Big Data Ethics

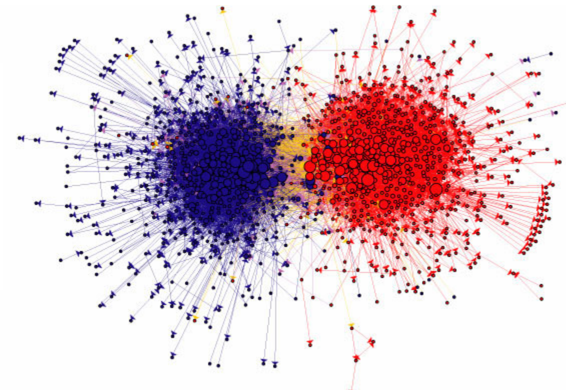
Anna Monreale

KDD LAB http://kdd.isti.cnr.it
Università di Pisa and ISTI-CNR

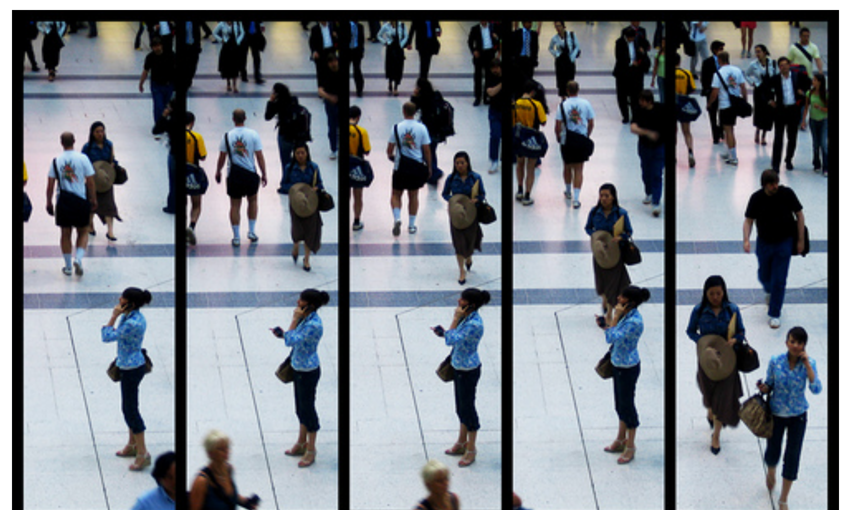
Big data "proxies" of social life

Shopping patterns & lifestyle

Relationships & social ties

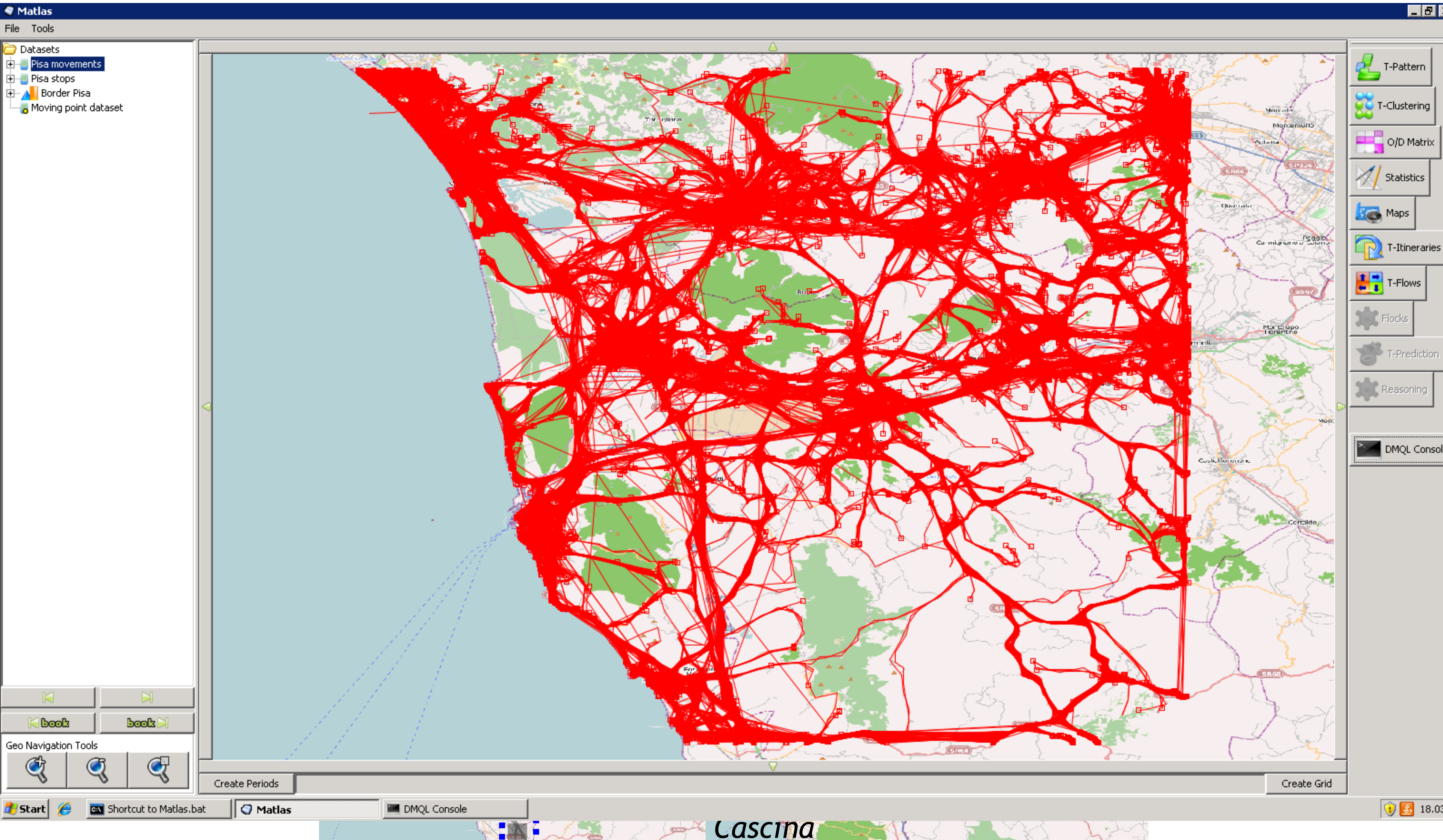


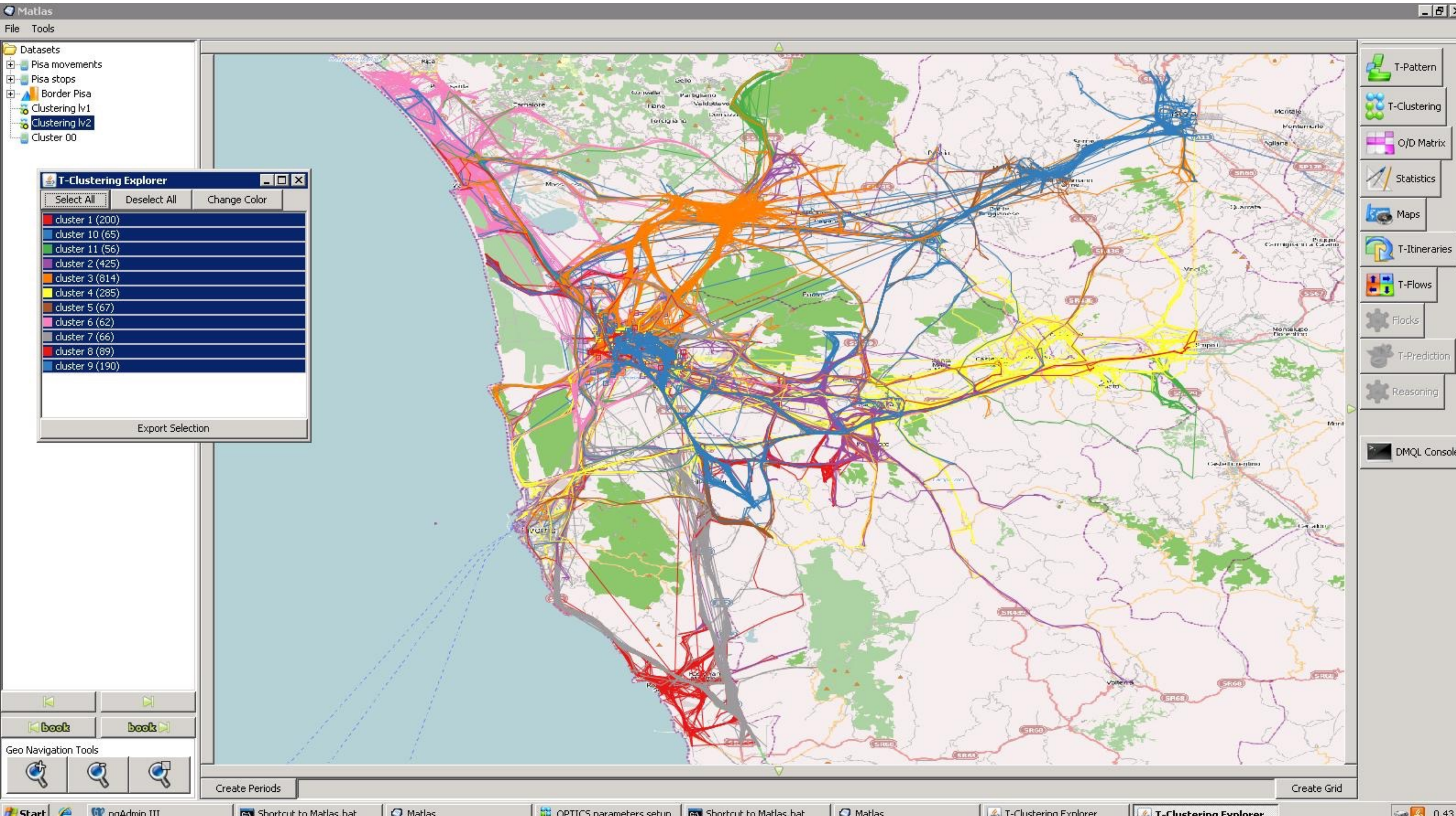
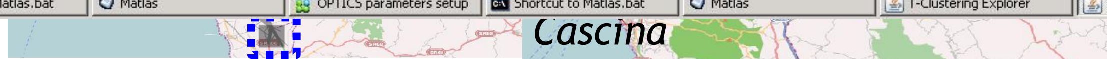
Movements



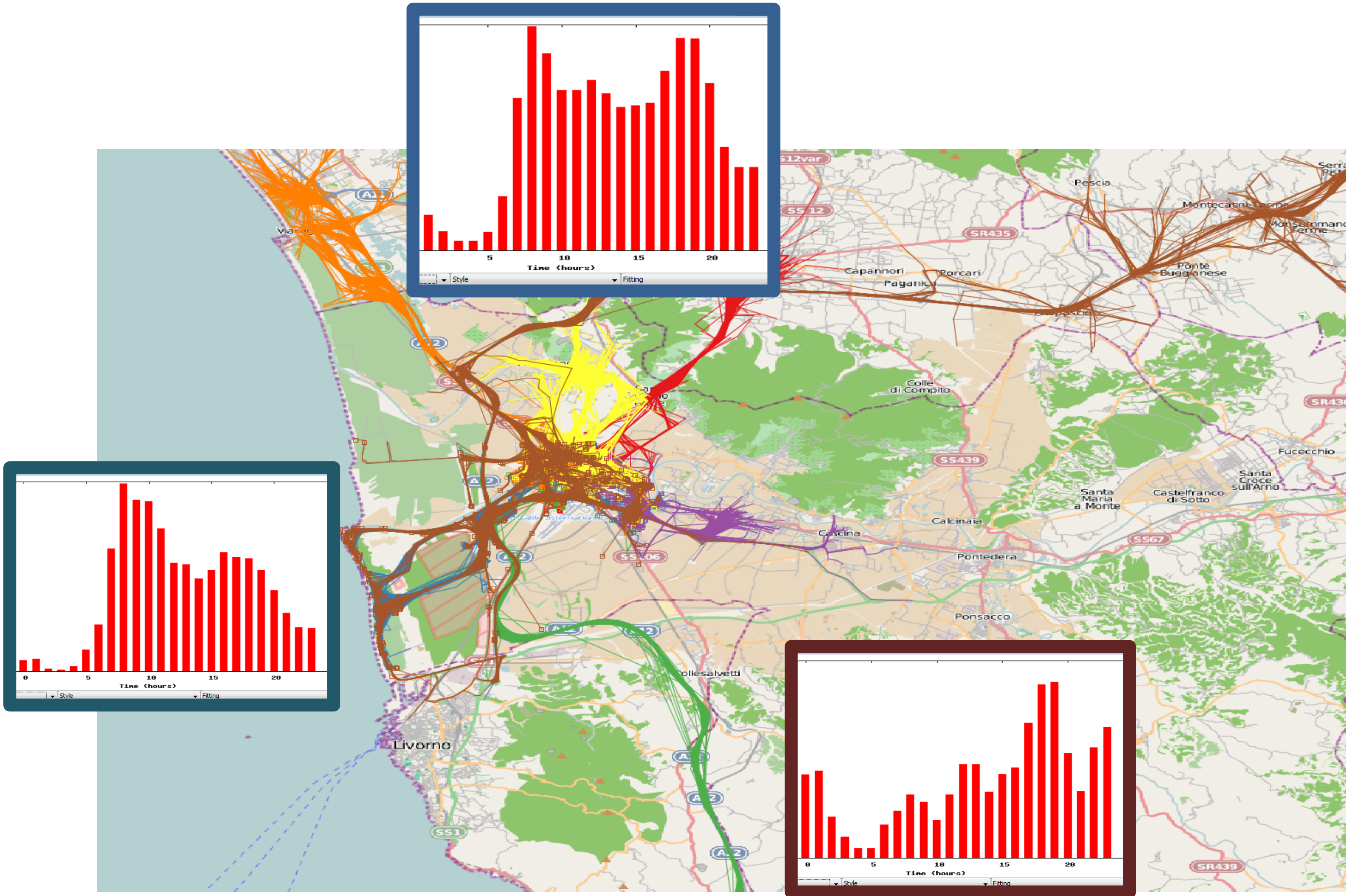
Desires, opinions, sentiments

WIKIPEDIA
The Free Encyclopedia

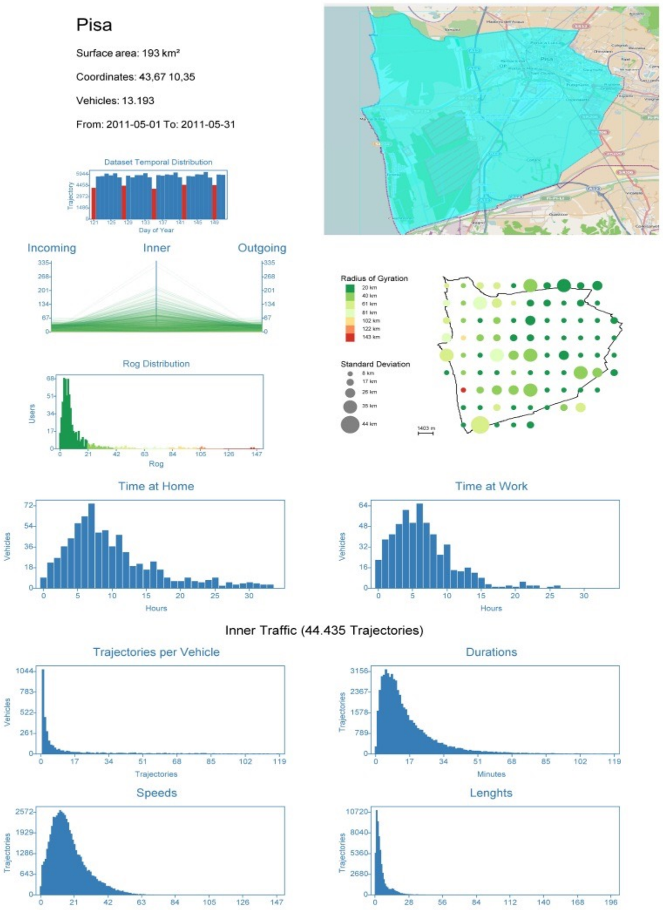

Matlas
File
Tools
Datasets
Pisa movements
Pisa stops
Border Pisa
Moving point dataset
book
book
Geo Navigation Tools
T-Pattern
T-Clustering
O/D Matrix
Statistics
Maps
T-Itineraries
T-Flows
Flocks
T-Prediction
Reasoning
DMQL Consol
Create Periods
Create Grid
Start
Shortcut to Matlas.bat
Matlas
DMQL Console
Cascina

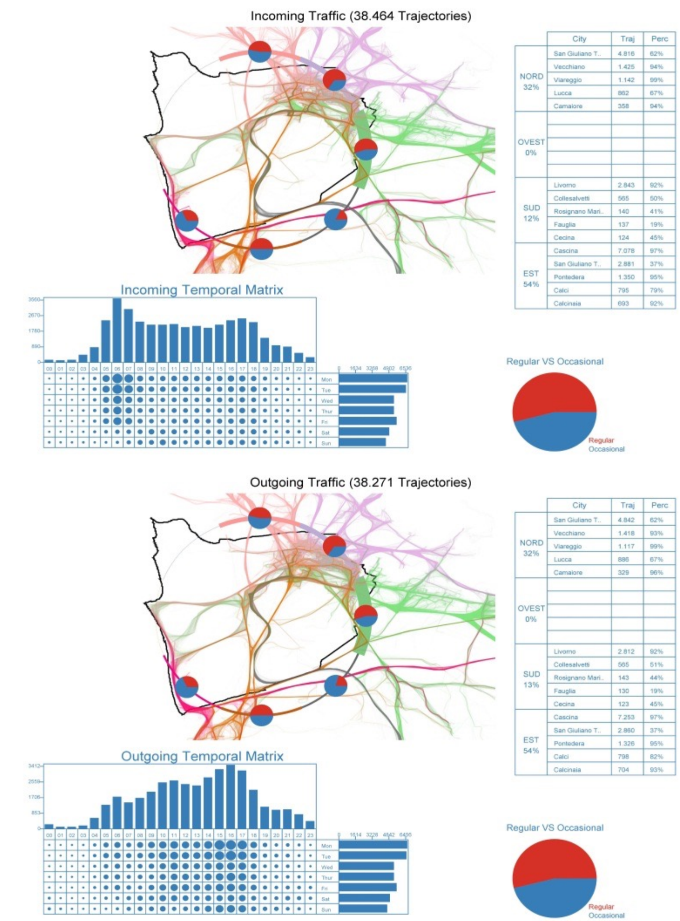

Matlas
File Tools
Datasets
Pisa movements
Pisa stops
Border Pisa
Clustering lv1
Clustering lv2
Cluster 00
T-Clustering Explorer
Select All
Deselect All
Change Color
cluster 1 (200)
cluster 10 (65)
cluster 11 (56)
cluster 2 (425)
cluster 3 (814)
cluster 4 (285)
cluster 5 (67)
cluster 6 (62)
cluster 7 (66)
cluster 8 (89)
cluster 9 (190)
Export Selection
book
book
Geo Navigation Tools
T-Pattern
T-Clustering
O/D Matrix
Statistics
Maps
T-Itineraries
T-Flows
Flocks
T-Prediction
Reasoning
DMQL Console
Create Periods
Create Grid
Start
pgAdmin III
Shortcut to Matlas.bat
Matlas
OPTICS parameters setup
Shortcut to Matlas.bat
Matlas
T-Clustering Explorer
T-Clustering Explorer
0.43

Cascina

City access paths



Mobility atlas of many cities





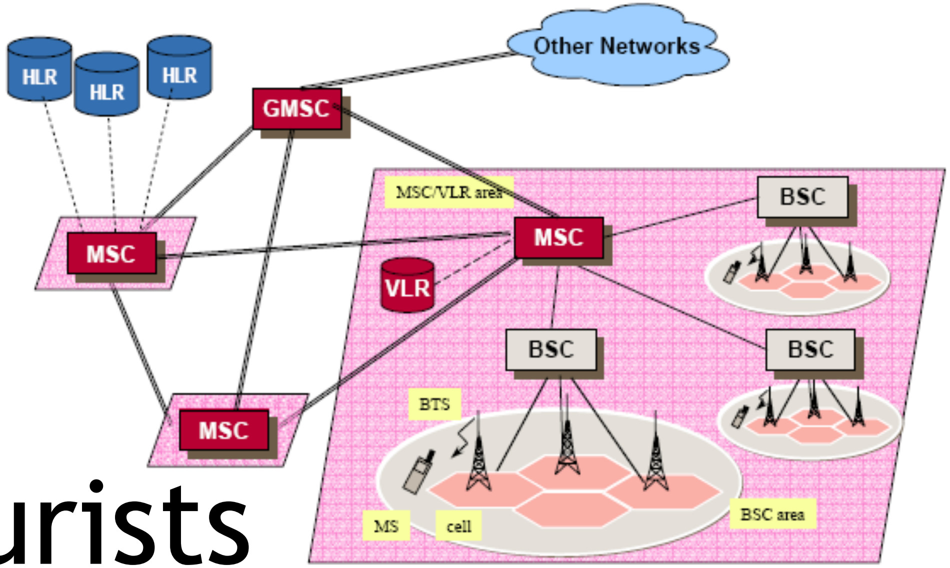
Mobile phone socio-meters

Analyze individual call habits to recognize profiles

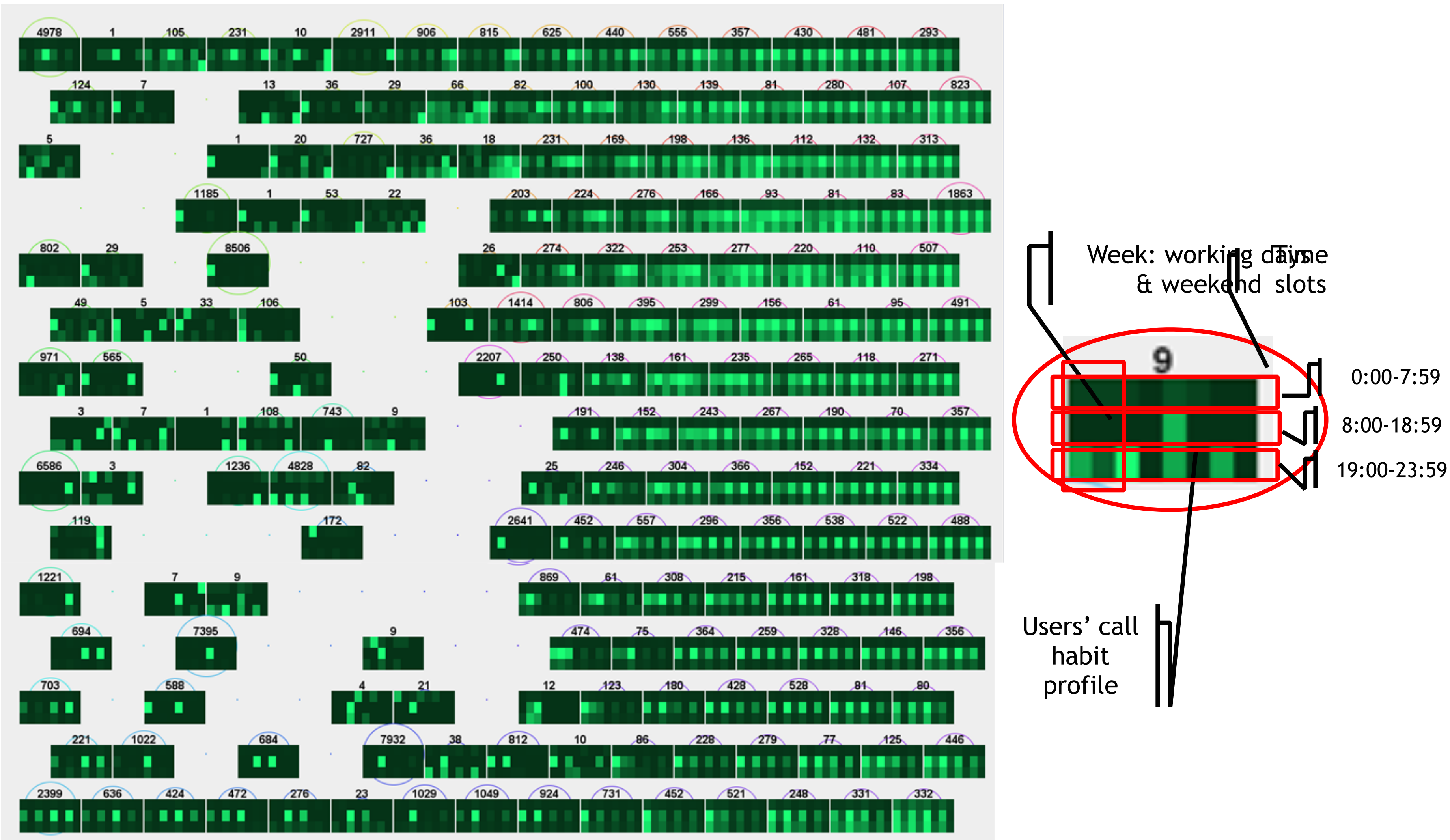
–Resident

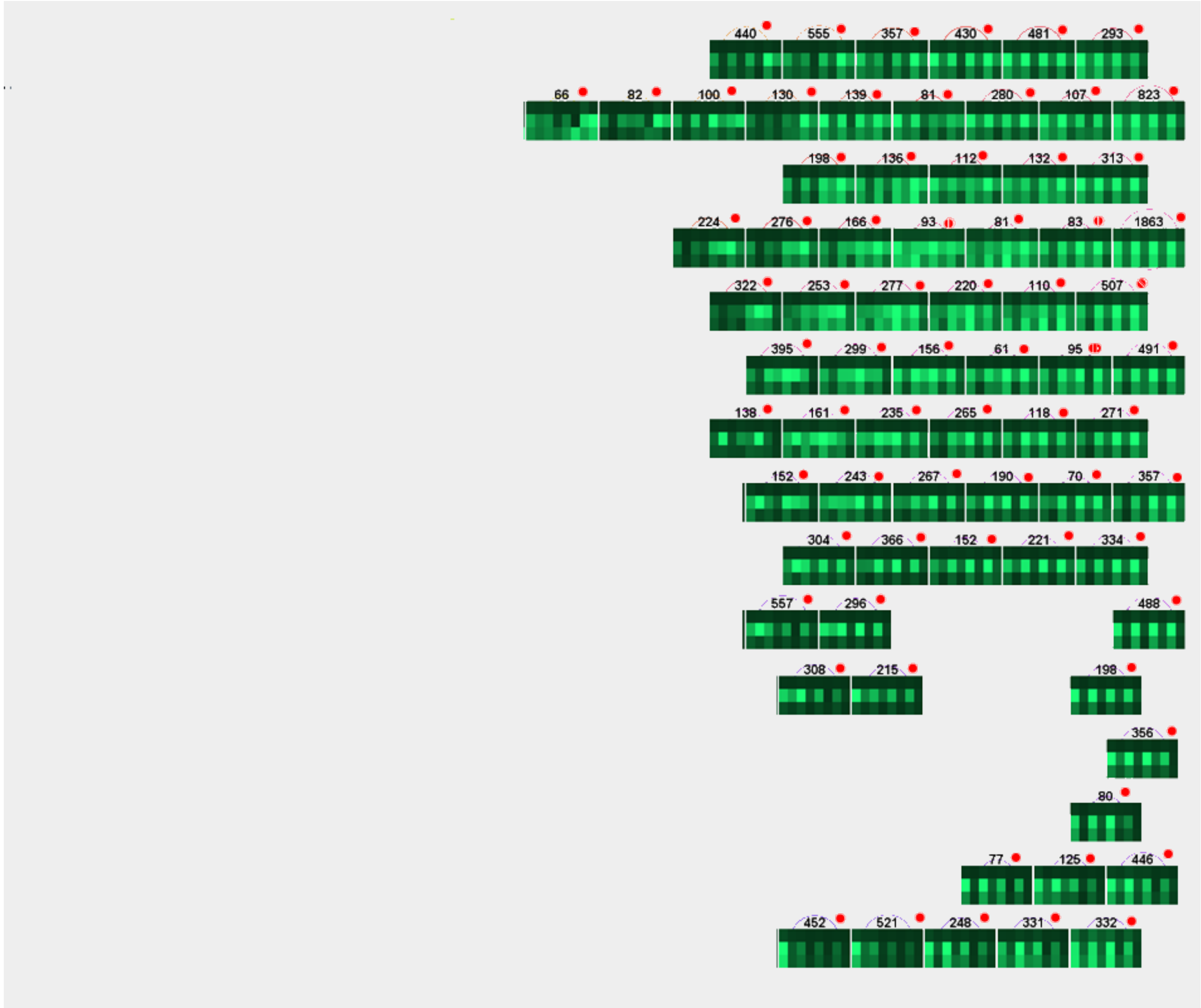
–Commuters

–Visitors/Tourists

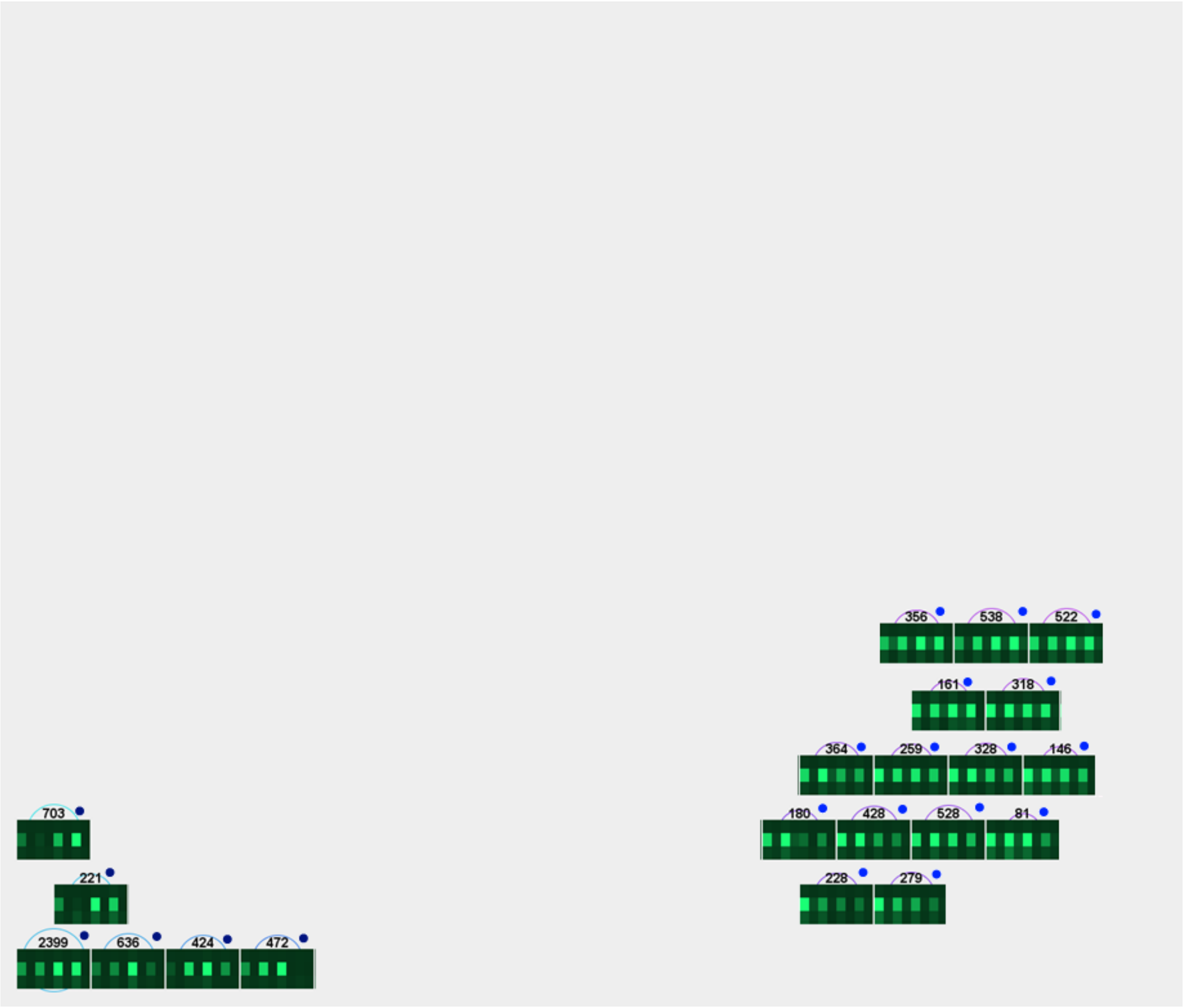


Call Habit Profiles

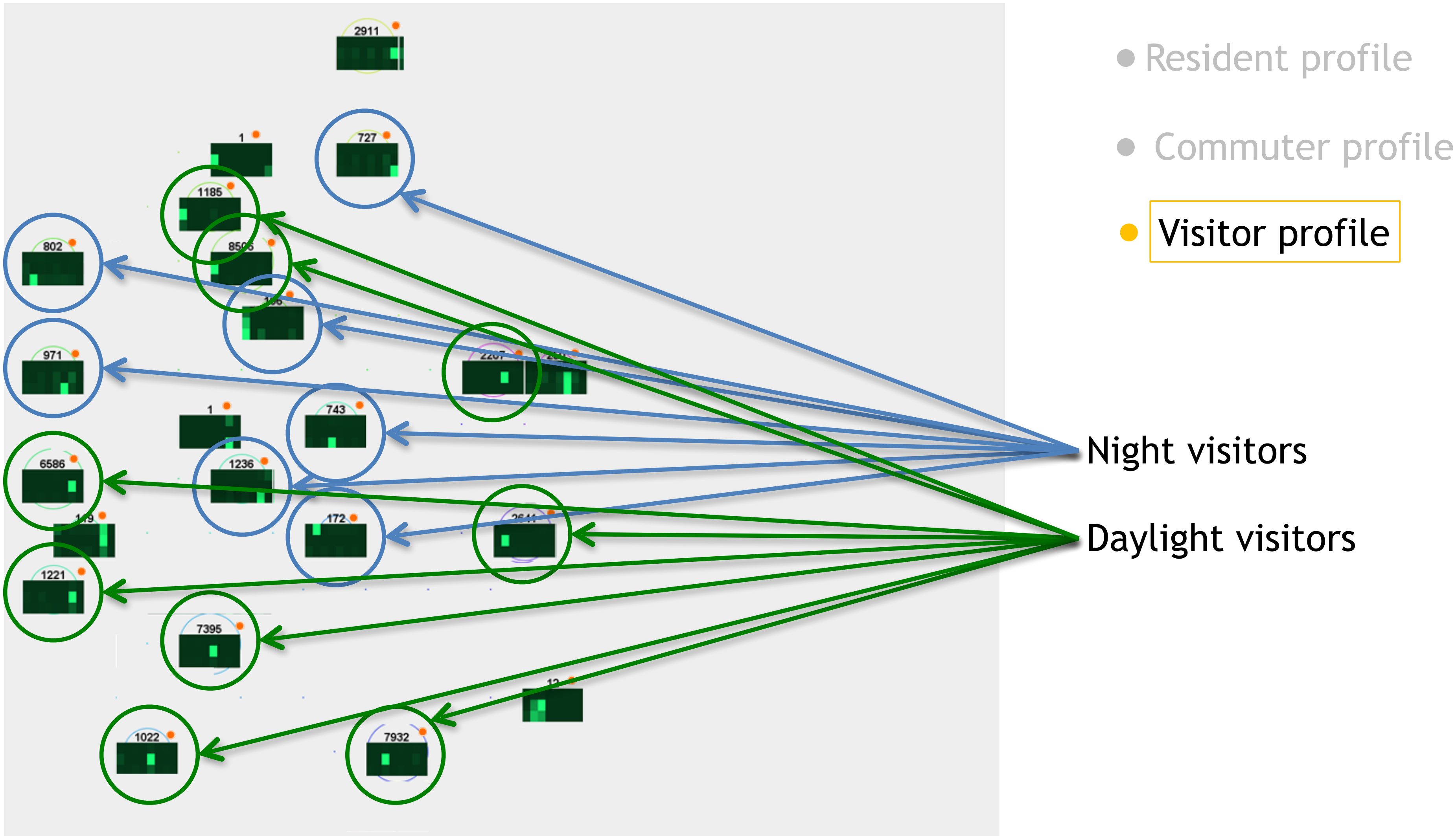




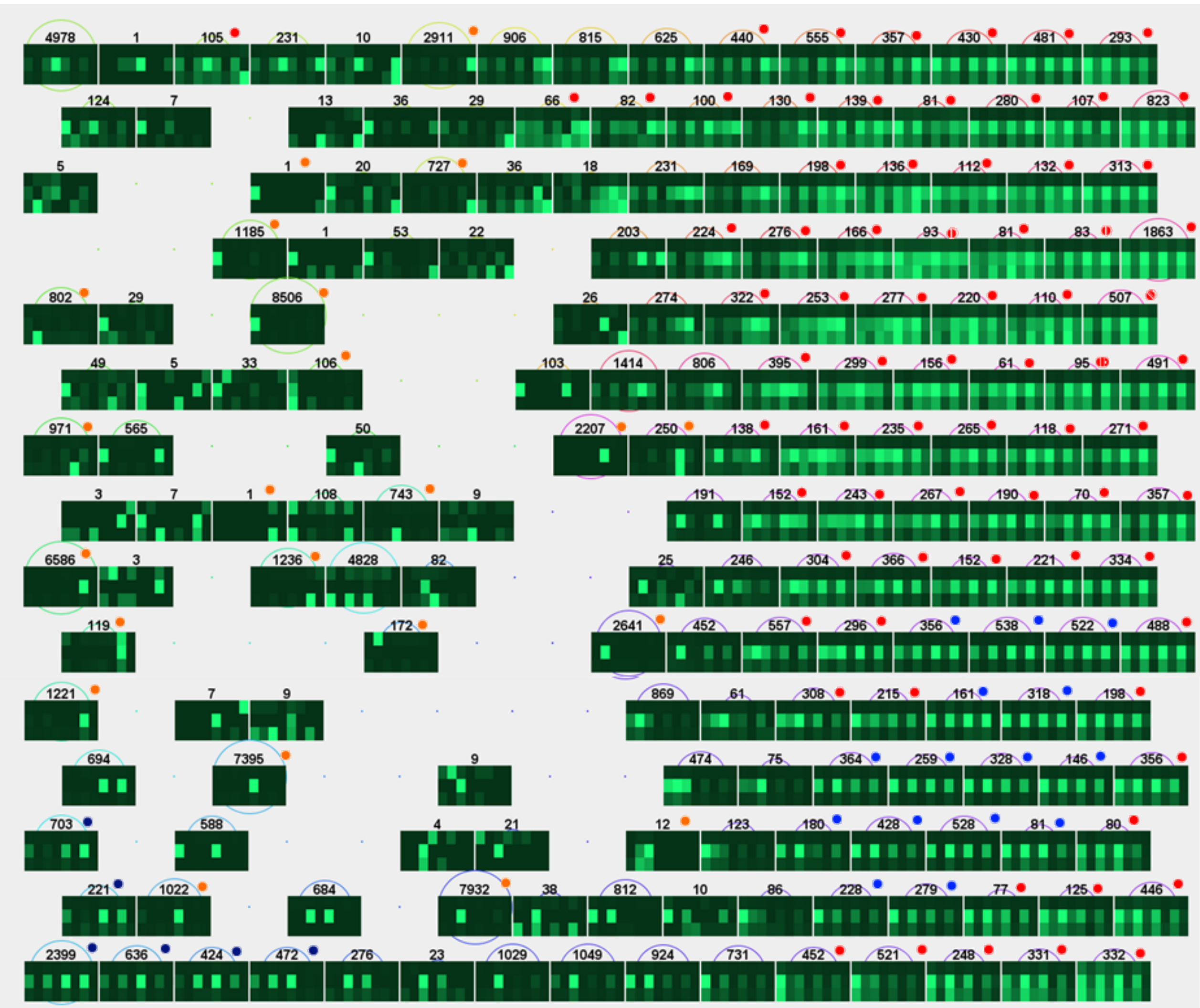
Resident profile



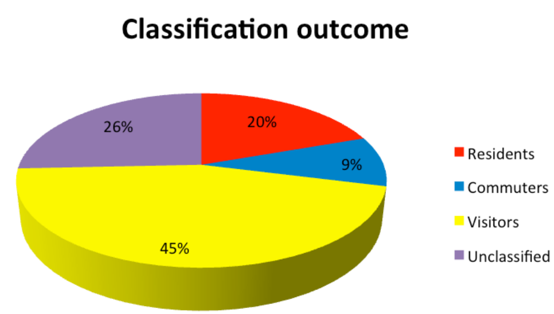
- Resident profile
- Commuter profile


Resident profile
Commuter profile
Visitor profile
Night visitors
Daylight visitors
2911
727
1185
802
971
743
6586
1236
172
1221
7395
1022
7932

User profile quantification



- Resident profile
- Commuter profile
- Visitor profile



Does current legal framework allow answering these new questions?

EU Legislation for protection of personal data

- European directives:
  - Data protection directive (95/46/EC)
  - ePrivacy directive (2002/58/EC) and its revision (2009/136/EC)
  - Proposal for a new EU Regulation (25 Jan 2012)

  http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm

EU: Personal Data

- **Personal data** is defined as any information relating to an identity or **identifiable** natural person.

- An **identifiable person** is one who can be identified, **directly or indirectly**, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.

Anonymity according to 1995/46/EC

- The principles of protection must apply to any information concerning an identified or identifiable person;
- To determine whether a person is identifiable, account should be taken of **all the means likely reasonably to be used** either by the controller or by any other person to identify the said person
- **The principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable**

EU Directive (95/46/EC) and new Proposal

- GOALS:
  - protection protection of individuals with regard to the processing of personal data
  - the free movement of such data

New Elements in the EU Proposal

- Principle of Transparency
- Data Portability
- Right of Oblivion
- Profiling
- Privacy by Design

Transparency & Data Portability

- **Transparency:**
  - Any information addressed to the public or to the data subject should be easily accessible and easy to understand
- **Data Portability:**
  - The right to transmit his/her personal data from an automated processing system, into another one

Oblivion & Profiling

- **Right to Oblivion:**
  - The data subject shall have the right to obtain the erasure of his/her personal data and the abstention from further dissemination of such data

- **Profiling:**
  - The right not to be subject to a measure which is based on profiling by means of automated processing

Privacy by Design Principle

- **Privacy by design** is an approach to protect privacy by inscribing it into the design specifications of information technologies, accountable business practices, and networked infrastructures, from the very start

- Developed by Ontario's Information and Privacy Commissioner, Dr. Ann Cavoukian, in the 1990s
  - as a response to the growing threats to online privacy that were beginning to emerge at that time.

Privacy by Design in EU

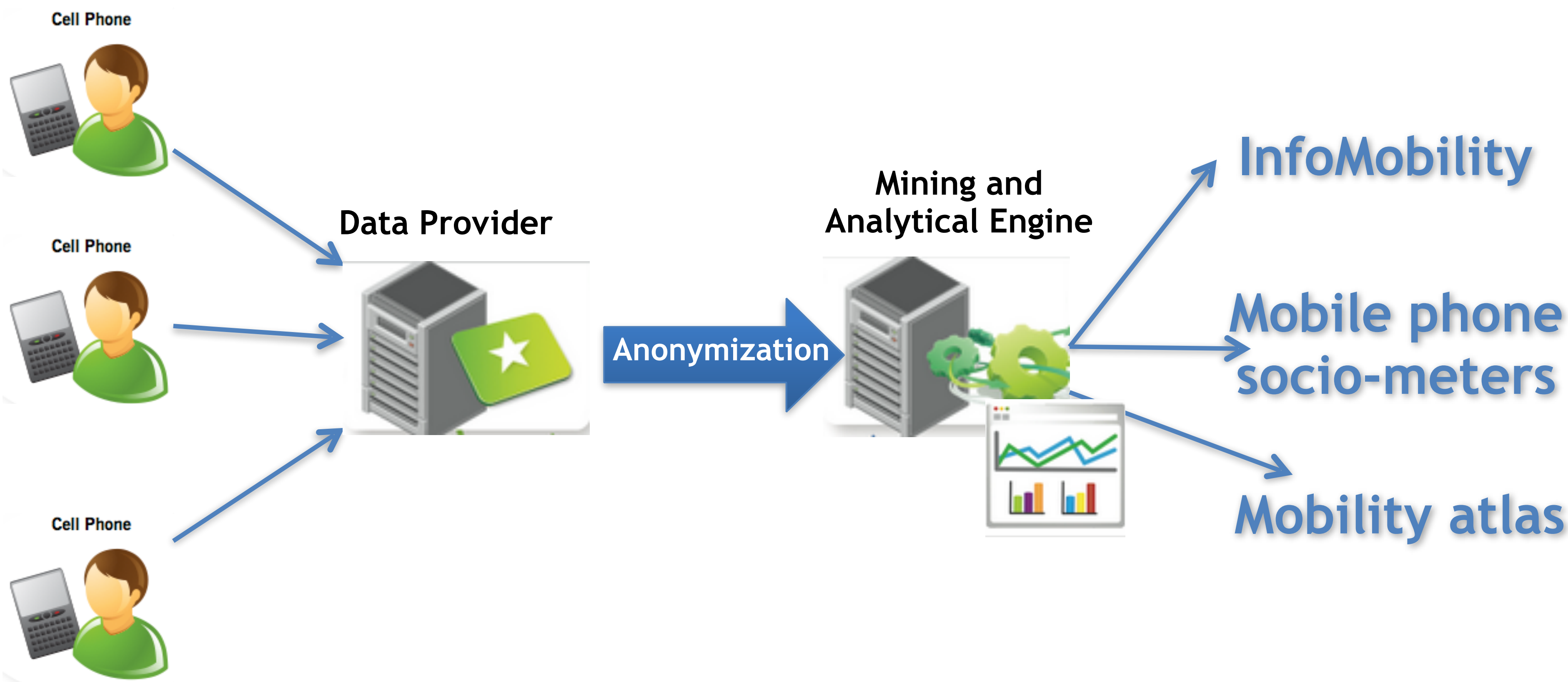
- In 2009, the EU Article 29 Data Protection Working Party and the Working Party on Police and Justice issued a joint Opinion, advocating **for incorporating the principles of Privacy-by-design into a new EU privacy framework**

- In the comprehensive reform of the data protection rules proposed on January 25, 2012 by the EC, the new data protection legal framework introduces the reference to **data protection by design and by default**

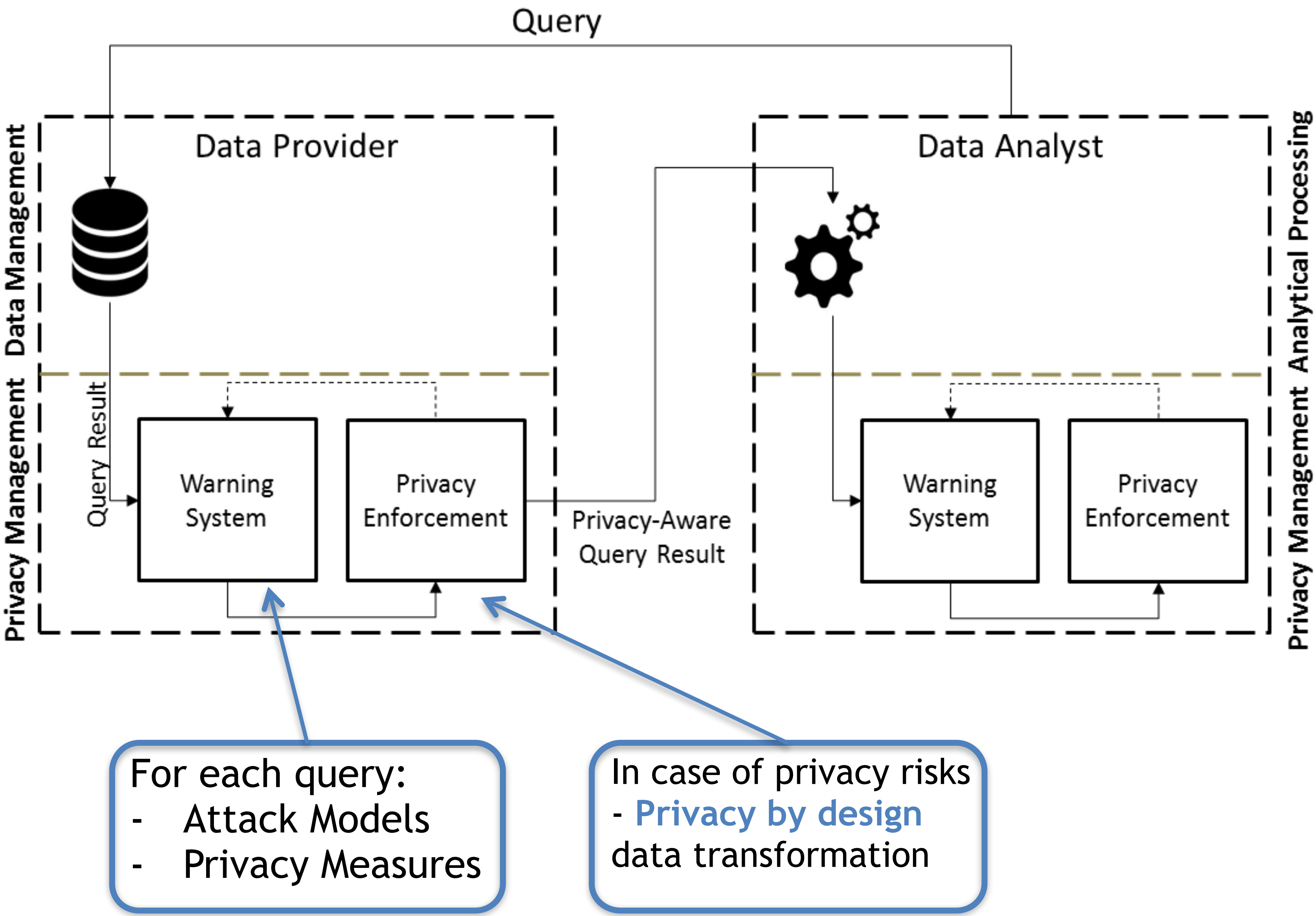
Privacy by Design in Big Data Analytics

- Design frameworks
  - to counter the threats of privacy violation
  - without obstructing the knowledge discovery opportunities of data analysis
- Trade-off between privacy quantification and data utility

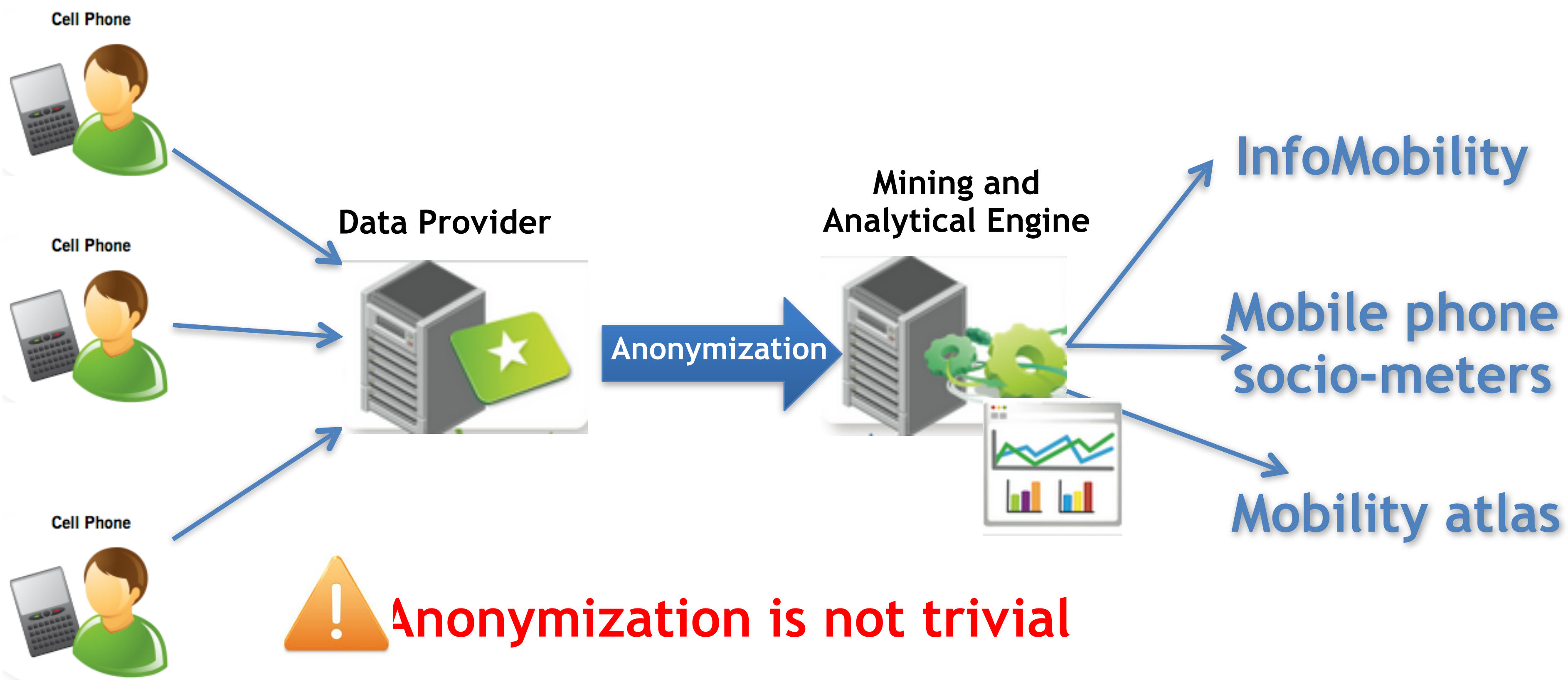
Privacy-by-Design in Big Data Analytics



Privacy-by-Design in Big Data Analytics



Privacy-by-Design in Big Data Analytics



Anonymization is not trivial

De-identification is not enough

EU Article 29 Data Protection Working Party: Opinion 05/2014

- Opinion 05/2014 on Anonymization Techniques
- Provides recommendations to handle these techniques by taking account of **the residual risk** of identification inherent in each of them.

Opinion 05/2014: Effective Anonymisation Solution

- Prevents all parties from
  - **Singling out** an individual in a dataset
  - **Linking** two records within a dataset (or between two separate datasets)
  - **Inferring** any information in such dataset

Opinion 05/2014: Techniques

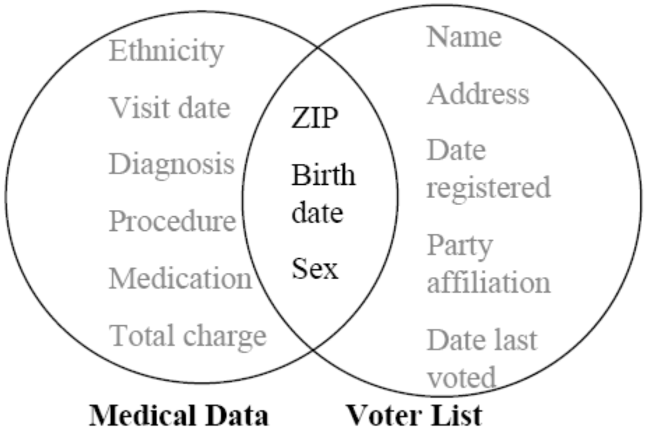
- Anonymity by randomization
- Anonynity by generalization
- Differential-privacy
- l-diversity
- t-closeness
- Pseudonymisation

Example of privacy attacks

Re-identification of Massachussetts' governor

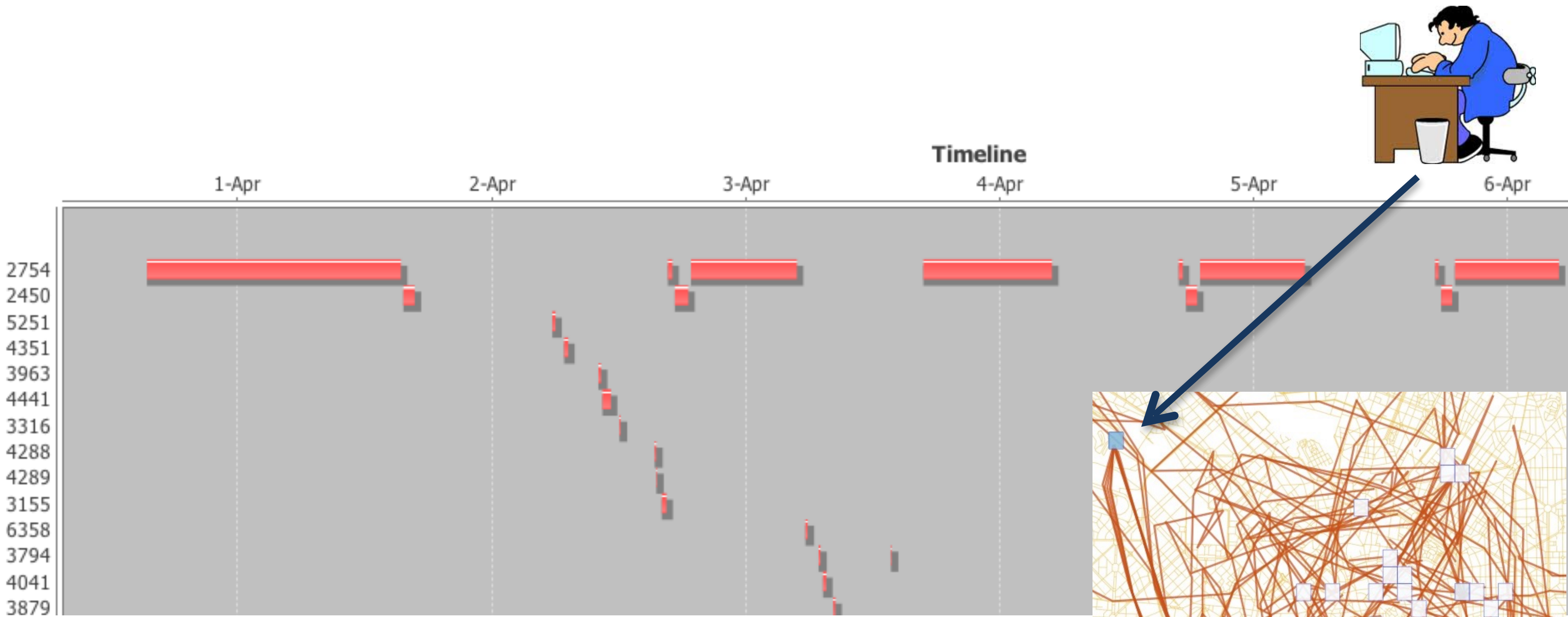
- Sweeney managed to re-identify the medical record of the governor of Massachussetts
  - MA collects and publishes sanitized medical data for state employees (microdata) left circle
  - voter registration list of MA (publicly available data) right circle

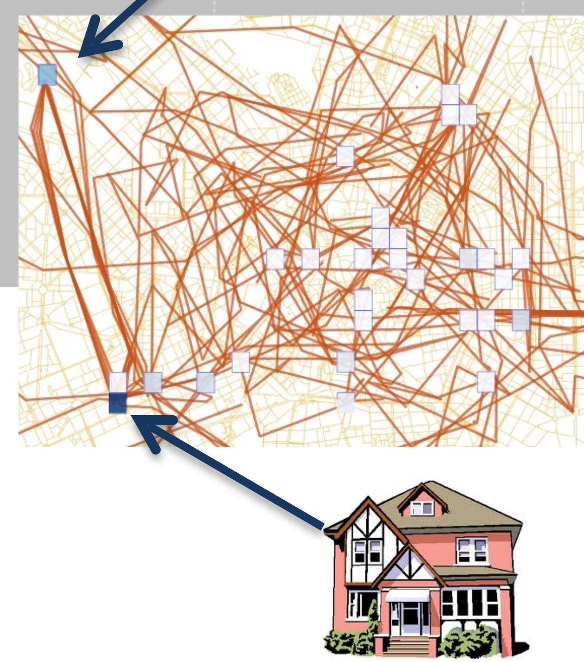
- looking for governor's record
- join the tables:
  - 6 people had his birth date
  - 3 were men
  - 1 in his zipcode



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

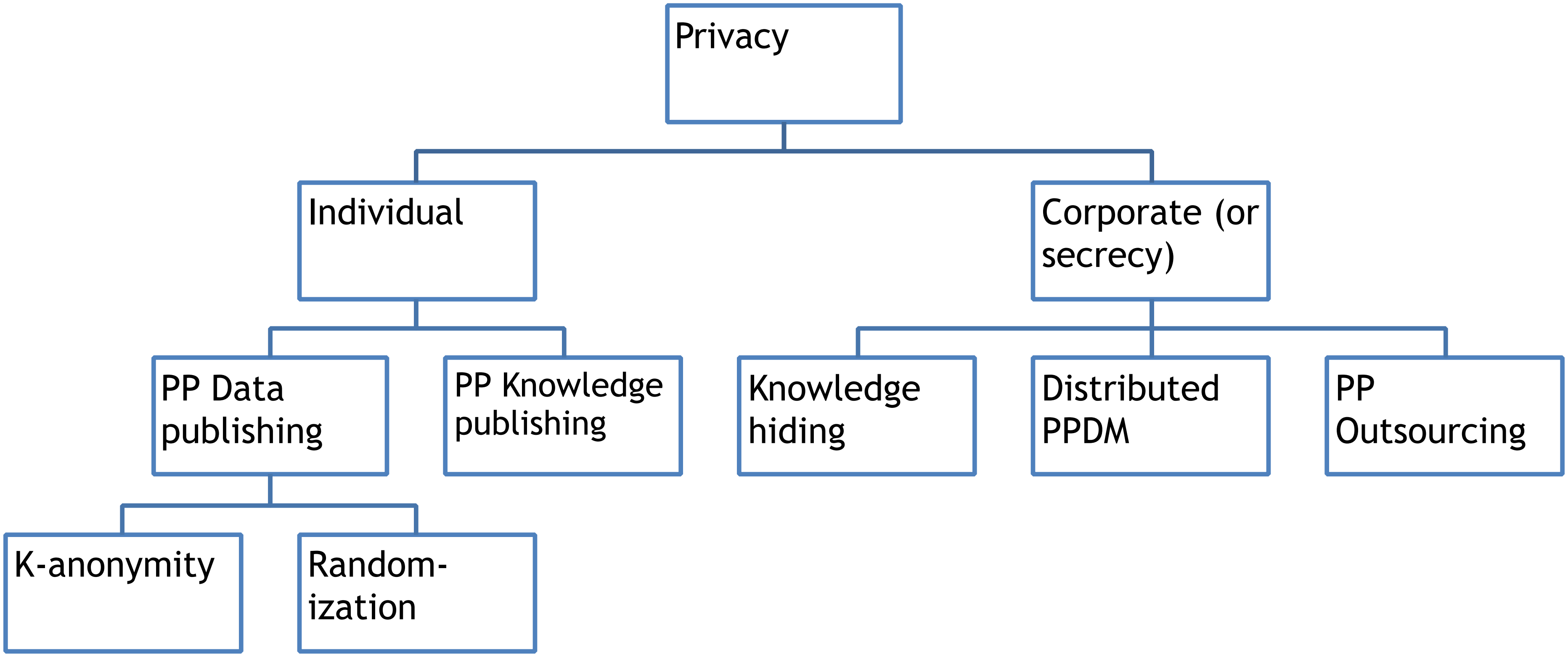
De-identified User Trajectory



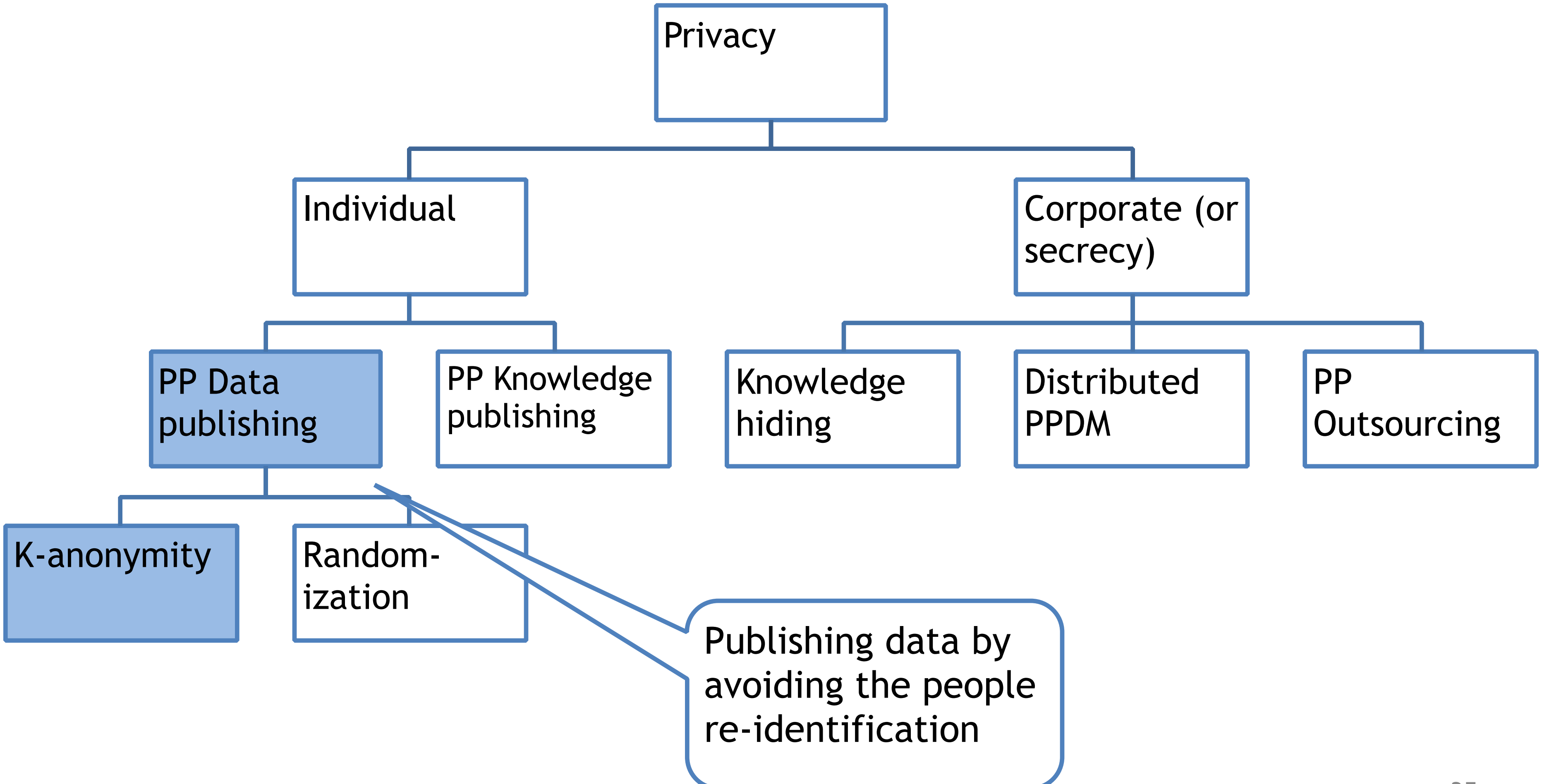


- Discovering persons living in that home and working in that company we can identify the user

Ontology of Privacy in Data Analysis



Ontology of Privacy in Data Analysis

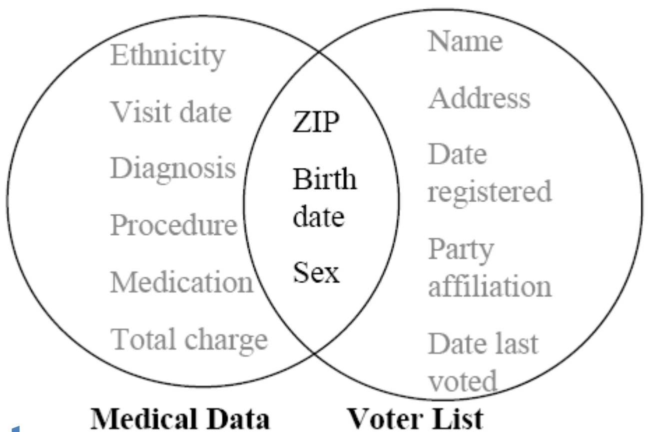


Data K-anonymity

- What is disclosed?
  - the data (modified somehow)
- What is hidden?
  - the real data
- How?
  - by transforming the data in such a way that it is not possible the re-identification of original database rows under a fixed anonymity threshold (individual privacy)

Linking Attack

- 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



- **regarding the US 1990 census data**
  - 87% of the population are unique based on (zipcode, gender, birth date)

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

K-Anonymity

- k-anonymity: hide each individual among k-1 others
  - each QI set should appear at least k times in the released data
  - linking cannot be performed with confidence > 1/k
- How to achieve this?
  - Generalization: publish more general values, i.e., given a domain hierarchy, roll-up
  - Suppression: remove tuples, i.e., do not publish outliers. Often the number of suppressed tuples is bounded
- Privacy vs utility tradeoff
  - do not anonymize more than necessary
  - Minimize the distortion
- Complexity?
  - NP-Hard!! [Meyerson and Williams PODS '04]

Classification of Attributes

Key Attribute	Quasi-Identifier			Sensive Attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

Example

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Figure 2 Example of *k*-anonymity, where *k*=2 and QI={*Race*, *Birth*, *Gender*, *ZIP*}

K-anonymity Vulnerability

- **k-anonymity** does not provide privacy if:
  - Sensitive values in an equivalence class lack diversity
  - The attacker has background knowledge
- This leads to the l-Diversity model:

Lack diversity

Bob	
Zipcode	***Age***
47678	27

Background Knowledge
(Carl's brother has heart disease)

Carl	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

l-Diversity

- Principle
  - Each equivalence class has at least *l* well-represented sensitive values
- Distinct *l*-diversity
  - Each equivalence class has at least *l* distinct sensitive values
  - Probabilistic inference

…	**Disease**
	…
	HIV
	HIV
	…
	HIV
	pneumonia
	bronchitis
	…

10 records

8 records have HIV

2 records have other values

Limitations of *l*-Diversity

l-Diversity is insufficient to prevent attribute disclosure.

Similarity Attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

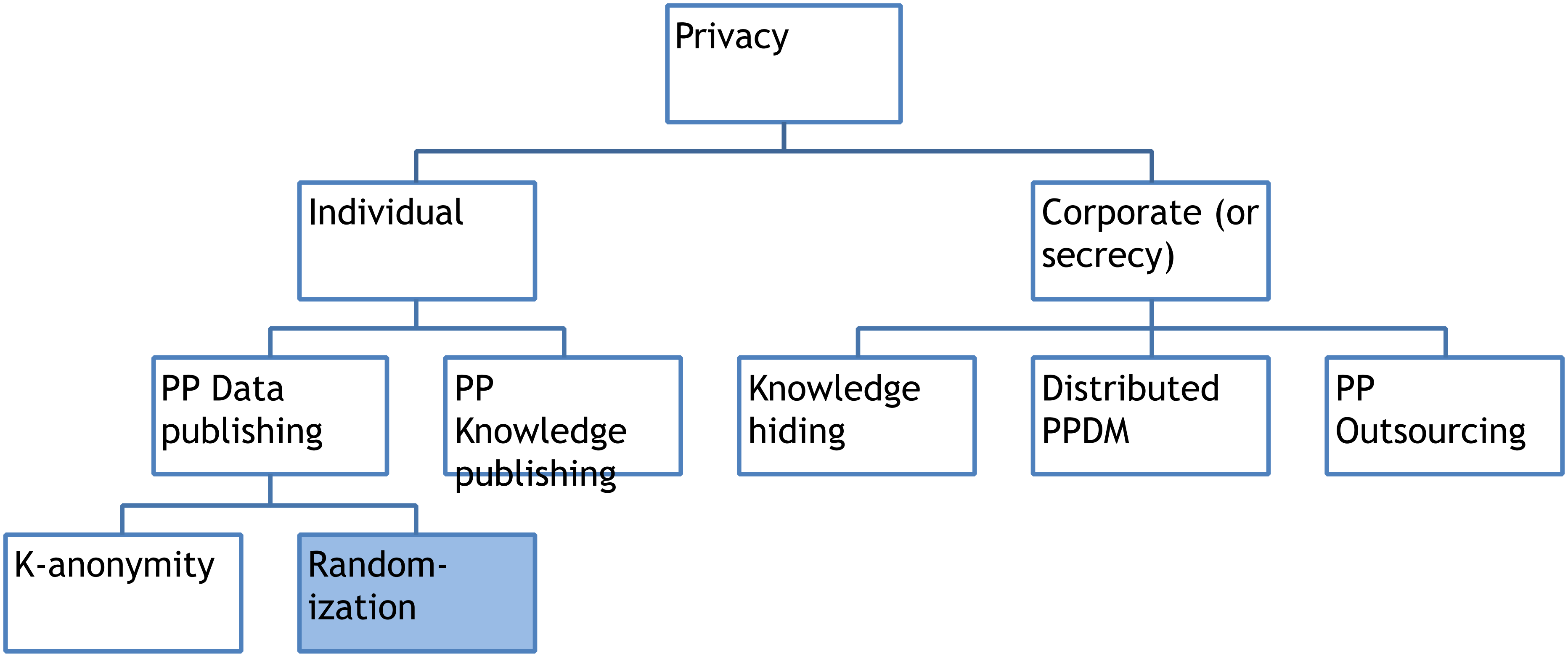
1. Bob's salary is in [20k,40k], which is relative low.
2. Bob has some stomach-related disease.

l-Diversity does not consider semantic meanings of sensitive values

K-Anonymity

- Samarati, Pierangela, and Latanya Sweeney. "Generalizing data to provide anonymity when disclosing information (abstract)."
  In PODS '98.
- Latanya Sweeney: k-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5): 557-570 (2002)
- Machanavajjhala, Ashwin, Daniel Kifer, Johannes Gehrke, and Muthuramakrish- nan Venkitasubramaniam. "*l*-diversity: Privacy beyond *k*-anonymity." *ACM Trans. Knowl. Discov. Data* 1, no. 1 (March 2007): 24.
- Li, Ninghui, Tiancheng Li, and S. Venkatasubramanian. "*t*-Closeness: Privacy Beyond *k*-Anonymity and *l*-Diversity." *ICDE 2007*.

Ontology of Privacy in Data Analysis



Randomization

- What is disclosed?
  - the data (modified somehow)
- What is hidden?
  - the real data
- How?
  - by perturbating the data in such a way that it is not possible the identification of original database rows (individual privacy), but it is still possible to extract **valid** knowledge (models and patterns).

  - A.K.A. ***"distribution reconstruction"***

Problem

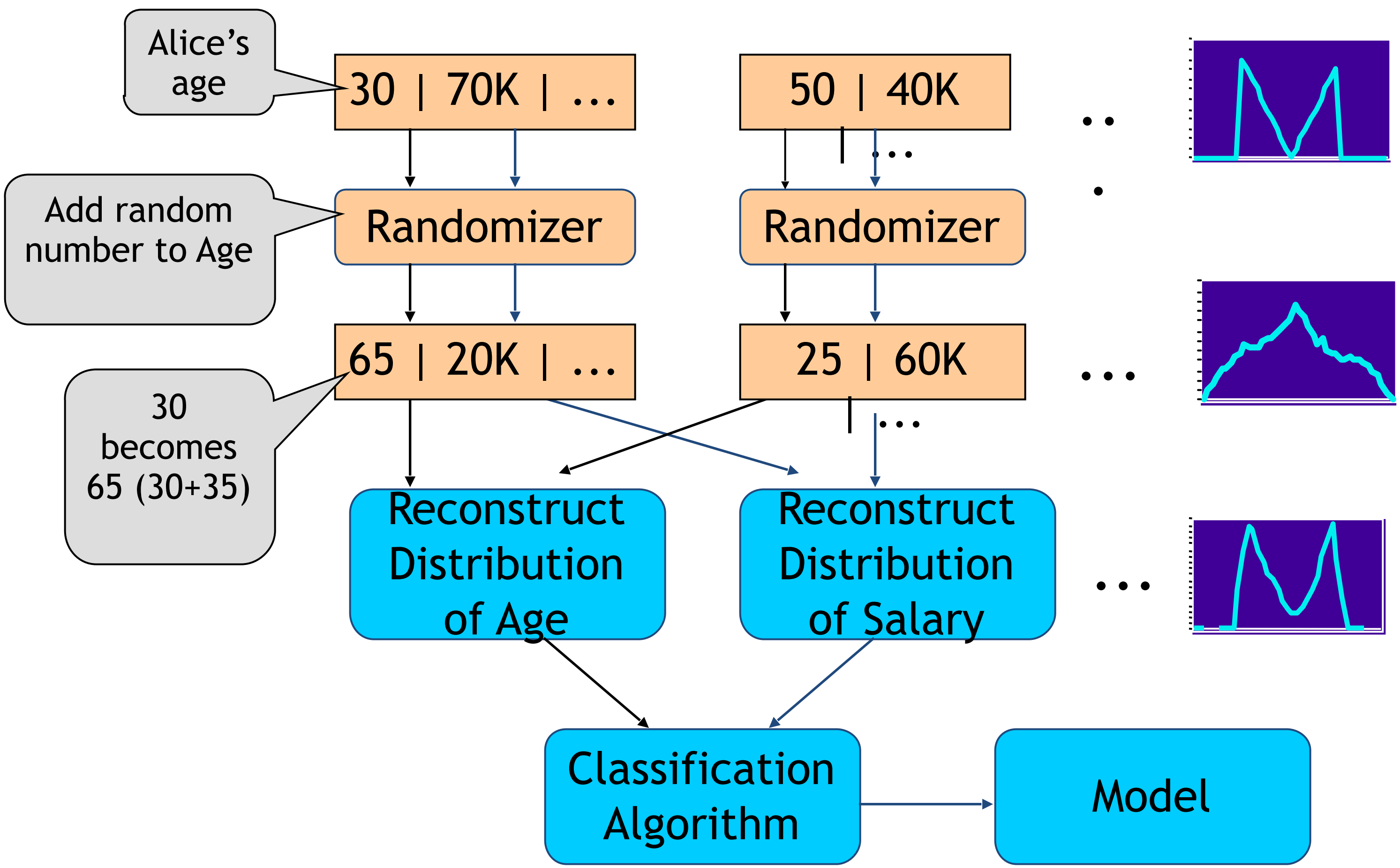
- **Original values $x_1, x_2, \ldots, x_n$**
  - from probability distribution X (unknown)
- **To hide these values, we use $y_1, y_2, \ldots, y_n$**
  - from probability distribution Y
    - Uniform distribution between $[-\alpha, \alpha]$
    - Gaussian, normal distribution with $\mu = 0, \sigma$

- Given
  - $x_1+y_1, x_2+y_2, \ldots, x_n+y_n$
  - the probability distribution of Y

Estimate the probability distribution of X.

R. Agrawal and R. Srikant. Privacy-preserving data mining. In Proceedings of SIGMOD 2000.

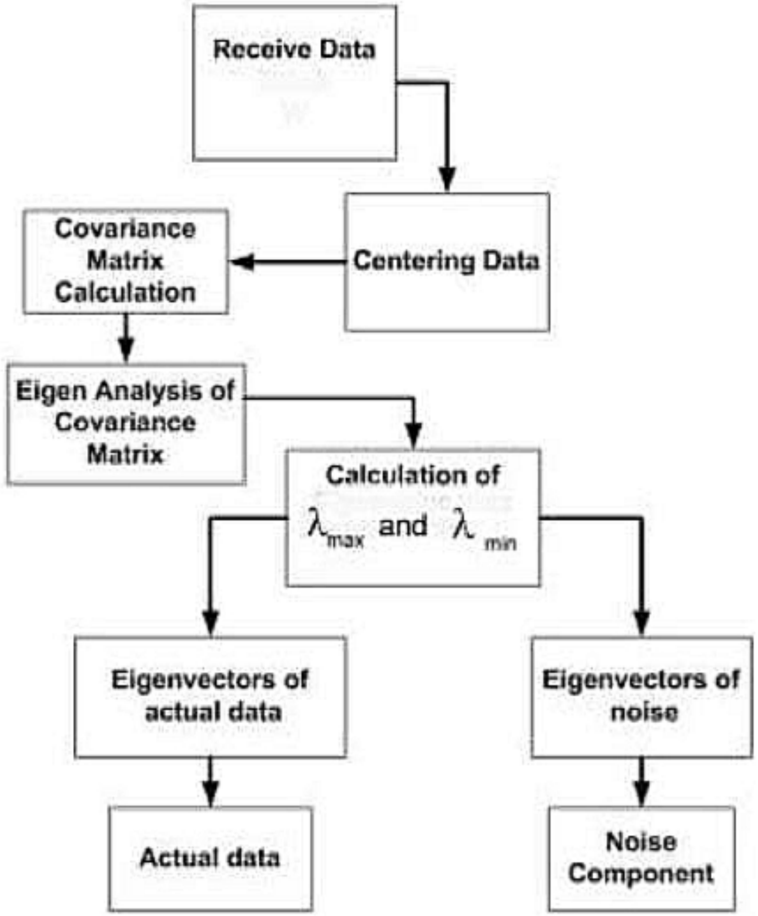
Randomization Approach Overview



Why is privacy preserved?

- Cannot reconstruct individual values accurately
- Can only reconstruct distributions

Weakness: Spectral Filtering Technique



Hillol Kargupta, Souptik Datta, Qi Wang, Krishnamoorthy Sivakumar: [On the Privacy Preserving Properties of Random Data Perturbation Techniques.](#) ICDM 2003:99-106

Assumptions of Spectral Filtering Technique

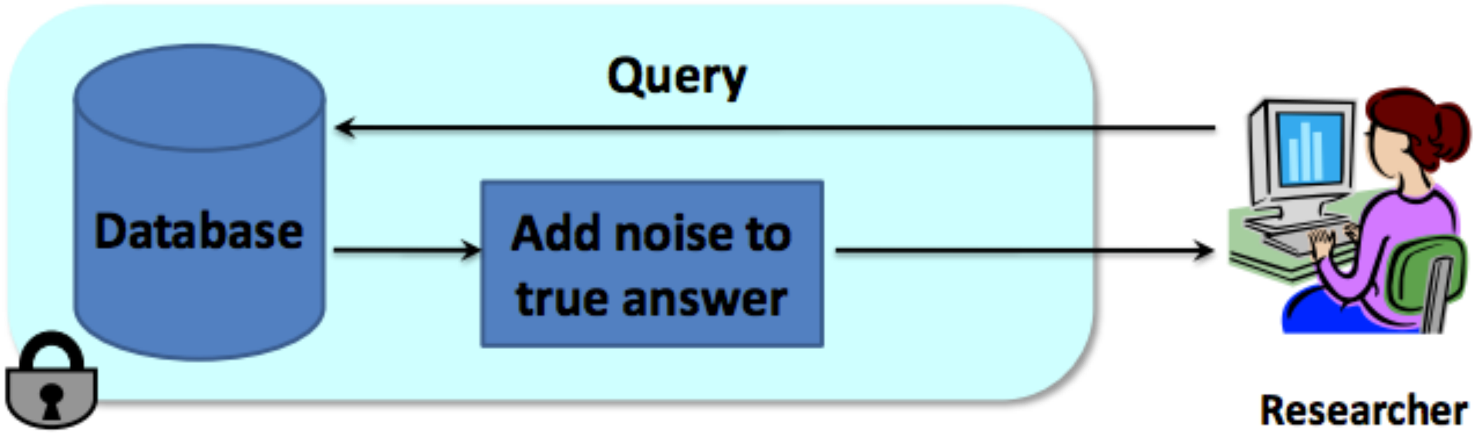
- This technique separates noise and original data in d-dimensional data, $(x_1, x_2, \ldots, x_d)$

- Two main assumptions:
  - **Correlation among attributes**
  - **Independence between noise and original data**

- The spectral filtering exploits the correlation among the attributes

Differential Privacy

- **Goal:** The risk to my privacy should not increase as a result of participating in a s



- Add noise to answers such that:
  - Each answer does not leak too much information about the database
  - Noisy answers are close to the original answers

Cynthia Dwork: Differential Privacy. ICALP (2) 2006: 1-12

Attack

Name	Has Diabetes
Alice	yes
Bob	no
Mark	yes
John	yes
Sally	no
Jack	yes

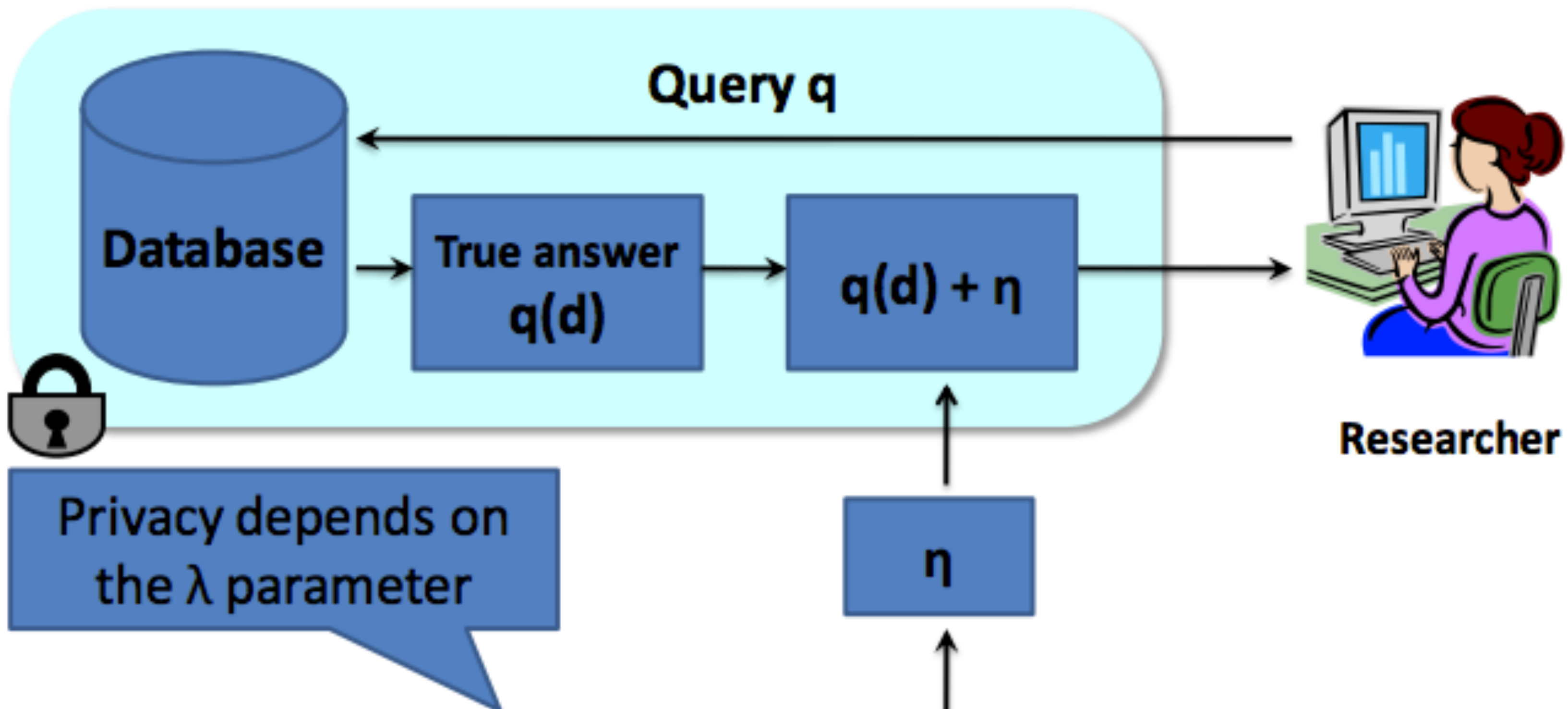
1) how many persons have Diabetes? **4**
2) how many persons, excluding Alice, have Diabetes? **3**

- **So the attacker can infer that Alice has Diabetes.**

- Solution: make the two answer similar

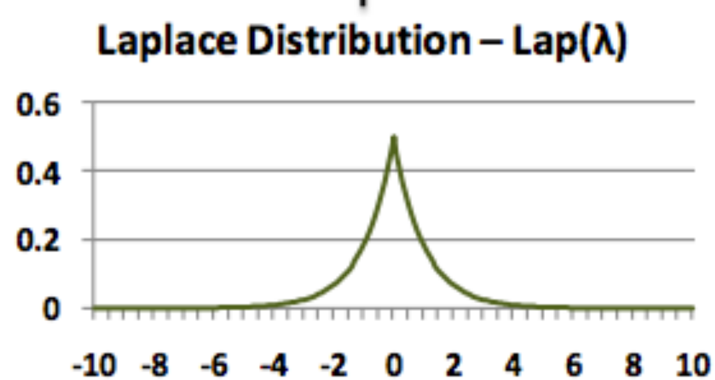
1) the answer of the first query could be 4+1 = 5
2) the answer of the second query could be 3+2.5=5.5

Differential Privacy



$$h(\eta) = \exp(-\eta / \lambda)$$

Mean: 0,
Variance: $2\lambda^2$



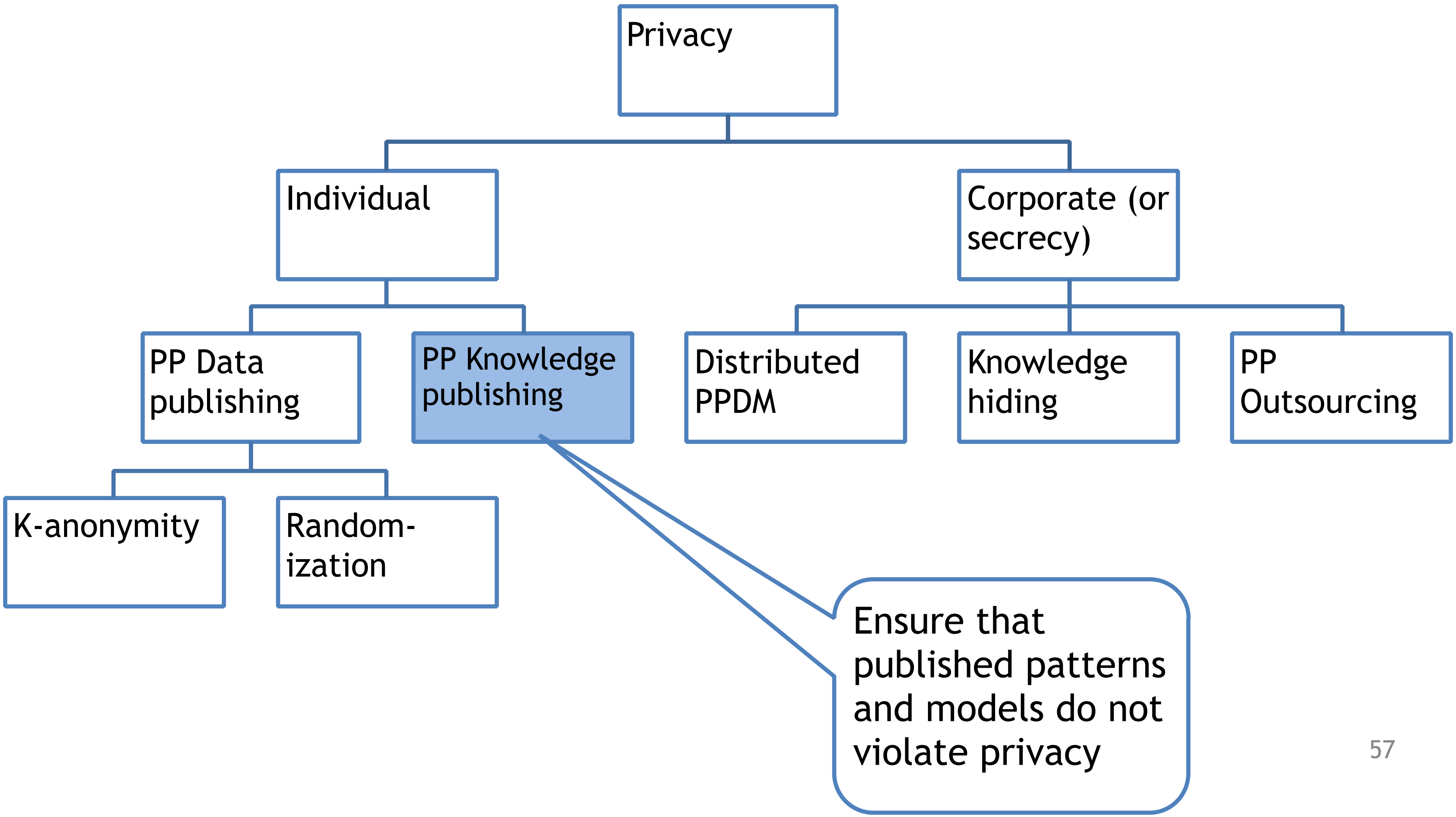
Randomization

- R. Agrawal and R. Srikant. Privacy-preserving data mining. In Proceedings of SIGMOD 2000.
- D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In Proceedings of PODS, 2001.
- W. Du and Z. Zhan. Using randomized response techniques for privacy-preserving data mining. In Proceedings of SIGKDD 2003.
- A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In Proceedings of PODS 2003.
- A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In Proceedings of SIGKDD 2002.
- K. Liu, H. Kargupta, and J. Ryan. Random Projection-based Multiplicative Perturbation for Privacy Preserving Distributed Data Mining. IEEE Transactions on Knowledge and Data Engineering (TKDE), VOL. 18, NO. 1.
- K. Liu, C. Giannella and H. Kargupta. An Attacker's View of Distance Preserving Maps for Privacy Preserving Data Mining. In Proceedings of PKDD'06

Differential Privacy

- Cynthia Dwork: Differential Privacy. ICALP (2) 2006: 1-12
- Cynthia Dwork: The Promise of Differential Privacy: A Tutorial on Algorithmic Techniques. FOCS 2011: 1-2
- Cynthia Dwork: Differential Privacy in New Settings. SODA 2010: 174-183

Ontology of Privacy in Data Analysis



Privacy-aware Knowledge Sharing

- What is disclosed?
  - the intentional knowledge (i.e. rules/patterns/models)
- What is hidden?
  - the source data

- The central question:

"do the data mining results themselves violate privacy"

Privacy-aware Knowledge Sharing

- Association Rules can be dangerous…

A: Age = 27, Postcode = 45254, Christian ⇒ American
(support = 758, confidence = 99.8%)

B: Age = 27, Postcode = 45254 ⇒ American
(support = 1053, confidence = 99.9%)

Since *sup(rule) / conf(rule) = sup(head)* we can derive:

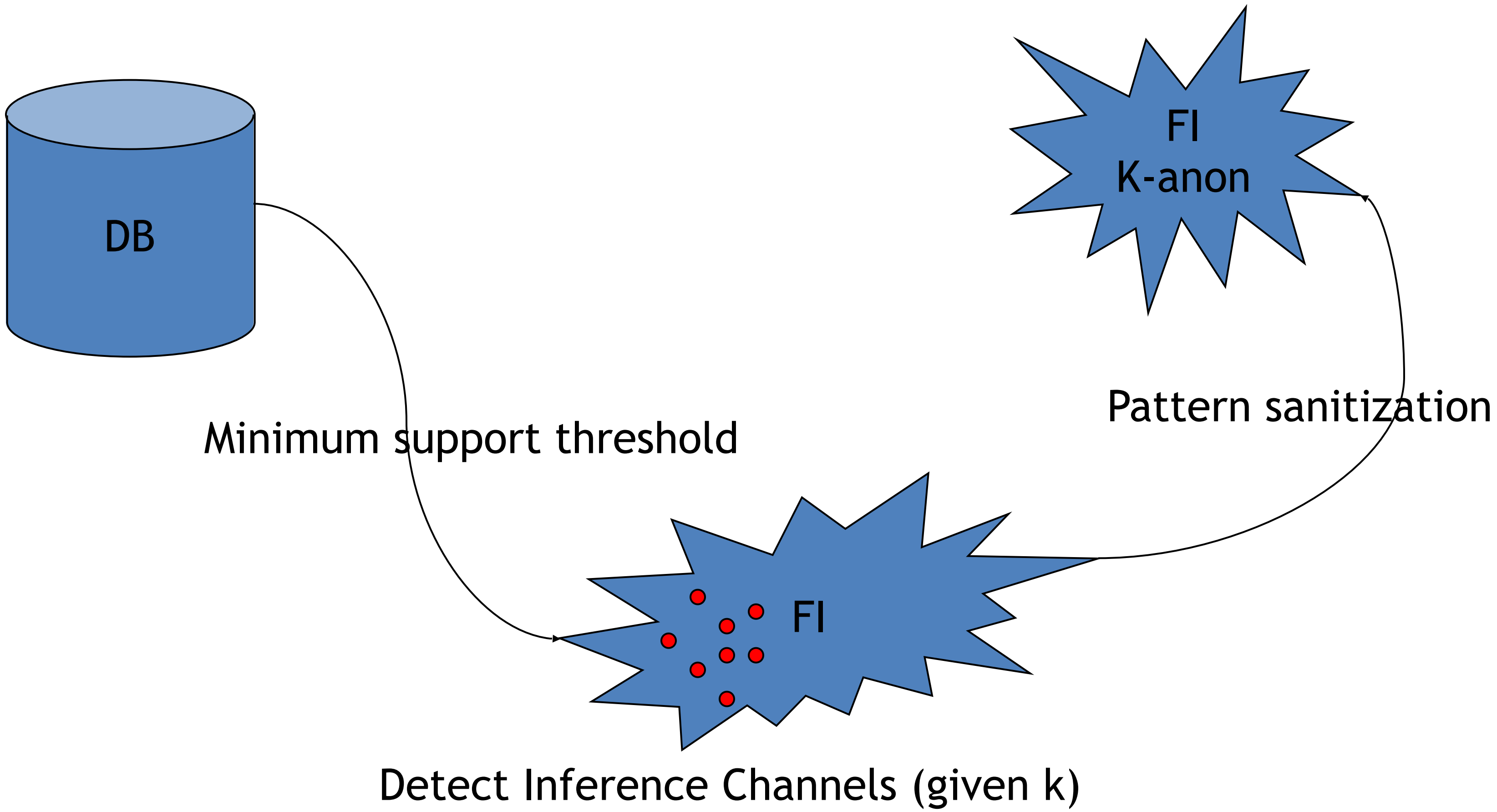
Age = 27, Postcode = 45254, not American
(support = 1,confidence = 100%)

Age = 27, Postcode = 45254, not American ⇒ Christian
(support = 1, confidence = 100.0%)

This information refers to my France neighbor…. he is Christian!

- How to solve this kind of problems?

The scenario



Privacy-aware Knowledge Sharing

- M. Kantarcioglu, J. Jin, and C. Clifton. When do data mining results violate privacy? In Proceedings of the tenth ACM SIGKDD, 2004.

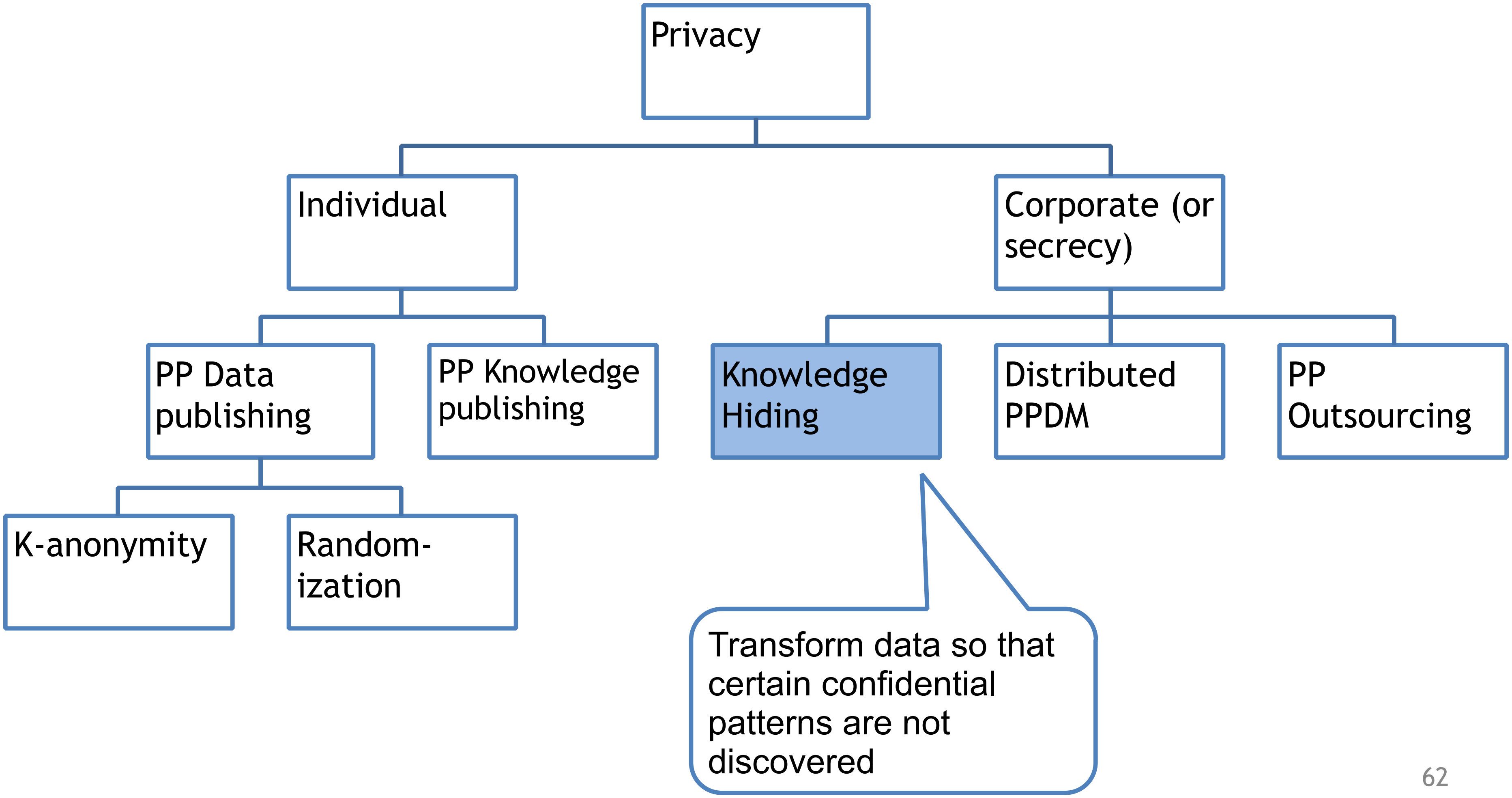
- S. R. M. Oliveira, O. R. Zaiane, and Y. Saygin. Secure association rule sharing. In Proc.of the 8th PAKDD, 2004.

- P. Fule and J. F. Roddick. Detecting privacy and ethical sensitivity in data mining results. In Proc. of the 27° conference on Australasian computer science, 2004.

- Maurizio Atzori, Francesco Bonchi, Fosca Giannotti, Dino Pedreschi: Anonymity preserving pattern discovery. VLDB J. 17(4): 703-727 (2008)

- A. Friedman, A. Schuster and R. Wolff. *k*-Anonymous Decision Tree Induction. In Proc. of PKDD 2006.

Ontology of Privacy in Data Analysis



Knowledge Hiding

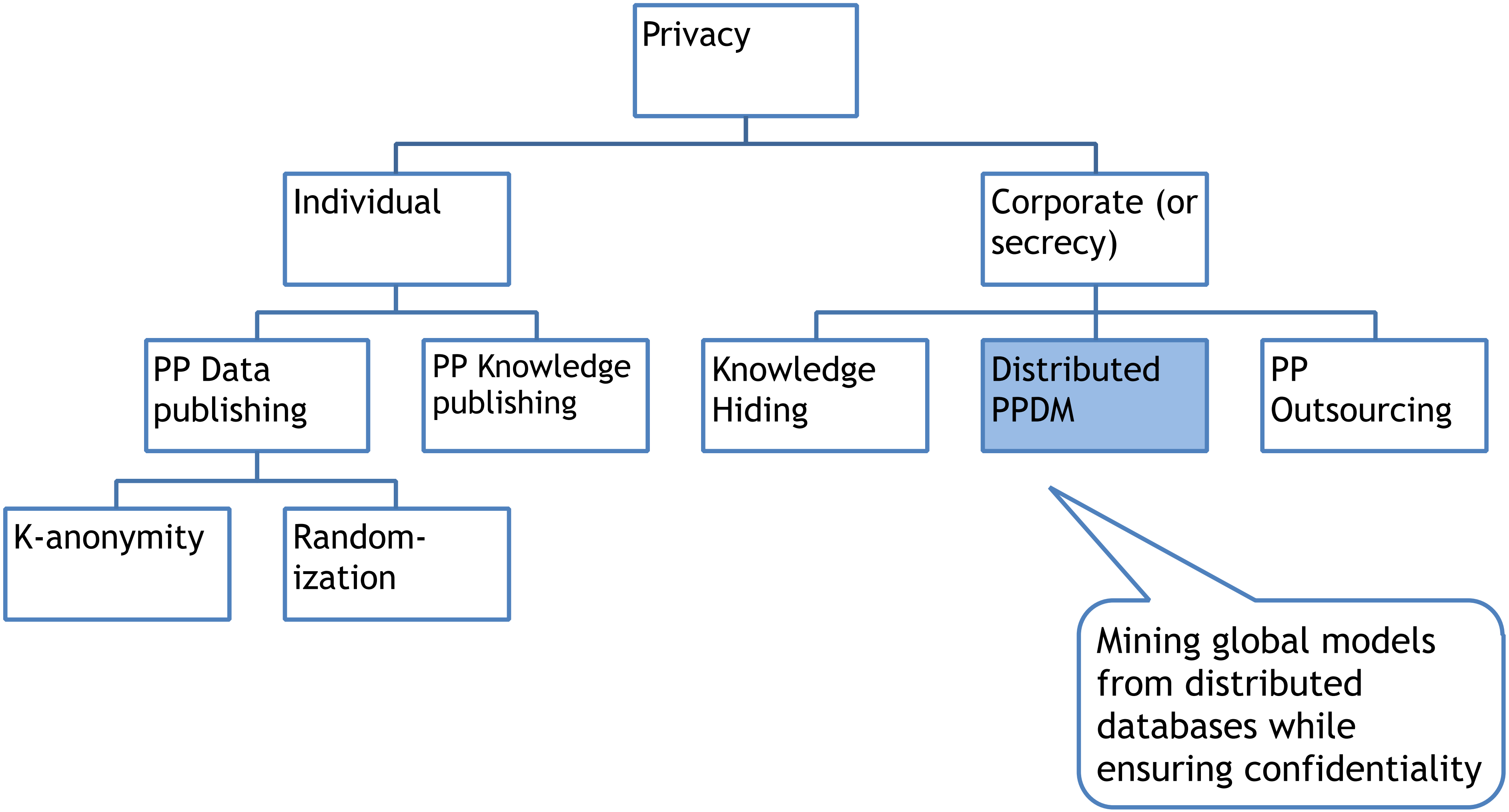
- What is disclosed?
  - the data (modified somehow)
- What is hidden?
  - some “sensitive” knowledge (i.e. secret rules/patterns)
- How?
  - usually by means of data **sanitization**
    - the data which we are going to disclose is modified in such a way that the sensitive knowledge can non longer be inferred
    - the original database is modified as less as possible.

Knowledge Hiding

- This approach can be instantiated to association rules as follows:
  - D source database;
  - R a set of association rules that can be mined from D;
  - R_h a subset of R which must be hidden.
- **Problem:** how to transform D into D' (the database we are going to disclose) in such a way that $R \setminus R_h$ can be mined from D'

- Typical solution is to reduce the confidence or support of rules

Ontology of Privacy in Data Analysis



Distributed Privacy Preserving Data Mining

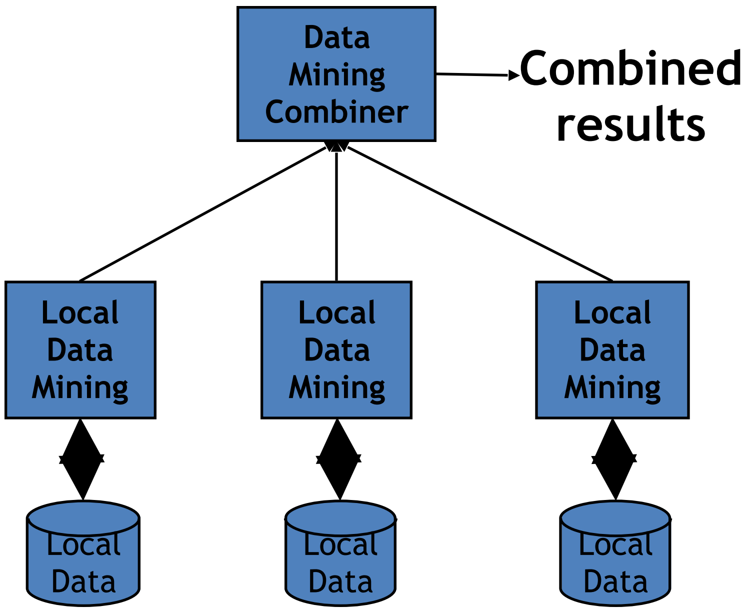
- Objective?
  - computing a valid mining model from several **distributed datasets**, where each party owing a dataset does not communicate its data to the other parties involved in the computation
- How?
  - cryptographic techniques

- A.K.A. “*Secure Multiparty Computation*”

Secure Multyparty Computation

How to compute the results without sharing data except the final result of the data mining result?

Many protocols for computation of
- secure sum
- secure set union
- secure size of intersection
- scalar product



Horizontal Partitioned Data

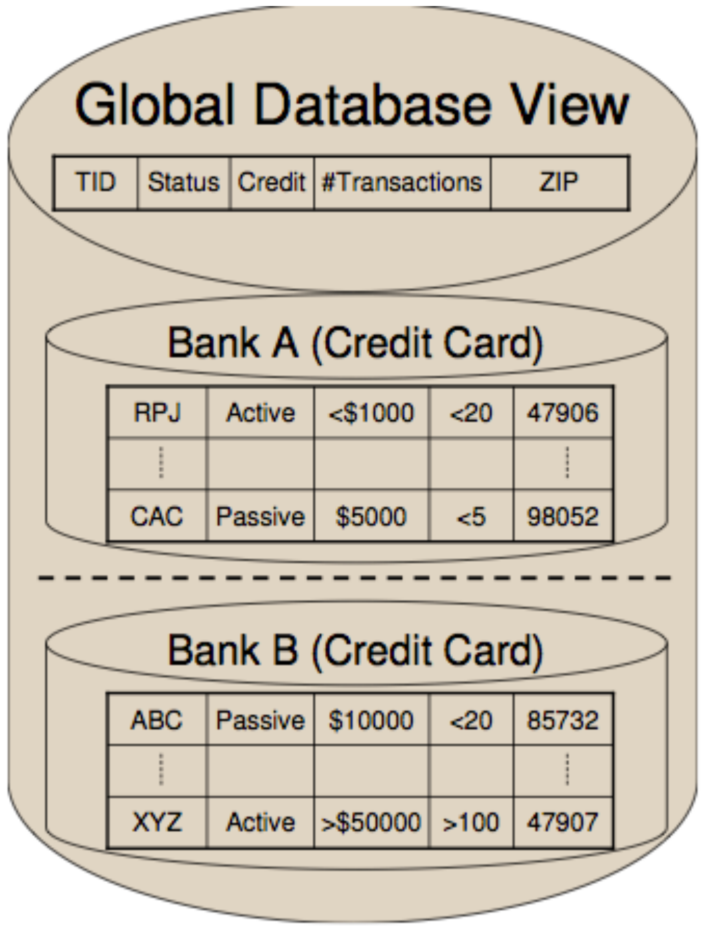


Figure 2.2. Horizontally partitioned database

Vertically Partitioned Data

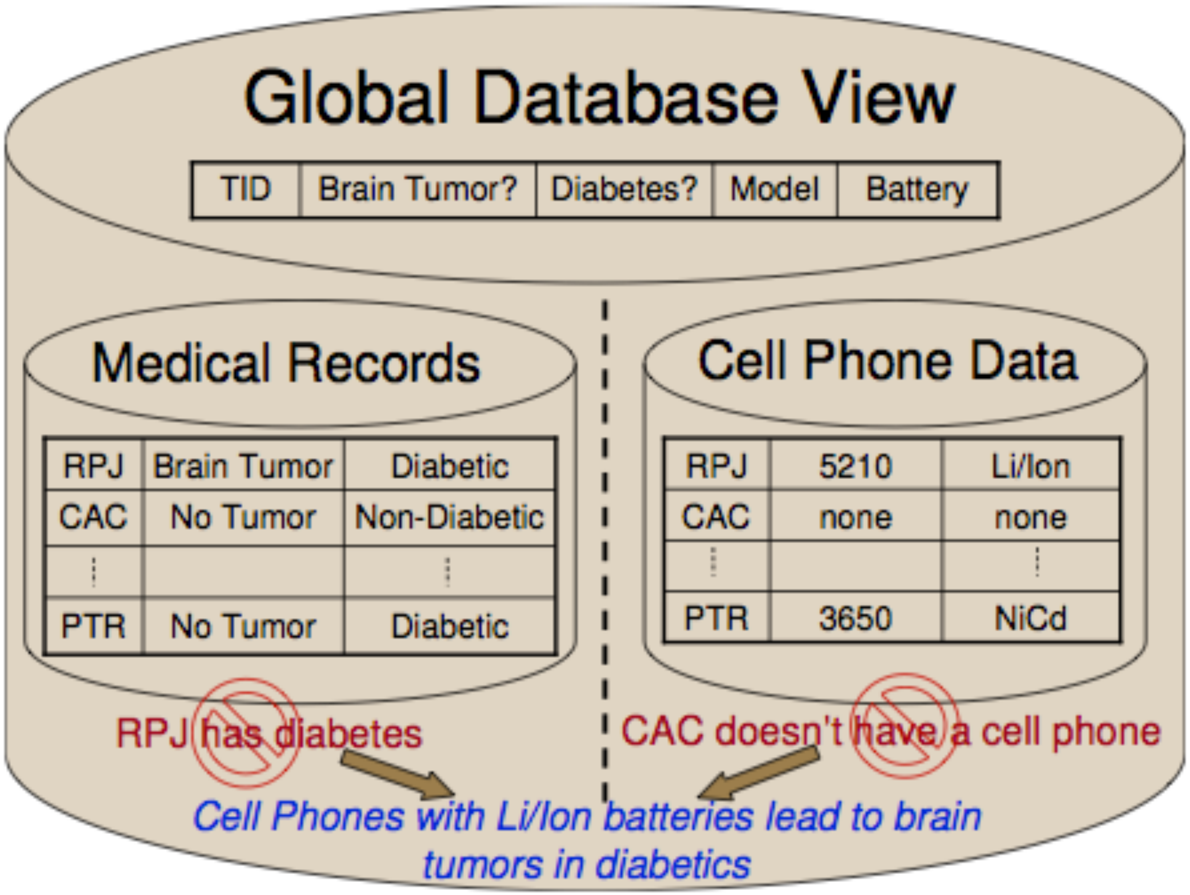


Figure 2.1. Vertically partitioned database

Distributed Privacy Preserving Data Mining

- C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y.Zhu. Tools for privacy preserving distributed data mining. SIGKDD Explor. Newsl., 4(2), 2002.

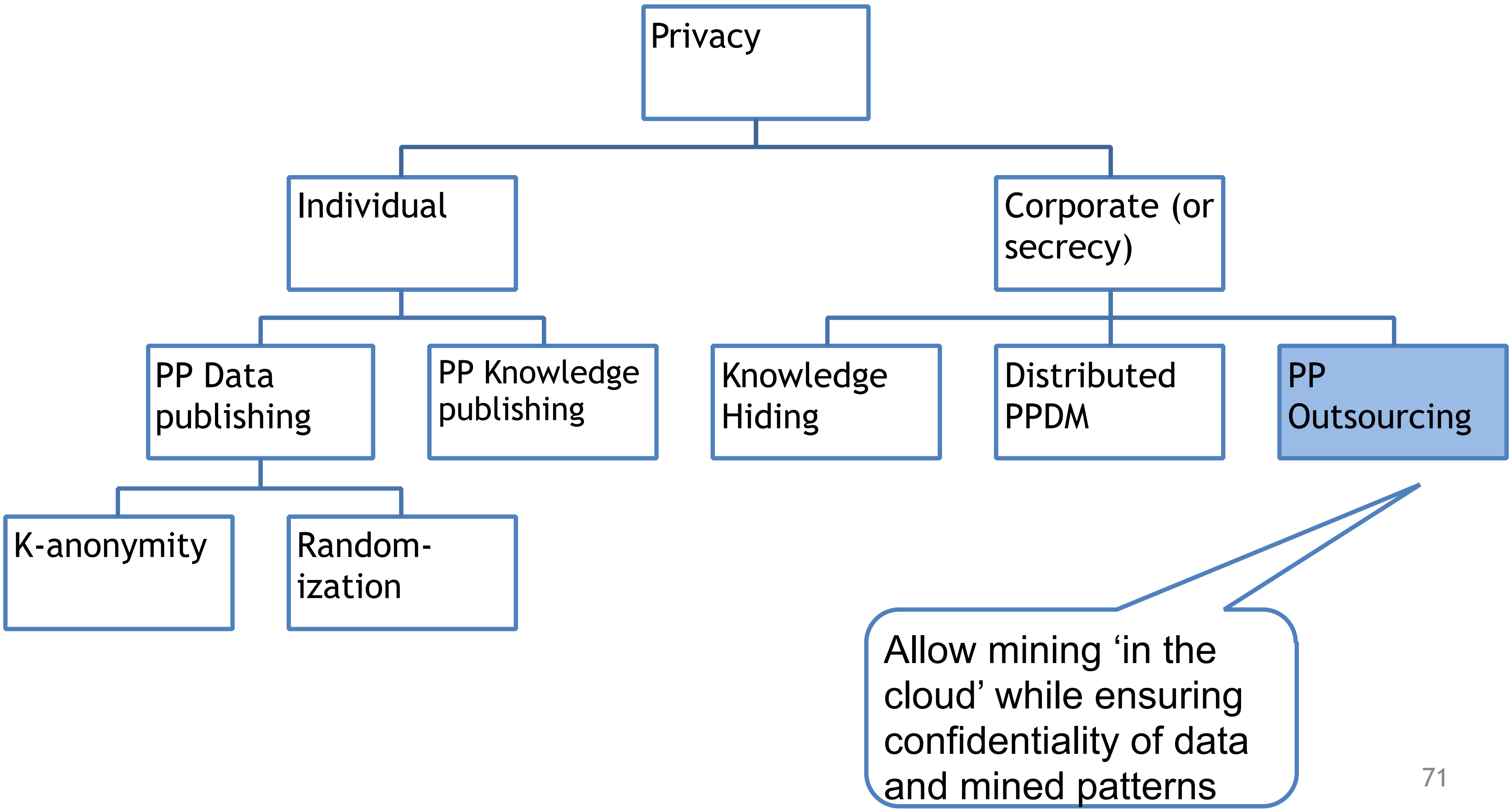
- M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. In SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'02), 2002.

- B. Pinkas. Cryptographic techniques for privacy-preserving data mining. SIGKDD Explor. Newsl., 4(2), 2002.

- J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In Proceedings of ACM SIGKDD 2002.

- Stavros Papadopoulos, Aggelos Kiayias, Dimitris Papadias: Secure and efficient in-network processing of exact SUM queries. 517-528, ICDE 2011

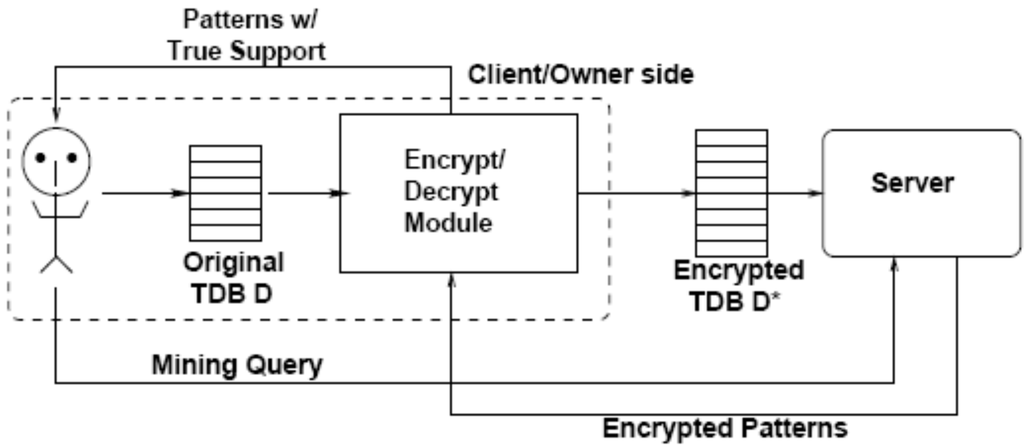
Ontology of Privacy in Data Analysis



Privacy-Preserving Outsourcing of DM

- Organizations could do not posses
  - **in-house expertise** for doing data mining
  - **computing infrastructure** adequate
- **Solution**: Outsourcing of data mining to a service provider
  - specific human resources
  - technological resources
- The server has access to data of the owner
- Data owner has the property of both
  - **Data** can contain personal information about individuals
  - **Knowledge** extracted from data can provide competitive advantages

Privacy-aware Outsourcing of FP



- The client encrypts its data using an encrypt/decrypt (ED) module
- ED module transforms the input data into an encrypted database
- The server conducts data mining and sends the patterns to the client
- The ED module recovers the true identity of the returned patterns

Fosca Giannotti, Laks V.S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui Wang. Privacy-preserving data mining from outsourced databases. CPDP, 2010.

Privacy-Preserving Outsourcing of DM

- W. K. Wong, David W. Cheung, Edward Hung, Ben Kao, and Nikos Mamoulis. Security in outsourcing of association rule mining. In *VLDB*, pages 111–122, 2007.
- C. Tai, P. S. Yu, and M. Chen. k-support anonymity based on pseudo taxonomy for outsourcing of frequent itemset mining. In *KDD*, pages 473–482, 2010.
- Fosca Giannotti, Laks V.S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui Wang. Privacy-preserving data mining from outsourced databases. In *SPCC2010, in conjunction with CPDP*, 2010.
- Ian Molloy, Ninghui Li, and Tiancheng Li. On the (in)security and (im)practicality of outsourcing precise association rule mining. In *ICDM*, pages 872–877, 2009.
- Ling Qiu, Yingjiu Li, and Xintao Wu. Protecting business intelligence and customer privacy while outsourcing data mining tasks. *Knowledge Information System*, 17(1):99–120, 2008.

Opinion 05/2014: Recomnendations

- Each above technique fails to meet with certainty the criteria of **effective anonymisation**. However as some of these risks may be met in whole or in part by a given technique, careful engineering is necessary in devising the application of an individual technique to the specific situation and in applying a **combination of those techniques** as a way to enhance the robustness of the outcome.
- The optimal solution should be decided on **a case-by-case basis**: a solution meeting the three criteria would be robust against identification performed by the most likely and **reasonable means** the data controller or any third party may employ.
- Whenever a proposal does not meet one of the criteria, a thorough **evaluation of the identification risks** should be performed. This evaluation should be provided to the authority if national law requires that the authority shall assess or authorise the anonymisation process.

Application of Privacy-by-Design

- Many companies are realizing the necessity to
  - consider privacy at every stage of their business
  - integrate privacy requirements "by design" into their business model.

- **The main problem is that in many contexts it is not completely clear which are the approaches for incorporating privacy- by-design**

Privacy by Design 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**
    - Taking into account the **Data Minimization Principle**
  - guarantee that the analytical questions can be answered correctly, within a **quantifiable** approximation that specifies the **data utility**

Privacy by Design in Mobility Atlas

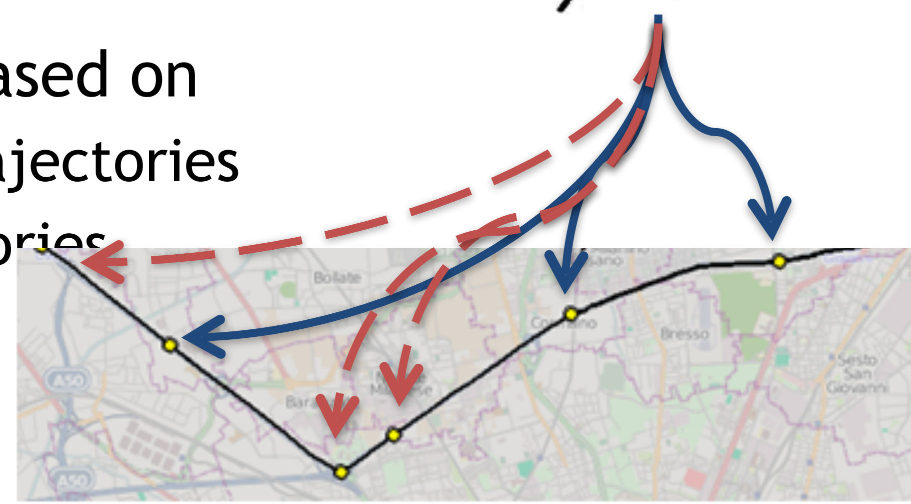
A. Monreale, G. Andrienko, N. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo

The Journal Transactions on Data Privacy, 2010

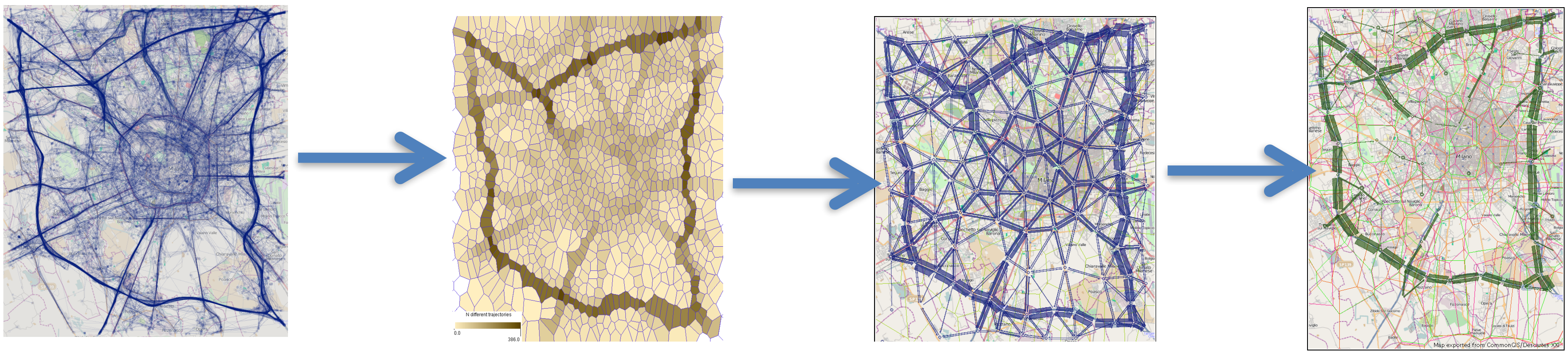
Knowledge Discovery and Delivery Lab
(ISTI-CNR & Univ. Pisa)
www-kdd.isti.cnr.it

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

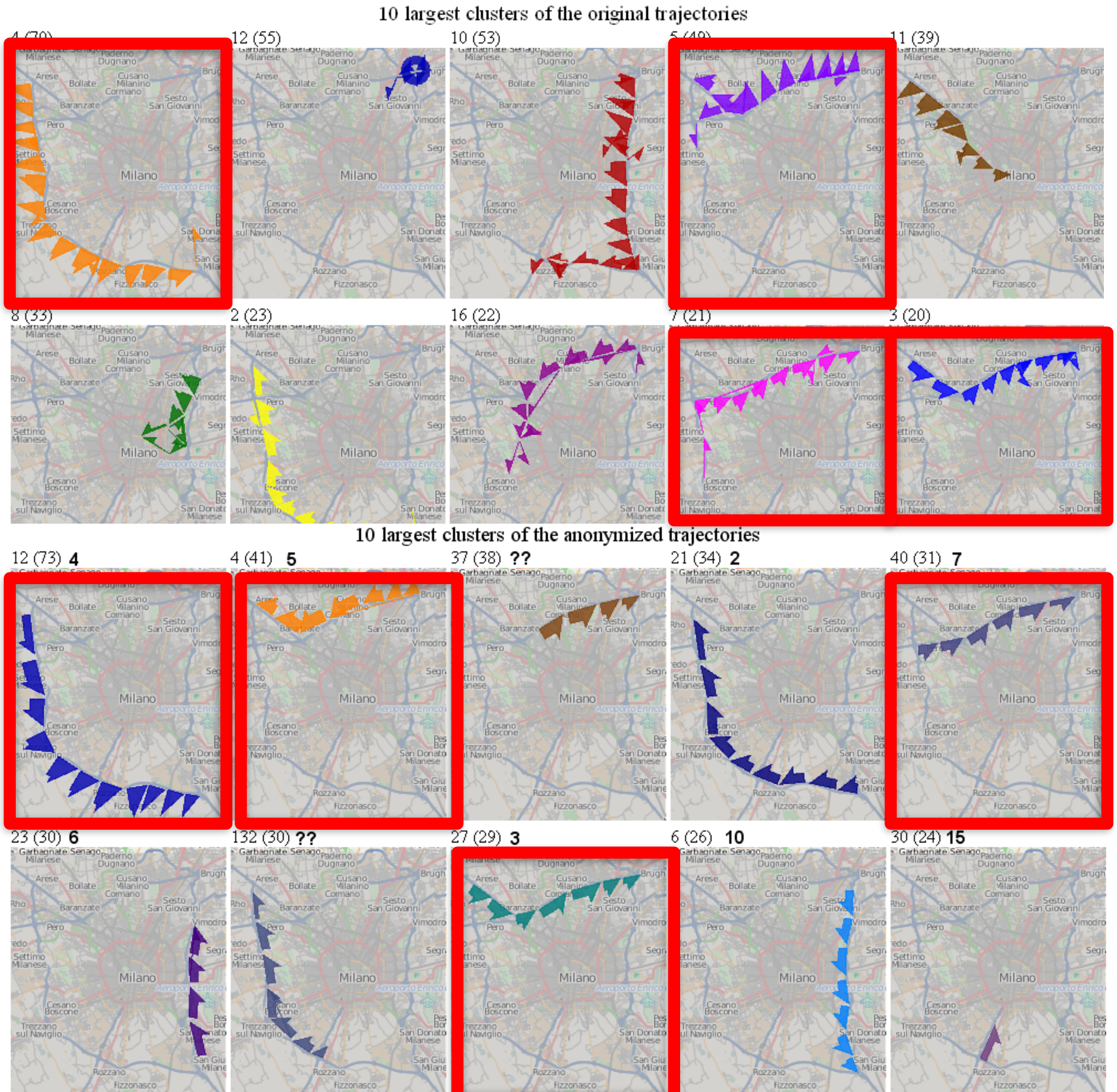


Trajectory Anonymization



- Given a 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



Probability of re-identification: k=16

Known Positions	Probability of re-identification
1 position	*98% trajectories have a P <= 0.03 (K=30)*
2 positions	*98% of trajectories have a P <= 0.05 (K=20)*
4 positions	*99% of trajectories have a P <= 0.06 (K=17)*
…..	

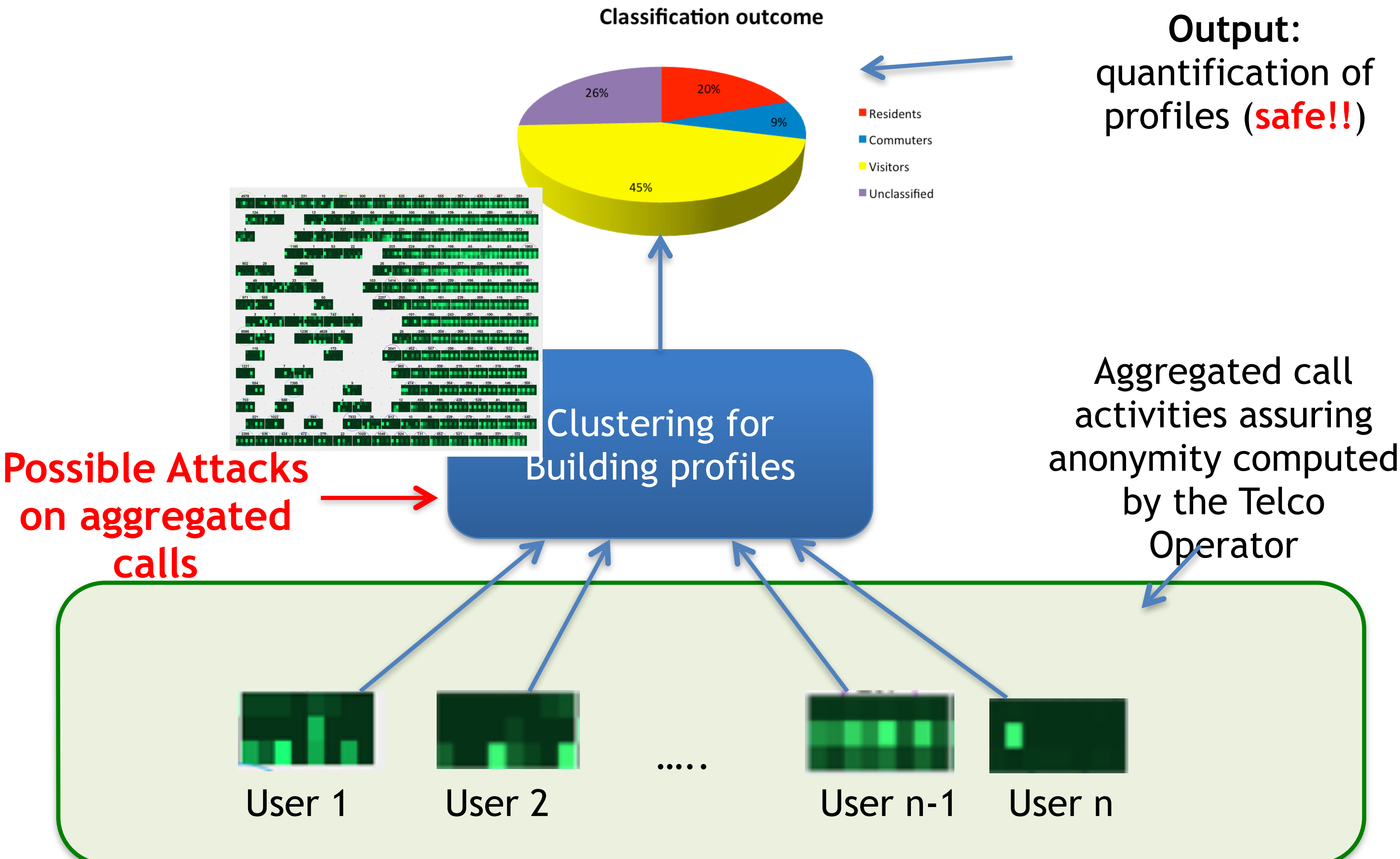
Privacy by Design in Mobile phone socio-meters Analysis

A. Monreale, F. Giannotti, D. Pedreschi, S. Rinzivillo

IEEE Big Data Conference, 2013

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

Privacy-Aware socio-meter



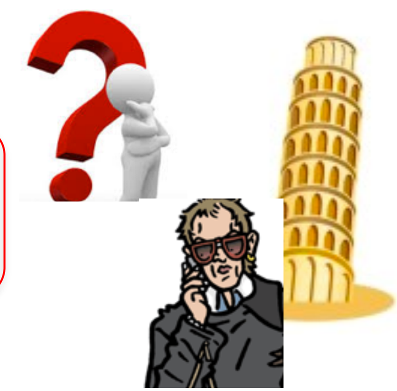
Attack risk based on Call Activites (Strong)

Analyst working on GSM data of 232K users with access to their call profiles

From: 02/04/14
To: 22/04/14

Apriori knowledge:
3 weeks of her boy-friend's call activity

Inference:
his activities in Pisa during the remaining week



From: 23/04/14
To: 29/04/14

Assumption: the attacker is not sure if the user is one of the profiles because he could not have any call activity in Pisa

Probability of re-identification for 4 weeks (232K GSM users)

Probability of re-identification		
% Users	**A priori Knowledge: 2 weeks**	**A priori Knowledge: 3 weeks**
30%	$P \leq 0.004\%$ (|C|=25,000)	$P \leq 0.006\%$ (|C|=15,000)
40%	$0.004\% < P \leq 0.02\%$ (|C|=5,000)	$0.006\% < P \leq 0.04\%$ (|C|=2,500)
20%	$0.02\% < P \leq 0.2\%$ (|C|=500)	$0.04\% < P \leq 0.4\%$ (|C|=250)
9,4%	$0.2\% < P \leq 0.8\%$ (|C|=125)	$0.4\% < P \leq 1\%$ (|C|=100)
0,6%	$0.8\% < P \leq 25\%$ (|C|=4)	$1\% < P \leq 50\%$ (|C|=2)

- |C| numbers of indistinguishable profiles

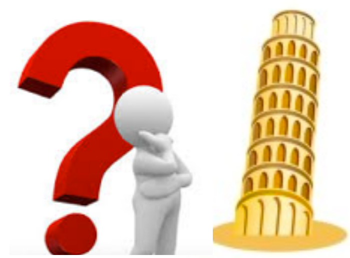
Attack risk based on User Presence (Reasonable)

Analyst working on GSM data of 232K users with access to their call profiles

From: 02/04/14
To: 22/04/14

Apriori knowledge:
For 3 weeks her boy-friend has been in Pisa

Inference:
was he in Pisa during the remaining 1 week?



From: 23/04/14
To: 29/04/14

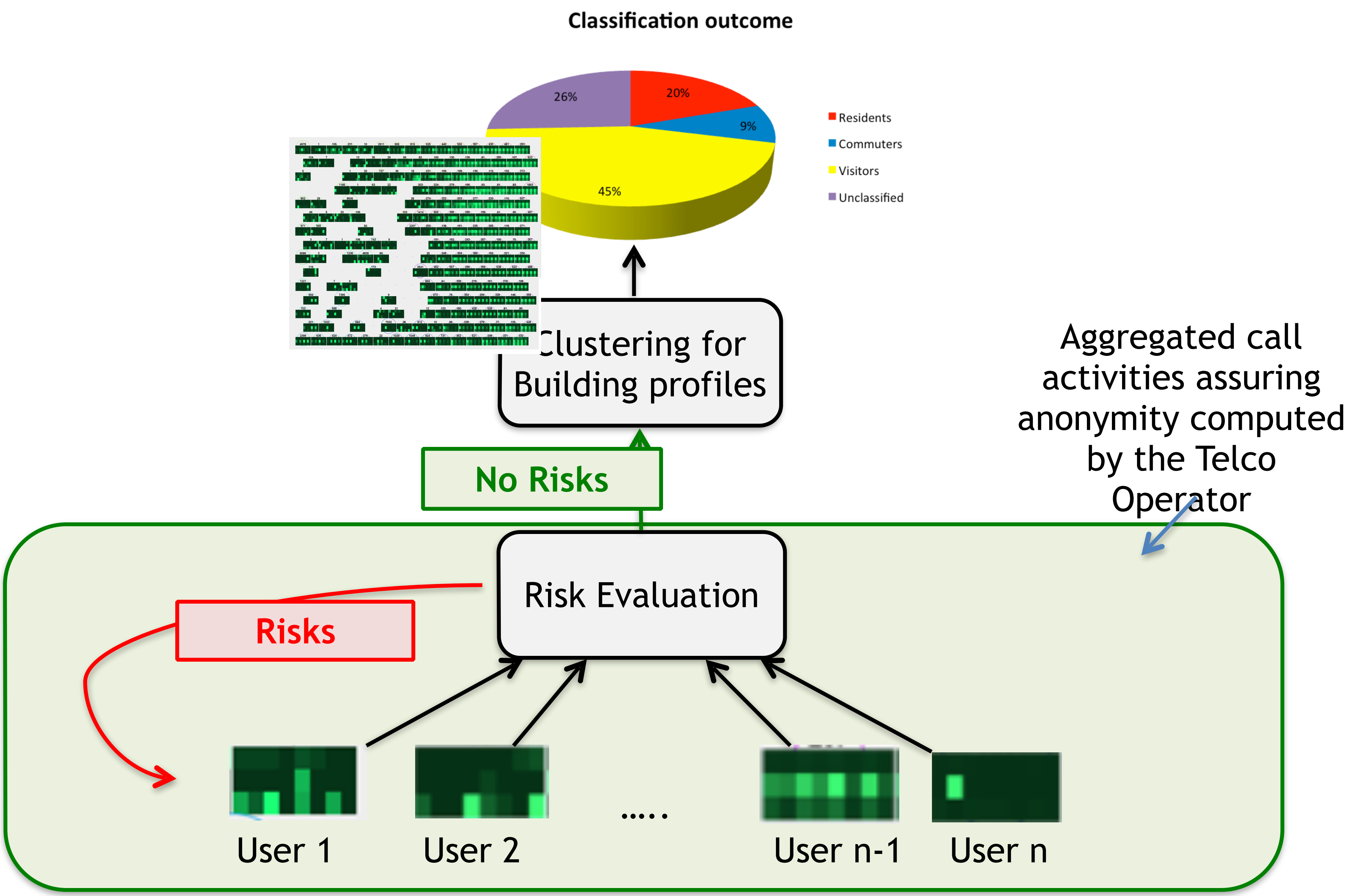
Assumption: the attacker is not sure if the user is one of the profiles because he could not have any call activity in Pisa

Probability of re-identification for 4 weeks (232K GSM users)

% Users	A priori Knowledge: 4 weeks
10%	$P \leq 0.003\%$ (|C|=33,000)
60%	$0.003\% < P \leq 0.017\%$ (|C|=5,800)
30%	$0.017\% < P \leq 0.025\%$ (|C|=4,000)

- |C| numbers of indistinguishable profiles

Privacy-Aware socio-meter



A change of perspective

A change of perspective

- The big data originate from the digital breadcrumbs of human activities
- Each person are becoming a statistical entity
- Only the single individual can link own digital breadcrumbs from his sources and extract a deep knowledge about himelf

Personal data as economic asset

- "*Personal data is the new oil of the Internet and the new currency of the digital world*"
  - Maglena Kuneva, former European Commissioner of Consumer Protection

- "*Personal data is emerging as a new economic asset class, a key resource for the 21st century that will touch all aspects of society*"
  - World Economic Forum report 2011

Liquid Data

- “Big data is a new asset,” says Alex Pentland, a computational social scientist and director of the Human Dynamics Lab at the M.I.T. “You want it to be liquid and to be used.”

WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

Personal Data: The Emergence of a New Asset Class

The new deal on data

- Quoting Alex (Sandy) Pentland (MIT) at WEF 2009

The first step toward open information markets is to give people ownership of their data*. The simplest approach to defining what it means to "own your own data" is to go back to Old English Common Law for the three basic tenets of ownership, which are the rights of*

- ***possession***,
- ***use***, *and*
- ***disposal***

WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

Industry Agenda

Unlocking the Value of Personal Data: From Collection to Usage

Prepared in collaboration with The Boston Consulting Group

February 2013

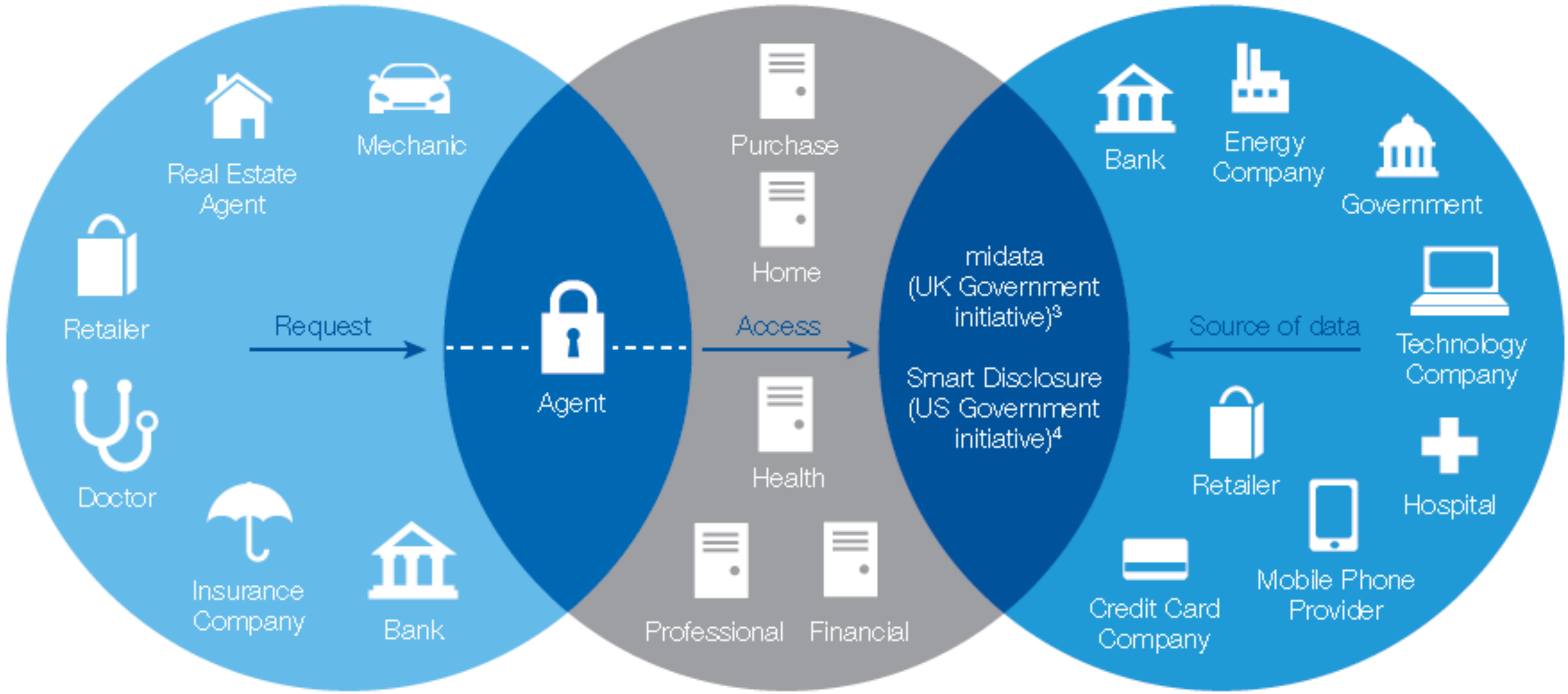
WEF's Key Concepts

- Shifting from governing the usage of data rather than the data itself
- Regulation has to take into account the context of data usage
- New ways to engage the individual, help them to understand and provide them the tools to make real choice based on clear valu exchange

Requesting party

Personal data store[1]

Data handback[2]



Companies who want to access data about individuals can request it through data agents

Several data stores are now up and running allowing individuals to exercise control over how data about them is used

Several governments are working with the private sector to give individuals access to a copy of data about them in a usable format which can then be stored in their locker and shared with other providers

Towards a new deal on personal data?

- **Full control of personal data / knowledge**
  - From informed consent to awareness, support for the management of own personal data and knowledge
- **Data liberation**
  - Right to withdraw personal data at any moment in full from any service provider
- **Oblivion**
  - Right to having personal data forgotten
- **Public good**
  - Right to have full access to the collective knowledge

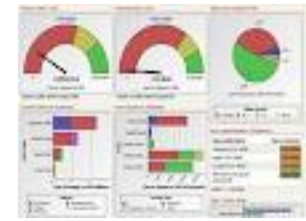
Individual knowledge and Collective knowledge

Collect
ive
Knowle
dge

Collective patterns

Individual profiles

Personal data store for self knowledge

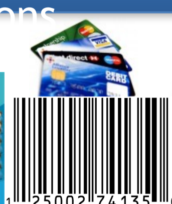


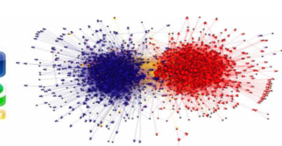
Self-awarenes

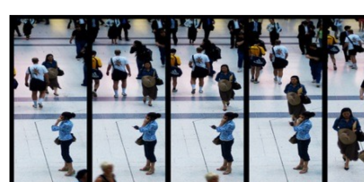
Novel Indicators

Social mining of many individuals (multidmensional, privacy and trust and donations

ocial mining of individual histories (multidimensiona

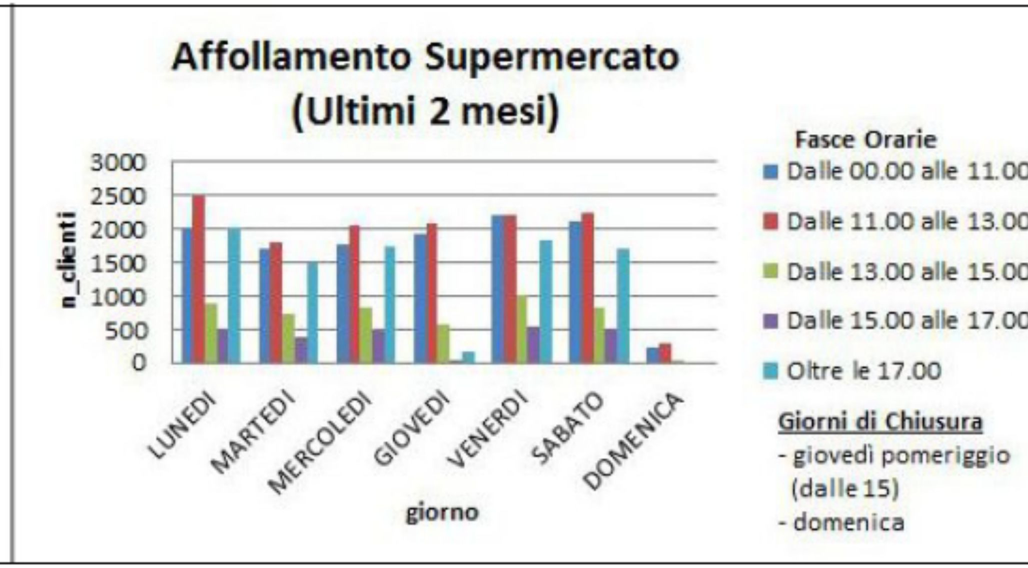






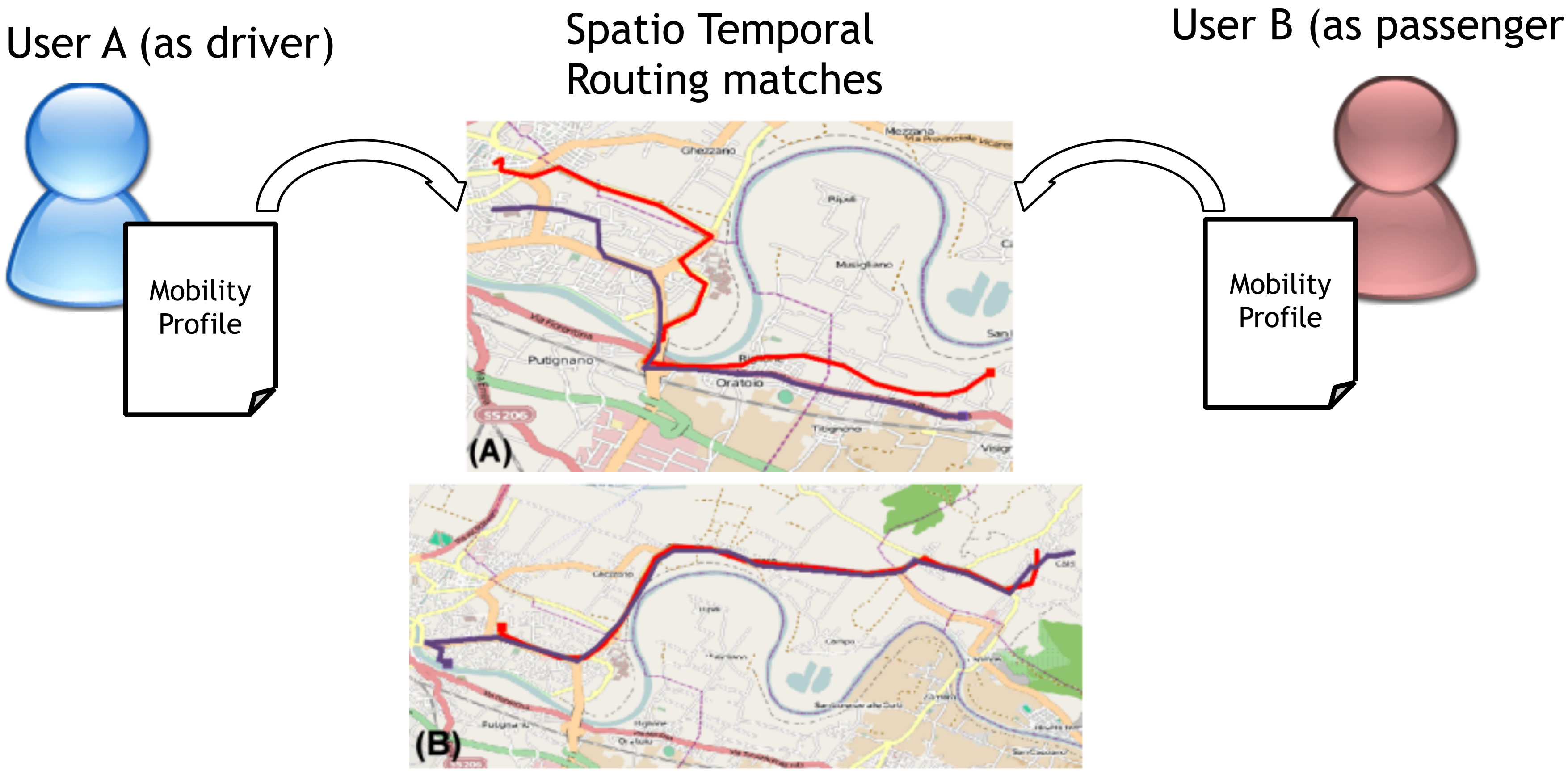
Traditional DB sources

When can I go to shopping?



giorno e ora (fascia oraria) di visita

Who can I share the car with?



New challenges for preserving privacy: User-Centric ecosystem

- How giving the control to individuals on the setting of the privacy level?

- How applying in this new context privacy-by-design?

- Which privacy model is suitable?

Mobility Analytics and Privacy in User-Centric Ecosystems

A.Monreale, H. Wang, F. Pratesi, D. Pedreschi, S. Rinzivillo, G. Andrienko, N. Andrieko

AGILE 2013

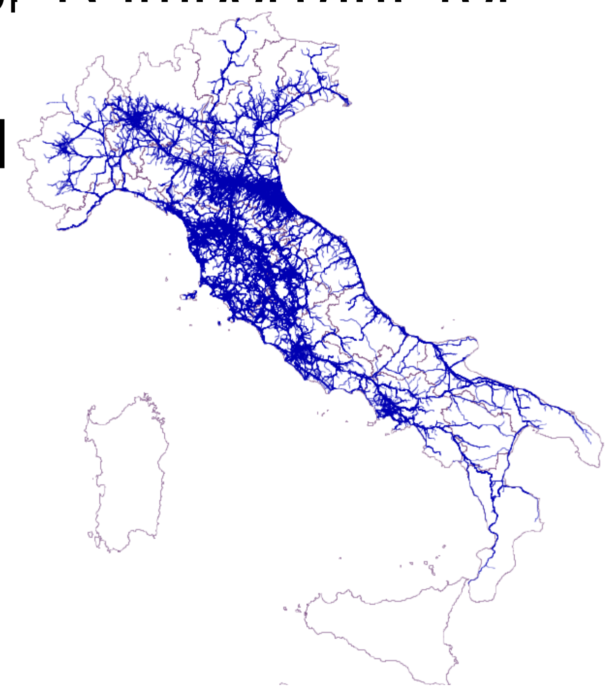
Knowledge Discovery and Delivery Lab
(ISTI-CNR & Univ. Pisa)
www-kdd.isti.cnr.it

Motivation

Availability of low cost GPS devices enables the collection of data about movements of people at a large scale

Understanding human mobility behavior is important for:

- improving the use of city space
- accessibility of various places and
- managing the traffic network
- reducing traffic jams



Generalization and summarization can help traffic data exploration

Distributed vs Centralized System

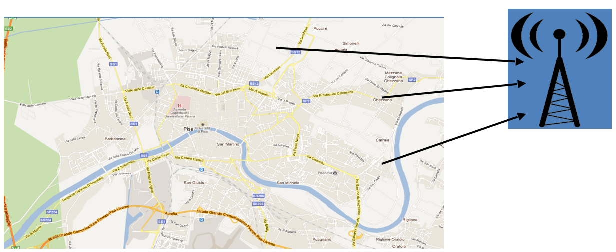
- Movement data of multiple individual devices can be collected and generalized and summarized by a central station.

- **Two important problems:**
  - computational resources
  - individual privacy at risk

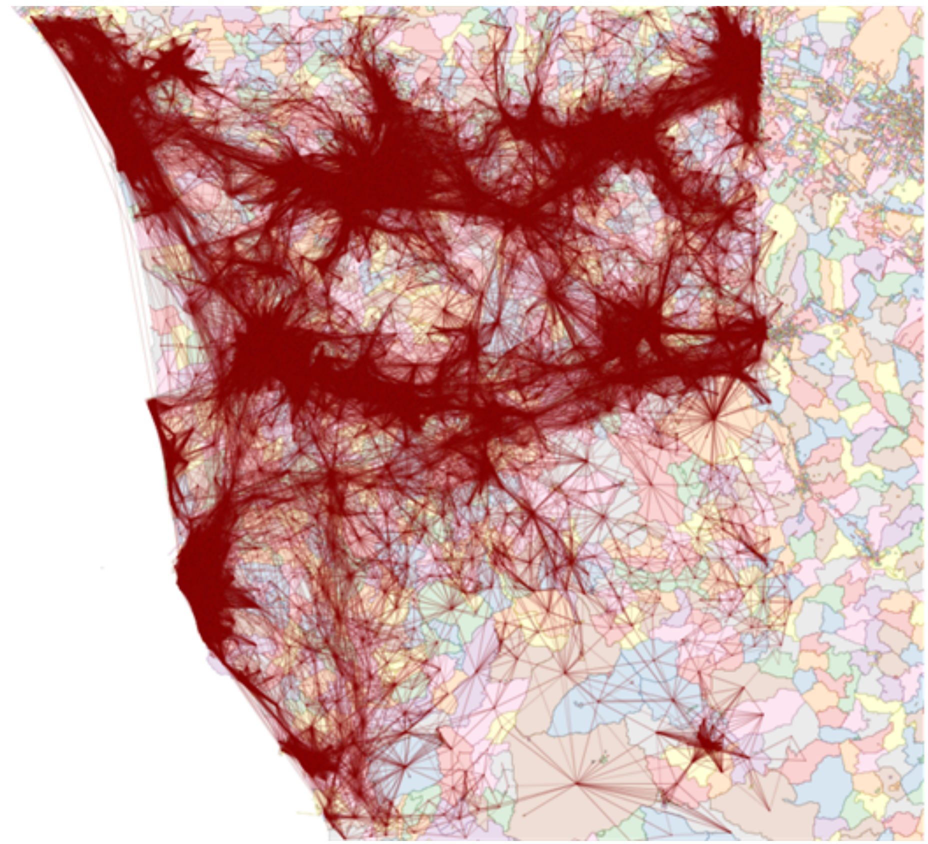
- **Solution:** Distributed Computation of data Generalization and summarization

Distributed Scenario

Vehicles collect **trajectories**, that can be transmitted (after a **generalization** step)



The coordinator computes a **data aggregation** describing the traffic flows



Trajectory Generalization

We start with a set of trajectories

We transform a trajectory in a **generalized trajectory**

We create a **frequency vector (similar to OD Matrix)**

f_j =

ab	ba	ac	ca	ad	da	bc	cb	bd	db	cd	dc
1	0	0	0	0	0	2	1	0	0	1	0

Privacy Issues

Privacy: From frequency vectors we can derive sensitive visits

- sometimes we can derive exactly trajectories
- the generalization it is not sufficient

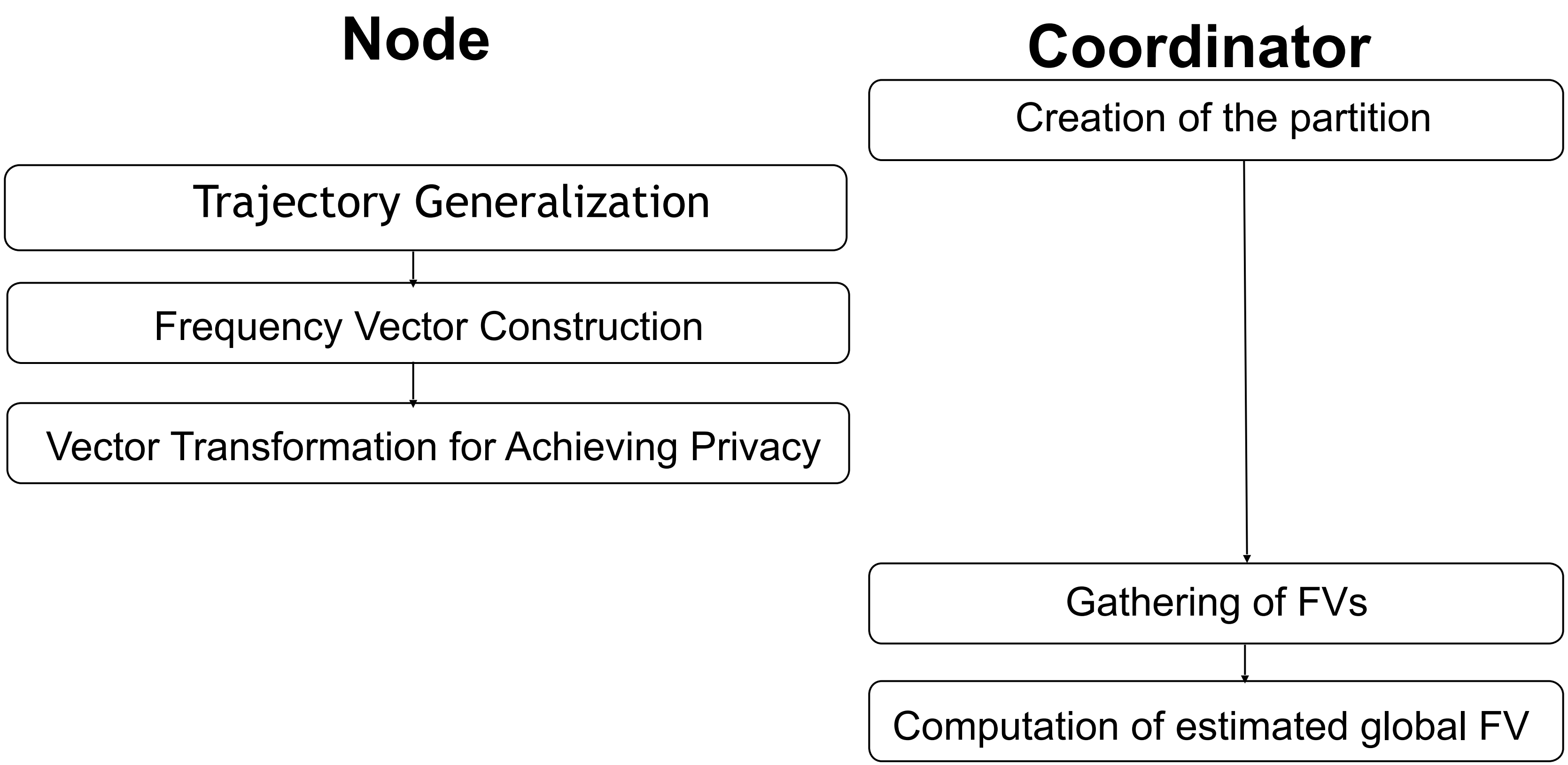
Privacy-Preserving Framework

- Distributed Randomization of individual OD matrix from GPS data while preserving global traffic flow

- **Linking Attack**: the attacker
  - wants to infer the movements from an area to another area of a specific user

- **Countermeasure** based on **Differential Privacy**

Mobility Analytics and Privacy in User-Centric Ecosystems



Mobility Analytics and Privacy in User-Centric Ecosystems

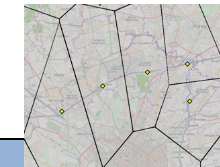
Node

Trajectory Generalization

Frequency Vector Construction

Vector Transformation for Achieving Privacy

Coordinator



Creation of the partition

Gathering of FVs

Computation of estimated global FV

Mobility Analytics and Privacy in User-Centric Ecosystems

Node

Trajectory Generalization

↓

Frequency Vector Construction

↓

Vector Transformation for Achieving Privacy

Coordinator

Creation of the partition

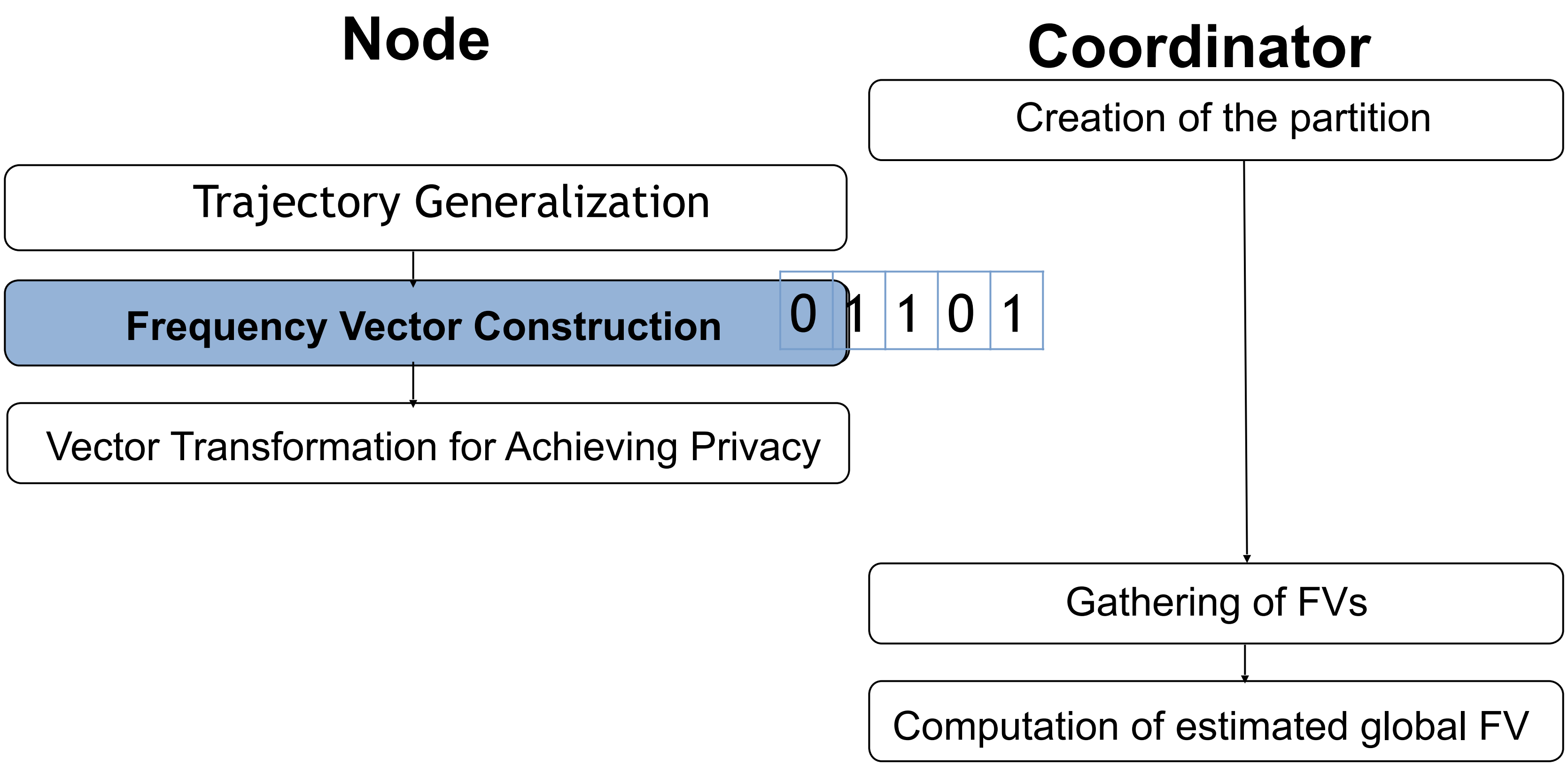
↓

Gathering of FVs

↓

Computation of estimated global FV

Mobility Analytics and Privacy in User-Centric Ecosystems



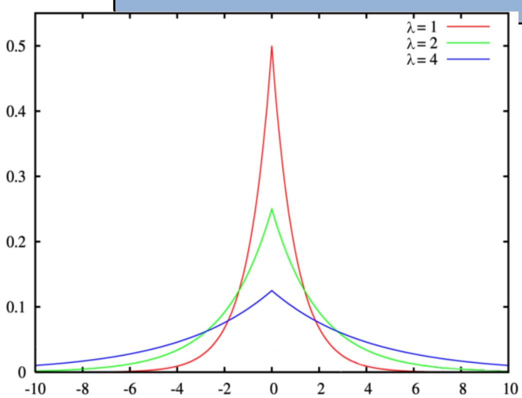
Mobility Analytics and Privacy in User-Centric Ecosystems

Node

Trajectory Generalization

Frequency Vector Construction

Vector Transformation for Achieving Privacy



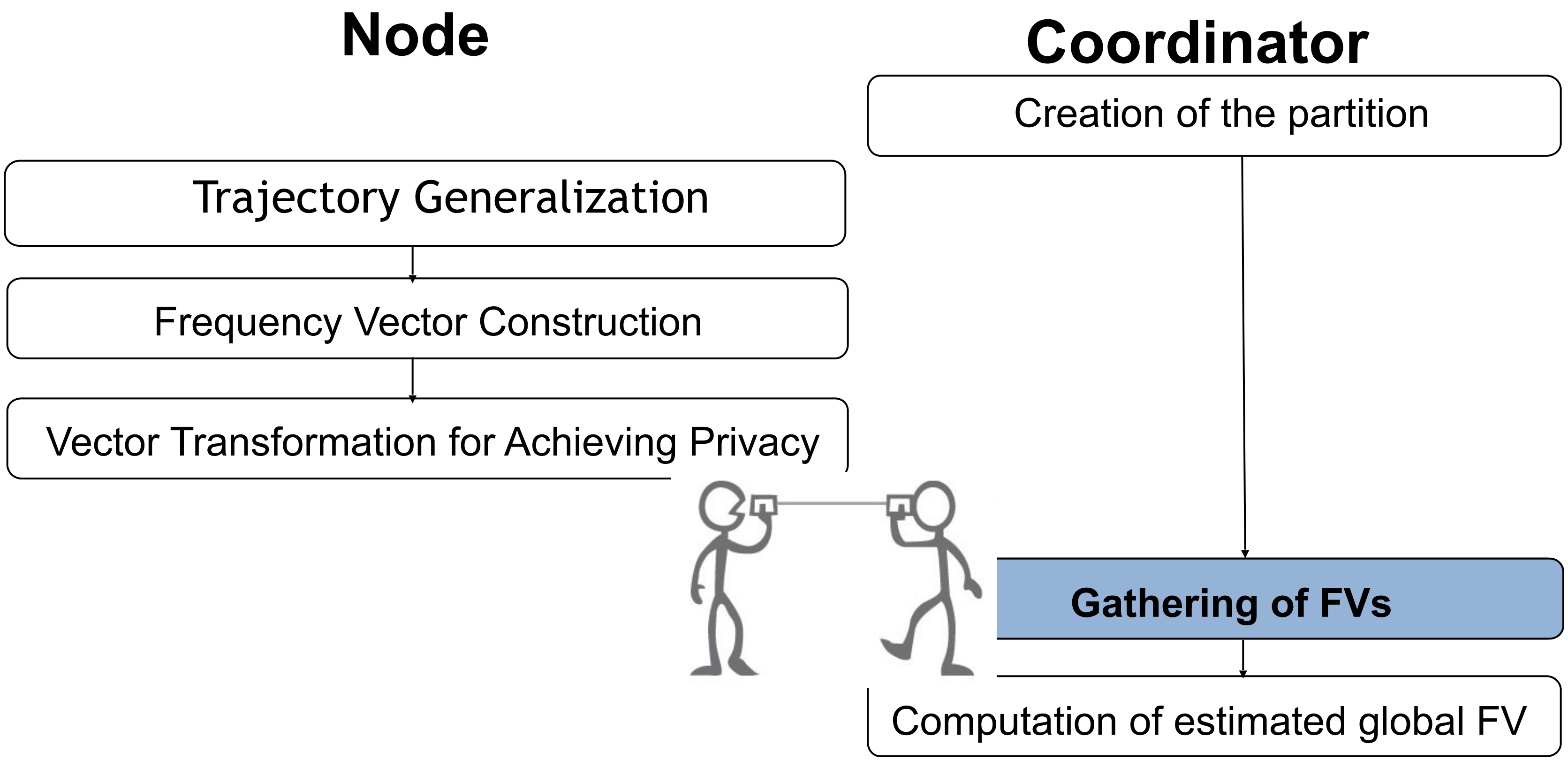
Coordinator

Creation of the partition

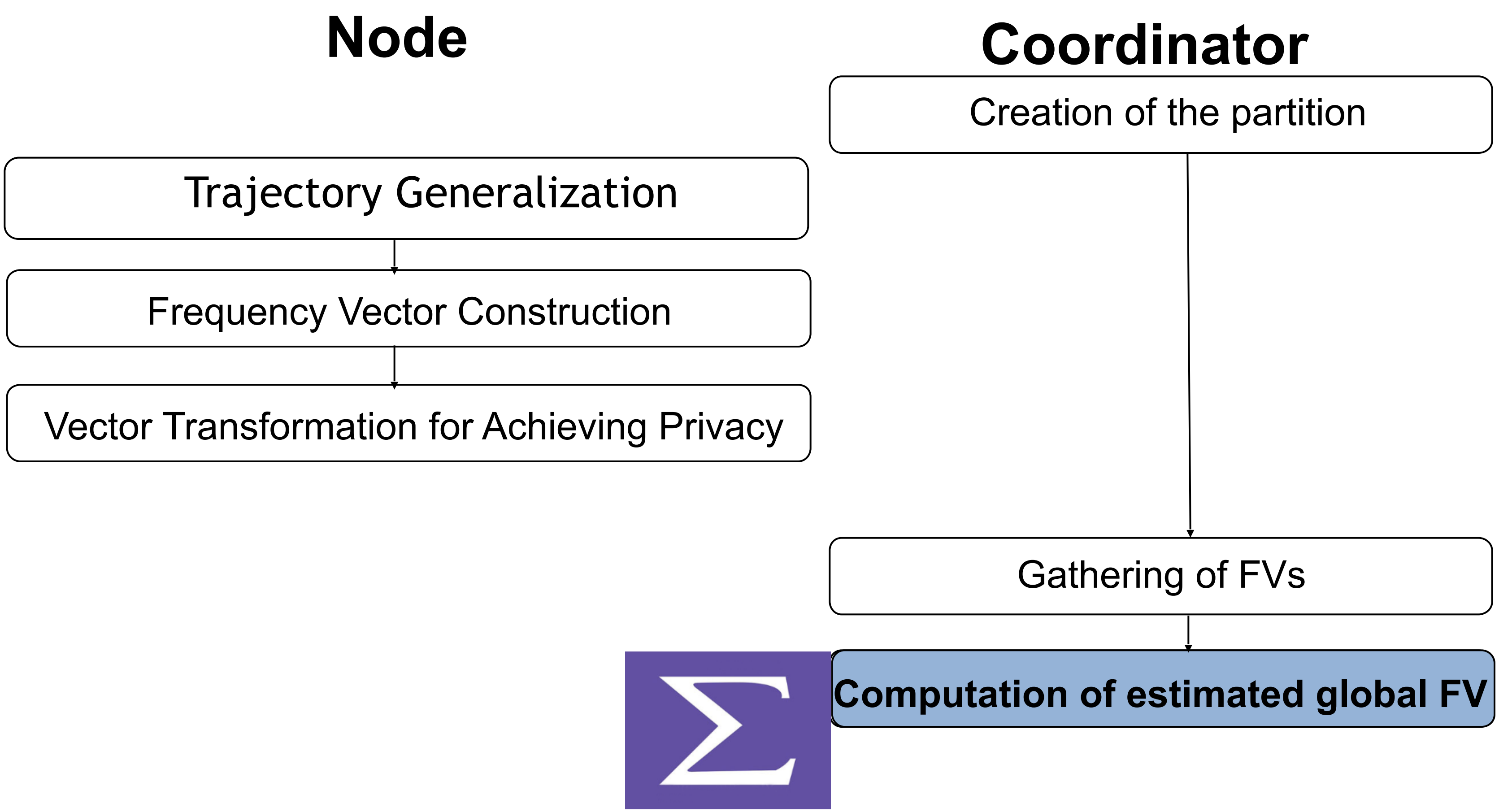
Gathering of FVs

Computation of estimated global FV

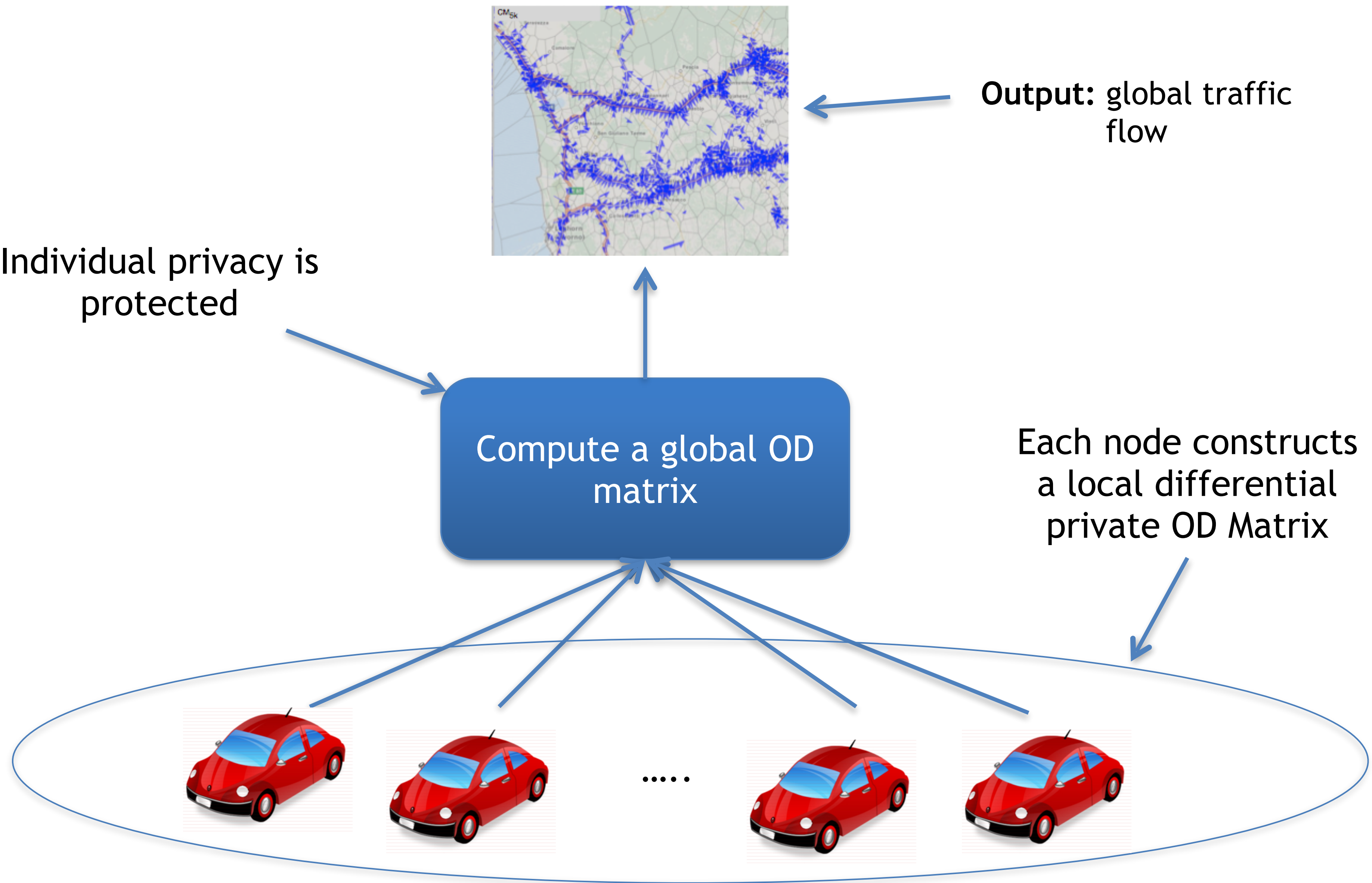
Mobility Analytics and Privacy in User-Centric Ecosystems



Mobility Analytics and Privacy in User-Centric Ecosystems



Privacy-aware Analytical Process



ε-Differential Privacy

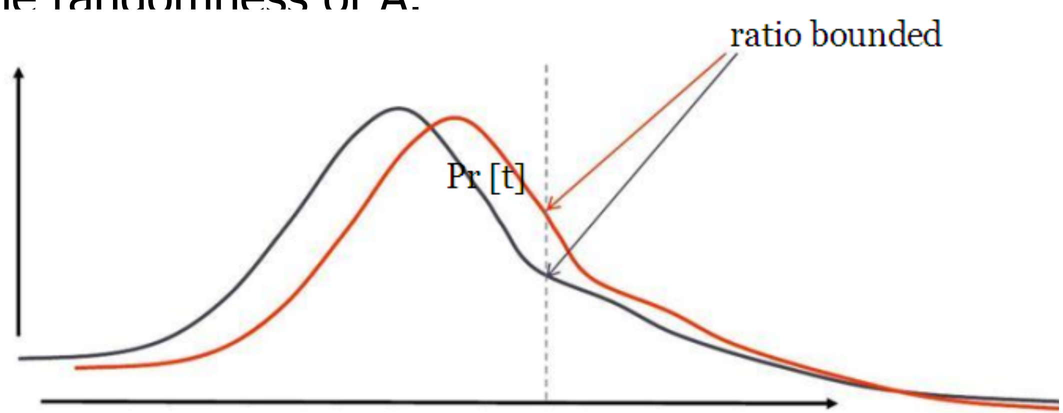
Goal: the ability of an adversary to inflict harm should be essentially the same, independently of whether any individual opts in to, or opts out of, the dataset.

ε-Differential Privacy [Dwork,2006]: A privacy mechanism A gives ε-Differential Privacy if for any dataset D_1 and D_2 differing on at most one record, and for any possible output D' of A we have

$$\Pr[A(D_1) = D'] \leq e^{\varepsilon} \times \Pr[A(D_2) = D']$$

where the probability is taken over the randomness of A.

$$\frac{\Pr[A(D_1) = D']}{\Pr[A(D_2) = D']} \leq e^{\varepsilon}$$

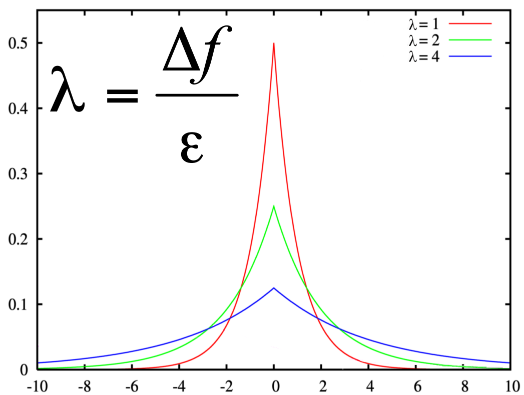


Sensitivity

Sensitivity: for any function $f : D \rightarrow \mathbf{R}^d$, the sensitivity of f is

$$\Delta f = \max_{D_1, D_2} \| f(D_1) - f(D_2) \|_1$$

for all D_1, D_2 differing in at most one record



For our purpose, the sensitivity is move-based: **how much adding or removing a single flow can affect the move frequency?**

- In our case the sensitivity is always =1

Sensitivity - Example

Example: Trajectories in the interval τ:

T1:(a,b)(b,c)(c,e)

T2:(f,g)(g,a)(a,b)(b,c)(c,a)(a,b)

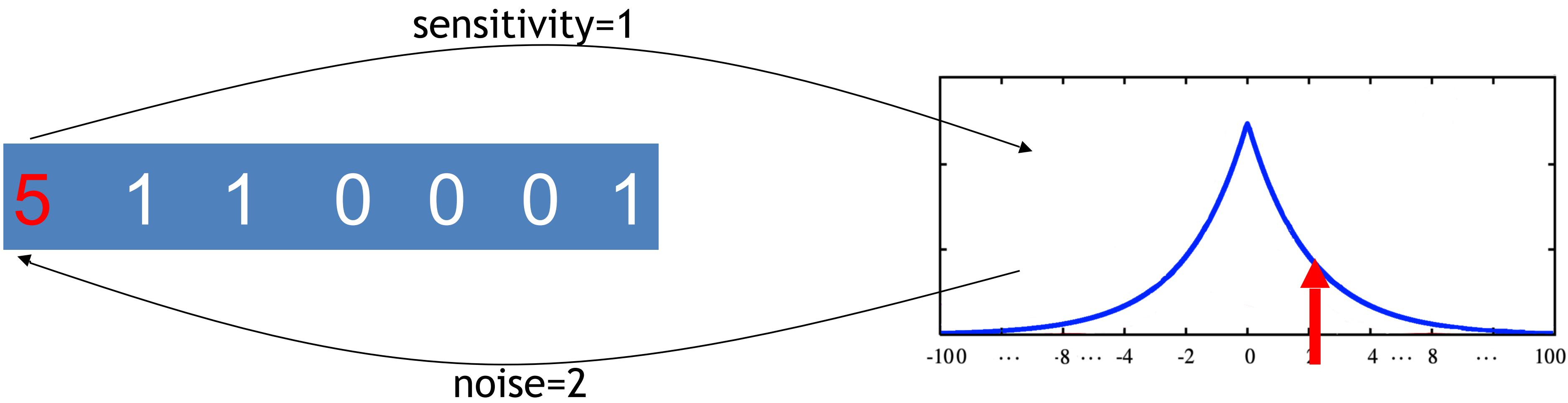
Move-based sensitivity:

D_1: (a,b),(b,c),(c,e),(f,g),(g,a),(a,b),(b,c),(c,a), (a,b)

D_2: (a,b),(b,c),(c,e),(f,g),(g,a),(a,b),(b,c),(c,a)

Sensitivity of the query (a,b) is 1.

ε-Differential Privacy

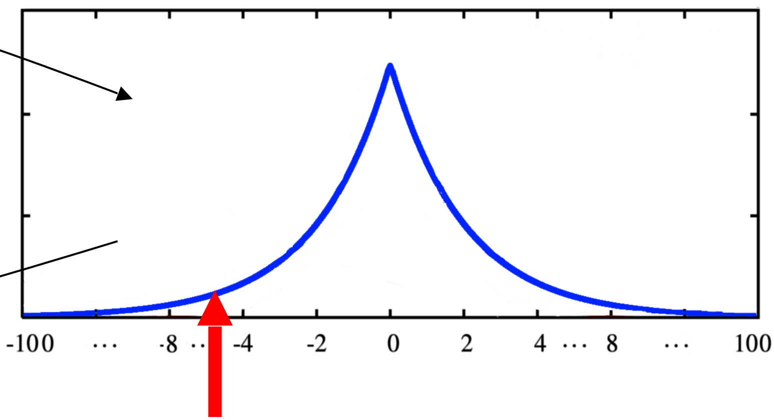


ε-Differential Privacy

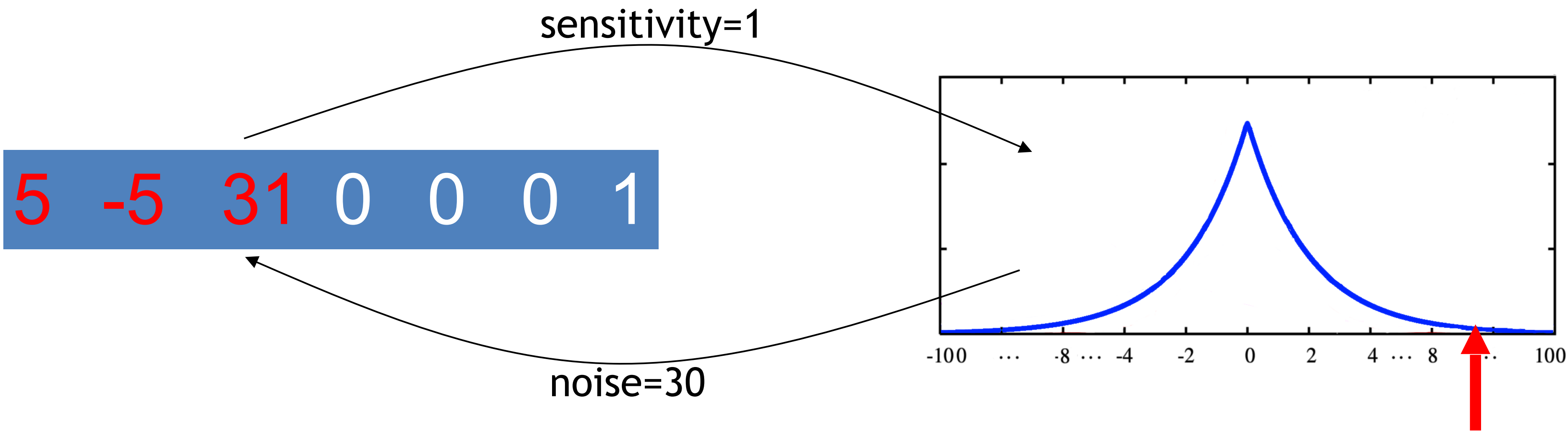
sensitivity=1

5	-5	1	0	0	0	1

noise=-6



ε-Differential Privacy



Problems:

- very big flows
- negative flows

(ε,δ)–Differential Privacy

(ε,δ)**-Differential Privacy:** A privacy mechanism A gives (ε,δ)-Differential Privacy if for any dataset D_1 and D_2 differing on at most one record, and for any possible output D′ of A we have

$$\Pr[A(D_1) = D'] \le e^{\varepsilon} \times \Pr[A(D_2) = D'] + \delta$$

where the probability is taken over the randomness of A.

δ **describes a specific privacy loss.**

(ε,δ)-Differential Privacy for avoiding negative flows

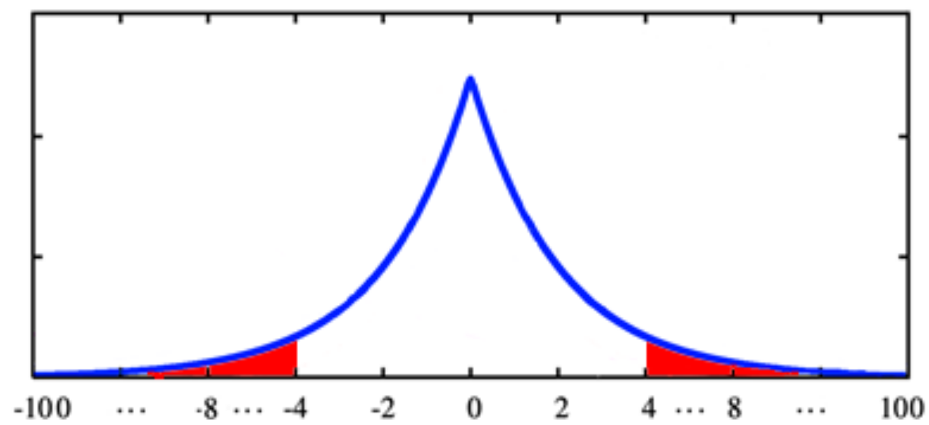
Bounding noise value to the interval [−m,m] where m is the value of the move count

- No too much noise and no negative flows
- Privacy leaks measured by δ → (ε,δ)-differential privacy
- δ depends on m

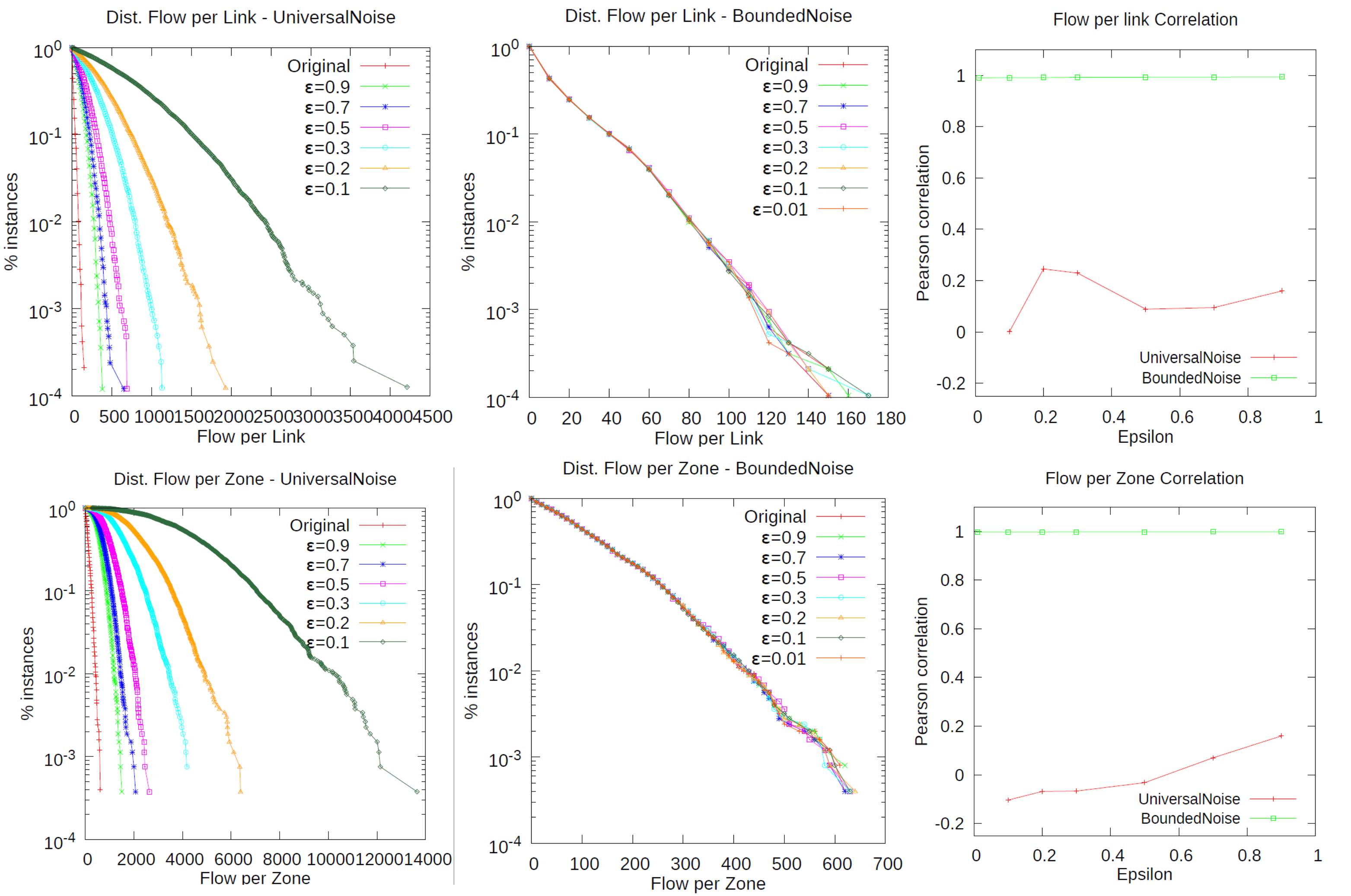
3 1 1 0 0 0 1

↓

6 0 2 0 0 0 1



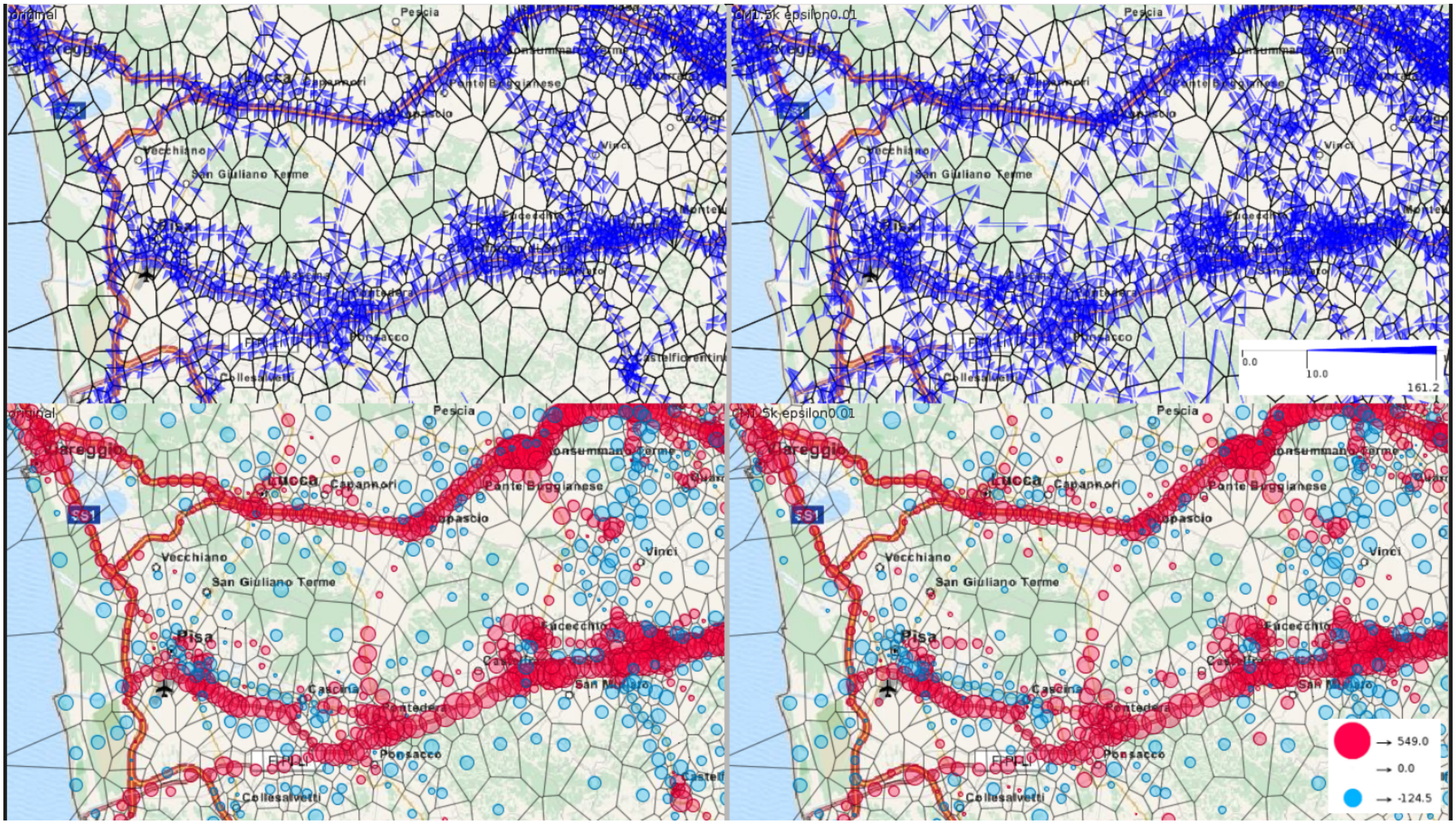
Quality of Network Measures



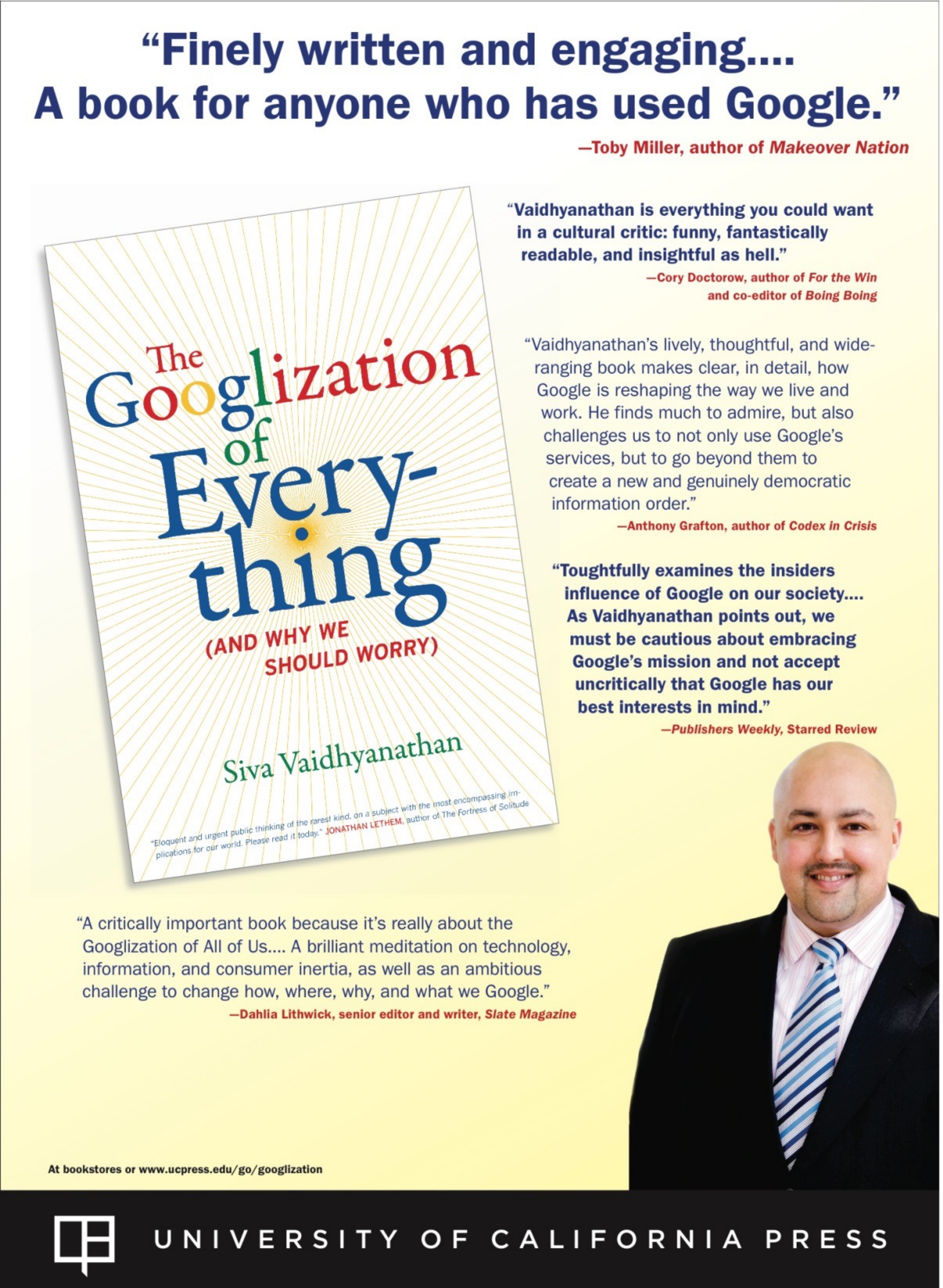
Mobility Analysis

Original Values

BoundedNoise (ε=0.01)



New Challenges in Big Data Era



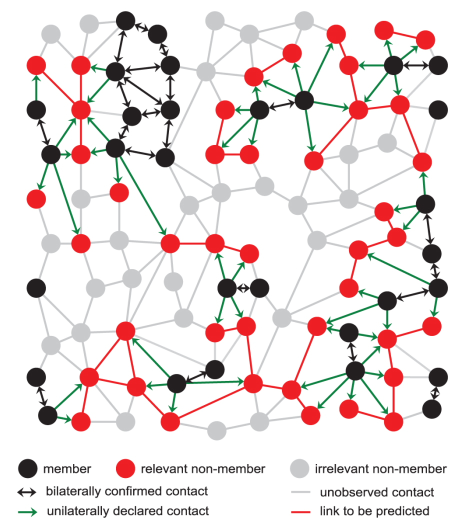
We are not Google's customers,
we are its products.

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

What do people know about me even if I don't have a Facebook profile?

- The danger to one's privacy posed by Facebook even if the person itself is not a Facebook user.

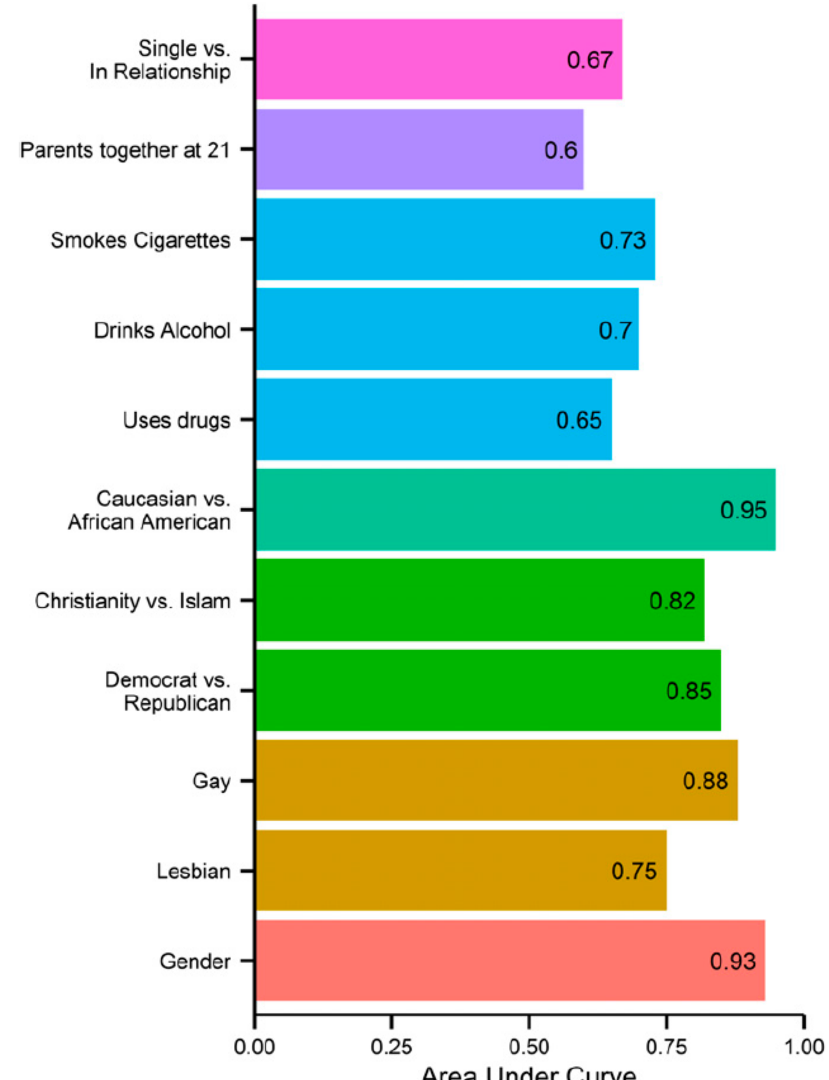
- By analyzing the profiles and communication of the friends of a set of persons without a Facebook account, the authors were able **to infer the relationship established between the "offline" persons.**



Horvát E-Á, Hanselmann M, Hamprecht FA, Zweig KA (2012) One Plus One Makes Three (for Social Networks). PLoS ONE 7(4): e34740. doi:10.1371/journal.pone.0034740

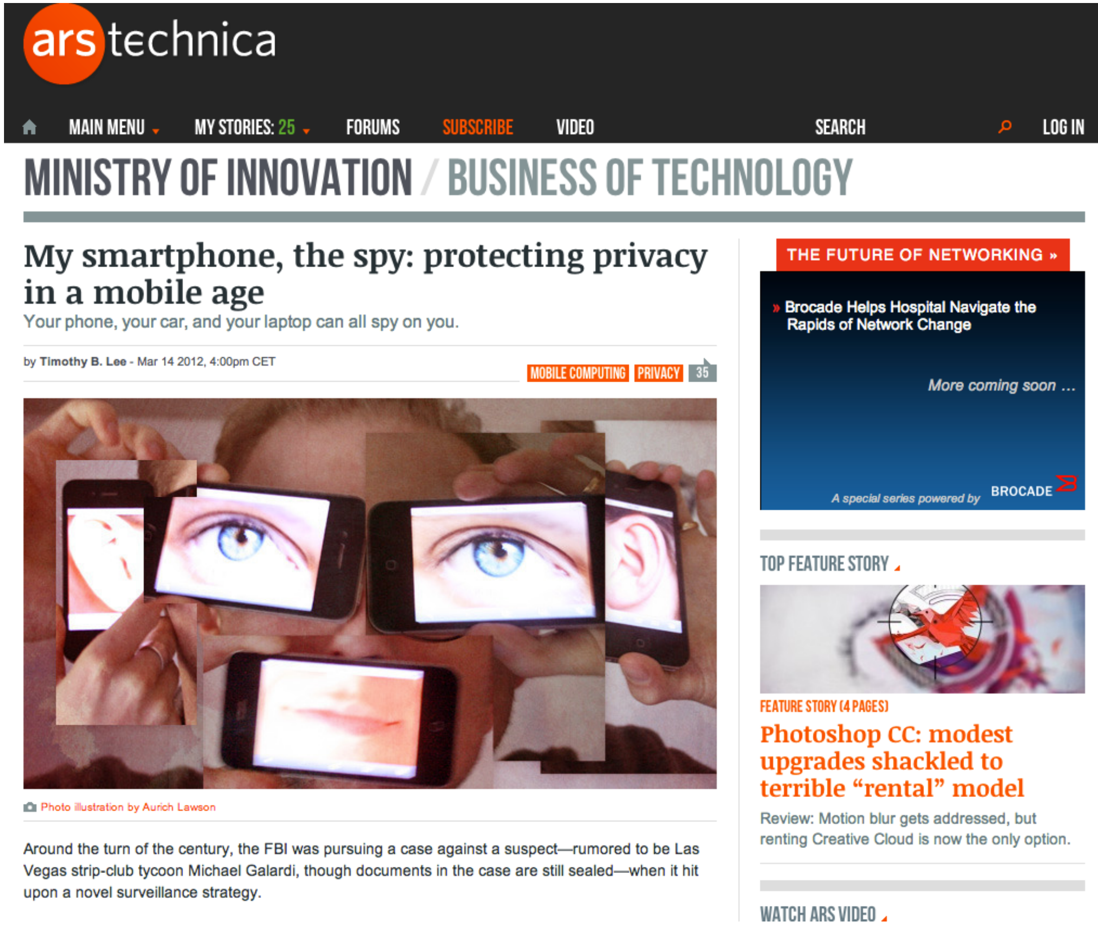
Private traits and attributes are predictable from digital records of human behavior

- Kosinski, Stillwell, Graepel
  - PNAS, March 2013
- «likes» in Facebook enable the inference of user sentitive data
- Web search data, web browsing histories, credit card records are very similar …



My smartphone, the spy: protecting privacy in a mobile age

- Your phone, your car, and your laptop can all spy on you
- Short essay about the capabilities of smartphones to be converted into spying devices at will, from the mother company.
- It opens with the report about an old case in which the FBI asked a company to turn on the microphone of the suspect's cellphone.



ars technica

MAIN MENU MY STORIES: 25 FORUMS SUBSCRIBE VIDEO SEARCH LOG IN

MINISTRY OF INNOVATION / BUSINESS OF TECHNOLOGY

My smartphone, the spy: protecting privacy in a mobile age

Your phone, your car, and your laptop can all spy on you.

by Timothy B. Lee - Mar 14 2012, 4:00pm CET

MOBILE COMPUTING PRIVACY 35

Photo illustration by Aurich Lawson

Around the turn of the century, the FBI was pursuing a case against a suspect—rumored to be Las Vegas strip-club tycoon Michael Galardi, though documents in the case are still sealed—when it hit upon a novel surveillance strategy.

THE FUTURE OF NETWORKING »

» Brocade Helps Hospital Navigate the Rapids of Network Change

More coming soon ...

A special series powered by BROCADE

TOP FEATURE STORY

FEATURE STORY (4 PAGES)

Photoshop CC: modest upgrades shackled to terrible "rental" model

Review: Motion blur gets addressed, but renting Creative Cloud is now the only option.

WATCH ARS VIDEO