DATA MINING 2
Ethics Principles: Privacy

Anna Monreale

a.a. 2019/2020

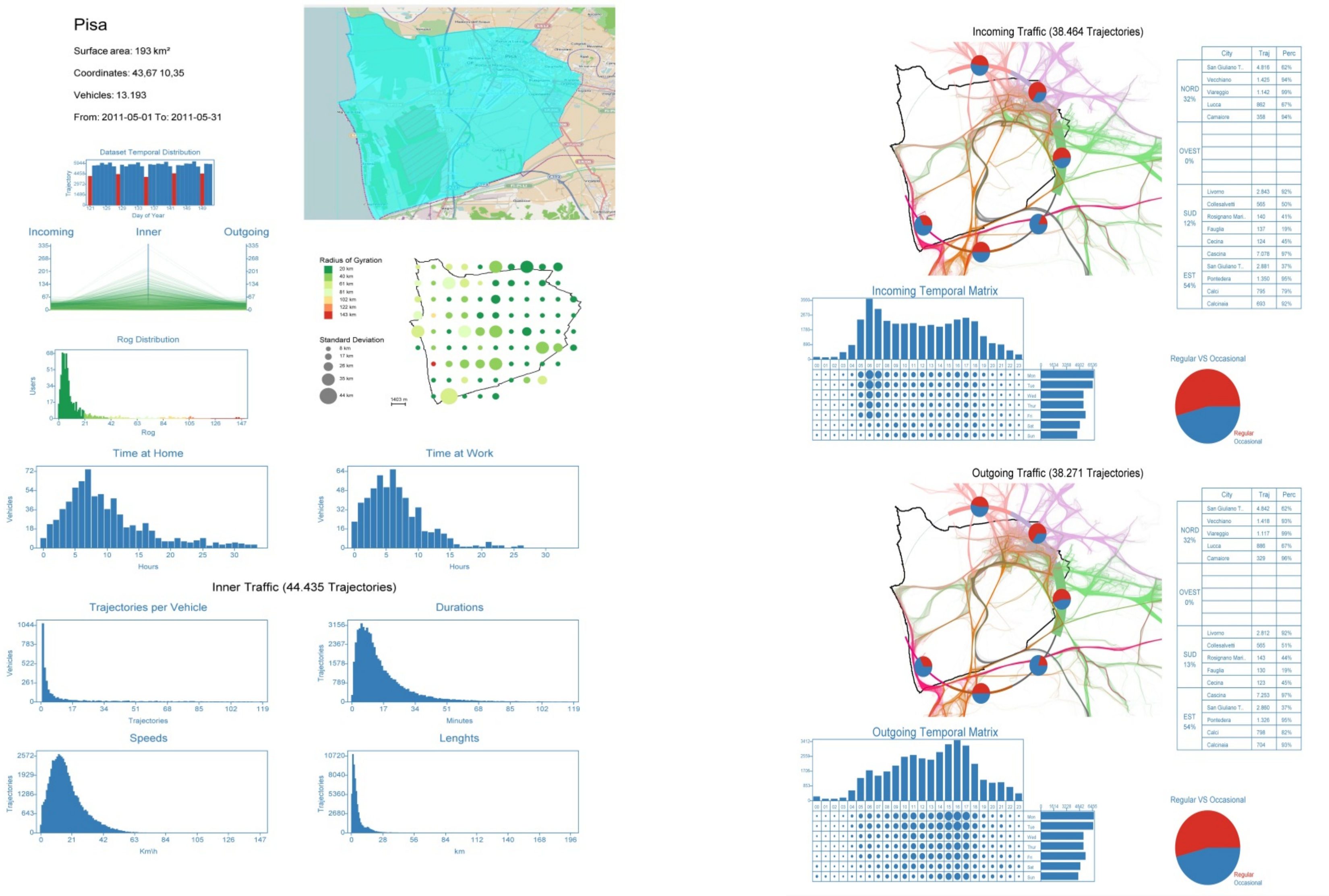
UNIVERSITÀ DI PISA

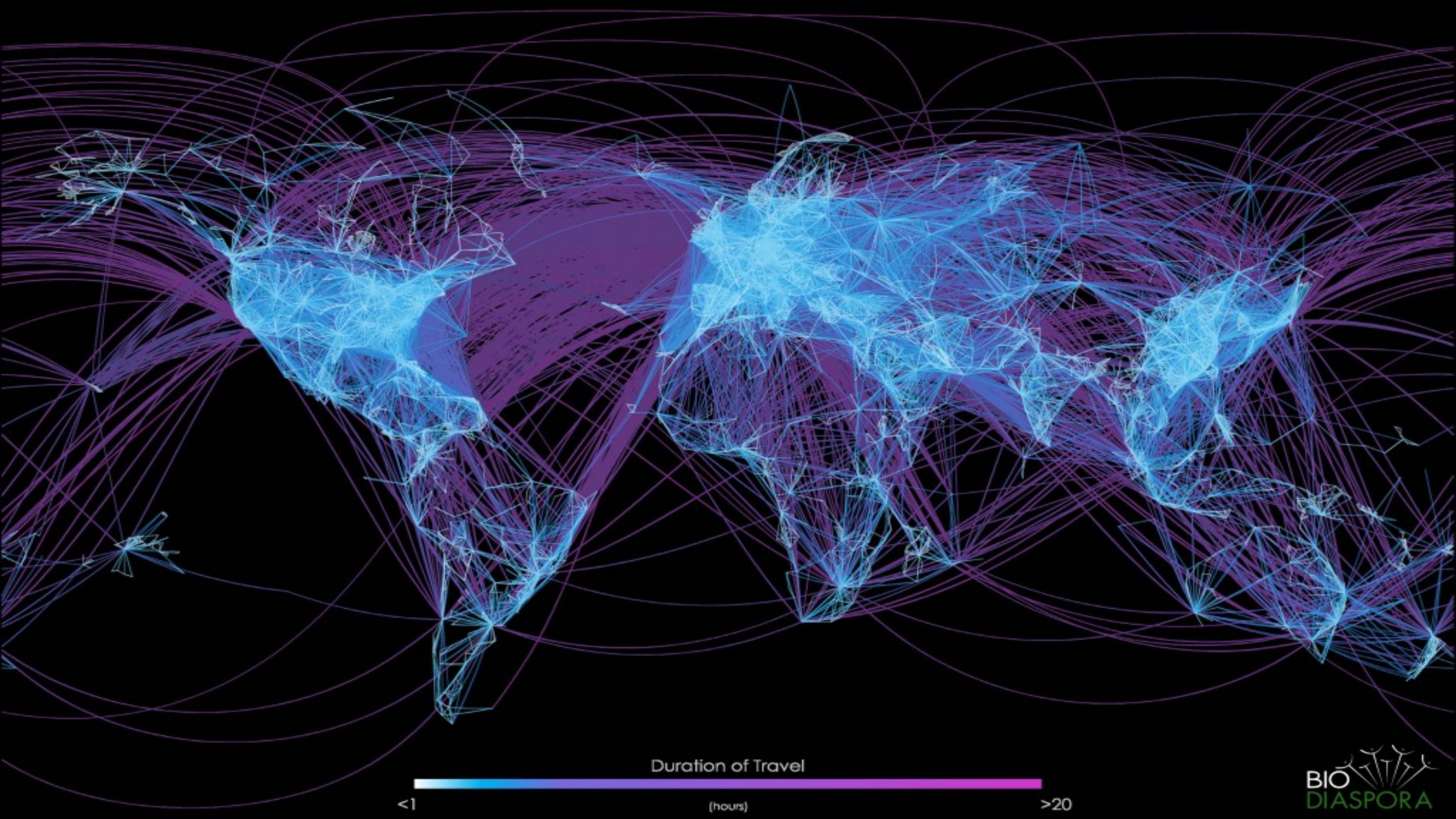
Our digital traces ….

- 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: New, More Carefully Targeted Financial Services

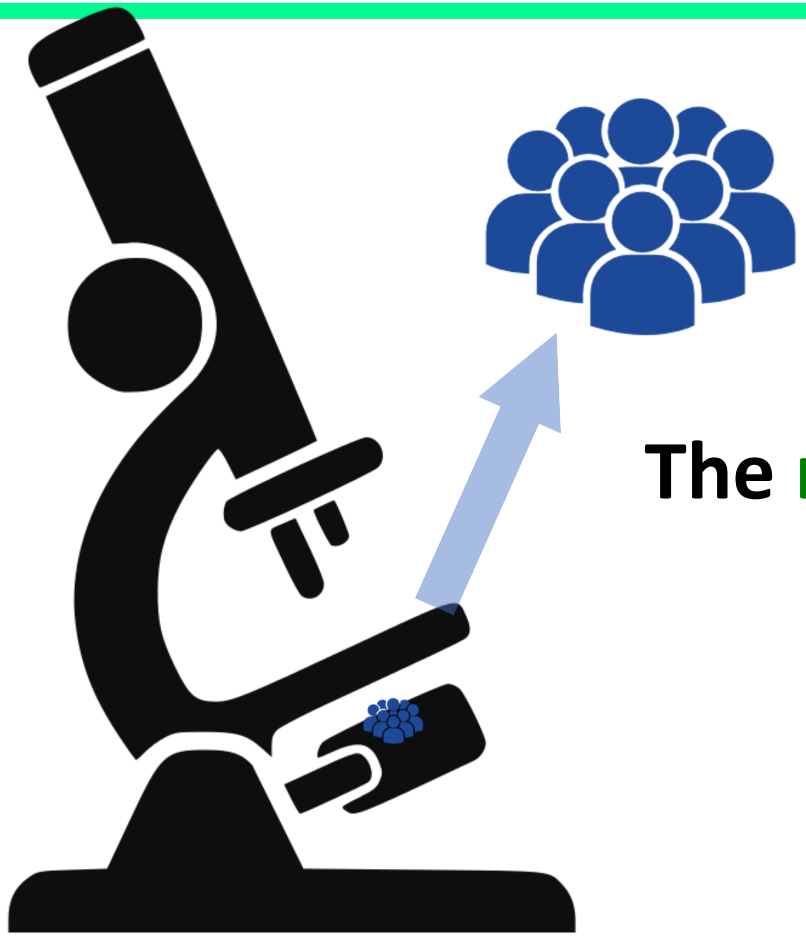
Mobility Atlas of Many Cities





Duration of Travel
<1
(hours)
>20
BIO
DIASPORA

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

EU Requirements for trustworthy AI

1. **Privacy:** avoid re-identification of people in data and sensitive inferences
2. **Transparency/Explainability**: transparency should be applied to every stage of the AI lifecycle, indeed it prescribes the possibility to have a complete view on the whole system
3. **Fairness**: avoid AI base their decision on sensitive attributes like gender, religion belief, etc.
4. **Robustness**: AI system developers should prevent system hacking and adversarial attacks.
5. **Accountability:** allow propriate mechanisms to identify the responsibility for AI systems' outcomes are put in place during their whole lifecycle
6. **Sustainability**: the design stage of an AI system there should be an environmental impact assessment (e.g., **climate** impact)

EU Requirements for trustworthy AI

1. **Privacy:** avoid re-identification of people in data and sensitive inferences
2. **Transparency/Explainability**: transparency should be applied to every stage of the AI lifecycle, indeed it prescribes the possibility to have a complete view on the whole system
3. **Fairness**: avoid AI base their decision on sensitive attributes like gender, religion belief, etc.
4. **Robustness**: AI system developers should prevent system hacking and adversarial attacks.
5. **Accountability:** allow propriate mechanisms to identify the responsibility for AI systems' outcomes are put in place during their whole lifecycle
6. **Sustainability**: the design stage of an AI system there should be an environmental impact assessment (e.g., **climate** impact)

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

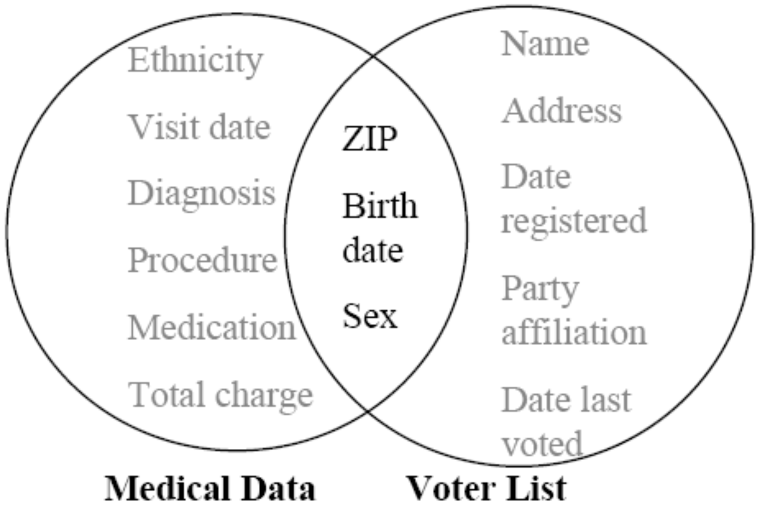
Is Pseudonymization enough for data protection?

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	Cancro
2	F	1960	300133	Gastrite
3	M	1950	300111	Infarto
4	M	1955	300112	Emicrania
5	F	1965	300200	Lussazione
6	M	1953	300115	Frattura

Which is the disease of the Governor?

Making Data Anonymous

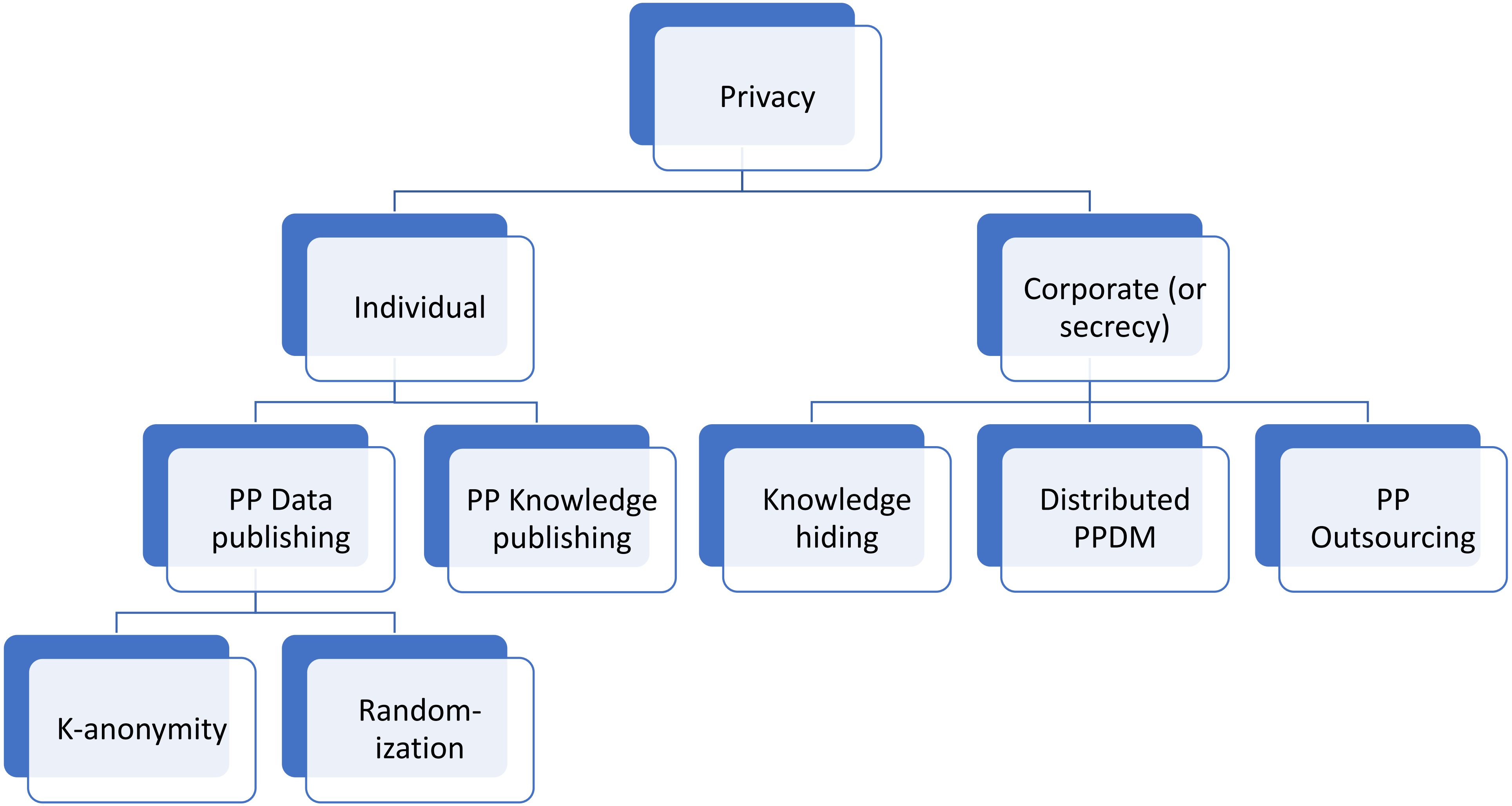
K-Anonymity

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

ID	Gender	YoB	ZIP	DIAGNOSIS
1	F	[1960-1965]	300***	Cancro
2	F	[1960-1965]	300***	Gastrite
3	M	[1950-1955]	30011*	Infarto
4	M	[1950-1955]	30011*	Emicrania
5	F	[1960-1965]	300***	Lussazione
6	M	[1950-1955]	30011*	Frattura

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	Cancro
3	F	1960	300133	Gastrite
2	M	1950	300111	Infarto
4	M	1955	300112	Emicrania
5	F	1965	300200	Lussazione
6	M	1953	300115	Frattura

K-Anonymity

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	Cancro
2	F	1960	300133	Gastrite
3	M	1950	300111	Infarto
4	M	1950	300111	Infarto
5	M	1950	300111	Infarto
6	M	1953	300115	Frattura

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	Cancro
3	F	1960	300133	Gastrite
2	M	1950	300111	Infarto
4	M	1950	300111	Emicrania
5	M	1950	300111	Lussazione
6	M	1953	300115	Frattura

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 & Differential Privacy

Randomization

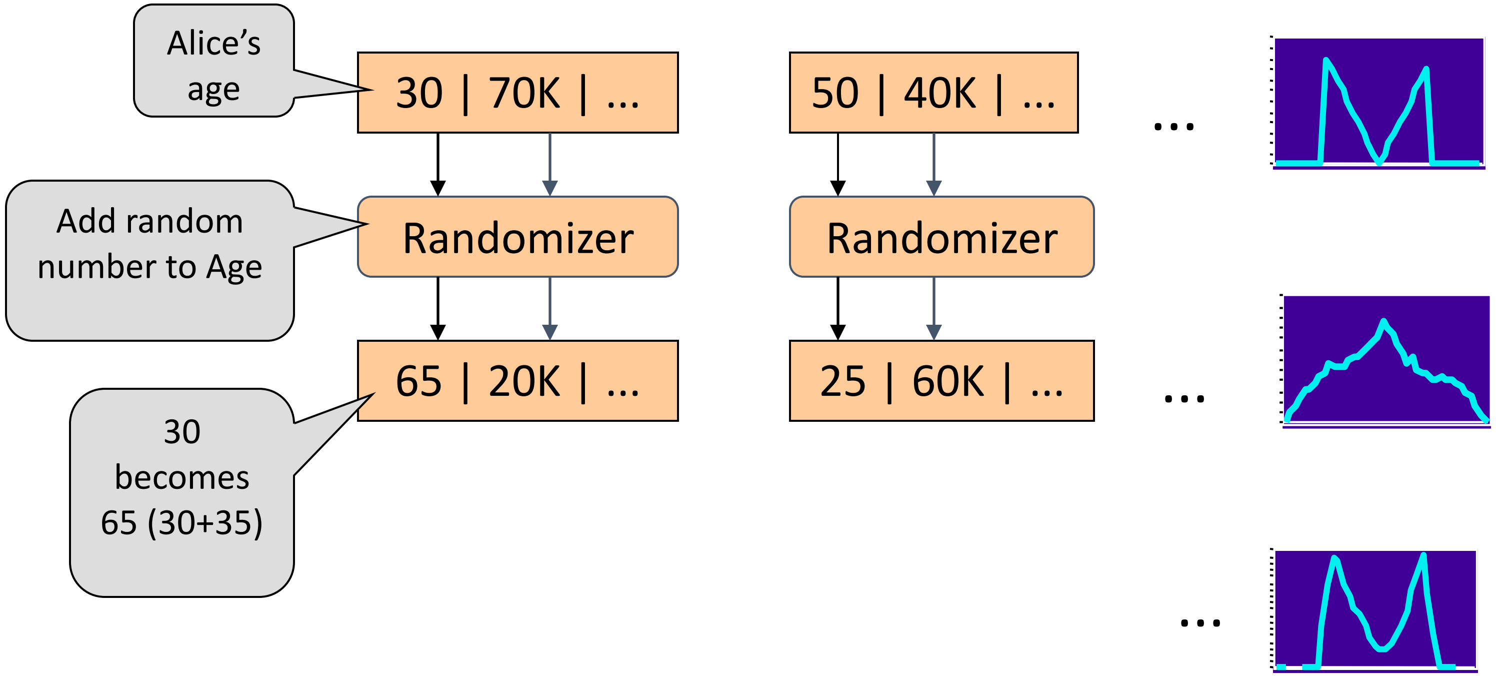
- **Original values $x_1, x_2, ..., x_n$**
  - from probability distribution X (unknown)
- **To hide these values, we use $y_1, y_2, ..., y_n$**
  - from probability distribution Y
    - Uniform distribution between $[-\alpha, \alpha]$
    - Gaussian, normal distribution with $\mu = 0, \sigma$

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

Estimate the probability distribution of X.

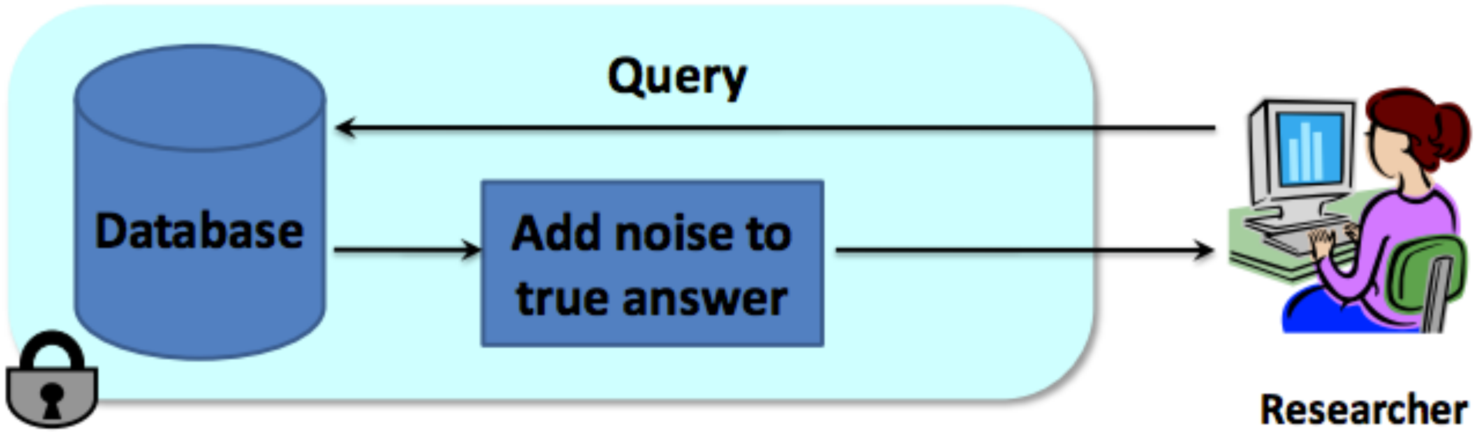
R. Agrawal and R. Srikant. Privacy-preserving data mining. In Proceedings of SIGMOD 2000.

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

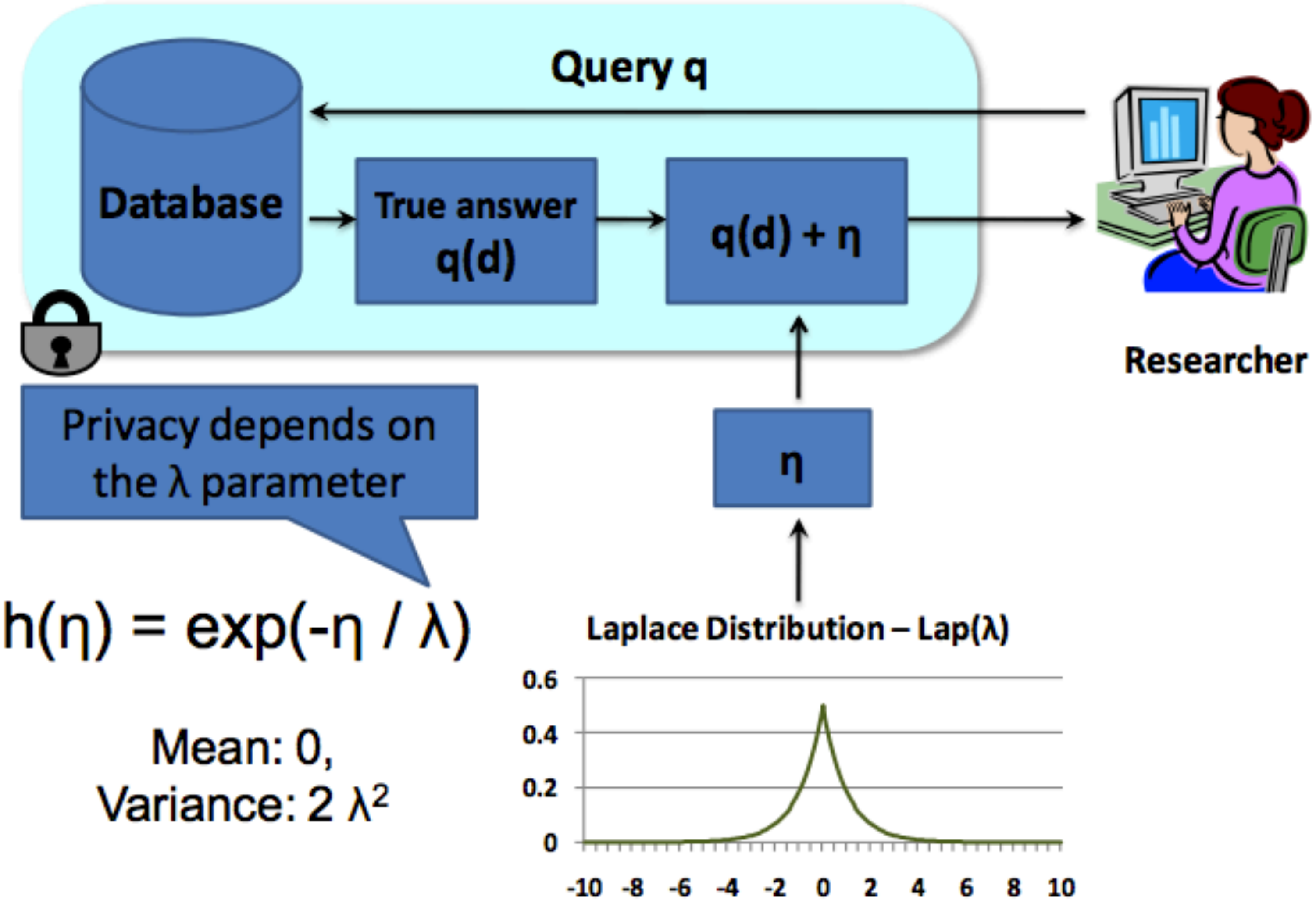
1) how many persons have Diabetes? **4**
2) how many persons, excluding Alice, have Diabetes? **3**

- **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



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

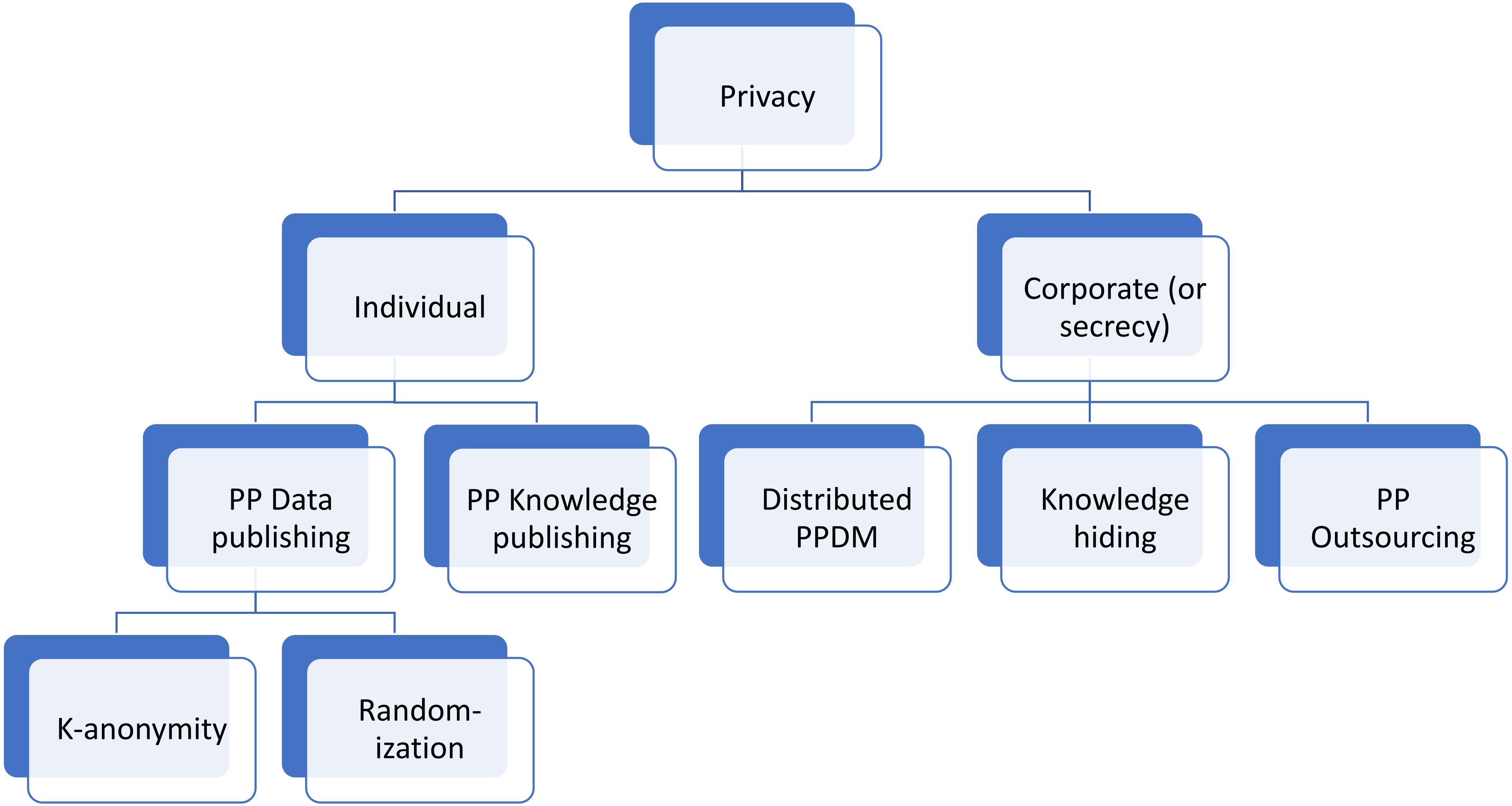
Mean: 0,
Variance: $2\lambda^2$

Randomization & Differential Privacy

- 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

- 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 by Design and Risk Assessment

Privacy by Design Methodology

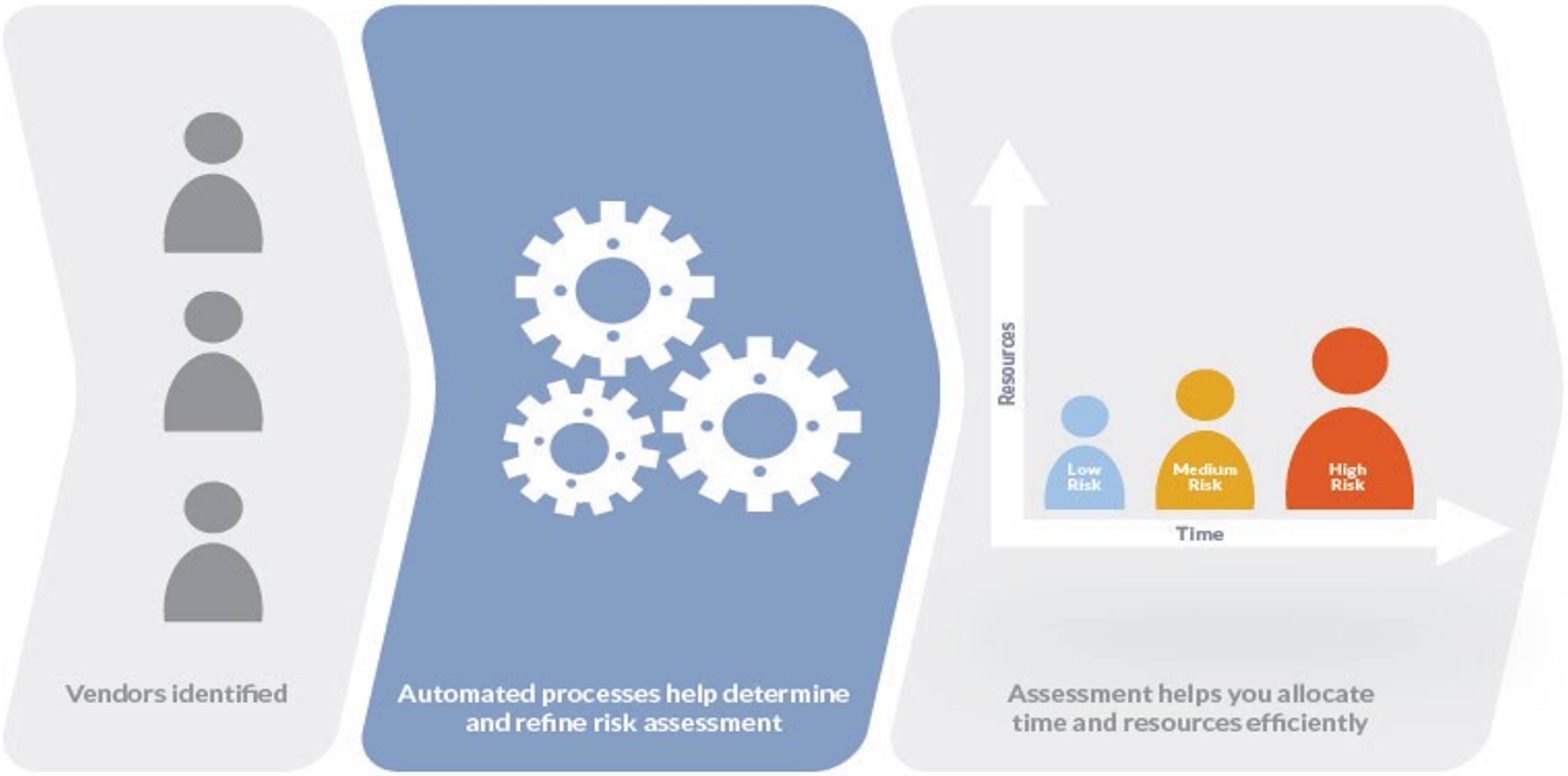
The framework is designed with assumptions about

- 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

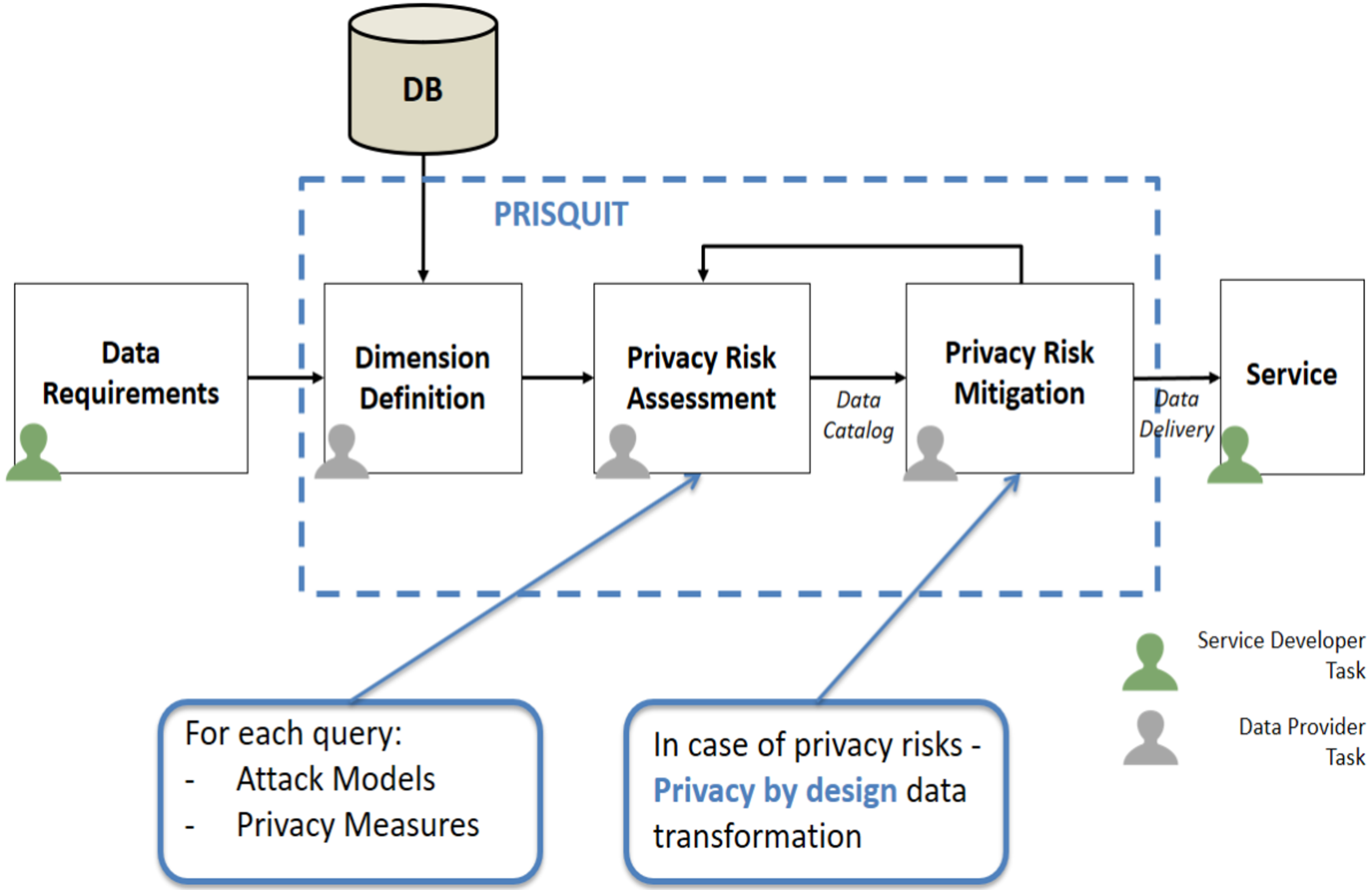
Design a privacy-preserving framework able to

- 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



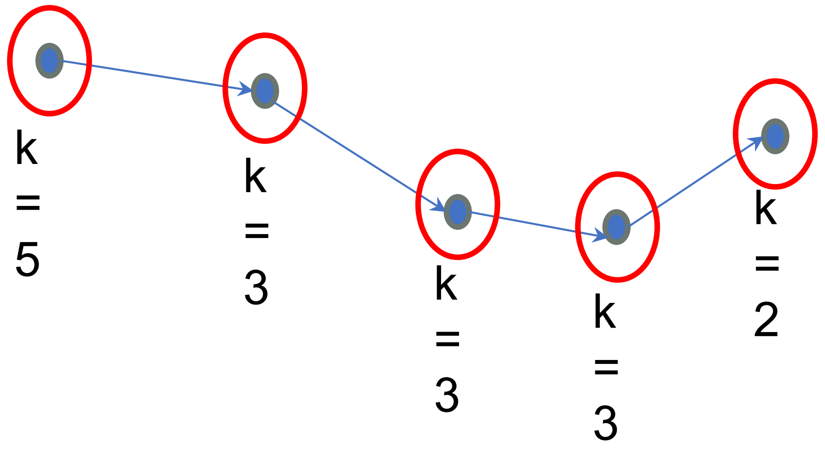
Privacy-by-Design in Big Data Analytics



Privacy Risk Measures

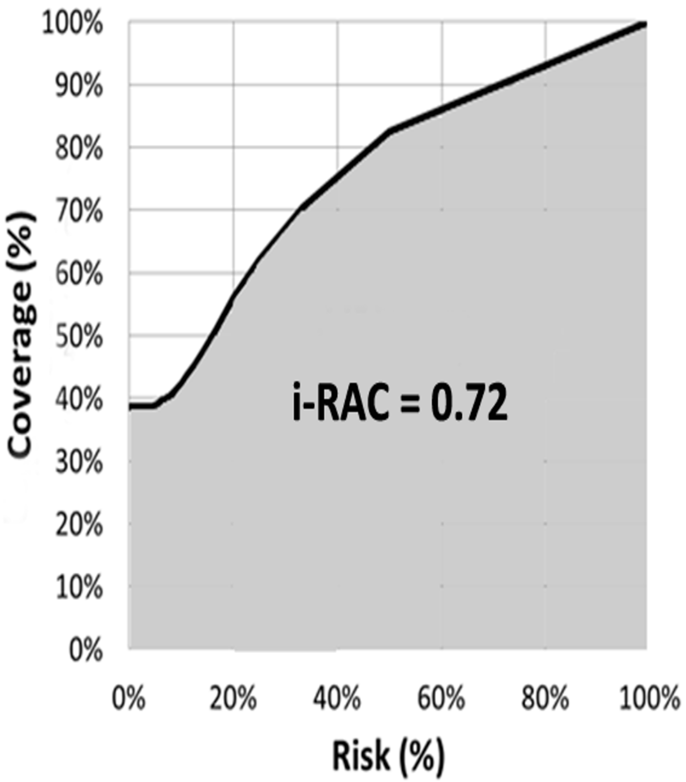
- **Probability of re-identification** denotes the probability to correctly associate a record to a unique identity, *given* a BK

- **Risk of re-identification** is the maximum probability of re-identification *given* a set of BK



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 <u>only</u> users having at most that risk

Attack Simulation

Tabular data

Background knowledge:

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

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:

All the possible sub-sequences!

Sequence:

$< a_1, t_1> <a_2, t_2> <a_3, t_3> <a_4, t_4> <a_5, t_4>$

The Approach

Suitable for any form of data: tabular, graphs, sequences

Key issue: the language of BK – how to specifies the set of possible attacks

Several kinds of data in each domain. Ex. in **mobility**:
- presence (individual frequent locations)
- trajectory (individual movements)
- road segment (collective frequent links)
- profiles (individual systematic movements)
- individual call profiles (from CDR data)

Purchasing Data

Basket
It is an ordered sequence of items.

$$\boldsymbol{b_p = \langle i_1, i_2, i_3, \dots, i_D \rangle}$$

Where $i_i \in I$ the set of items.

Historical baskets
It is the concatenation of the temporally ordered basket of a customer.

$$\boldsymbol{Basket_u = b_1 \cdot b_2 \cdot b_3 \cdots b_m}$$

Where m is the total number of baskets of the customer u in the dataset.

Adversary Attack: Item Sequence Attack

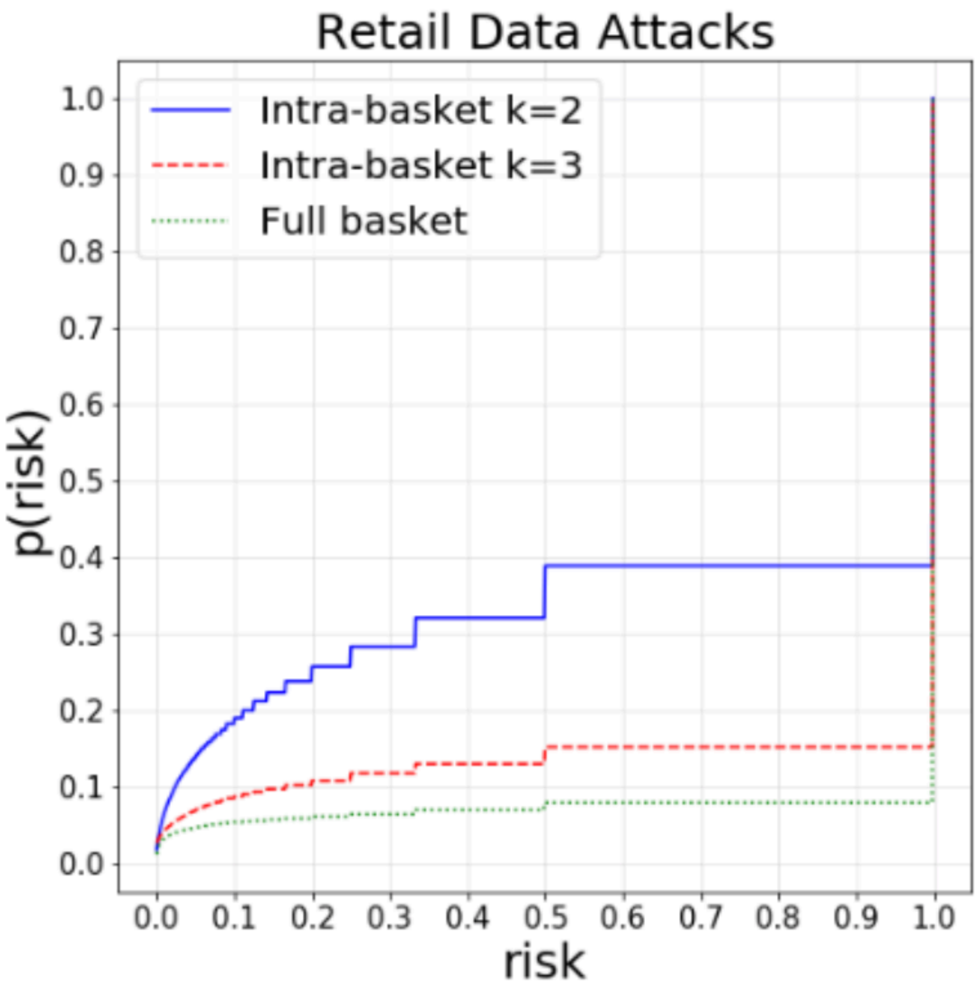
- The adversary knows a ***subset of items*** purchased by the customer and their ***temporal order***

- On ***historical baskets*** (temporally ordered concatenation of the customer's baskets).

- k number of items i_i of an individual u known by the adversary;
- ***Item sequence background knowledge***: a set of configurations based on k items $B_k = I^{u,k}$
- The ***matching function*** is defined as

$$matching(d, b) = \begin{cases} true, & if\ b \subseteq Basket_u \\ false, & otherwise \end{cases}$$

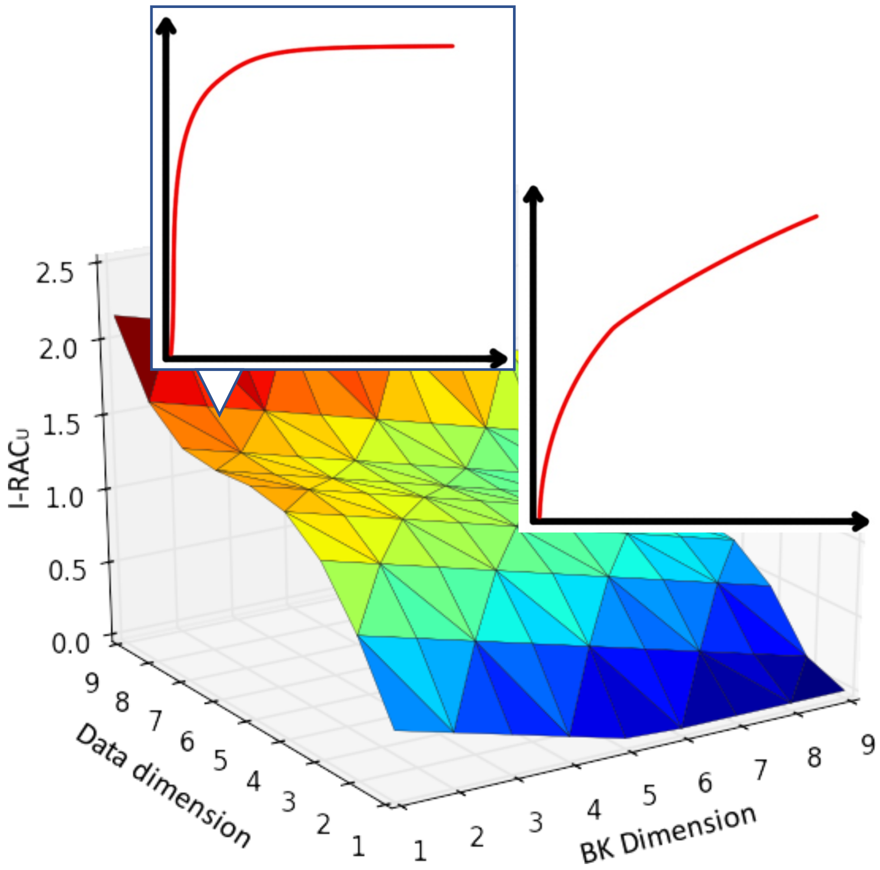
Simulation Attack Model



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!



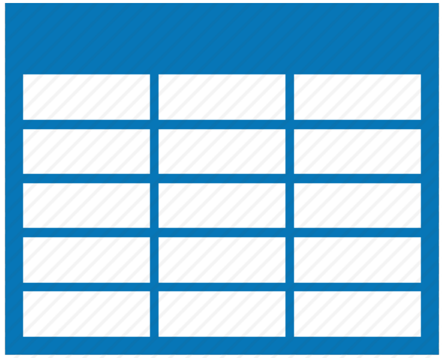
Data Mining Approach

Using classification techniques to predict the privacy risks of individuals.

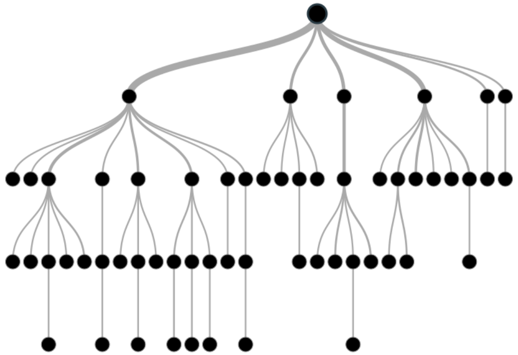
1. Simulate the risk of each individual *R*
2. Extract from the dataset a set of individual features *F*
3. Construct a training dataset (F,R)
4. Learning a classifier/regressor to predict the risk/risk level

Data Mining 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

Features

symbol	name	symbol	name
I	Total number of items	$\overline{I}^{daily}_{max}$	Maximum number of products in a day divided by the total products
I_{unique}	Total number of unique items	$\overline{I}^{daily}_{avg}$	Average number of products in a day divided by the total products
I_{avg}	Total number of items averaged over time	E_{i_j}	Product entropy
I^{d}_{max}	Maximum number of items bought in a day	w_{i_j}	Frequency of the product
I^{d}_{avg}	Average number of items bought per day	$w^{avg}_{i_j}$	Average frequency of the product
E	Purchasing entropy	U_{i_j}	Number of users who bought the product
$Locs$	Distinct locations	$U^{avg}_{i_j}$	Average number of users who bought the product
I^{avg}_{unique}	Total number of unique items averaged over time		

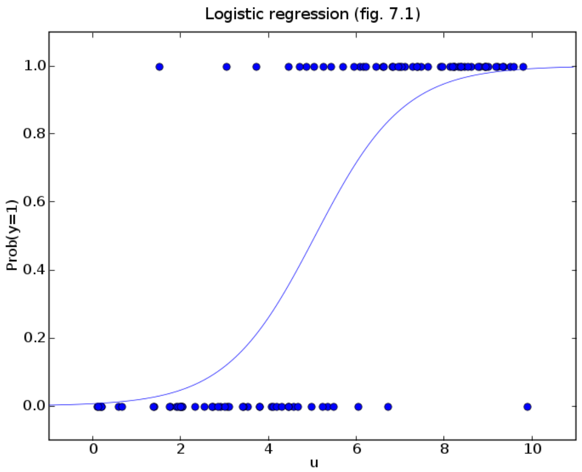
Privacy risk prediction: example of training data

UserId	Procuct Entropy	Unique Items	Num. Items	Prurchase Entropy	Risk
u_1	0.9	9	280	0.9	1.0
u_2	1	13	400	1	1.0
u_3	0.12	2	58	0.12	0.15
u_4	0.09	2	61	0.09	0.075
u_5	0.22	4	120	0.22	0.25

Feature-based Predictor

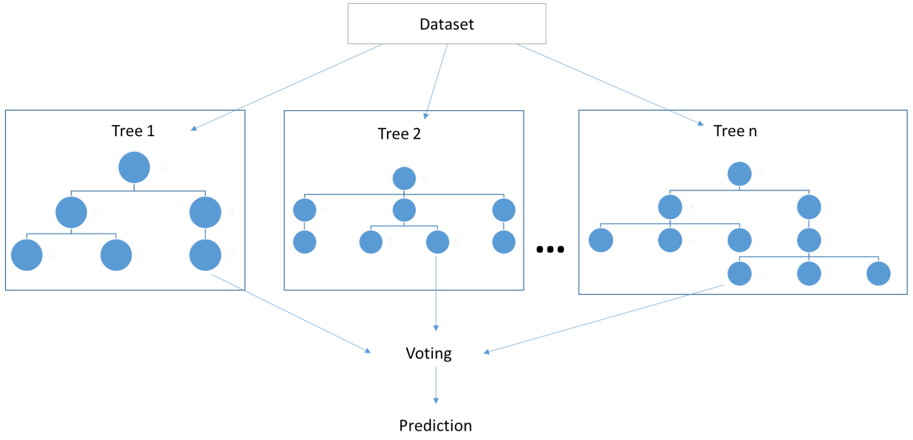
Logistic regression

- A probability model;
- First, it applies a linear function; then a sigmoid function.



Random forest

- Ensemble model composed of decision trees;
- Random sampling for the creation of a tree;
- Majority vote for the final output.

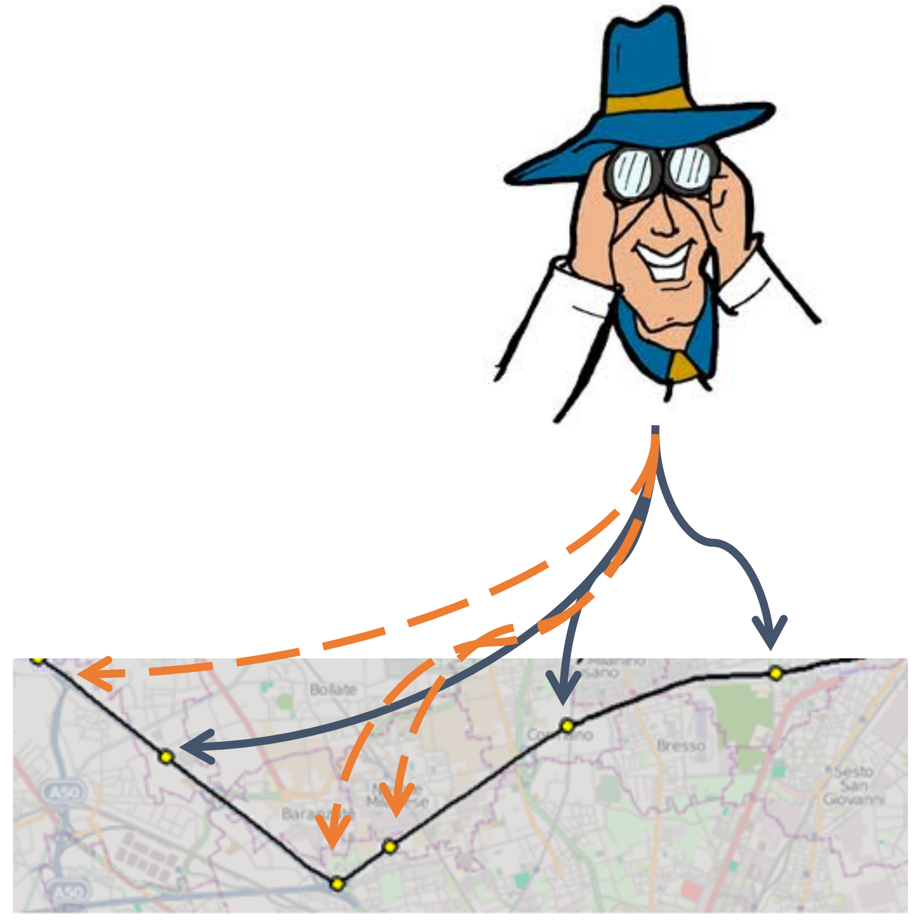


Mitigation Strategy

- Anonymization of movement data while preserving clustering

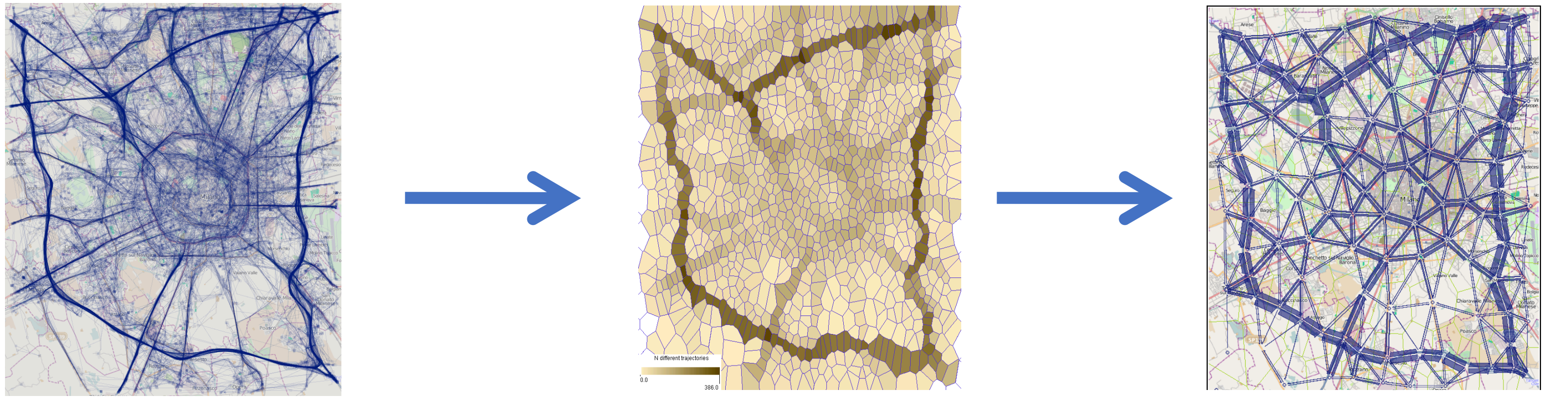
- **Trajectory Linking Attack**: the attacker
  - knows some points of a given trajectory
  - and wants to infer the whole trajectory

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



Trajectory Generalization

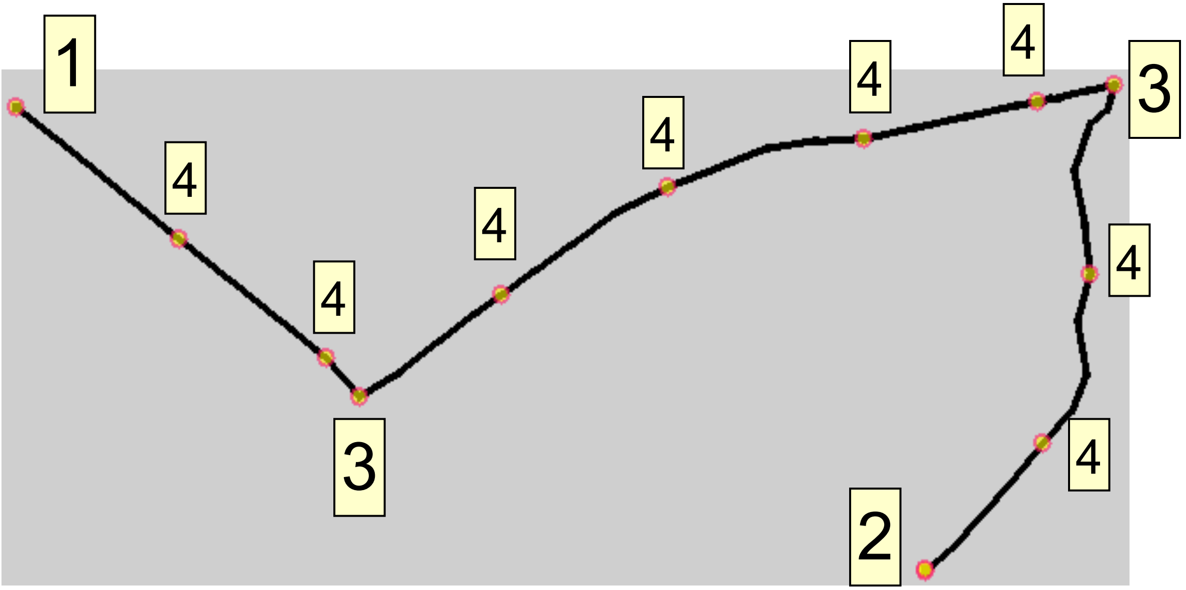
- Given a trajectory dataset
  1. Partition of the territory into **Voronoi cells**
  2. Transform trajectories into sequence of cells



Partition of territory: Characteristic points

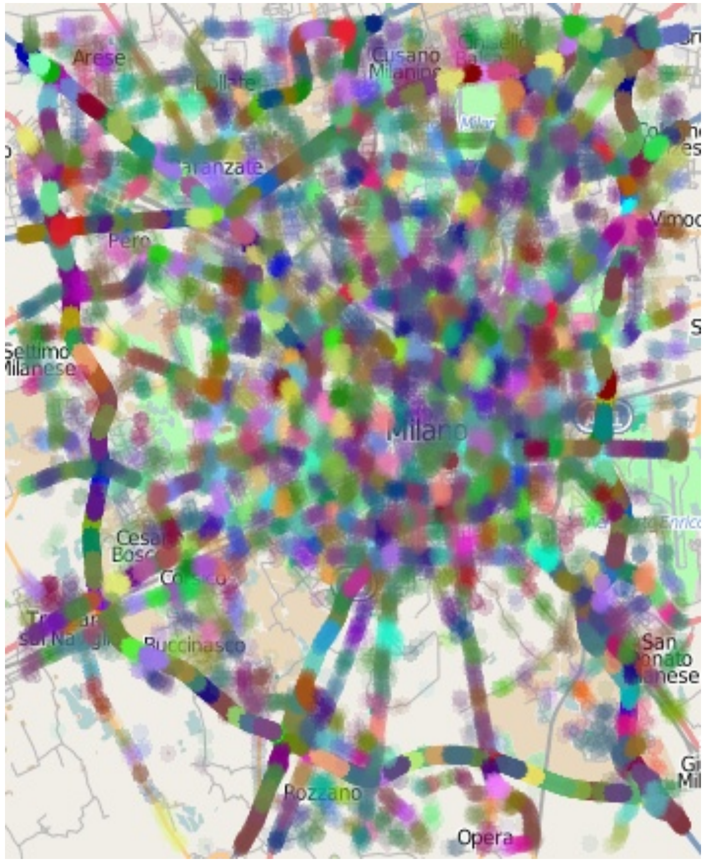
Characteristic points extraction:

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



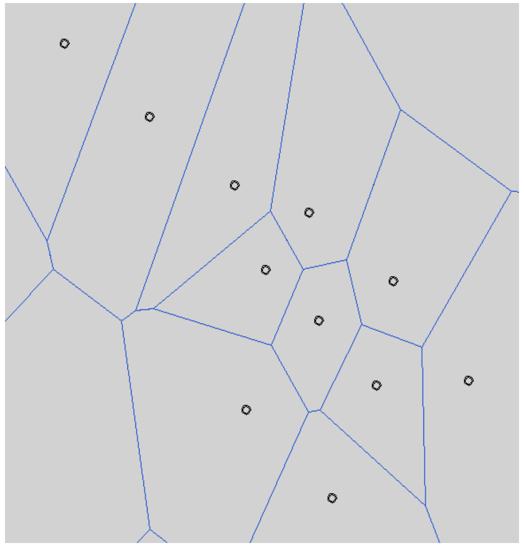
Partition of territory: spatial clusters

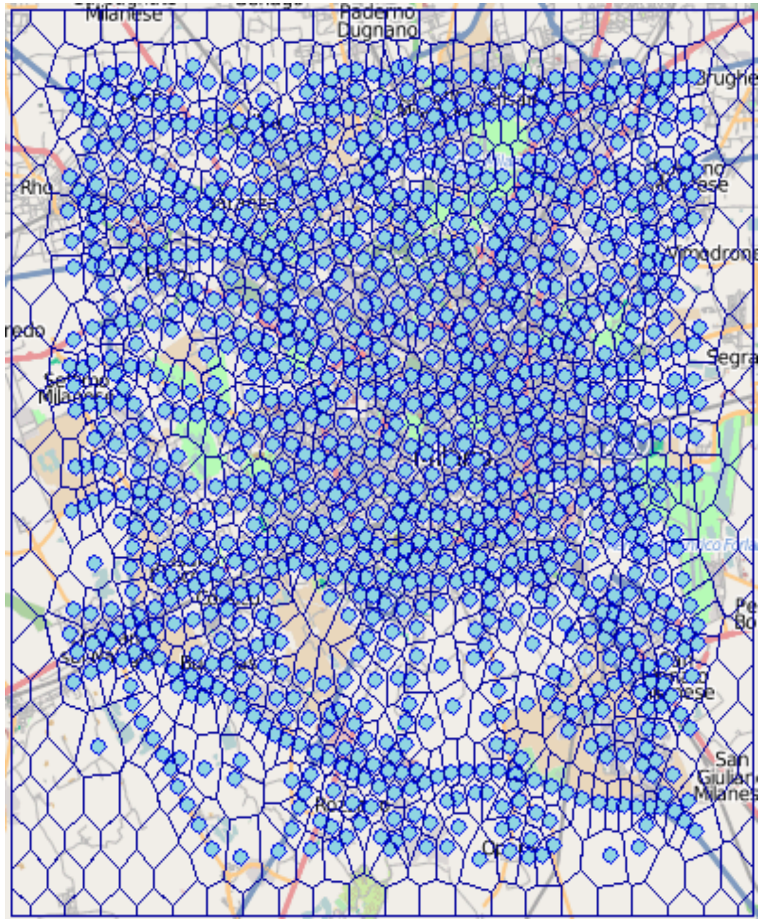
- Group the extracted points in **Spatial Clusters** with desired spatial extent
- **MaxRadius**: parameter to determine the spatial extent and so the degree of the generalization



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:

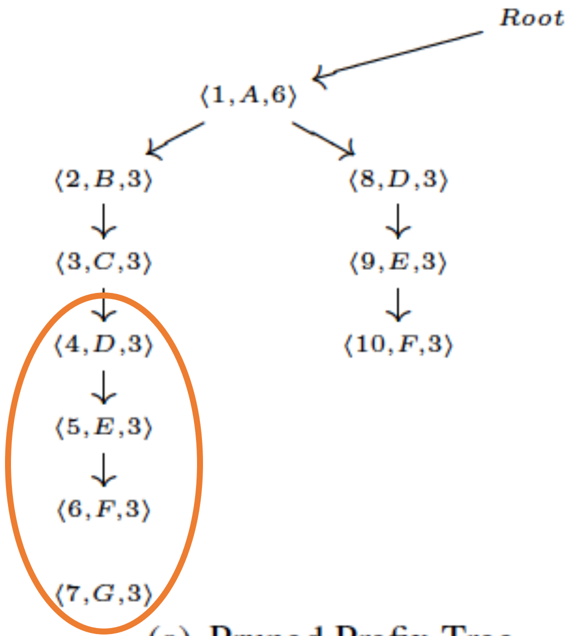
- the area a_1 containing its first point p_1 is found
- The following points are checked
- If a point p_i is not contained in a_1 for it the containing area a_2 is found
- and so on …

Generalized trajectory: From sequence of areas to sequence of centroids of areas

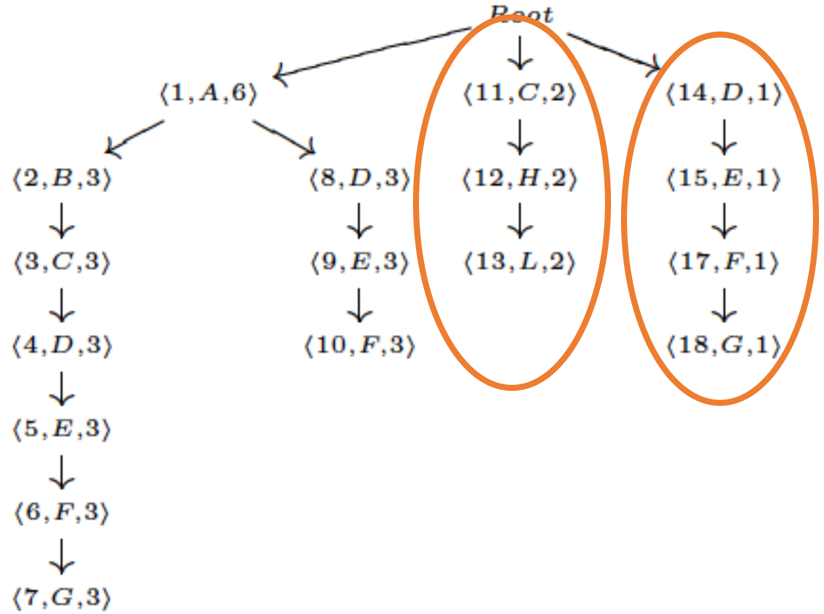
Generalization vs k-Anonymity

- 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?
- Data transformation strategy
  - recovering portions of trajectories which are frequent at least k times
  - without introducing noise

Privacy Transformation: Example



(a) Pruned Prefix Tree



(b) Anonymized Prefix Tree

$$\mathcal{L}_{cut}$$

$$(CHL, 1)$$

$$(DEJFG, 1)$$

$$(DECHL, 1)$$

Clustering on Anonymized Trajectories