Funded by the European Union

RECITALS
Privacy-Preserving Data Sharing & ID Management

Privacy-Preserving Record Linkage: Past, Present and Yet-to-Come

EDBT 2026
Tampere, Finland

L. Stetsikas, D. Karapiperis, G. Papadakis,
and M. Koubarakis

Outline

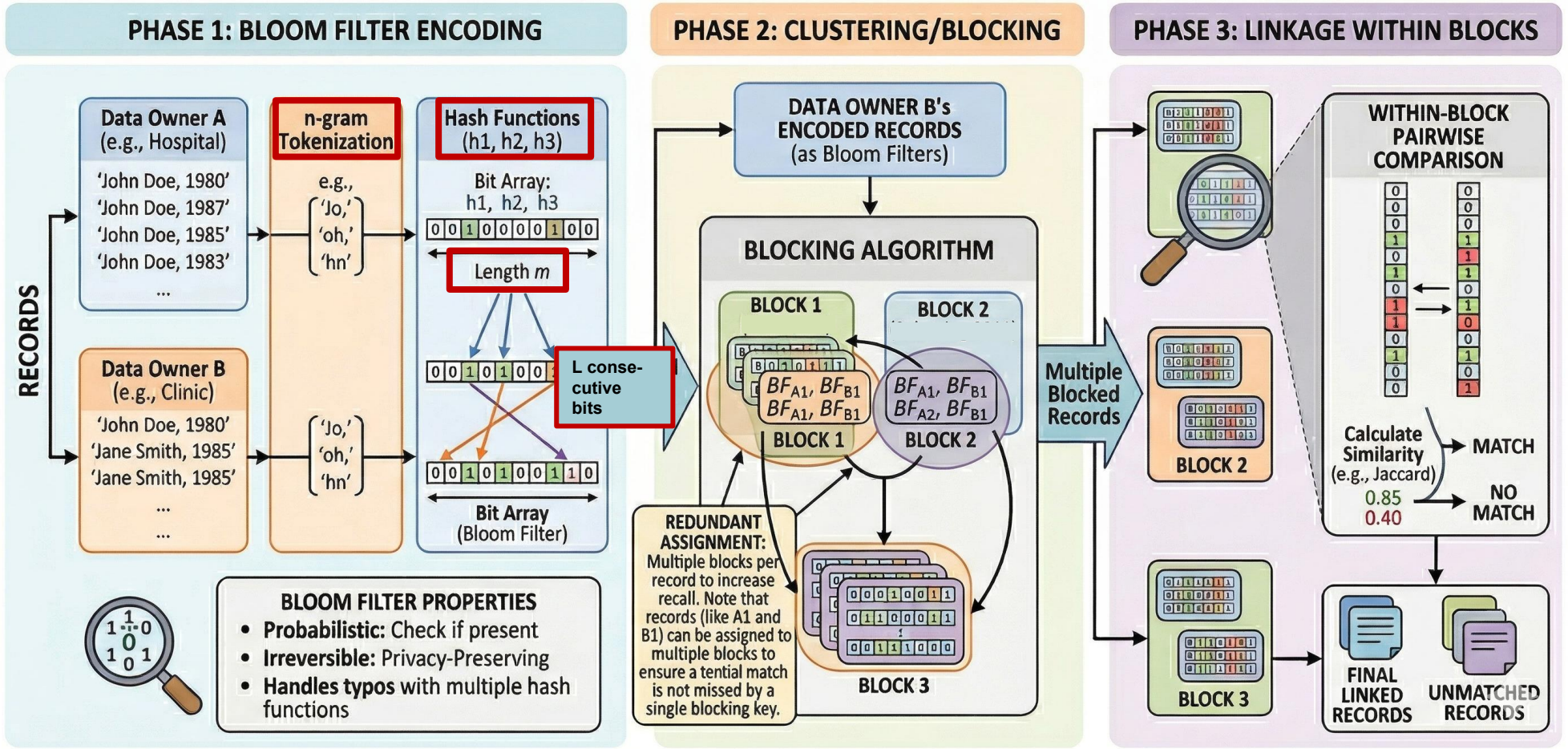
- **Part IV: The Future of PPRL**
- **Part V: Experimental results**
- **Part VI: Hands-on Session: PPRL tools**
- **Part VII: Conclusions and Final Remarks**

Funded by
the European Union

RECITALS
Privacy-Preserving Data
Sharing & ID Management

The Future of PPRL

Redundancy-based blocking



Funded by the European Union
RECITALS Privacy-Preserving Data Sharing & ID Management

Drawbacks of redundancy-based blocking

Two types of “useless” candidate pairs:

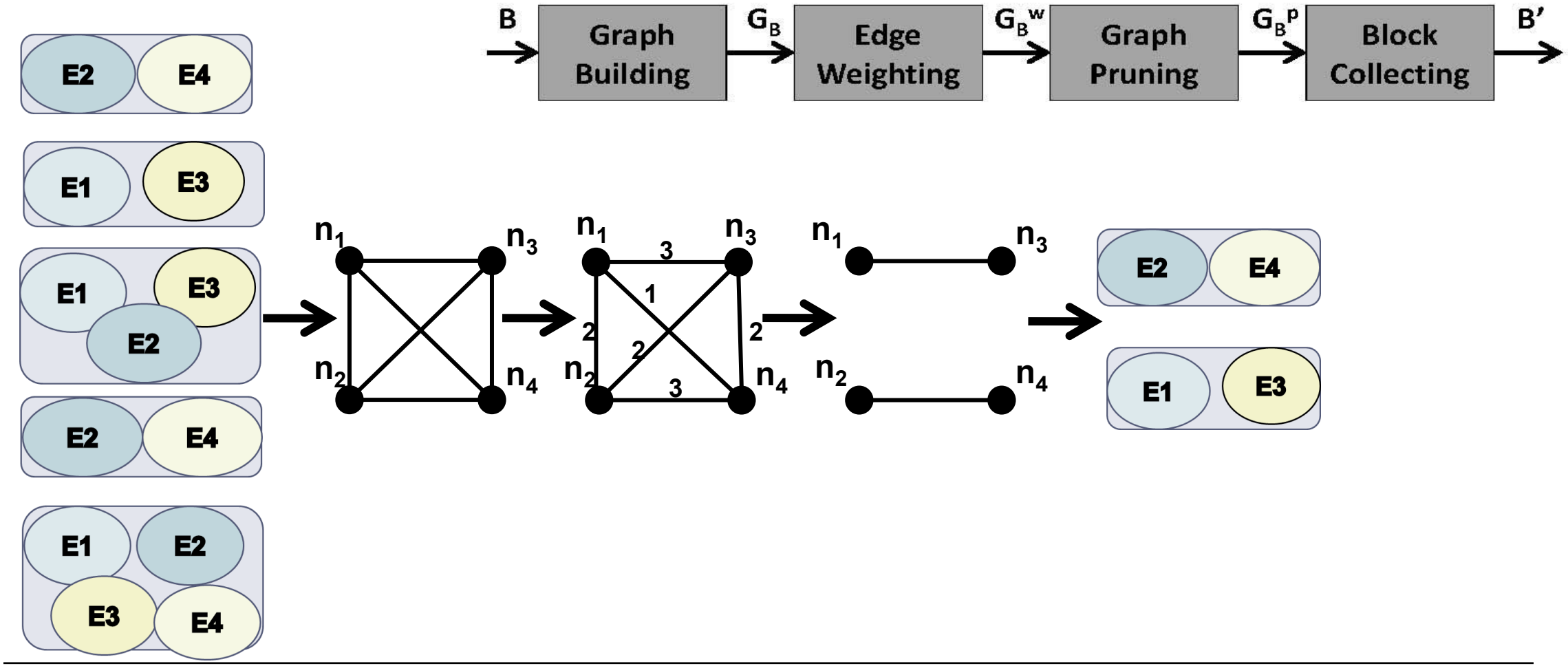
- Redundant pairs
  - repeated across different blocks
- Superfluous pairs
  - Involve non-matching entities

Blocking can be improved by:

1. eliminating *all* redundant pairs
2. avoiding *most* superfluous pairs

Funded by the European Union
RECITALS Privacy-Preserving Data Sharing & ID Management

Solution 1: Meta-blocking



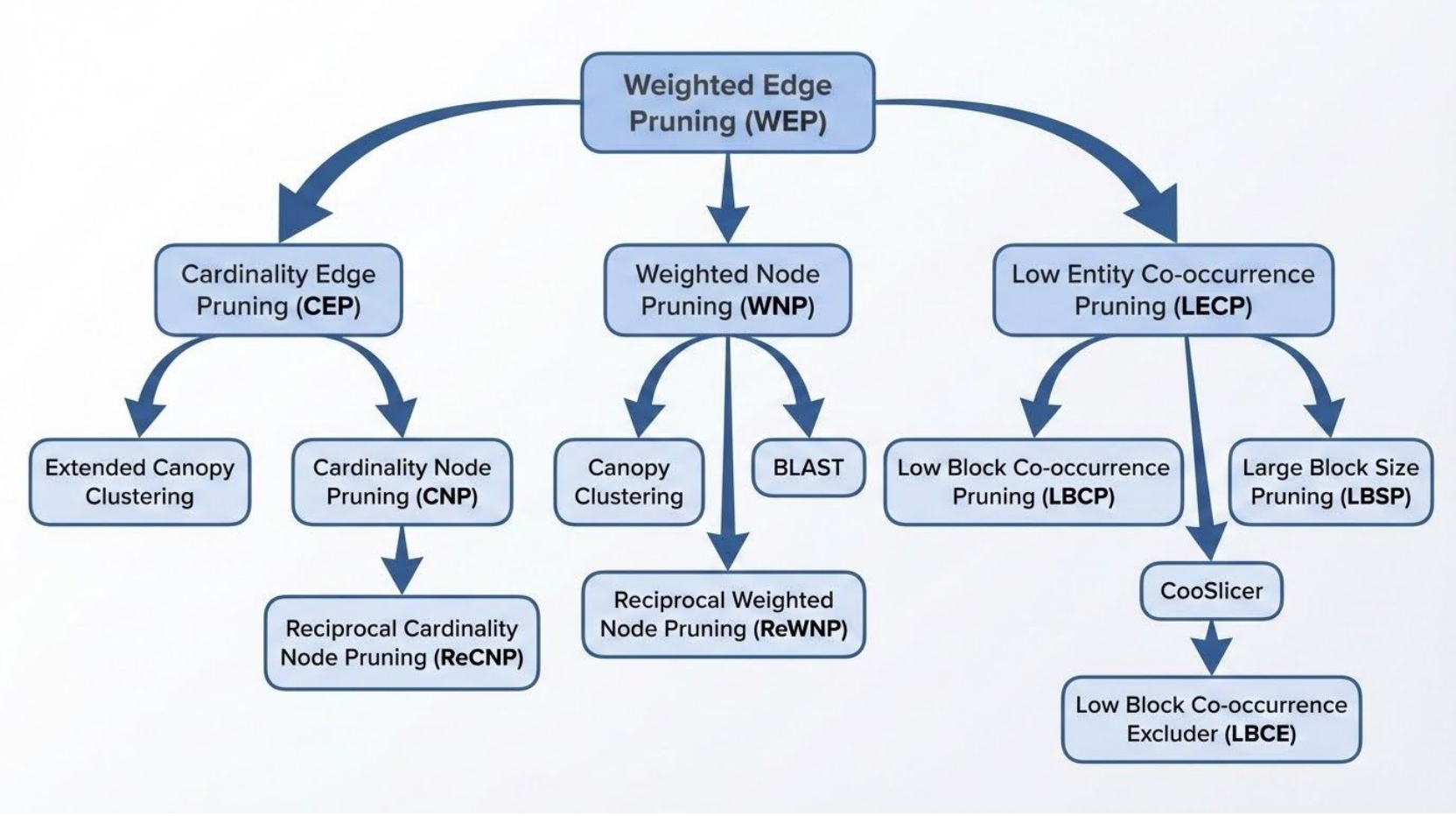
Funded by the European Union
RECITALS Privacy-Preserving Data Sharing & ID Management

Meta-blocking in practice

- Flexible framework that is defined by:
  1. Pruning algorithm
  2. Weighting scheme
  3. Pruning threshold

- High time efficiency and scalability
  - O(|D1|+|D2|) instead of O(|C|), where C is the set of candidate pairs
  - Inverted indices

Pruning algorithms



Weighting Schemes – Part I

1. Aggregate Reciprocal Comparisons Scheme (ARCS)

$$e_{i,j}.weight = \sum_{b_k \in B_{i,j}} \frac{1}{\|b_k\|}$$

2. Common Blocks Scheme (CBS)

$$e_{i,j}.weight = |B_{i,j}|$$

3. Enhanced Common Blocks Scheme (ECBS)

$$e_{i,j}.weight = |B_{i,j}| \cdot \log\left(\frac{|B|}{|B_i|}\right) \cdot \log\left(\frac{|B|}{|B_j|}\right)$$

4. Jaccard Scheme (JS)

$$e_{i,j}.weight = \frac{|B_{i,j}|}{|B_i| + |B_j| - |B_{i,j}|}$$

5. Enhanced Jaccard Scheme (EJS)

$$e_{i,j}.weight = \frac{|B_{i,j}|}{|B_i| + |B_j| - |B_{i,j}|} \cdot \log\left(\frac{|E_B|}{|v_i|}\right) \cdot \log\left(\frac{|E_B|}{|v_j|}\right)$$

Funded by the European Union
RECITALS Privacy-Preserving Data Sharing & ID Management

Weighting Schemes – Part II

Similarity Functions

Cosine

$$\frac{|B_i \cap B_j|}{\sqrt{|B_i| \cdot |B_j|}} = \frac{\text{CBS}}{\sqrt{|B_i| \cdot |B_j|}}$$

Dice

$$2 \cdot \frac{|B_i \cap B_j|}{|B_i| + |B_j|} = \frac{2 \cdot \text{CBS}}{|B_i| + |B_j|}$$

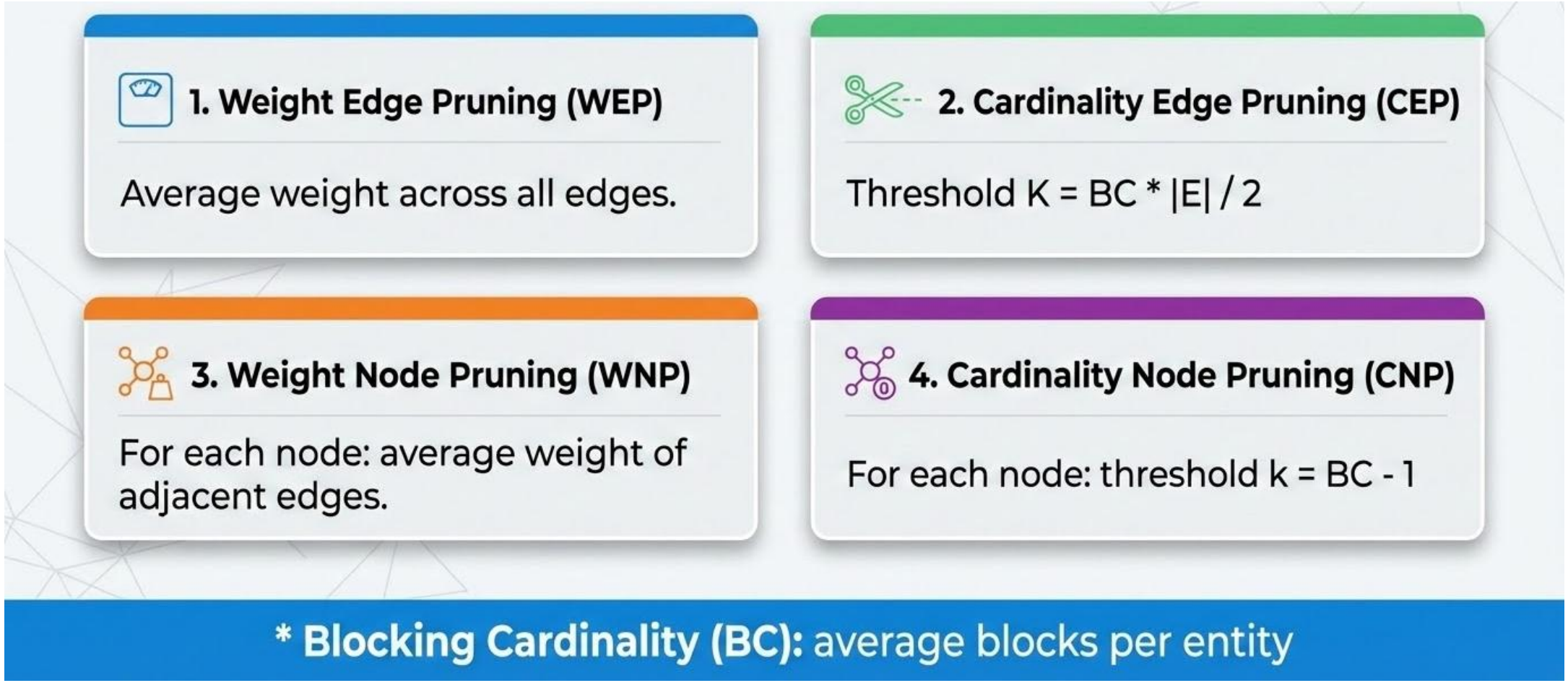
Size-normalized Weighting Schemes

- **SN-CBS** = $\sum_{b \in B_i \cap B_j} \frac{1}{|b|}$
- **SN-Cosine** = $\frac{\text{SN-CBS}}{\sqrt{\text{SN-}B_i \cdot \text{SN-}B_j}}$
- **SN-Dice** = $2 \cdot \frac{2 \cdot \text{SN-CBS}}{\text{SN-}B_i + \text{SN-}B_j}$
- **SN-Jaccard** = $\frac{\text{SN-CBS}}{\text{SN-B}_i + \text{SN-B}_j - \text{SN-CBS}}$

Cardinality-normalized Weighting Schemes

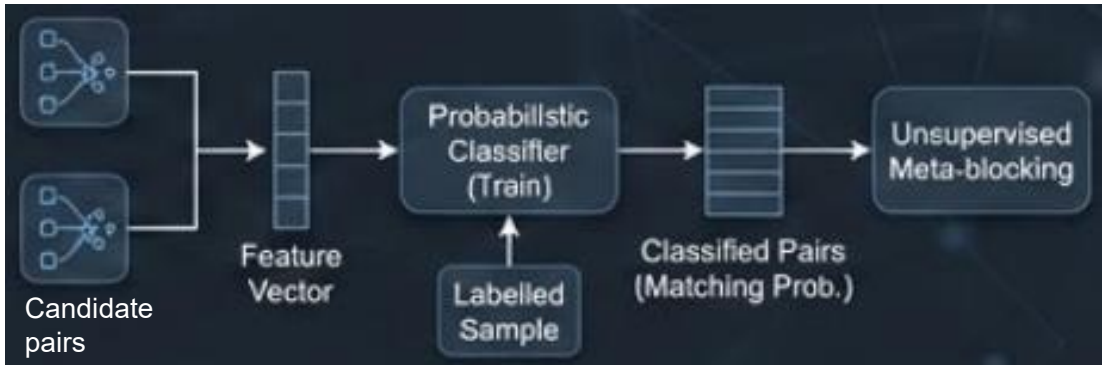
- **CN-CBS** = $\sum_{b \in B_i \cap B_j} \frac{1}{||b||}$
- **CN-Cosine** = $\frac{\text{CN-CBS}}{\sqrt{\text{CN-}B_i \cdot \text{CN-}B_j}}$
- **CN-Dice** = $2 \cdot \frac{2 \cdot \text{CN-CBS}}{\text{CN-}B_i + \text{CN-}B_j}$
- **CN-Jaccard** = $\frac{\text{CN-CBS}}{\text{CN-B}_i + \text{CN-B}_j - \text{CN-CBS}}$

Pruning thresholds



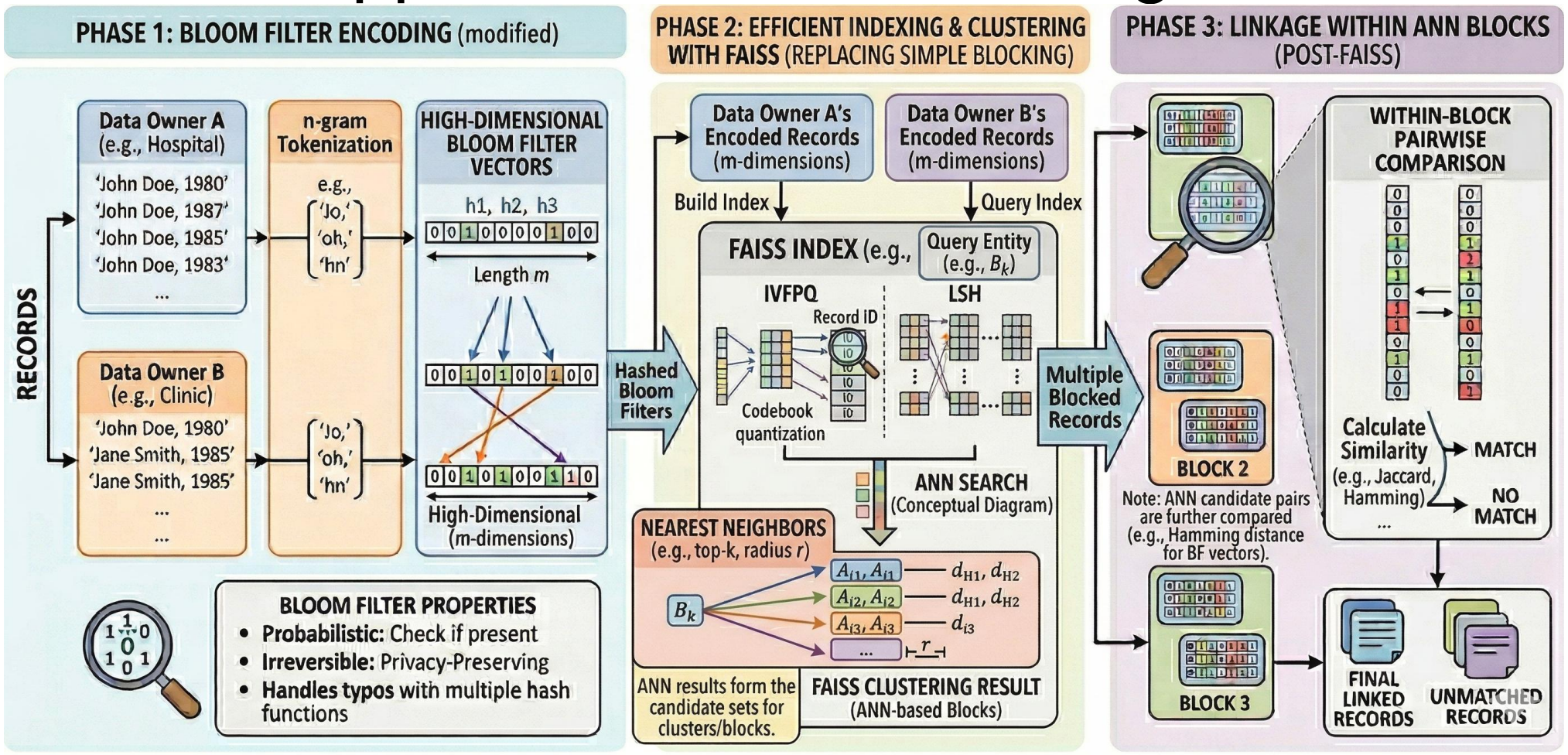
Funded by the European Union
RECITALS Privacy-Preserving Data Sharing & ID Management

Supervised Meta-blocking

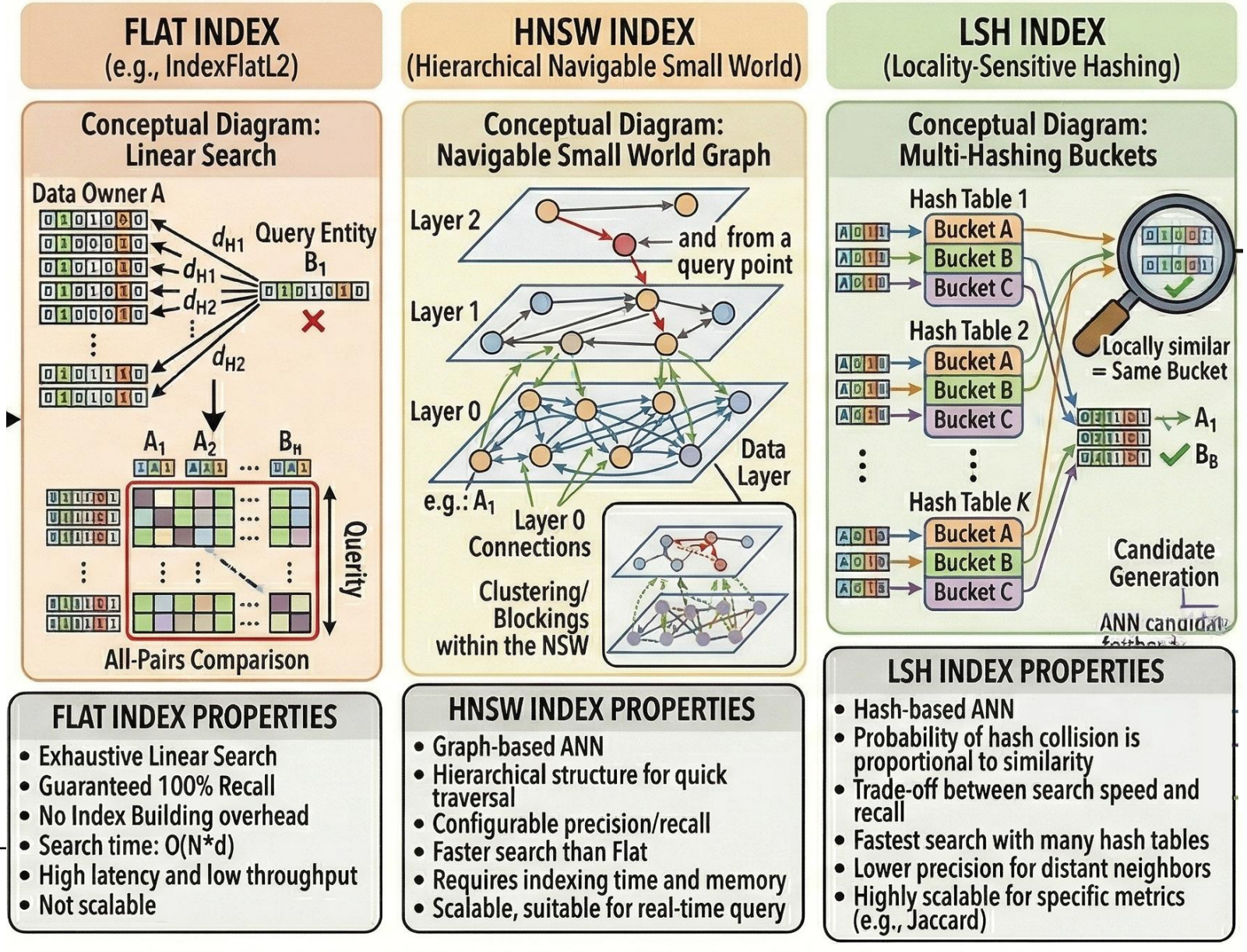


- Based on a binary probabilistic classifier
  - Each candidate pair is transformed into a feature vector
  - Each feature corresponds to a weighting scheme
    - Feature selection is optional
  - A sample of the candidate pairs is labelled to train a probabilistic classification model (Logistic Regression, SVC, etc)
    - 50 instances are sufficient in RL
  - All other candidate pairs are classified by the trained classifier, which assigns a matching probability
  - We then apply any unsupervised meta-blocking algorithm on these pairs

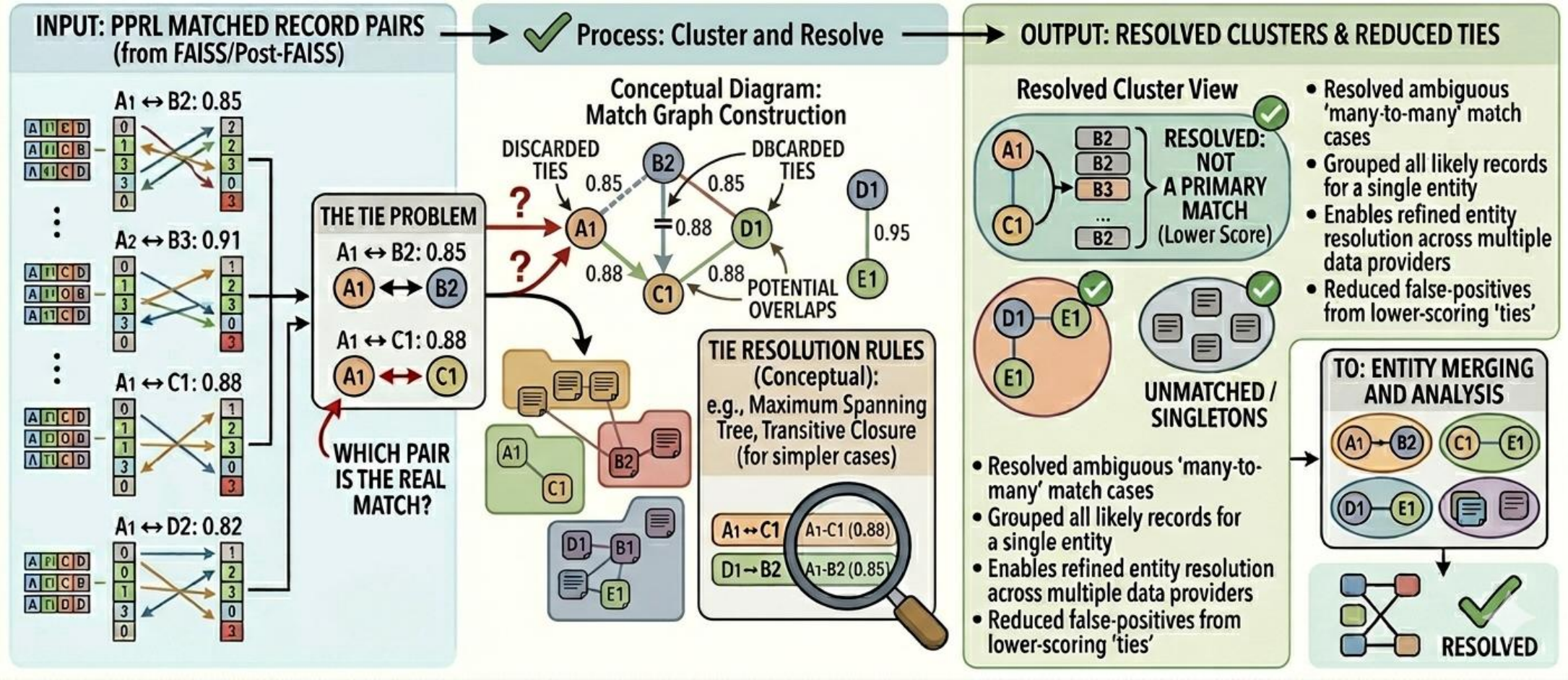
Solution 2: Approximate Nearest Neighbor Search



FAISS indices



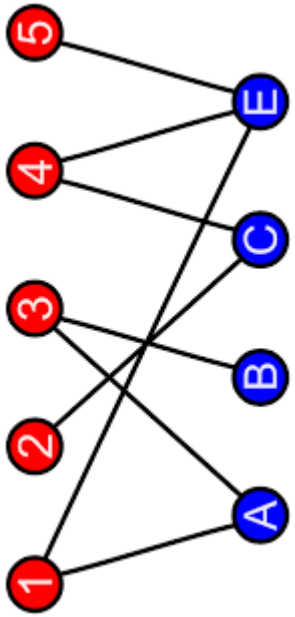
Post-matching with Clustering



Funded by the European Union
RECITALS Privacy-Preserving Data Sharing & ID Management

Clustering algorithms

- Goal: resolve conflicts, which violate the 1-1 constraint
  - E.g., “A matches B” and “B matches C”
- Solution:
  - The matching results correspond to a bipartite graph
  - Apply established bipartite graph matching algorithms
  - Relevant tasks:
    - Stable marriage
    - Assignment problem
  - Strict requirements for applicable algorithms



Selection criteria

We consider algorithms that:

- Have a learning-free functionality
  - We perform fine-tuning based on the ground-truth
- Time complexity ≤ $O(n^2)$
  - n stands for the number of input entities
  - E.g., the Hungarian algorithm is excluded, due to its cubic complexity, $O(n^3)$
- Space complexity is $O(n+m)$
  - m denotes the number of edges

Funded by the European Union
RECITALS Privacy-Preserving Data Sharing & ID Management

Selected algorithms

CNC – $O(m)$
Connected components

RSR – $O(n\,m)$
Ricochet Sequential Rippling Clustering

RCA – $O(|V_1||V_2|)$
Row Column Assignment Clustering

BAH
Best Assignment Heuristic
BAH: Additional configuration parameter: 10,000 search steps or 2 min. run-time

BMC – $O(m)$
Best Match Clustering
BMC: Additional configuration parameter: the node partition used as basis

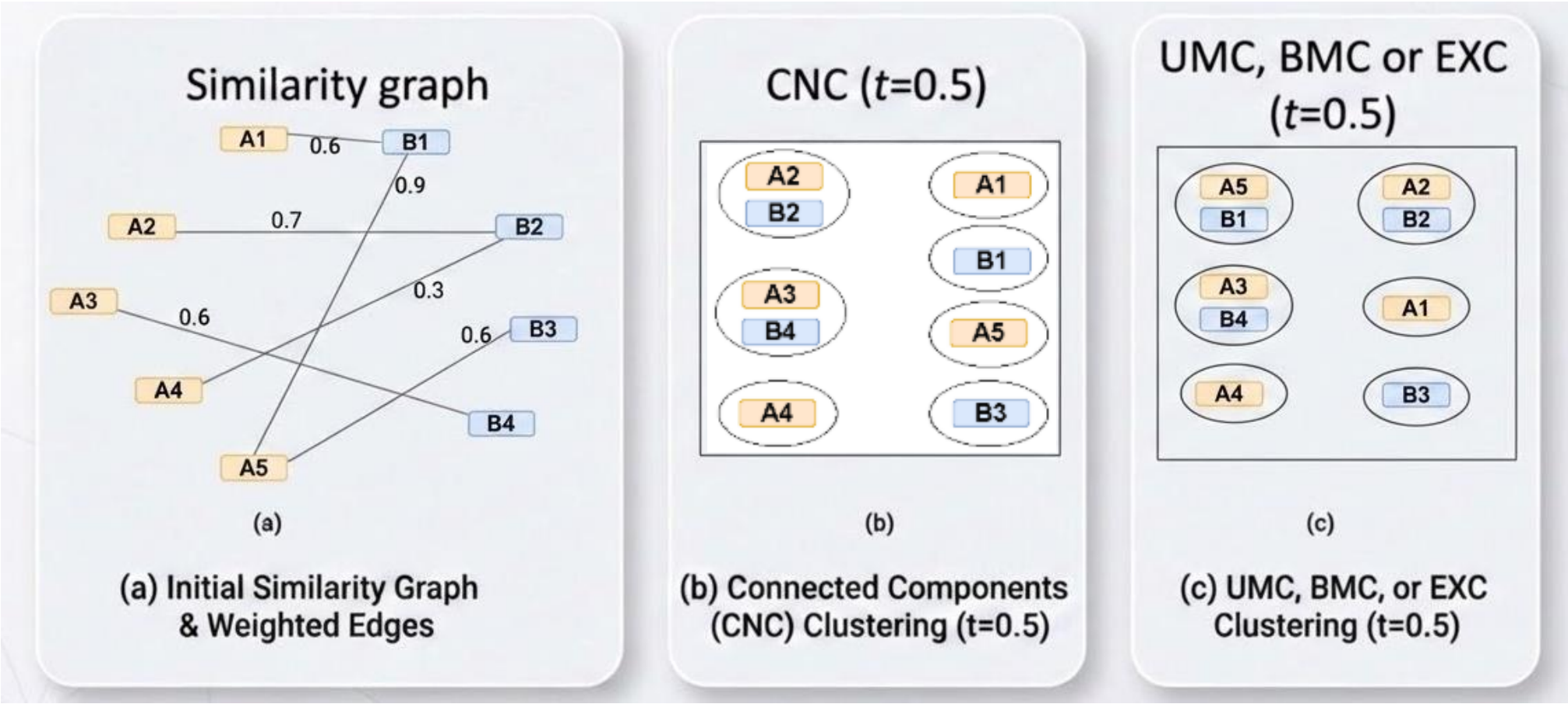
EXC – $O(n\,m)$
Exact Clustering

KRC – $O(n + m \log m)$
Király's Clustering

UMC – $O(m \log m)$
Unique Mapping Clustering

Common configuration parameter: **similarity threshold t**

Clustering algorithms in action



Experimental Results

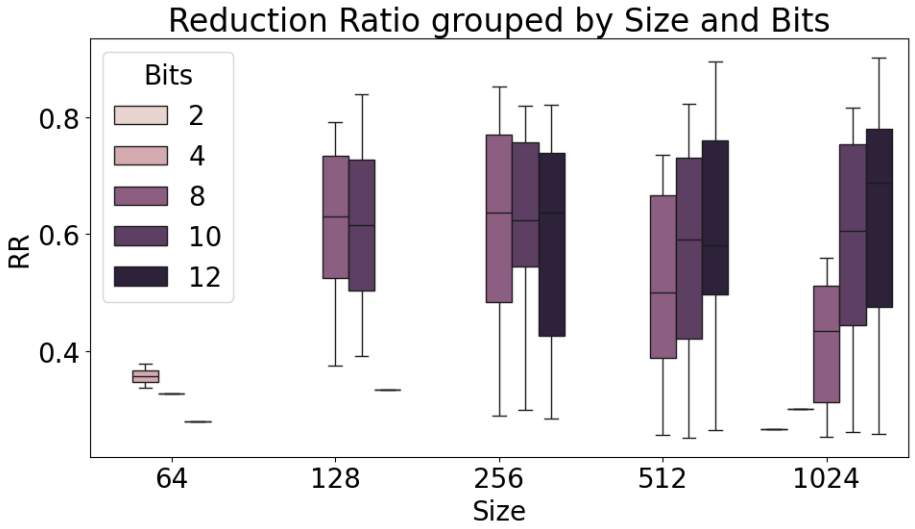
Evaluation Measures

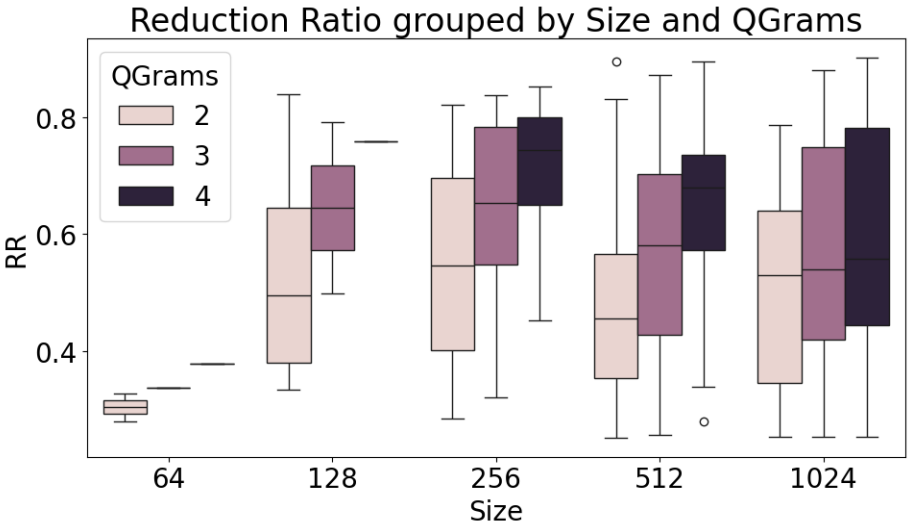
- Linkage Quality (i.e., Effectiveness)
  - Blocking step
    - Pairs Completeness (**PC**): portion of matching pairs captured in the candidate set
    - Pairs Quality (**PQ**): ratio of matching pairs to total candidate pairs
    - Reduction Ratio (**RR**): reduction in the number of candidate pairs compared to the Cartesian product
    - Goal: High completeness (i.e., maximize PC and PQ) with minimal candidate pairs (i.e., maximize RR) → trade-off!
  - Matching step
    - Precision (**Pr**): Proportion of predicted matches that are correct
    - Recall (**Re**): Proportion of true matches that are identified
    - F-Measure (**F1**): Harmonic mean of precision and recall (balanced metric)
    - Goal: the higher the value for each measure, the better

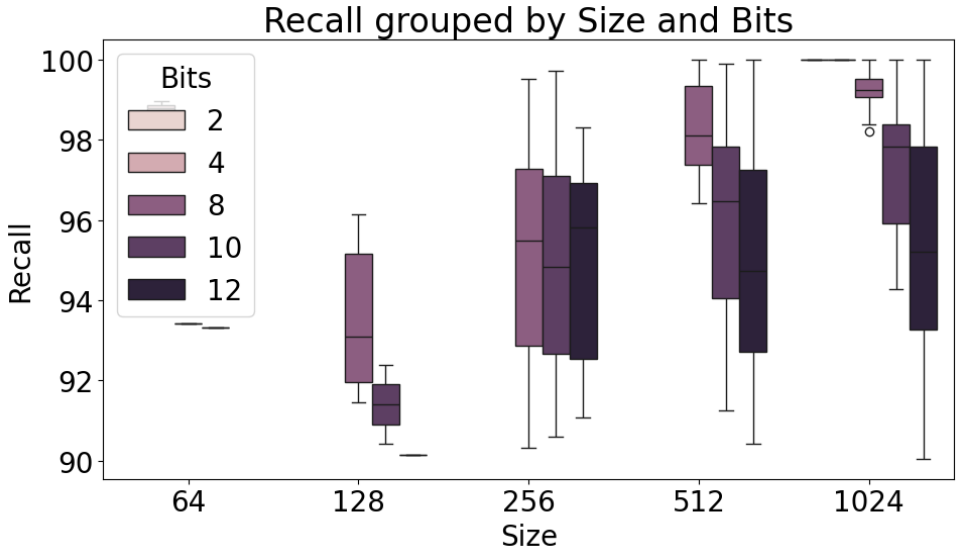
Funded by the European Union
RECITALS
Privacy-Preserving Data Sharing & ID Management

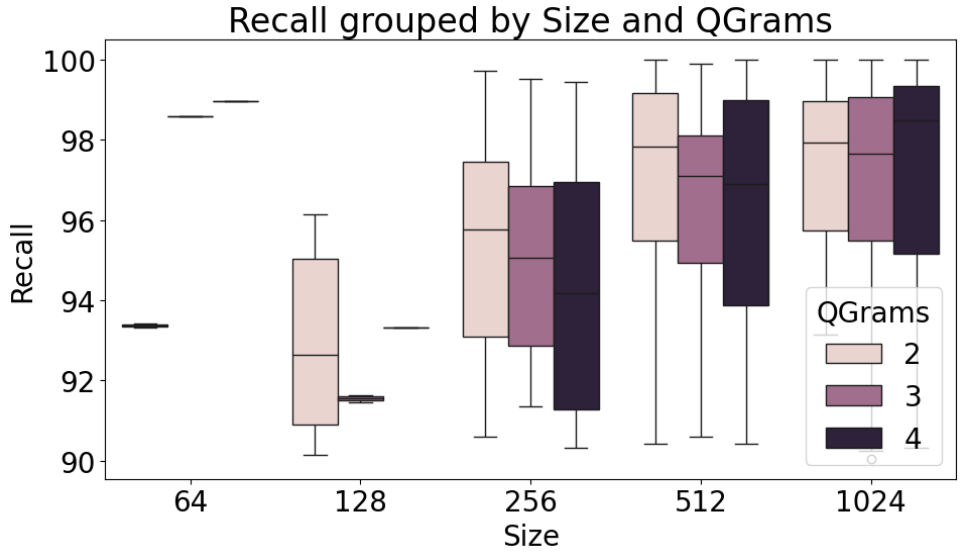
Blocking

Abt-Buy dataset matching on name





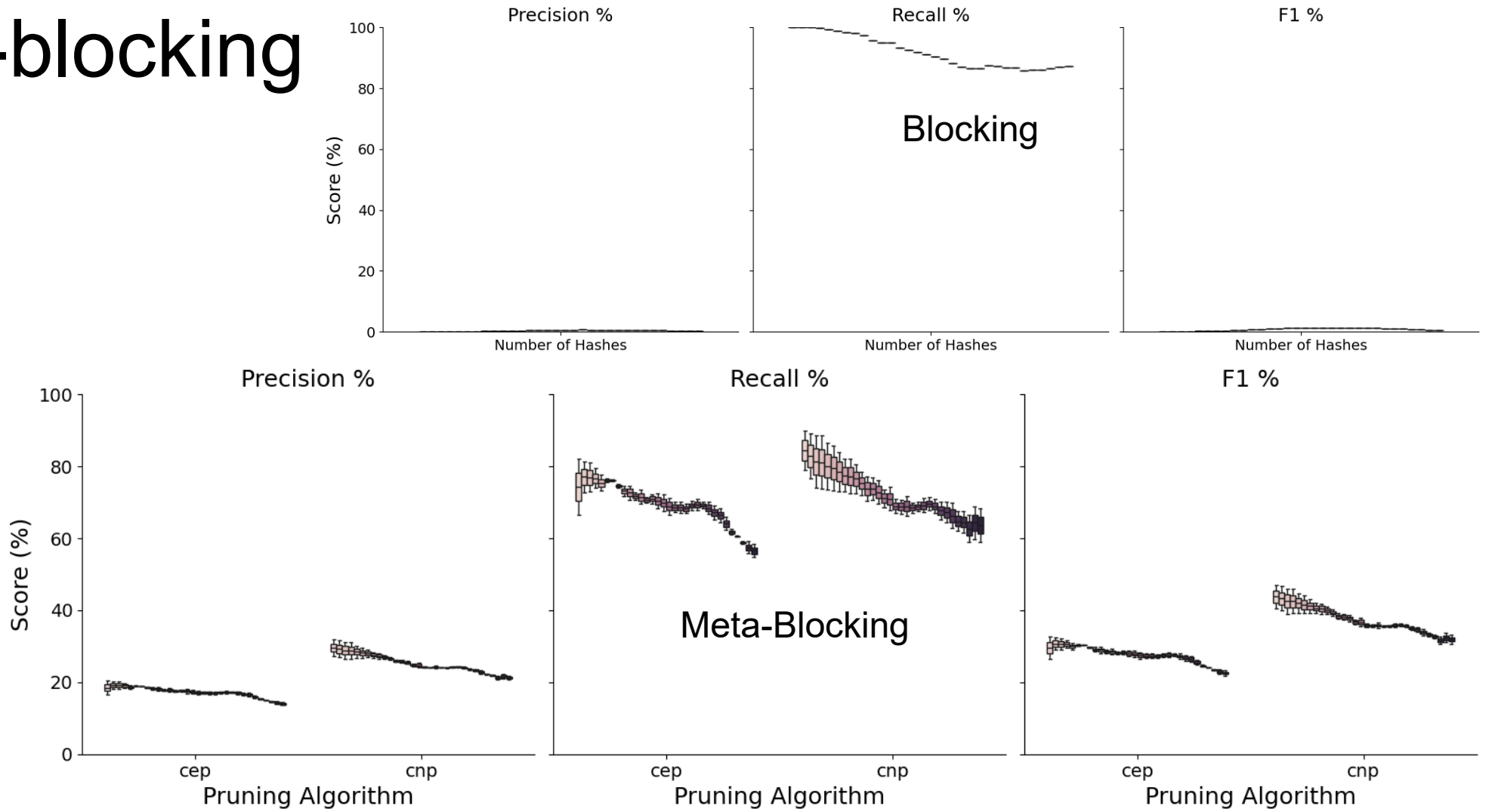




Funded by the European Union

RECITALS
Privacy-Preserving Data Sharing & ID Management

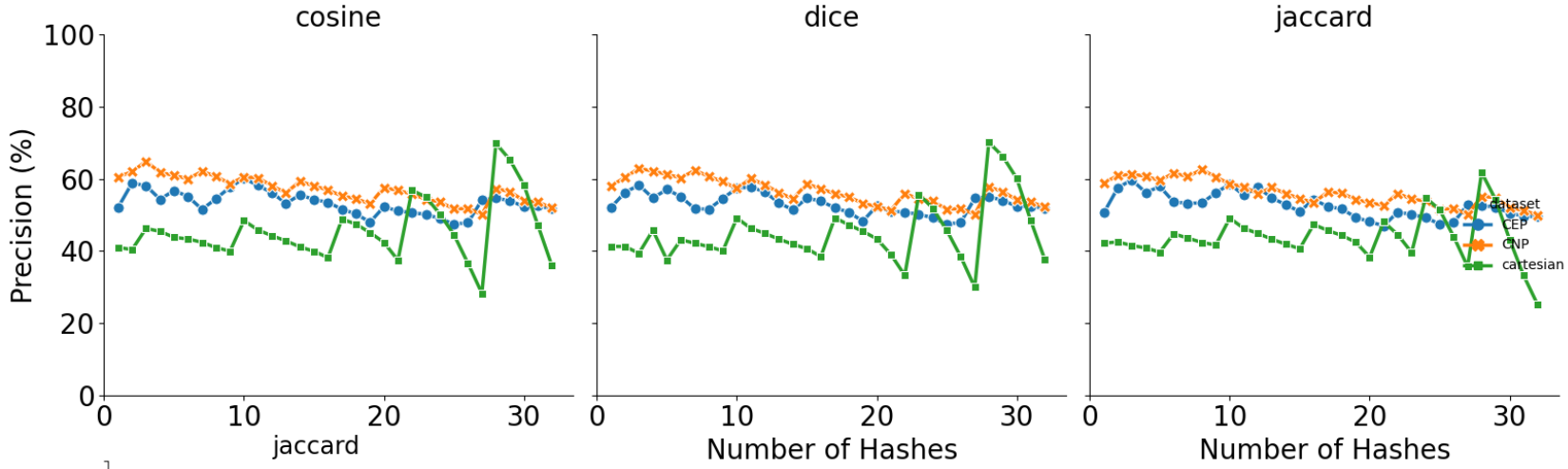
Meta-blocking

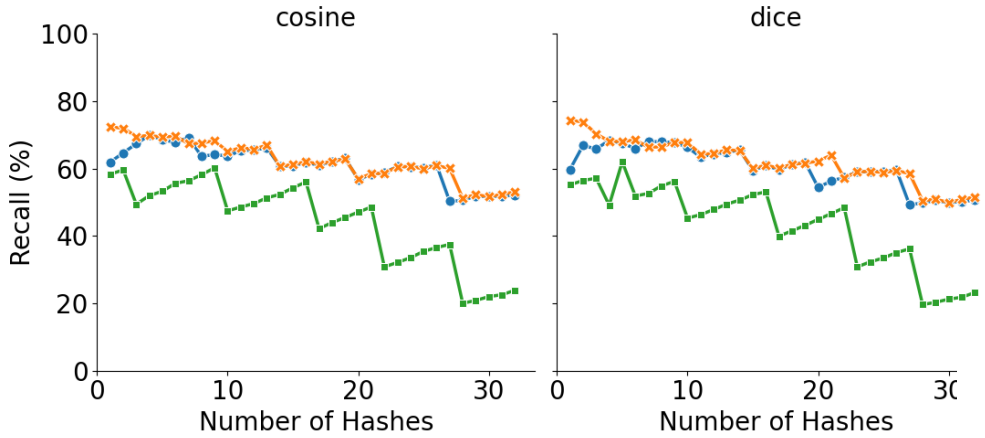


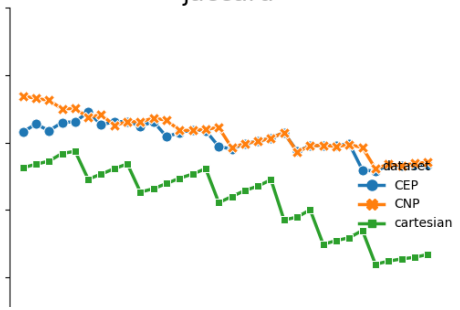
Funded by the European Union

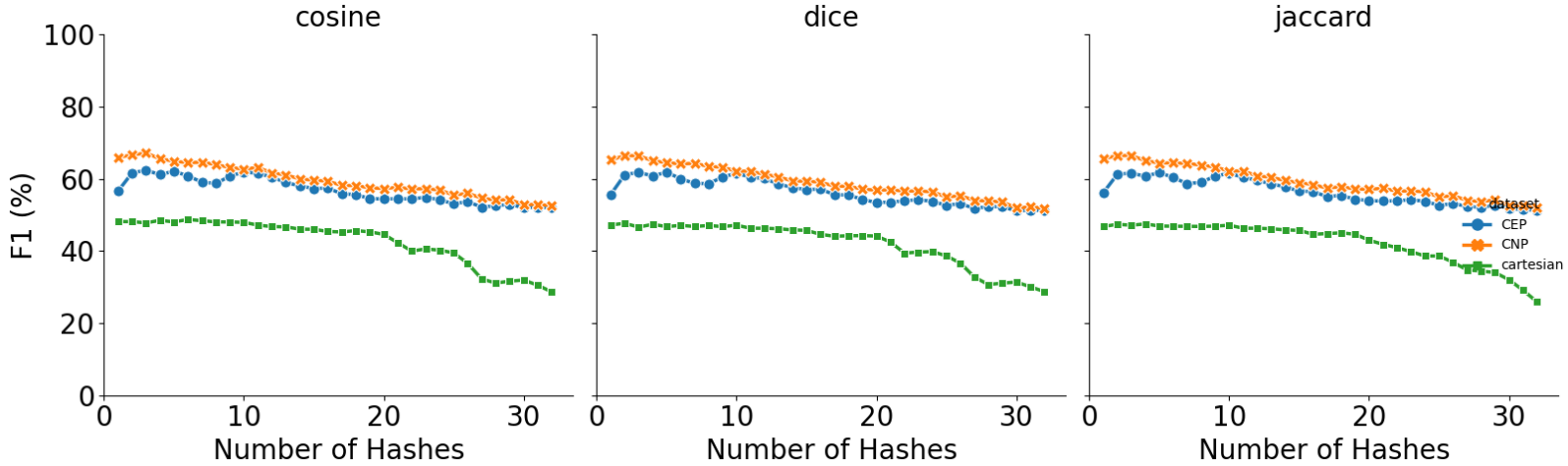
RECITALS
Privacy-Preserving Data Sharing & ID Management

Matching on top of Meta-blocking

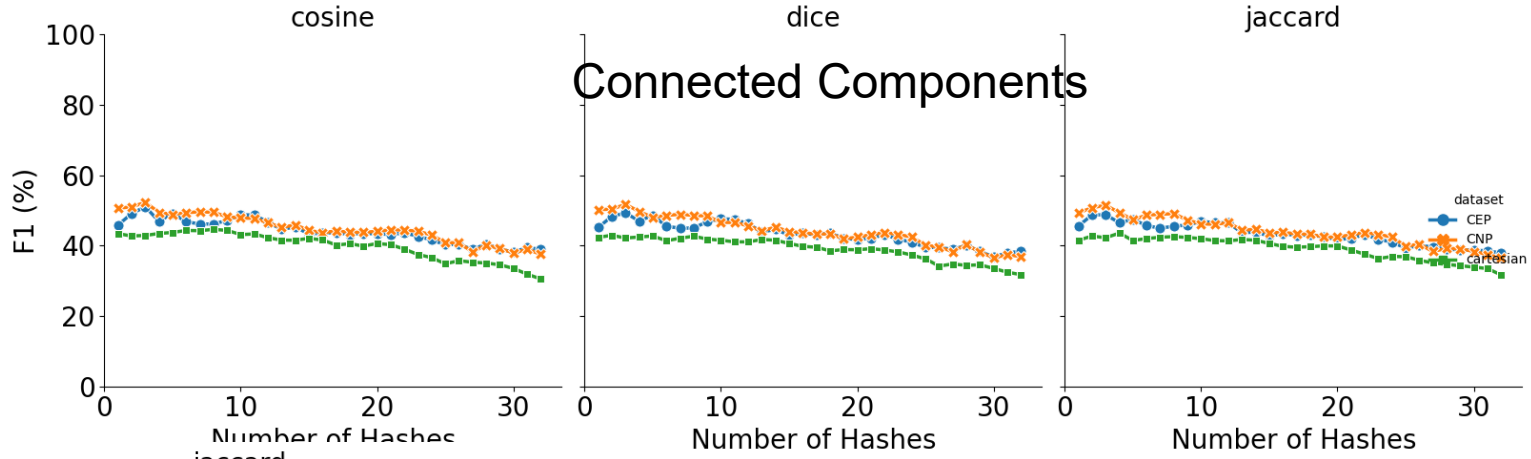


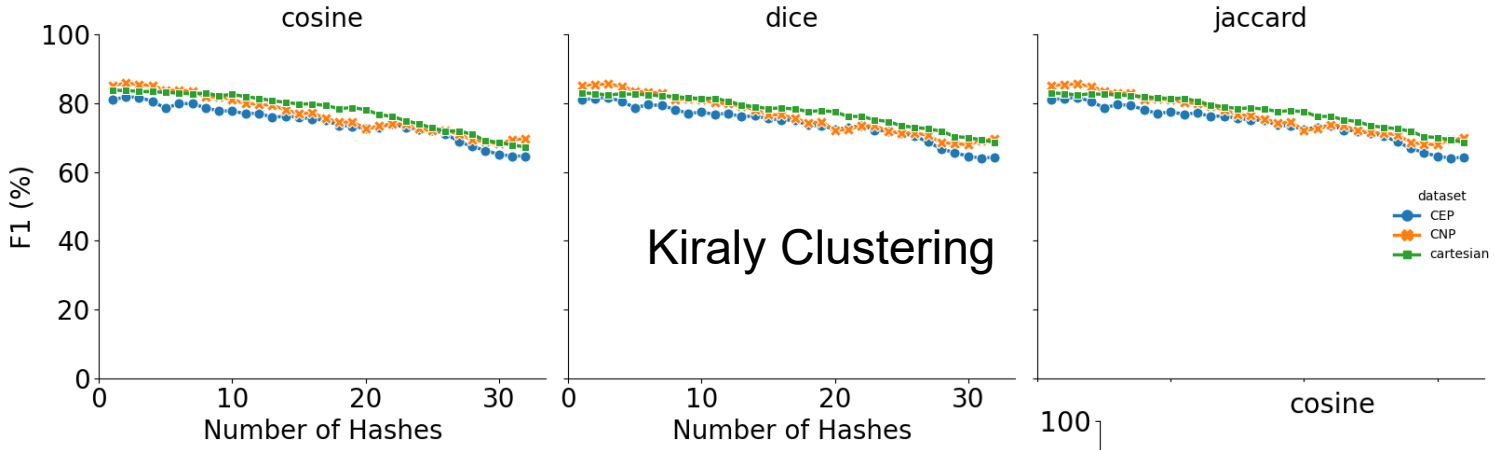


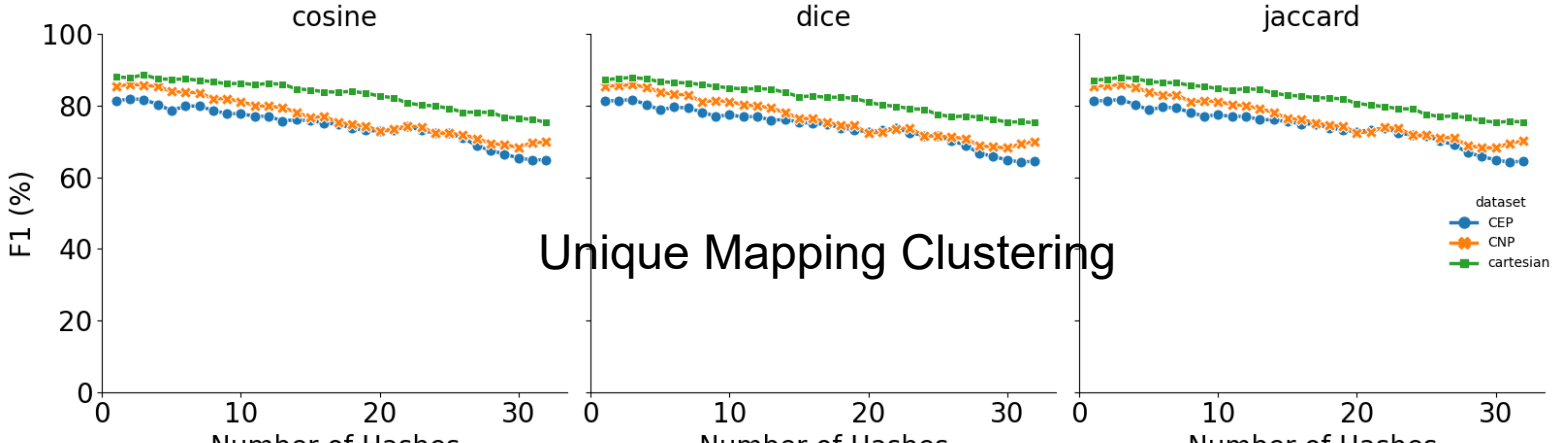




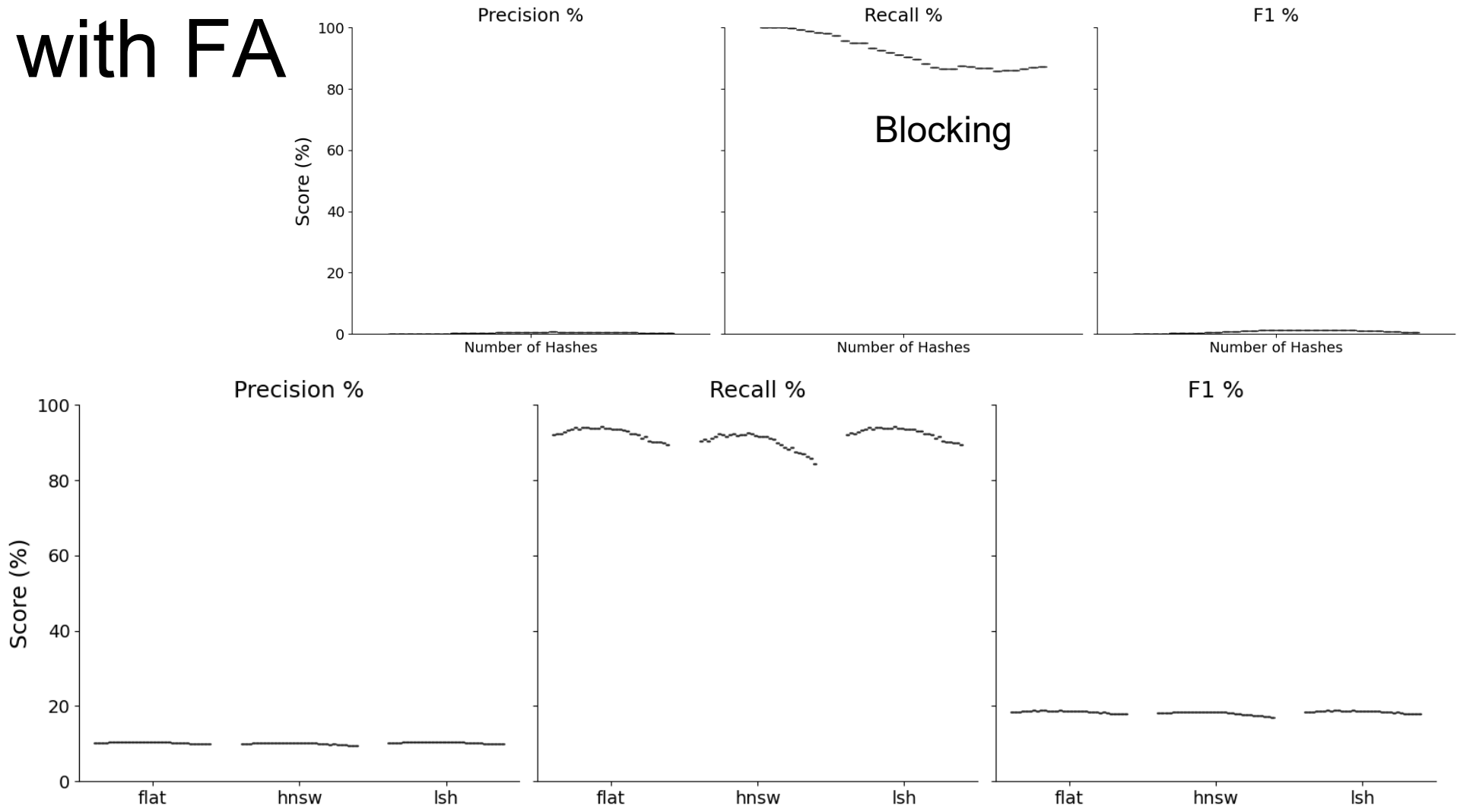
Clustering on top of Meta-blocking







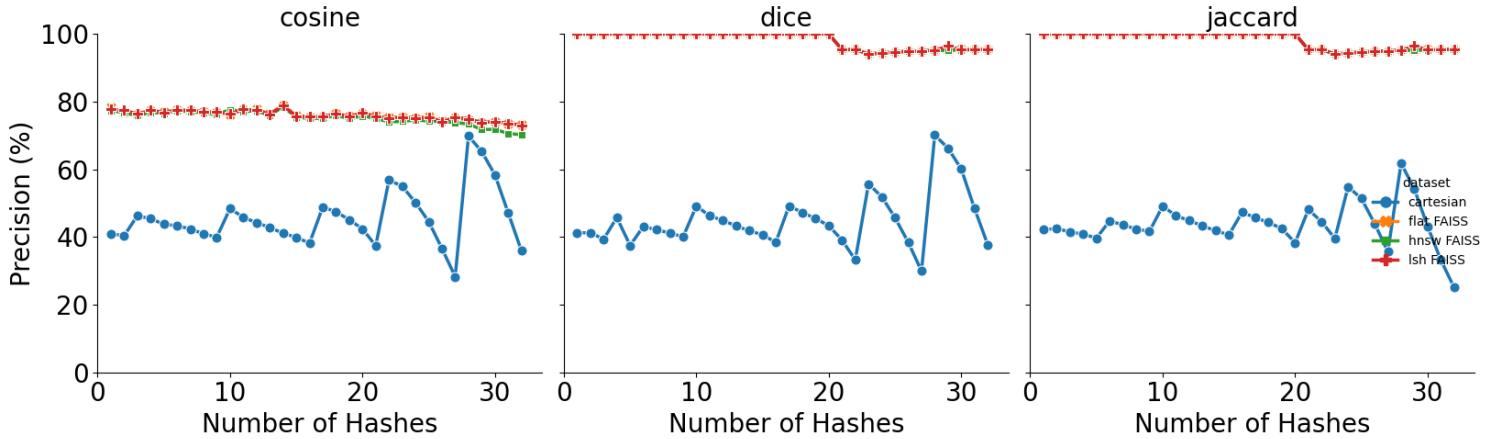
ANNS with FA

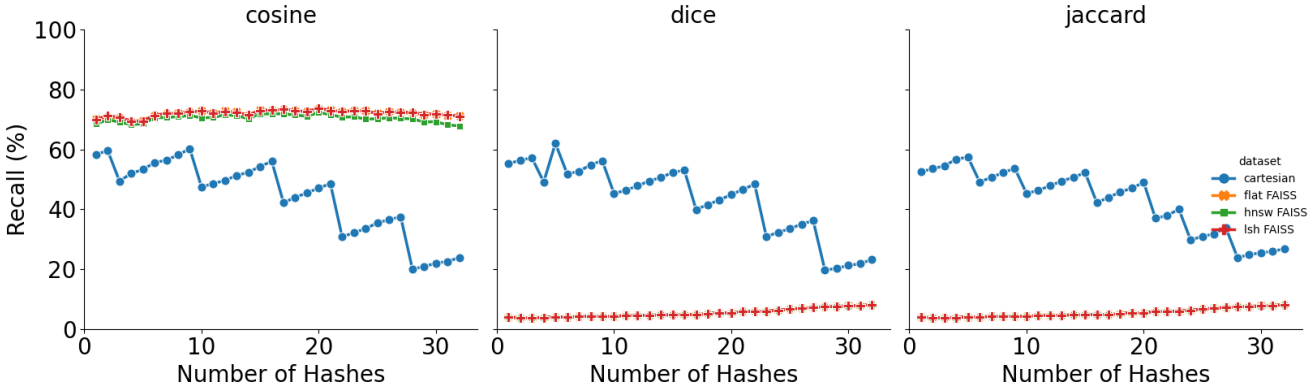

Precision %
Recall %
F1 %
Blocking
Score (%)
Number of Hashes
Precision %
Recall %
F1 %
Score (%)
flat
hnsw
lsh

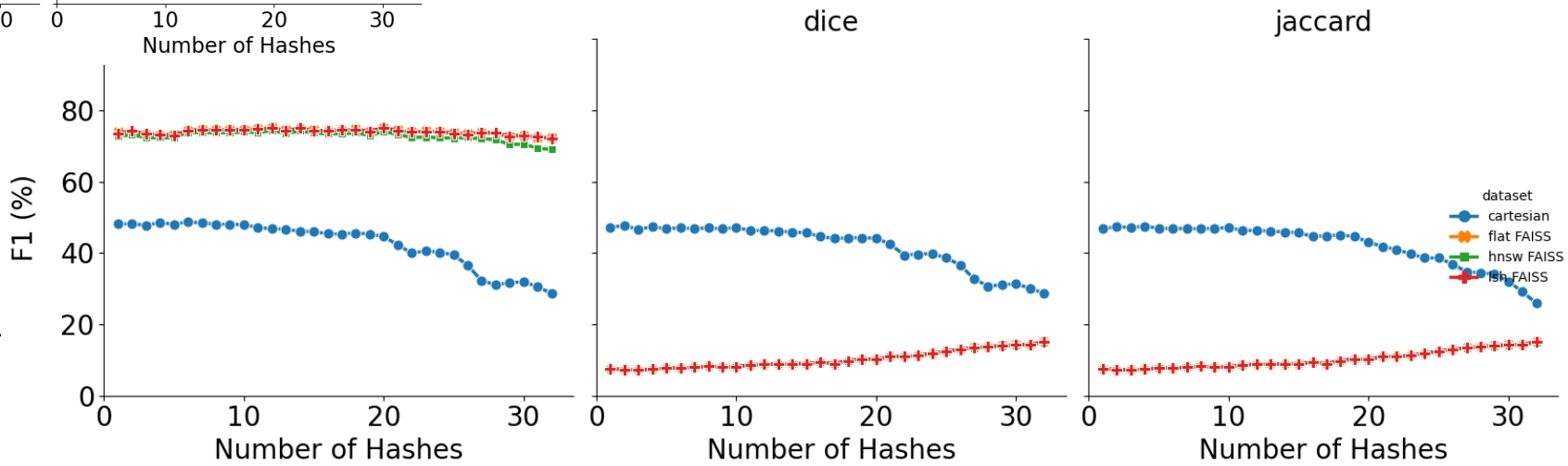
Funded by the European Union

RECITALS
Privacy-Preserving Data Sharing & ID Management

Matching on top of ANNS



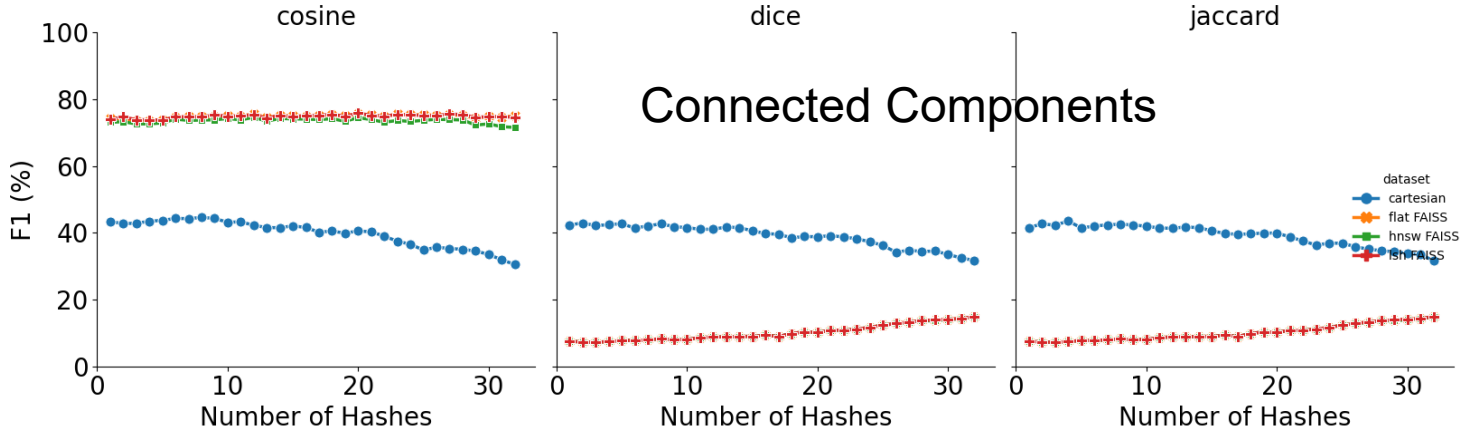


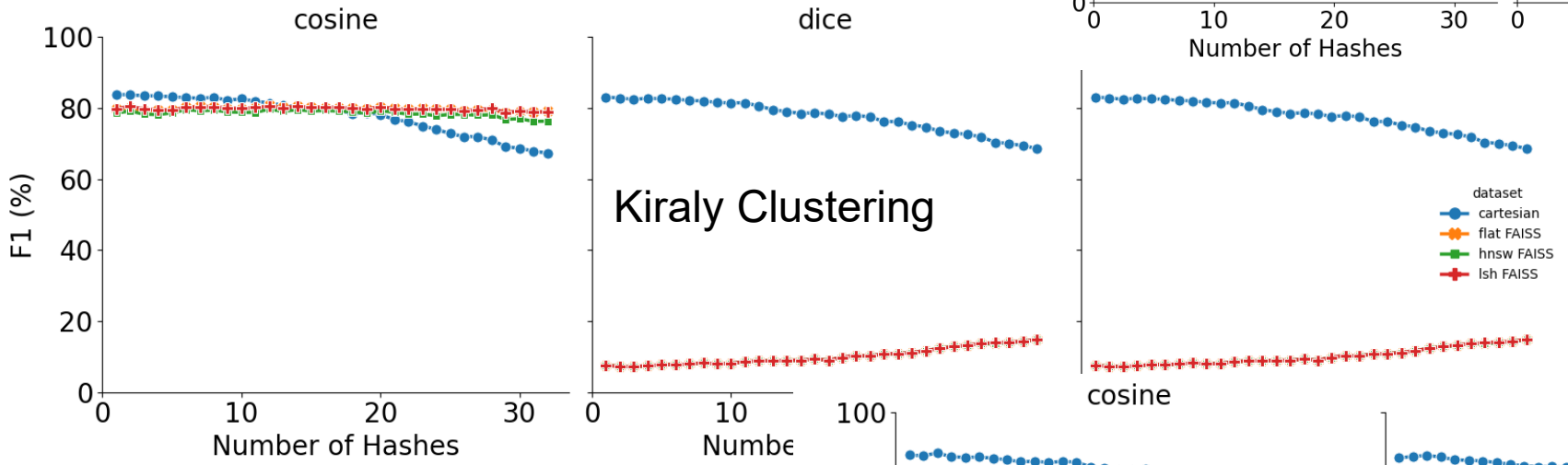


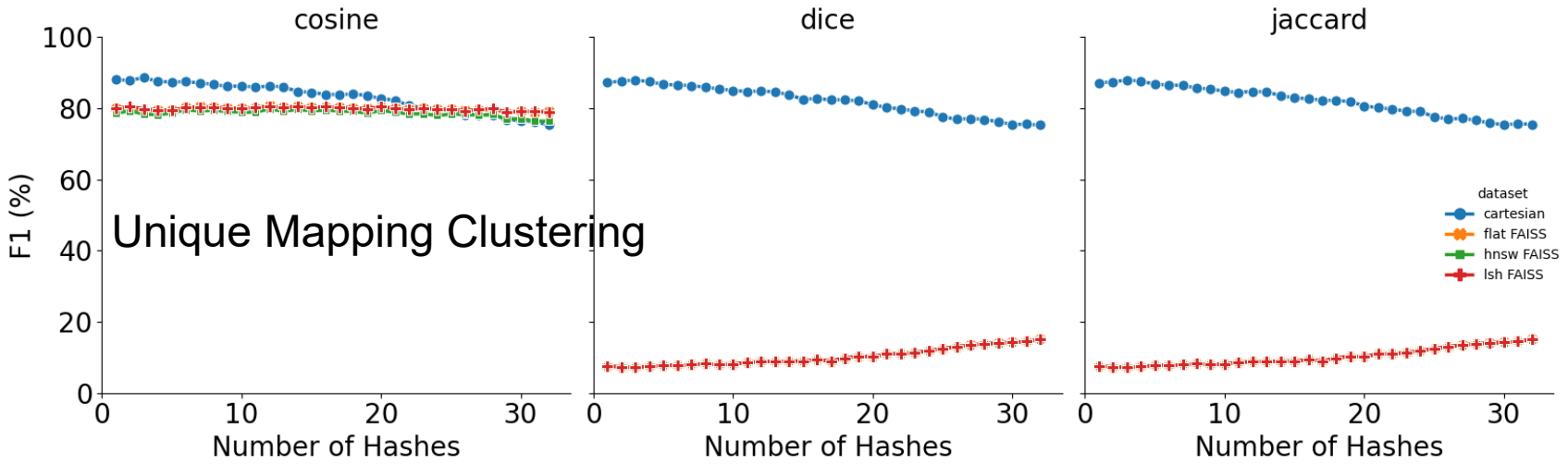
Funded by the European Union

RECITALS
Privacy-Preserving Data Sharing & ID Management

Clustering on top of ANNS







THANK YOU!

gpapadis@di.uoa.gr

RECITALS

Privacy-Preserving Data Sharing & ID Management

Funded by the European Union

Q&A

Funded by
the European Union

RECITALS
Privacy-Preserving Data
Sharing & ID Management

Conclusions
& Final Remarks

Past & Present of PPRL

- From Theory to Necessity
  - PPRL has matured from a cryptographic application to a practical requirement driven by GDPR, HIPAA, and cross-institutional data needs.
- Bloom Filters Dominate – But Are Not Enough
  - BF-based encoding enables approximate matching at scale, yet remains vulnerable to cryptanalysis, frequency, and composition attacks.
- The Privacy-Quality-Scalability Triangle
  - No single method optimises all three simultaneously; deployment requires explicit trade-off decisions.
- Deep Learning Opens New Frontiers
  - SNNs, federated learning, and autoencoders offer powerful new tools but introduce new challenges around labelled data and compute costs.
- Ongoing Research Is Essential
  - Evolving regulations, new attack vectors, and real-time requirements demand continuous innovation across encoding, blocking, and matching.

Future of PPRL

- Improve efficiency and scalability of PPRL through:
  - Meta-blocking
  - ANNs
- Improve effectiveness of PPRL through:
  - Post-matching clustering (bipartite graph matching)
- **privJedAI**: the first comprehensive, open-source library of PPRL tools
  - Modular architecture for end-to-end PPRL
  - Comprises state-of-the-art techniques for each step

Challenges lying ahead

Improved tools for PPRL

- Simpler GUI: no coding for users of any expertise
- Guidelines for creating effective solutions

Simplified pipeline configuration

- Current state:
  - Several parameters in every method
  - Performance sensitive to internal configuration
  - Manual fine-tuning required
- Target state:
  - Automatic data-driven configuration

THANK YOU!

gpapadis@di.uoa.gr

RECITALS

Privacy-Preserving Data Sharing & ID Management

Funded by the European Union

Q&A