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

EDBT 2026
**Tampere, Finland**

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

Funded by
the European Union

# Outline

- **Part I: Introduction & Core Concepts**
- **Part II: Foundations of Privacy-Preserving Record Linkage (The Past: 1998 – 2020)**
- **Part III: Recent Advances: 2021 – 2025**

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

# Preliminaries: What is Record Linkage?

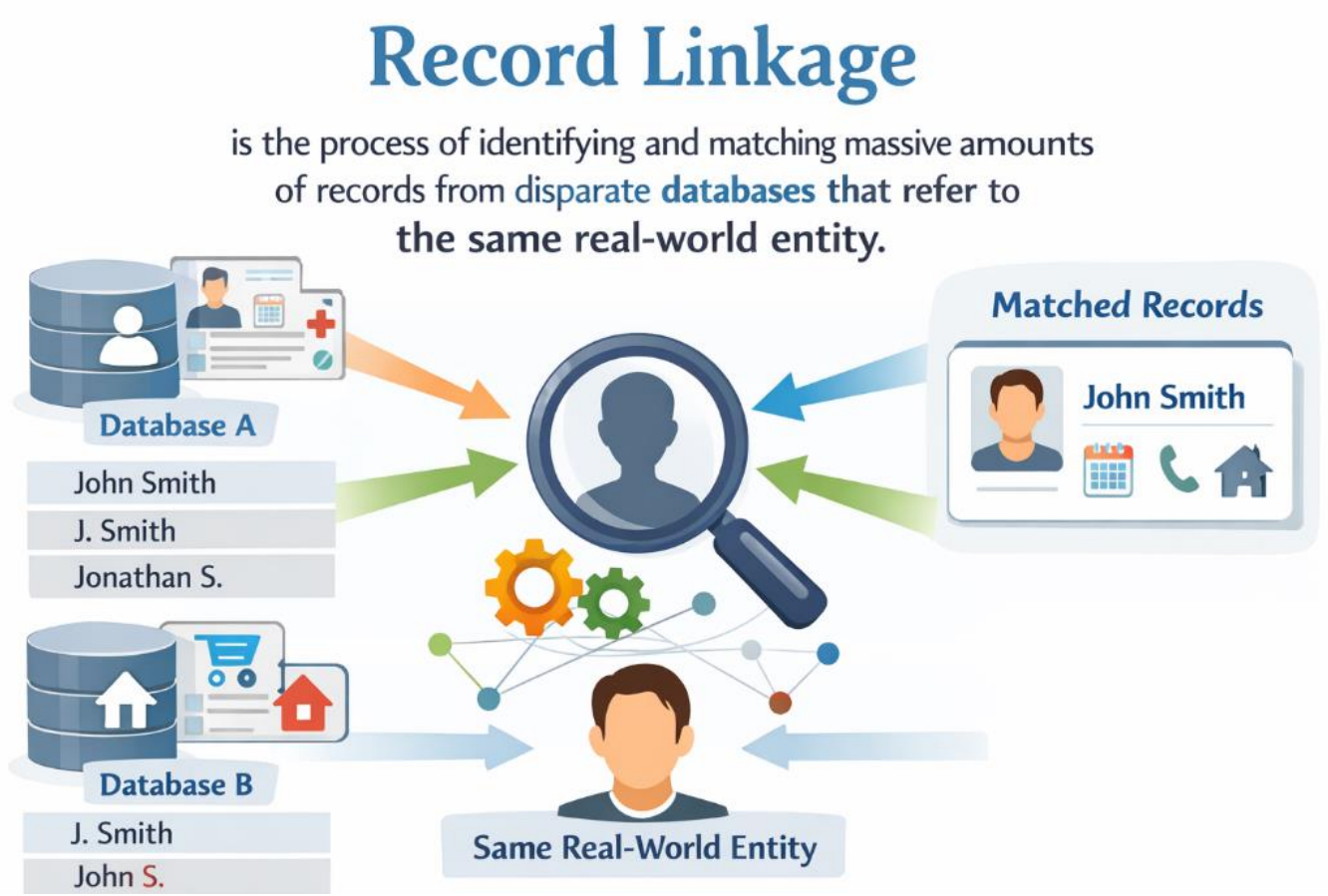



Image created by Gemini

# The Core Problem of Record Linkage: Lack of unique global identifiers

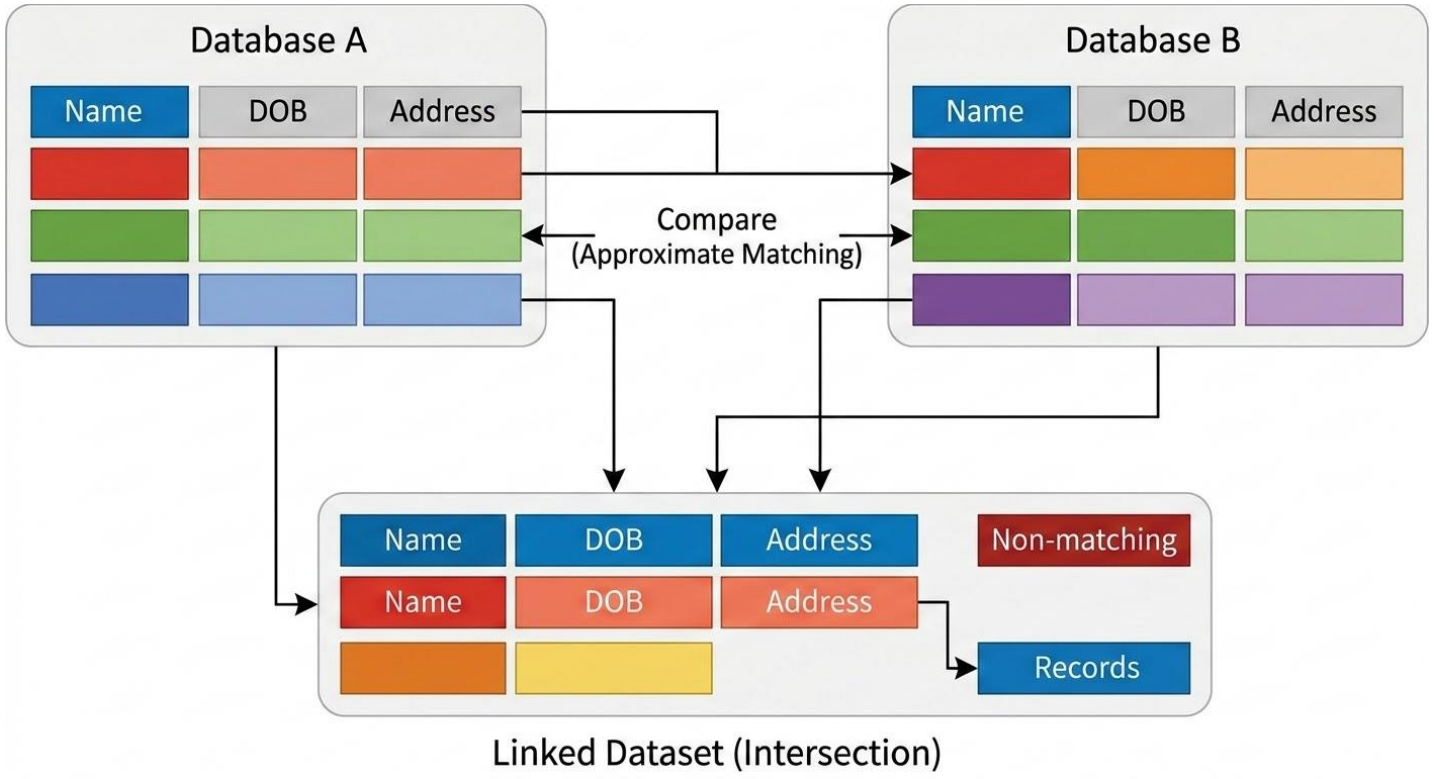



Image created by Gemini

# The Need for Privacy in Record Linkage

- Personal information is highly sensitive. Sharing it in plain text violates legal regulations (e.g., GDPR, HIPAA) and erodes public trust.
- **The Question:** How do we link records across organizations without revealing the sensitive data itself?

# Preliminaries: What is PPRL?

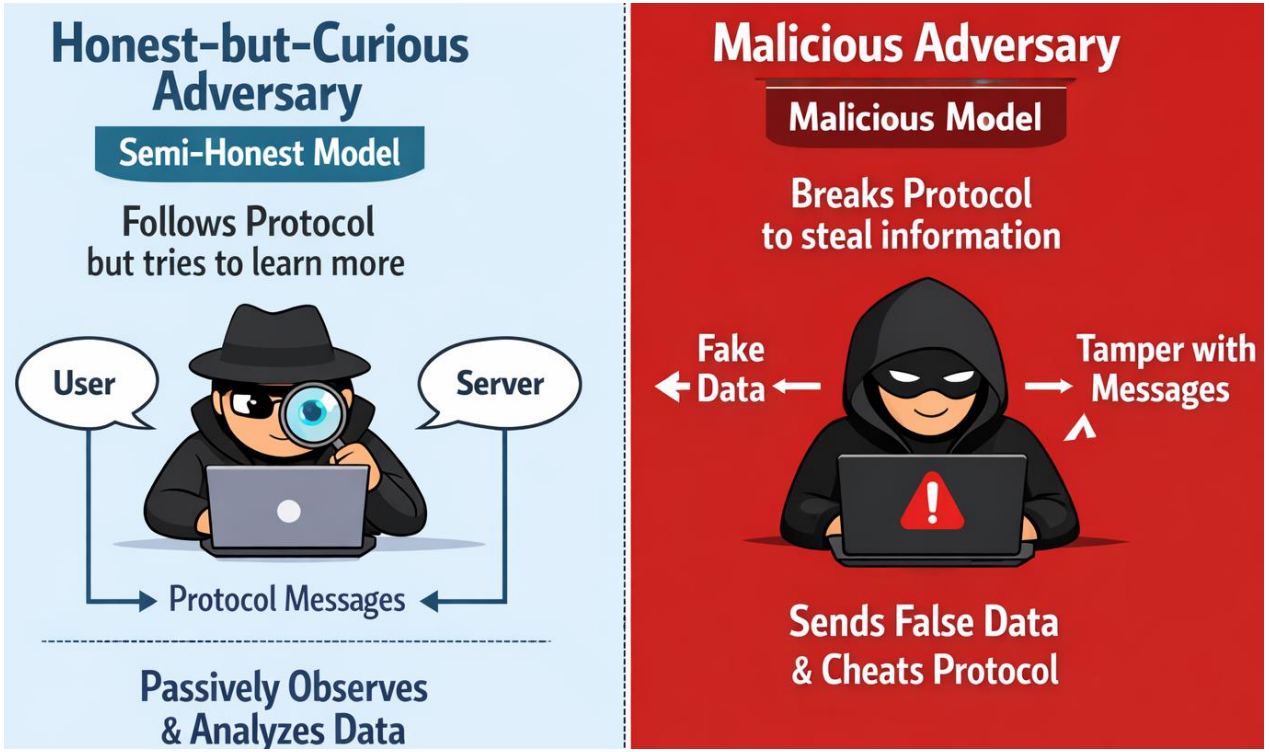

Image created by Gemini

# Foundational Challenges of PPRL

- **Privacy (data remains secure)**
- **Linkage Quality (typos, missing values, and variations)**
- **Scalability (compare millions of records)**

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

# Adversary models




Created by Gemini

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

# The End-to-End PPRL Workflow

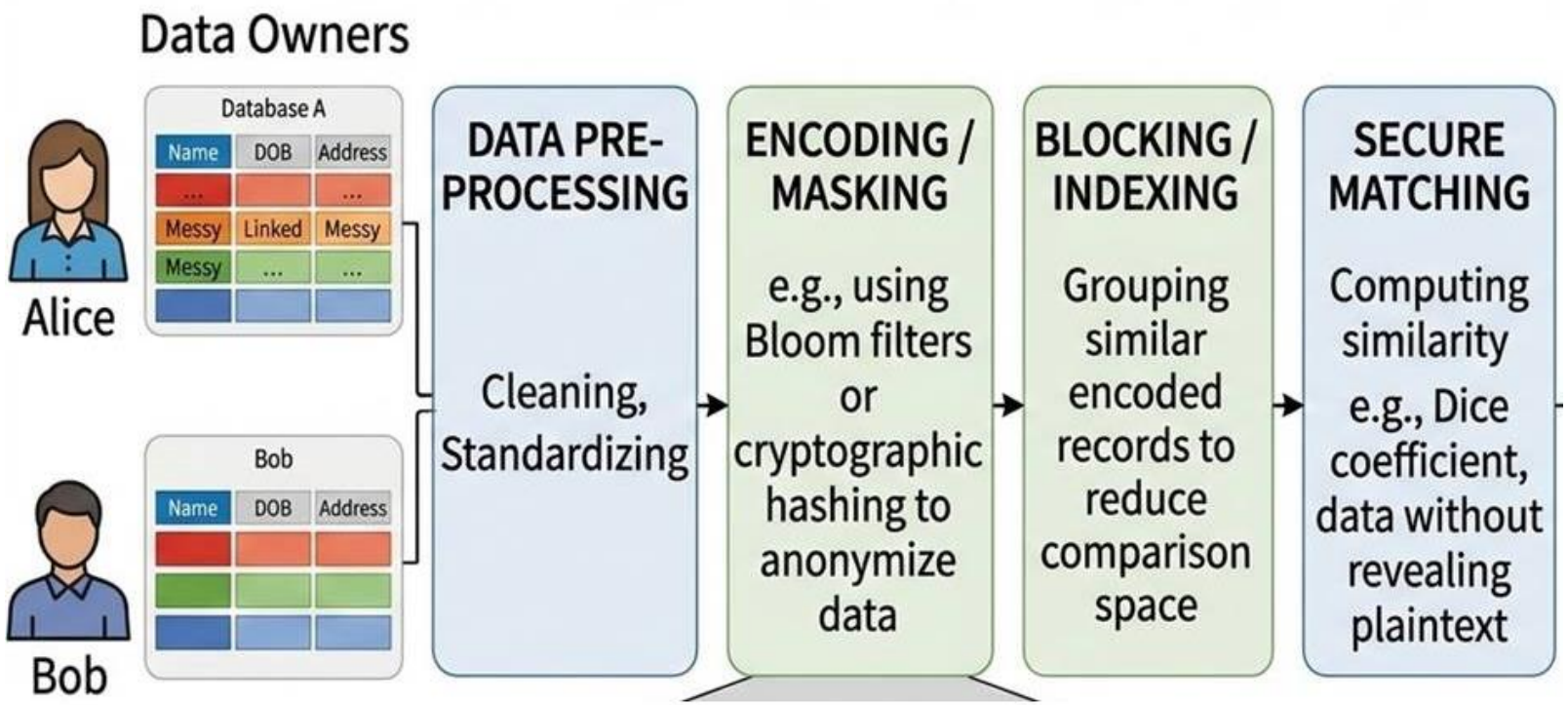
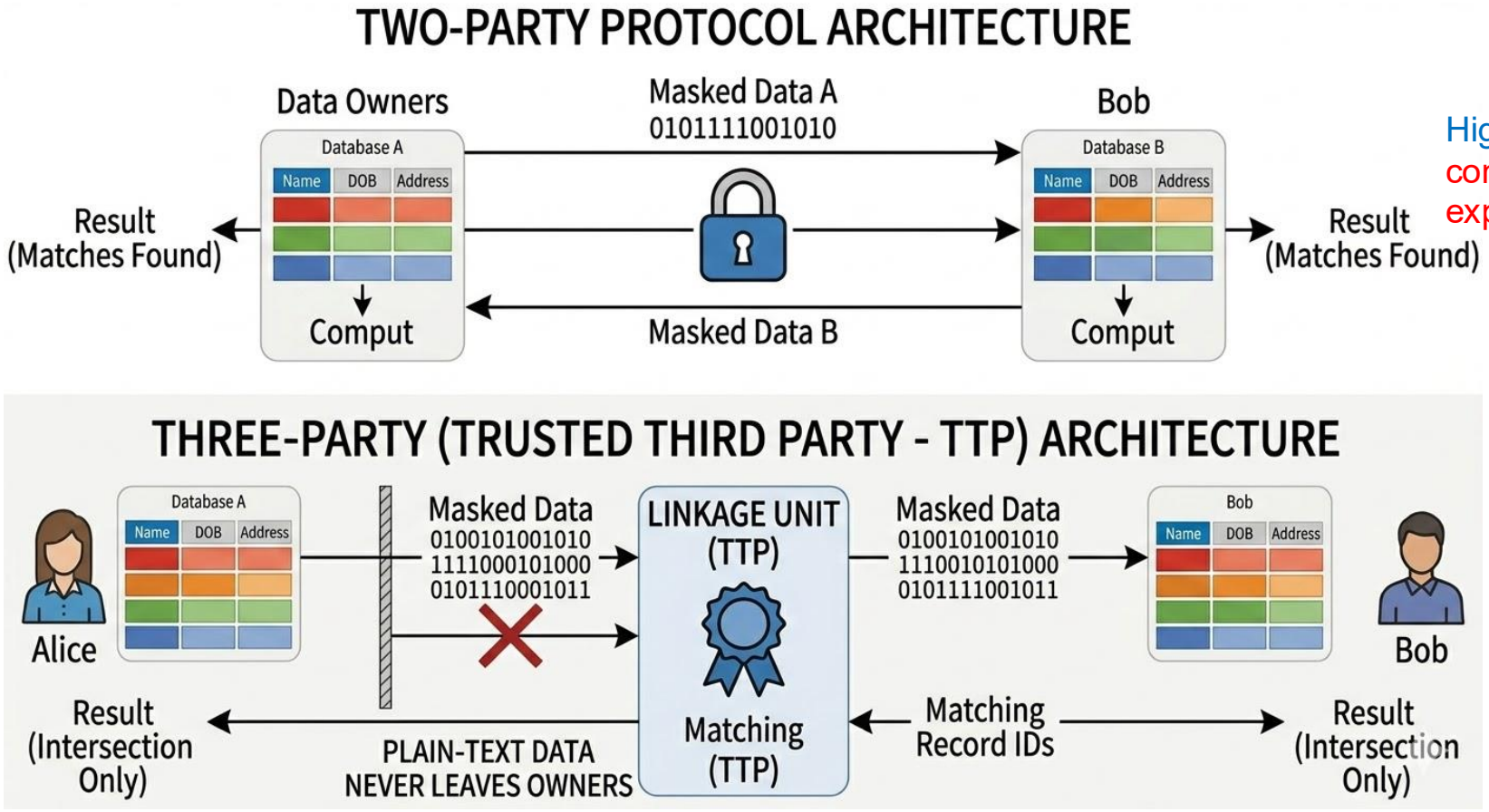



Image created by Gemini

# Architectural Protocols




Highly secure but computationally expensive.

The TTP never sees the plaintext data; The owners never see each other's data. Highly scalable and the most common practical architecture.

Image created by Gemini

# Practical Applications

- **Electronic Health Records (EHR):** Linking patient trajectories.
- **Public Health Surveillance:** Cross-referencing travel registries and hospital admissions during a pandemic.
- **National Security & Crime:** Identifying suspects by securely linking law enforcement databases across different countries.
- **Fraud Detection:** Banks collaborating to identify overlapping suspicious transaction patterns without violating client confidentiality.

# Foundations of PPRL (The Past)

# Period: 1998 - 2020

**Core Areas Covered:**

- Encoding Techniques (Exact, Approximate, and Multi-modal)
- Blocking and Indexing Strategies (Scaling the comparison space)
- Matching Techniques (Non-learning & Learning-based Methods)

**Objective:** To trace the development of PPRL from early exact-matching cryptographic protocols through the data-driven algorithms of 2020.

# SMC & Cryptographic Hashes

(Van Eycken et al. 2000, Weber et al. 2012)

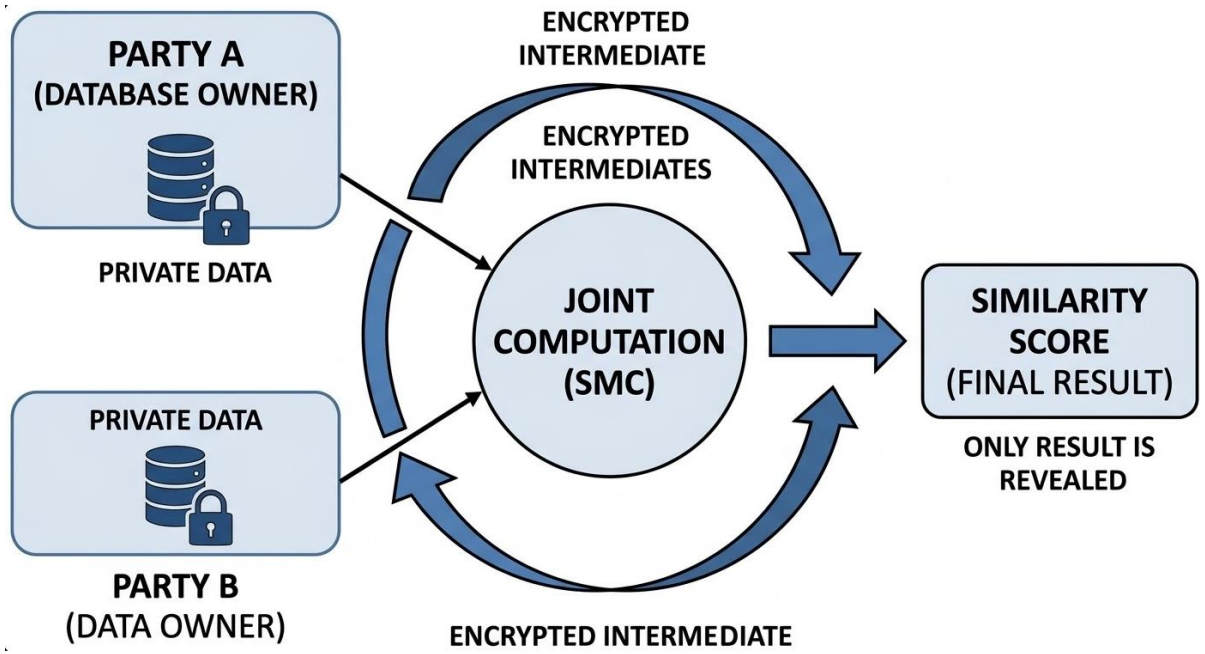



Image created by Gemini

# Limitations

- While these methods securely masked data, they only allowed for exact matching.
- They were computationally expensive.

# Approximate Matching

- The transition was motivated by the reality of dirty data; exact matching missed too many true links due to typos and lexical variations.

# Phonetics (Karakasidis and Verykios 2009)

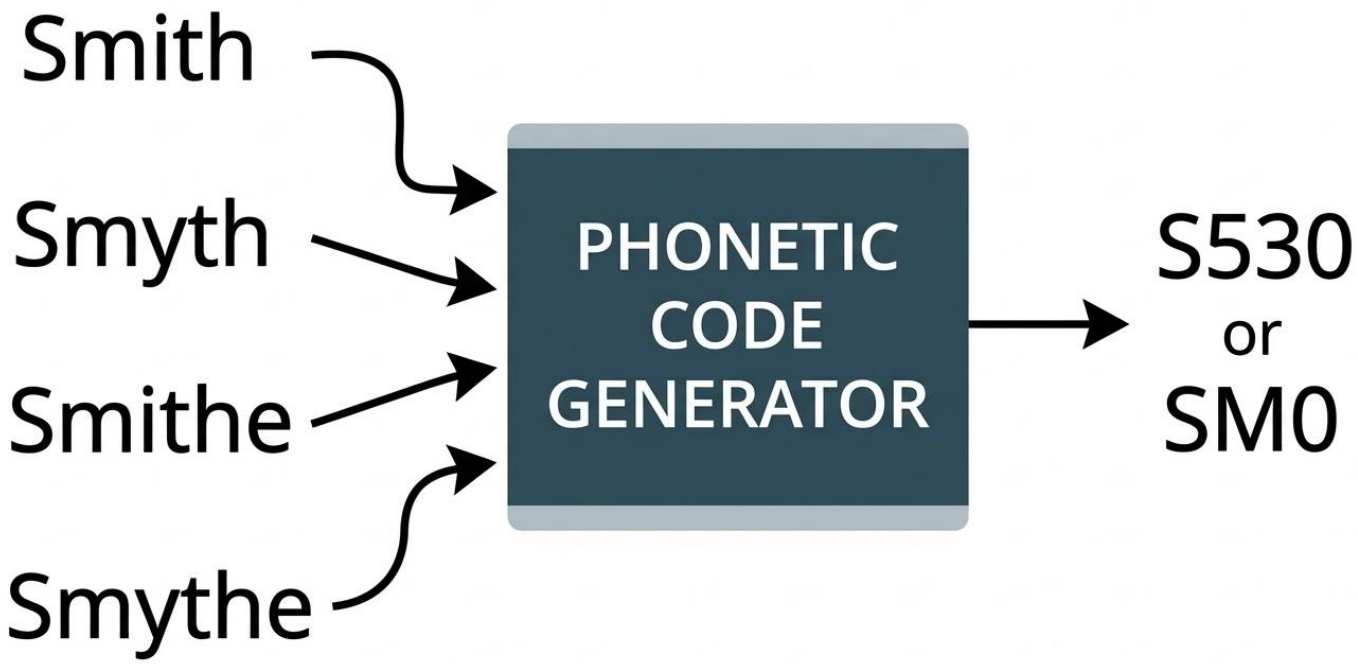
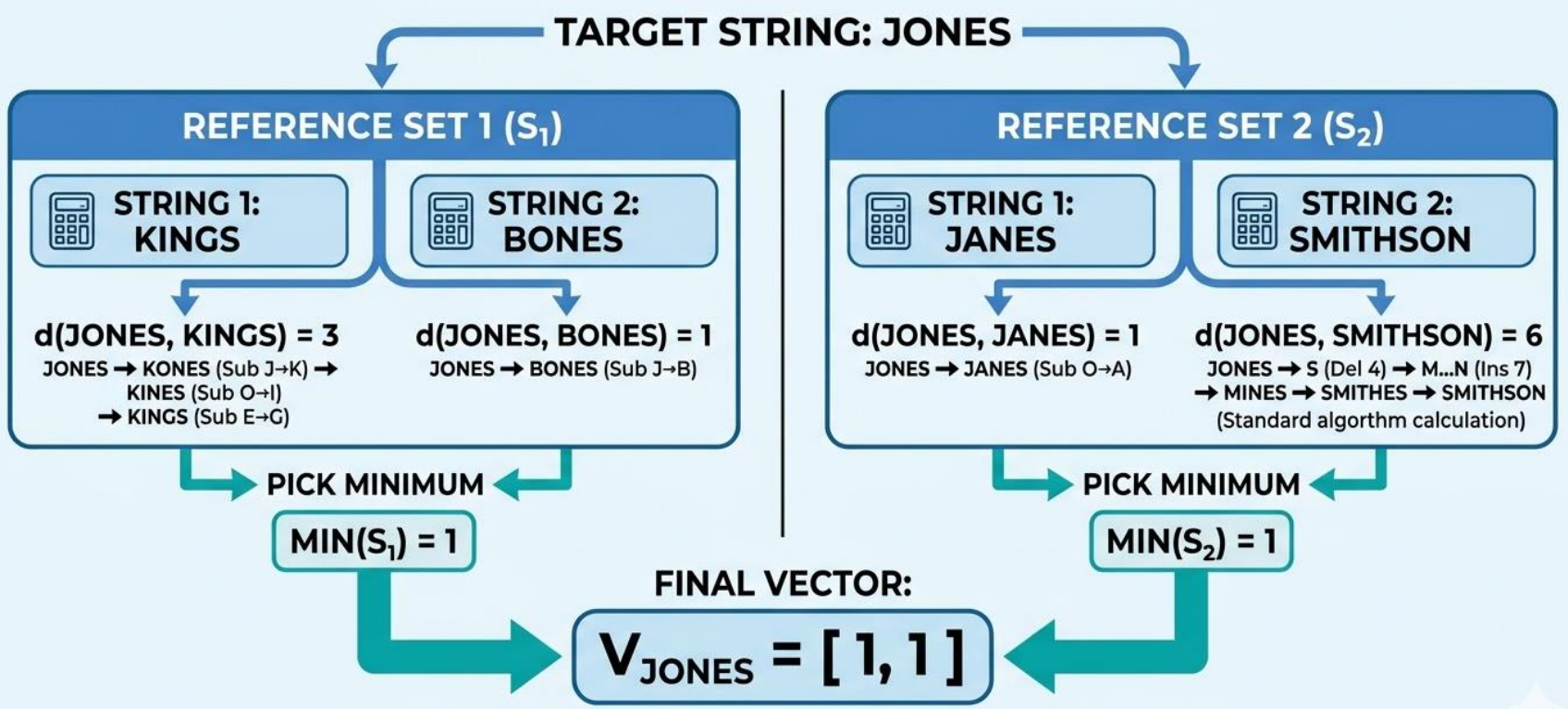



Image created by Gemini

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

# Distance-Preserving Embeddings

## (Scannapieco et al. 2007)




Each number in the vectors represents the minimum edit distance between our string and a specific element in a public reference set.

Image created by Gemini

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

# Limitations

- While fast, both approaches suffer from an inability to capture similar strings effectively, and a vulnerability to frequency attacks.

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

# The Advent of Bloom Filters
(Schnell et al. 2009, Durham et al. 2013)

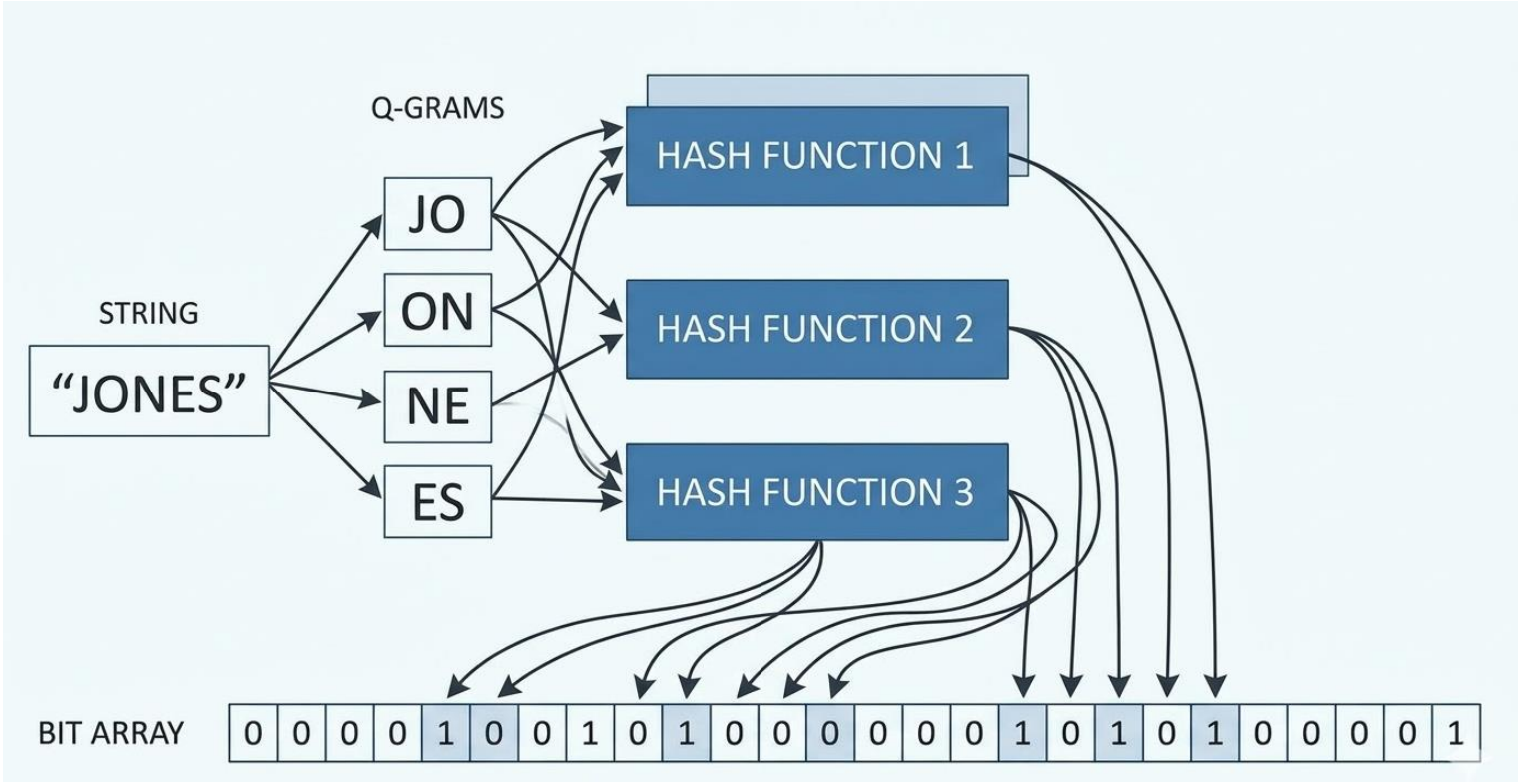



Image created by Gemini

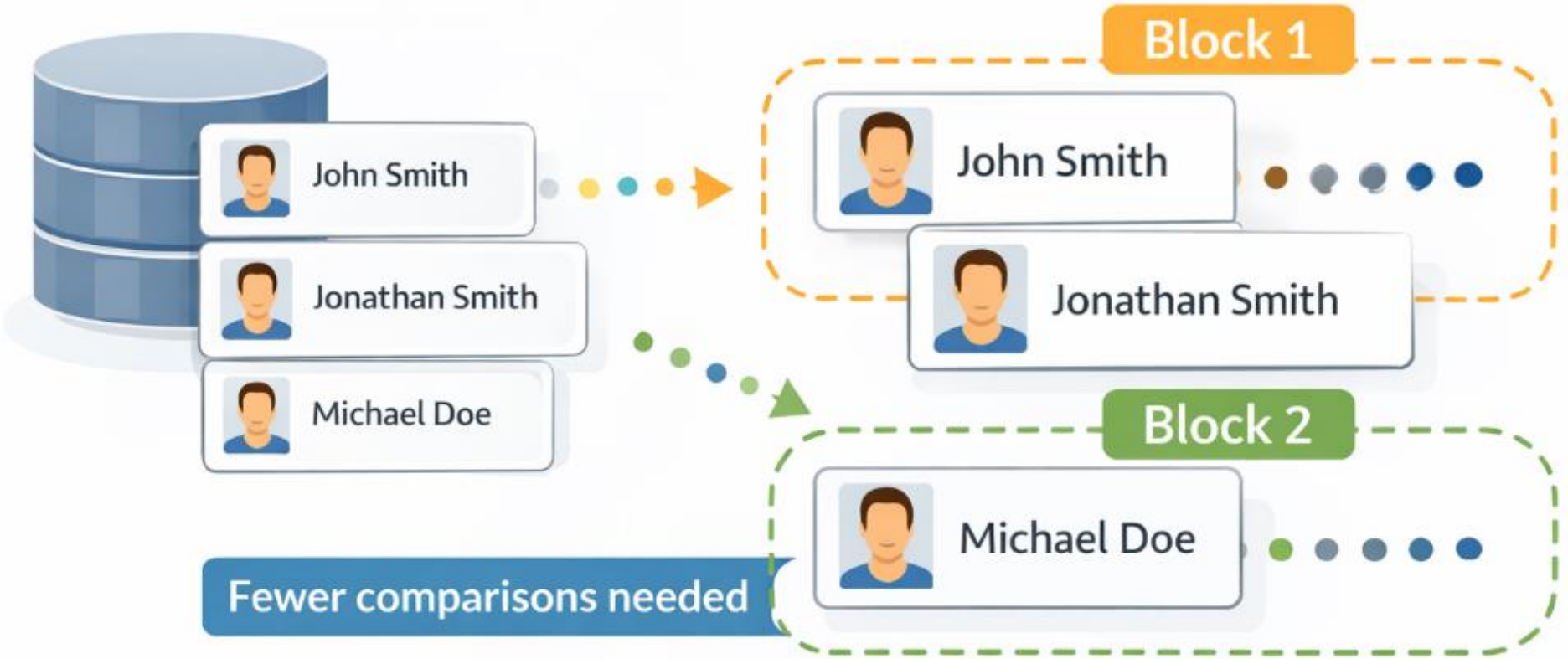
# Limitations

- Despite their immense utility, Bloom filters were eventually shown to be susceptible to cryptanalysis attacks, driving further research in later years (Kuzu et al. 2011, Niedermeyer et al. 2014).

# Beyond Strings

- **Numerical Data**: Bloom filters were proposed to securely encode **integers** and **floating-point numbers** by hash-mapping a set of neighboring values (Vatsalan and Christen 2014, Karapiperis et al. 2017).
- **Sequence & Image Data**: Order information is preserved within the Bloom filter space by appending monotonically increasing **time stamps** directly to the individual data points (Xue et al. 2020).

# Private Blocking/Indexing




Blocking techniques partition records into smaller blocks based on selected attributes to drastically reduce the search space and accelerate the linkage process, while maintaining strict privacy.

Image created by Gemini

# Spatial indexes & k-Anonymity

- **k-Anonymity Generalization**: Categorized records into generalized hierarchies based on semantics, ensuring each group contained at least ***k*** records to protect identities (Inan et al. 2008).
- **Spatial indexes**: KD-Trees, R-Trees, and Ball-Trees were employed to partition the comparison space and improve search efficiency (Scannapieco et al. 2007 , Inan et al. 2010).
- **Limitations**: Spatial indexes suffered from the curse of dimensionality while generalization techniques caused load imbalance problems with reduced linkage accuracy.

# Private Sorted Neighborhood (Vatsalan et al. 2013)

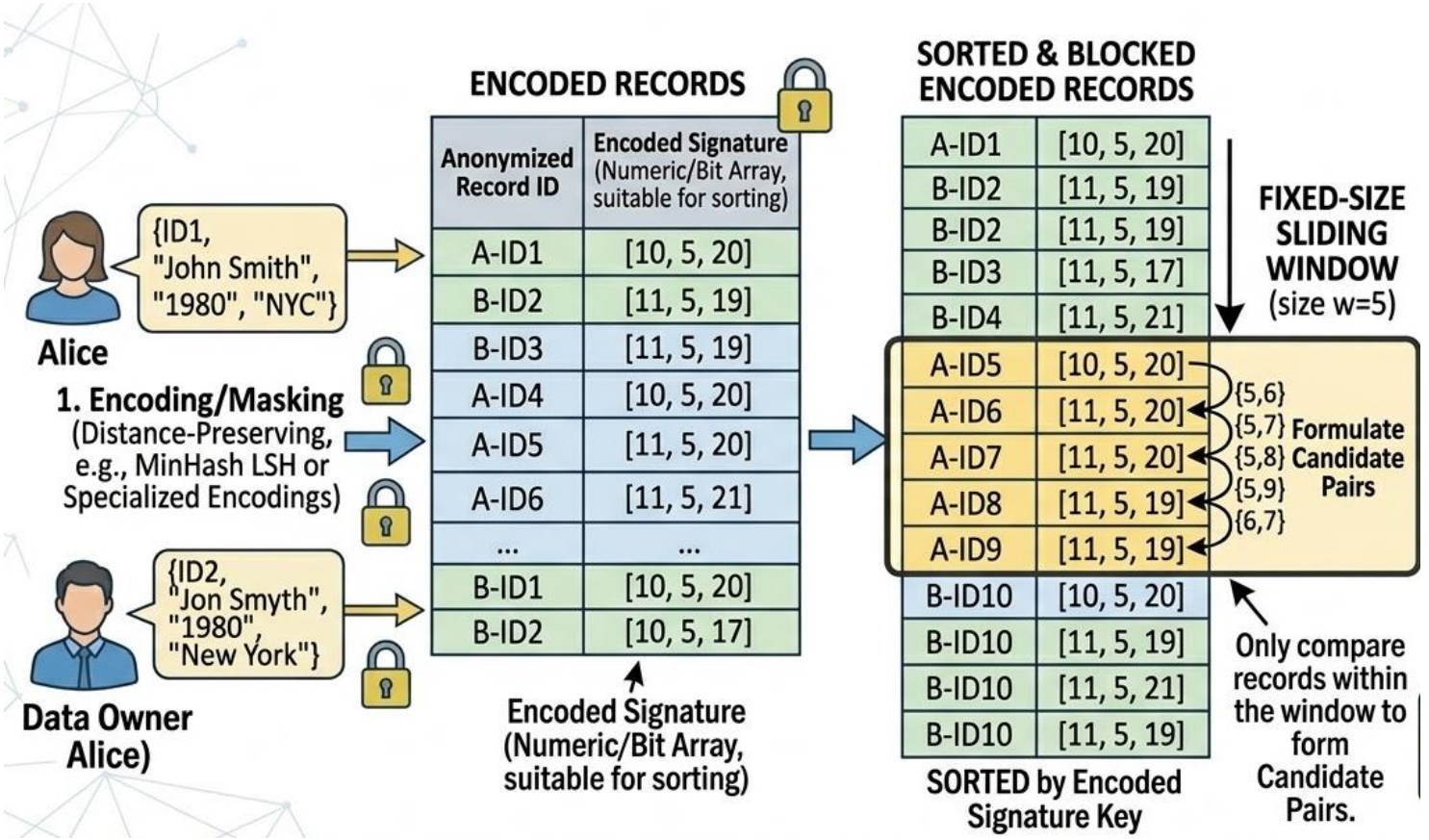



Image created by Gemini

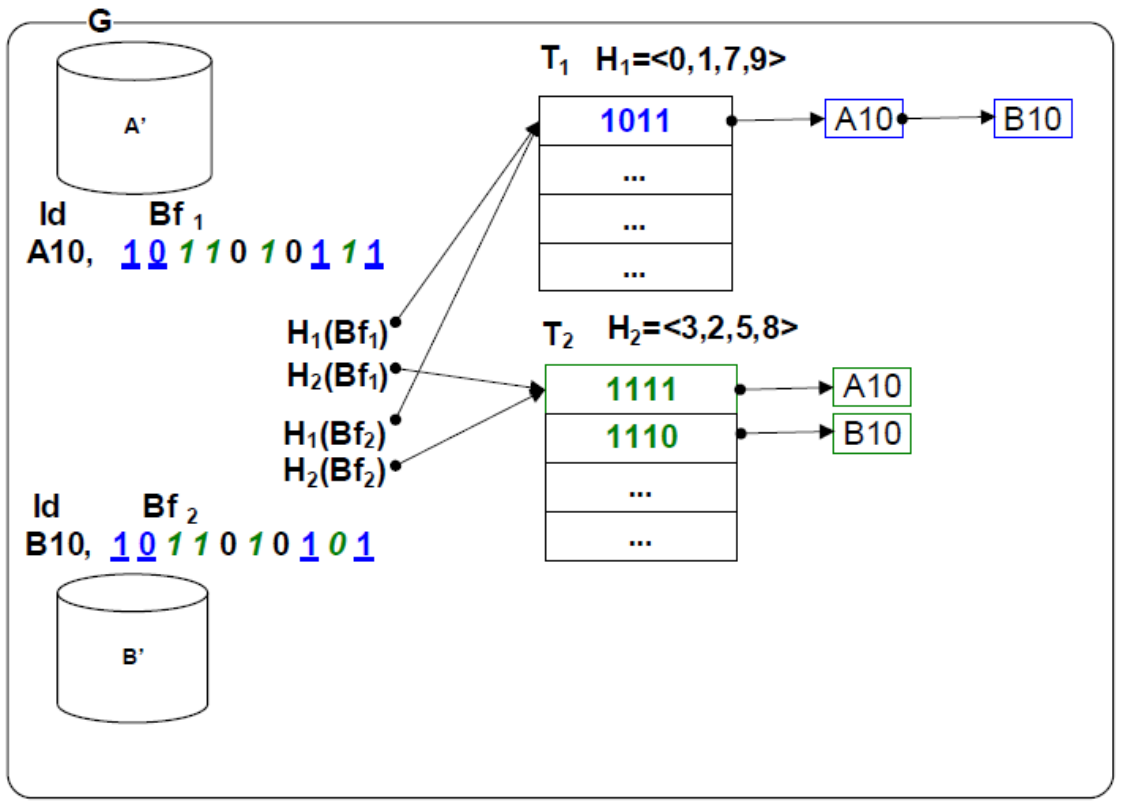
# Limitations

- Typos in the attributes used for the sorting key will push true matches far apart, resulting in missed links.
- Larger windows decrease efficiency, while smaller windows reduce matching accuracy.

# Locality-Sensitive Hashing (LSH)

- LSH became highly popular for embedding string values into metric spaces (Karapiperis and Verykios 2015).
- LSH provided **theoretical guarantees** for identifying similar pairs *in the metric space* with high probability.

# Hamming LSH with Bloom filters (Karapiperis and Verykios 2016)




**Hamming LSH** simply selects a random subset of bit indices from a Bloom filter to formulate index keys.
Bloom filters that share the exact same index key are placed into the same bucket.

Image taken from the paper

# Limitations

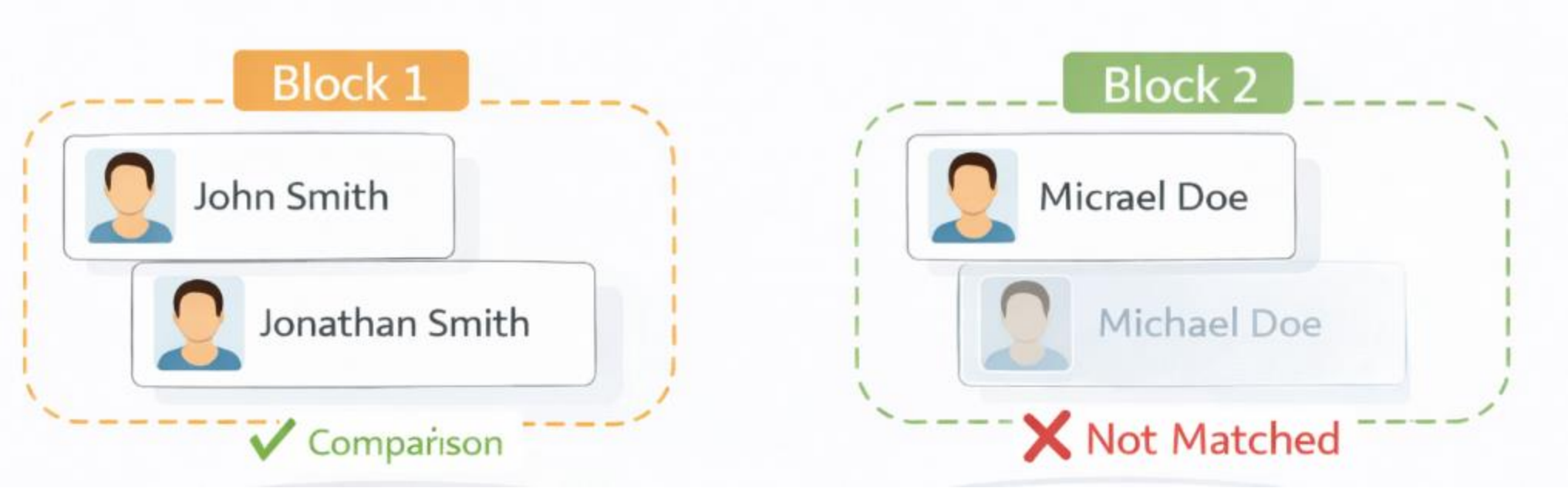
- LSH requires large amounts of memory to host its multiple indexes.
- It also requires careful tuning of data-dependent parameters and is largely restricted to specific encodings like Bloom filters.

# Further Blocking Strategies

- **Bit Tree** blocking allows multiple data owners to securely store encoded records in tree nodes without relying on a single linkage unit (Ranbaduge et al. 2015).
- **Meta-Blocking,** borrowed from traditional Entity Resolution, restructures block collections and prune unnecessary, redundant comparisons using graphs(Karakasidis et al. 2015, Ranbaduge et al. 2016).

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

# Secure Matching




Matching evaluates the encoded candidate pairs within each block to classify them as either matches or non-matches.

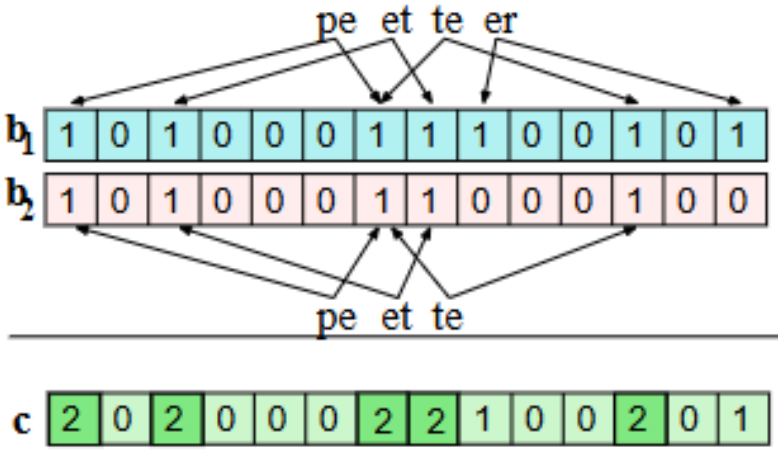
Image created by Gemini

# Matching: Early Non-Learning Methods

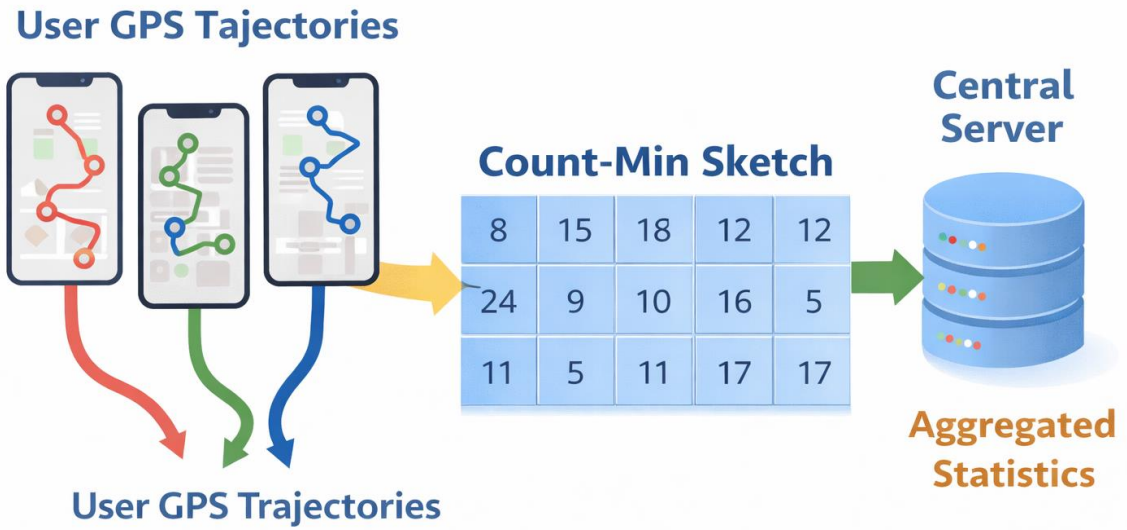
- **Secure set-intersection** and **stochastic scalar product** protocols were used to find common elements and measure Euclidean distances securely (Karapiperis et al. 2015).
- **Limitation:** While highly secure, these cryptographic protocols added **massive computational overhead**.
- **Similarity or distance measures** (e.g., Dice coefficient, Jaccard and Hamming distances) are calculated directly on the bit arrays of the Bloom filters (Durham et al. 2013, Karapiperis and Verykios 2015).
- **Limitation:** Inherent hash collisions within Bloom filters cause the measured distance in the encoded space to be smaller than the true distance in the plaintext space, degrading linkage accuracy.

# Aggregation-Based Matching

- **Counting Bloom Filters** (Vatsalan et al. 2017) and **Count-Min Sketches** (Yang et al. 2020)




Owners perform a secure summation protocol to aggregate Bloom filters into a single Counting Bloom Filter




A Count-Min sketch stores users locations, which allows a central server to query aggregated statistics

Image created by Gemini

# Limitations

- **Counting Bloom filters** face a strict trade-off where tuning parameters to increase privacy inherently degrades linkage quality.
- **Count-Min sketches** are restricted to categorical data, making them unsuitable for approximate string matching.

# The Shift to Learning-Based Methods

- This era integrated **Machine Learning** directly on encoded data (Erlingsson et al 2014, Vatsalan and Christen 2016, Xue et al. 2020).
- Because distances are preserved in Bloom filter encoded spaces, they were successfully used as inputs for **Support Vector Machines** and **random forests** classifiers.
- **Limitations:** Supervised models require labeled ground-truth data for training, which is not easy to obtain in strict privacy-preserving environments.

Funded by the European Union

RECITALS
Privacy-Preserving Data Sharing & ID Management

# The Current Landscape of PPRL

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

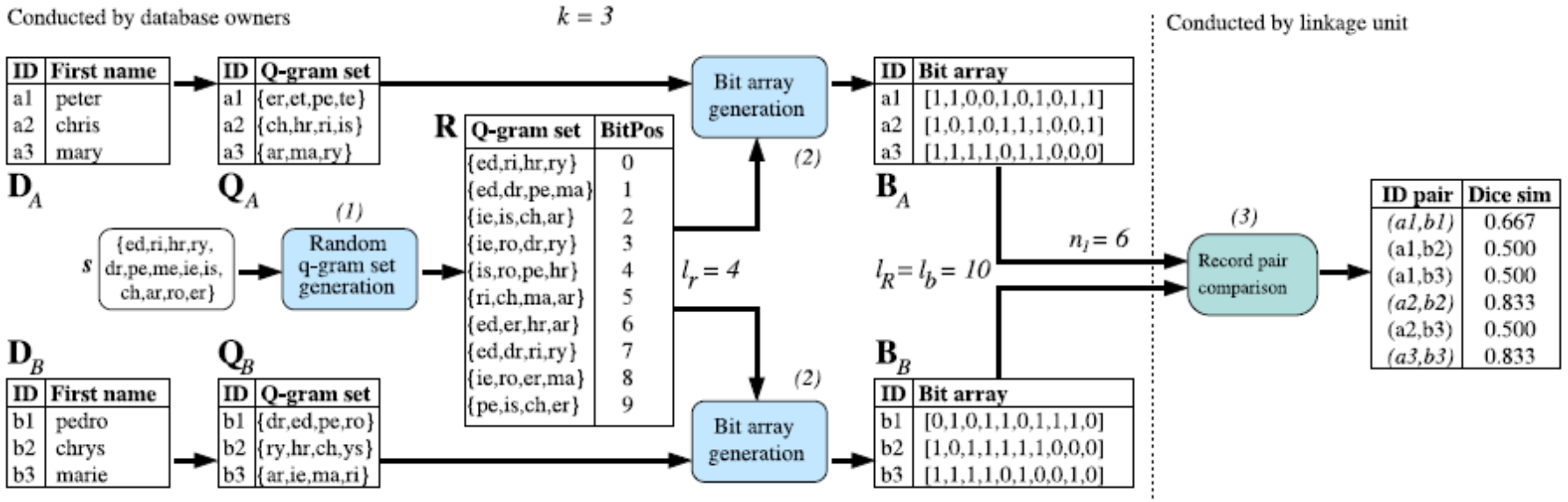
# Recent Advances: 2021–2025

**Core Areas Covered:**

- Advanced Cryptanalysis Defenses (Rank Swapping & Autoencoders)
- Circuit-Based Protocols for Fuzzy Matching
- Deep Learning Integration (Representation Learning & SNNs)
- Differential Privacy in Localized Training

**Objective:** To explore how the newest generation of PPRL utilizes deep learning and advanced protocols to address privacy and matching accuracy.

# Defending Against Frequency Attacks (Ziyad et al. 2025)



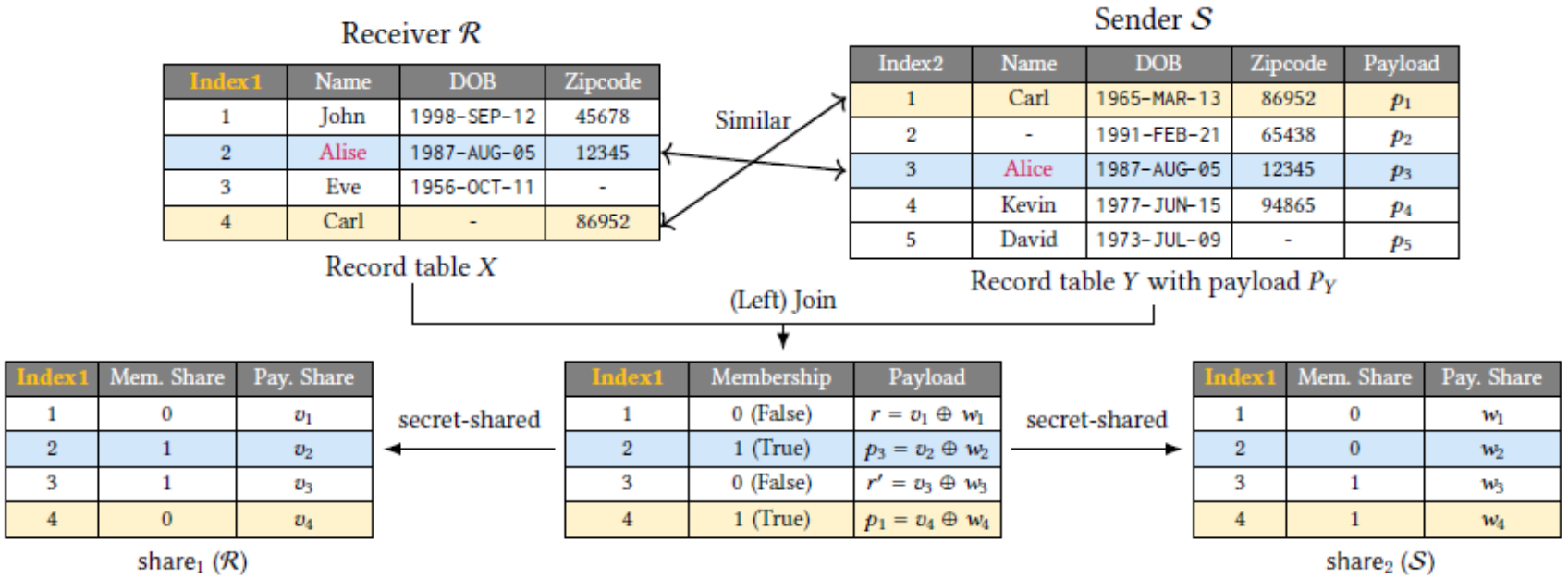

Owners locally convert their records into q-grams and encode them into **fixed-weight bit arrays** based on their similarity to a shared, random reference set.

Image taken from the paper

# Limitations

- Comparing every local record against a massive, shared reference set requires significant computational effort prior to linkage.
- Forcing outputs into fixed-weight arrays discards precise similarity values, potentially increasing false positives.

# A Circuit-Based Protocol for Fuzzy Matching (Han et al. 2025)



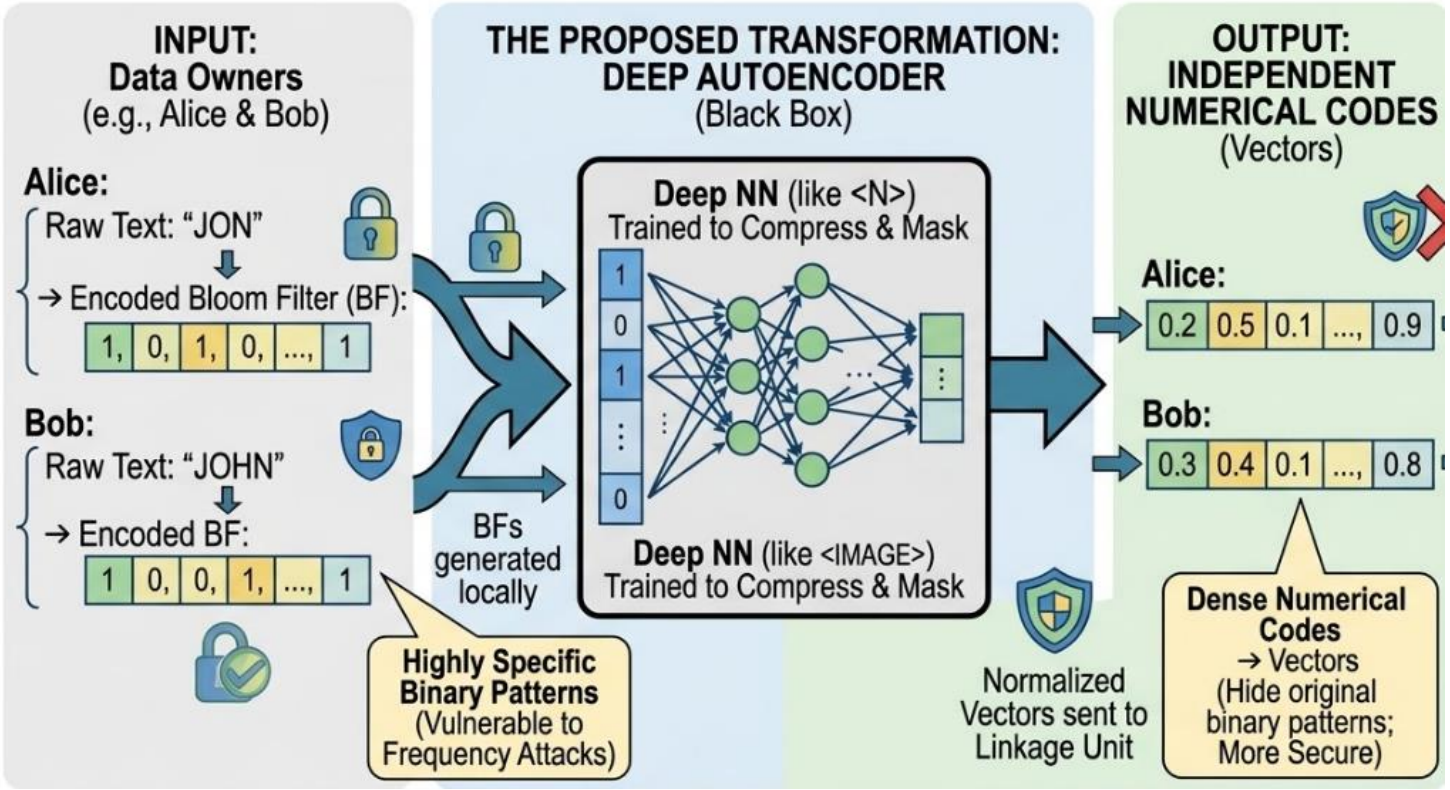

Two parties locally encode their records, and then use **cuckoo hashing** for blocking. They compare these records using a secure circuit protocol, ultimately outputting only **secret shares** of the final matches, so that neither party learns anything about the other's unlinked data.

Image taken from the paper

# Limitation

- The protocol requires heavy, multi-round communication between the two parties to securely compute the circuit.

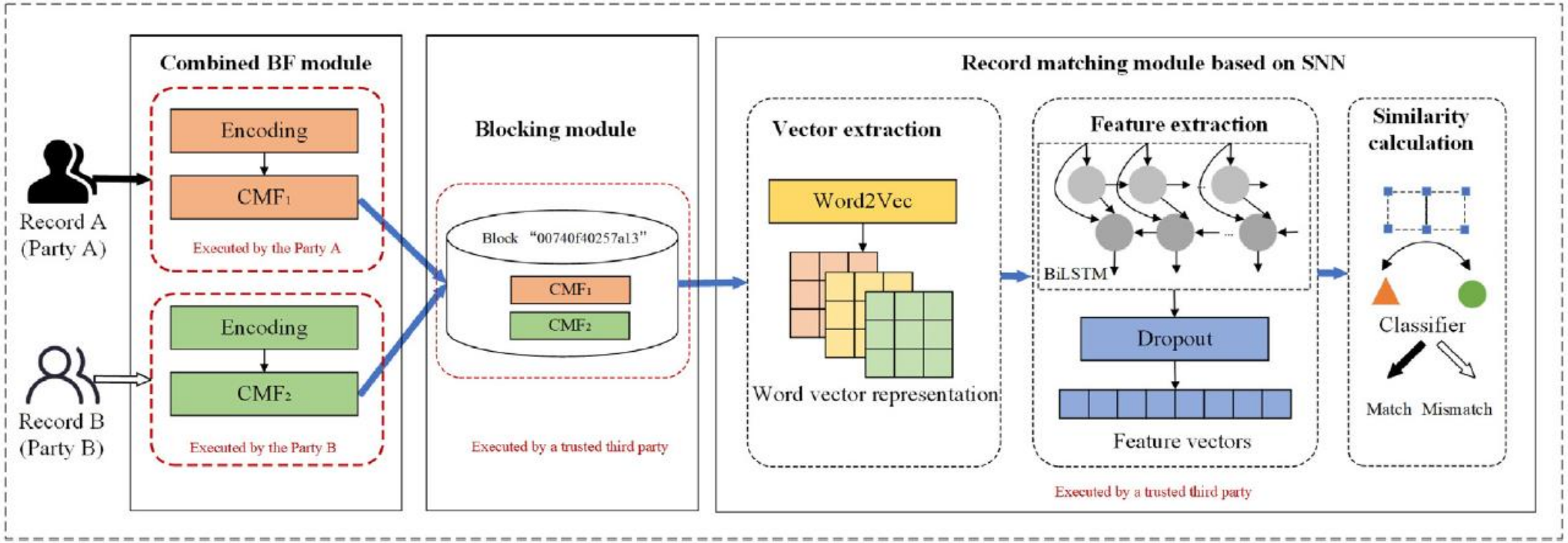
# Autoencoders (Christen et al. 2023)




Bloom filters are fed into an autoencoder neural network, which compresses them into lower-dimensional dense embeddings. The owners send ***only* these dense embeddings** to the Linkage Unit. They completely hide the underlying bit patterns.

Image created by Gemini

# Deep Learning for Matching Accuracy (Yao et al. 2023)

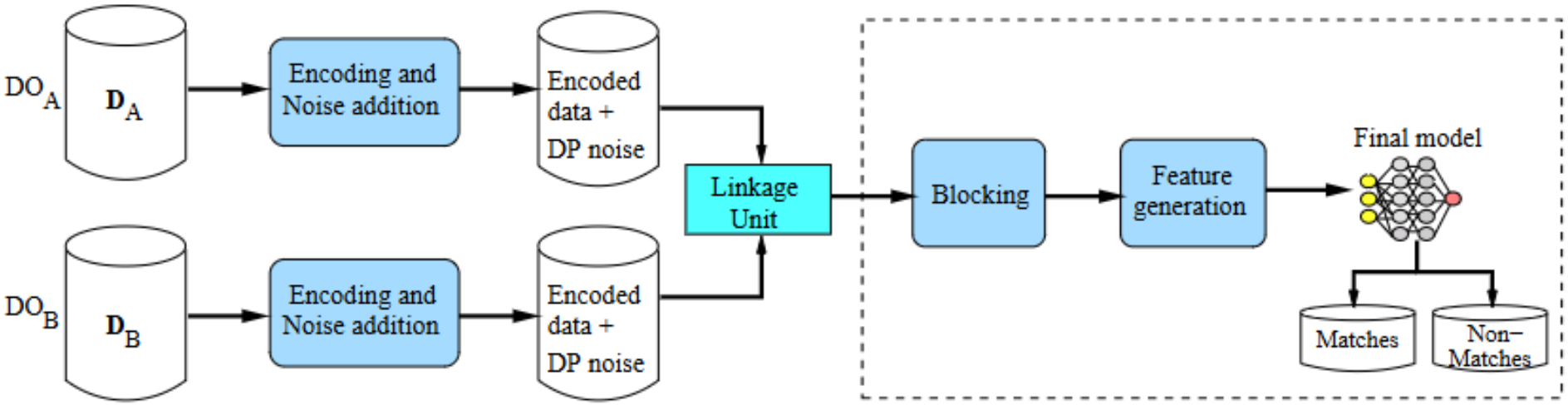



Owners first secure their records locally using Bloom filters.
A trusted third party feeds the candidate pairs into a Siamese Neural Network to generate dense embeddings, using Word2Vec, which are then processed through a BiLSTM (RNN) neural network to classify the pairs as matches or non-matches.



Image taken from the paper

# Differential Privacy & Localized Training (Ranbaduge et al. 2024)




Data Owners train their models locally and transmit the model weights to a Secure Aggregator, which averages them to build a **global model**.
The owners encode their records into Bloom filters, inject Differentially Private noise, and send them to a Linkage Unit, which uses the global model to definitively classify the pairs as matches or non-matches.

Image taken from the paper

# Limitations

- Advanced models (especially supervised SNNs) require vast amounts of labeled ground truth data, which is rarely accessible across privacy-preserving environments.
- Training autoencoders, generating complex embeddings, and communicating federated weights require immense processing power and bandwidth.
- Injecting Differentially Private noise guarantees mathematical privacy but inherently degrades the accuracy of the global model.

# THANK YOU!

dkarapiperis@ihu.edu.gr

**RECITALS**

Privacy-Preserving Data Sharing & ID Management

**Funded by the European Union**

# Q&A

# References


- O. Goldreich, S.Micali, and A.Wigderson, "How to play any mental game," STOC, 1987, pp. 218–229.
- E. Van Eycken, K. Haustermans, F. Buntinx, A. Ceuppens, J. Weyler, E. Wauters, H. Van Oyen, M. De Schaever, D. Van den Berge, and M. Haelterman, "Evaluation of the encryption procedure and record linkage in the belgian national cancer registry," Archives of public health, vol. 58, no. 6, pp. 281–294, 2000.
- S. Weber, H. Lowe, A. Das, and T. Ferris, "A simple heuristic for blindfolded record linkage," Journal of the American Medical Informatics Association," 2012.
- A. Karakasidis and V. Verykios, "Privacy preserving record linkage using phonetic codes," BCI, 2009, pp. 101 – 106.
- M. Scannapieco, I. Figotin, E. Bertino, and A. Elmagarmid, "Privacy preserving schema and data matching," SIGMOD, 2007, pp. 653 – 664.
- R. Schnell, T. Bachteler, and J. Reiher, "Privacy-preserving record linkage using Bloom filters," Central Medical Inf. and Decision Making," vol. 9, 2009.
- E. Durham, M. Kantarcioglu, Y. Xue, C. Toth, M. Kuzu, and B. Malin, "Composite Bloom filters for secure record linkage," TKDE, 2013.

- M. Kuzu, M. Kantarcioglou, E. Durham, and B. Malin, "A constraint satisfaction cryptanalysis of Bloom filters in private record linkage," PETS, 2011, pp. 226 – 245.
- F. Niedermeyer, S. Steinmetzer, M. M. Kroll, and R. Schnell, "Cryptanalysis of Basic Bloom Filters Used for Privacy Preserving Record Linkage," JPC, vol. 6, no. 2, 2014.
- D. Vatsalan and P. Christen, "Privacy-preserving matching of similar patients," JBI, vol. 59, pp. 285–298, 2016.
- D. Karapiperis, A. Gkoulalas-Divanis, and V. S. Verykios, "Distance aware encoding of numerical values for privacy-preserving record linkage," ICDE, 2017, pp. 135–138.
- W. Xue, D. Vatsalan, W. Hu, and A. Seneviratne, "Sequence data matching and beyond: New privacy-preserving primitives based on Bloom filters," TIFS, vol. 15, pp. 2973–2987, 2020.
- Q. Yang, Y. Shen, D. Vatsalan, J. Zhang, M. A. Kaafar, and W. Hu, "P4mobi: A probabilistic privacy-preserving framework for publishing mobility datasets," TVT, 2020.
- A. Inan, M. Kantarcioglou, E. Bertino, and M. Scannapieco, "A hybrid approach to private record linkage," ICDE, 2008, pp. 496 – 505.

- A. Inan, M. Kantarcioglu, G. Ghinita, and E. Bertino, "Private record matching using differential privacy," EDBT, 2010.
- D. Vatsalan, P. Christen, and V. Verykios, "Efficient two-party private blocking based on sorted nearest neighborhood clustering," CIKM, 2013, pp. 1949 – 1958.
- D. Karapiperis and V. Verykios, "An LSH-based Blocking Approach with a Homomorphic Matching Technique for Privacy-Preserving Record Linkage," TKDE, vol. 27, no. 4, pp. 909–921, 2015.
- U. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," SIGSAC, 2014.
- T. Ranbaduge, P. Christen, and D. Vatsalan, "Clustering-based scalable indexing for multi-party privacy-preserving record linkage," PAKDD, Springer LNAI, Hanoi, 2015.
- A. Karakasidis, G. Koloniari, and V. S. Verykios, "Scalable blocking for privacy preserving record linkage," SIGKDD, 2015, pp. 527–536.
- T. Ranbaduge, D. Vatsalan, and P. Christen, "Scalable block scheduling for efficient multi-database record linkage," ICDM, 2016, pp. 1161–1166.
- D. Karapiperis and V. S. Verykios, "A fast and efficient hamming LSH-based scheme for accurate linkage," KAIS, vol. 49, no. 3, pp. 861–884, 2016.

- D. Karapiperis, A. Gkoulalas-divanis, and V. Verykios, "Distance-aware encoding of numerical values for privacy-preserving record linkage," ICDE, 2017.
- D. Vatsalan, P. Christen, and E. Rahm, "Scalable Multi-Database Privacy-Preserving RecordLinkage using Counting Bloom Filters," ICDMW, 2017.
- D. Karapiperis, D. Vatsalan, V. S. Verykios, and P. Christen, "Large scale multi-party counting set intersection using a space efficient global synopsis," DASFAA, 2015.
- A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings," PETS, vol. 2015, no. 1, pp. 92–112, 2015.
- Q. Bui-Nguyen, Q. Wang, J. Shao, and D. Vatsalan, "Repairing of record linkage: Turning errors into insight," EDBT, 2019, pp. 638–641.
- A. Vidanage, P. Christen, T. Ranbaduge, and R. Schnell, "A Vulnerability Assessment Framework for Privacy-preserving Record Linkage," TPS, 26, 3, 2023.
- S. Ziyad, P. Christen, A. Vidanage, C. Nanayakkara, and R. Schnell, "Privacy-preserving record linkage using reference set based encoding: A single parameter method," Information Systems 133, 2025.

- V. Christen, T. Häntschel, P. Christen, and E. Rahm, "Privacy-preserving record linkage using autoencoders," International Journal of DataScience and Analytics 15, 2023.
- K. Han, S. Kim, and Y. Son, "Private Computation on Common Fuzzy Records," PETS, 567–583, 2025.
- S. Yao, Y. Ren, D. Wang, Y. Wang, W. Yin, and L. Yuan, "SNN-PPRL: A secure record matching scheme based on siamese neural network.", Journal of Information Security and Applications," 76, 2023.
- T. Ranbaduge, D. Vatsalan, and M. Ding, "Privacy-Preserving Deep Learning Based Record Linkage," TKDE, 36, 11, 2024.