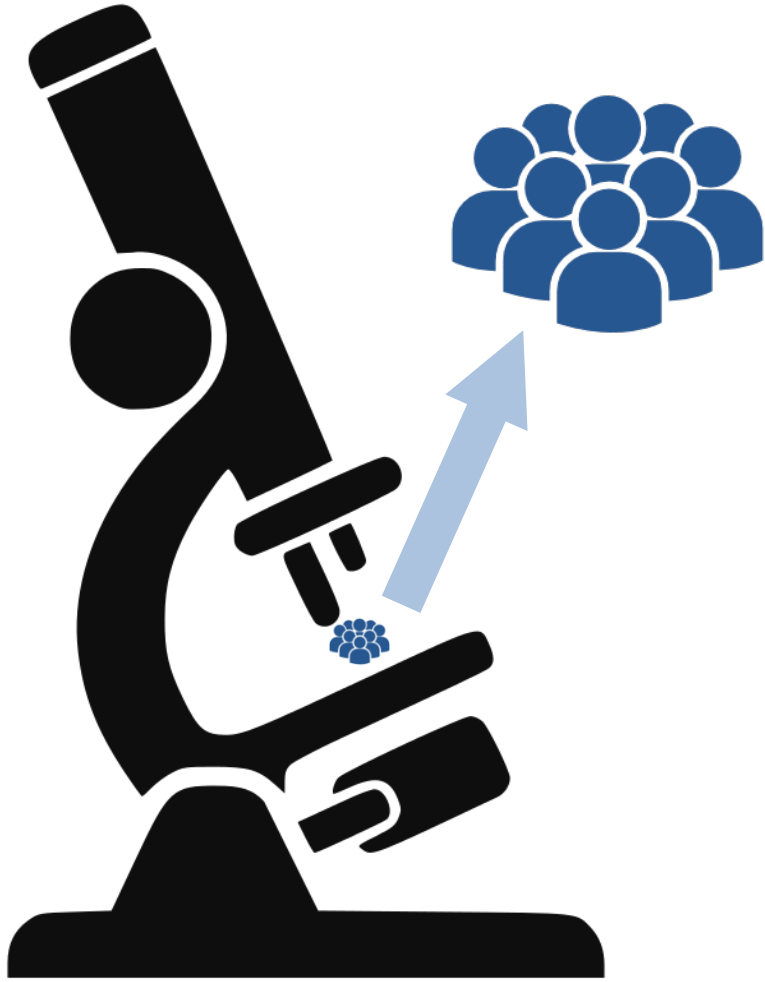


Desires, opinions, sentiments

Movements

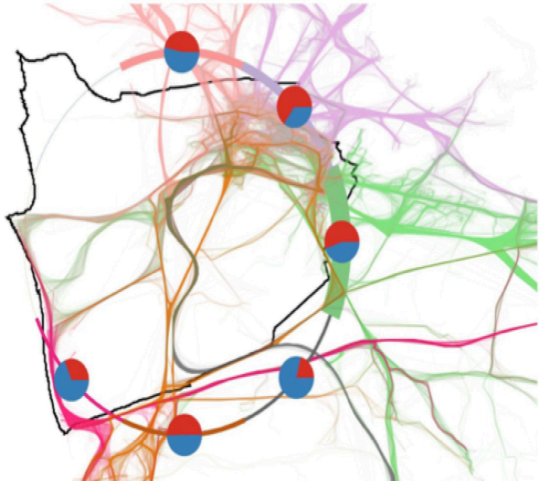


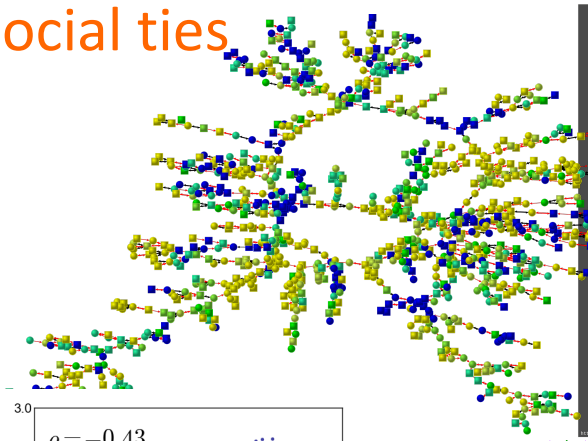
Social mining: making sense of big data to understand society

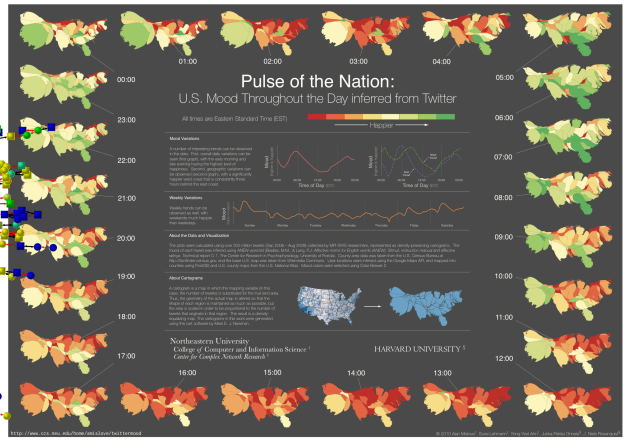


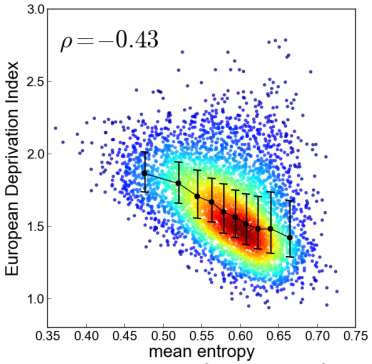
What is Social Mining

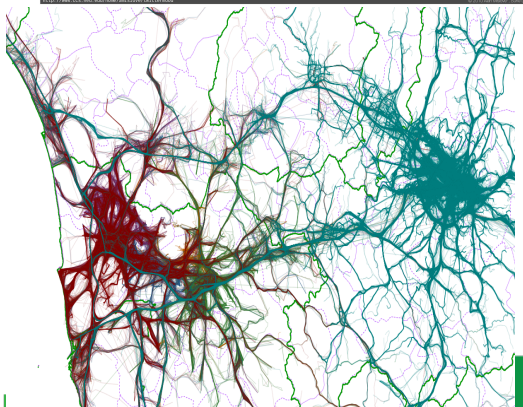
- Automated discovering patterns and models of human behaviour across the various social dimensions that have big data “proxies”
  - desires and opinions
  - relationships and social ties
  - life-styles
  - mobility



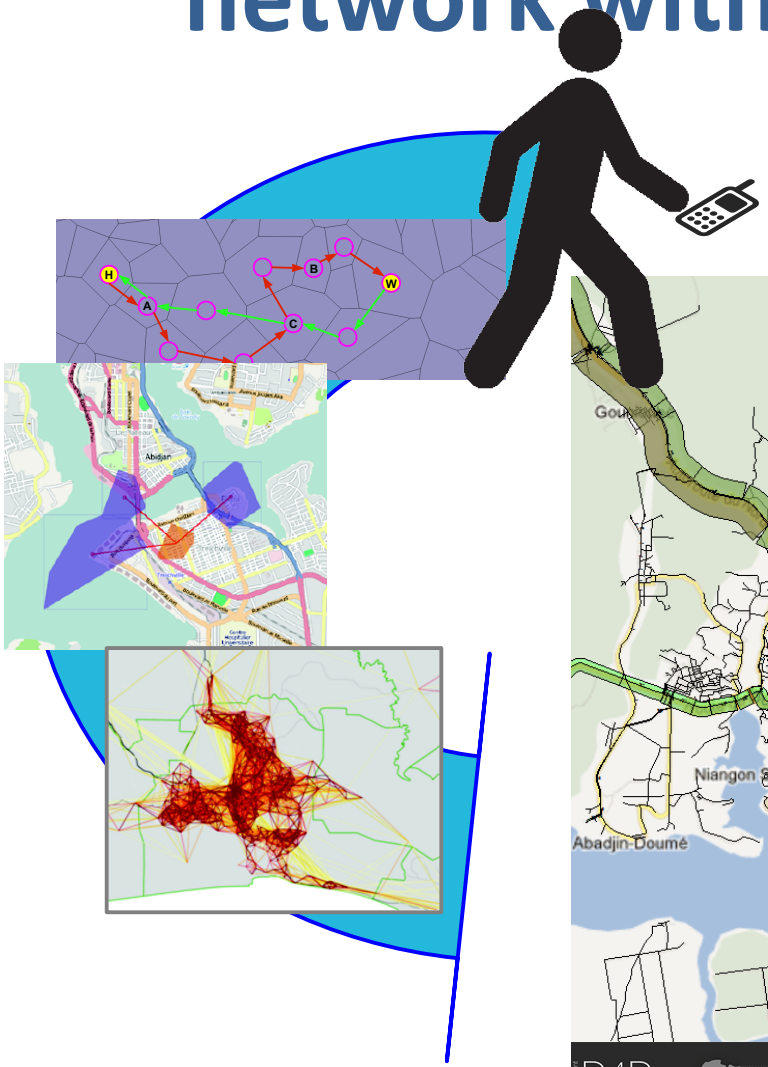


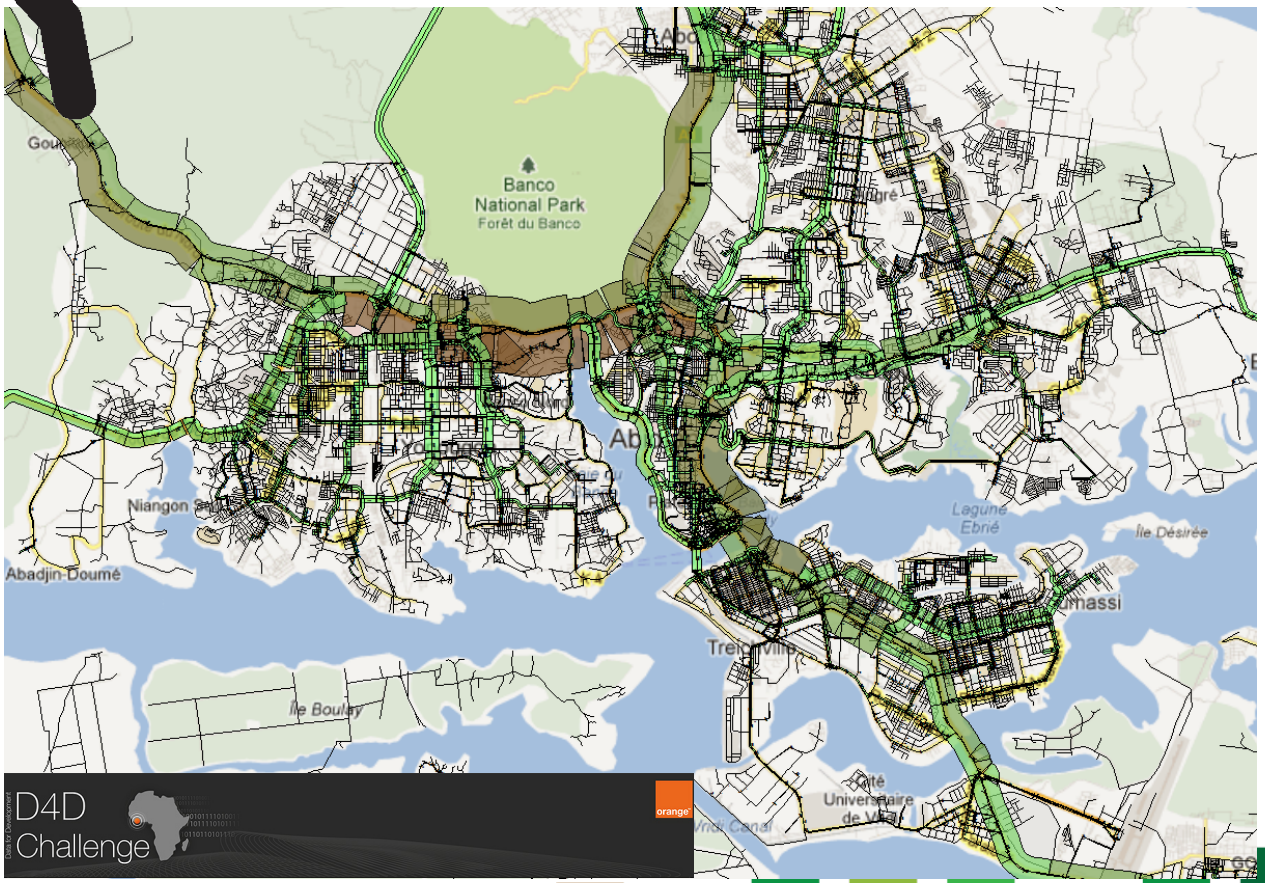






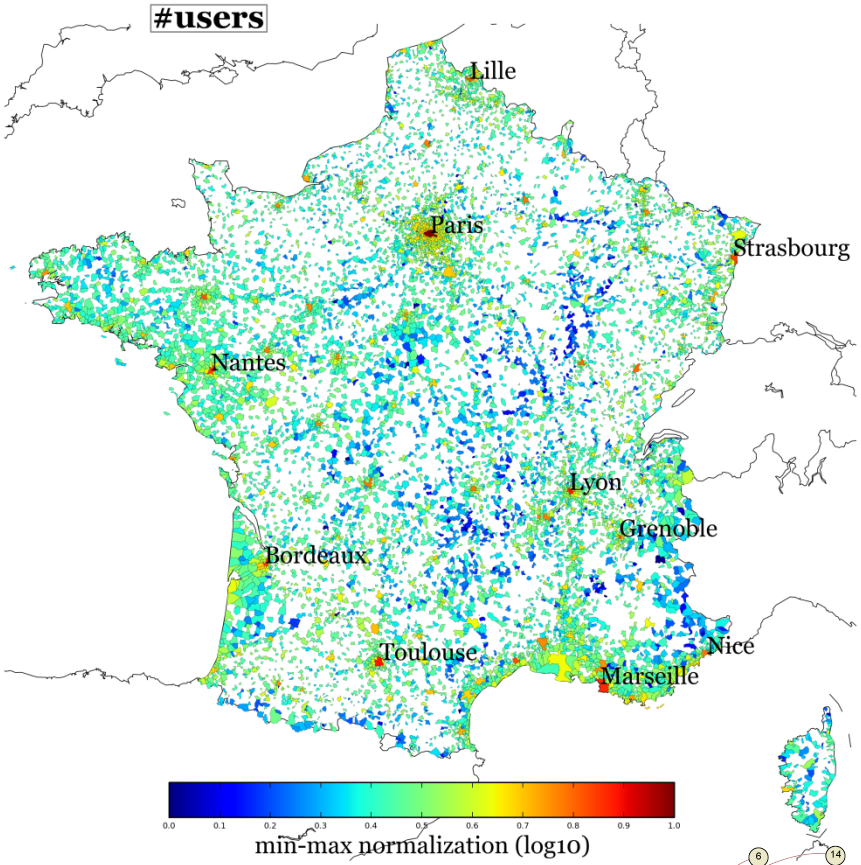
Estimating traffic fluxes on road network with mobile phone data

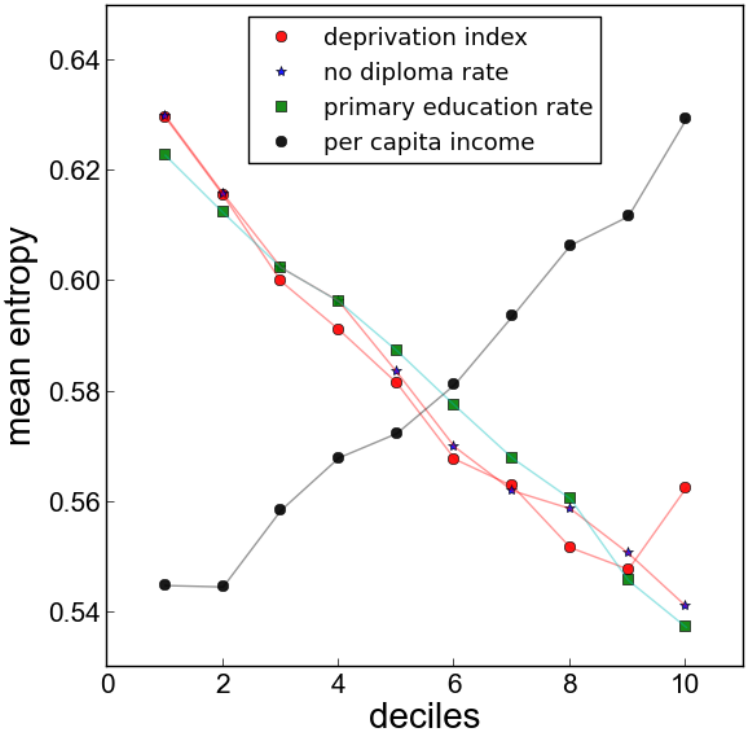


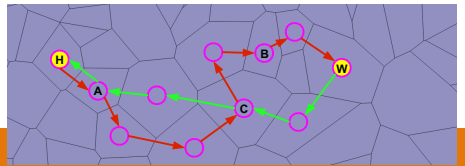


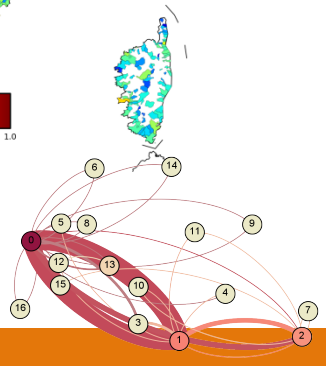
Estimating Country well-being with mobility

Diversity









$$d_i^{(n)} = \sum_{j=1}^{|V|} \frac{1}{k_j} M_{ij} p_j^{(n-1)} \forall i$$

$$p_j^{(n)} = \sum_{i=1}^{|U|} \frac{1}{k_i} M_{ij} d_i^{(n-1)} \forall j$$

Social Sensing

Real-time crisis mapping

TABLE 6. Results of evaluating our system's ability to detect areas damaged by the Emilia earthquake.

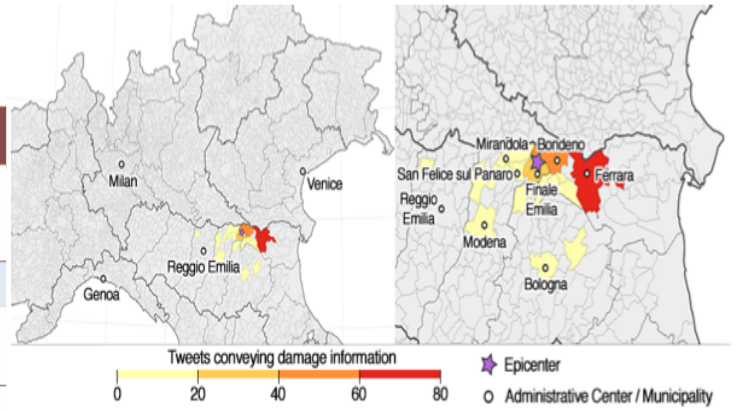
Task	Evaluation metrics					
	Precision	Recall	Specificity	Accuracy	F-measure	MCC*
Detect all damaged areas	0.895	0.202	0.992	0.797	0.330	0.365
Detect areas that suffered significant damage	0.867	0.813	0.992	0.982	0.839	0.830

*MCC: Matthews correlation coefficient

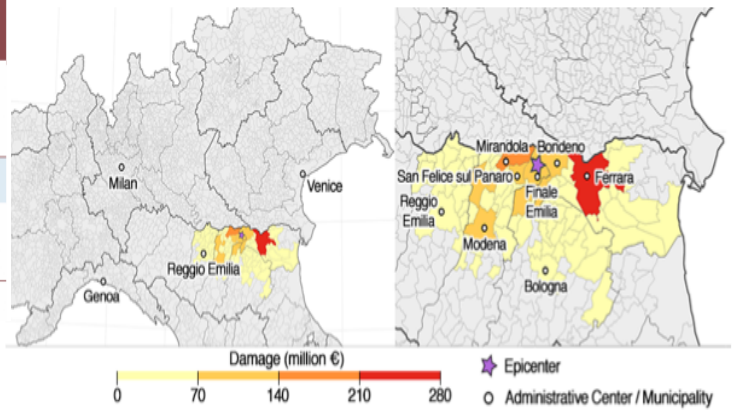
TABLE 7. Results of evaluating our system's ability to detect areas damaged by the Sardegna flood.

Task	Evaluation metrics					
	Precision	Recall	Specificity	Accuracy	F-measure	MCC*
Detection of all damaged areas	0.640	0.410	0.973	0.915	0.500	0.470
Detection of areas that suffered significant damage	0.500	0.643	0.973	0.960	0.563	0.545

*MCC: Matthews correlation coefficient



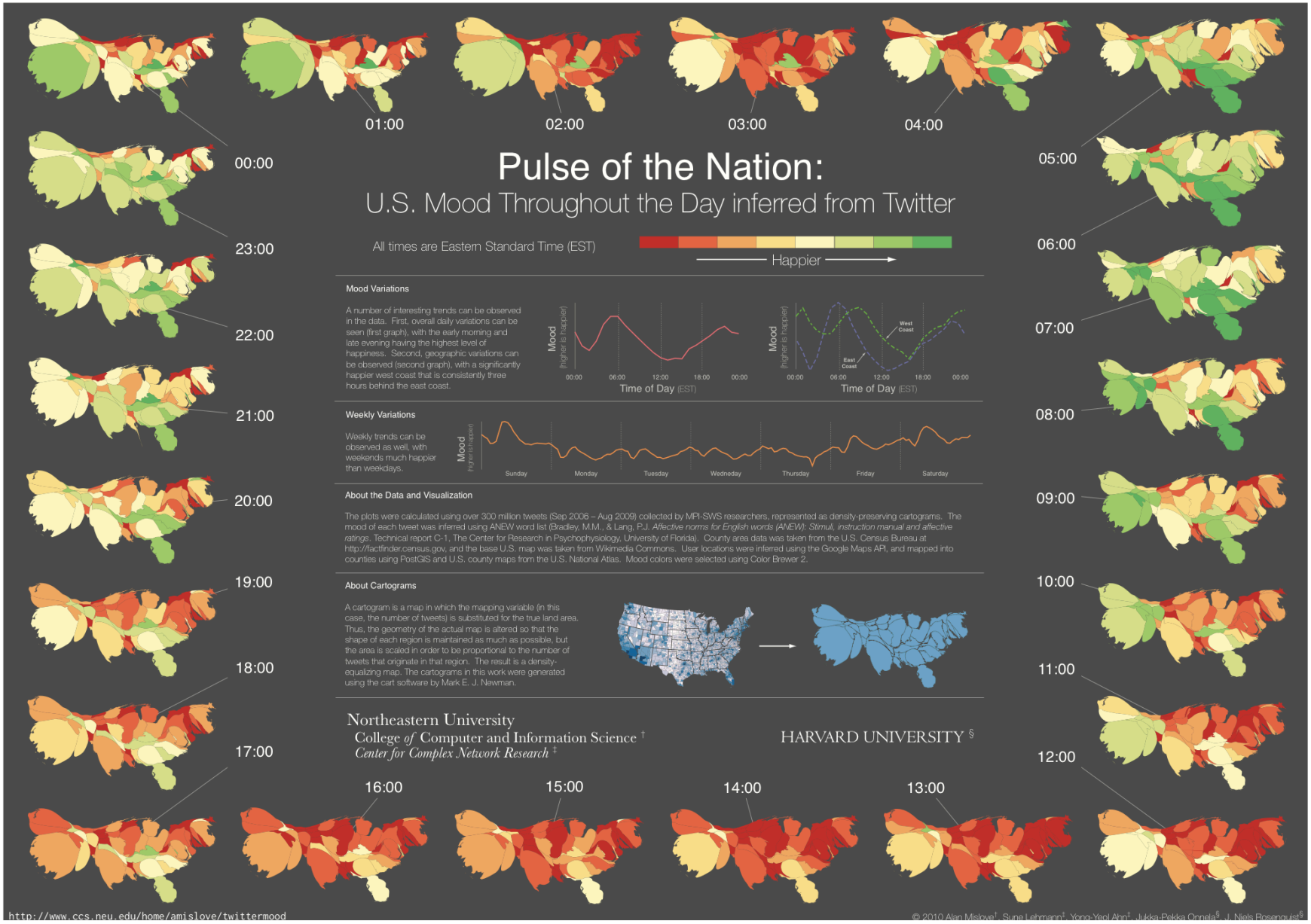
Overview and detail of the crisis map generated by our system.



Overview and detail of the authoritative post-event damage assessment map. Authoritative data about the economic losses are provided by Emilia Romagna regional district (http://www.openricostruzione.it).

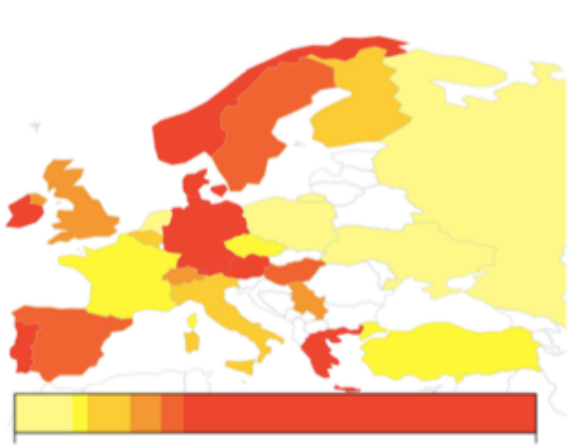
Avvenuti, M., Cresci, S., Del Vigna, F., & Tesconi, M. (2016). Impromtu crisis mapping for prioritized emergency response. *IEEE Computer*, (to appear in May 2016).

Measuring happiness with twitter data

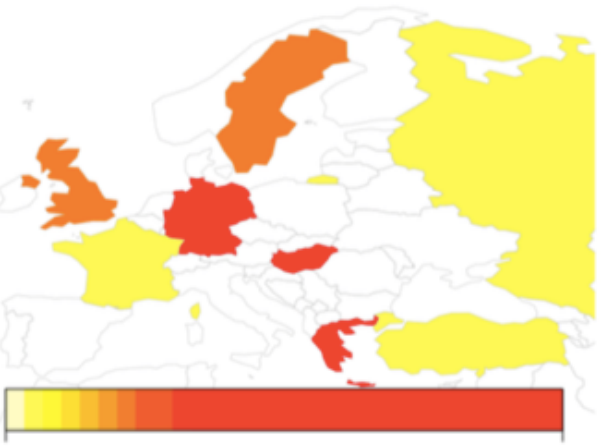


Sentiment Analysis

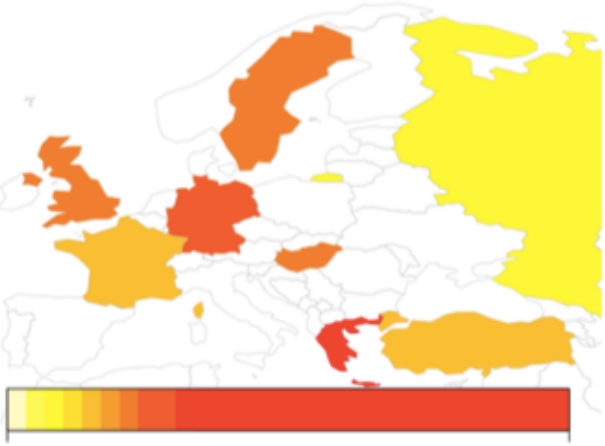
- Internal and external perception by country
  - Index **ρ** - the ratio between pro refugees users and against refugees users
  - Red means a higher predominance of positive sentiment, higher ρ



\- + (a) Overall.



\- + (b) Internal perception.



\- + (c) External perception.

RONGER IN
HOW BRITAIN TALKED BREXIT
NHS
IMMIGRATION
BREAKING POINT
We send the EU £350 million a week
let's fund our NHS instead
Vote Leave
ECONOMY
LAW
SOVEREIGI

3 Million Brexit Tweets Reveal Leave Voters Talked About Immigration More Than Anything Else

Groundbreaking analysis shows immigration, not sovereignty or the NHS, dominated the conversation – and making British judges responsible for British law was a key theme for Leave supporters.

James Ball
BuzzFeed Special Correspondent

Chris Applegate
Editorial Developer, UK

posted on Dec. 9, 2016, at 2:03 p.m.

https://www.buzzfeed.com/jamesball/3-million-brexit-tweets-reveal-leave-voters-talked-about-imm?utm_term=.jmDQE9JNR#.fuOOrb145

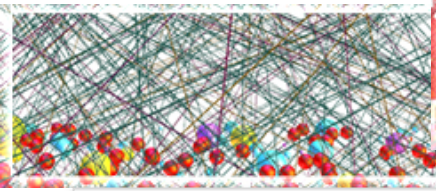
What's special in social mining?

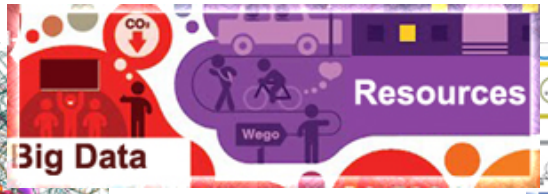
- Any **data science experiment** is composed by:
  - data acquiring (open data, crowdsourcing, crowdsensing,)
  - model building (data mining, machine learning, network science, …and very complex validation phase),
  - creation of an exploration scenario (what-if analysis) (different validation setting),
- ….similar to many other data-driven science process,…**but data are produced by humans**

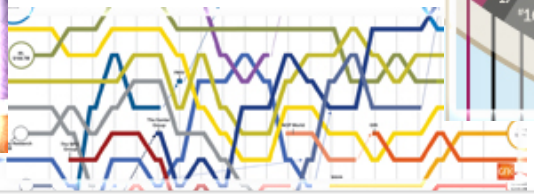
What is needed

- responsibility and trust
  - **FACT**: Fairness, Accuracy, Confidentiality and Transparency
- harness social mining for scientific advancement and for the social good
  - **FAIR**: Findable, Accessible, Interoperable, Reproducible
- **responsible open data science**

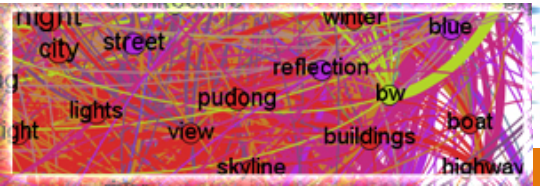
SoBigData GOAL is…

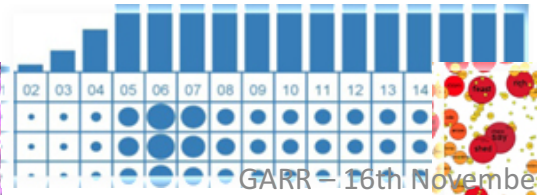


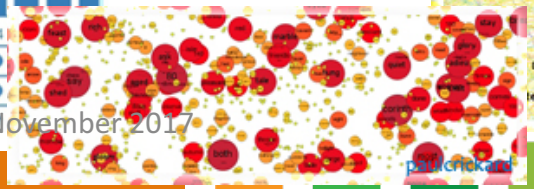


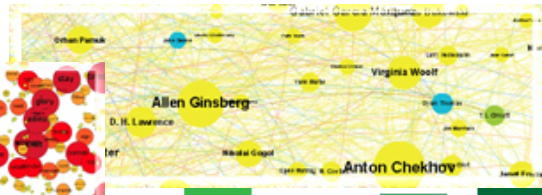


TO CONSTRUCT THE **Multidisciplinary European Infrastructure on Big Data and Social Data Mining (the Social Mining CERN)** providing an integrated ecosystem **for ethic-sensitive scientific discoveries** and advanced applications of social data mining on the various dimensions of social life, as recorded by "big data".









The **pillars** for reaching the goal

- **an ever-growing, distributed data ecosystem for procurement, access and curation of big social data, within an ethic-sensitive context**, based on
  - innovative strategies for acquiring social big data for research purposes,
  - using both opportunistic means offered by social sensing technologies and
  - participatory means based on user involvement as prosumers of social data and knowledge.

The **pillars** for reaching the goal

- **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.

The **pillars** for reaching the goal

- **Building the Social Mining community of scientific, industrial, and other stakeholders (e.g. policy makers),**

The path to achieve the goals

- **Integrate European national infrastructures** and centres of excellence in big data analytics, social mining and data science
  1. ***Text and Social Media Mining*** **(TSMM)**
  2. ***Social Network Analysis*** **(SNA)**
  3. ***Human Mobility Analytics*** **(HMA)**
  4. ***Web Analytics*** **(WA)**
  5. ***Visual Analytics*** **(VA)**
  6. ***Social Data*** **(SD)**

Integrating national research Infrastructures

GATE 011 general architecture for text engineering

LIVINGARCHIVE Lectives

SoBigData Euro Lab on Big Data Analytics & Social Mining

TARTU ÜLIKO 1632

L3S Research C

Fraunhof

The Consortium

Consiglio Nazionale delle Ricerche

The University Of Sheffield.

Università di Pisa

Fraunhofer FHR

TARTU ÜLIKOOL 1632

IMT Institute for Advanced Studies Lucca

Leibniz Universität Hannover

King's College London

Scuola Normale Superiore

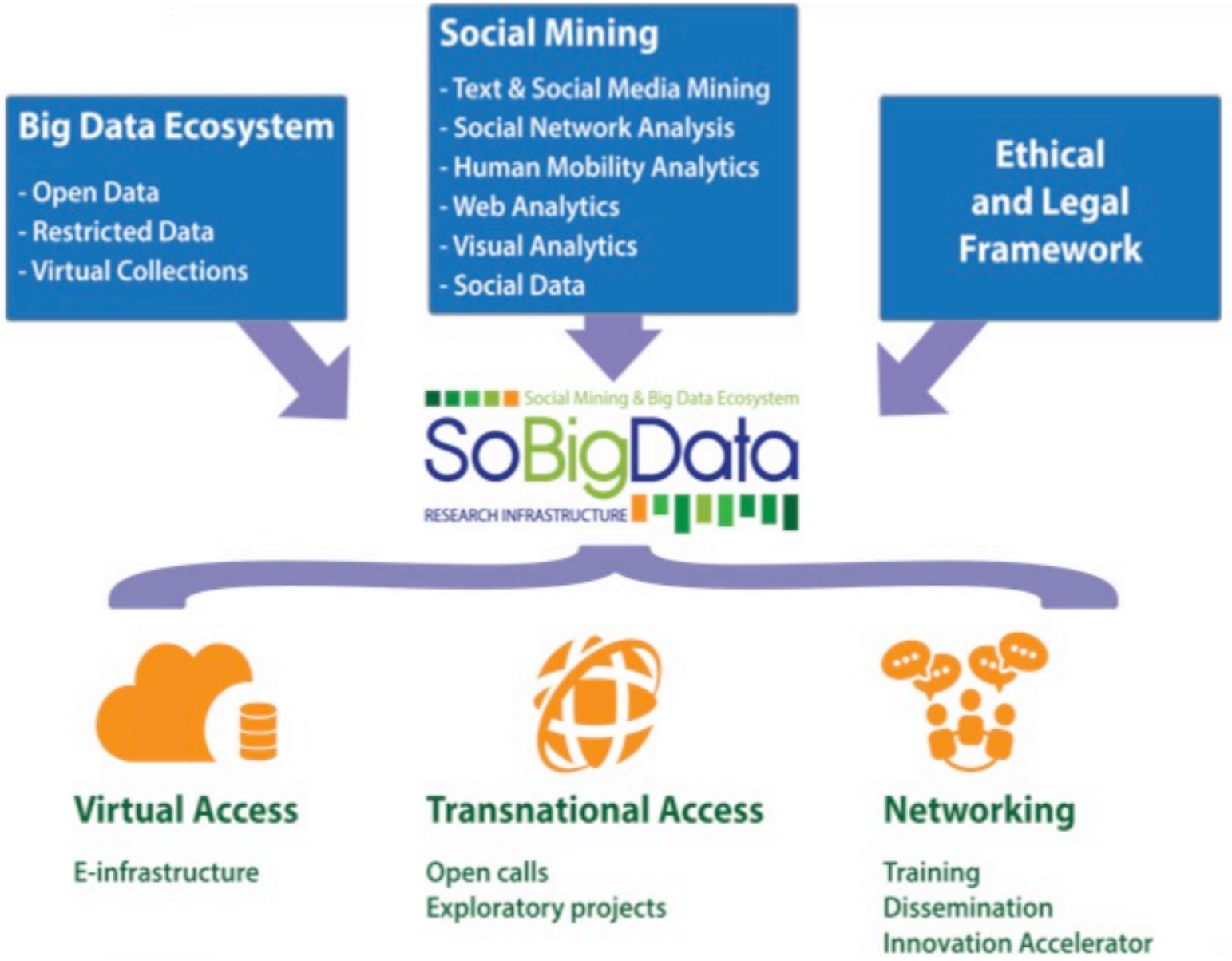
Aalto University

ETH Zürich

TU Delft Delft University of Technology

The path to achieve the goals

- **Grant 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.
- **joint research, and extensive networking and innovation actions**

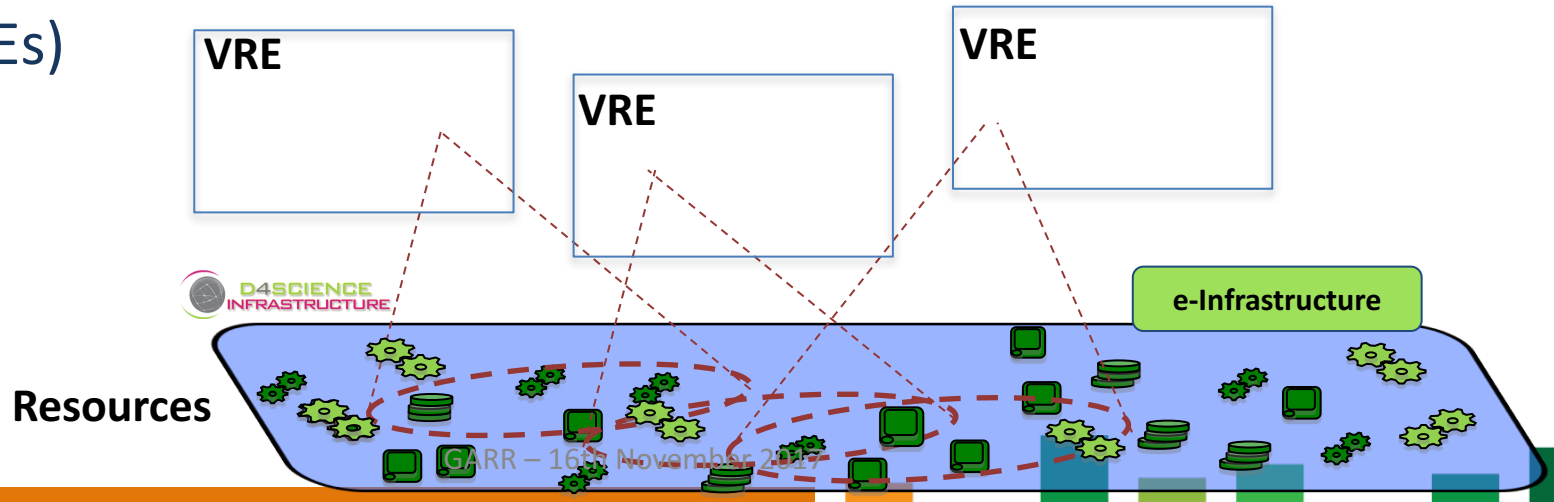

Social Mining
- Text & Social Media Mining
- Social Network Analysis
- Human Mobility Analytics
- Web Analytics
- Visual Analytics
- Social Data
Big Data Ecosystem
- Open Data
- Restricted Data
- Virtual Collections
Ethical and Legal Framework
Social Mining & Big Data Ecosystem
SoBigData
RESEARCH INFRASTRUCTURE
Virtual Access
E-infrastructure
Transnational Access
Open calls
Exploratory projects
Networking
Training
Dissemination
Innovation Accelerator

Lines of actions of the RI

- Overcome data and methods fragmentation
- Attract critical mass of users
- Multidisciplinary community Building
- Creating a new generation of DataScientists
- Improve SME competiveness

SoBigData e-infrastructure

- It is powered by the **D4Science platform**
  - Used as a production system by other communities
    - Users: +3600 (SoBigData is 16.5%, fast growing), Computing: 1500 cores, Storage: 400 TB)
- Supports basic services
  - **Authentication, Authorization, Accounting framework**
  - **Resource Catalogue**
  - **Virtual Research Environments management framework** (VREs)



SoBigData VREs

Exploratories Access Training Dissemination Consortium EU Project Blog

Virtual Access

Home » Access

The Virtual Access offers online services for big data and social mining research. A web front-end comprising of a catalogue of SoBigData resources (**data** and **services**) and a set of SoBigData **Virtual Research Environments** (VREs).

VREs are web-based working environments equipped with a number of applications, enabling scientists to have access to the set of data, services, and algorithms needed to perform their investigation in a **collaborative way**.

Go to the **SoBigData LAB** to perform your experiments

Go to VRE **Exploratories** to see the thematic environments of SoBigData

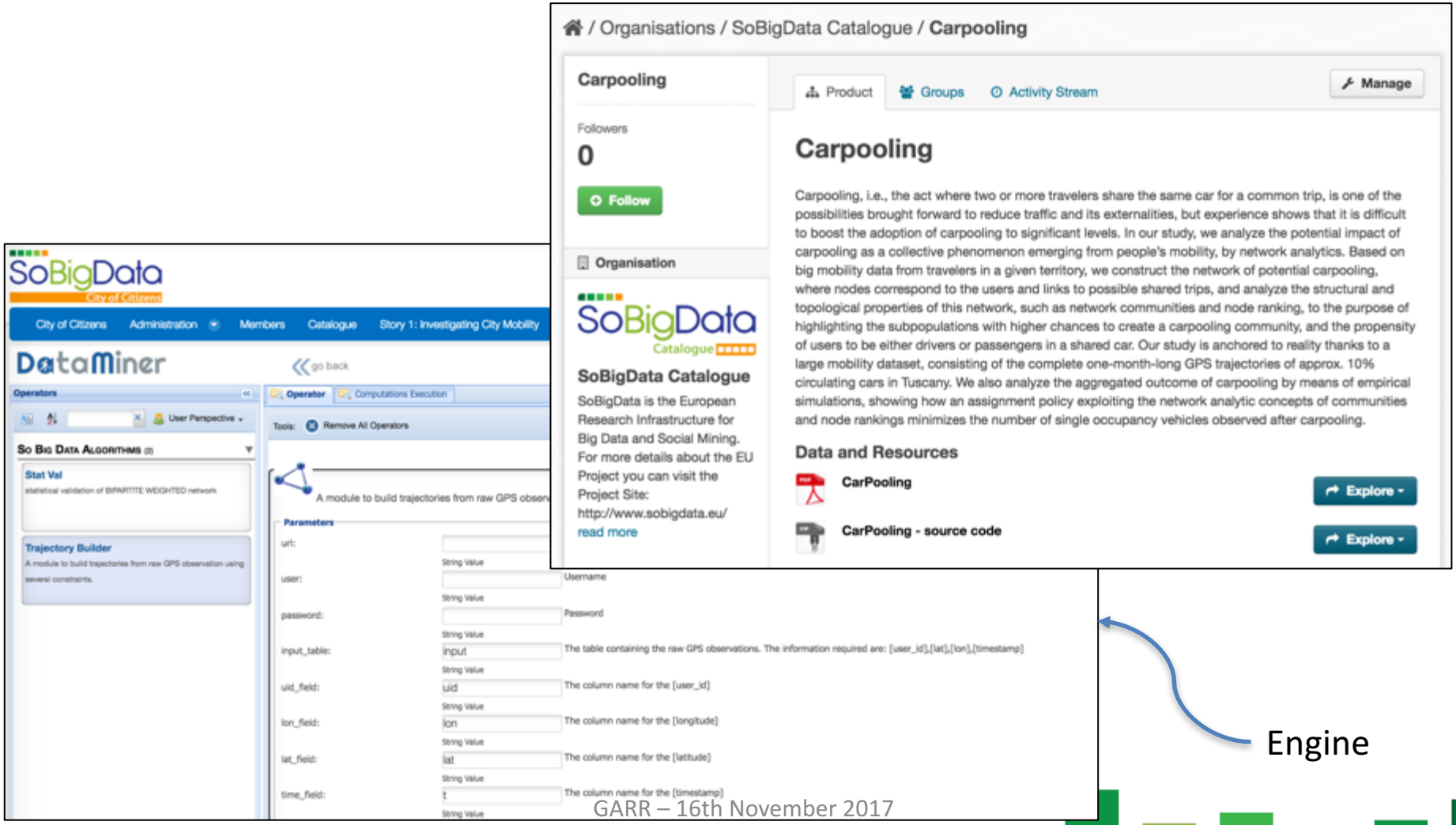
- City of Citizens
- Societal Debates
- Wellbeing
- Migration
- Sport Science

Go to the **Catalogue** to have a list of our methods and datasets

Go to SoBigData **Virtual Research Environments - VREs**

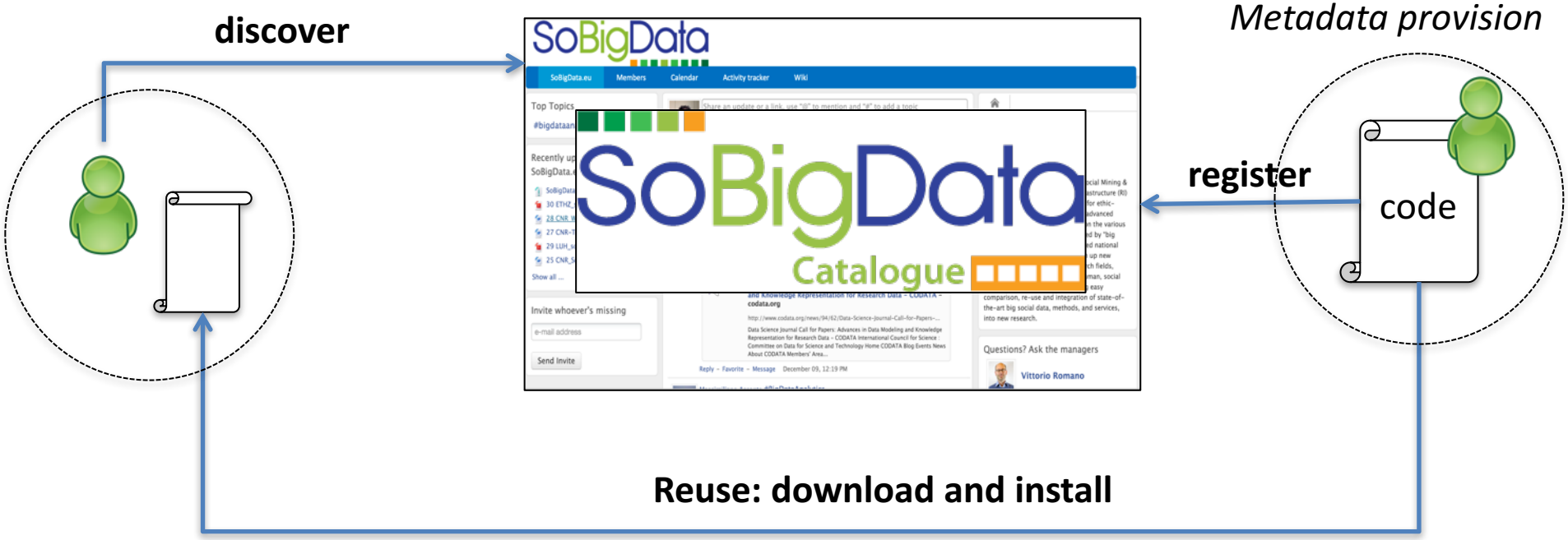
- Tag-me
- Smaph

SoBigData Lab and Catalogue VREs

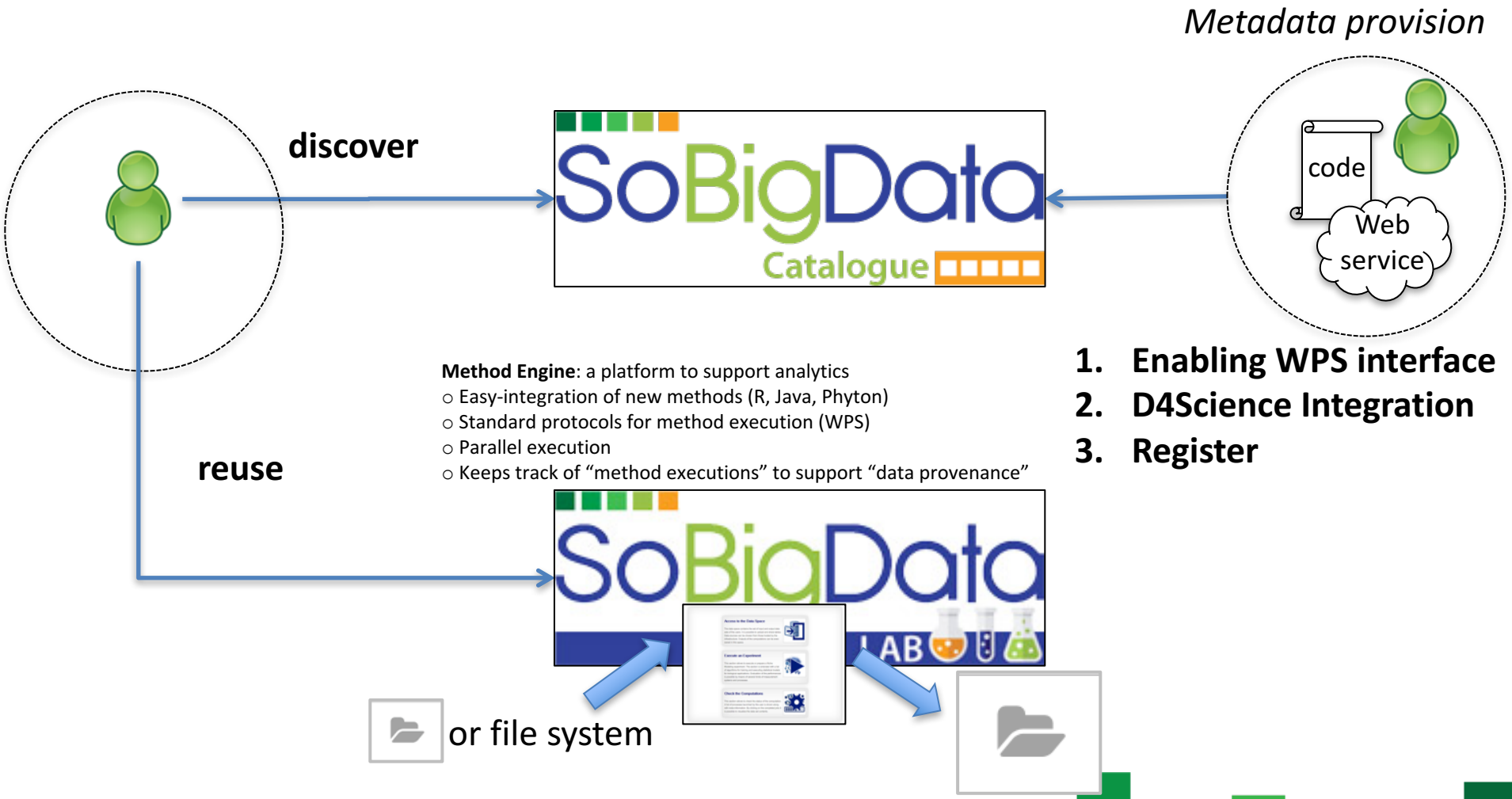


Engine

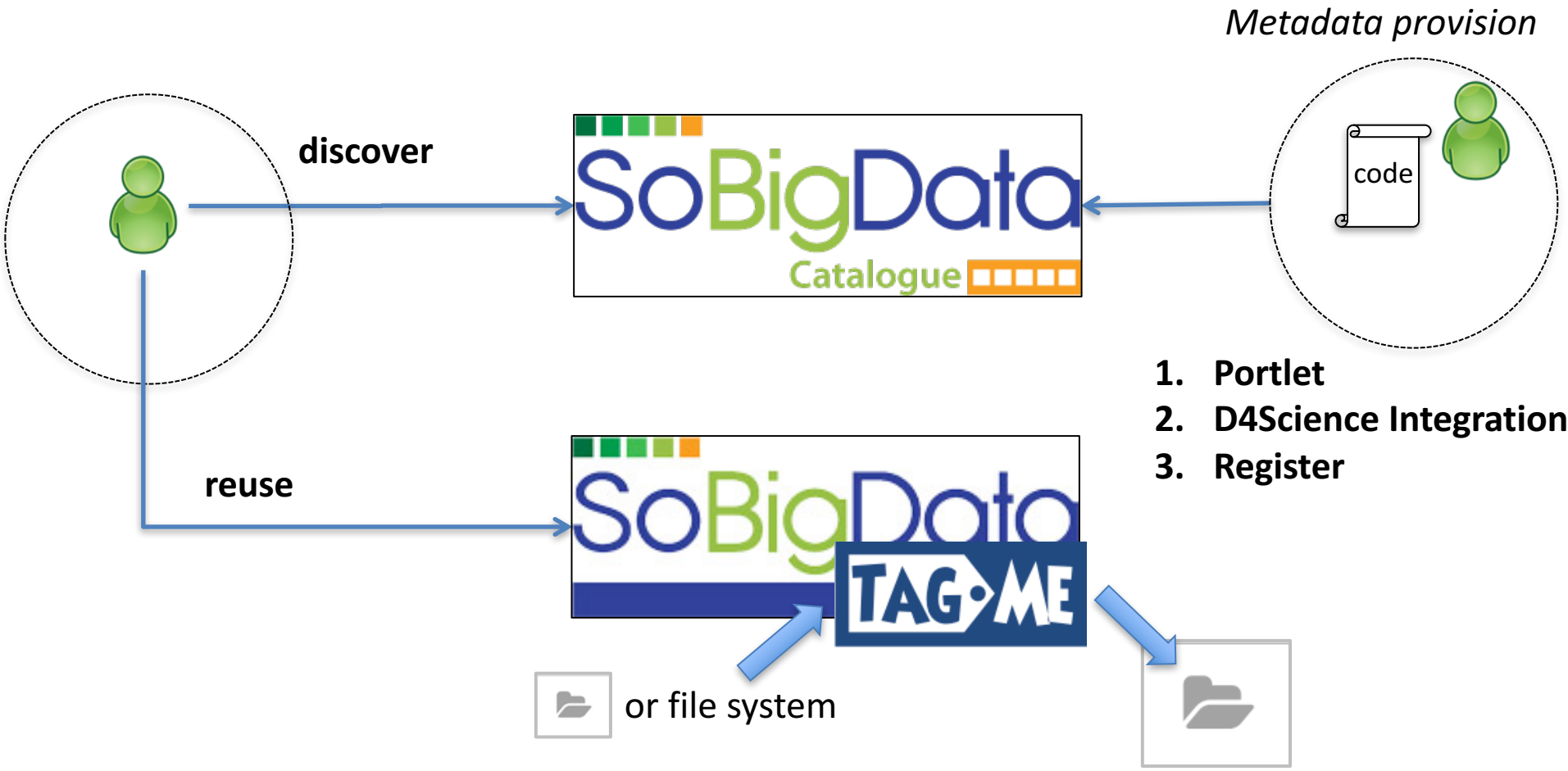
Methods as Software: patterns of sharing and reuse



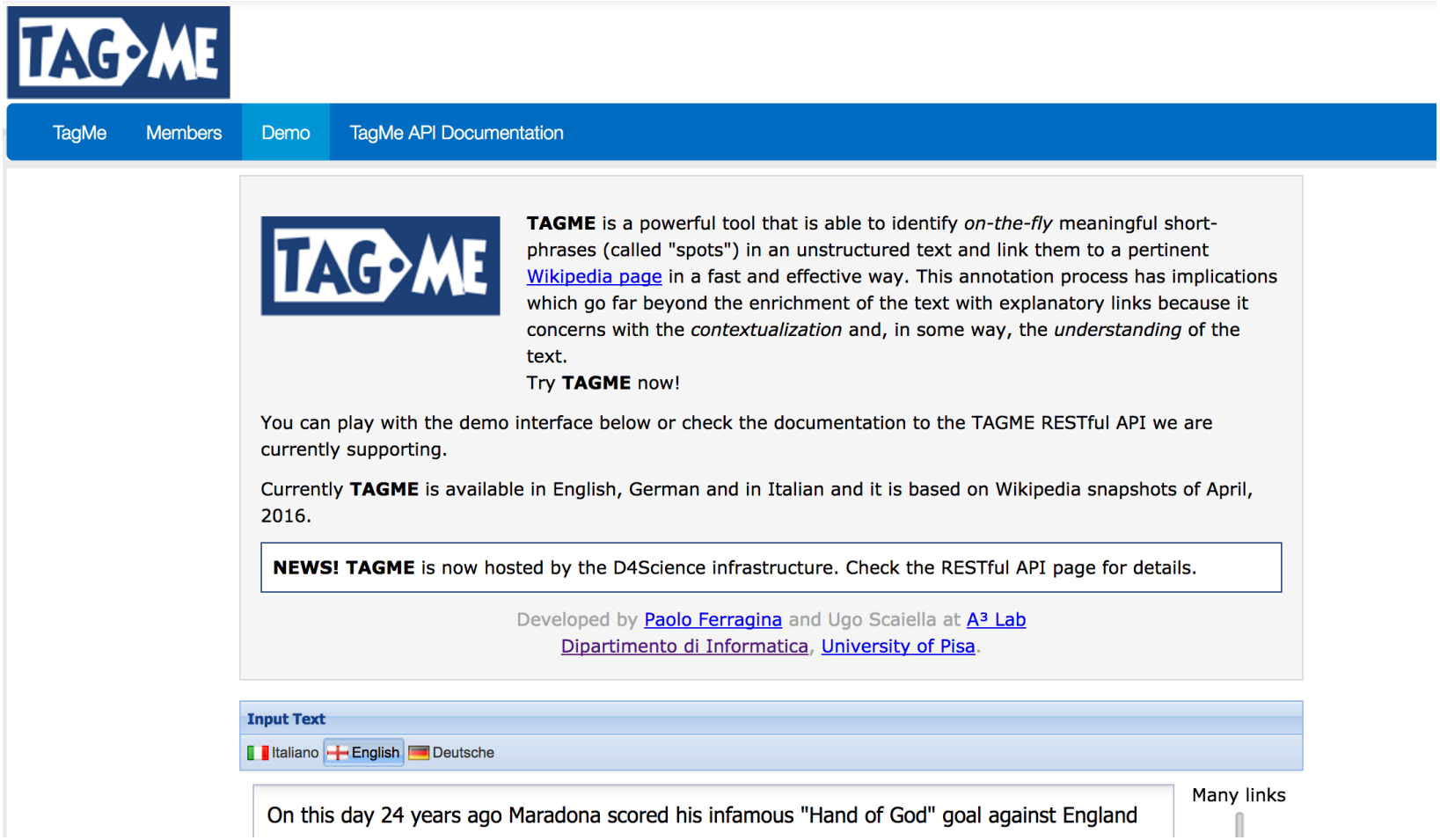
Method as a Service: patterns of sharing and reuse



Web application: patterns of sharing and reuse

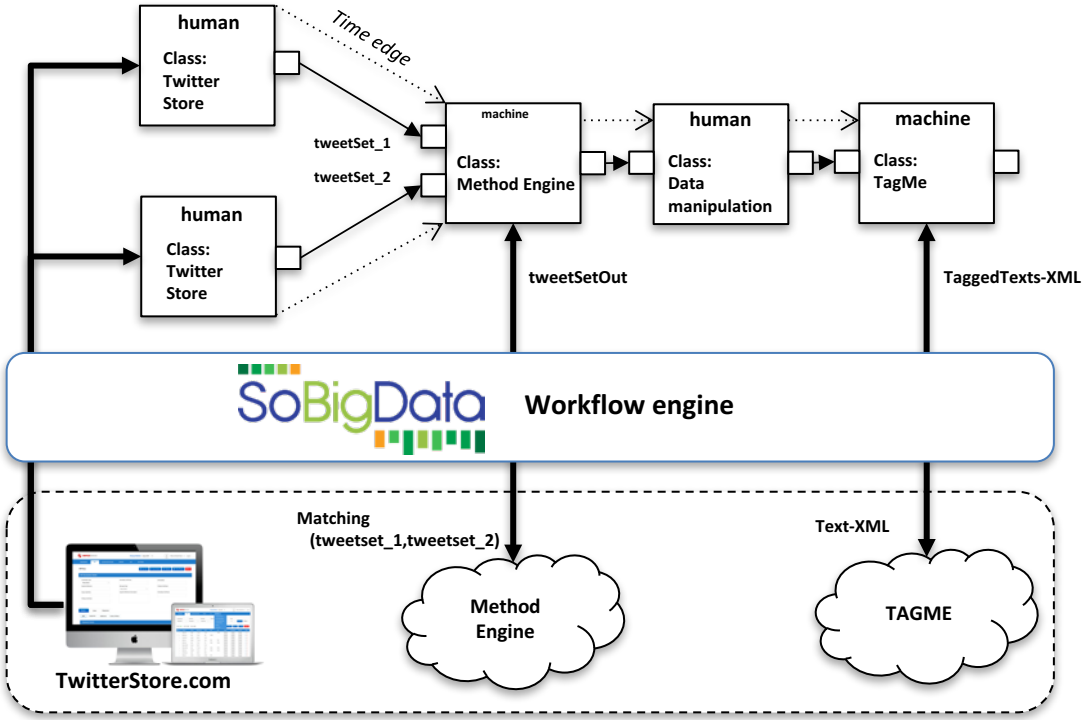


Application VRE



(Hybrid) Workflow Language

- **Issue**: **repeatability is limited to individual methods as a service**
- **Goals**: Workflow language to model and share scientific processes to enable their repeatability
- **Challenges**: enabling repeatability by modeling combinations of manual and autonomic steps



Ongoing research: HyWare: a HYbrid Workflows lAnguage for Research E-infrastructures

Leonardo Candela, Fosca Giannotti, Valerio Grossi, Paolo Manghi, Roberto Trasarti

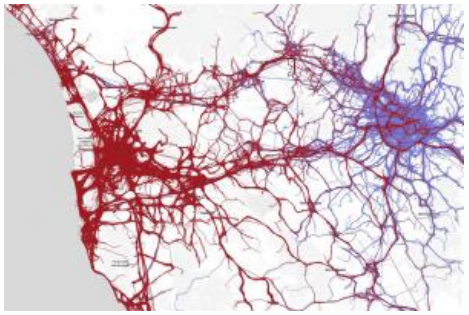
published in DLib Magazine Journal, 2016

Exploratories

Exploratories

Virtual Research Environments tailored on specific domains where Social Mining is applied

- Promotes results sharing among scientists and communities
- Promotes the use of RI through Virtual and Transnational Access

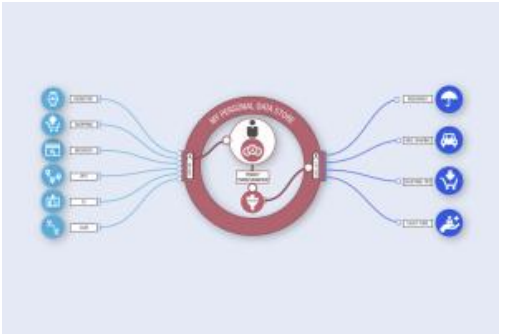


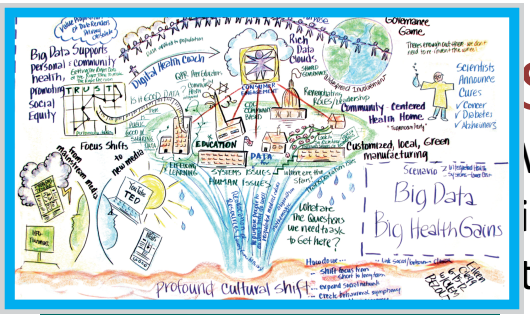
City of Citizens

This exploratory tells stories about cities and people living in it. We describe those territories by means of data, statistics and models.

Well-being & Economic Performance

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.





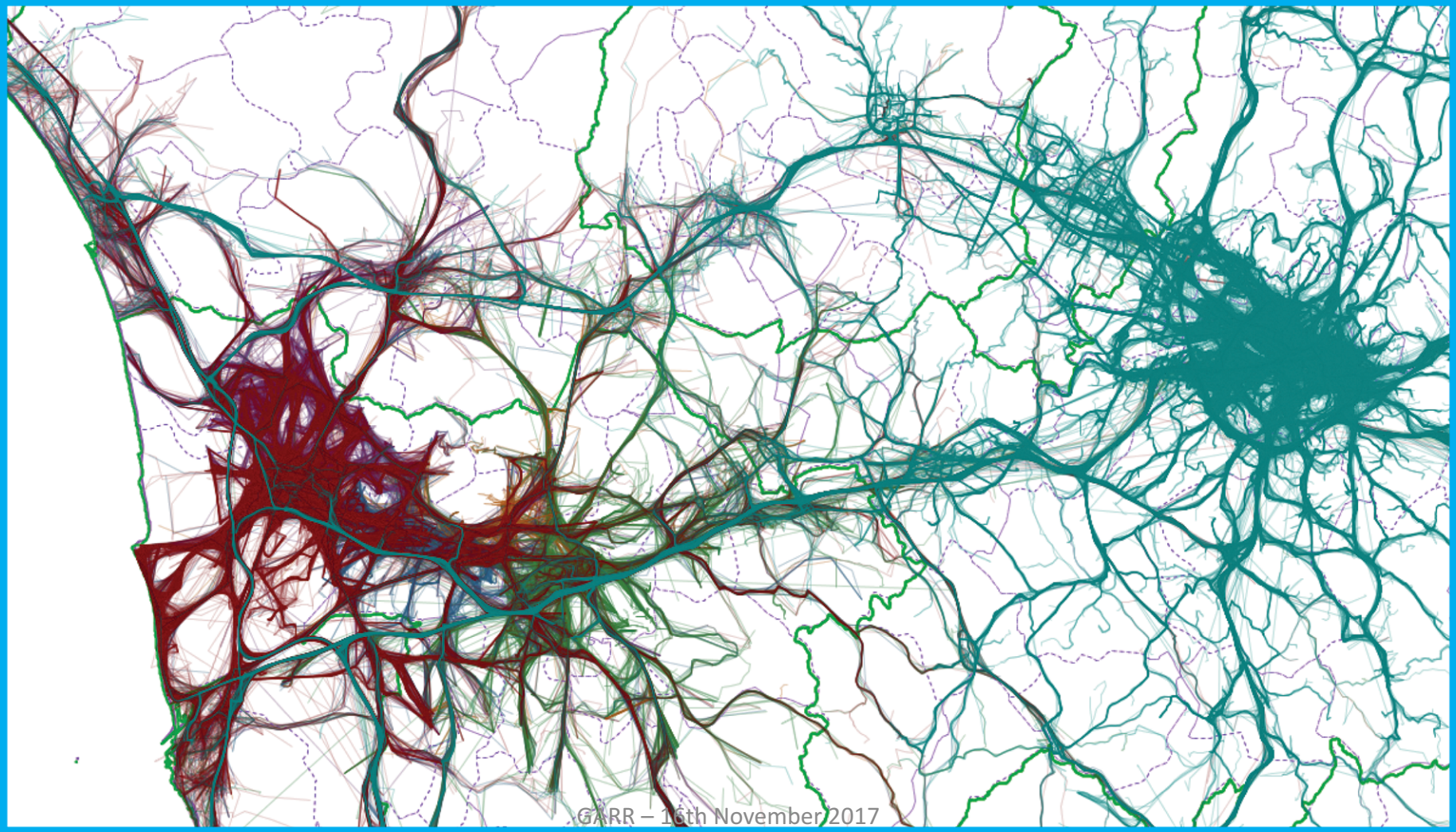
Societal Debates

We study public debates on social media and newspaper. We can identify 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.

Exploratory: Big Data for City of Citizens



Investigating City Mobility

How do people move into the city? How does the traffic change during the day? And how does it vary during the week? How does the turism presence affect the traffic? Our data scientits already study the traffic in the Italian cities of Pisa and Florence by analyzing Big Data sources such as mobile phone traces, veicular gps and social media data as proxy of human behaviour.

The results could be useful for both local administrators and citizens. The local administrators could have a tool to quantify accurately city's traffic and understand city's usage, so they could take better decisions to manage mobility. Citizens could take informations to know traffic situation in real time and they could choose the best and fastest way.

Our studies could be useful in carpooling, too. Indeed, Big Data analysis can suggest to citizens who can share the travel with them.

Datasets | Methods & Tools | Thematic Clusters | People | Contact

- GPS data of private cars
- CDR data from a telecommunication company
- Foursquare check-ins
- Flickr geo-localized photos
- Official administrative information (OAI)

Living the story in SoBigData Virtual Research Environment

(only alpha version available - final service will be available in September 2016)

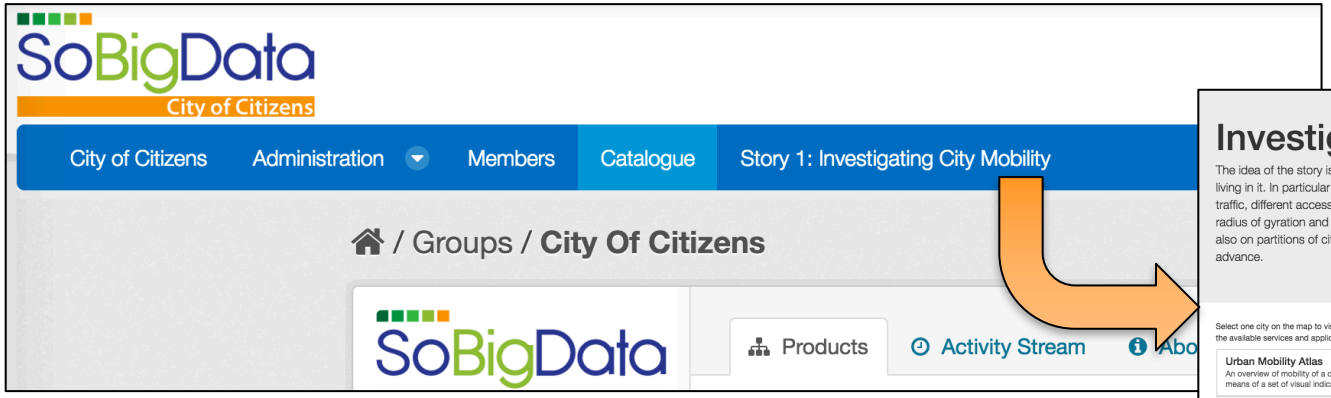
Datasets | Methods & Tools | Workflows | Thematic Clusters | People | Contact

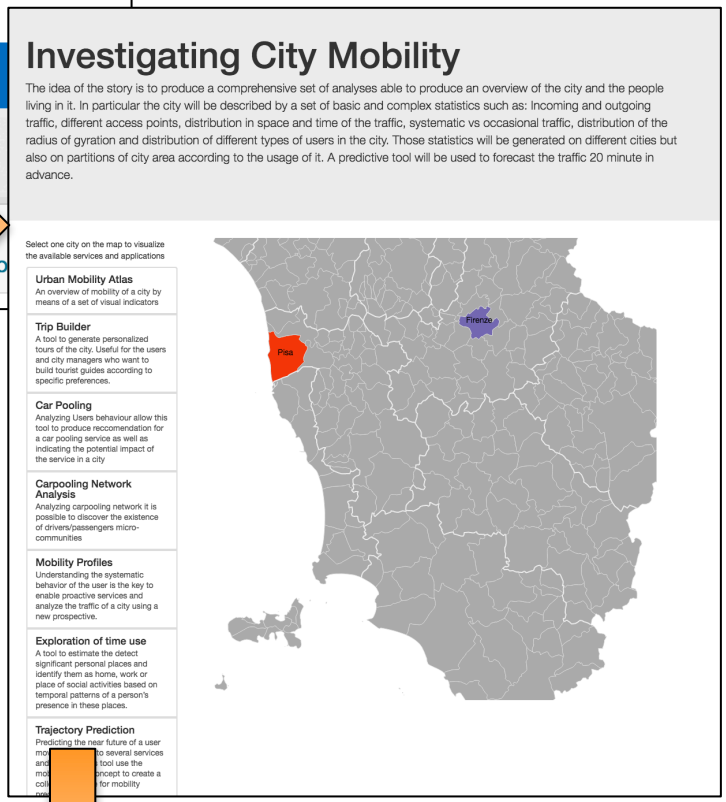
Method	Partner	SoBigData RI - Integration
Urban Profiles	AALTO	Service hosted, Download
Urban Mobility Atlas	CNR	Web Service
Trajectory Builder	CNR	Service hosted, Download
Borders	CNR	
Sociometer	CNR	Download
Trip Builder	CNR	Web Page
Car Pooling	CNR	Service hosted, Download
MyWay	CNR	Service hosted, Download
Privacy Risk	CNR	Service hosted, Download
O/D Matrix	CNR	Web Service
Mobility Profiles	CNR	Service hosted, Download
Exploration of Time	FRH	Download
Statistical Validation	SNS	

Living the story in SoBigData Virtual Research Environment

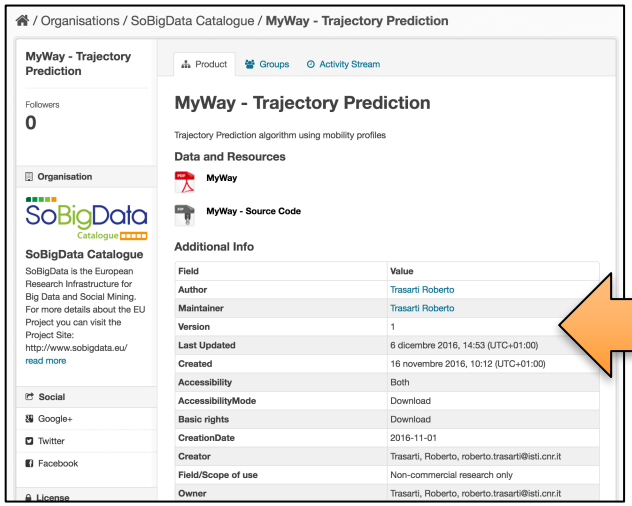
(only alpha version available - final service will be available in September 2016)

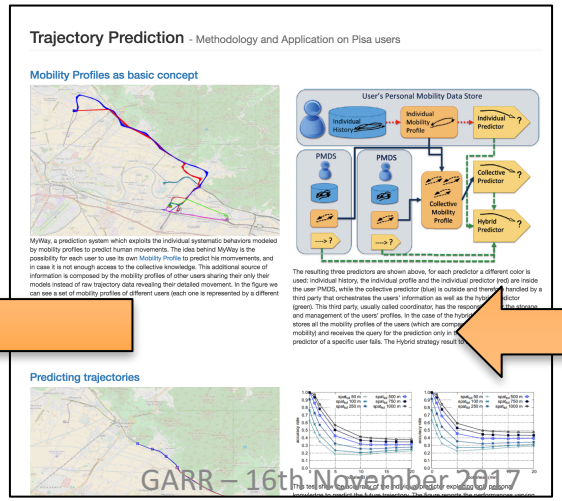
Exploratory VREs



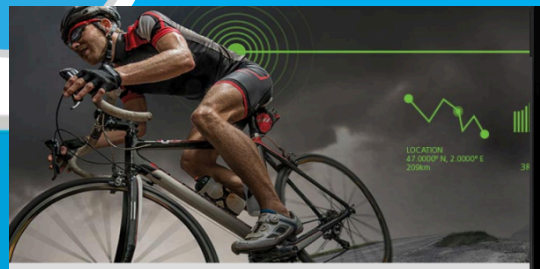


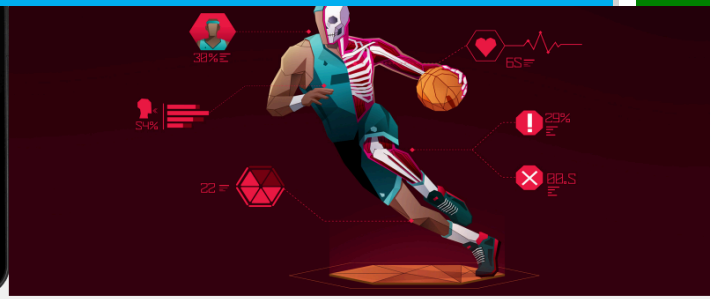
The frontal page shows the workflows, the tools and examples on how to use them in real cases. Each tool is also connected to the catalogue in order to let the user download or access to it.





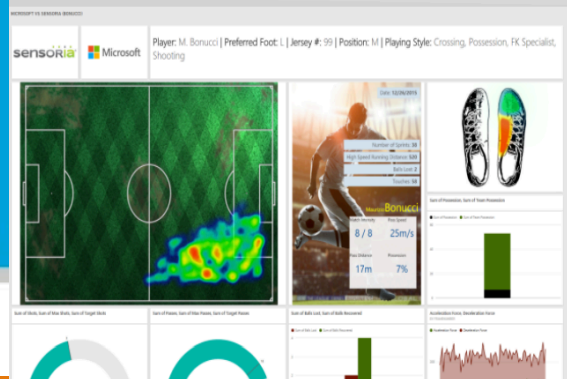
Big Data for Sport Science

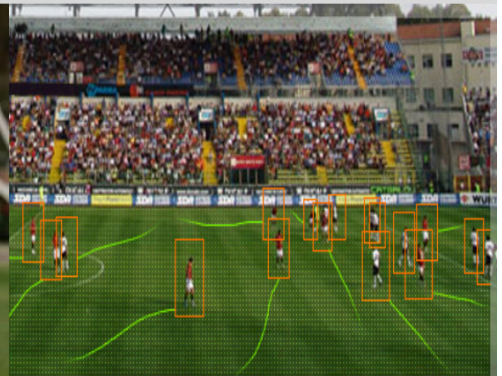




SoBigData

Data Science for sports analytics





Ethics and trust

Legal and Ethical framework

Define and implement the legal and ethical framework of the SoBigData RI, in accordance with the European and national legislations

Monitor of research

Monitor the compliance of experiments and research protocols with the framework

Privacy-by-design

The development of big data analytics and social mining tools with Value-Sensitive Design and privacy-by-design methodologies

The ethics of SoBigData

- Gathering large quantities of data may have serious consequences:
  - consequences range from personal harm,
  - to issues of autonomy, injustice and inequality.
- SoBigData adheres to a *value-sensitive design* approach:
  - design solutions to overcome ethical dilemma's, in this case those between **the utility of the data gathered vs. the protection of the individuals subject to the research.**

The ethics of SoBigData

- How do we create an infrastructure in which such methods can be disseminated and improved upon?
- Data Management Plan plays a key role:
  - Each data has its privacy requirements and fact checks and responsibility
- Anonymization techniques are part of the research
- Researchers will be trained in applying the necessary procedural safeguards

Monreale et al. *EPJ Data Science* 2014, **2014**:10
http://www.epjdatascience.com/content/2014/1/10

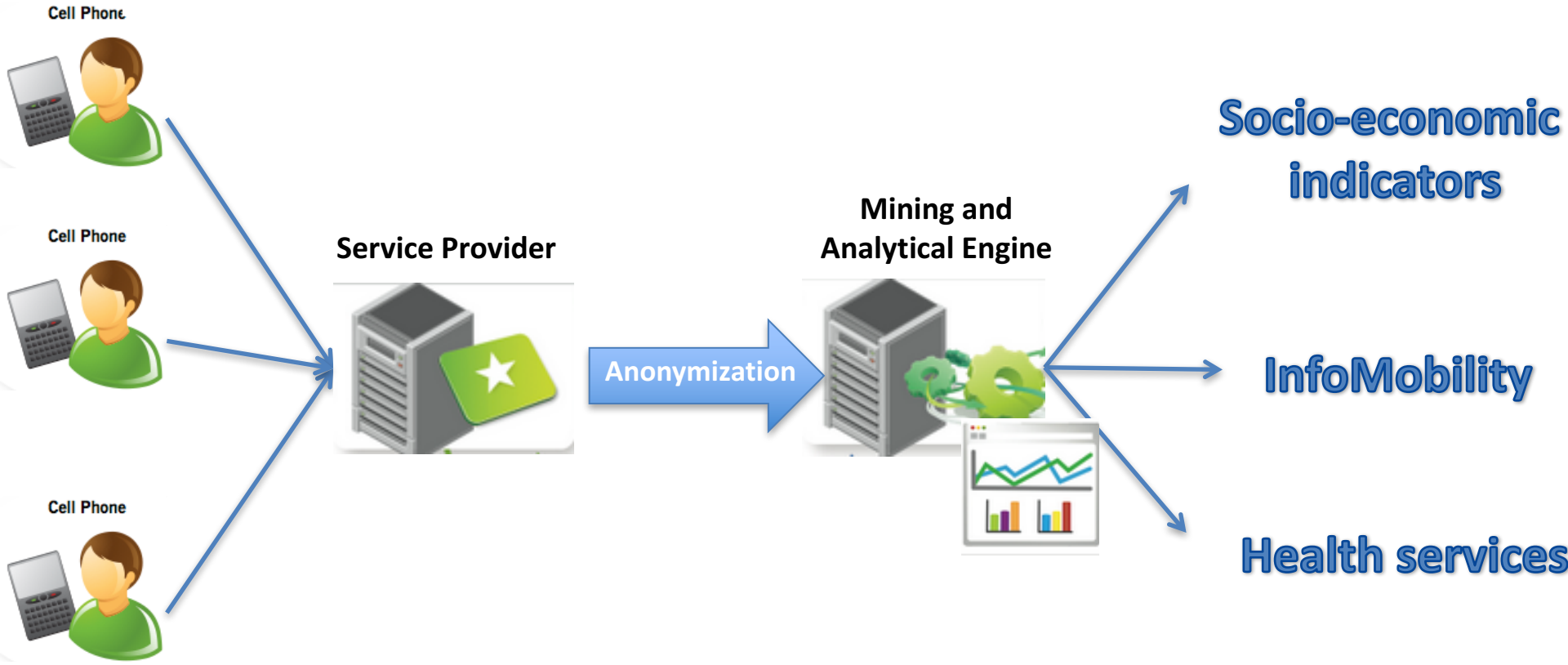
EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Privacy-by-design in big data analytics and social mining

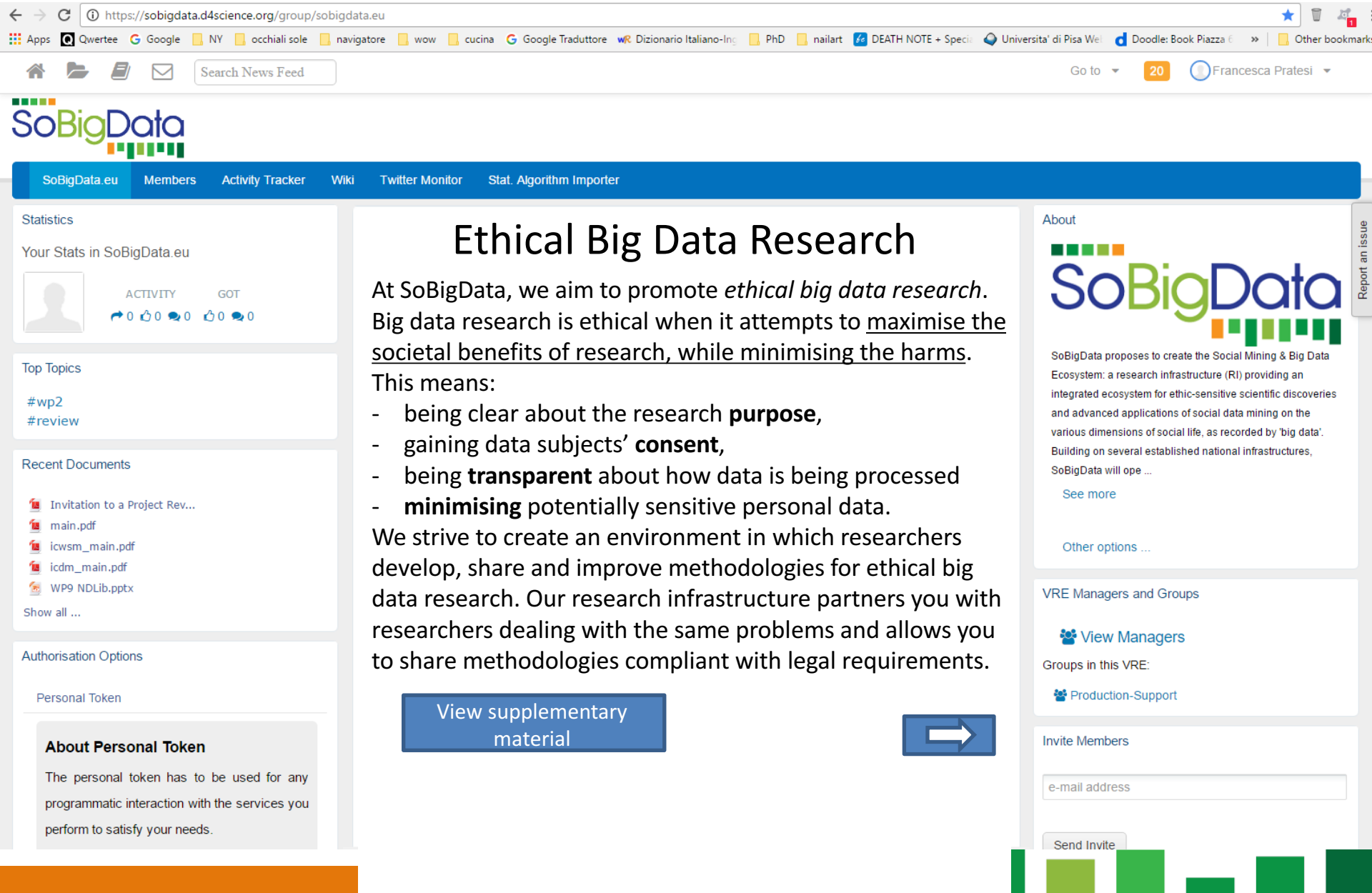
Anna Monreale[1,2*], Salvatore Rinzivillo[2], Francesca Pratesi[1,2], Fosca Giannotti[2] and Dino Pedreschi[1]



Educating the responsible data scientists

Based on a cooperation between ethicists and computer

1. A **M**assive **O**nline **O**pen **C**ourse (MOOC) which instructs all prospective researchers about the legal and ethical dangers of big data research and the steps they can take to minimise these;
2. A set of workflows that outline the steps researchers can take when designing their approach;
3. Information pop-ups which redirect researchers to state-of-the-art ethical methods.


https://sobigdata.d4science.org/group/sobigdata.eu
Search News Feed
Go to
20
Francesca Pratesi
SoBigData
SoBigData.eu
Members
Activity Tracker
Wiki
Twitter Monitor
Stat. Algorithm Importer
Statistics
Your Stats in SoBigData.eu
ACTIVITY
GOT
Top Topics
#wp2
#review
Recent Documents
Invitation to a Project Rev...
main.pdf
icwsm_main.pdf
icdm_main.pdf
WP9 NDLib.pptx
Show all ...
Authorisation Options
Personal Token
About Personal Token
The personal token has to be used for any programmatic interaction with the services you perform to satisfy your needs.
Ethical Big Data Research
At SoBigData, we aim to promote ethical big data research. Big data research is ethical when it attempts to maximise the societal benefits of research, while minimising the harms.
This means:
- being clear about the research purpose,
- gaining data subjects' consent,
- being transparent about how data is being processed
- minimising potentially sensitive personal data.
We strive to create an environment in which researchers develop, share and improve methodologies for ethical big data research. Our research infrastructure partners you with researchers dealing with the same problems and allows you to share methodologies compliant with legal requirements.
View supplementary material
About
SoBigData
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 ope ...
See more
Other options ...
VRE Managers and Groups
View Managers
Groups in this VRE:
Production-Support
Invite Members
e-mail address
Send Invite
Report an issue

https://sobigdata.d4science.org/group/sobigdata.eu

Search News Feed

Go to 20 Francesca Pratesi

SoBigData

SoBigData.eu | Members | Activity Tracker | Wiki | Twitter Monitor | Stat. Algorithm Importer

Statistics

Your Stats in SoBigData.eu

ACTIVITY GOT

0 0 0 0 0 0

Top Topics

#wp2
#review

Recent Documents

Invitation to a Project Rev...
main.pdf
icwsm_main.pdf
icdm_main.pdf
WP9 NDLib.pptx

Show all ...

Authorisation Options

Personal Token

About Personal Token

The personal token has to be used for any programmatic interaction with the services you perform to satisfy your needs.

Data Protection Law in the EU

At present, the centerpiece of the European data protection legislation is the Data Protection Directive (DPD) and implementing national laws. From 25 May 2018 a new legal instrument - the General Data Protection Regulation (**GDPR**) – will be directly applicable in all Member States.

View 95/46/EC Dir. | View GDPR

Personal data are “any information relating to an identified or identifiable natural person ('data subject') […]” (GDPR - Article 4(1))
Sensitive data are […]

About

SoBigData

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 ope ...

See more

Other options ...

VRE Managers and Groups

View Managers

Groups in this VRE:

Production-Support

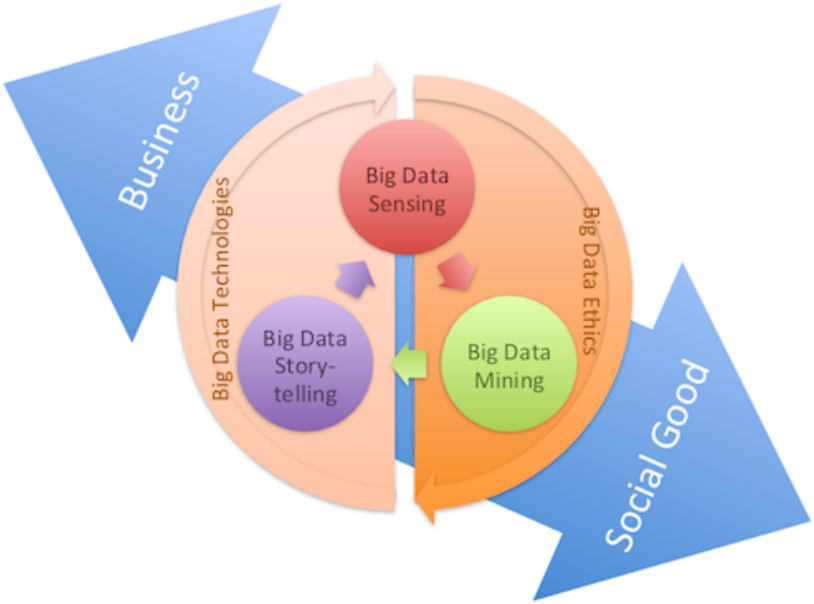
Invite Members

e-mail address

Send Invite

Report an issue

Creating a new generation of data scientists



10 events and 537 trainees

Datathons

- Zurich and Delft : young startuppers (40 participants)
- London: journalists
- Tartu: social scientist (220 participants)
- Pisa: High School students
  - II level Master Bigdata Analytics (4th edition)
  - Master degree in data science (since 2004)
- Summer School
- Phd programme in Data Science in Pisa

An initiative devoted to **Tuscan** enterprises interested in exploring the potential of **big data** in improving the production processes, their market or their interaction with customers and suppliers.

The Tuscan Data Challenge gives opportunity to all the Tuscan companies, that usually do not use their **data**, to enter the world of Big Data and exploit its potential.

Who is Involved

Consiglio Nazionale delle Ricerche

Università di Pisa

IMT Institute for Advanced Studies Lucca

Scuola Normale Superiore

Tuscan Big Data Challenge 2016

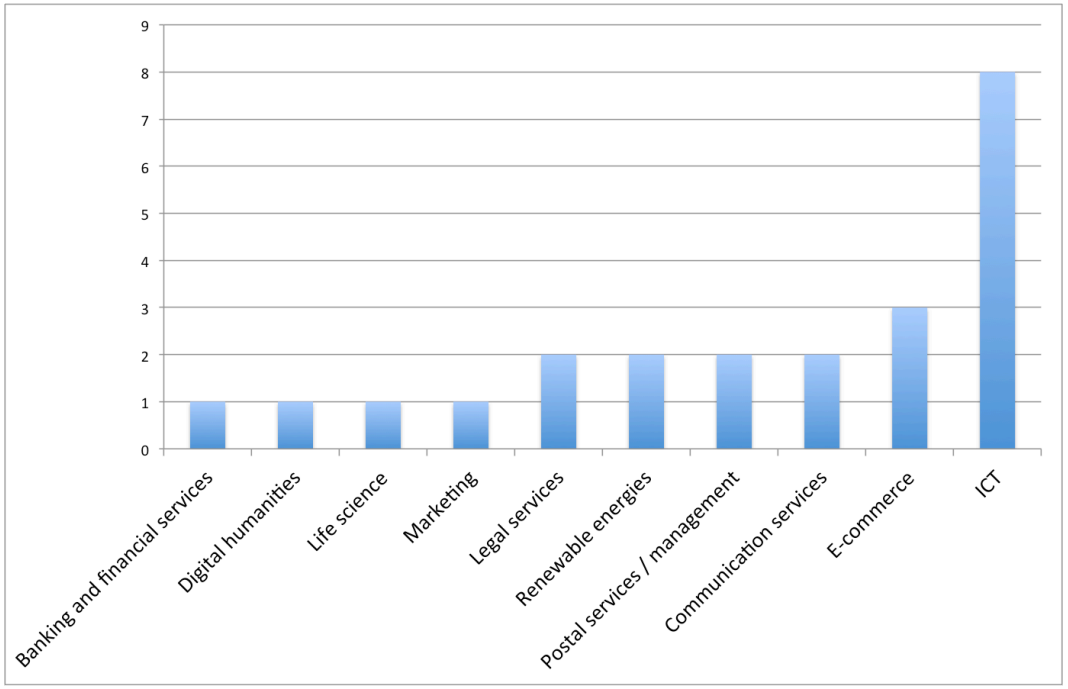
Results in Short

Participants: 23 Tuscan SMEs

Selected: 9

Type:

- Legal services
- Banking and financial services
- E-commerce
- ICT
- Renewable energies
- Postal services
- Digital humanities
- Life science
- Communication services
- Marketing



SoBigData TRANSNATIONAL ACCESS

Transnational Access supports **short term scientific mission** (between 2 weeks and 2 months) of researchers or teams at one of the installations of SoBigData that will provide **big data computing platforms**, **big social data resources**, and cutting-edge **computational methods**:

- Interact with the local experts, discuss research questions
  Run experiments on non-public datasets and algorithms
  Present results at workshops/seminars
- **to enable** multi-disciplinary social mining experiments with the Research Infrastructure **assets**: *big data sets, analytical tools, services and skills*.

Site: **http://www.sobigdata.eu/access/transnational**

Contacts: ***Gate*** - Kalina Bontheva <k.bontcheva@sheffield.ac.uk>
SoBigData.it - Claudio Lucchese <claudio.lucchese@isti.cnr.it>
Fraunhofer - Thorsten May <thorsten.may@igd.fraunhofer.de>
UT - Jaan Ubi <jaan.ubi@ut.ee>
L3S - Thomas Risse <risse@L3S.de>
AALTO UNIVERSITY - Michael Mathioudakis <michael.mathioudakis@aalto.fi>
Nervousnet - Iza Moise <izabela.moise@gess.ethz.ch>

Installations

SoBigData Euro Lab on Big Data Analytics & Social Mining
GATE general architecture for text engineering
Fraunhofer IGD
L3S Research Center
Aalto University
Tartu Ülikool 1632 Universitas Tartuensis
nervousnet

KEY DATES

Theme-driven exploratory calls:

Publish Call for Proposals: ***mid October 2017***
Proposal Submission deadline: ***mid November 2017***
Applicant Notification: ***December 2017***
Visits period: ***January – June 2018***

www.sobigdata.eu
SoBigData SoBigData

Consiglio Nazionale delle Ricerche · Università di Pisa · Scuola Normale Superiore · IMT Institute for Advanced Studies Lucca · The University of Sheffield. · King's College London

Fraunhofer FHR · Leibniz Universität Hannover · Tartu Ülikool 1632 · Aalto University · ETH zürich · TU Delft

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

SoBigData Research Infrastructure

Become a SoBigData Supporter

- Share and make findable your data science results
- Become part of the scientific network
- http://www.sobigdata.eu/join-us