SoBigData Research Infrastructure

Social Mining & Big Data Analytics

RRI & Data Science
Anna Monreale

Dipartimento di Informatica
Università di Pisa

KDD Lab

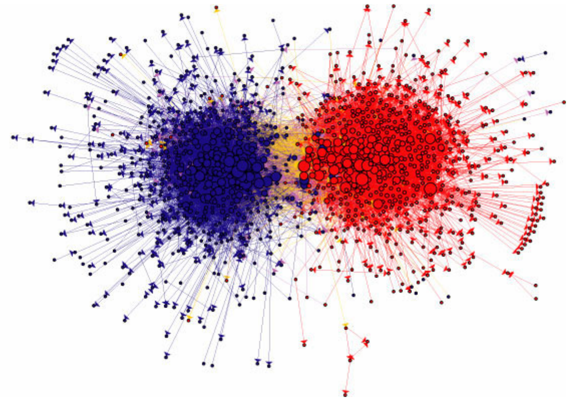
Our digital traces ….

- We produce an unthinkable amount of data while running our daily activities.
- How can we manage all these data? Can we get an added value from them?

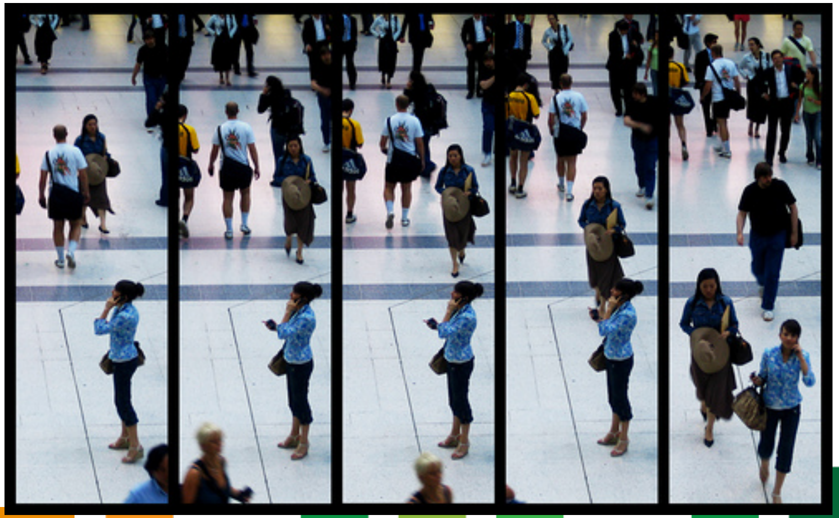
Big data “proxies” of social life

Shopping patterns & lifestyle

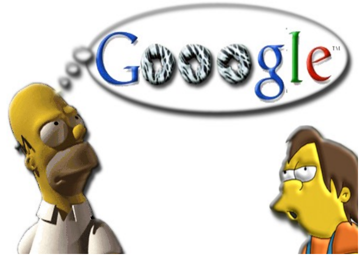
Relationships & social ties

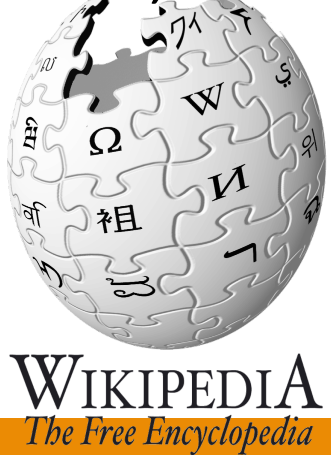


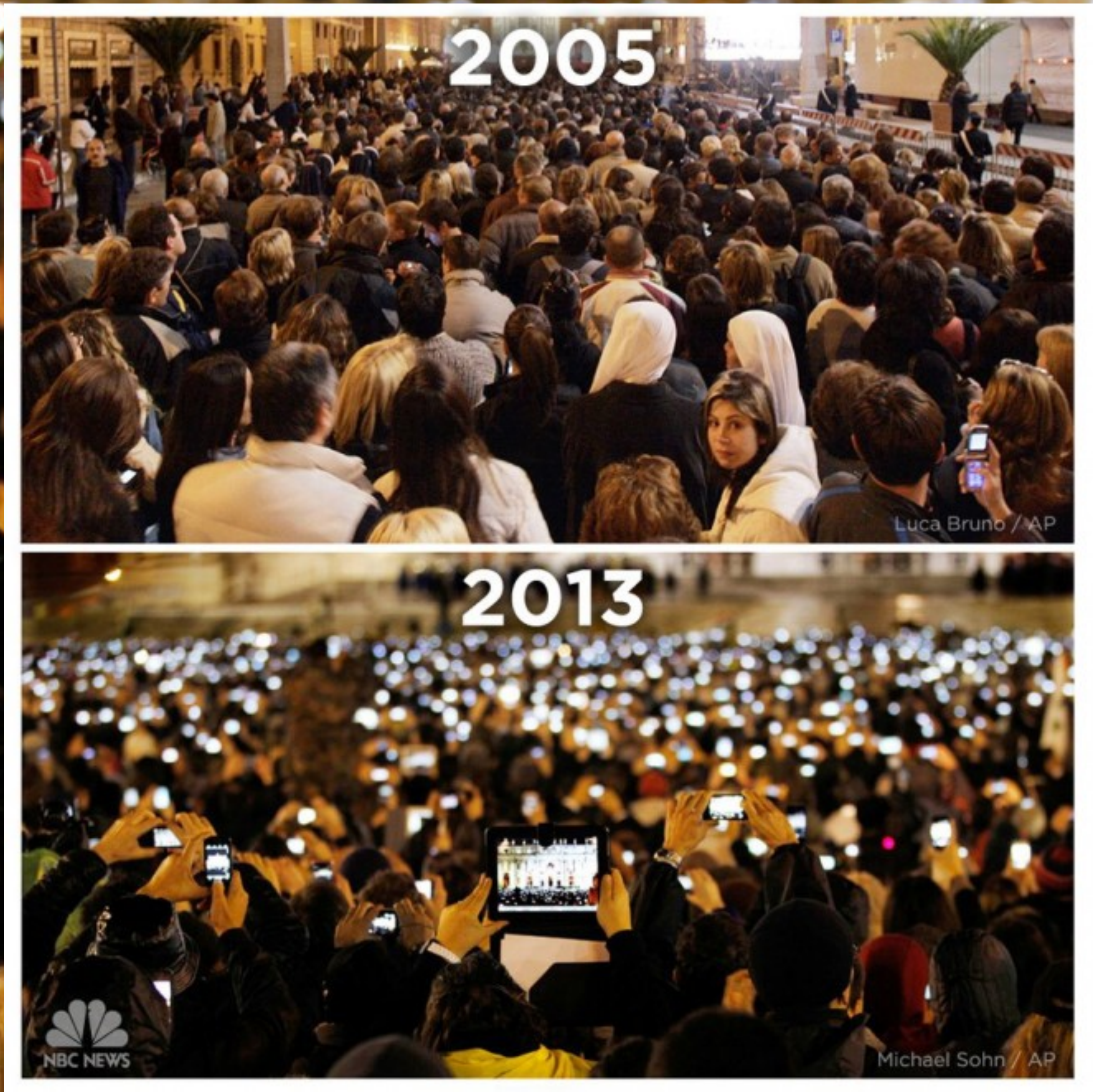
Movements



Desires, opinions, sentiments

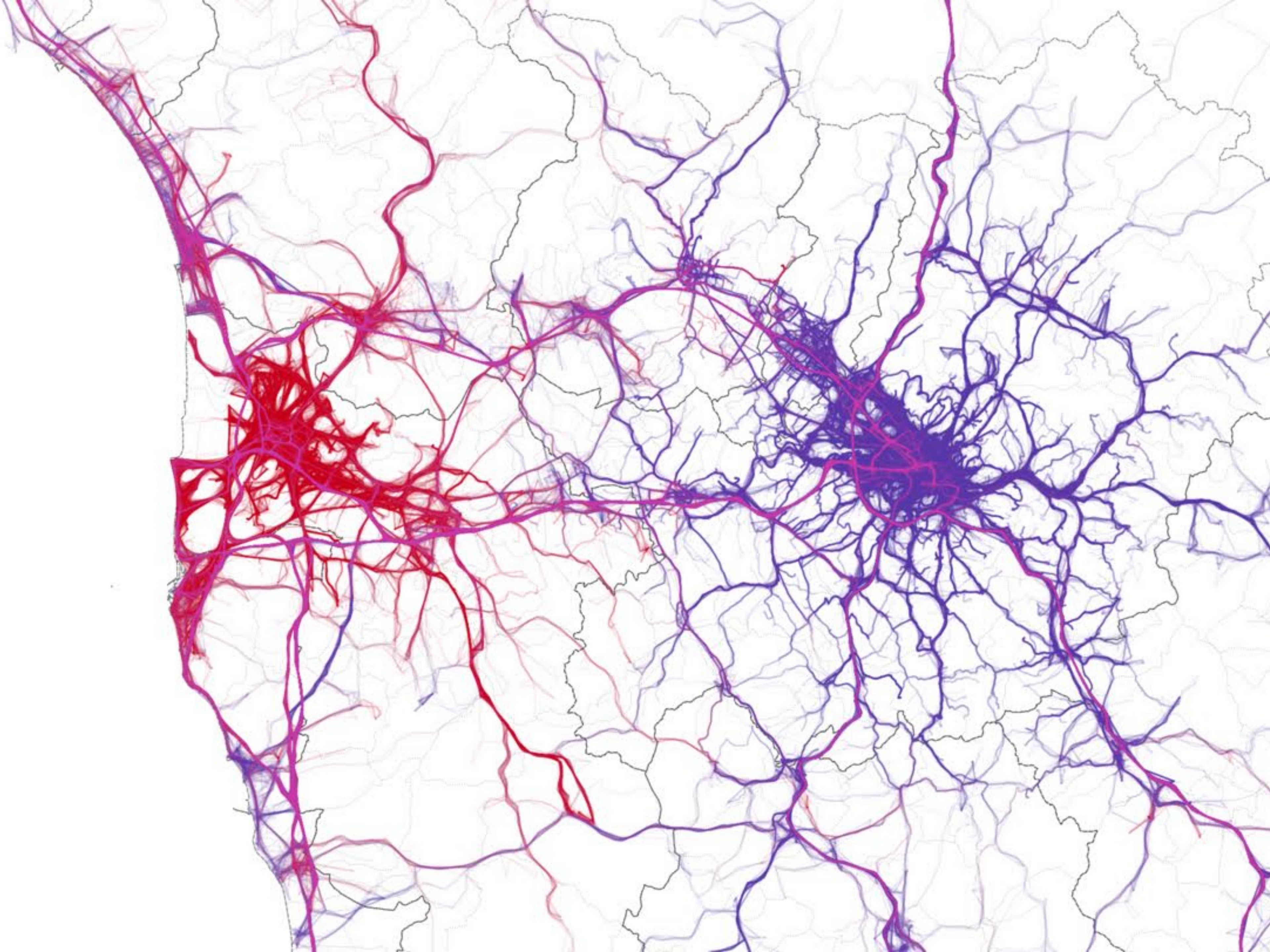




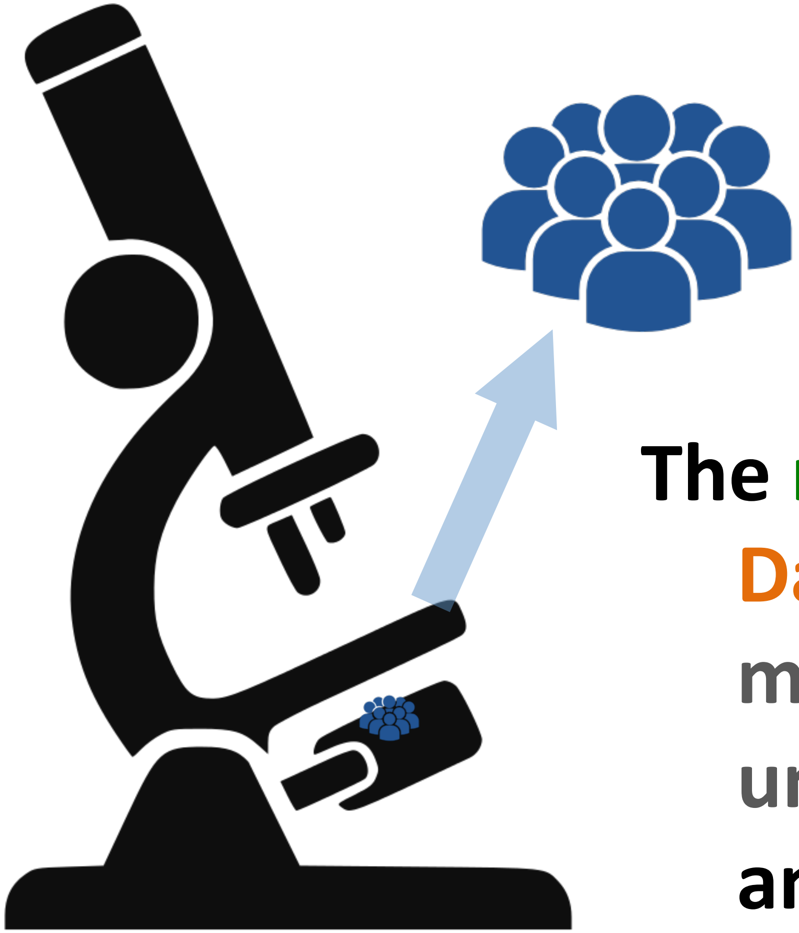


Luca Bruno / AP

Michael Sohn / AP

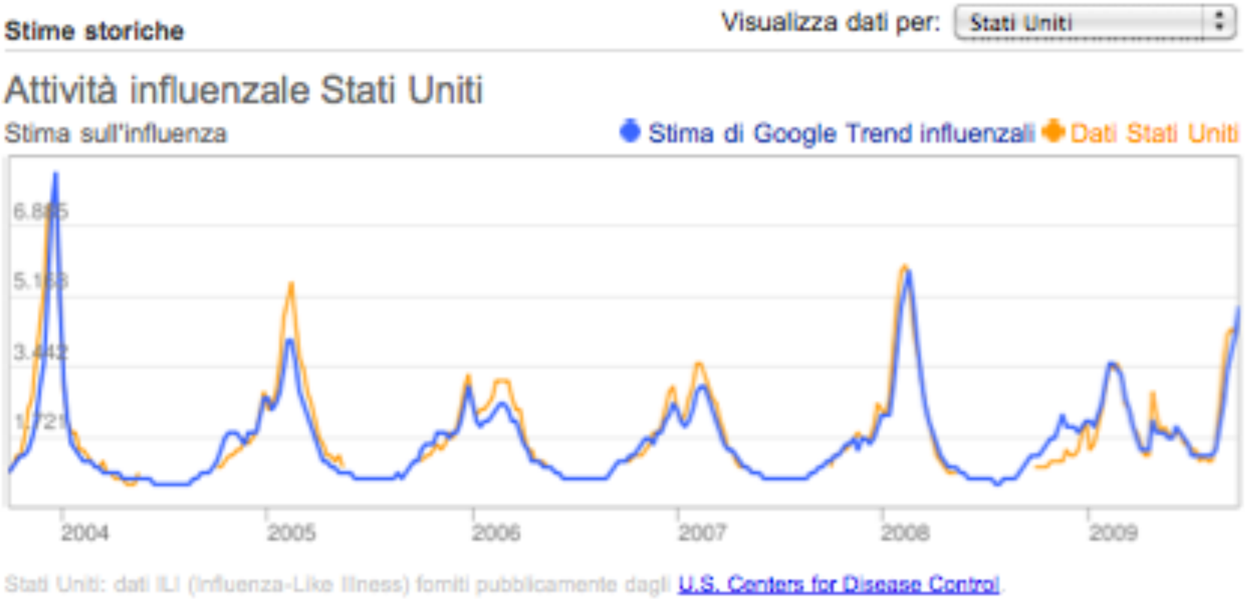


Big Data Analytics & Social Mining



The main tool for a Data Scientist to measure, understand, and possibly predict human behavior

Nowcasting epidemics



Stime storiche

Visualizza dati per: Stati Uniti

Attività influenzale Stati Uniti

Stima sull'influenza

Stati Uniti: dati ILI (Influenza-Like Illness) forniti pubblicamente dagli U.S. Centers for Disease Control.

Google

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

[1]Google Inc. [2]Centers for Disease Control and Prevention

Nature 457, 1012-1014 (2009)

Mobility Atlas of a City

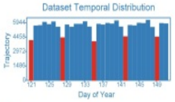
Pisa

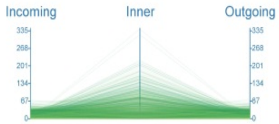
Surface area: 193 km²

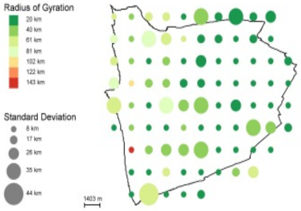
Coordinates: 43,67 10,35

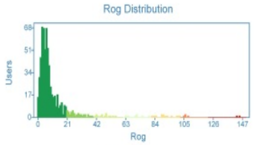
Vehicles: 13.193

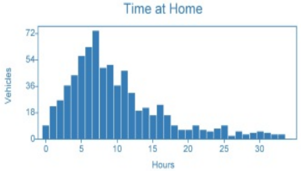
From: 2011-05-01 To: 2011-05-31

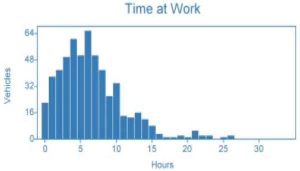

Dataset Temporal Distribution



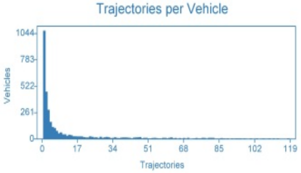


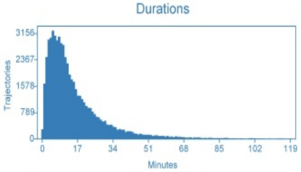

Rog Distribution

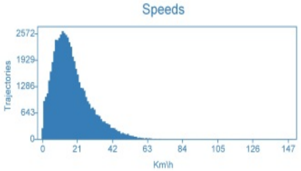

Time at Home

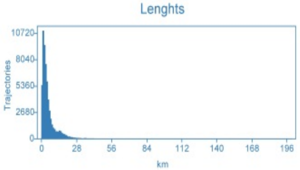

Time at Work

Inner Traffic (44.435 Trajectories)

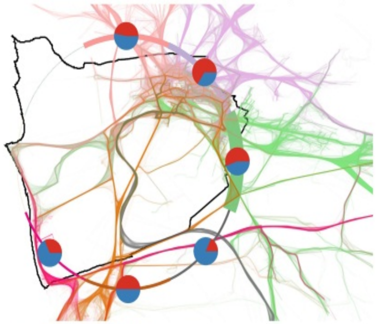

Trajectories per Vehicle


Durations

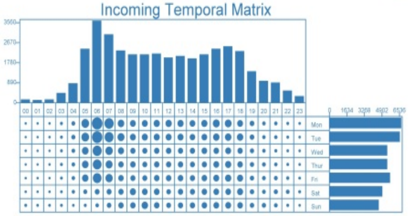

Speeds

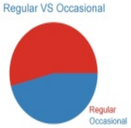

Lenghts

Incoming Traffic (38.464 Trajectories)

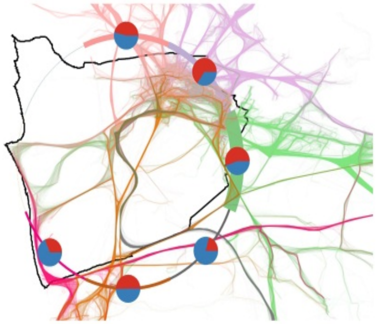


	City	Traj	Perc
NORD 32%	San Giuliano T..	4.816	62%
	Vecchiano	1.425	94%
	Viareggio	1.142	99%
	Lucca	862	67%
	Camaiore	358	94%
OVEST 0%			
SUD 12%	Livorno	2.843	92%
	Collesalvetti	565	50%
	Rosignano Mari..	140	41%
	Fauglia	137	19%
	Cecina	124	45%
EST 54%	Cascina	7.078	97%
	San Giuliano T..	2.881	37%
	Pontedera	1.350	95%
	Calci	795	79%
	Calcinaia	693	92%

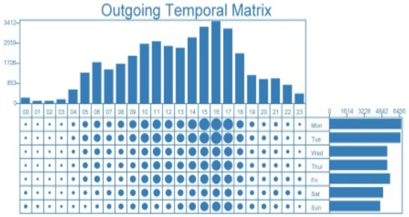

Incoming Temporal Matrix

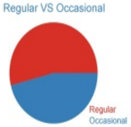

Regular VS Occasional

Outgoing Traffic (38.271 Trajectories)

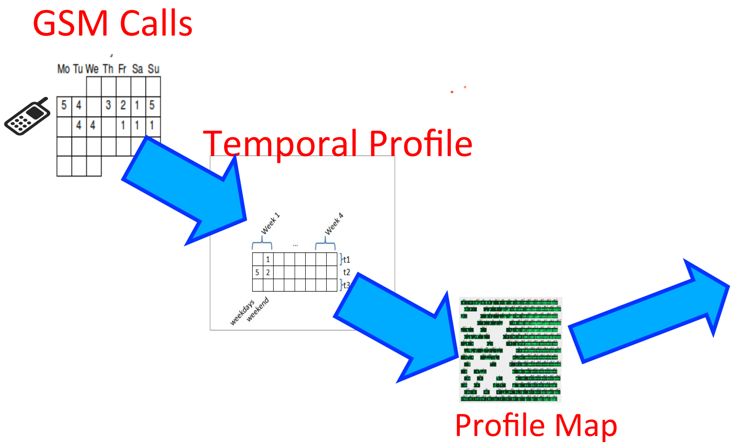


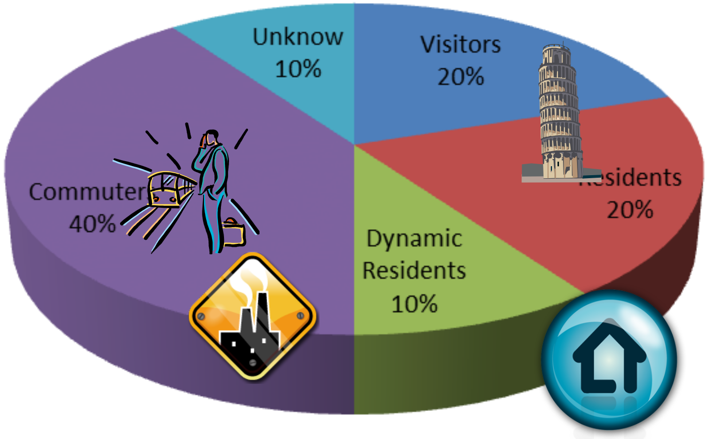
	City	Traj	Perc
NORD 32%	San Giuliano T..	4.842	62%
	Vecchiano	1.418	93%
	Viareggio	1.117	99%
	Lucca	886	67%
	Camaiore	329	96%
OVEST 0%			
SUD 13%	Livorno	2.812	92%
	Collesalvetti	565	51%
	Rosignano Mari..	143	44%
	Fauglia	130	19%
	Cecina	123	45%
EST 54%	Cascina	7.253	97%
	San Giuliano T..	2.860	37%
	Pontedera	1.326	95%
	Calci	798	82%
	Calcinaia	704	93%


Outgoing Temporal Matrix


Regular VS Occasional

Estimation of users city categories by mobile phone data



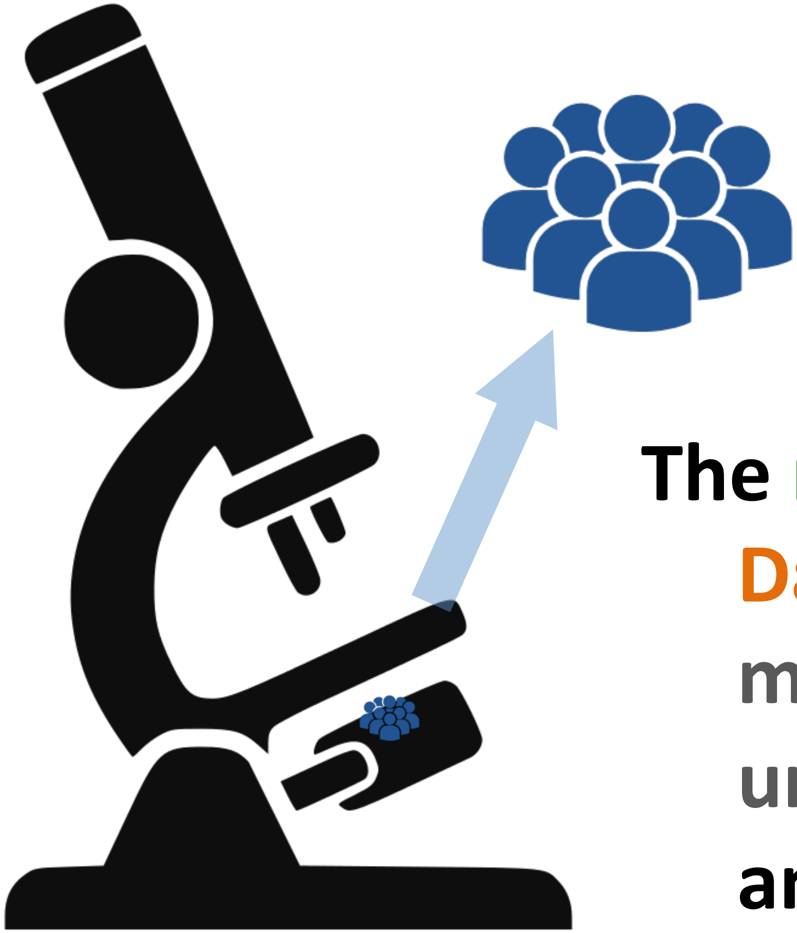


ISTITUTO DI SCIENZA E TECNOLOGIE
DELL'INFORMAZIONE "A. FAEDO"

UNIVERSITÀ DI PISA

WIND
Più vicini.

Big Data Analytics & Social Mining



The main tool for a
Data Scientist to
measure,
understand,
and possibly predict
human behavior

Data Scientist needs to take into account ethical and legal aspects and social impact of data science

The perils of big data & data science

- Not only “**privacy**” – (lack of) **protection of personal data**
- Lack of **transparency** on use of data
- Huge **asymmetry** between users’ and company’s information
- **Access to data**, even own’s personal digital traces
- Potential **discrimination** due to profiling

Legislation Knowledge

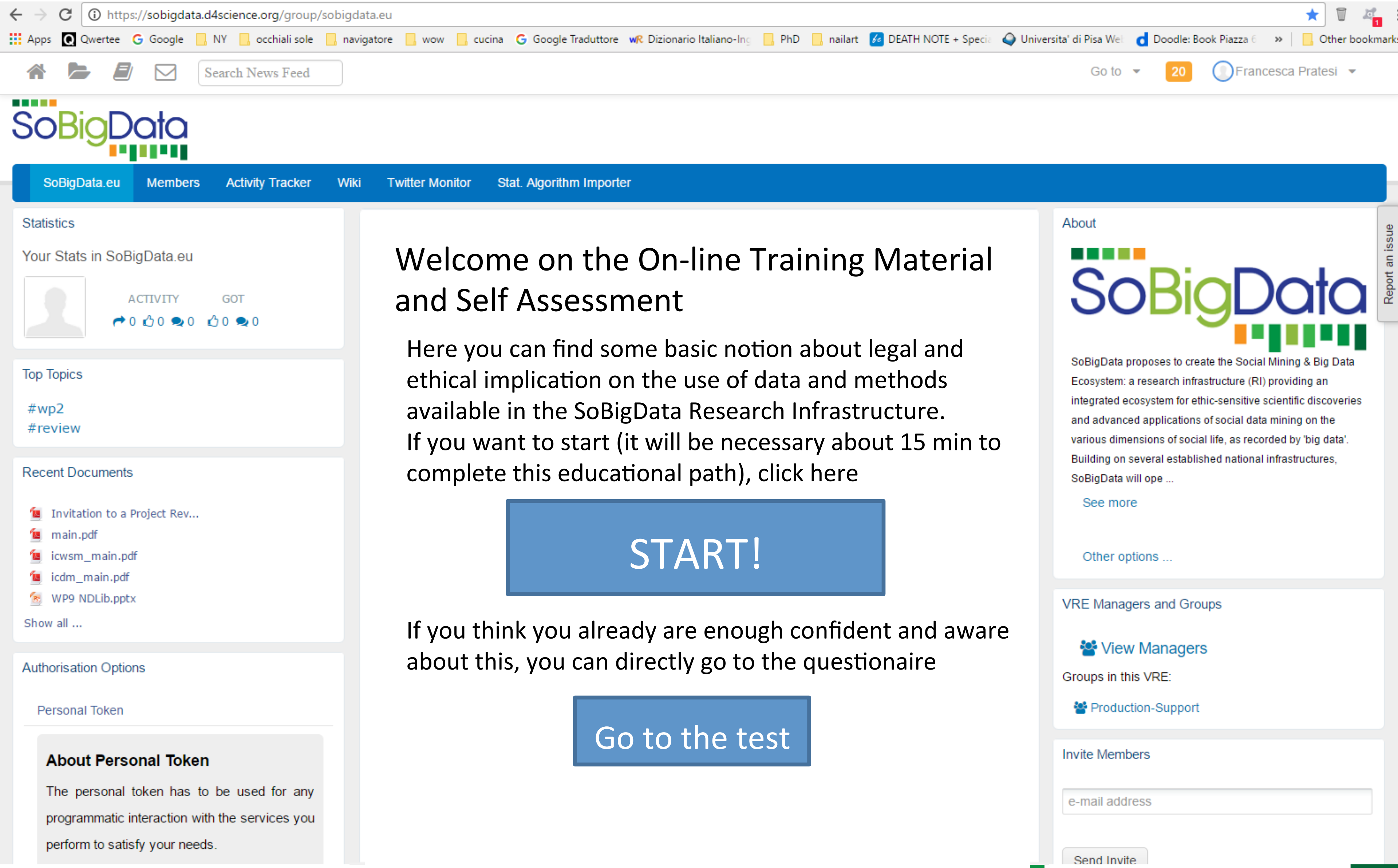
- The data scientist should know the **notions, concepts and principles** of the data protection legislation
  - General Data Protection Regulation, European Data Protection Directive
- The data scientist should know the **responsibilities** designed by the laws
  - Data Controller, Data Processor
- The data scientist should act in **compliance with data protection law principles**
  - fair and lawful processing, purpose limitation, **privacy by design** and by default
  - Processing personal data only on a legal basis (e.g. consent, research exceptions etc.)
  - Implement appropriate technical and organizational measures to protect the data and guarantee the privacy right of individuals

How to address these issues?

SoBigData Data Strategies

- Strategies for supporting actors (Data Scientists) in SoBigData RI
  - **Online Training Material & Compliance Self-assessment**
  - **Tool for the privacy risk assessment**

- On-line Training Material
- Self Assessment



Welcome on the On-line Training Material and Self Assessment

Here you can find some basic notion about legal and ethical implication on the use of data and methods available in the SoBigData Research Infrastructure.
If you want to start (it will be necessary about 15 min to complete this educational path), click here

START!

If you think you already are enough confident and aware about this, you can directly go to the questionaire

Go to the test

Statistics

Your Stats in SoBigData.eu

ACTIVITY GOT

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.

About

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

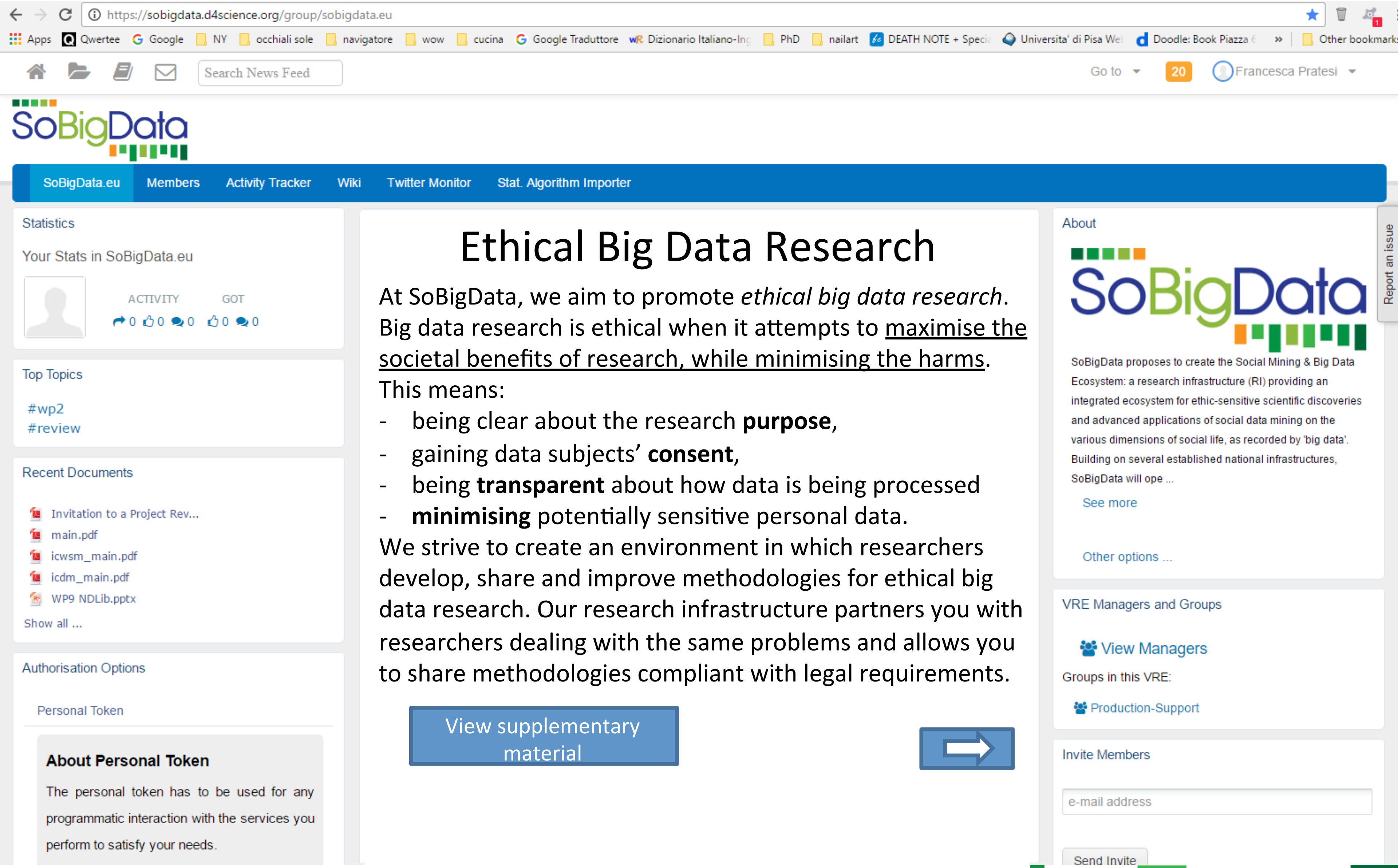
Groups in this VRE:

Production-Support

Invite Members

e-mail address

Send Invite



Statistics

Your Stats in SoBigData.eu

ACTIVITY GOT

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.

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 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

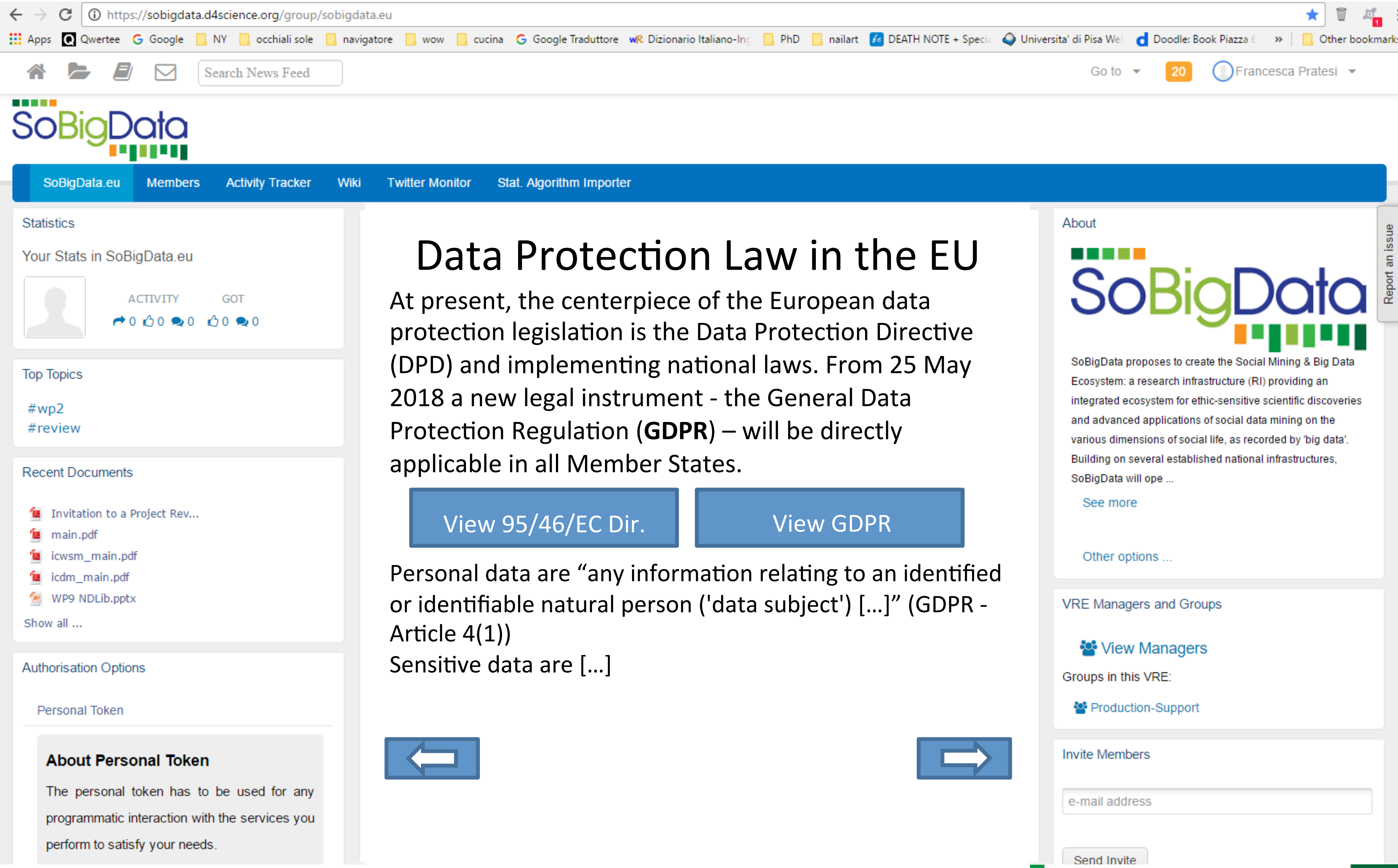
Groups in this VRE:

Production-Support

Invite Members

e-mail address

Send Invite



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

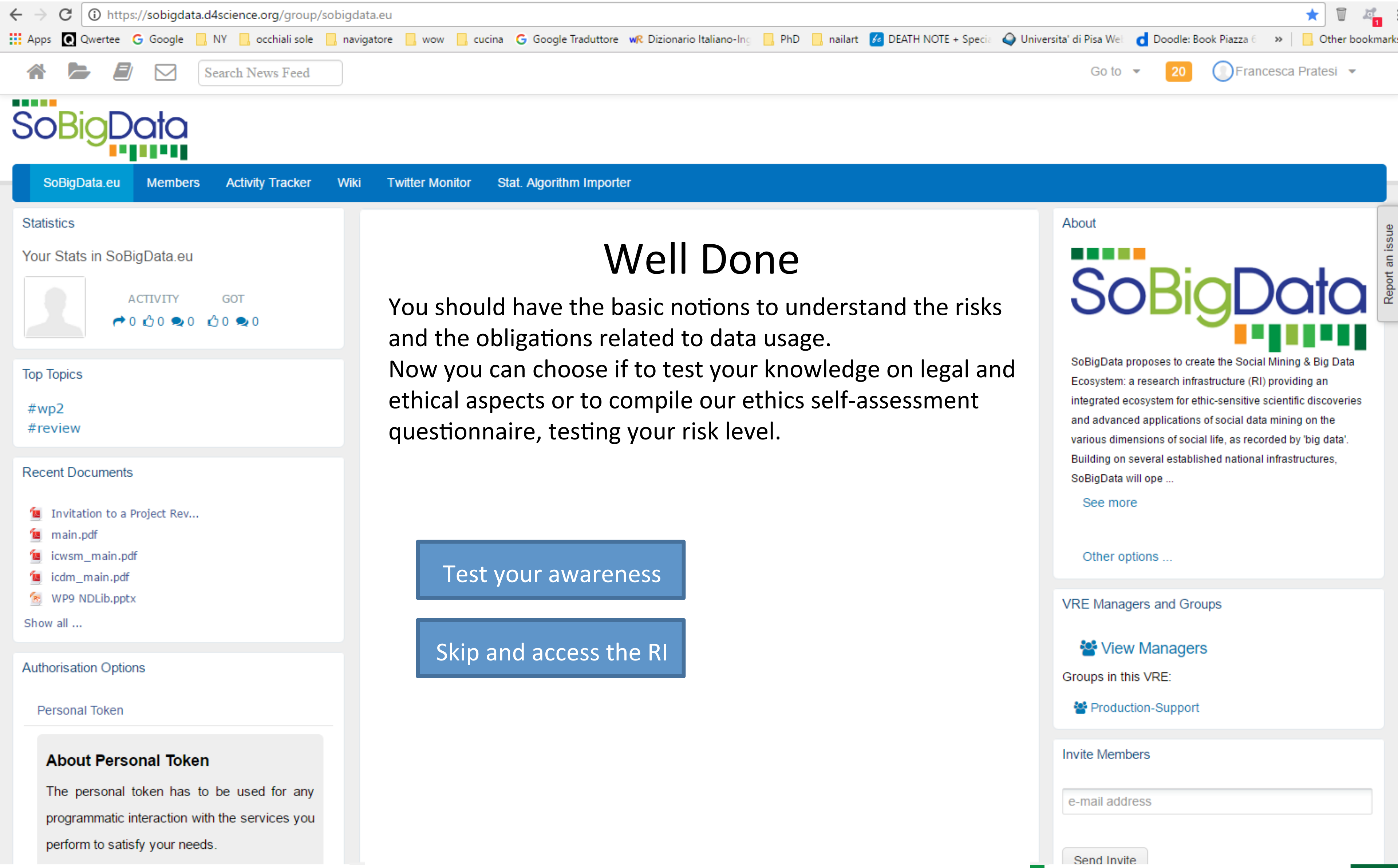
Personal data are “any information relating to an identified or identifiable natural person ('data subject') […]” (GDPR - Article 4(1))

Sensitive data are […]

About Personal Token

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

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 ...



Statistics

Your Stats in SoBigData.eu

ACTIVITY GOT

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.

Well Done

You should have the basic notions to understand the risks and the obligations related to data usage.
Now you can choose if to test your knowledge on legal and ethical aspects or to compile our ethics self-assessment questionnaire, testing your risk level.

Test your awareness

Skip and access the RI

About

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

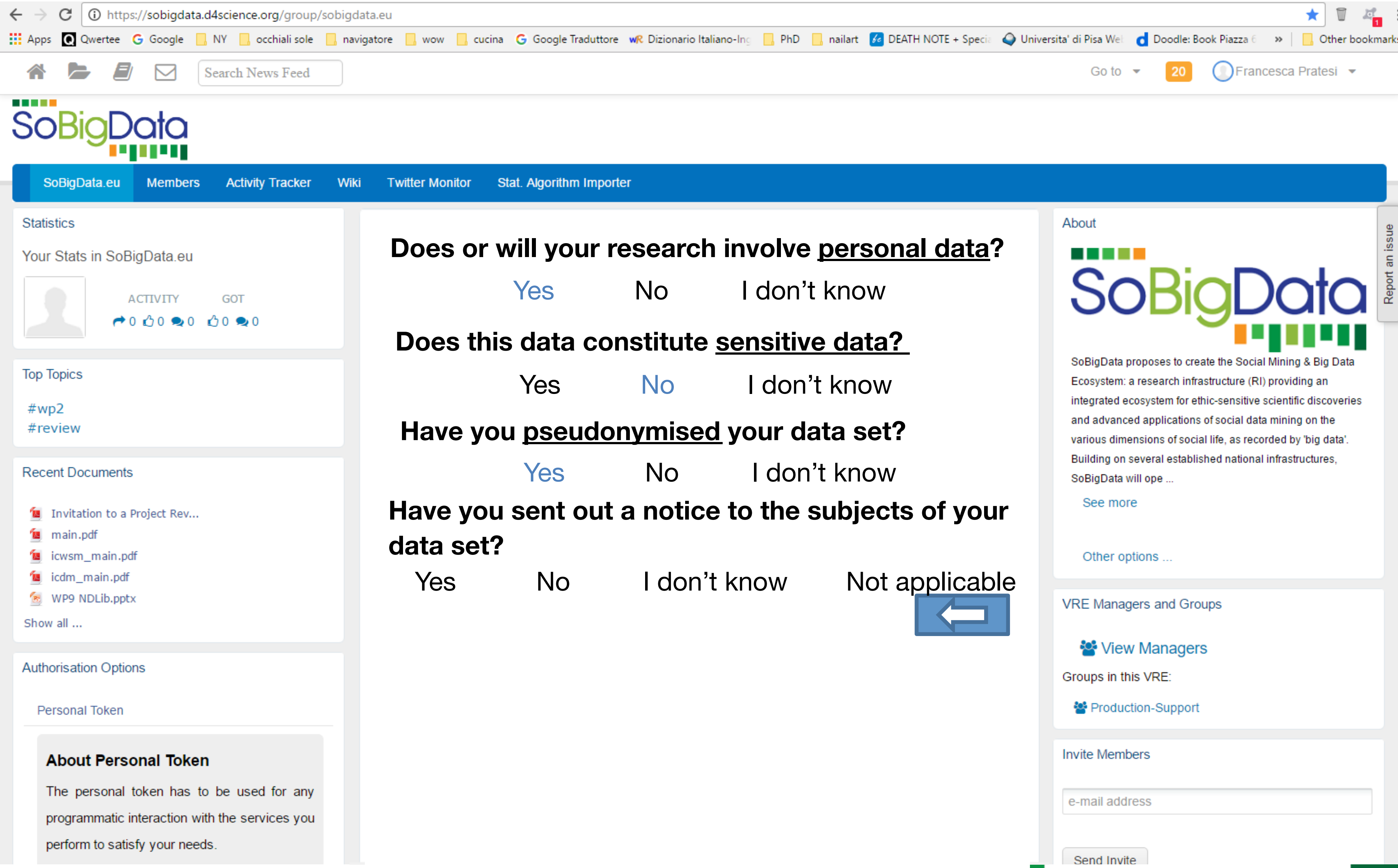
Groups in this VRE:

Production-Support

Invite Members

e-mail address

Send Invite



Statistics

Your Stats in SoBigData.eu

ACTIVITY 0 0 0 GOT 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.

Does or will your research involve personal data?

[✓ Yes] No I don’t know

Does this data constitute sensitive data?

Yes [✓ No] I don’t know

Have you pseudonymised your data set?

[✓ Yes] No I don’t know

Have you sent out a notice to the subjects of your data set?

Yes No I don’t know Not applicable

About

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

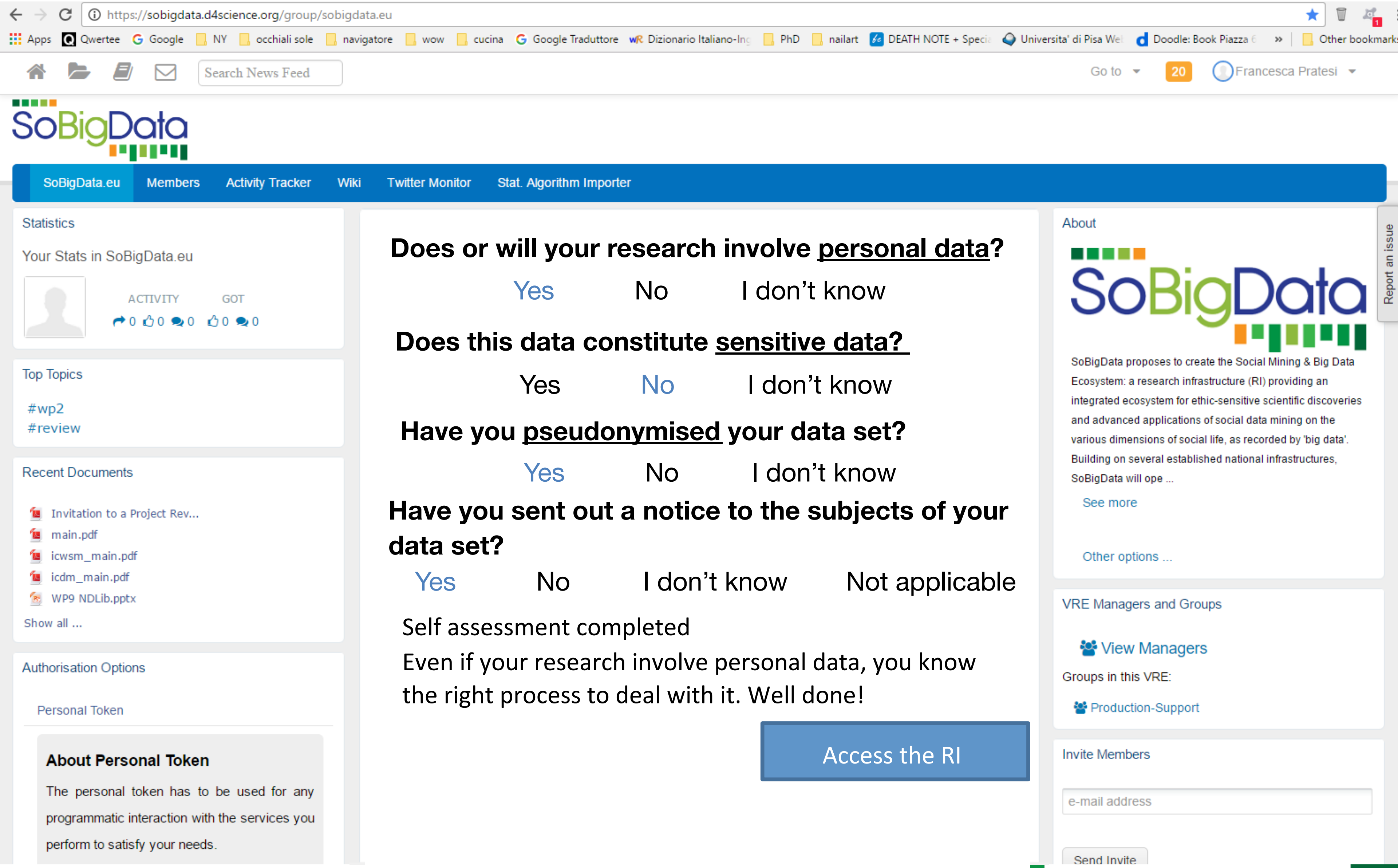
Groups in this VRE:

Production-Support

Invite Members

e-mail address

Send Invite



Statistics

Your Stats in SoBigData.eu

ACTIVITY GOT
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.

Does or will your research involve personal data?

[✓ Yes] No I don’t know

Does this data constitute sensitive data?

Yes [✓ No] I don’t know

Have you pseudonymised your data set?

[✓ Yes] No I don’t know

Have you sent out a notice to the subjects of your data set?

[✓ Yes] No I don’t know Not applicable

Self assessment completed
Even if your research involve personal data, you know the right process to deal with it. Well done!

Access the RI

About

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

Privacy Risk Assessment

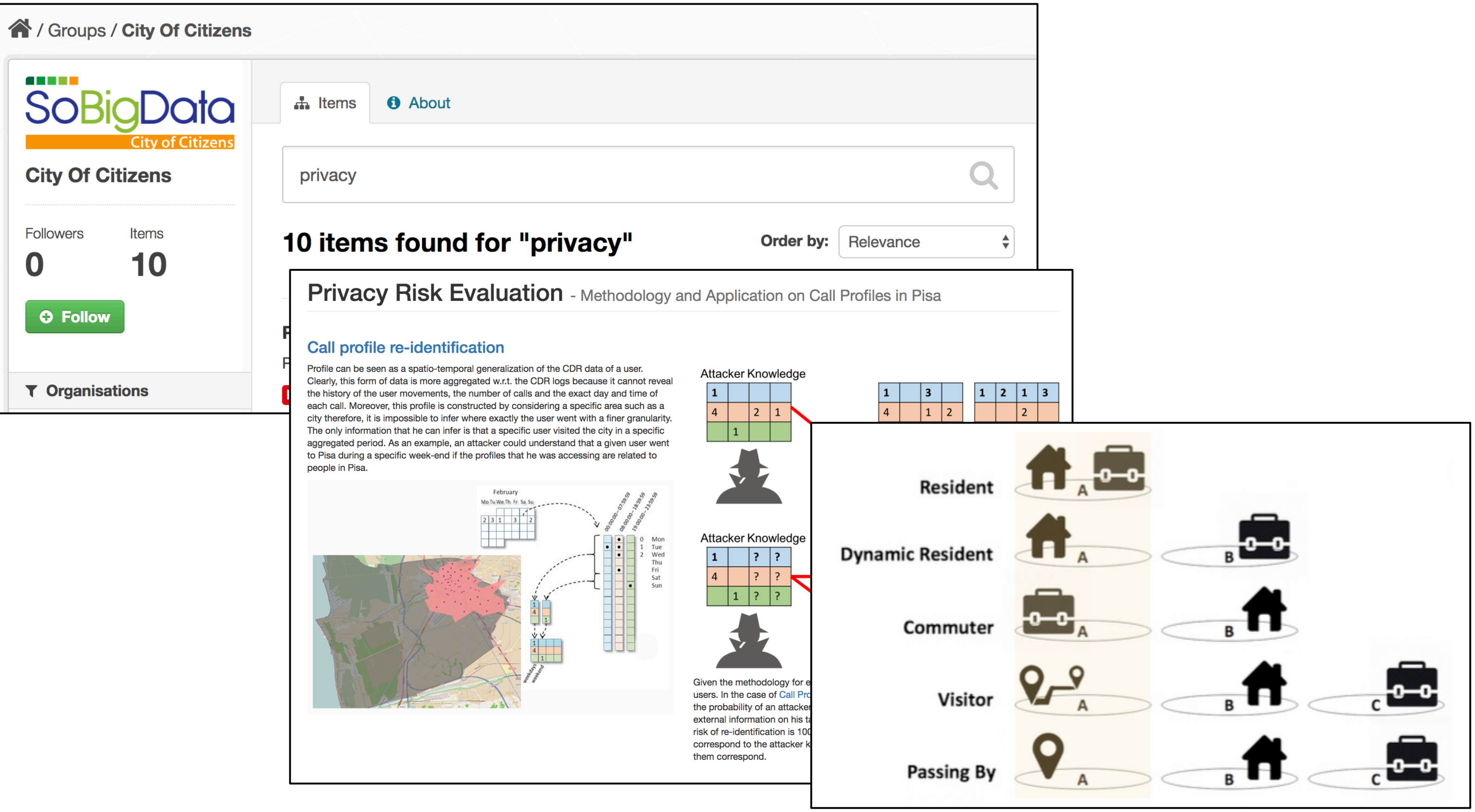
Methodology for PRA

- Service and data format definition
- External information definition
- Simulation of privacy harmful Inferences
- Vulnerability (Risk) quantification
- (Risk mitigation)

Methods

- Privacy Risk Assessment
  - **Human call habits data (CDR aggregation)**
  - Trajectory data (from GPS observations)
  - Individual shop habits
  - …
- Privacy Risk Mitigation

Definition of service
Classifying user behaviour (Sociometer)



/ Groups / **City Of Citizens**

City Of Citizens

Followers 0

Items 10

Items · About

privacy

10 items found for "privacy"

Order by: Relevance

Organisations

Privacy Risk Evaluation - Methodology and Application on Call Profiles in Pisa

Call profile re-identification

Profile can be seen as a spatio-temporal generalization of the CDR data of a user. Clearly, this form of data is more aggregated w.r.t. the CDR logs because it cannot reveal the history of the user movements, the number of calls and the exact day and time of each call. Moreover, this profile is constructed by considering a specific area such as a city therefore, it is impossible to infer where exactly the user went with a finer granularity. The only information that he can infer is that a specific user visited the city in a specific aggregated period. As an example, an attacker could understand that a given user went to Pisa during a specific week-end if the profiles that he was accessing are related to people in Pisa.

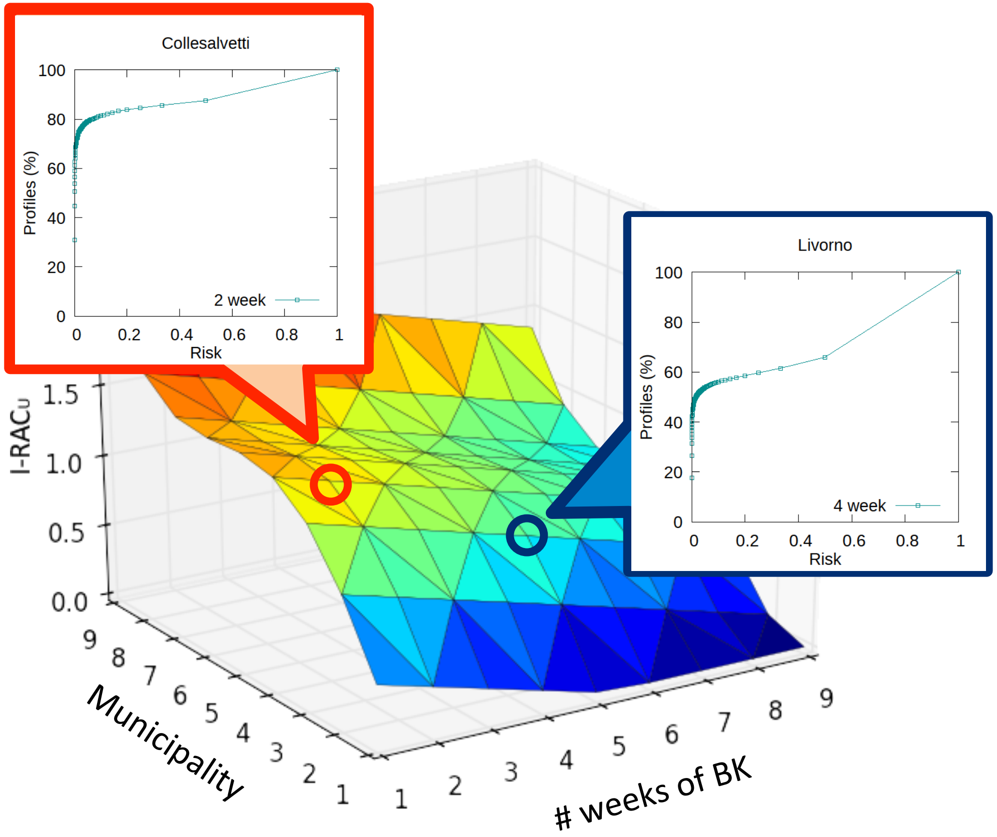
Given the methodology for e
users. In the case of Call Pro
the probability of an attacker
external information on his ta
risk of re-identification is 100
correspond to the attacker k
them correspond.

Simulation of privacy harmful Inferences

Data dimension:
The spatial area in which the analysis is performed.

Background Knowledge dimension:
The temporal window (in weeks) in which the attacker recorded the user activity.

I-RACu:
An indicator of the risk of re-identification of the users



RRI & Data Scientist

- Data Scientist is a professional figure with a mix of competence and knowledge
  - On methods and technologies for managing large amount of data
  - On analytical techniques and modelling of data and data mining
  - On story-telling techniques and data visualization
  - On **ethical and legal aspects** and social impact of data science
  - On appropriate technical and organizational measures for data protection

Thank you!!!!

Anna Monreale

Dipartimento di Informatica

Università di Pisa

anna.monreale@unipi.it

Knowledge Discovery and Delivery Lab
(ISTI-CNR & Univ. Pisa)
www-kdd.isti.cnr.it