Your Data, Your Creativity!

BIGDATA TECH 2017

Intelligent Application of Big Data: Big Opportunities and Big Risks

Anna Monreale

Dipartimento di Informatica

Università di Pisa

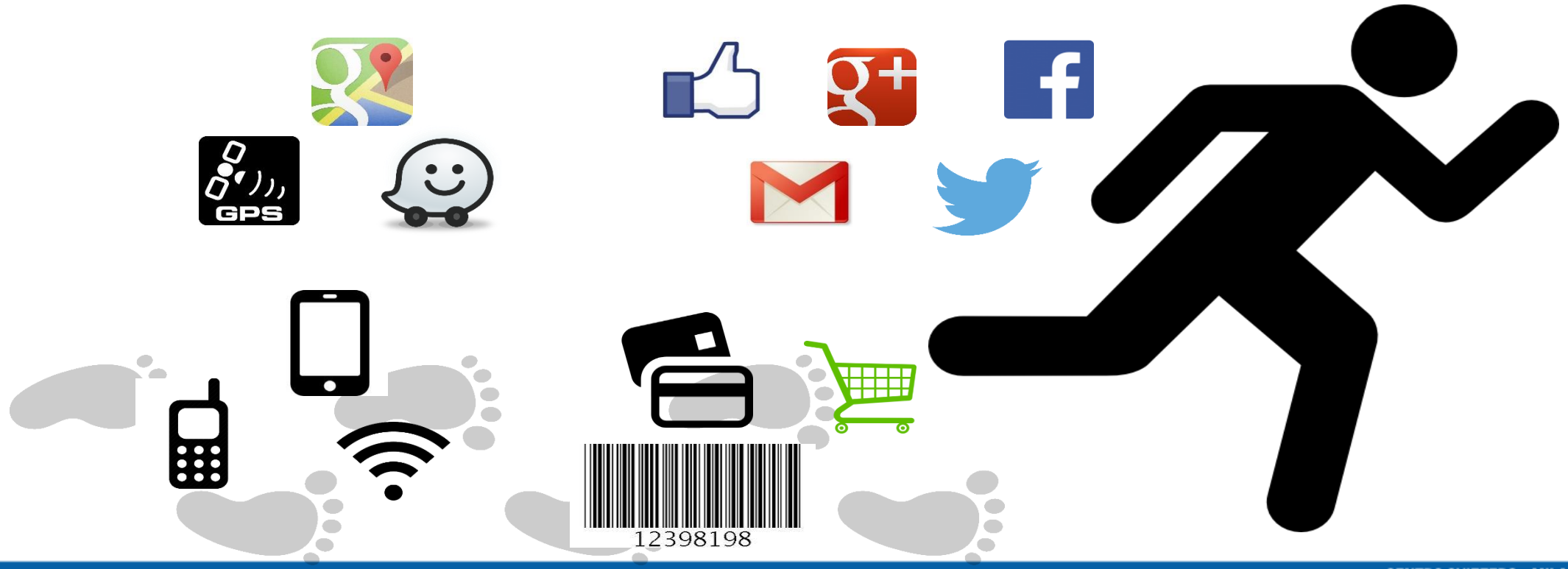
KDD Lab

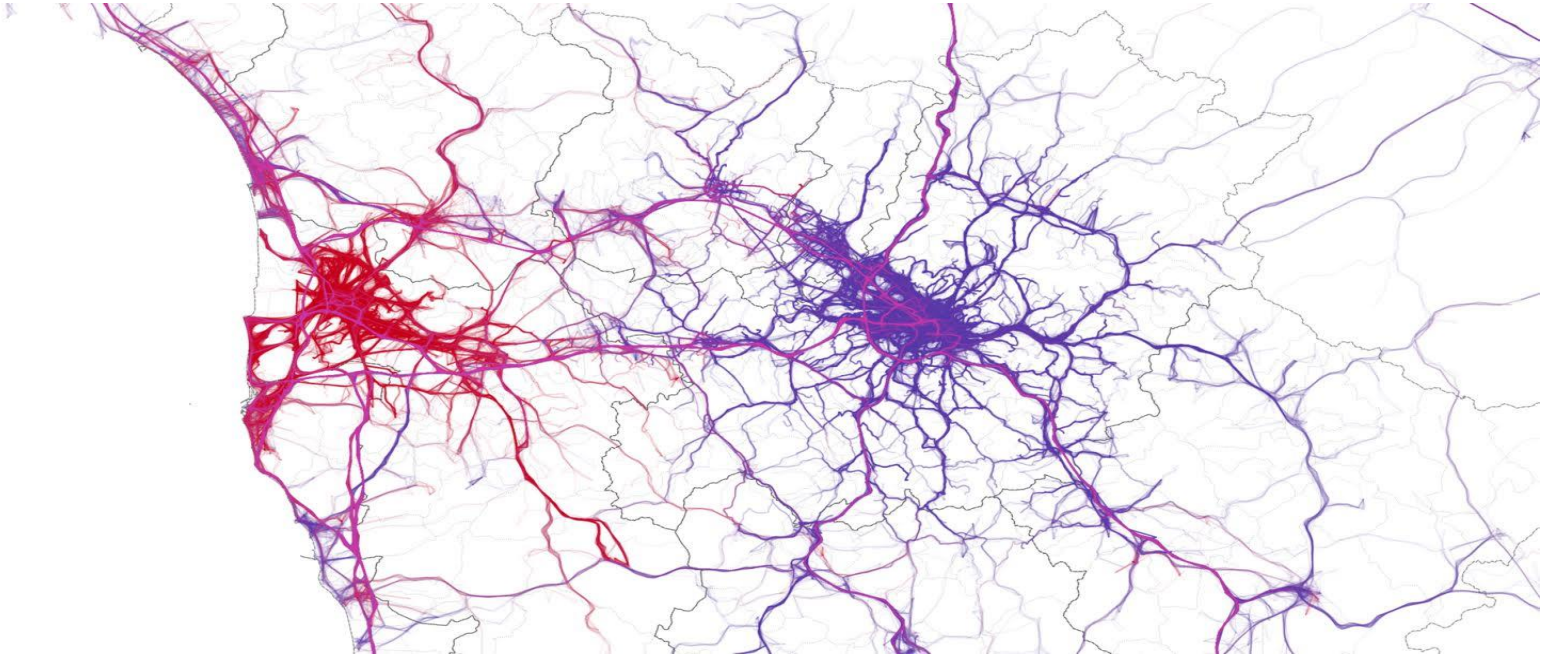
CENTRO SVIZZERO - MILANO - 12 OTTOBRE 2017


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

Our digital traces ….

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





Big data opportunities

Mobility atlas of many cities

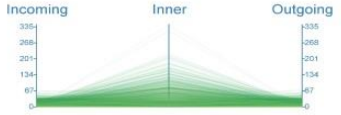
Pisa

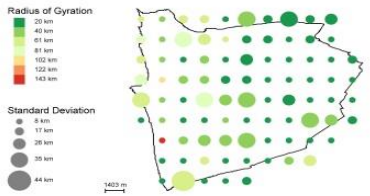
Surface area: 193 km²

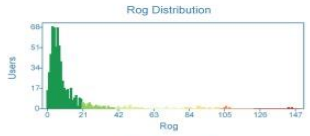
Coordinates: 43,67 10,35

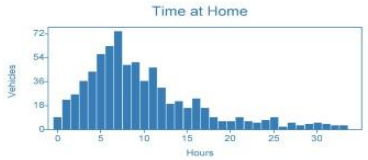
Vehicles: 13.193

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

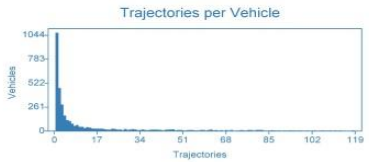


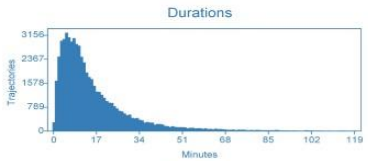


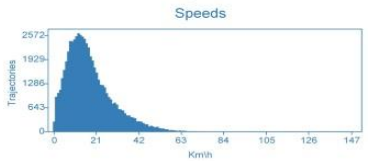


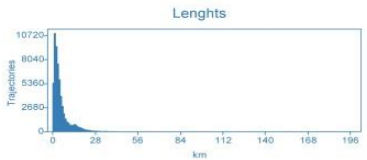


Inner Traffic (44.435 Trajectories)

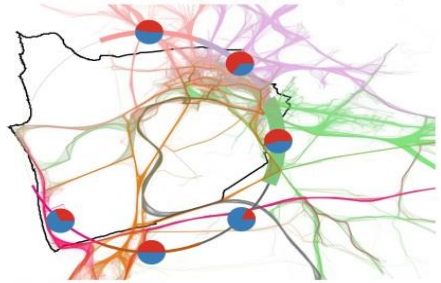






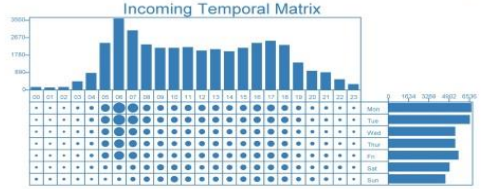


Incoming Traffic (38.464 Trajectories)

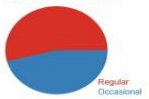


	City	Traj	Perc
NORD 32%	San Giuliano T..	4.916	62%
	Vecchiano	1.425	94%
	Viareggio	1.142	99%
	Lucca	862	67%
	Camaiore	358	94%
OVEST 0%			
SUD 12%	Livorno	2.843	92%
	Collesalvetti	565	50%
	Rosignano Mari..	140	41%
	Fauglia	137	19%
	Cecina	124	45%
EST 54%	Cascina	7.078	97%
	San Giuliano T..	2.881	37%
	Pontedera	1.350	95%
	Calci	795	79%
	Calcinaia	693	92%

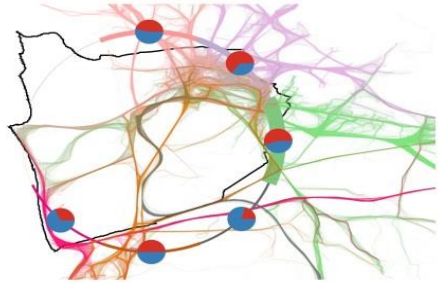
Incoming Temporal Matrix



Regular VS Occasional

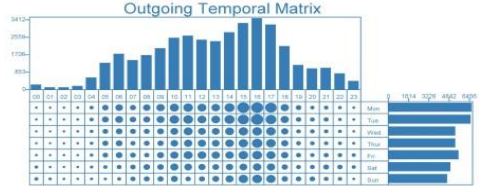


Outgoing Traffic (38.271 Trajectories)

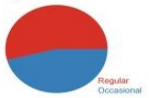


	City	Traj	Perc
NORD 32%	San Giuliano T..	4.842	62%
	Vecchiano	1.418	93%
	Viareggio	1.117	99%
	Lucca	886	67%
	Camaiore	329	96%
OVEST 0%			
SUD 13%	Livorno	2.812	92%
	Collesalvetti	565	51%
	Rosignano Mari..	143	44%
	Fauglia	130	19%
	Cecina	123	45%
EST 54%	Cascina	7.253	97%
	San Giuliano T..	2.860	37%
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

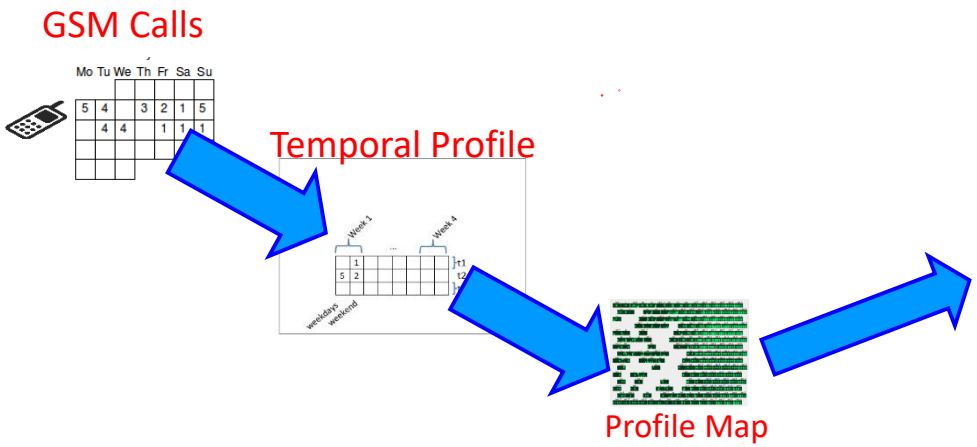
Outgoing Temporal Matrix

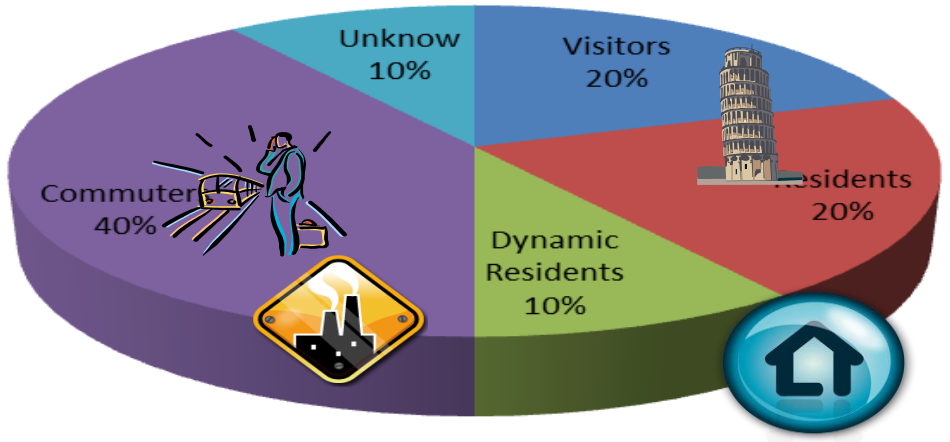


Regular VS Occasional



A Sociometer based on Mobile Phone Data for Real Time Demographics





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

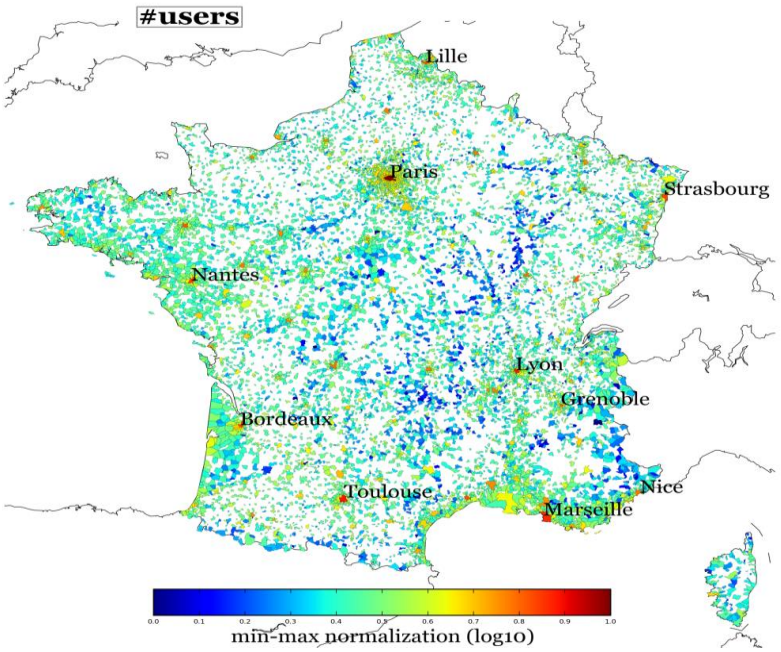
KDD Lab

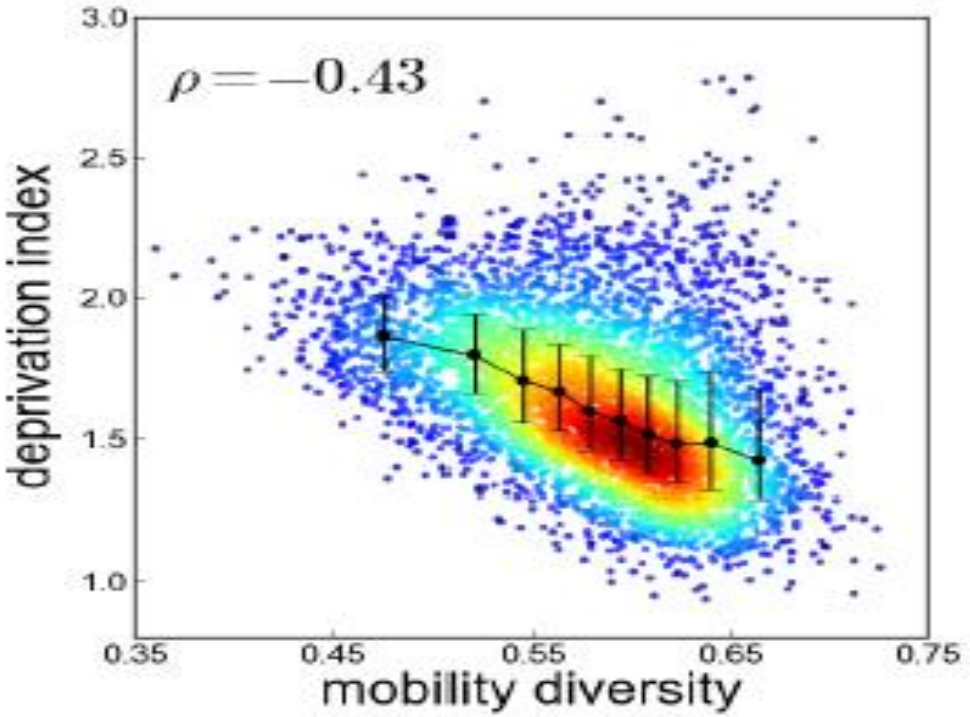
UNIVERSITÀ DI PISA

Istat

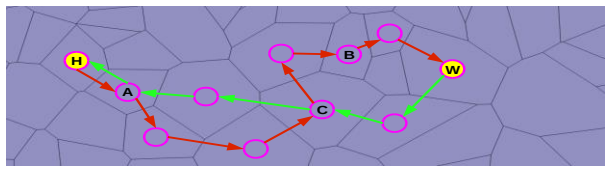
WIND
Più vicini.

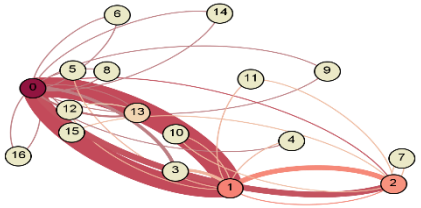
Big Data: Diversity and economic development



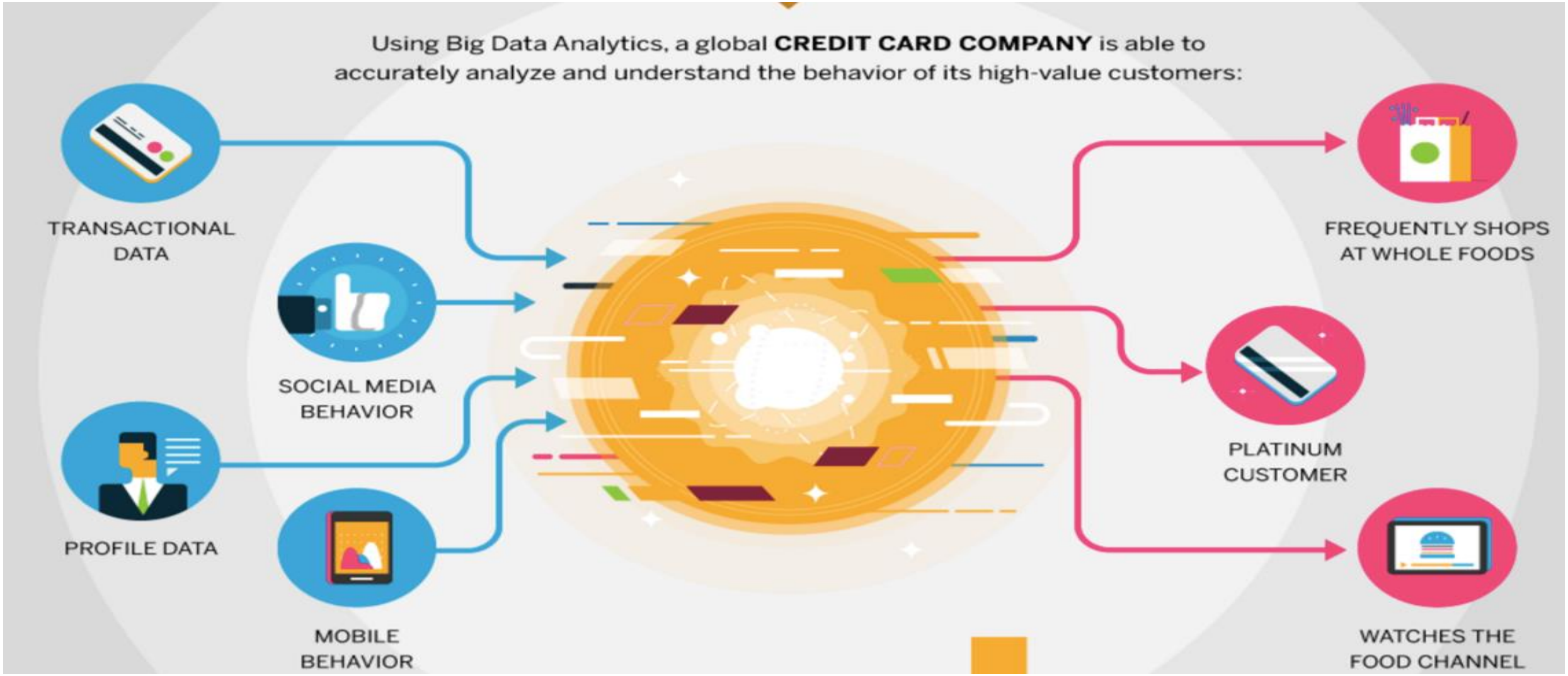


Mobility diversity is correlated with **wellbeing indicators** and **socio-economic development** (Income, Deprivation Index, Education level).

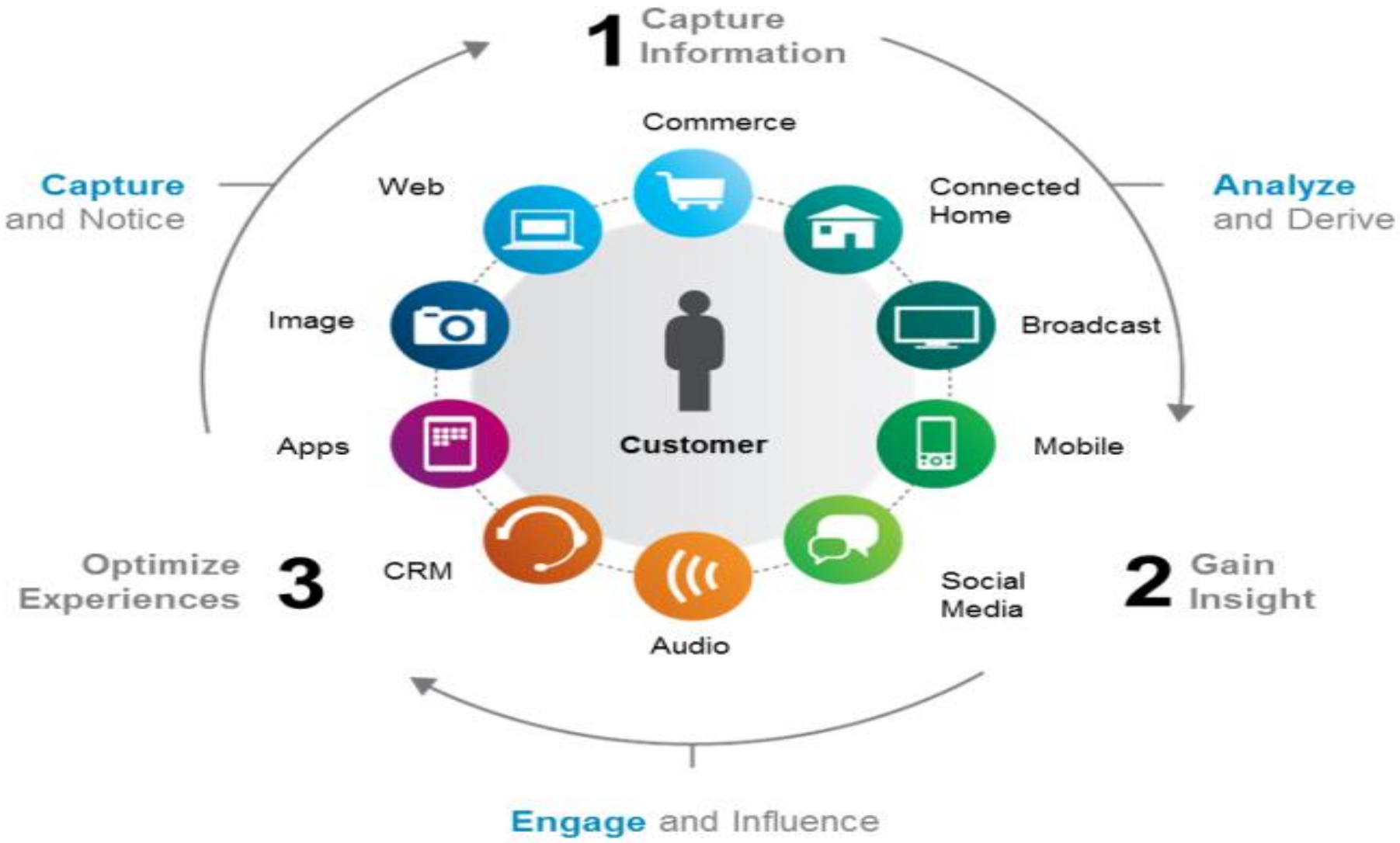




Big Data: from credit card to customer habits



Big Data: new, more carefully targeted financial services



Big Data Analytics & Social Mining

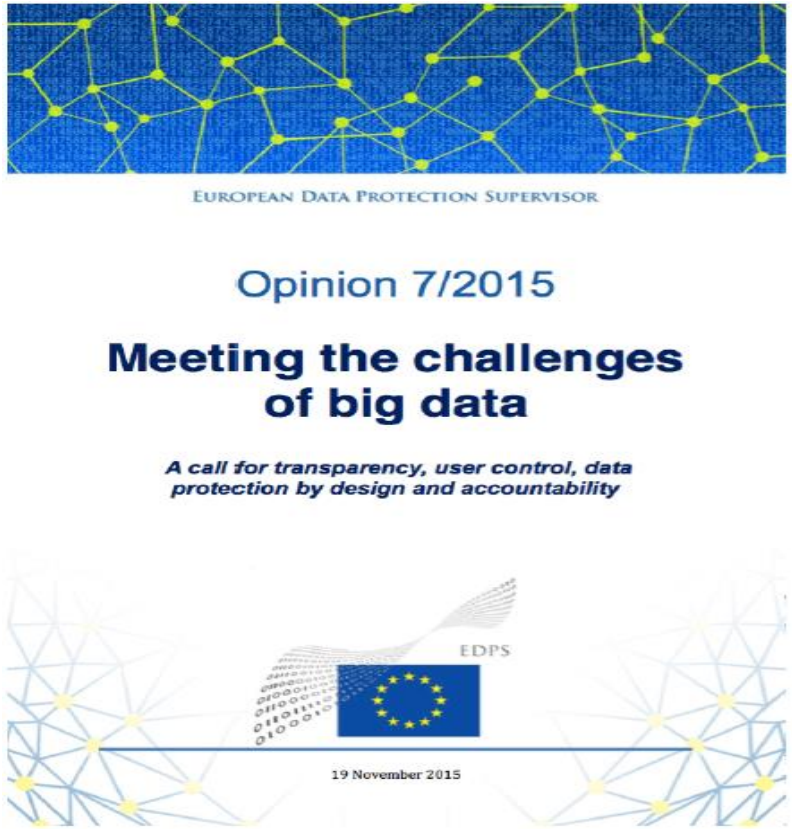
The **main tool** for a
Data Scientist to
measure,
understand,
and possibly predict
human behavior

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

Risks: Privacy, fairness

E.U. - EDPS

- secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2015/15-11-19_Big_Data_EN.pdf



General Data Protection Regulation
April 2018

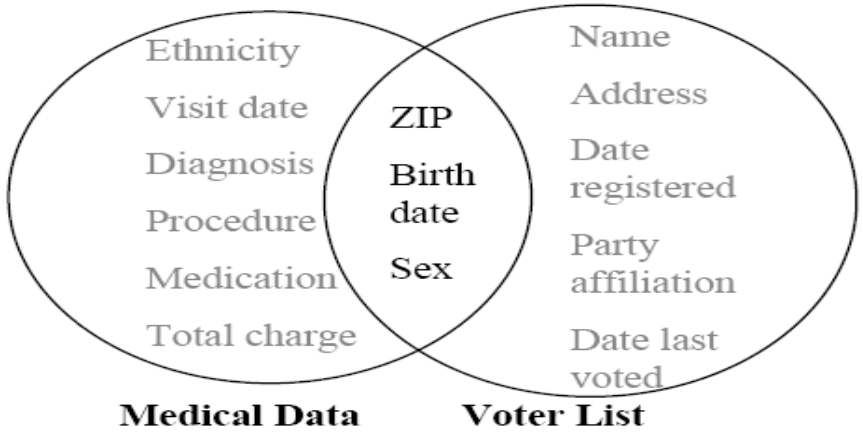
Big Data risks: Privacy

- Any individual has the right to privacy protection
  - The right to be **directly or indirectly non-identifiable**
- Analyze this kind of data also combining them can bring to individual privacy violation
- The new EU Privacy Regulation requires that the data Controller maintains an updated report on the privacy risk assessment on perosnal data collected

Re-identification of Massachussetts' governor

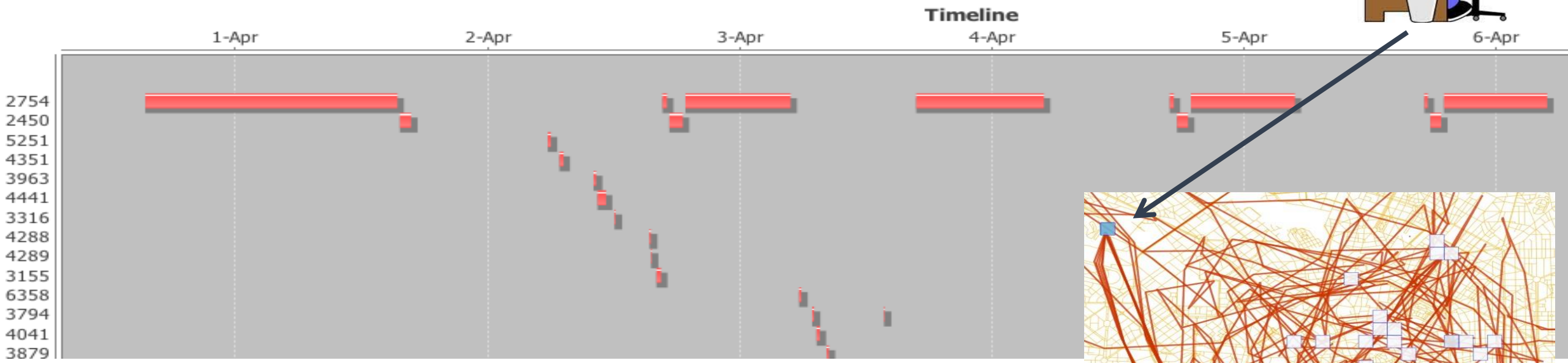
- Sweeney managed to re-identify the medical record of the governor of Massachussetts
  - MA collects and publishes sanitized medical data for state employees (microdata) left circle
  - voter registration list of MA (publicly available data) right circle

- looking for governor's record
- join the tables:
  - 6 people had his birth date
  - 3 were men
  - 1 in his zipcode



Latanya Sweeney: k-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5): 557-570 (2002)

De-identified User Trajectory



Discovering persons living in that home and working in that company we can identify the user

The need of Privacy by design big data analytics

- Design analytical process that implement the **privacy-by-design** principle

- **Consider privacy at every stage of their business**
- **Integrate privacy requirements “by design” into their business model.**

Privacy by Design Methodology in Big Data Analytics

- The framework is designed with assumptions about
  - The **sensitive data** that are the subject of the analysis
  - The **attack model**, i.e., the knowledge and purpose of a malicious party that wants to discover the sensitive data
  - The **target analytical questions** that are to be answered with the data

- Design a privacy-preserving framework able to
  - transform the data into an anonymous version with a **quantifiable privacy guarantee**
  - guarantee that the analytical questions can be answered correctly, within a **quantifiable** approximation that specifies the **data utility**

Your Data, Your Creativity!

Privacy by Design in Mobility Atlas

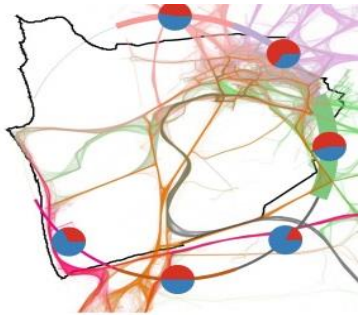
BIGDATA TECH 2017

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

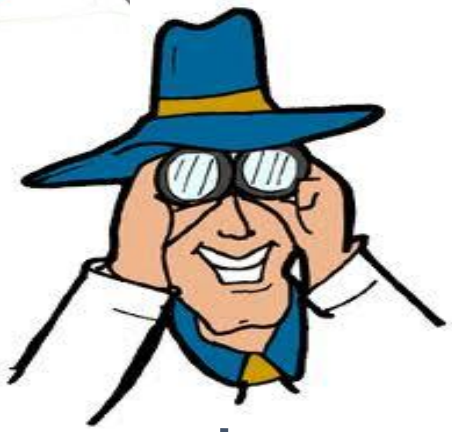
CENTRO SVIZZERO - MILANO - 12 OTTOBRE 2017

Privacy-Preserving Framework

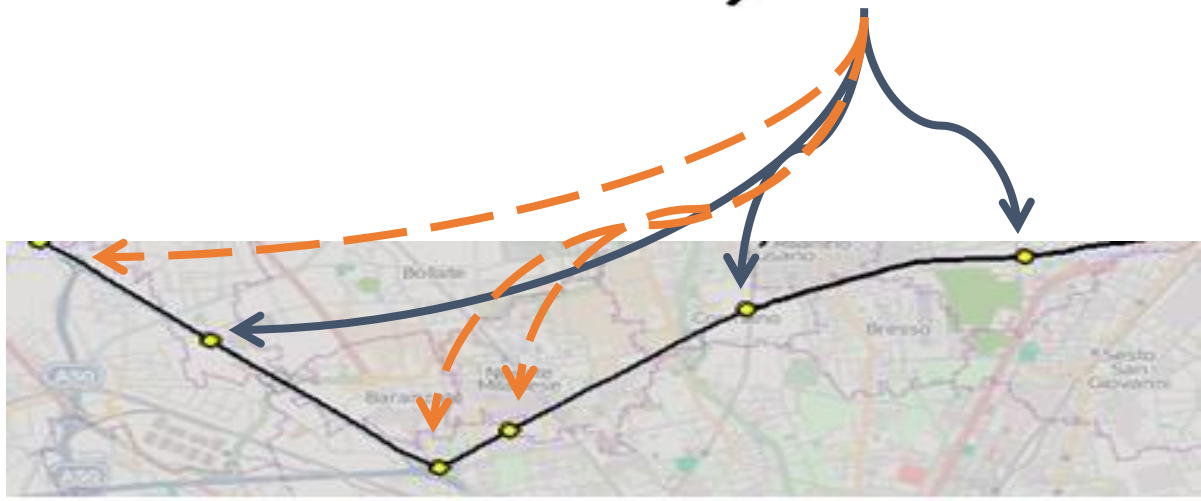
- Anonymization of movement data while preserving clustering



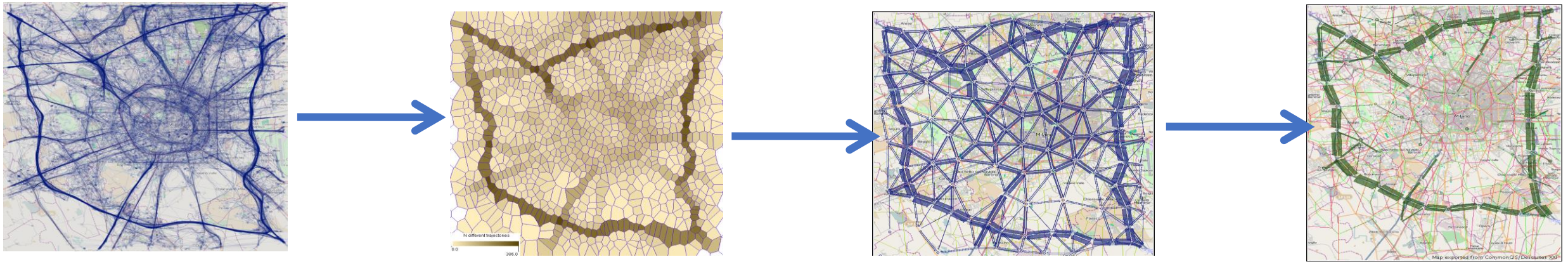
- **Trajectory Linking Attack**: the attacker
  - knows some points of a given trajectory
  - and wants to infer the whole trajectory



- **Countermeasure**: method based on
  - **spatial generalization** of trajectories
  - **k-anonymization** of trajectories

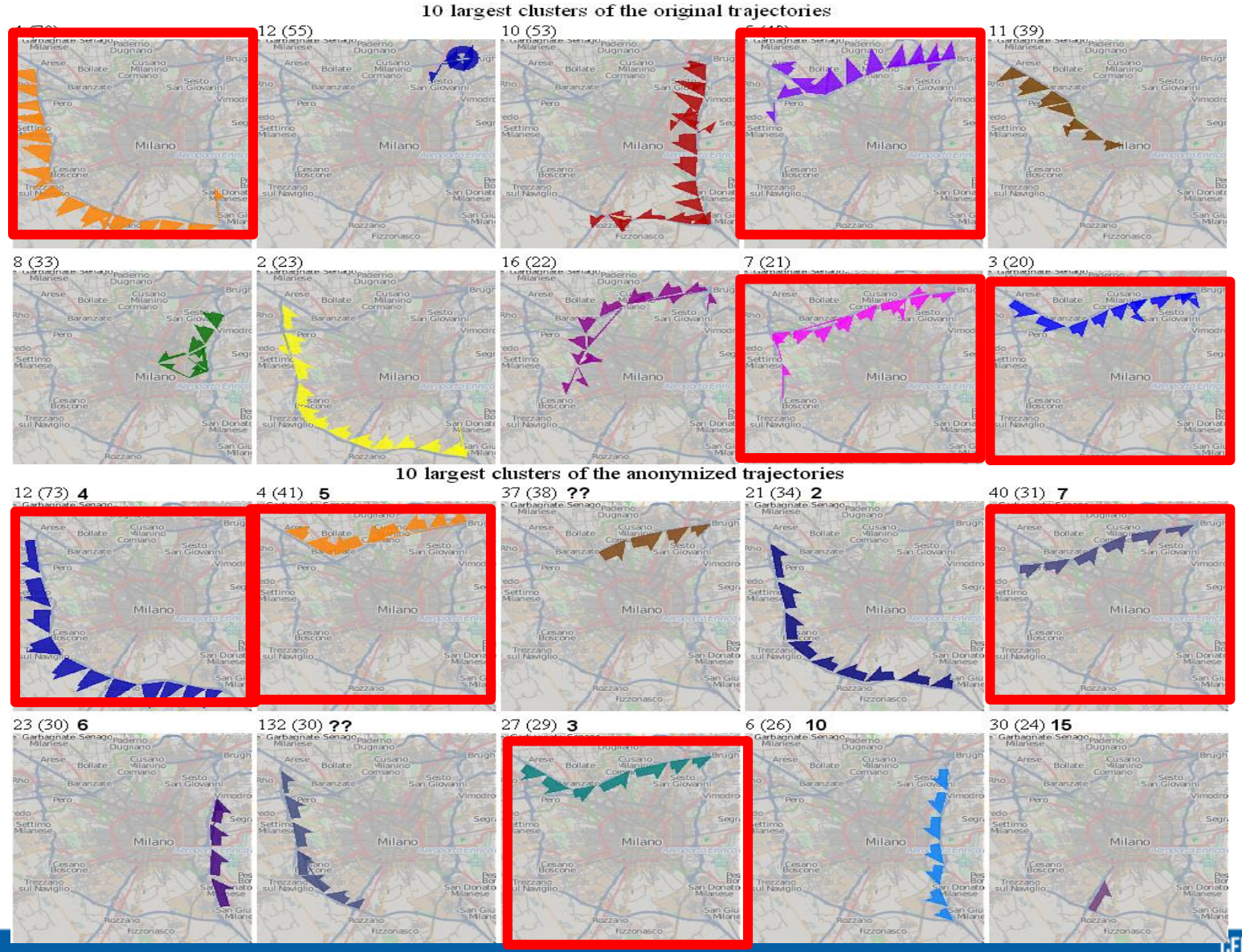


Trajectory Anonymization



- Given a trajectoy dataset
  1. Partition of the territory into **Voronoi cells**
  2. Transform trajectories into sequence of cells
  3. Ensure k-anonymity:
     - ➢ For each generalized trajectory there exist at least others k-1 different people with the same trajectory? If not transform data in similar ones.

Clustering on Anonymized Trajectories



Your Data, Your Creativity!

Data-Driven Privacy Risk Assessment

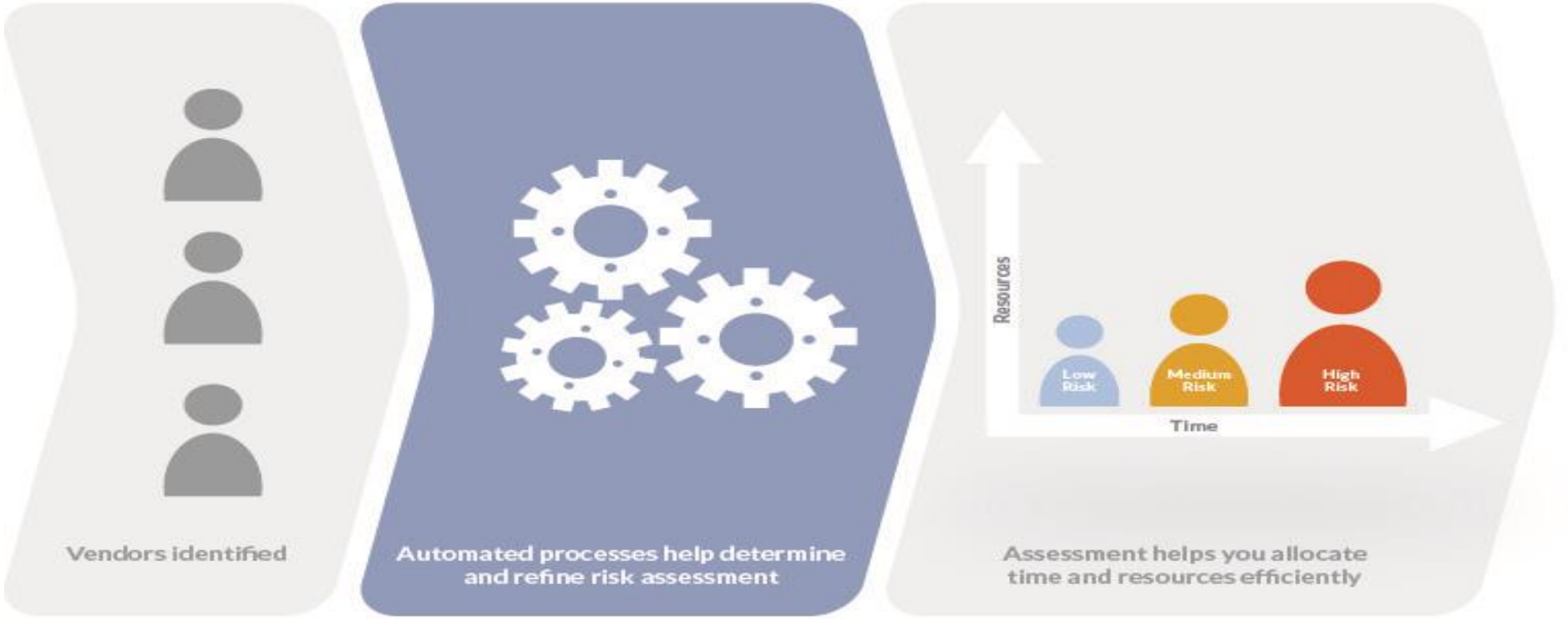
BIGDATA TECH 2017

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

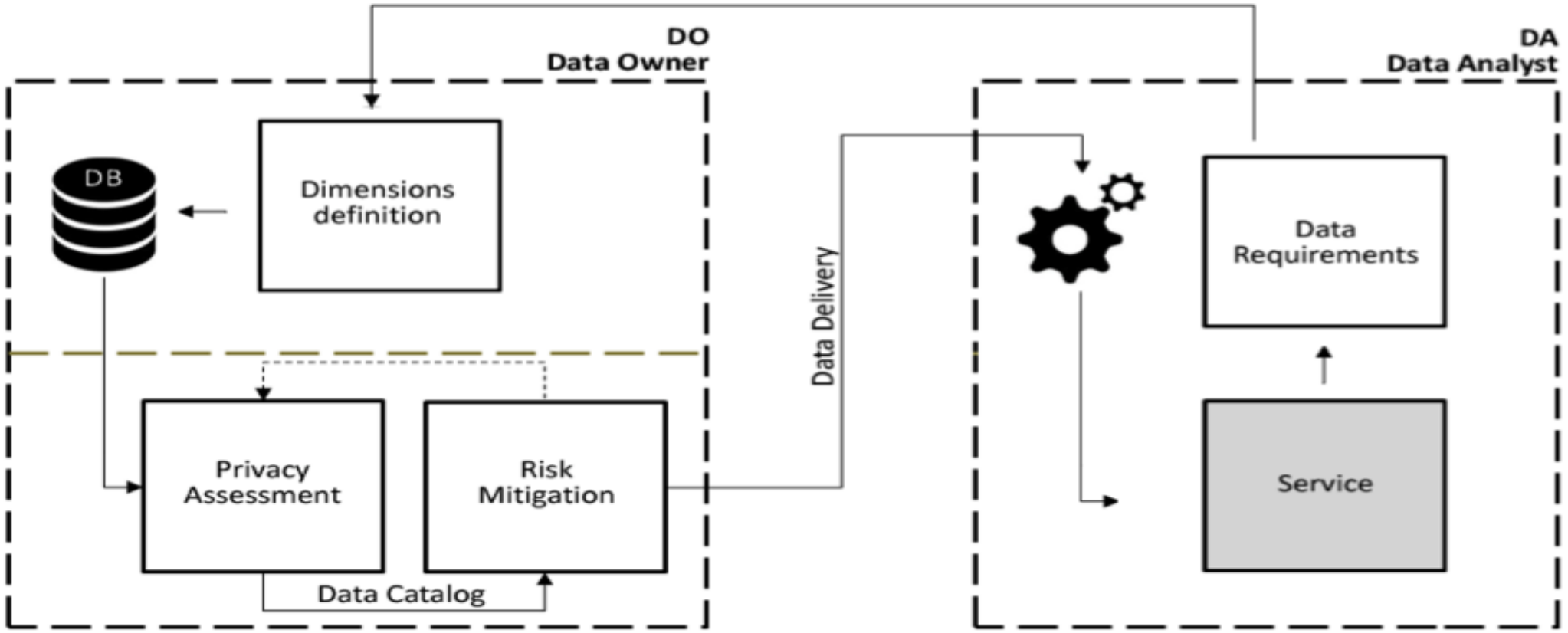
CENTRO SVIZZERO - MILANO - 12 OTTOBRE 2017

Privacy Risk Assessment

- Tools for a systematic assessment of the privacy risks on personal data that could be exploited for improving or offering ne knowledge-based services



Privacy Risk Assessment Framework for Data Sharing



Privacy Risk Assessment Framework for Data Sharing

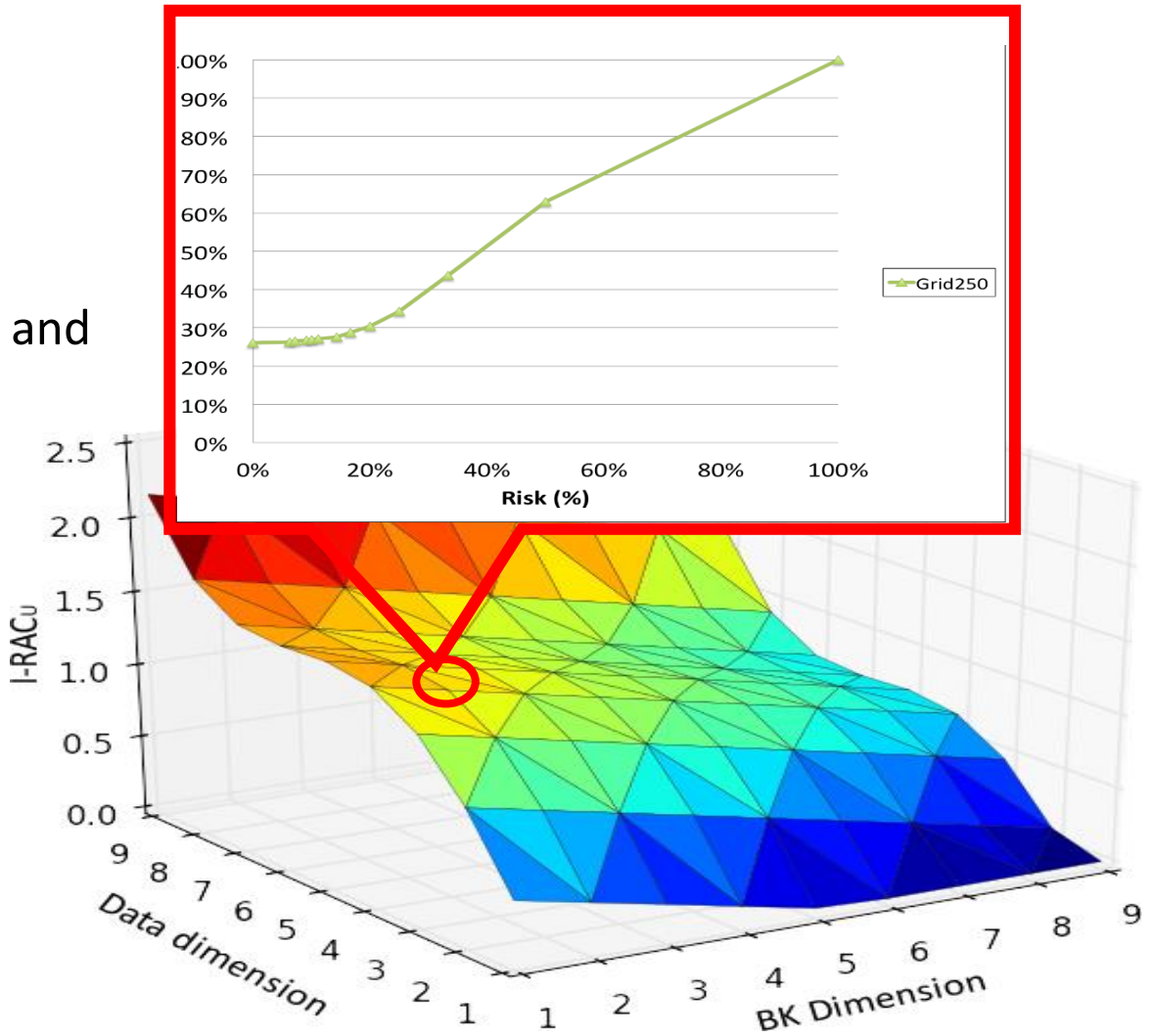
Data Catalog

For each:

- **Data Format**, i.e., the data needed for the service
- **Risk Assessment Setting**, i.e., the set of pre-processing and privacy attacks

The Data Catalog provides:

- **Quantification of Privacy Risk**, i.e., the evaluation of the real risk of re-identification
- **Quantification of Data Quality**, i.e., the quality level we can achieve with private data, compared with the data quality of original data.



Your Data, Your Creativity!

Fairness in Decision Systems and Profiling

BIGDATA TECH 2017

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

CENTRO SVIZZERO - MILANO - 12 OTTOBRE 2017

Big Data Risks: Fairness

More at www.propublica.org/article/what-we-know-about-the-computer-formulas-making-decisions-in-your-life

When Algorithms Discriminate

The New York Times, July 9, 2015

A recent Carnegie Mellon study found that Google was showing ads for high-paying jobs to more men than women. Another study from Harvard showed that Google searches for "black-sounding" names yielded suggestions for arrest-record sites more often than other types of names. Algorithms are often described as "neutral" and "mathematical," but as these experiments suggest, they can also reproduce and even reinforce bias.

Discrimination-aware Data Mining

Dino Pedreschi Salvatore Ruggieri Franco Turini
Dipartimento di Informatica, Università di Pisa
L.go B. Pontecorvo 3, 56127 Pisa, Italy
{pedre,ruggieri,turini}@di.unipi.it

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

WIRED

Who do you blame when an algorithm gets you fired?

Profiling

Profile:

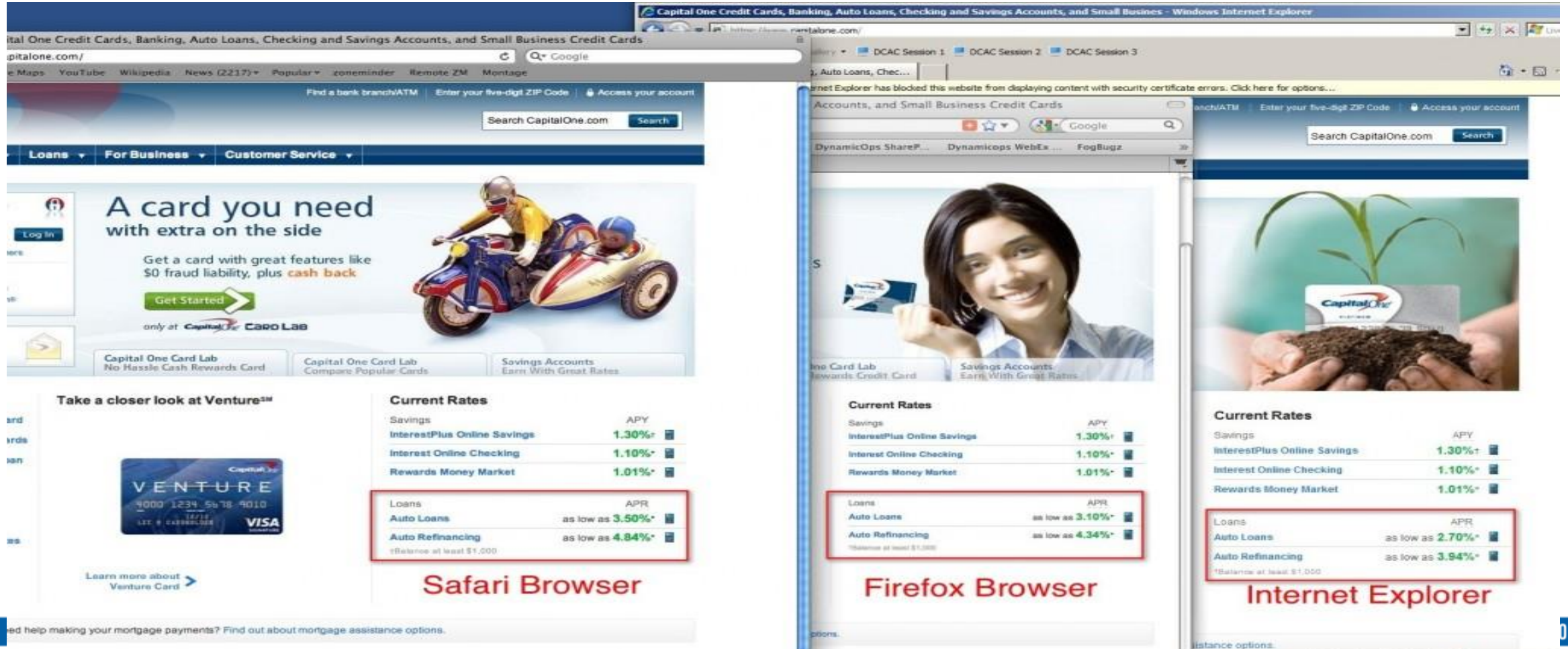
- a set of data characterising a category of individuals that is intended to be applied to an individual

Profiling

An automatic data processing technique that consists of applying a “profile” to an individual, particularly in order to take decisions concerning her or him or for analysing or predicting her or his personal preferences, behaviours and attitudes

Fairness

- Formal equality
  - «People that are alike should be treated alike» (Aristotele)
    - Equality of opportunities



Fairness

- Formal equality
  - «People that are alike should be treated alike» (Aristotele)
    - Equality of opportunities
- Discrimination
  - An unjustified distinction of individuals based on their membership, or perceived membership, in a certain group or category.
  - Human rights laws prohibit discrimination on the grounds of
    - sex, gender, sexual orientation, race, ethnicity , skin color, social origin, genetic features, language, religion or belief, political or other personal opinion, membership of a national minority, property, birth, parentage, disability, illness, marital status, or age.

Discrimination

What is discrimination?

- An unjustified distinction of individuals based on their membership, or perceived membership, in a certain group or category, without regard to individual characteristics
  - Protected-by-law groups on the grounds of sex, gender, sexual orientation, race, ethnicity, skin, language, religion, personal opinions, membership of a minority, disability, illness, marital status, age, …

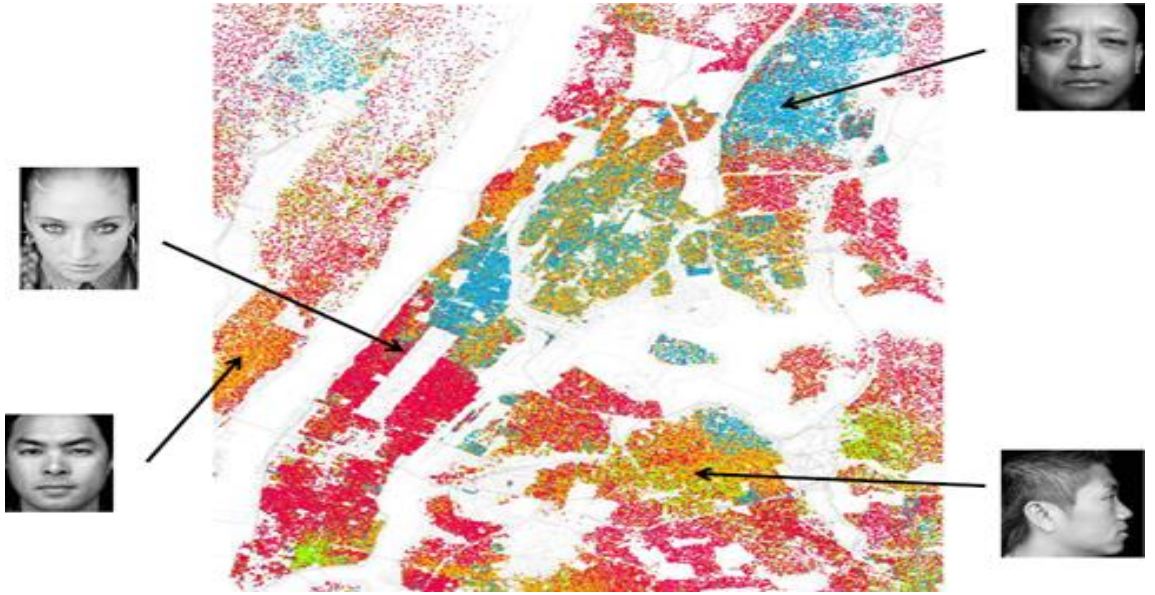
- Direct discrimination consists of rules or procedures that explicitly impose disproportionate burdens on minority or disadvantaged groups

- Indirect discrimination consists of rules or procedures that, while not explicitly mentioning discriminatory attributes, intentionally or not impose the same disproportionate burdens.

Indirect discrimination: Redlining

- Racial segregation
  - imply that ZIP and race are correlated

- Banks have (sometimes) exploited this
  - to make restrictions on loans to minorities by restricting loans to neighb.
  - even if this means loosing some good customer (*reverse tokenism*)



Your Data, Your Creativity!

New Challenge:
The Right of explanation

BIGDATA TECH 2017

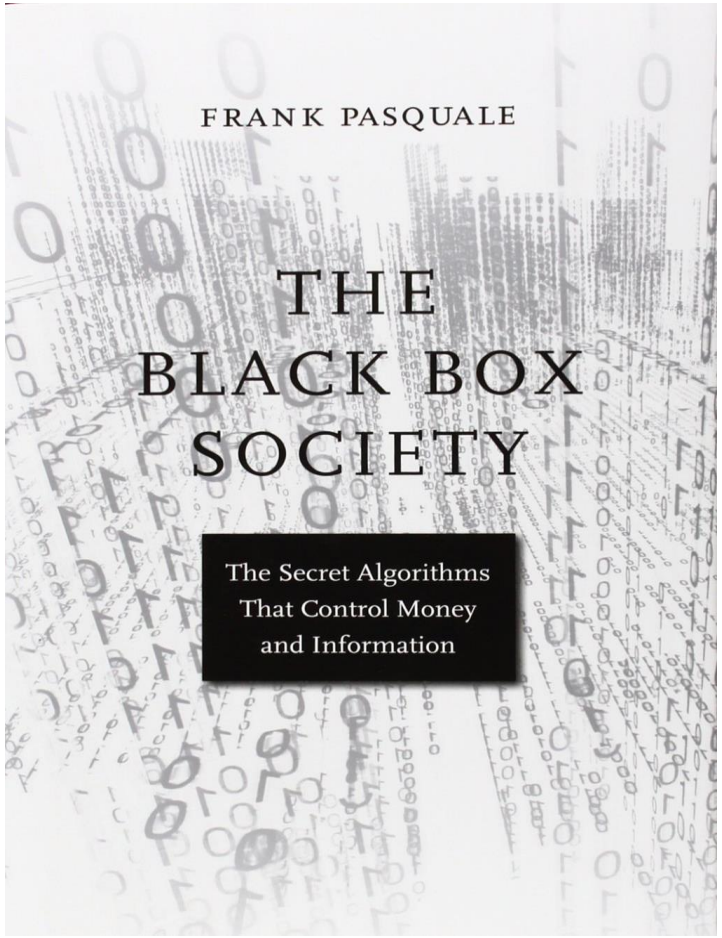
Knowledge Discovery and Delivery Lab
(ISTI-CNR & Univ. Pisa)
www-kdd.isti.cnr.it

Right of explanation

- Applying AI within many domains requires **transparency** and **responsibility**:
  - health care
  - finance
  - surveillance
  - autonomous vehicles
  - Government

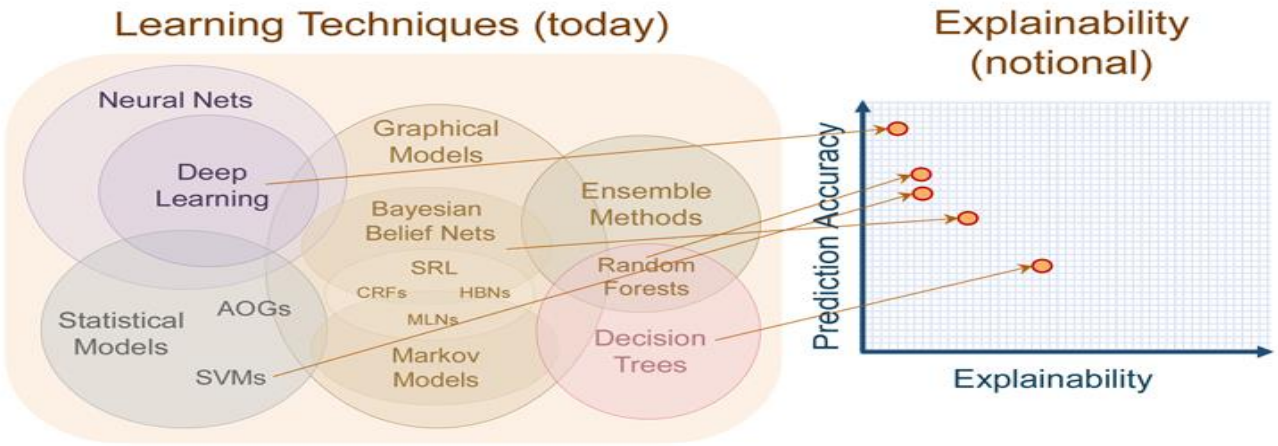
- EU General Data Protection Regulation (April 2016) establishes a right of explanation for all individuals to obtain "meaningful explanations of the logic involved" when automated (algorithmic) individual decision-making, including profiling, takes place.

- In sharp contrast, (big) data-driven mining models are *black boxes*.



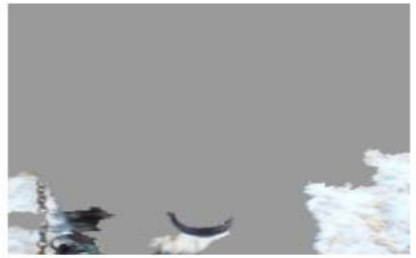
Is data mining Permanently Inscrutable?

- nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable



- "Why Should I Trust You?" Explaining the Predictions of Any Classifier
- KDD 2016 Conference Paper

(a) Husky classified as wolf



(b) Explanation

Accountability

- “Why exactly was my loan application rejected?”
- “What could I have done differently so that my application would not have been rejected?”

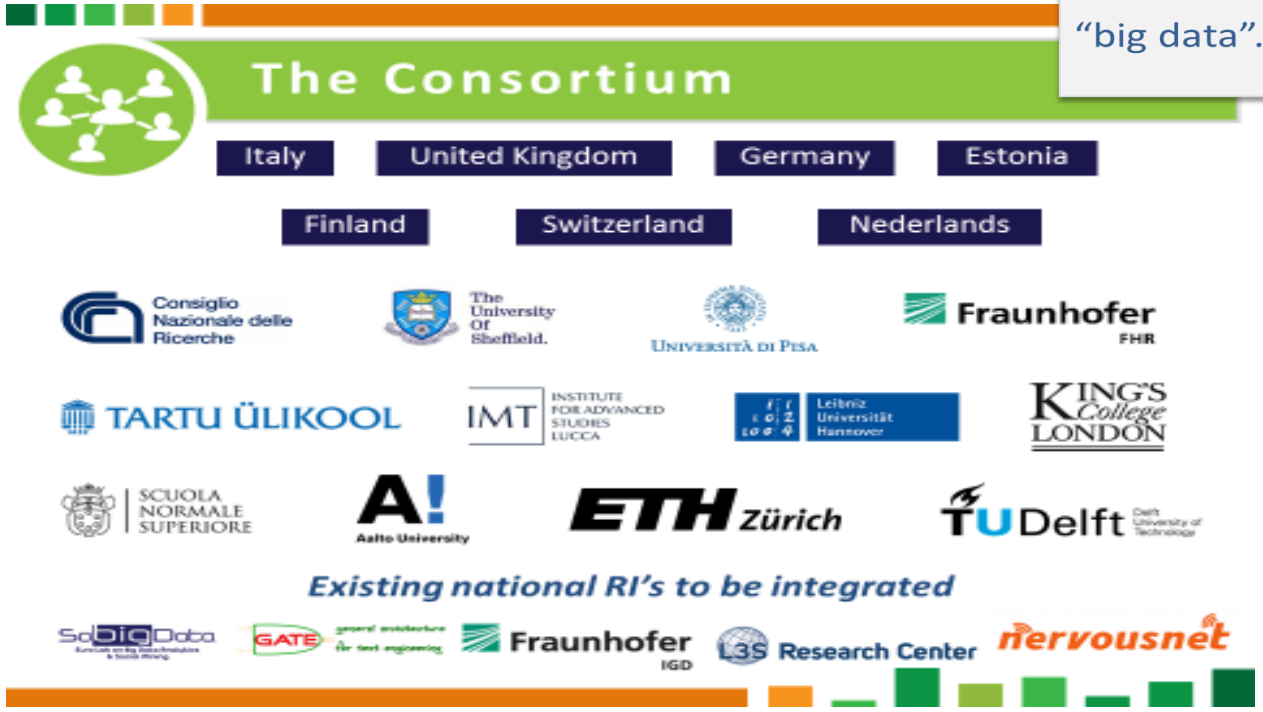
What we need …

- Black boxes map user's features into a class or a score without exposing the reasons why
- Worrying for possible biases and prejudices hidden in the training data and learned by the algorithms or, worse, for unfair rules maliciously introduced by humans.
- Solving the problem of the explanation of black-box decision making
- Develop a logical/statistical framework of data-driven discovery of explanatory rules providing understandable, sound, solid and succinct explanations of black boxes
- A repertoire of tools for explanatory rule discovery, validation, and visualization

EU Projects: SoBigData

Social Mining & Big Data Ecosystem project (SoBigData, H2020-INFRAIA-2014-2015, duration: 2015-2019, www.sobigdata.eu

A **Multidisciplinary** European **Infrastructure** for **Big Data** and **Social Data Mining** providing an integrated ecosystem for **ethically sensitive scientific discoveries** and advanced applications of social data mining on the various dimensions of social life, as recorded by "big data".



Coordinator:
Fosca Giannotti,
ISTI-CNR, Pisa

Data Ethics Literacy

- Rapporto MIUR su Big Data, 28 Luglio 2016
  - www.istruzione.it/allegati/2016/bigdata.pdf
- Master UNIPI in Big Data Analytics & Social Mining
  - www.masterbigdata.it

http://www.masterbigdata.it

Deadline for Applications:

17 November 2017

Knowledge Discovery & Data Mining Lab
http://kdd.isti.cnr.it

Thank you!


BIG DATA
TRANSPARENCY
USER CONTROL
DIGNITY
FREEDOM
PRIVACY BY DESIGN
ACCOUNTABILITY
PRIVACY