Karlsruhe Institute of Technology

Privacy-Enhancing Technologies

Module 3: Database Anonymization

Thorsten Strufe

30.05.2022 – hybrid, KIT and TUD

Disclaimer: This lecture was prepared in cooperation with

Patricia Arias-Cabarcos, Javier Parra-Arnau, and input from the people at the chair

KASTEL Security Research Labs

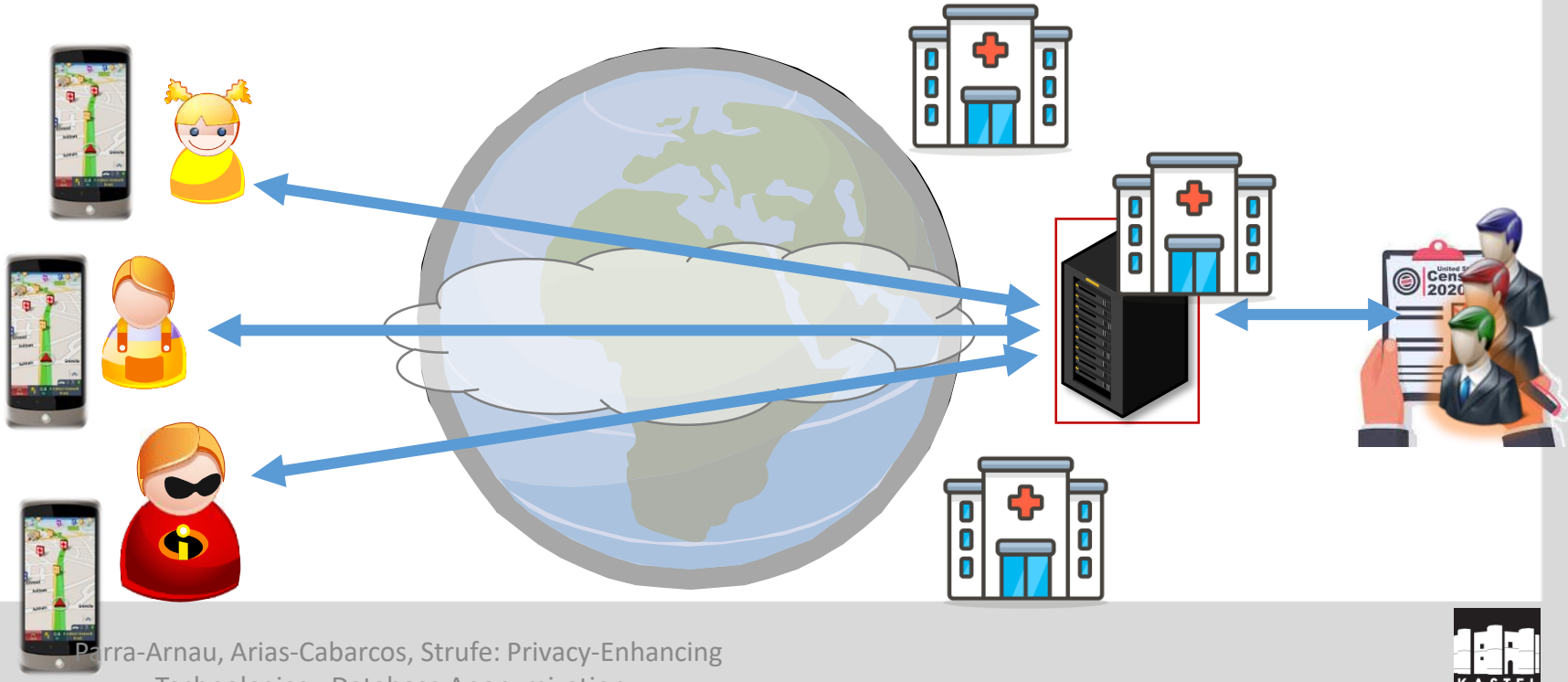
Statistical Disclosure Control

- Statistical disclosure control (SDC) is the field that protects statistical databases so that they can be released without revealing confidential information that can be linked to specific individuals among those to which the data correspond

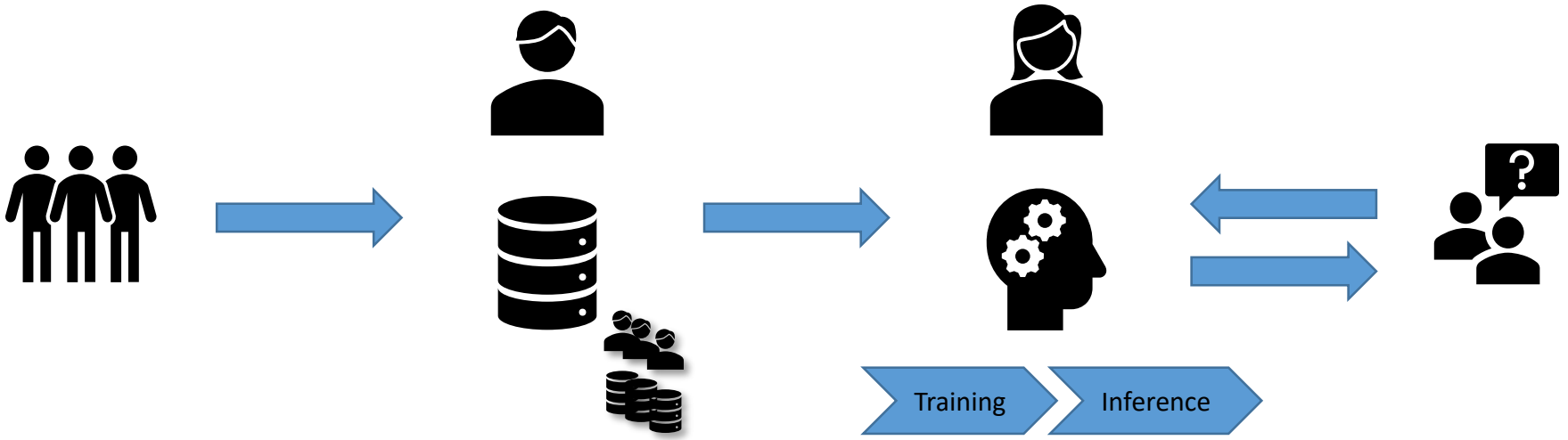
- Seek to provide useful statistical information while guaranteeing respondent privacy is not compromised

- Three formats
  - Tabular
  - Queryable
  - Microdata

- Techniques
  - Syntactic SDC
  - Semantic SDC



SDC vs PPDM vs PIR



- SDC aims to provide respondent privacy
- Privacy-preserving data mining (PPDM) seeks database owner privacy
- Private information retrieval (PIR) aims for user/analyst privacy

[1] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer and P.-P. de Wolf (2012) Statistical Disclosure Control, Wiley.

SDC applications

- Areas of application of SDC techniques include:
  - Official statistics
    - US Census Bureau
  - Health information
    - HIPAA in the U.S. and similar rules in other western countries
    - Increasing push towards medical data exchange (genomics, biosignals, …)
  - E-commerce
    - Secondary purpose is restricted

Database formats

- Tabular data
  - publish static aggregate information without disclosing confidential information on specific individuals
- Queryable databases
  - Aggregate information obtained by an analyst should not reveal information at the individual level
- Microdata
  - Perturbed the original database so as to keep the analytical usefulness of the data, while avoid respondent linkage

Tabular data protection

- Tabular data
  - Magnitude table. Sum of a particular response across a subset of respondents. E.g., turnover of all businesses of a particular industry within a region
  - Frequency table. Number of respondents satisfying certain criteria. E.g., number of respondents in a city who suffer from a given condition

Disclosure attacks on tabular data

- External attack

frequency table

Job_i x Town_j	Job_j x Town_j
1	36

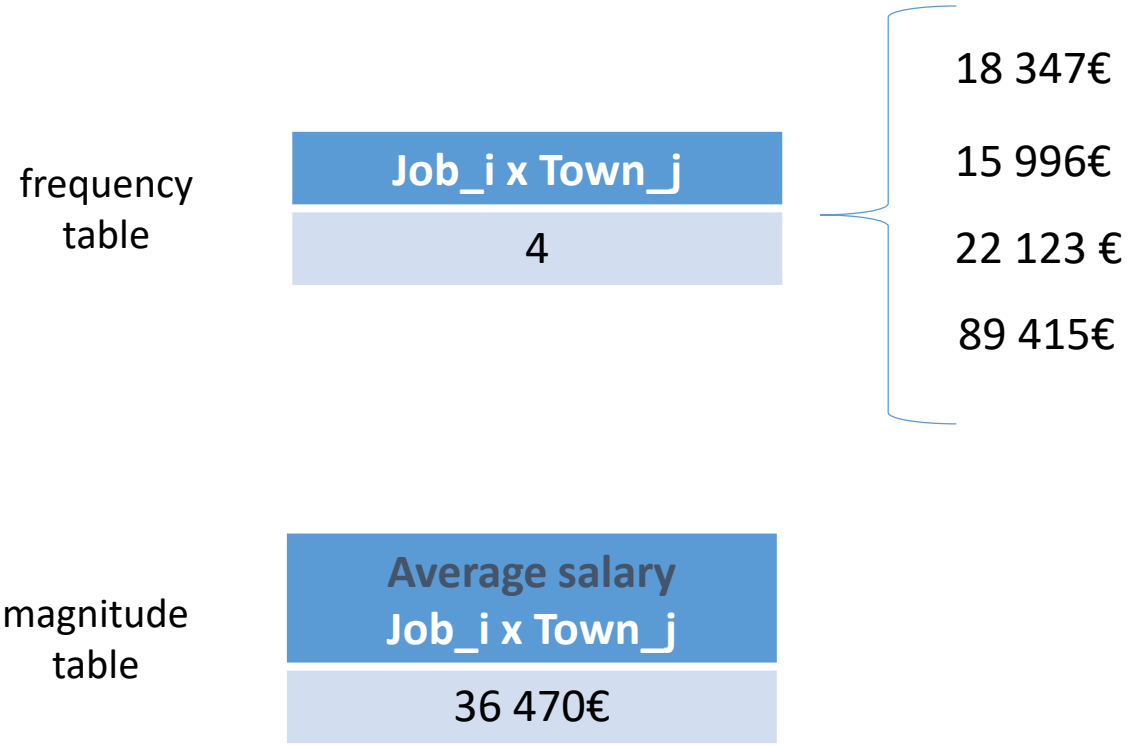
magnitude table

Average salary Job_i x Town_j	Average salary Job_j x Town_j
18 347€	45 000€

- Internal attack
  - Having two respondents is not enough

Disclosure attacks on tabular data

- Dominance attack



k-Anonymity and Disclosure Attacks on Microdata

Karlsruhe Institute of Technology

Identifying Attribute

Quasi-identifier

Sensitive attribute

a tuple

Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Female	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	53806	Flu
Eric	2/28/76	Female	53806	Hang Nail

- The information for each respondent contained in the released data set cannot be distinguished from at least $k - 1$ individuals
- Each tuple of quasi-identifier values in the released table must appear in at least *k* records

k-Anonymity

Karlsruhe Institute of Technology

original table

date of birth

Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Female	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	53806	Flu
Eric	2/28/76	Female	53806	Hang Nail

2-anonymous table

DOB	Gender	Zipcode	Disease
*	Female	5371*	Heart Disease
*	Female	5371*	Hepatitis
*	Male	5370*	Brochitis
*	Male	5370*	Broken Arm
*	Female	538**	Flu
*	Female	538**	Hang Nail

Limitations of k-anonymity

Karlsruhe Institute of Technology

Original microdata

Alice →

Naroto →

QID			SA
Zipcode	Age	Sex	Disease
47676	27	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Ovarian Cancer
47905	43	M	Heart disease
47909	52	F	Cancer
47906	47	M	Cancer

3-anonymous table

QID			SA
Zipcode	Age	Sex	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
4790*	[43,52]	*	Heart disease
4790*	[43,52]	*	Cancer
4790*	[43,52]	*	Cancer

- Suppose that the adversary knows Alice's combination of quasi-identifier attributes is (47676, 27, F). The attacker does not know which of the first 3 records corresponds to Alice's record, but learns her health condition is cancer
  - **Homogeneity attack**

- Suppose that the adversary knows Naroto's combination of quasi-identifier attributes is (47905, 47, M). The attacker learns the last record is probably Naroto's as Japanese people have low incidence of heart attacks
  - **Background knowledge attack**

KASTEL

Limitations of k-anonymity

- It prevents identity disclosure
  - The attacker cannot find out which record corresponds to a given respondent
  - however, from the previous examples, it is prone to homogeneity and background-knowledge attacks - **no privacy at all**
- But not (sensitive or confidential) attribute disclosure
  - The adversary cannot tell that a given person has a certain sensitive attribute
- Assumes which information is available for linkage or which not

p-Sensitive, k-anonymity

3-sensitive, 6-anonymous table

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

at least 3 different values of the confidential attribute

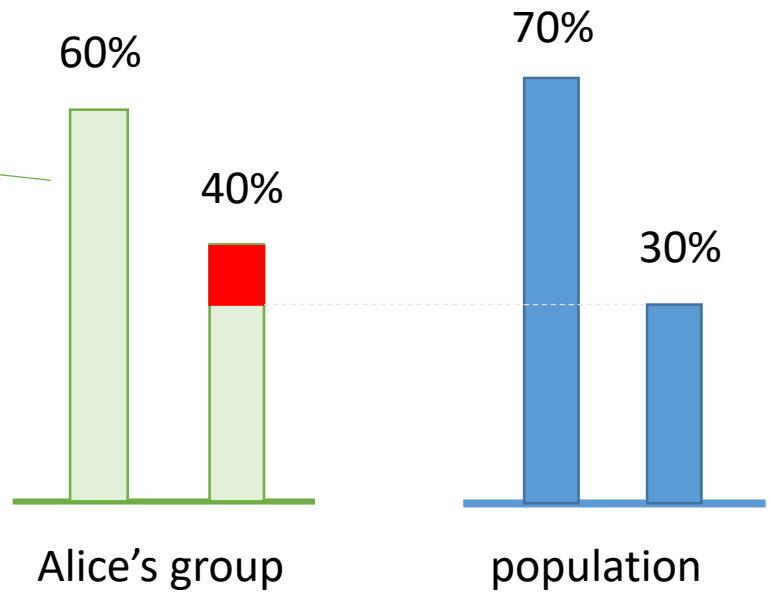
- Aimed to protect against confidential attribute disclosure
- The idea is to have at least p different sensitive values of the confidential attribute within each k-anonymous class

Limitations of p-sensitive, k-anonymity

- Prone to skewness attacks

Alice

ZIP	Age	Disease
089**	<30	N
089**	<30	Y
089**	<30	N
089**	<30	N
089**	<30	Y
283**	[30-40]	N
283**	[30-40]	N
283**	[30-40]	Y
283**	[30-40]	N
283**	[30-40]	N



l-Diversity

- The idea is that the sensitive attributes are “diverse” within each k-anonymous group
- Each equivalence class has at least l well-represented sensitive values
- Different meanings of “well-represented” values, in addition to distinct l-diversity
  - Entropy l-diversity. The entropy of the distribution of sensitive values in each equivalence class is at least log l

$$H(Z|X=x) = -\sum_{z} p_{Z|X}(z|x) \log p_{Z|X}(z|x) \geq \log l \quad \text{for all class } x$$

entropy of the confidential attribute Z on the equivalent class x

parameter

Limitations of l-diversity

- Still vulnerable to skewness attacks
- And similarity attacks…

3-diverse, 3-anonymous table

QID			SA
Zipcode	Age	Sex	Disease
476**	2*	*	Lung Cancer
476**	2*	*	Prostate Cancer
476**	2*	*	Bladder Cancer
4790*	[43,52]	*	Heart disease
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Diabetes

t-Closeness

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

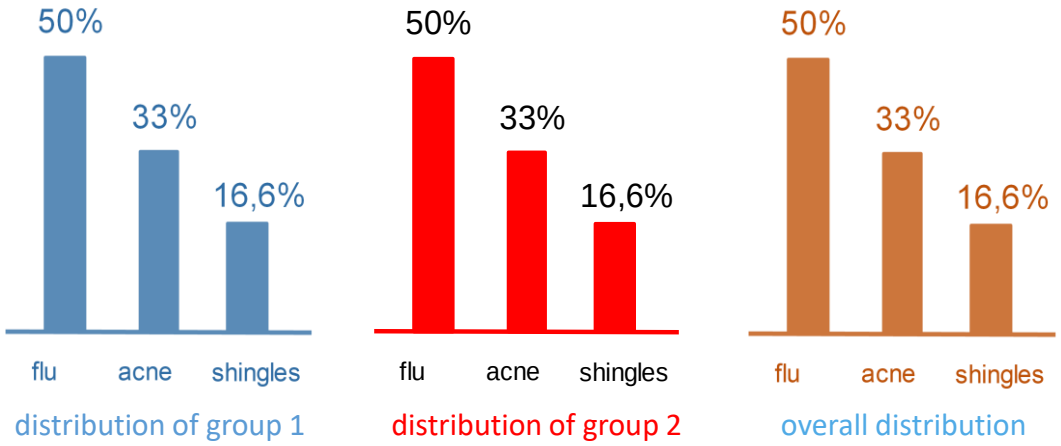
overall distribution

- The idea is that the distribution of confidential attributes given perturbed key attributes observed must be close to the entire distribution of the confidential attribute

$$\left| d(p_{Z|X}(z|x), p_Z(z)) \right| \leq t$$

confidential

group or equivalence class



Karlsruhe Institute of Technology

KASTEL

But how!?

Mechanisms to enforce data similarity (syntactic privacy notions)

Queryable databases protection

- Query perturbation
  - Deterministically correct answers not needed
  - Input vs output perturbation
- Query restriction
  - Deterministically correct answers and exact are needed
  - Refuse to answer to sensitivity queries
- Camouflage
  - Deterministically correct answers but non-exact are okay
  - Small interval answers of each confidential value

Brief overview of methods for tabular data

- Non-perturbative
  - Do not perturb or modify the values in the tables, but may eliminate or suppress them. Example include cell suppression through sensitive rules
- Perturbative
  - Output a table with some modified values. Examples include controlled rounding and controlled tabular adjustment

	Italian	Spanish	Total
City1	2	7	9
City2	5	12	17
City3	12	0	12
Total	19	19	38

	Italian	Spanish	Total
City1	0	10	10
City2	5	10	20
City3	10	0	10
Total	15	20	40

rounding base 5

Microdata protection

- Microdata are matrices of respondents per attributes
  - Numerical (e.g., weight, salary) or categorical (e.g., gender, job)

	Key Attributes		Confidential Attributes
Identifiers	Height	Weight	High Cholesterol
John Smith	5'4"	158	Y
Tang Lee	5'3"	162	Y
Luis Melo	5'6"	161	Y
Anna Frank	5'8"	157	N

Microdata protection

- Identifiers are removed, obviously
- QIs can be used to record linkage but they possess high analytical value
- Therefore anonymization algorithms must address QIs
  - privacy-utility trade-off

Key Attributes		Confidential Attributes
Height	Weight	High Cholesterol
5’4”	158	Y
5’3”	162	Y
5’6”	161	Y
5’8”	157	N

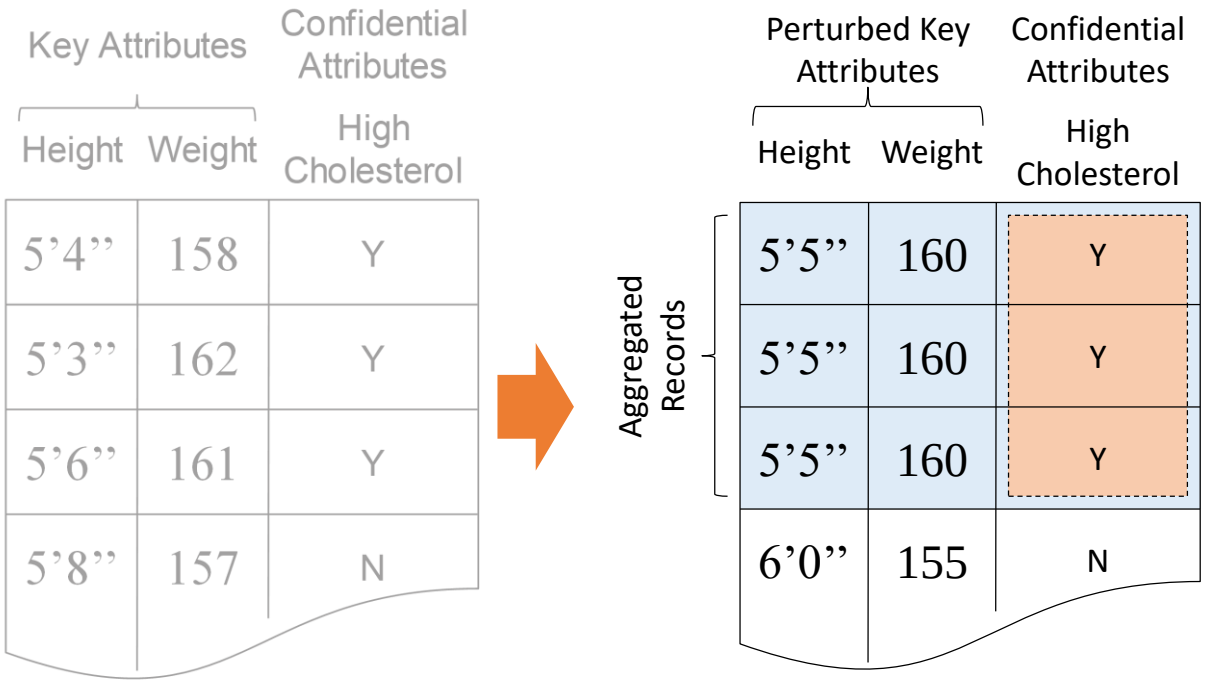
Methods for microdata protection

- **Masking** methods: generate a modified version of the original data
  - Perturbative: modify data
    - Noise addition, microaggregation, rank swapping, microdata rounding, and resampling
  - Non-perturbative: do not modify the data but rather produce partial suppressions or reductions of detail in the original dataset
    - Sampling, global recoding, top and bottom coding, and local suppression

- **Synthetic** methods: generate synthetic or artificial data with similar statistical properties

Perturbation: Microaggregation

- Mask by grouping and replacement by “mean” value



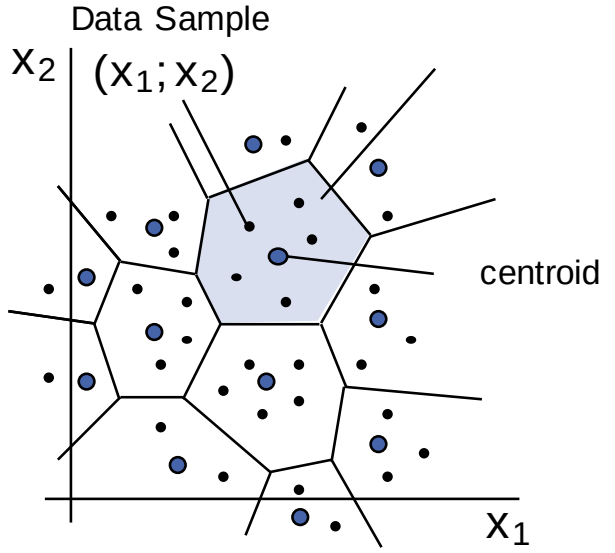
Perturbation: Microaggregation (ctd.)

- The optimal k-partition is the one maximizing the within-group homogeneity
  - The higher the within-group homogeneity, the lower the information loss

- A typical criterion to measure homogeneity in clustering is the sum of squared errors (SSE)

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)'(x_{ij} - \bar{x}_i)$$

centroid



Perturbation: Data swapping

- The idea is to transform a database by exchanging values of confidential attributes among individual records
- Information less is not reduced but may refrain participants to contribute their data

Variable	Original data			After perturbing the data		
ID	Gender	Region	Education	Gender	Region	Education
1	female	rural	higher	female	rural	higher
2	female	rural	higher	female	rural	*lower*
3	male	rural	lower	male	rural	lower
4	male	rural	lower	*female*	rural	lower
5	female	urban	lower	*male*	urban	*higher*
6	female	urban	lower	female	urban	lower

Perturbative masking – Noise addition

- Uncorrelated noise addition
  - Neither variances nor correlations are preserved
- Correlated noise addition
  - Means and correlations can be preserved
- Noise addition and linear transformation
- Noise addition and non-linear transformation

Perturbation: Differential privacy for microdata

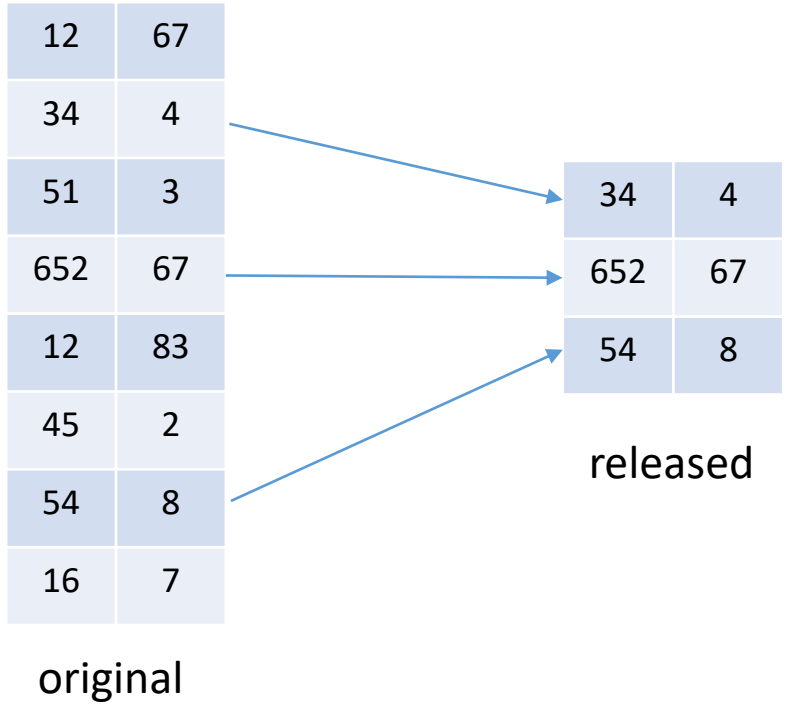
- Non-interactive scenario. Microdata available for any use without restrictions (publish tables of microdata with DP guarantees)

We'll go there later, just briefly:

- DP mechanisms are tied to query functions, depend on their sensitivity
- Naïve approach to generate DP microdata with the identity function
  - Collection of responses to the query "What's the content of the i-th record of the microdata for $i = 1, \dots, n$?"
  - L1-sensitivity of the identity function?

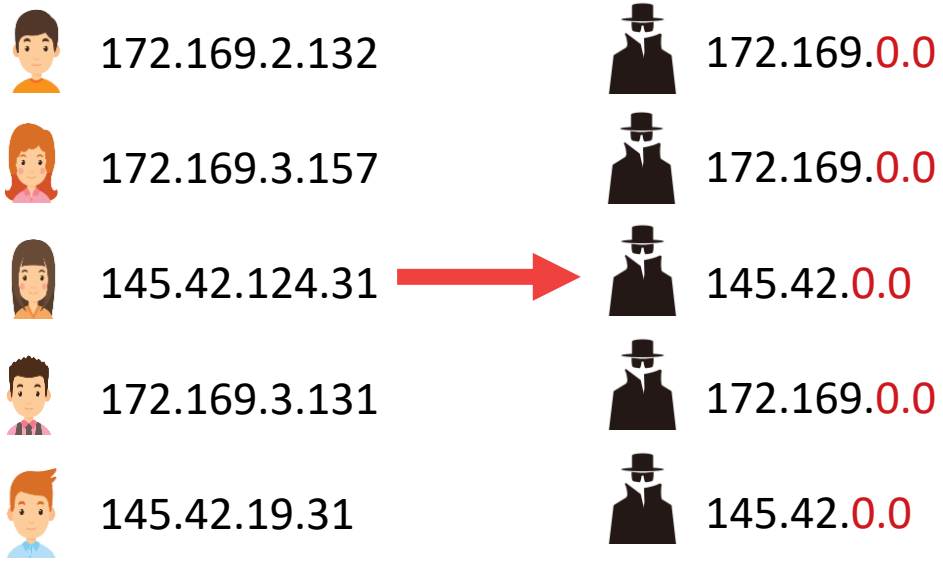
Non-Perturbative Masking: Sampling

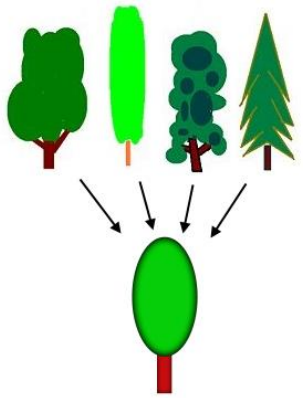
- Publish random sample of the original set of records
- Correlation determines which properties are retained (uncorrelated: none)
- Continuous numerical data need further protection



Non-Perturbative: Generalization/Coarsening

- Reduce detail of information
- Remove least-significant parts, preserve significant, but general information





Non-perturbative Masking: Global recoding

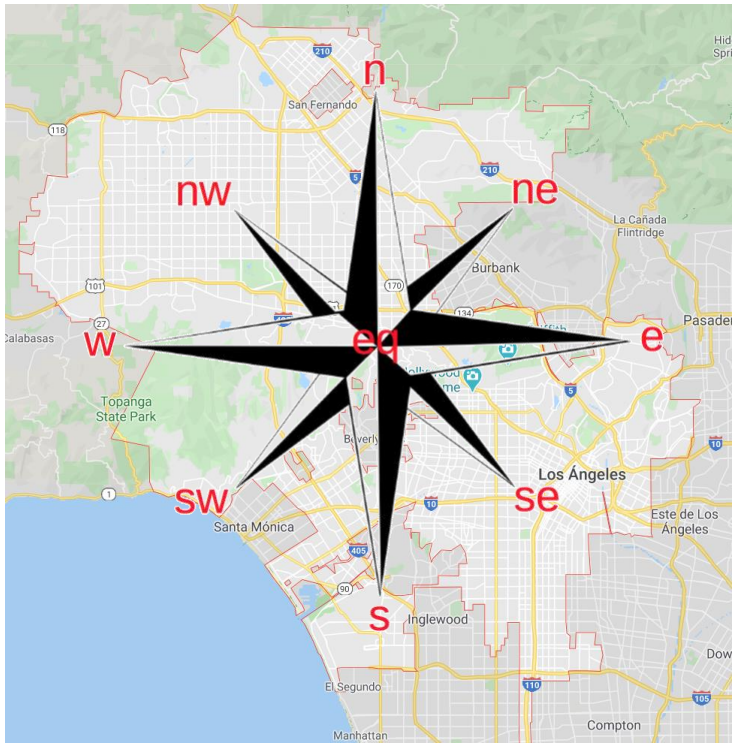
- Combine several categories of a categorical attribute or construct intervals for continuous variables
- Reduction of the level of detail and potentially the disclosure risk

A–L [edit]

- Angelino Heights[TG]
- Angeles Mesa
- Angelus Vista
- Arleta[MLA][TG]
- Arlington Heights[MLA]
- Arts District[1]
- Atwater Village[MLA]
- Baldwin Hills[TG]
- Baldwin Hills/Crenshaw[MLA]
- Baldwin Village[TG]
- Baldwin Vista[2]
- Beachwood Canyon[3]
- Bel Air, Bel-Air or Bel Air Estates[MLA][TG]
- Benedict Canyon[4]
- Beverly Crest[MLA]
- Beverly Glen[TG]
- Beverly Grove[MLA]
- Beverly Hills Post Office[5]
- Beverly Park[6]
- Beverlywood[MLA]
- Boyle Heights[MLA][TG]
- Brentwood[MLA][TG]
- Brentwood Circle[7]
- Brentwood Glen[8]
- Broadway-Manchester[MLA]
- Brookside
- Bunker Hill[1]
- Cahuenga Pass[TG]
- Canoga Park[MLA][TG]
- Canterbury Knolls[9]
- Carthay[MLA]
- Castle Heights
- Central-Alameda[MLA]
- Central City[TG]
- Century City[M
- Chatsworth[MLA
- Chesterfield Sc
- Cheviot Hills[M
- Chinatown[MLA
- Civic Center[10
- Crenshaw[TG]
- Crestwood Hill
- Cypress Park[
- Del Rey[MLA][TG
- Downtown[MLA
- Eagle Rock[TG]
- East Gate Bel
- East Hollywoo
- East Los Ange
- Echo Park[MLA
- Edendale[12]

M–Z [edit]

- Manchester Square[MLA]
- Mandeville Canyon[26]
- Marina Peninsula[27]
- Mar Vista[MLA][TG]
- Melrose Hill[28]
- Mid-City[MLA][TG]
- Mid-Wilshire[MLA]
- Miracle Mile[TG]
- Mission Hills[MLA][TG]
- Montecito Heights[MLA][TG]
- Monterey Hills[29]
- Mount Olympus[TG]
- Mount Washington[MLA][TG]
- Nichols Canyon[30]
- NoHo Arts District[31]
- North Hills[MLA][TG]
- North Hollywood[MLA][TG]
- Northridge[MLA][TG]
- North University Park[TG]
- Old Bank District[1]
- Outpost Estates[32]
- Pacific Palisades[MLA][TG]
- Pacoima[MLA][TG]
- Palms[MLA][TG]
- Panorama City[MLA][TG]
- Park La Brea[TG]
- Picfair Village[33]
- Pico Robertson[34]
- Pico-Union[MLA]
- Platinum Triangle[citation needed]
- Playa del Rey[TG]
- Playa Vista[MLA]
- Porter Ranch[MLA][TG]
- Rancho Park[MLA][TG]
- Reseda[MLA][TG
- Reynier Village
- Rose Hills
- Rustic Canyon
- San Pedro[MLA
- Sawtelle[MLA][T
- Shadow Hills[M
- Sherman Oaks
- Sherman Villag
- Silver Lake[MLA
- Skid Row[37]
- Solano Canyo
- South Central,
- South Park[MLA
- South Roberts
- Spaulding Squ
- Studio City[MLA



Source: Wikipedia, Google Maps

Non-perturbative Masking: Local suppression

- Eliminate certain values of individual attributes so as to increase the number of records sharing a combination of key-attribute values
- Oriented to categorical attributes

North	[20-30]
South	[30-40]
South	[20-30]
South-West	[20-30]
East	[30-40]
South-West	[30-40]
South	[50-60]
East	[20-30]

***	[20-30]
South	[30-40]
South	[20-30]
South-West	[20-30]
East	[30-40]
South-West	[30-40]
South	[50-60]
East	[20-30]

Intermediate Summary of Masking

- Attempt to adapt microdata to achieve privacy notion
  - (Recall k-anonymity, …)
- Potentially quite complex optimization problem with many degrees of freedom

Alice

Naroto

QID			SA
Zipcode	Age	Sex	Disease
47676	27	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Ovarian Cancer
47905	43	M	Heart disease
47909	52	F	Cancer
47906	47	M	Cancer

QID			SA
Zipcode	Age	Sex	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
4790*	[43,52]	*	Heart disease
4790*	[43,52]	*	Cancer
4790*	[43,52]	*	Cancer

Synthetic microdata generation

- **Masking** methods: generate a modified version of the original data
  - Perturbative: modify data
    - Noise addition, microaggregation, rank swapping, microdata rounding, and resampling
  - Non-perturbative: do not modify the data but rather produce partial suppressions or reductions of detail in the original dataset
    - Sampling, global recoding, top and bottom coding, and local suppression

- **Synthetic** methods: generate synthetic or artificial data with similar statistical properties

Synthetic microdata generation

- Extract chosen, preserved statistics from microdata (probabilities, distributions, ML models)
- Randomly generate data (sampling, transformation)

- Pros:
  - Possibility to generate “unlimited” data sets
  - seem to address the reidentification problem, as data are “synthetic”
- Cons:
  - Published synthetic records can match an individual’s data, if model is not private
  - Data utility limited to the statistics captured by the model

Summary

- Tabular, queryable and microdata formats
- SDC aims to protect individuals privacy while providing useful statistical information

- A common classification for mechanisms is perturbative and non-perturbative
  - Perturbative mechanisms modify the data, while non-perturbative mechanisms produce partial suppressions or reductions of detail of the data
  - Generation of synthetic data (not usually private)