



Your Data, Your Creativity!

Intelligent Application of Big Data: Big Opportunities and Big Risks

BIGDATA
TECH 2017

Anna Monreale

Dipartimento di Informatica

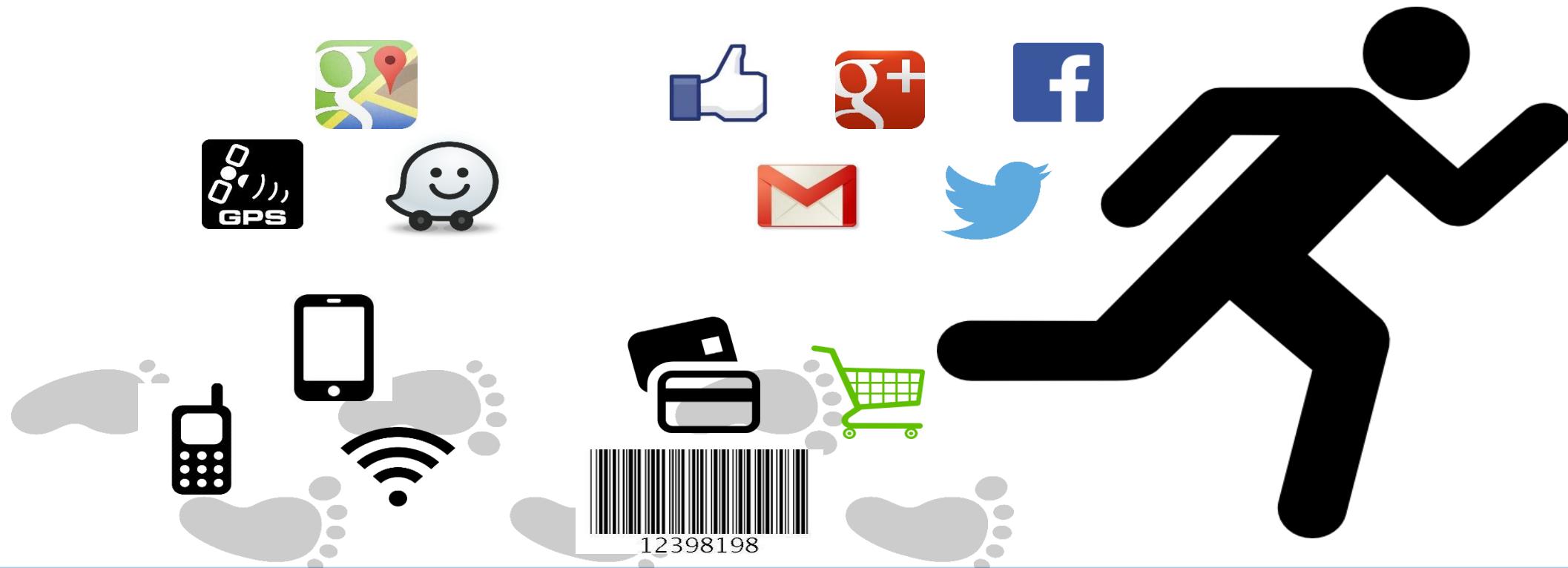
Università di Pisa

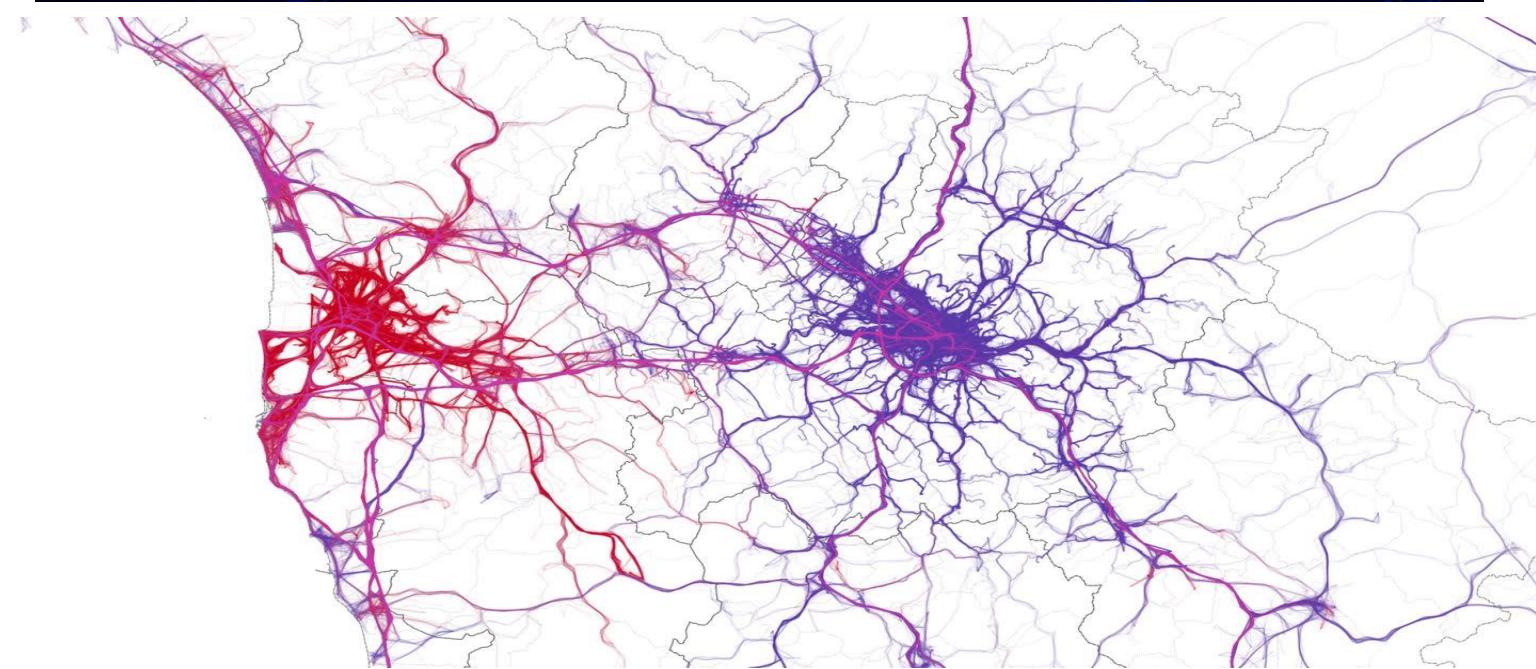




Our digital traces

- 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

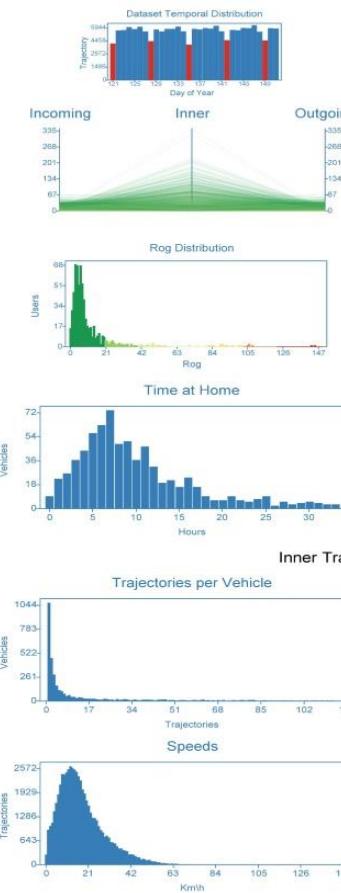
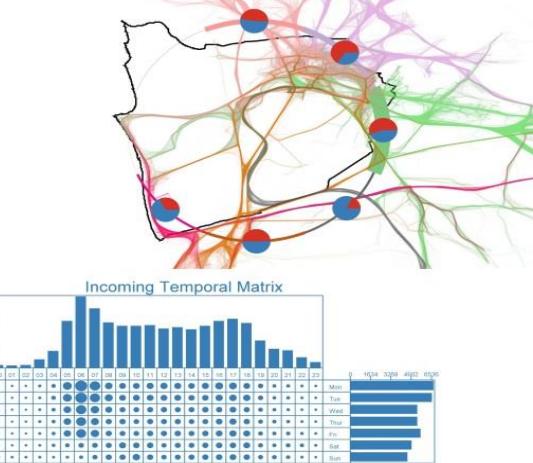
Mobility atlas of many cities

PisaSurface area: 193 km²

Coordinates: 43.67 10.35

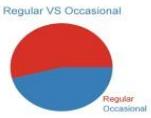
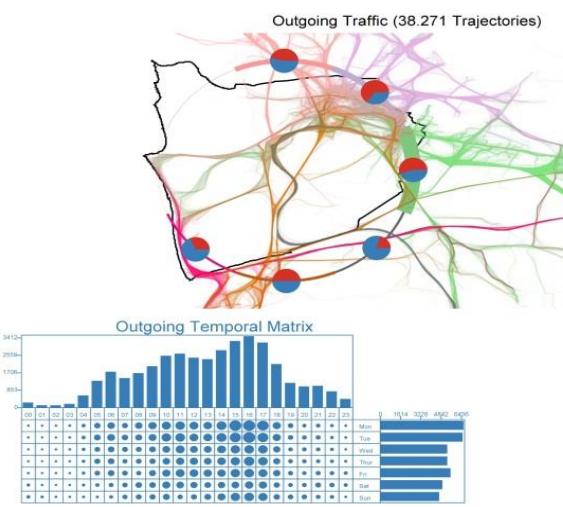
Vehicles: 13.193

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

**Incoming Traffic (38.464 Trajectories)**

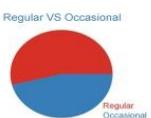
City	Traj	Perc
San Giuliano T.	4.916	62%
Vecciano	1.425	94%
Viareggio	1.142	99%
Lucca	892	87%
Camaiore	358	94%

NORD 32%
OVEST 0%
SUD 12%
EST 54%

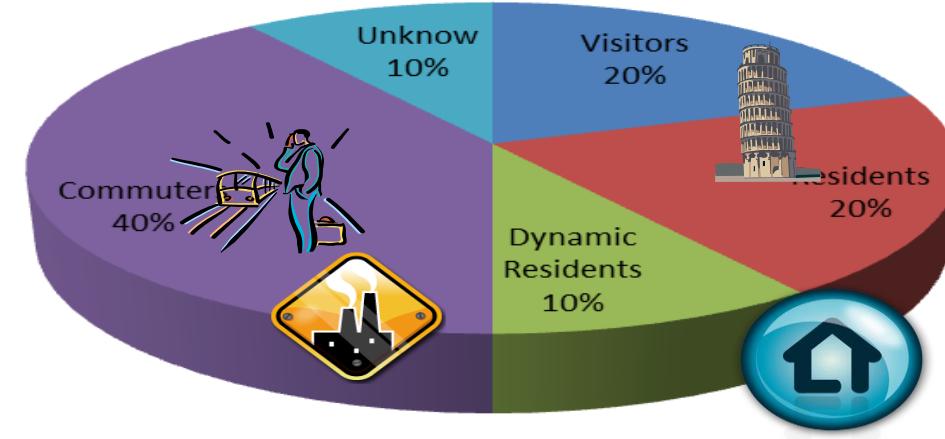
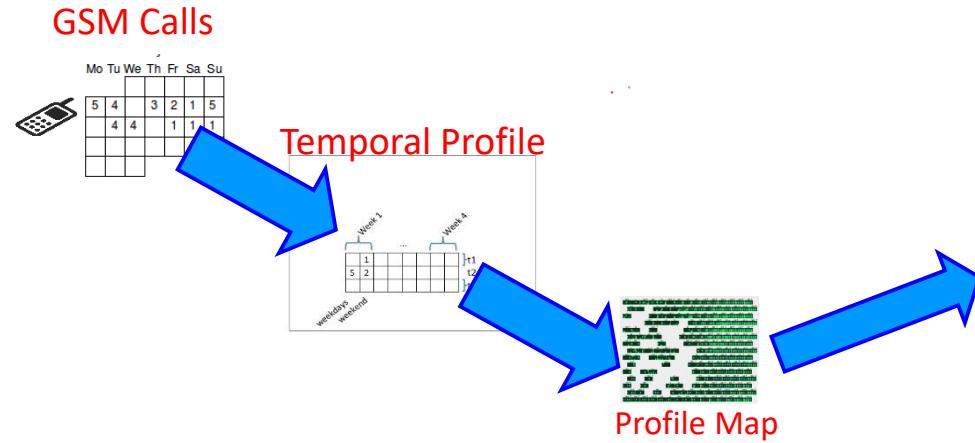
**Outgoing Traffic (38.271 Trajectories)**

City	Traj	Perc
San Giuliano T.	4.942	62%
Vecciano	1.416	93%
Viareggio	1.117	99%
Lucca	895	87%
Camaiore	329	96%

NORD 32%
OVEST 0%
SUD 13%
EST 54%



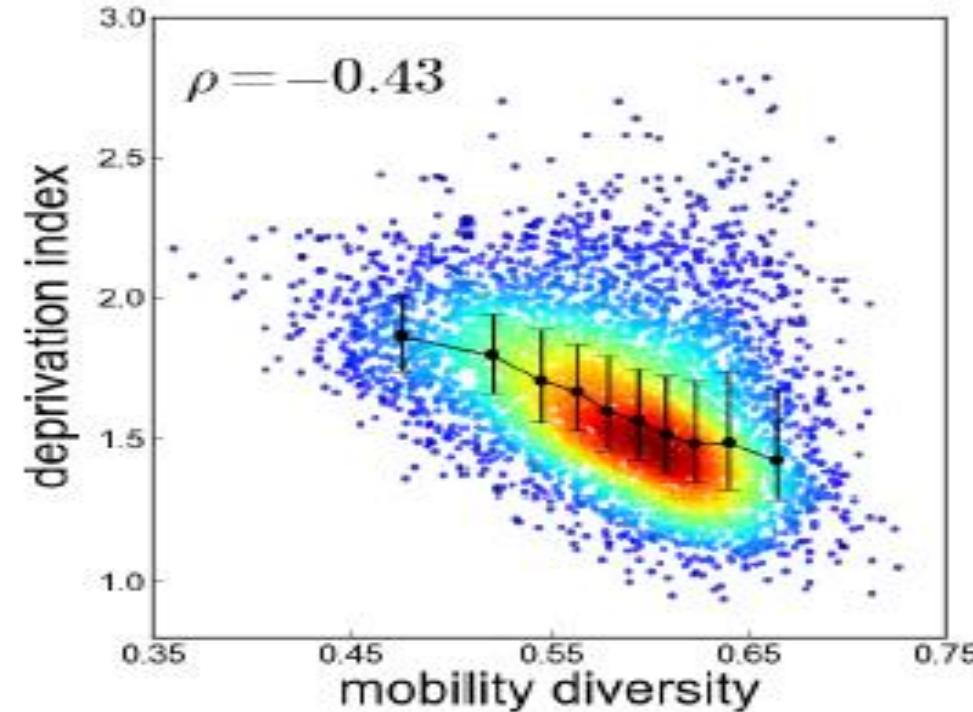
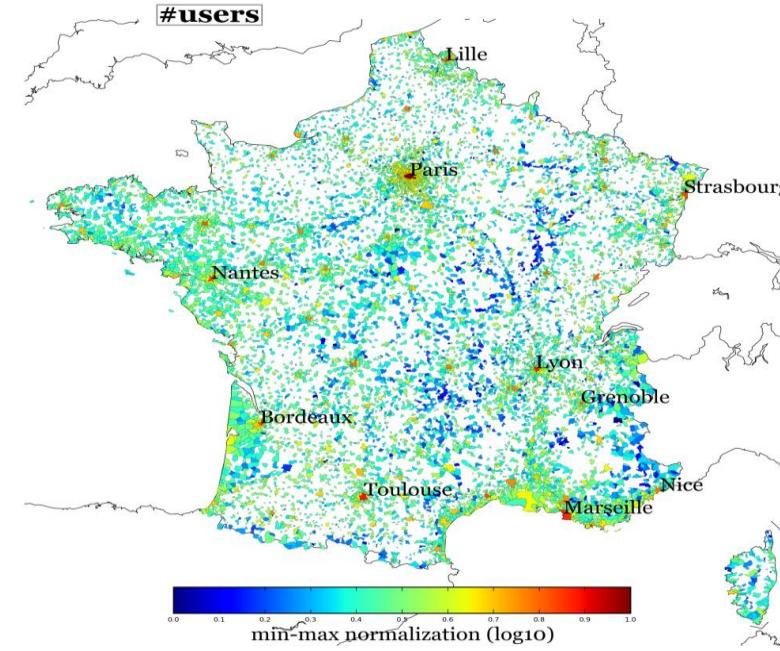
A Sociometer based on Mobile Phone Data for Real Time Demographics



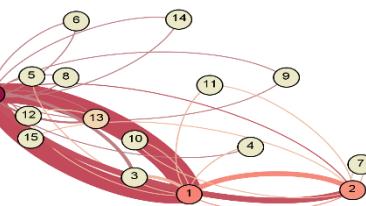
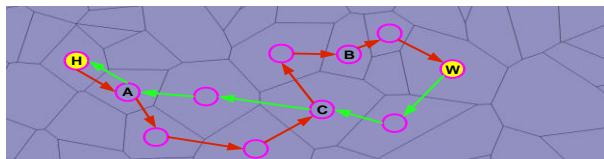
ISTITUTO DI SCIENZA E TECNOLOGIE
DELL'INFORMAZIONE "A. FAEDO"



Big Data: Diversity and economic development

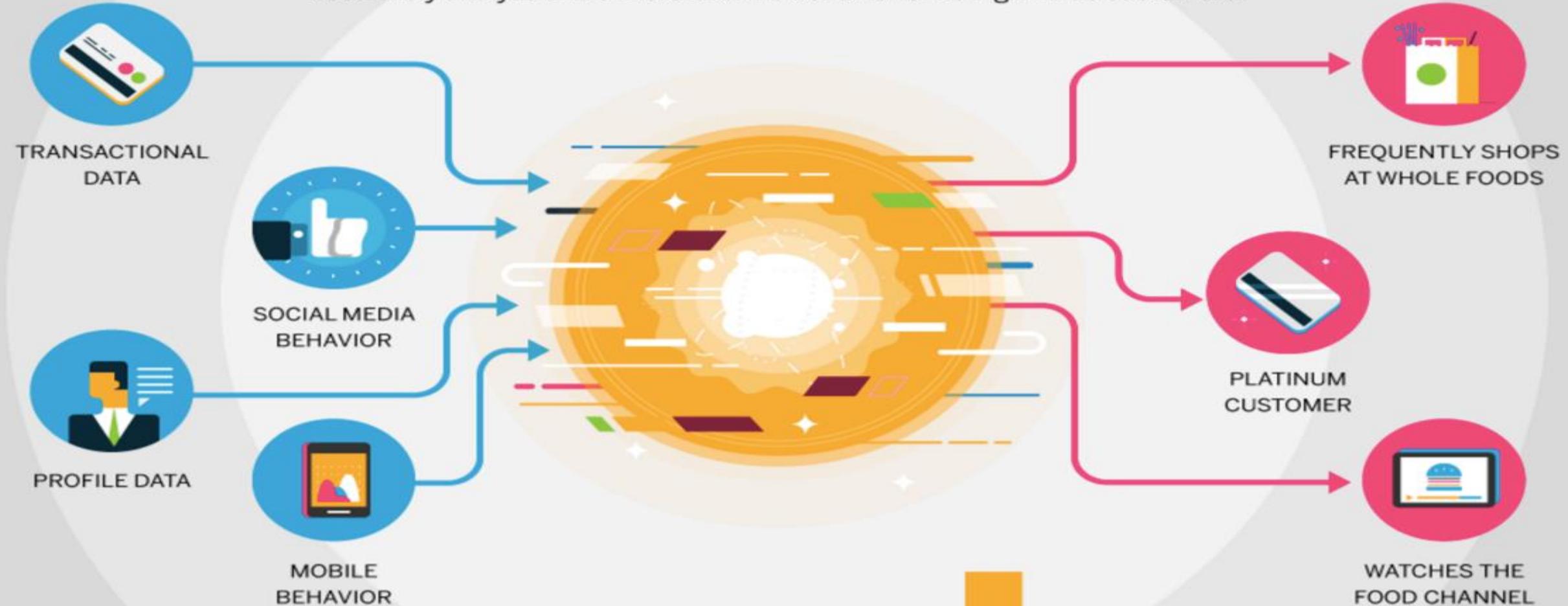


Mobility diversity is correlated with **wellbeing indicators** and **socio-economic development** (Income, Deprivation Index, Education level).

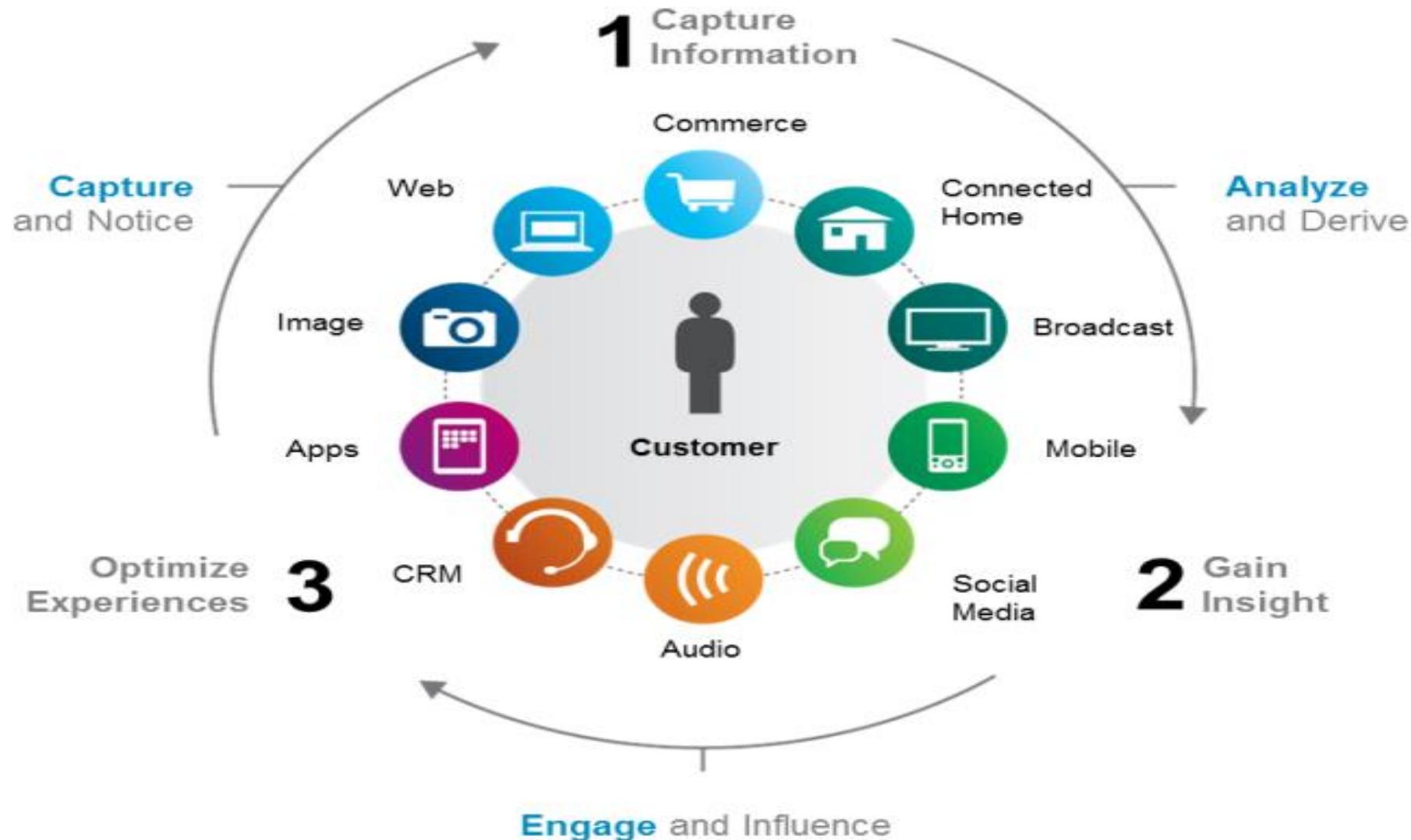


Big Data: from credit card to customer habits

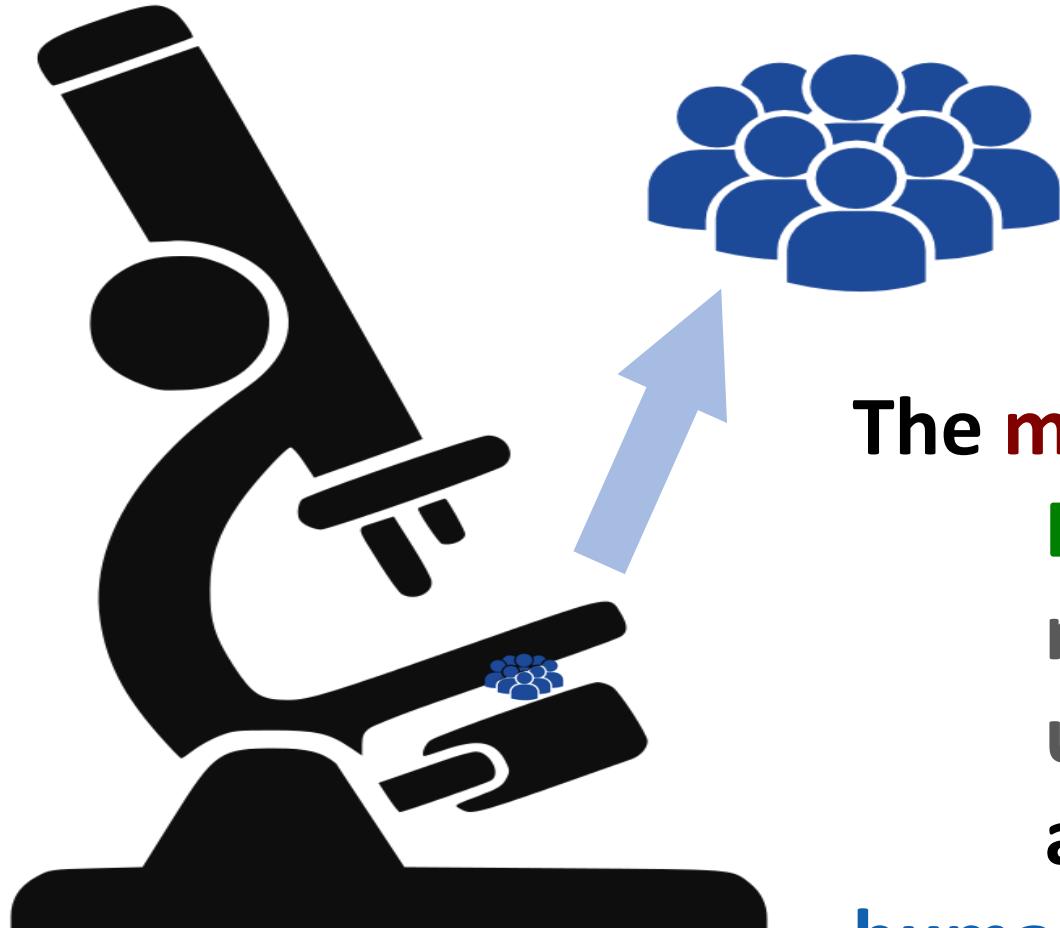
Using Big Data Analytics, a global **CREDIT CARD COMPANY** is able to accurately analyze and understand the behavior of its high-value customers:



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](http://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2015/15-11-19%20Big%20Data%20EN.pdf)



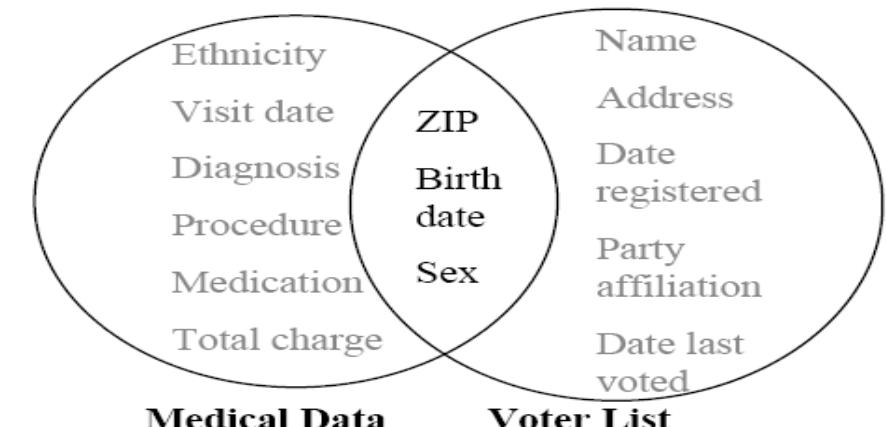
**General Data Protection Regulation
April 2018**

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 personal data collected

Re-identification of Massachusetts' governor

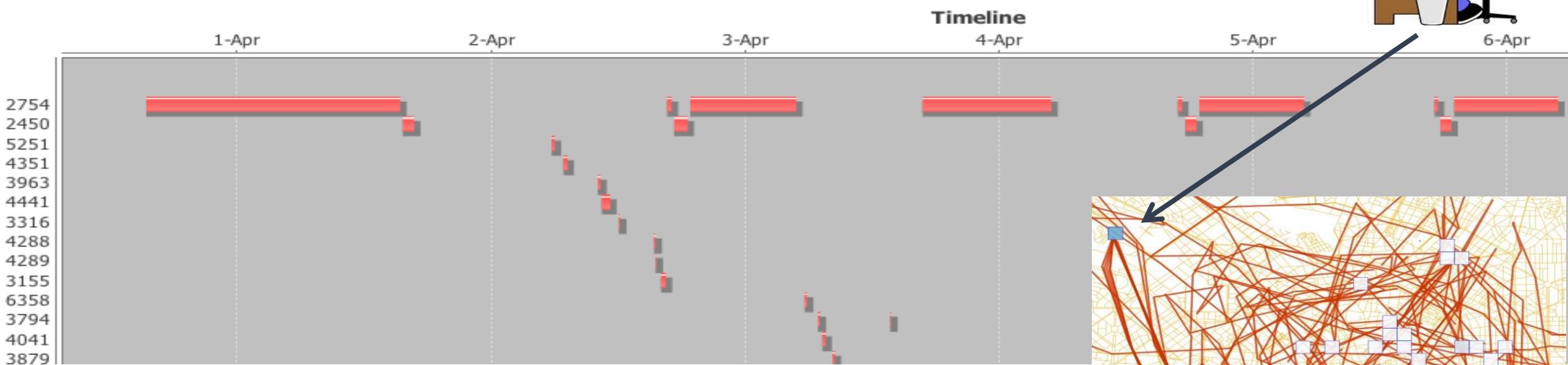
- Sweeney managed to re-identify the medical record of the governor of Massachusetts
 - MA collects and publishes sanitized medical data for state employees (microdata) **left circle**
 - voter registration list of MA (publicly available data) **right circle**
- looking for governor's record
- join the tables:
 - 6 people had his birth date
 - 3 were men
 - 1 in his zipcode



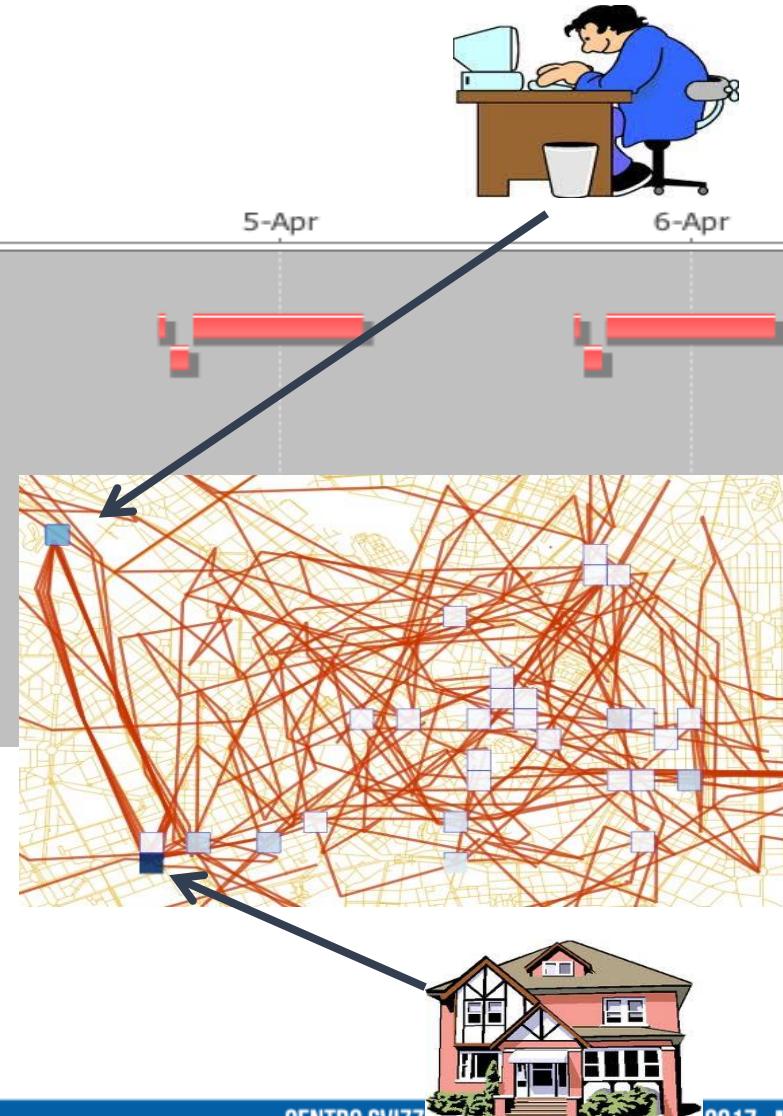
Latanya Sweeney: *k-Anonymity: A Model for Protecting Privacy*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5): 557-570 (2002)



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

CENTRO SVIZZERO - MILANO - 12 OTTOBRE 2017

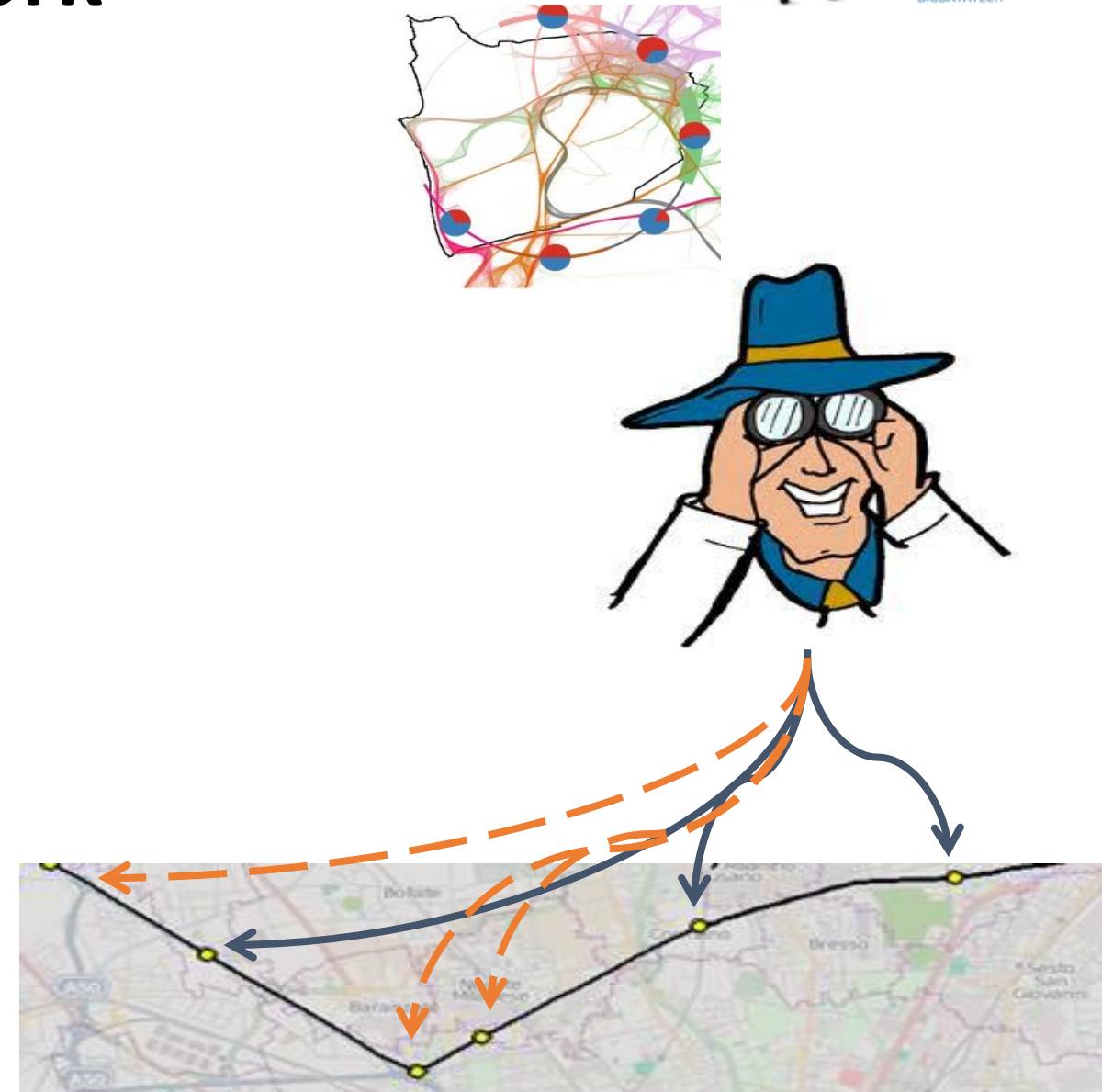


Knowledge Discovery and Delivery Lab
(ISTI-CNR & Univ. Pisa)
www-kdd.isti.cnr.it

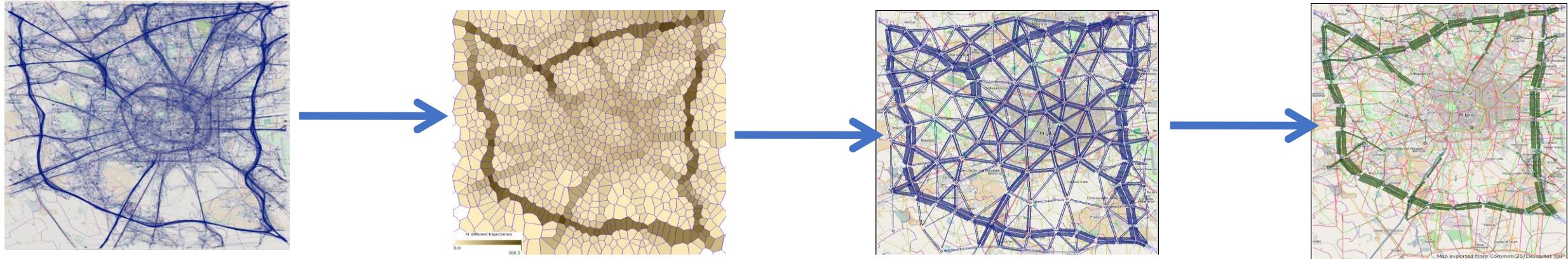


Privacy-Preserving Framework

- 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



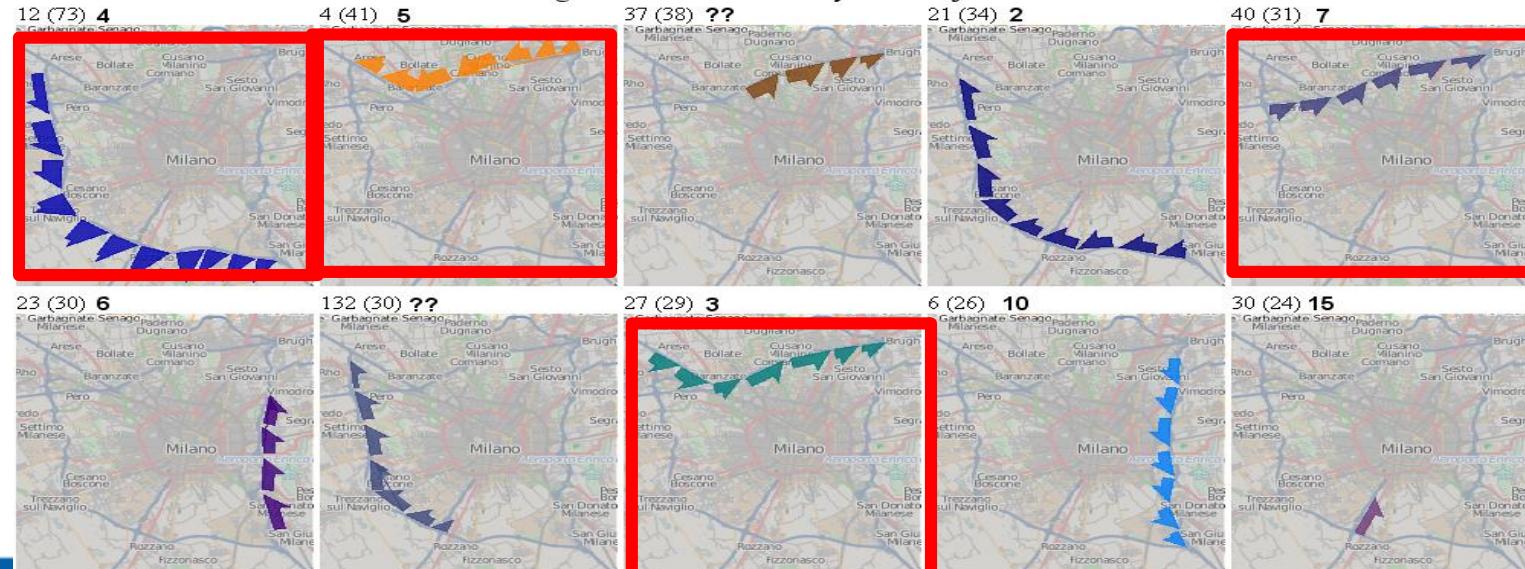
- Given a trajectory dataset
 1. Partition of the territory into **Voronoi cells**
 2. Transform trajectories into sequence of cells
 3. Ensure k-anonymity:
 - For each generalized trajectory there exist at least others $k-1$ different people with the same trajectory? If not transform data in similar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

Clustering on Anonymized Trajectories

10 largest clusters of the original trajectories



10 largest clusters of the anonymized trajectories





Your Data, Your Creativity!



BIGDATA
TECH 2017

CENTRO SVIZZERO - MILANO - 12 OTTOBRE 2017

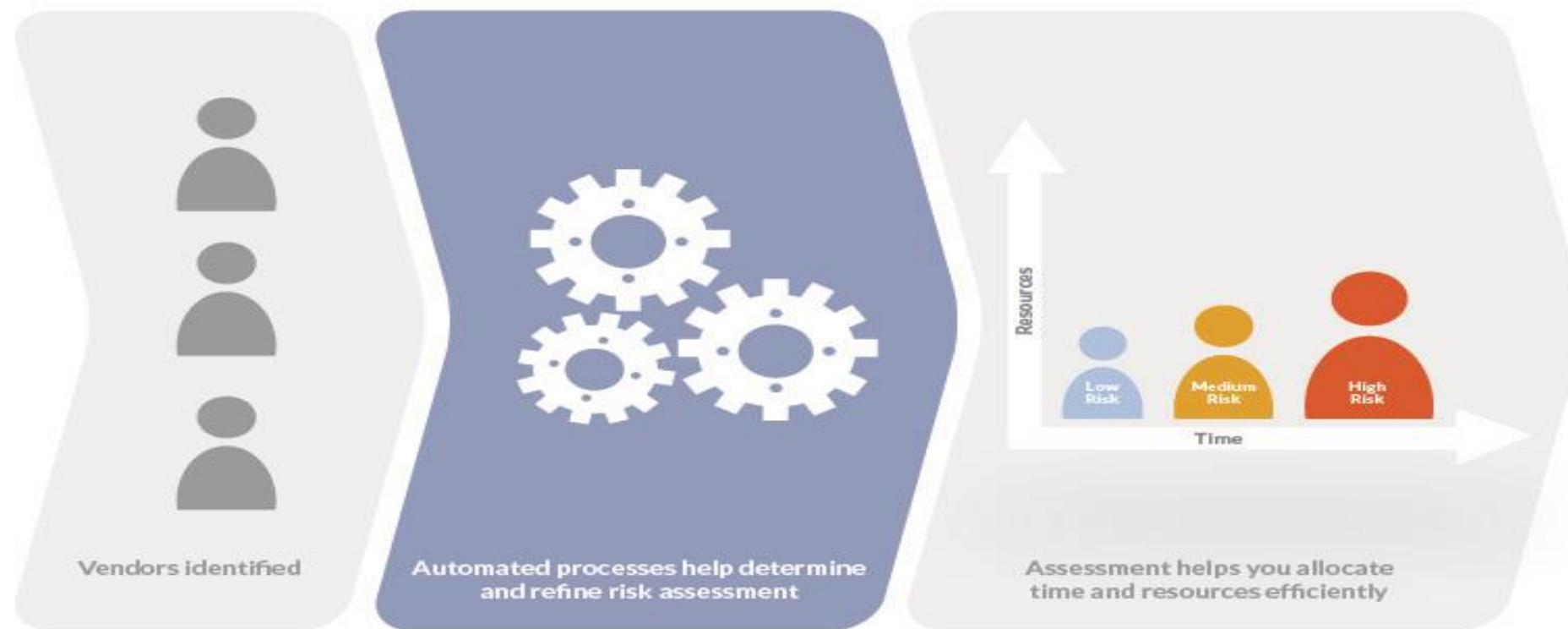
Data-Driven Privacy Risk Assessment



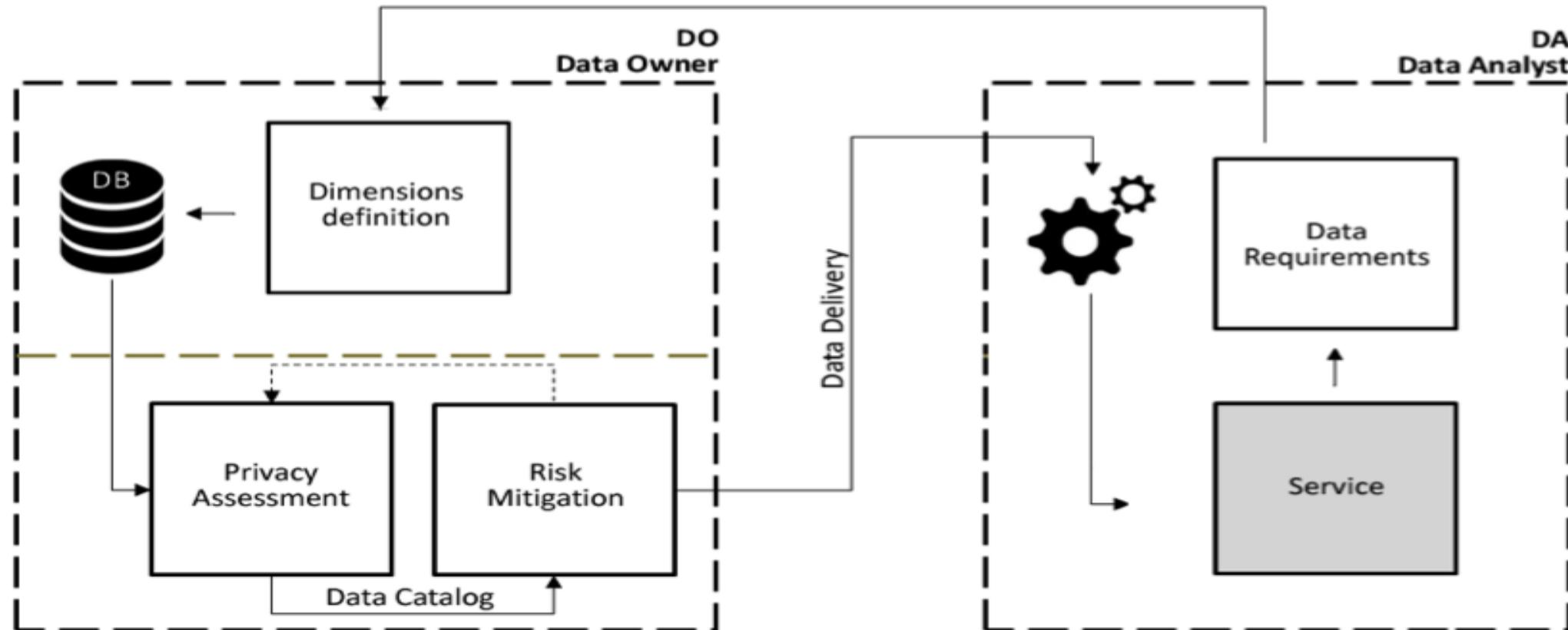
Knowledge Discovery and Delivery Lab
(ISTI-CNR & Univ. Pisa)
www-kdd.isti.cnr.it

Privacy Risk Assessment

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



Privacy Risk Assessment Framework for Data Sharing



Privacy Risk Assessment Framework for Data Sharing

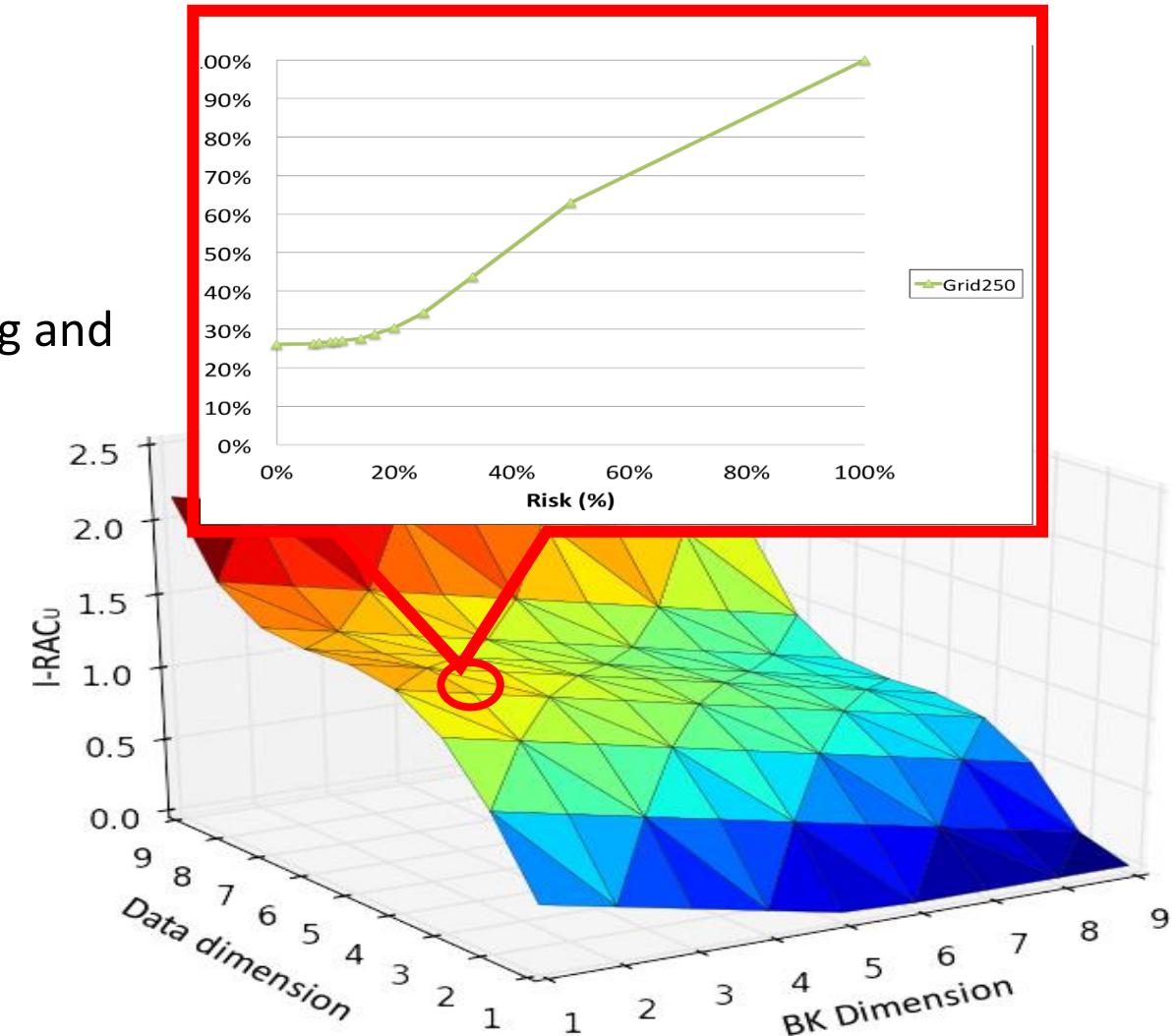
Data Catalog

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

CENTRO SVIZZERO - MILANO - 12 OTTOBRE 2017



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

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.
Copyright 2008 ACM 978-1-60558-193-4/08/08 ...\$5.00.

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

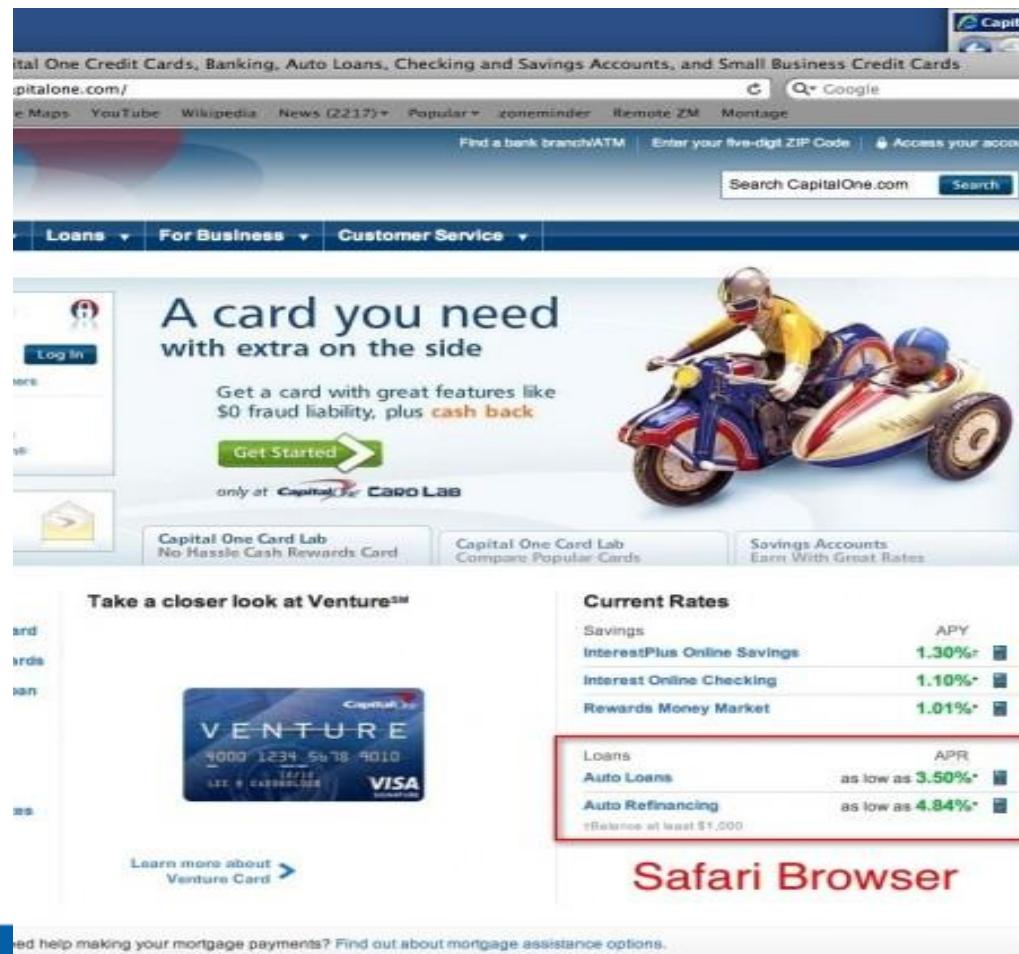
An automatic data processing technique that consists of applying a “profile” to an individual, particularly in order to take decisions concerning her or him or for analysing or predicting her or his personal preferences, behaviours and attitudes



Fairness

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

Safari Browser



A card you need with extra on the side

Get a card with great features like \$0 fraud liability, plus cash back

Get Started >

only at CapitalOne Card Lab

Capital One Card Lab
No Hassle Cash Rewards Card

Capital One Card Lab
Compare Popular Cards

Savings Accounts
Earn With Great Rates

Take a closer look at VentureSM

VENTURE

InterestPlus Online Savings **1.30%**

Interest Online Checking **1.10%**

Rewards Money Market **1.01%**

Auto Loans **as low as 3.50%**

Auto Refinancing **as low as 4.84%**

Learn more about Venture Card >

ed help making your mortgage payments? Find out about mortgage assistance options.

Firefox Browser



A card you need with extra on the side

Get a card with great features like \$0 fraud liability, plus cash back

Get Started >

only at CapitalOne Card Lab

Capital One Card Lab
Rewards Credit Card

Savings Accounts
Earn With Great Rates

Current Rates

Savings **APY 1.30%**

InterestPlus Online Savings **1.30%**

Interest Online Checking **1.10%**

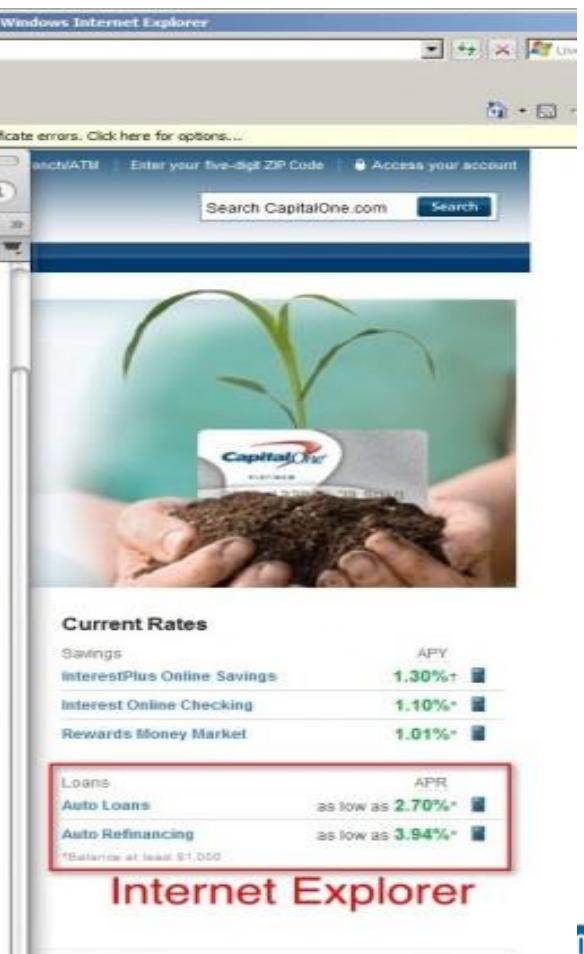
Rewards Money Market **1.01%**

Auto Loans **APR as low as 3.10%**

Auto Refinancing **as low as 4.34%**

*Balance of least \$1,000

Internet Explorer



A card you need with extra on the side

Get a card with great features like \$0 fraud liability, plus cash back

Get Started >

only at CapitalOne Card Lab

Capital One Card Lab
Rewards Credit Card

Savings Accounts
Earn With Great Rates

Current Rates

Savings **APY 1.30%**

InterestPlus Online Savings **1.30%**

Interest Online Checking **1.10%**

Rewards Money Market **1.01%**

Auto Loans **APR as low as 2.70%**

Auto Refinancing **as low as 3.94%**

*Balance of least \$1,000

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.

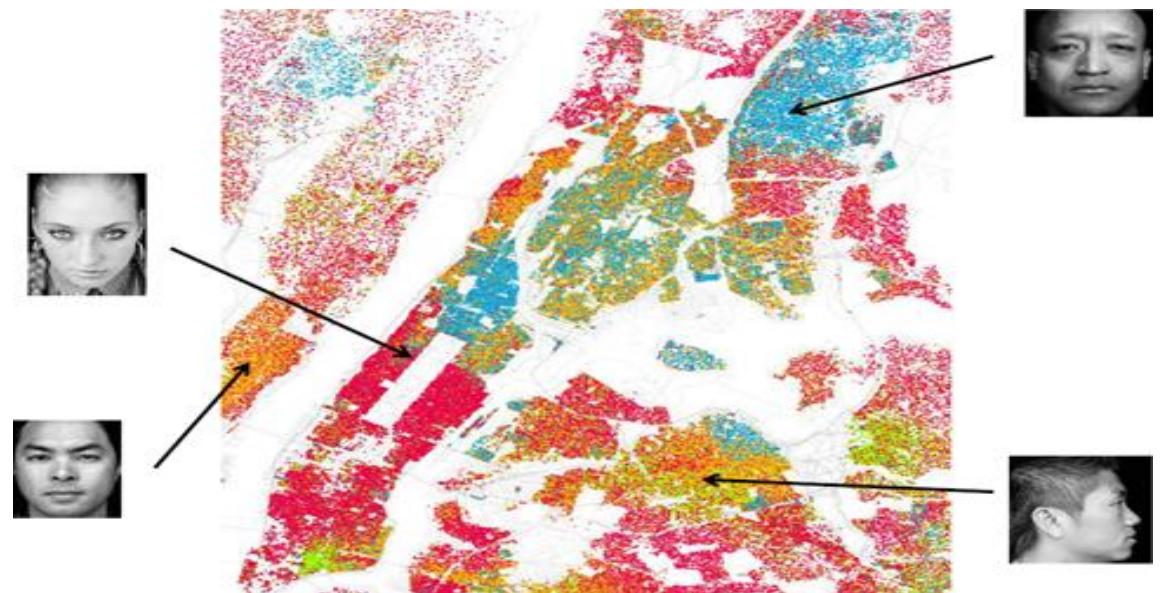
Discrimination

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

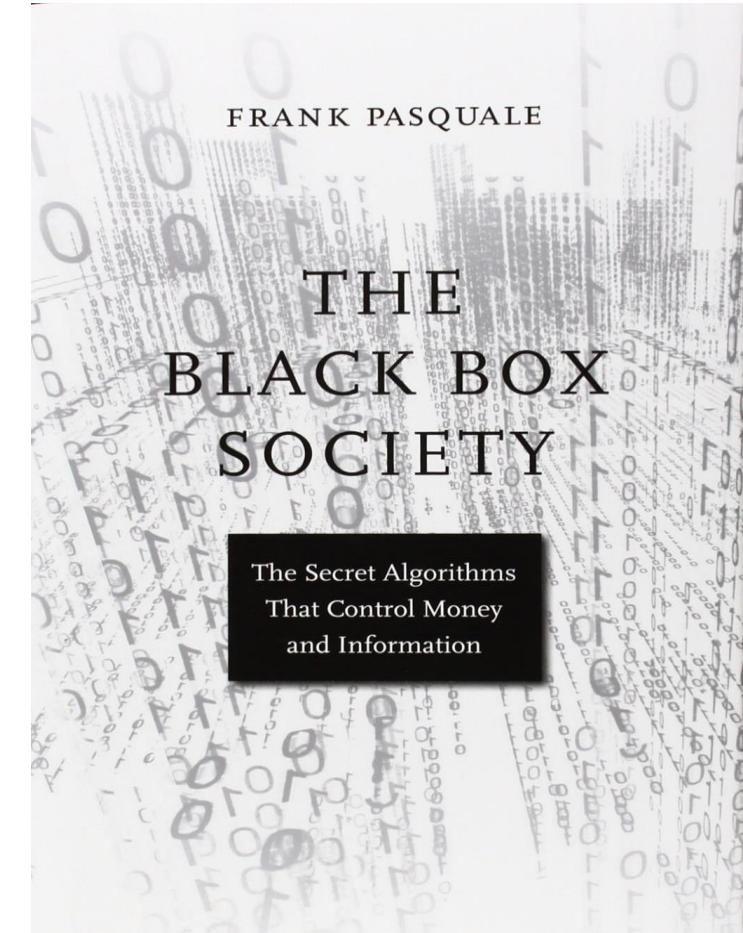
CENTRO SVIZZERO - MILANO - 12 OTTOBRE 2017



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

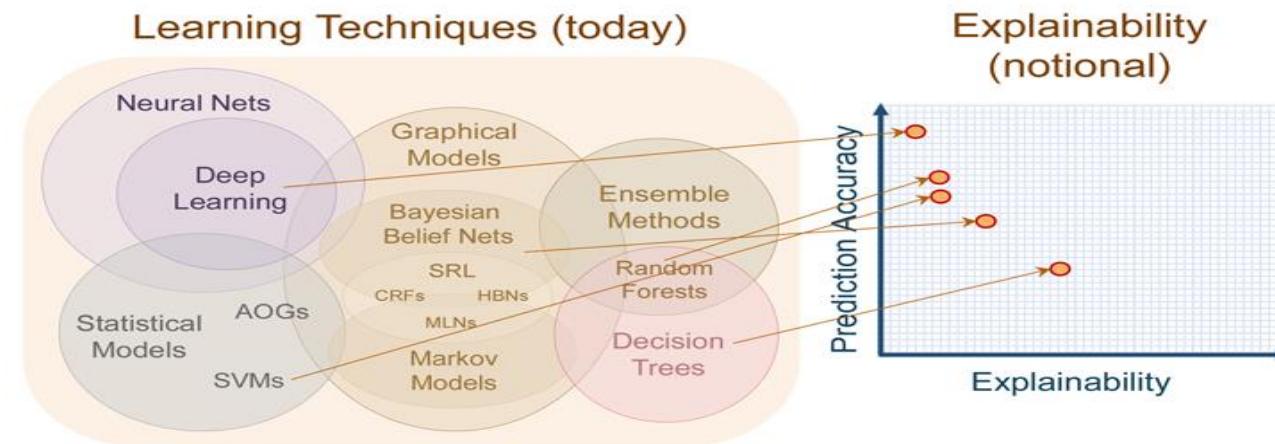
Right of explanation

- 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?”



The screenshot shows the homepage of the journal *nature*. The header includes the title "nature" and the subtitle "International weekly journal of science". Below the header is a navigation bar with links: Home, News & Comment, Research, Careers & Jobs, Current Issue, Archive, Audio & Video, and For Authors. A secondary navigation bar at the bottom shows the current path: Archive > Volume 537 > Issue 7621 > Editorial > Article. The main content area features a large, bold title: "More accountability for big-data algorithms". Below the title is a sub-headline: "To avoid bias and improve transparency, algorithm designers must make data sources and profiles public." The date of publication is listed as 21 September 2016. At the bottom right of the page are sharing icons for social media.

NATURE | EDITORIAL



More accountability for big-data algorithms

To avoid bias and improve transparency, algorithm designers must make data sources and profiles public.

21 September 2016

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



The Consortium

Italy United Kingdom Germany Estonia

Finland Switzerland Nederlands

 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

 A!

 ETH Zürich

 TU Delft

Existing national RI's to be integrated

 SoBigData
Bureau for Social Mining & Data Mining

 GATE general
architecture
for test
engineering

 Fraunhofer
IGD

 L3S Research Center

 nervousnet

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



Master Big Data
Università di Pisa
Consiglio Nazionale delle Ricerche



Master Universitario Di II Livello
Big Data Analytics E Social Mining

Aree tecnico-scientifiche:

- Big Data Technology
- Big Data Sensing & Procurement
- Big Data Mining
- Big Data Story Telling
- Big Data Ethics

Aree di innovazione socio-economica:

- Big Data for Social Good
- Big Data for Business

Il Master Big Data ha l'obiettivo di formare "data scientists", dei professionisti dotati di un mix di competenze multidisciplinari che permettono non solo di acquisire dati ed estrarne conoscenza, ma anche di raccontare "storie" attraverso questi dati, a supporto delle decisioni, della creatività e dello sviluppo di servizi innovativi, e di saper gestire le ripercussioni etiche e legali dei Big Data, che spesso contengono informazioni personali e suscitano problematiche relative alla privacy, alla trasparenza, alla consapevolezza.

BIG DATA

SoBigData
UNIVERSITÀ DI PISA



<http://www.masterbigdata.it>
Deadline for Applications:
17 November 2017



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

Thank you!

