ETICHIS & PRIVACY

Anna Monreale
Università di Pisa

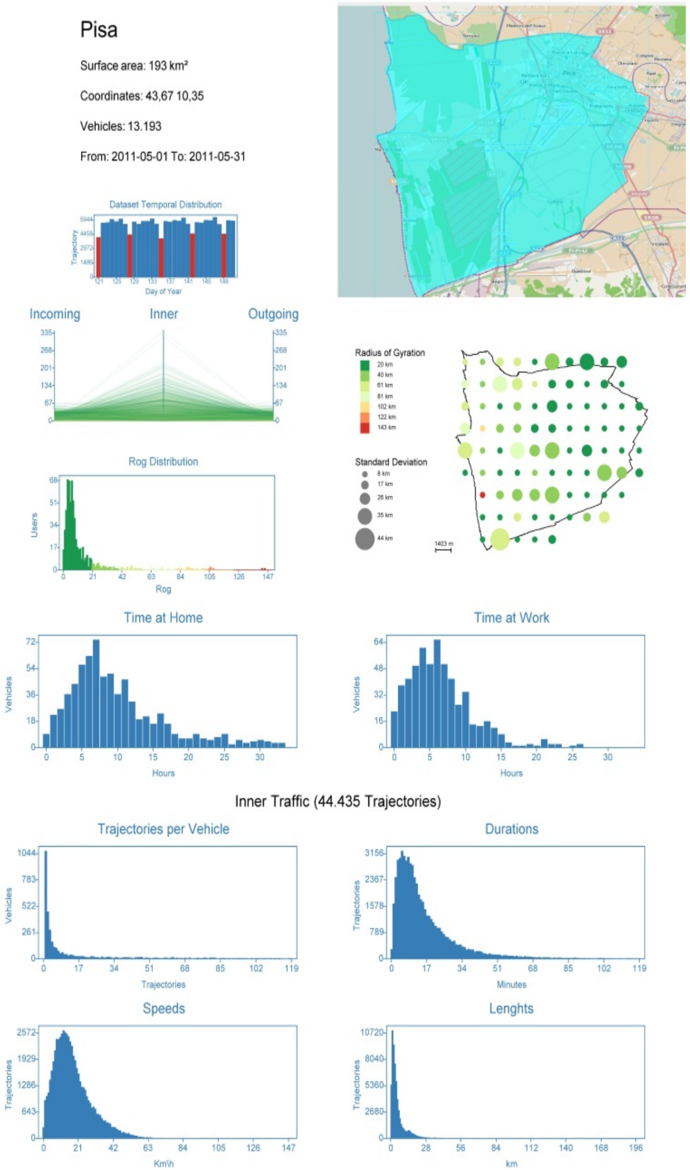
Knowledge Discovery and Delivery Lab
(ISTI-CNR & Univ. Pisa)
www-kdd.isti.cnr.it

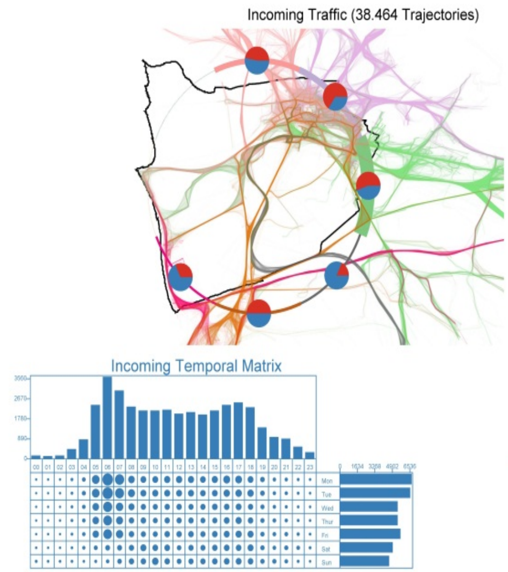
Our digital traces ….

- 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?

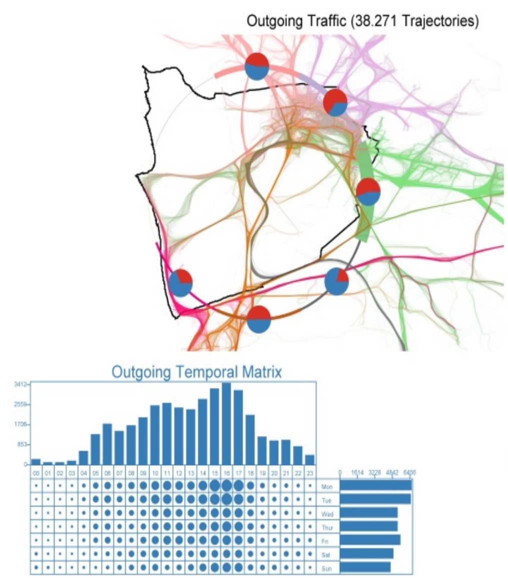
Big Data: new, more carefully targeted financial services

Mobility atlas of many cities



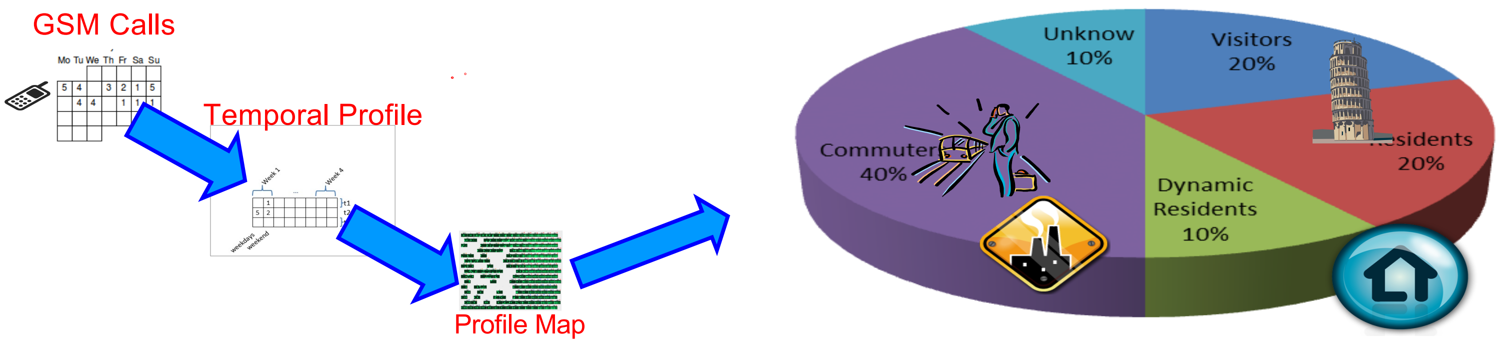


	City	Traj	Perc
NORD 32%	San Giuliano T..	4.816	62%
	Vecchiano	1.425	94%
	Viareggio	1.142	99%
	Lucca	862	67%
	Camaiore	358	94%
OVEST 0%			
SUD 12%	Livorno	2.843	92%
	Collesalvetti	565	50%
	Rosignano Mari..	140	41%
	Fauglia	137	19%
	Cecina	124	45%
EST 54%	Cascina	7.078	97%
	San Giuliano T..	2.881	37%
	Pontedera	1.350	95%
	Calci	795	79%
	Calcinaia	693	92%



	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%

A Sociometer based on Mobile Phone Data for Real Time Demographics



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

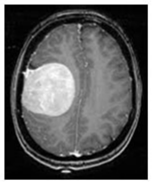
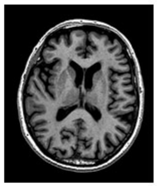
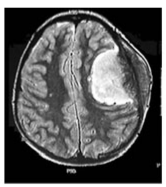
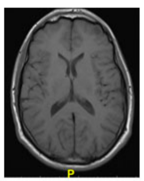
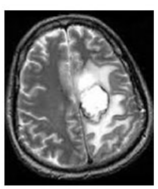
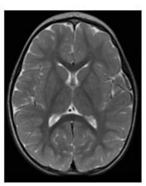
KDD Lab

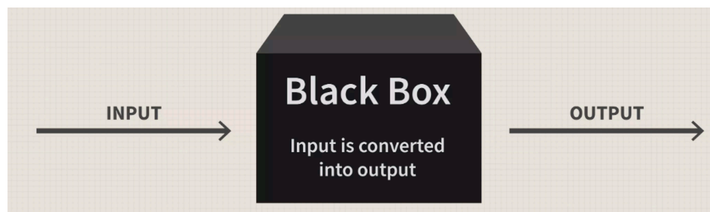
UNIVERSITÀ DI PISA

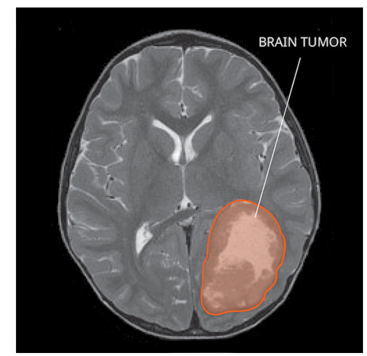
Istat

WIND
Più vicini.

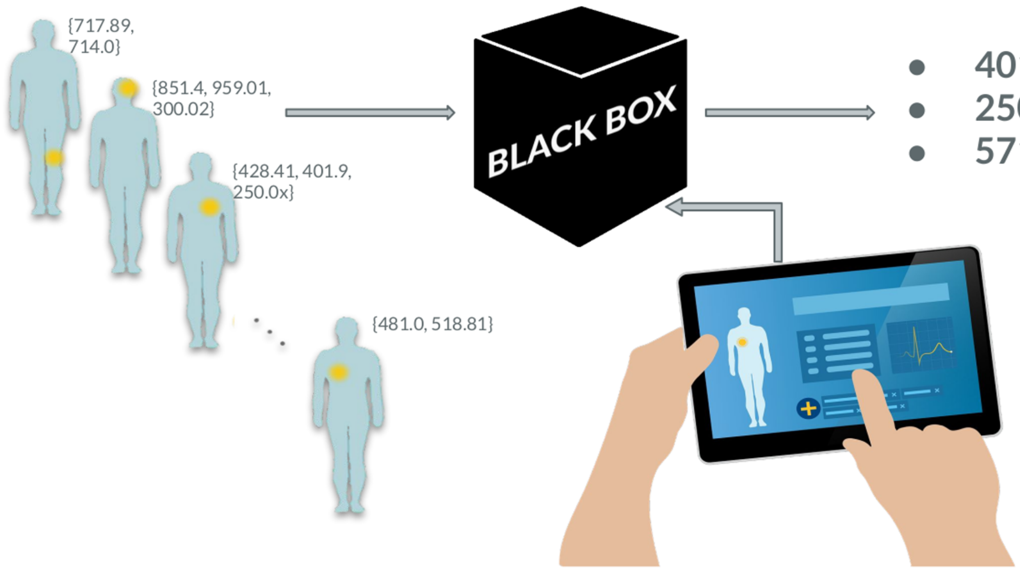
AI in healthcare

Brain Tumor Image	Brain Non Tumor Image
	
	
	



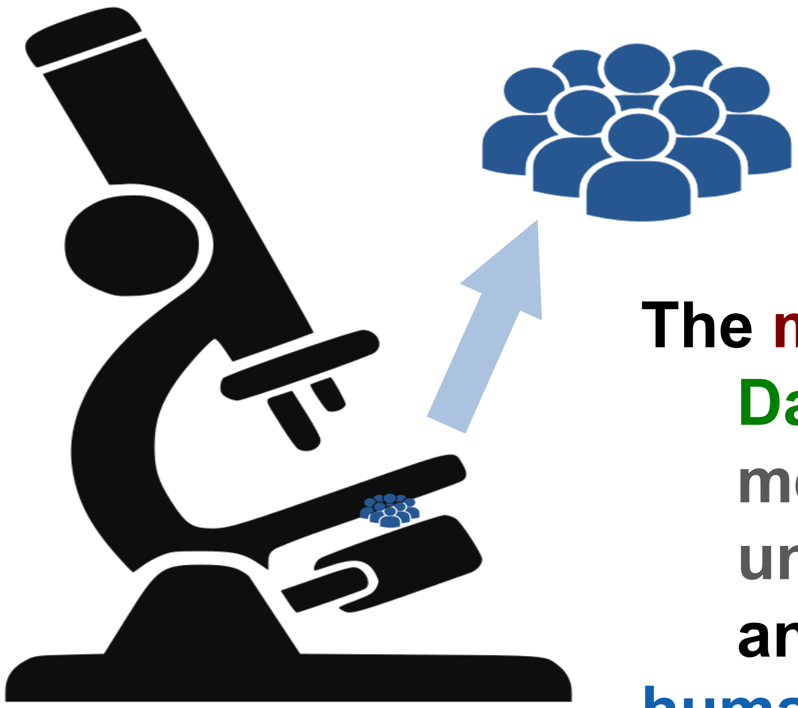


AI in healthcare



- **401.0**, Hypertension
- **250.0x**, Diabetes
- **571.8**, Nonalcoholic liver disease

AI, Big Data Analytics & Social Mining



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

Artificial Intelligence: what is it now?

From **encoding** intelligent behavior

↓

To **discovery** and **capture**
intelligent behavior from **data**

Especially (but not only) **personal data**

Artificial Intelligence
=
Collective Intelligence!!

- **Learning from many examples**

- Provide **support for decision making**
  - Enabling nowcasting, what-if simulations based on big data analytics & modeling

Learning from experience

- Data mining & machine learning + big data are the **fulcrum of AI**
- Big data = record the (human) experience
- IoT will facilitate this trend

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

EU Ethics Guidelines for AI – (2019)

Human-centric approach: AI as a means, not an end

Trustworthy AI as our foundational ambition, with three components

Lawful AI	complying with all applicable laws and regulations
Ethical AI	ensuring adherence to ethical principles and values
Robust AI	perform in a **safe, secure** and **reliable** manner, both form technical and a social perspective, with safeguards to foresee and prevent unintentional harm

Requirements

1. **Human agency and oversight**
   - Fundamental rights
   - Human agency
   - Human oversight

2. **Technical robustness**
   - Resilience to attack and security
   - Safety
   - Accuracy
   - Reliability and reproducibility

3. **Privacy and data governance**
   - Privacy and data protection
   - Quality and integrity of data
   - Access to data

4. **Transparency**
   - Traceability
   - Explainability

Requirements

5. Diversity, non-discrimination and fairness

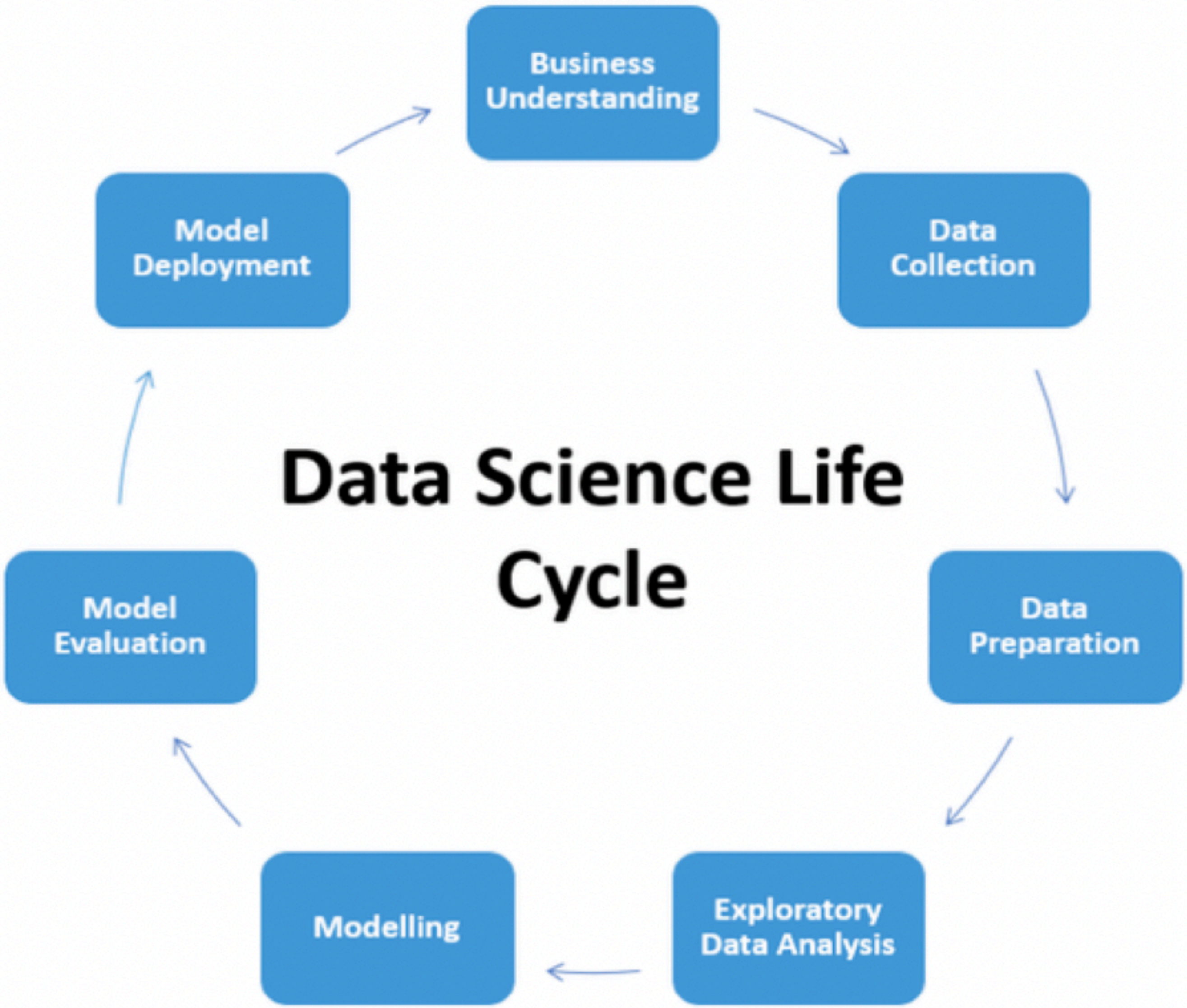
- Avoidance of unfair bias
- Accessibility and universal design
- Stakeholder Participation

6. Societal and environmental well-being

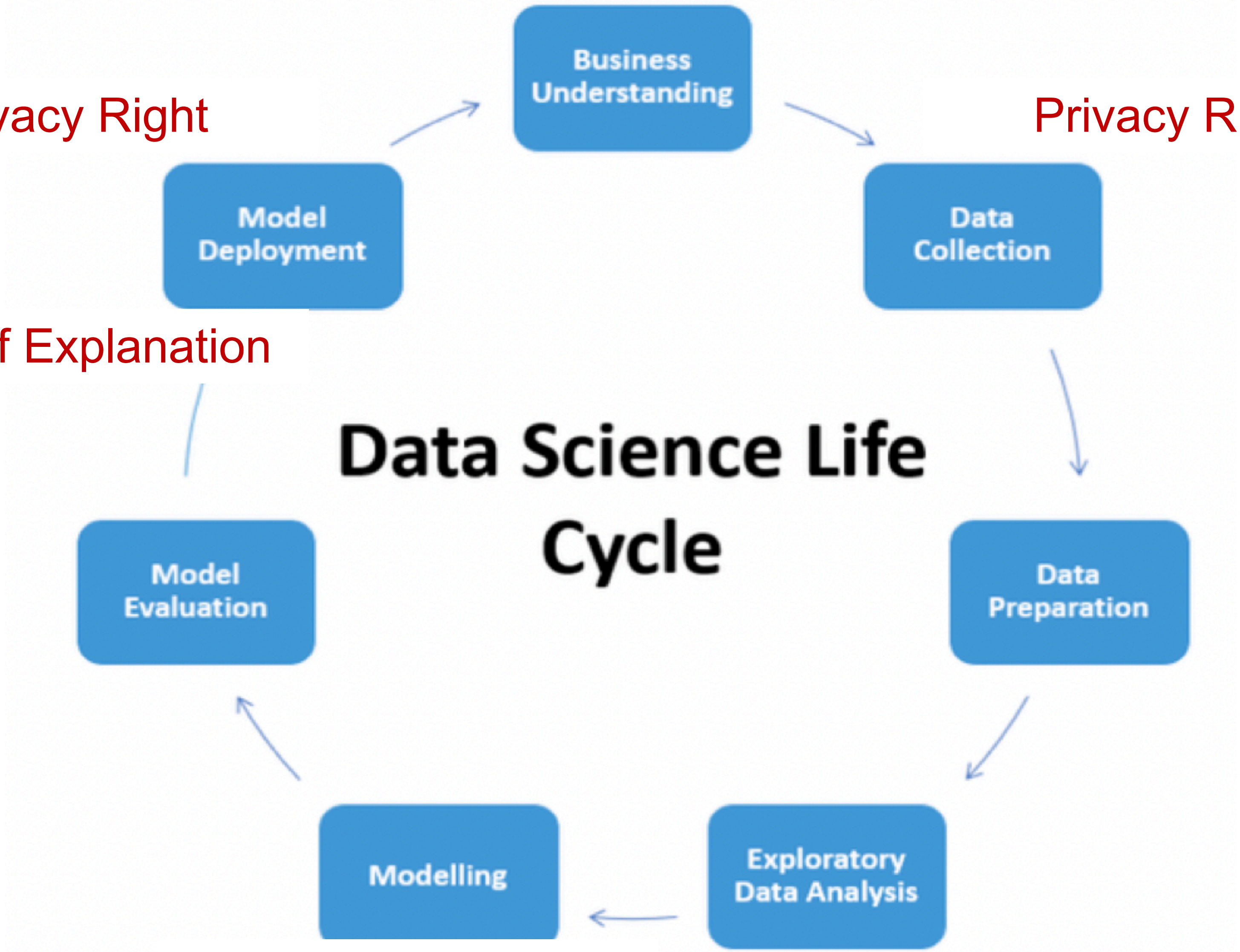
- Sustainable and environmentally friendly AI
- Social impact
- Society and Democracy

7. Accountability

- Minimisation and reporting of negative impacts
- Auditability
- Minimisation and reporting of negative impacts
- Trade-offs


Business Understanding
Data Collection
Data Preparation
Exploratory Data Analysis
Modelling
Model Evaluation
Model Deployment
Data Science Life Cycle


Privacy Right
Right of Explanation



Privacy Right

Privacy Right

Right of Explanation

Privacy Right

Right of Explanation

PRIVACY & DATA PROTECTION

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)
  - General Data Protection Regulation (May 2018)

http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=IT

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.

Personal Data

- Your name
- Home address
- Photo
- Email address
- Bank details
- Posts on social networking websites
- Medical information,
- Computer or mobile IP address
- Mobility traces
- ………

Sensitive Data

- Sensitive personal data is a specific set of "**special categories**" that must be treated with extra security
  - Racial or ethnic origin
  - Political opinions
  - Religious or philosophical beliefs
  - Trade union membership
  - Genetic data
  - Biometric data

EU Directive (95/46/EC) and GDPR

- **GOALS**:
  - protection protection of individuals with regard to the **processing** of personal data
  - the free movement of such data
  - User control on personal data
- The term “process” covers anything that is done to or with personal data:
  - collecting
  - recording
  - organizing, structuring, storing
  - adapting, altering, retrieving, consulting, using
  - disclosing by transmission, disseminating or making available, aligning or combining, restricting, erasing, or destroying data.

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**

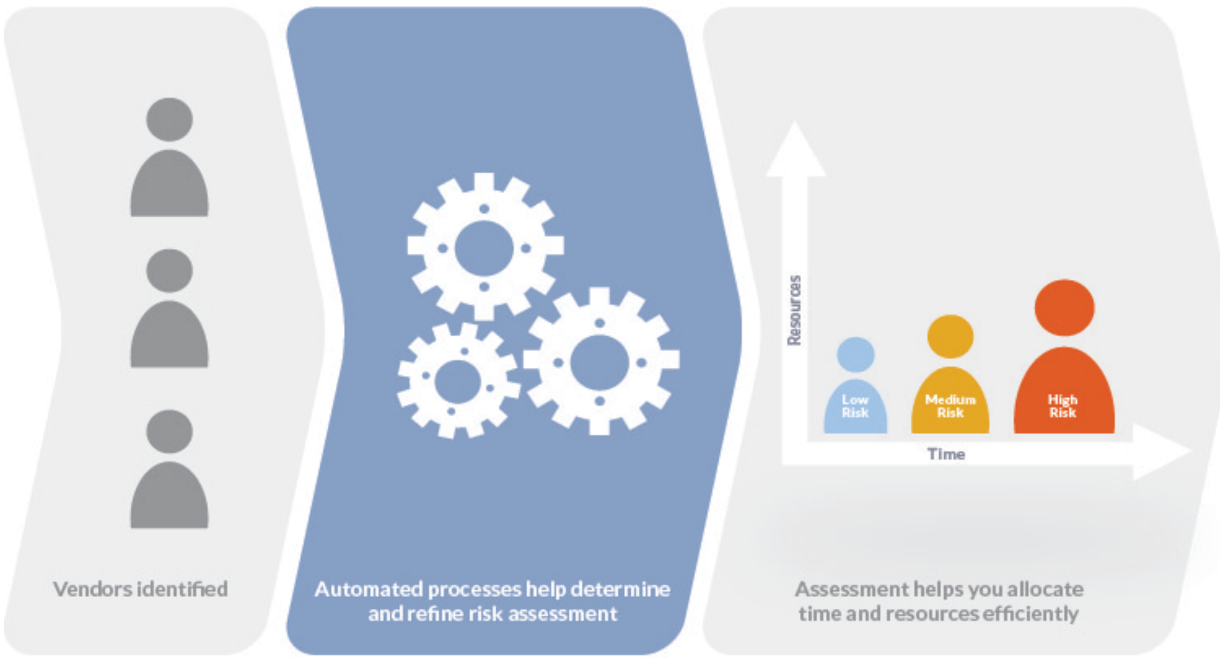
Privacy by Design Principle

- **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 Risk Assessment

- GDPR requires that data controllers maintain an updated report on the privacy risk assessment on perosnal data collected



PSEUDONYMIZATION & ANONYMIZATION

Anonymization vs Pseudonimization

- Pseudonymization and Anonymization are two distinct terms often confused

- Anonymized data and pseudonymized data fall under very different categories in the regulation

- **Anonymization guarantees data protection** against the (direct and indirect) data subject re-identification

- **Pseudonymization substitutes the identity** of the data subject in such a way that additional information is required to re-identify the data subject

Pseudonymization

Substitute an **identifier** with a surrogate value called **token**

Substitute unique names, fiscal code or any attribute that identifies uniquely individuals in the data

Example of Pseudonymization

Name	Gender	DoB	ZIP Code	Diagnosis
Anna Verdi	F	1962	300122	Cancro
Luisa Rossi	F	1960	300133	Gastrite
Giorgio Giallo	M	1950	300111	Infarto
Luca Nero	M	1955	300112	Emicrania
Elisa Bianchi	F	1965	300200	Lussazione
Enrico Rosa	M	1953	300115	Frattura

ID	Gender	DoB	ZIP CODE	DIAGNOSIS
11779	F	1962	300122	Cancro
12121	F	1960	300133	Gastrite
21177	M	1950	300111	Infarto
41898	M	1955	300112	Emicrania
56789	F	1965	300200	Lussazione
65656	M	1953	300115	Frattura

Properties of a Surrogate Value

- Irreversible without private information
- Distinguishable from the original value

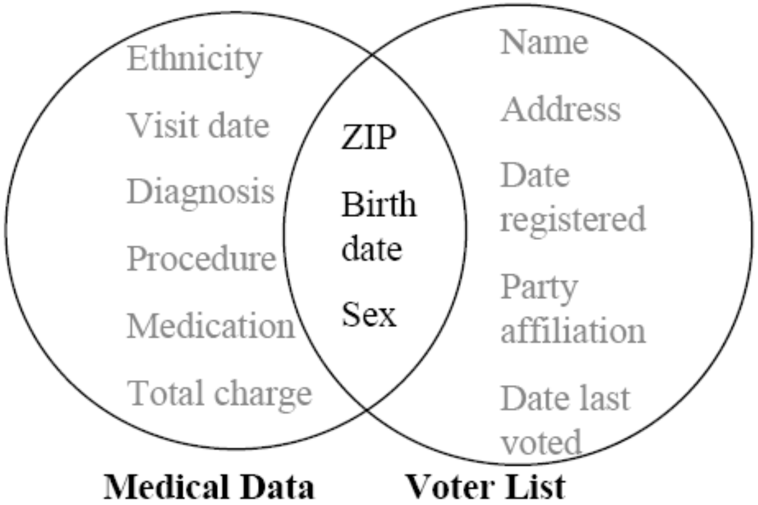
Is Pseudonymization enough for data protection?

Pseudonymized data are still Personal Data!!

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)

Linking Attack

Governor: **birth date = 1950,** CAP = **300111**

ID	Gender	YoB	ZIP	DIAGNOSIS
1	F	1962	300122	Cancer
3	F	1960	300133	Gastritis
2	M	1950	300111	Heart Attack
4	M	1955	300112	Headache
5	F	1965	300200	Dislocation
6	M	1953	300115	Fracture

Which is the disease of the Governor?

Making data anonymous

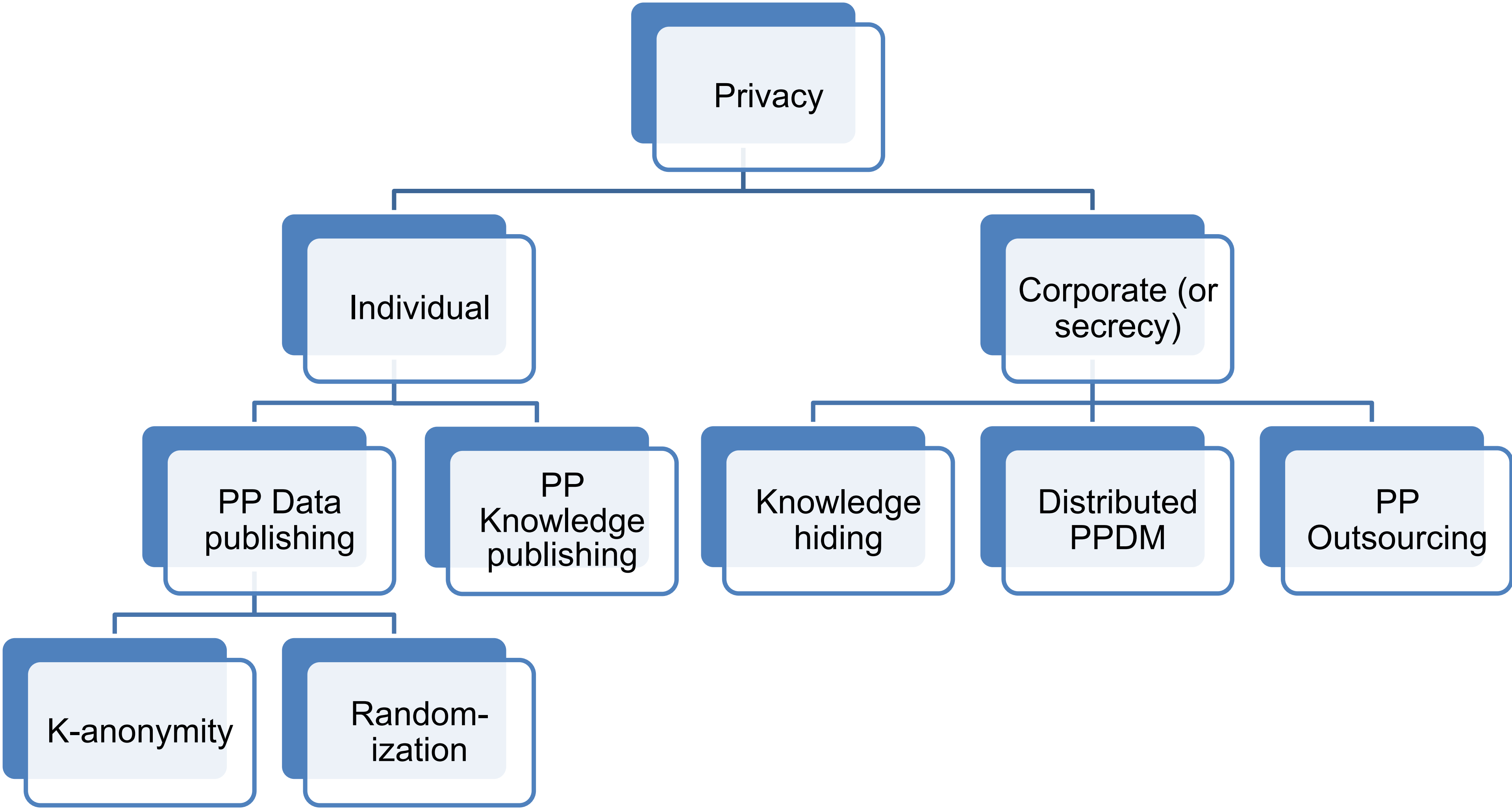
K-anonymity

Governor: Birth Date **= 1950,** CAP = **300111**

ID	Gender	YoB	ZIP	DIAGNOSIS
1	F	[1960-1956]	300***	Cancer
3	F	[1960-1956]	300***	Gastritis
2	M	[1950-1955]	30011*	Heart Attack
4	M	[1950-1955]	30011*	Headache
5	F	[1960-1956]	300***	Dislocation
6	M	[1950-1955]	30011*	Fracture

Which is the disease of the Governor?

Ontology of Privacy in Data Mining



Attribute classification

Identifiers | Quasi-identifiers | Sensitive

ID	Gender	YoB	ZIP	DIAGNOSIS
1	F	1962	300122	Cancer
3	F	1960	300133	Gastritis
2	M	1950	300111	Heart Attack
4	M	1955	300112	Headache
5	F	1965	300200	Dislocation
6	M	1953	300115	Fracture

K-Anonymity

- **k-anonymity hides 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

Vulnerability of K-anonymity

ID	Gender	DoB	ZIP	DIAGNOSIS
1	F	1962	300122	Cancer
3	F	1960	300133	Gastritis
2	M	1950	300111	Heart Attack
4	M	1950	300111	Heart Attack
5	M	1950	300111	Heart Attack
6	M	1953	300115	Fracture

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

ID	Gender	DoB	ZIP	DIAGNOSIS
1	F	1962	300122	Heart Attack
3	F	1960	300133	Headache
2	M	1950	300111	Dislocation
4	M	1950	300111	Fracture
5	M	1950	300111	Heart Attack
6	M	1953	300115	Headache

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.*

Randomization

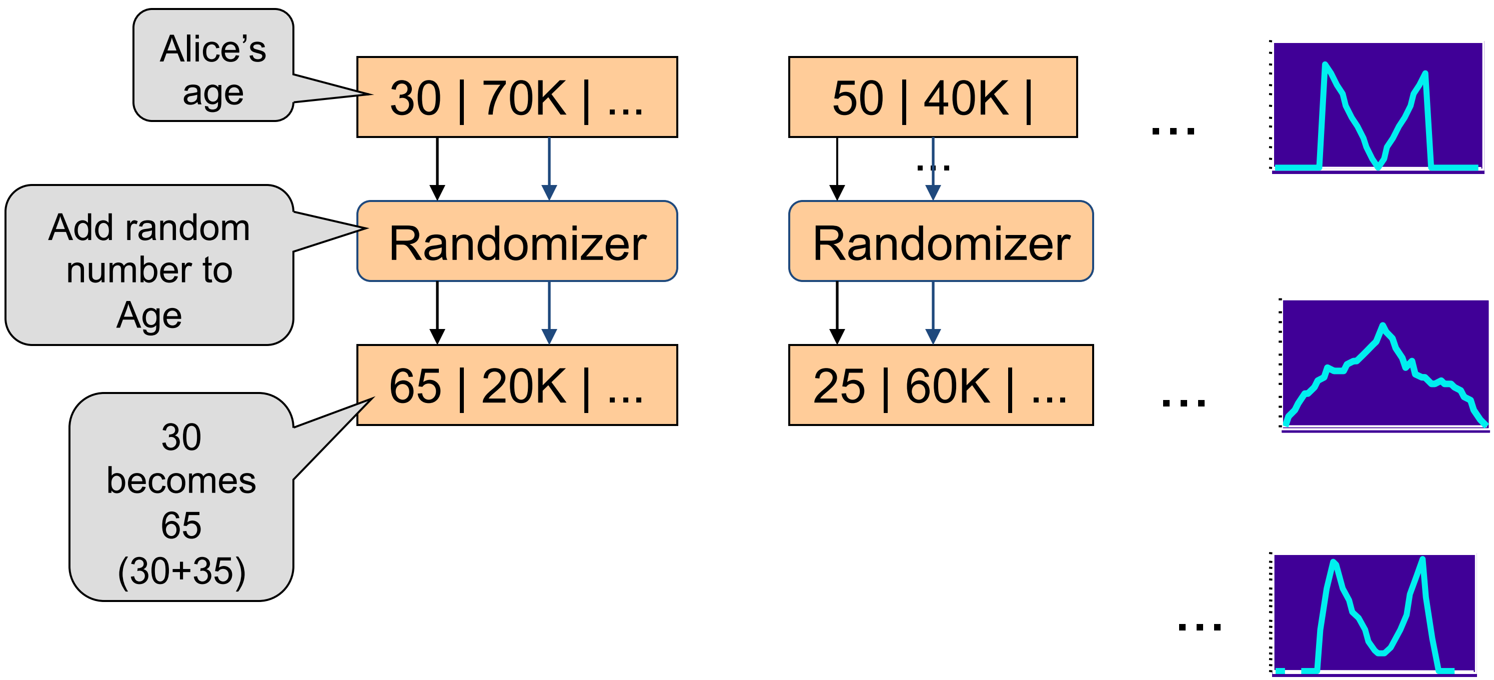
- **Original values $x_1, x_2, ..., x_n$**
  - from probability distribution X (unknown)
- **To hide these values, we use $y_1, y_2, ..., y_n$**
  - from probability distribution Y
    - Uniform distribution between $[-\alpha, \alpha]$
    - Gaussian, normal distribution with $\mu = 0, \sigma$

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

Estimate the probability distribution of X.

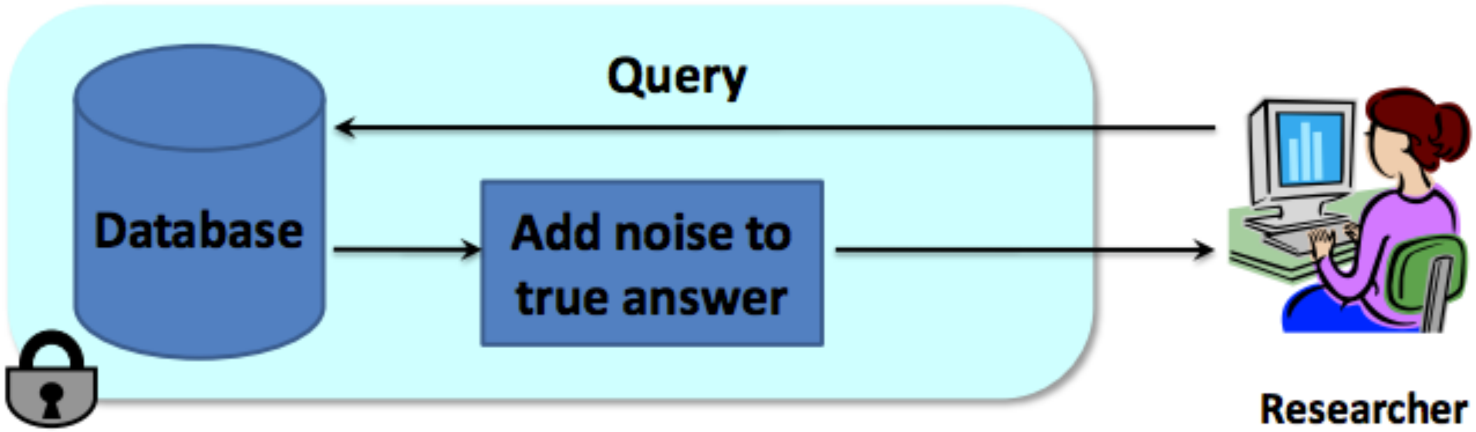
R. Agrawal and R. Srikant. Privacy-preserving data mining. In Proceedings of SIGMOD 2000.

Randomization Approach Overview



Differential Privacy

- **The risk to my privacy should not increase as a result of participating in a statistical database**



- 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

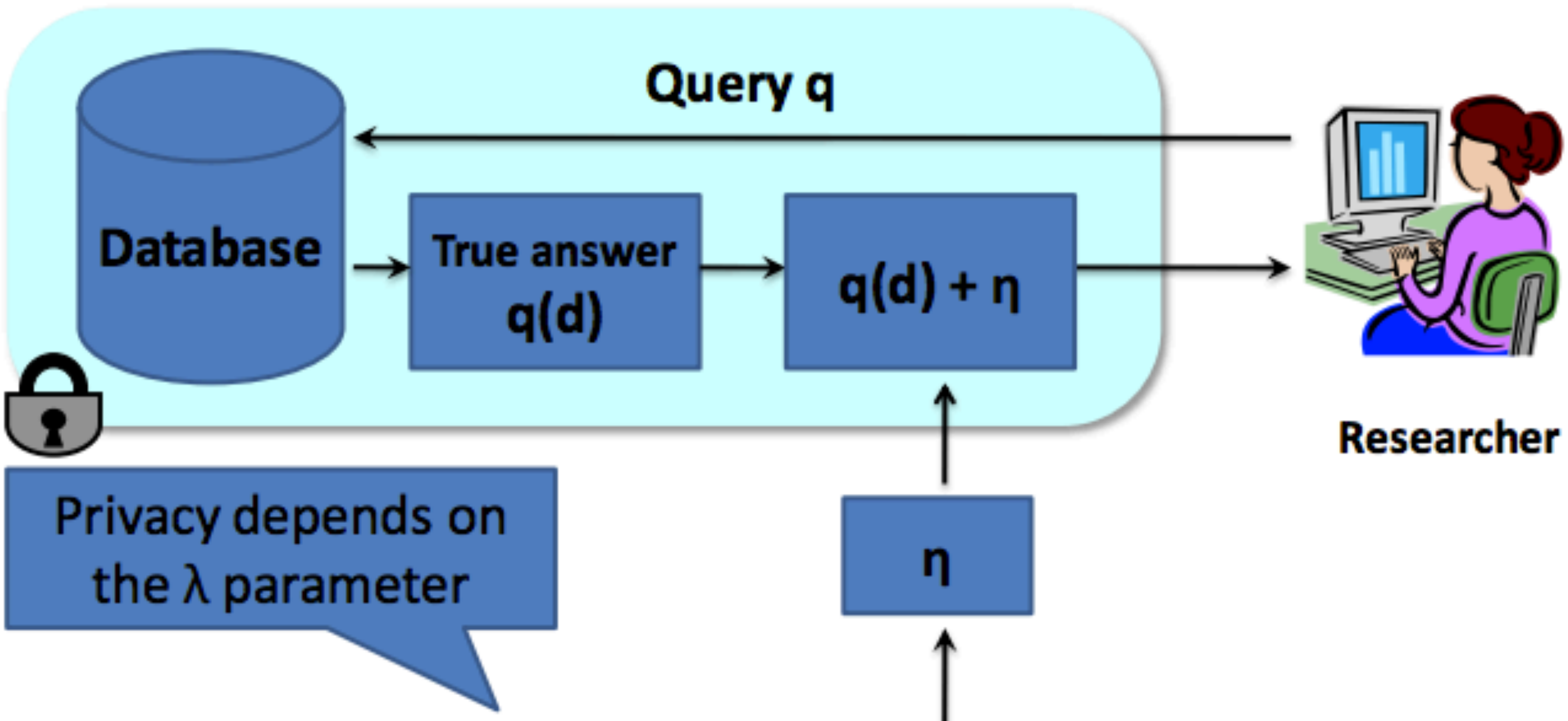
1) how many persons have Diabetes? **4**
2) how many persons, excluding Alice, have Diabetes? **3**

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

- **Solution**: make the two answers similar

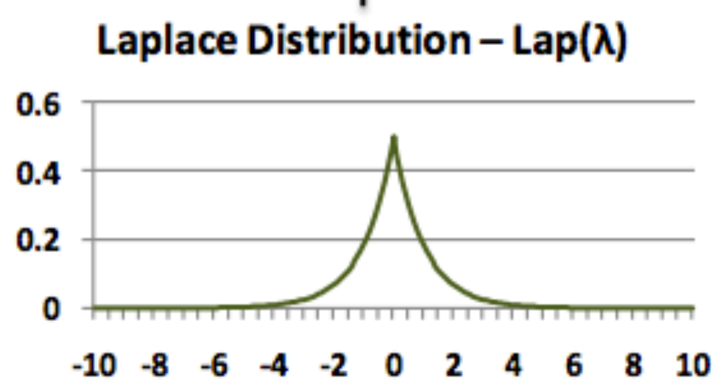
1) the answer of the first query could be 4+1 = 5
2) the answer of the second query could be 3+2.5=5.5

Differential Privacy



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

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



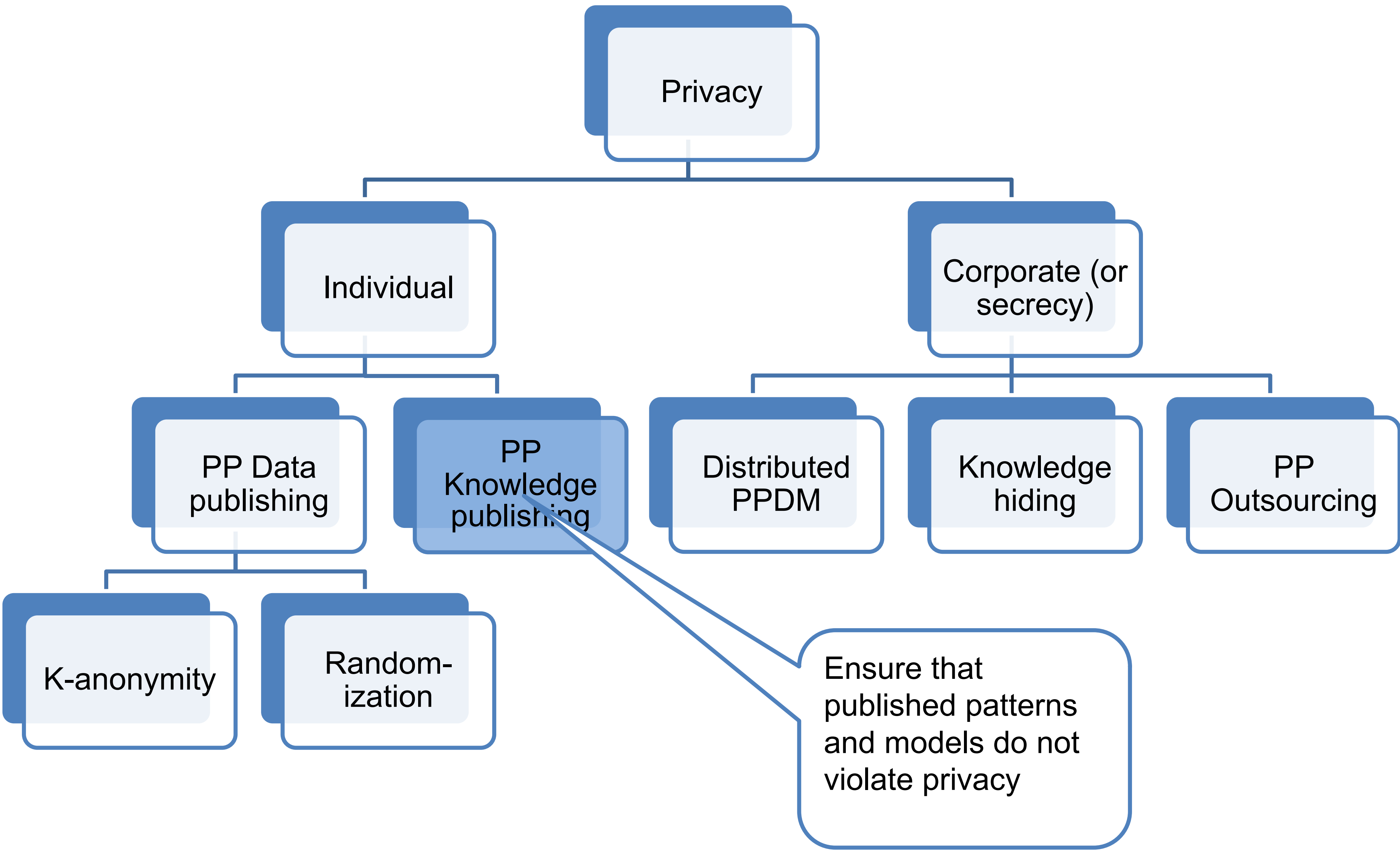
Randomization

- 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 Mining



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, Religion=Christian ⇒ Country=American
(support = 758, confidence = 99.8%)

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

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

Age = 27, Postcode = 45254, Country=not American
(support = 1)

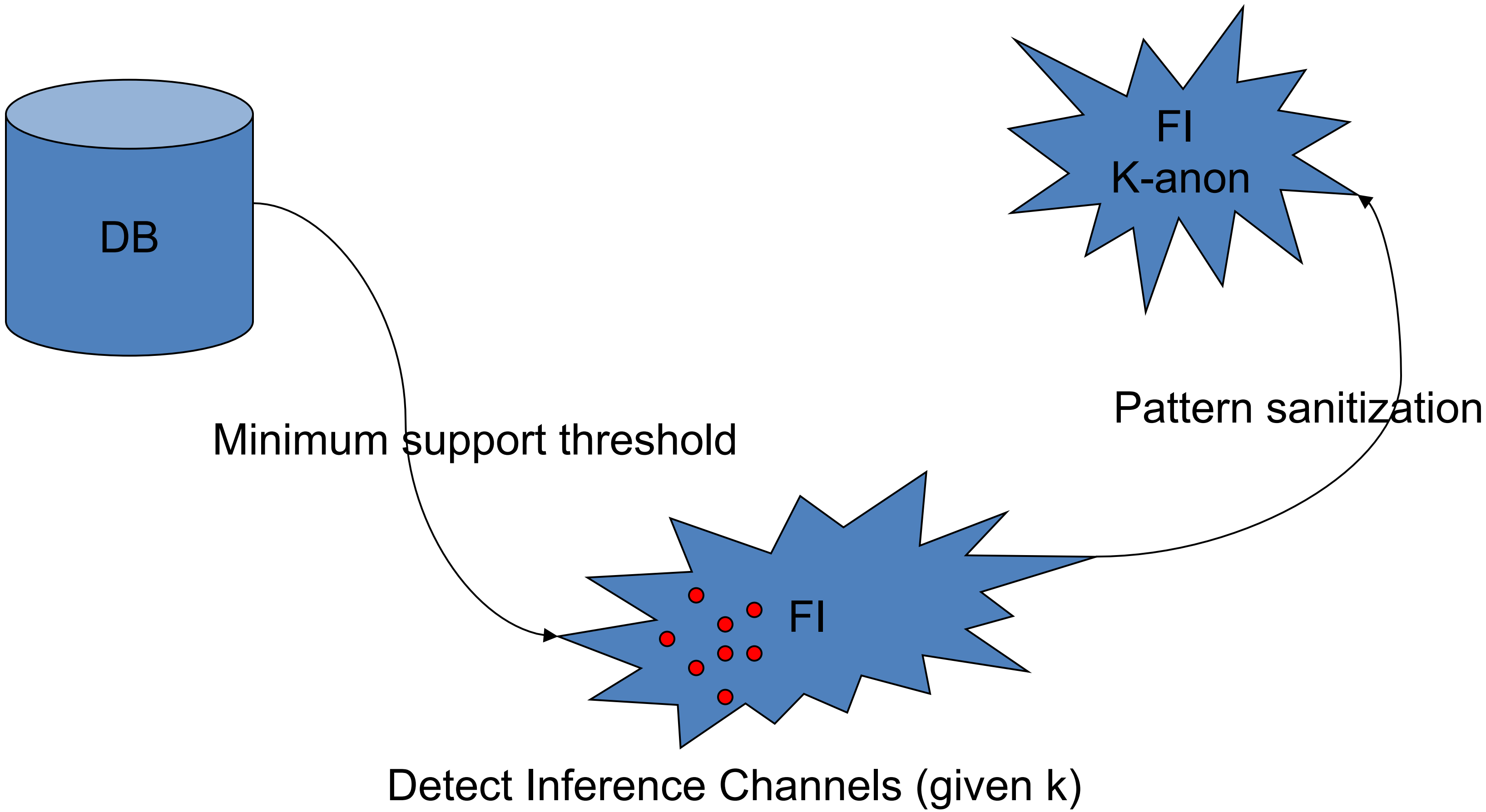
Age = 27, Postcode = 45254, Country=not American, Religion=Christian
(support = 1)

Age = 27, Postcode = 45254, Country=not American ⇒ Religion=Christian
(support = 1, confidence=1100%)

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.

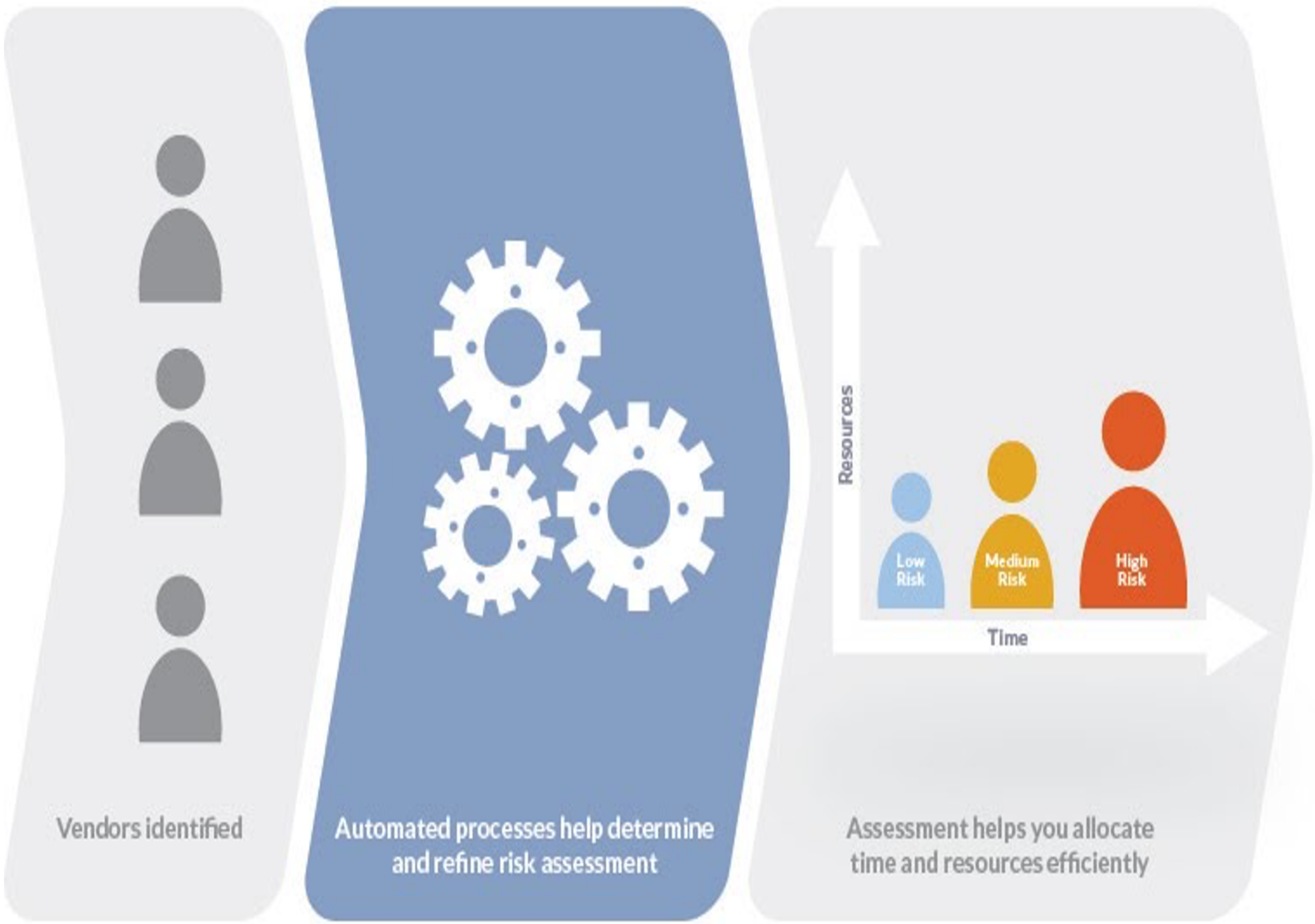
New Regulation

- Privacy by Design
- Privacy Risk Assessment

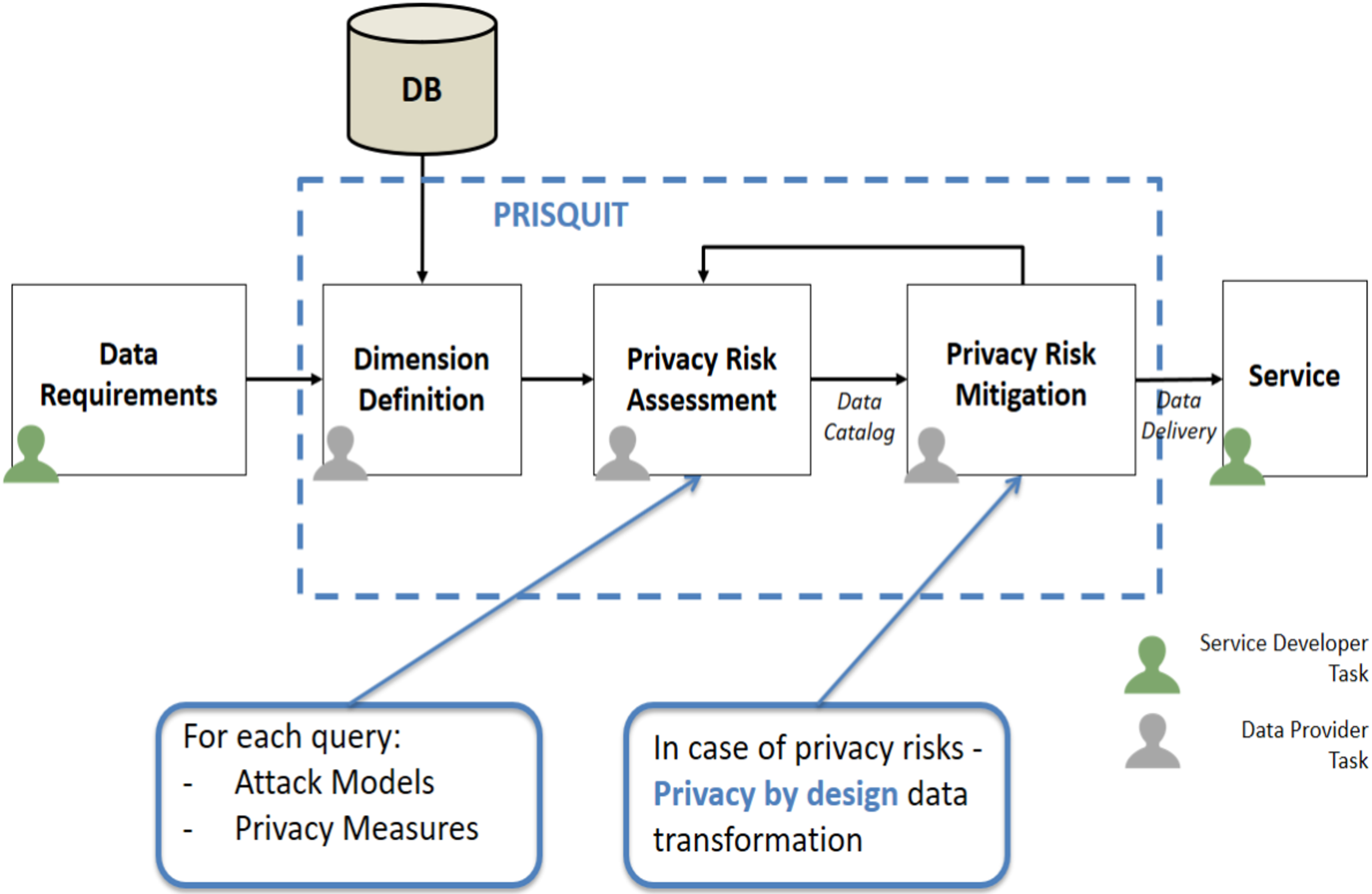
Privacy by design Methodology

- 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**

Privacy Risk Assessment



Privacy-by-Design in Big Data Analytics



Attack Simulation

Tabular data

ID	Gender	DoB	ZIP	DIAGNOSIS
1	F	1962	300122	Cancro
3	F	1960	300133	Gastrite
2	M	1950	300111	Infarto
4	M	1950	300111	Infarto
5	M	1950	300111	Infarto
6	M	1953	300115	Frattura

Background knowledge:

1. Gender, DoB, Zip
2. Gender, DoB
3. Gender, Zip
4. DoB, Zip
5. Gender
6. DoB
7. Zip

Background knowledge:

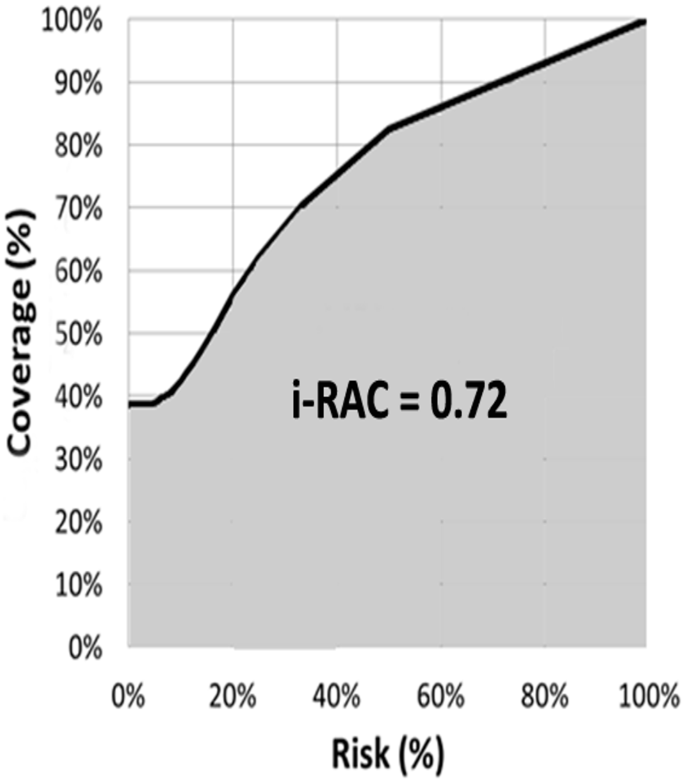
Sequences and Trajectories

All the possible sub-sequences!

$<loc_1, t_1> <loc_2, t_2> <loc_3, t_3> <loc_4, t_4> <loc_5, t_4>$

Risk and Coverage (RaC) curve

- A diagram of coverage (% of data preserved) at varying values of risk
- Concept has analogies with ROC curves.
- Each curve can be summarized by a single measure, e.g. AUC (area under the curve) – the closer to 1, the better



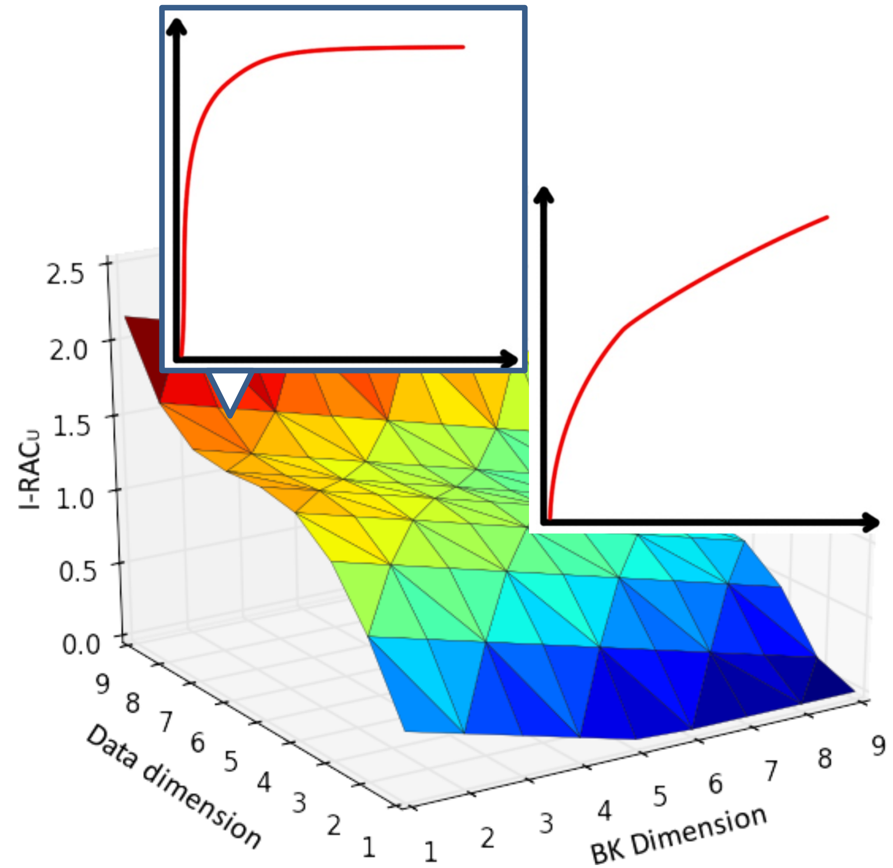
RAC_U →for each risk value, quantifies the percentage of users in U having that risk

RAC_D → for each risk value, quantifies the data in D covered by only users having at most that risk

Empirical Privacy Risk Assessment

- Defining a set of attacks based on common data formats
- Simulates these attacks on experimental data to **calculate privacy risk**

Time complexity is a problem!

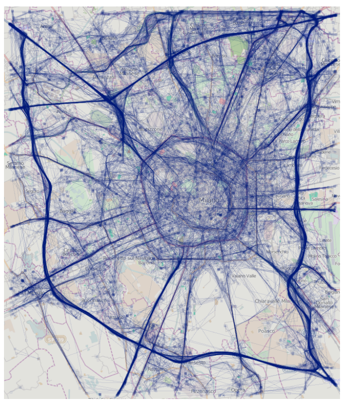


DATA MINING APPROACH

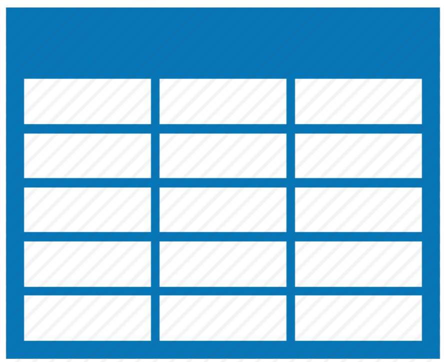
- Using classification techniques to predict the privacy risks of individuals.

1. Simulate the risk of each individual *R*
2. Extract from the dataset a set of individual features *F*
3. Construct a training dataset (F,R)
4. Learning a classifier/regressor to predict the risk/risk level

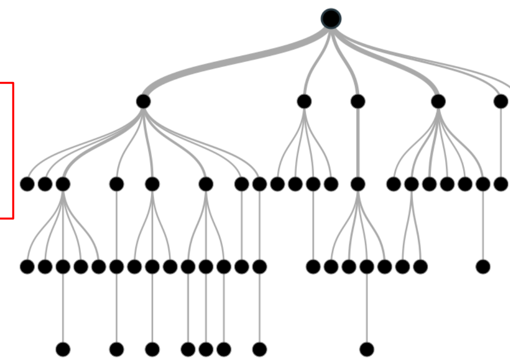
Approach



- Features extraction from raw data
- Privacy Risks values by attack simulation



Learning a classifier



For each new user extracting **Features** and using the classifier to predict the risk

Mobility Data

- GPS provided by Octo-Telematics May 2011, Tuscany

- **Two datasets:**
  - Florence: 9715 trajectories
  - Pisa: 2280 trajectories

- **Classification**:
  - Random Forest Classifier
  - Evaluation by accuracy of classification and weighted average F-measure

Experiments on Mobility Data

<table>
<tr><th>symbol</th><th>name</th><th>structures</th><th>attacks</th></tr>
<tr><td>V</td><td>visits</td><td rowspan="5">trajectory</td><td rowspan="6">LOCATION
LOCATION SEQUENCE
VISIT</td></tr>
<tr><td>$\overline{V}$</td><td>daily visits</td></tr>
<tr><td>D_{max}</td><td>max distance</td></tr>
<tr><td>D_{sum}</td><td>sum distances</td></tr>
<tr><td>$\overline{D}_{sum}$</td><td>D_{sum} per day</td></tr>
<tr><td>D_{max}^{trip}</td><td>D_{max} over area</td><td>trajectory
location set</td></tr>
<tr><td>$Locs$</td><td>distinct locations</td><td>frequency vector</td><td rowspan="2">FREQUENT LOCATION
FREQUENT LOC. SEQUENCE</td></tr>
<tr><td>$Locs_{ratio}$</td><td>$Locs$ over area</td><td>frequency vector
location set</td></tr>
<tr><td>R_g</td><td>radius of gyration</td><td rowspan="2">probability vector</td><td rowspan="3">PROBABILITY</td></tr>
<tr><td>E</td><td>mobility entropy</td></tr>
<tr><td>E_i</td><td>location entropy</td><td>probability vector
probability vector dataset</td></tr>
<tr><td>U_i</td><td>individuals per location</td><td rowspan="5">frequency vector,
frequency vector dataset</td><td rowspan="5">FREQUENCY
PROPORTION
HOME AND WORK</td></tr>
<tr><td>U_i^{ratio}</td><td>U_i over individuals</td></tr>
<tr><td>w_i</td><td>location frequency</td></tr>
<tr><td>w_i^{pop}</td><td>w_i over overall frequency</td></tr>
<tr><td>$\overline{w}_i$</td><td>daily location frequency</td></tr>
</table>

configuration			Florence		Pisa		FI → PI		PI → FI	
			ACC	F	ACC	F	ACC	F	ACC	F
Visit	locations with timestamps	$k=2$	0.94	0.94	0.93	0.93	0.93	0.92	0.93	0.93
		$k=3$	0.94	0.94	0.93	0.93	0.93	0.93	0.93	0.93
		$k=4$	0.94	0.94	0.93	0.93	0.93	0.93	0.92	0.92
		$k=5$	0.94	0.94	0.92	0.92	0.93	0.93	0.91	0.92
	avg baseline		0.82	0.81	0.81	0.80				
Frequency	locations with frequencies	$k=2$	0.90	0.89	0.83	0.82	0.79	0.79	0.76	0.70
		$k=3$	0.94	0.93	0.89	0.89	0.84	0.86	0.83	0.79
		$k=4$	0.92	0.93	0.89	0.89	0.85	0.86	0.85	0.85
		$k=5$	0.93	0.93	0.89	0.89	0.71	0.73	0.85	0.82
	avg baseline		0.53	0.53	0.41	0.41				
HW	two most frequent locations		0.62	0.59	0.57	0.54	0.57	0.55	0.51	0.49
	avg baseline		0.37	0.37	0.28	0.29				
Location	locations without sequence	$k=2$	0.93	0.92	0.86	0.86	0.87	0.87	0.85	0.81
		$k=3$	0.95	0.95	0.91	0.91	0.87	0.87	0.87	0.82
		$k=4$	0.95	0.95	0.91	0.91	0.89	0.89	0.89	0.86
		$k=5$	0.95	0.95	0.91	0.91	0.89	0.90	0.87	0.85
	avg baseline		0.57	0.56	0.44	0.44				
Freq.Loc. Sequence	locations with sequence	$k=2$	0.93	0.92	0.88	0.87	0.88	0.87	0.86	0.83
		$k=3$	0.94	0.94	0.88	0.89	0.90	0.89	0.73	0.66
		$k=4$	0.94	0.94	0.89	0.89	0.85	0.87	0.86	0.82
		$k=5$	0.93	0.94	0.89	0.89	0.90	0.90	0.86	0.83
	avg baseline		0.58	0.57	0.46	0.45				
Frequent Location	locations without sequence	$k=2$	0.81	0.79	0.71	0.69	0.73	0.74	0.65	0.62
		$k=3$	0.86	0.85	0.8	0.78	0.81	0.81	0.75	0.72
		$k=4$	0.87	0.86	0.81	0.79	0.83	0.83	0.79	0.75
		$k=5$	0.87	0.87	0.81	0.8	0.82	0.83	0.78	0.75
	avg baseline		0.65	0.65	0.56	0.55				

Measure importance

	Florence		Pisa			Florence		Pisa	
	measure	**impo.**	**measure**	**impo.**		**measure**	**impo.**	**measure**	**impo.**
1	$\overline{V}$	3.66	$Locs_{ratio}$	3.24	15	U_2^{ratio}	0.96	U_2^{ratio}	0.92
2	E	2.92	D_{sum}	3.22	16	U_n	0.88	U_n	0.88
3	D_{sum}	2.75	$\overline{V}$	2.87	17	w_n^{pop}	0.83	r_g	0.87
4	$Locs_{ratio}$	2.51	E	2.62	18	E_n	0.79	E_n	0.79
5	V	1.91	V	1.69	19	E_2	0.74	E_2	0.75
6	w_1^{pop}	1.77	$Locs$	1.66	20	D_{max}	0.68	w_n^{pop}	0.73
7	$Locs$	1.67	w_1^{pop}	1.62	21	D_{max}^{trip}	0.63	D_{max}^{trip}	0.67
8	U_1	1.44	U_1	1.46	22	r_g	0.61	D_{max}	0.58
9	U_1^{ratio}	1.32	U_1^{ratio}	1.40	23	w_1	0.42	$\overline{w}_1$	0.48
10	$\overline{D}_{sum}$	1.19	U_2	1.16	24	$\overline{w}_2$	0.40	w_1	0.44
11	U_2	1.12	U_n^{ratio}	1.09	25	$\overline{w}_1$	0.36	$\overline{w}_2$	0.36
12	w_2^{pop}	1.07	w_2^{pop}	1.07	26	w_n	0.13	w_n	0.15
13	E_1	1.05	E_1	1.06	27	$\overline{w}_n$	0.12	w_2	0.13
14	U_n^{ratio}	0.99	$\overline{D}_{sum}$	0.98	28	w_2	0.10	$\overline{w}_n$	0.13

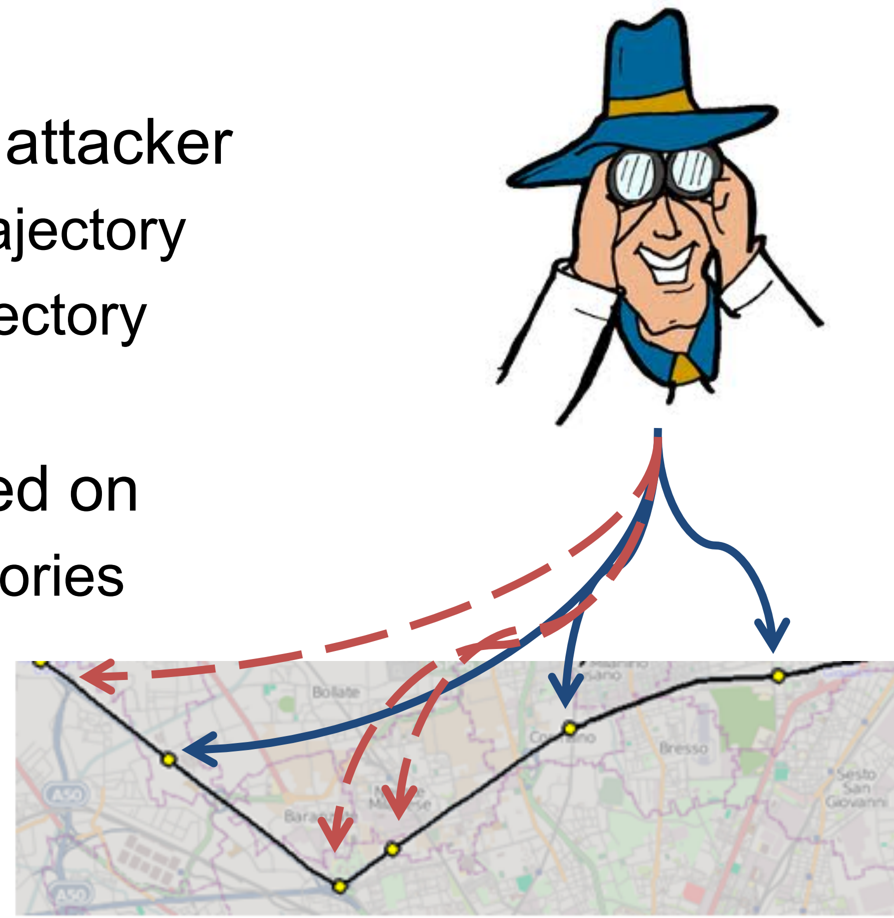
Privacy by Design in Mobility Atlas

A. Monreale, G. Andrienko, N. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo
The Journal Transactions on Data Privacy, 2010

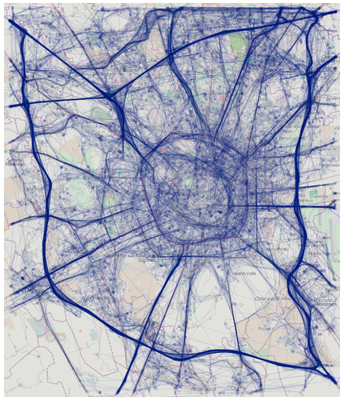
Knowledge Discovery and Delivery Lab
(ISTI-CNR & Univ. Pisa)
www-kdd.isti.cnr.it

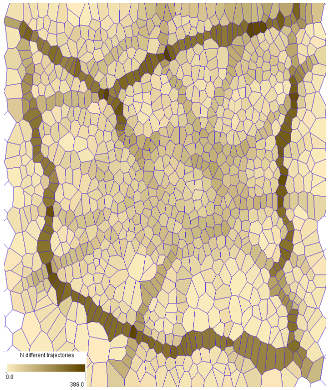
Privacy-Preserving Framework

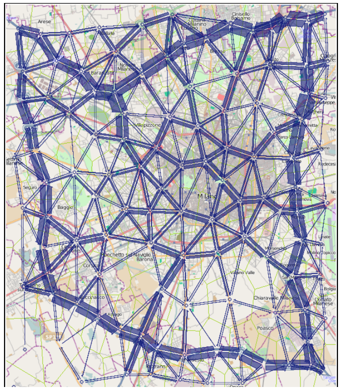
- Anonymization of movement data while preserving clustering
- **Trajectory Linking Attack**: the attacker
  - knows some points of a given trajectory
  - and wants to infer the whole trajectory
- **Countermeasure**: method based on
  - **spatial generalization** of trajectories
  - **k-anonymization** of trajectories



Trajectory Generalization



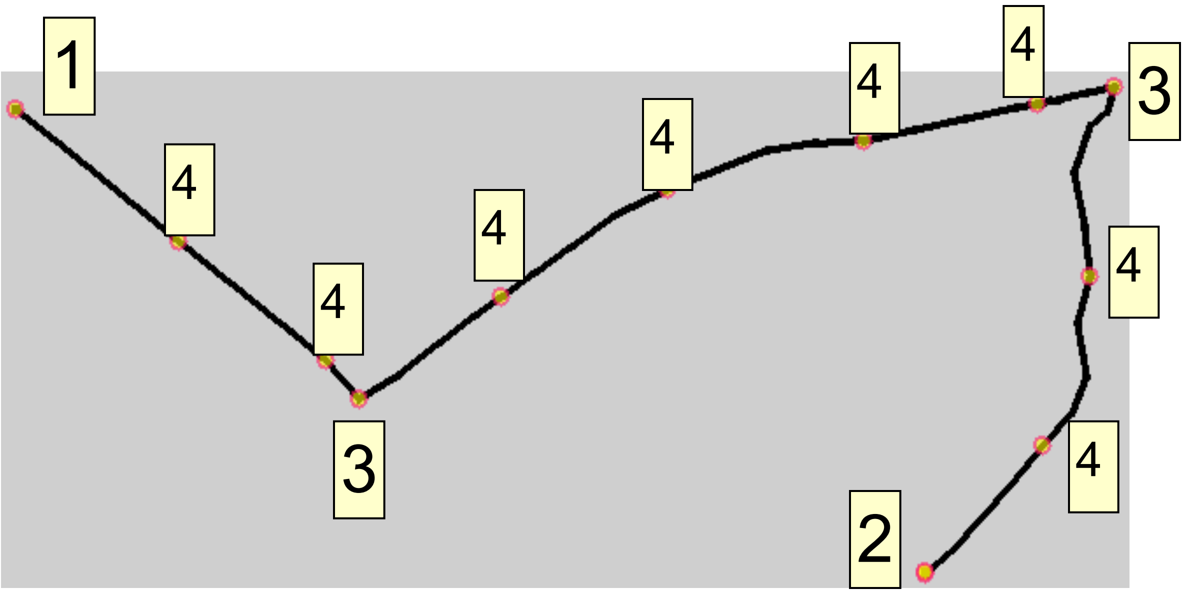




- Given a trajectory dataset
  1. Partition of the territory into **Voronoi cells**
  2. Transform trajectories into sequence of cells

Partition of territory: Characteristic points

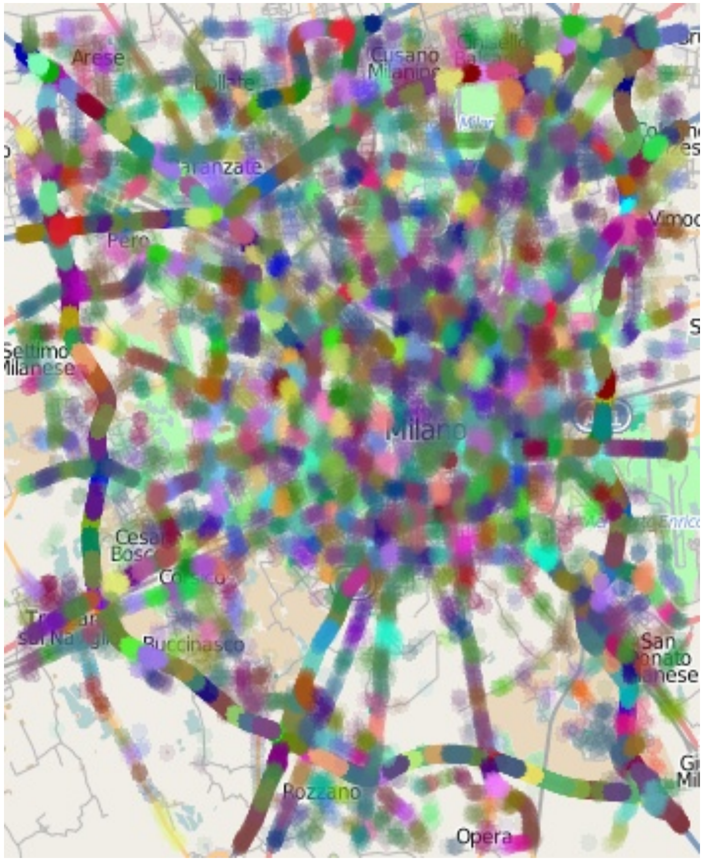
- Characteristic points extraction:
  - Starts (1)
  - Ends (2)
  - Points of significant turns (3)
  - Points of significant stops,and representative points from long straight segments (4)



Partition of territory: spatial clusters

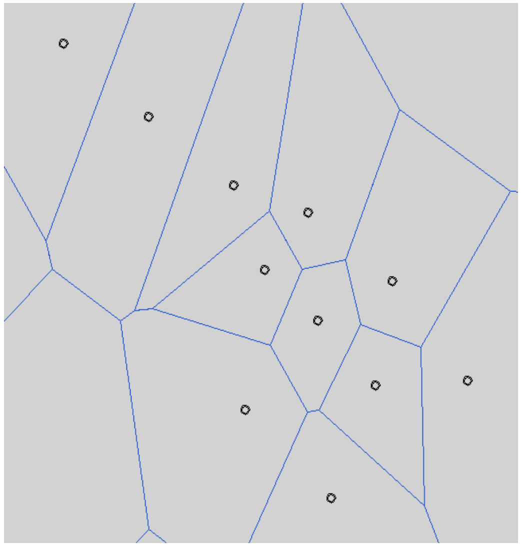
- Group the extracted points in **Spatial Clusters** with desired spatial extent

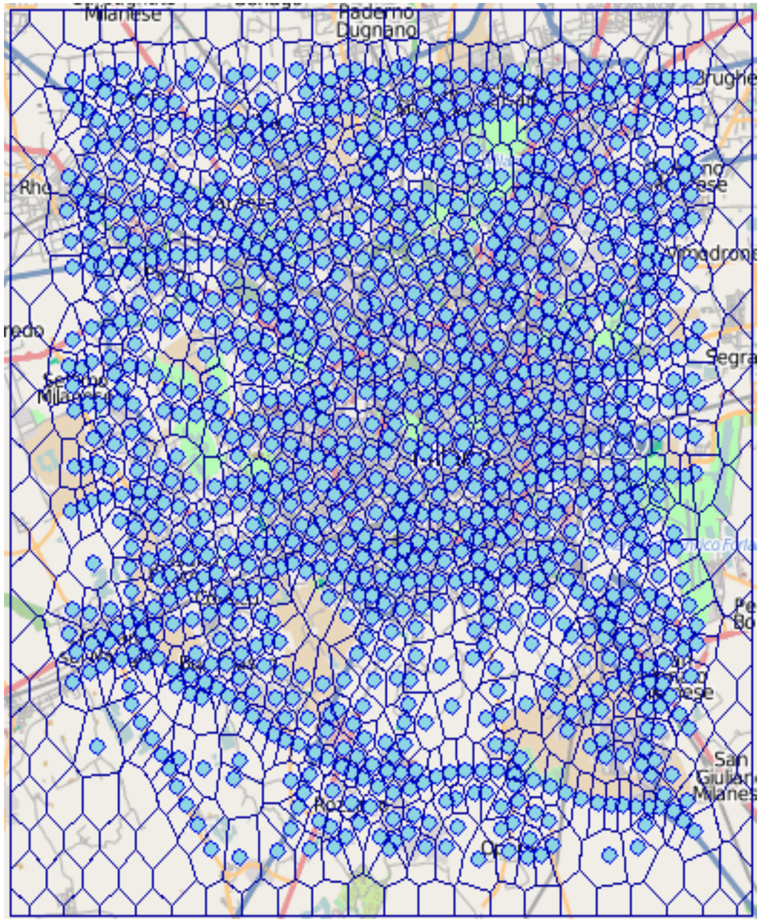
- **MaxRadius**: parameter to determine the spatial extent and so the degree of the generalization



Partition of territory: Voronoi Tessellation

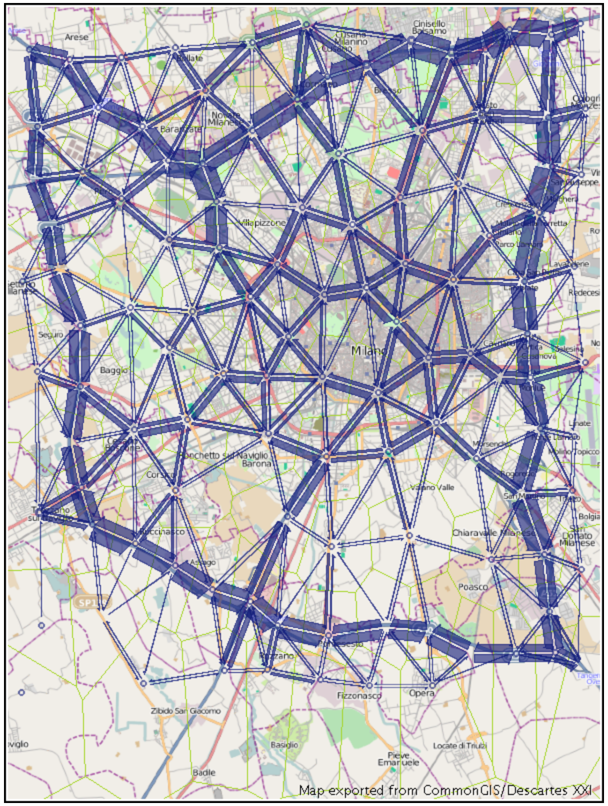
- Partition the territory into **Voronoi cells**
- The **centroids** of the spatial clusters used as generating points





Generation of trajectories

- Divide the trajectories into segments that link Voronoi cells
- For each trajectory:
  - the area a_1 containing its first point p_1 is found
  - The following points are checked
  - If a point p_i is not contained in a_1 for it the containing area a_2 is found
  - and so on …
- **Generalized trajectory**: From sequence of areas to sequence of centroids of areas

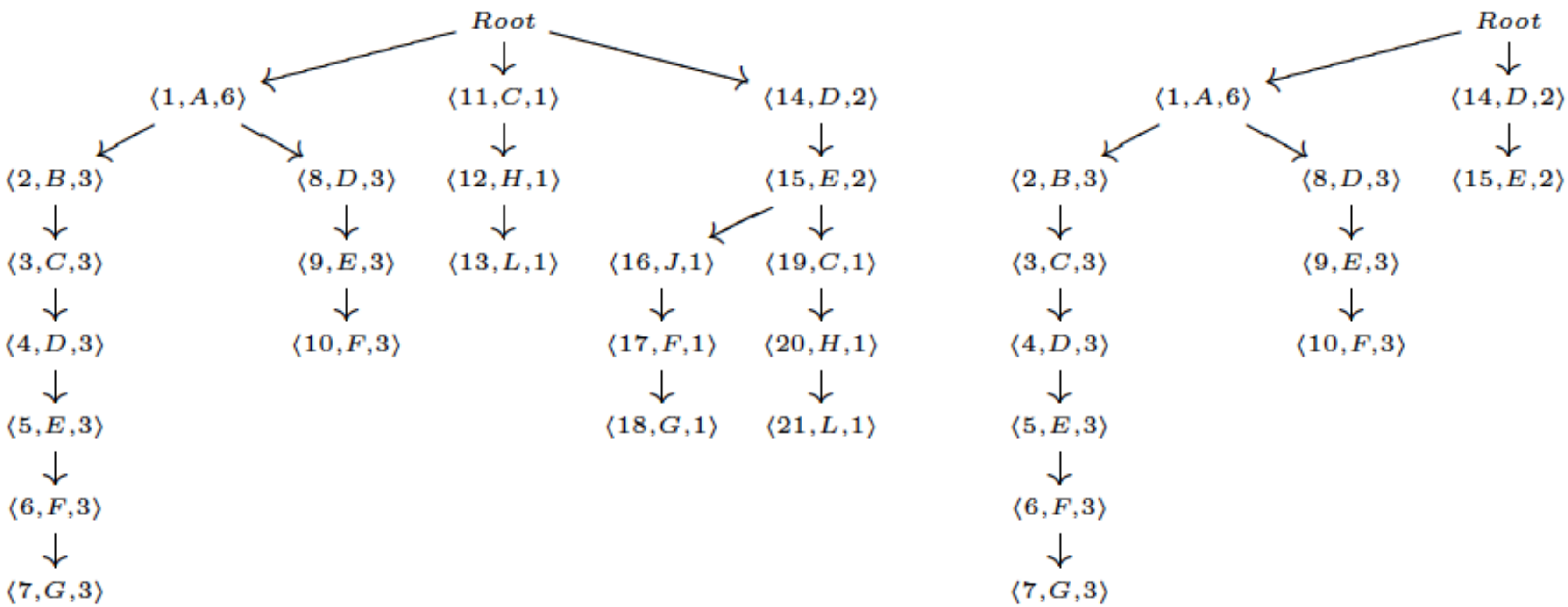


Generalization vs k-anonymity

- Generalization could not be sufficient to ensure k-anonymity:
  - For each generalized trajectory there exist at least others k-1 different people with the same trajectory?
- Two transformation strategies
  - KAM-CUT
    - publishing only the k-frequent prefixes of the generalized trajectories
  - KAM-REC
    - recovering portions of trajectories which are frequent at least k times
    - without introducing noise

KAM-CUT Approach

- The prefix tree is anonymized w.r.t. a threshold k
  - all the trajectories whose support is less than k are pruned from the prefix tree



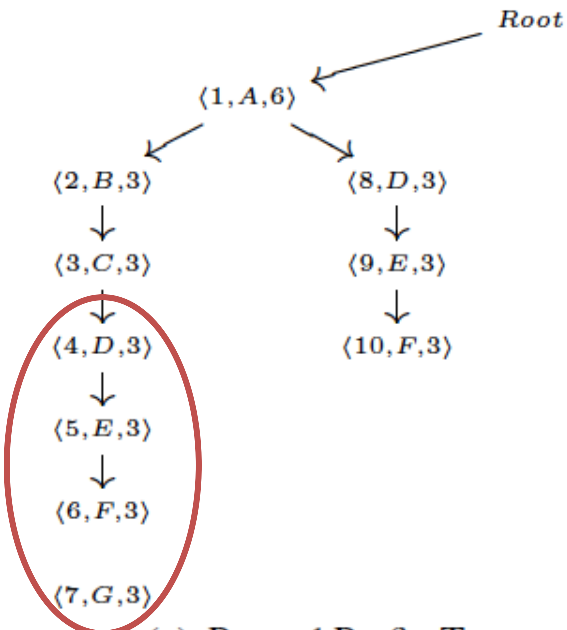
(a) Prefix Tree Construction

(b) Anonymized Prefix Tree

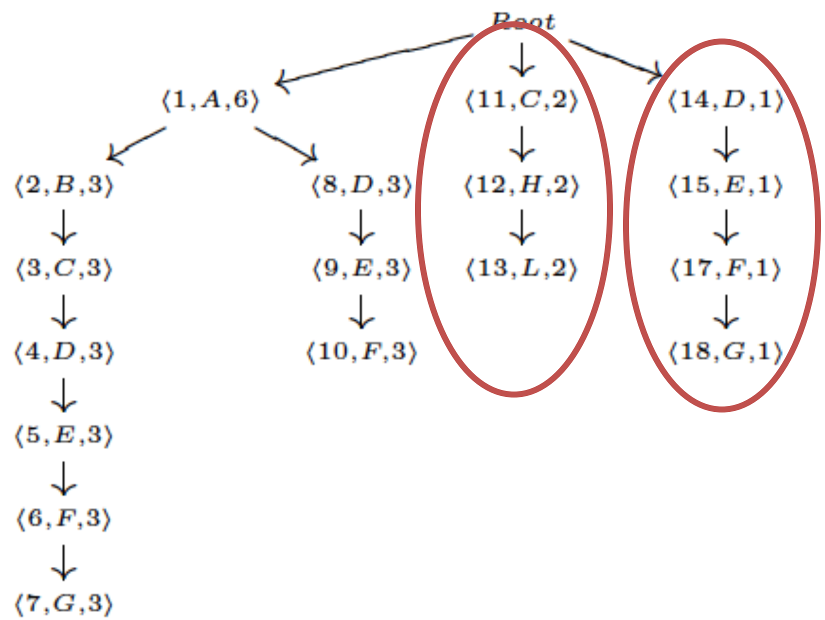
KAM-REC Approach

- The prefix tree is anonymized w.r.t. a threshold k
  - all the trajectories with support less than k are pruned from the prefix tree and put into a list
  - A subtrajectory is recovered and appended to the root if
    - appears in the prefix tree
    - appears in at least k different trajectories in the list

KAM-REC: Example



(a) Pruned Prefix Tree



(b) Anonymized Prefix Tree

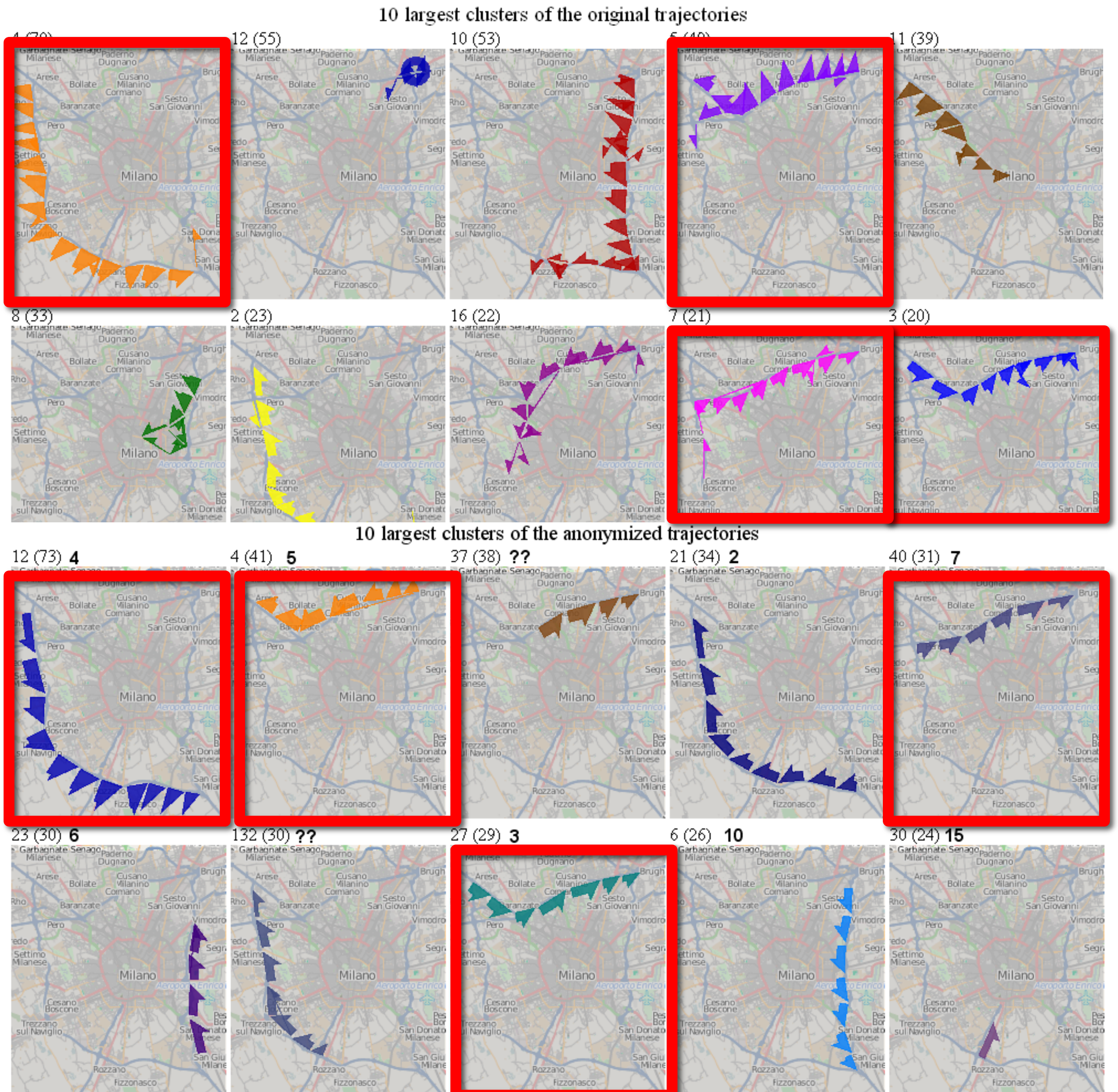
$\mathcal{L}_{cut}$

$(CHL, 1)$

$(DEJFG, 1)$

$(DECHL, 1)$

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)
…..	