FSTA 2024

Fuzzy clustering and fuzzy measures in data privacy

Vicenç Torra

January, 2024

Dept. CS, Umeå University, Sweden

Outline

1. Preliminaries: Data privacy

2. Data protection: microaggregation

3. Information loss: clustering

4. Disclosure risk: Worst case scenario

5. Summary

Preliminaries

A context:

Data-driven machine learning/statistical models

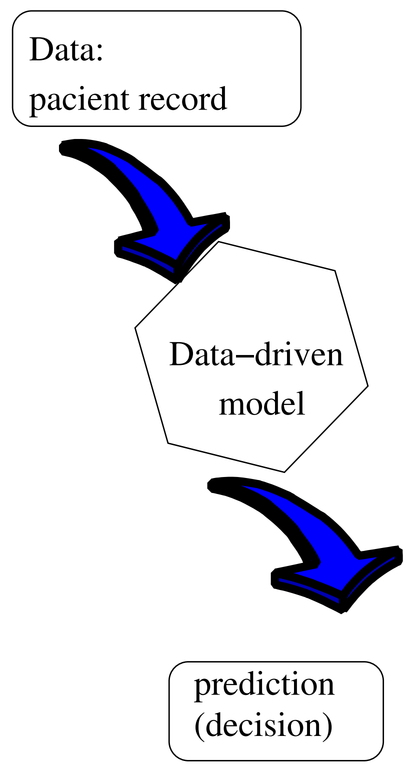
Prediction using (machine learning/statistical) models

- Data is collected to be used
  (otherwise, better not to collect them[1])

[1] Concept: Data minimization (see Privacy by Design and GDPR)

Prediction using (machine learning/statistical) models

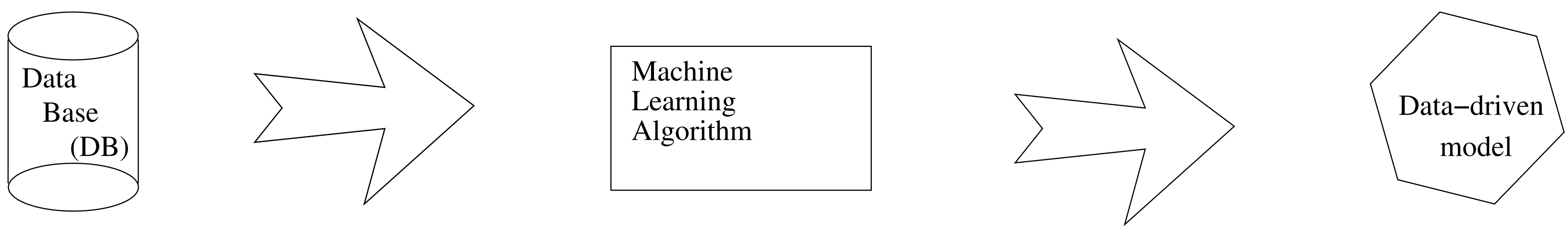
- Application of a model for decision making
  data $\Rightarrow$ prediction/decision



- Example: predict the length-of-stay at admission

Data-driven machine learning/statistical models

- From (huge) databases, build the "decision maker"
  - Use (logistic) regression, deep lerning, neural networks, . . . classification algorithms, decision trees, . . .



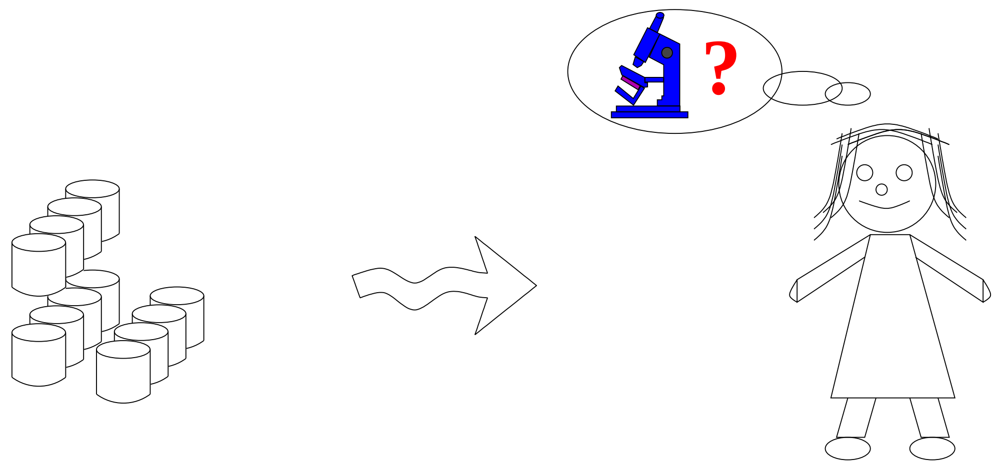
- Example: build a predictor from hospital historical data about length-of-stay at admission

Privacy for machine learning and statistics:

Data-driven machine learning/statistical models

Data is sensitive

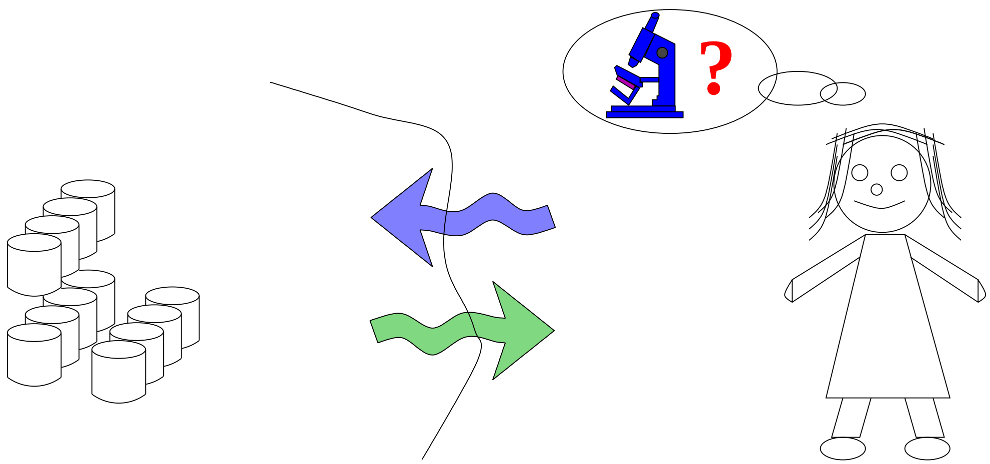
- Who/how is going to create this model (this "decision maker")?
- Case #1. Sharing (part of the data)



Data is sensitive

- Who/how is going to create this model (this "decision maker")?

- Case #2. Not sharing data, only querying data



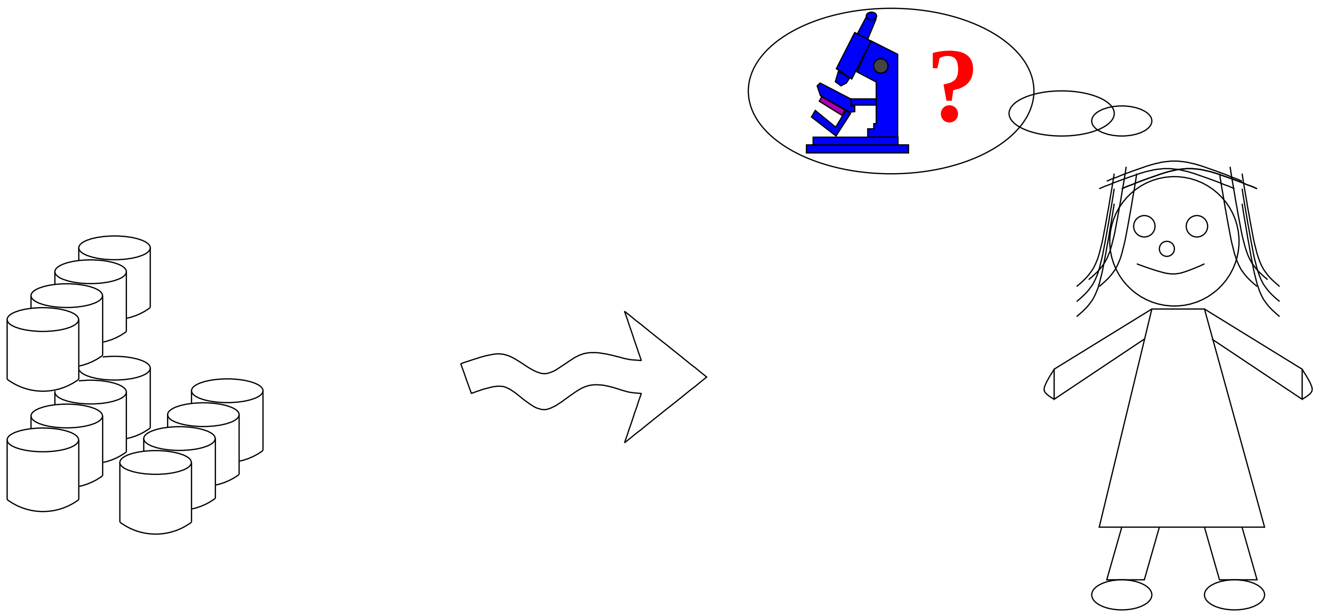
Data is sensitive

- Case #1. Sharing (part of the data)
- Q: How different children ages and diagnoses affect this length of stay? Average length of stay is decreasing in the last years due to new hospital policies?
- Data: Existing database with previous admissions (2010-2019). To avoid disclosure a view of the DB restricting records to children born before 2019 and only providing for these records year of birth, town, year of admission, illness, and length of stay.
  - ~~Anna Božena~~, Liptovská Sielnica[2], illness-1, 120 days

[2]Obyvateľstvo: 604 (2022, wikipedia)

Context: Data privacy

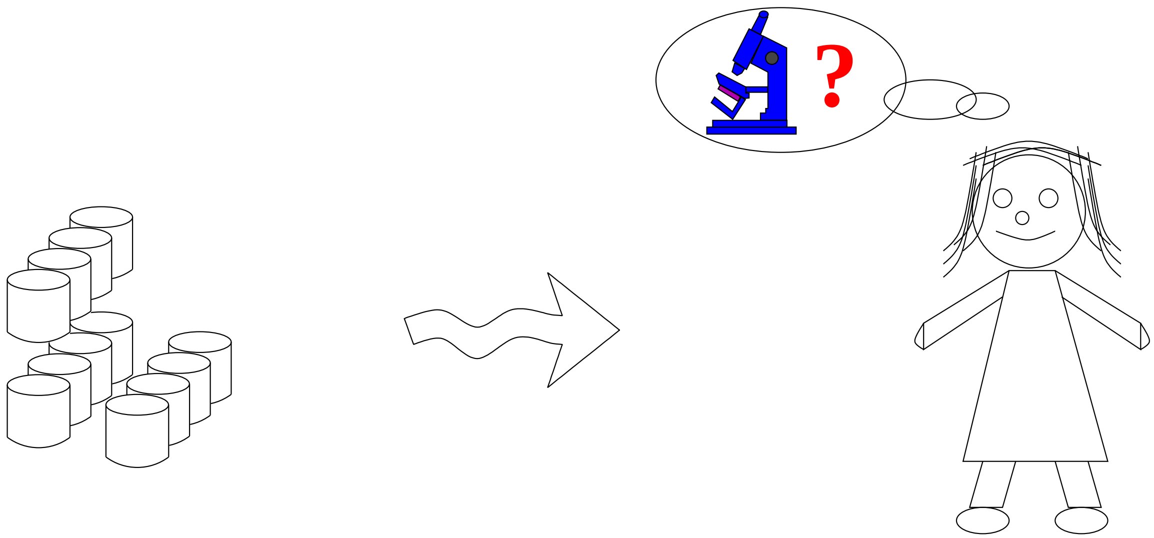
Data privacy in context. A researcher wants to analyze data



DB = {(*Hana*, *Age* = 40, *Town*=*Liptovský Ján*, *salary*=*1800 EUR*), ...}

Context: Data privacy

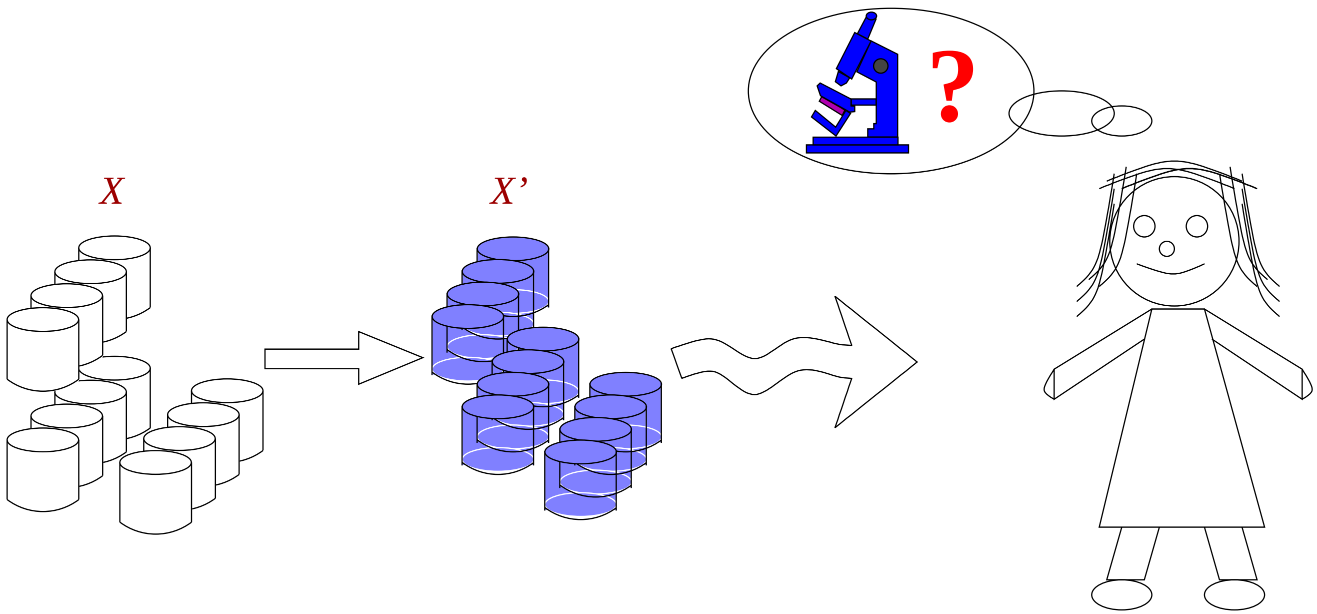
- Identity disclosure, find Hana in the database



$DB = \{($~~Hana~~$, Age = 40, Town=$Liptovský Ján$, salary=1800\ EUR), \ldots\}$

Context: Data privacy

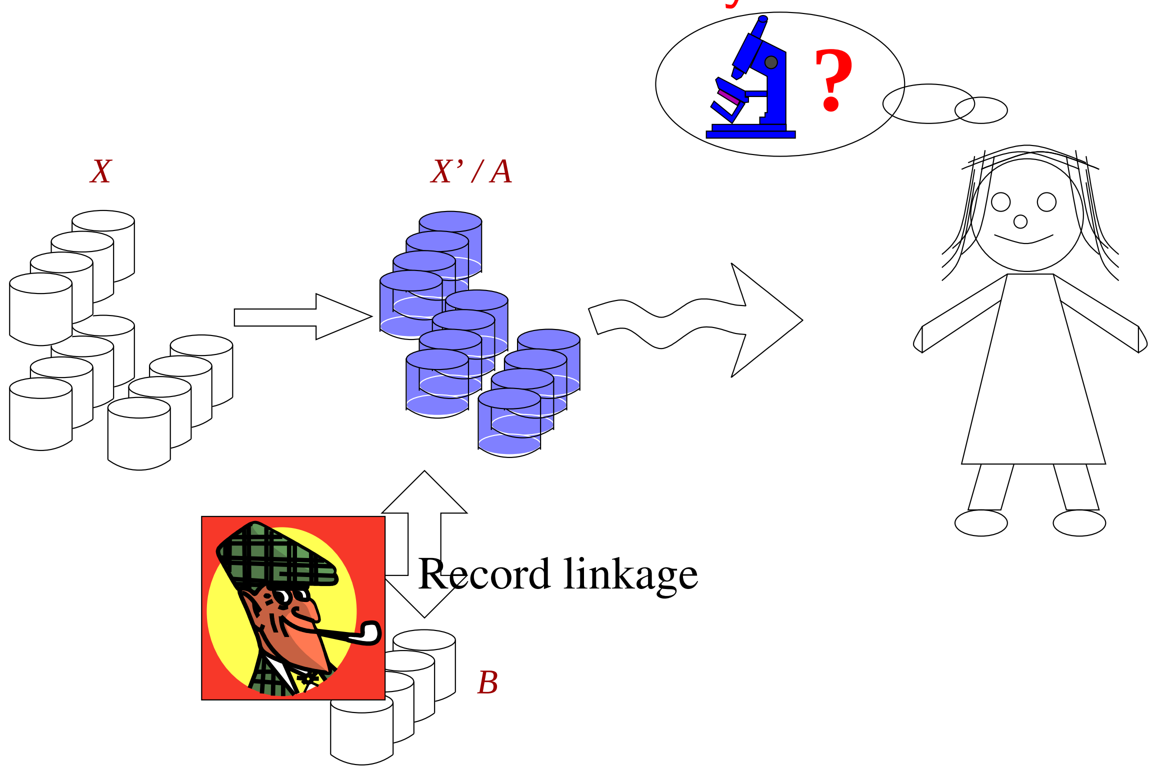
- To avoid disclosure, remove identifiers, anonymize records / modify records



DB = {(~~*Hana*~~, *Age* = *41*, *Town*=*Liptovský Mikuláš district*, *salary*=*1800 EUR*), ...}

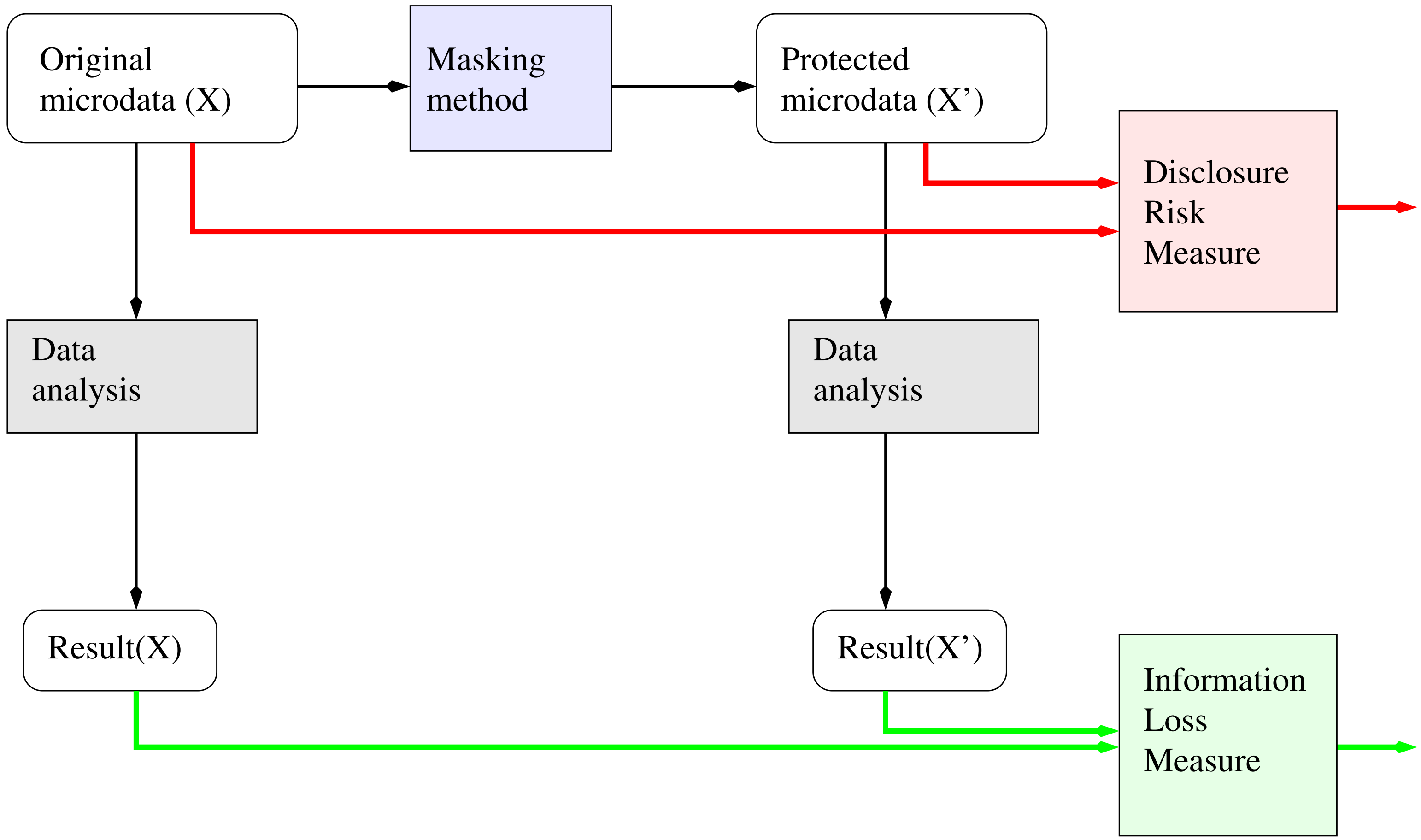
Context: Identity disclosure risk in data privacy

- Q1: Protection: How to obtain X'?
- Q2: Identity disclosure risk by modeling an intruder attack
  - How many records in B can be correctly linked to X'



- Q3: Is data useful? Information loss measures

Data-driven protection methods



Data protection

Microaggregation

Microaggregation

- Informal definition. Small clusters are built for the data, and then each record is replaced by a representative.

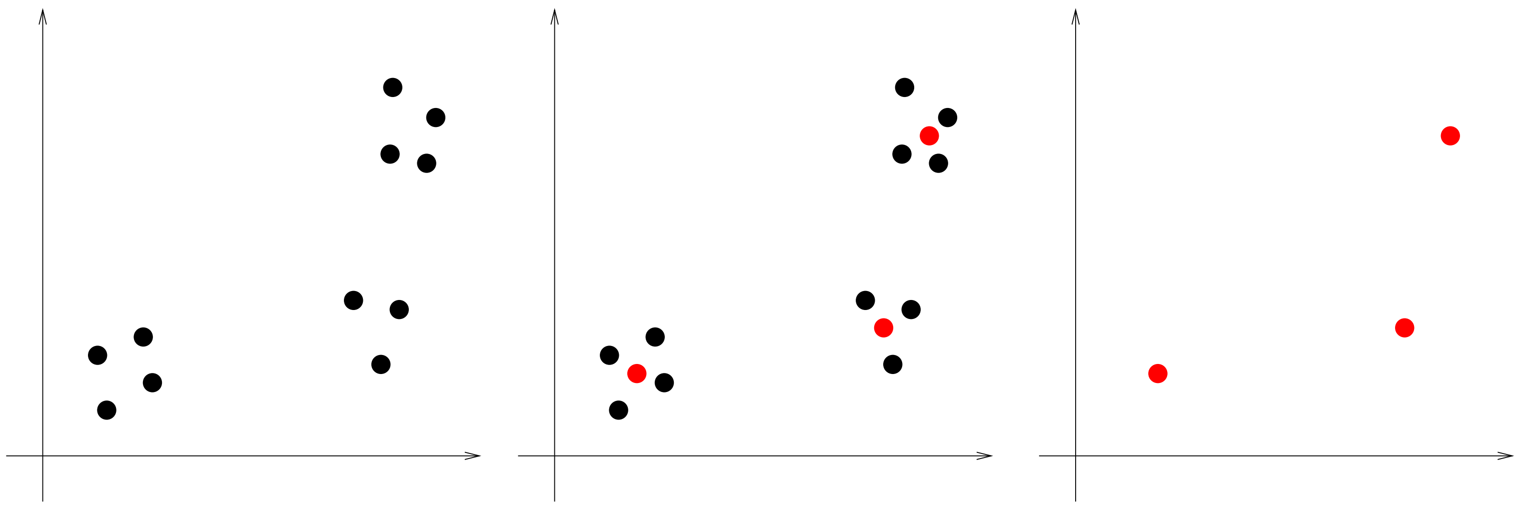
Microaggregation

- Informal definition. Small clusters are built for the data, and then each record is replaced by a representative.

- Disclosure risk and information loss
  - Low disclosure is ensured requiring k records in each cluster
  - Low information loss is ensured as clusters are small

Microaggregation

- Graphical representation of the process.



Microaggregation

- Formalization. u_{ij} to describe the partition of the records in X. That is, $u_{ij} = 1$ if record j is assigned to the ith cluster. v_i be the representative of the ith cluster.

- k is the minimum **size** of the cluster
  $c = |X|/k$ (approx.)

Minimize $SSE = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}(d(x_j, v_i))^2$

Subject to $\sum_{i=1}^{c} u_{ij} = 1$ for all $j = 1, \ldots, n$

$2k \geq \sum_{j=1}^{n} u_{ij} \geq k$ for all $i = 1, \ldots, c$

$u_{ij} \in \{0, 1\}$

Microaggregation

- Discussion
  - A good method in terms of the privacy-utility trade-off
  - Similar as k means with a constraint on k
  - Small k: low privacy, low information loss
  - Large k: high privacy, large information loss

- Inconvenient:
  - Easy to attack, given some information one can guess the cluster
  - Independent microaggregation of variables + intersection attacks: it can lead to reidentification

Fuzzy microaggregation

Fuzzy microaggregation

- Goal
  - Make membership to a cluster uncertain
  - As a side effect, outliers weight to cluster centers will be reduced
  - Provide a transparency-aware protection mechanism

Fuzzy microaggregation

- Introduce fuzziness in the clusters
  - **Approach 1.** Methods trying to keep the constraint on the number of records k. Recursive partitive methods. Partitioning large clusters into smaller ones, until an appropriate size is achieved.
  - **Approach 2.** Simple method based on fuzzy c-means.

Fuzzy microaggregation

- Introduce fuzziness in the clusters (FCM-like)

Minimize $SSE = \sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij})^m (d(x_j, v_i))^2$

Subject to $\sum_{i=1}^{c} u_{ij} = 1$ for all $j = 1, \ldots, n$

$u_{ij} \in [0, 1]$

Fuzzy microaggregation

- Introduce fuzziness in the clusters (FCM-like)

  Minimize $SSE = \sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij})^m (d(x_j, v_i))^2$
  Subject to $\sum_{i=1}^{c} u_{ij} = 1$ for all $j = 1, \ldots, n$
  $u_{ij} \in [0, 1]$

- m is the degree of fuzziness
  - $m = 1$ crisp solution
  - $m >> 1$ very much fuzzy solution

Fuzzy microaggregation

- Introduce fuzziness in the clusters (FCM-like)

  Minimize $SSE = \sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij})^m (d(x_j, v_i))^2$
  Subject to $\sum_{i=1}^{c} u_{ij} = 1$ for all $j = 1, \ldots, n$
  $u_{ij} \in [0, 1]$

- m is the degree of fuzziness
  - $m = 1$ crisp solution
  - $m >> 1$ very much fuzzy solution

- Solved using (iterative) alternate optimization: (1) u_{ij}, (2) v_i

Fuzzy microaggregation

- Introduce fuzziness in the clusters (FCM-like)
- m is the degree of fuzziness
- When computing the solution:
  - $m = 1$ crisp solution, clusters are clearly disjoint, data only affects the nearest cluster centroid
  - $m >> 1$ all clusters are overlapping all data affects all cluster centroids (and, thus, $v_i = v_j = \bar{X}$)

Fuzzy microaggregation

- Introduce fuzziness in the clusters (FCM-like)
- m is the degree of fuzziness
- When using the solution as classification rule:
  - $m = 1$ crisp solution, a point is only classified to a single class
  - $m >> 1$ a point assigned to all classes with membership $u_{ij} = 1/c$
- i.e., classification rule:

$$u_i(x) = \left(\left(\sum_{r=1}^{c} \frac{||x - v_i||^2}{||x - v_r||^2} \right)^{\frac{1}{m-1}} \right)^{-1}$$

Fuzzy microaggregation

- Introduce fuzziness in the clusters (FCM-like)
- m is the degree of fuzziness
- We decouple m in clustering with m in membership computation
  - m_1 for computing clusters and cluster centers
  - m_2 for membership assignment

Fuzzy microaggregation

- Algorithm
  - Apply FCM with m_1
  - Recompute membership of points to clusters with m_2
  - Assign points to clusters probabilistically (using membership functions)
  - Replace original data by cluster centers ($X' = \rho(X)$)

Fuzzy microaggregation

- Properties
  - Maximum utility, no protection. $m_1 = 1$, $m_2 = 1$, $c = |X|$
  - The larger the m_1, the larger the protection, larger info. loss
    $X' = \bar{X}$
  - The larger the m_2, the larger the protection, larger info. loss
    x_j can be assigned to any cluster (same probability $1/c$).
    k-anonymity is probabilistically satisfied
  - The smaller the c, the larger the protection, larger info. loss
  - Isolated points can cause problems,
    fuzzy cluster robust to outliers
  - Experiments: $m_1 = 1.1$, $m_2 = 1.2$ were quite good

Fuzzy microaggregation and constraints

Fuzzy microaggregation with constraints

- Properties
  - Constraints on the data

$$net + tax = gross$$

  - Protection needs to satisfy constraints $X = \rho(X)$
  - Even if data does not satisfy constraints, protected data should
- Several approaches for different type of protection mechanisms
  - Noise addition
  - Approach based on functional equations[3]
  - Microaggregation (FCM-based) with constraints

[3] VT (2008) Constrained Microaggregation: Adding Constraints for Data Editing, Trans. Data Privacy

Fuzzy microaggregation with constraints

- New optimization problem

Minimize $SSE = \sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij})^m (d(x_j, v_i))^2$

Subject to $\sum_{i=1}^{c} u_{ij} = 1$ for all $j = 1, \ldots, n$

$\alpha \cdot v_i = A$ for all $i = 1, \ldots, c$

$u_{ij} \in [0, 1]$

Fuzzy microaggregation with constraints

- New optimization problem

$$\begin{array}{ll} \text{Minimize} & SSE = \sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij})^m (d(x_j, v_i))^2 \\ \text{Subject to} & \sum_{i=1}^{c} u_{ij} = 1 \text{ for all } j = 1, \ldots, n \\ & \alpha \cdot v_i = A \text{ for all } i = 1, \ldots, c \\ & u_{ij} \in [0, 1] \end{array}$$

- m is the degree of fuzziness
- α are the coefficients of the constraints
  $\alpha \cdot v_i = A$

Fuzzy microaggregation with constraints

- Optimization problem, to be solved using an alternate optimization algorithm
  - Mimizing w.r.t. u_{ij}

$$u_{ij} = \left(\left(\sum_{r=1}^{c} \frac{||x_j - v_i||^2}{||x_j - v_r||^2} \right)^{\frac{1}{m-1}} \right)^{-1}$$

  - Minimizing w.r.t. v_{is} (s is the sth position in vector v_i)

$$v_{is} = \frac{\sum_{k=1}^{n}(u_{ik})^m x_{ks} - \alpha_s \frac{\sum_{k=1}^{n}(u_{ik})^m[\alpha^T x_k - A]}{\alpha^T \alpha}}{\sum_{k=1}^{n}(u_{ik})^m}$$

Fuzzy microaggregation with constraints

- Properties
  - When $\alpha_s = 0$, the Equation reduces to FCM case for s
  - When data already satisfies linear constraints, the Equation reduces to FCM case

Fuzzy microaggregation with constraints

- Properties (similar as before)
  - Maximum utility, no protection. $m_1 = 1$, $m_2 = 1$, $c = |X|$
  - The larger the m_1, the larger the protection, larger info. loss $X' = \bar{X}$
  - The larger the m_2, the larger the protection, larger info. loss x_j can be assigned to any cluster (same probability $1/c$). k-anonymity is probabilistically satisfied
  - The smaller the c, the larger the protection

Fuzzy microaggregation with constraints

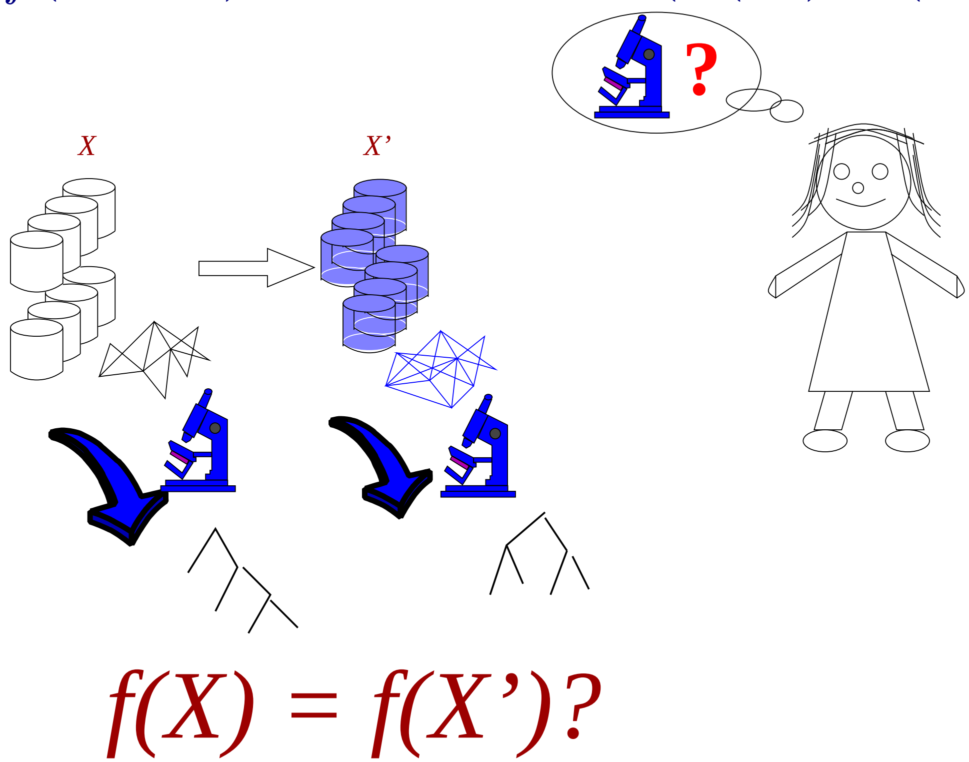
- Applied the same approach for Entropy-based Fuzzy c-Means

Information loss

Information loss

- Fuzziness in Information loss.
  - Compare X and X' w.r.t. analysis (f)

$$IL_f(X, X') = divergence(f(X), f(X'))$$



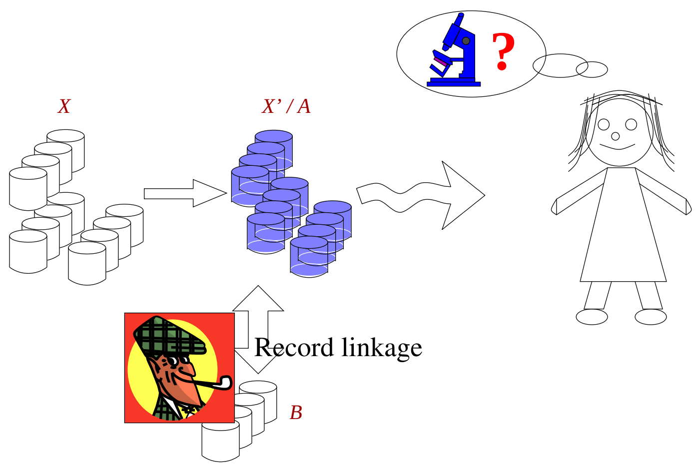
  - f is fuzzy clustering.
  - Difficulty: How to compare fuzzy clusters? (fuzzy clust. suboptimal)

Information loss

- Fuzziness in Information loss.
  - Compare X and X' w.r.t. analysis (f)[4]
    - ▷ *X = {(Hana, Age = 40, Town=Liptovský Ján, salary=1800 EUR), ...}*
    - ▷ *X' = {(~~Hana~~, Age = 41, Town=Liptovský Mikuláš district, ...}*
  - $IL_{FCM}(X, X')$=divergence(fuzzy clustering(X), fuzzy clustering(X'))

[4] V Torra, Y Endo, S Miyamoto (2009) On the Comparison of Some Fuzzy Clustering Methods for Privacy Preserving Data Mining: Towards the Development of Specific Information Loss Measures, Kybernetika 45:3 548-560

Disclosure risk assessment



Context: Identity disclosure risk in data privacy

- Identity disclosure risk measure
  - Worst case scenario = the most conservative estimation of risk
  - Worst case scenario / maximum knowledge:
    - ▷ Best information $B = X$
    - ▷ Best knowledge on the protection process: transparency attacks
    - ▷ Best record linkage algorithm:
      - Best record linkage algorithm: distance-based record linkage
      - Best parameters: distance
  - Best means: the most possible number of reidentifications
    The more the better (for an intruder)

Context: Identity disclosure risk in data privacy

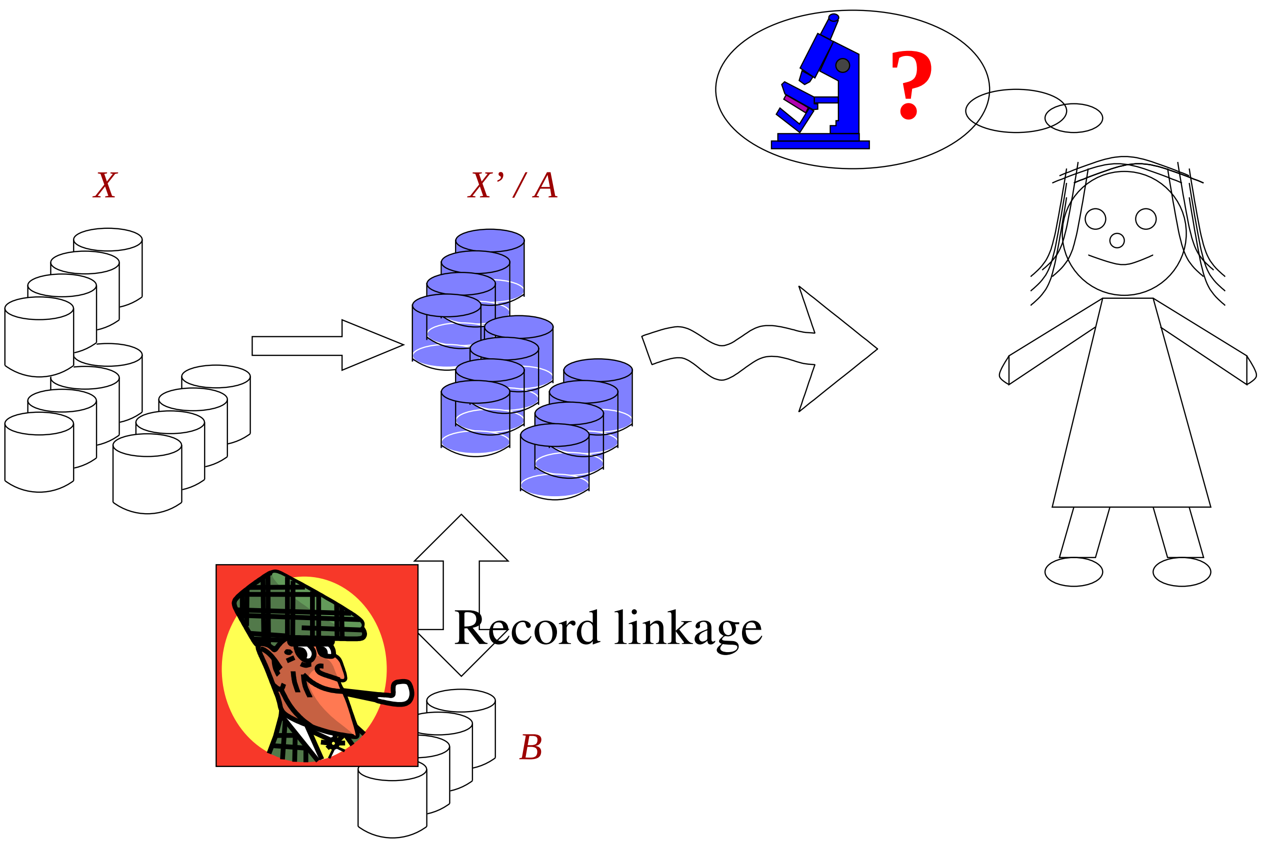
- Can we do better than with the Euclidean distance?
- Other options:
  - Weighted Euclidean distance (weights w) d_w
  - Mahalanobis distance (using covariance matrix Q)
- But also
  - Choquet integral (measure μ) d_μ
  - Bilinear forms (using positive definite matrix Q) d_Q

Context: Identity disclosure risk in data privacy

- Can we do better than with the Euclidean distance?
- Other options:
  - Weighted Euclidean distance (weights w) d_w
  - Mahalanobis distance (using covariance matrix Q)
- But also
  - Choquet integral (measure μ) d_μ
  - Bilinear forms (using positive definite matrix Q) d_Q
- Num. Reidentifications $d_\mu \geq$ Num. Reid. $d_w \geq d$

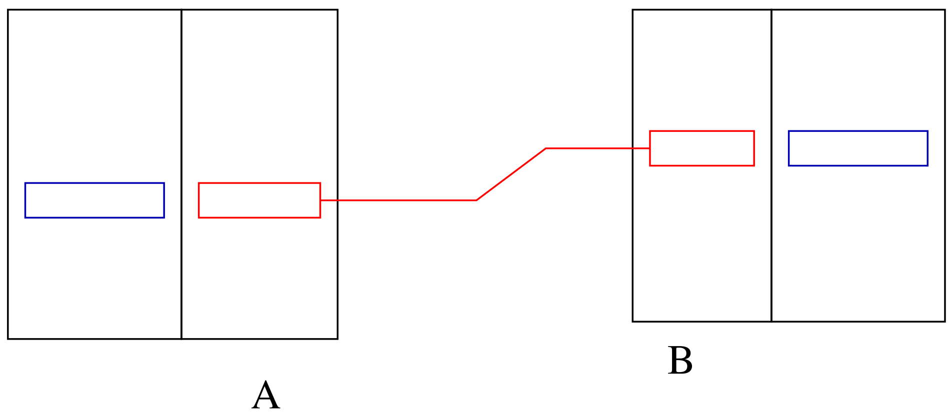
Context: Identity disclosure risk in data privacy

- How to find these parameters (μ and Q)?
- For risk analysis of a protected file X', we know both X and $A = X'$
- So, find best parameters using optimization (and $B = X$)



Context: Identity disclosure risk in data privacy

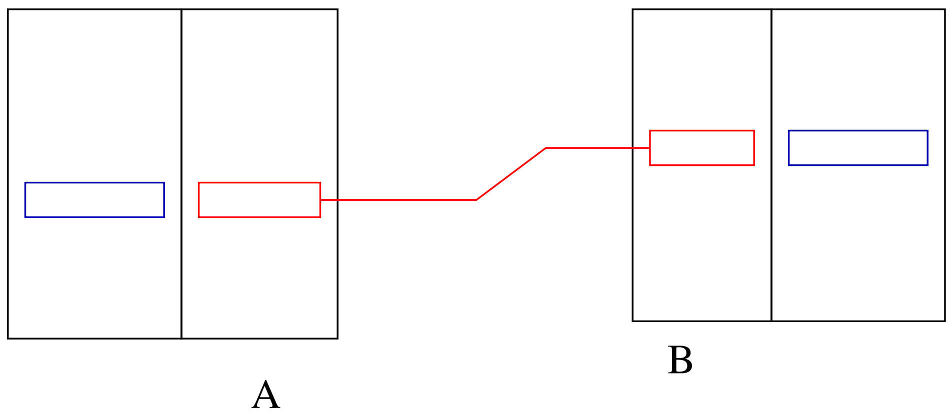
- Distance based record linkage: $d(A_i, B_i)$



- Find the *nearest* record (*nearest* in terms of a distance)
- Formally, 2 sets of vectors
  $A_i = (a_1, \dots, a_N)$,
  (a_i protected version of b_i)
  $B_i = (b_1, \dots, b_N)$
- $V_k(a_i)$: kth variable, ith record
- Distance $d(V_k(a_i), V_k(b_j))$ for all pairs (a_i, b_j).

Context: Identity disclosure risk in data privacy

- Distance based record linkage: $d(A_i, B_i)$



  - Find the *nearest* record (*nearest* in terms of a distance)
  - Formally, 2 sets of vectors $A_i = (a_1, \ldots, a_N)$, (a_i protected version of b_i) $B_i = (b_1, \ldots, b_N)$
  - $V_k(a_i)$: kth variable, ith record
  - Distance $d(V_k(a_i), V_k(b_j))$ for all pairs (a_i, b_j).

- Distance based on aggregation functions $\mathbb{C}$
  E.g., $\mathbb{C} = CI$ (Choquet integral)

- Worst-case scenario: learn weights/fuzzy measure
  $\rightarrow$ Optimization problem

Context: Identity disclosure risk in data privacy

- Distance based record linkage: $d(A_i, B_i)$
  - Main constraint: for a given i, for all j

$$\sum_{k=1}^{N} p_i d(V_k(A_i), V_k(B_j)) > \sum_{k=1}^{N} p_i d(V_k(A_i), V_k(B_i))$$

For aligned files A and B (i.e., A_i corresponds to B_i)

- As this is sometimes impossible to satisfy for all i, introduce K_i which means $K_i = 1$ incorrect linkage, and then

$$\sum_{k=1}^{N} p_i (d(V_k(A_i), V_k(B_j)) - d(V_k(A_i), V_k(B_i))) + CK_i > 0$$

Context: Identity disclosure risk in data privacy

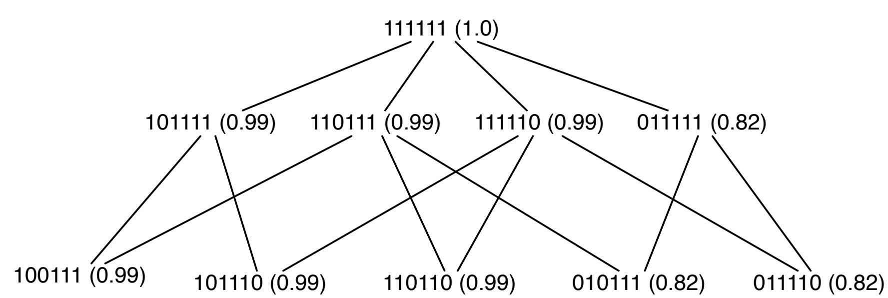
- Case $\mathbb{C} = WM$:

$$
\begin{aligned}
Minimise \quad & \sum_{i=1}^{N} K_i \\
Subject\ to: & \\
& \sum_{k=1}^{N} p_i (d(V_k(a_i), V_k(b_j)) - d(V_k(a_i), V_k(b_i))) + CK_i > 0 \\
& K_i \in \{0, 1\} \\
& \sum_{i=1}^{N} p_i = 1 \\
& p_i \geq 0
\end{aligned}
$$

- Similar with $\mathbb{C} = CI$ (Choquet integral) and μ
- Extensive work comparing different scenarios and $\mathbb{C}$.

Context: Identity disclosure risk in data privacy

- Results give:
  - number reidentifications in the worst-case scenario
  - Importance of weights (or sets of weights in fuzzy measures)
- Examples:
  - Choquet integral



  - Weighted Mean (WM):
    - ▷ V_1 0.016809573957189, V_2 0.00198841786482128, V_3 0.00452923777074791
    - ▷ V_4 0.138812880222131, V_5 0.835523953314578, V_6 0.00233593687053289

Identity disclosure

- **Privacy from re-identification**. Worst-case scenario.
  - ML for DBRL parameters: Distances considered $\mathbb{C}$
    - Weighted mean.
      Weights: importance to the attributes
      Parameter: weighting vector $n = \#$ attributes

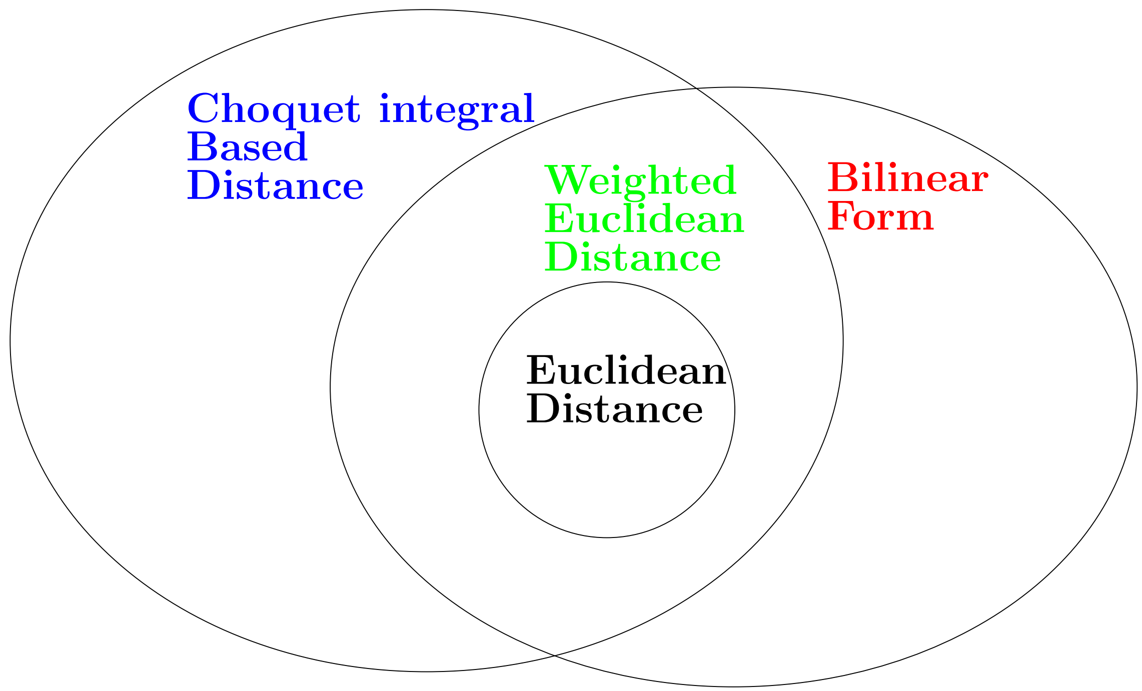
Identity disclosure

- **Privacy from re-identification**. Worst-case scenario.
  - ML for DBRL parameters: Distances considered $\mathbb{C}$
    - ▷ Weighted mean.
      Weights: importance to the attributes
      Parameter: weighting vector $n =\#$ attributes
    - ▷ OWA - linear combination of order statistics (weighted):
      Weights: to discard lower or larger distances
      Parameter: weighting vector $n =\#$ attributes
    - ▷ Bilinear form - generalization of Mahalanobis distance
      Weights: interactions between pairs of attributes
      Parameter: square matrix: $n \times n$ ($n =\#$ attributes)
    - ▷ Choquet integral.
      Weights: interactions of sets of attributes ($\mu : 2^X \to [0,1]$)
      Parameter: non-additive measure: $2^n - 2$ ($n =\#$ attributes)

Identity disclosure

Distances used in record linkage based on aggregation operators

- Graphically



Bilinear form. Quadratic form that generalizes Mahalanobis distance. Choquet integral. A fuzzy integral w.r.t. a fuzzy measure (non-additive measure). CI generalizes Lebesgue integral. Interactions.

Summary

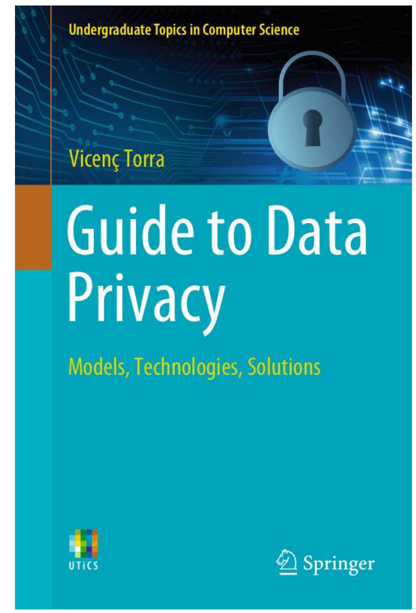
Summary

- Results presented
  - Fuzzy clustering for data protection (microaggregation)
  - Information loss using fuzzy clustering
  - Distance for fuzzy measures (reidentification, disclosure risk)

References

References

- V. Torra, G. Navarro-Arribas (2020) Fuzzy meets privacy: a short overview, Proc. INFUS 2020.
- V. Torra (2022) Guide to Data Privacy, Springer.



Thank you