Big Data Anonymisation Techniques

Claude Castelluccia
INRIA PRIVATICS
November 2016

Inria
PRIVATICS

BIG DATA is Useful

- Analysis and publication of large datasets are essential to research and business

- Very useful to:
  - predict flu
  - improve transportation, Logistic
  - improve knowledge and efficiency
  - Improve services….

BIG DATA: The Privacy Risks

- Singling-out/ Re-Identification:
  - Adversary (ADV) is able to identify the target's record in the published dataset… from some know information
- Attribute Inference
  - ADV can infer (or guess) private/sensitive attributes from released dataset
  - Because of cross-attributes and cross-users correlation!
- Example:
  - a dataset reveals that all users who went to points A, B, C, also went to D (for example an hospital).
  - I know that if a yarget was at A, B, C… i can then infer that target was also in D!

BIG DATA

The Risks of **Identity** Inference: The AOL Case

- In 2006, AOL released 20 million search queries for 650.000 users
- « (pseudo)-Anonymized » by removing AOL id and IP address
- Easily de-anonymized in a couple of days by looking at queries

The AOL Case

The New York Times

Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

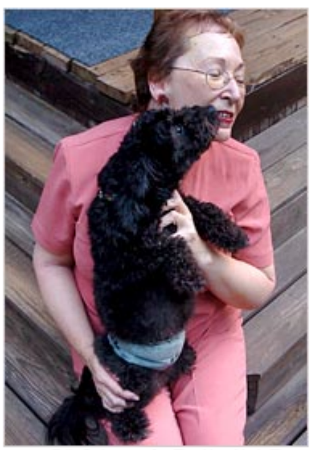
CAMCORDERS CAMERAS CELLPHONES COMPUTERS HANDHELDS HOME VIDEO MUSIC PERIPHERA

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-MAIL THIS
PRINT
REPRINTS

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.



Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

AOL removed the search data from its site over the

BIG DATA

The Risks of **Attribute** Inference: The Target Case

- Target identified about 25 products that, when analyzed together, allowed him to assign each shopper a “pregnancy prediction” score.
- More important, he could also estimate her due date to within a small window
- Target could (and does) send coupons timed to very specific stages of her pregnancy.

Source: How Companies Learn Your Secrets, NYTimes, Feb. 2012

TARGET

Other Examples..

1997: The case of Massachusetts' Governor

2009: Netflix prize

2013: NYC Taxi dataset

…

Datasets need to be Well Sanitized/Anonymized…

Sanitization : *process which increases the uncertainty in the data in order to preserve privacy..*
⇒ Inherent trade-off between the desired level of privacy and the utility of the sanitized data.
Typical example : public release of data.

Examples drawn from the “sanitization” entry on Wikipedia

What is Data Anonymization for Computer Scientists?

*Data are anonymised if all identifying elements (all quasi-identifiers) have been eliminated from a set of personal data. No element may be left in the information which could, by exercising **reasonable** effort, serve to re-identify the person(s) concerned.*

- *Where data have been successfully anonymised, they are no longer personal data.*

A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

Produced by FUTURE OF PRIVACY FORUM
FPF.ORG
In collaboration with EY

What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability.

This is a primer on how to distinguish different categories of data.

DEGREES OF IDENTIFIABILITY

Information containing direct and indirect identifiers.

PSEUDONYMOUS DATA

Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.

DE-IDENTIFIED DATA

Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.

ANONYMOUS DATA

Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.

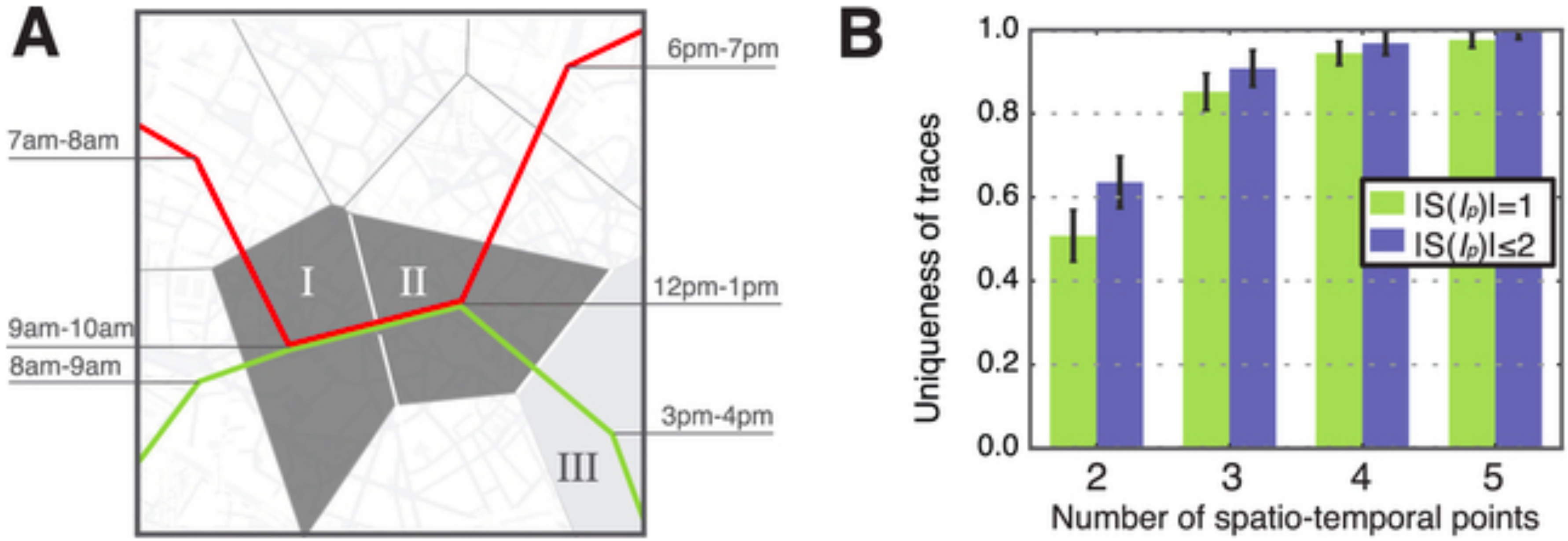
	EXPLICITLY PERSONAL	POTENTIALLY IDENTIFIABLE	NOT READILY IDENTIFIABLE	KEY CODED	PSEUDONYMOUS	PROTECTED PSEUDONYMOUS	DE-IDENTIFIED	PROTECTED DE-IDENTIFIED	ANONYMOUS	AGGREGATED ANONYMOUS
DIRECT IDENTIFIERS Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)	INTACT	PARTIALLY MASKED	PARTIALLY MASKED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
INDIRECT IDENTIFIERS Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)	INTACT	INTACT	INTACT	INTACT	INTACT	INTACT	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
SAFEGUARDS and CONTROLS Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals	NOT RELEVANT due to nature of data	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	NOT RELEVANT due to nature of data	NOT RELEVANT due to high degree of data aggregation
SELECTED EXAMPLES	Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)	Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68:A8:6D:35:65:03)	Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses & legal representations)	Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)	Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = 5L7T LX619Z) (unique sequence not used anywhere else)	Same as Pseudonymous, except data are also protected by safeguards and controls	Data are suppressed, generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 = 3.0-3.5, gender: female = gender: male)	Same as De-Identified, except data are also protected by safeguards and controls	For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)	Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)

Why is Data Anonymization Difficult?

- Quasi-identifiers are difficult to identify exhaustively
- Many combination of attributes can be used to « single-out » a user
- We are all unique by different ways, we are full of Q.I.
  - See « Unicity me! *»
  - Mobility pattern, webhistory, .
  - Data (content) and meta-data
    - i.e. timing can betray you!
    - Google search timing pattern can tell when you were away!

*Unicity Me! American Scientific, http://www.americanscientist.org/libraries/documents/20142614253010209-2014-03CompSciHayes.pdf

Unique in the Crowd [Nature2013]



- Only 4 spatio-temporal points are necessary to uniquely identify a user with a probability > 95% !

Why is Data Anonymization Difficult?

Anonymisation is a utility/privacy optimization

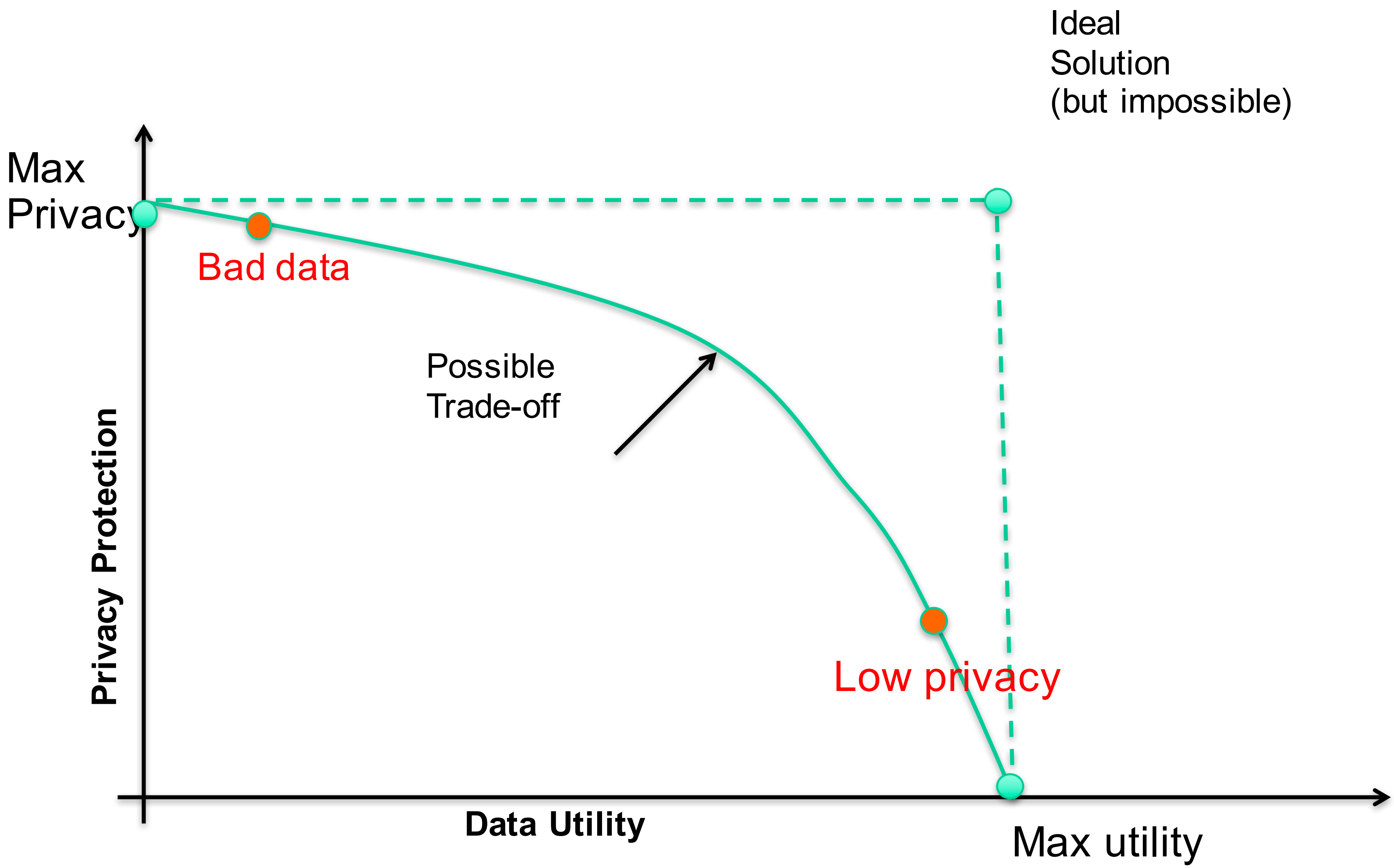
No generic solution that optimizes utility and privacy!

Anonymisation should be performed case by case…. According to:

- Type of data
- Sensitivity of data
- Type of release
- Adversary models
- ….

Risk-based approach….

Privacy vs Utility Tradeoff



Pseudo-Anonymization

- Anonymisation is NOT pseudo-anonymization!
- What is Pseudo-Anonymization?
  - *Personal information contains identifiers, such as a name, date of birth, sex and address. When personal information is pseudonymised, the identifiers are replaced by one pseudonym. Pseudonymisation is achieved, for instance, by encryption of the identifiers in personal data.*

Microdata

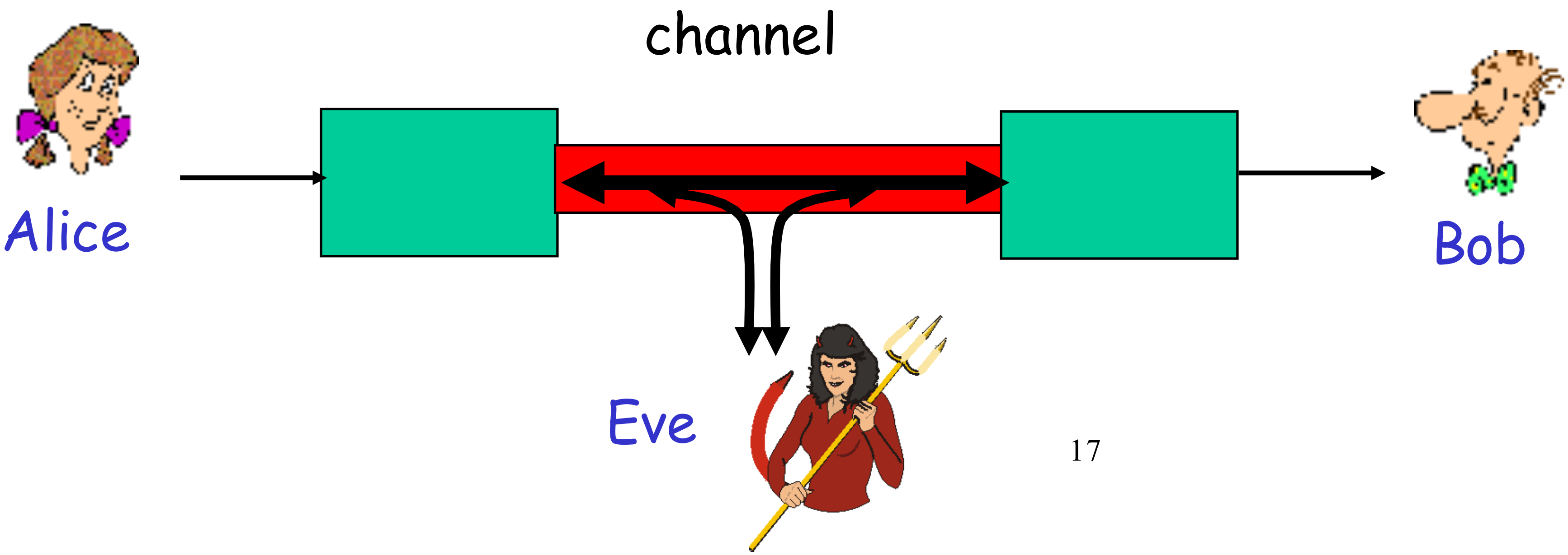
Name	Zipcode	Age	Sex	Disease
[illegible]	47677	29	F	Ovarian Cancer
[illegible]	47602	22	F	Ovarian Cancer
[illegible]	47678	27	M	Prostate Cancer
[illegible]	47905	43	M	Flu
[illegible]	47909	52	F	Heart Disease
[illegible]	47906	47	M	Heart Disease

Pseudo-Anonymization

- Why is Pseudo-Anonymization not good Enough?
  - It does not compose, i.e. several Pseudo-Anonymized data can be combined to de-anonymize…
  - External Information can also be exploited.
  - See previous examples
- We need schemes that also alter the quasi-identifiers (not only the identifiers)
  - K-anonymity
  - Differential Privacy
  - …

Why not using Cryptography?

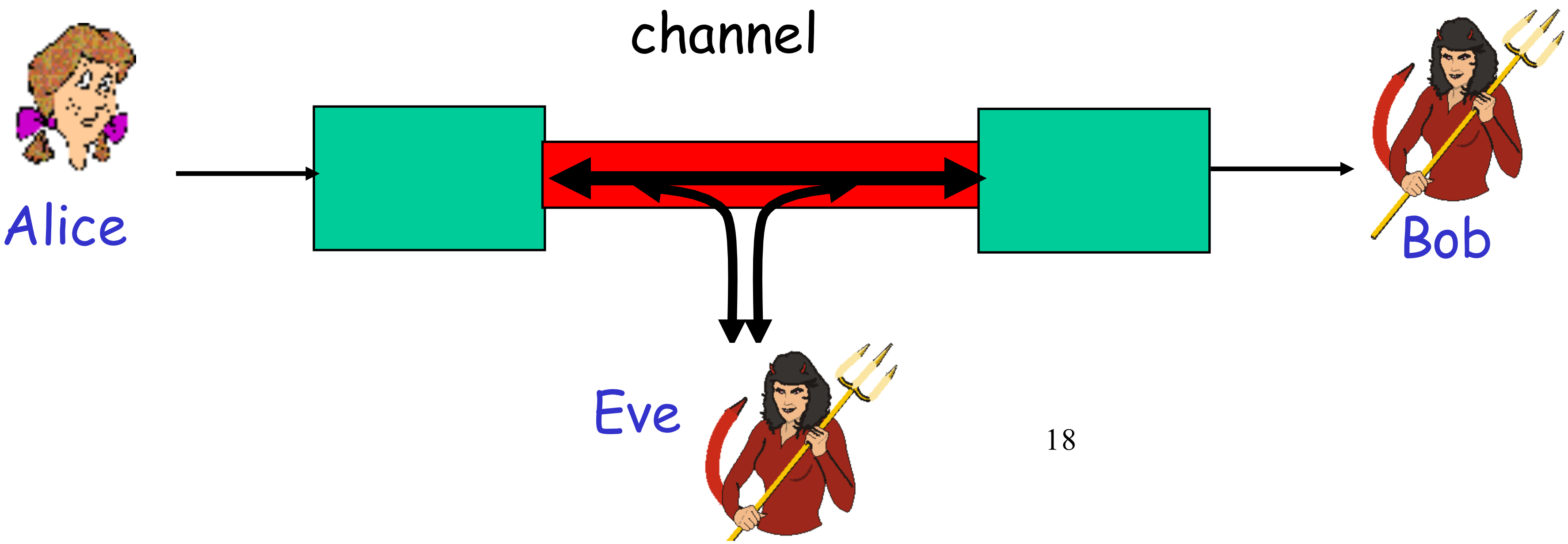
- The Trust models are different!
- In cryptography, sender and receiver trust each others:
  - Alice sends a dataset to Bob
  - Alice encrypts to protect from Eve, the eavesdropper
  - But Bob can decrypt and recover the original dataset!
  - The **adversary is Eve**!



Cryptography and Anonymization

With Data anonymization, the sender does not trust the receiver

- Alice anonymized a dataset to “hide” some (usually personal) information and sends it to Bob (possibly after encryption).
- Bob recovers the anonymized dataset. It can process it to compute some statistics/inferences…but can’t recover the hidden information (identity or attribute).
- Bob is also the adversary!



Anonymization As A Security Measure

- Anonymization is often presented in order to protect privacy (personal information), to be in conformity with the Law
- Note that similar techniques can be used to improve security or protect intellectual properties
  - A bank might want to hide the names of his customers to his employees (to avoid data leakage)
  - A company that is exchanging some files with another company might want to hide some “sensitive/important” information (not necessary personal information)

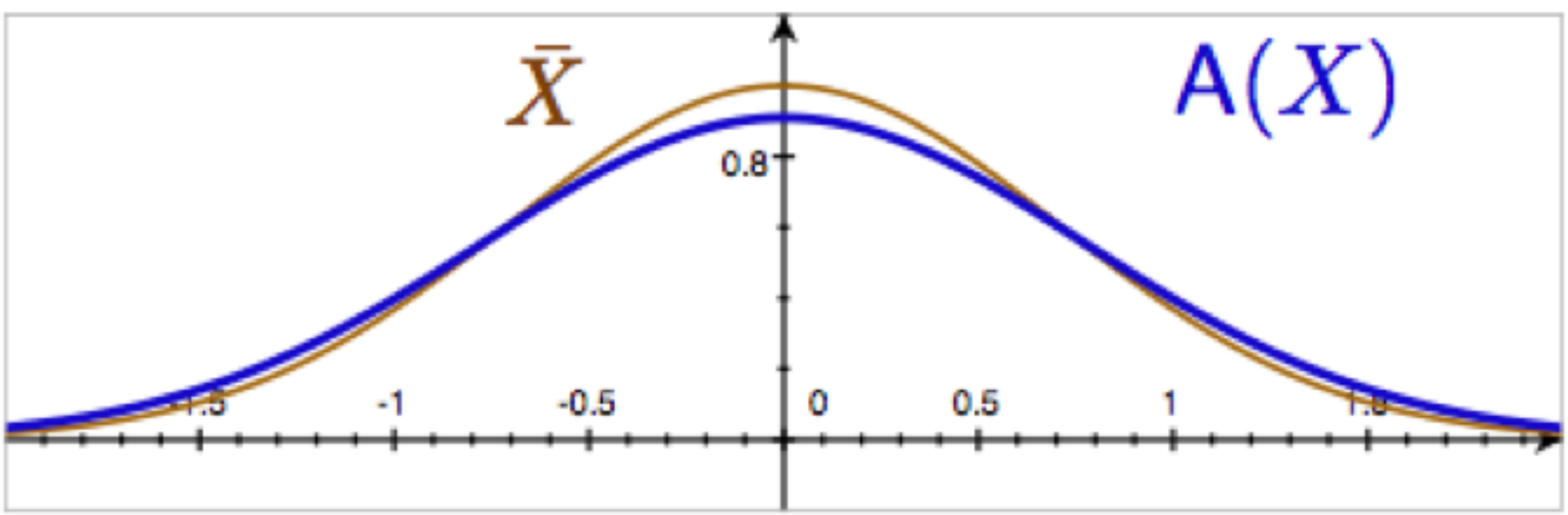
Some Data anonymization methods…

- Random perturbation
  - Input perturbation
  - Output perturbation
- Generalization
  - The data domain has a natural hierarchical structure.
- Suppression
- Permutation
  - Destroying the link between identifying and sensitive attributes that could lead to a privacy leakage.

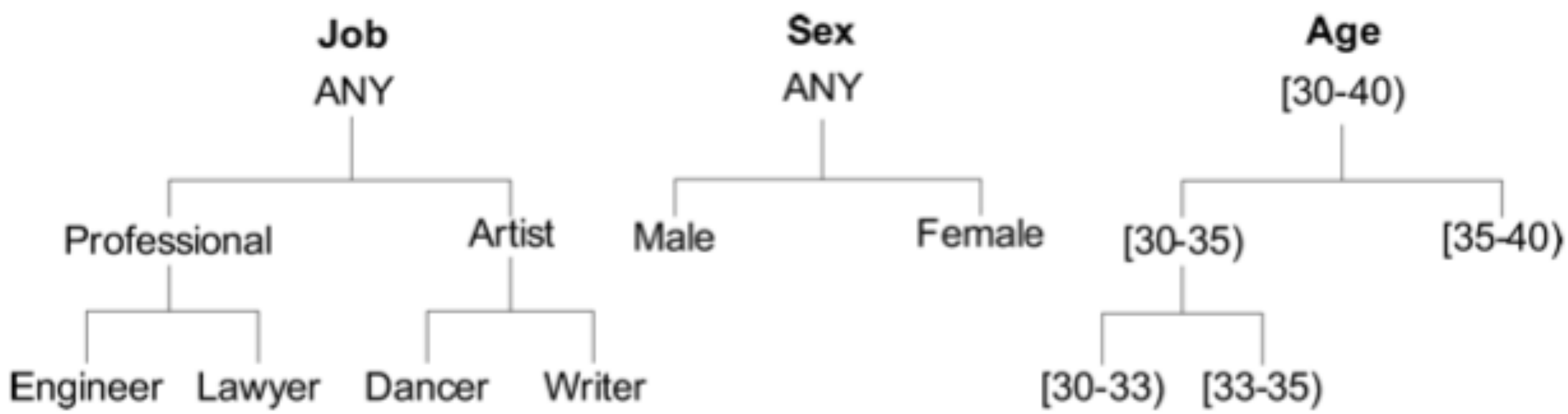
Randomization Methods

Randomization : add independent noise (such as Gaussian or uniform) to the values transmitted.

Goal : hide the specific values of attributes while preserving the joint distribution of the data.



Generalization Methods



Suppression Methods

Age	Sex	Disease (sensitive)
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	Flu
38	Female	Flu
38	Female	Heart
38	Female	Heart

Age	Sex	Disease (sensitive)
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	~~HIV~~
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	Flu
38	Female	Flu
38	Female	Heart
38	Female	Heart

K-anonymity

- Privacy guarantee : in each group of the sanitized dataset, each invidivual will be identical to a least $k - 1$ others.
- Reach by a combination of generalization and suppression.
- Example of use : sanitization of medical data.

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

But K-Ano. does not compose ☹!

- Question : suppose that Alice's employer knows that she is 28 years old, she lives in ZIP code 13012 and she visits both hospitals. What does he learn?

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

(a)

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

(b)

But K-ANO does not compose ☹!

- Question : suppose that Alice's employer knows that she is 28 years old, she lives in ZIP code 13012 and she visits both hospitals. What does he learn?

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

(a)

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

(b)

Other Attacks on k-Anonymity

- k-Anonymity does not provide privacy if
  - Sensitive values in an equivalence class lack diversity
  - The attacker has background knowledge

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Carl does not have heart disease	
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

Some Other Anonymization Schemes

- *l*-diversity (MKGV[1] 07) : maintain the diversity for each group with respect to the possible values of the sensible attributes.
- Can be instancied by a metric based on *entropy*.
- Prevent against attacks based on homogeneity and some other attacks.
- *t*-closeness (LLV[2] 07) : the distribution of the attributes in each group must be close to that on the global population.
- *t* is a threshold that should not be exceed and which represents the proximity between distributions.

l-Diversity: Preventing the Homogeneity attack

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Sensitive attributes must be "diverse" within each quasi-identifier equivalence class

Distinct l-Diversity

- Each equivalence class has at least l well-represented sensitive values
- Doesn’t prevent probabilistic inference attacks

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

10 records

8 records have HIV

2 records have other values

t-Closeness

[Li et al. ICDE '07]

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Distribution of sensitive attributes within each quasi-identifier group should be "close" to their distribution in the entire original database

Why publish quasi-identifiers at all??

Toward « Provable » Anonymization

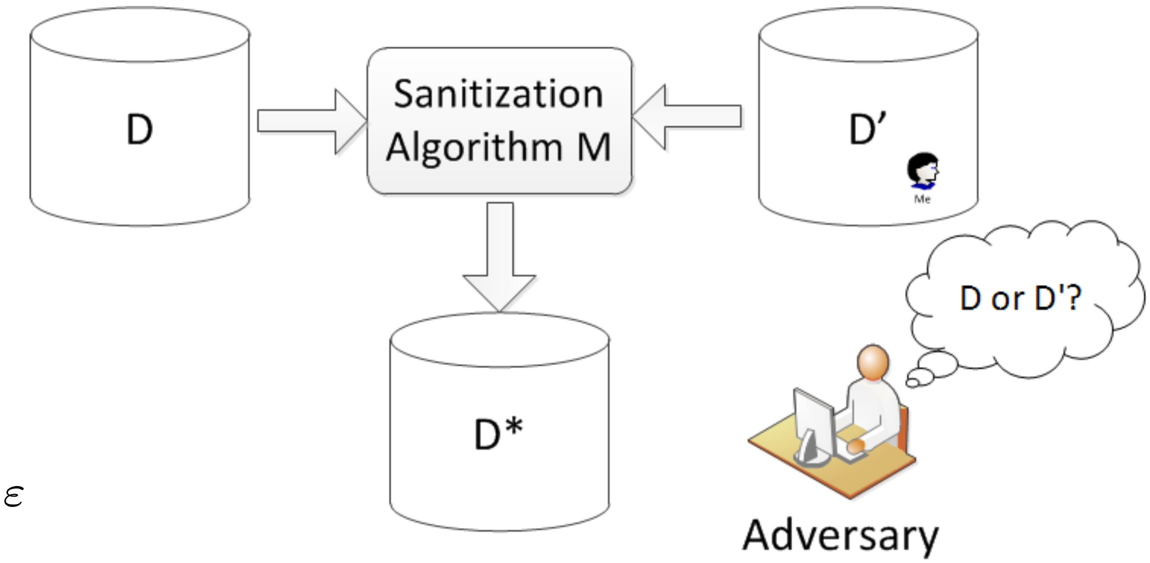
- Stronger schemes are necessary

- Differential Privacy (DP)
  - Provides some strong and measurable guarantees
  - Secures even with external sources of data
  - Composes

- Intuition of DP:
  - Changes to my data not noticeable
  - Output is “independent” of my data

Privacy Model

- **Differential privacy**

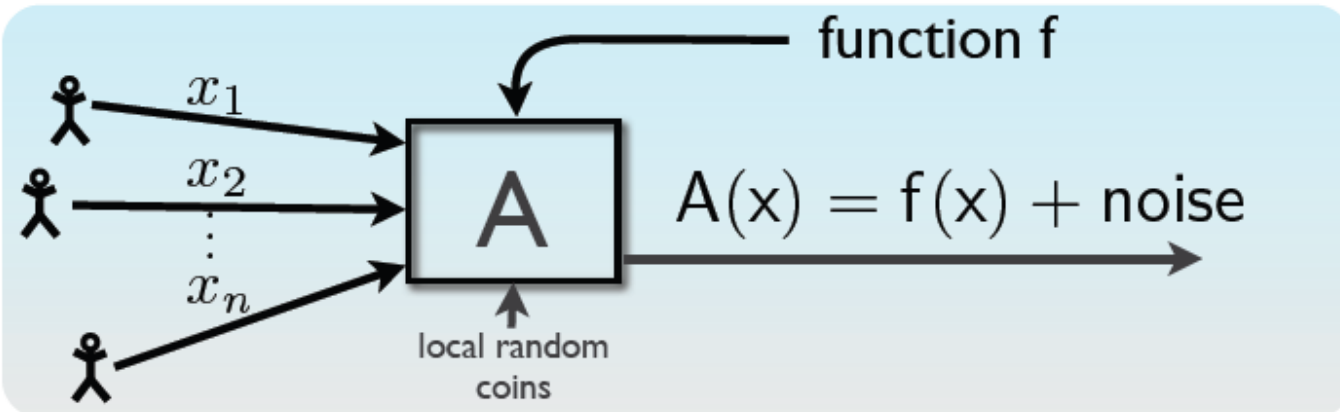


$$e^{-\varepsilon} \leq \frac{\Pr(M(D) = D^*)}{\Pr(M(D') = D^*)} \leq e^{\varepsilon}$$

- ***composes securely***: retain privacy guarantees in the presence of independent releases[1]
- Secure even with arbitrary external knowledge!

[1] S.R. Ganta, S. Kasiviswanathan, A. Smith. *Composition Attacks and Auxiliary Information in Data Privacy*. KDD'08

Differential Privacy

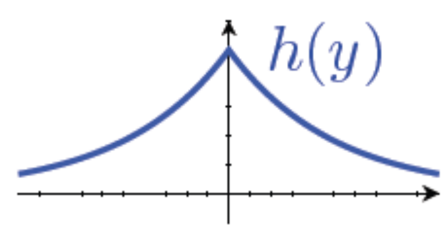


- **Global Sensitivity:** $GS_f = \max_{\text{neighbors } x, x'} \|f(x) - f(x')\|_1$
  - ➢ Example: $GS_{\text{proportion}} = \frac{1}{n}$

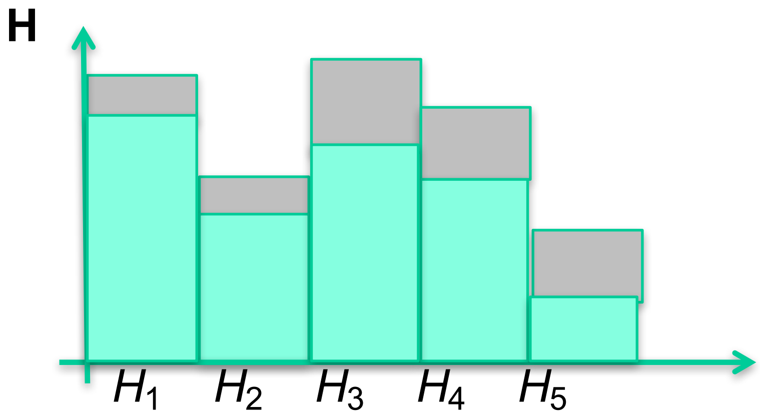
Theorem: If $A(x) = f(x) + \text{Lap}\left(\frac{GS_f}{\epsilon}\right)$, then A is ϵ-differentially private.

➢ Laplace distribution $\text{Lap}(\lambda)$ has density

$$h(y) \propto e^{-|y|/\lambda}$$

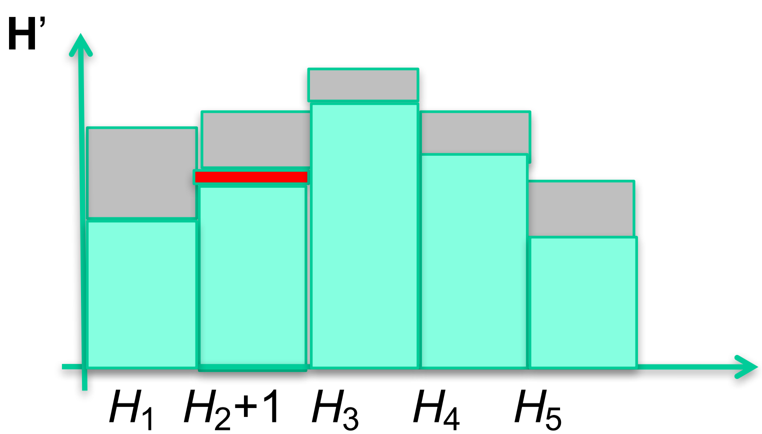


Histogram Release with Laplace Mechanism



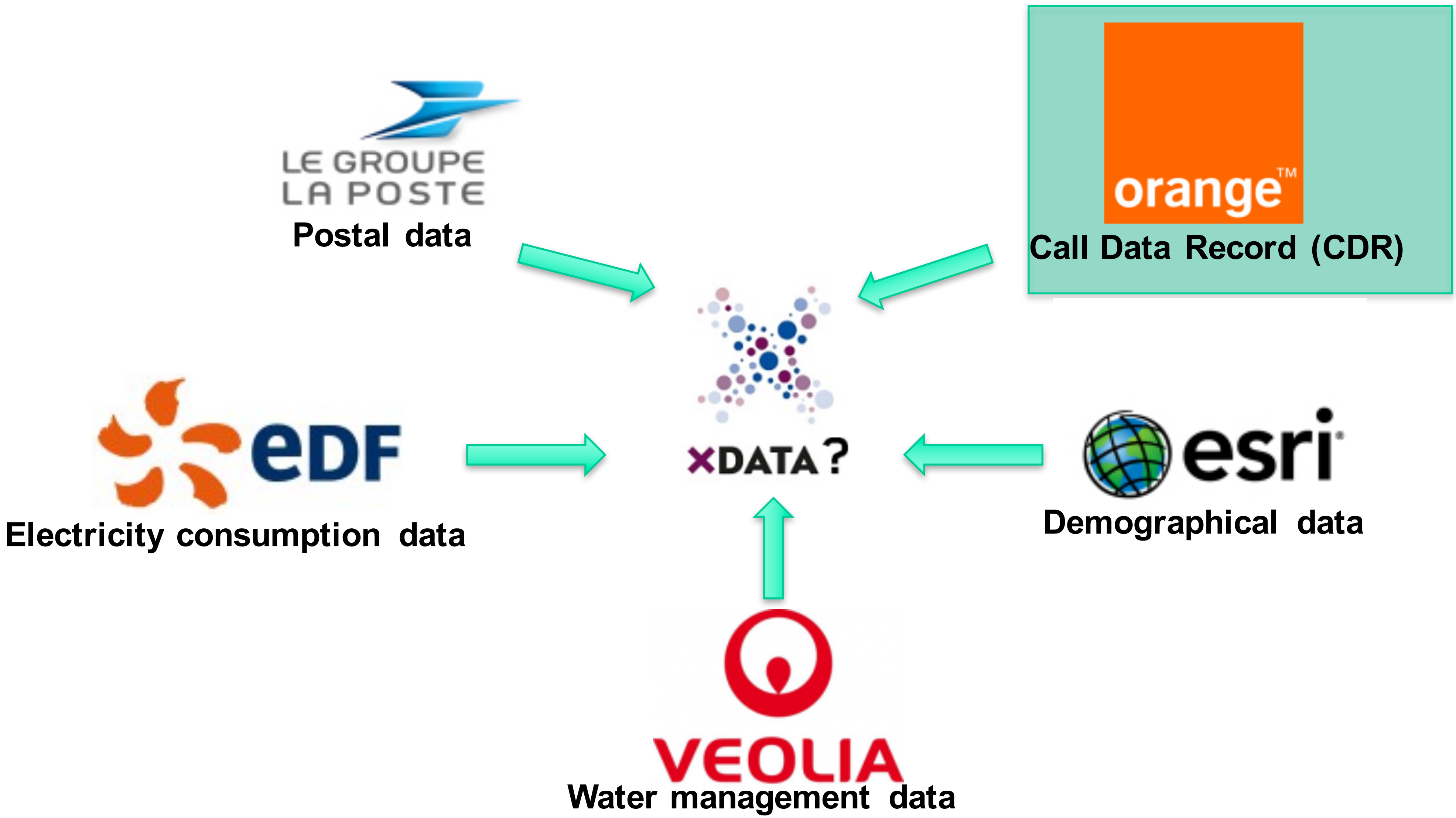
Add random Laplace noise to each bin before publishing!

$$\frac{\prod_i \Pr(H_i + Laplace(\lambda) = H_i^*)}{\prod_i \Pr(H_i' + Laplace(\lambda) = H_i^*)} \leq \exp\left(\frac{\sum_i |H_i - H_i'|}{\lambda}\right) = e^{\frac{1}{\lambda}}$$



- **Global sensitivity:**
  $\Delta H = \Sigma |H_i - H_i'|$
- For histograms: $\Delta H = 1$
- If $\lambda = \Delta H / \varepsilon$, we have *ε-differential privacy*

Example: Spatio-temporal density from CDR



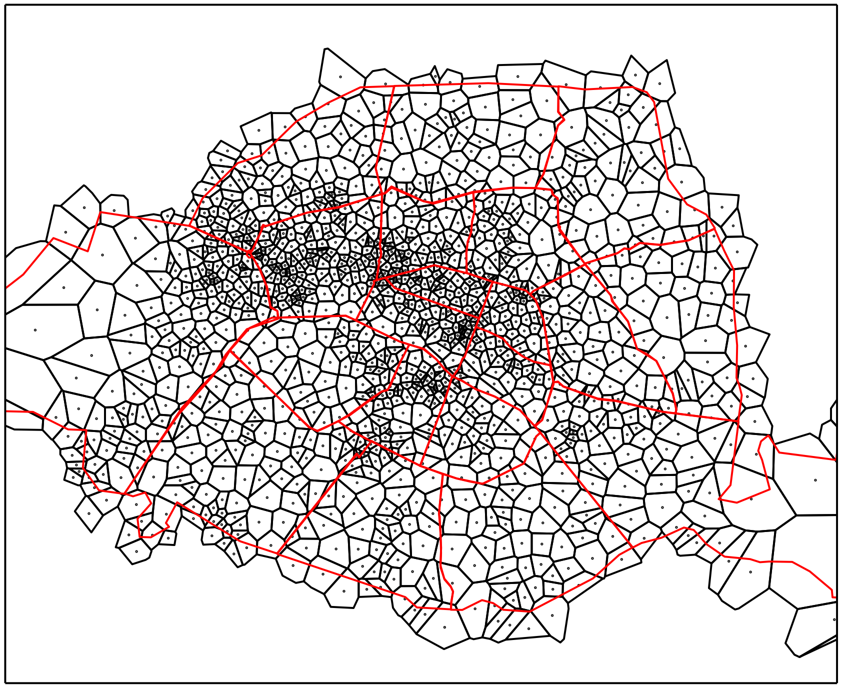
(Simplified) Call Data Record

Rec #	Phone	Lat	Lon	Time	Event
1	0644536701	46.345	2.32	13:34:12 01/09/2007	Incoming SMS
2	0634556702	47.123	1.65	14:31:02 02/09/2007	Outgoing Call
…	…				

- 4 types of events:
  - Incoming SMS/Call
  - Outgoing SMS/Call
- Phone numbers are scrambled (No Personal Data in the dataset)

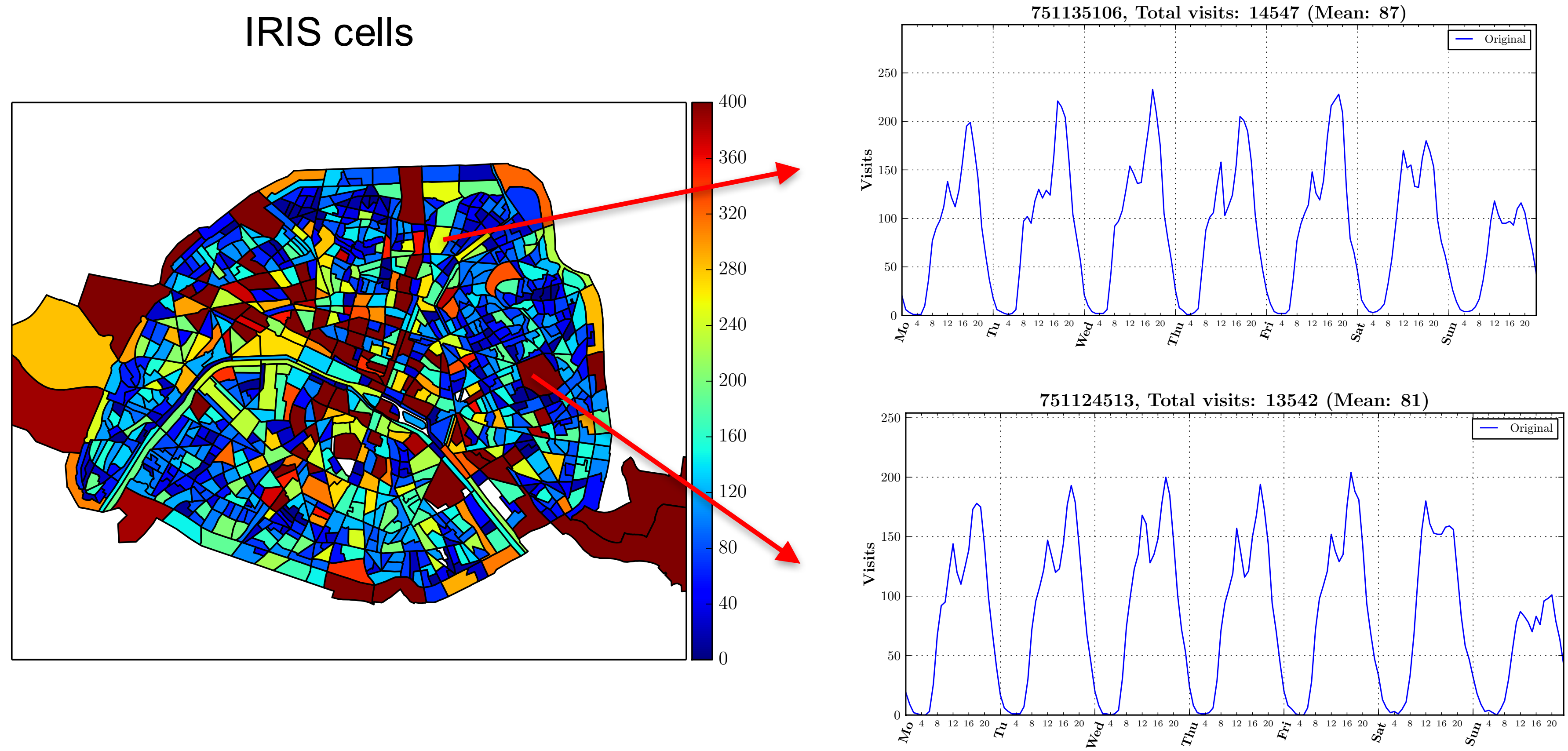
Paris CDR (provided by Orange™)

- 1,992,846 users
- 1303 towers
- 10/09/2007 – 17/09/2007
- Mean trace length: 13.55 (std.dev: 18)
- Max. trace length: 732



Goal: Release spatio-temporal density (and not CDR)

Number of individuals at a given hour at any IRIS cell in Paris



Overview of our approach

1. Sample *x (≈ 30)* visits per user uniformly at random (to decrease sensitivity)
2. Create time-series: map tower cell counts to IRIS cell counts
3. Perturb these time-series to guarantee differential privacy

Overview of our approach

1. Sample *x (≈ 30)* visits per user uniformly at random
2. Create time-series: map tower cell counts to IRIS cell counts
3. Perturb these time-series to guarantee differential privacy

Perturbation of time series

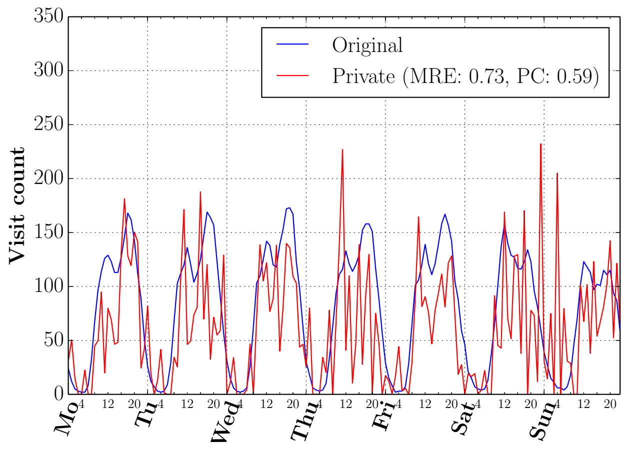
Naïve solution: add properly calibrated Laplace noise to each count of the IRIS cell (one count per hour over 1 week)

Problem: *Counts are much smaller than the noise!*

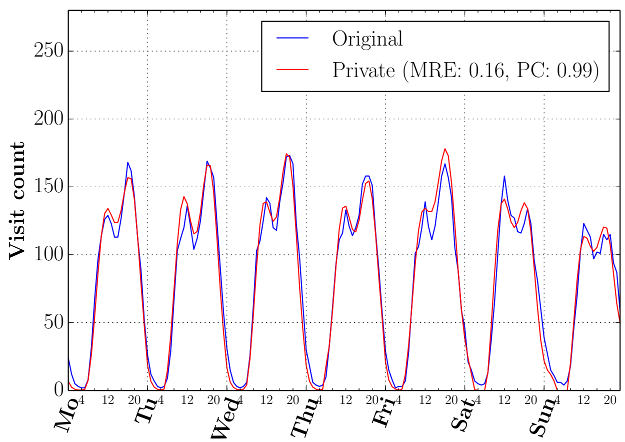
Our approach:

1. cluster nearby less populated cells until their aggregated counts become sufficiently large to resist noise.
2. perturb the aggregated time series by adding noise to their largest Fourier coefficients
3. scale back with the (noisy) total number of visits of individual cells to get the individual time series

Naïve approach (ε=0.3)



Our approach (ε=0.3)



Conclusion :
There are no universal solutions!

- **There are no "universal" anonymization solutions that fit all applications**
  - in order to get the best accuracy, they have to be customized to the application and the public characteristics of the dataset
    - specific context
    - specific utility/privacy tradeoff
    - Specific ADV models
    - Specific impacts..
- Anonymization is all about utility/efficiency trade-off!
  - Full-proof security is not always necessary (and probably impossible)!
  - It has to be performed with a PRA (Privacy Risk Analysis)

Conclusion :
Anonymization does not solve everything!

- Sanitization schemes protect against re-identification, **not inference**!
- You can learn and infer a lot from data
  - You can infer religion from Mobility data!
  - Interest from Google search requests
- You can learn and infer a lot from meta-data!
  - Who communicated with whom?
  - Is a user away/active?
- It is up to the society to decide what is acceptable or not!
  - By balancing the benefits with the risks*.

**Benefit-Risk Analysis for Big Data Projects, http://www.futureofprivacy.org/wp-content/uploads/FPF_DataBenefitAnalysis_FINAL.pdf*

Thanks for your attention!

Claude.castelluccia@inria.fr

Inria PRIVATICS