

Big Data Analytics

FOSCA GIANNOTTI AND ROBERTO TRASARTI

[HTTP://DIDAWIKI.DI.UNIPI.IT/DOKU.PHP/BIGDATAANALYTICS/BDA/](http://didawiki.di.unipi.it/doku.php/bigdataanalytics/bda/)

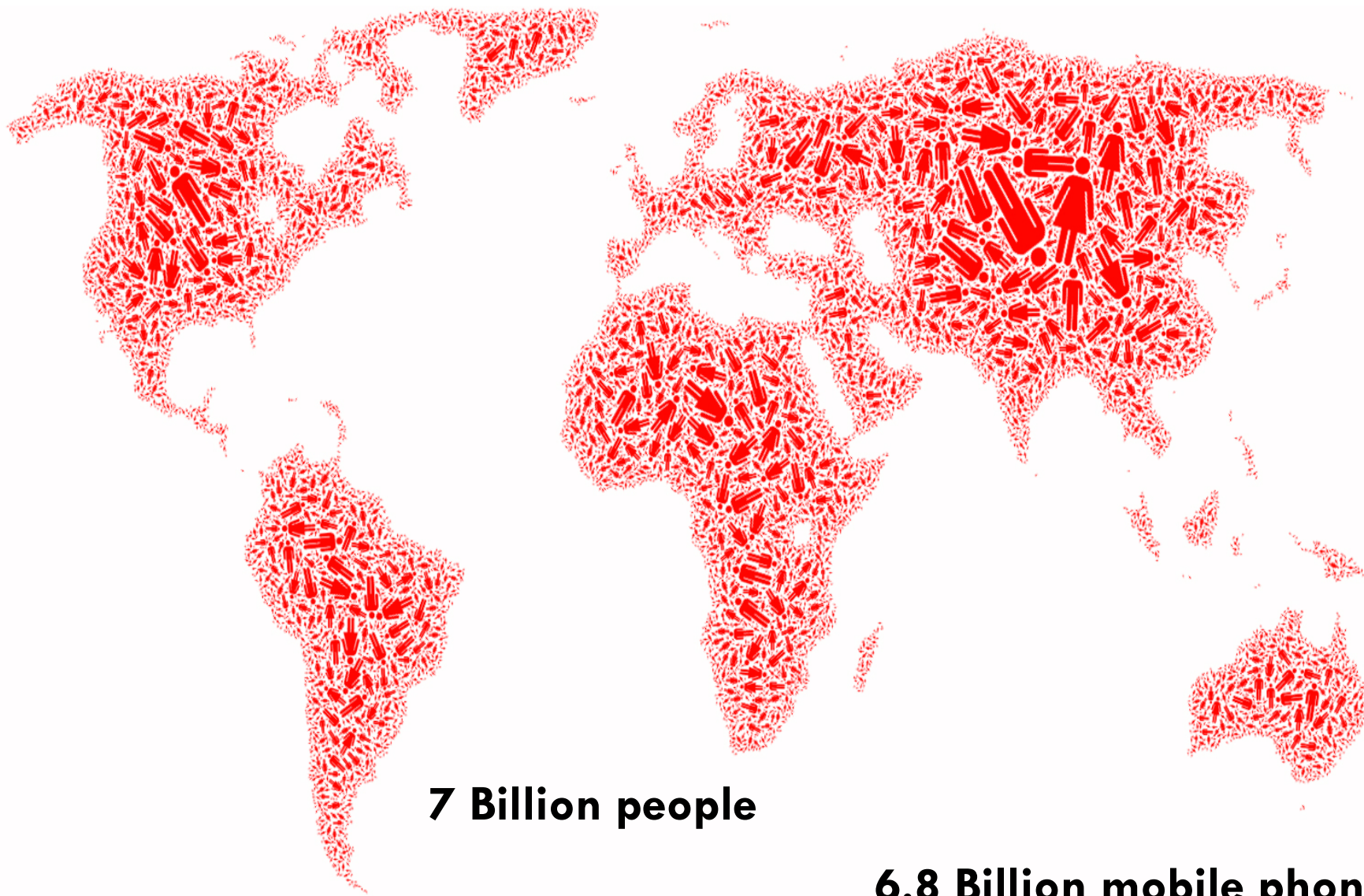
DIPARTIMENTO DI INFORMATICA - Università di Pisa
anno accademico 2017/2018

Big Data from smart environments

We live in an era where ubiquitous digital devices are able to broadcast rich information about human lives in real-time and at a high rate. The reality is that we just began to recognize significant research challenges across a spectrum of topics.







7 Billion people

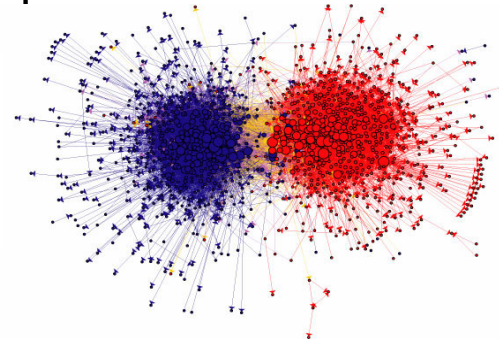
6.8 Billion mobile phones

Digital Footprints of Human Activities

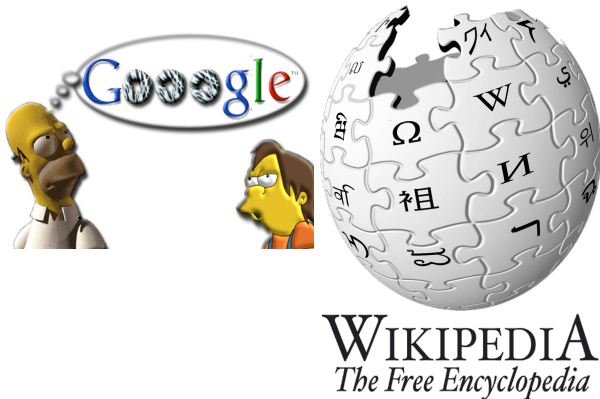
Shopping patterns & lifestyle



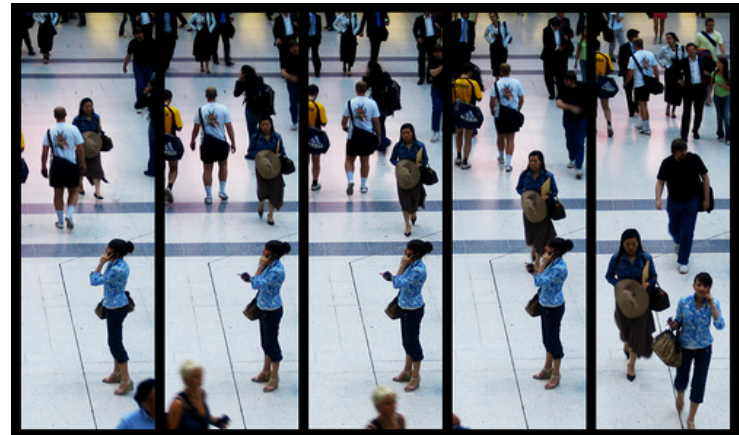
Relationships & social ties

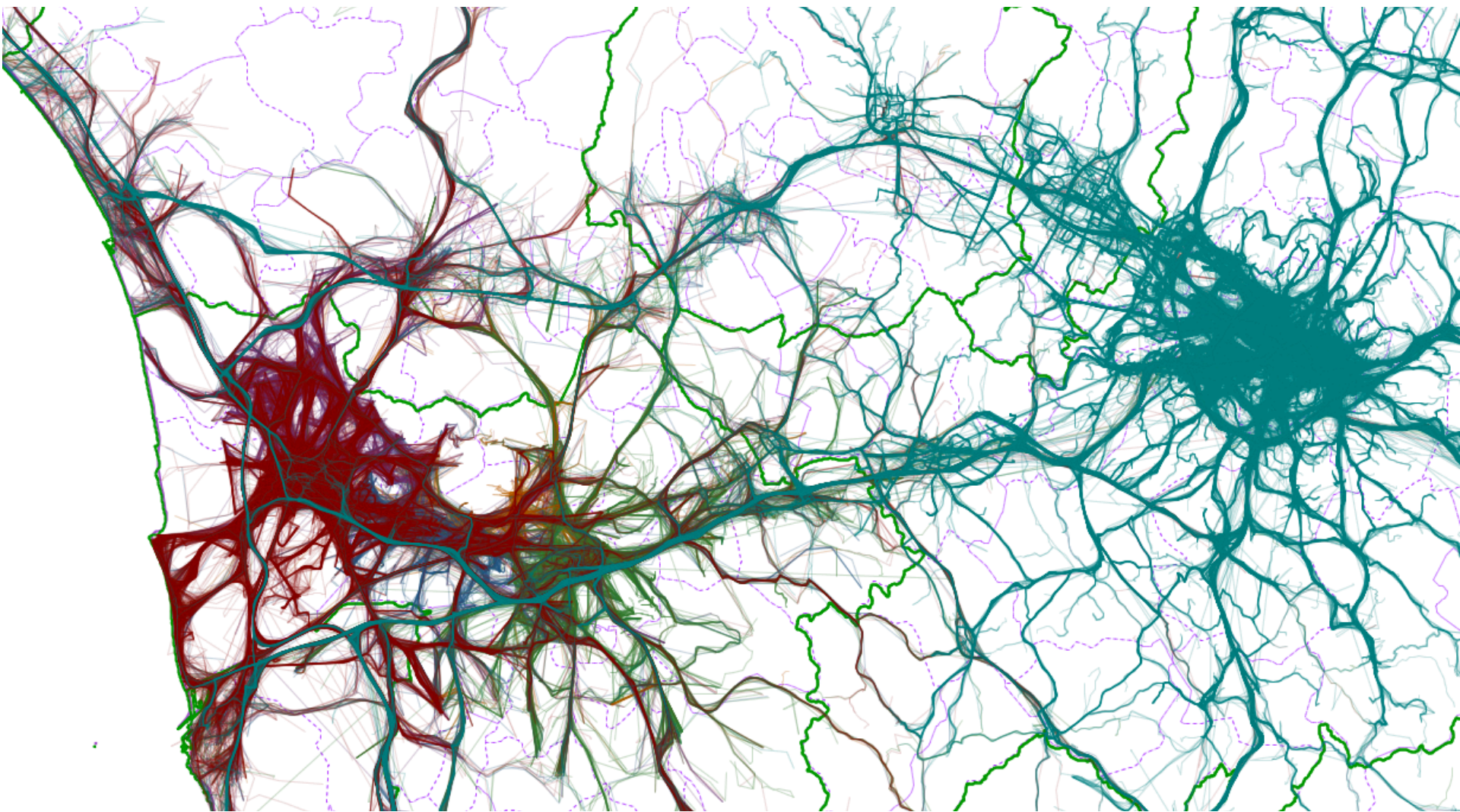


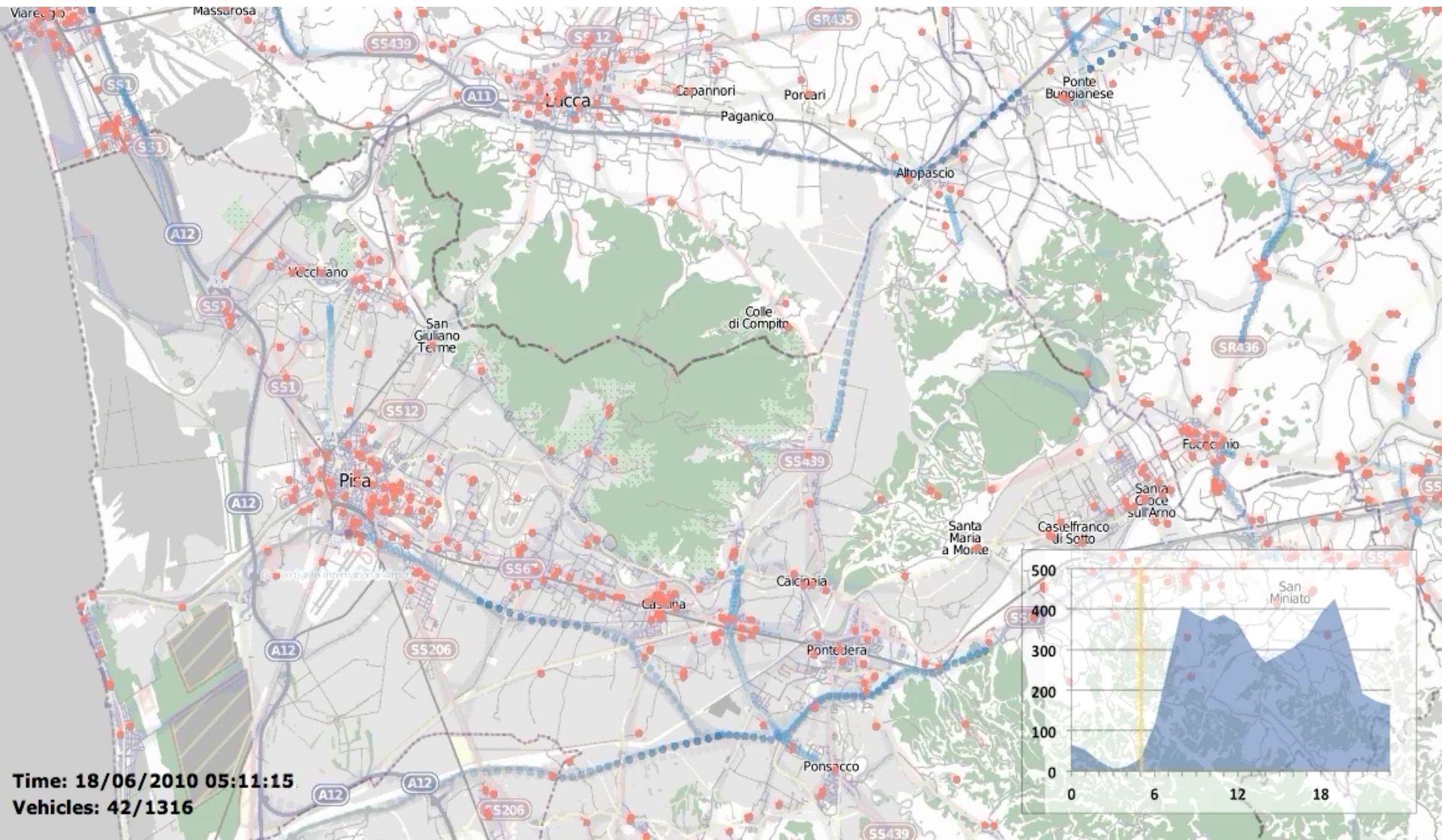
Interests, opinions, sentiments

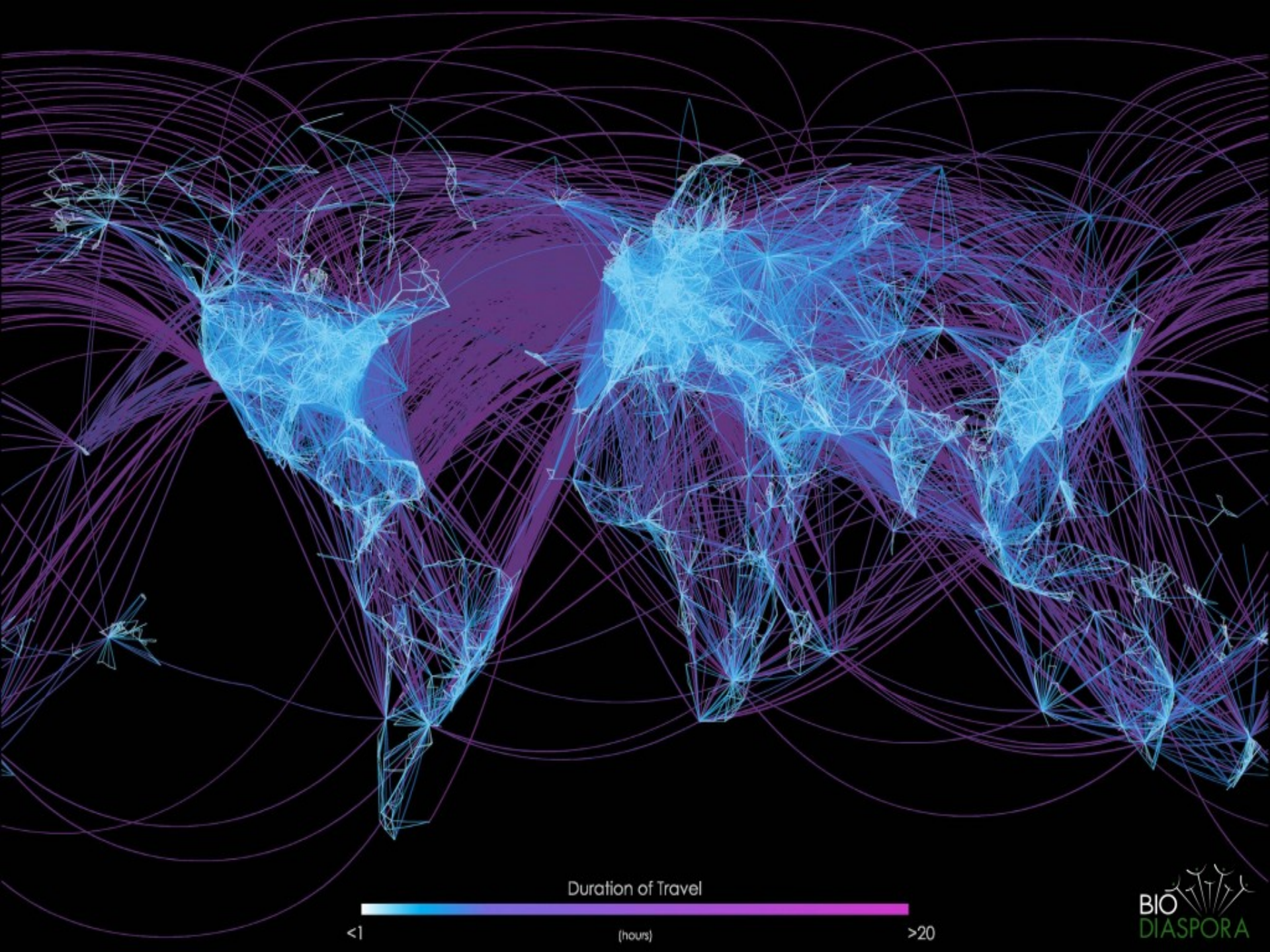


Movements









Duration of Travel

<1

(hours)

>20

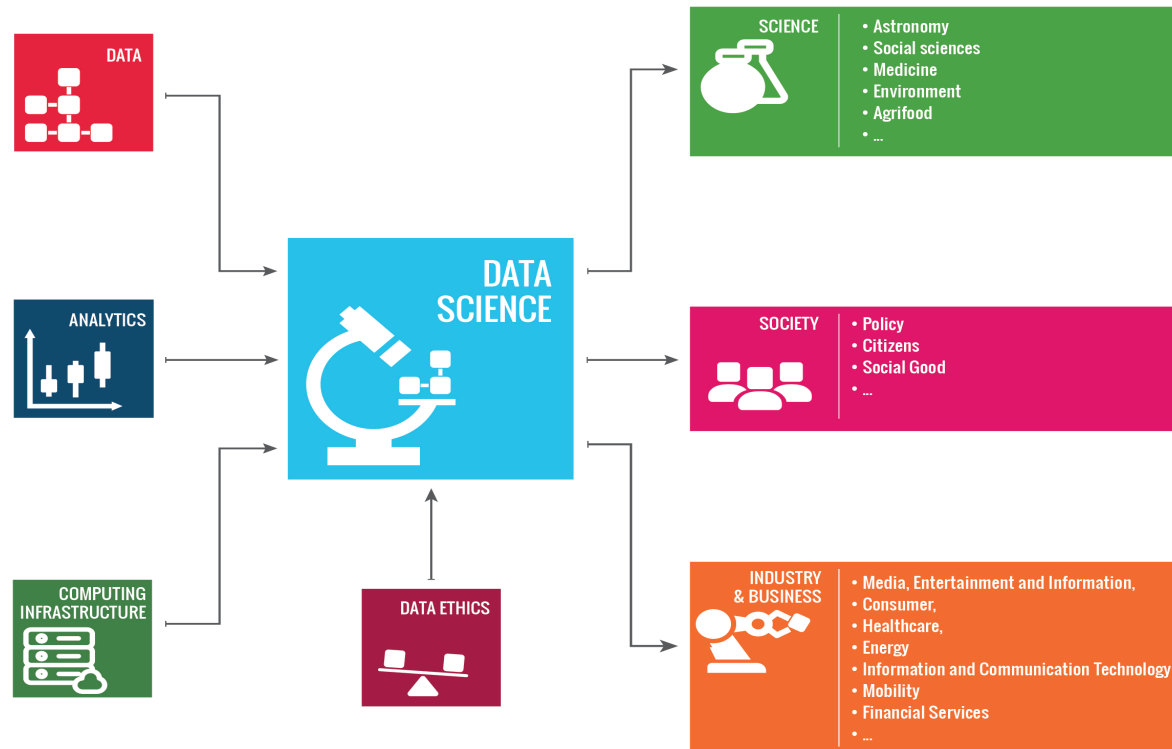




pollicini digitali

LA VITA NOVA, E-MAGAZINE DE IL SOLE 24
ORE

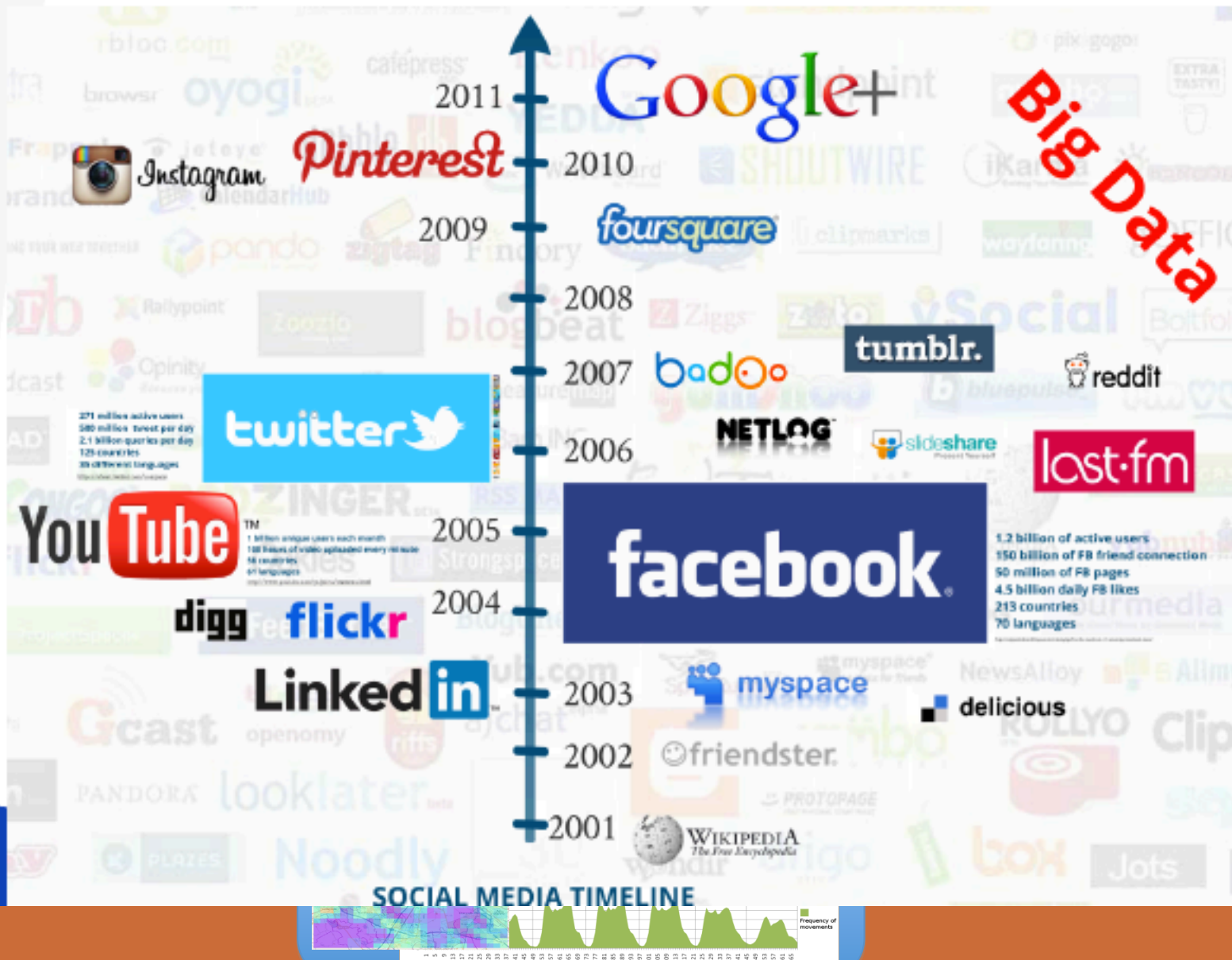
data availability, sophisticated analysis techniques, and scalable infrastructures brought what we call today “Data Science”



- “Data Science and BigData: a Game-changer for Science and Innovation” Document for G7 Academy, March 2017,
- “Realizing our Digital Future and shaping its impact on Knowledge, Industry, and WorkForce Document for G7 Academy, March 2018:

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)

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

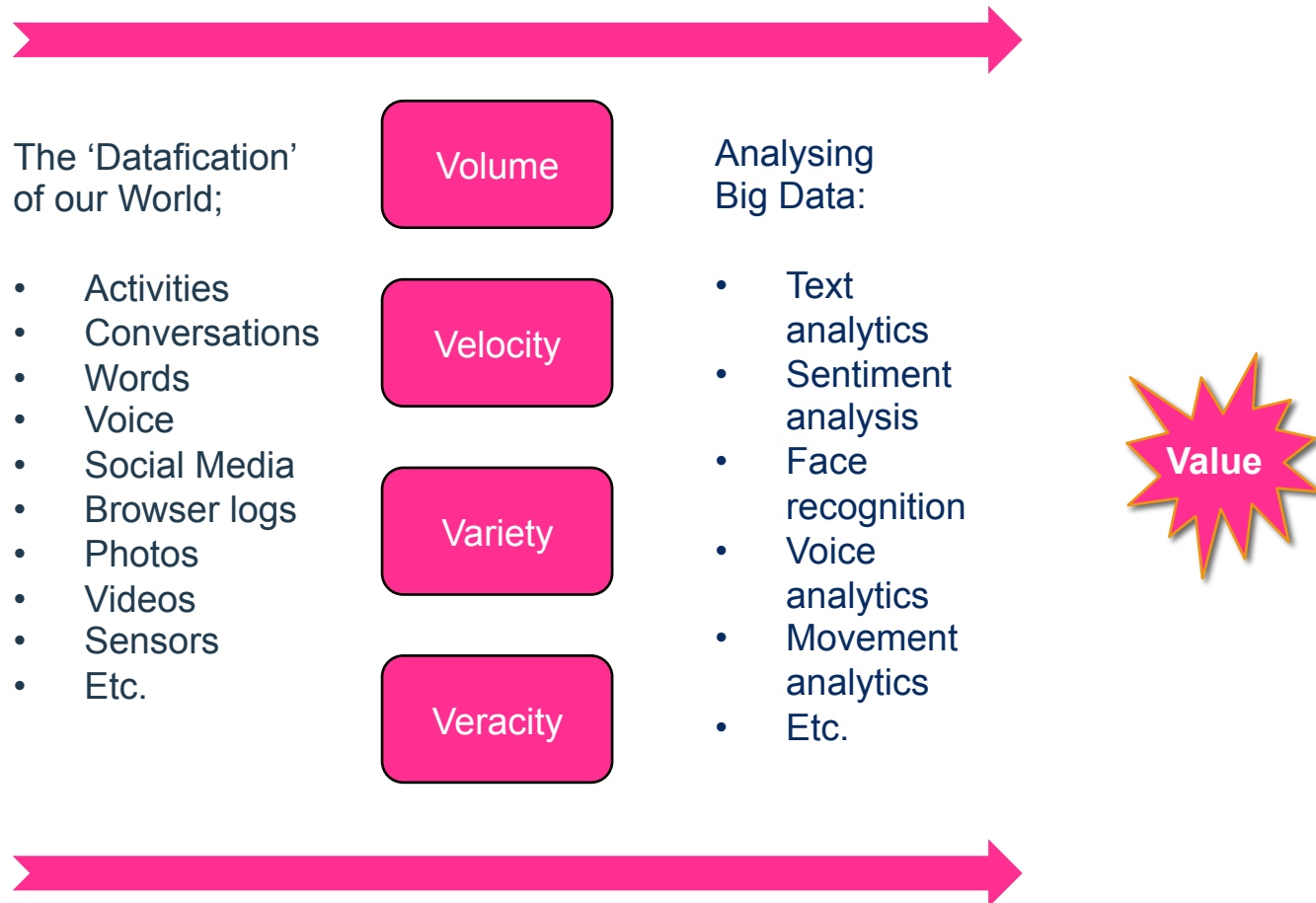
Variety of data in all respects and the challenges of combining variety

Veracity related to aspects such as trust in dealing with data, i.e. statistical significance.

Value

Privacy

Turning Big Data into Value:



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

The Future of Jobs

Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution

January 2016

New and Emerging Roles

Our research also explicitly asked respondents about new and emerging job categories and functions that they expect to become critically important to their industry by the year 2020, and where within their global operations they would expect to locate such roles.

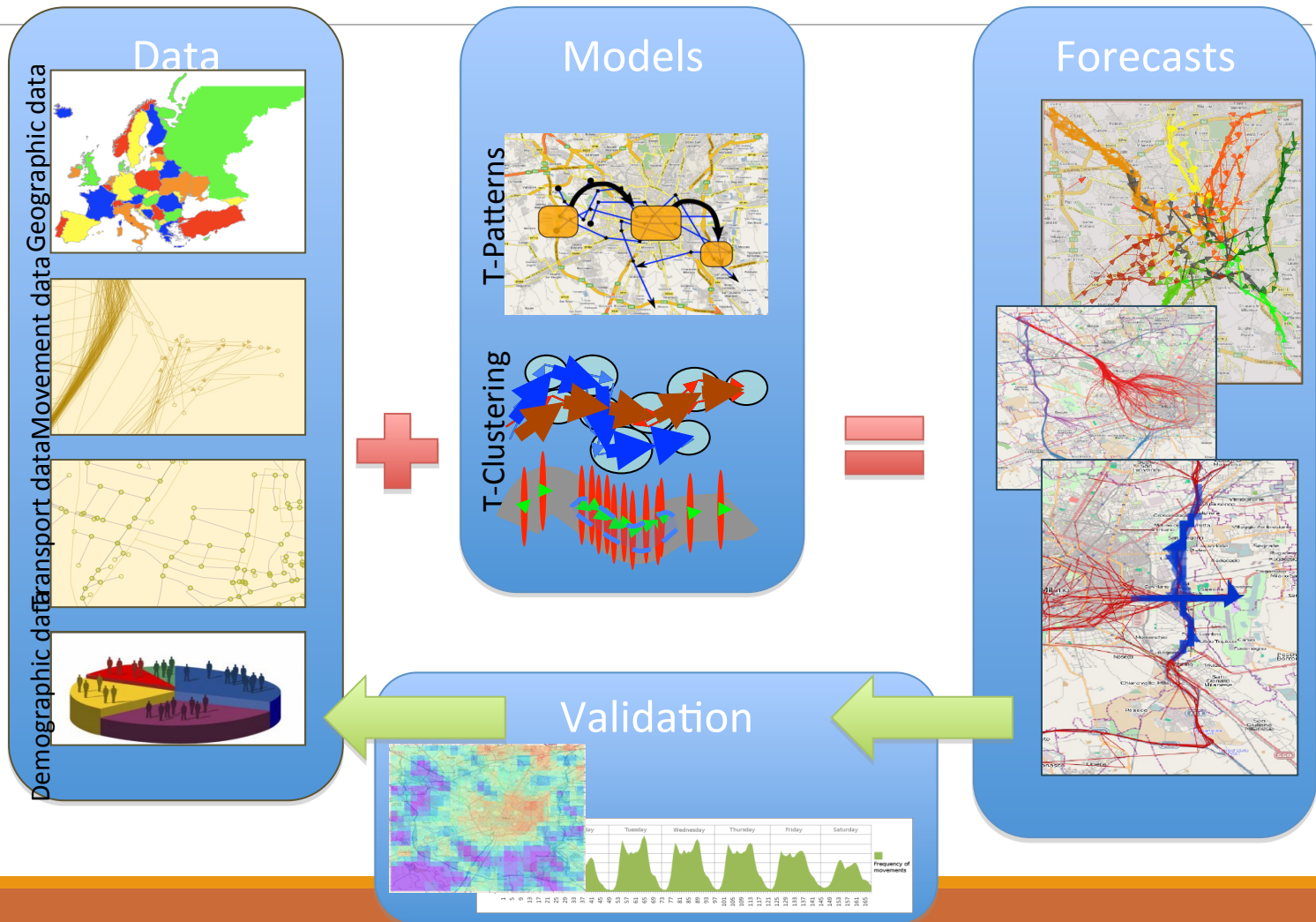
Two job types stand out due to the frequency and consistency with which they were mentioned across practically all industries and geographies. The first are data analysts, as already frequently mentioned above, which companies expect will help them make sense and derive insights from the torrent of data generated by the technological disruptions referenced above. The second



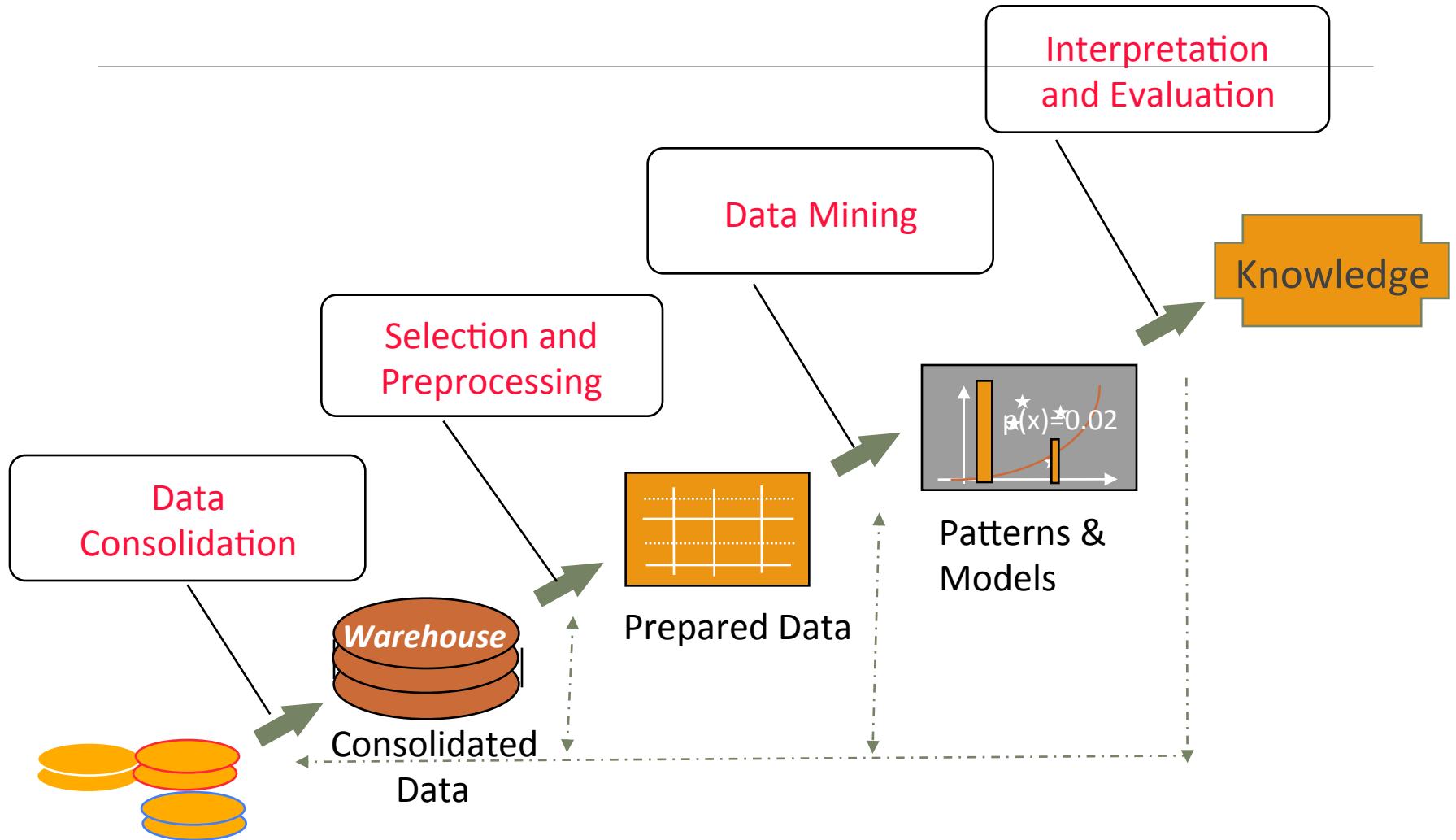
http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf

How to develop a big data analytics project

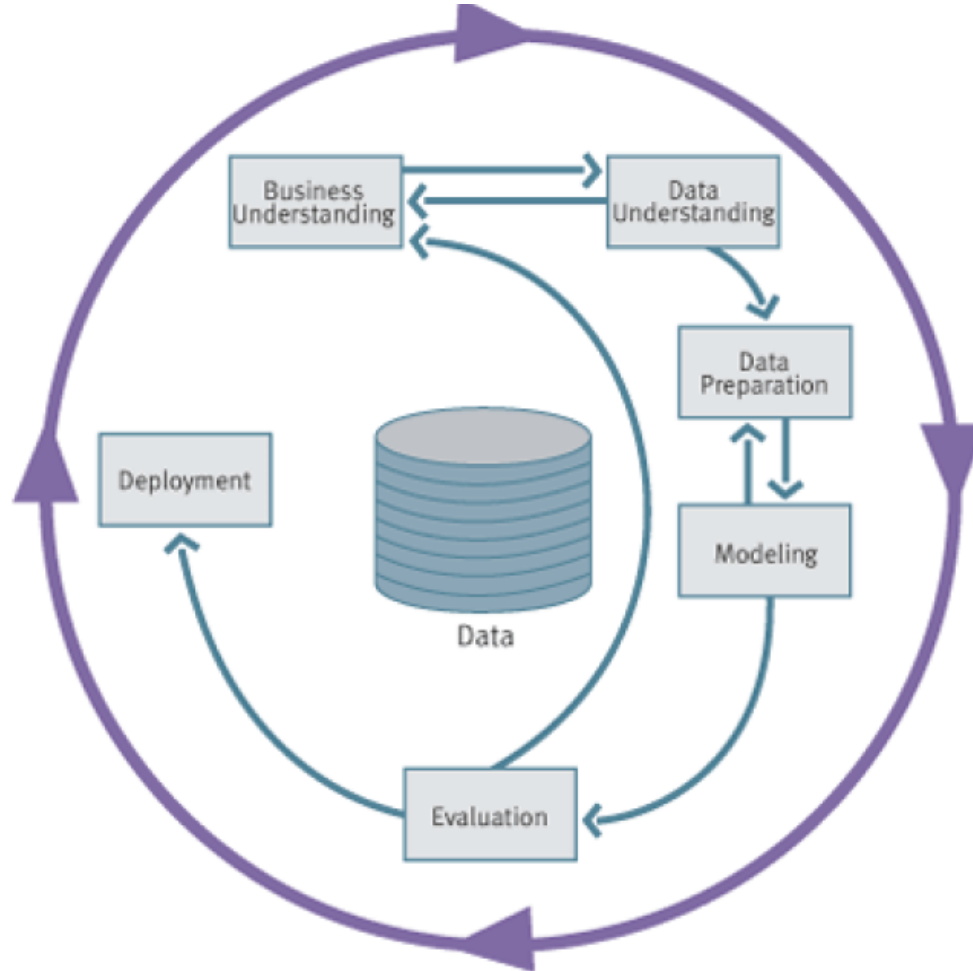
From DATA to KNOWLEDGE



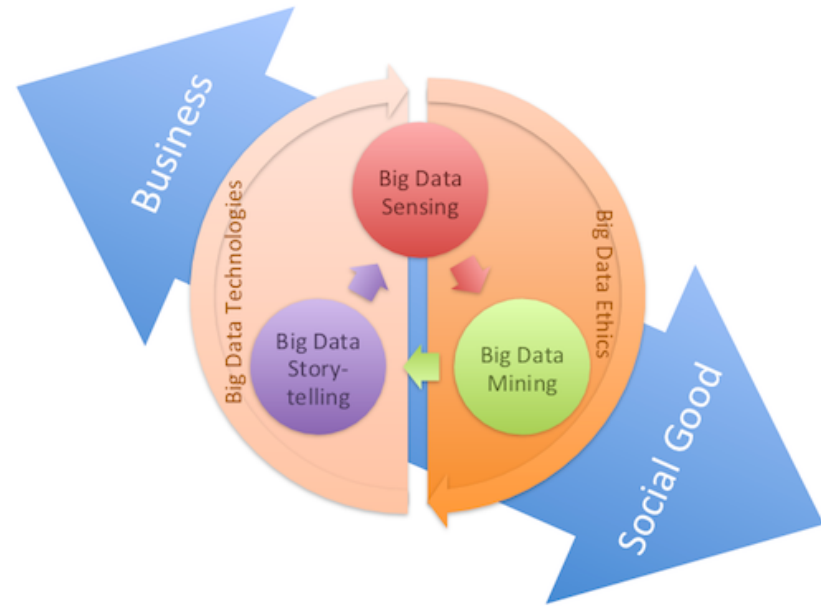
The KDD process



CRISP Model

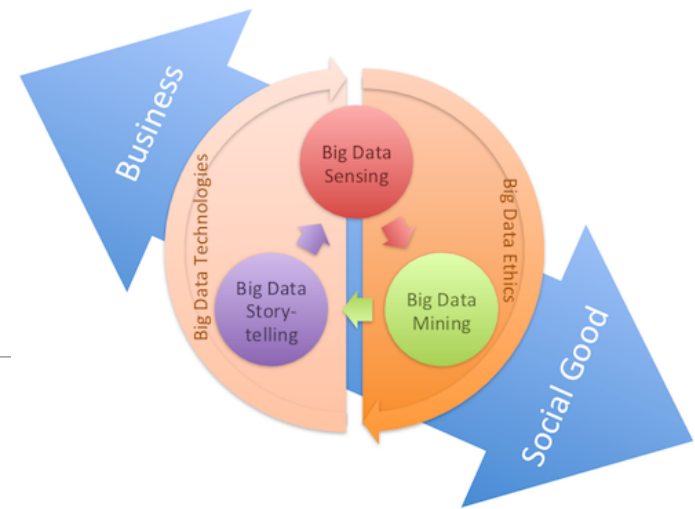


The modern data scientist!!!



Big Data Sensing & Procurement

Big data sources, crowdsourcing, crowdsensing
Web Search Engines and Information Retrieval
Analytical Crawling, Text Annotation



Big Data Mining

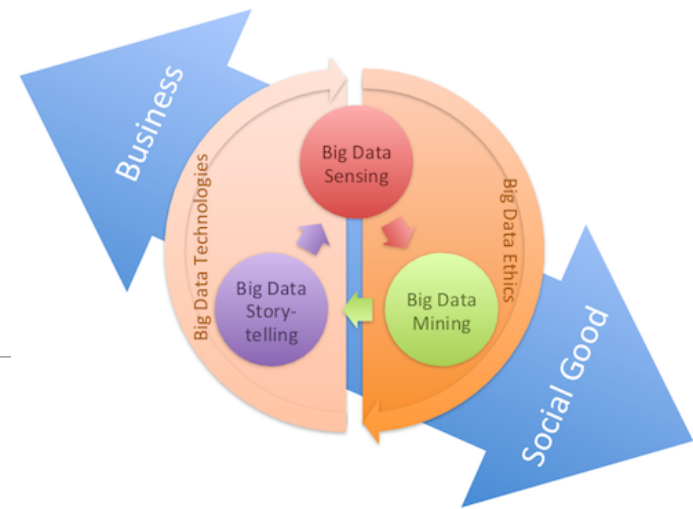
Data Mining & Machine Learning

Mobility Data Analysis

Social Network Analysis

Web Mining & Nowcasting

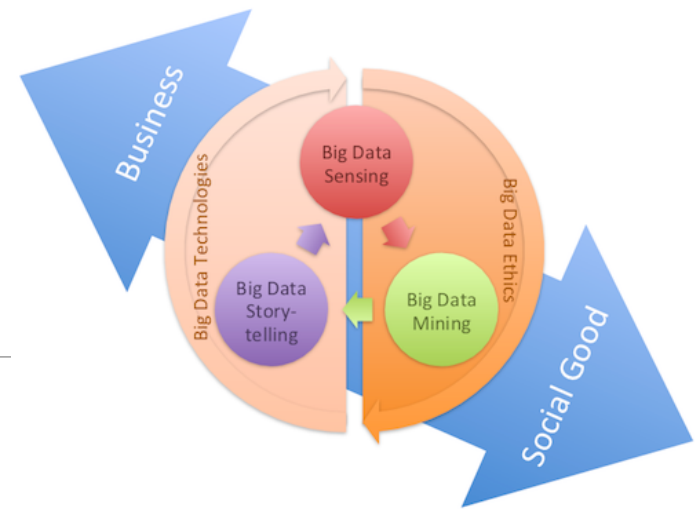
Sentiment Analysis & Opinion Mining



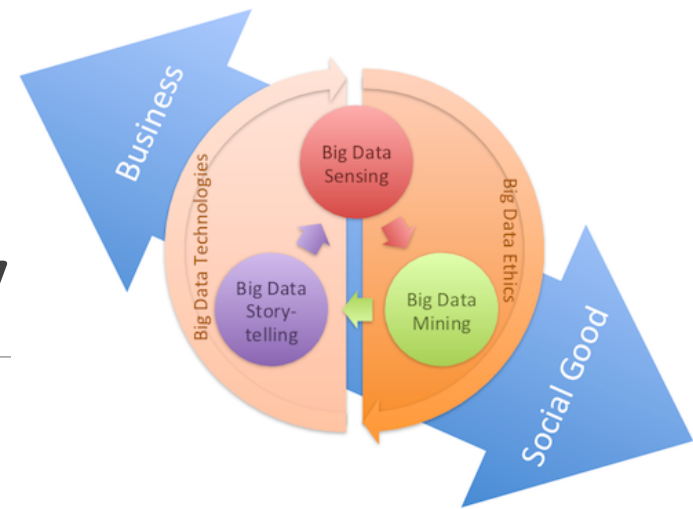
Big Data Story Telling

Data Visualization & Visual analytics

Data Journalism & Story Telling



Big Data Technology



Data Management for Business Intelligence

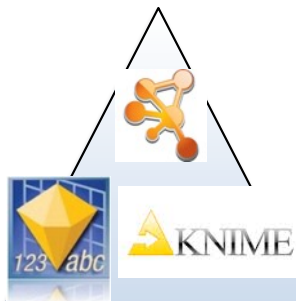
High Performance & Scalable Analytics, NO-SQL Big Data Platforms



Data Science



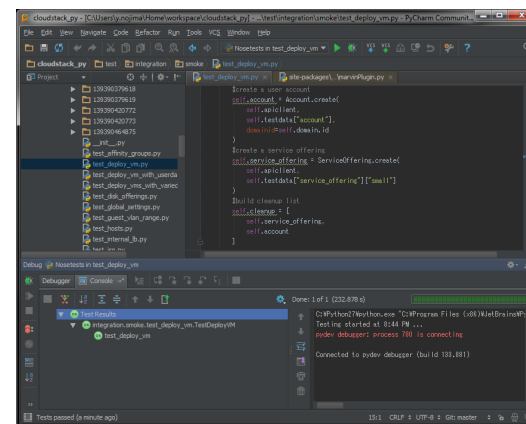
Visual Tools



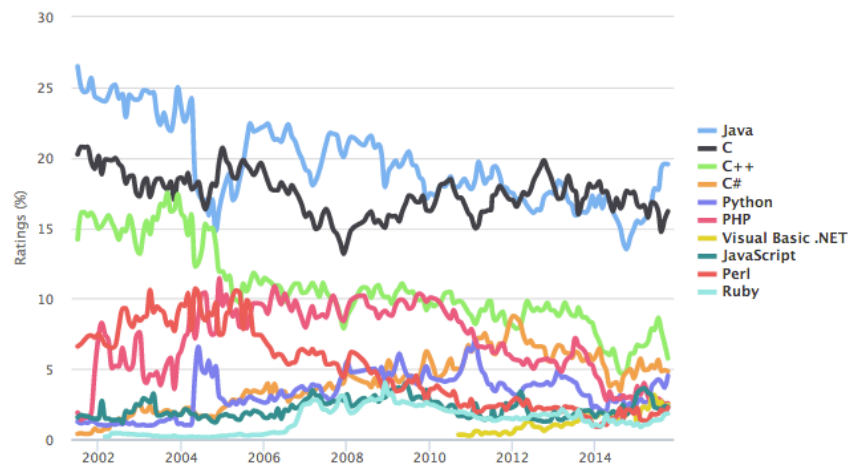
Specialized Libraries



Programming Languages



TIOBE Programming Community Index
Source: www.tiobe.com



Data Mining/ML



BI/Visualization



ETL/DW



Data Processing



Data Storage



Course Goals

This course is an introduction to the emergent field of big data analytics and social mining, aimed at acquiring and analyzing big data from multiple sources to the purpose of discovering the patterns and models of human behavior that explain social phenomena.

Course Focus

The focus is on:

- **new challenges in implementing a knowledge Discovery process ...when data are Big Data.**
- **what can be learnt from big data in different domains:** mobility and transportation, urban planning, demographics, economics, social relationships, opinion and sentiment, sport etc.;
- **the analytical and mining methods and methodology** that can be used to realize Big Data analytics projects .
- an **introduction to basic technologies** to collect, manipulate and process big data.

Module1: Methodological scenarios

lectures:

Lecture 1-2: What is possible to observe with Mobile Phone Data? Novel questions: Estimating Presence, estimating Origin-Destination Matrix, understanding city dynamics, classifying city users, observing unemployment, gender distribution, Nowcast Wellbeing.

Data preparation, Model Construction and Validation

Lecture 3-4: What is possible to observe with GPS data? Mobility Data mining methods in a nut shell: Trajectory patter mining, Mobility profiles, Next Location Prediction. Novel questions: Understanding human mobility, Understanding travel demand, Predicting travel purpose, Building territory indicators. **Data preparation, Model Construction and Validation**

Lecture 5-6: What is possible to observe with Social Media Data? Combining Space and Sentiment: measuring happyness with twitter data. Quantification. **Data preparation, Model Construction and Validation**

Lecture 7: What is possible to observe with IoT Data? Sensor data in sport and training. Predicting athlets injuries. **Data preparation, Model Construction and Validation**

Lecture 8: Paper presentation from students and peer-to-peer discussion (one presenter and two discussants)

Module2: Technologies lectures:

1. Python for Data Science
2. The Jupyter Notebook: developing open-source and reproducible data science
3. MongoDB: fast querying and aggregation in NoSQL databases
4. GeoPandas: analyze geo-spatial data with Python
5. Scikit-learn: programming tools for data mining and analysis
6. M-Atlas: a toolkit for mobility data mining

Module 3: Laboratory for interactive project development

1. Data Understanding and Project Formulation
2. Mid Term Project Results
3. Final Project results

Exam:

The two **mid-terms will be 40%** of the final grade, the remaining **60% is the evaluation of the Project and the Discussion**. There is the possibility to do the a final test about technologies if the Mid-Terms are not sufficient.

02/10 - Datasets presentation

30/10 - Mid-term Tech I

20/11 - Discussing the final project proposal - Collective discussion (not evaluated)

18/12 - Mid-term Tech II and Final Project proposal

15/01 & 16/02 - Final Project and Discussion

Project steps:

- data set presentation and projects will be presented on 2/10
 - the students are required to submit a proposal submission. A preliminary collective discussion is planned on 20/11
 - proposal submission is a report on data understanding that can be realized in team and a proposal for **each member of an analytical objective to be investigated individually**: not more than 8 pages. Proposal submission planned on 18/12
- (Collaborations are welcome, but at the end any student has to demonstrate her/his effort in realizing the project)
- the project report is presented before the oral exam and discussed individually on 15/01 or 16/02

Big Data Analytics- Evaluation

Ongoing projects (on small datasets) or seminars on research papers with presentation to the class

Final (Team) Project

- Team of 2-3 person.
- Unique grade.
- Projects consist into the realization of some complete analytical processes on a given problem and a given dataset, aimed at realizing some novel services
- A final report followign the CRISP standard describing all steps: exploration, preparation and anaysis and final evaluation.
- Project presentation .ppt

Individual Project Discussion

Big data & new questions to ask

Big Data & Social Mining



The Social
Microscope: **a tool to
measure, understand,
and possibly predict
human behavior**

Google Flu Trends

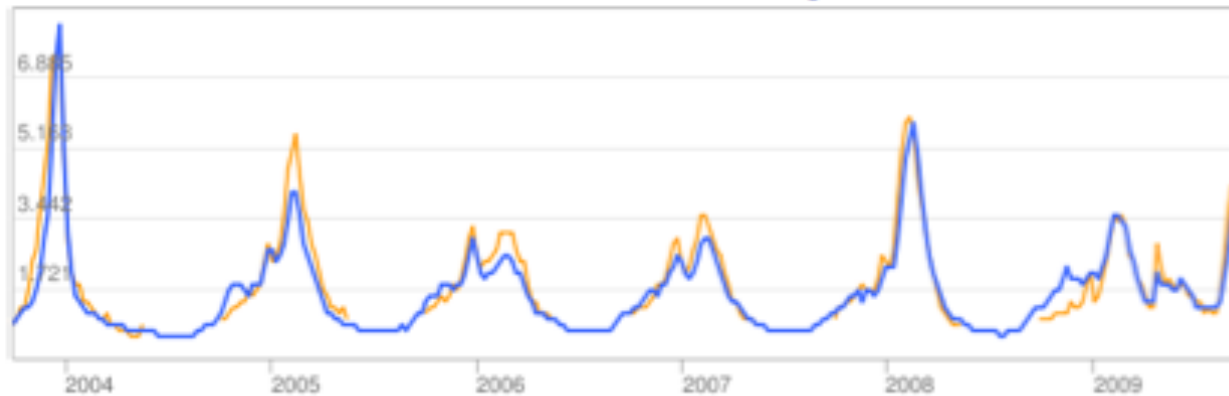
Stime storiche

Visualizza dati per: Stati Uniti

Attività influenzale Stati Uniti

Stima sull'influenza

● Stima di Google Trend influenzali ● Dati Stati Uniti



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

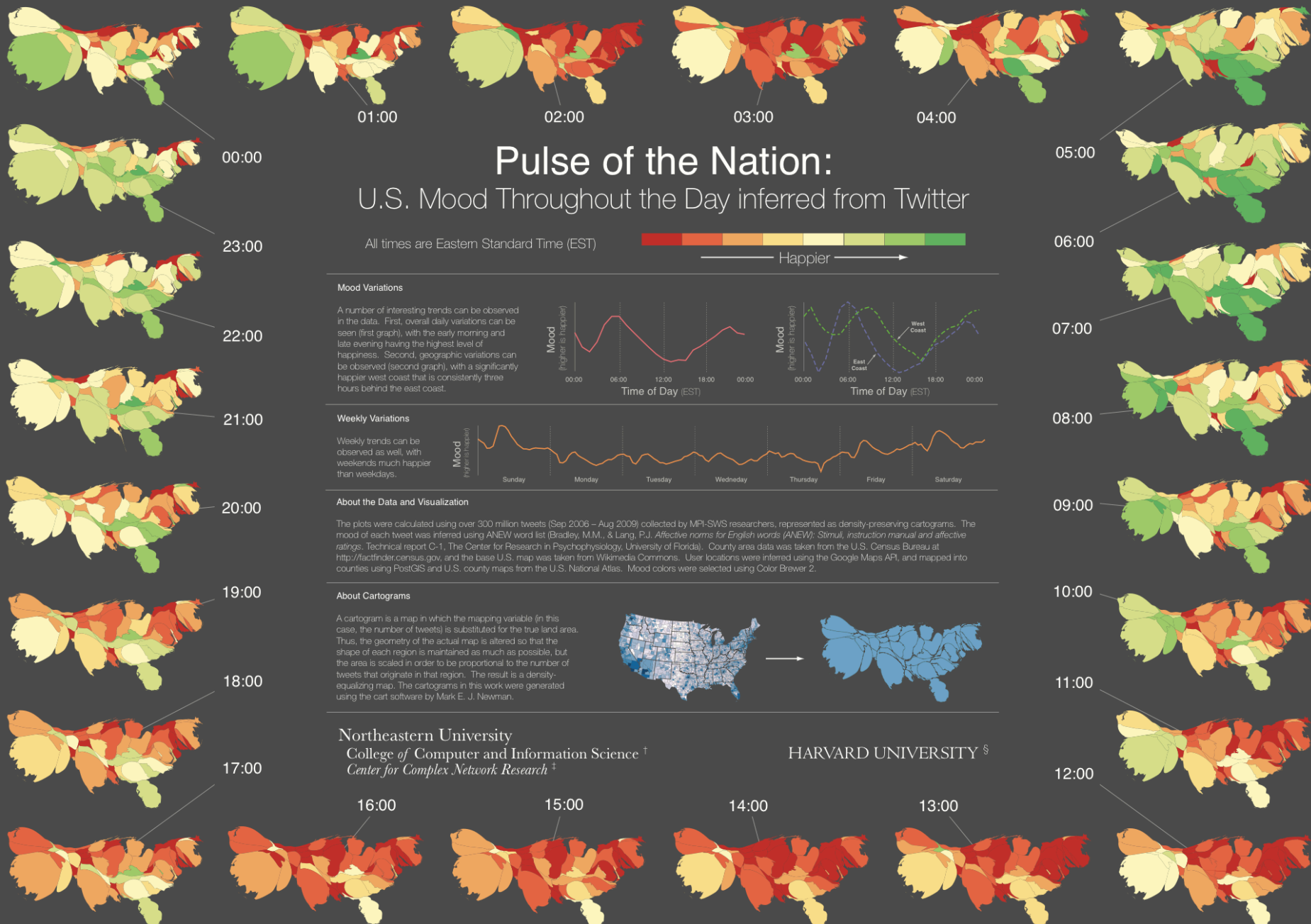


Detecting influenza epidemics using search engine query data

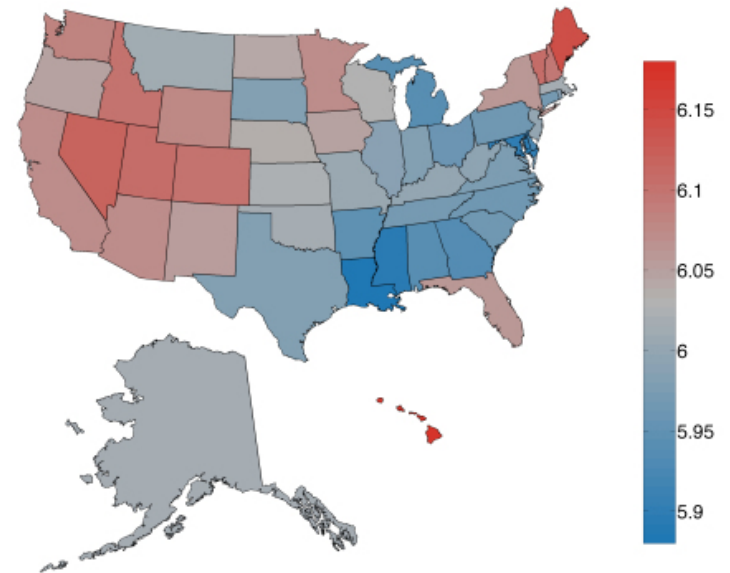
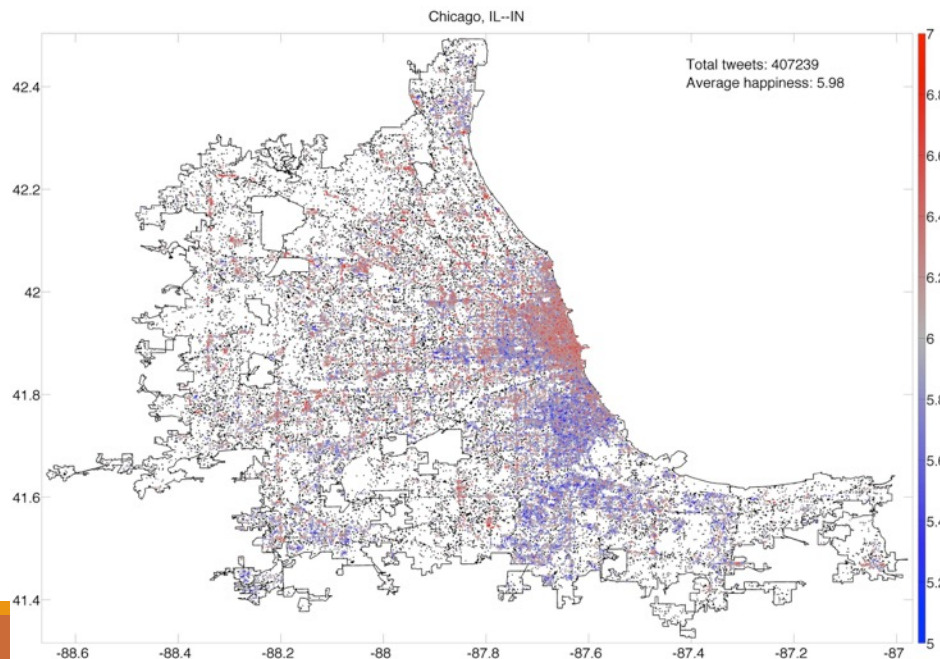
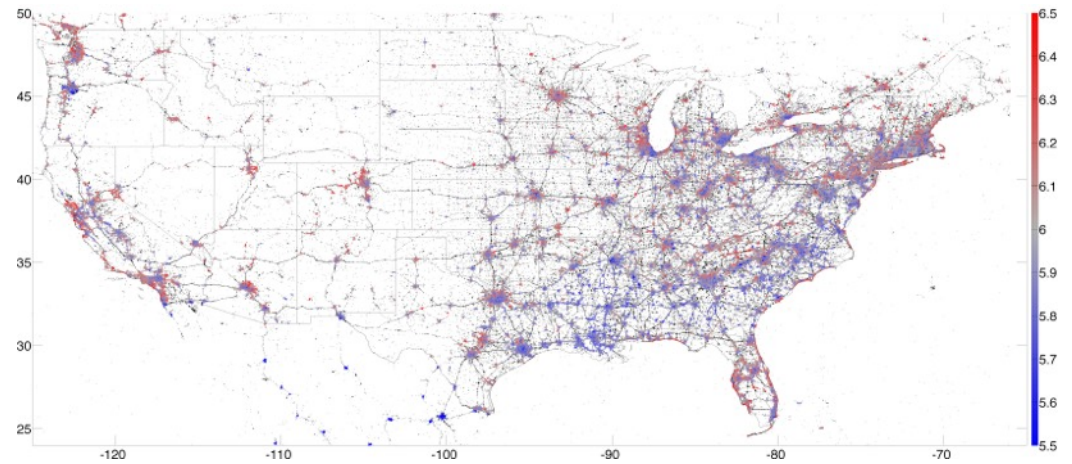
Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer², Mark S. Smolinski¹ & Larry Brilliant¹

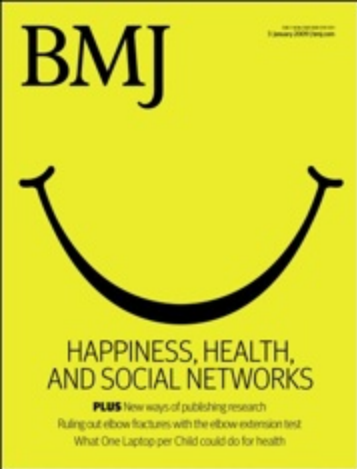
¹Google Inc. ²Centers for Disease Control and Prevention

Nature 457, 1012-1014 (2009)

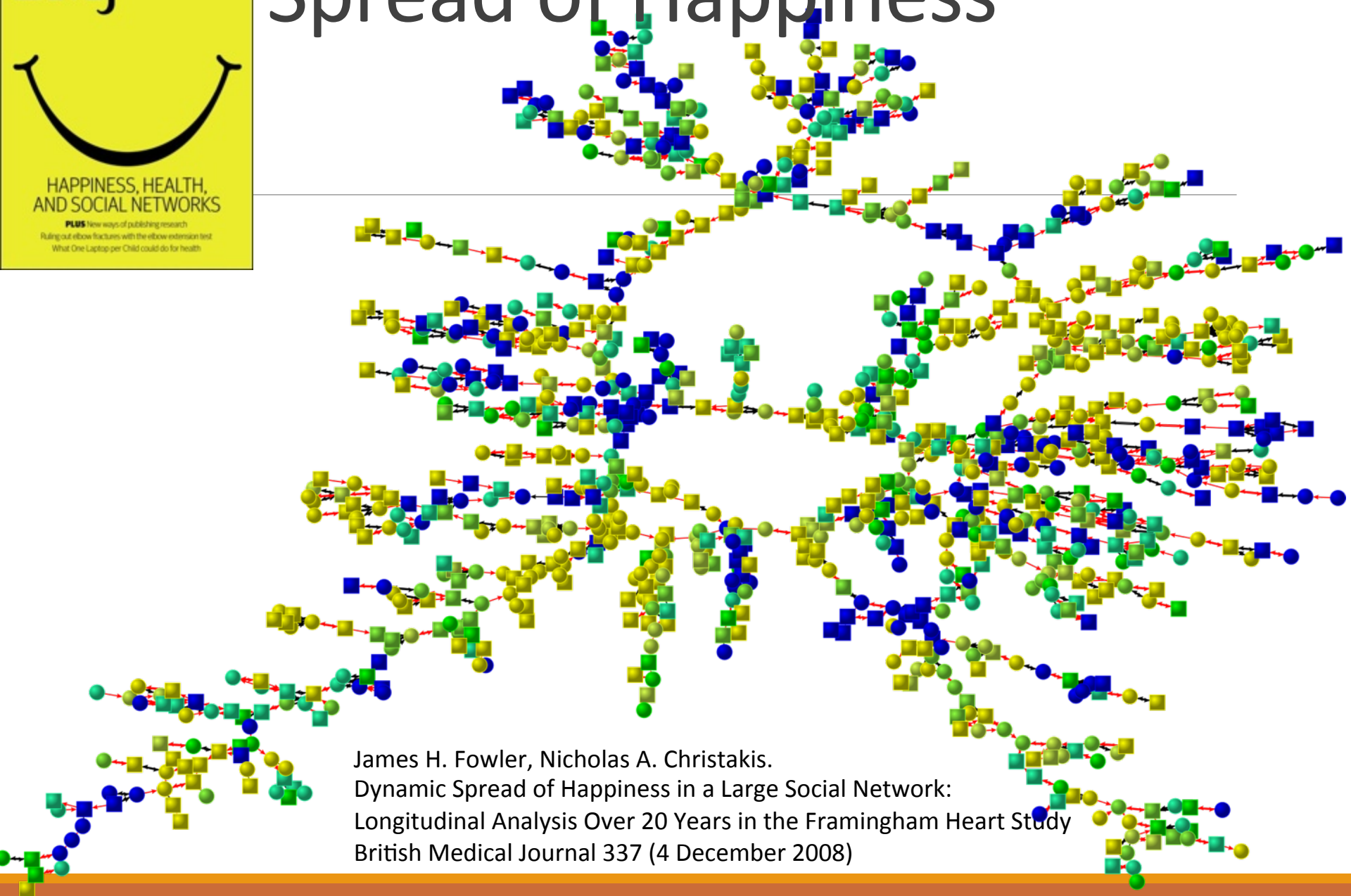


Searching the most happy city in US



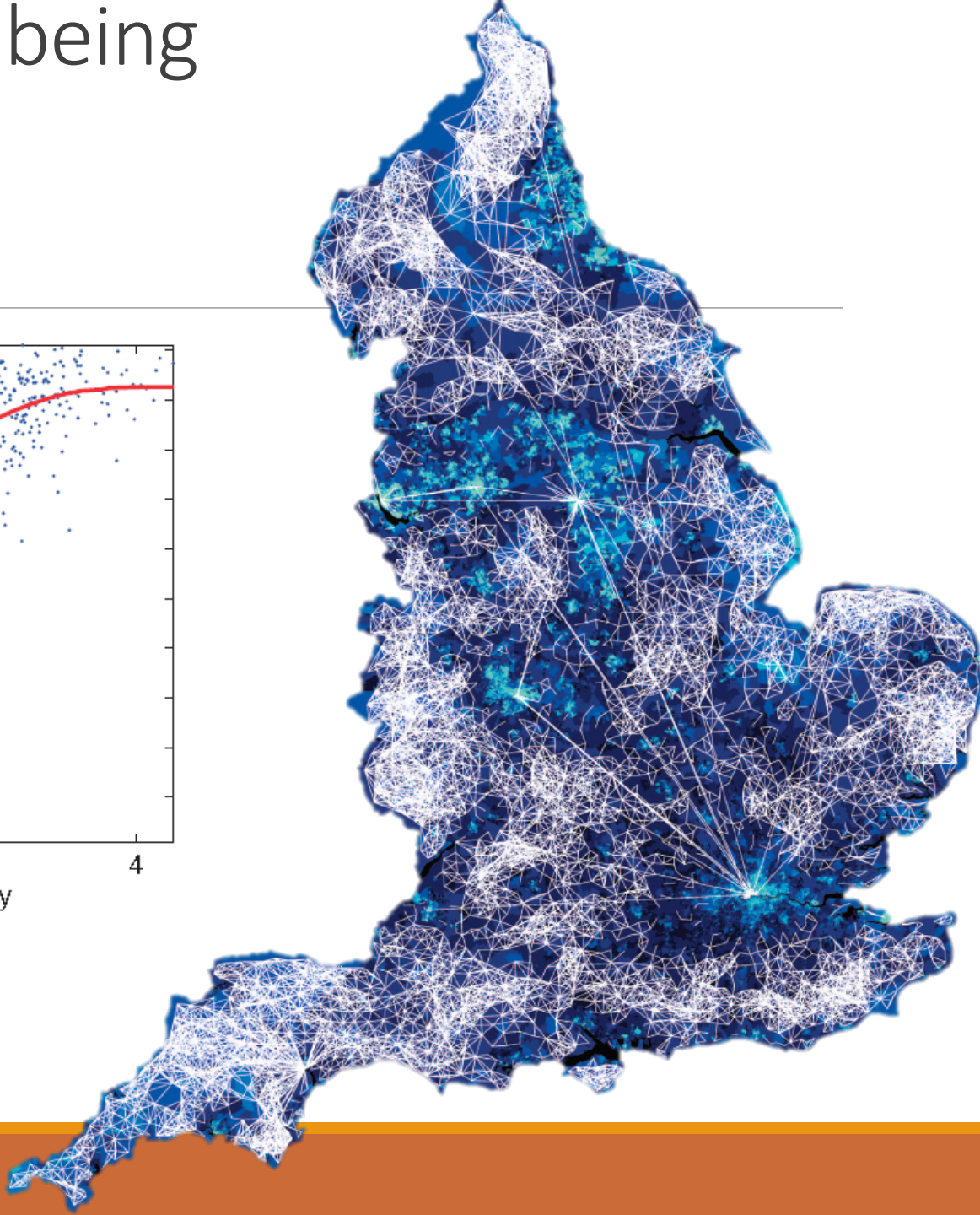
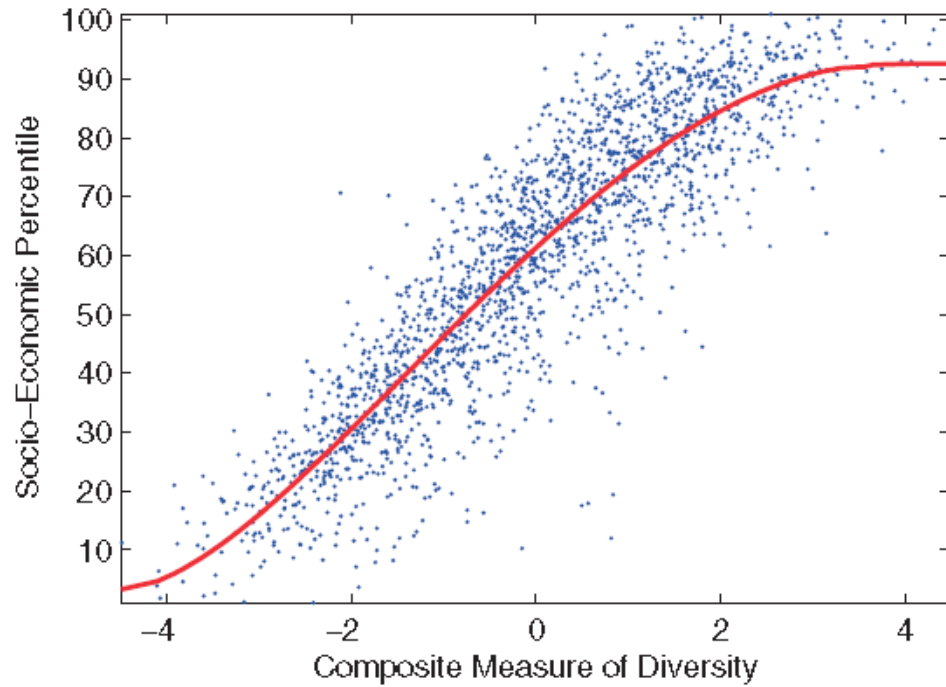


Spread of Happiness

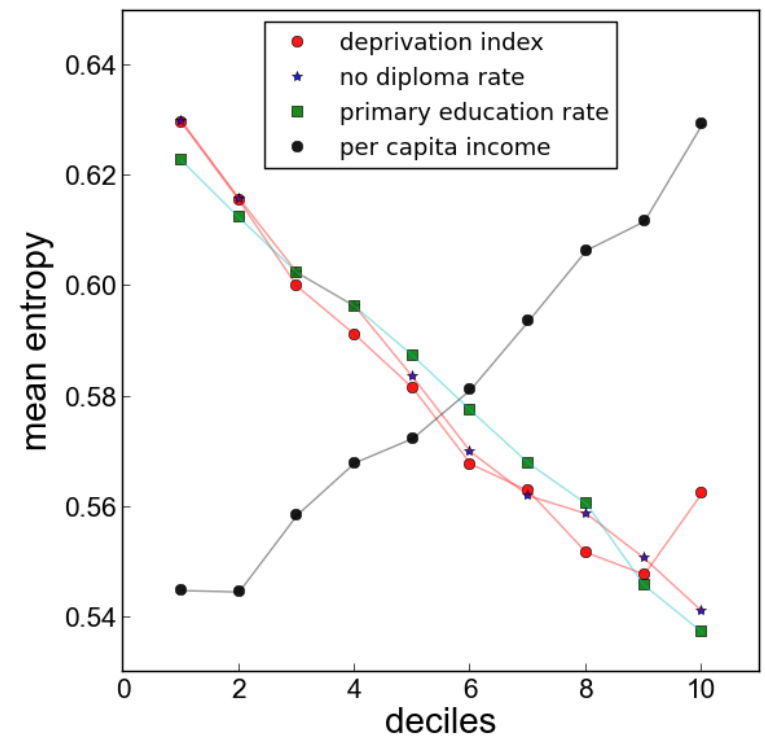
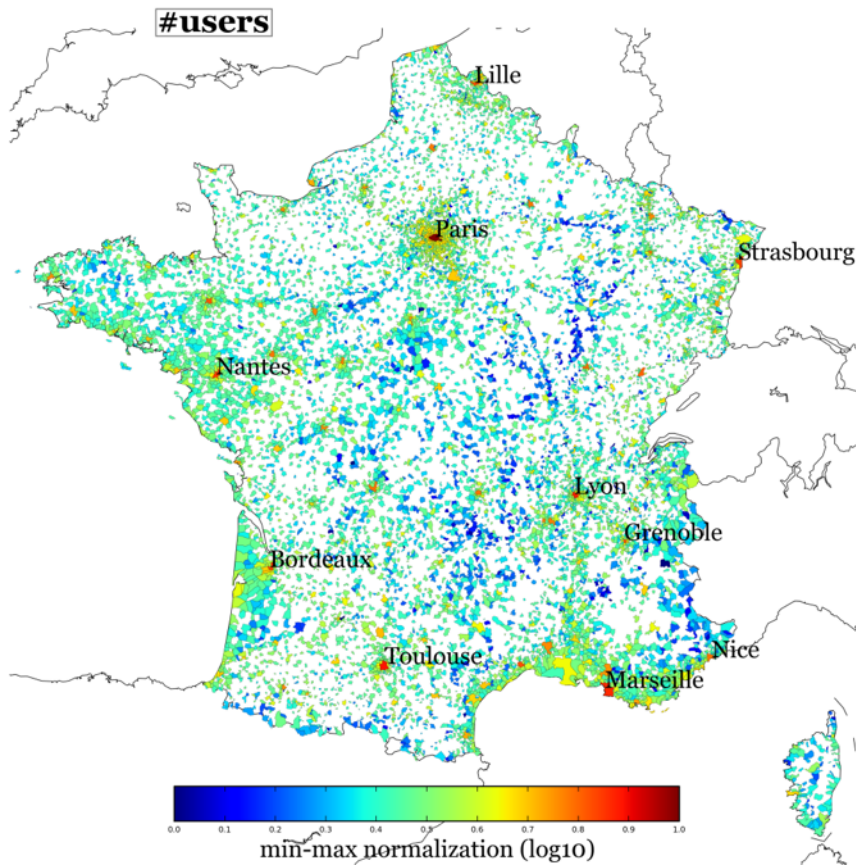


James H. Fowler, Nicholas A. Christakis.
Dynamic Spread of Happiness in a Large Social Network:
Longitudinal Analysis Over 20 Years in the Framingham Heart Study
British Medical Journal 337 (4 December 2008)

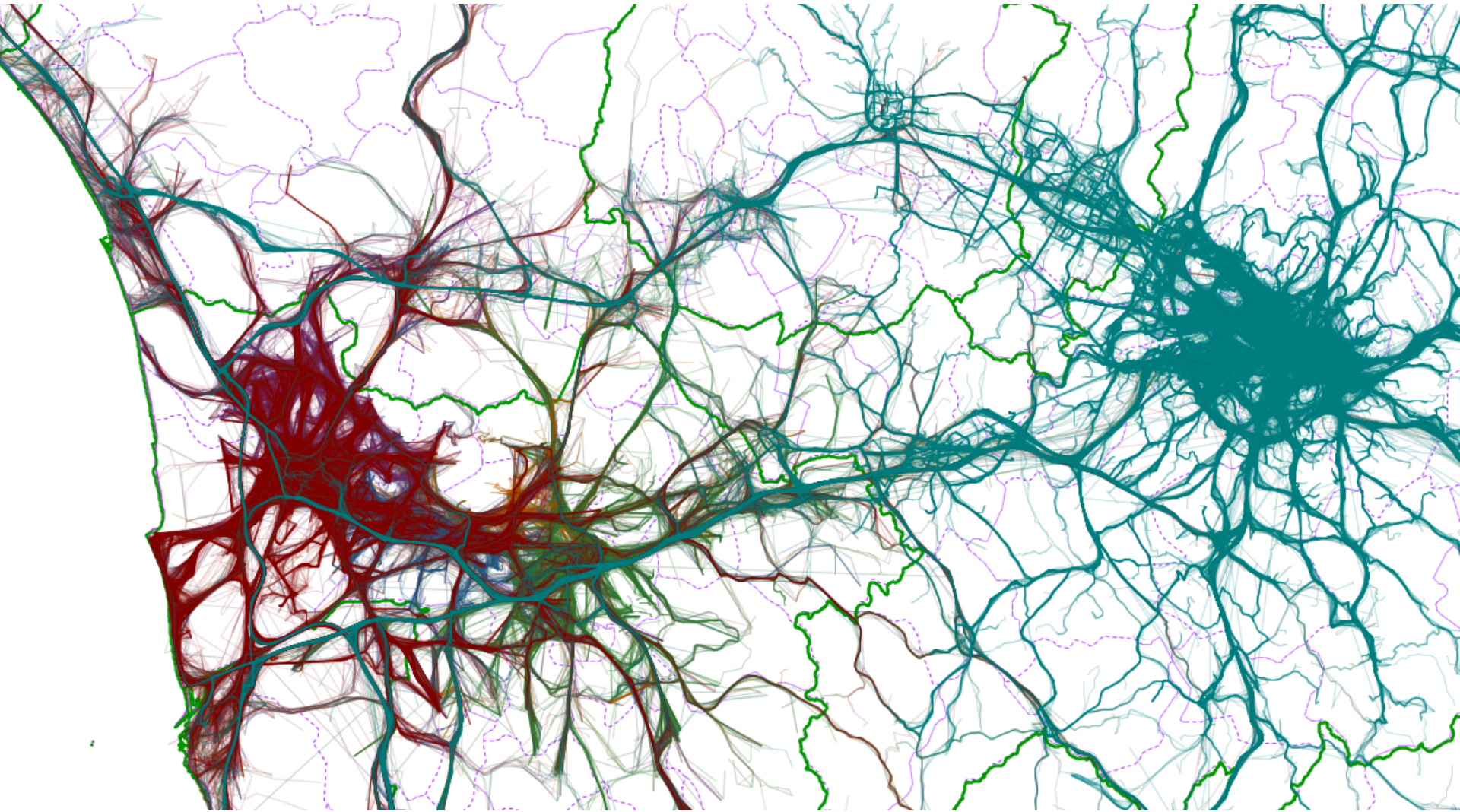
Diversity and Wellbeing (phone calls)



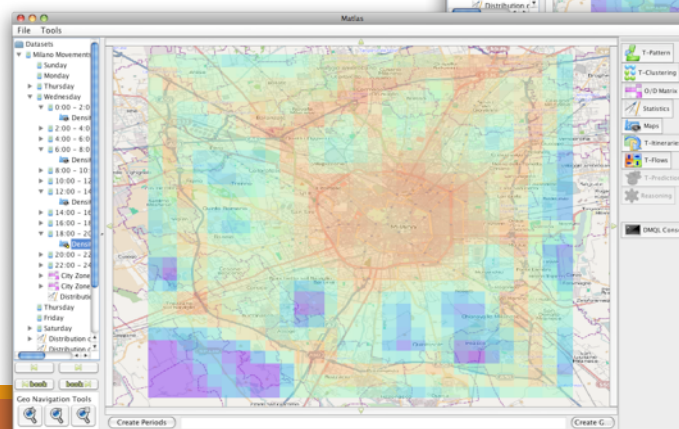
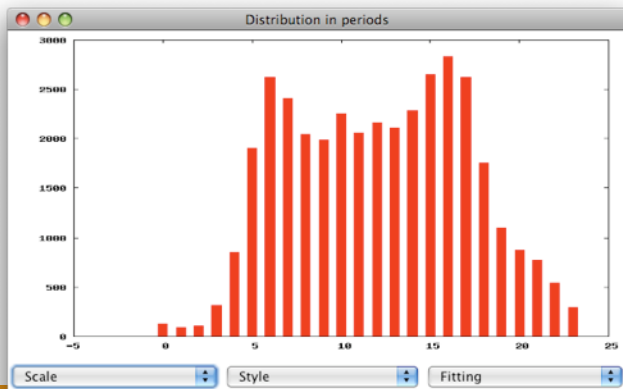
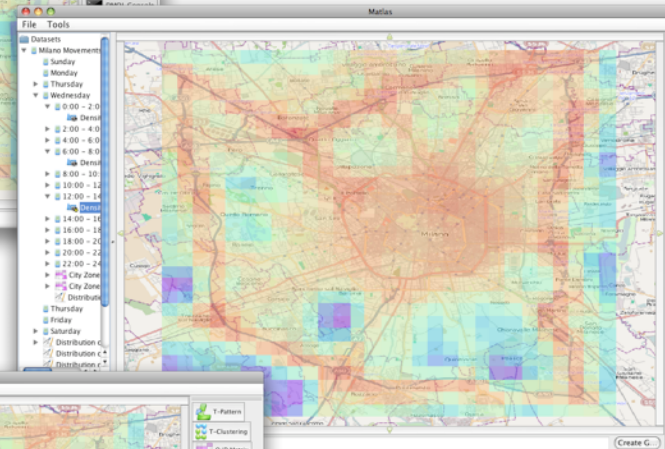
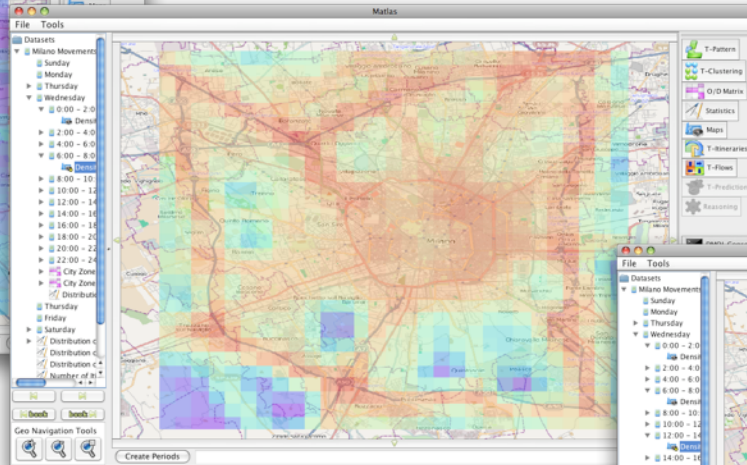
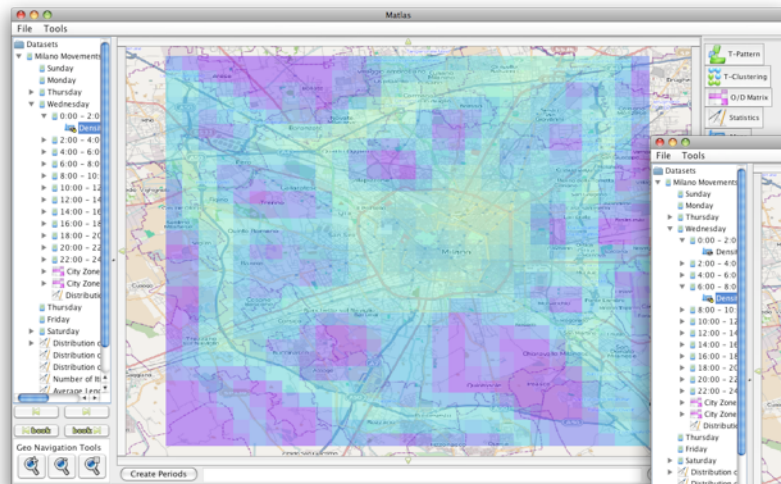
Diversity and Wellbeing (Mobility)



Big Data for smart cities



How people use the city during the day?



Fingerprint of the city

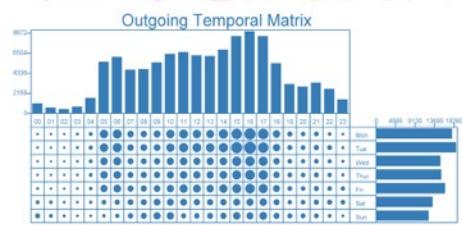
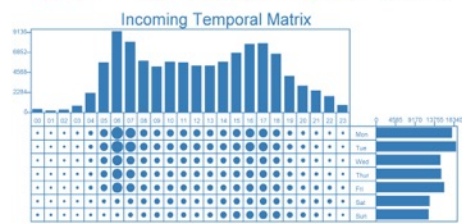
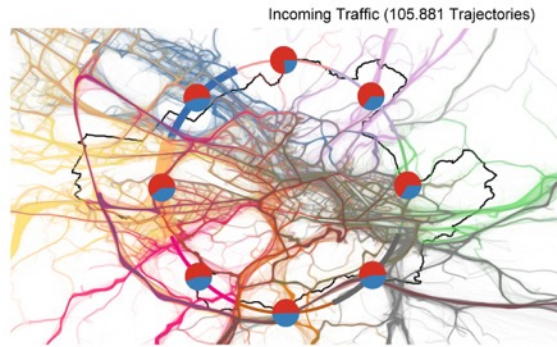
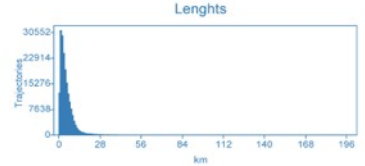
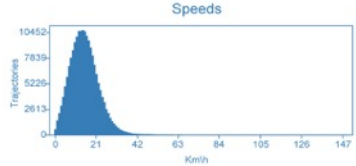
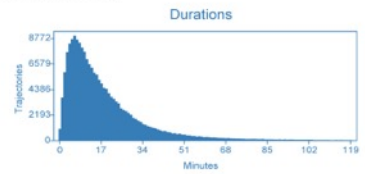
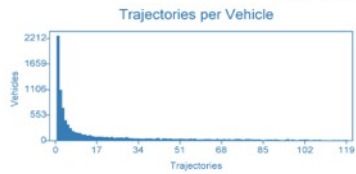
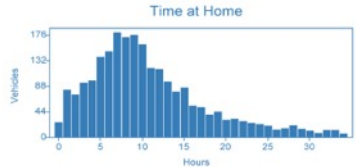
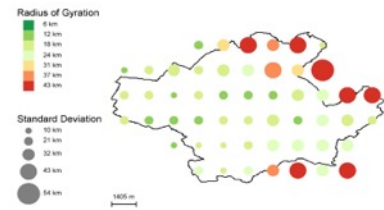
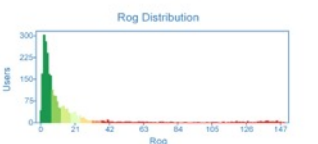
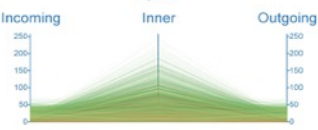
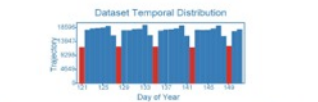
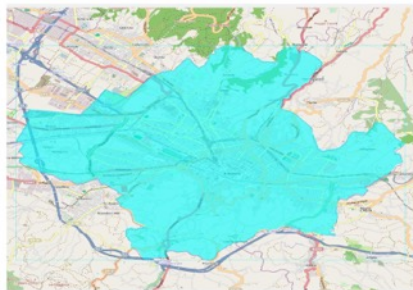
Firenze

Surface area: 106 km²

Coordinates: 43,78 11,24

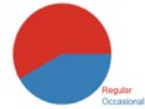
Vehicles: 32.752

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



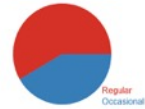
	City	Traj	Perc
NORD 13%	Sesto Fiorent.	9.123	95%
	Catanzaro	1.434	44%
	Viaglia	758	81%
	Campi Bisenzio	435	87%
	Borgo San Loro	380	48%
OVEST 50%	Scandico	13.447	98%
	Campi Bisenzio	8.058	92%
	Prato	8.048	94%
	Sesto Fiorent.	4.824	34%
	Lastra a Signe	2.342	95%
SUD 16%	Impruneta	3.863	87%
	San Casciano I.	1.838	75%
	Figline Valdar.	1.190	81%
	Greve in Chan.	868	36%
	Tavarnelle Val.	744	92%
EST 19%	Bagno a Ripol.	7.314	92%
	Firenze	3.970	95%
	Portofino	2.797	97%
	Greve in Chan.	1.516	92%
	Rignano sull'A.	774	92%

Regular VS Occasional

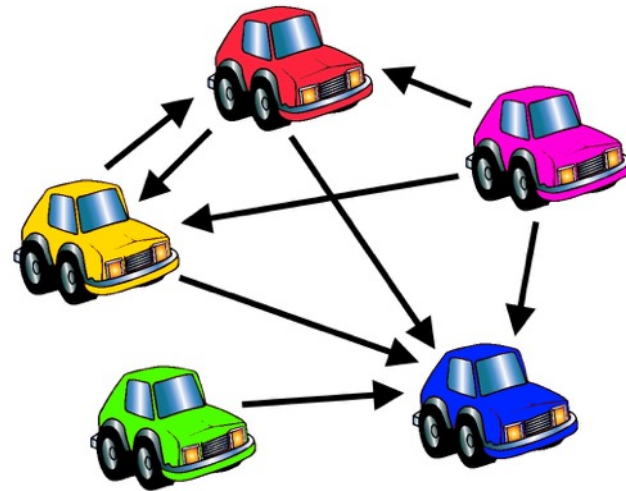
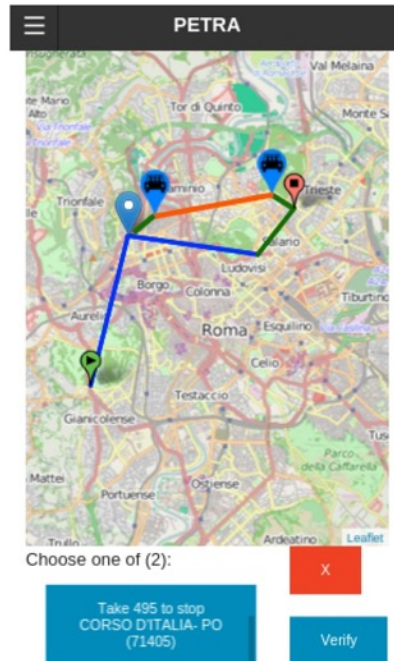
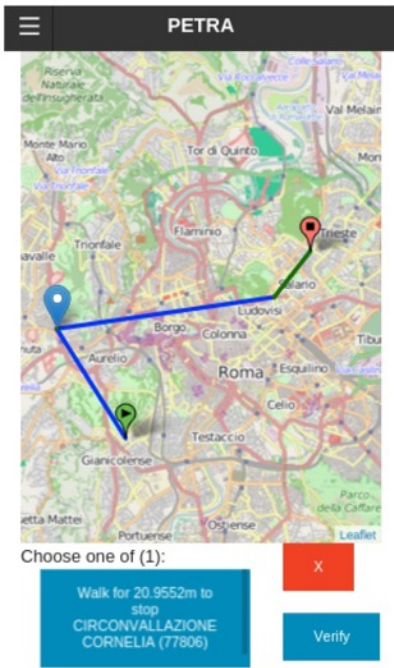


	City	Traj	Perc
NORD 12%	Sesto Fiorent.	8.058	95%
	Catanzaro	1.235	36%
	Viaglia	845	87%
	Campi Bisenzio	487	71%
	Borgo San Loro	458	54%
OVEST 52%	Scandico	13.439	98%
	Prato	6.168	95%
	Campi Bisenzio	5.846	92%
	Sesto Fiorent.	5.521	39%
	Lastra a Signe	2.423	98%
SUD 14%	Impruneta	3.965	95%
	San Casciano I.	1.801	72%
	Figline Valdar.	1.155	77%
	Tavarnelle Val.	742	92%
	Greve in Chan.	735	29%
EST 20%	Bagno a Ripol.	7.701	95%
	Firenze	3.682	94%
	Portofino	2.606	98%
	Greve in Chan.	1.670	87%
	Rignano sull'A.	818	96%

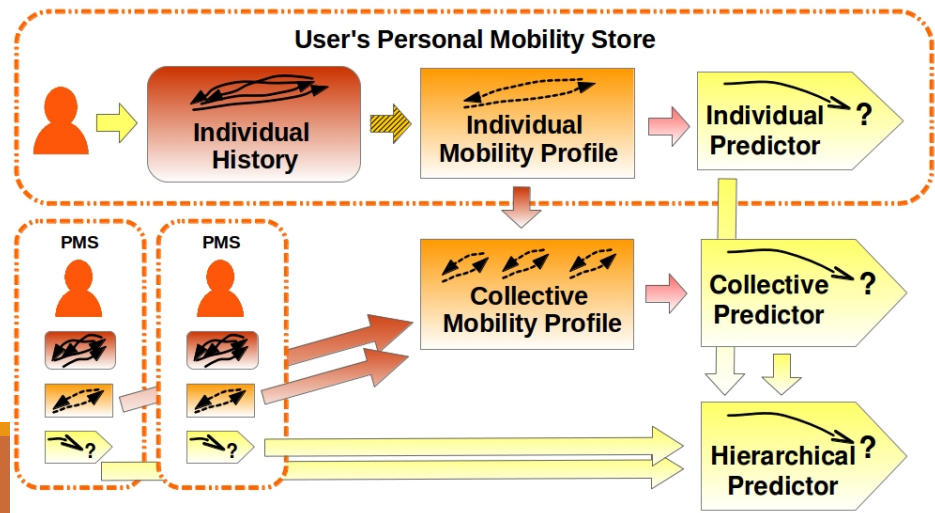
Regular VS Occasional



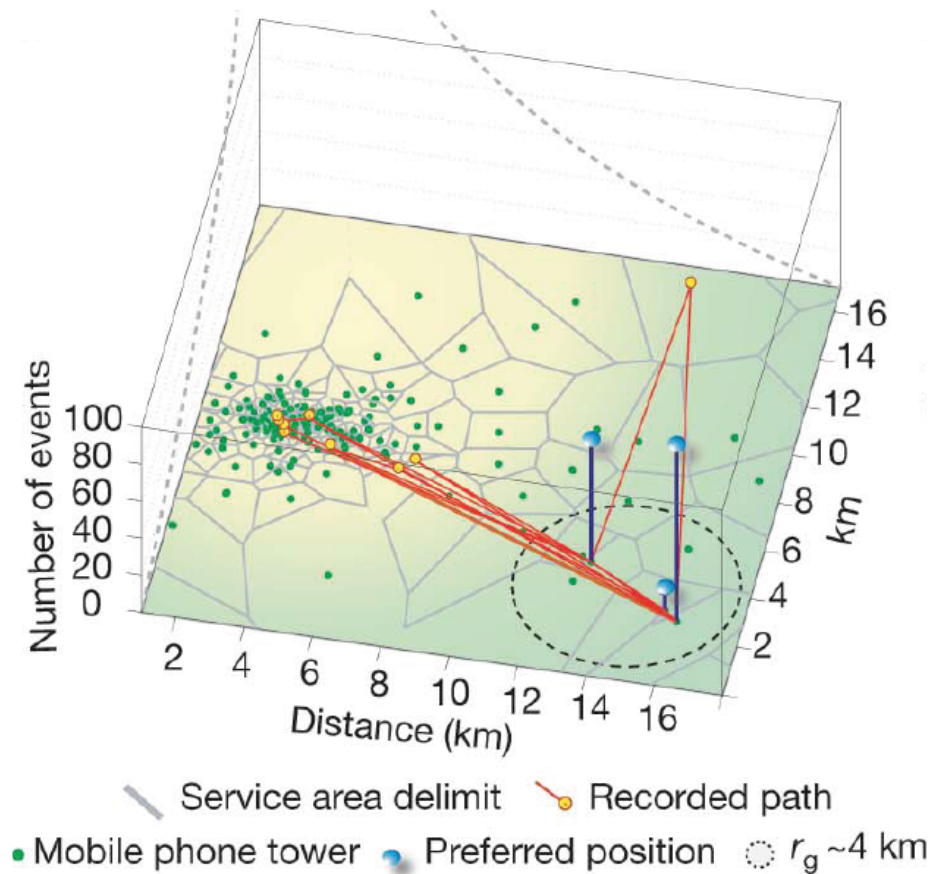
Personal mobility assistant



**Carpooling
Network**



Call Data Records... when, where and who



when
you
call

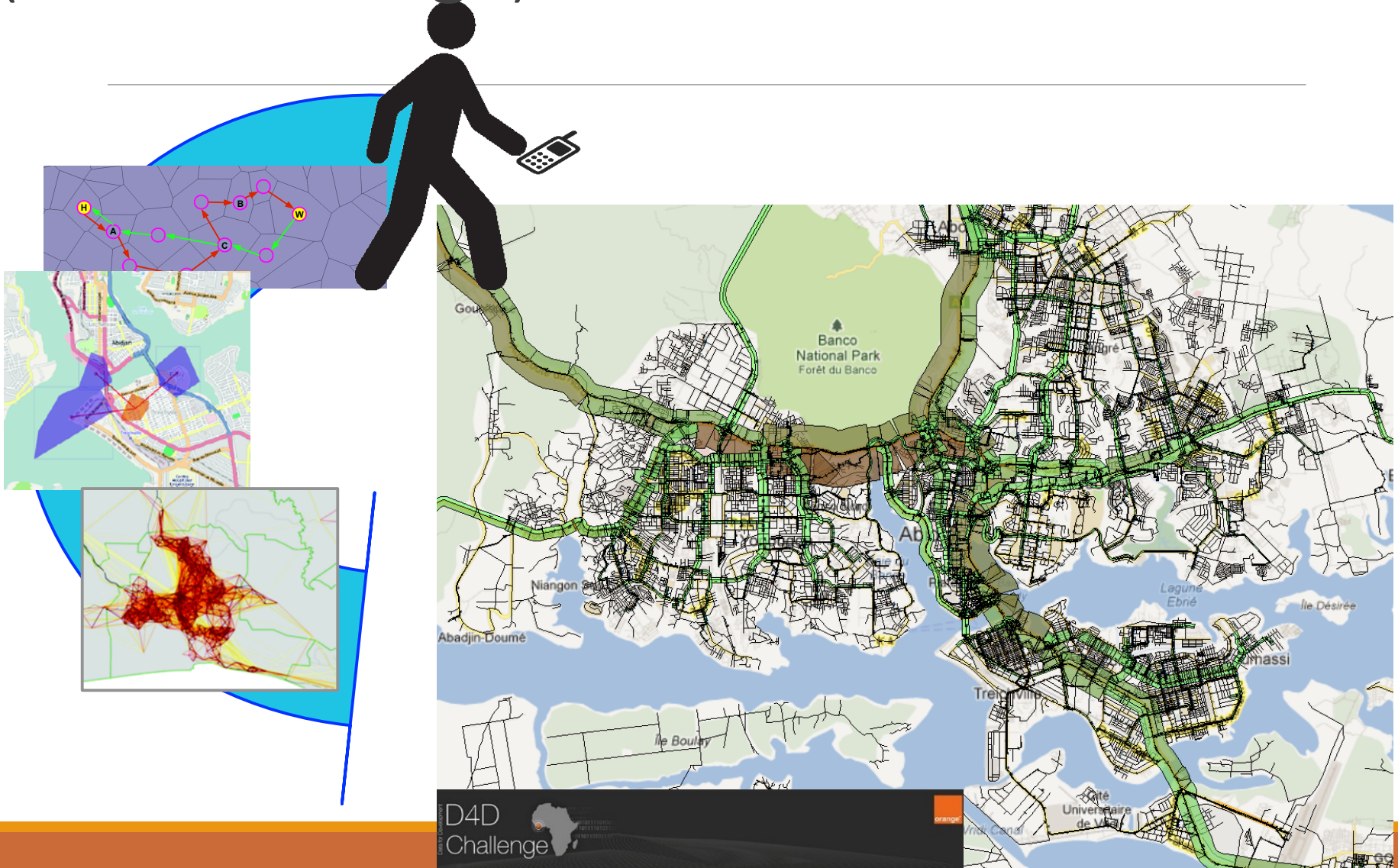


where
you
call

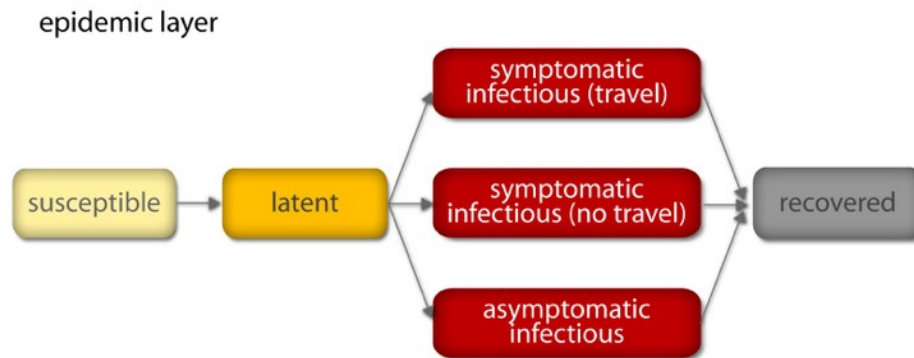
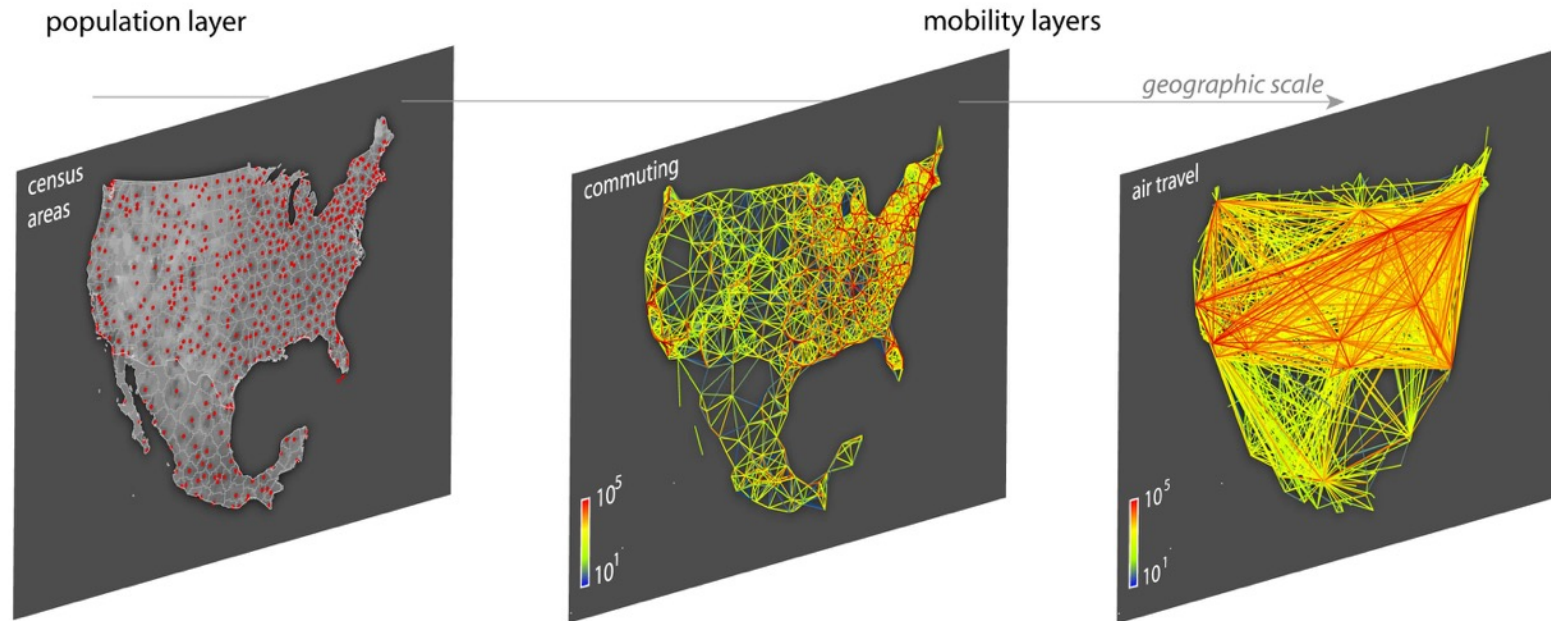


who
you
call

Call Data Records for Developing Countries (D4D Challenge)

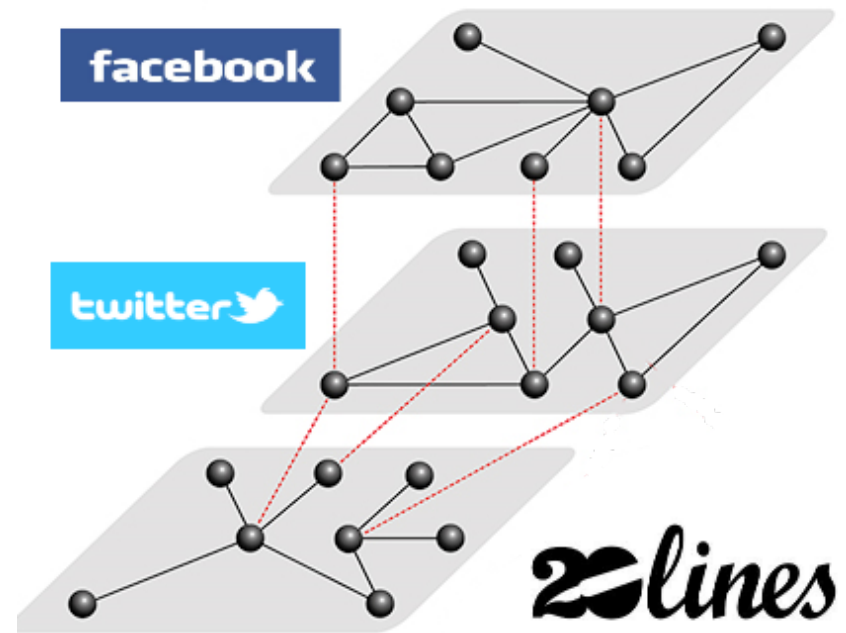
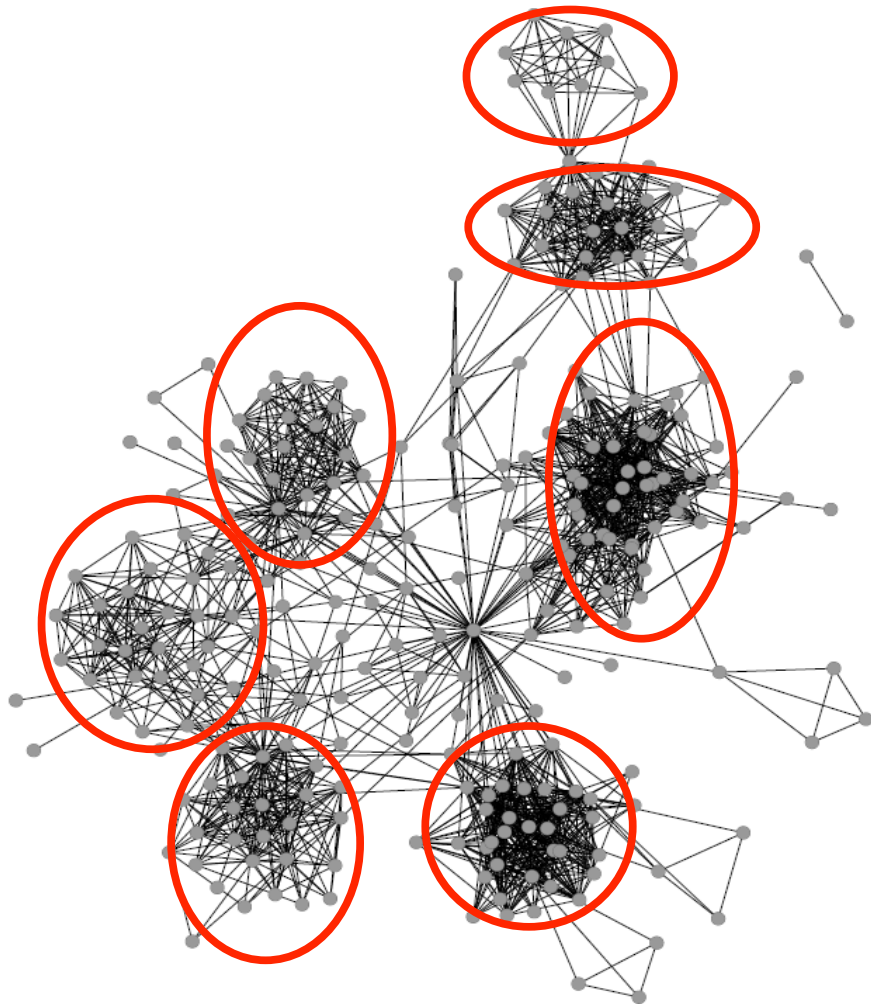


Epidemics simulations

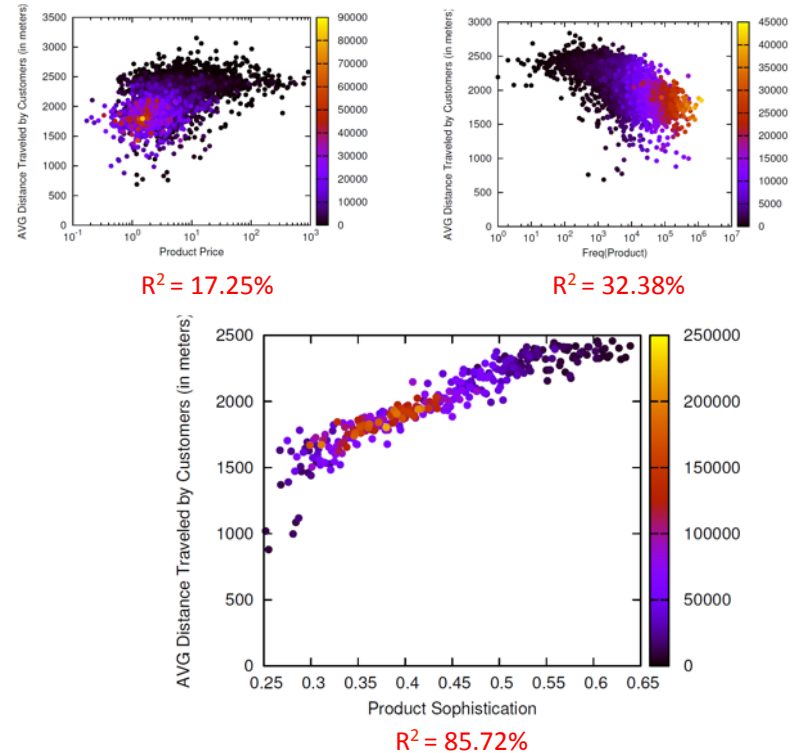


Parameter	Value	Description
β	from R_0	transmission probability
ε^{-1}	1.9 [1.1-2.5] d	average latency period
μ^{-1}	3 [3-5] d	average infectious period
p_t	50%	probability of traveling for infectious individuals
p_a	33%	probability of being asymptomatic
r_β	50%	relative infectiousness of asymptomatic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

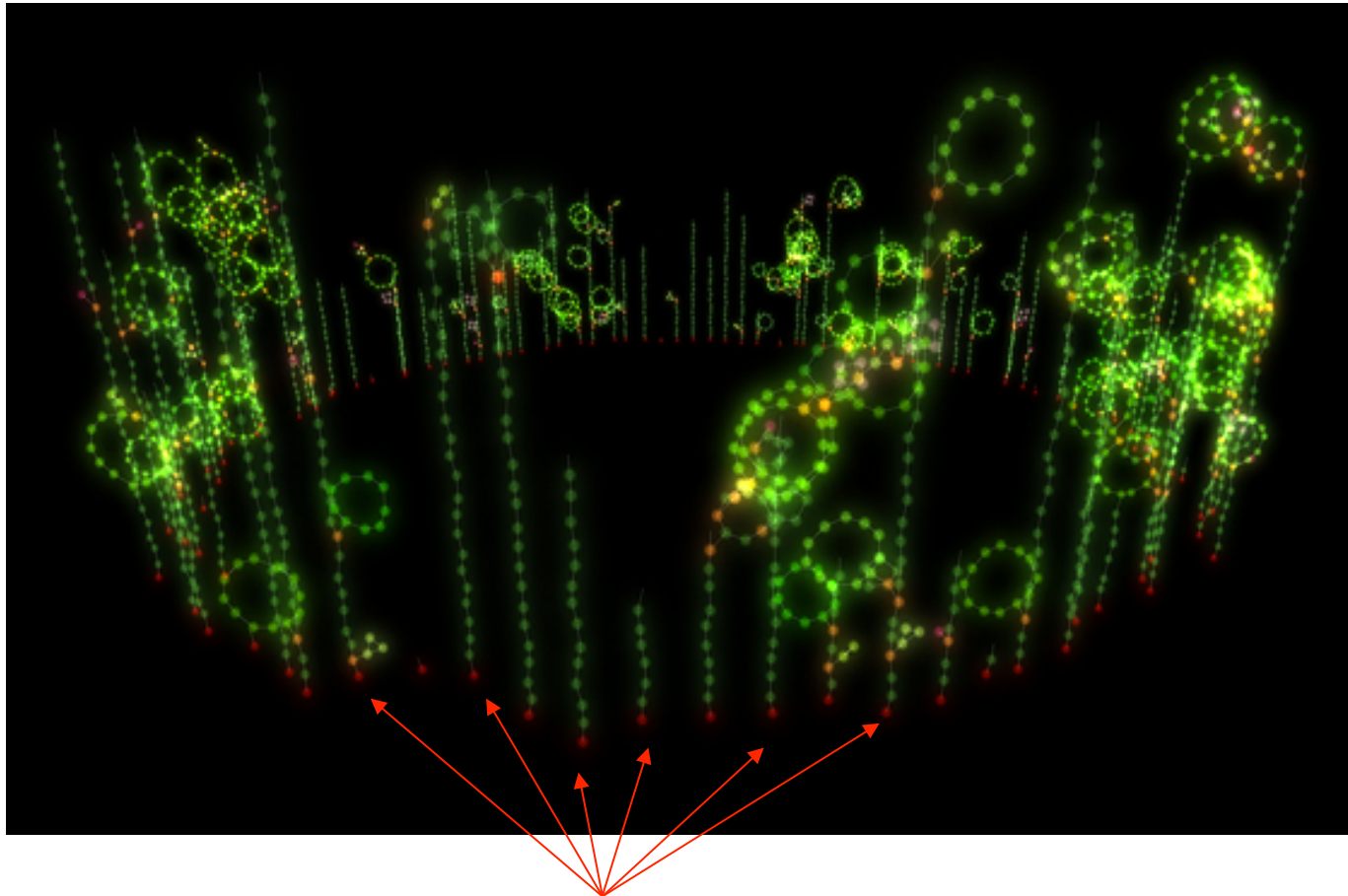
Community Discovery, Evolution, Diffusion, Multidimensionality,...



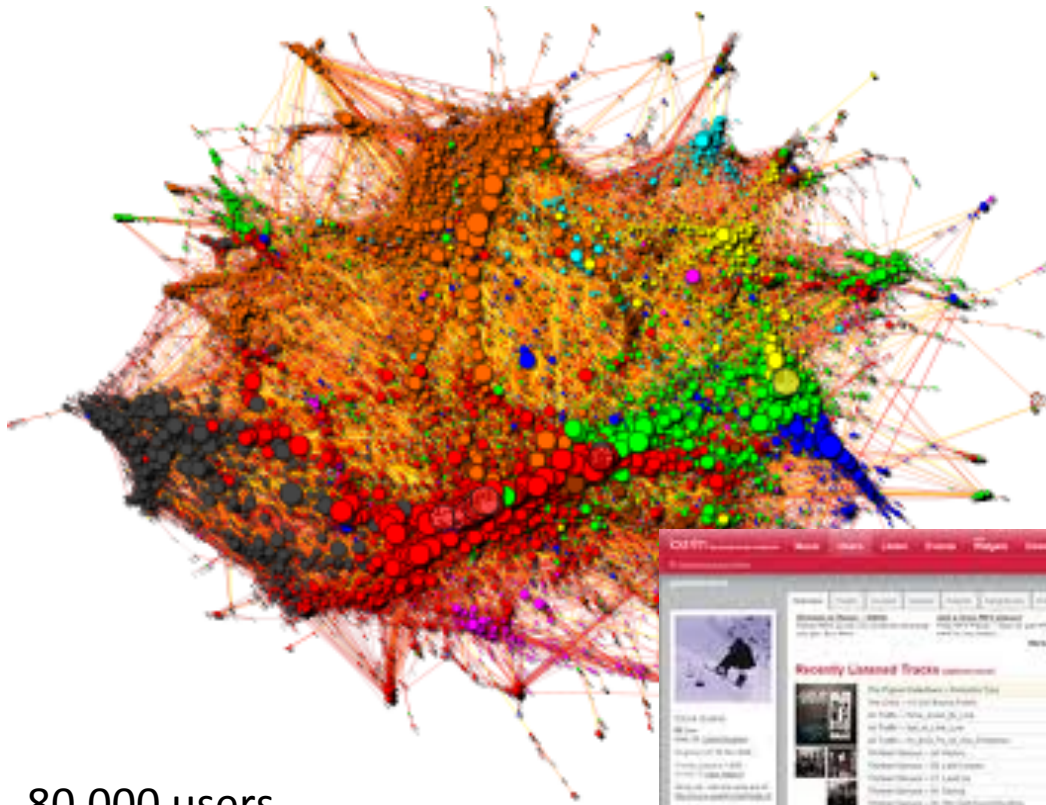
Retail Market as Complex system



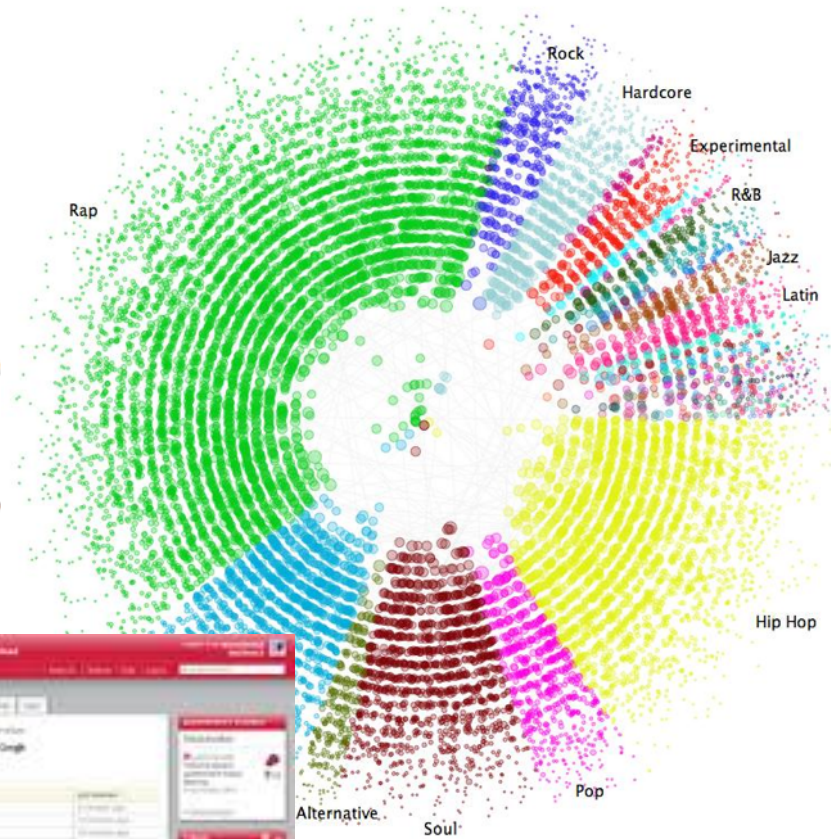
Social Influence: Leaders



Ask to LAST.FM



80.000 users,
4000.000 connections

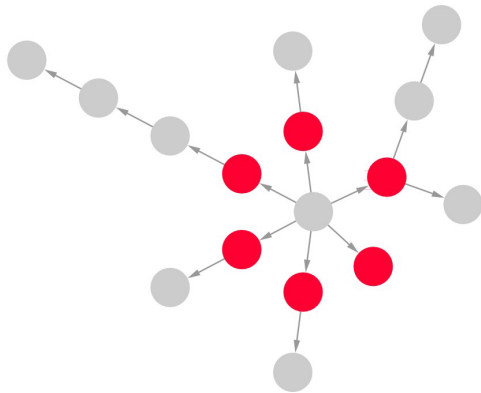


What is Social Prominence?

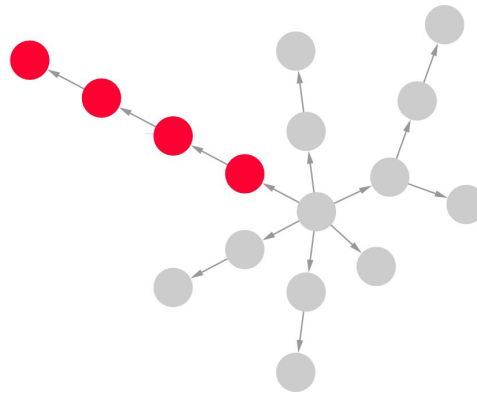
It has been observed that a small set of users in a Social Network is able to anticipate (or influence) the behavior of the entire network

We detected 3 possible scenarios:

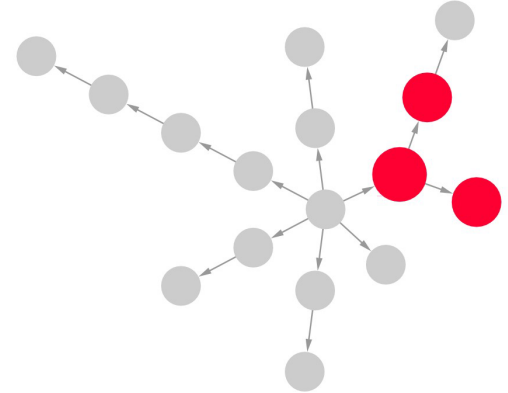
width



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**:

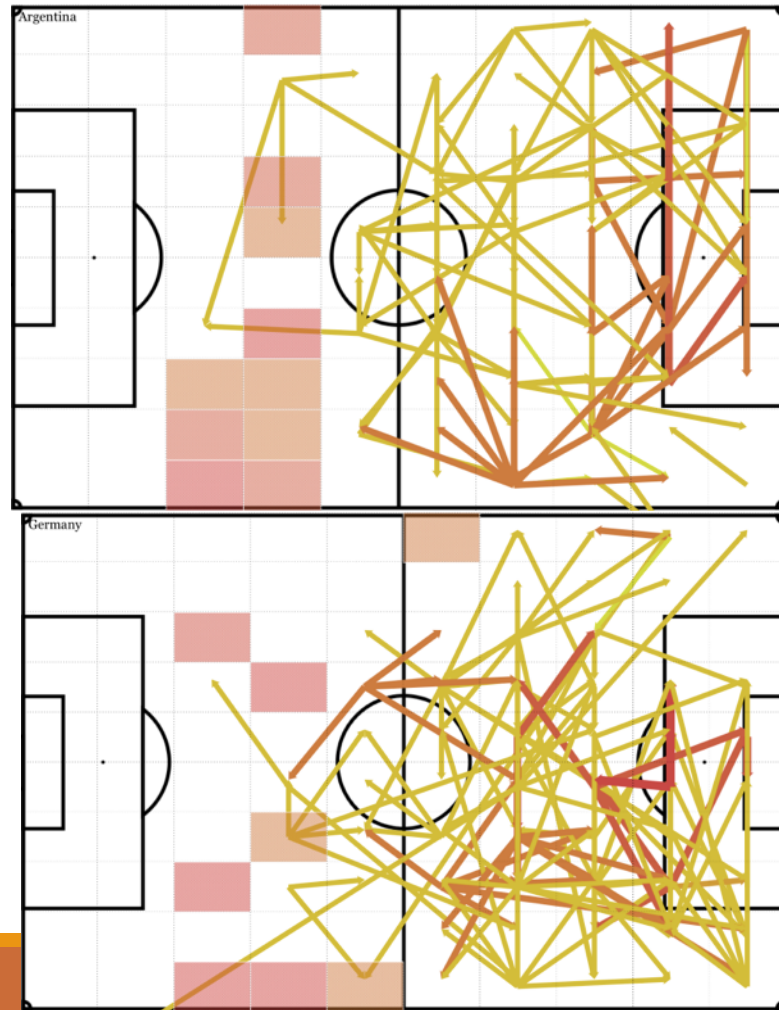
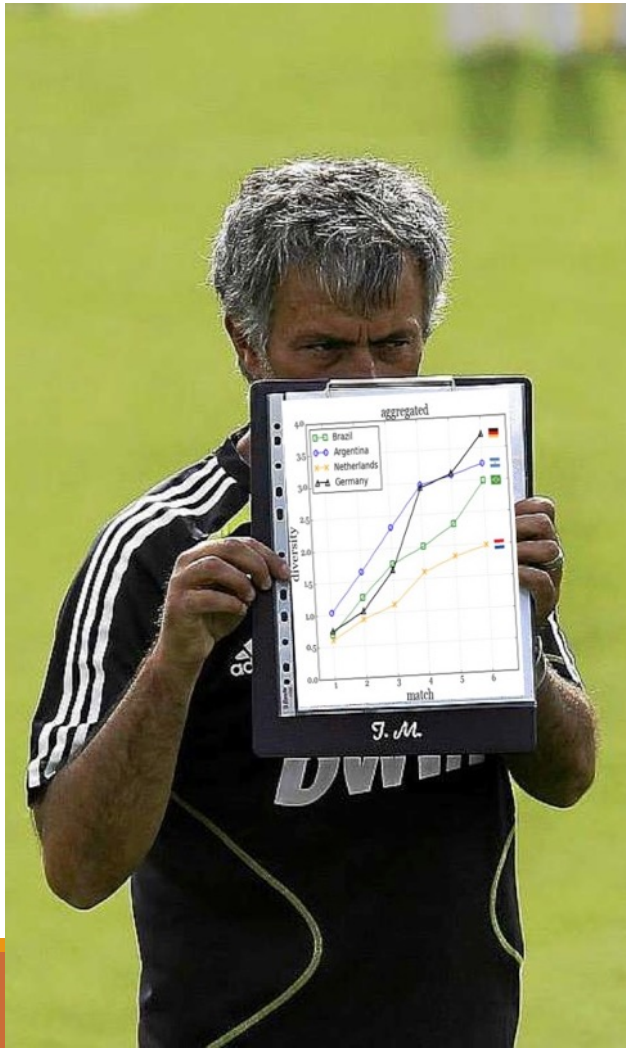
- data from Strava.com
- How you train is fundamental
- A confirmation of the “overcompensation” theory



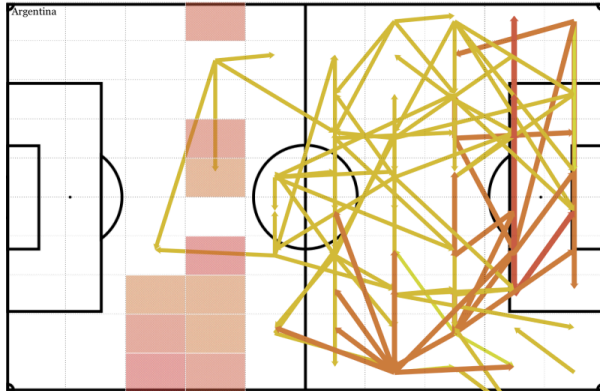
The patterns of success in Sports

“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)

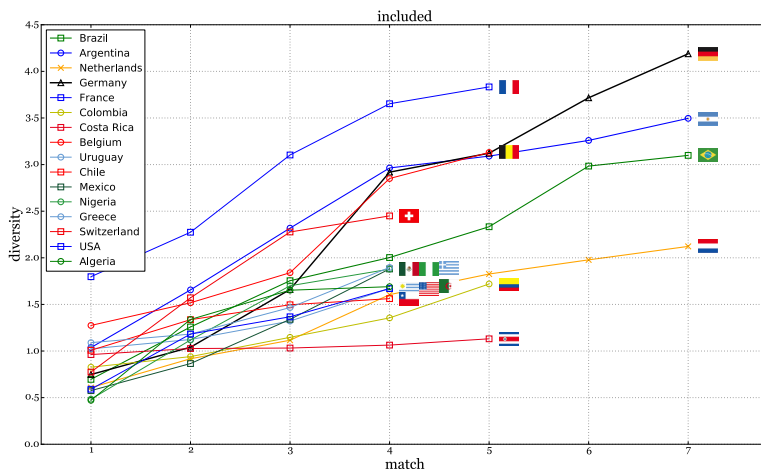


Big Data: the way of **Success**



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



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



Data from Opta:

All events during the match

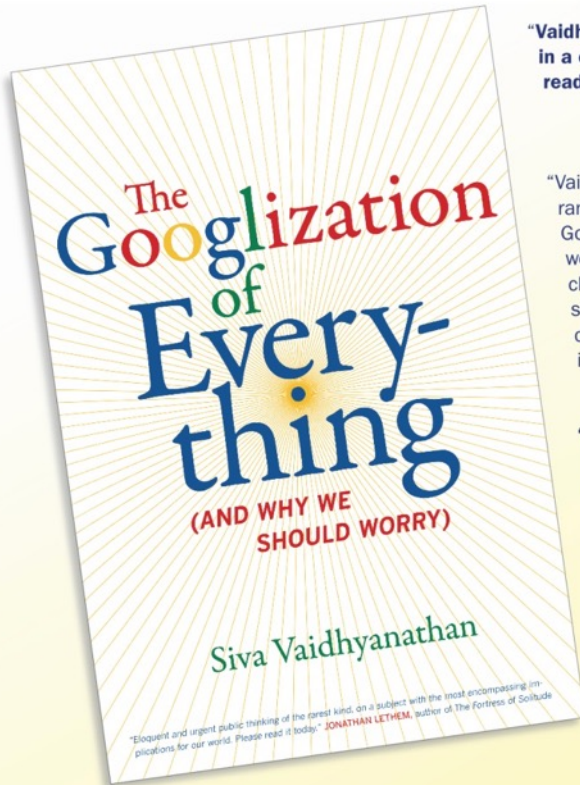
```
...  
<tackle,15.4,41.1,112>  
<pass,25.0,67.1,113>  
<pass,65.0,87.1,115>  
<assist,82.1,35.8,120>  
<goal attempt,82.1,35.8,121>  
.....
```


Big Data Analytics & Social Mining



**“Finely written and engaging....
A book for anyone who has used Google.”**

—Toby Miller, author of *Makeover Nation*



“Vaidhyanathan is everything you could want in a cultural critic: funny, fantastically readable, and insightful as hell.”

—Cory Doctorow, author of *For the Win* and co-editor of *Boing Boing*

“Vaidhyanathan’s lively, thoughtful, and wide-ranging book makes clear, in detail, how Google is reshaping the way we live and work. He finds much to admire, but also challenges us to not only use Google’s services, but to go beyond them to create a new and genuinely democratic information order.”

—Anthony Grafton, author of *Codex in Crisis*

“Thoughtfully examines the insiders influence of Google on our society.... As Vaidhyanathan points out, we must be cautious about embracing Google’s mission and not accept uncritically that Google has our best interests in mind.”

—Publishers Weekly, Starred Review

“A critically important book because it’s really about the Googlization of All of Us.... A brilliant meditation on technology, information, and consumer inertia, as well as an ambitious challenge to change how, where, why, and what we Google.”

—Dahlia Lithwick, senior editor and writer, *Slate Magazine*



At bookstores or www.ucpress.edu/go/googlization

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

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



UNIVERSITY OF CALIFORNIA PRESS

\$300 billion

potential annual value to US health care—more than double the total annual health care spending in Spain

€250 billion

potential annual value to Europe's public sector administration—more than GDP of Greece

\$600 billion

potential annual consumer surplus from using personal location data globally

McKinsey Global Institute



60% potential increase in
retailers' operating margins
possible with big data

140,000–190,000

more deep analytical talent positions, and

1.5 million
more data-savvy managers
needed to take full advantage
of big data in the United States

McKinsey Global Institute





ARTICLE PREVIEW To read the full article, **sign-in** or **register**. HBR subscribers, click **here** to
for **FREE** access »

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Today

"Data Scientist" Job Trends

"Data Scientist" x

+ Add Term

Find Trends

Scale: Absolute | Relative

Job Postings



Data scientist



... a new kind of professional has emerged, the **data scientist**, who combines the skills of **software programmer, statistician** and **storyteller/artist** to extract the nuggets of gold hidden under mountains of data.



Kashmir Hill, Forbes Staff
Welcome to The Not-So Private Parts where technology & privacy collide
[+ Follow](#) (1,410) [Follow](#) 216k

TECH | 2/16/2012 @ 11:02AM | 2,106,633 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

318 comments, 169 called-out [+ Comment Now](#) [+ Follow Comments](#)

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. [Target](#), for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the [New York Times](#) how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel,



Target has got you in its aim

Most Read on Forbes

NEWS [People](#) [Places](#) [Companies](#)

Why McDonald's Employee Budget Has Everyone Up In Arms +154,303 views


The NY Times Tries -- And Fails -- To Protect Obamacare From Health Insurance 'Rate Shock' +94,764 views

If You Haven't Seen A Windows Phone Lately, It's Because They're Practically Disappearing +52,878 views


Nintendo Surprises Fans With 'Earthbound' For Wii U, Out Today +45,418 views

Why Are Walmart Stores Such A

Predicting who could be persuaded and



MIKE GUALTIERI'S BLOG



Forrester Blogs > Business Technology > Application Development & Delivery Professionals > Mike Gualtieri

HOW THE OBAMA CAMPAIGN USED PREDICTIVE ANALYTICS TO INFLUENCE VOTERS

Posted by [Mike Gualtieri](#) on June 27, 2013

58 Recommendations


Print

Email

0 comments

Tweet 35

The Obama 2012 campaign famously used [big data predictive analytics](#) to influence individual voters. They hired more than 50 analytics experts, including [data scientists](#), to predict which voters will be positively persuaded by political campaign contact such as a call, door knock, flyer, or TV ad. Uplift modeling (aka persuasion modeling) is one of the hottest forms of predictive analytics, for obvious reasons — most organizations wish to persuade people to do something such as buy! In this special episode of Forrester TechnoPolitics, [Mike](#) interviews Eric Siegel, Ph.D., author of [Predictive Analytics](#), to find out: 1) What exactly is uplift modeling? and 2) How did the Obama 2012 campaign use it to persuade voters? (< 4 minutes)



Why should you develop a mobile-first strategy?

Watch the webinar **The Way We Develop Is Changing** with analyst Michael Facemire

Are you extracting value from big data?

Listen to the webinar **Big Data — Gold Rush Or Illusion?** with Holger Kisker and Martha Bennett

Are your mobile apps ready for customer demand?

Download **the first report** from the Mobile App Development Playbook

Start

Vision papers

F Giannotti, D Pedreschi, A Pentland, P Lukowicz, D Kossmann, J Crowley, D Helbing. **A planetary nervous system for social mining and collective awareness.** The European Physical Journal Special Topics 214 (1), 49-75, 2012

M Batty, KW Axhausen, F Giannotti, A Pozdnoukhov, A Bazzani, M Wachowicz. **Smart cities of the future.** The European Physical Journal Special Topics 214 (1), 481-518, 2012

Editors: Mirco Nanni, Costantino Thanos, Fosca Giannotti and Andreas Rauber, **Big Data Analytics: towards a European research agenda**, ERCIM White paper,
<http://www.ercim.eu/images/stories/pub/white-paper-BigDataAnalytics.pdf>

G7 Academies Meeting - Rome, 23-25 March 2017
Joint Statement on New economic growth: the role of science, technology, innovation and infrastructure, Position Paper on Data Science by Fabio Beltram, Fosca Giannotti, Dino Pedreschi