Supervised Learning Approach for Distance Based Record Linkage as Disclosure Risk Evaluation

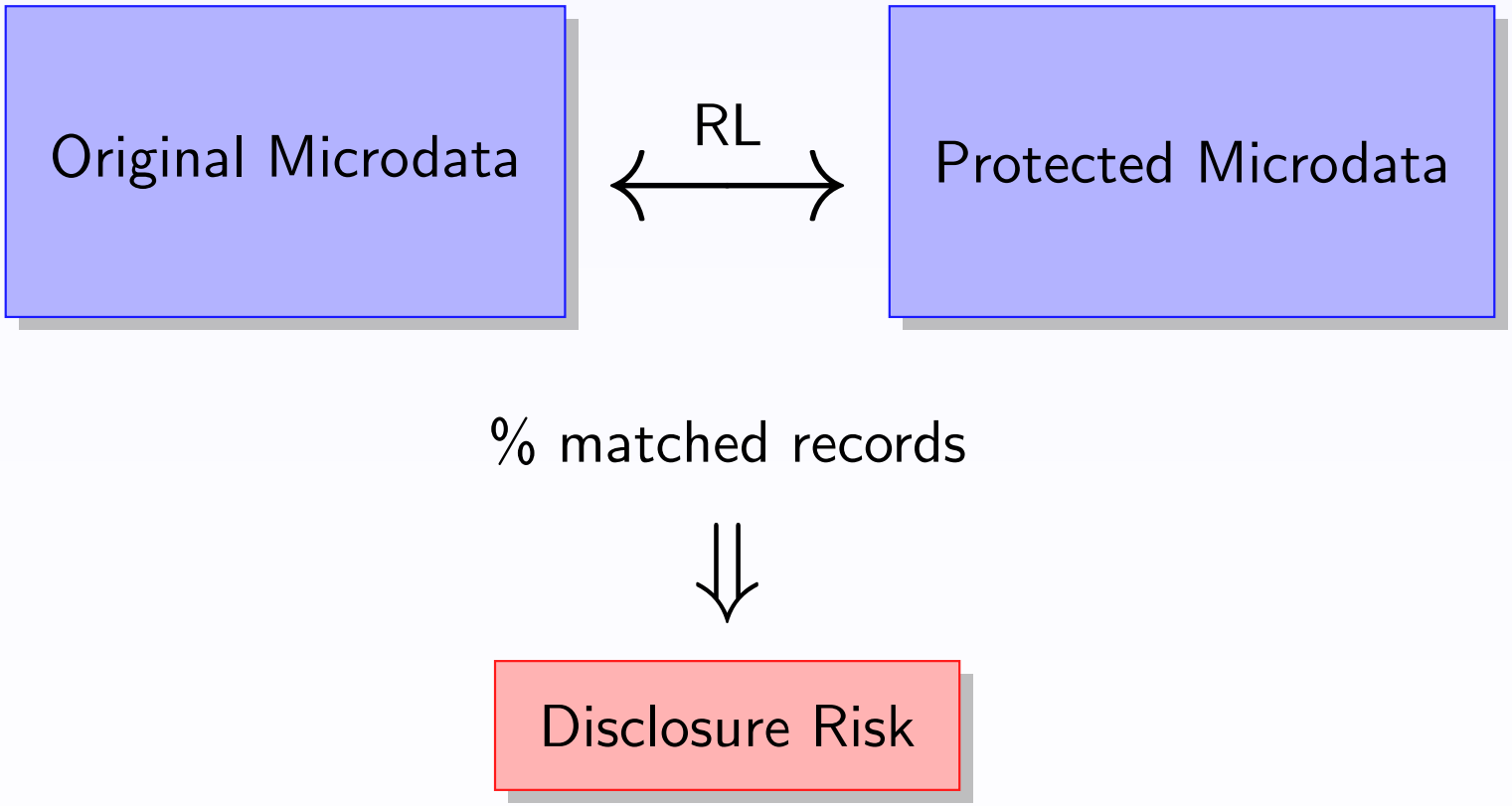
Vicenç Torra[1] Guillermo Navarro-Arribas[2] Daniel Abril[1]

[1]Artificial Intelligence Research Institute (IIIA),
Spanish council for Scientific Research (CSIC)

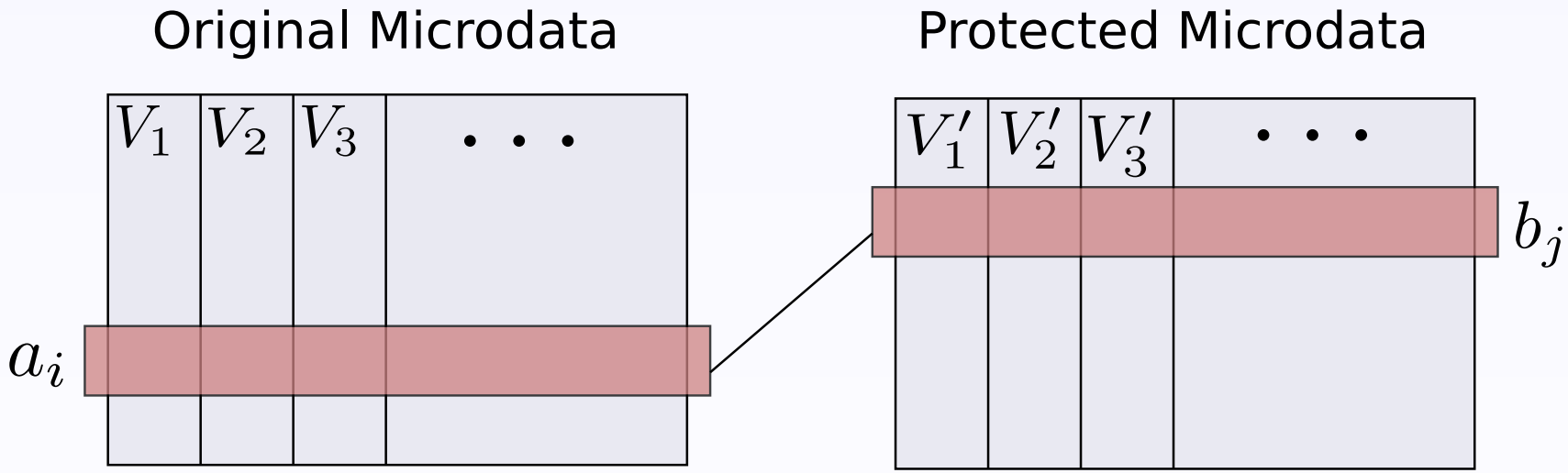
[2]Department of Information and Communications Engineering (DEIC), Universitat Autònoma de Barcelona (UAB)

Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality
Tarragona, 26-28 October 2011

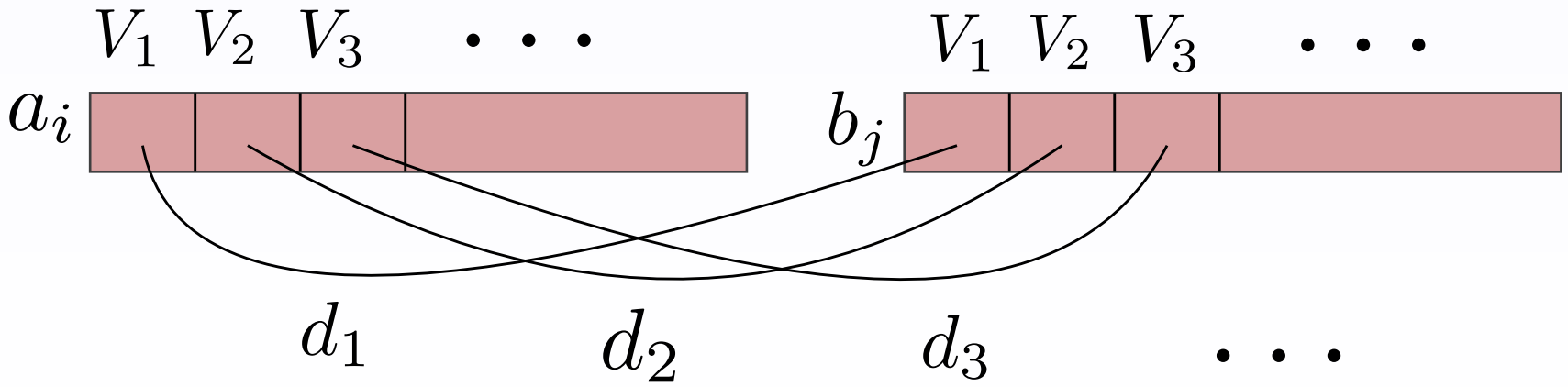
Record Linkage (RL) for Disclosure Risk evaluation



Distance-based Record Linkage (DBRL)



Per-attribute distance d_i:



Aggregation of distances

Record distance

$$d(a,b)^2 = \mathbb{C}(d_1(a,b)^2, \ldots, d_n(a,b)^2).$$

$\mathbb{C}$	Variable weighting
Arithmetic Mean (d^2AM)	None
Weighted Mean (d^2WM)	Uniform
Choquet Integral (d^2CI)	Fuzzy measure
Mahalanobis Distance (d^2MD)	Covariance-like matrix

Results

- Determine weights by **supervised learning**.
- Improves the re-identification percentage: best results for sets where attributes have different protection degrees.

	d^2AM	d^2WM	d^2CI	d^2MD
M5-38	0.3975	0.905	0.9125	0.9225
M6-385	0.78	0.9925	0.9975	0.9975

- Learning process determines **key attributes** (more weighted).
- Computation time has to be considered.