Privacy ed Etica in data science

Dino Pedreschi, Fosca Giannotti
Pisa KDD Lab, ISTI-CNR & Univ. Pisa
http://www-kdd.isti.cnr.it/

Master MAINS 2018

Ethics & GDPR

Data Scientists have an obligation to take into account the ethical and legal aspects and the social impact of Data Science

SoBigData Research Infrastructure

Social Mining &
Big Data Analytics

H2020 - www.sobigdata.eu

September 2015- August 2019

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

Istituto di Informatica e Telematica del CNR

IN SUPREMÆ DIGNITATIS 1343 UNIVERSITÀ DI PISA

SCUOLA NORMALE SUPERIORE

Institutions Markets Technologies IMT INSTITUTE FOR ADVANCED STUDIES LUCCA

Ethics and Security

Legal and Ethical framework

Define and implement the legal and ethical framework of the SoBigData RI, in accordance with the European and national legislations

Monitor of research

Monitor the compliance of experiments and research protocols with the framework

Privacy-by-design

The development of big data analytics and social mining tools with Value-Sensitive Design and privacy-by-design methodologies

Big Data risks: Privacy

- Any individual has the right to privacy protection
  - The right to be **directly or indirectly non-identifiable**

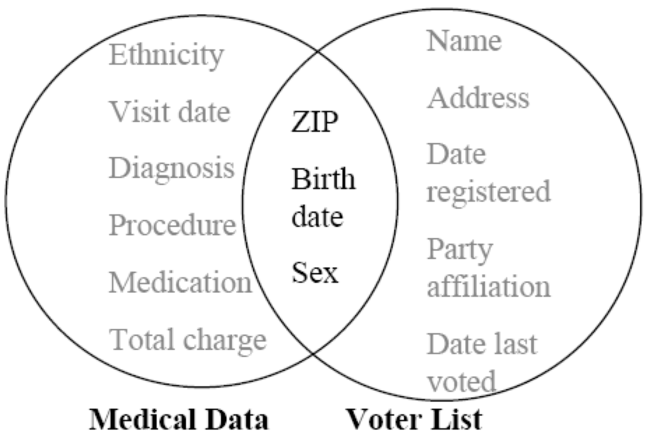
- Analyze this kind of data also combining them can bring to individual privacy violation

- The new EU Privacy Regulation requires that the data Controller maintains an updated report on the privacy risk assessment on perosnal data collected

Re-identification of Massachussetts' governor

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

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



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

The “old” European legislation for protection of personal data

- European directives:
  - Data protection directive (95/46/EC) and proposal for a new EU directive (25 Jan 2012)
  - http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm
  - ePrivacy directive (2002/58/EC) and its revision (2009/136/EC)

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.

The GDPR: the rights of the **digital persona**

- Will enter into force on 25 May 2018
- Introduces important novelties
  - New Obligations
  - New Rights


EUROPEAN DATA PROTECTION SUPERVISOR

Opinion 7/2015

Meeting the challenges of big data

A call for transparency, user control, data protection by design and accountability

EDPS

19 November 2015

Privacy-by-design

Data Protection Impact Assessment

Privacy-by-Design

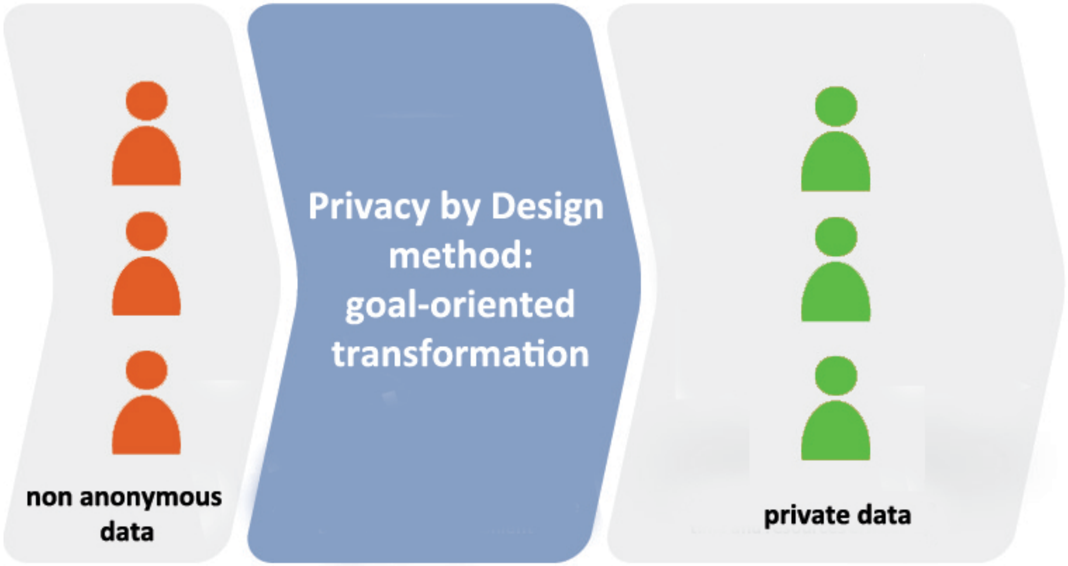
1. **Proactive** not reactive; preventative not remedial
2. Privacy as the **default** setting
3. Privacy **embedded** into design
4. **Full functionality** – positive-sum, not zero-sum
5. End-to-end security – **full lifecycle protection**
6. Visibility and **transparency** – keep it open
7. Respect for user privacy – keep it **user-centric**

Privacy-by-Design (2)

- Proactive approach
- Permits to reach a good trade-off between privacy and quality results
- Necessary assumptions; we need to define:
  1) personal data
  2) attack model
  3) analytical queries

Privacy by design big data analytics

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



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

Privacy by Design Methodology in Big Data Analytics

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

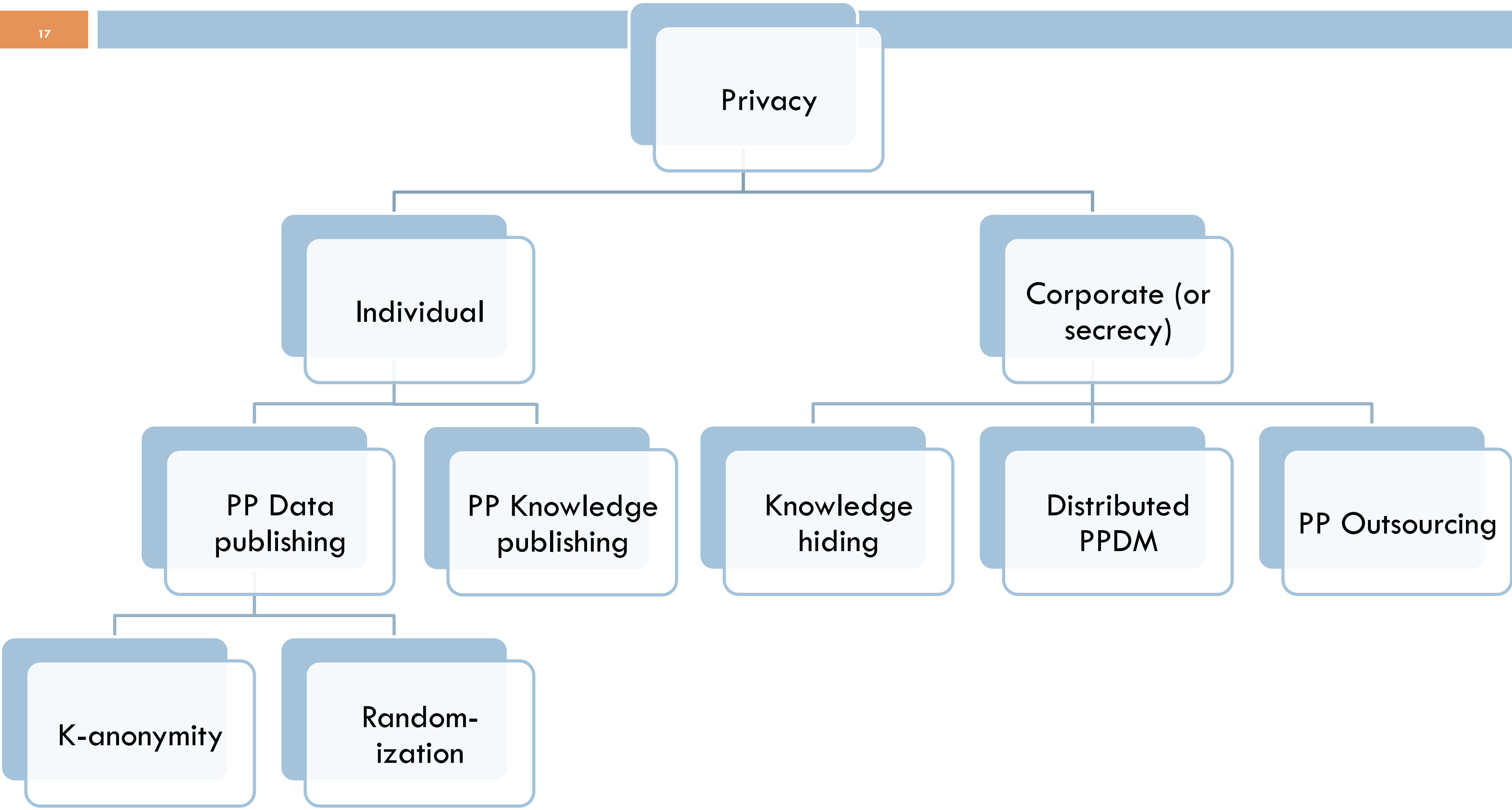
Design a privacy-preserving framework able to

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

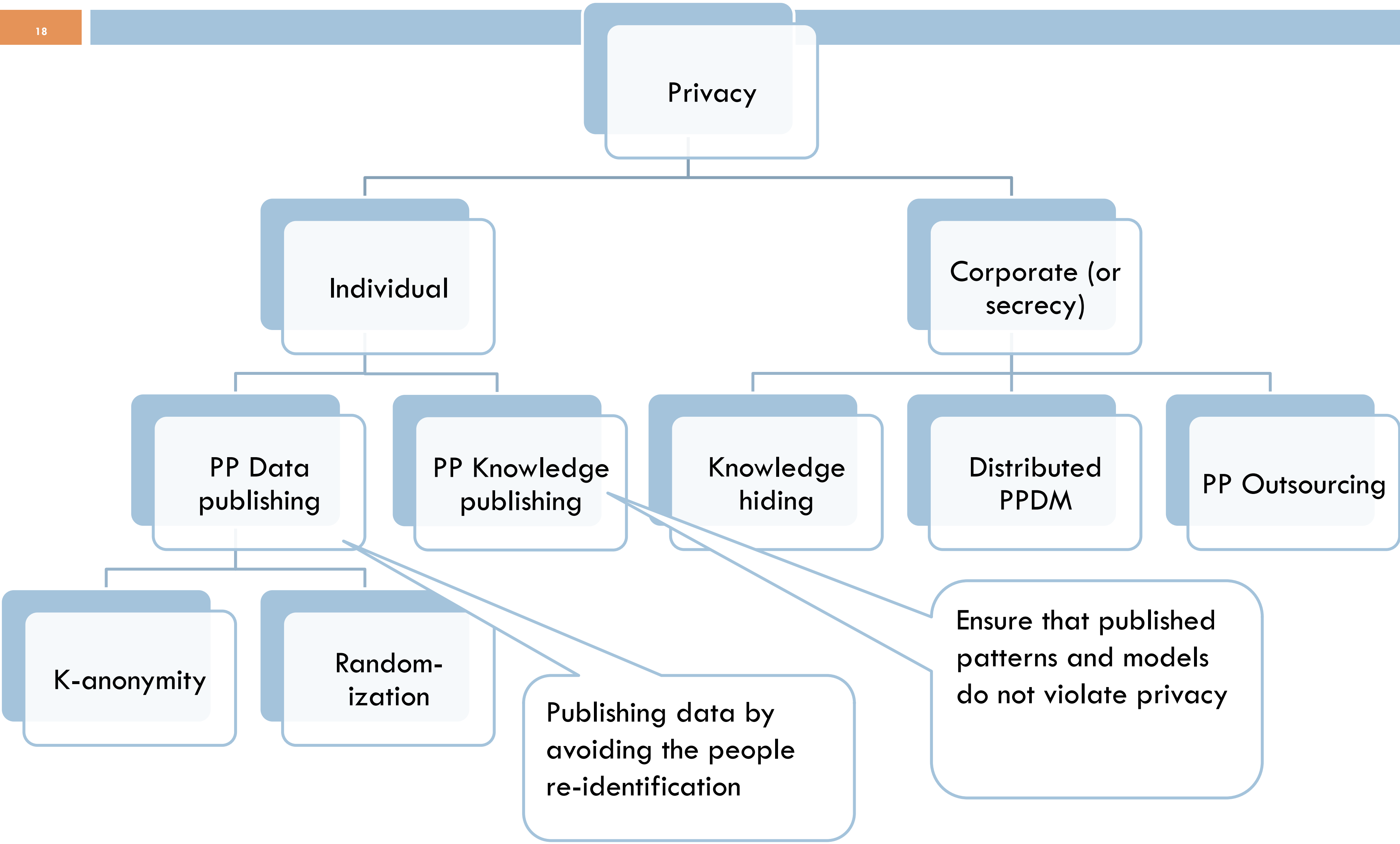
Privacy enhancing TECNOLOGIE (PET)

Short overview

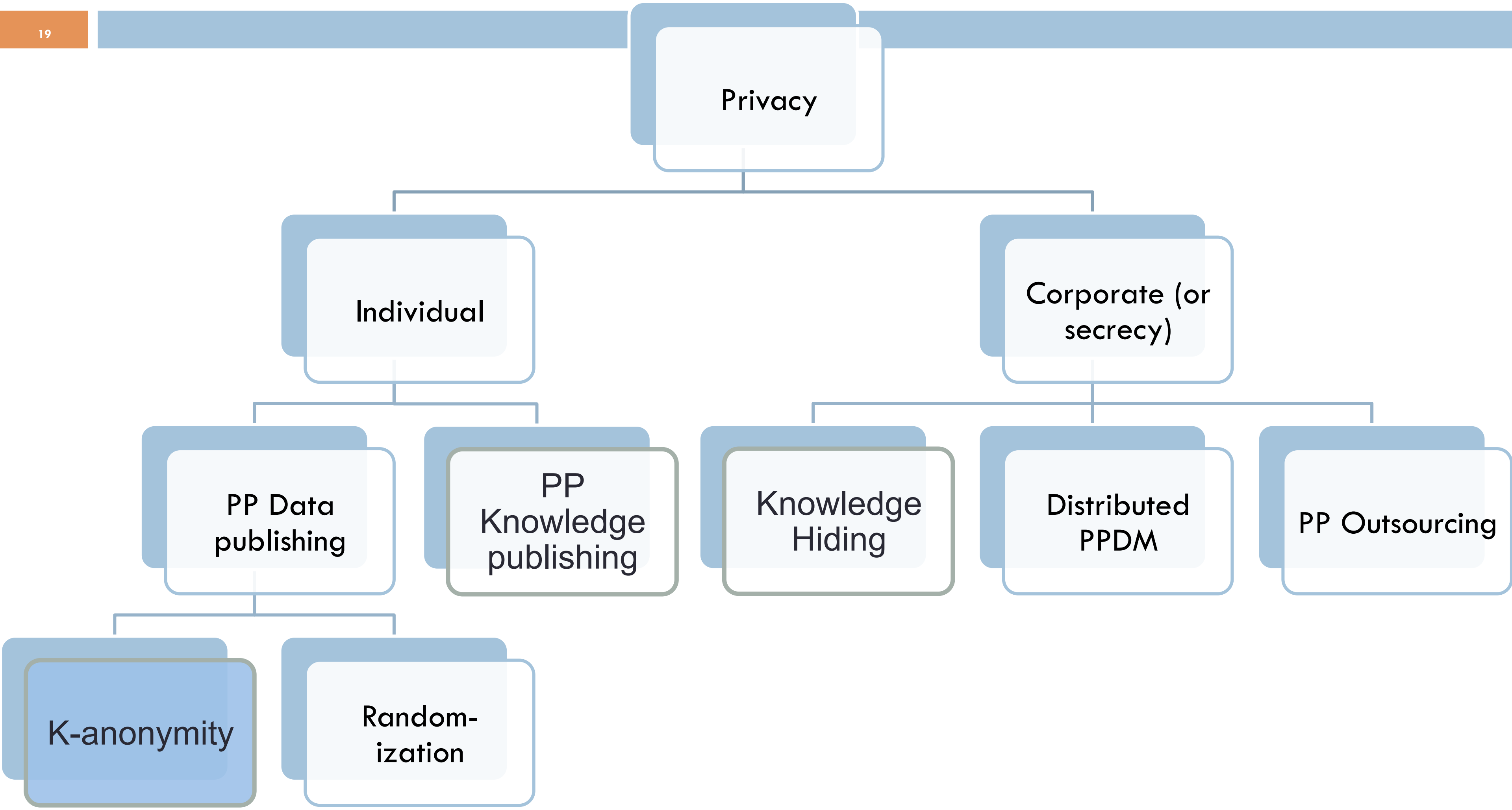
Ontology of Privacy in Data Analysis



Ontology of Privacy in Data Analysis



Ontology of Privacy in Data Analysis



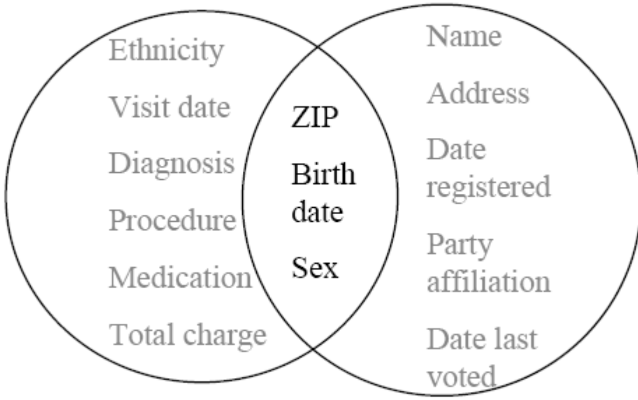
Data K-anonymity

- 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

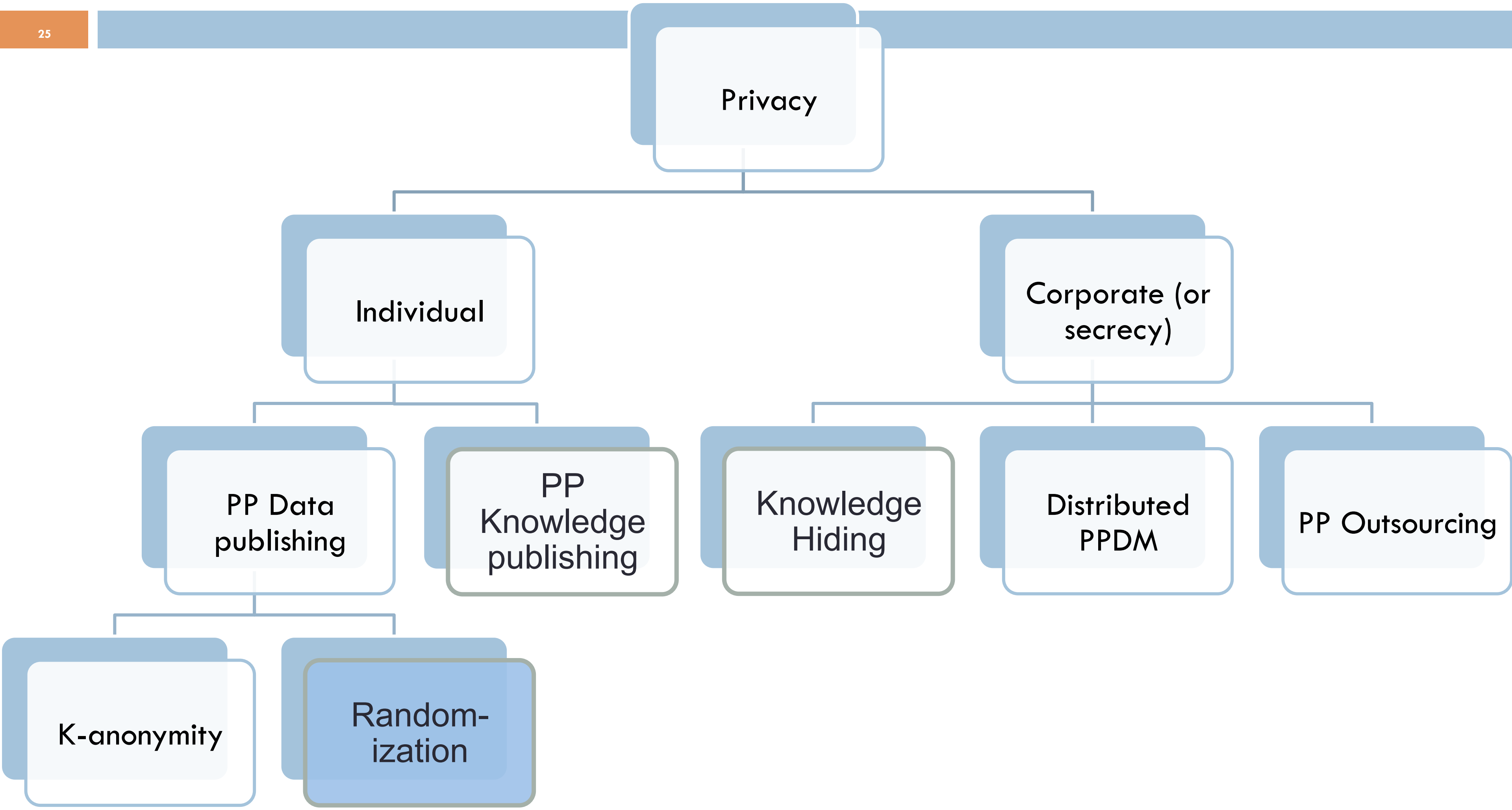
Key Attribute	Quasi-Identifier			Sensive Attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

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*}

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

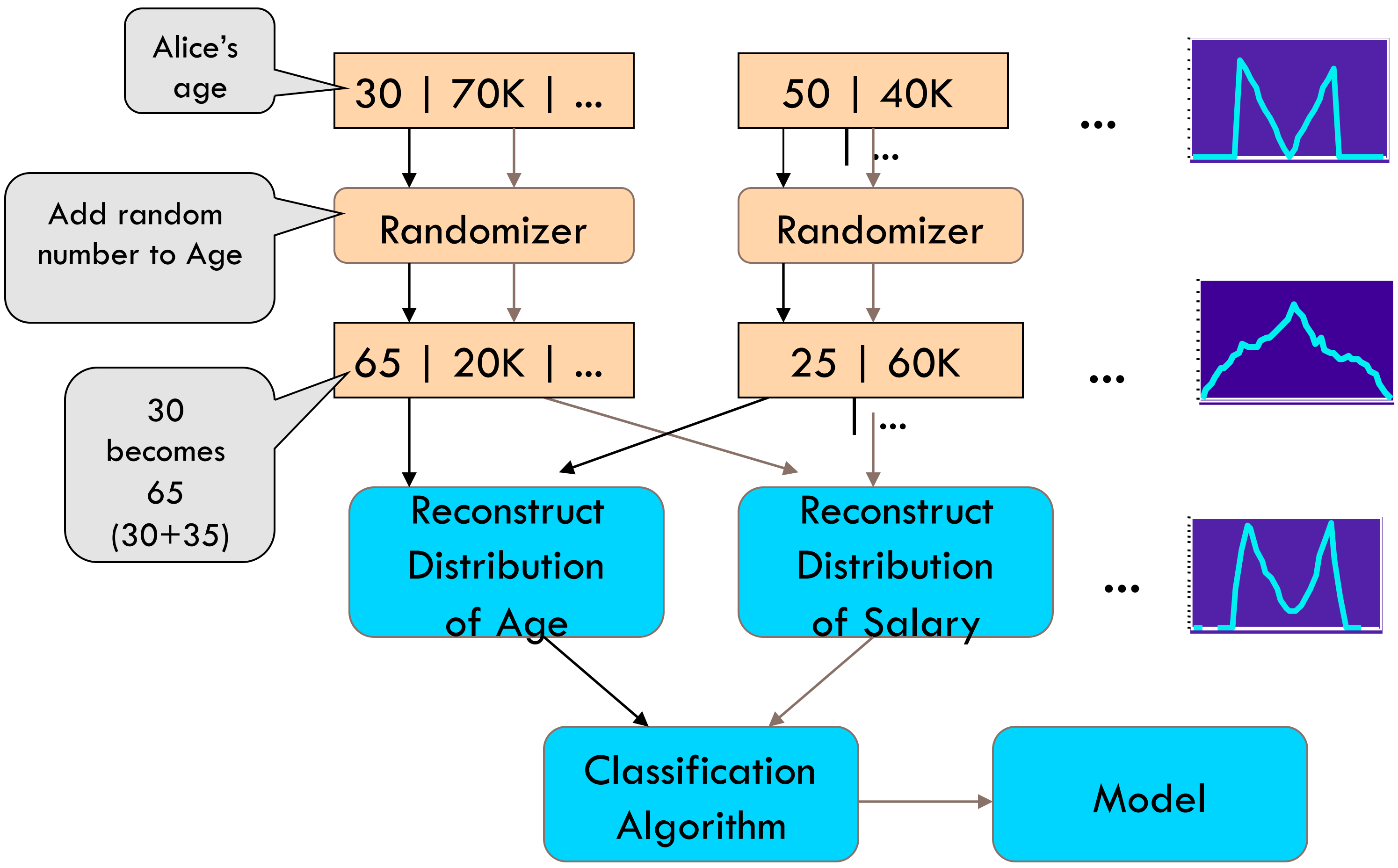
- **Original values $x_1, x_2, \ldots, x_n$**
  - from probability distribution X (unknown)
- **To hide these values, we use $y_1, y_2, \ldots, y_n$**
  - from probability distribution Y
    - Uniform distribution between $[-\alpha, \alpha]$
    - Gaussian, normal distribution with $\mu = 0, \sigma$

- 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.

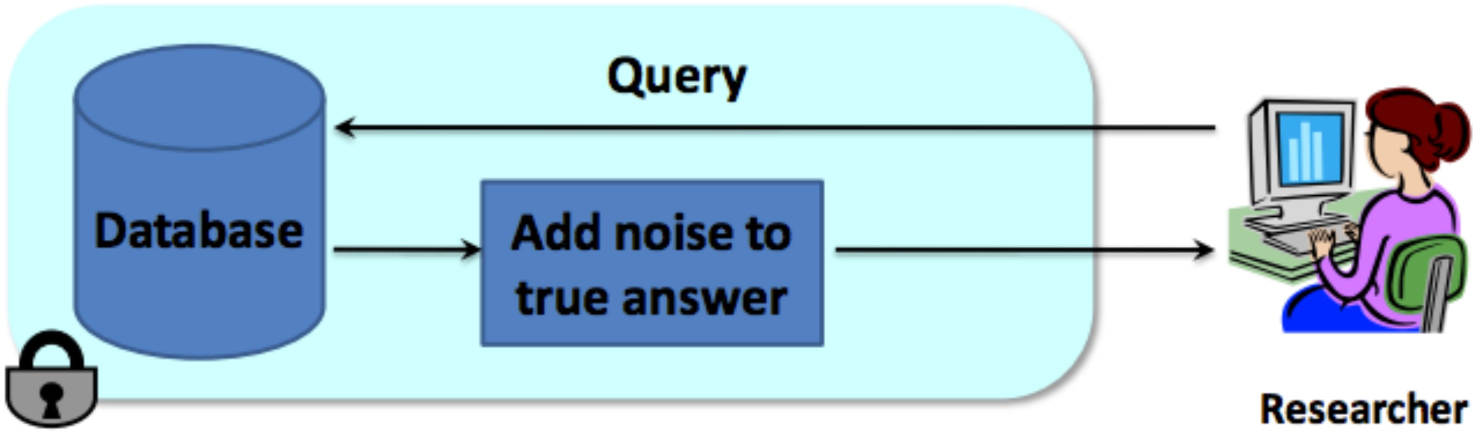
R. Agrawal and R. Srikant. Privacy-preserving data mining. In Proceedings of SIGMOD 2000.

Randomization Approach Overview



Differential Privacy

- **Goal:** 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

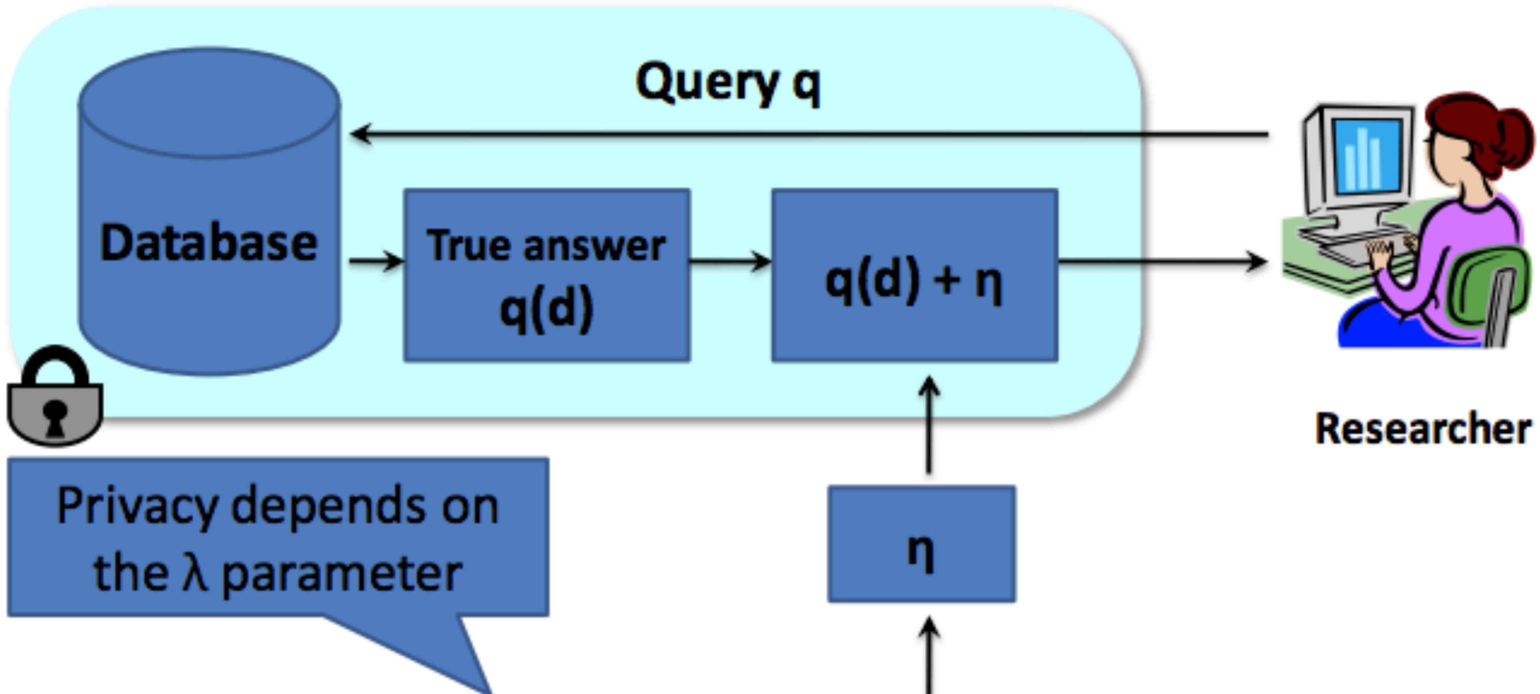
1) how many persons have Diabetes? **4**
2) how many persons, excluding Alice, have Diabetes? **3**

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

- Solution: make the two answer similar

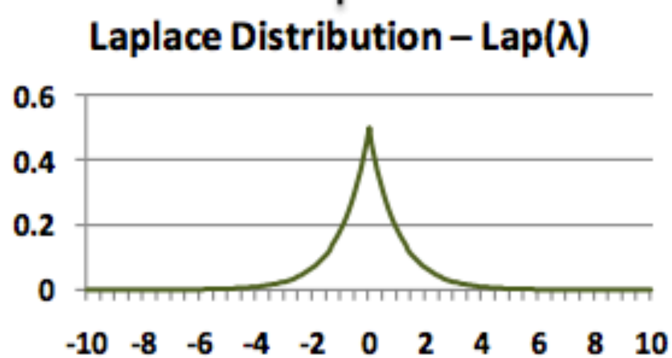
1) the answer of the first query could be 4+1 = 5
2) the answer of the second query could be 3+2.5=5.5

Differential Privacy



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

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



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

..summarizing

What is mainly done

Sanitisation
Encryption or removal of sensitive information

Anonymisation
Encryption or removal of personally identifiable information

Multi-party computation
Distribution of data and processing tasks over multiple parties

Encryption
Encoding of information so that only authorised parties can access it

Policy enforcement
Enforcement of rules for the use and handling of resources

Access control
Selective restriction of access to places or resources

What we need

Accountability
Evaluation of compliance with policies and provision of evidence

Transparency
Explication of information collection and processing

What is coming up

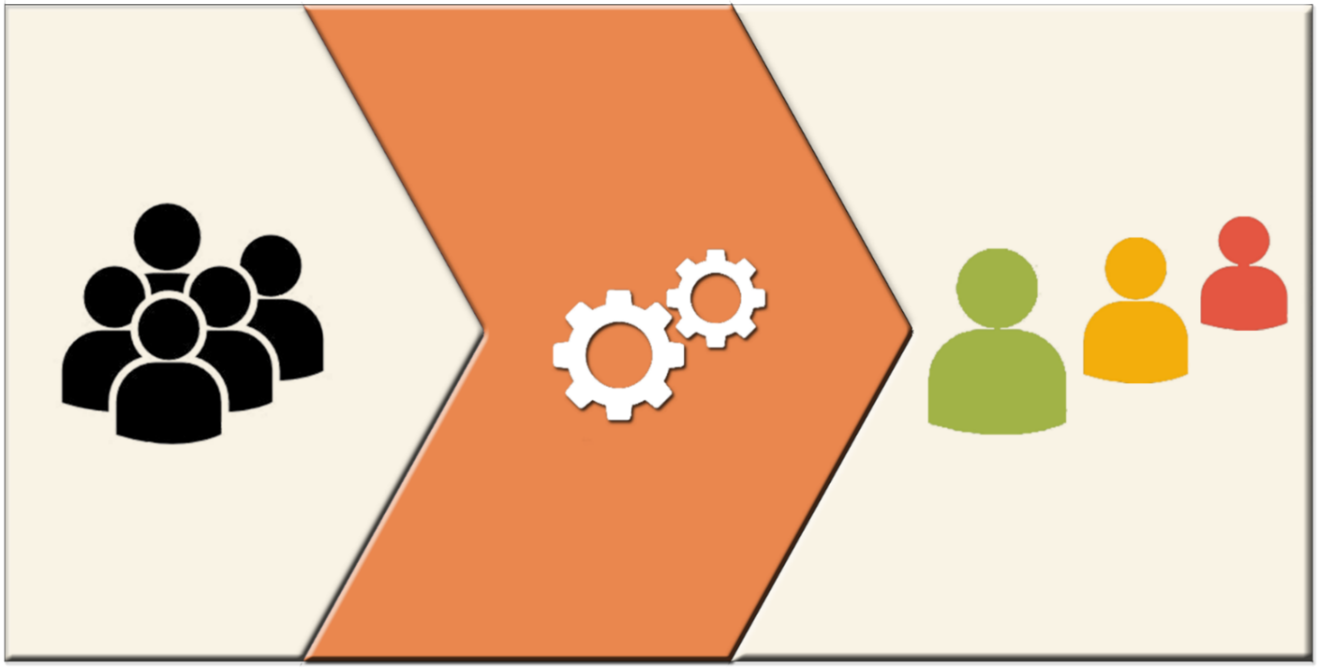
Data provenance
Attesting of the origin and authenticity of information

Access and portability
Facilitating the use and handling of data in different contexts

User control
Specification and enforcement of rules for data use and handling

Privacy Risk Assessment

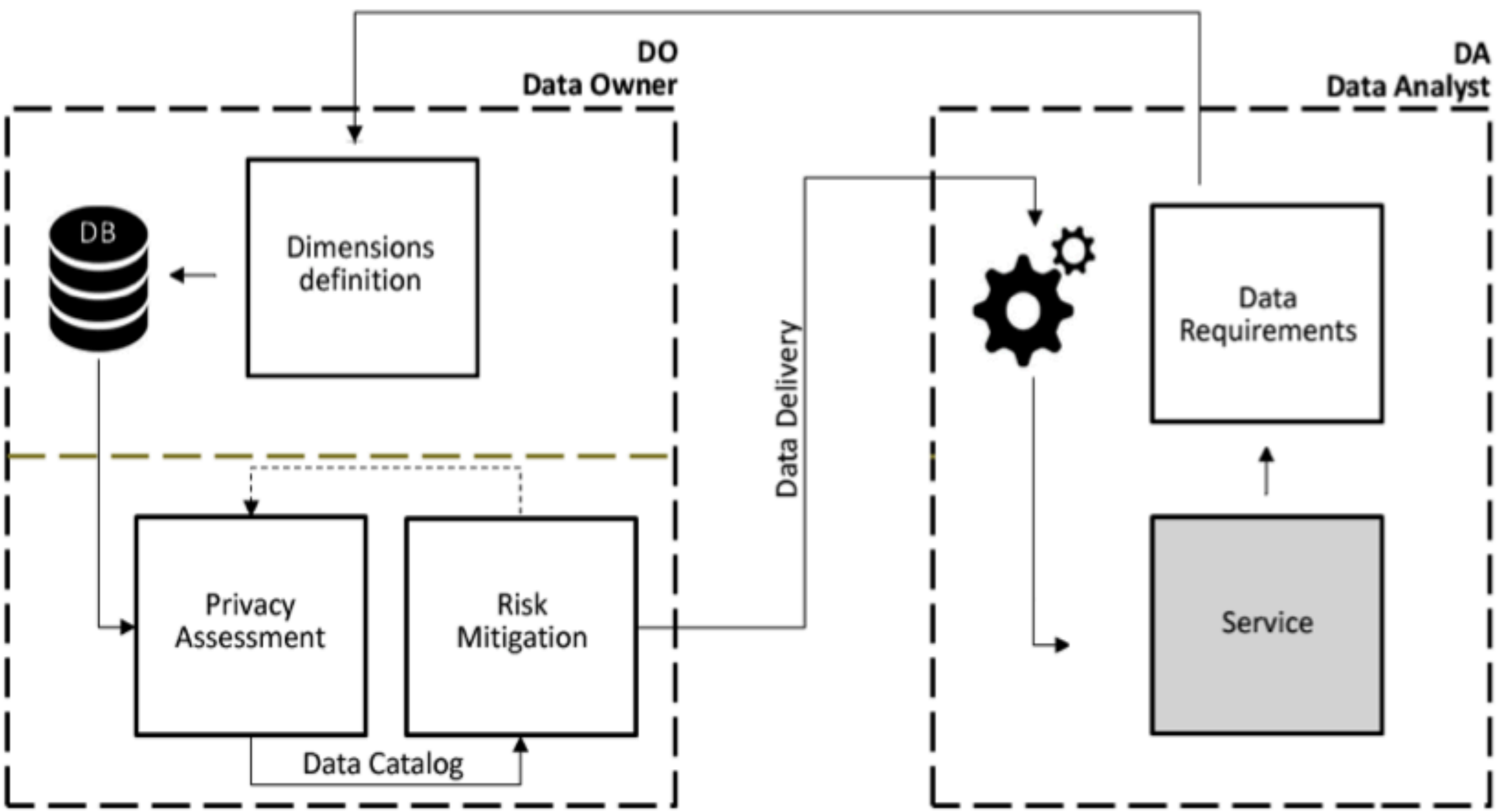
Privacy Risk Assessment



What a risk is?

- Risk is the chance (understood as a probabilistic notion) that a danger (i.e., an event with harmful consequences) will happen
- Or (more technically) :
- Risk is an **objective measurable entity** combining the probability of an adverse event and the magnitude of its consequences.

Privacy Risk Assessment Framework



PRIVACY-AWARE FRAMEWORK FOR DATA SHARING

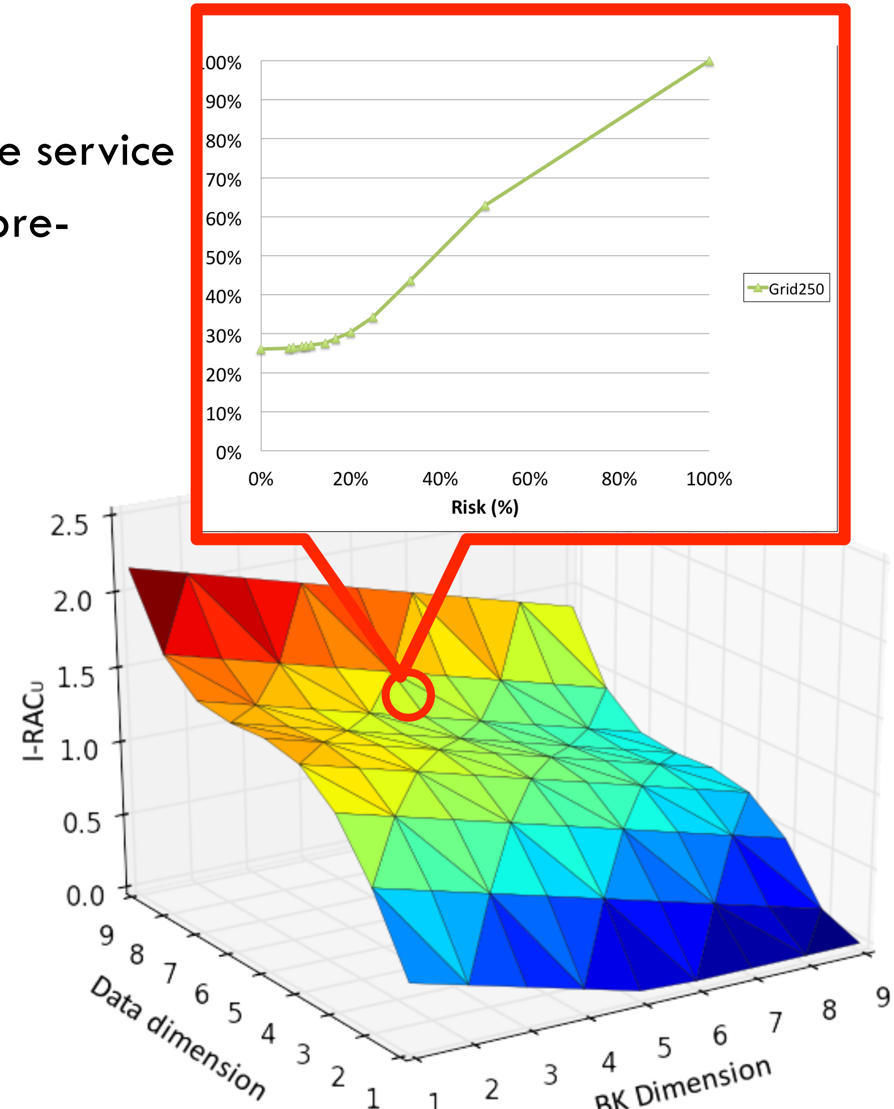
Data Catalog

For each:

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

The Data Catalog provides:

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

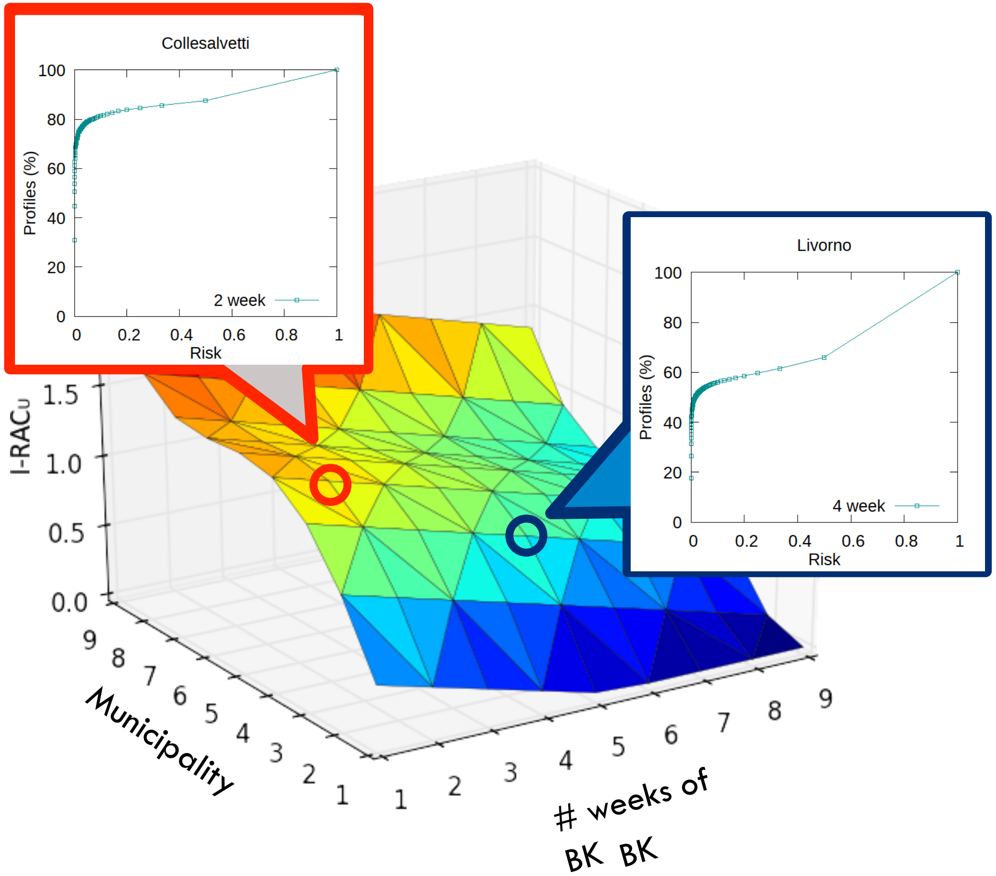


Simulation of privacy harmful Inferences

Data dimension:
The spatial area in which the analysis is performed.

Background Knowledge dimension:
The temporal window (in weeks) in which the attacker recorded the user activity.

I-RACu:
An indicator of the risk of re-identification of the users



PRIVACY RISKS IN IoT

Risk and countermeasures

Why is it important to quantify risk?

- Beyond the GDPR, we must consider what happens when we **share** some data
- An example in IoT is to collect and share data about sensors

Sensors are very useful

- Sensors can help us managing our personal life:
  - Checking our messages
  - Saving statistics about our training sessions
  - Monitoring our sleep
  - Checking the food contained in our fridge
  - Switching on/off our heating system
  - […]

But the related data are very personal!

- Sensors data might reveal information about:
  - Habits
    - Movements
    - Personal tastes
    - Social networks
  - Health status

We want to «anonymize» these data

- We want to limit the quantity/quality of information we share
- Several ways, which are strongly related to the service we want to obtain
  - For example, a training optimization app do not need the precise information about location were we are
  - And probably also the time can be (slightly) shifted

Just a quick recap

- We have some standard techniques:
  - Generalization
  - Suppression
  - Randomization
- We have (at least) two dimensions that can be explored:
  - Time
  - Space
- The key point is to study which is the **minimum information** needed (data minimization principle)

How we can discover if we are safe?

We can:

- quantify the privacy risk we have sharing the original data
- apply one (or more) of the previous techniques and
- then, quantify the new privacy risk

How can we quantify the risk?

- We need to find a possible measure
- For example the risk of re-identification
  - Which is the probability to correctly associate a record to a single individual?
- Another possible privacy risk is the risk of inference
  - What could an attacker discover about his/her target?

Example of risk of re-identification

An attacker (Alice) gain access to a camera survelling system

Knowing that her target (Bob) wears a black&white striped

Example of risk of re-identification

Four persons wear a striped sweater, so Alice can say that Bob is one of those ones → **probability of re-identification=1/4**

The Amnesia Tool

- Visit the website: https://amnesia.openaire.eu/
- Let’s see a demo

Some practical examples

How is it possible to define services GDPR compliant?

Services that need for GPS data

- Parking Assistance
- Geolocalized Marketing Advices
- Traffic jam analysis and prevention
- Navigation systems development
- Route/destination prediction
- Selection of the best location where to open a new facility
  - franchise store
  - fuel station
  - shopping mall
- […]

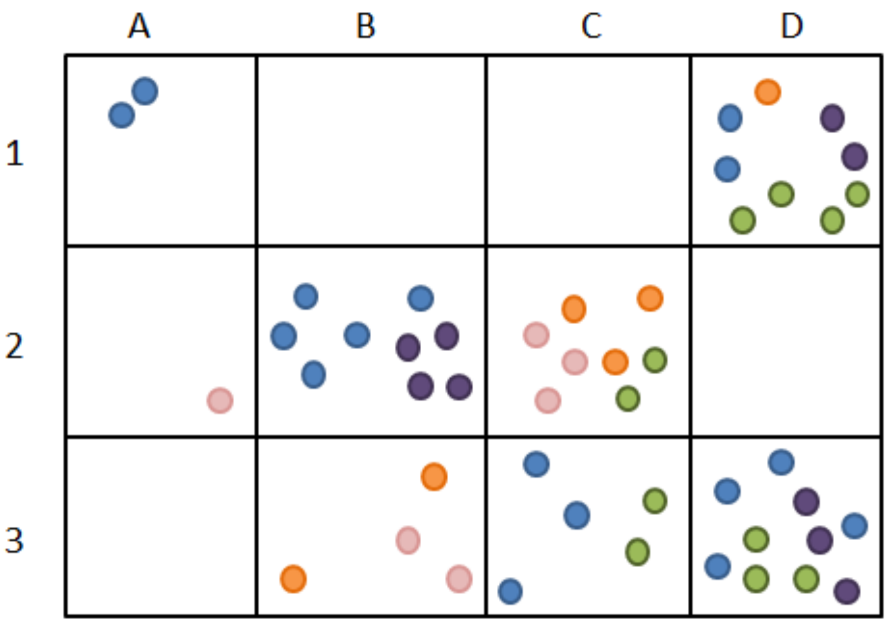
Example1: Individual Presences

- Possible services:
  - Developing Parking Assistance
  - Geolocalized Marketing Advices
- These services do not need for all individual trajectories: specific movements are not necessary
- The only information needed is the last position of an individual (and maybe the time)

Data description

- **For each user, list of locations (grid cells) that the user has frequently visited (#visit>threshold)**

User id Cell id



Blue: <B2,5>,<D3,4>,<C3,3>,<A1,2>,<D1,2>
Green: <D1,4>,<D3,3>,<C2,2>,<C3,2>
Orange: <C2,3>,<B3,2>
Purple: <B2,4>,<D3,3>,<D1,2>
Pink: <C2,3>,<B3,2>

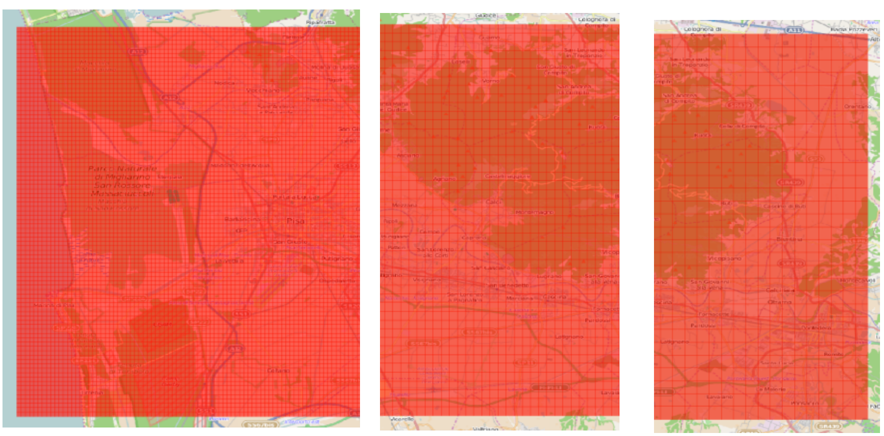
Data Dimensions

Grid size: defines the granularity of the spatial information released about each user

Frequency threshold: defines a filter on the data DO can distribute

Spatial granularity used:

Grids (cell side): 250, 500 and 750 meters



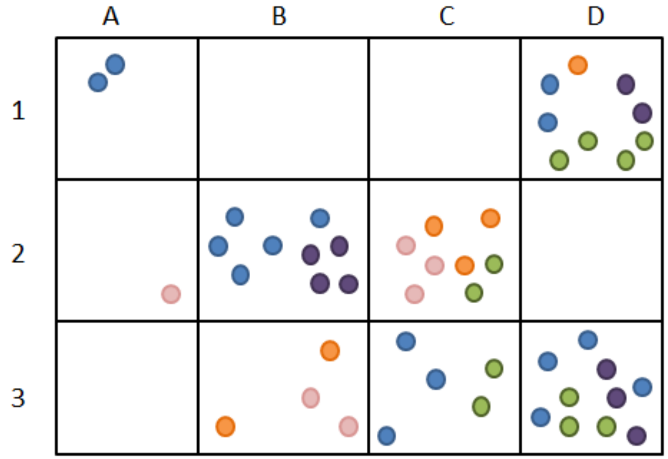
Frequency threshold: 1, 4, 7, 10, 13

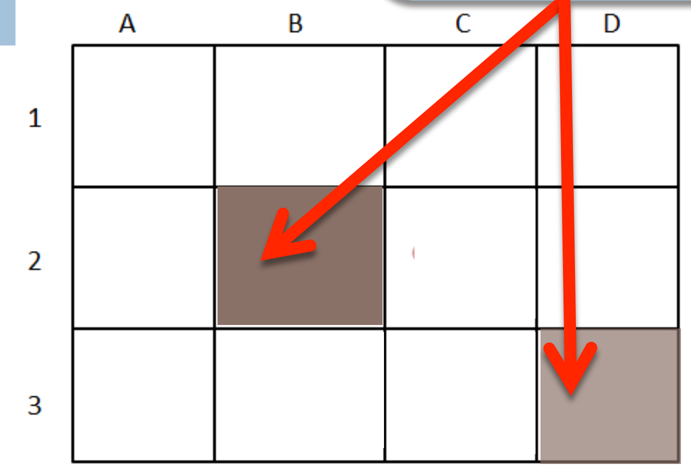
Possible Attacks

- We need to define which is a **reasonable background knowledge** that an attacker can have
- This is compliant with the Privacy-by-Design paradigm
- We analyzed different levels of background knowlege

Attack 1: Top-k places

Background Knowledge:
Top-k places





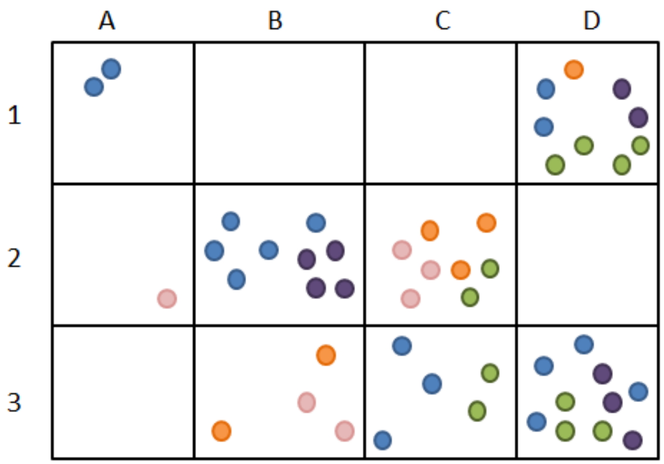
The attacker knows the first k location(s) of his target

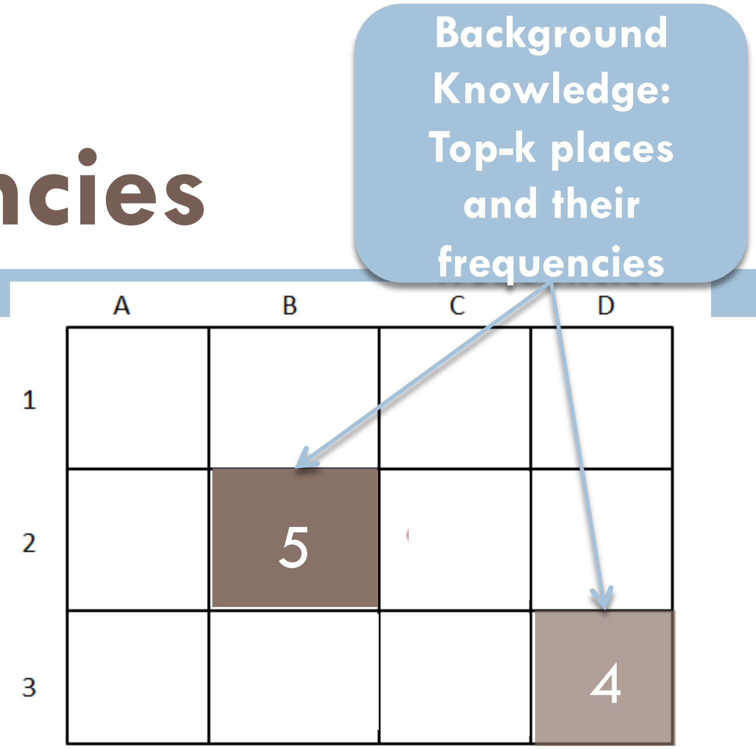
Background Knowledge Dimension:

- Number of locations known (h = 1, 2, 3)

E.g., Mr. Smith lives in B2 and works in D3

Attack 2: Add frequencies





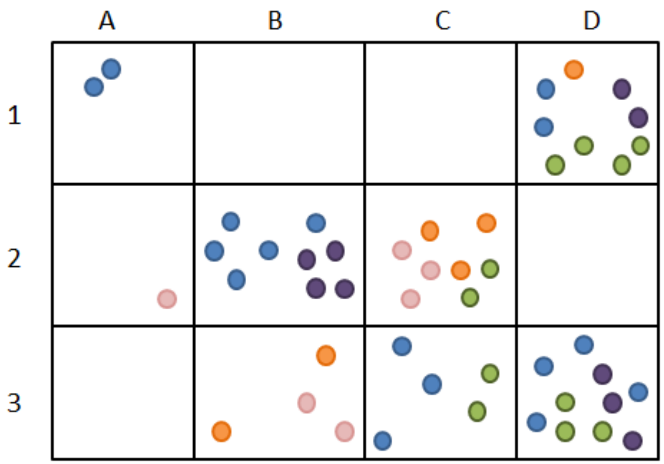
The attacker knows the first k location(s) of his target, and also the exact frequency

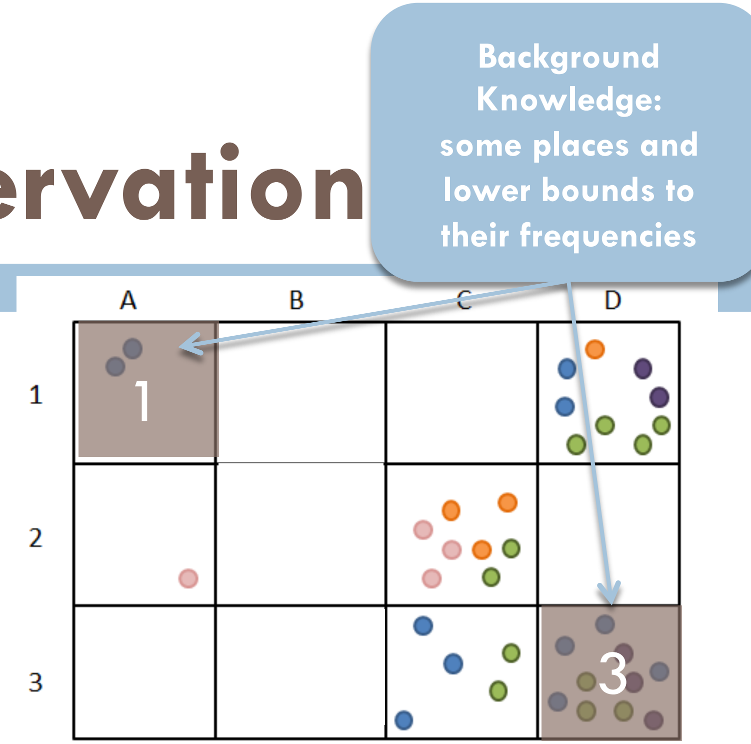
Background Knowledge Dimension:

- Number of locations known (h = 1, 2, 3)

E.g., Mr. Smith lives in B2 (and he parked there 5 times) and works in D3 (and he went to work 4 times)

Attack 3: Casual observation





The attacker knows some location(s) with minimum frequencies

Background Knowledge Dimensions:

- Number of locations known (h = 1, 2, 3)
- Minimum frequency associate to the known locations (100% of original freq, 50% of original freq, only presence)

E.g., Mr. Smith was seen once in A1 and 3 times in D3

Simulation of Attack

- We simulate the chosen attack (or all of them)
- At the end we obtain a list of individuals with their own probability of re-identification

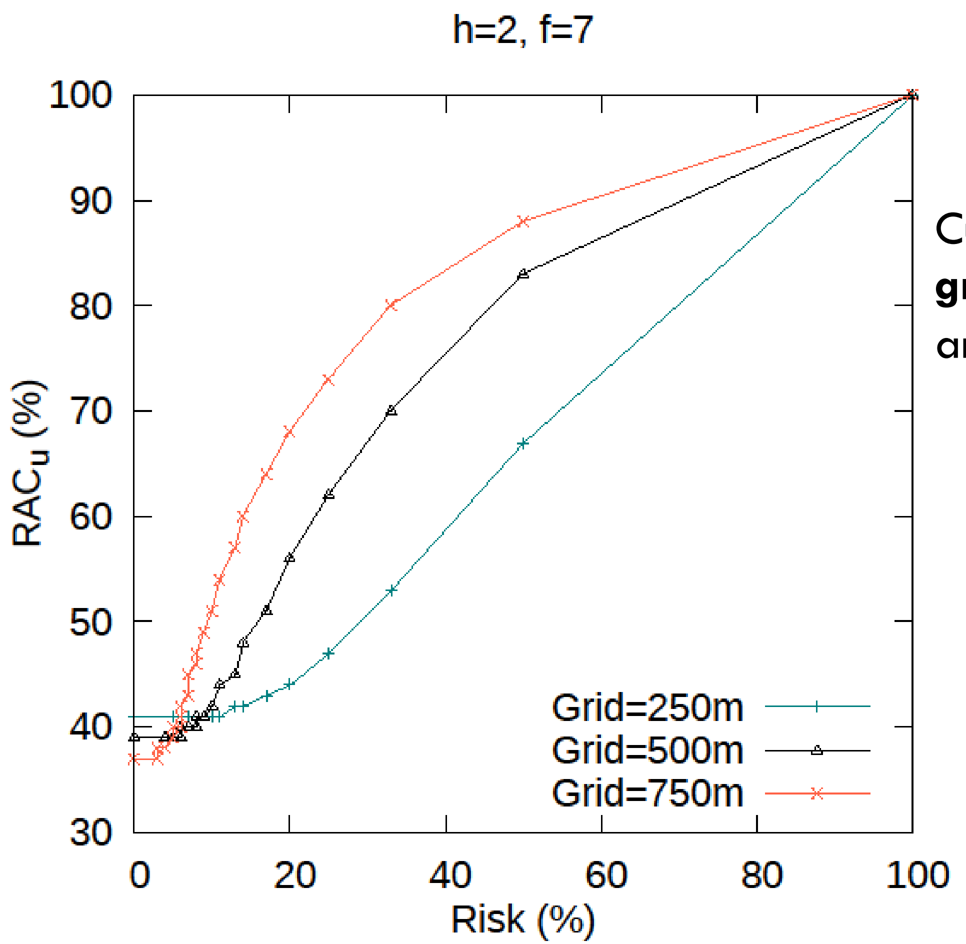
Pseudo ID	Probability
100	1/3
101	1/10
102	1/50
203	1/30
205	1/25
…	…
452	1/30

What next?

- Having in mind a privacy threshold (e.g., 1/20)
- We see that many of our individual are already safe
- We can act (a… … n the other ones (e.g., 100&10…

Pseudo ID	Probability
100	1/3
101	1/10
102	1/50
203	1/30
205	1/25
…	…
452	1/30

Real Experimental Results (Attack 3)



Cumulative risk varying the **grid** and fixing #location and frequency

Example2: Individual movements

- Possible services:
  - Traffic jam analysis and prevention
  - Navigation systems development
  - Route/destination prediction
- Now the movements are fundamental
- But we still can generalize position and time

Data Description

- Vehicle ID → replaced with pseudo-ID
- Location → replaced with link-ID (the street) and potentially generalized
- Time → generalized in time windows

Dimensions

- ## Data Dimensions
  - **temporal dimension** that defines the temporal granularity of the times associate to each link; e.g. approximation to 30min, 60min, 5hours

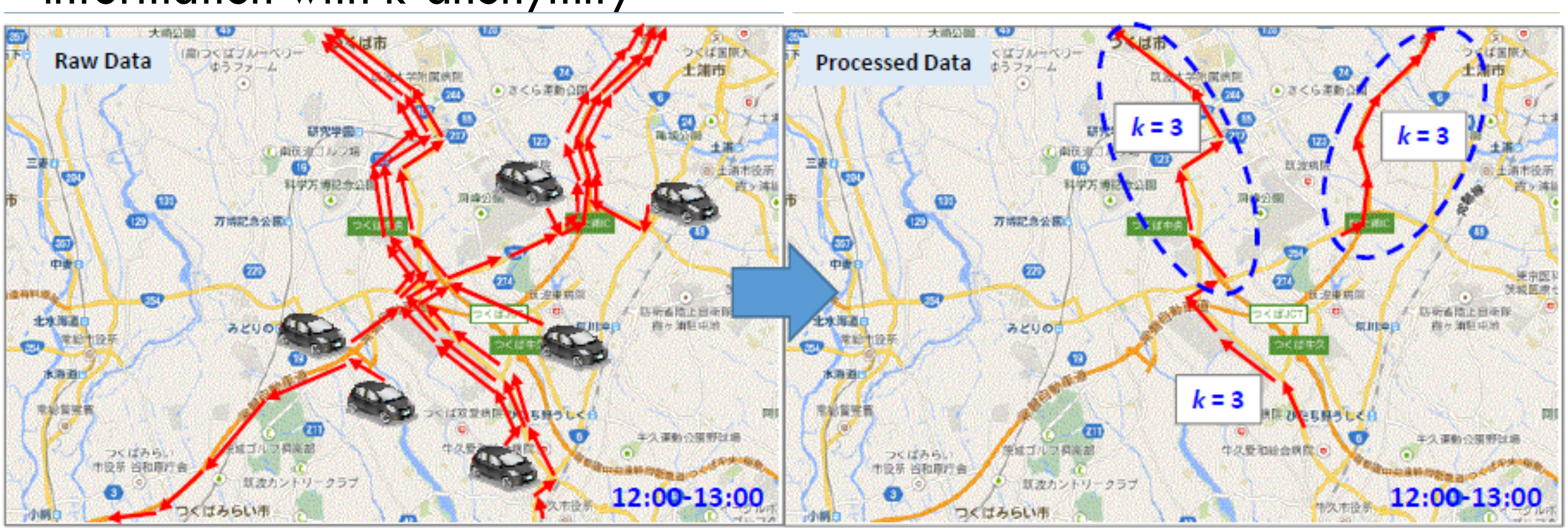
- ## BK Dimension
  - Number of links crossed by the user known by the attacker; we used h=1,2,the whole trajectory.

Analysis at collective level

- Possible services:
  - Selection of the best location where to open a new facility
    - franchise store
    - fuel station
    - shopping mall
- Fort he deployment of these services we do not need individual movements, but we can only analyze movements that are frequent at **collective level**
  - E.g., to establish if (x,y) is a good position for a fuel station, we can analyze how many vehicles usually travel in the nearby

Data description

- Areas that have many vehicles in a specific time period (non personal if we remove origin/destination and anonymize trips)
- Purpose of movements (work, leisure, …) (non personal)
- Shopping Malls will probably need demography of age, gender and family distribution. This information can be transforment to non-personal information with k-anonymity



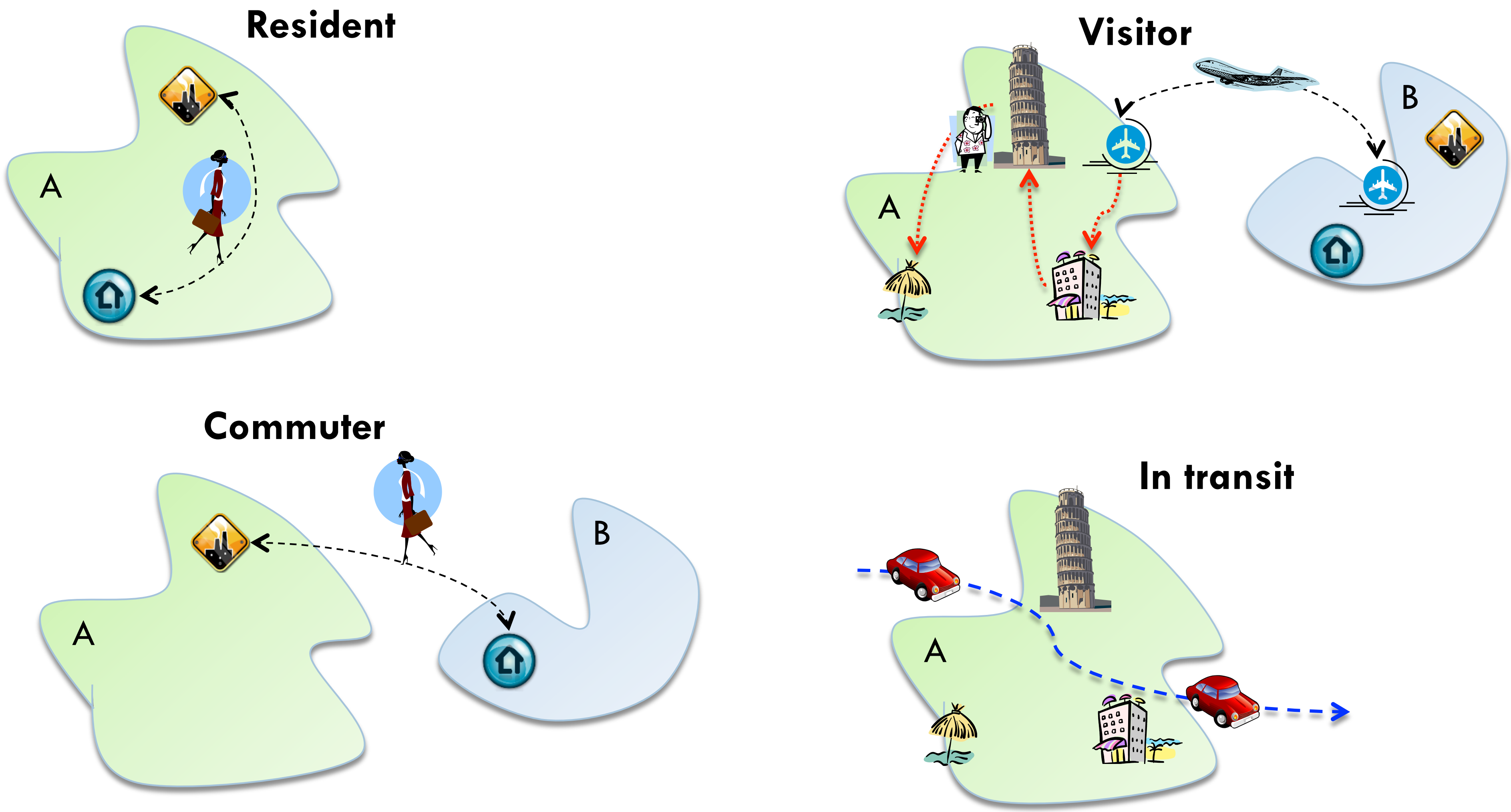
Dimensions

- Data Dimensions
  - **temporal dimension** that defines the temporal granularity of the time-window; e.g. 1 hour, 4 hour ….
  - **frequency threshold** that defines a filter on the links to be distributed
  - **spatial tolerance** of the clustering that affects the clusters composition (eps).
- BK Dimension
  - It is fixed!

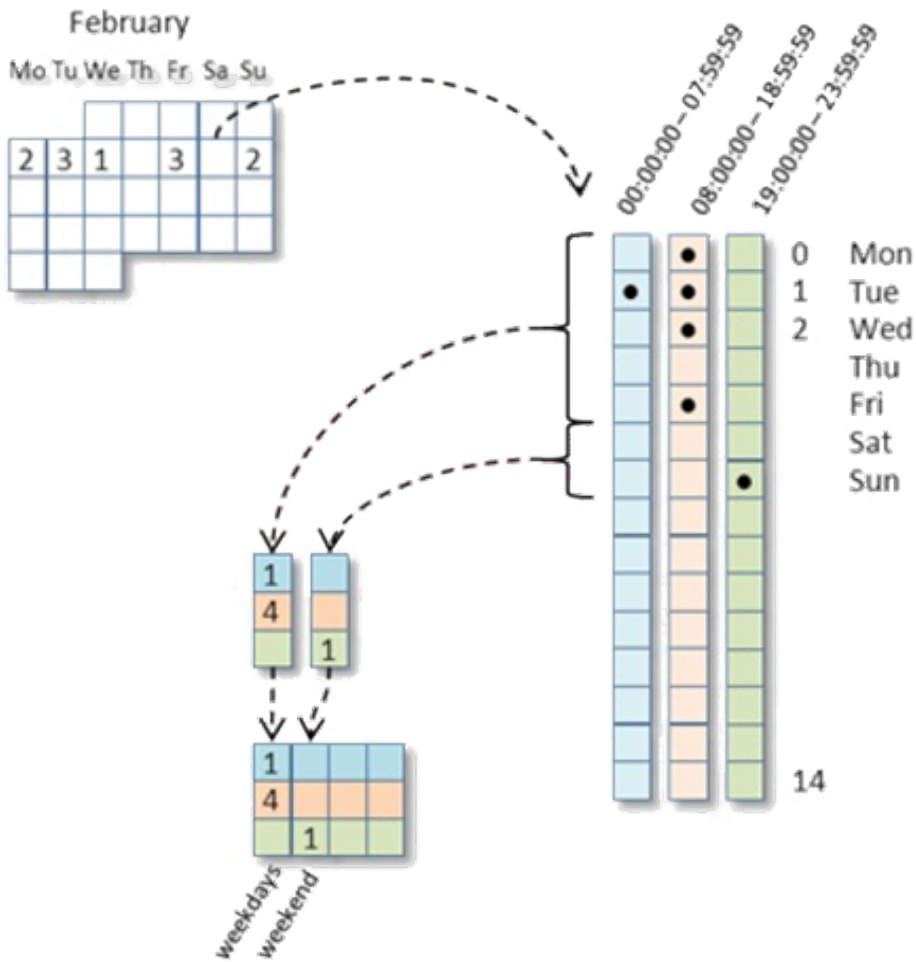
Services that need for mobile phone data

- Mobile phone data are very pervasive
  - In 2014, nearly 60% of the population worldwide already owned a mobile phone
  - Mobile phone penetration is forecasted to reach 67% by 2019
- Mobile phone data offer many new opportunities
  - Estimating presence in real-time
  - Peak detection & Event detection
  - Quantification of individuals based on their phone activity
  - Social Mining Analyses
  - […]

Sociometer: Estimating User Category



Sociometer: Data Definition



Attack based on Call Activities

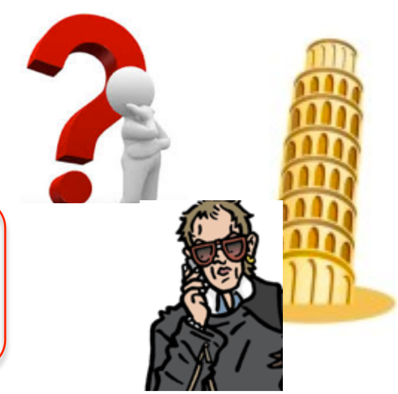
Analyst working on mobile phone data with access to their call profiles

From: 02/11/15

To: 22/11/15

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

Inference:
his activities in Pisa during the remaining week



From: 23/11/15

To: 29/11/15

Assumption: the attacker knows the user is one of the profiles

Example of the attack

Attacker knows *exactly* the call made by U in the first 3 weeks

	week 1		week 2		week 3		week 4	
morning	1						?	?
afternoon		2			1	1	?	?
evening	1		3	1	2		?	?

Example of the attack

Attacker knows *exactly* the call made by U in the first 3 weeks

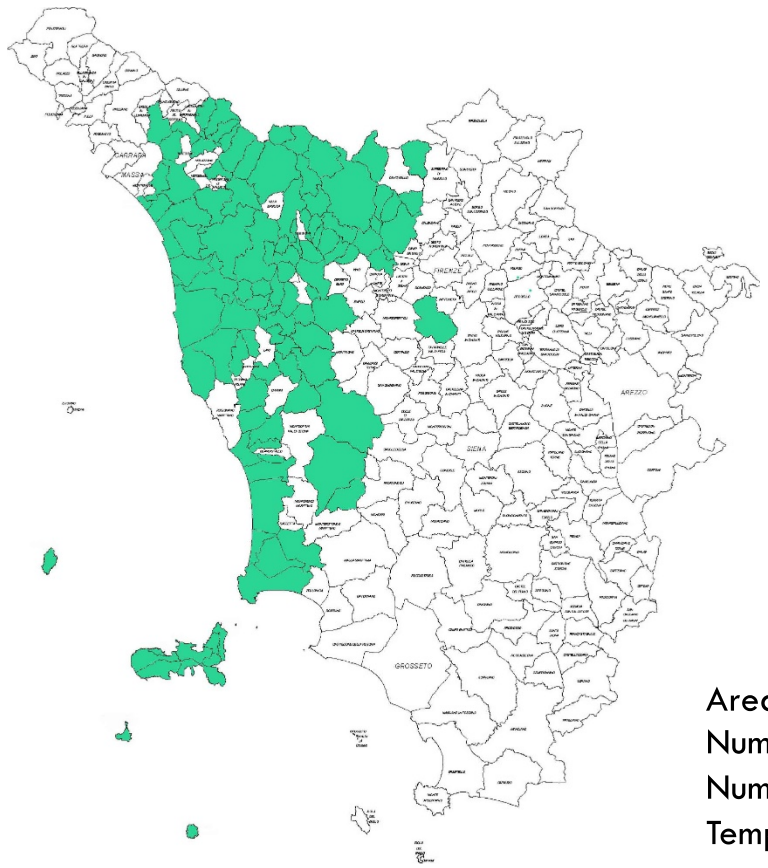
	week 1		week 2		week 3		week 4	
morning	1						?	?
afternoon		2			1	1	?	?
evening	1		3	1	2		?	?

1							
	2			1	1	2	
1		3	1	2			1

1							2
	2			1	1		
1		3	1	2		3	

K=2

Data Statistics



Area Covered: 106 municipalities out of 276
Number of calls: *51millons*
Number of active users: *181k*
Temporal window: 17/2/2014 – 23/3/2014

What next?

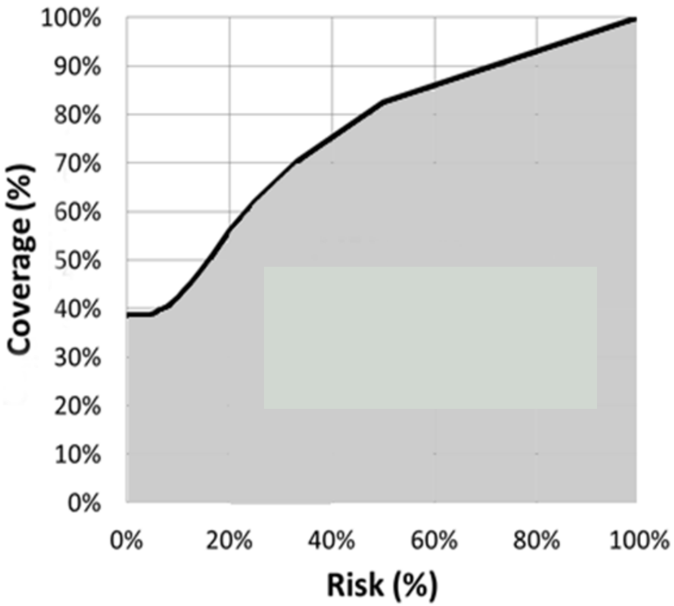
- Having in mind a privacy threshold (e.g., 1/20)
- We see that many of our individual are already safe
- We can act (a[…] n the other (e.g., 100&101)

Pseudo ID	Probability
100	1/3
101	1/10
102	1/50
203	1/30
205	1/25
…	…
452	1/30

An aggregated output

- An aggregated visualization can be useful to have a **global vision** of a dataset
- Two possible (equivalent) outputs

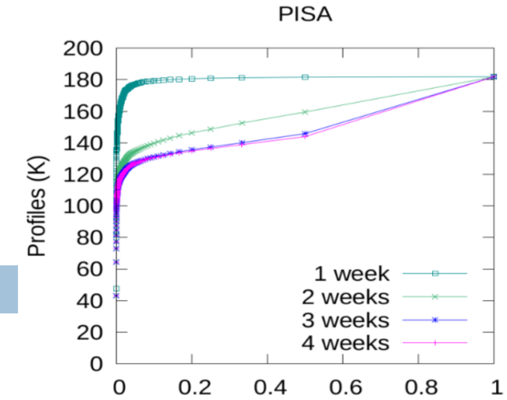
A diagram of coverage



A table

Risk (r)	% users
r<=1/20=0.05=5%	40
r<=1/5=0.2=20%	58
r<=1/3=0.33=33%	70
r<=1/2=0.5=50%	82
r<=1=100%	100

Real Experiments Results



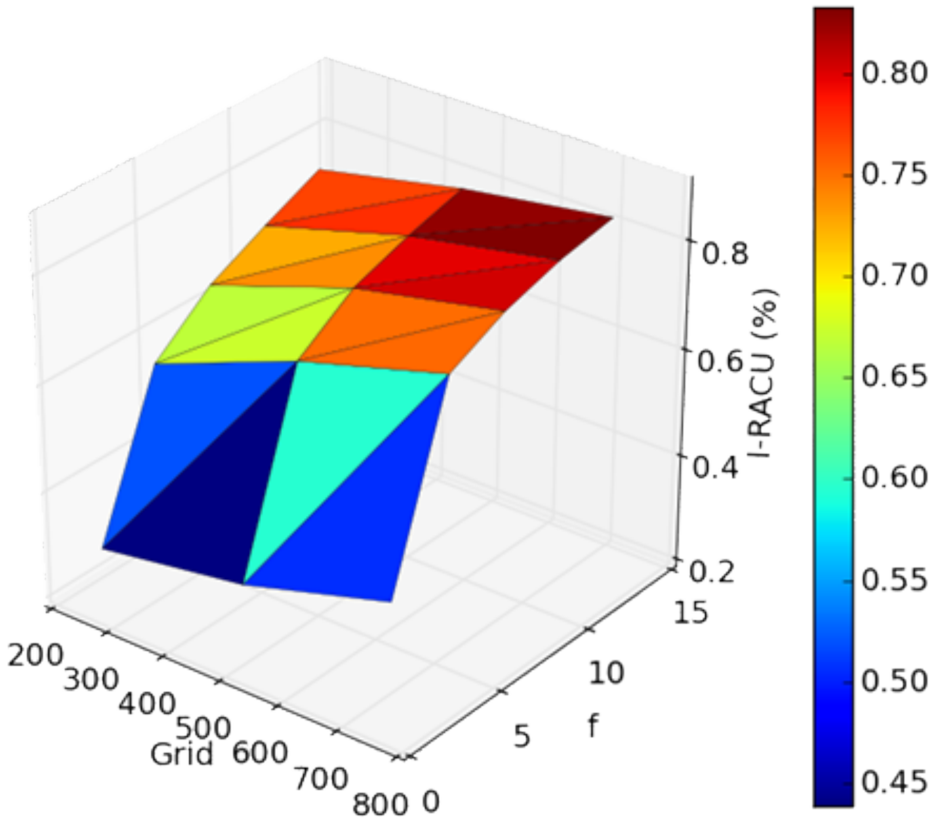
		bk: 1 week		bk: 2 weeks		bk: 3 weeks		bk: 4 weeks	
Risk (r)	**K**	**% users**	**# users**	**% users**	**# users**	**% users**	**# users**	**% users**	**# users**
r<=0.01%	K>=10.000	50	91.613	40	73.141	40	73.001	40	73.001
0.01%<r r<=0.1%	1000<=K K<10.000	22	40.514	16	29.595	14	26.311	14	26.328
0.1%<r r<=1%	100<=K K<1.000	16	30.179	11	19.707	9,6	17.494	9,5	17.381
1%<r r<=2%	50<=K K<100	4,8	8.688	2,7	4.953	2,3	4.244	2,3	4.225
2%<r r<=10%	10<=K K<50	4,6	8.434	6,8	12.322	5,5	10.031	5,3	9.741
10%<r r<=20%	5<=K K<10	0,7	1.213	3,6	6.574	2,5	4.586	2,3	4.170
r>20%	1<=K K<5	0,7	1.225	19	35.574	25	46.199	25	47.000

Risk of Re-IDENtification

A Practical Tool

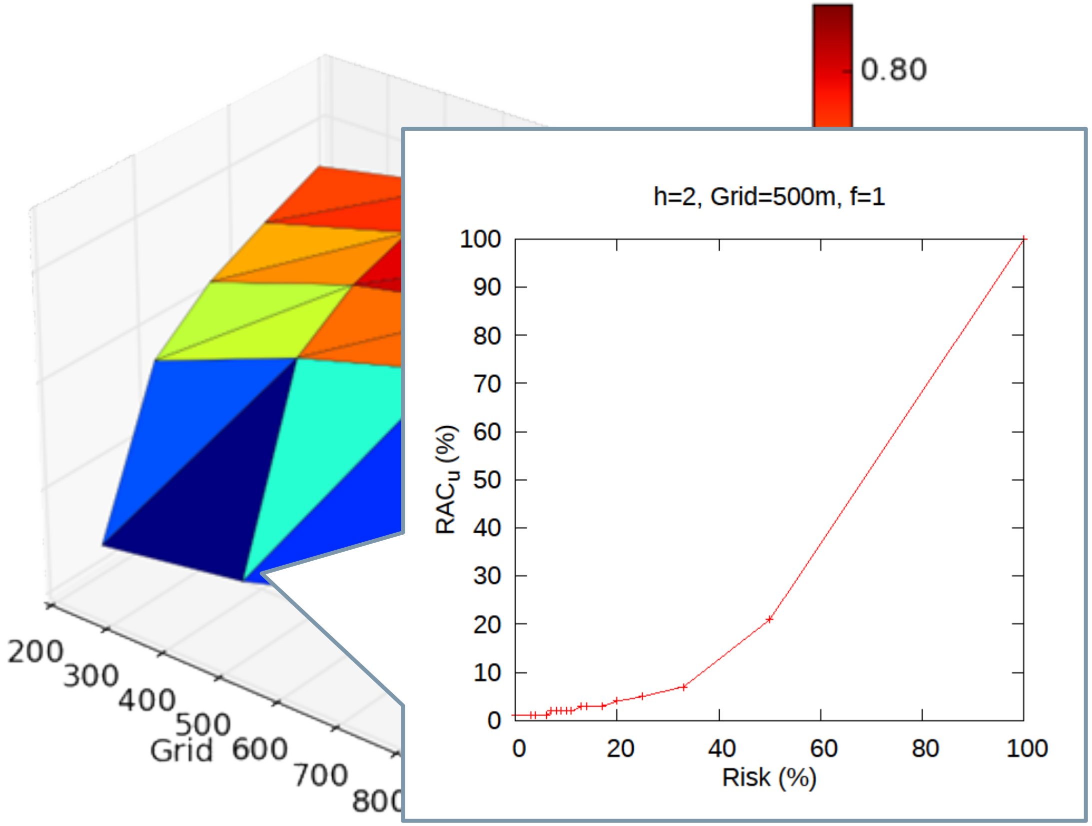
Simulation of a real case

Risk for each combination of min_frequency and grid, setting the #location to 2



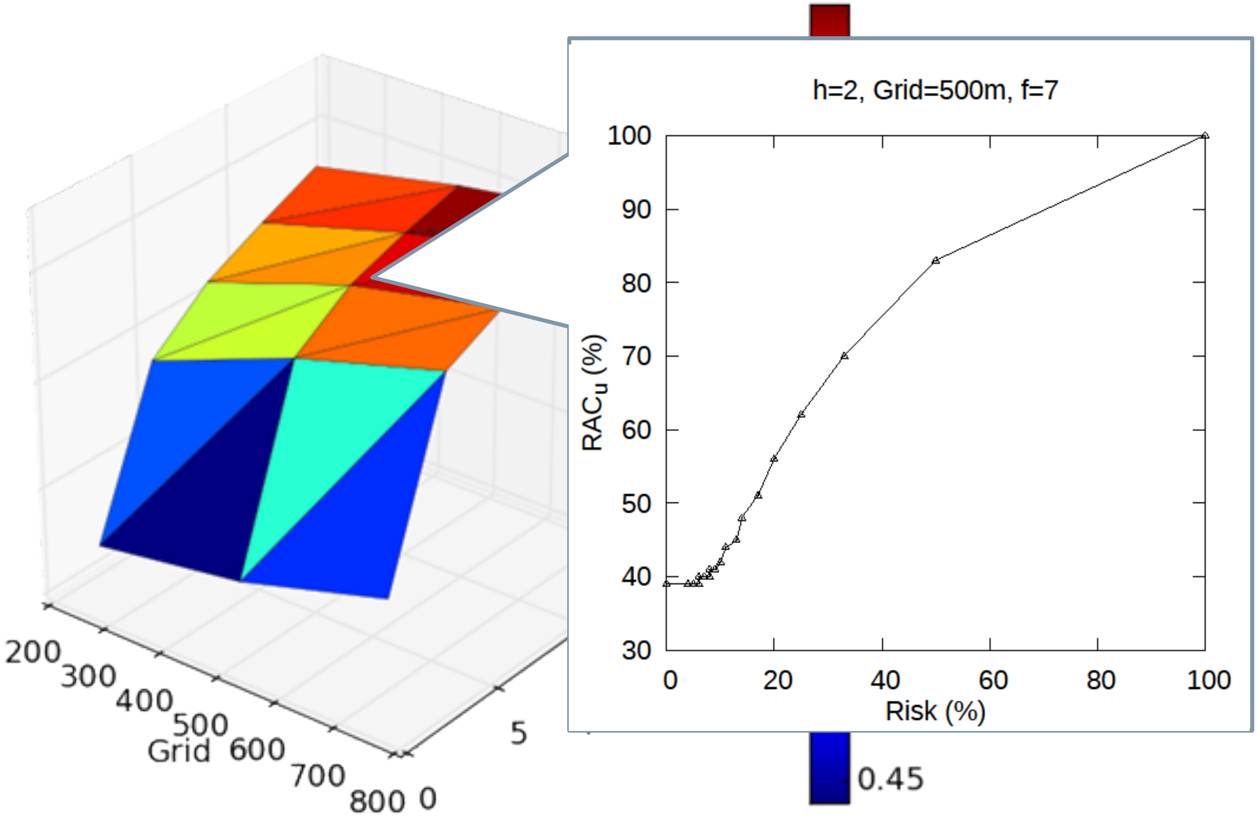
Simulation of a real case

Risk for each combination of min_frequency and grid, setting the #location to 2

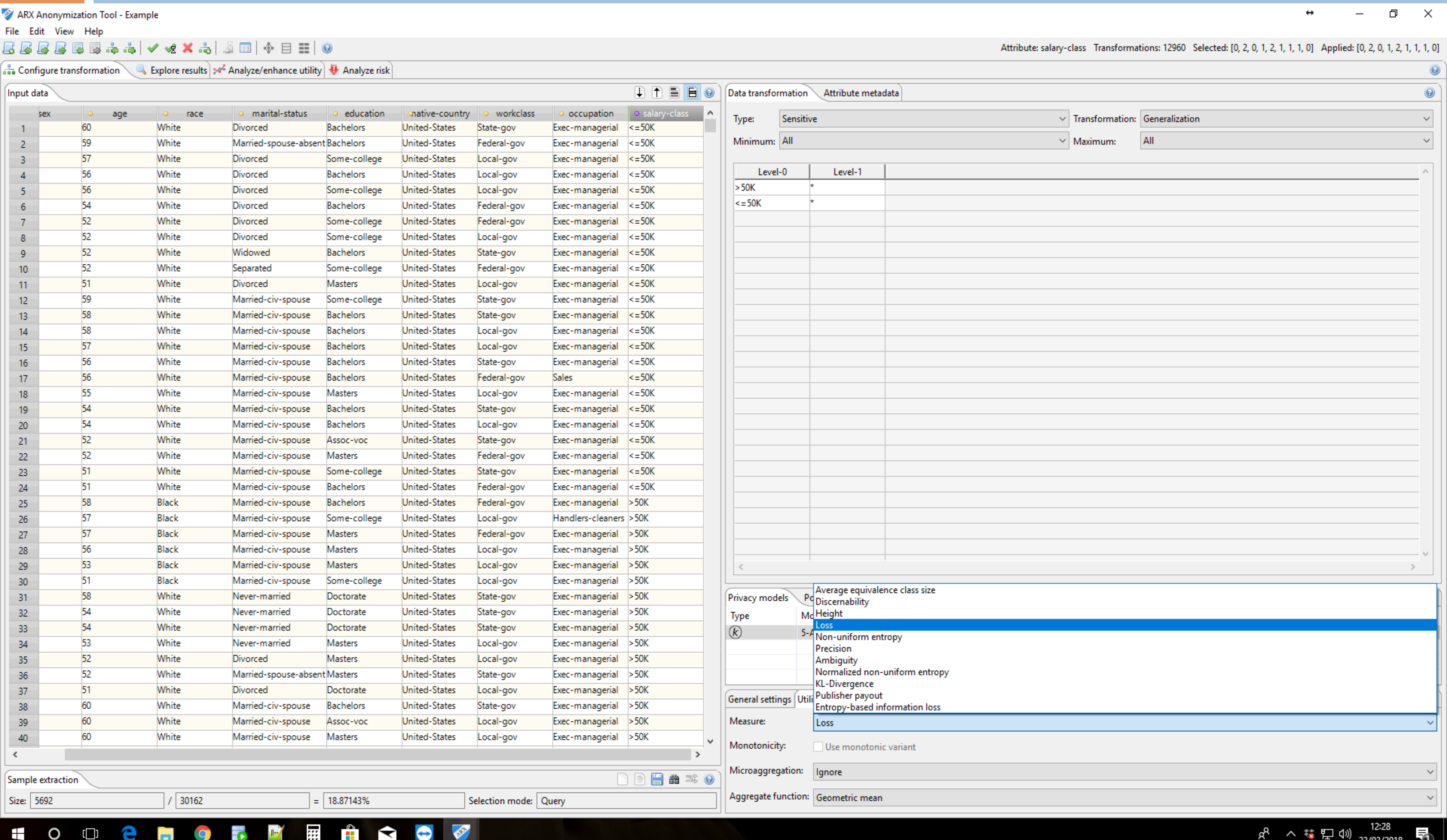


Simulation of a real case

Risk for each combination of min_frequency and grid, setting the #location to 2



http://arx.deidentifier.org/anonymization-tool/



The Amnesia Tool

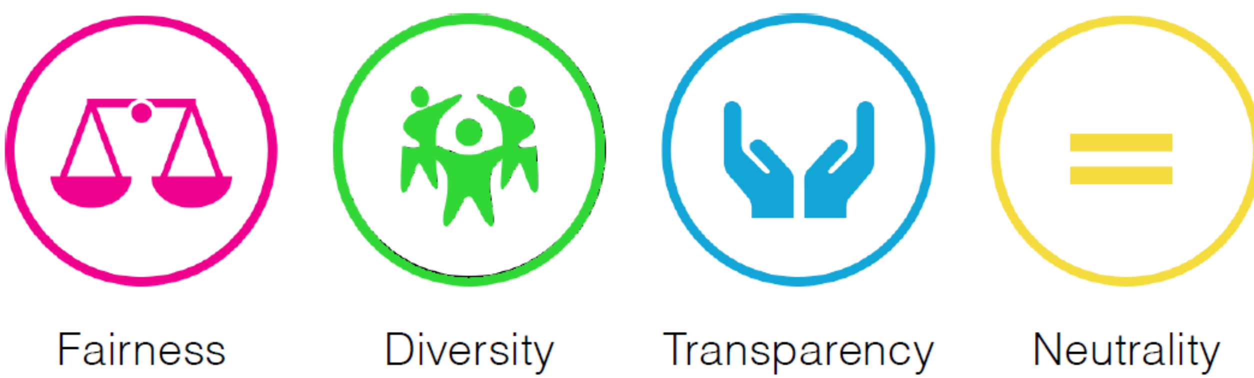
- Visit the website: https://amnesia.openaire.eu/

- Let's see a demo

Right of explanation

Big Data, Big Risks

- **Big data is algorithmic, therefore it cannot be biased!** And yet…
- All traditional evils of social discrimination, and many new ones, exhibit themselves in the big data ecosystem
- Because of its tremendous **power**, massive data analysis must be used **responsibly**
- Technology alone won't do: also need **policy**, **user involvement** and **education** efforts



- By 2018, 50% of business ethics violations will occur through improper use of big data analytics
- [source: Gartner, 2016]

The danger of black boxes

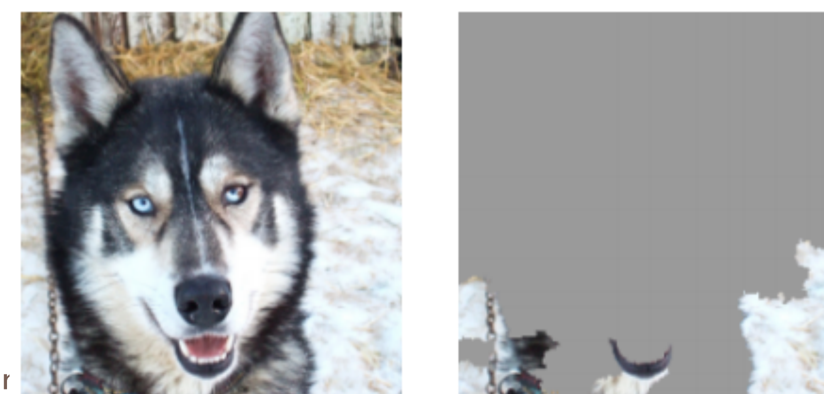
- The COMPAS score (Correctional Offender Management Profiling for Alternative Sanctions)
- A 137-questions questionnaire and a predictive model for "risk of crime recidivism." The model is a proprietary secret of Northpointe, Inc.
- The data journalists at propublica.org have shown that the model has a strong ethnic bias
  - blacks who did not reoffend are classified as high risk twice as much as whites who did not reoffend
  - whites who did reoffend were classified as low risk twice as much as blacks who did reoffend.

The danger of black boxes

- An accurate but untrustworthy classifier may result from an accidental bias in the training data.
- In a task of discriminating wolves from huskies in a dataset of images, the resulting deep learning model is shown to classify a wolf in a picture based solely on …

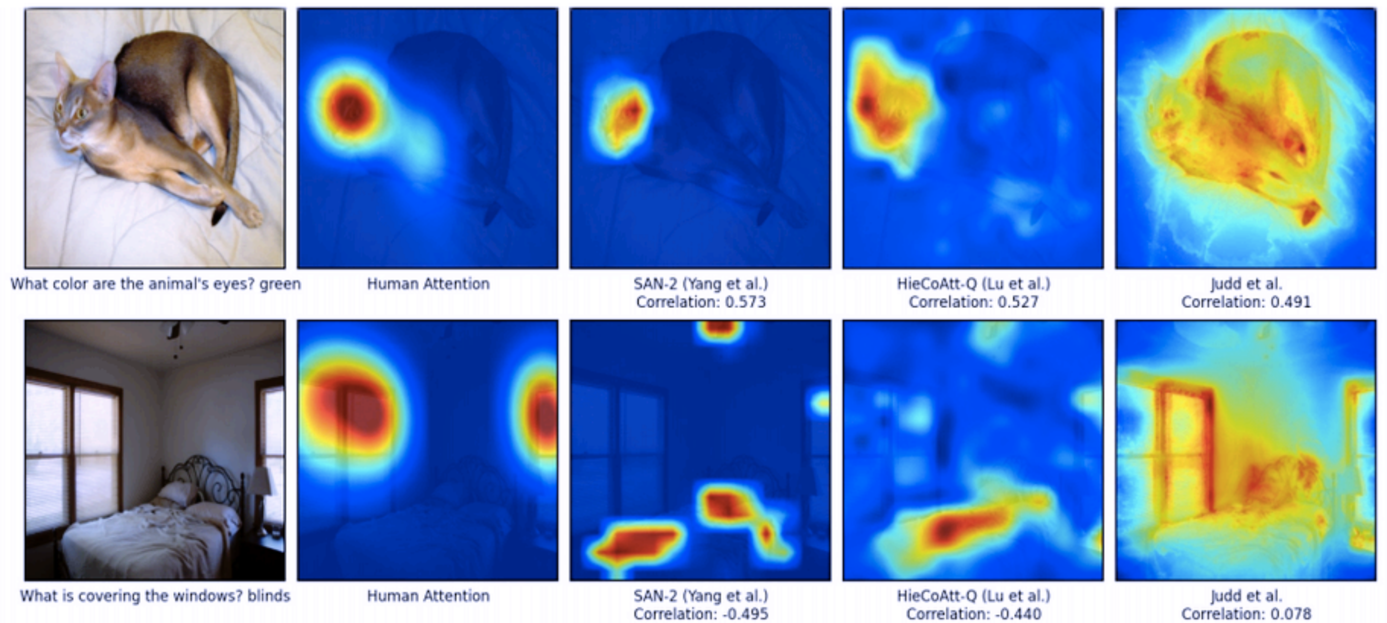
The danger of black boxes

- An accurate but untrustworthy classifier may result from an accidental bias in the training data.
- In a task of discriminating wolves from huskies in a dataset of images, the resulting deep learning model is shown to classify a wolf in a picture based solely on … **the presence of snow in the background!**



(a) Husky classified as wolf (b) Explanation

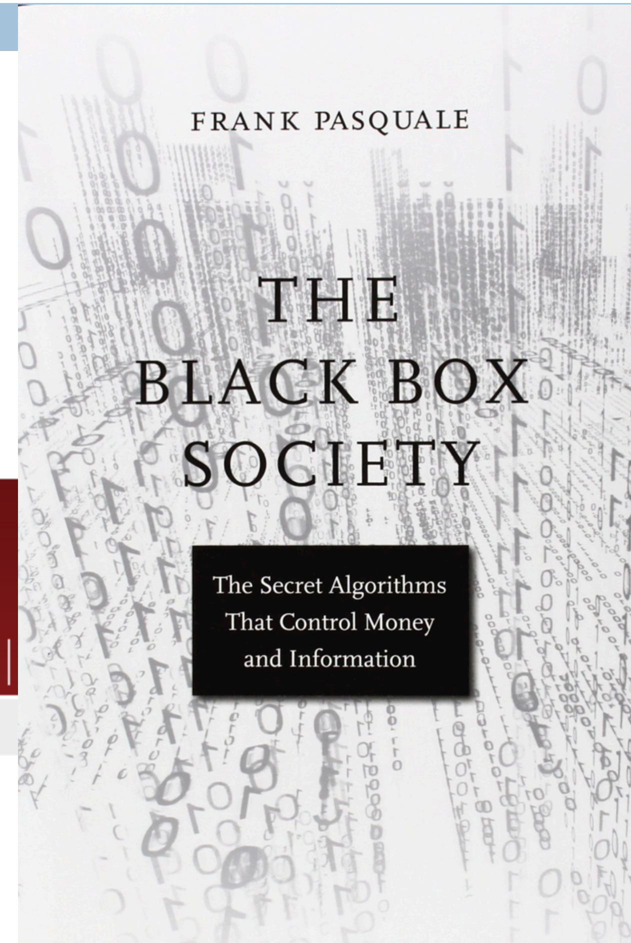
Deep learning is creating computer systems we don't fully understand



"THEY'RE PICKING [ANSWERS] BASED ON BIASES IN THE DATA SETS, RATHER THAN FROM FACTS ABOUT THE WORLD."

Transparent algorithms to build trust

- **Systems that recommend humans making a decision should explain why**



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video |

Archive > Volume 537 > Issue 7621 > Editorial > Article

NATURE | EDITORIAL

More accountability for big-data algorithms

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

21 September 2016

Right of explanation

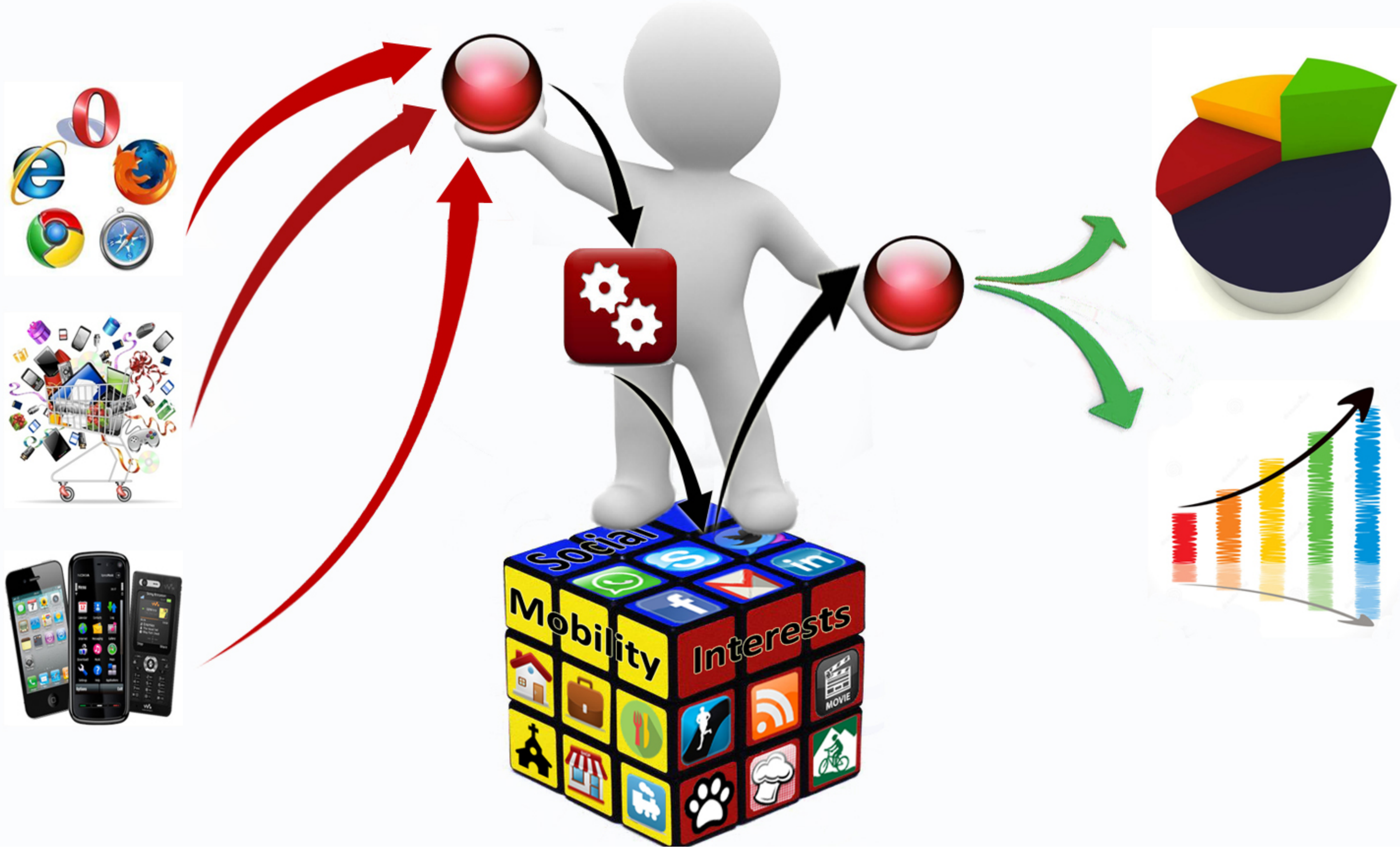
- Two key elements of a decision-making process must be highlighted:
  - its inherent functionality (strict logic)
  - and its contextual use, because it is needed to distinguish the technical *architecture* of an algorithm from the contextual *implementation* of the decision-making in which that algorithm is employed.
  - Following Article 15(1)(h), we can assert that the architecture represents algorithm functionality and the 'logic involved' in the automated processing, while implementation represents the overall decision-making process and thus the context in which the architecture works, i.e. the *significance* of a decision-making and its *envisaged consequences*.

TOWARDS a Personal DATA ecosystem

- An avalanche of personal information that, in most cases, gets lost – *like tears in rain*.

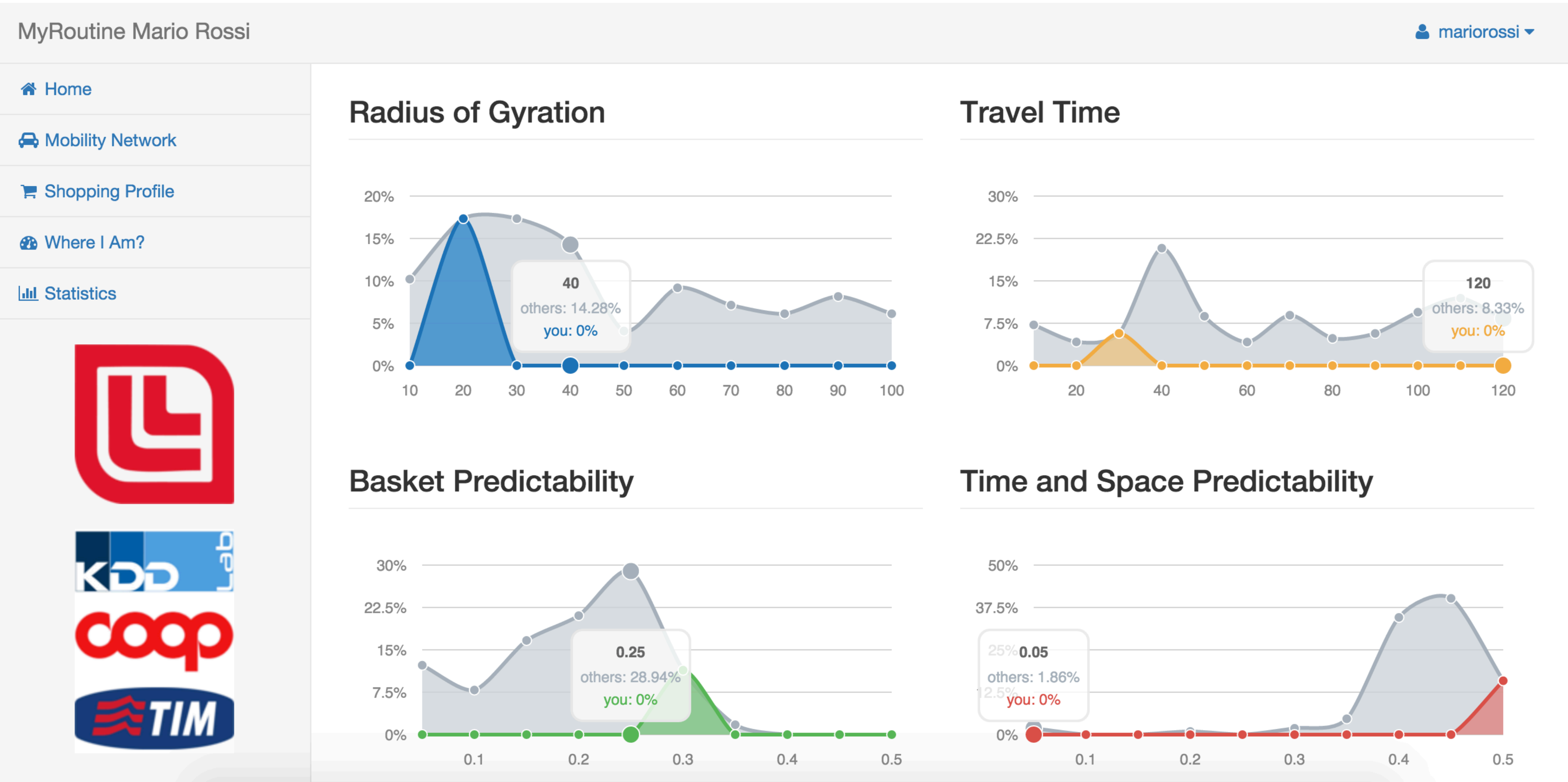
- Yet, only each one of us, individually, has the power to connect all this personal information into a personal data repository – and make sense of it.

A user-centric ecosystem for personal big data



Personal Data Ecosystem

Where am I? Comparison with the community



- We need a Personal Data Ecosystem
  - to acquire, integrate and make sense of our own data
  - and to connect with our peers and the surrounding urban community and infrastructure
- to the purpose of developing the **collective awareness** needed to face our grand challenges

A smart city is a city of participating, aware citizens

LAST POINT

Disseminate our knowledge

- We strongly believe that some basic knowledge about privacy and ethics pillars must be shared
- To this aim:
  - We created an online course
    - You can find it at: 146.48.83.51/moodle/
  - We wrote a manual:
    - http://www.shopping24.ilsole24ore.com/sh4/catalog/products/libri.jsp?productId=prod251016&categoryId=SH246094774&locale=informativa_cookies.html&productPage=true

TAKE HOME MESSAGE

Non bisogna aver paura ma stare un poco attenti

Le 10 regole per responsible data science

1. Acknowledge that data are people and can do harm
2. Recognize that privacy is more than a binary value
3. Guard against the reidentification of your data
4. Practice ethical data sharing
5. Consider the strengths and limitations of your data; big does not automatically mean better
6. Debate the tough, ethical choices
7. Develop a code of conduct for your organization, research community, or industry
8. Design your data and systems for auditability
9. Engage with the broader consequences of data and analysis practices
10. Know when to break these rules

Source: Zook M, Barocas S, boyd d, Crawford K, Keller E, Gangadharan SP, et al. (2017) Ten simple rules for responsible big data research. PLoS Comput Biol 13(3): e1005399. https://doi.org/10.1371/journal.pcbi.1005399


Thank
you

The Amnesia Tool

- Visit the website: https://amnesia.openaire.eu/

- Let's see a demo