Big Data

Paolo Ferragina & Dino Pedreschi

Dipartimento di Informatica

Università di Pisa

KDD LAB http://kdd.isti.cnr.it

ACUBE LAB http://acube.di.unipi.it

Siamo tutti pollicini digitali

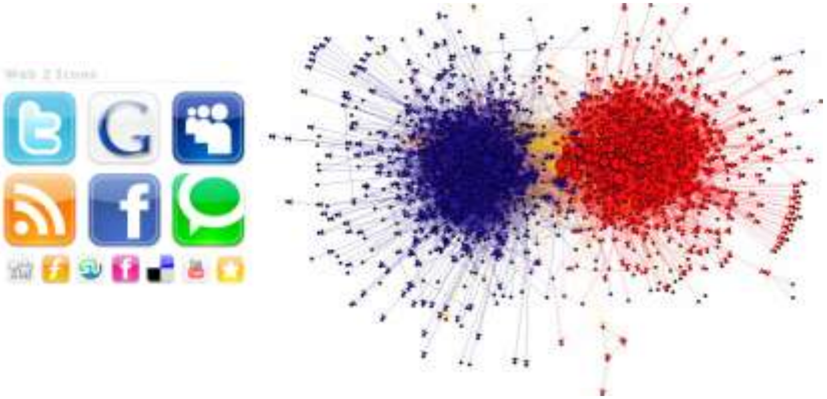
- Plenty of digital breadcrumbs behind us
- La Vita Nova, e-magazine de Il Sole 24 Ore
- Fosca Giannotti, Dino Pedreschi
- Dicembre 2012
- Everyone is becoming a «statistical entity»

Big data "proxies" of social life

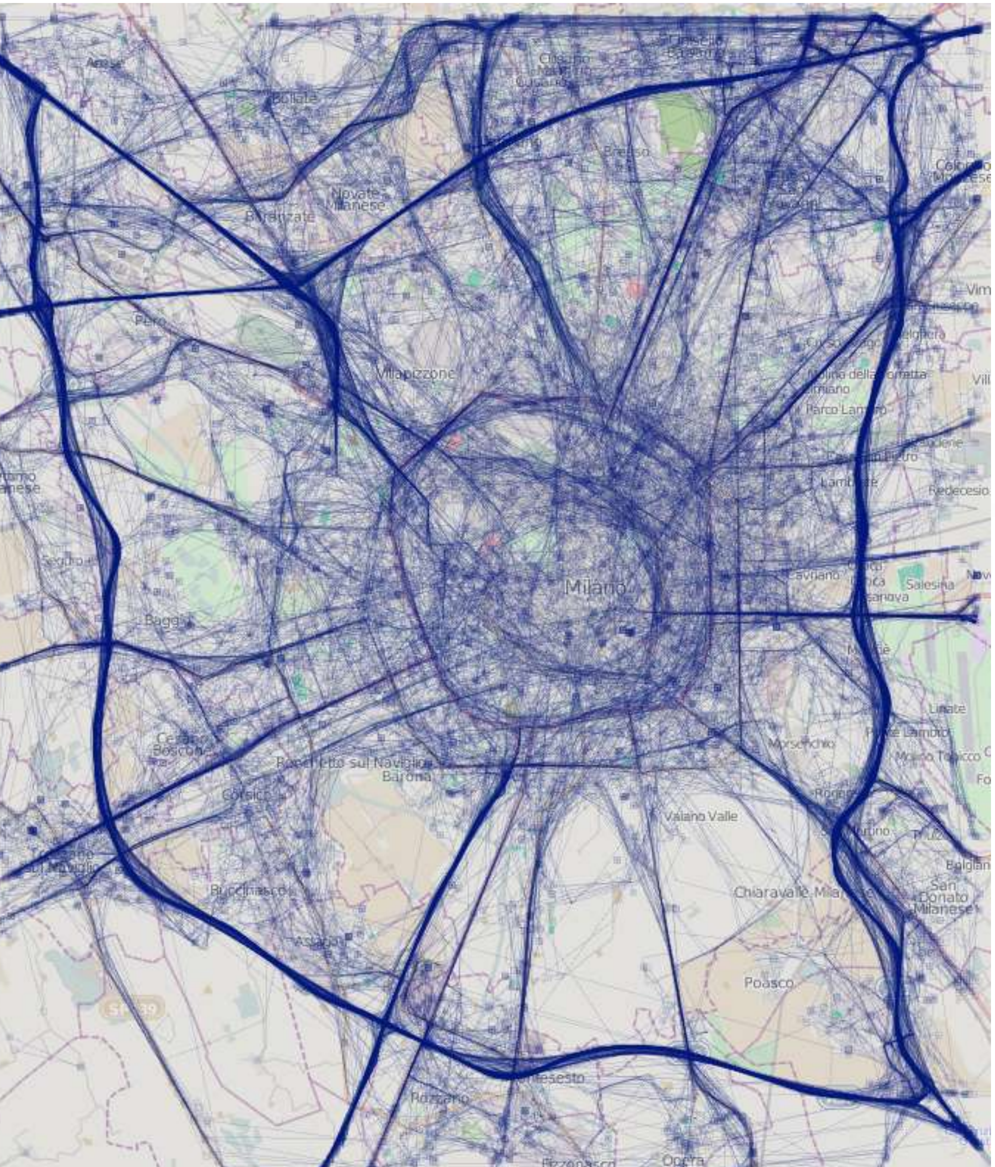
Shopping patterns & lifestyle

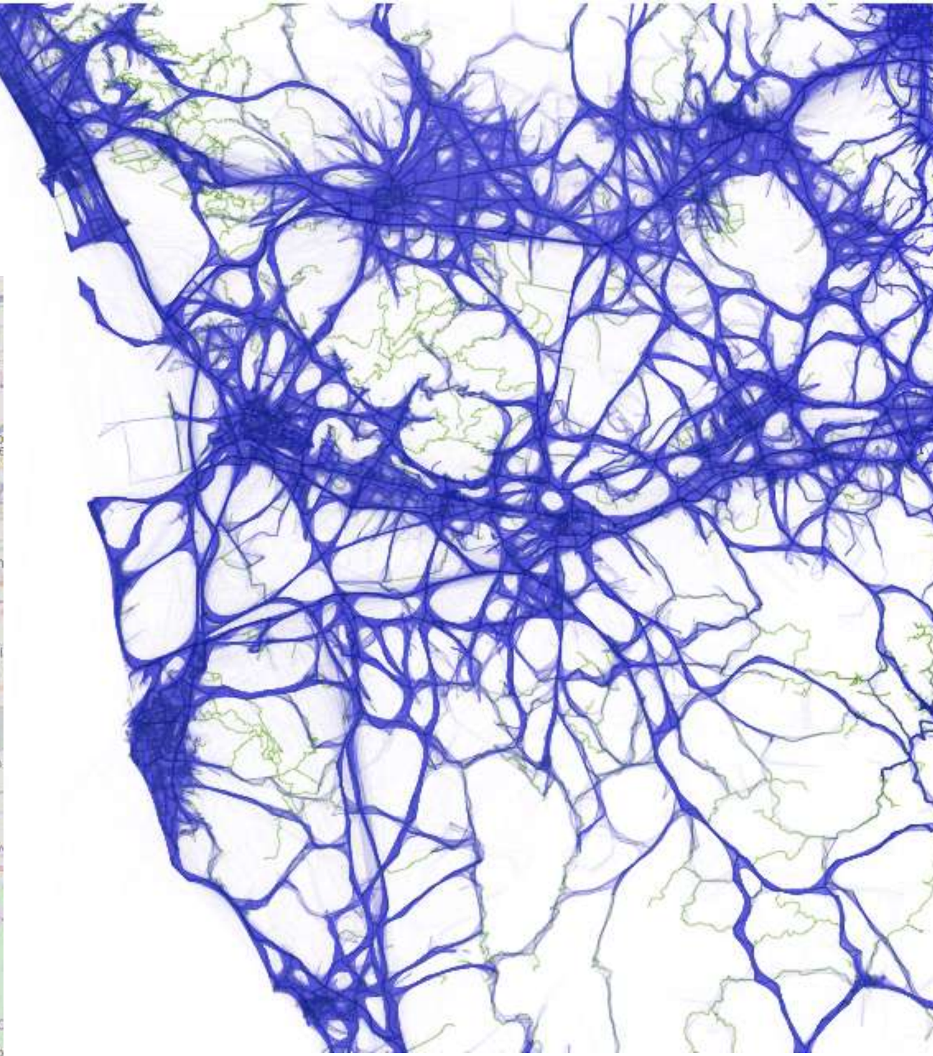
Desires, opinions, sentiments

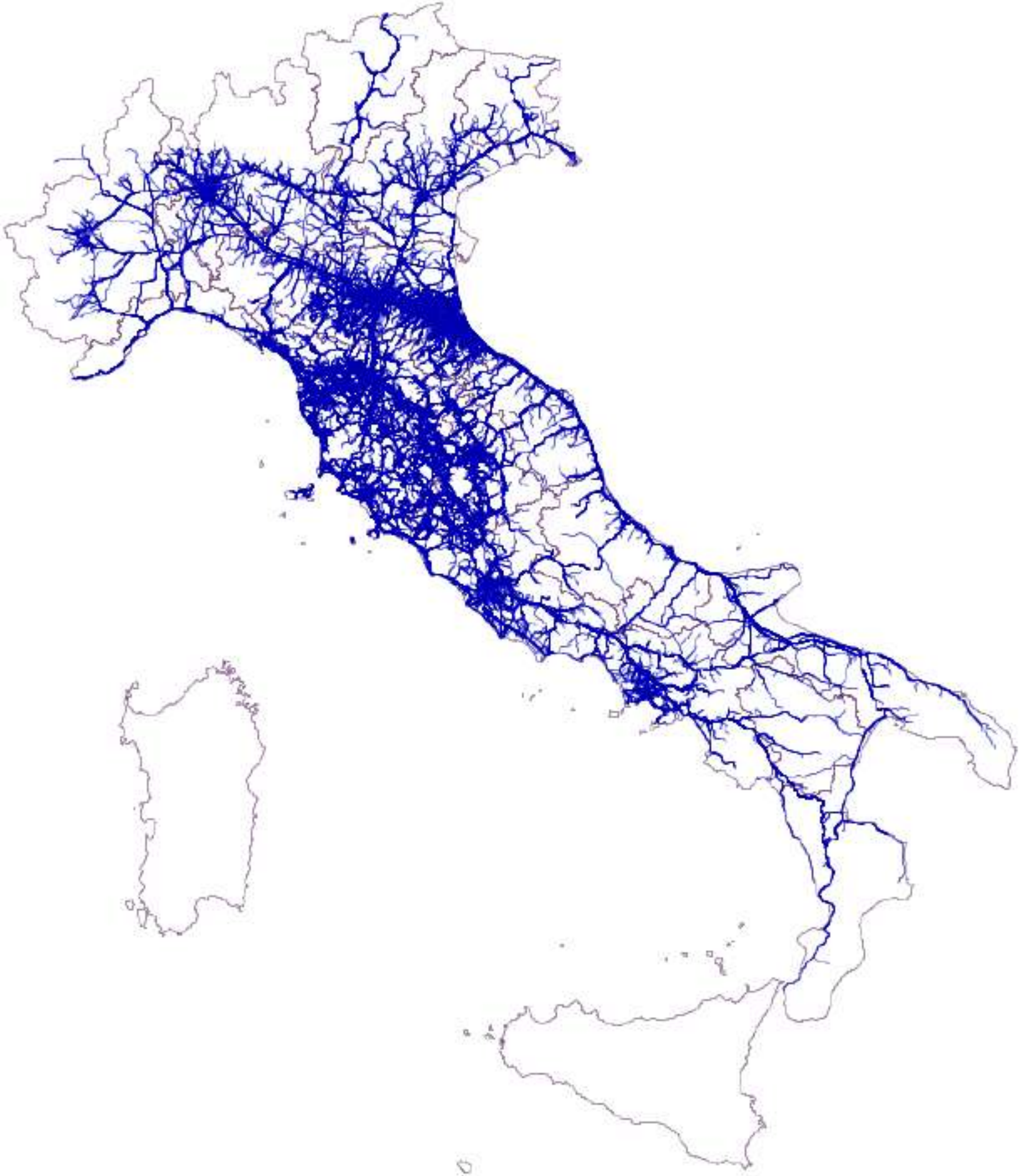
Relationships & social ties



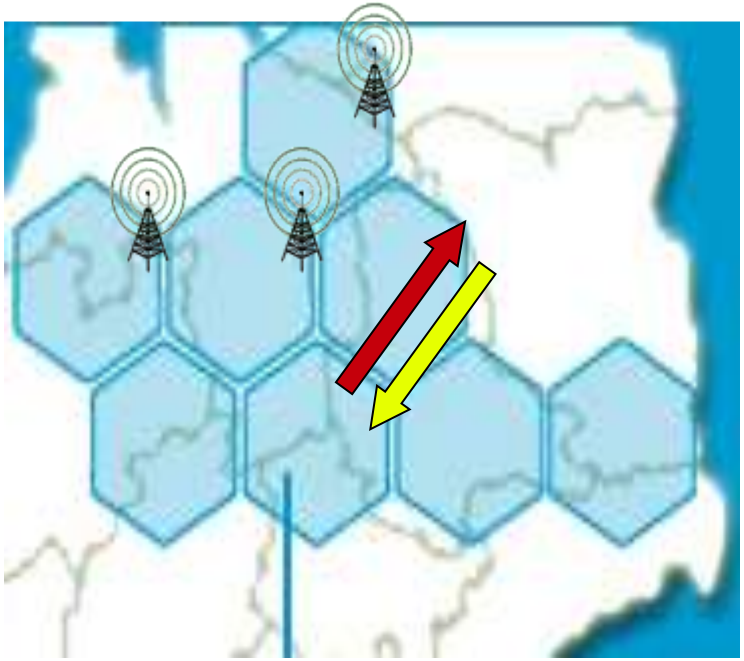
Movements

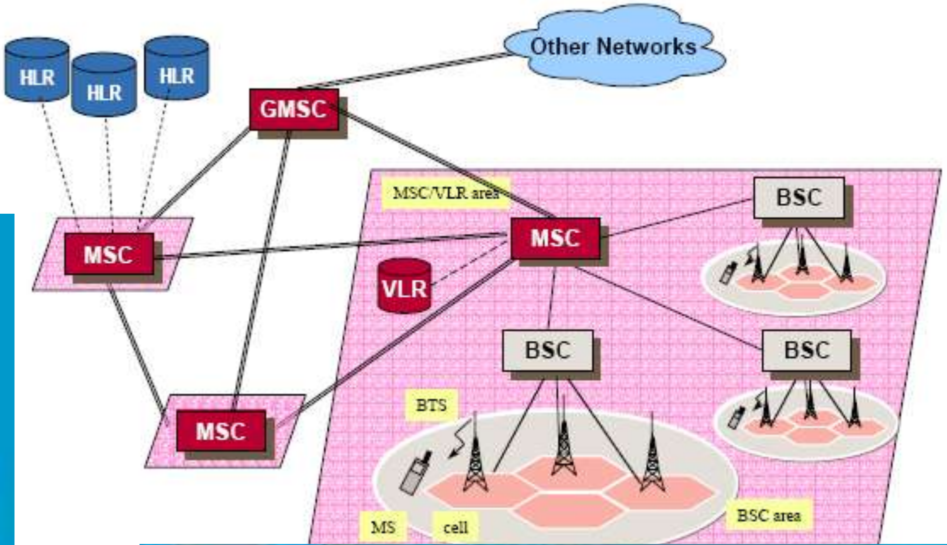

Arese
Bollate
Cinisello Balsamo
Cusano Milanino
Bresso
Novate Milanese
Baranzate
Cologno Monzese
Vimodrone
Pero
Villapizzone
Segrate
Parco Lambro
Lambrate
Redecesio
Milano
Baggio
Linate
Cesano Boscone
Ronchetto sul Naviglio
Barona
Corsico
Vaiano Valle
Buccinasco
Chiaravalle Milanese
San Donato Milanese
Assago
Poasco
Noverasco
Rozzano
Opera

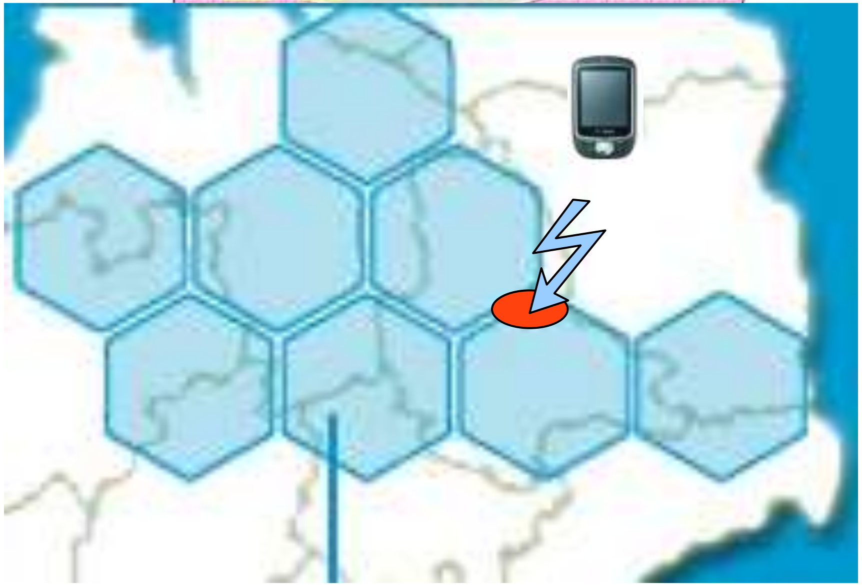




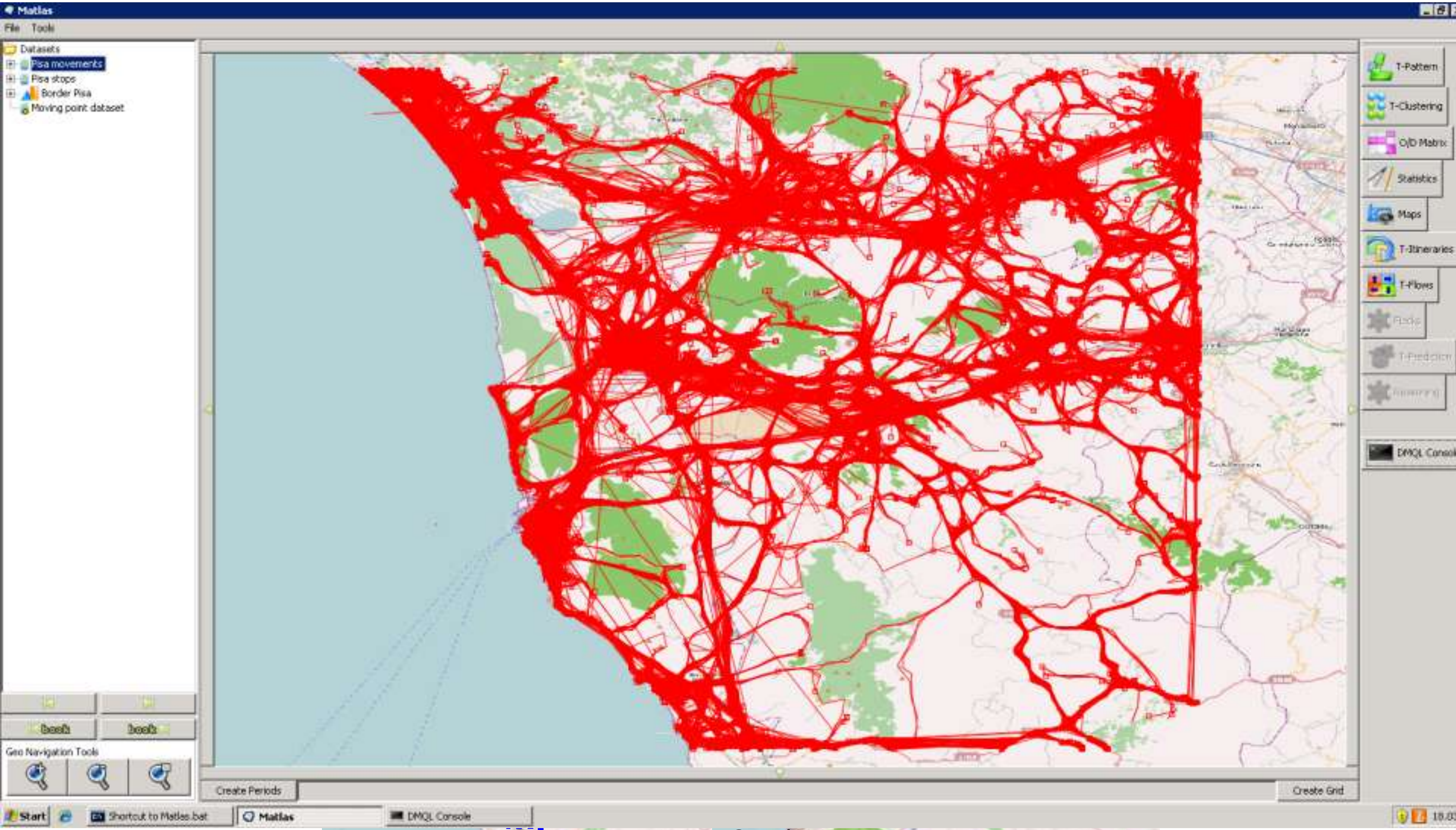
GSM data: CDR – call detail records





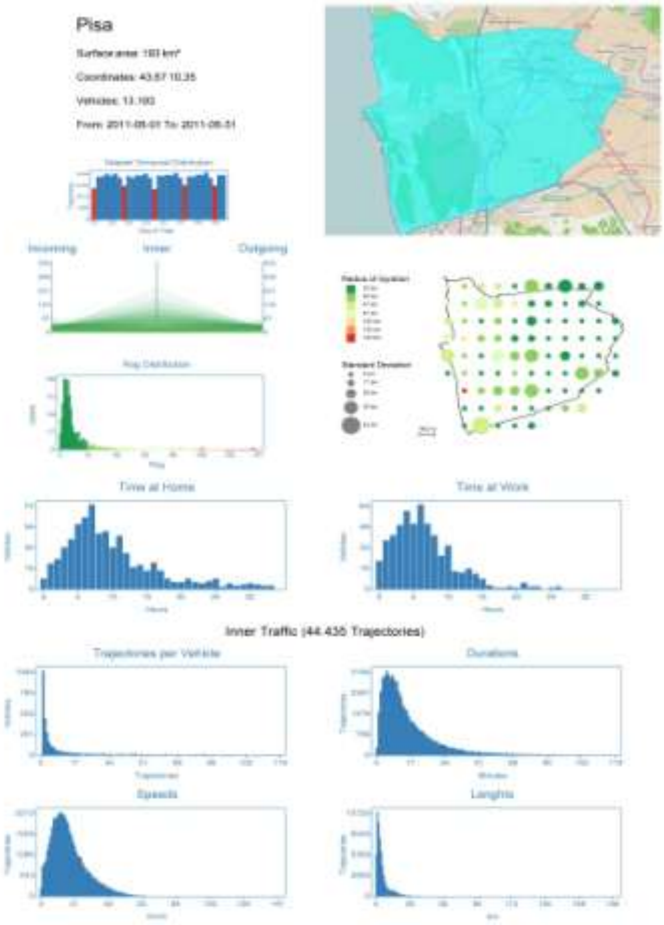


Understanding human mobility @ KDD LAB (Unipi + ISTI-CNR)


Matlas
File
Tools
Datasets
Pisa movements
Pisa stops
Border Pisa
Moving point dataset
book
book
Geo Navigation Tools
T-Pattern
T-Clustering
O/D Matrix
Statistics
Maps
T-Itineraries
T-Flows
Flocks
T-Prediction
Reasoning
DMQL Console
Create Periods
Create Grid
Start
Shortcut to Matlas.bat
Matlas
DMQL Console
18.03

Cascina

Mobility atlas of many cities



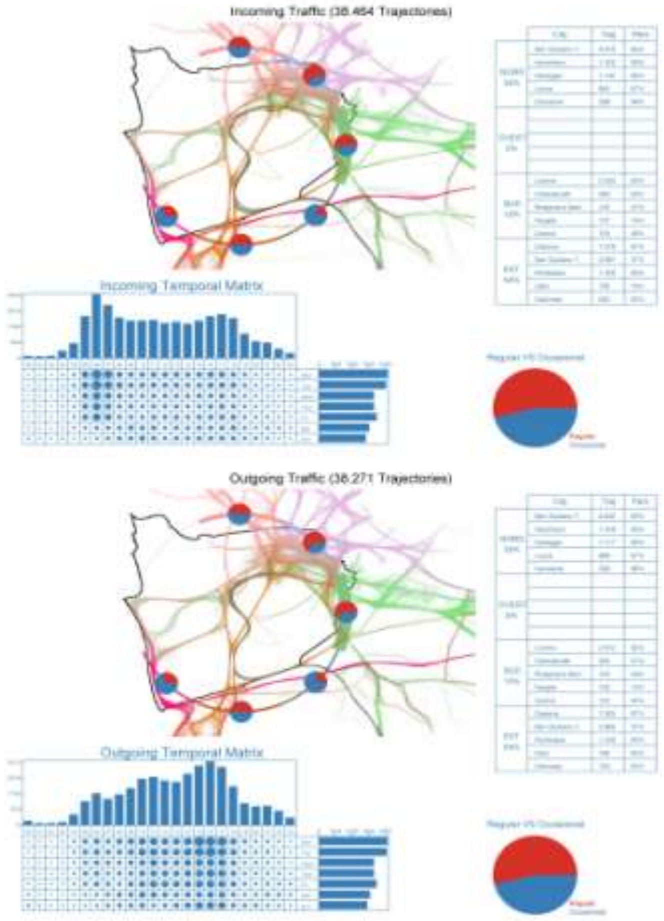
Pisa

Surface area: 193 km²

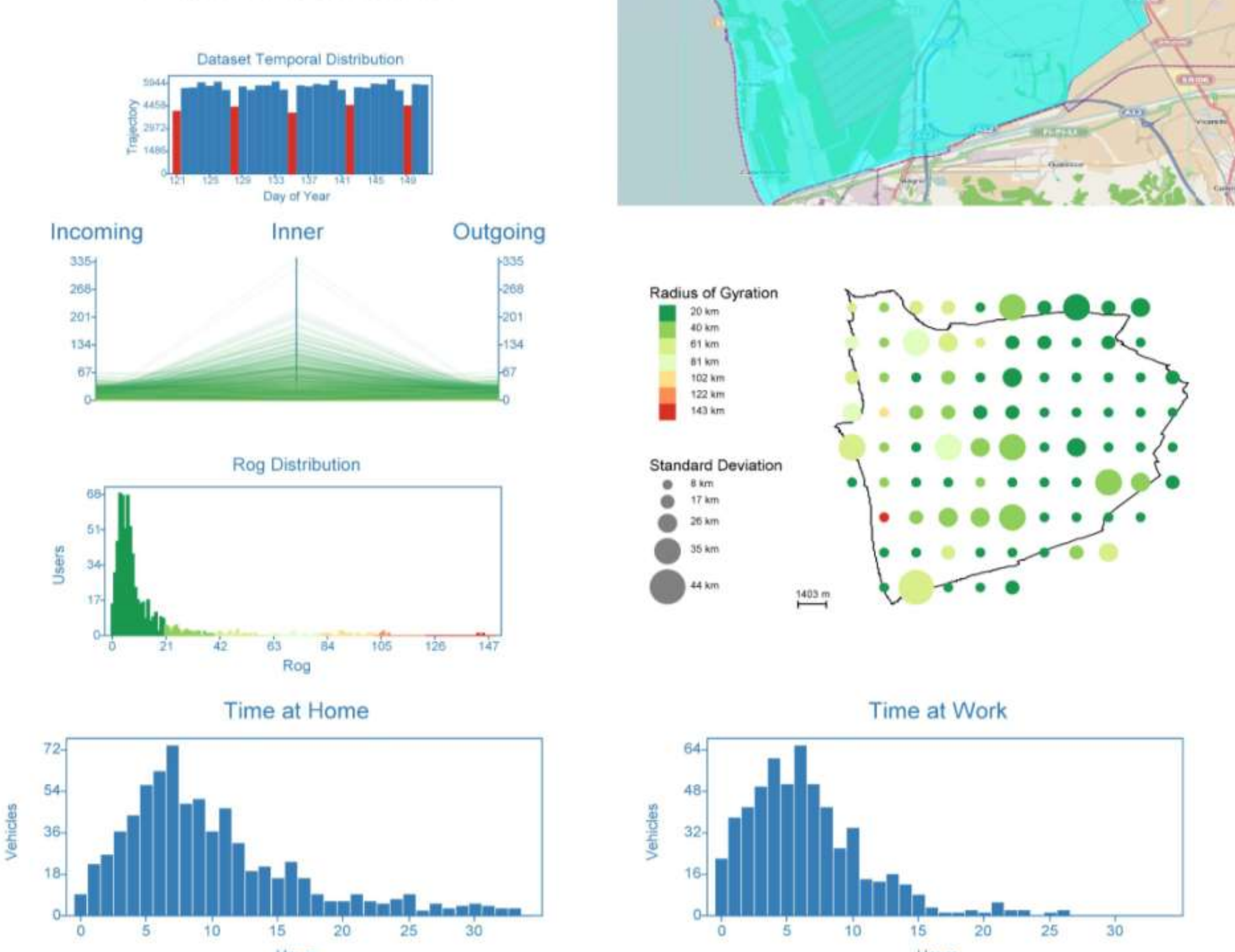
Coordinates: 43.67 10.35

Vehicles: 13.193

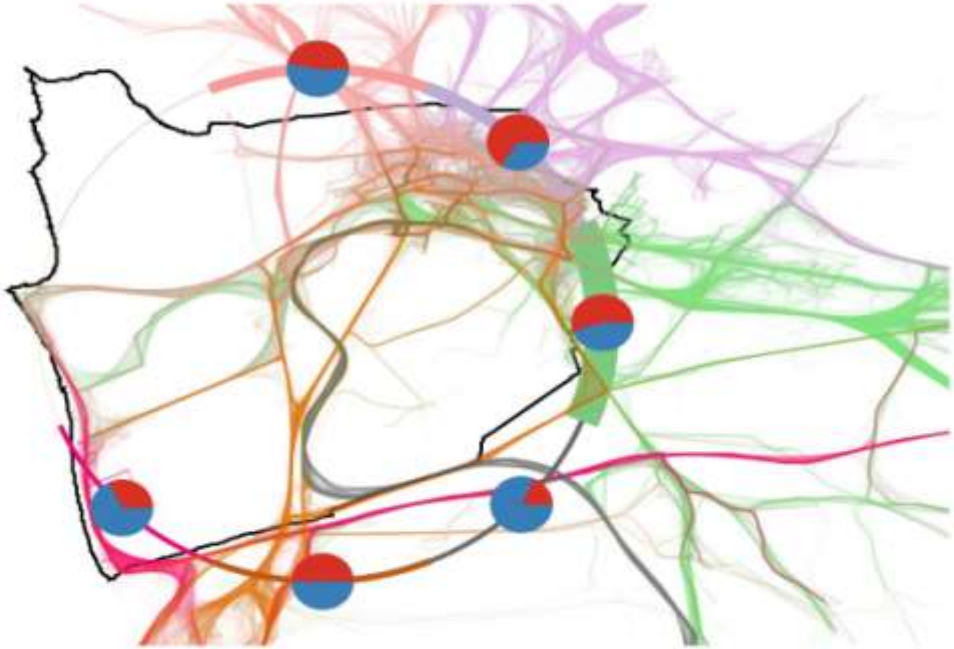
From: 2011-05-01 To: 2011-05-31



OCTO Telematics

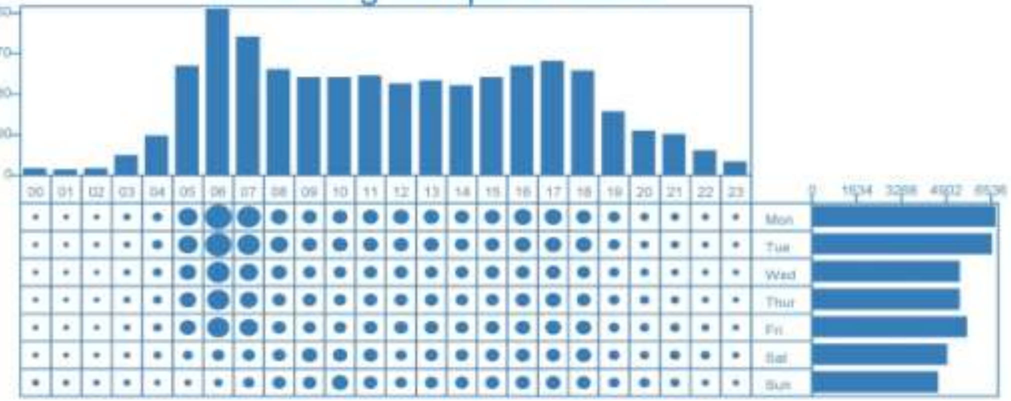

Dataset Temporal Distribution
Trajectory
5944
4458
2972
1486
0
121
125
129
133
137
141
145
149
Day of Year
Incoming
Inner
Outgoing
335
268
201
134
67
0
Rog Distribution
Users
68
51
34
17
0
21
42
63
84
105
126
147
Rog
Radius of Gyration
20 km
40 km
61 km
81 km
102 km
122 km
143 km
Standard Deviation
8 km
17 km
26 km
35 km
44 km
1403 m
Time at Home
Vehicles
72
54
36
18
0
5
10
15
20
25
30
Time at Work
Vehicles
64
48
32
16
0
5
10
15
20
25
30

Incoming Traffic (38.464 Trajectories)

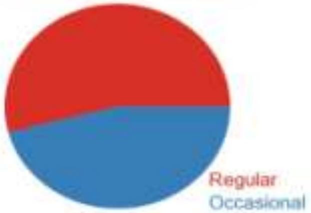


	City	Traj	Perc
NORD 32%	San Giuliano T..	4.816	62%
	Vecchiano	1.425	94%
	Viareggio	1.142	99%
	Lucca	882	67%
	Camaiore	358	94%
OVEST 0%			
SUD 12%	Livorno	2.843	92%
	Collesalvetti	565	50%
	Rosignano Mari..	140	41%
	Fauglia	137	19%
	Cecina	124	45%
EST 54%	Cascina	7.078	97%
	San Giuliano T..	2.881	37%
	Pontedera	1.350	95%
	Calci	795	79%
	Calcinaia	693	92%

Incoming Temporal Matrix



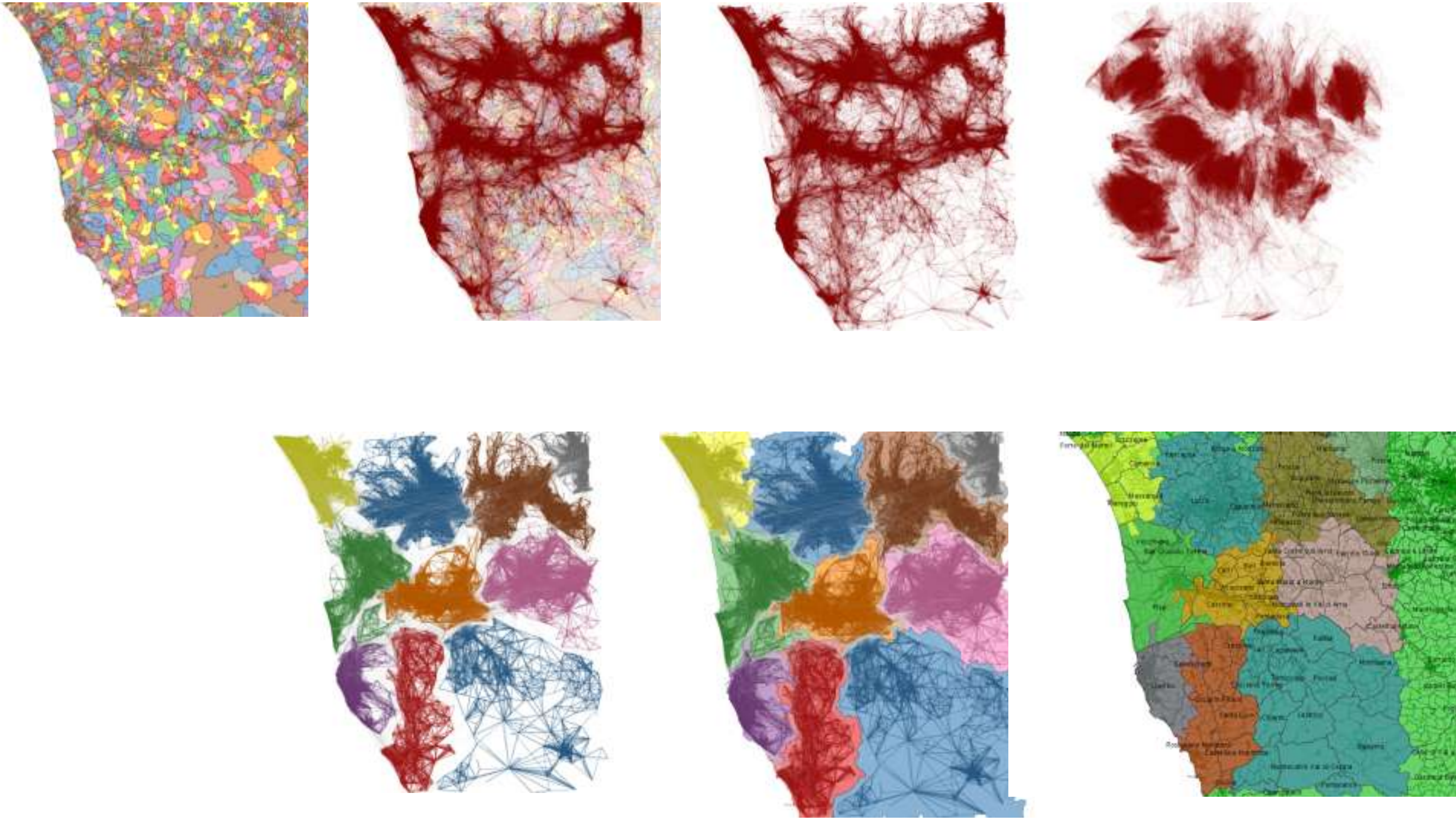
Regular VS Occasional



Outgoing Traffic (38.271 Trajectories)

	City	Traj	Perc
	San Giuliano T..	4.842	62%
	Vecchiano	1.418	93%

Discover the borders of mobility



Estimate O-D matrix from phone data

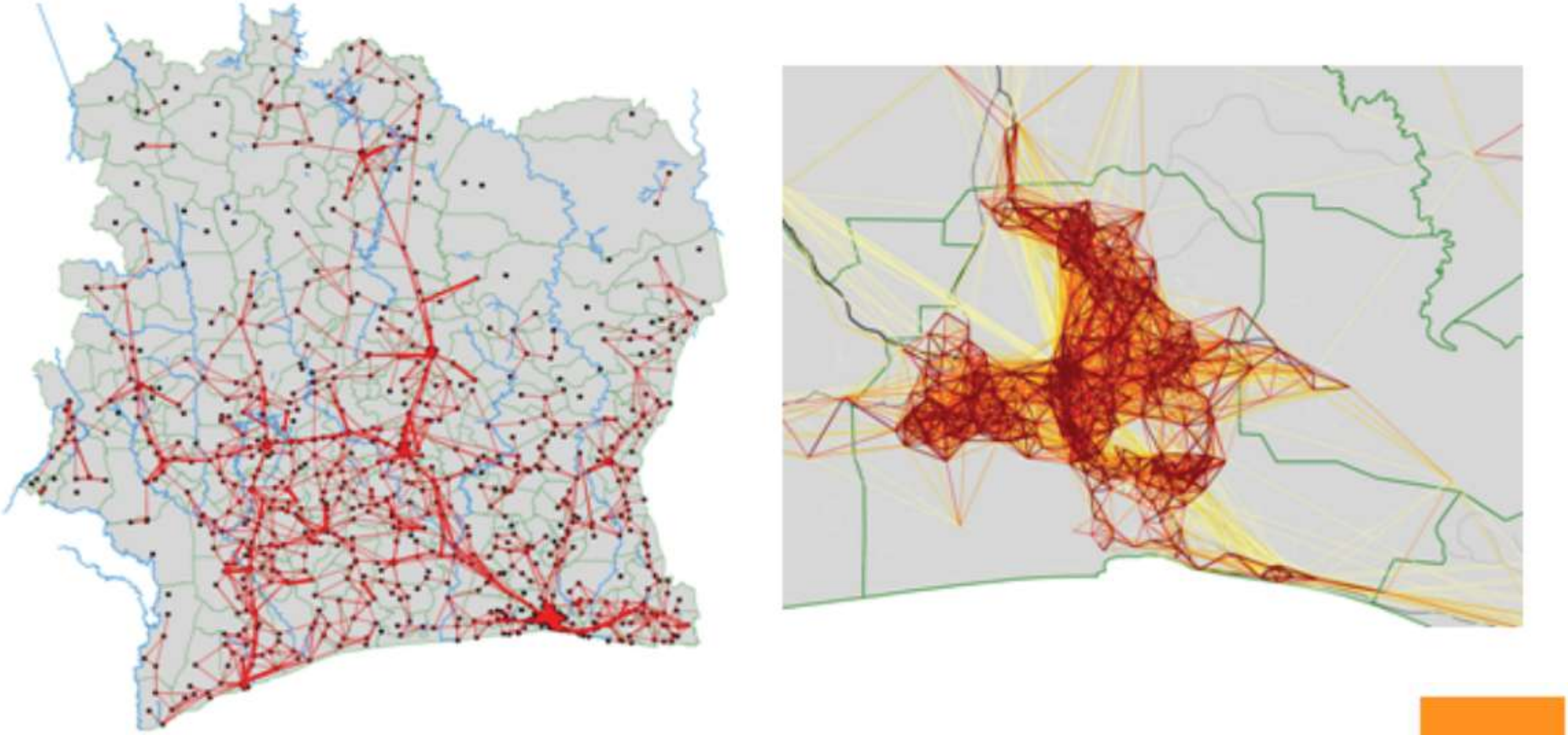
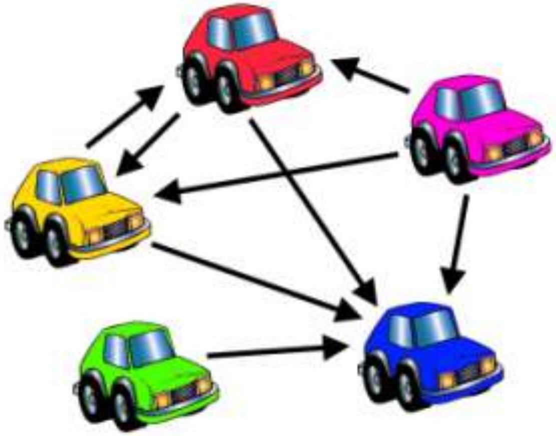


Figure 12: Mobile phone movements in Ivory Coast and Abidjan.

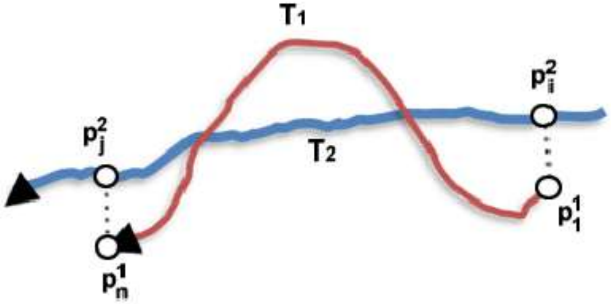
Mirco Nanni, Roberto Trasarti, et al.:
MP4-A Project: Mobility Planning for Africa. “Data for Development” Orange challenge, 2013

proactive car pooling

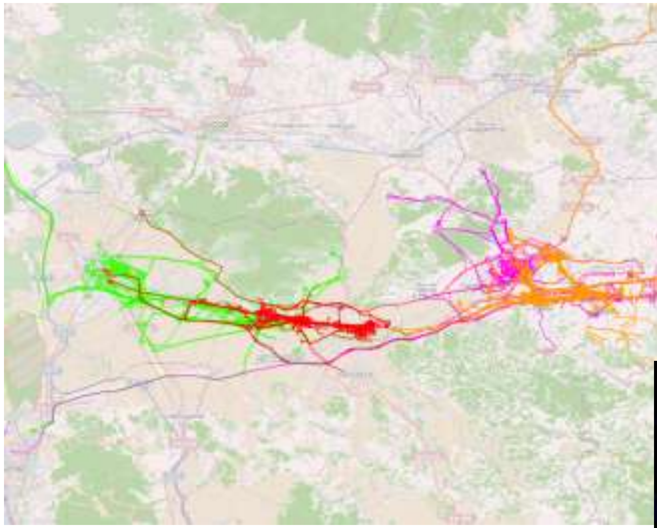
Car Pooling

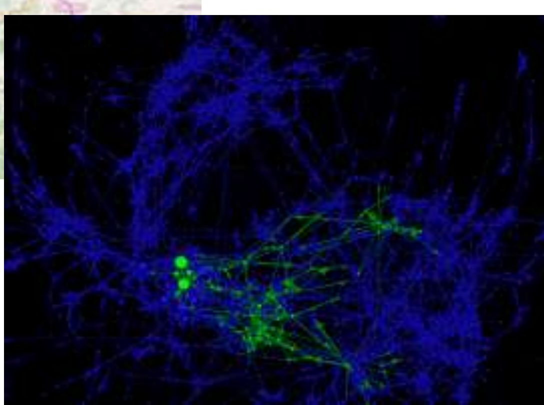


Carpooling Network



Trajectory matching





Carpooling Communities

Carpooling potential

PISA

16467 users, 357137 trips

20% are systematic trips

40% of them are matching trips

142.740.060 saved Kms

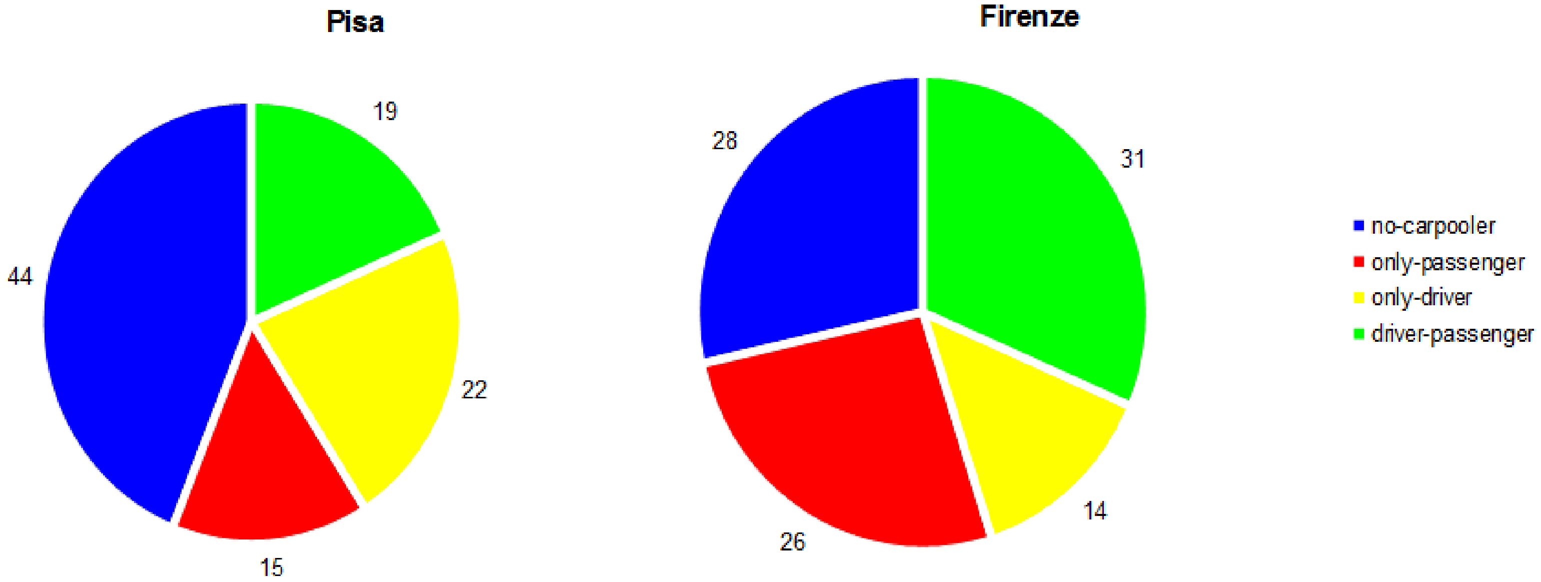
FIRENZE

34864 users, 1040872 trips

9% are systematic trips

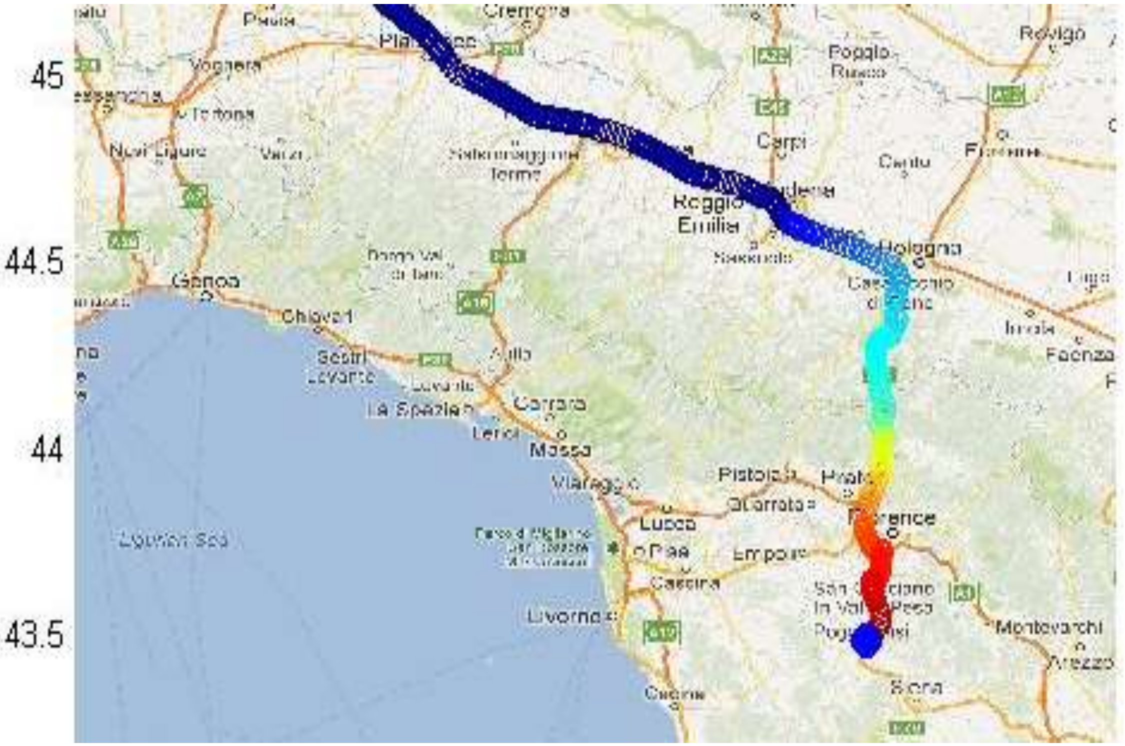
58% of them are matching trips

220.802.240 saved Kms



Electrifiability

Electrifiability



In Tuscany **75% users** have a daily mobility covered at 100% by an electrical car (home charging only)

In Pisa 90.5% of daily trips are electrifiable:
562.061 km electrificable

City users Sociometer

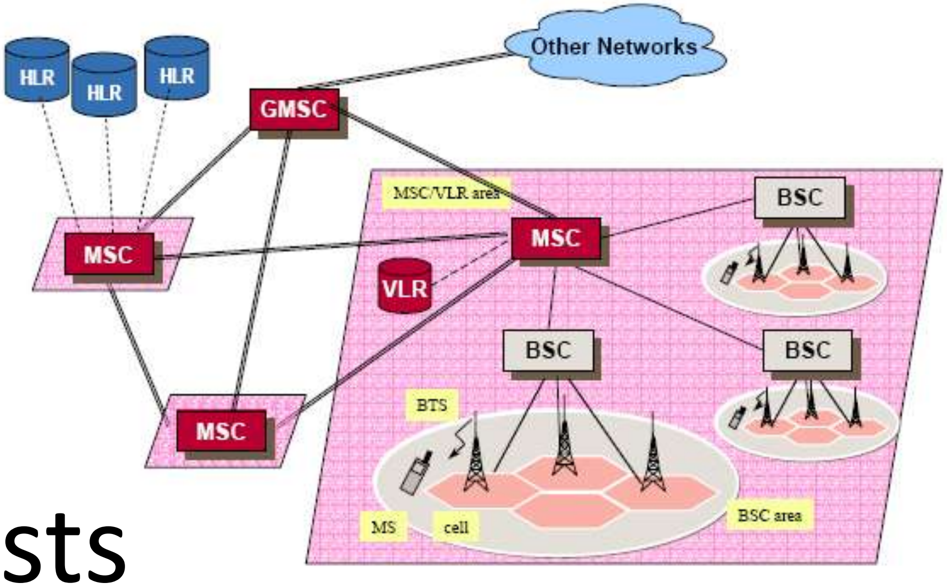
Mobile phone socio-meter

Analyze individual call habits to recognize profiles

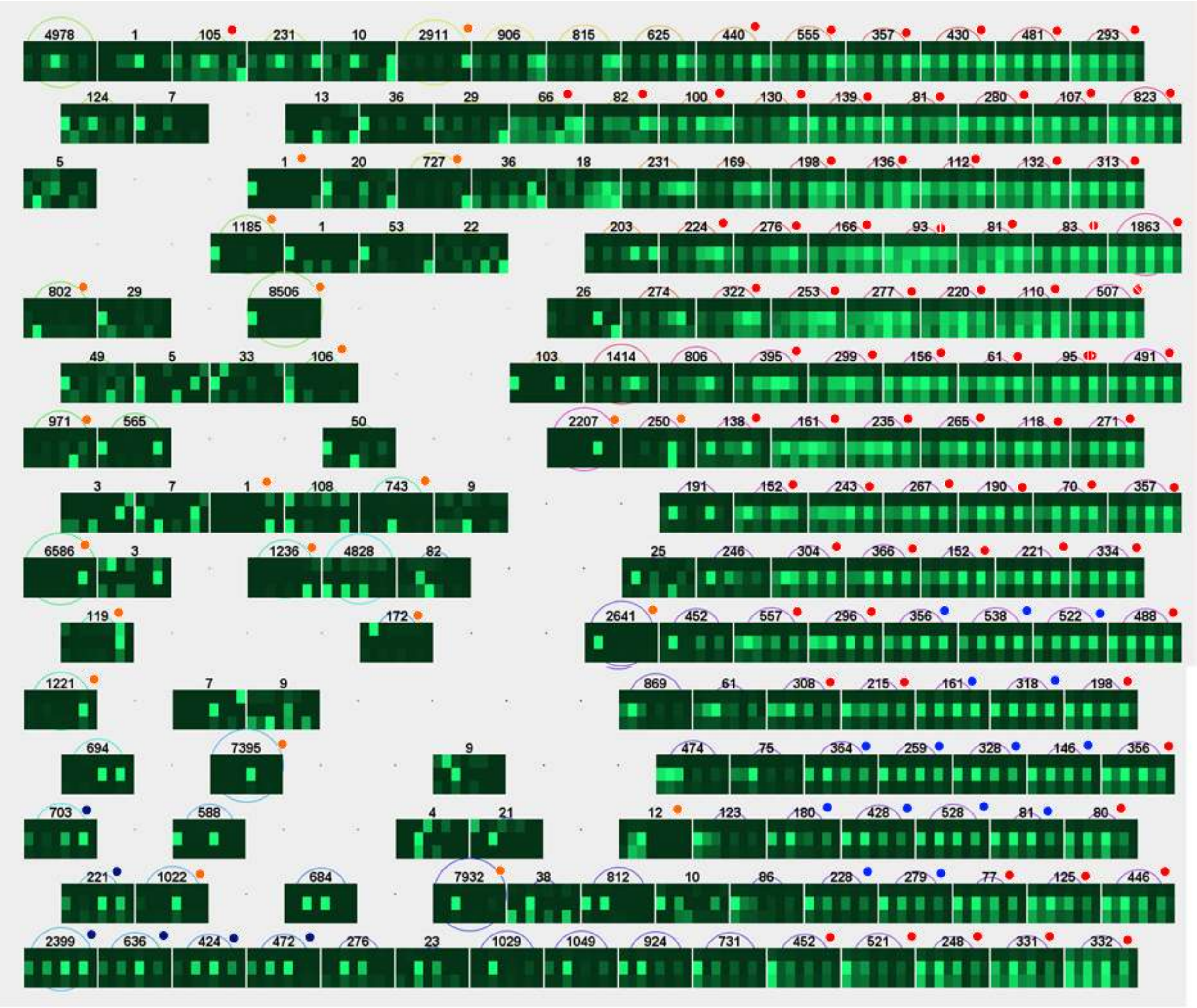
–Resident

–Commuters

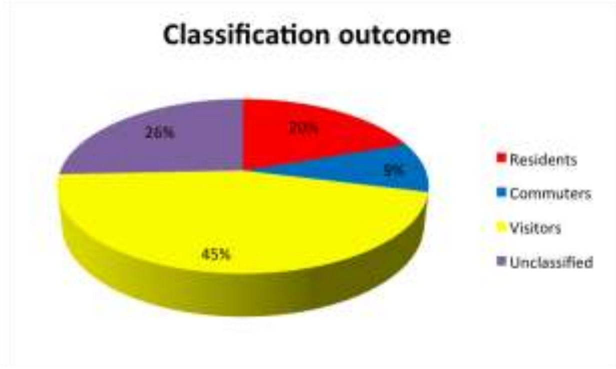
–Visitors/Tourists



City user profile **quantification**

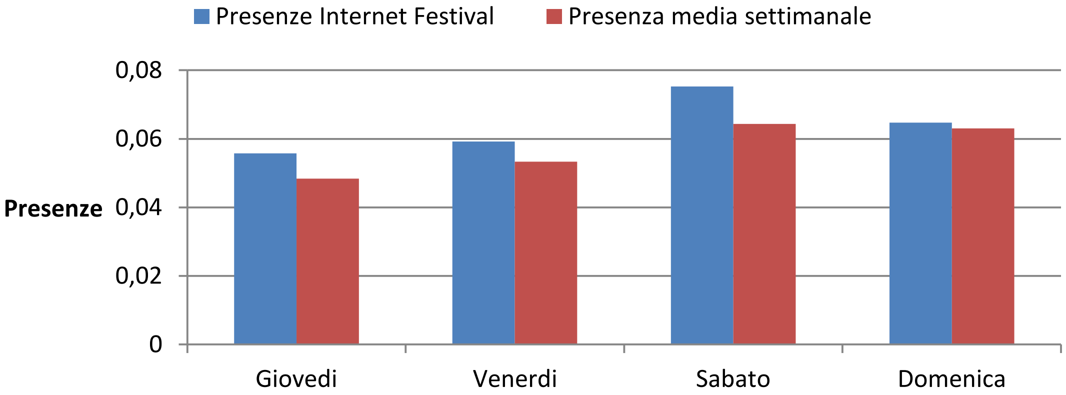


- ● Resident profile
- ● Commuter profile
- ● Visitor profile



Monitoring big events

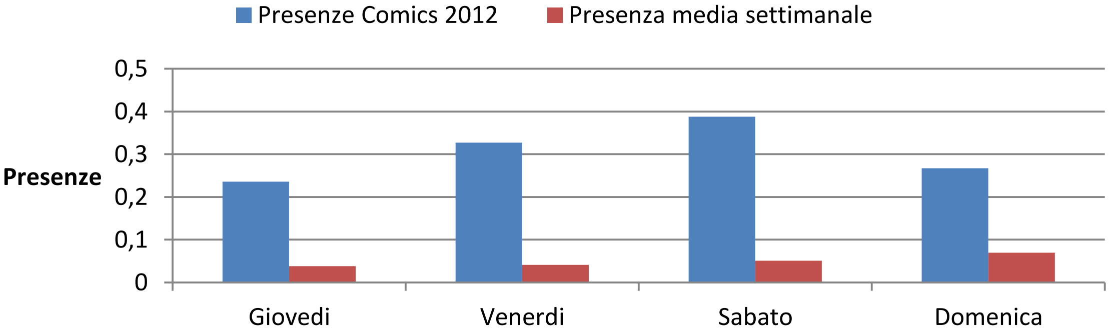
Presence of Visitors GSM - Pisa - Historical Center



INTERNET FESTIVAL 2013
FORME DI FUTURO

LUCCA COMICS AND GAMES 2012

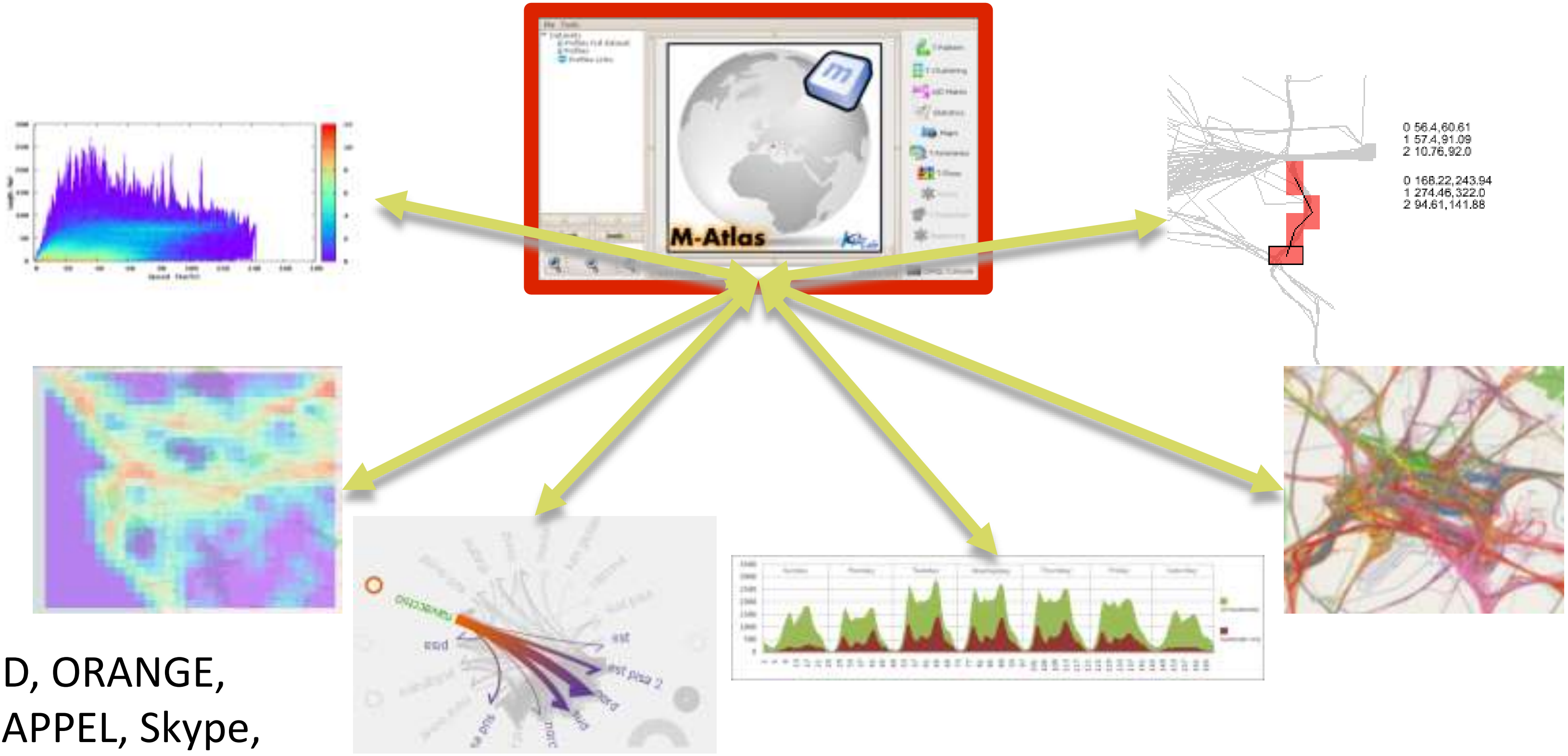
Presence of Visitors GSM - Lucca Comics 2012



Enabling technologies for Big Data analytics

M-Atlas

An analytic platform to extract, store, combine different kinds of models to build mobility knowledge discovery processes.

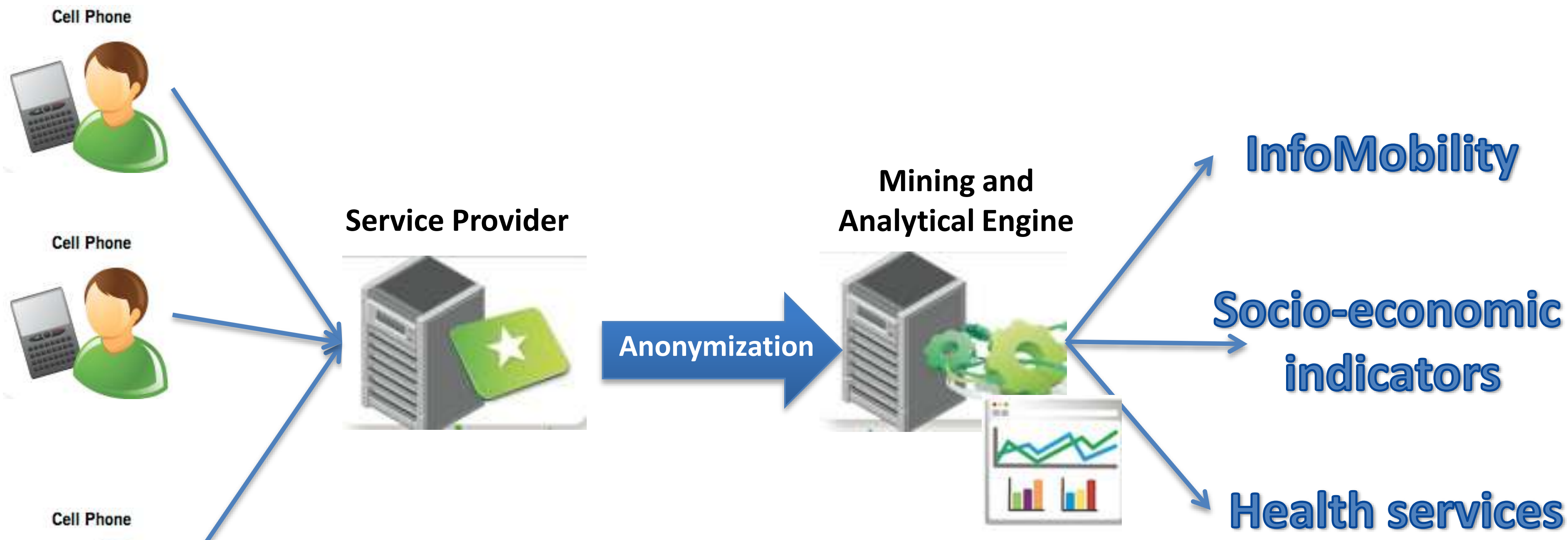


WIND, ORANGE, GOUDAPPEL, Skype, OCTOTelematics, Telecom Italia, Toyota, ENEL, ISTAT IBM, Local administrations

Google
FACULTY AWARD

One of best EU-FET results, invited for exhibition at Strasburg Parliament

Privacy-by-Design in Big Data Analytics



Cell Phone

Cell Phone

Cell Phone

Service Provider

Anonymization

Mining and Analytical Engine

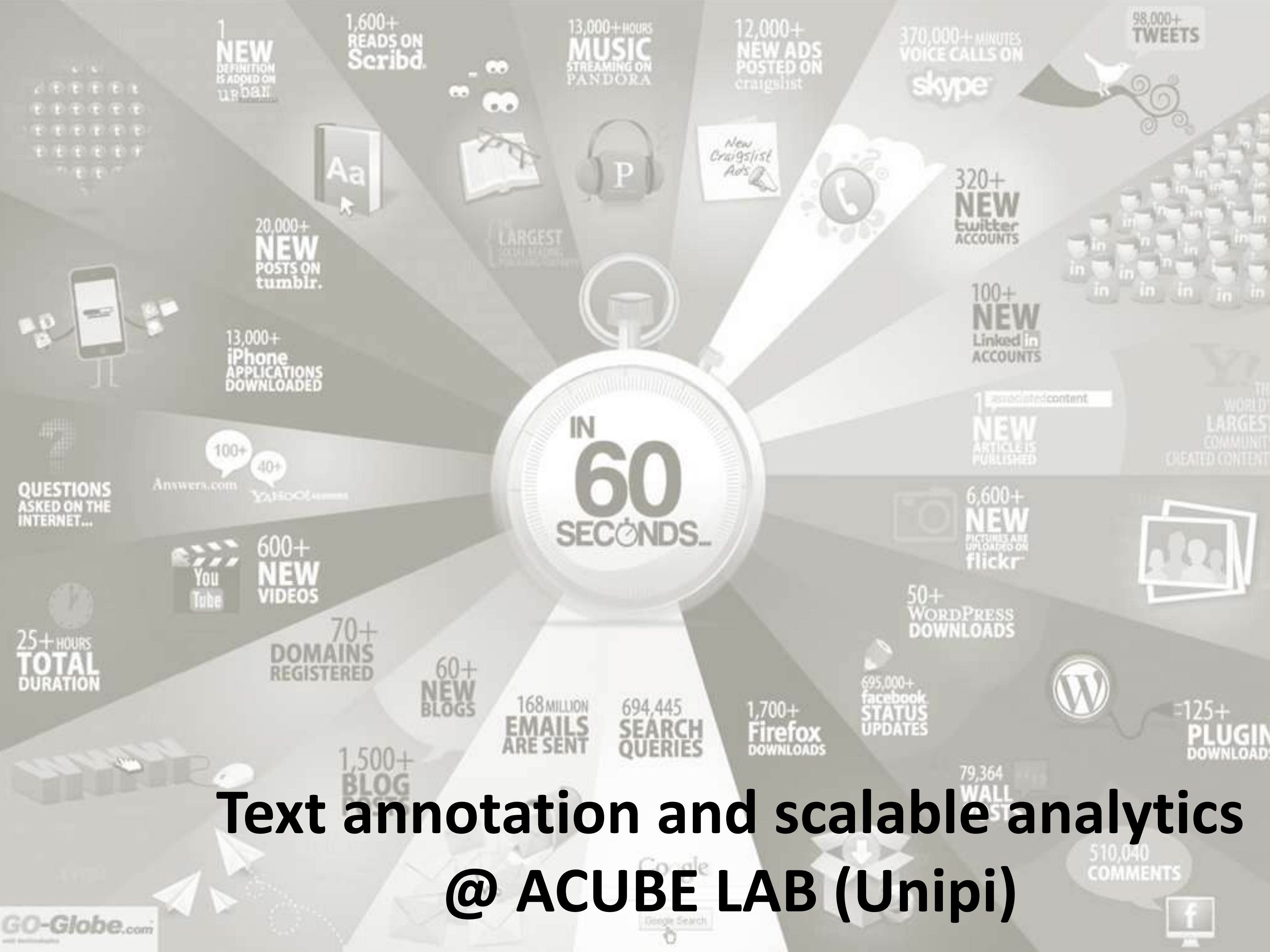
InfoMobility

Socio-economic indicators

Health services

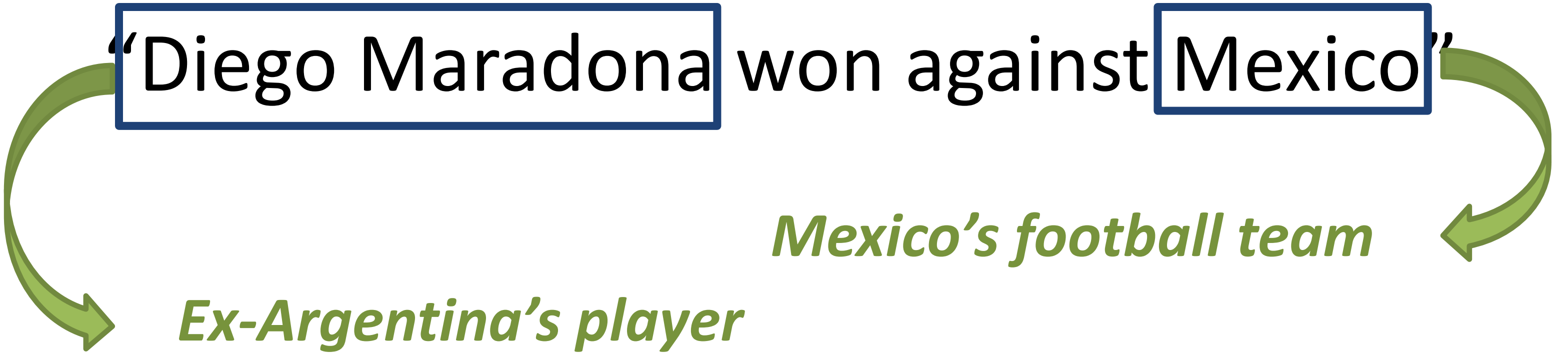
Anonymization is not trivial

De-identification is not enough


1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
New Craigslist Ads
Aa
P
LARGEST
20,000+ NEW POSTS ON tumblr.
320+ NEW twitter ACCOUNTS
100+ NEW Linked in ACCOUNTS
13,000+ iPhone APPLICATIONS DOWNLOADED
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT
IN 60 SECONDS...
100+ Answers.com
40+ YAHOO! ANSWERS
QUESTIONS ASKED ON THE INTERNET...
6,600+ NEW PICTURES ARE UPLOADED ON flickr
600+ NEW VIDEOS
You Tube
25+ HOURS TOTAL DURATION
50+ WORDPRESS DOWNLOADS
70+ DOMAINS REGISTERED
60+ NEW BLOGS
168 MILLION EMAILS ARE SENT
694,445 SEARCH QUERIES
1,700+ Firefox DOWNLOADS
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
1,500+ BLOG POSTS
79,364 WALL POSTS
510,040 COMMENTS
Google
Google Search
GO-Globe.com

Text annotation and scalable analytics @ ACUBE LAB (Unipi)

Topic-based annotation



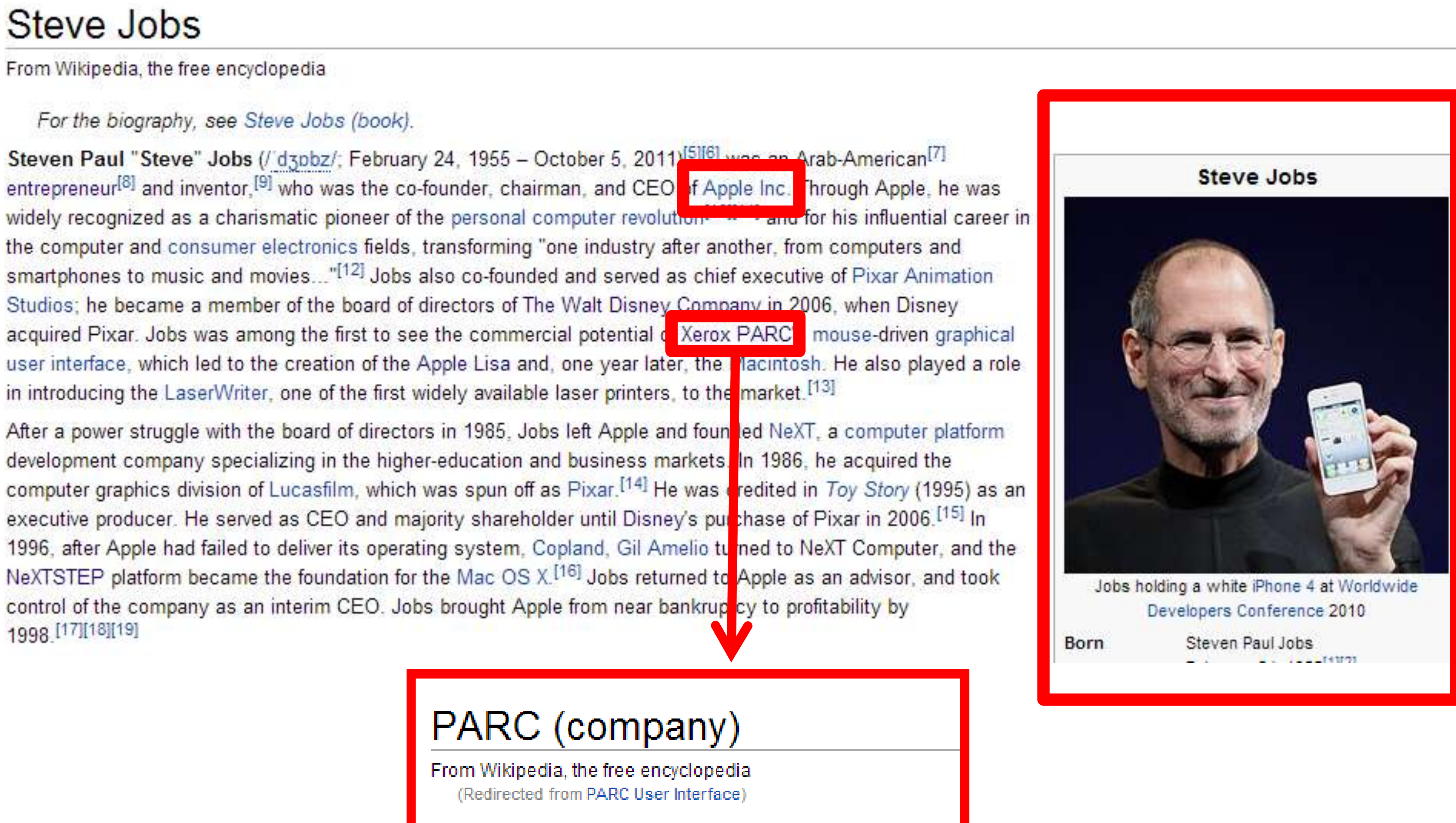
"Diego Maradona won against Mexico"

Mexico's football team

Ex-Argentina's player

Find **anchors** and annotate them with **articles** drawn from **Wikipedia!**

Wikipedia is a rich source of instances



Steve Jobs

From Wikipedia, the free encyclopedia

For the biography, see Steve Jobs (book).

Steven Paul "Steve" Jobs (/ˈdʒɒbz/; February 24, 1955 – October 5, 2011)[5][6] was an Arab-American[7] entrepreneur[8] and inventor,[9] who was the co-founder, chairman, and CEO of Apple Inc. Through Apple, he was widely recognized as a charismatic pioneer of the personal computer revolution[10][11] and for his influential career in the computer and consumer electronics fields, transforming "one industry after another, from computers and smartphones to music and movies..."[12] Jobs also co-founded and served as chief executive of Pixar Animation Studios; he became a member of the board of directors of The Walt Disney Company in 2006, when Disney acquired Pixar. Jobs was among the first to see the commercial potential of Xerox PARC's mouse-driven graphical user interface, which led to the creation of the Apple Lisa and, one year later, the Macintosh. He also played a role in introducing the LaserWriter, one of the first widely available laser printers, to the market.[13]

After a power struggle with the board of directors in 1985, Jobs left Apple and founded NeXT, a computer platform development company specializing in the higher-education and business markets. In 1986, he acquired the computer graphics division of Lucasfilm, which was spun off as Pixar.[14] He was credited in *Toy Story* (1995) as an executive producer. He served as CEO and majority shareholder until Disney's purchase of Pixar in 2006.[15] In 1996, after Apple had failed to deliver its operating system, Copland, Gil Amelio turned to NeXT Computer, and the NeXTSTEP platform became the foundation for the Mac OS X.[16] Jobs returned to Apple as an advisor, and took control of the company as an interim CEO. Jobs brought Apple from near bankruptcy to profitability by 1998.[17][18][19]

Steve Jobs

Jobs holding a white iPhone 4 at Worldwide Developers Conference 2010

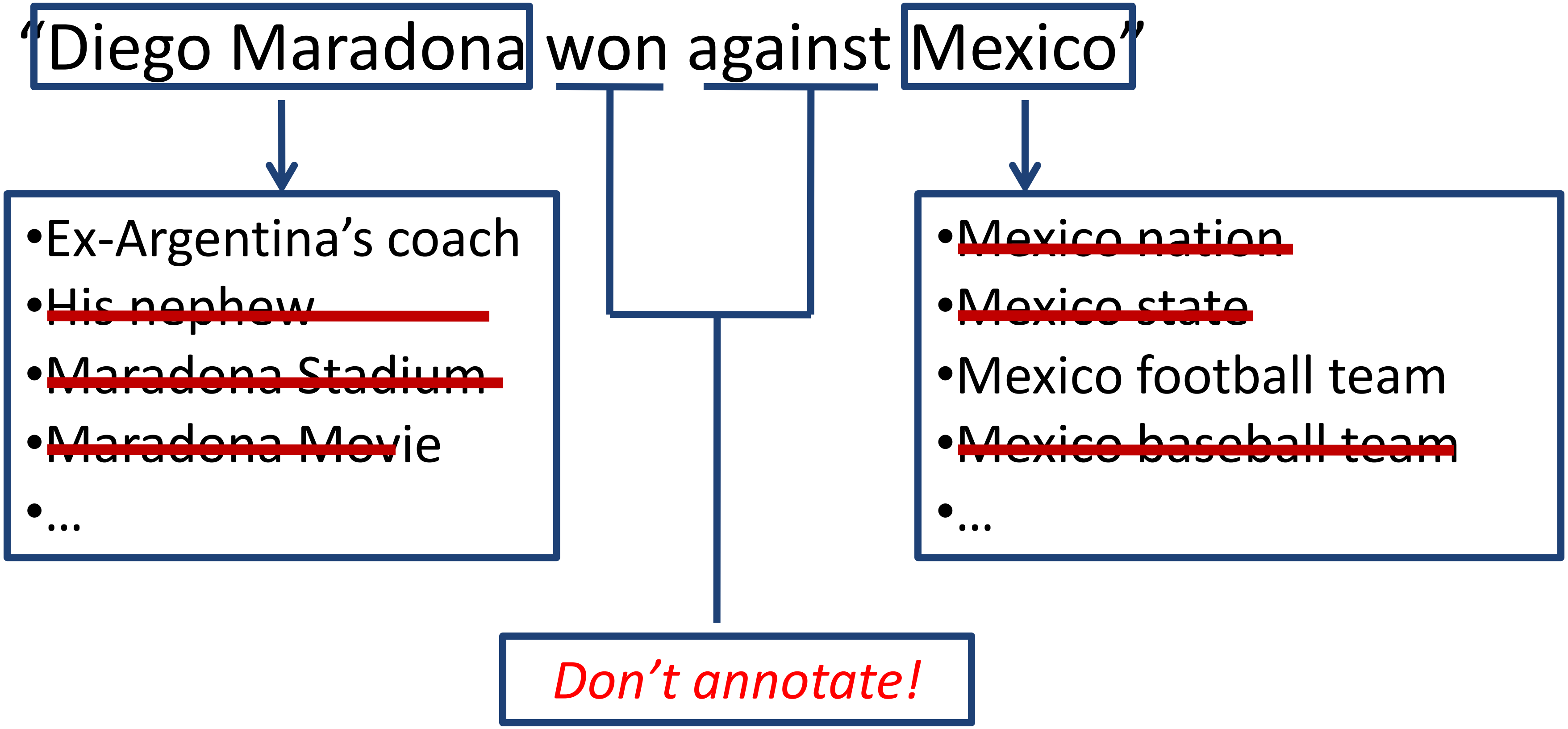
Born Steven Paul Jobs

PARC (company)

From Wikipedia, the free encyclopedia

(Redirected from PARC User Interface)

Why is it a difficult problem?

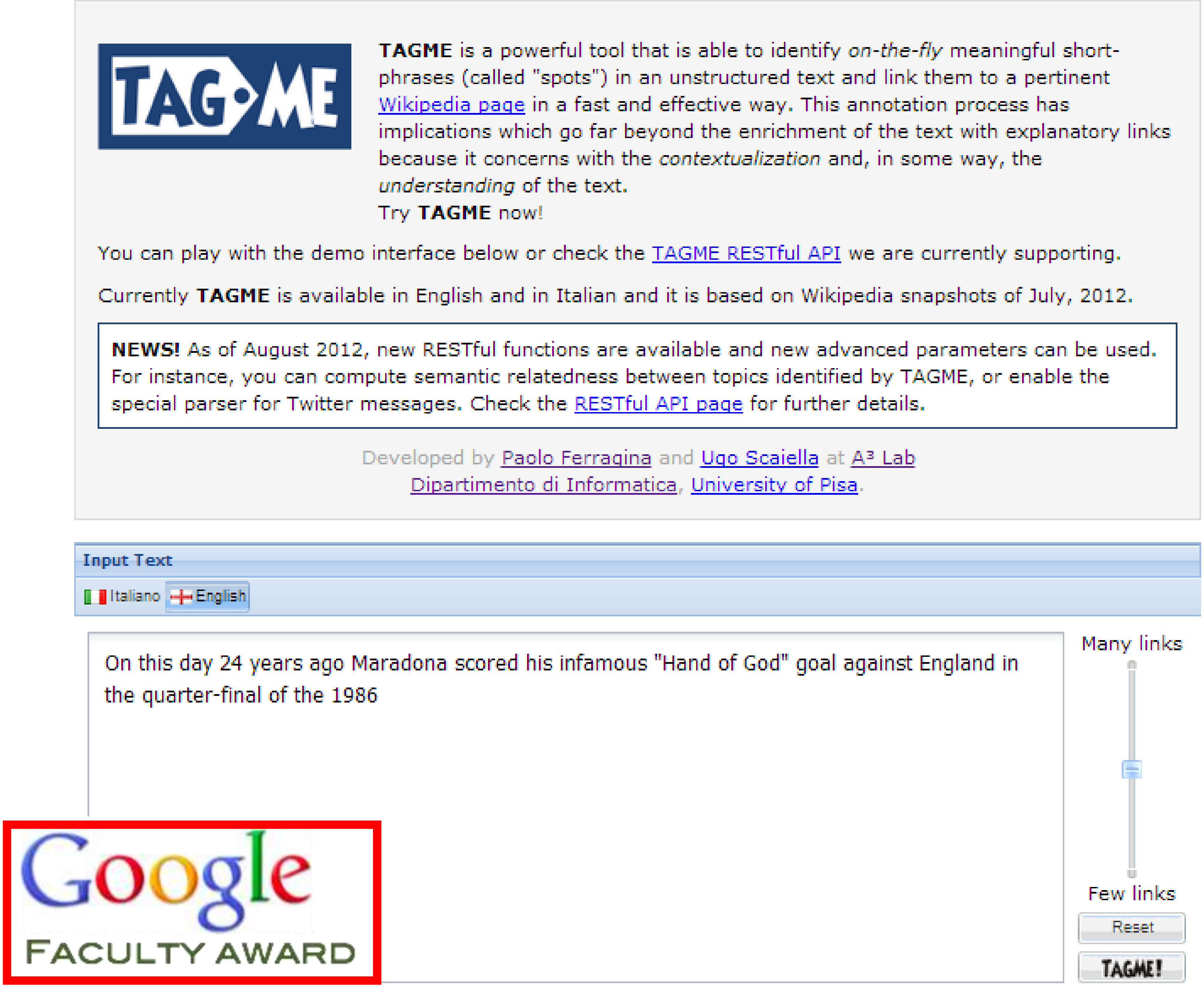


“Diego Maradona won against Mexico”

- Ex-Argentina’s coach
- ~~His nephew~~
- ~~Maradona Stadium~~
- ~~Maradona Mov~~ie
- …

- ~~Mexico nation~~
- ~~Mexico state~~
- Mexico football team
- ~~Mexico baseball team~~
- …

Don’t annotate!



TAGME is a powerful tool that is able to identify *on-the-fly* meaningful short-phrases (called "spots") in an unstructured text and link them to a pertinent Wikipedia page in a fast and effective way. This annotation process has implications which go far beyond the enrichment of the text with explanatory links because it concerns with the *contextualization* and, in some way, the *understanding* of the text.
Try **TAGME** now!

You can play with the demo interface below or check the TAGME RESTful API we are currently supporting.

Currently **TAGME** is available in English and in Italian and it is based on Wikipedia snapshots of July, 2012.

NEWS! As of August 2012, new RESTful functions are available and new advanced parameters can be used. For instance, you can compute semantic relatedness between topics identified by TAGME, or enable the special parser for Twitter messages. Check the RESTful API page for further details.

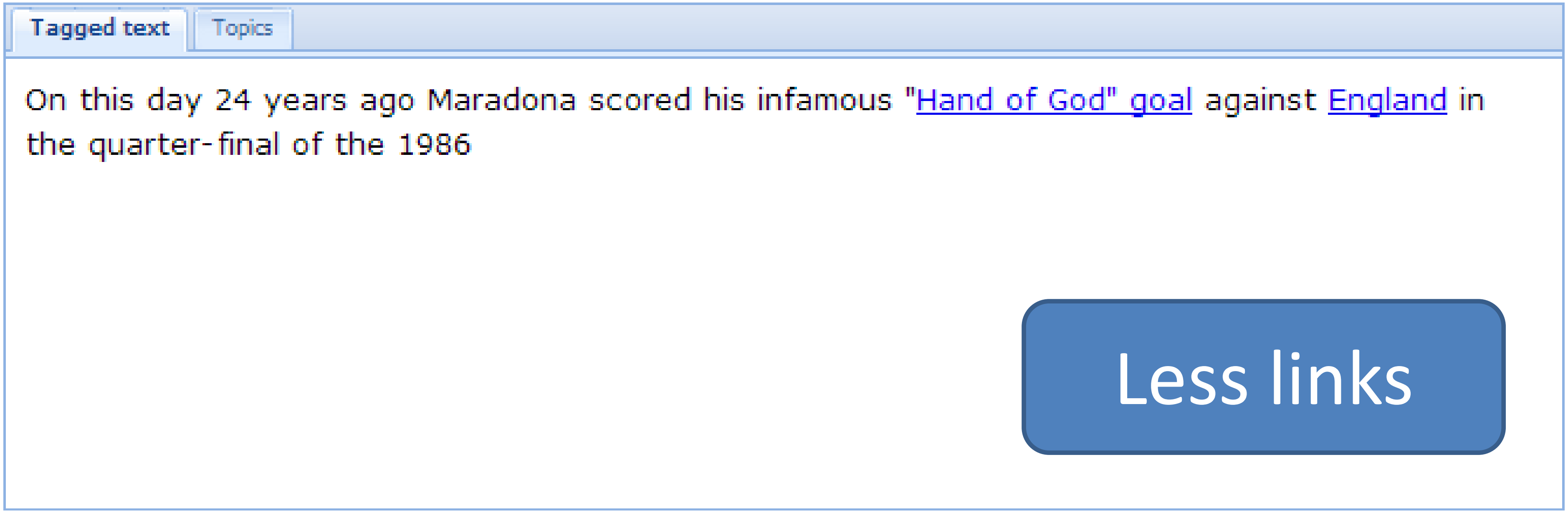
Developed by Paolo Ferragina and Ugo Scaiella at A³ Lab
Dipartimento di Informatica, University of Pisa.

Input Text

Italiano English

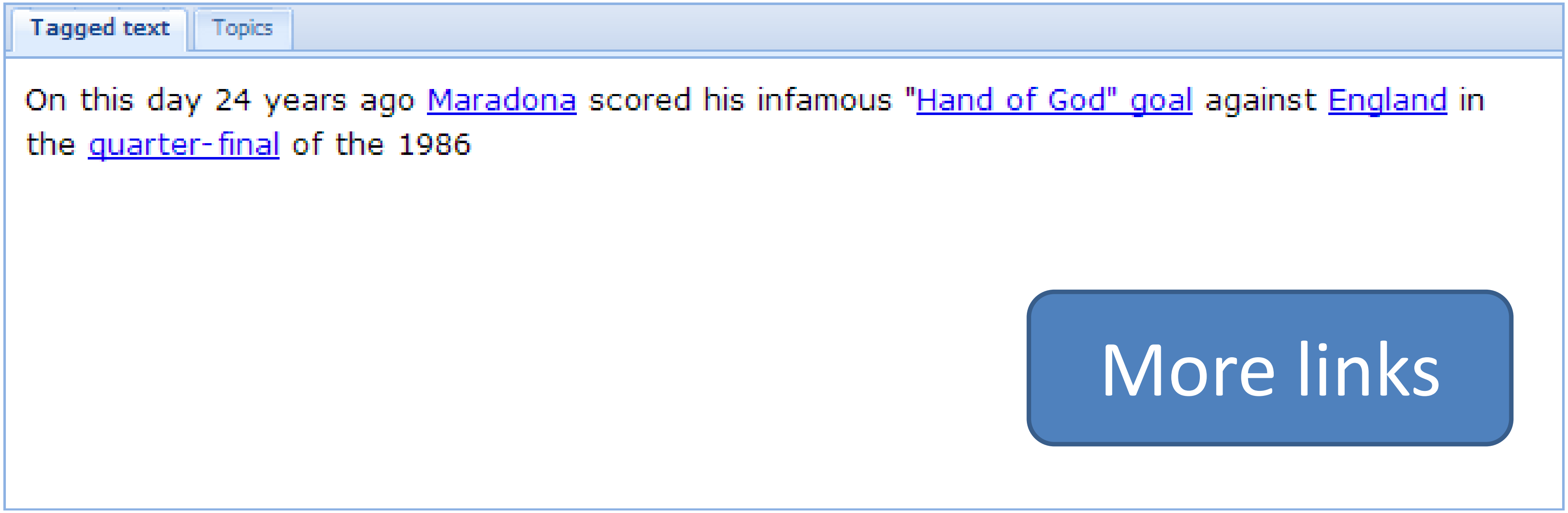
On this day 24 years ago Maradona scored his infamous "Hand of God" goal against England in the quarter-final of the 1986

Google
FACULTY AWARD



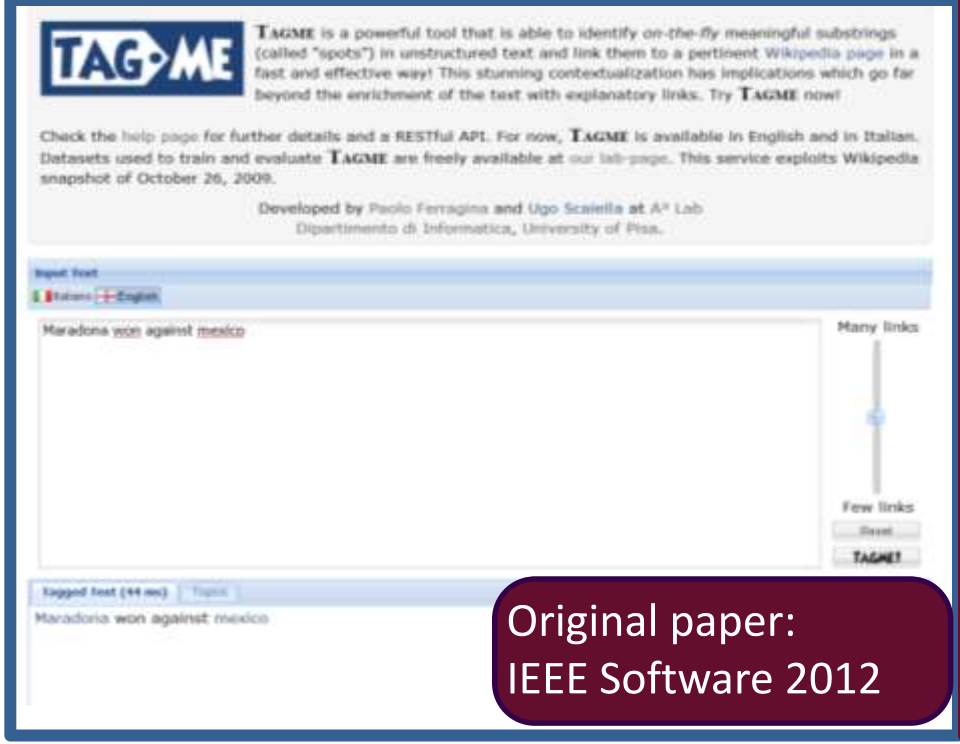
On this day 24 years ago Maradona scored his infamous "Hand of God" goal against England in the quarter-final of the 1986

Less links



On this day 24 years ago Maradona scored his infamous "Hand of God" goal against England in the quarter-final of the 1986

More links



TAGME is a powerful tool that is able to identify *on-the-fly* meaningful substrings (called "spots") in unstructured text and link them to a pertinent Wikipedia page in a fast and effective way! This stunning contextualization has implications which go far beyond the enrichment of the text with explanatory links. Try **TAGME** now!

Check the help page for further details and a RESTful API. For now, **TAGME** is available in English and in Italian. Datasets used to train and evaluate **TAGME** are freely available at our lab-page. This service exploits Wikipedia snapshot of October 26, 2009.

Developed by Paolo Ferragina and Ugo Scaiella at A³ Lab
Dipartimento di Informatica, University of Pisa.

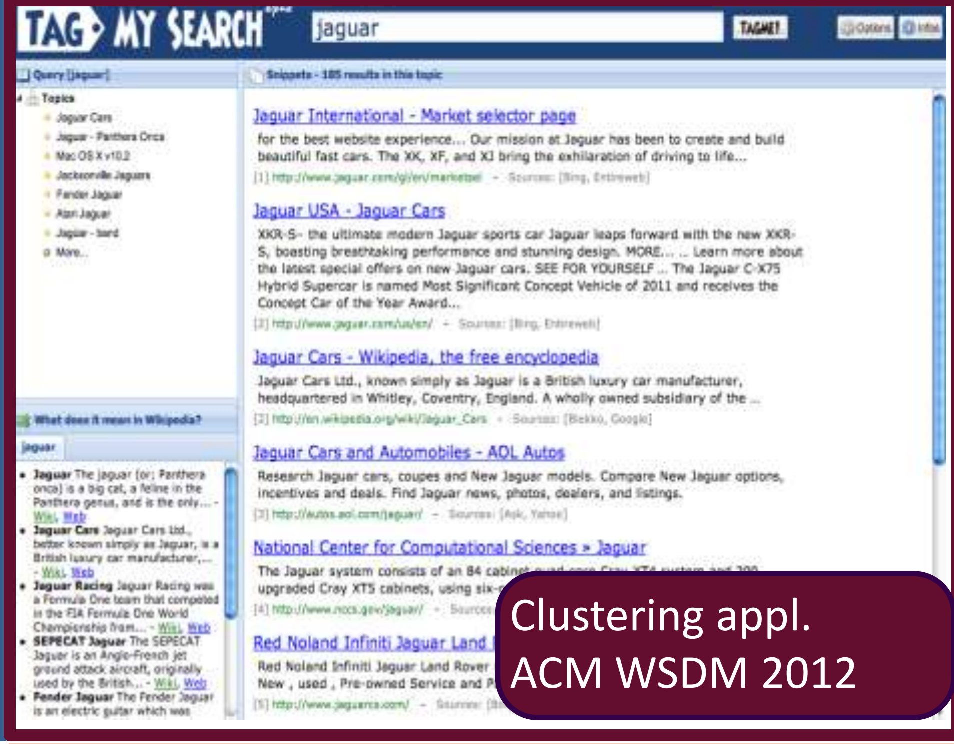
Input Text

Italiano | English

Maradona won against mexico

Maradona won against mexico

Original paper:
IEEE Software 2012



jaguar

Query [jaguar]

Topics
- Jaguar Cars
- Jaguar - Panthera Onca
- Mac OS X v10.2
- Jacksonville Jaguars
- Fender Jaguar
- Atari Jaguar
- Jaguar - bard
- More...

What does it mean in Wikipedia?

jaguar

- **Jaguar** The jaguar (or; Panthera onca) is a big cat, a feline in the Panthera genus, and is the only... - Wiki, Web
- **Jaguar Cars** Jaguar Cars Ltd., better known simply as Jaguar, is a British luxury car manufacturer,... - Wiki, Web
- **Jaguar Racing** Jaguar Racing was a Formula One team that competed in the FIA Formula One World Championship from... - Wiki, Web
- **SEPECAT Jaguar** The SEPECAT Jaguar is an Anglo-French jet ground attack aircraft, originally used by the British... - Wiki, Web
- **Fender Jaguar** The Fender Jaguar is an electric guitar which was

Snippets - 185 results in this topic

Jaguar International - Market selector page
for the best website experience... Our mission at Jaguar has been to create and build beautiful fast cars. The XK, XF, and XJ bring the exhilaration of driving to life...
[1] http://www.jaguar.com/gl/en/marketsel - Sources: [Bing, Entireweb]

Jaguar USA - Jaguar Cars
XKR-S- the ultimate modern Jaguar sports car Jaguar leaps forward with the new XKR-S, boasting breathtaking performance and stunning design. MORE... .. Learn more about the latest special offers on new Jaguar cars. SEE FOR YOURSELF ... The Jaguar C-X75 Hybrid Supercar is named Most Significant Concept Vehicle of 2011 and receives the Concept Car of the Year Award...
[2] http://www.jaguar.com/us/en/ - Sources: [Bing, Entireweb]

Jaguar Cars - Wikipedia, the free encyclopedia
Jaguar Cars Ltd., known simply as Jaguar is a British luxury car manufacturer, headquartered in Whitley, Coventry, England. A wholly owned subsidiary of the ...
[2] http://en.wikipedia.org/wiki/Jaguar_Cars - Sources: [Bekko, Google]

Jaguar Cars and Automobiles - AOL Autos
Research Jaguar cars, coupes and New Jaguar models. Compare New Jaguar options, incentives and deals. Find Jaguar news, photos, dealers, and listings.
[3] http://autos.aol.com/jaguar/ - Sources: [Ask, Yahoo]

National Center for Computational Sciences » Jaguar
The Jaguar system consists of an 84 cabinet quad-core Cray XT4 system and 200 upgraded Cray XT5 cabinets, using six-c
[4] http://www.nccs.gov/jaguar/ - Source

Red Noland Infiniti Jaguar Land
Red Noland Infiniti Jaguar Land Rover
New , used , Pre-owned Service and P
[5] http://www.jaguarcs.com/ - Source: [B

Clustering appl.
ACM WSDM 2012

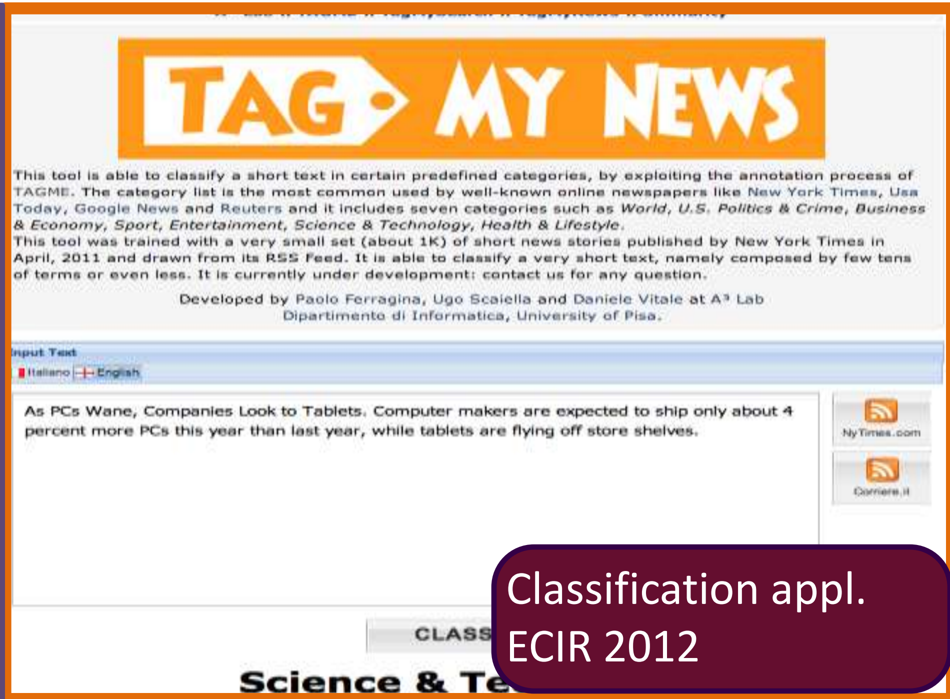
Details on…

http://acube.di.unipi.it/tagme

Regional project with:

net7

SPAZIODATI



This tool is able to classify a short text in certain predefined categories, by exploiting the annotation process of TAGME. The category list is the most common used by well-known online newspapers like New York Times, Usa Today, Google News and Reuters and it includes seven categories such as *World, U.S. Politics & Crime, Business & Economy, Sport, Entertainment, Science & Technology, Health & Lifestyle.*
This tool was trained with a very small set (about 1K) of short news stories published by New York Times in April, 2011 and drawn from its RSS Feed. It is able to classify a very short text, namely composed by few tens of terms or even less. It is currently under development: contact us for any question.

Developed by Paolo Ferragina, Ugo Scaiella and Daniele Vitale at A³ Lab
Dipartimento di Informatica, University of Pisa.

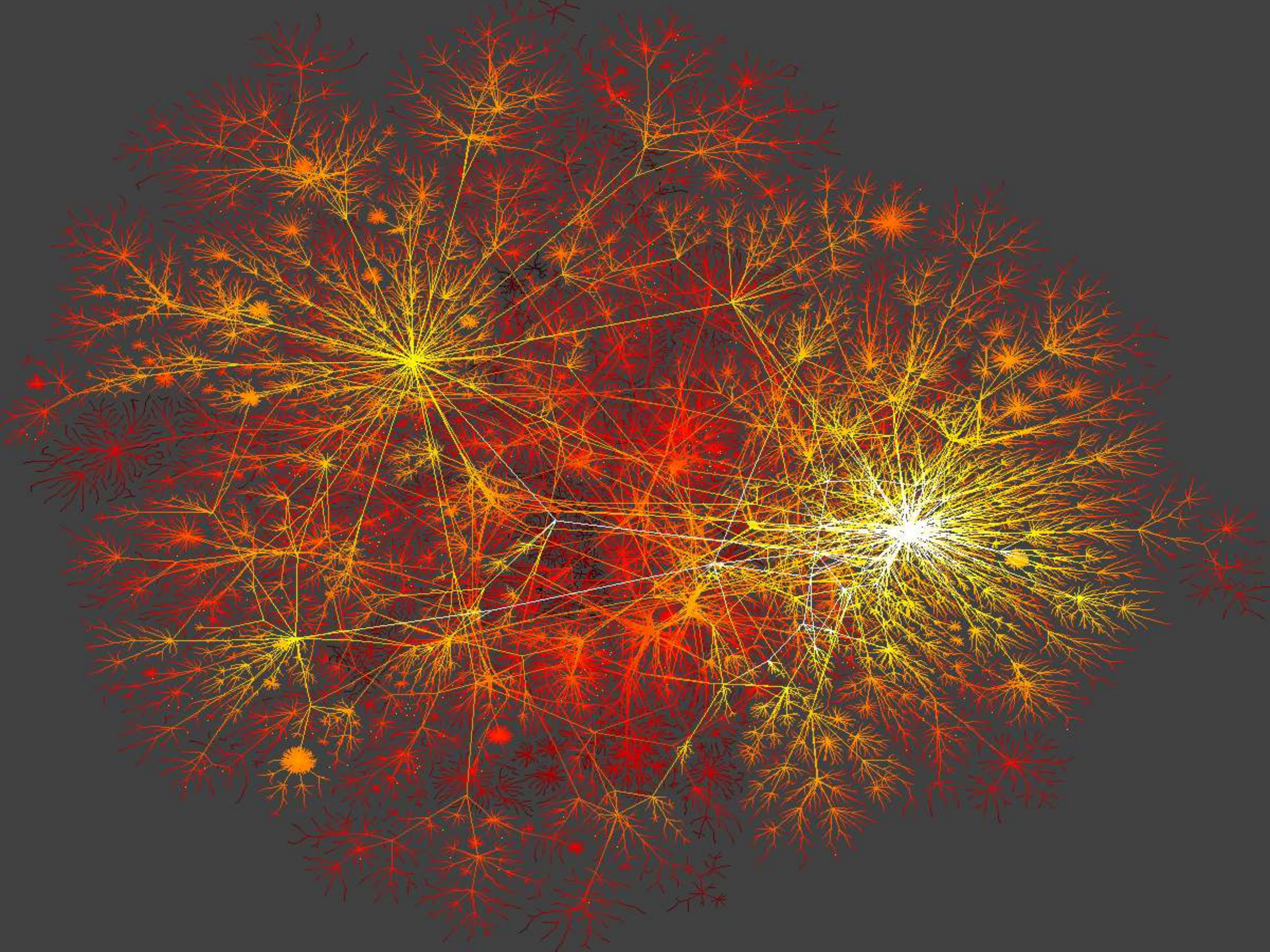
Input Text

Italiano | English

As PCs Wane, Companies Look to Tablets. Computer makers are expected to ship only about 4 percent more PCs this year than last year, while tablets are flying off store shelves.

Science & Te

Classification appl.
ECIR 2012



- Nodes ≅ users, entities **(~ 1 bil)**
- Edges *explicit* = friend, follower… **(~ 10 bil)**
- Edges *implicit* = similarity, click… **(» 100 bil)**
- Textual Data = post, tweet, news… **(> Ptb)**

The Knowledge Graph

Learn more about one of the key breakthroughs behind the future of search.

See it in action

Discover answers to questions you never thought to ask, powered by the Knowledge Graph.

New searching algorithms

The goal is:

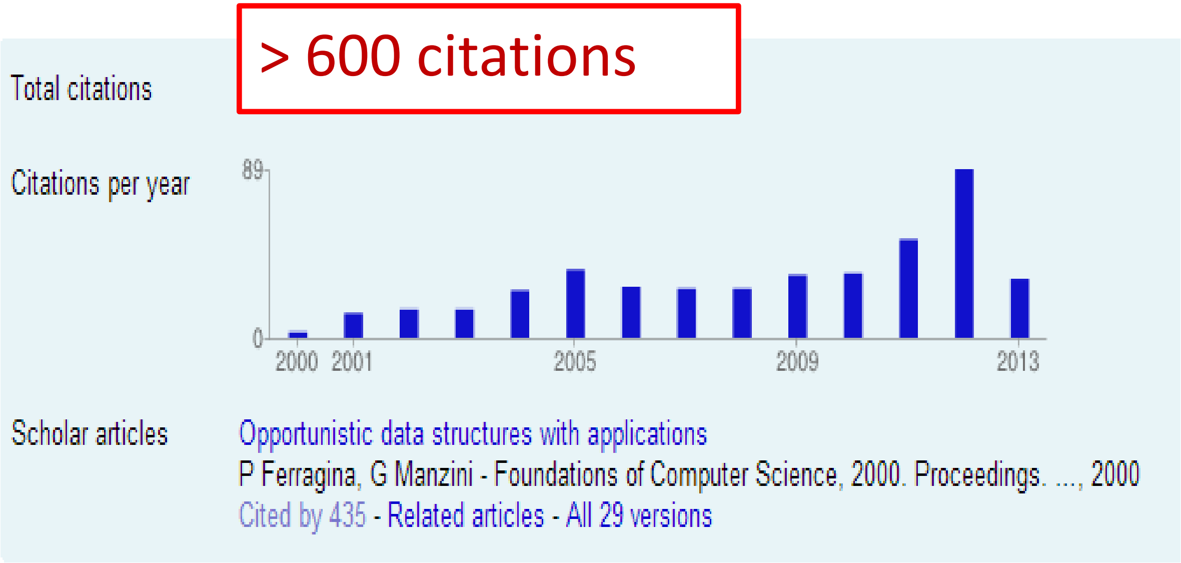
- Minimize the occupied space
- Maximize the **substring-search** throughput

2007

Under US-patenting

We were the first to show how to search *bzip*-ed data

> 600 citations



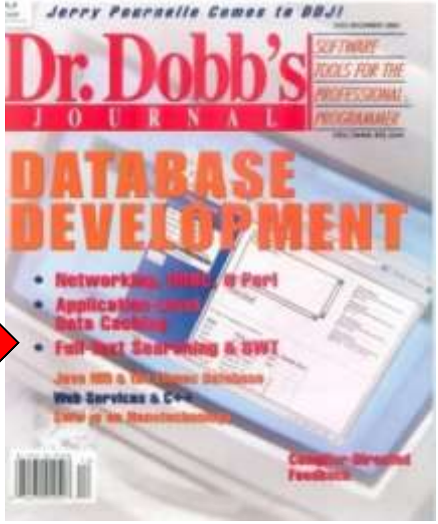
Total citations

Citations per year

Scholar articles

Opportunistic data structures with applications
P Ferragina, G Manzini - Foundations of Computer Science, 2000. Proceedings. ..., 2000
Cited by 435 - Related articles - All 29 versions

2003


Jerry Pournelle Comes to DDJ!
Dr. Dobb's
JOURNAL
SOFTWARE TOOLS FOR THE PROFESSIONAL PROGRAMMER
DATABASE DEVELOPMENT
Networking, ..., & Perl
Application-... Data Caching
Full-Text Searching & BWT
Web Services & C++

Collaborations

Google
FACULTY AWARD

tiscali.

Google

Y!™

Bassilichi

net7

SPAZIODATI

4 submitted US patents: Yahoo and NY Univ.
2 accepted US patents: Rutgers Univ. and AT&T-Lucent

SoBigData

Bootstrap Workshop

Towards a European Laboratory

on Big Data Analytics and Social Mining

18 July 2013, h **10:00-17:30**
Auditorium of the National Research Council
Area della Ricerca **CNR**, Via Moruzzi 1, **Pisa**

Program

10:00: Reception and refreshments

10:30-11:30: Setting the stage

Welcome address D. Laforenza (President CNR Research Campus), N. De Francesco (Univ. Pisa, Vice-Rector)
Opening remarks C. Montani (ISTI-CNR, Director), F. Turini (Dip. Informatica Univ. Pisa, Director).
Towards a Euro Lab on Big Data Analytics & Social Mining M. Conti (DIITET-CNR, Director)
Big Data Analytics & Social Mining for Science and Society F. Giannotti (ISTI-CNR)
Democratizing big data: the ethical challenges of social mining D. Pedreschi (Univ. Pisa)

11:30-12:00: Keynote

Big data big insights: the coming age of computational social science
D. Lazer, Professor of Political Science and Computer and Information Science, Northeastern University, Boston, MA. Director, Program on Networked Governance, Harvard University

12:00-13:30: Panel – Big data & social mining: new models for participation and policy making

S. Targetti (Regione Toscana, Vice-President), F. Accordino (EC), F. Sestini (EC),E. Baldacci(ISTAT), C. Comella (Garante Privacy), F. Marzano (Stati Innovazione)
Moderator: L. De Biase (Il Sole 24 Ore)

13:30-14:30: Lunch break

14:30-16:00: SoBigData pills

Monitoring trend and engagement with social media mining M. Tesconi (IIT-CNR)
Exploring the structure of society
A. Passarella (IIT-CNR)
Sentiment quantification and opinion surveys
F. Sebastiani (ISTI-CNR)
Good answers for difficult questions R. Perego (ISTI-CNR)
Understanding human mobility C. Renso (ISTI-CNR | Univ. Pisa)
Big data in finance and economics
F. Lillo (Scuola Normale Superiore), G. Caldarelli (IMT Lucca)
Big data and official statistics – monitoring poverty/well-being at any scale
M. Pratesi (Univ. Pisa), F. Maggino (Univ. Firenze)
Do you need a big computer or a great algorithm ? P. Ferragina (Univ. Pisa)

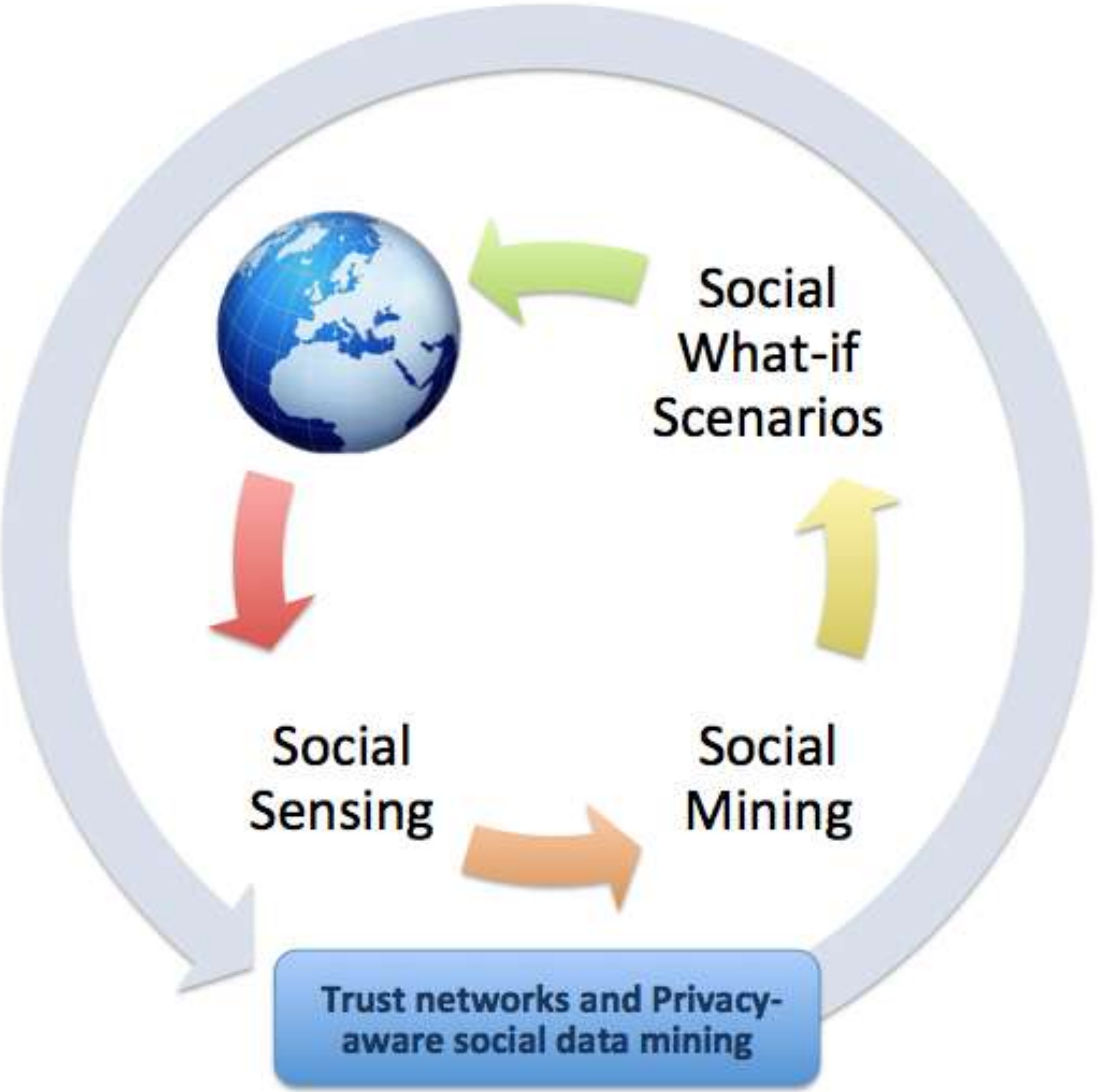
16:00-17:30: Panel – Big data & social mining: new models for social innovation and business

R. Soru (Tiscali), O. Cicchetti (Telecom Italia), T. Martino (Octotelematics), G. Gigliucci (ENEL Ricerca), A. Di Benedetto (CNA, Ass. Giovani Industriali)
Moderator: L. De Biase (Il Sole 24 Ore)

17:30: Conclusion

Registration: www.sobigdata.eu/registration

DIITET-Cnr



Social What-if Scenarios

Social Sensing

Social Mining

Trust networks and Privacy-aware social data mining

ACUBE Laboratory | acube.di.unipi.it

Knowledge Discovery & Data Mining Laboratory | kdd.isti.cnr.it