Big Data Analytics

FOSCA GIANNOTTI AND LUCA PAPPALARDO

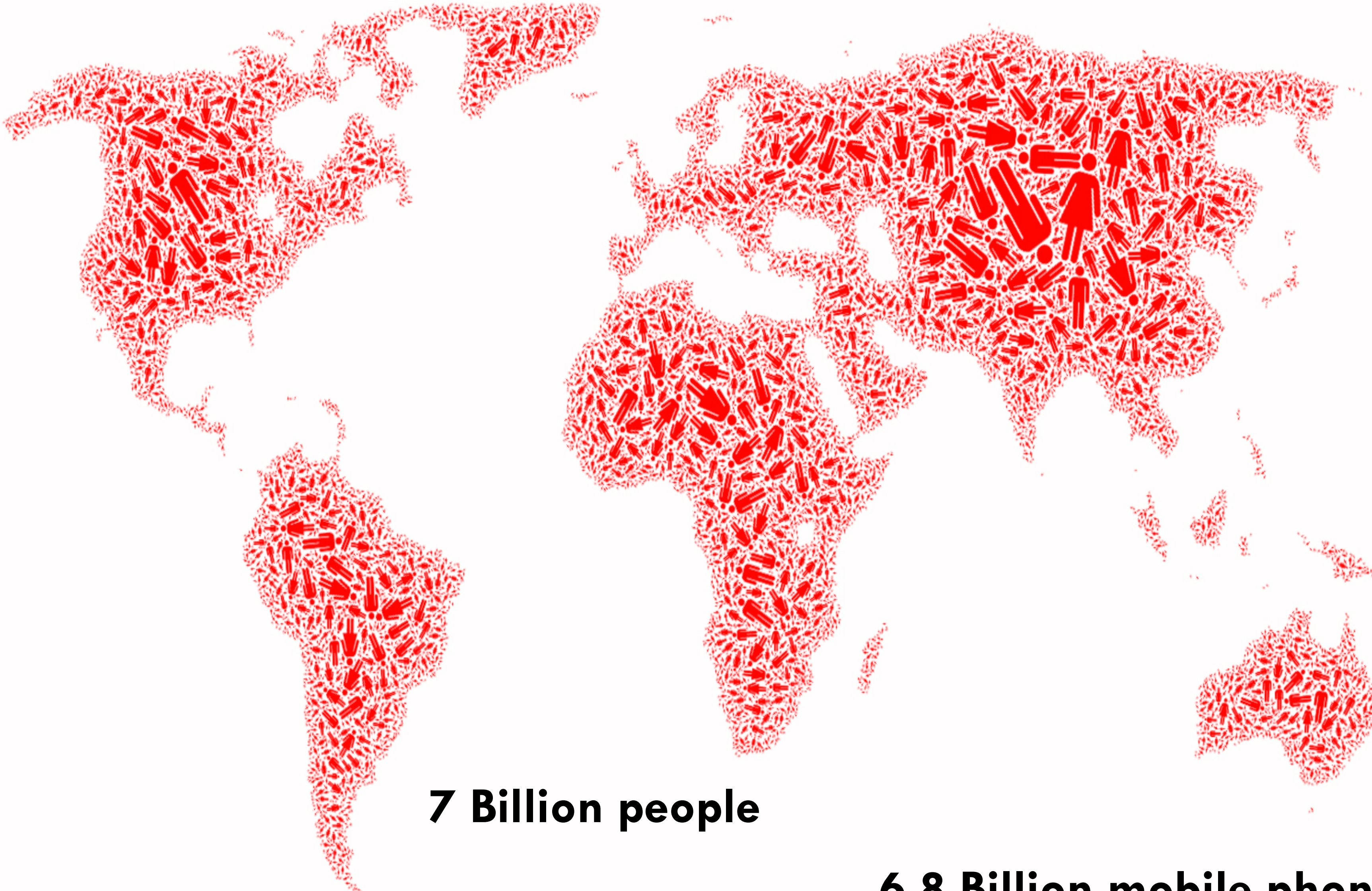
HTTP://DIDAWIKI.DI.UNIPI.IT/DOKU.PHP/BIGDATAANALYTICS/BDA/

DIPARTIMENTO DI INFORMATICA - Università di Pisa
anno accademico 2018/2019

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.


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



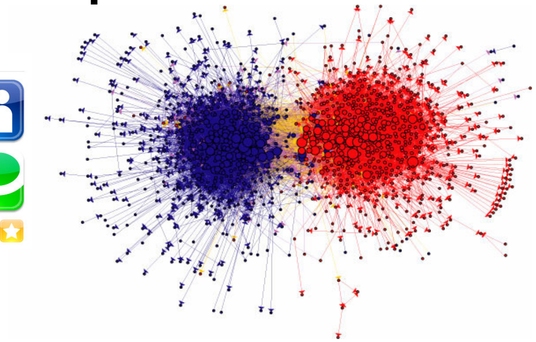
7 Billion people

6.8 Billion mobile phones

Digital Footprints of Human Activities

Shopping patterns & lifestyle

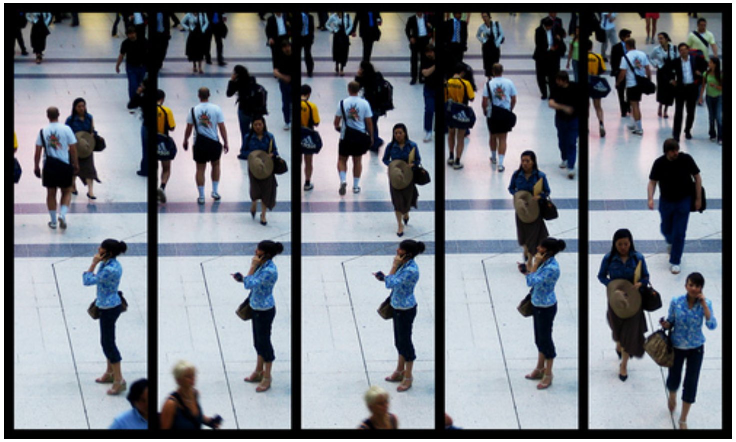
Relationships & social ties

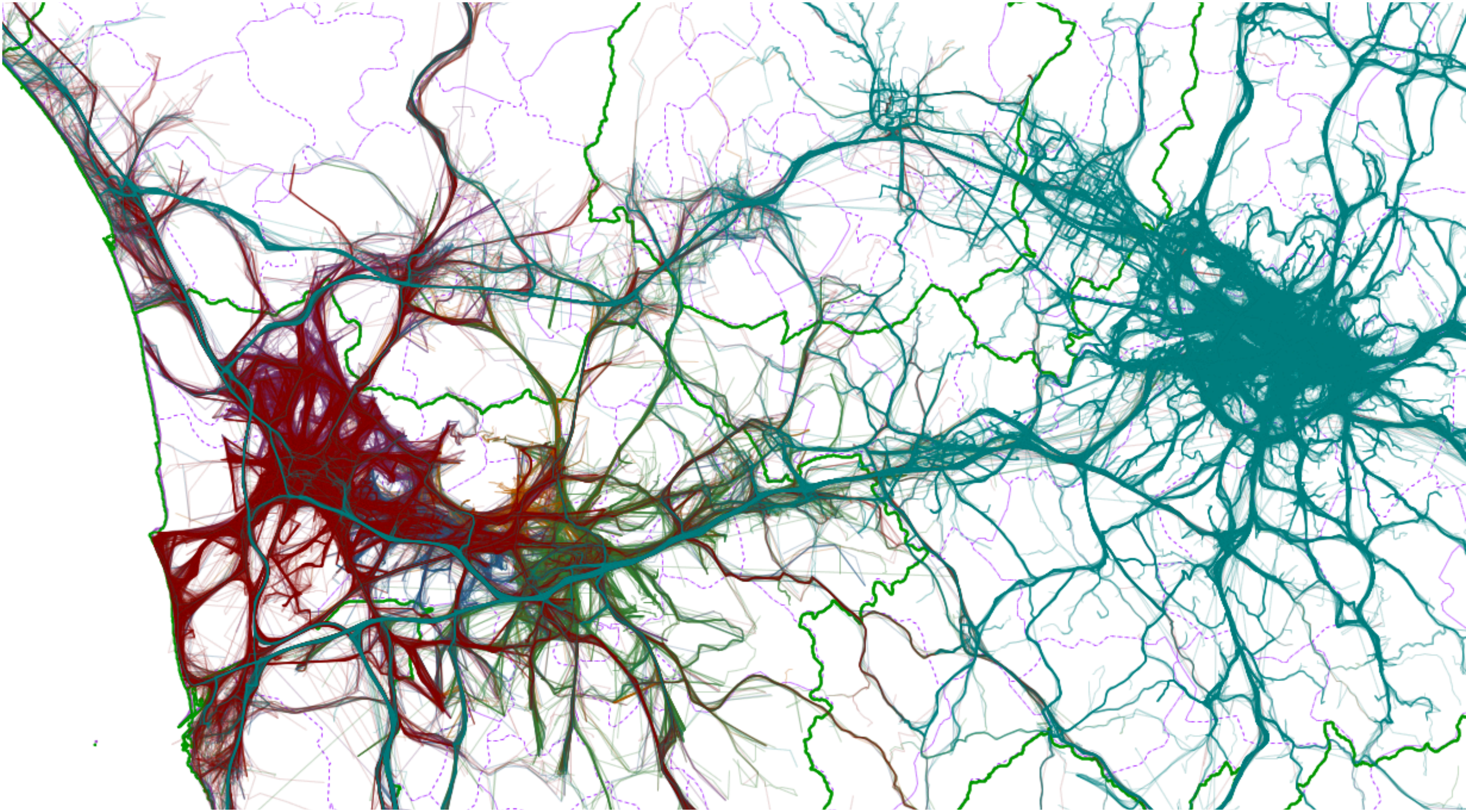


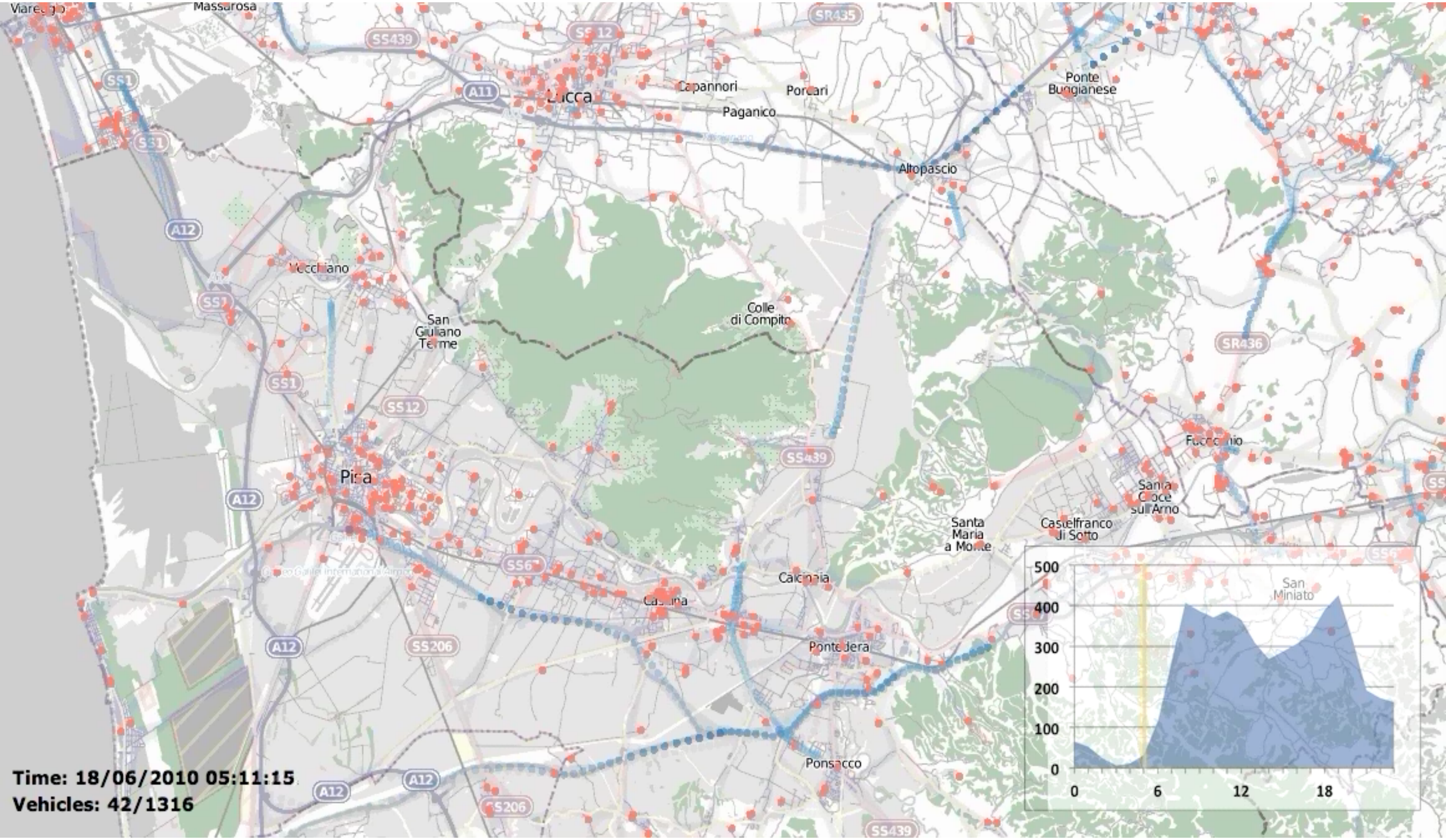
sires, opinions, sentiments

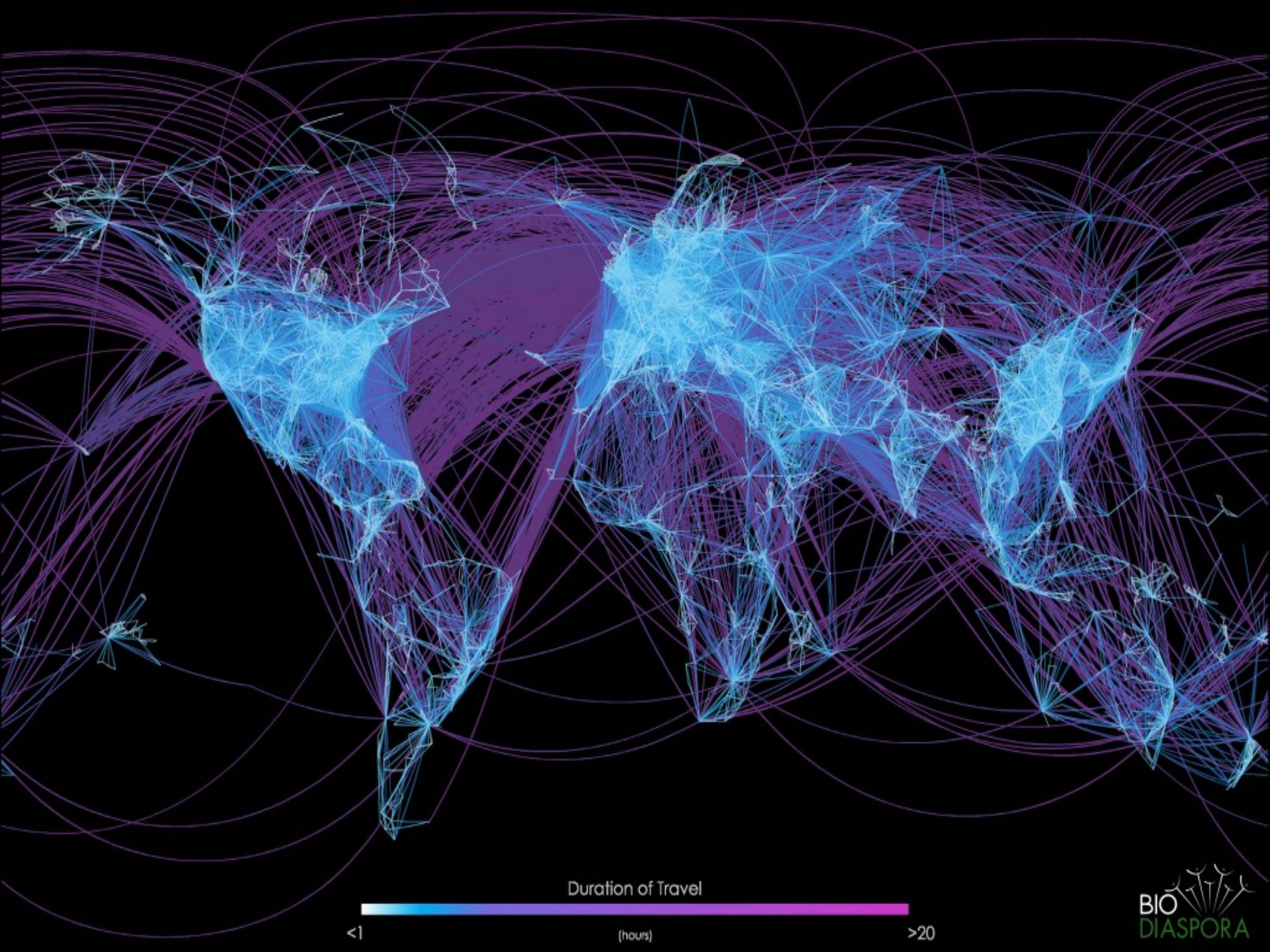
WIKIPEDIA
The Free Encyclopedia

Movements





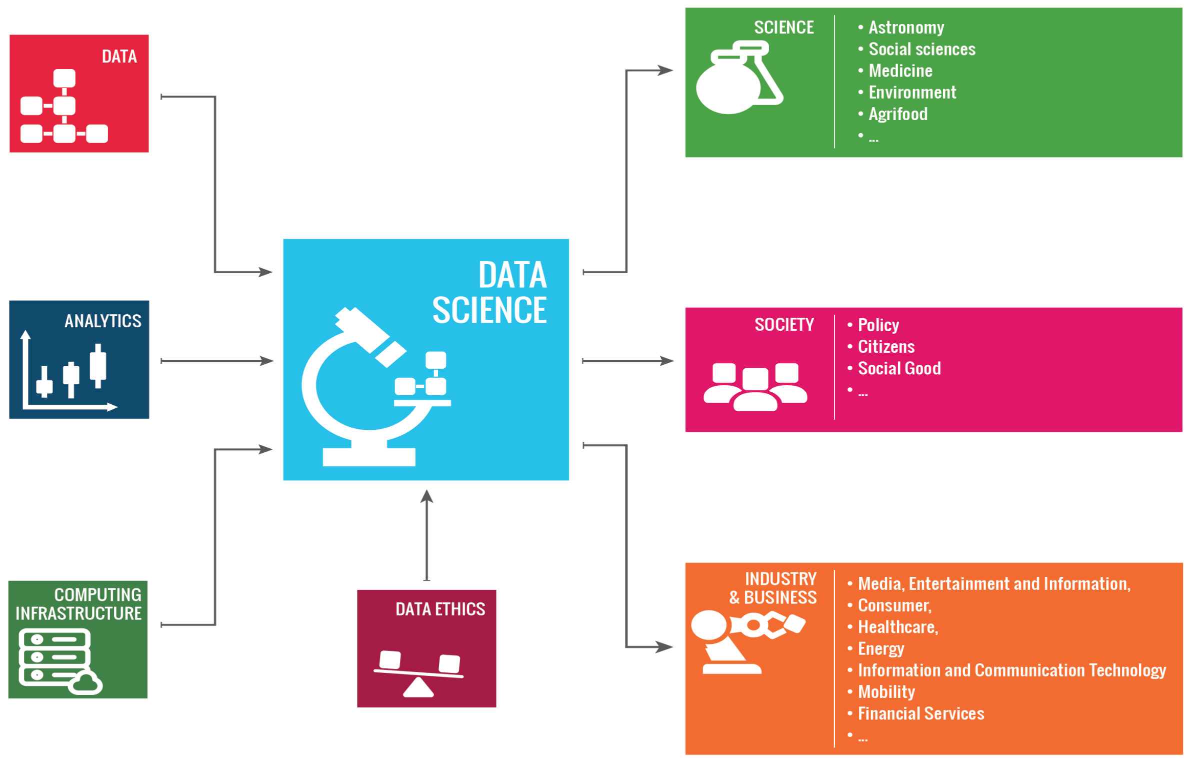

Massarosa
SS439
SS12
SR435
A11
Capannori
Porcari
Paganico
Ponte Buggianese
Altopascio
SS1
A12
San Giuliano Terme
Colle di Compito
SR436
SS439
Santa Maria a Monte
Castelfranco di Sotto
Santa Croce sull'Arno
Calcinaia
SS206
Pontedera
Ponsacco
500
400
300
200
100
0
0
6
12
18
San Miniato
Time: 18/06/2010 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

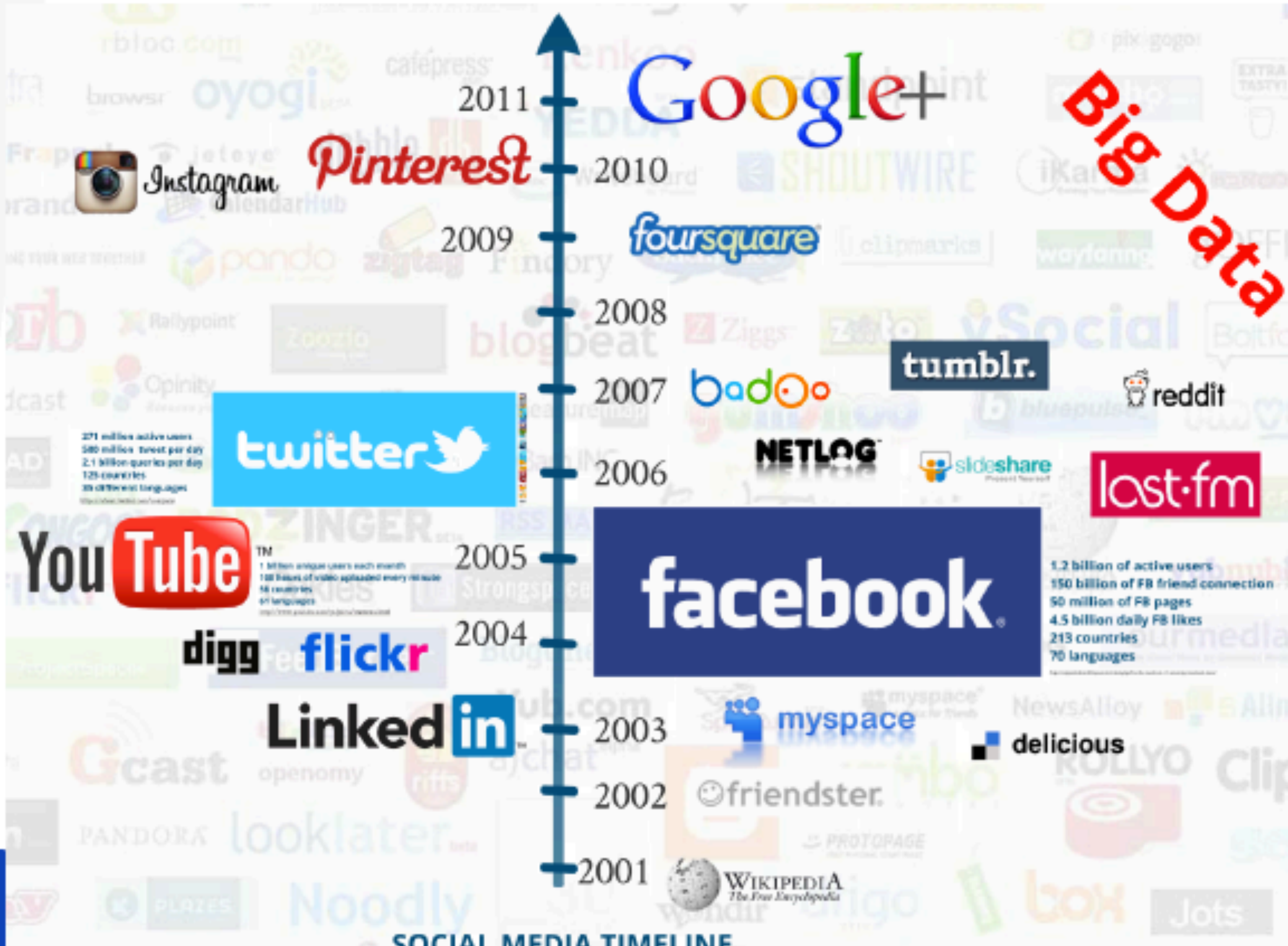
data availability, sophisticated analysis techniques, and scalable infrastructures brought what we call today "Data Science"

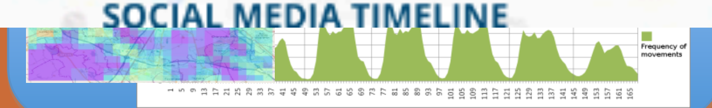


- "**Data Science and BigData: a Game-changer for Science and Innovation**" Document for G7 Academy, March 2017,
- "**Realizing our Digital Future and shaping its impact on Knowledge, Industry, and WorkForce** Document for G7 Academy, March 2018:

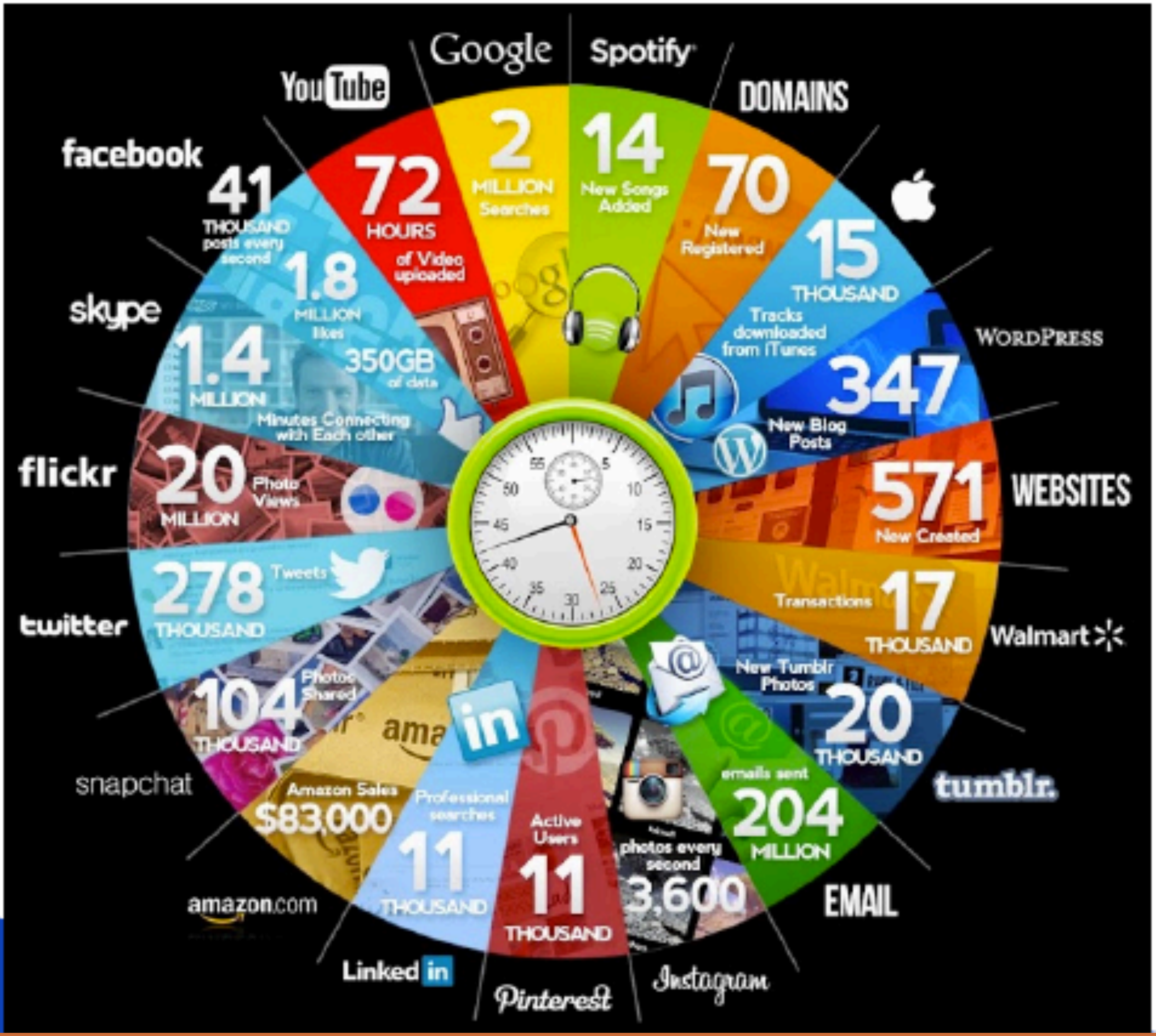
Big Data Number

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)

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

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.

Value

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

WORLD ECONOMIC FORUM

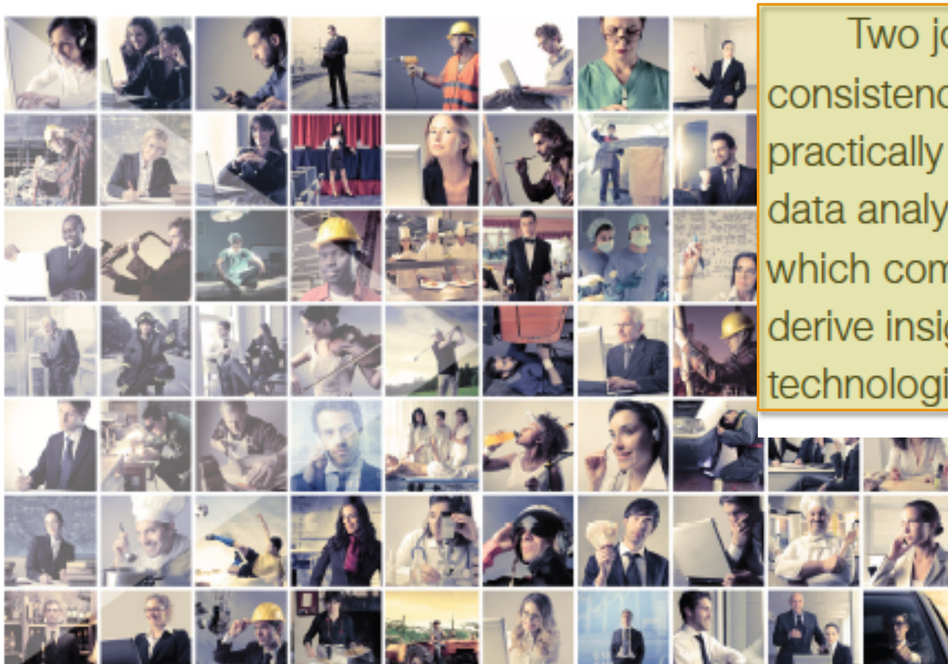
COMMITTED TO IMPROVING THE STATE OF THE WORLD

Global Challenge Insight Report

The Future of Jobs

Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution

January 2016



New and Emerging Roles

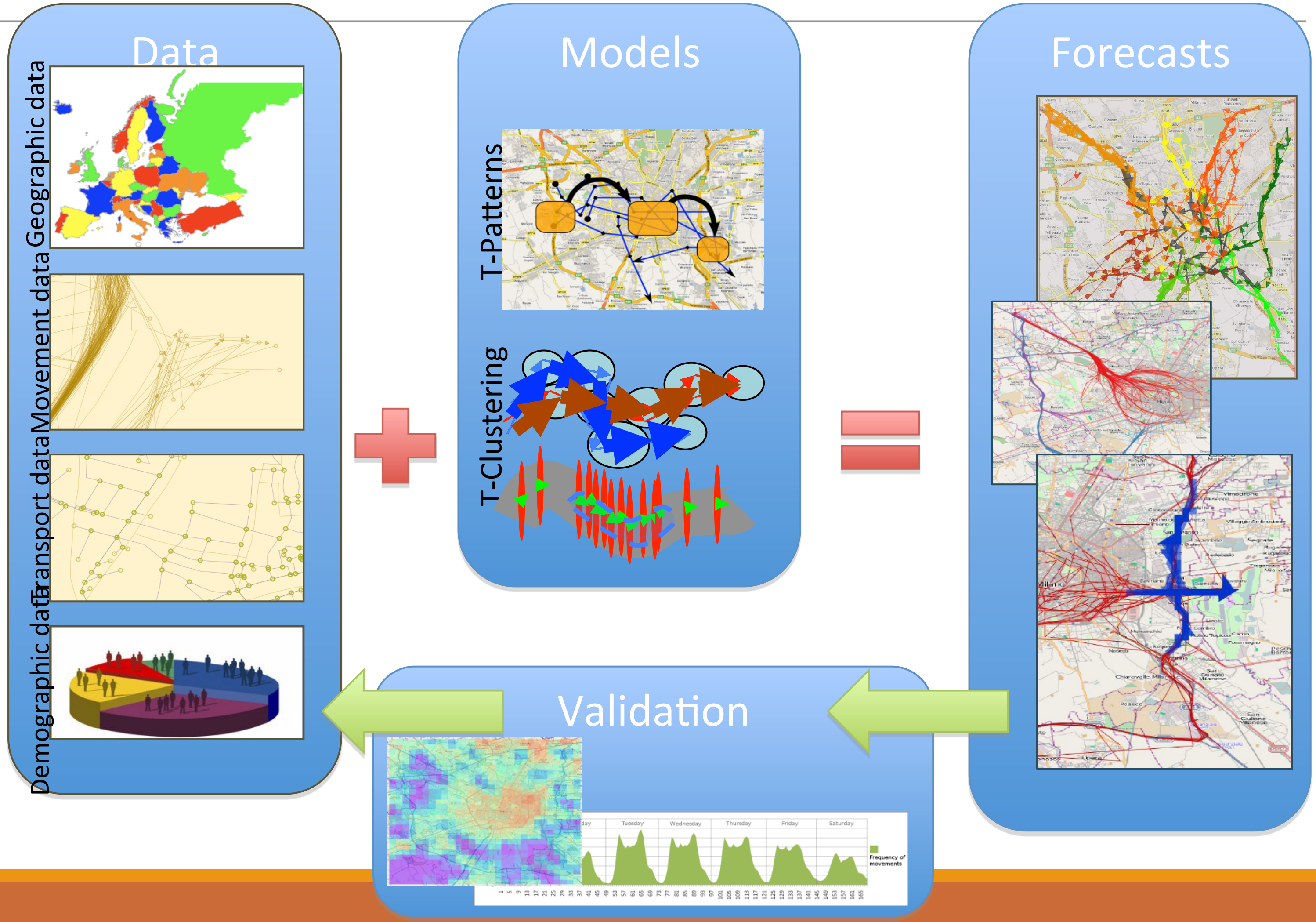
Our research also explicitly asked respondents about new and emerging job categories and functions that they expect to become critically important to their industry by the year 2020, and where within their global operations they would expect to locate such roles.

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

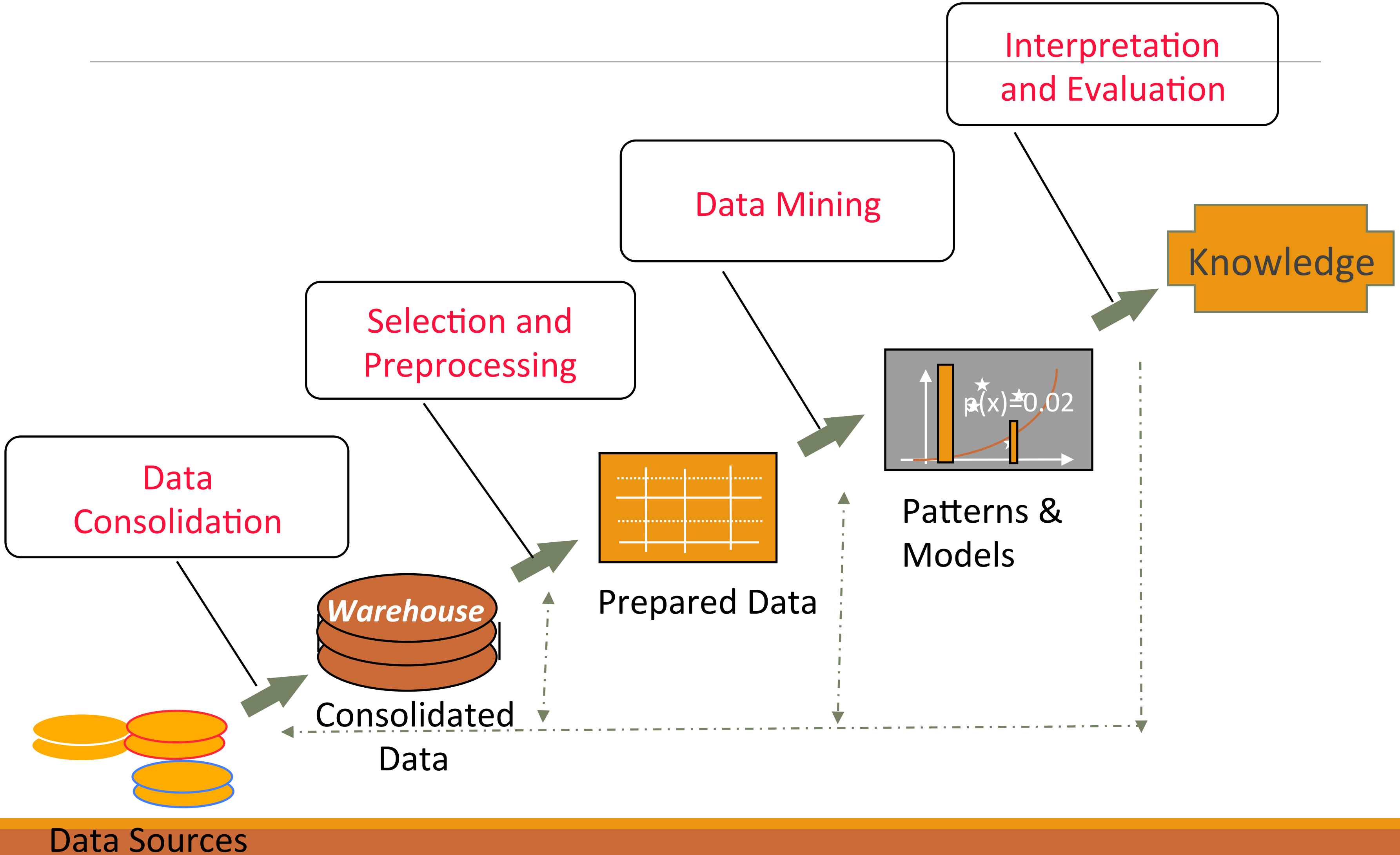
http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf

How to develop a big data analytics project

From DATA to KNOWLEDGE



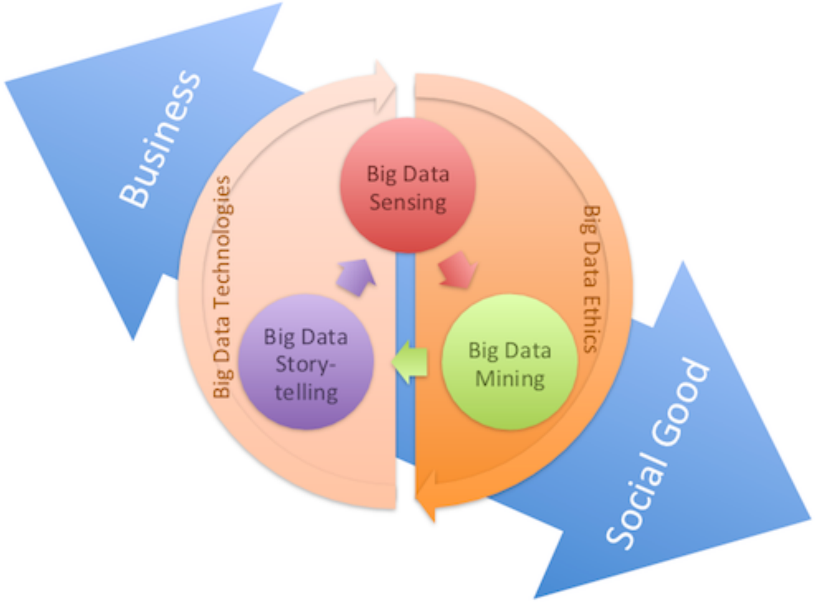
The KDD process



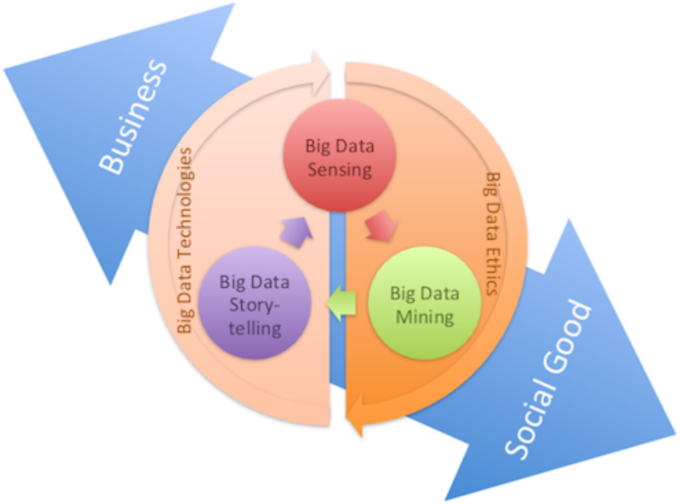
CRISP Model

Business Understanding
Data Understanding
Data Preparation
Deployment
Data
Modeling
Evaluation

The modern data scientist!!!



Big Data Sensing and Procurement

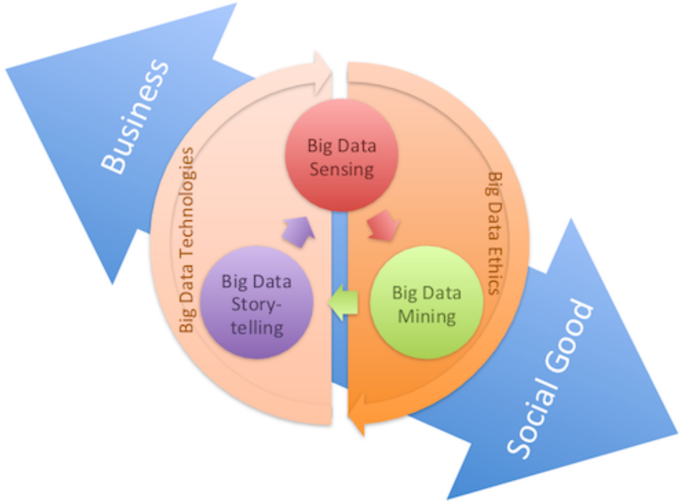


Big data sources, crowdsourcing, crowdsensing

Web Search Engines and Information Retrieval

Analytical Crawling, Text Annotation

Big Data Mining



Data Mining & Machine Learning

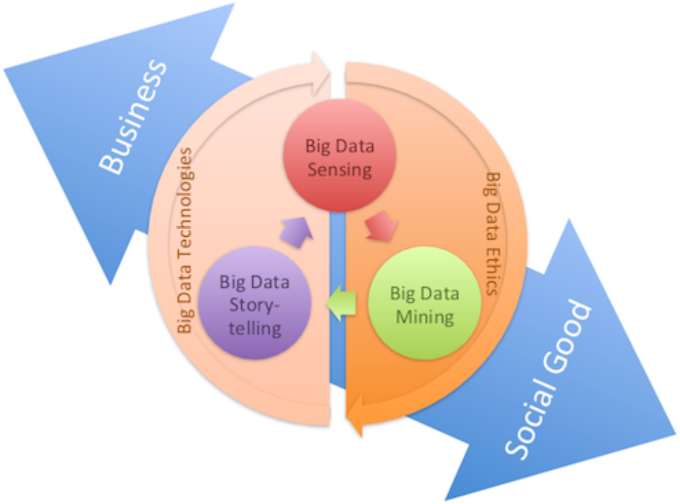
Mobility Data Analysis

Social Network Analysis

Web Mining & Nowcasting

Sentiment Analysis & Opinion Mining

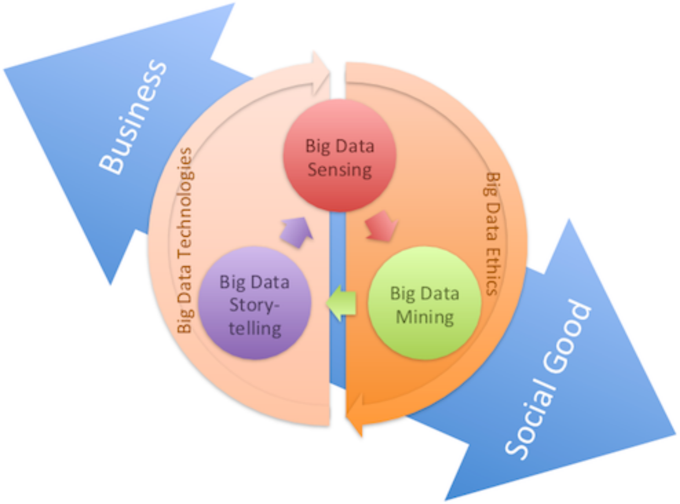
Big Data Story Tell



Data Visualization & Visual analytics

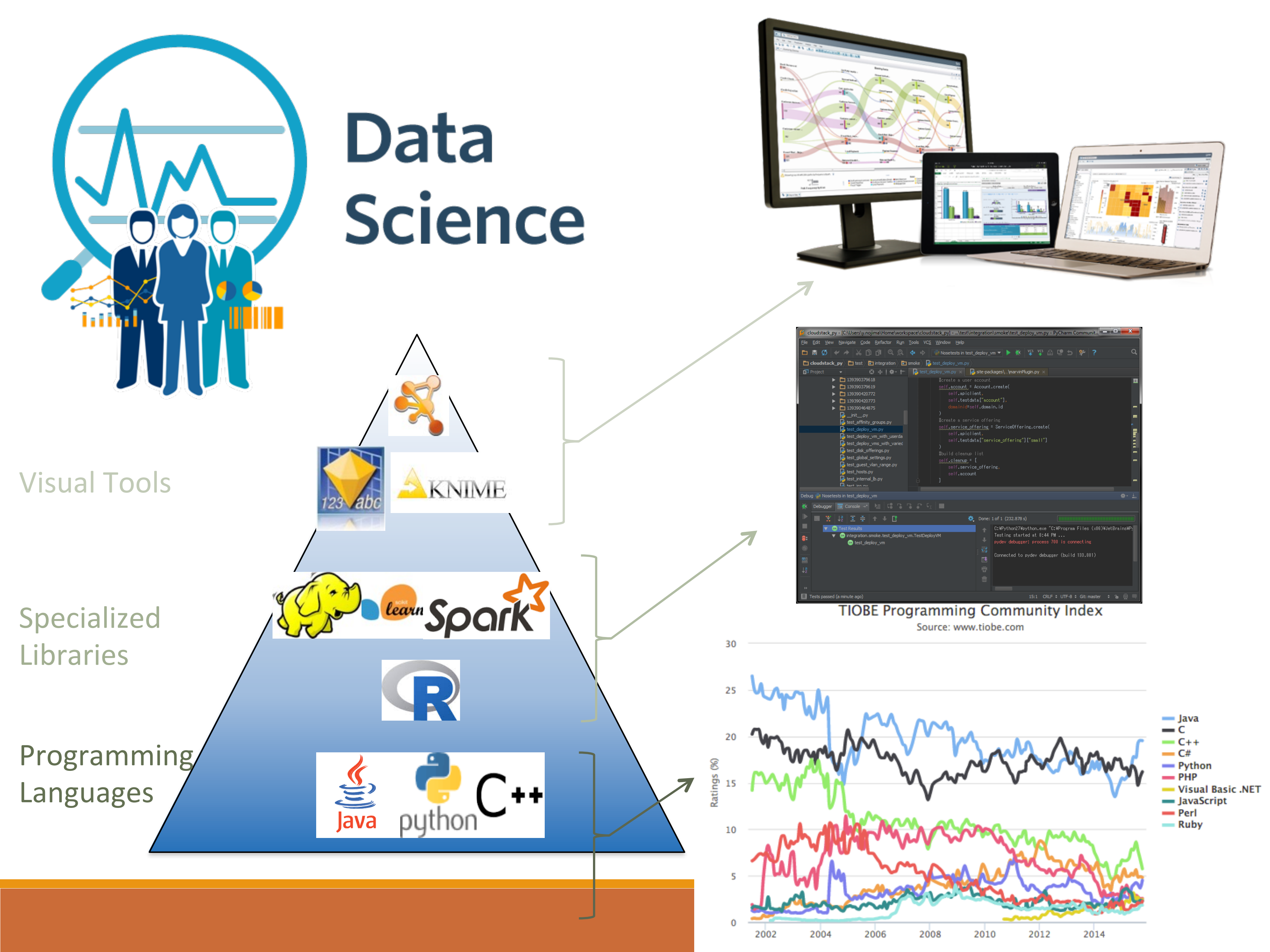
Data Journalism & Story Telling

Big Data Technol



Data Management for Business Intelligence

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



Data
Science
Visual Tools
Specialized
Libraries
Programming
Languages
KNIME
learn
Spark
R
Java
python
C++
TIOBE Programming Community Index
Source: www.tiobe.com
Ratings (%)
30
25
20
15
10
5
0
2002
2004
2006
2008
2010
2012
2014
Java
C
C++
C#
Python
PHP
Visual Basic .NET
JavaScript
Perl
Ruby

Data Mining/ML

WEKA The University of Waikato
sas
orange
cloudera Oryx
rapidminer
Angoss Predict. Act. Perform
SALFORD SYSTEMS
KNIME
Alpine DATA LABS
AZURE ML
APACHE SAMOA
StatSoft STATISTICA
IBM SPSS
R
python
mahout
Spark MLlib
H2O

BI/Visualization

Spotfire TIBCO Software
sas
TIBCO Jaspersoft
pentaho
chartio
looker
Yellowfin
MicroStrategy
COGNOS
QlikView
birst
board
DOMO
ZOOMDATA
POWER BI
Information Builders The Standard for Enterprise Business Intelligence
ORACLE BUSINESS INTELLIGENCE
SAP BusinessObjects
tableau software
GoodData
INSIGHTSQUARED
SISENSE
phocas Keep It Simple

ETL/DW

IBM Infosphere Information Server
ORACLE DATA INTEGRATOR
ELIXIR REPERTOIRE 8 Integrated Business Intelligence Suite
QlikView
HIVE
PowerCenter 8 INFORMATICA
MicroStrategy
SAP BusinessObjects
Ab Initio
presto
pentaho
ORACLE DATA WAREHOUSE
CloverETL
sas
Microsoft SQL Server Integration Services
talend open integration solutions
cloudera IMPALA

Data Processing

SAP HANA
BigMemory
STORM
Apache Flink
H2O
hadoop Map Reduce
TEZ
apache Ignite
Spark
GridGain
Apache Hama

Data Storage

Microsoft Access
SQLite
NETEZZA
Microsoft Azure
Cloudant an IBM Company
Neo4j
GREENPLUM A DIVISION OF EMC
amazon DynamoDB
Google BigTable
VERTICA
ORACLE DATABASE
IBM DB2
mongoDB
cloudera IMPALA
redis
CouchDB
PostgreSQL
ORACLE NOSQL DATABASE
OrientDB
MySQL
Microsoft SQL Server
SAP SYBASE
TERADATA
APACHE HBASE
cassandra
riak
MariaDB

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: esploration, preparation and anaysis and final evaluation.
- Project presentation .ppt

Individual Project Discussion

	Feature Construction	Model Construction	Validation	Interpretation and story telling
Required (all) **Grade range: 18-24**	Study of existing features, correlation analysis, selection of the interesting ones, transformation, construction of useful features	Select a modeling task appropriate to the analytical objective that the student has proposed	Provide a discussion on the base of objective measures of the methods about performance: SSE, Accuracy, ROC, Lift, Support, Confidence,	Discuss the achievements w.r.t the potential usage of the model and discuss the potential improvements
Advanced (all the required plus at least one) **Grade range: 24-28**	Integration with external (new) sources	Combine several models or adopt more advanced to archive better explanation or better performances (for example combine clustering with pattern mining or do ensemble methods or sophisticated classification as multilabel classification or cascade)	also consider possible domain dependent cost function	also discuss the potential improvements w.r.t a comparison w.r.t a quantitative baseline
Challenging (Advanced plus at least one) **Grade range: 28-30L**	Invention of new features	Compare your results with those obtained with other models and algorithms (e.g. compare a decision tree model with SVM-CNN trained on the same dataset)	Also discuss w.r.t a ground truth obtained by a null model or a human generated labelling , or other "true" source	also discuss the potential improvements w.r.t a comparison w.r.t a quantitative baseline

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

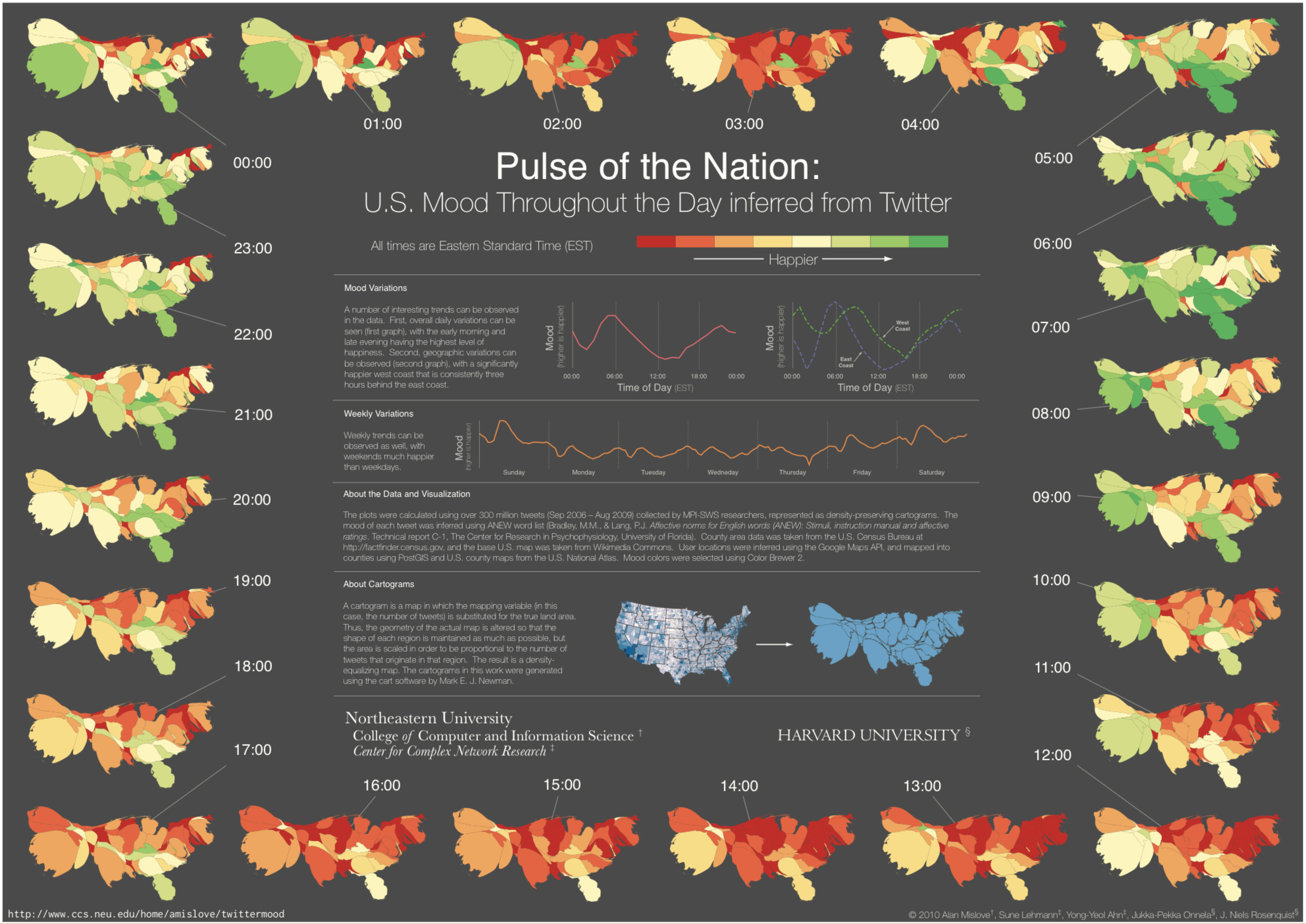
Google

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]

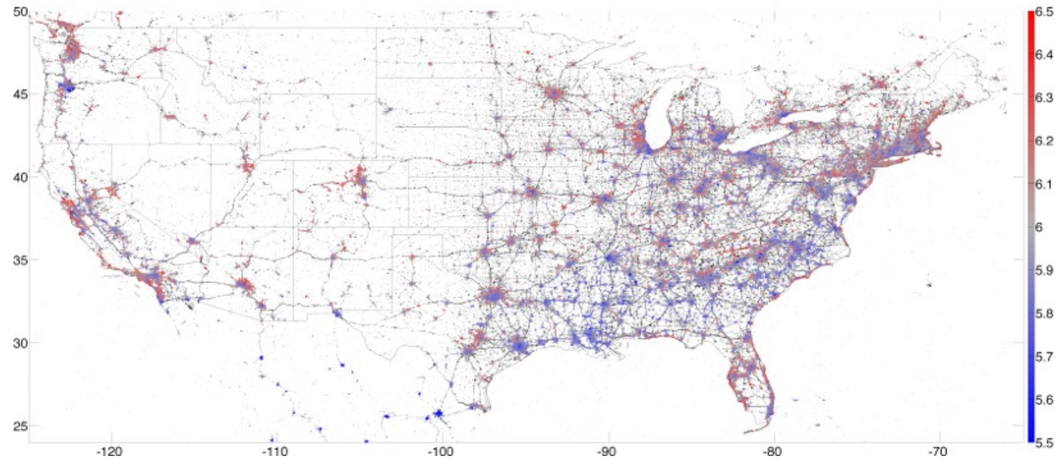
[1]*Google Inc.* [2]*Centers for Disease Control and Prevention*

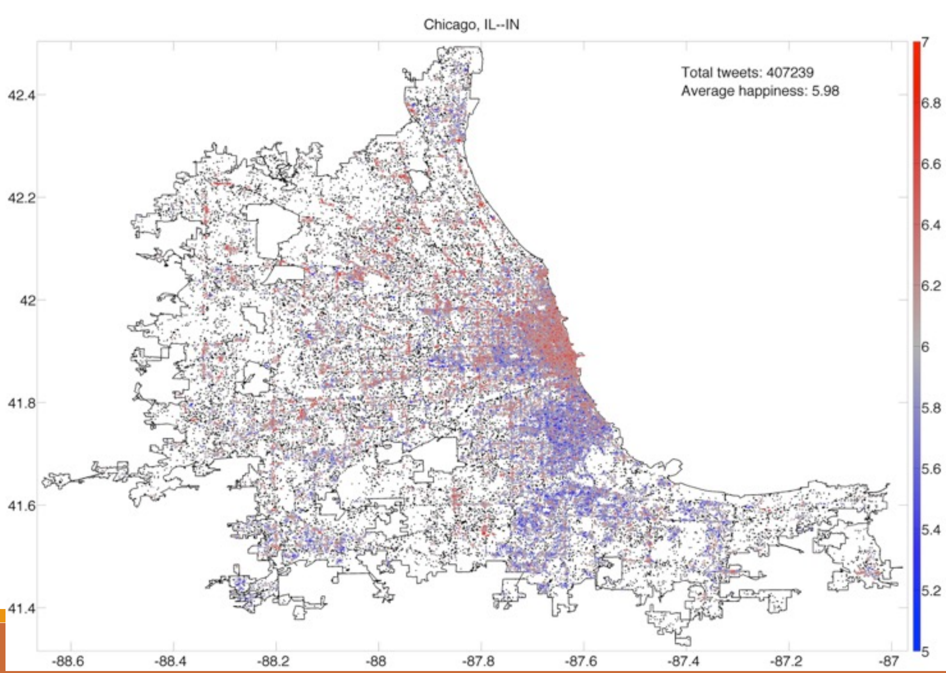
Nature 457, 1012-1014 (2009)

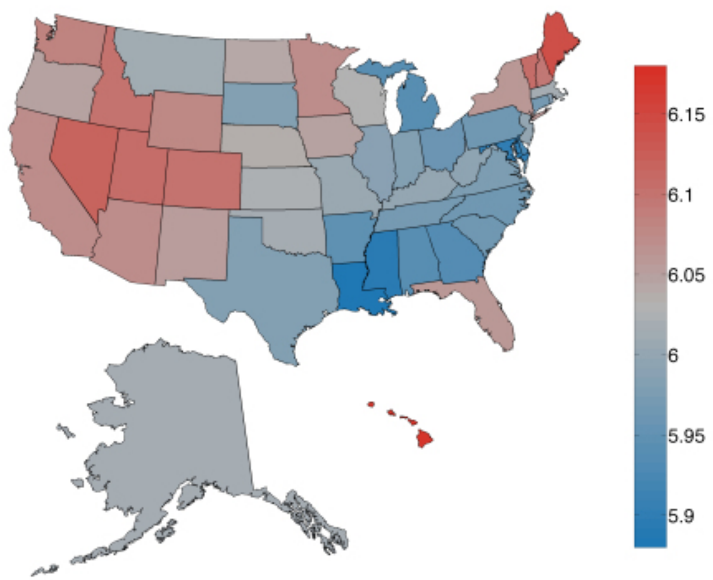


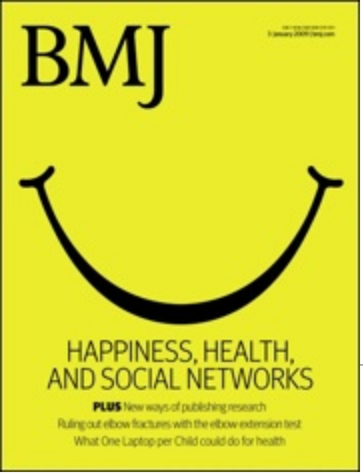
Pulse of the Nation:
U.S. Mood Throughout the Day inferred from Twitter
All times are Eastern Standard Time (EST)
Happier
00:00
01:00
02:00
03:00
04:00
05:00
06:00
07:00
08:00
09:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Mood Variations
A number of interesting trends can be observed in the data. First, overall daily variations can be seen (first graph), with the early morning and late evening having the highest level of happiness. Second, geographic variations can be observed (second graph), with a significantly happier west coast that is consistently three hours behind the east coast.
Mood (higher is happier)
00:00 06:00 12:00 18:00 00:00
Time of Day (EST)
West Coast
East Coast
Weekly Variations
Weekly trends can be observed as well, with weekends much happier than weekdays.
Mood (higher is happier)
Sunday Monday Tuesday Wedneday Thursday Friday Saturday
About the Data and Visualization
The plots were calculated using over 300 million tweets (Sep 2006 – Aug 2009) collected by MPI-SWS researchers, represented as density-preserving cartograms. The mood of each tweet was inferred using ANEW word list (Bradley, M.M., & Lang, P.J. *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings*. Technical report C-1, The Center for Research in Psychophysiology, University of Florida). County area data was taken from the U.S. Census Bureau at http://factfinder.census.gov, and the base U.S. map was taken from Wikimedia Commons. User locations were inferred using the Google Maps API, and mapped into counties using PostGIS and U.S. county maps from the U.S. National Atlas. Mood colors were selected using Color Brewer 2.
About Cartograms
A cartogram is a map in which the mapping variable (in this case, the number of tweets) is substituted for the true land area. Thus, the geometry of the actual map is altered so that the shape of each region is maintained as much as possible, but the area is scaled in order to be proportional to the number of tweets that originate in that region. The result is a density-equalizing map. The cartograms in this work were generated using the cart software by Mark E. J. Newman.
Northeastern University
College *of* Computer and Information Science †
Center for Complex Network Research ‡
HARVARD UNIVERSITY §
http://www.ccs.neu.edu/home/amislove/twittermood
© 2010 Alan Mislove†, Sune Lehmann‡, Yong-Yeol Ahn‡, Jukka-Pekka Onnela§, J. Niels Rosenquist§

Searching the most happy city in US

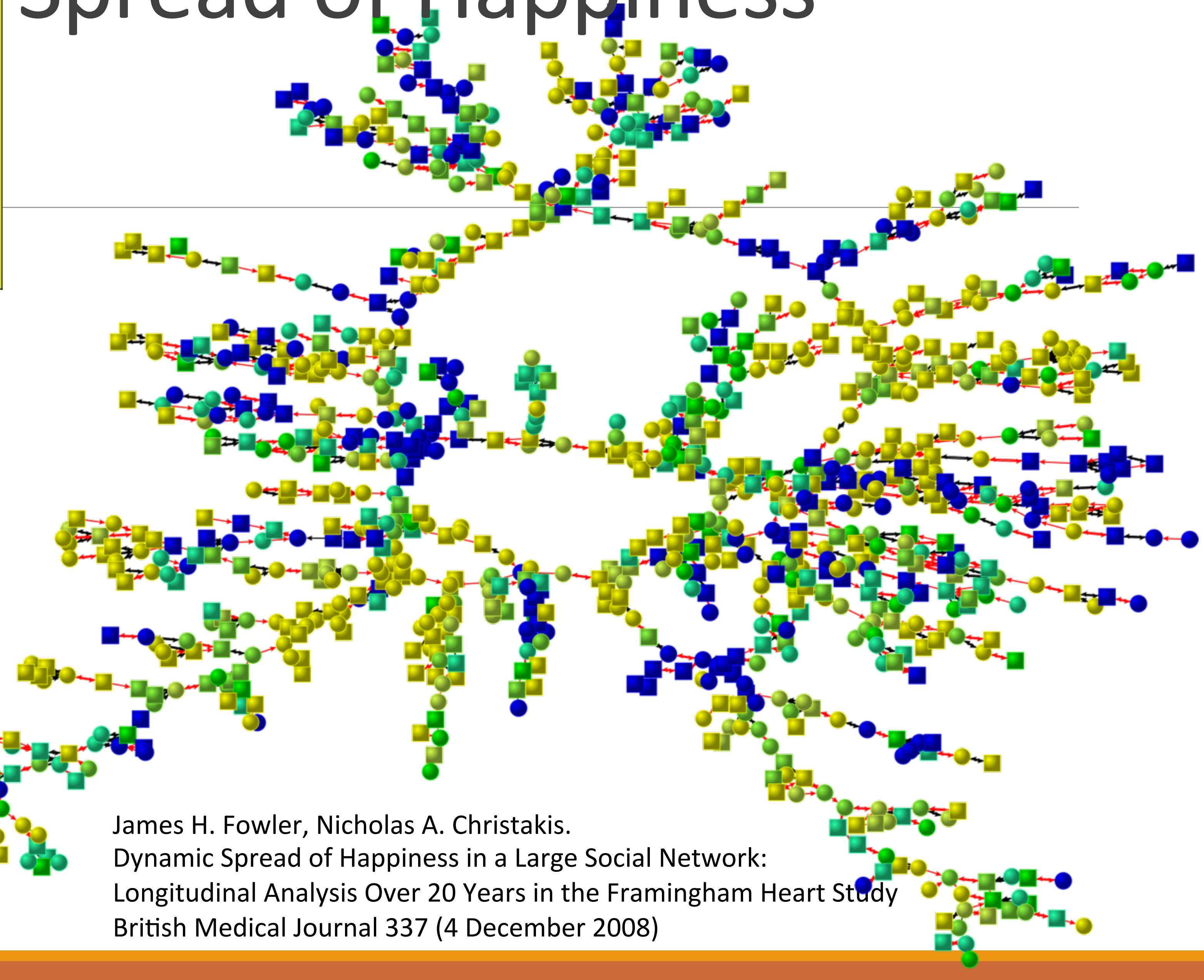






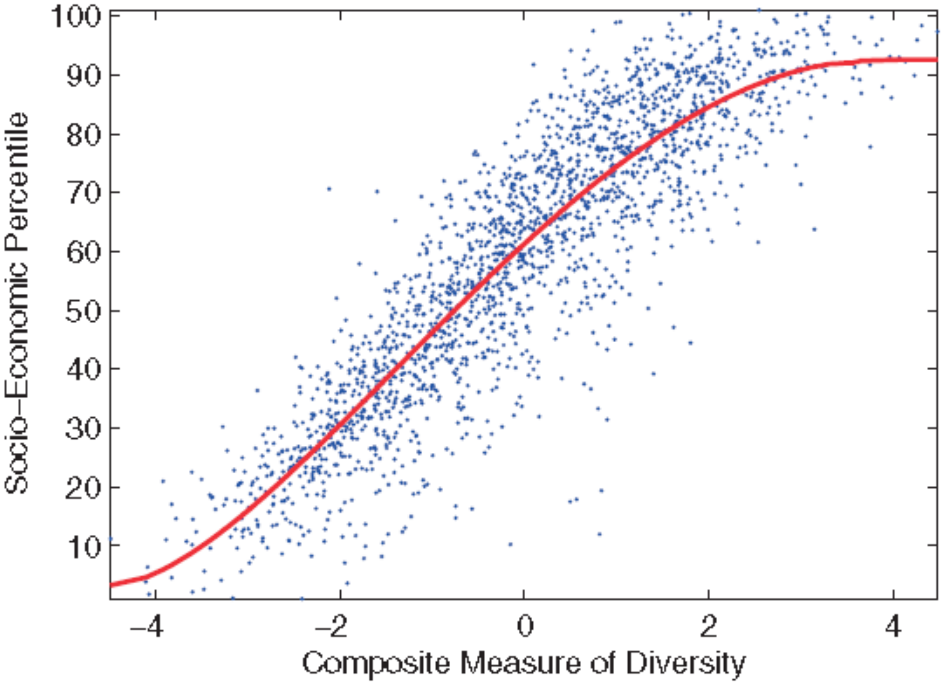


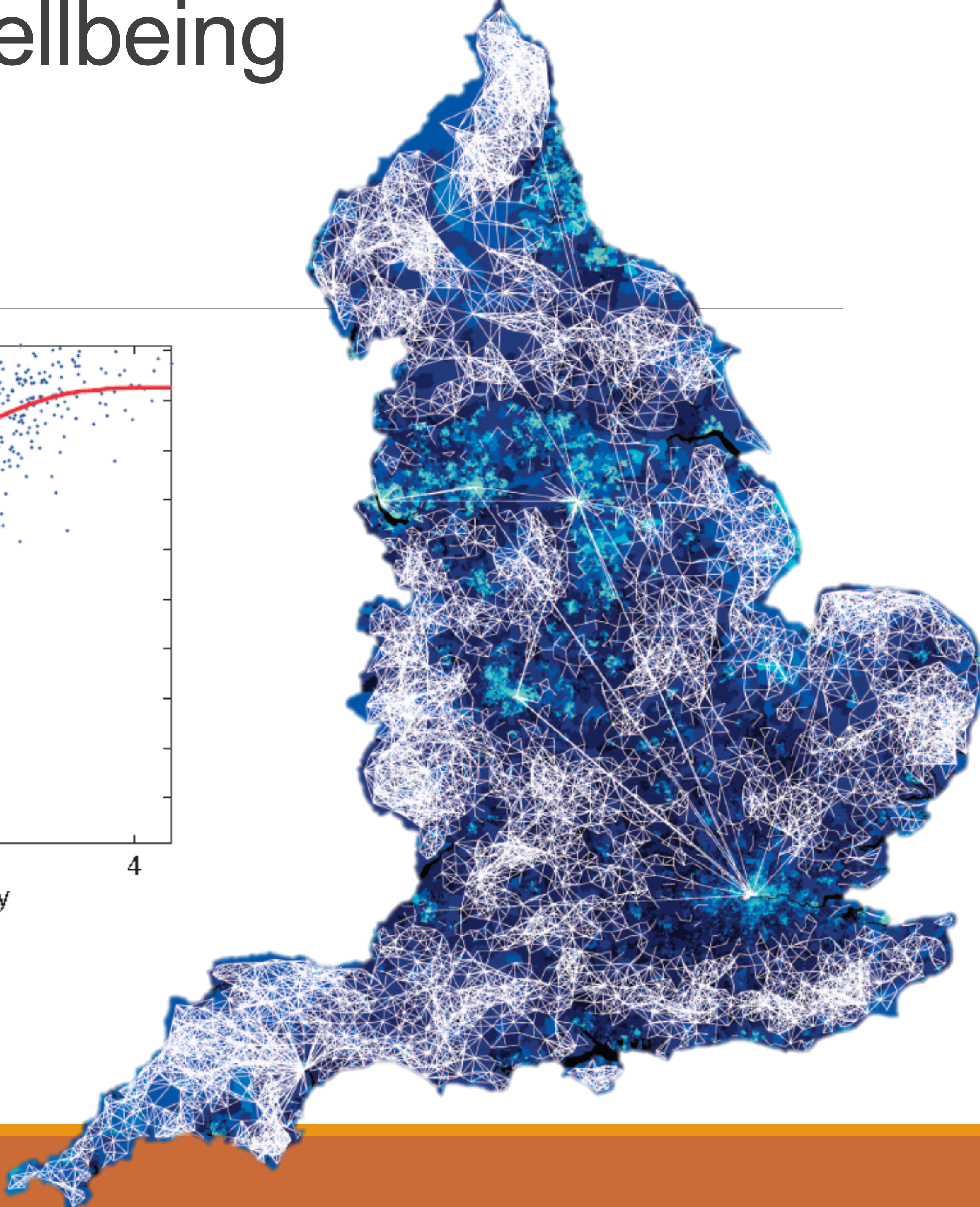
Spread of Happiness



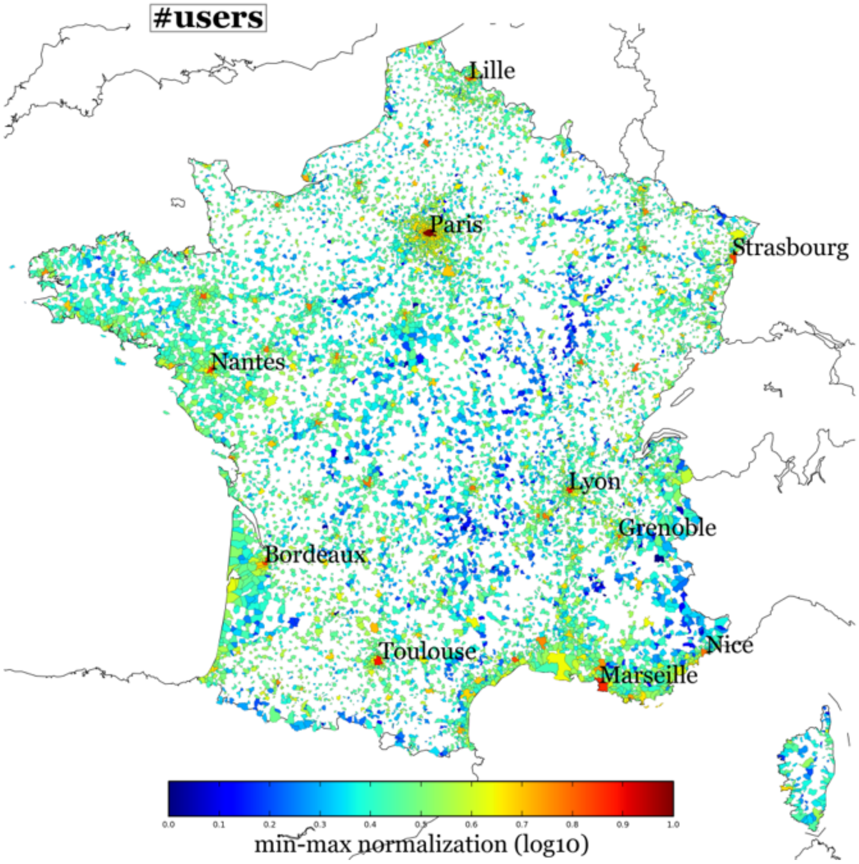
James H. Fowler, Nicholas A. Christakis.
Dynamic Spread of Happiness in a Large Social Network: Longitudinal Analysis Over 20 Years in the Framingham Heart Study
British Medical Journal 337 (4 December 2008)

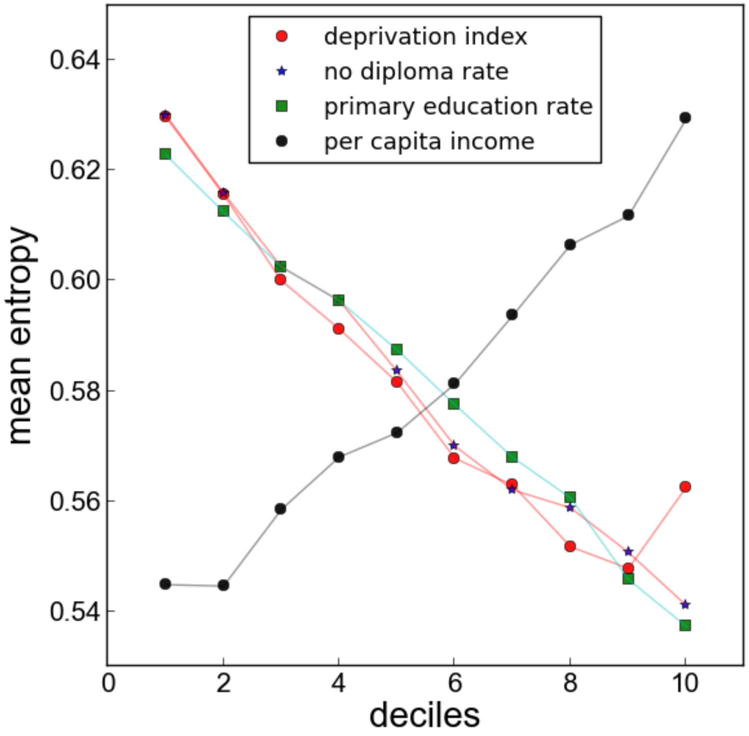
Diversity and Wellbeing (phone calls)



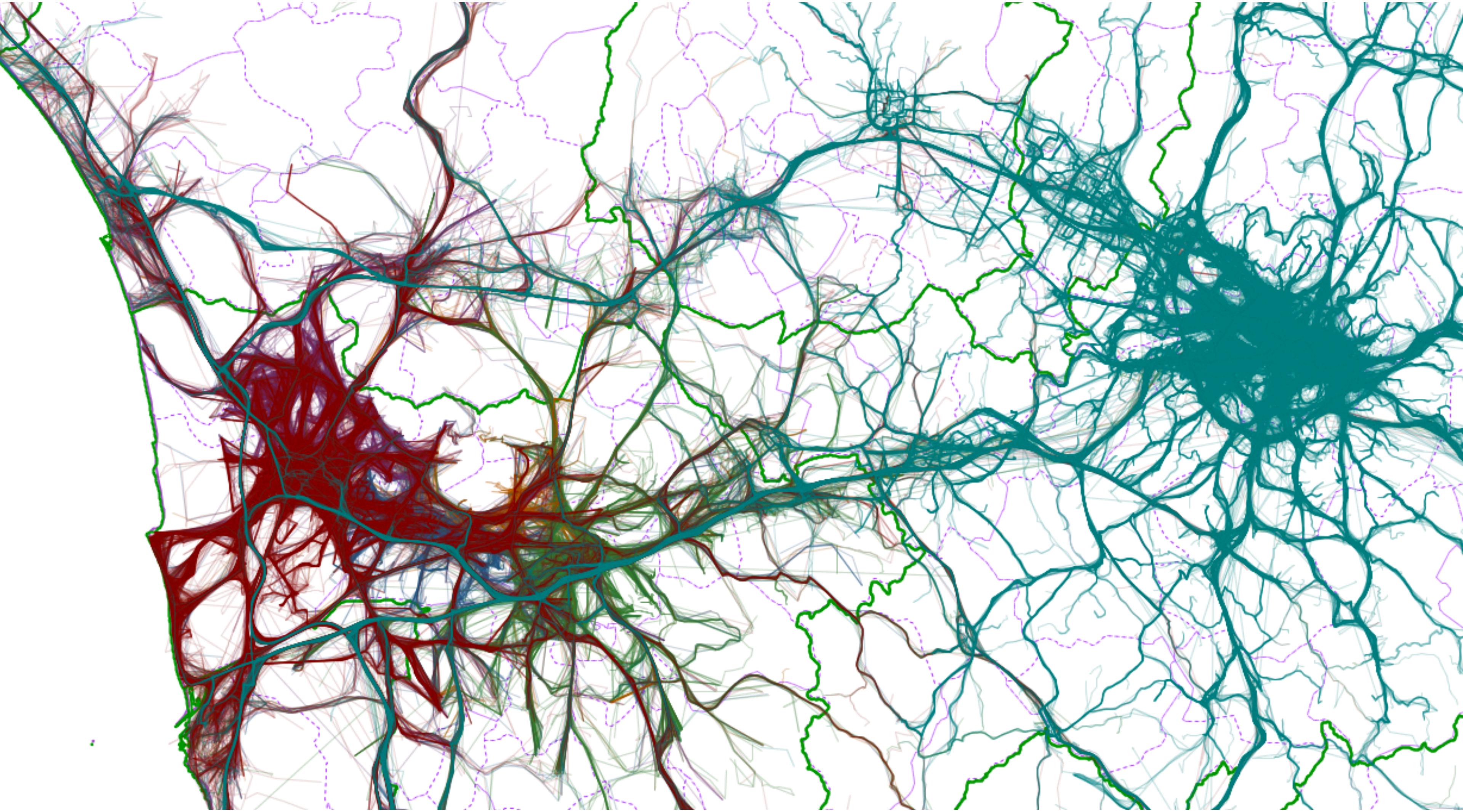


Diversity and Wellbeing (Mobility)

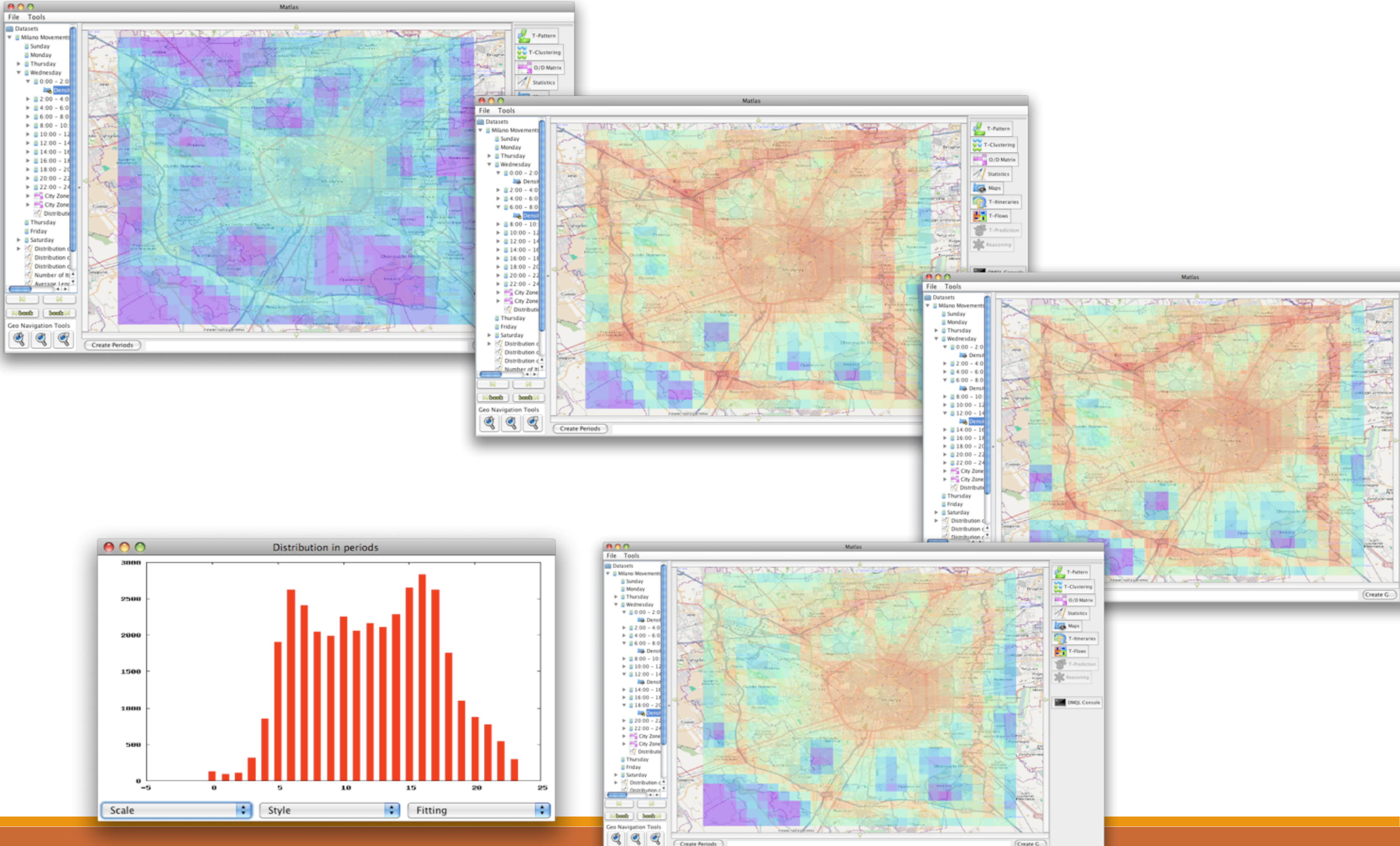




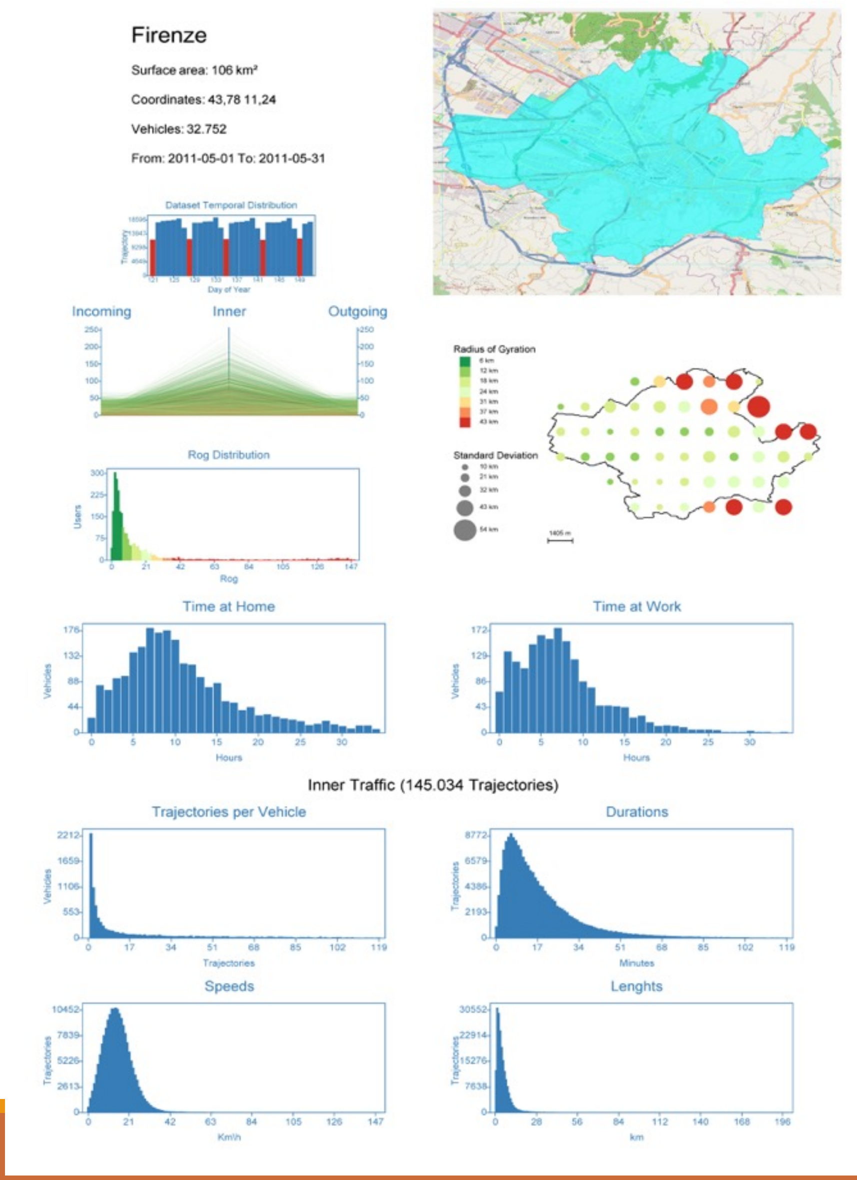
Big Data for smart cities

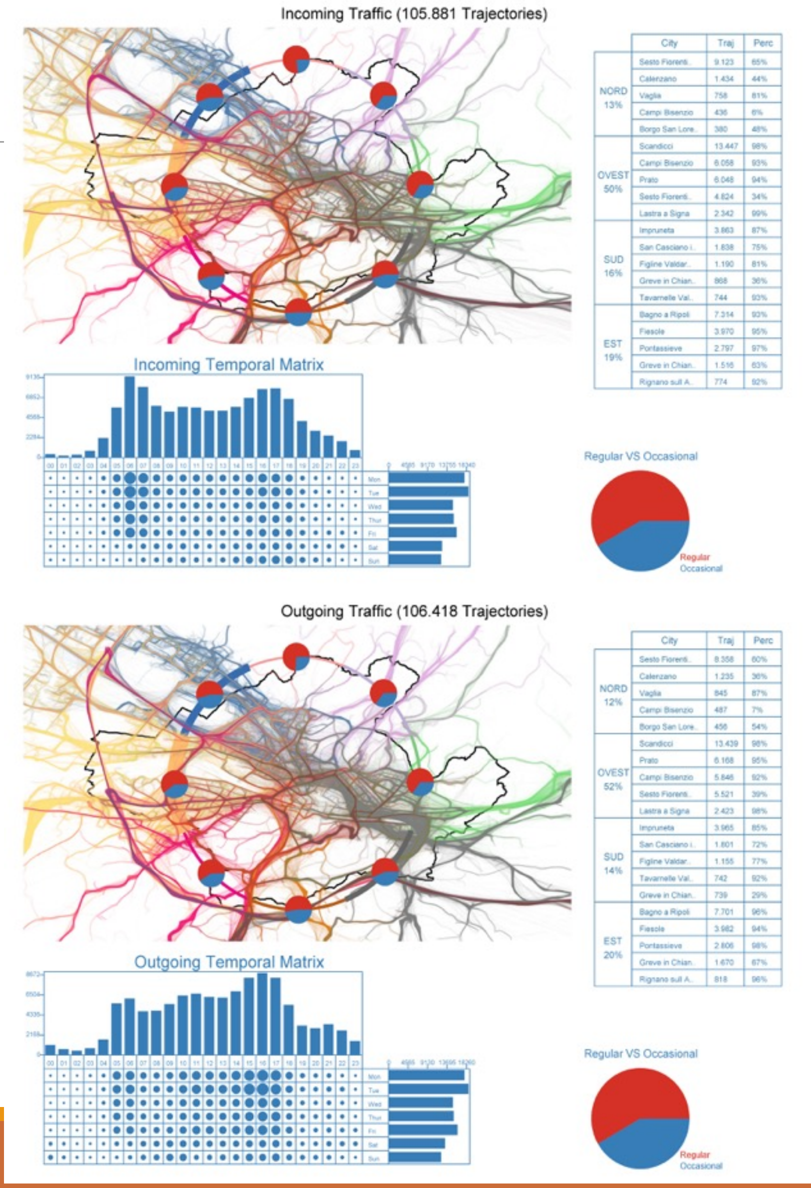


How people use the city during the day?



Fingerprint of the city

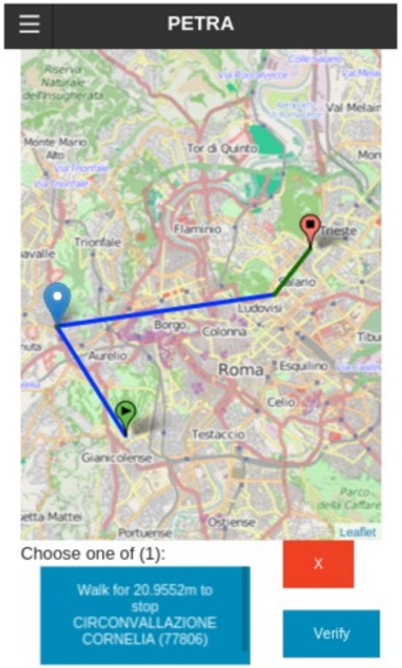


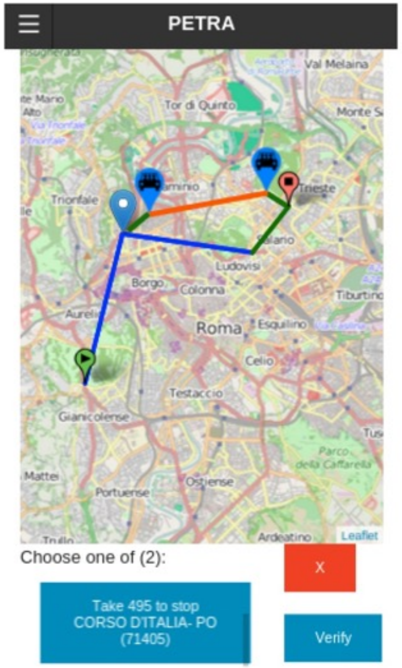


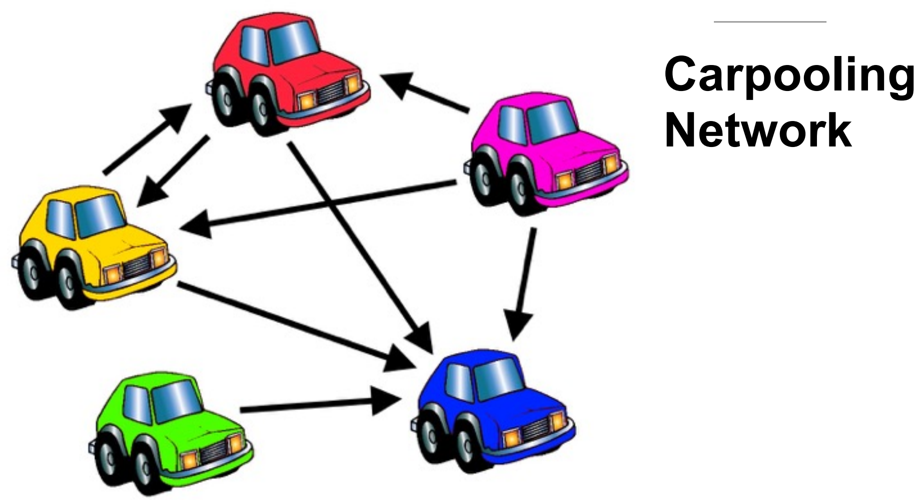
	City	Traj	Perc
NORD 13%	Sesto Fiorenti..	9.123	65%
	Calenzano	1.434	44%
	Vaglia	758	81%
	Campi Bisenzio	436	6%
	Borgo San Lore..	380	48%
OVEST 50%	Scandicci	13.447	98%
	Campi Bisenzio	6.058	93%
	Prato	6.048	94%
	Sesto Fiorenti..	4.824	34%
	Lastra a Signa	2.342	99%
SUD 16%	Impruneta	3.863	87%
	San Casciano 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%

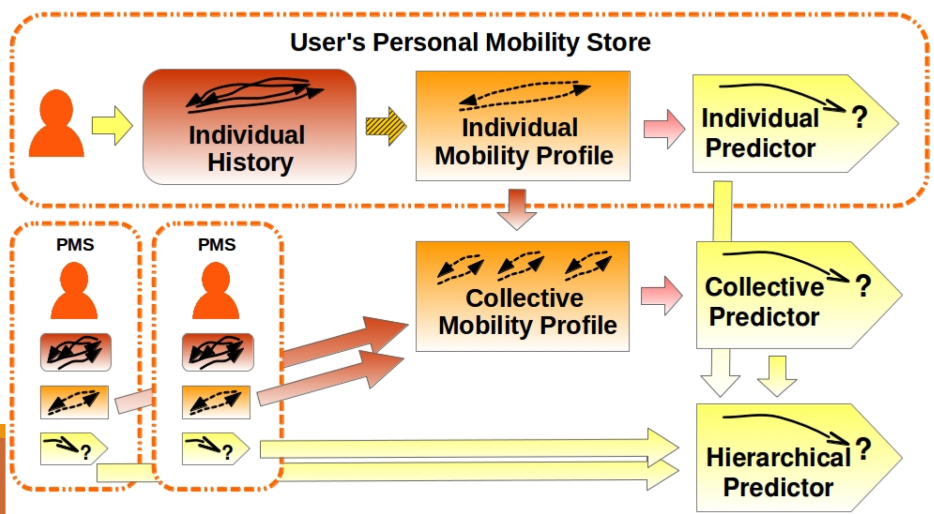
	City	Traj	Perc
NORD 12%	Sesto Fiorenti..	8.358	60%
	Calenzano	1.235	36%
	Vaglia	845	87%
	Campi Bisenzio	487	7%
	Borgo San Lore..	456	54%
OVEST 52%	Scandicci	13.439	98%
	Prato	6.168	95%
	Campi Bisenzio	5.846	92%
	Sesto Fiorenti..	5.521	39%
	Lastra a Signa	2.423	98%
SUD 14%	Impruneta	3.965	85%
	San Casciano 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

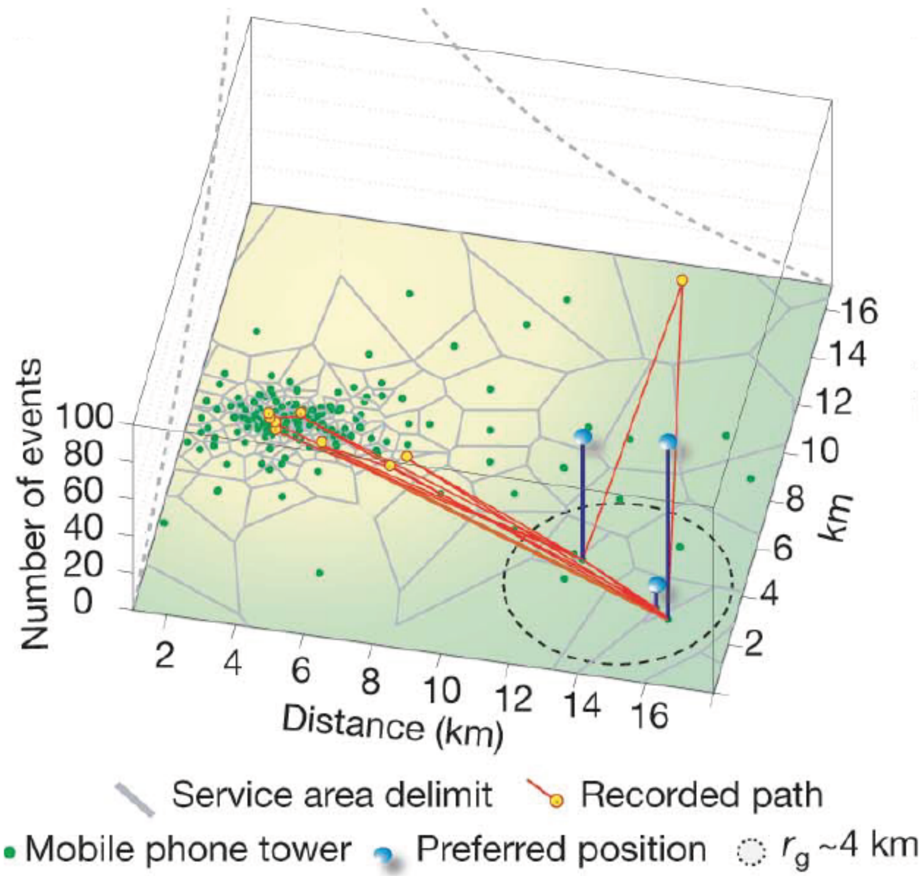








Call Data Records… when, where and who

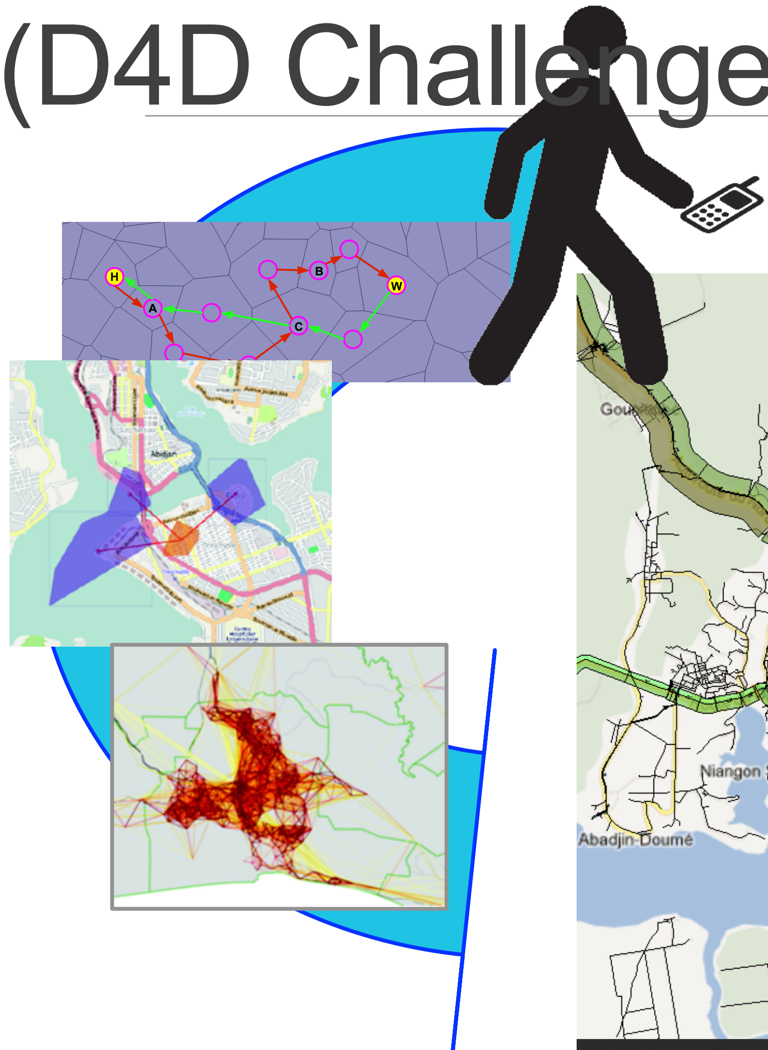


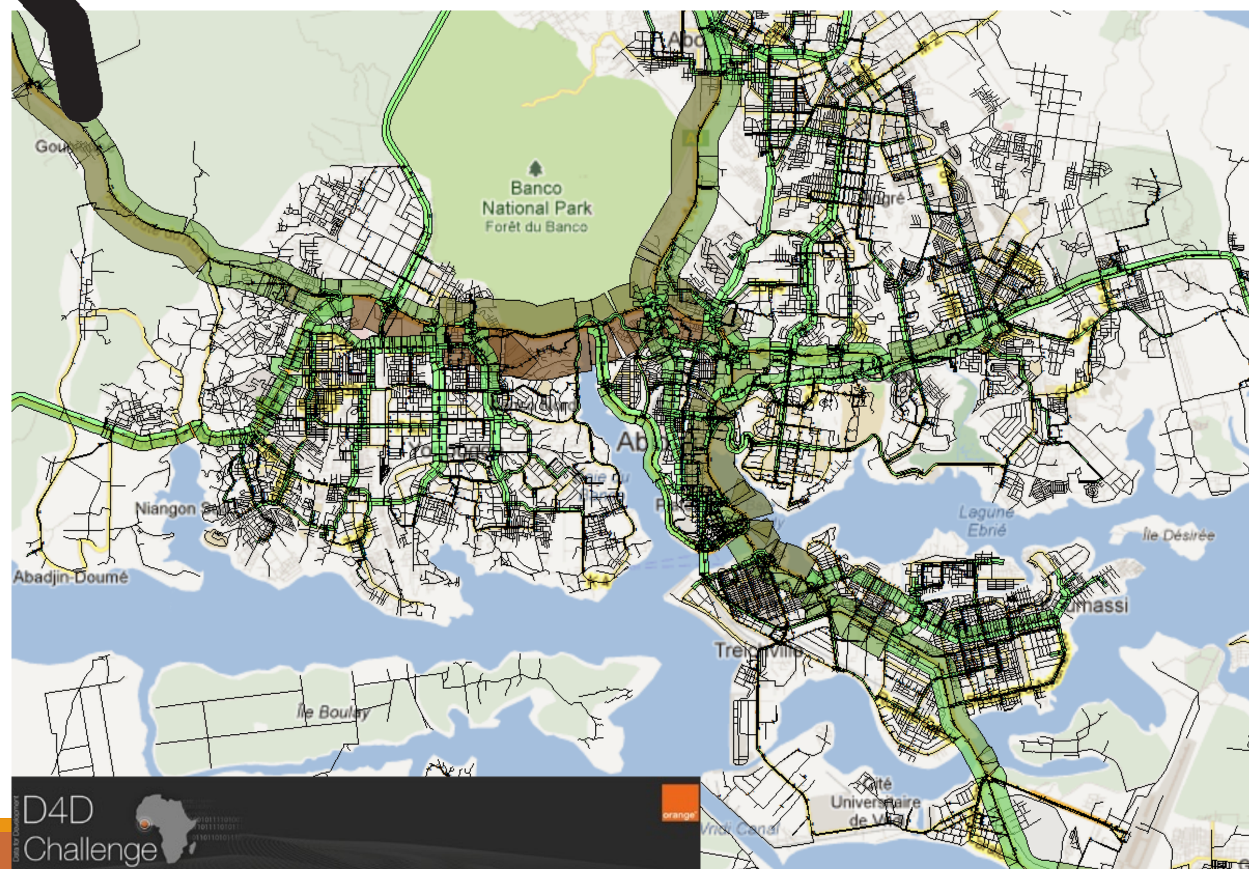
when you call

where you call

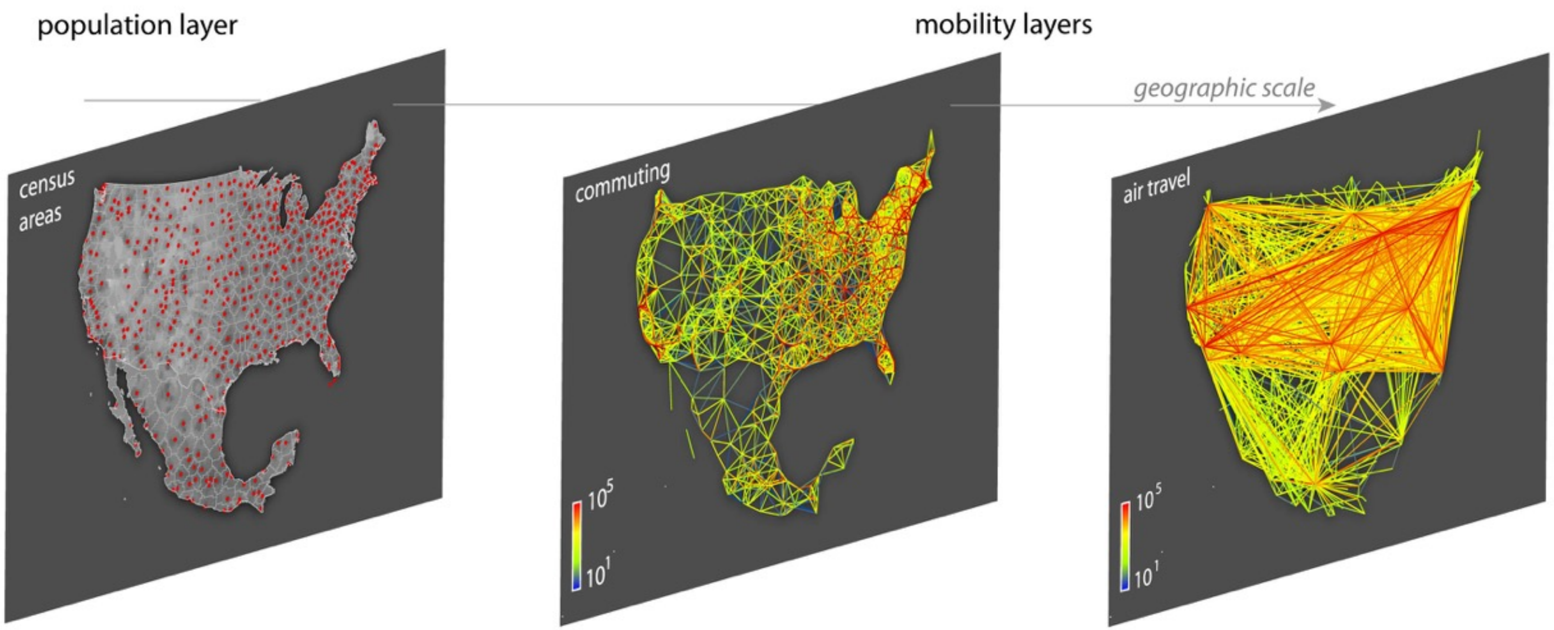
who you call

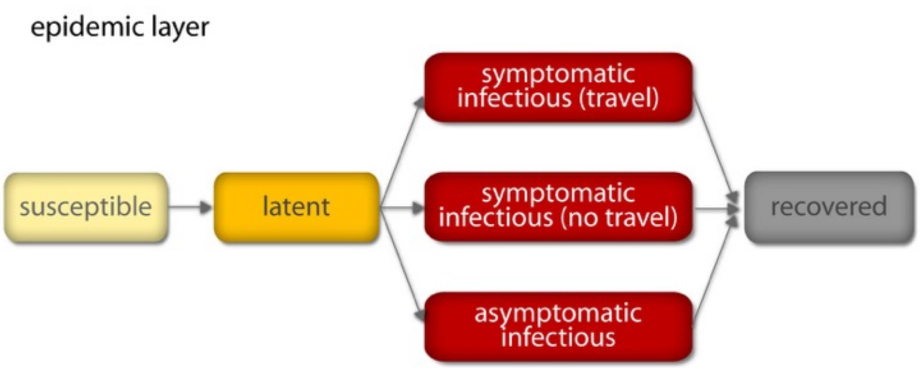
Call Data Records for Developing Countries (D4D Challenge)





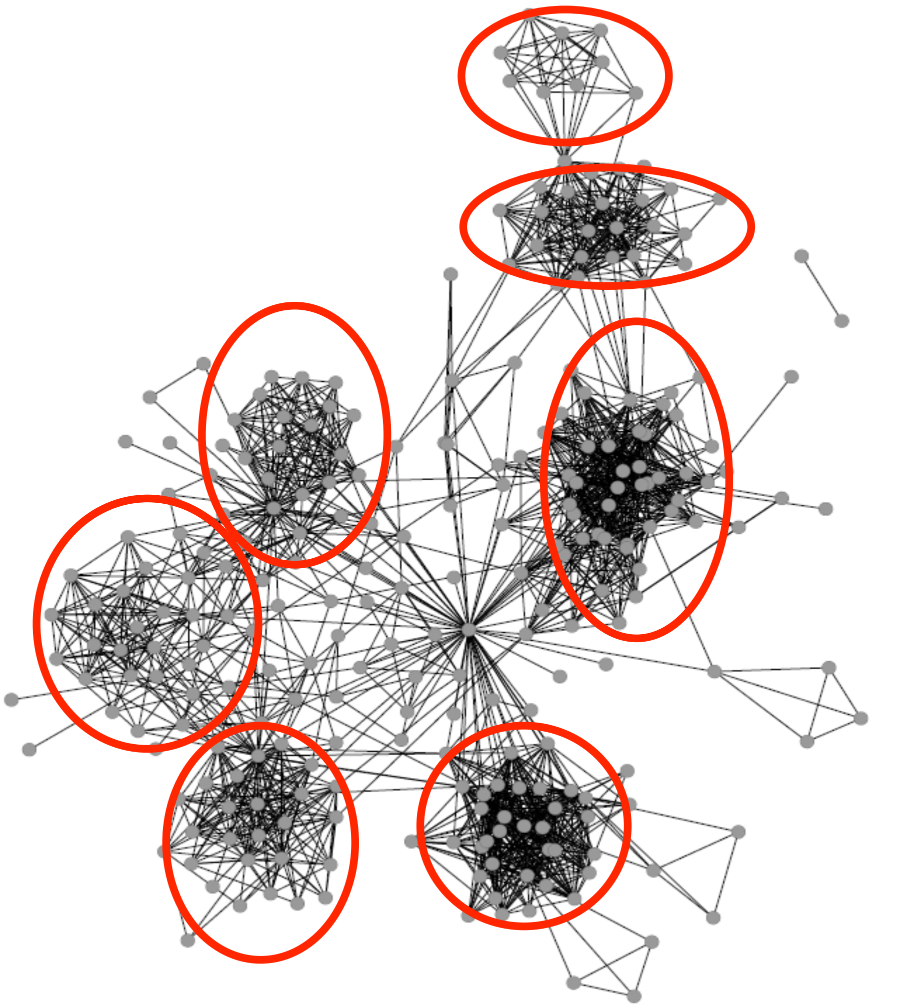
Epidemics simulations

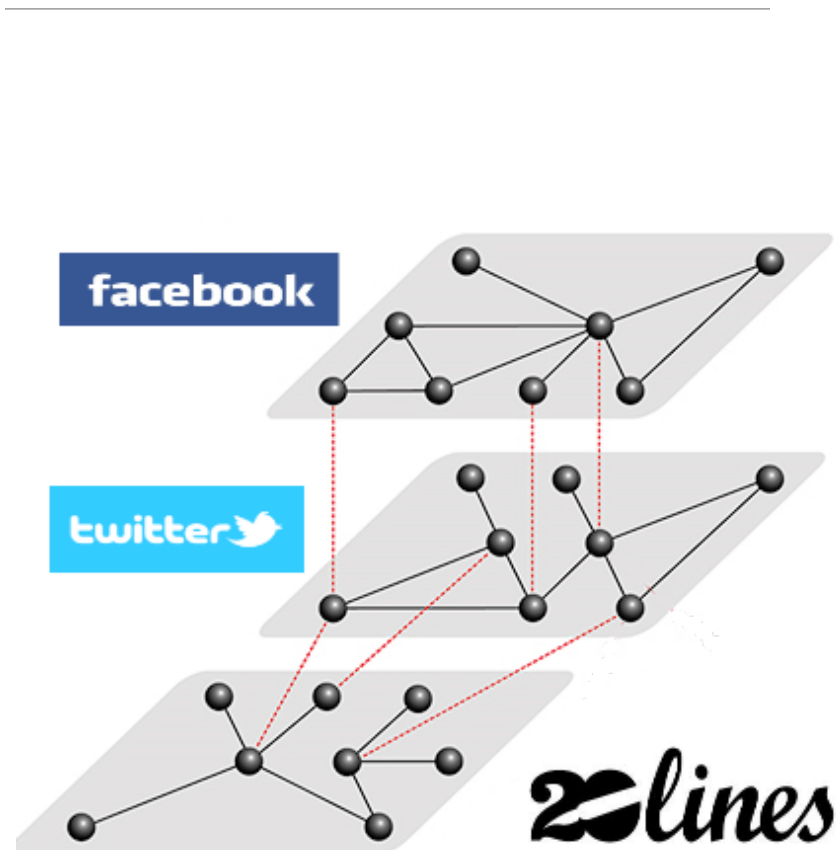




Parameter	Value	Description
β	from R_0	transmission probability
ε^{-1}	1.9 [1.1-2.5] d	average latency period
μ^{-1}	3 [3-5] d	average infectious period
p_t	50%	probability of traveling for infectious individuals
p_a	33%	probability of being asymptomatic
r_β	50%	relative infectiousness of asymptomatic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

Community Discovery, Evolution, Diffusion, Multidimensionality,…



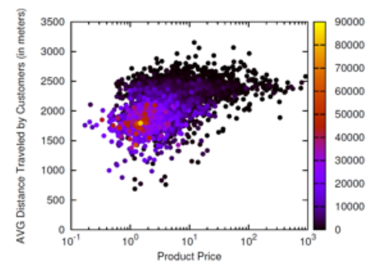


Michele Coscia, Giulio Rossetti, Fosca Giannotti, Dino Pedreschi: **Uncovering Hierarchical and Overlapping Communities with a Local-First Approach**. TKDD 9(1): 6 (2014)

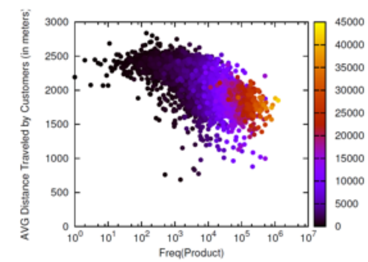
Retail Market as Complex system

generic utility function (rationality)

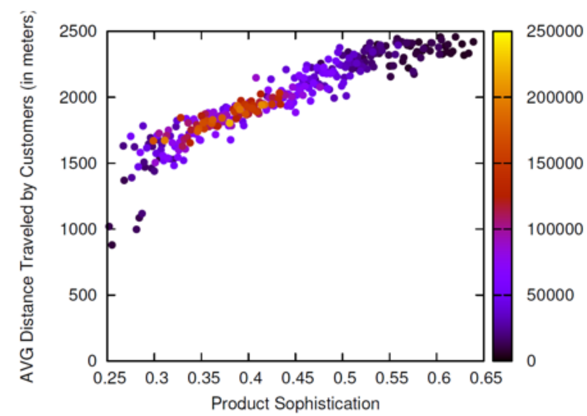
personal utility function (diversity)



$R^2 = 17.25\%$



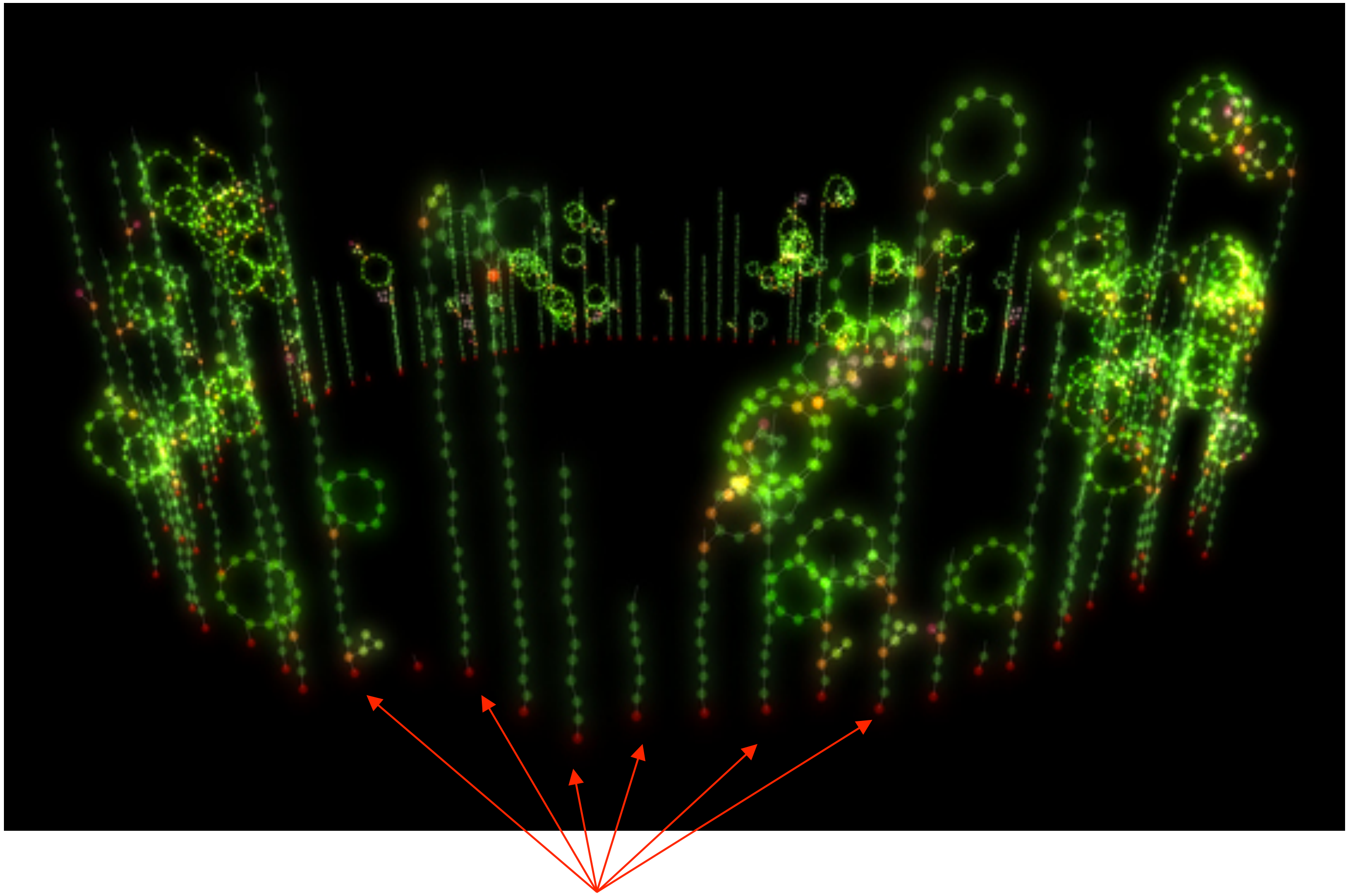
$R^2 = 32.38\%$



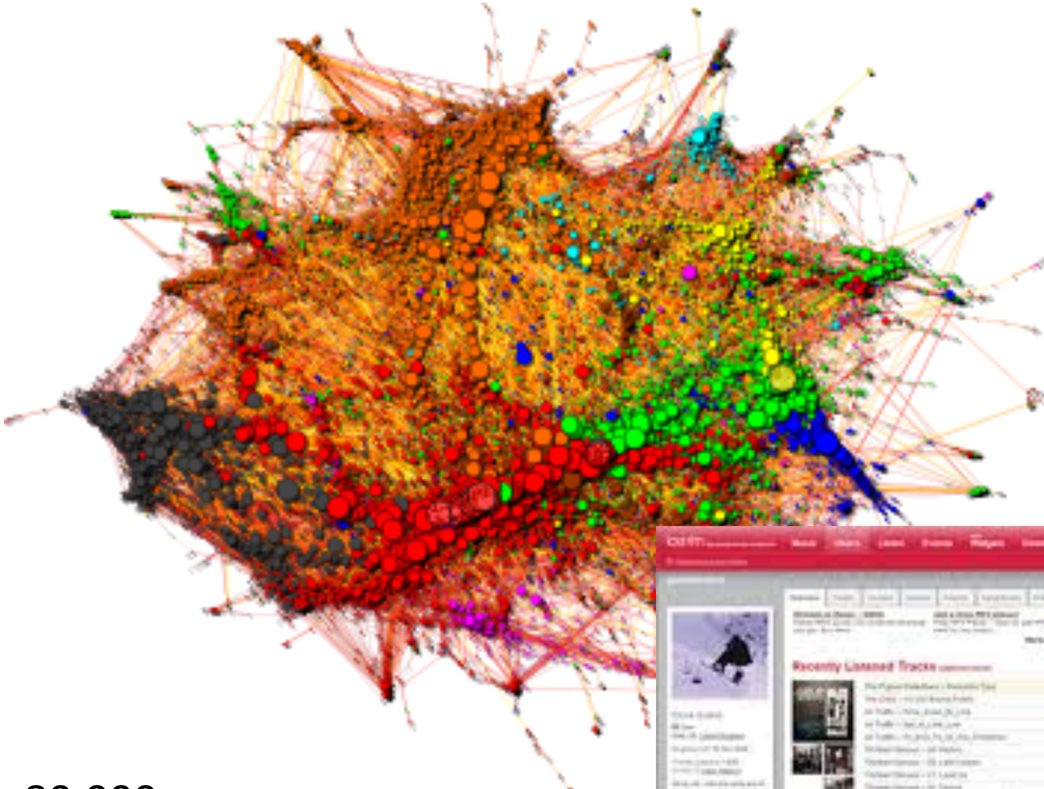
$R^2 = 85.72\%$

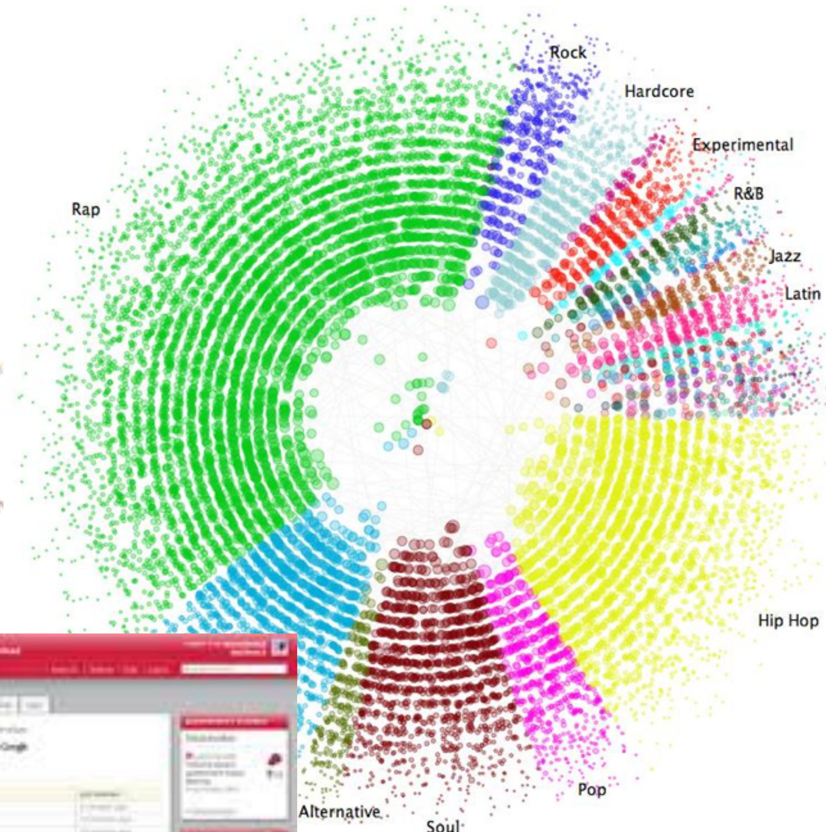
Pennacchioli, D., Coscia, M., Rinzivillo, S., Giannotti, F. and Pedreschi, D., **The retail market as a complex system**. In EPJ Data Science, 2014.

Social Influence: Leaders



Ask to LAST.FM





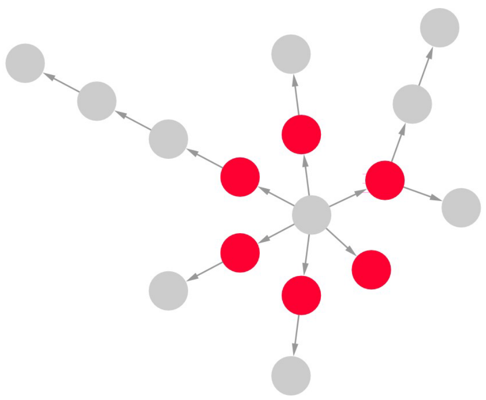
80.000 users,
4000.000 connections

What is Social Prominence?

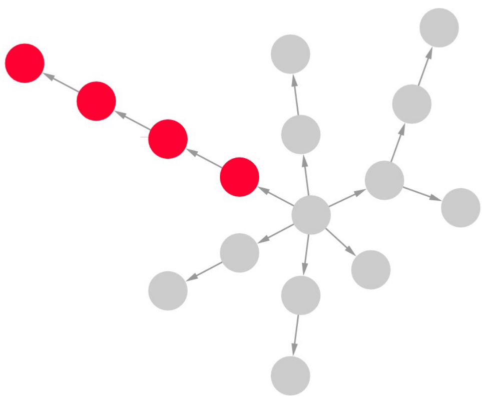
It has been observed that a small set of users in a Social Network is able to anticipate (or influence) the behavior of the entire network

We detected 3 possible scenarios:

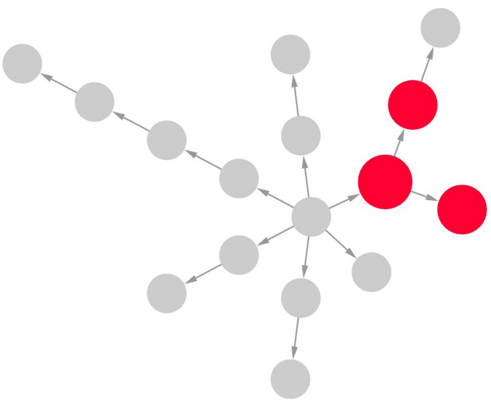
width



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

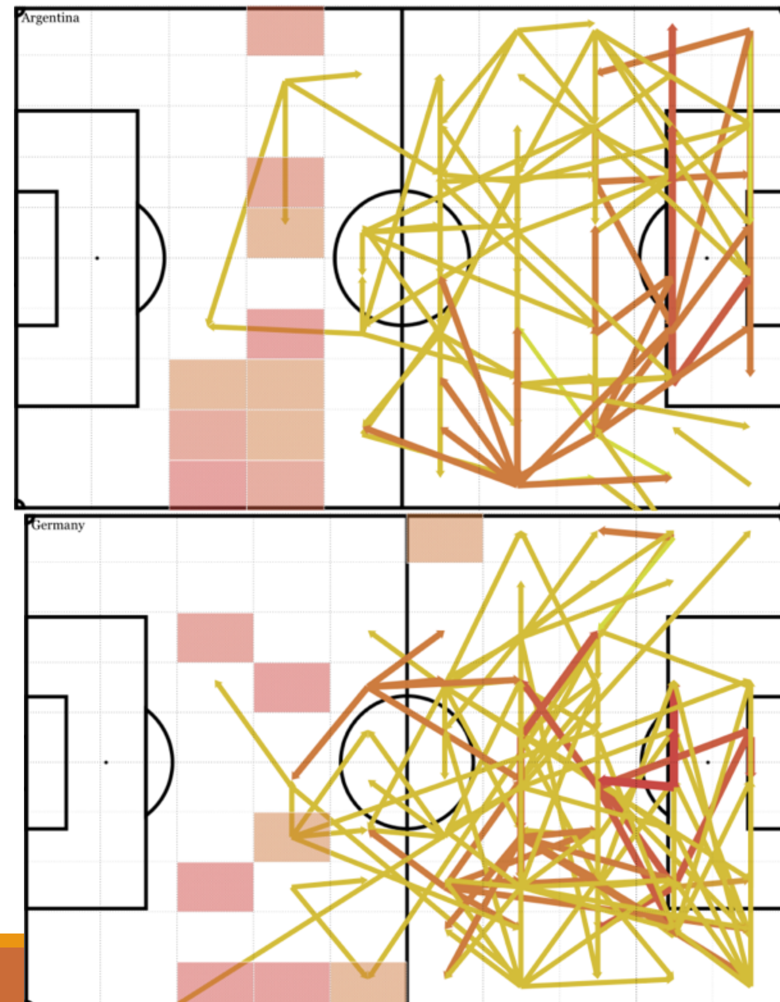
The patterns of success in cycling:

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

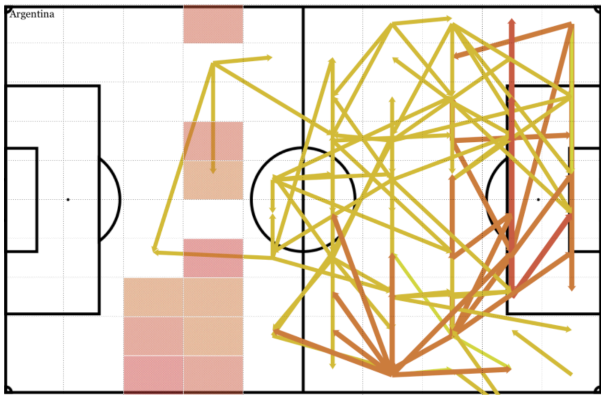
Sports

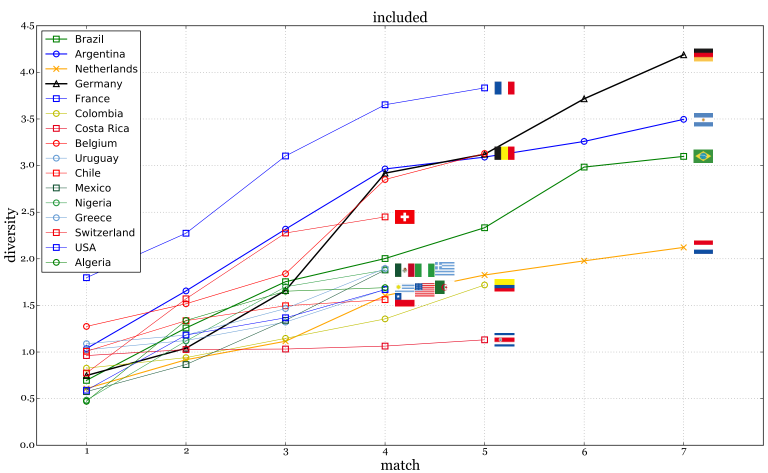
“Football is a simple game: 22 men chase a ball for 90 minutes and at the end, the Germans always win”

-- Gary Lieneker (after Italy 1990 Final)



Big Data: the way of Success





The patterns of success in football:

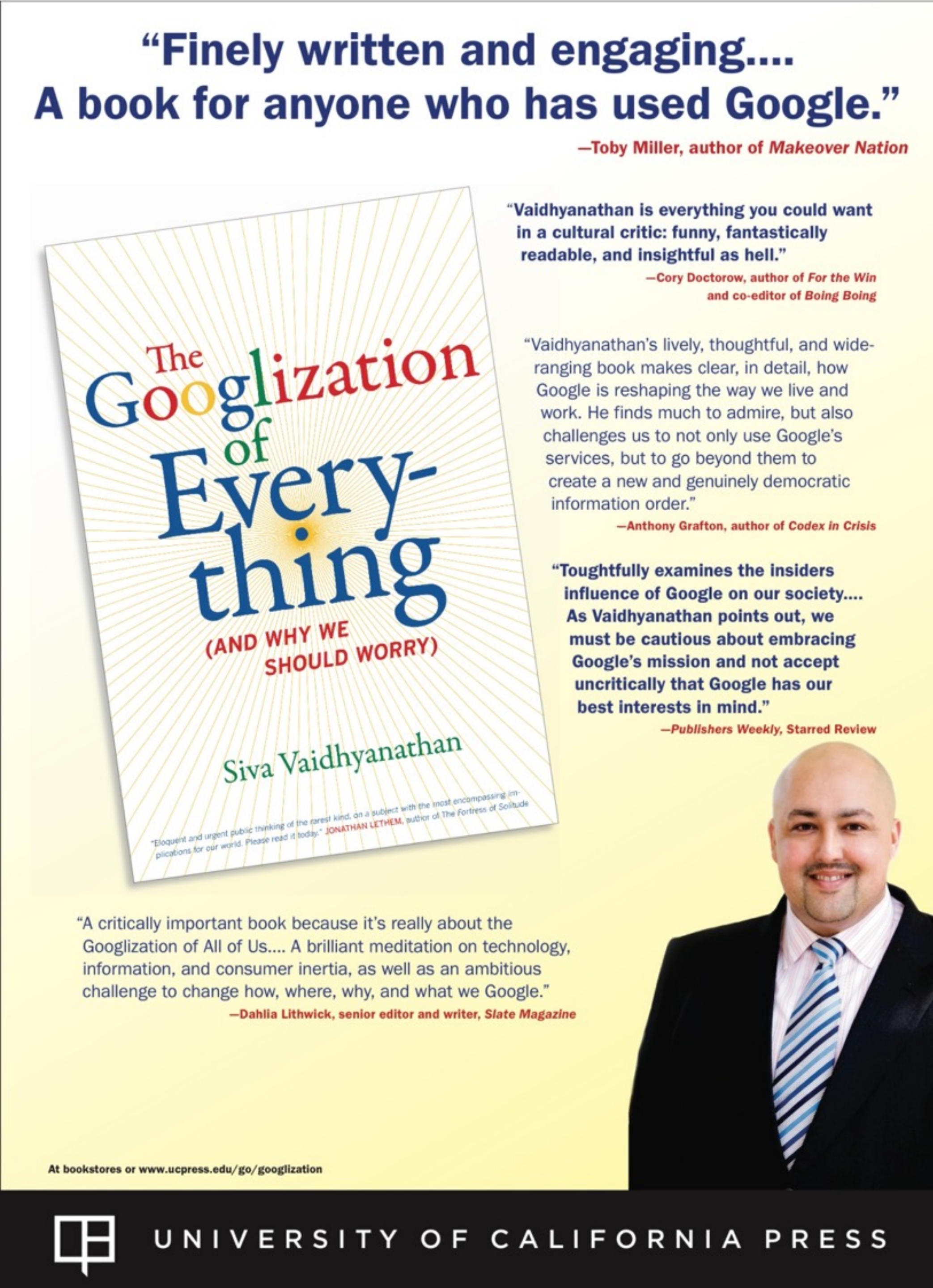
- detailed data on every match (trajectories, passes, goals, …)
- a network approach to study the strategy of teams
- a data mining approach to study the performance of players

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



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

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

$300 billion

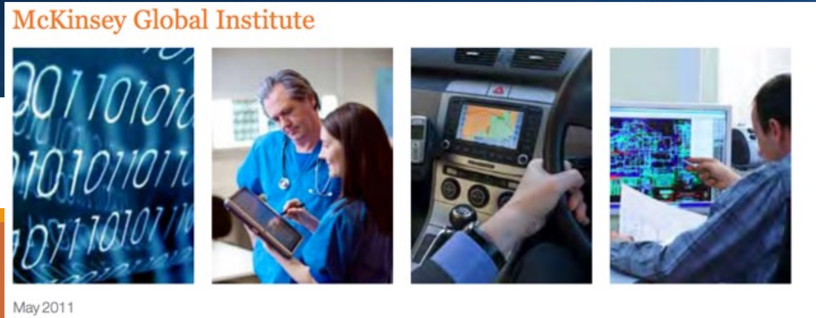
potential annual value to US health care—more than double the total annual health care spending in Spain

€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



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

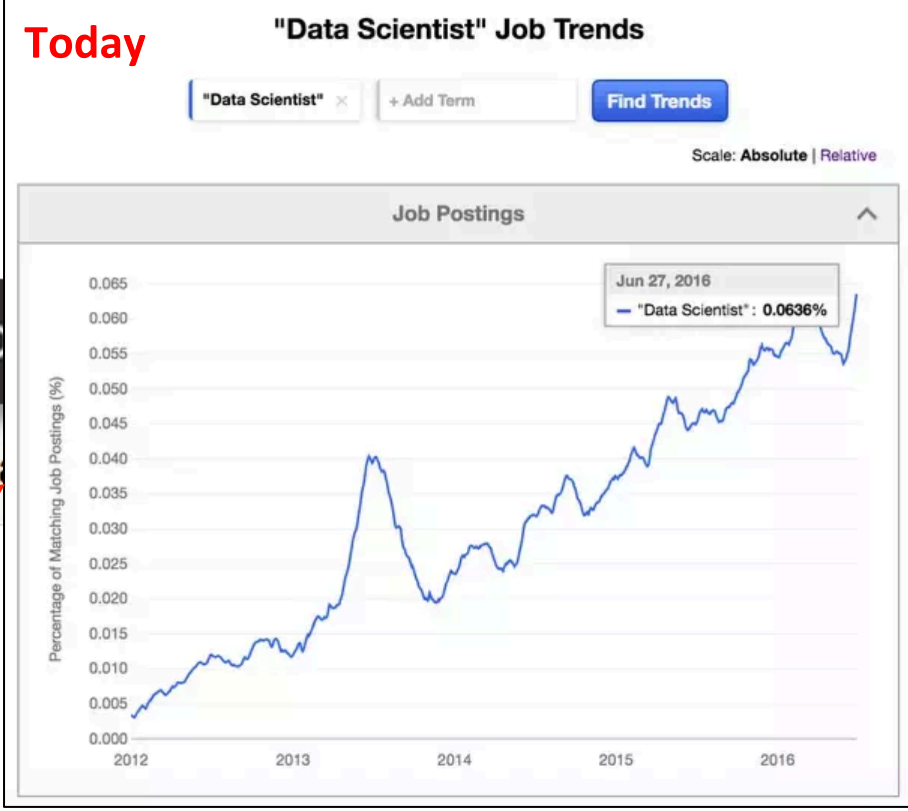
McKinsey Global Institute

May 2011

Harvard Business Review

THE MAGAZINE BLOGS VID

Guest Subscribe today



THE MAGAZINE

October 2012

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

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

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

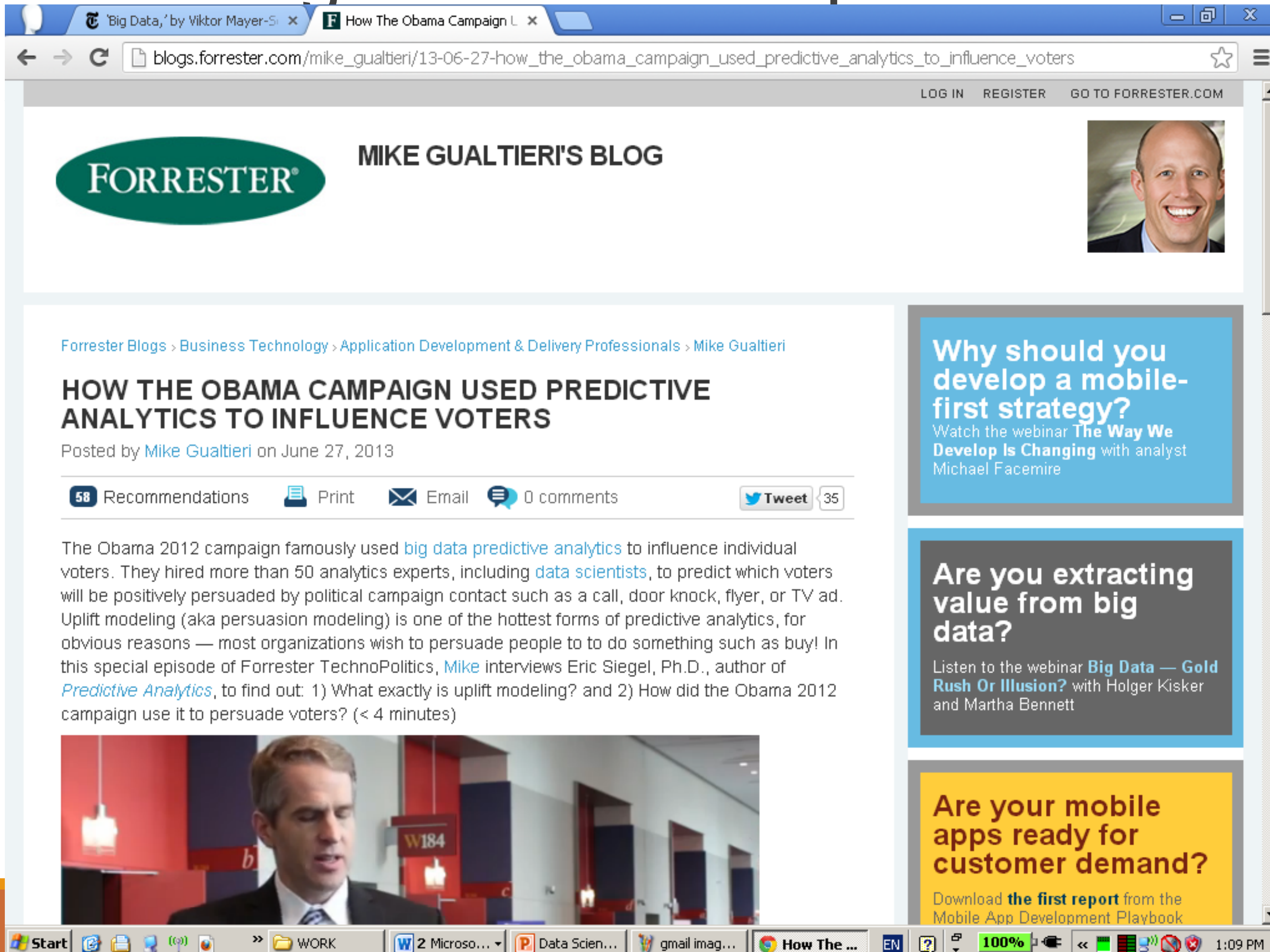
Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine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

TARGET

Target has got you in its aim

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,

Predicting who could be persuaded



blogs.forrester.com/mike_gualtieri/13-06-27-how_the_obama_campaign_used_predictive_analytics_to_influence_voters

LOG IN REGISTER GO TO FORRESTER.COM

FORRESTER®

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 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)

By correlating subjects' Facebook Likes with their OCEAN scores—a standard-bearing personality questionnaire used by psychologists—the team was able to identify an individual's gender, sexuality, political beliefs, and personality traits based only on what they had liked on Facebook.

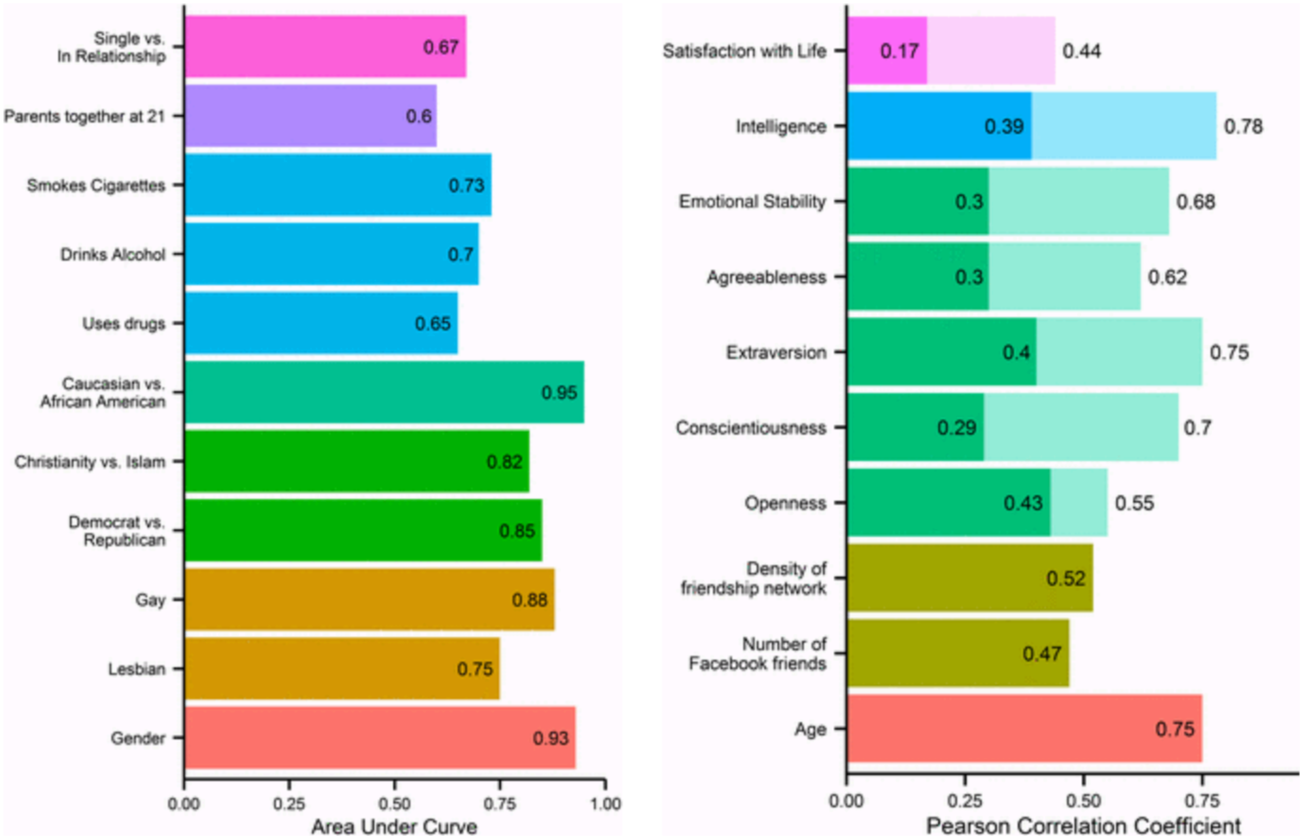
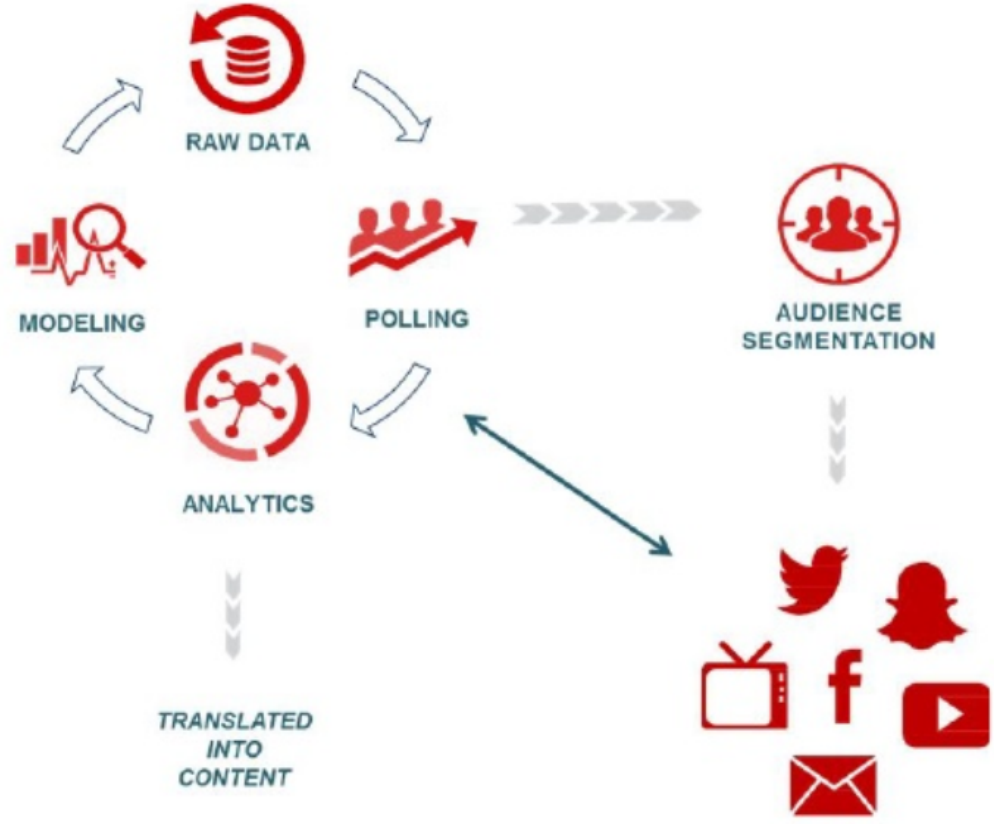


Image Credit: Michal Kosinski, David Stillwell, and Thore Graepel

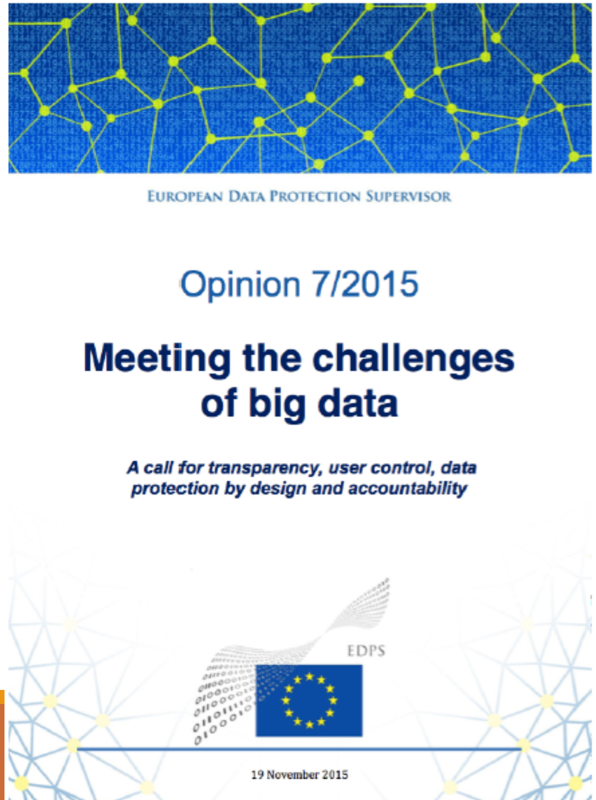
Persuasion Digital Marketing: Process

- Ingested data and audience profiles from the data team
- Devised communications to best promote a story to these individuals
- Executed digital ad buys across 30+ inventory sources delivering 1.5 billion impressions



The GDPR

- Will enter into force on 25 May 2018
- Introduces important novelties
  - New Obligations
  - New Rights



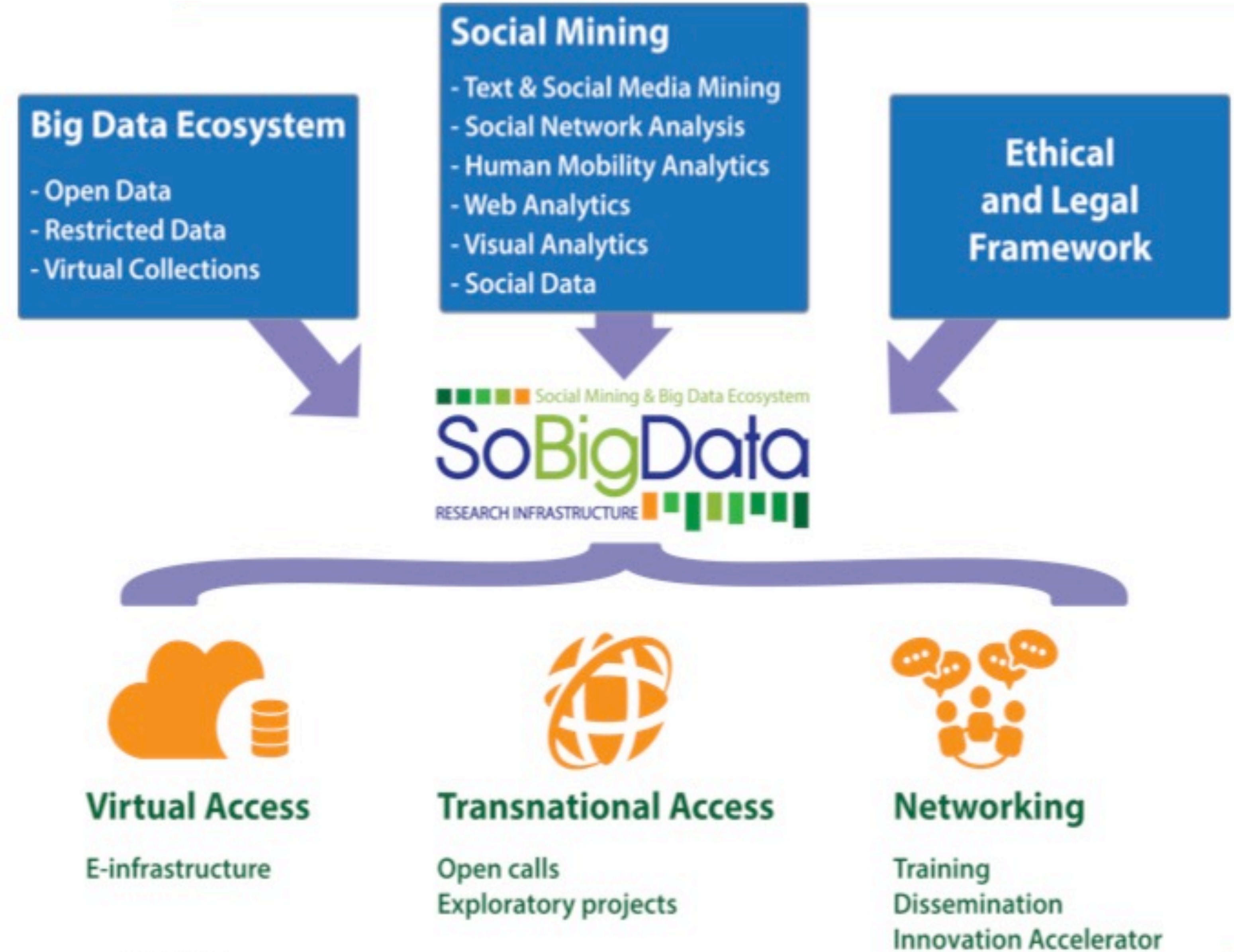
SoBigData Research Infrastructure

SOCIAL MINING & BIG DATA ECOSYSTEM

H2020 - [WWW.SOBIGDATA.EU](http://www.sobigdata.eu)

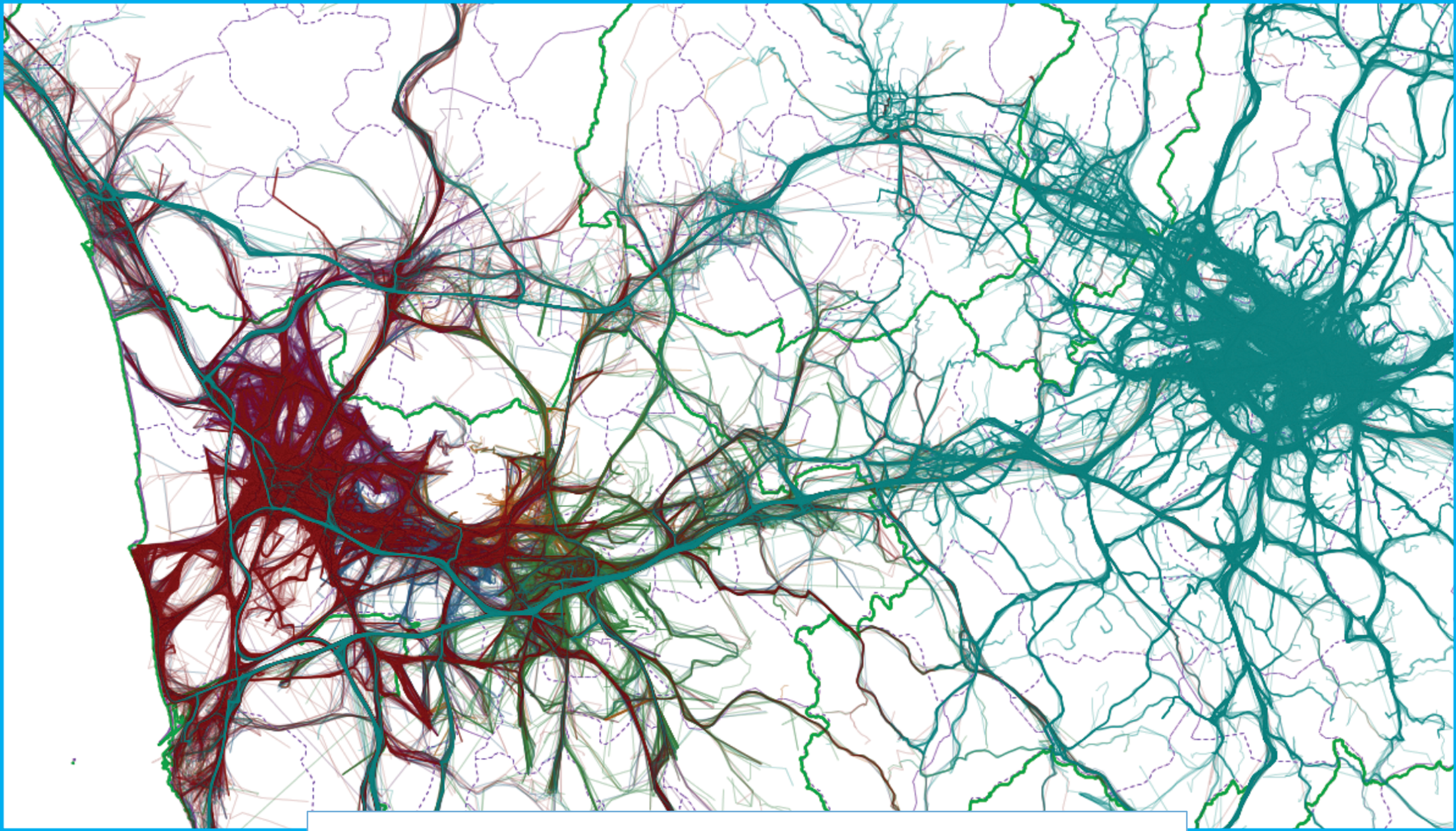
SEPTEMBER 2015- AUGUST 2019

Consiglio Nazionale delle Ricerche
The University Of Sheffield.
Università di Pisa
Fraunhofer FHR
Tartu Ülikool
IMT Institute for Advanced Studies Lucca
Leibniz Universität Hannover
King's College London
Scuola Normale Superiore
Aalto University
ETH Zürich
TU Delft Delft University of Technology


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

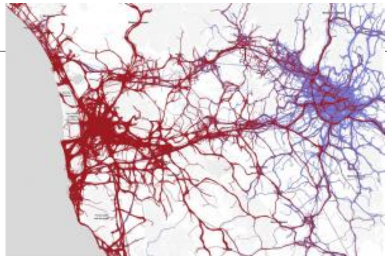
Exploratory:

Big Data for City of Citizens



Personal Mobility, Social + Mobility, Personal Sensing

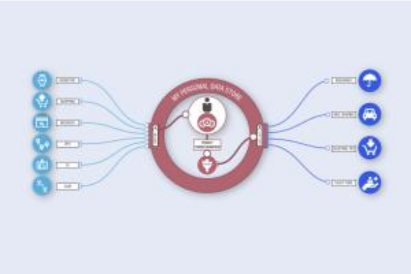
Exploratories



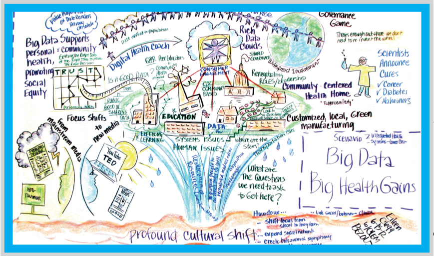
City of Citizens

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

Well-being & Economic Performance



Can Big Data help us to understand relationships between economy and daily life habits? We use data of purchases in supermarkets and investigate people's behavior.



Societal Debates

We study public debates on social media and newspaper. We can identify themes, following the discussions around them and tracking them through time and space.

Migration Studies

Could Big Data help to understand the migration phenomenon? We try to answer to some questions about migrations in Europe and in the world.

Vision papers

1. 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
2. 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
3. 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