BIG DATA ANALYTICS

Fosca Giannotti and Roberto Trasarti
Pisa KDD Lab, ISTI-CNR & Univ. Pisa
http://www-kdd.isti.cnr.it/

DIPARTIMENTO DI INFORMATICA – Università di Pisa

anno accademico 2015/2016

BRIGHT 2015 - La notte dei ricercatori in Toscana

••• 25 settembre 2015 •••

http://nottedeiricercatori.pisa.it/ • via Giuseppe Moruzzi 1 - 56124 PISA

- 17-19 “O Bog data è meglio Pelè”
- 18:30-19:30 La COOP sei tu..dimmi cosa compri ti dirà chi sei

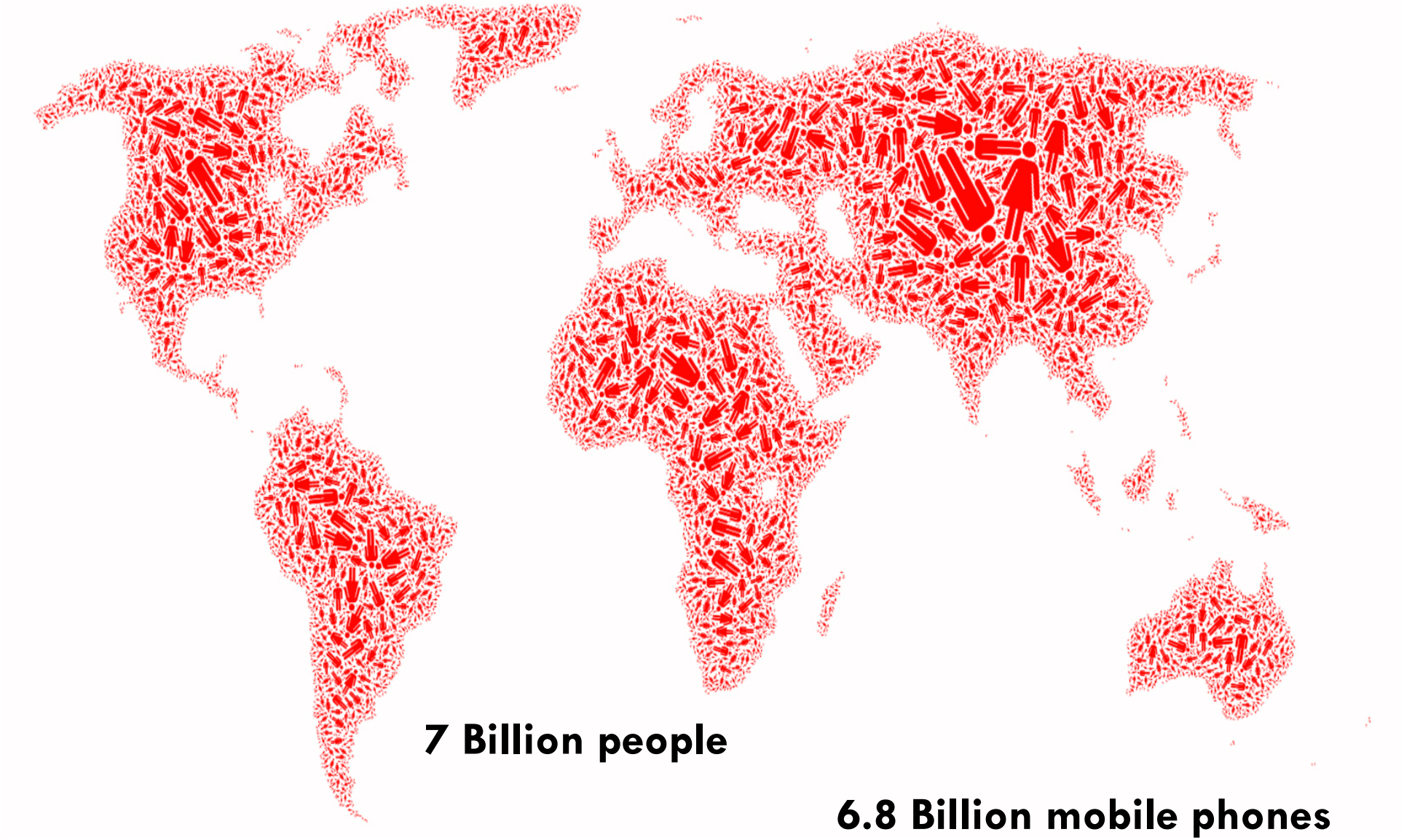
LECTURE 1:
SOCIAL MINING & BIG DATA

OPPORTUNITY & RISKS

FOSCA GIANNOTTI
KNOWLEDGE DISCOVERY & DATA MINING LAB. – ISTI CNR, PISA

Ars2015 - Capri April 30

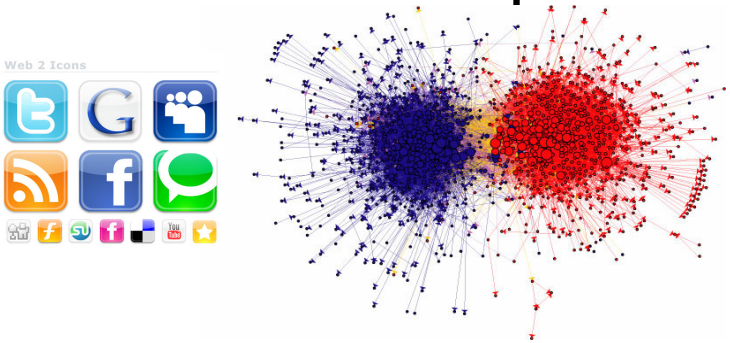

2005
Luca Bruno / AP
2013
NBC NEWS
Michael Sohn / AP


7 Billion people
6.8 Billion mobile phones

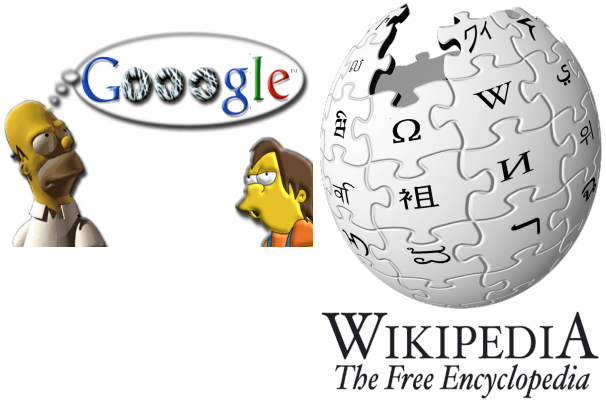
Digital Footprints of Human Activities

Shopping patterns & lifestyle

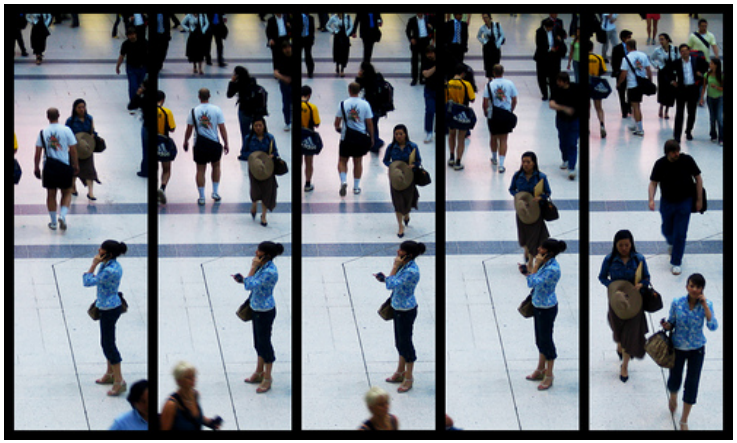
Relationships & social ties

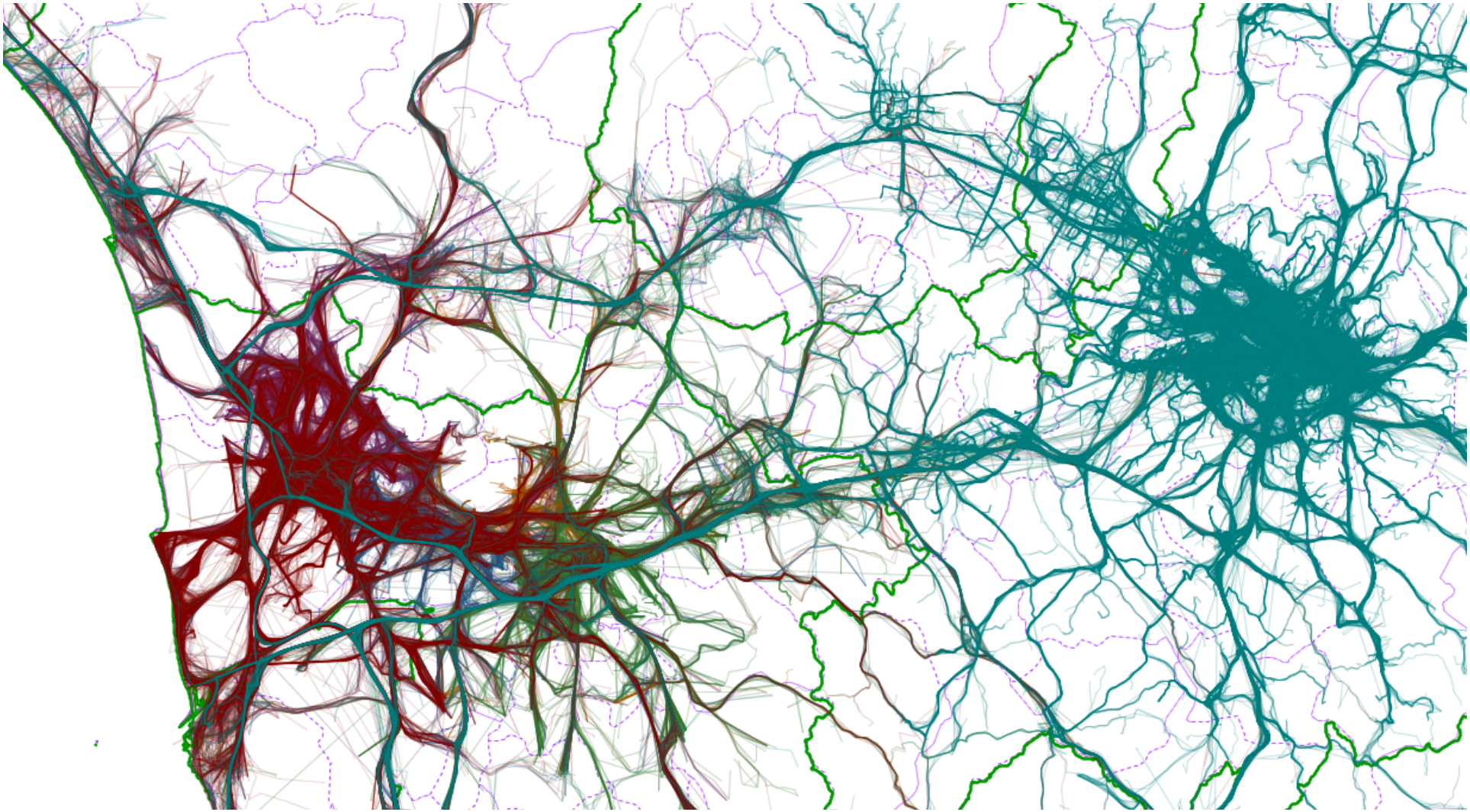


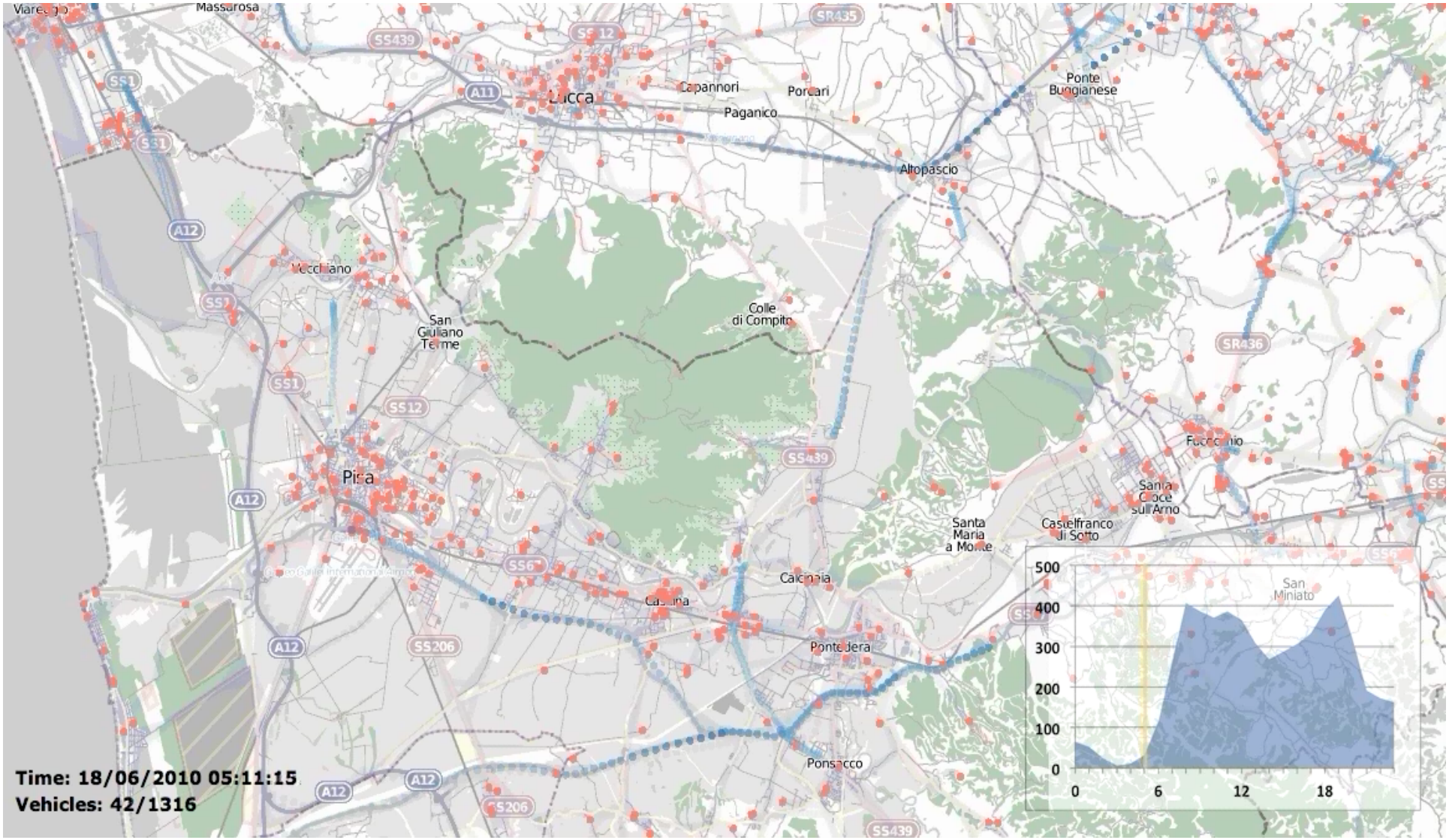
Desires, opinions, sentiments

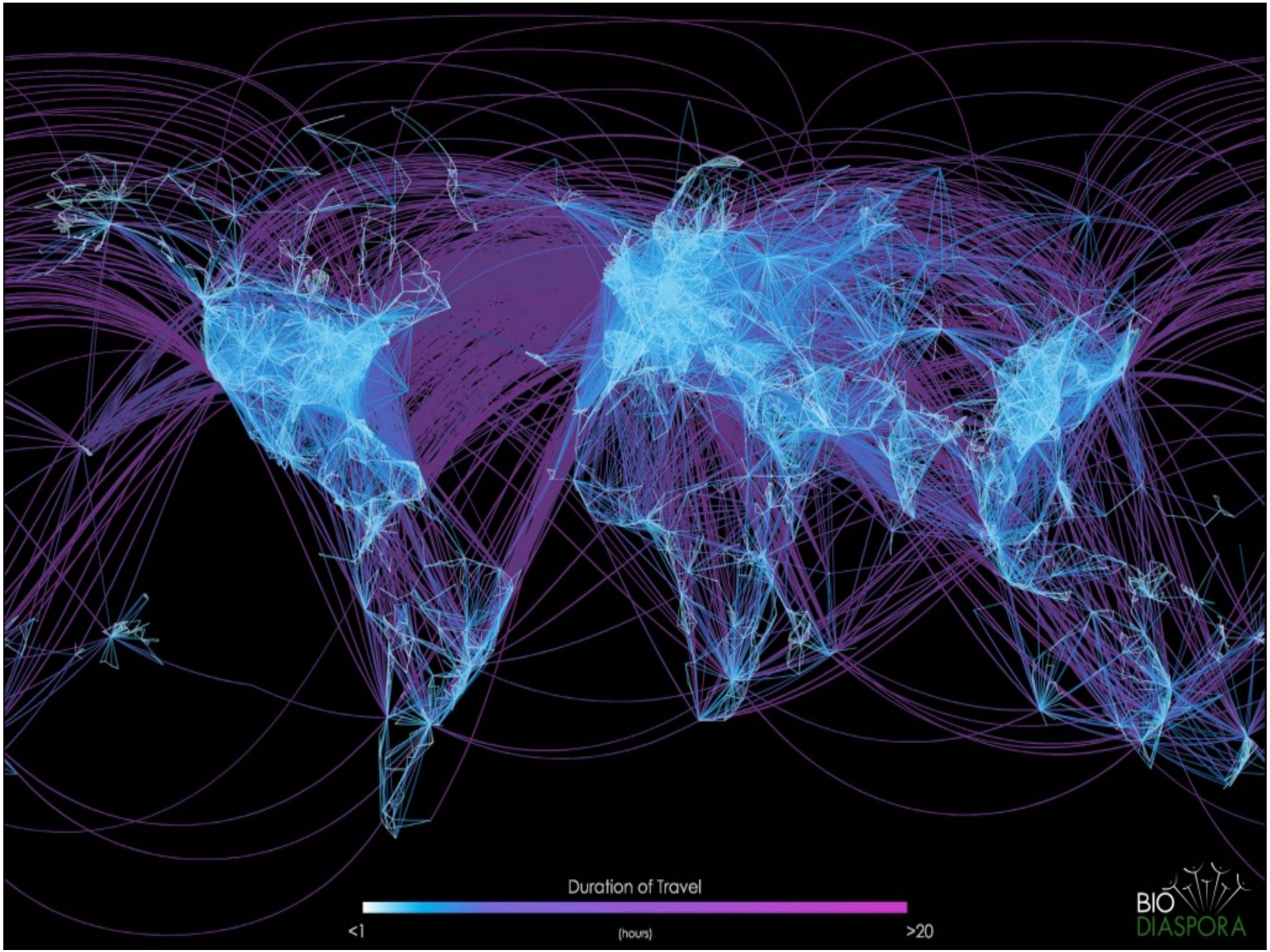


Movements






Massarosa
SS439
SS12
SR435
SS1
A11
Lucca
Capannori
Porcari
Paganico
Ponte
Buggianese
SS1
Altopascio
A12
Vecchiano
SS1
San
Giuliano
Terme
Colle
di Compito
SR436
SS1
SS12
Pisa
SS439
A12
Santa
Maria
a Monte
Caselfranco
di Sotto
SS6
Calcinaia
Cascina
SS206
Pontedera
A12
Ponsacco
A12
SS206
SS439
500
400
300
200
100
0
San
Miniato
0
6
12
18
Time: 18/06/2010 05:11:15
Vehicles: 42/1316


Duration of Travel
<1
(hours)
>20
BIO
DIASPORA

POLLICINI DIGITALI

La Vita Nova, e-magazine de Il Sole 24 Ore

Fosca Giannotti, Dino Pedreschi

Big Data & Social Mining

The Social Microscope:

a tool to

measure, understand,

and possibly predict

human behavior

SOCIAL MINING: MAKING SENSE OF BIG DATA

BIG DATA & NEW QUESTIONS TO ASK

Google Flu Trends

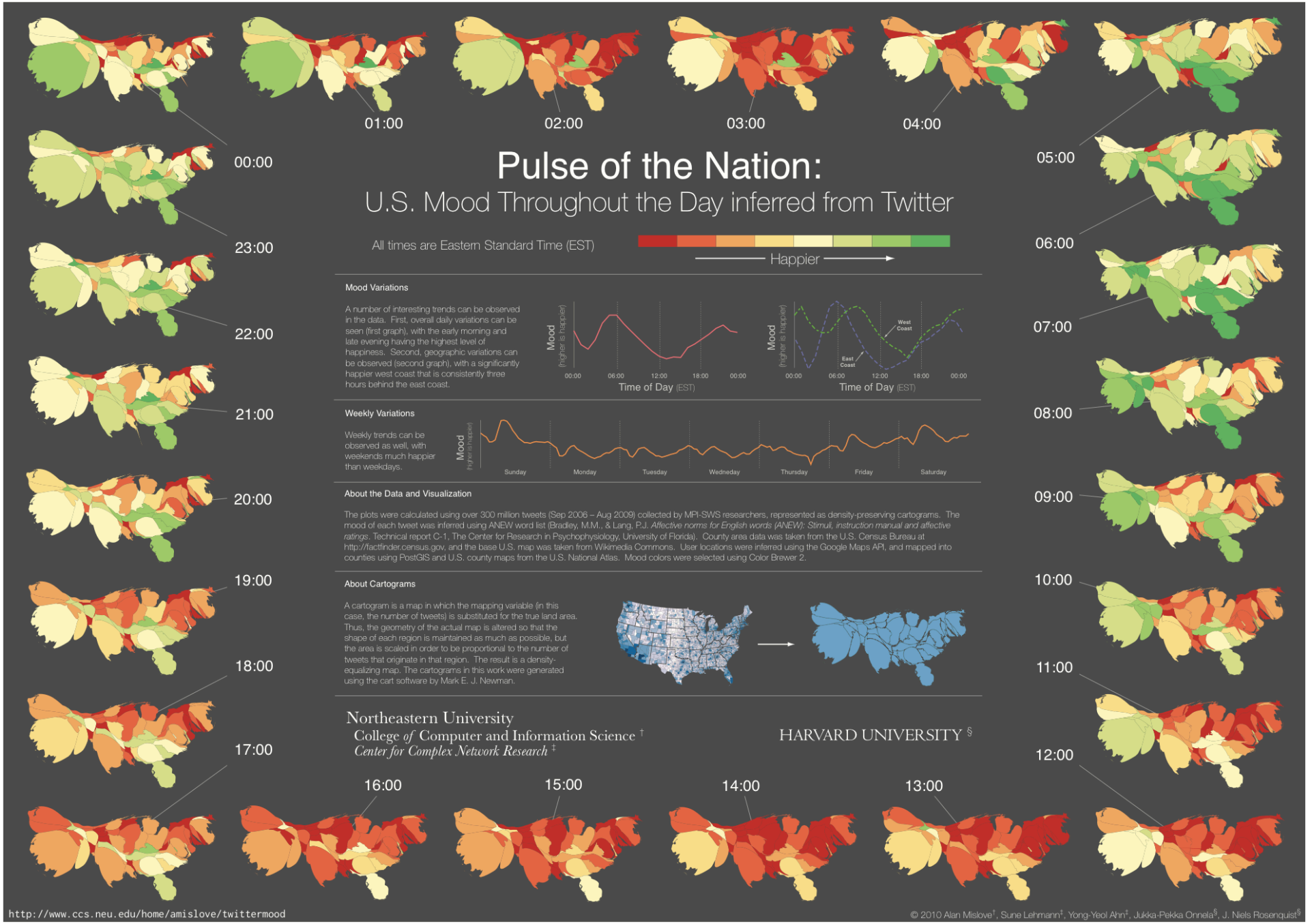
Google

Detecting influenza epidemics using search engine query data

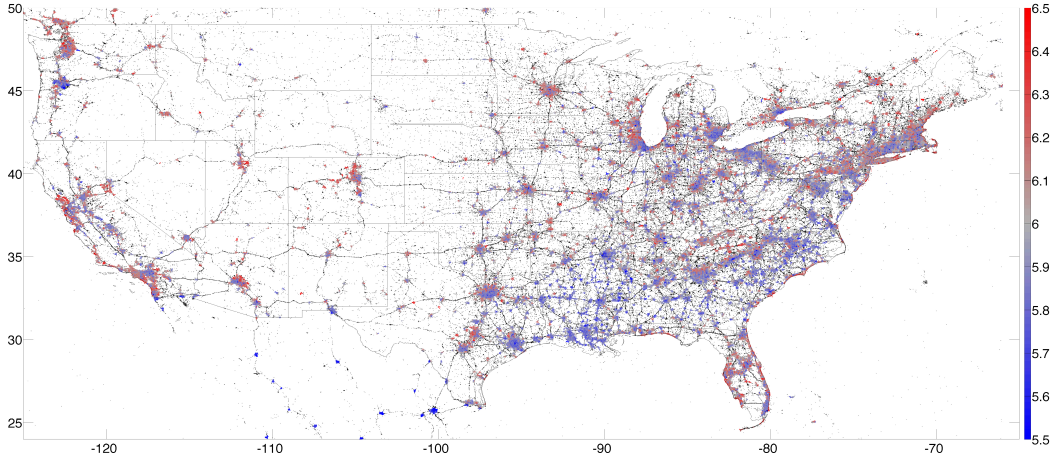
Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

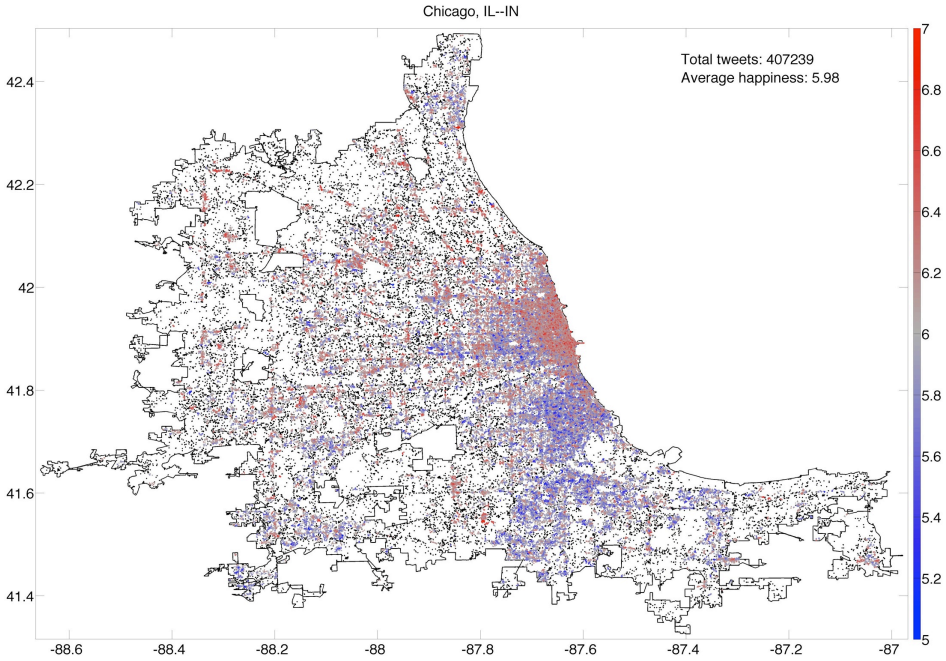
[1]*Google Inc.* [2]*Centers for Disease Control and Prevention*

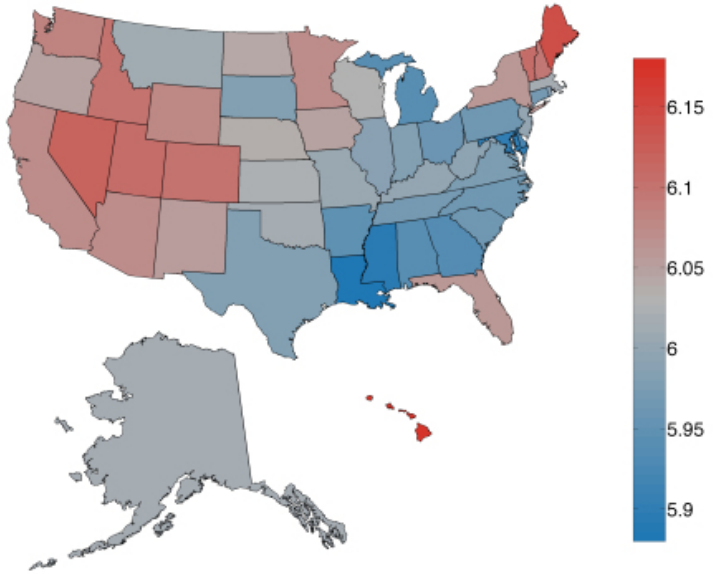
Nature 457, 1012-1014 (2009)



where-is-the-happiest-city-in-the-usa?

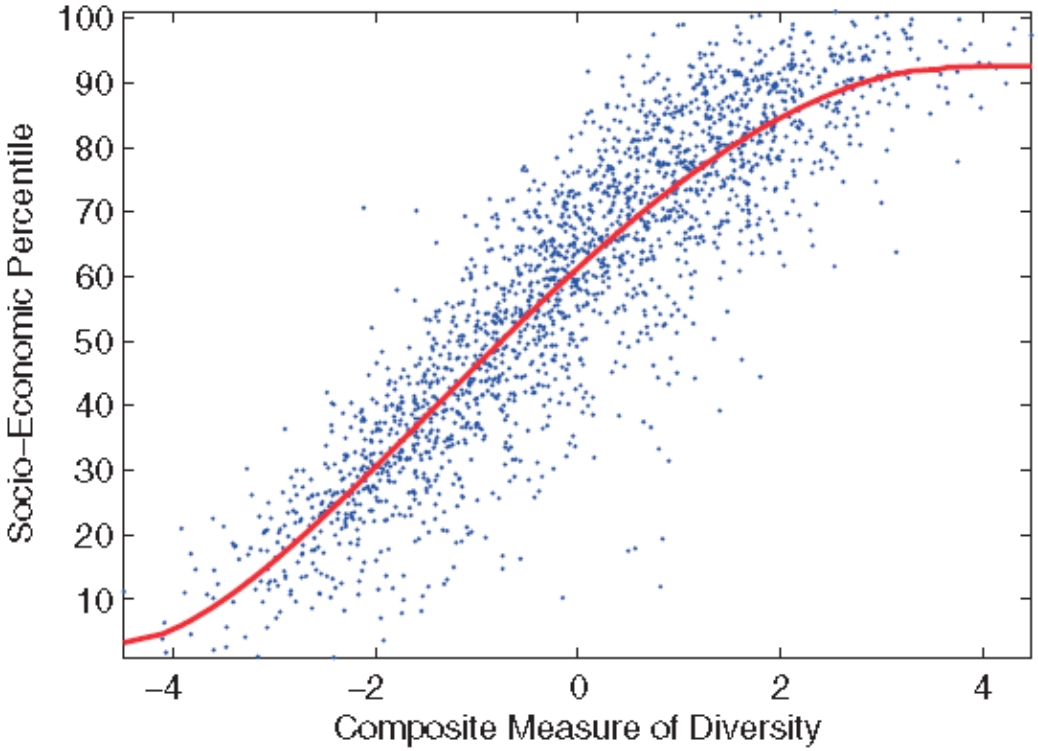


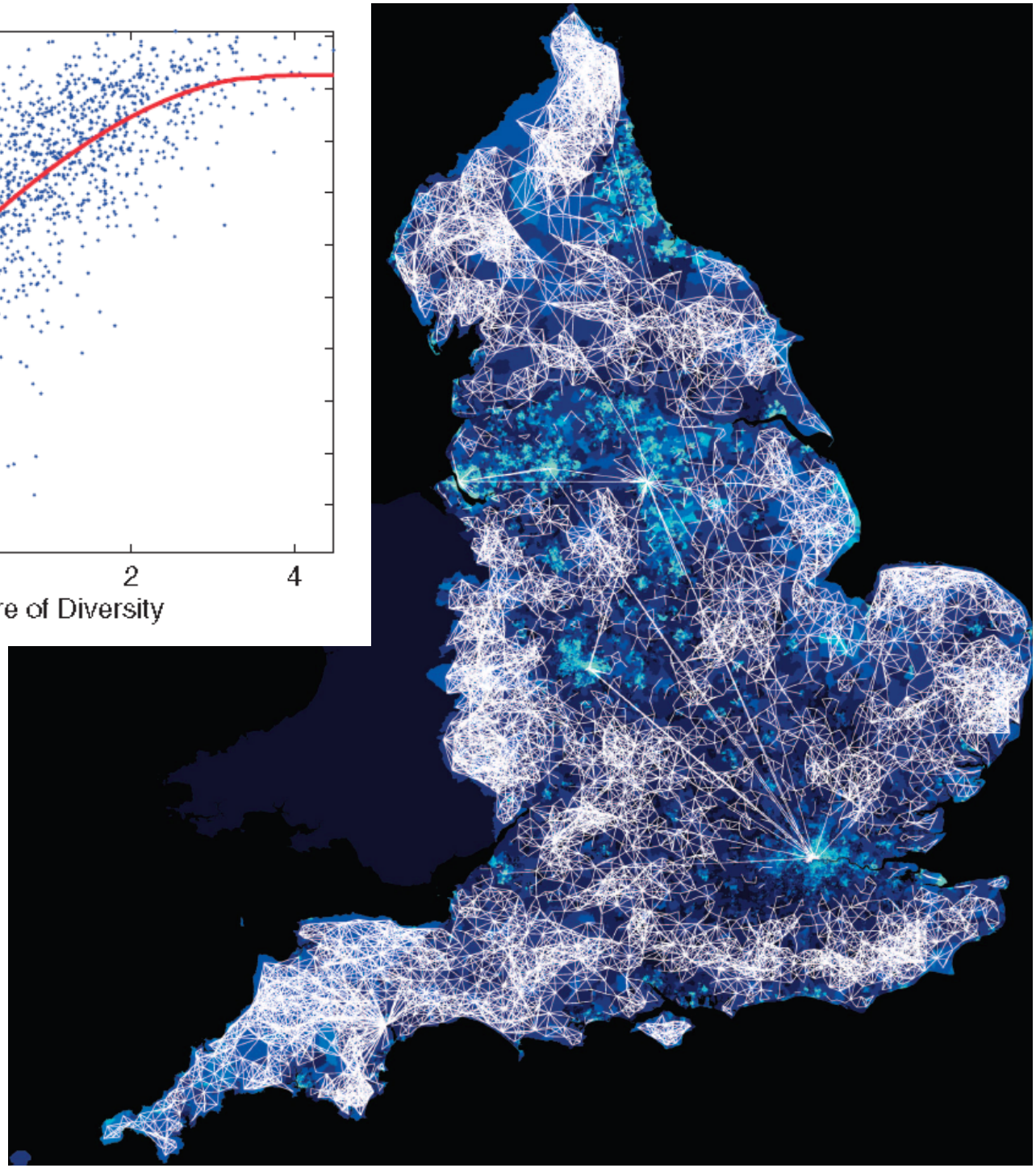


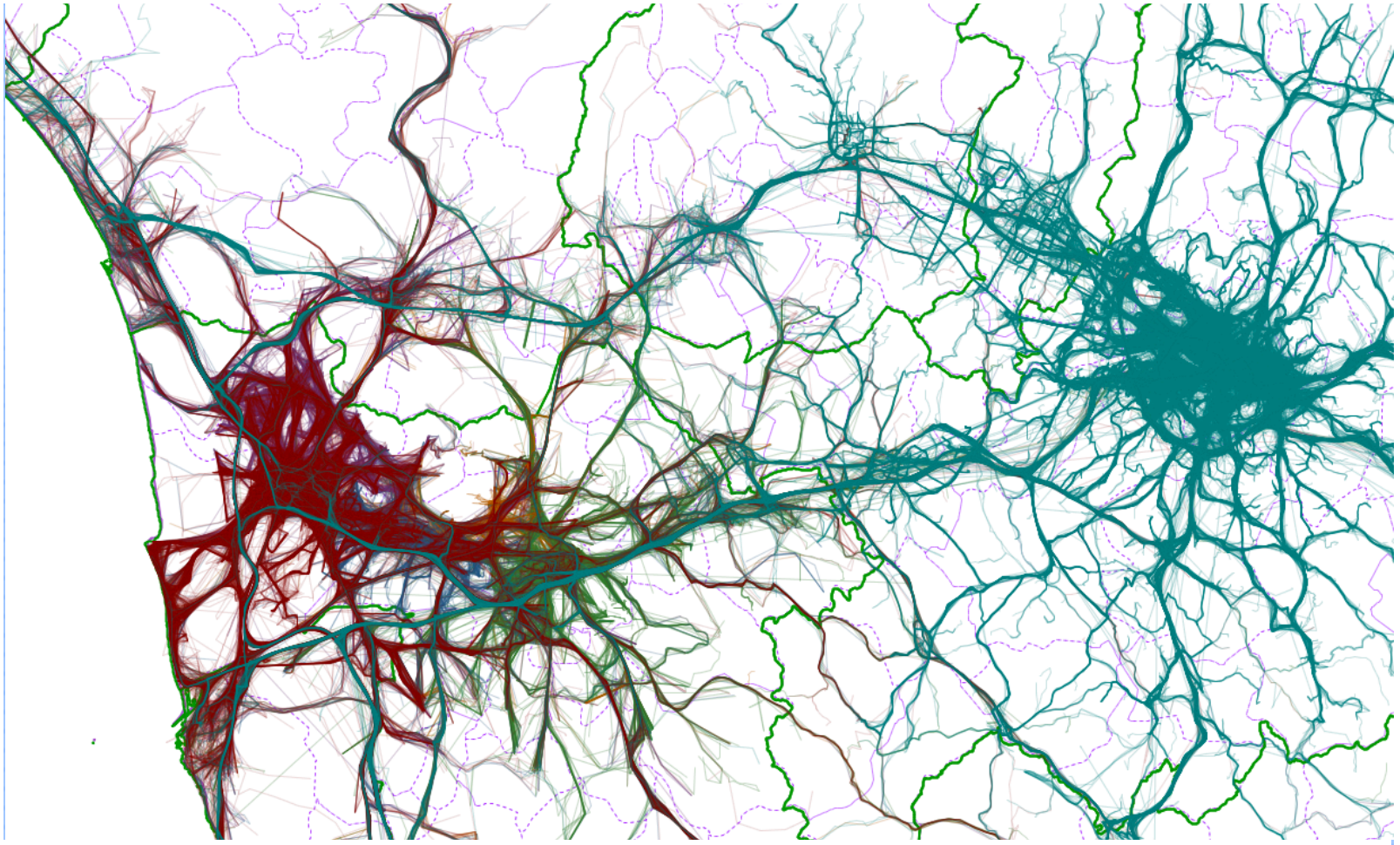


Spread of Happiness

James H. Fowler, Nicholas A. Christakis.
Dynamic Spread of Happiness in a Large Social Network:
Longitudinal Analysis Over 20 Years in the Framingham Heart Study
British Medical Journal 337 (4 December 2008)

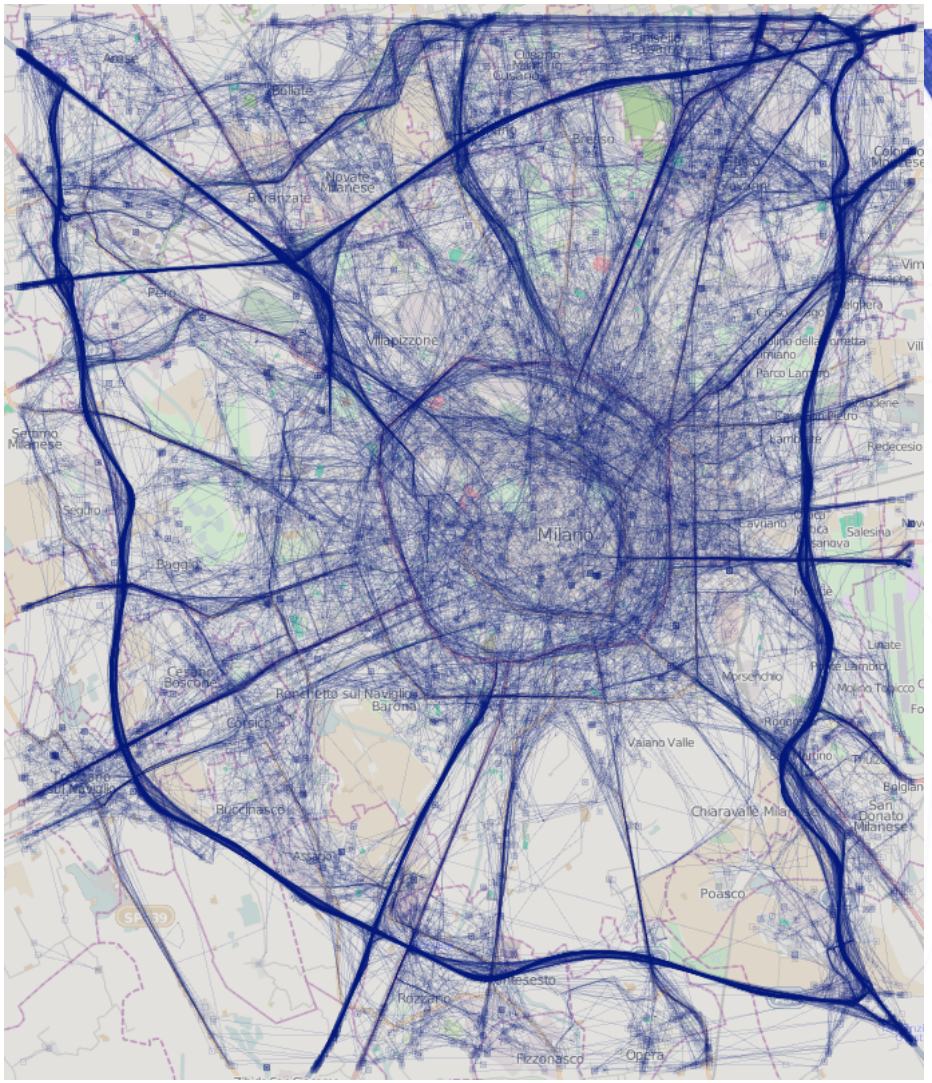

Socio–Economic Percentile
100
90
80
70
60
50
40
30
20
10
–4
–2
0
2
4
Composite Measure of Diversity

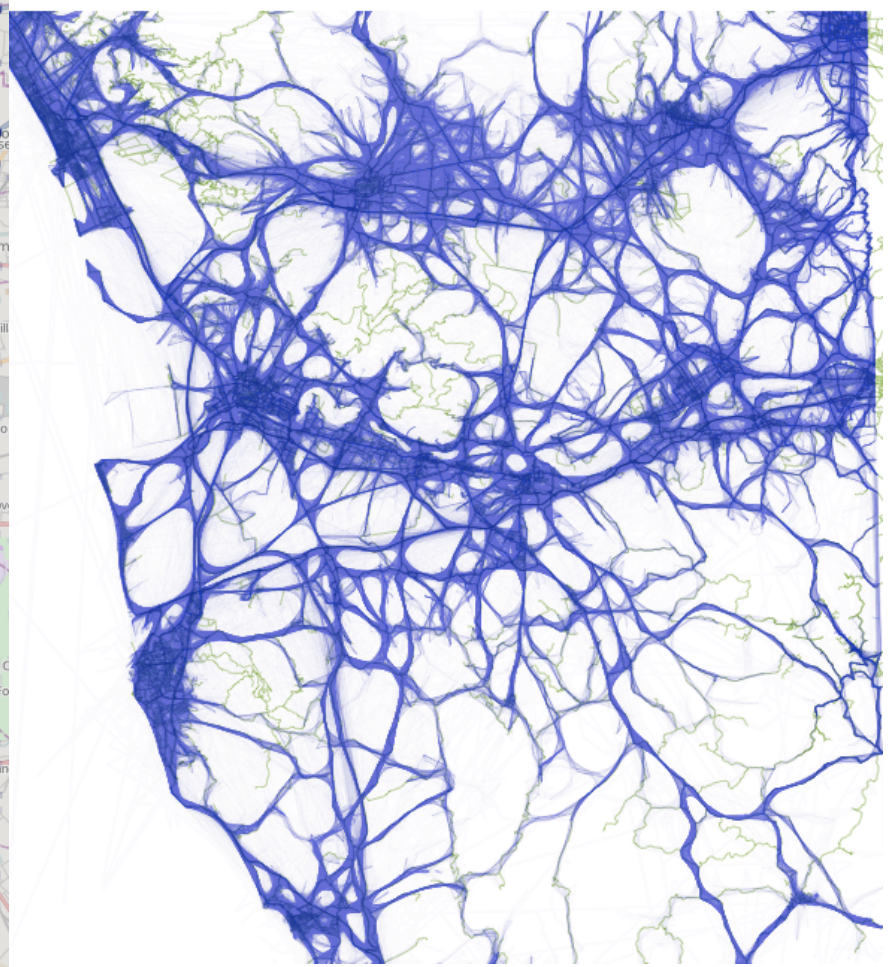




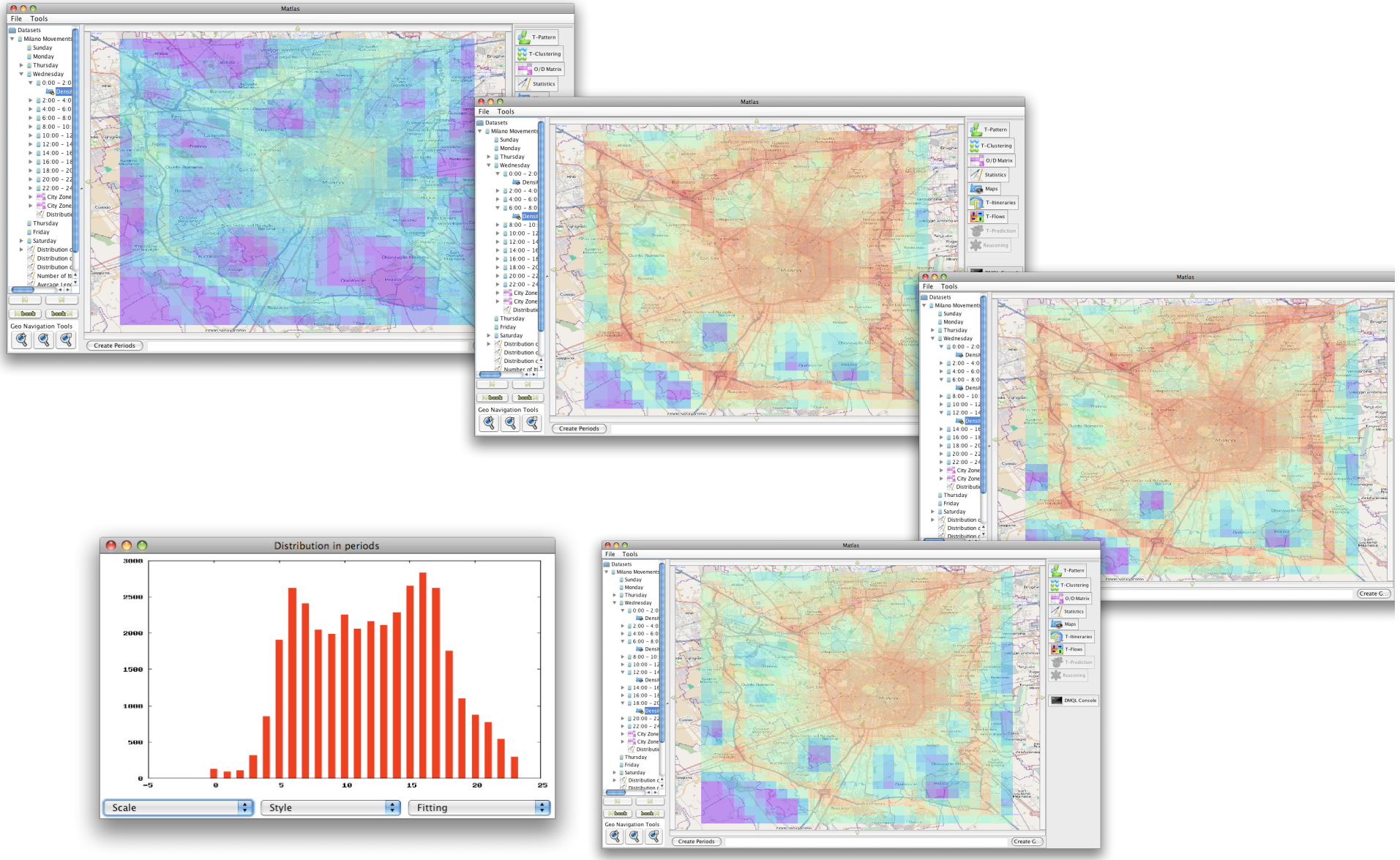
Mobility Analytics for Science of the Cities

Big Data for Smart Cities

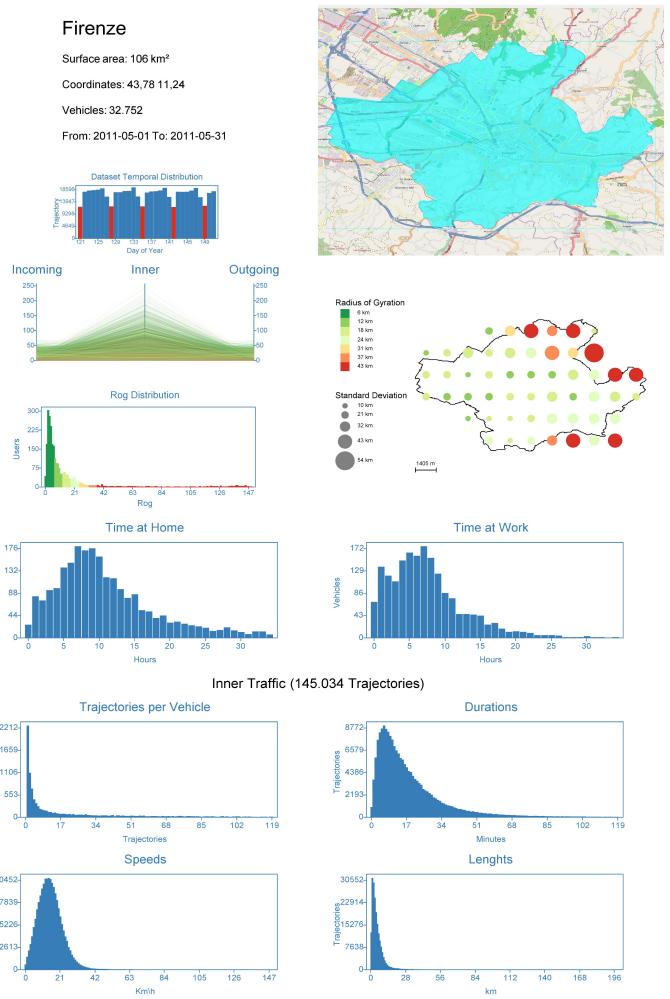


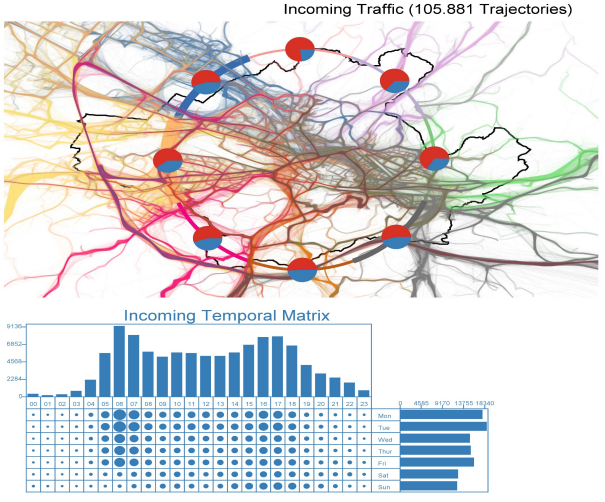


How do people move during the day?

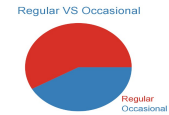


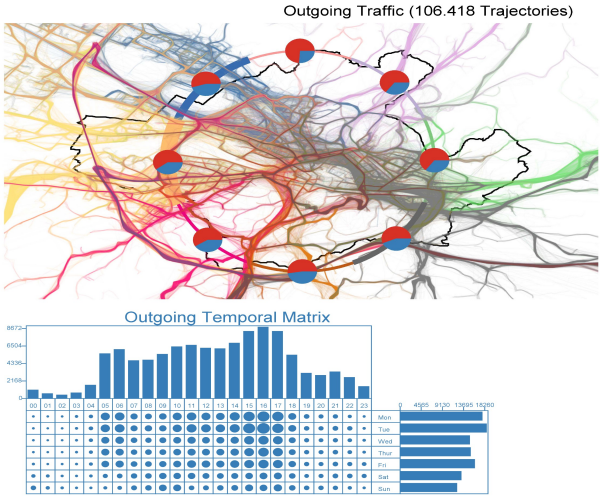
Big Data for Urban Mobility Atlas



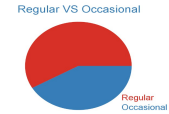


	City	Traj	Perc
NORD 13%	Sesto Fiorenti..	9.123	65%
	Calenzano	1.434	44%
	Vaglia	758	81%
	Campi Bisenzio	436	6%
	Borgo San Lore..	380	48%
OVEST 50%	Scandicci	13.447	98%
	Campi Bisenzio	6.058	93%
	Prato	6.048	94%
	Sesto Fiorenti..	4.824	34%
	Lastra a Signa	2.342	99%
SUD 16%	Impruneta	3.863	87%
	San Casciano i..	1.838	75%
	Figline Valdar..	1.190	81%
	Greve in Chian..	868	36%
	Tavarnelle Val..	744	93%
EST 19%	Bagno a Ripoli	7.314	93%
	Fiesole	3.970	95%
	Pontassieve	2.797	97%
	Greve in Chian..	1.516	63%
	Rignano sull A..	774	92%

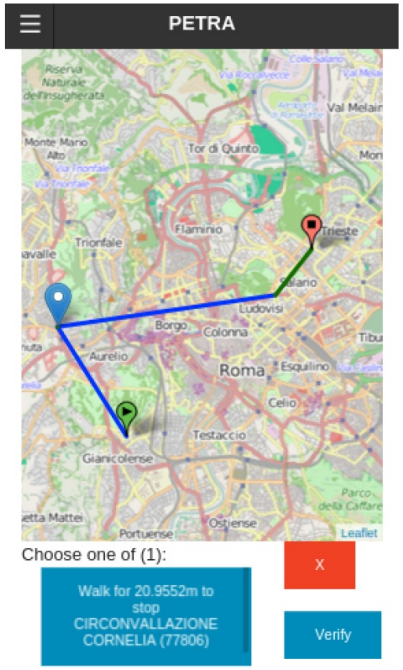


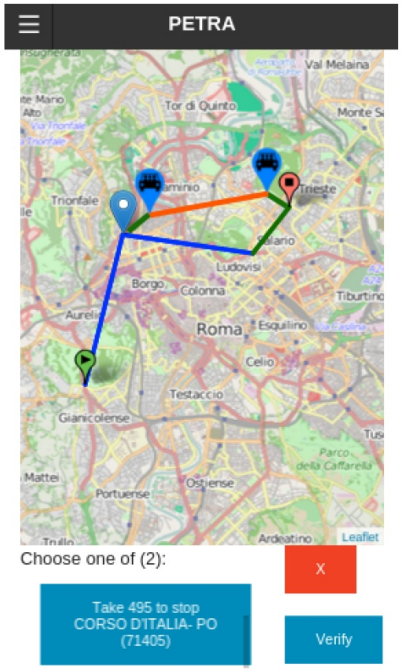


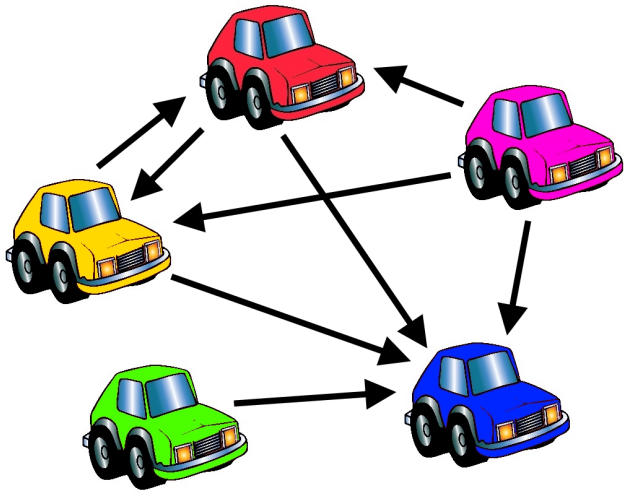
	City	Traj	Perc
NORD 12%	Sesto Fiorenti..	8.358	60%
	Calenzano	1.235	36%
	Vaglia	845	87%
	Campi Bisenzio	487	7%
	Borgo San Lore..	456	54%
OVEST 52%	Scandicci	13.439	98%
	Prato	6.168	95%
	Campi Bisenzio	5.846	92%
	Sesto Fiorenti..	5.521	39%
	Lastra a Signa	2.423	98%
SUD 14%	Impruneta	3.965	85%
	San Casciano i..	1.801	72%
	Figline Valdar..	1.155	77%
	Tavarnelle Val..	742	92%
	Greve in Chian..	739	29%
EST 20%	Bagno a Ripoli	7.701	96%
	Fiesole	3.982	94%
	Pontassieve	2.806	98%
	Greve in Chian..	1.670	67%
	Rignano sull A..	818	96%



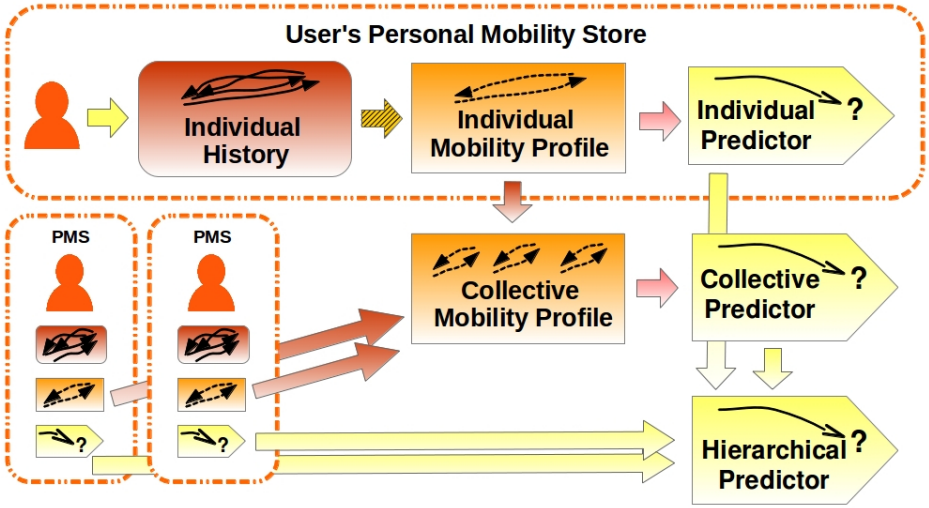
Personal mobility assistant



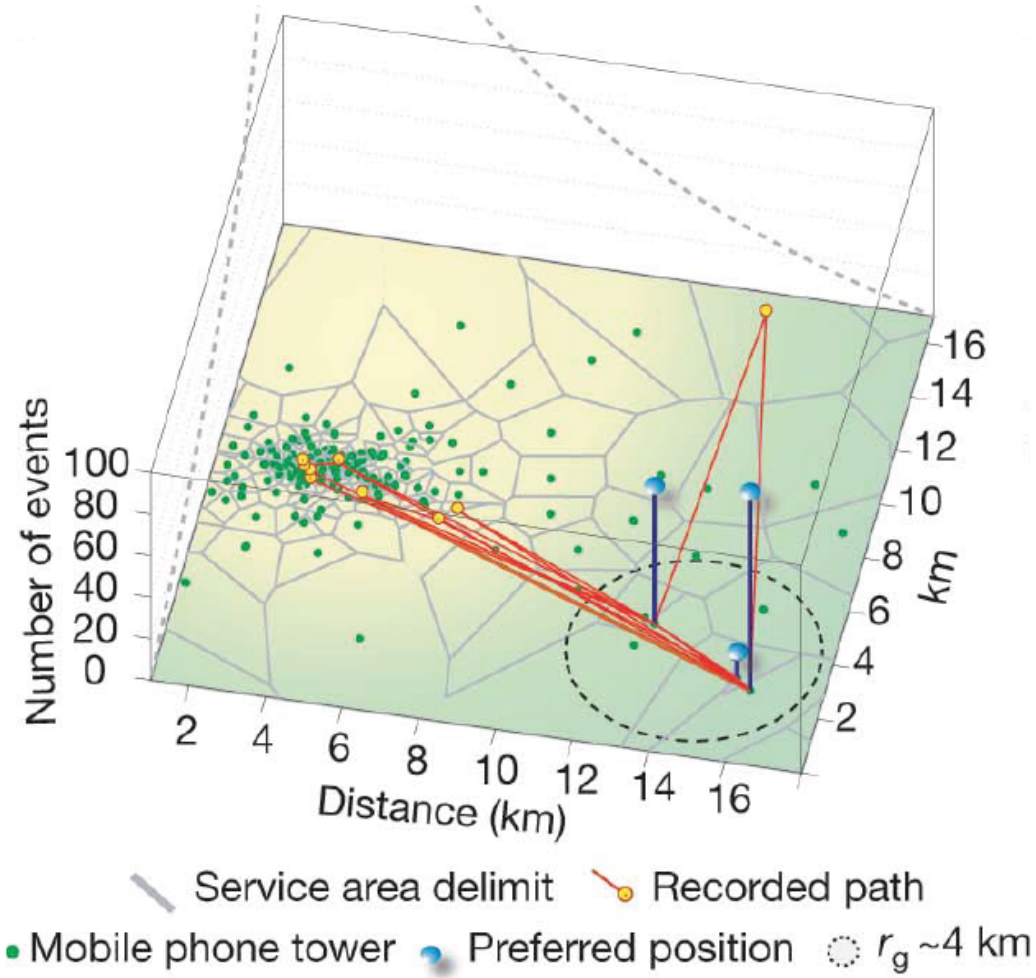




Carpooling Network



Focus on country-wide CDR data

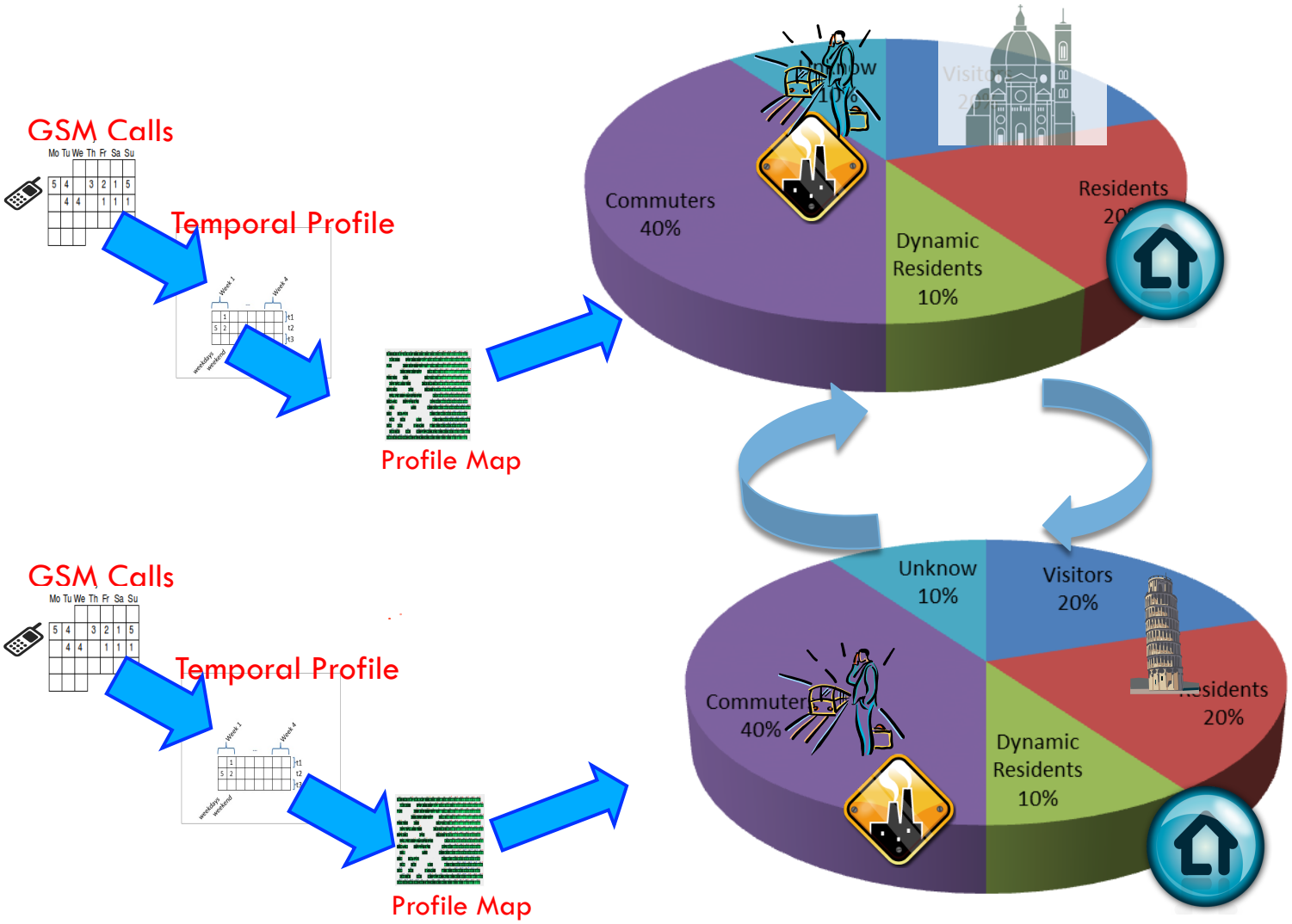


when you call

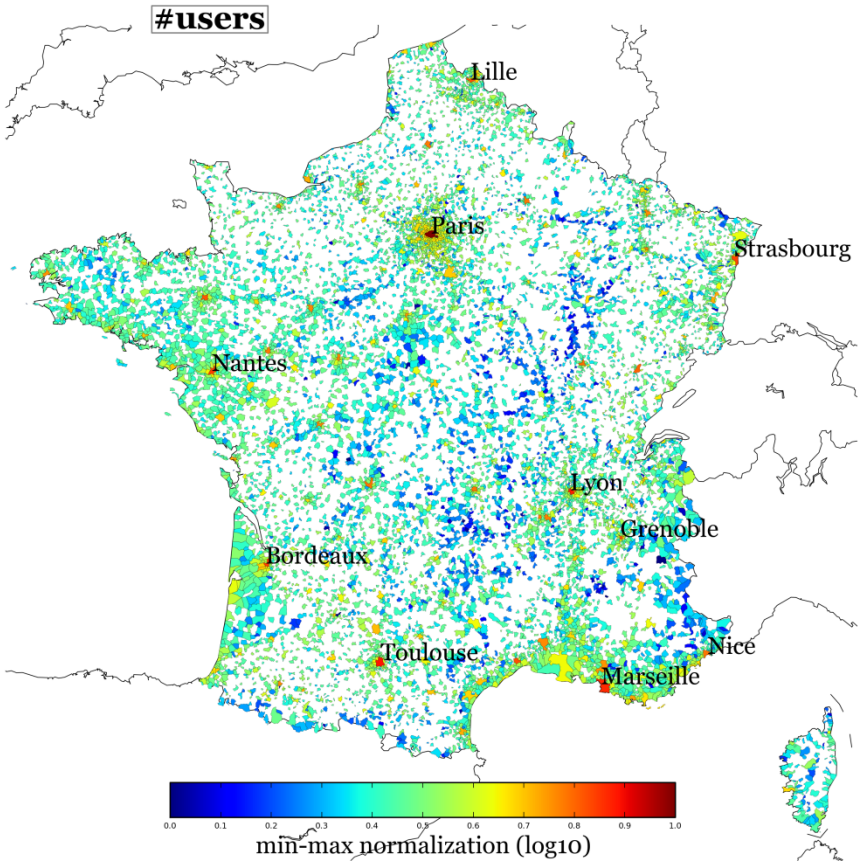
where you call

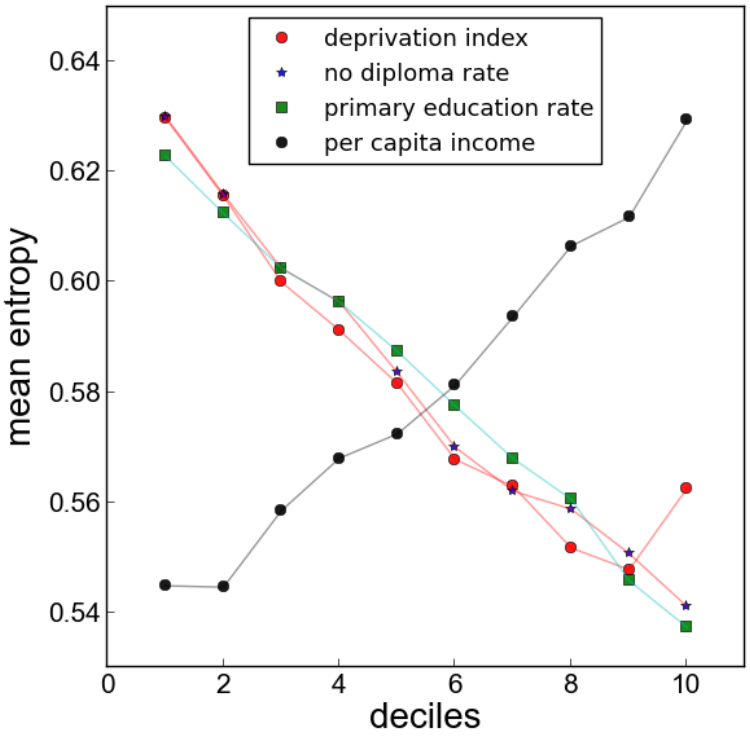
who you call

Big Data in Official Statistics: Persons & Places

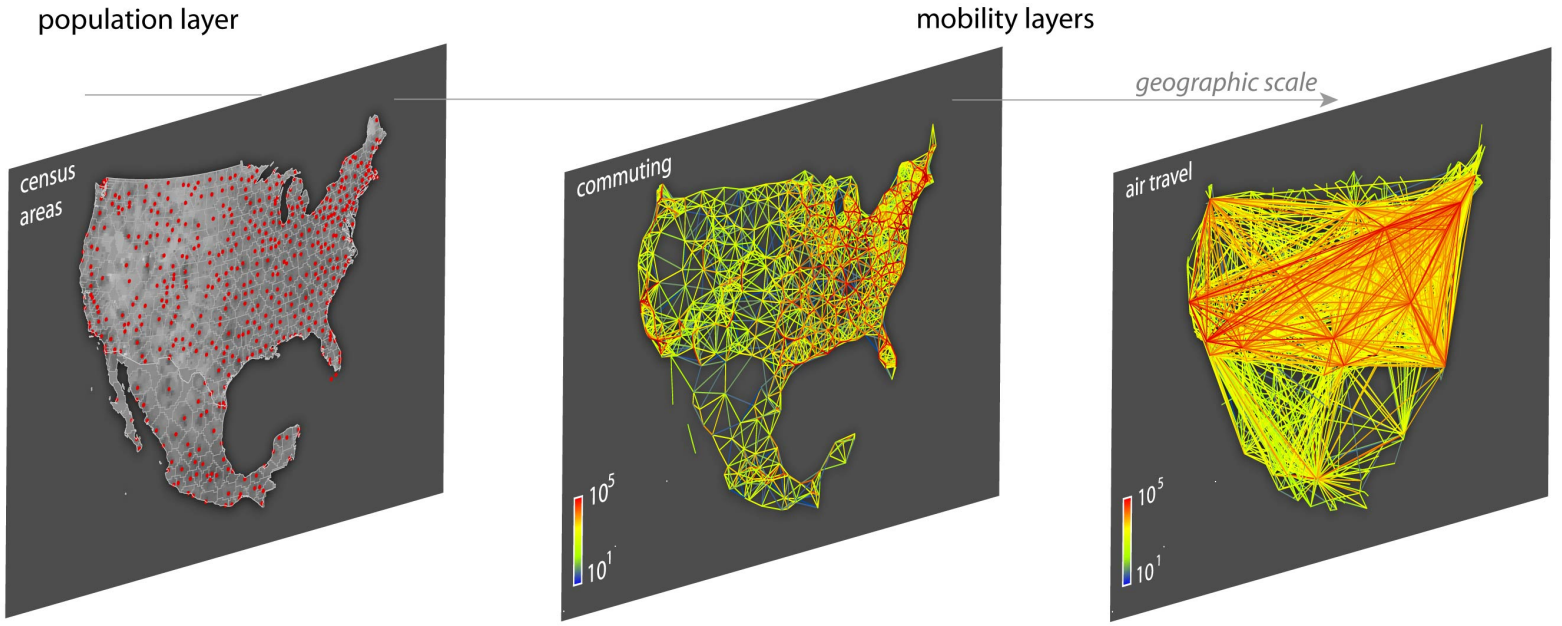


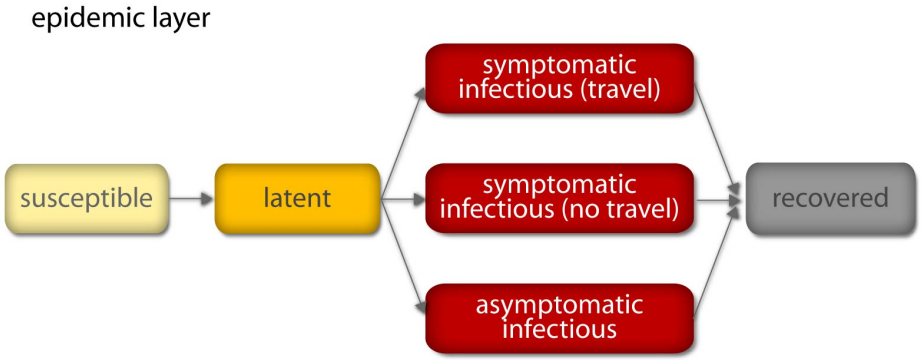
Big Data: mobility diversity and wellbeing





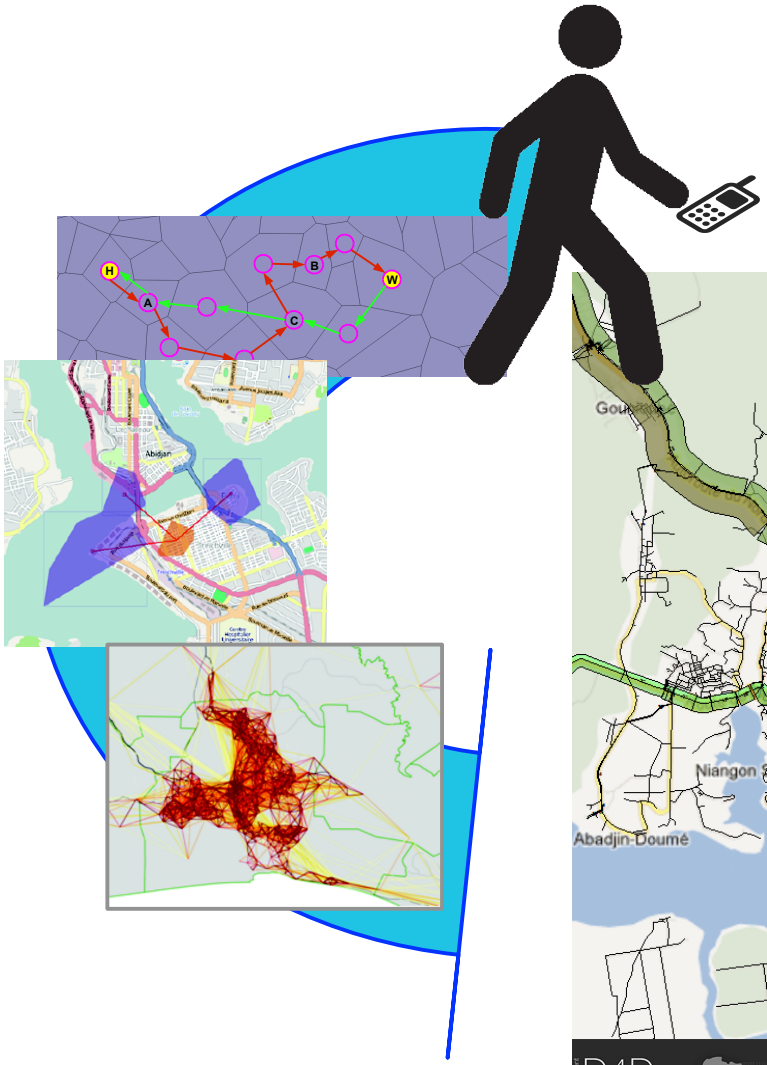
Big Data and epidemics

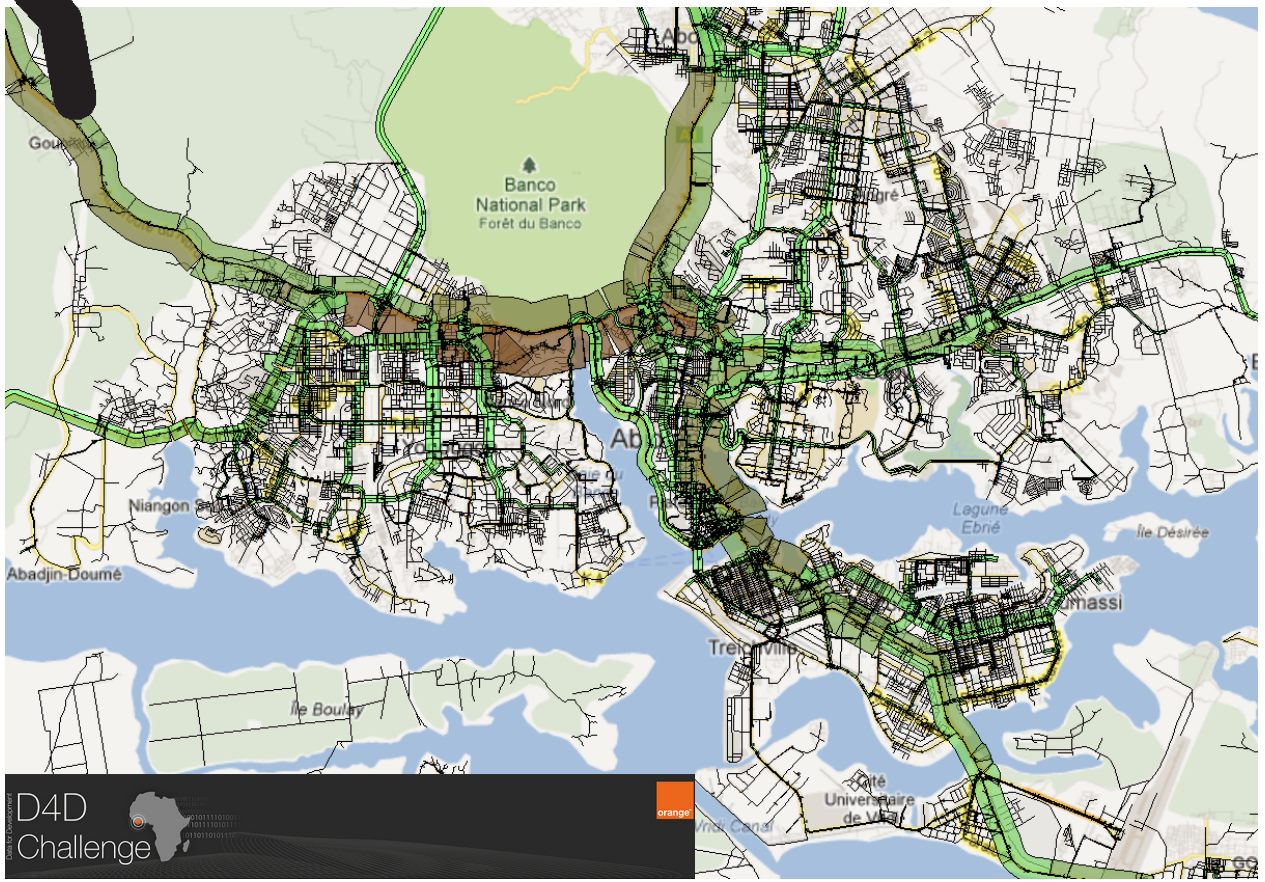


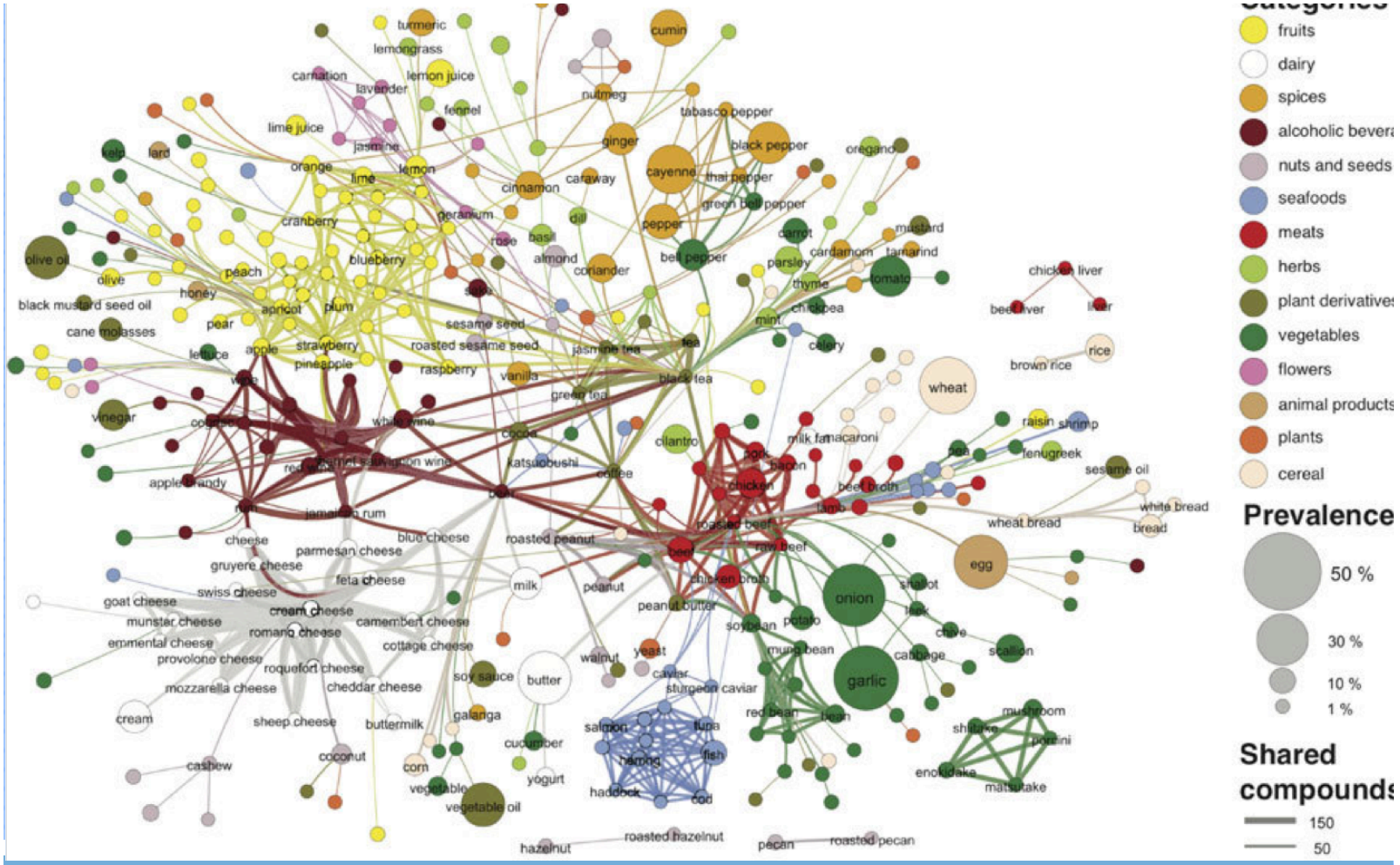


Parameter	Value	Description
β	from R_0	transmission probability
ε^{-1}	1.9 [1.1-2.5] d	average latency period
μ^{-1}	3 [3-5] d	average infectious period
p_t	50%	probability of traveling for infectious individuals
p_a	33%	probability of being asymptomatic
r_β	50%	relative infectiousness of asymptomatic infectious individuals

Big Data for Developing Countries

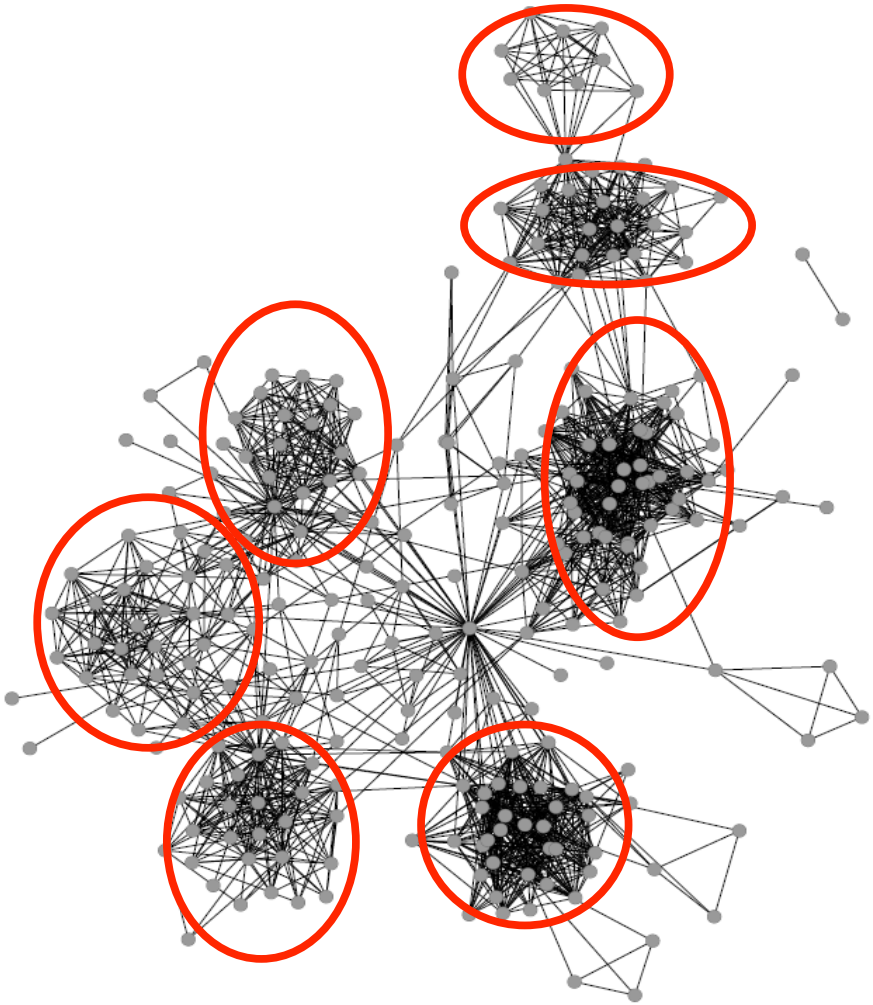


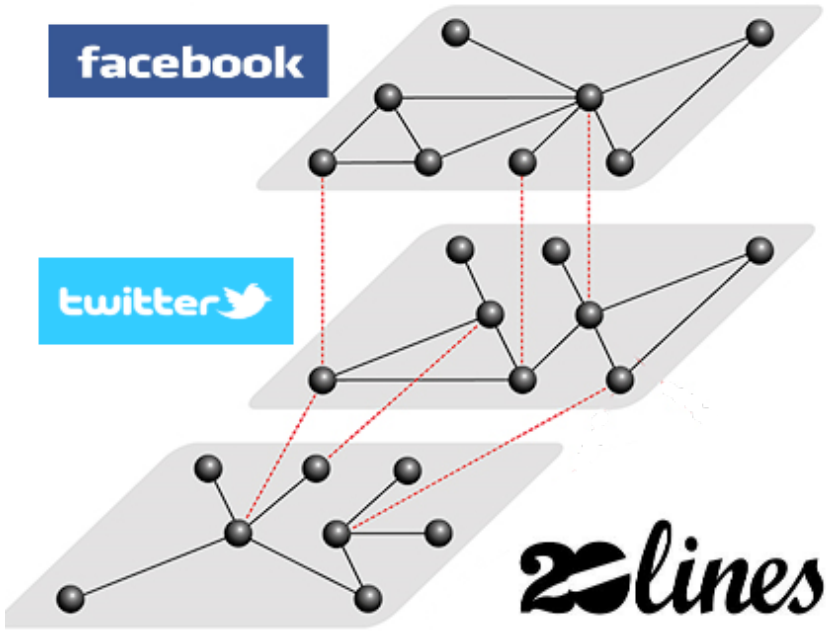




Social Network Analysis

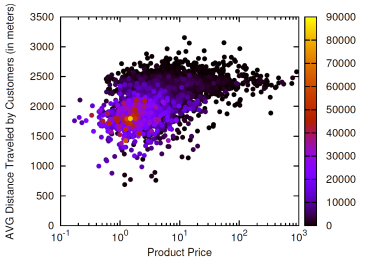
Community Discovery, Evolution, Diffusion, Multidimensionality,…



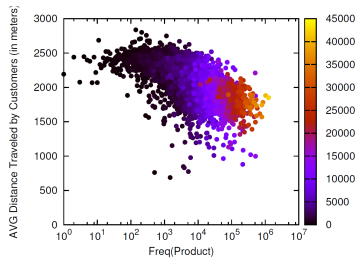


Michele Coscia, Giulio Rossetti, Fosca Giannotti, Dino Pedreschi: **Uncovering Hierarchical and Overlapping Communities with a Local-First Approach.** TKDD 9(1): 6 (2014)

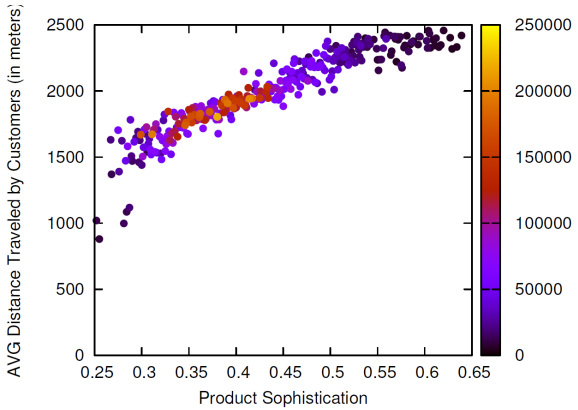
Retail Market as Complex system



$R^2 = 17.25\%$



$R^2 = 32.38\%$

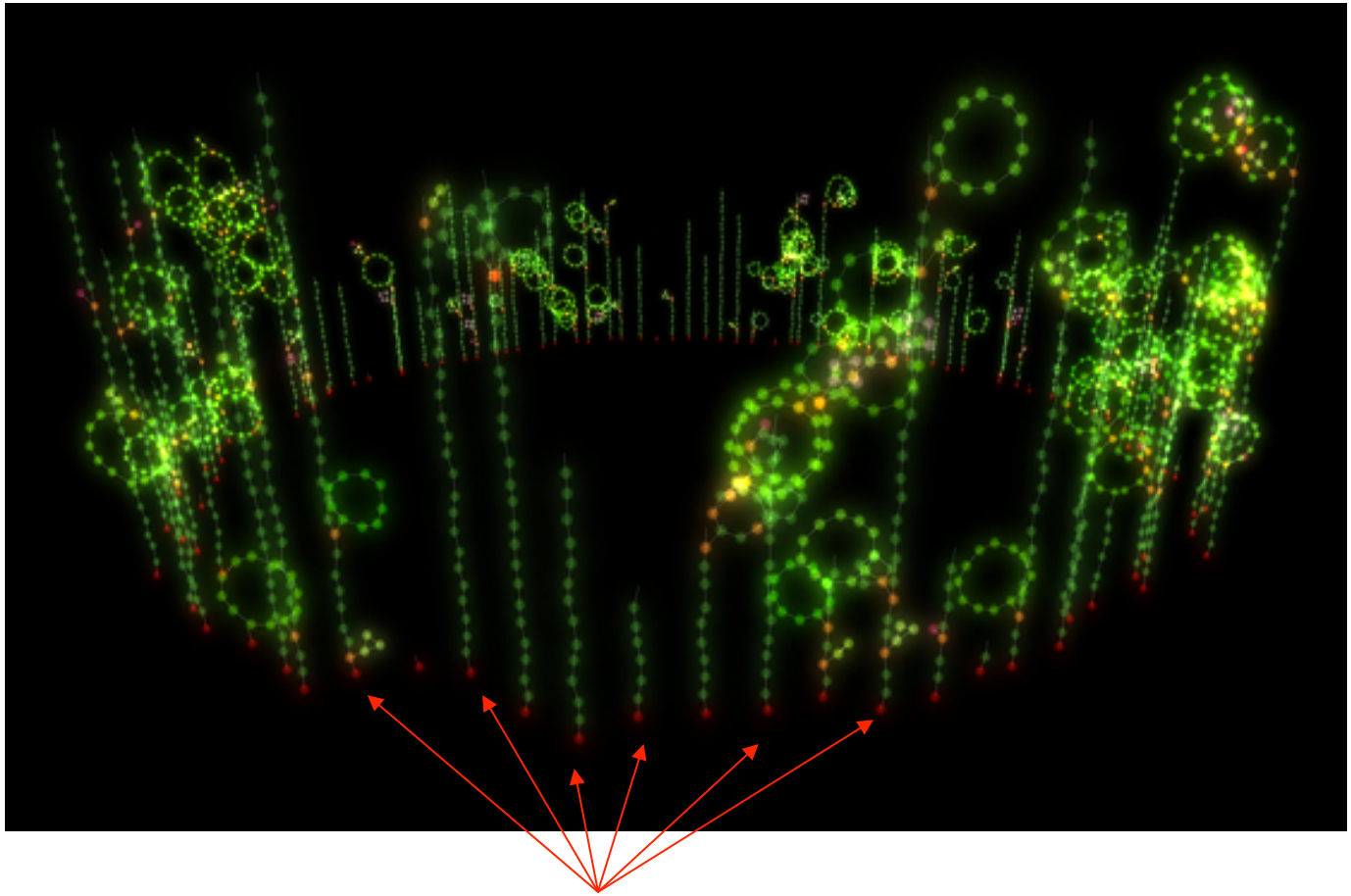


$R^2 = 85.72\%$

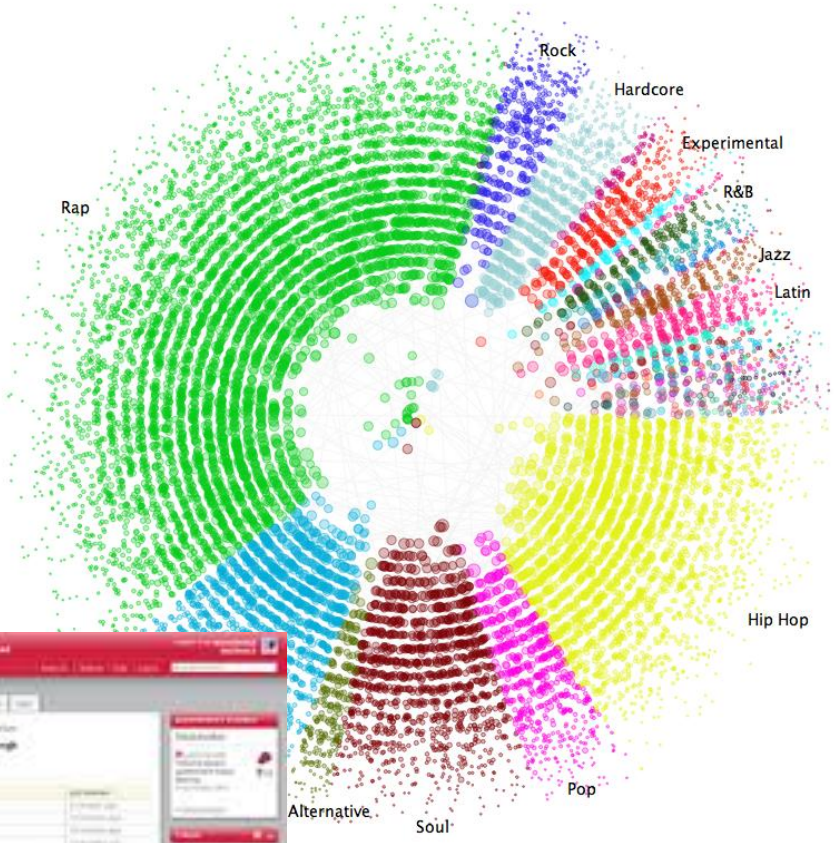
Pennacchioli, D., Coscia, M., Rinzivillo, S., Giannotti, F. and Pedreschi, D., **The retail market as a complex system**. In EPJ Data Science, 2014.

THE PARADOX OF SOCIAL INFLUENCE

Social Influence: Leaders

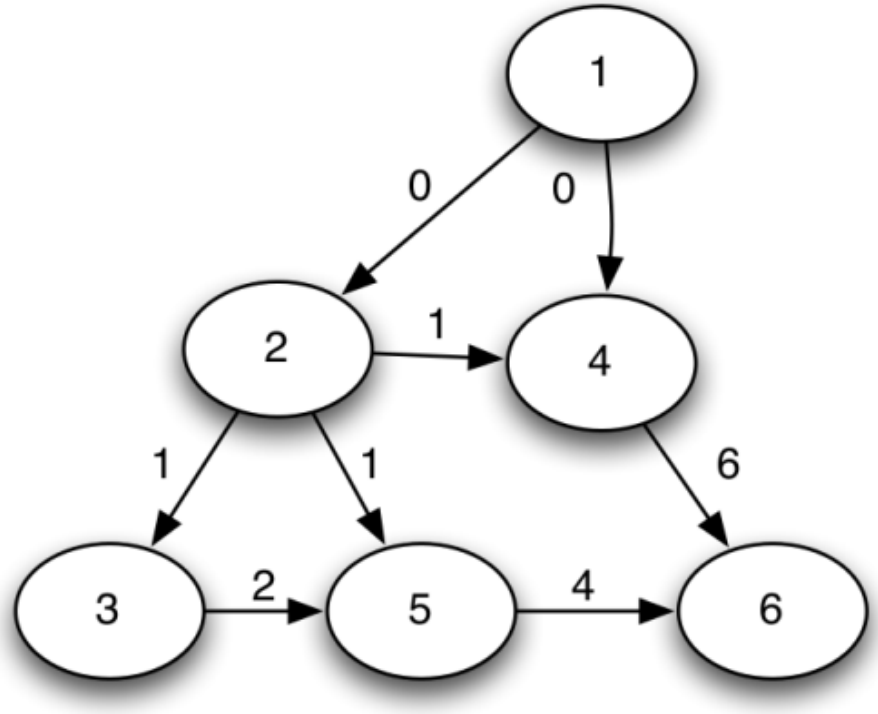


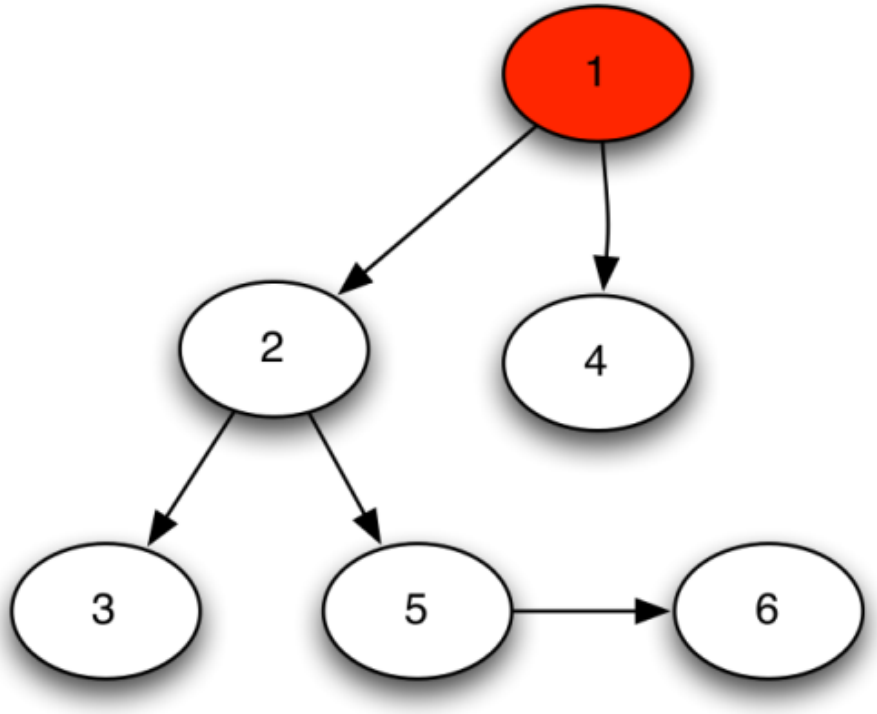
Ask to LAST.FM



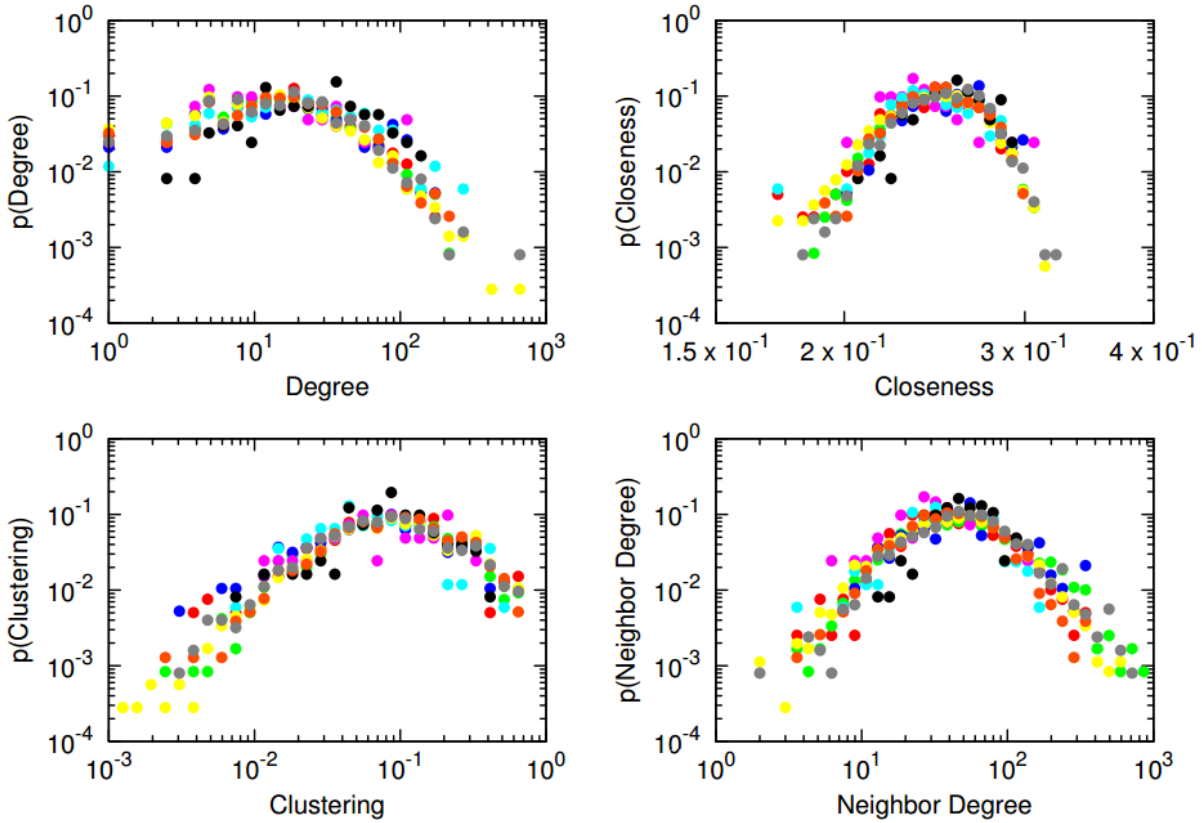
80.000utenti, 4000.000 connessioni

Leader finding





From BigData...true influenzer are not leaders

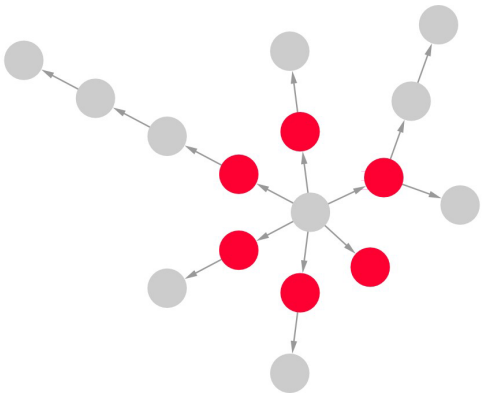


… abbiamo scoperto che i leader teorici, quelli che avrebbero in teoria il potere di influenzare la rete sociale, non hanno una grande influenza pratica sulla rete.

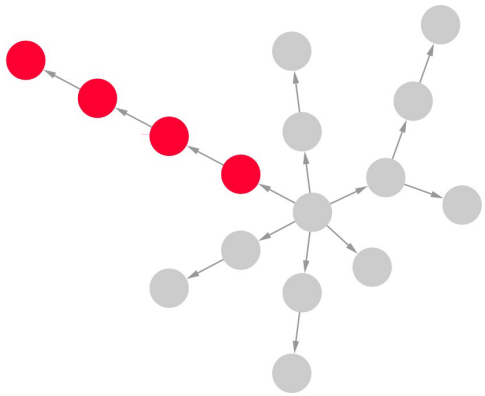
What is Social Prominence?

- It has been observed that a small set of users in a Social Network is able to anticipate (or influence) the behavior of the entire network
- We detected 3 possible scenarios:

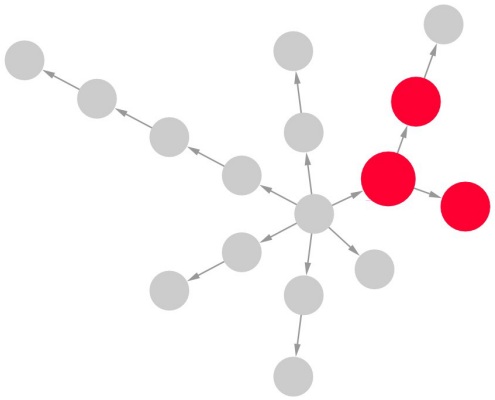
width



length



strength



No limits to creativity

If data are available, then any phenomenon becomes measurable, quantifiable and possibly predictable … including human behaviour

Big Data: the way of Success

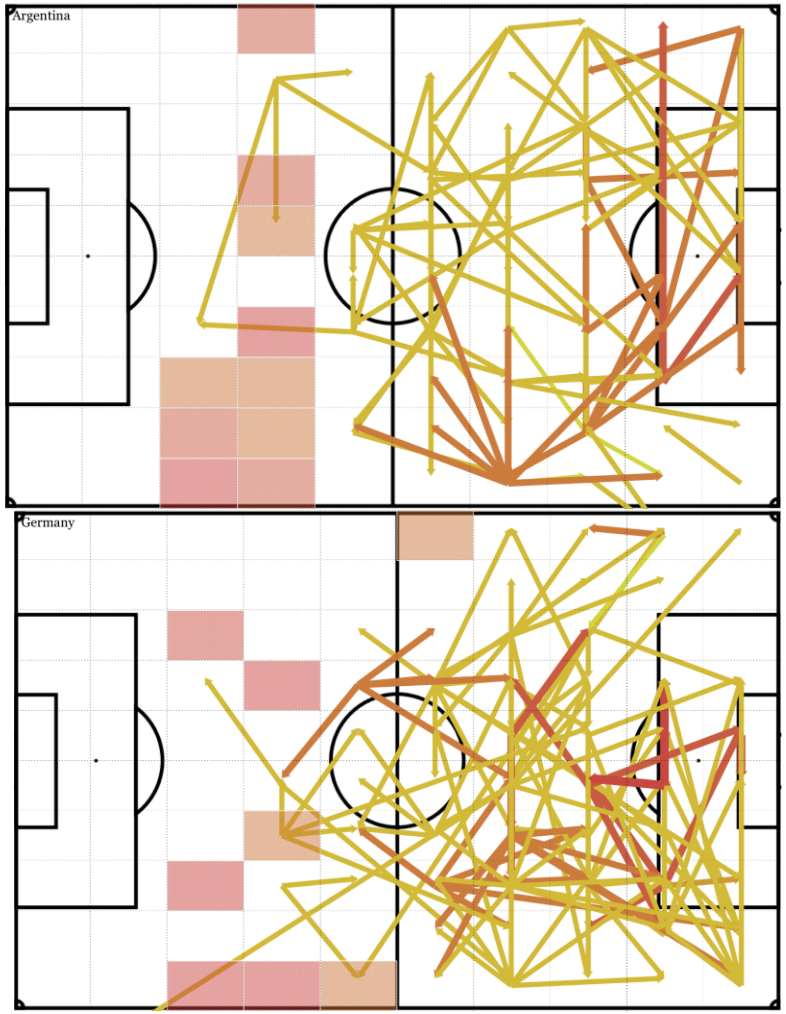
The patterns of success in cycling:

- data from Strava.com
- How you train is fundamental
- A confirmation of the "overcompensation" theory

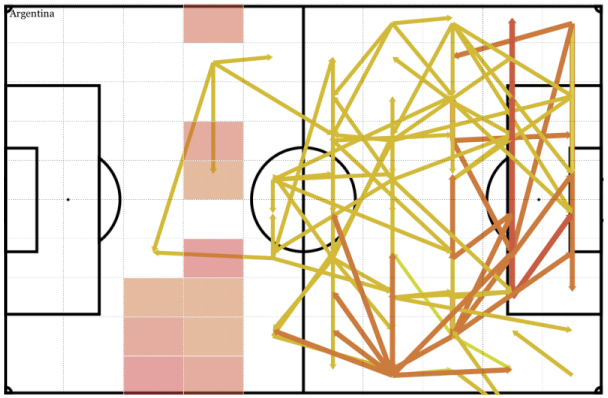
The patterns of success in Sports

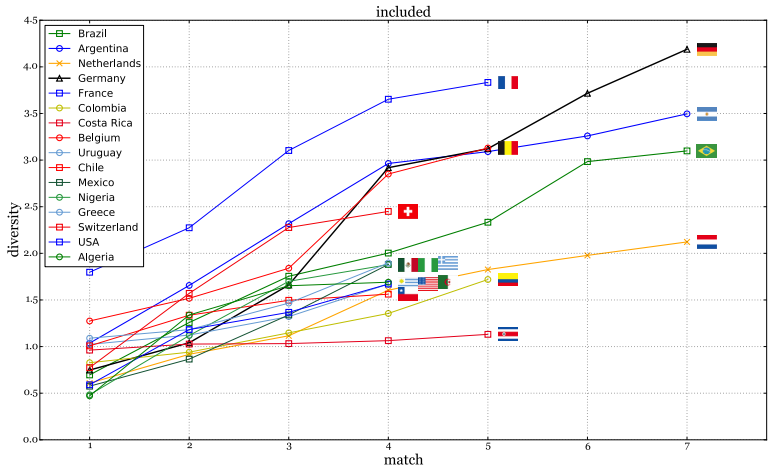
“Football is a simple game: 22 men chase a ball for 90 minutes and at the end, the Germans always win”

-- Gary Lieneker (after Italy 1990 Final)



Big Data: the way of Success





According to our models the final will be Germany-Argentina. Are our data-driven models correct ? Let's see what happens!!! #WorldCup2014

9:00 PM - 8 Lug 2014 Pisa, Italia

1 RETWEET 2 FAVORITES

The patterns of success in football:

- detailed data on every match (trajectories, passes, goals, …)
- a network approach to study the strategy of teams
- a data mining approach to study the performance of players

Data from Opta:

All events during the match

```
...
<tackle,15.4,41.1,112>
<pass,25.0,67.1,113>
<pass,65.0,87.1,115>
<assist,82.1,35.8,120>
<goal attempt,82.1,35.8,121>
.....
```

Big Data Analytics & Social Mining

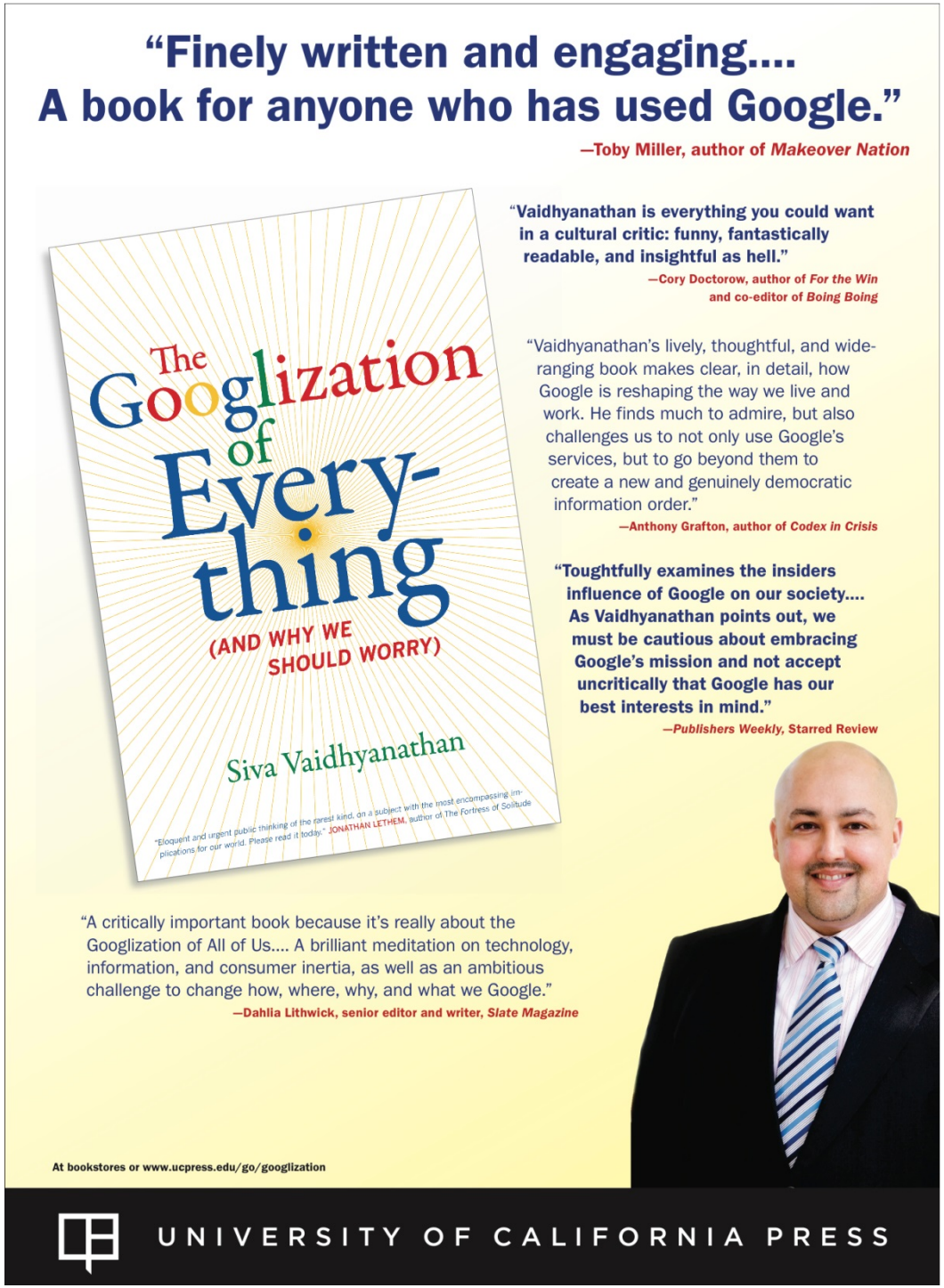
Who's got the benefits of big data so far?

- A few **latifundists of data**
- **GAFA**

  - Profiling for behavioral advertising and target marketing

- **NSA**

  - Profiling discovering potential threats to homeland security
  - Mass surveillance



We are not Google's customers,
we are its products.

We – our fancies, fetishes, predilections, and preferences – are what Google sells to advertisers.

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

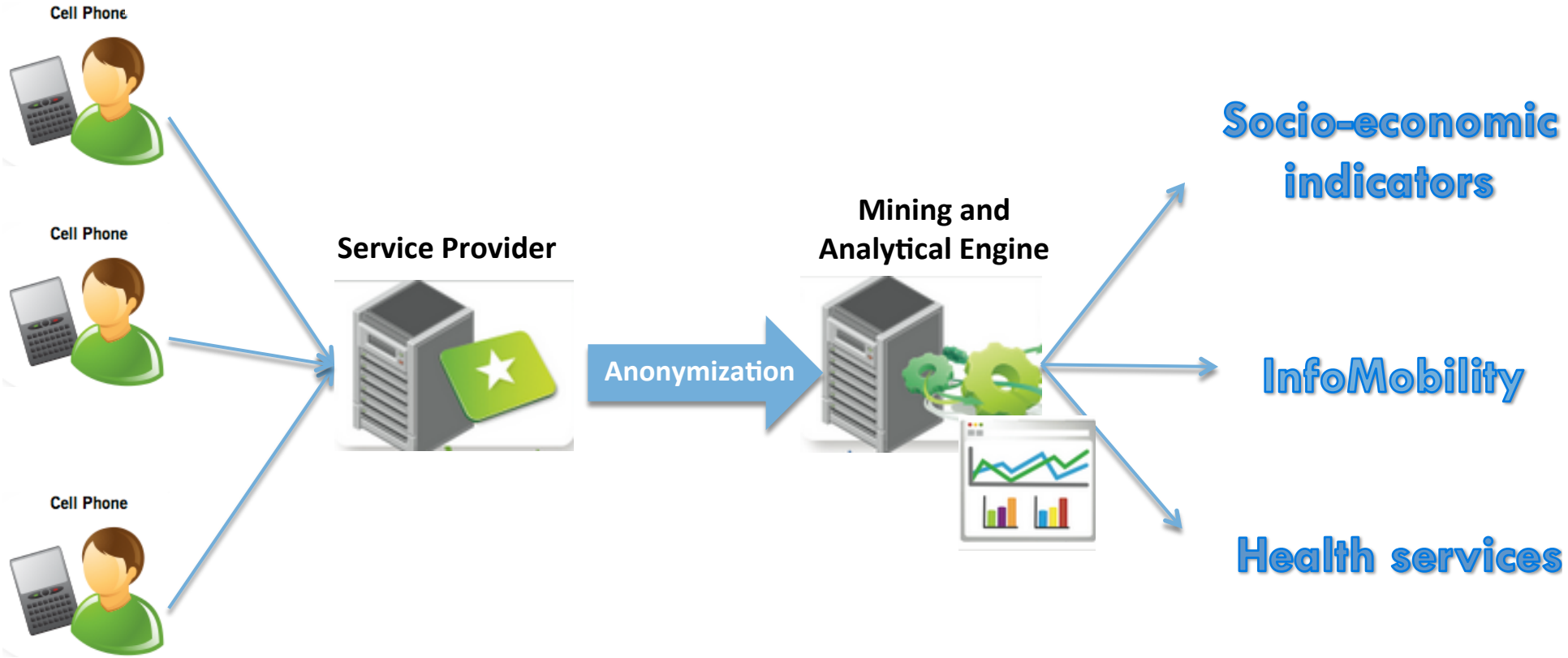
EPJ Data Science
a SpringerOpen Journal

EPJ.org

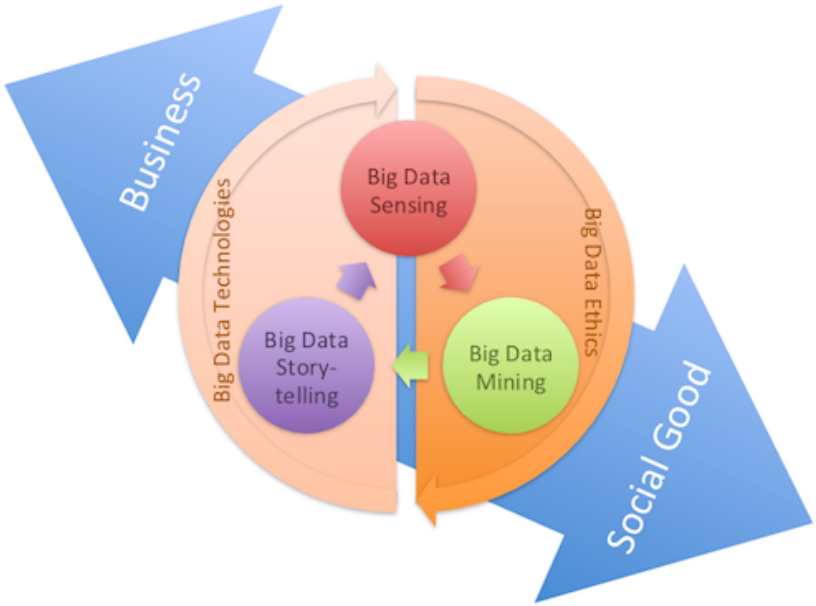
REGULAR ARTICLE Open Access

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]

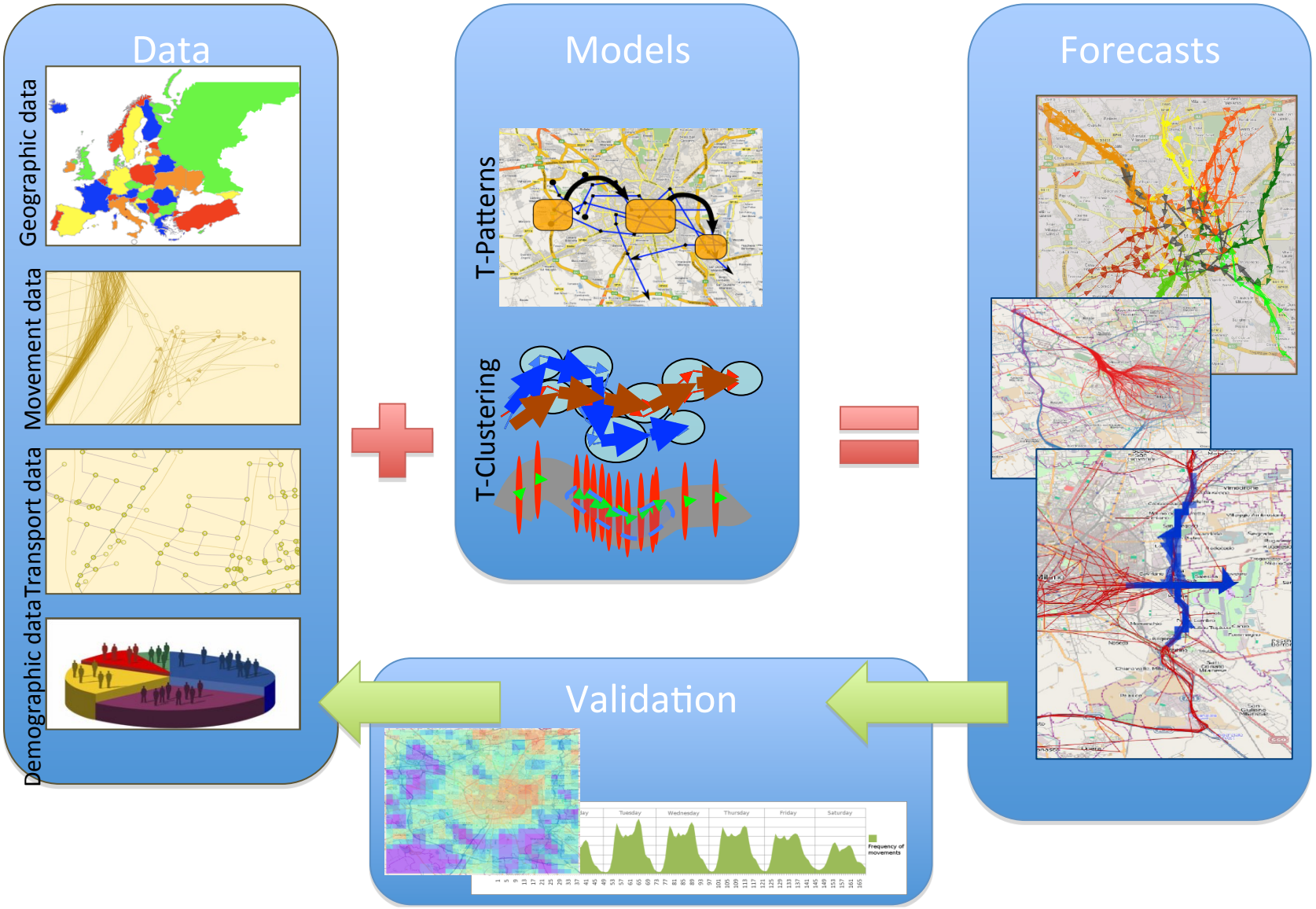


The modern data scientist!!!

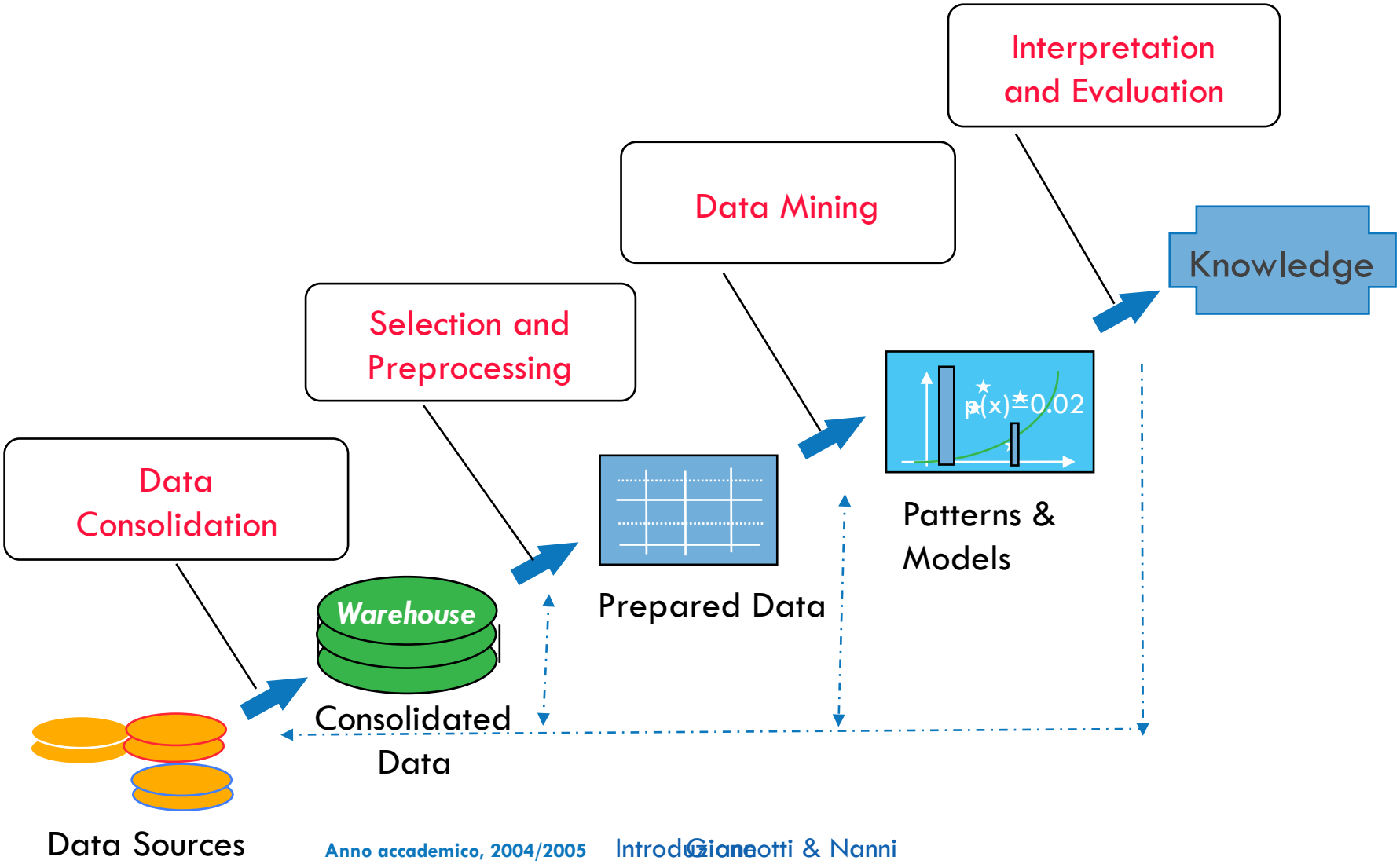


How to develop a big data analytics project

From DATA to KNOWLEDGE

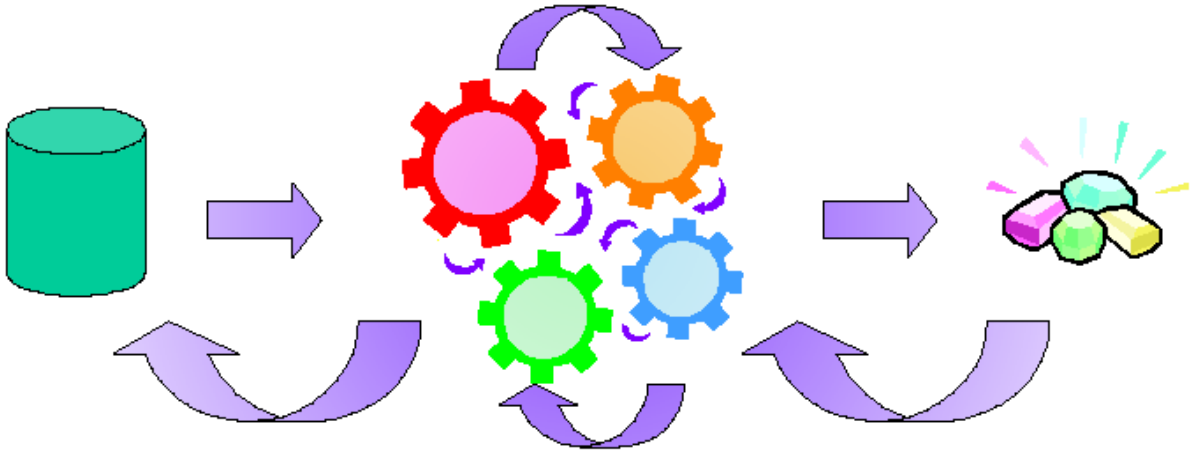


The KDD process

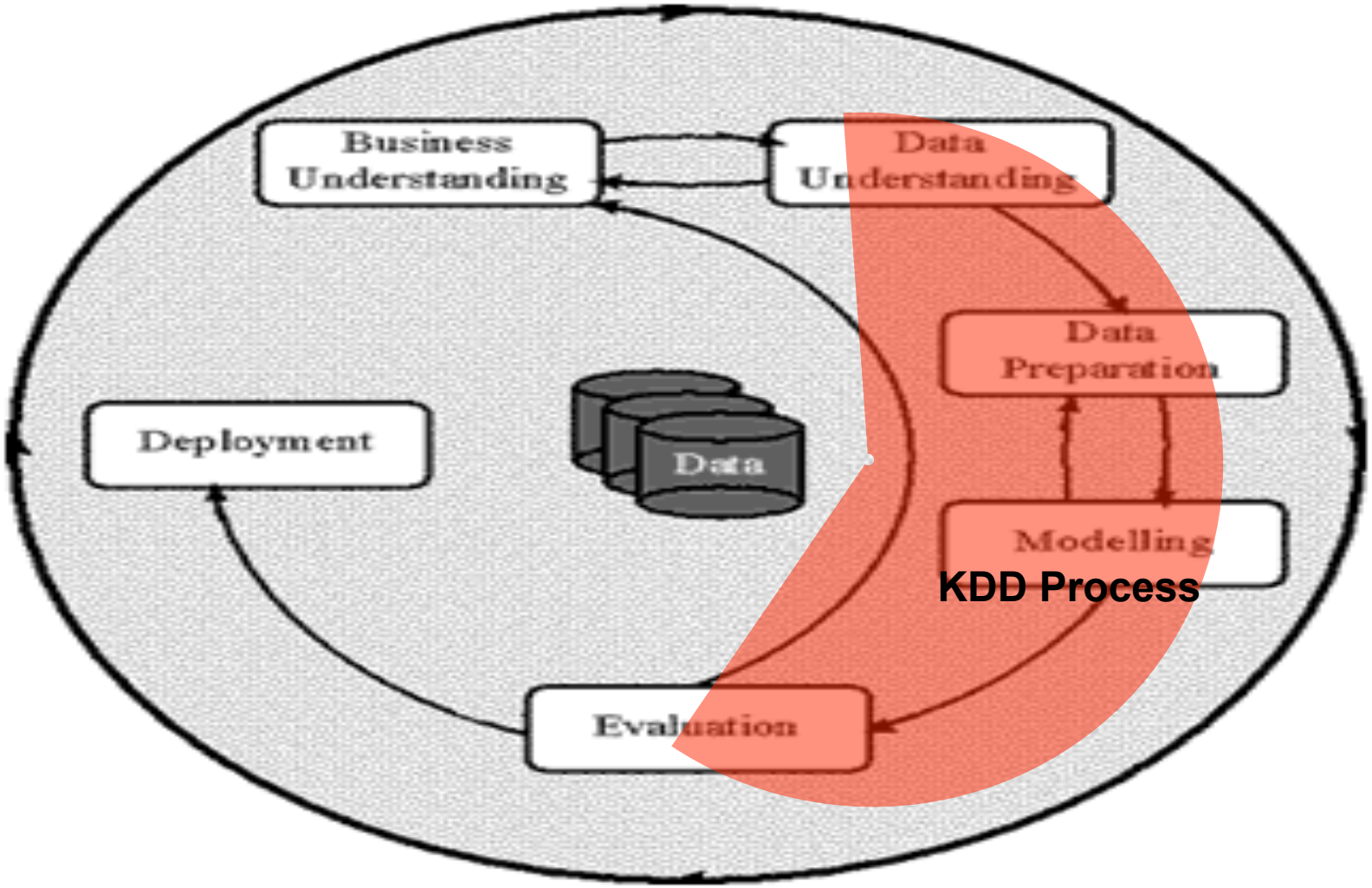


The KDD Process in Practice

- KDD is an Iterative Process
  - art + engineering rather than science

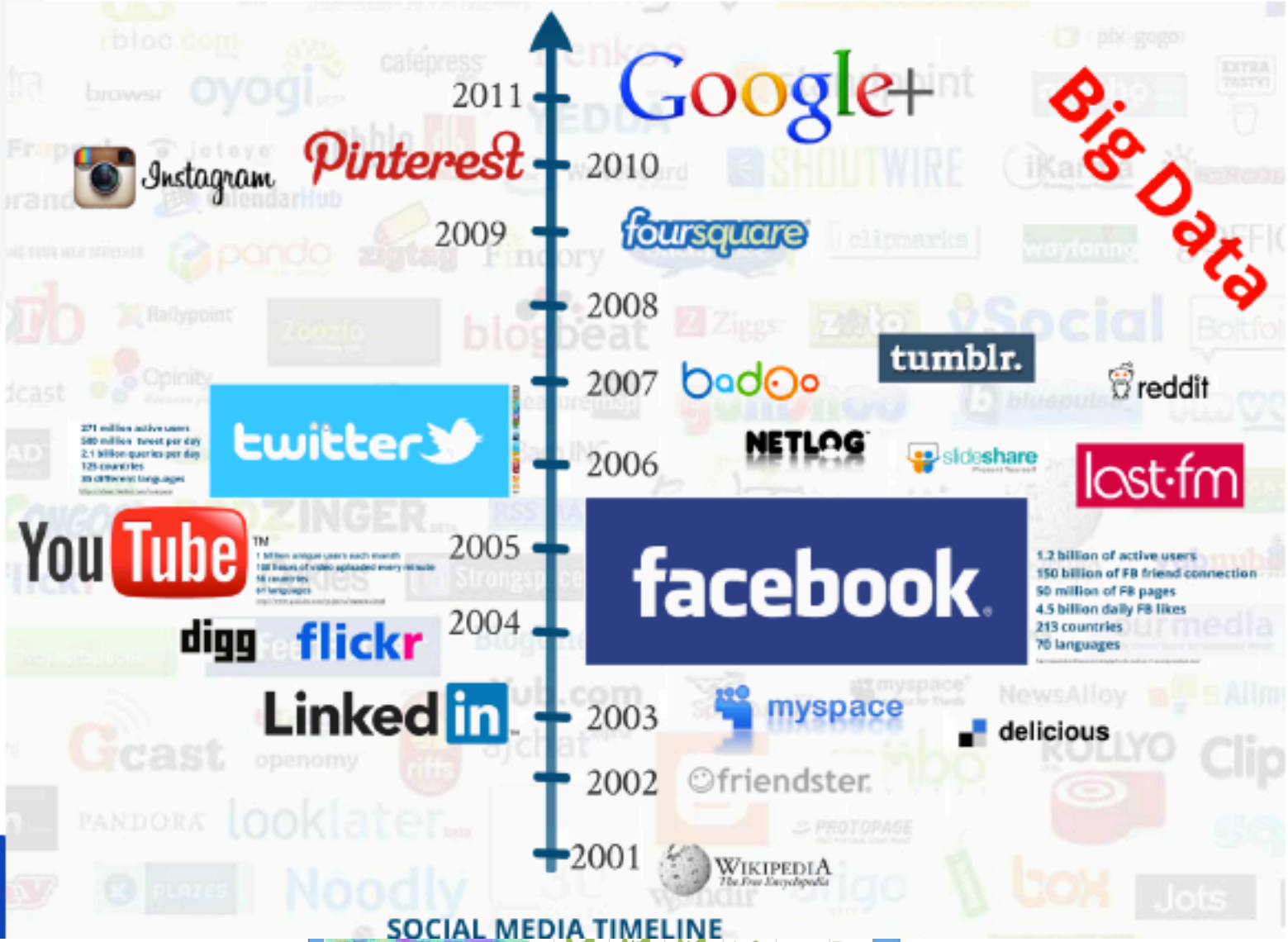


CRISP-DM: The life cicle of a data mining project



Big Data Number

Social Media Timeline



Every minute in Social Media

Data….

- 1,200,000,000,000,000,000,000 bytes

of data

- Facebook - 1,150 million users
- Gmail – 425 million users
- Skype – 300 million users
- Tweeter – 500 million users (200M active)
- WhatsApp – 300+ million users
- Youtube – 1,000 million users (4 B daily views)
- Instagram - 150 million users

Sources:
http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/ September 15, 2013

Data….

- Waze – 50 million users
- Amazon – 209 million users
- Ebay - 120 million users
- Paypal - 132 million users
- Google searches – ~12 billion (monthly, US alone)

Sources:
http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/ September 15, 2013

Big Data and Vs

- **Volume and complexity** of data is increasing. “complexity”: it refers to the context of data (creation, provenance, relations) in which it exists and which must be considered when interpreting or re-using the data.
- **Velocity** with which data is being created and characterised is changing
- **Variety** of data in all respects and the challenges of combining variety
- Veracity related to aspects such as trust in dealing with data, i.e. statistical significance.
- Value
- Privacy

With the datafication comes big data, which is often described using the four Vs:

- Volume
- Velocity
- Variety
- Veracity

Bernad Marr Bigdata: using Smart BigData analytics and metrics
To make better decisions

Volume…

…refers to the vast amounts of data generated every second. We are not talking Terabytes but Zettabytes or Brontobytes. If we take all the data generated in the world between the beginning of time and 2008, the same amount of data will soon be generated every minute. New big data tools use distributed systems so that we can store and analyse data across databases that are dotted around anywhere in the world.

Bernad Marr Bigdata: using Smart BigData analytics and metrics
To make better decisions

Velocity…

…refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds. Technology allows us now to analyse the data while it is being generated (sometimes referred to as in-memory analytics), without ever putting it into databases.

Bernad Marr Bigdata: using Smart BigData analytics and metrics
To make better decisions

Variety…

refers to the different types of data we can now use. In the past we only focused on structured data that neatly fitted into tables or relational databases, such as financial data. In fact, 80% of the world's data is unstructured (text, images, video, voice, etc.) With big data technology we can now analyse and bring together data of different types such as messages, social media conversations, photos, sensor data, video or voice recordings.

Bernad Marr Bigdata: using Smart BigData analytics and metrics
To make better decisions

Veracity…

…refers to the messiness or trustworthiness of the data. With many forms of big data quality and accuracy are less controllable (just think of Twitter posts with hash tags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of content) but technology now allows us to work with this type of data.

Bernad Marr Bigdata: using Smart BigData analytics and metrics
To make better decisions

Turning Big Data into Value:

The datafication of our world gives us unprecedented amounts of data in terms of Volume, Velocity, Variety and Veracity. The latest technology such as cloud computing and distributed systems together with the latest software and analysis approaches allow us to leverage all types of data to gain insights and add value.

Bernad Marr Bigdata: using Smart BigData analytics and metrics
To make better decisions

Turning Big Data into Value:

The 'Datafication' of our World;

- Activities
- Conversations
- Words
- Voice
- Social Media
- Browser logs
- Photos
- Videos
- Sensors
- Etc.

Volume

Velocity

Variety

Veracity

Analysing Big Data:

- Text analytics
- Sentiment analysis
- Face recognition
- Voice analytics
- Movement analytics
- Etc.

Bernad Marr Bigdata: using Smart BigData analytics and metrics To make better decisions

First exercise: look for open datasets

Esempi – dati governativi

- Data.gov
- US Census Bureau
- Open Data Portal dell'Unione Europea
- Data.gov.uk
- the CIA World Factbook
- Healthdata.gov
- Dati.gov.it

DwB
Data without Boundaries

Search.. SEARCH

About DwB

- Work Packages
- Deliverables
- Participants

Metadata Services

- CIMES for European National Official Statistics Microdata
- MISSY for Integrated European Official Statistics Microdata

Access Services

- Legal Frameworks
- National Accreditation & Access
- Transnational Access
- Guides
- Routines for Integrated Microdata
- Synthetic data tools

Activities and Events

- European Data Access Forums
- Users' Conferences
- Training Events and Material
- Staff Visits to Research Data

Data without Boundaries – DwB

The Data without Boundaries – DwB – project came to a formal end on 30 April 2015. The project had a mission to support equal and easy access to the rich resources of official microdata for the European Research Area, within a structured framework where responsibilities and liability would be equally shared. During its four-year lifespan the DwB worked towards preparing a comprehensive European service with better and friendly metadata, a more harmonized transnational accreditation and a secure infrastructure that would allow transnational access to the highly detailed and confidential microdata, both national and European, so that the European Union would be able to continuously produce cutting-edge research and reliable policy evaluations.

The resulting output is presented on this website as well as the tools and services that are maintained and developed by project partners beyond the project end. Please, use the navigation on the left to learn more. The main event to display and discuss the project work was the Second European Data Access Forum (EDAF) held in March 2015 in Luxembourg. All the presentation slides and audio files are available from the EDAF event page.

Bridging Three Communities

The DwB Formal End

The project came to a formal end on 30 April 2015. [More...]

NEW Deliverable Available!

D5.5 (Final report & recommendations for the continuation of services for European OS microdata) is now publicly available online [More...]

New Service: Visualisation Tool

Legal frameworks for official statistics microdata access demonstrated by Visualisation Tool. [More...]

Microdata Information System MISSY

Systematically structured metadata for official statistics. [More...]

New Service: CIMES

Overview of metadata of official microdata for research purposes. [More...]

UNITED NATIONS GLOBAL PULSE

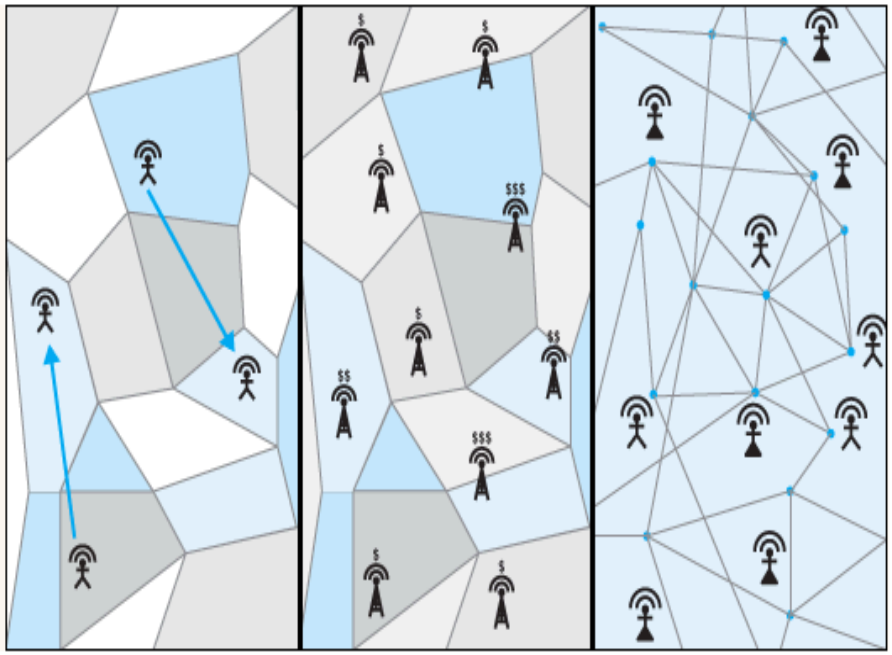
Search SEARCH

GLOBAL PULSE

- HOME
- RESEARCH
- LABS
- BLOG
- MULTIMEDIA
- ABOUT
- CONTACT

MOBILE PHONE NETWORK DATA FOR DEVELOPMENT

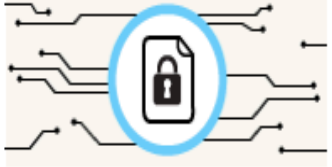
A synthesis of a growing body of research on mobile phone data analysis in development or humanitarian contexts. Read More /



SUBSCRIBE TO OUR NEWSLETTER

email address GO

RESEARCH PROJECTS



PRIVACY PRINCIPLES

ROADMAP

PULSE LABS

NEWS

Thoughts on the Google Flu Trends debate

TWITTER

Global Pulse 4 Apr

A WORLD THAT COUNTS

MOBILISING THE DATA REVOLUTION FOR SUSTAINABLE DEVELOPMENT