USENIX Security Symposium 2025

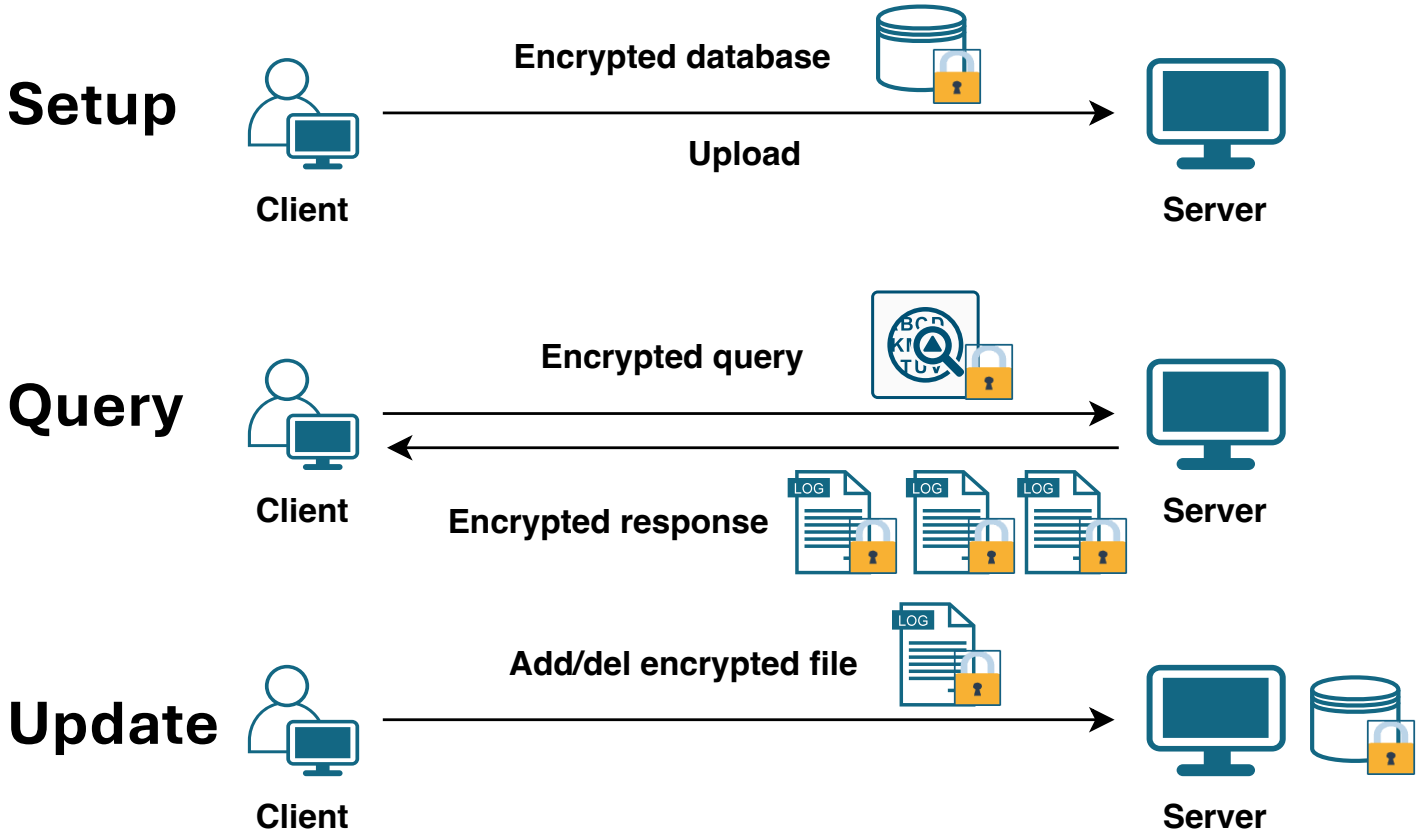
ALERT: Machine Learning-Enhanced Risk Estimation for Databases Supporting Encrypted Queries

Longxiang Wang[1*], Lei Xu[2,1*], Yufei Chen[1], Ying Zou[2], Cong Wang[1]

[1] City University of Hong Kong
[2] Nanjing University of Science and Technology

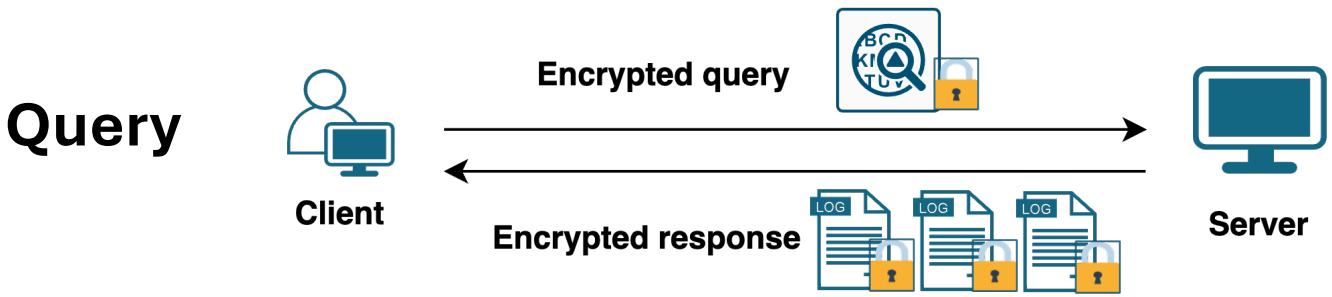
Background & Motivation

- **Dynamic Searchable Symmetric Encryption (DSSE):** Allows for searching over encrypted data with sublinear complexity.



Background & Motivation

- Encryption database **still leak metadata**
  - Leakage Abuse Attacks (LAA):



I can see
- The client access a file
- The file is #1
- #1's access control list

Background & Motivation

Our Observations

- Defensive approaches require a trade-off between security and efficiency

Background & Motivation

Our Observations

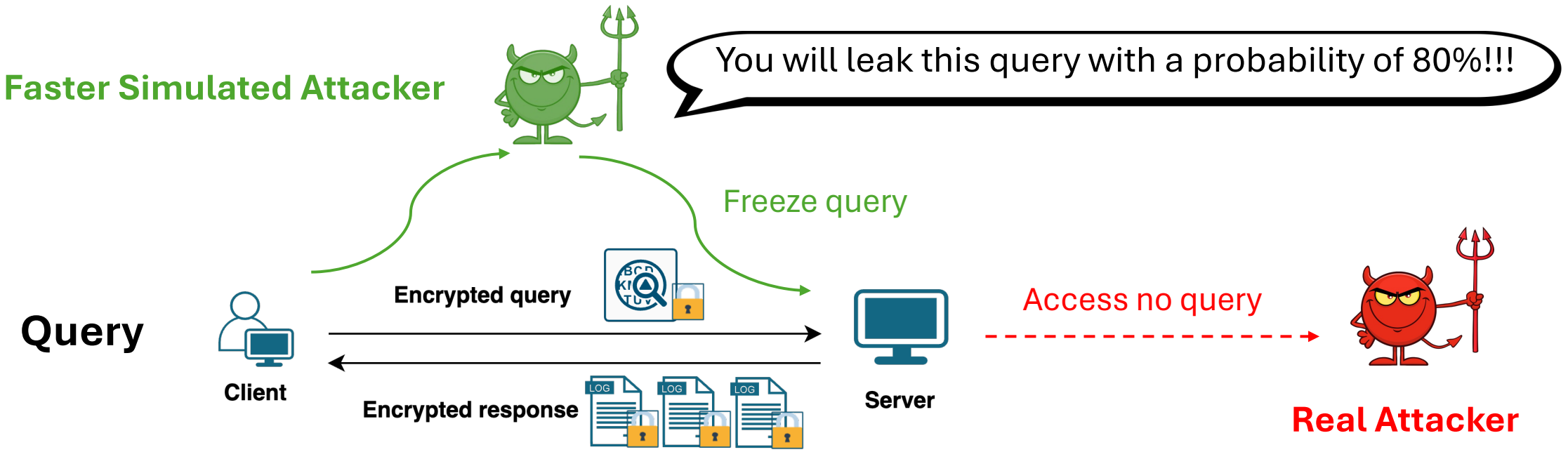
- Defensive approaches require a trade-off between security and efficiency

Is there a way that can prevent LAAs without sacrificing system performance?

Background & Motivation

Our Observations

- Defensive approaches require a trade-off between security and efficiency
- LAA is a natural way to assess leakage risks



Background & Motivation

Our Observations

- Defensive approaches require a trade-off between security and efficiency
- LAA is a natural way to assess leakage risks
- Previous LAAs[1-3] or leakage analysis approaches[4-6] focus on exploring new vulnerability vectors , not on efficiency

[1] Nie Hao et al. Query Recovery from Easy to Hard: Jigsaw Attack against SSE. USENIX Security 2024.
[2] Simon Oya et al. IHOP: Improved Statistical Query Recovery against Searchable Symmetric Encryption through Quadratic Optimization. USENIX Security 2022.
[3] Marc Damie, Florian Hahn, and Andreas Peter. A highly accurate query-recovery attack against searchable encryption using non-indexed documents. USENIX Security 2021.
[4] Evgenios M. Kornaropoulos, Nathaniel Moyer, Charalampos Papamanthou, and Alexandros Psomas. Leakage inversion: Towards quantifying privacy in searchable encryption. In Proc. of ACM CCS, 2022.
[5] Seny Kamara and Tarik Moataz. Bayesian leakage analysis: A framework for analyzing leakage in encrypted search. IACR ePrint., 2023.
[6] Alexandra Boldyreva, Zichen Gui, and Bogdan Warinschi. Understanding leakage in searchable encryption: a quantitative approach. Proc. Priv. Enhancing Technol., 2024(4):503–524, 2024.

Background & Motivation

Our Observations

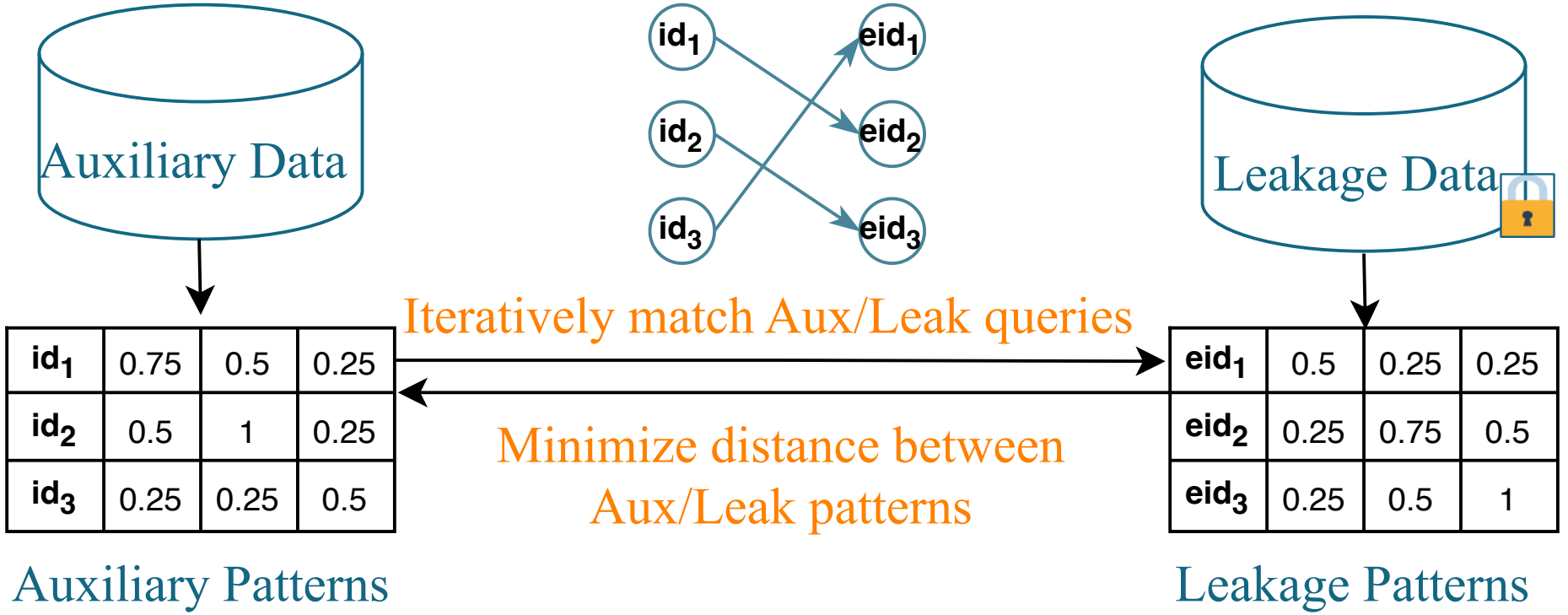
- Defensive approaches require a trade-off between security and efficiency
- LAA is a natural way to assess leakage risks
- Previous LAAs[1-3] or leakage analysis approaches[4-6] focus on exploring new vulnerability vectors , not on efficiency

How to build a LAA that achieves comparable performance while significantly reducing risk assessment latency?

[1] Nie Hao et al. Query Recovery from Easy to Hard: Jigsaw Attack against SSE. USENIX Security 2024.
[2] Simon Oya et al. IHOP: Improved Statistical Query Recovery against Searchable Symmetric Encryption through Quadratic Optimization. USENIX Security 2022.
[3] Marc Damie, Florian Hahn, and Andreas Peter. A highly accurate query-recovery attack against searchable encryption using non-indexed documents. USENIX Security 2021.
[4] Evgenios M. Kornaropoulos, Nathaniel Moyer, Charalampos Papamanthou, and Alexandros Psomas. Leakage inversion: Towards quantifying privacy in searchable encryption. In Proc. of ACM CCS, 2022.
[5] Seny Kamara and Tarik Moataz. Bayesian leakage analysis: A framework for analyzing leakage in encrypted search. IACR ePrint., 2023.
[6] Alexandra Boldyreva, Zichen Gui, and Bogdan Warinschi. Understanding leakage in searchable encryption: a quantitative approach. Proc. Priv. Enhancing Technol., 2024(4):503–524, 2024.

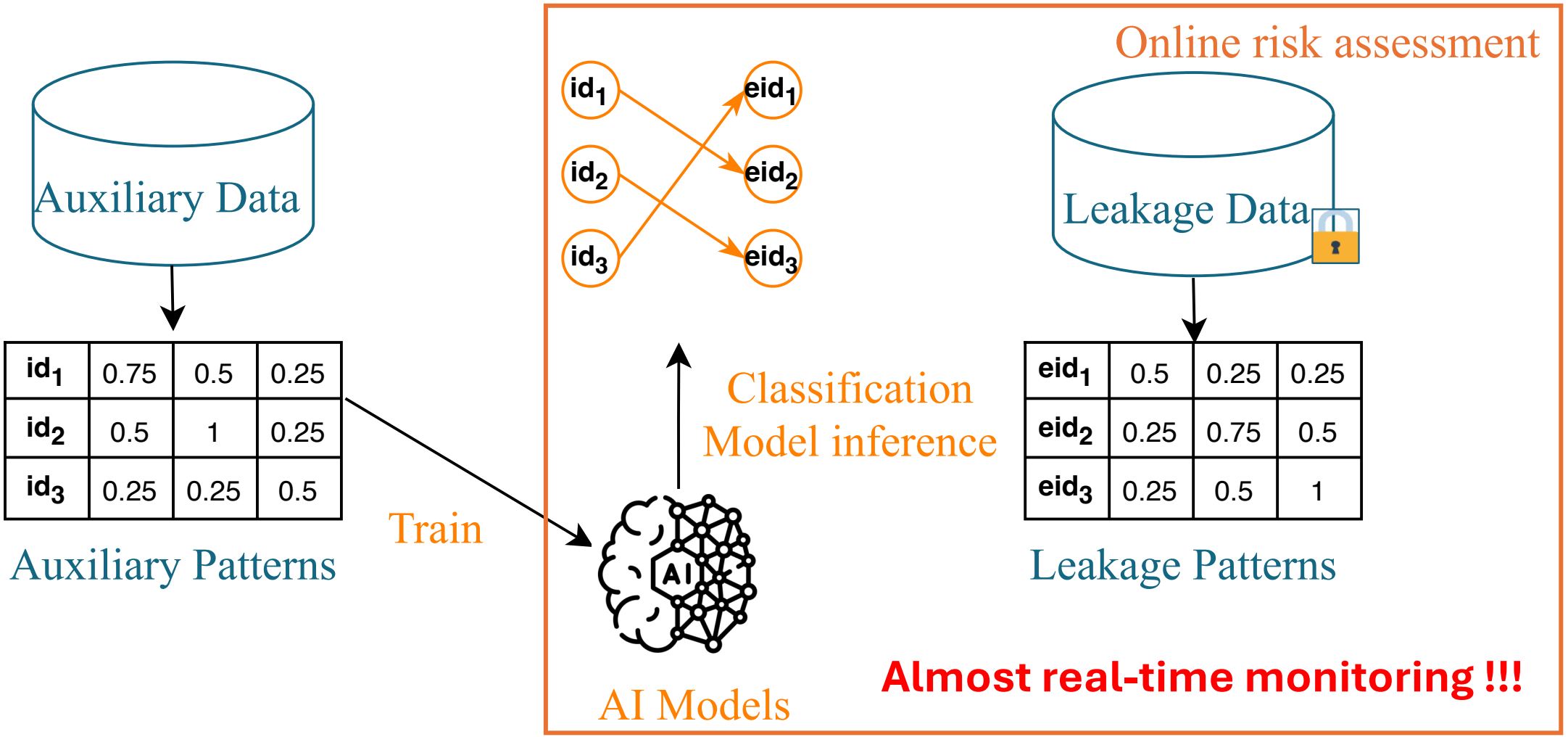
Problem Description

- Reformulate from heuristic optimization problems to machine learning multi-classification problems



Why and how to use AI for LAA

- Previous LAAs: Solve heuristic optimization problems
- **ALERT:** Learn and memorize auxiliary data in models



Technical Challenges

- Reformulate from heuristic optimization problems to machine learning multi-classification problems
  - **Training data misalignment:** Two types of data inconsistency
  - **Training scalability challenge:** Large number of classes (queries)
  - **Computational efficiency challenge:** Costly co-occurrence matrix update

ALERT– Module1: Data preparation

- Volume pattern, co-occurrence pattern
- **Training data misalignment:** Two types of data inconsistency
  - **Inconsistency1**: files across different timestamps

Timestamp1

	id_1	id_2	id_3
id_1	3	2	1
id_2	2	4	1
id_3	1	1	2

	id_1	id_2	id_3
id_1	1	0	-1
id_2	-0.22	1.09	-0.87
id_3	-0.58	-0.58	1.15

Vector Normalization

	id_x	id_y	id_z
id_x	30	20	10
id_y	20	40	10
id_z	10	10	20

Timestamp2

	id_1	id_2	id_3
id_1	1	0	-1
id_2	-0.22	1.09	-0.87
id_3	-0.58	-0.58	1.15

ALERT– Module1: Data preparation

- Volume pattern, co-occurrence pattern
- **Training data misalignment:** Two types of data inconsistency
  - **Inconsistency1**: files across different timestamps
  - **Inconsistency2:** unknown order among queries

Timestamp1

	id_1	id_2	id_3
id_1	3	2	1
id_2	2	4	1
id_3	1	1	2

↓

	id_1	id_2	id_3
id_1	3	2	1
id_2	4	2	1
id_3	2	1	1

Vector resorting

Timestamp2

	id_x	id_y	id_z
id_x	2	1	1
id_y	1	3	2
id_z	1	2	4

↓

	Id_x	id_y	id_z
id_x	2	1	1
id_y	3	2	1
id_z	4	2	1

ALERT– Module2: System Training

- **Naïve solution:** training a multi-classification model
  - Gradient Boosting Decision Trees (GBDT) model: Catboost[1]
- **Training scalability challenge:** large number of classes (queries) → Increased latency and low risk assessment performance

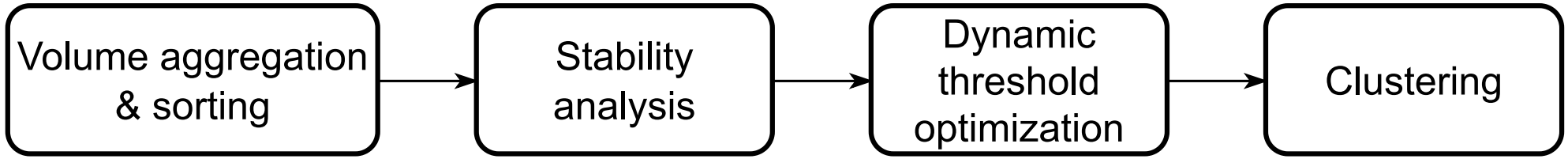
[1] Prokhorenkova et al. CatBoost: unbiased boosting with categorical features. NuerIPS, 2018.

ALERT– Module2: System Training

- **Naïve solution:** training a multi-classification model
  - Gradient Boosting Decision Trees (GBDT) model: Catboost[1]
- **Training scalability challenge:** large number of classes (queries) → Increased latency and low risk assessment performance

↓

- **Solution: pre-classify queries into clusters**
  - Base on historical volume information
  - Dynamic minimize threshold θ while constraining weighted average standard deviation within acceptable bounds

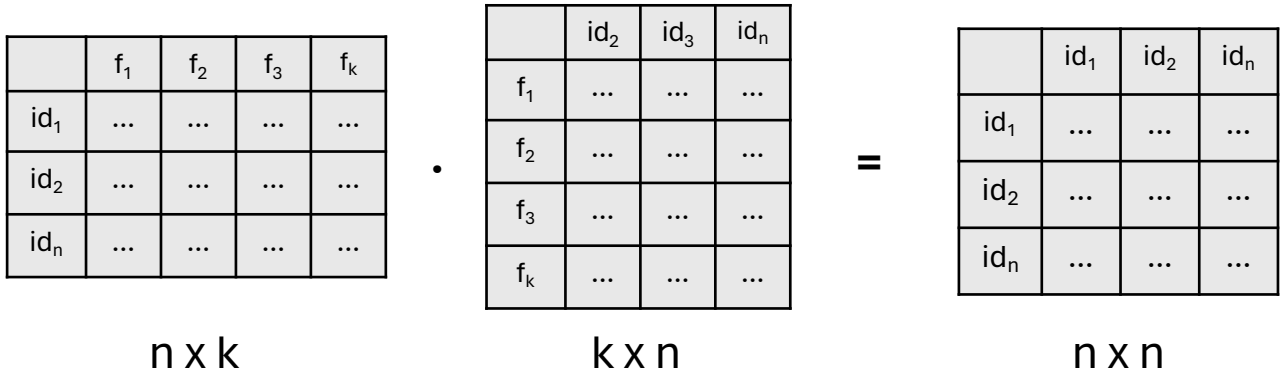


[1] Prokhorenkova et al. CatBoost: unbiased boosting with categorical features. NuerIPS, 2018.

ALERT– Module3: Risk Assessment

- **Computation efficiency challenge:** Costly co-occurrence matrix update
  → Concatenation approach

Original calculation

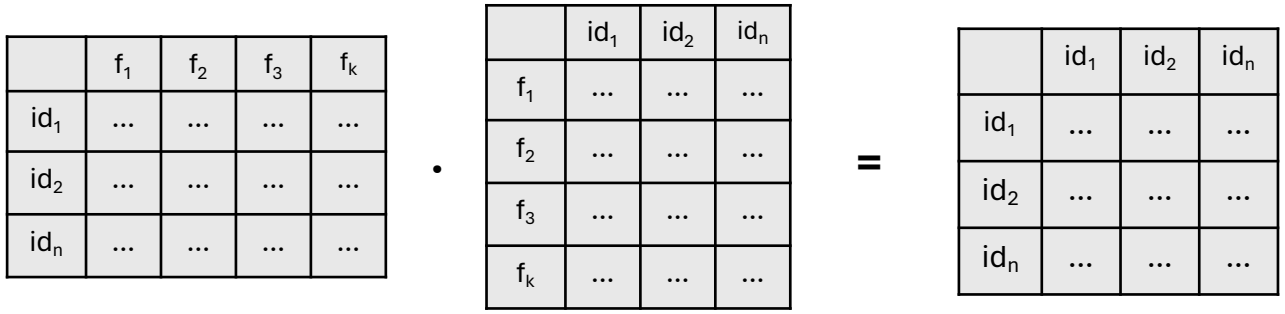


ALERT– Module3: Risk Assessment

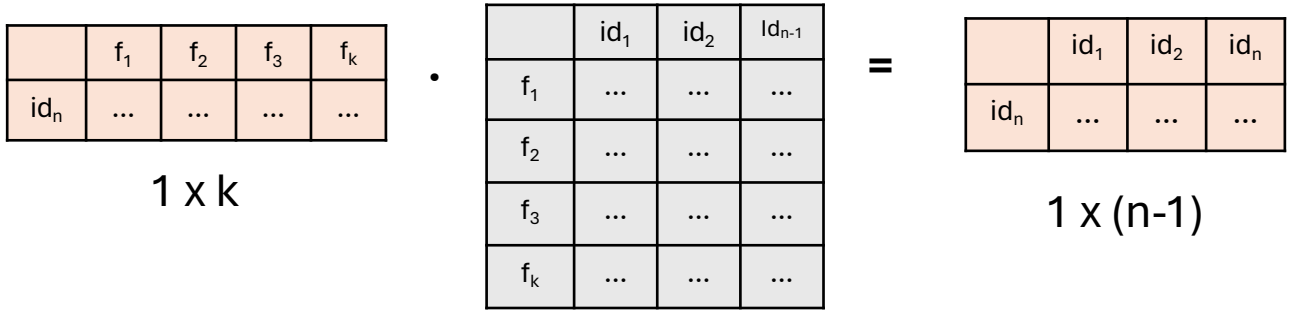
➢ **Computation efficiency challenge:** Costly co-occurrence matrix update

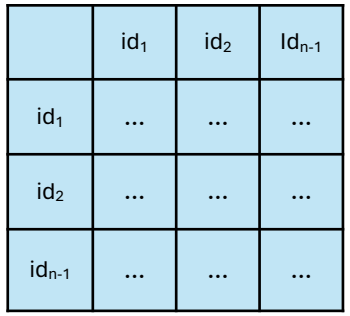
→ Concatenation approach

Original calculation

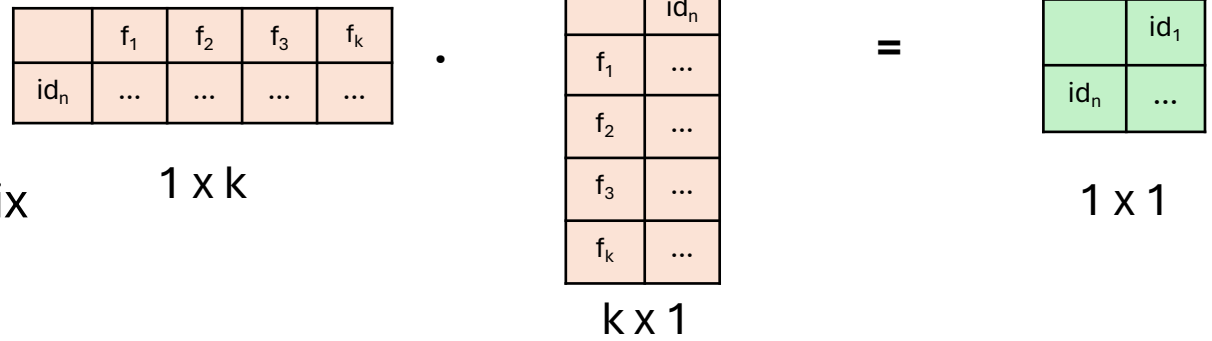


Optimized calculation





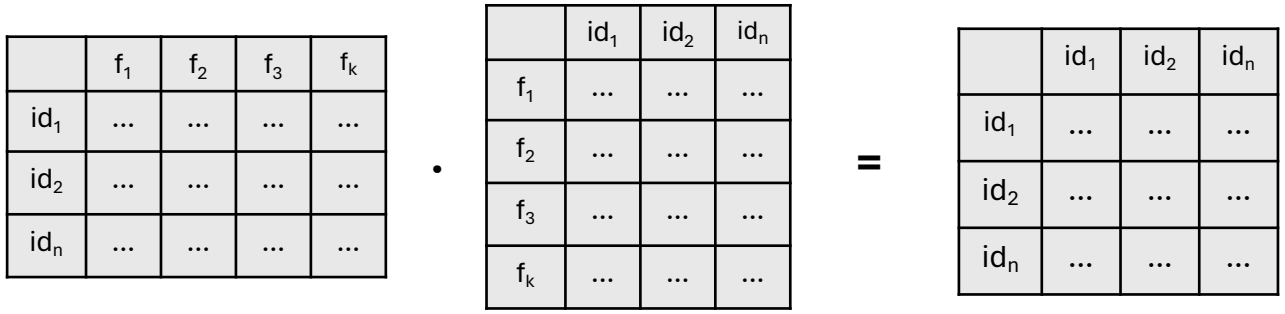
Previous co-occurrence matrix



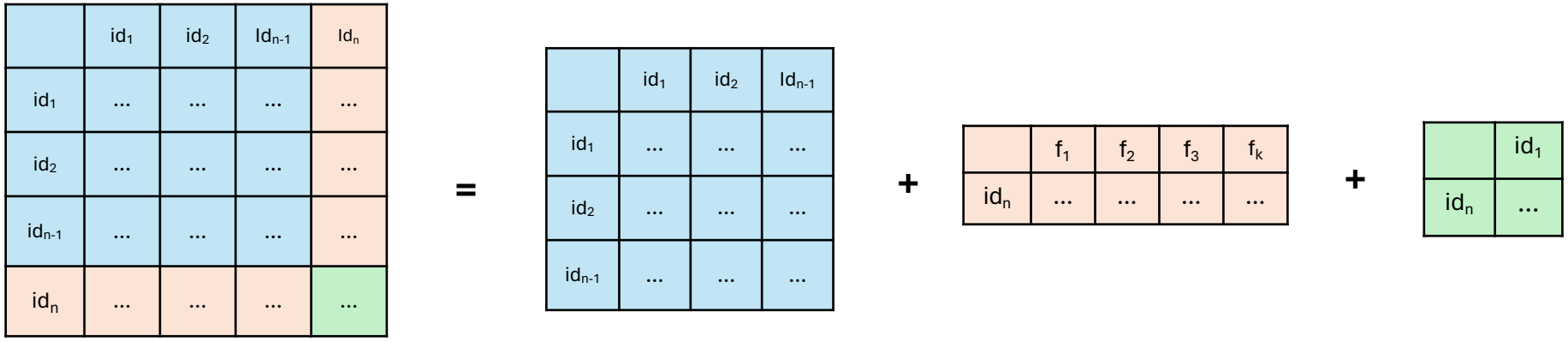
ALERT– Module3: Risk Assessment

- **Computation efficiency challenge:** Costly co-occurrence matrix update
  → Concatenation approach

Original calculation



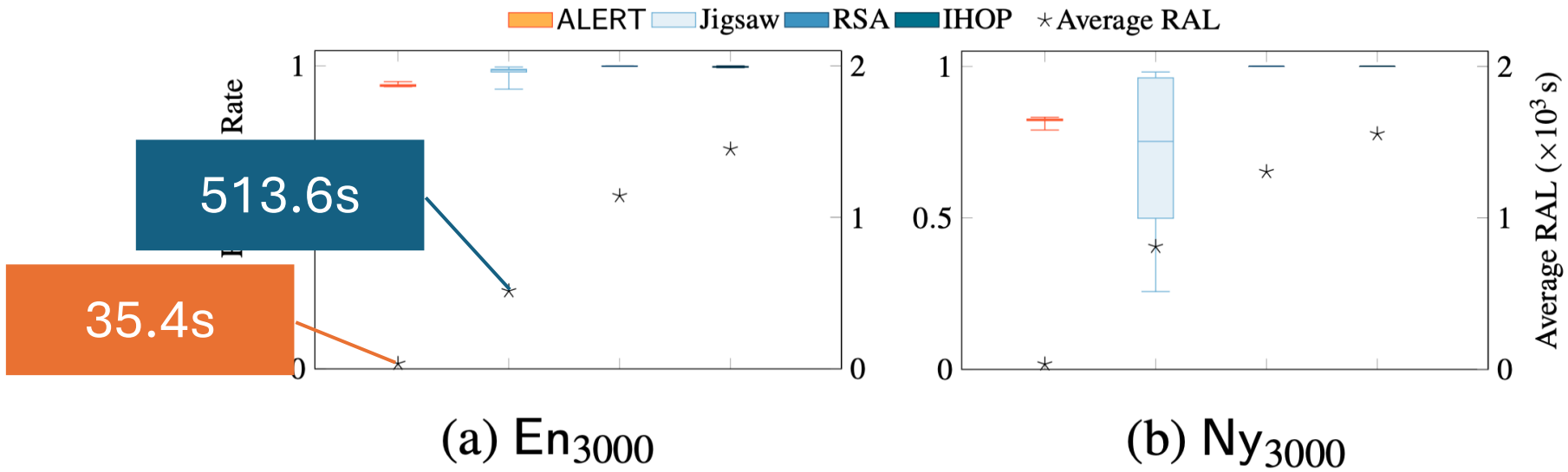
Optimized calculation



Time complexity: $O(kn^2)$ → $O(kn)$

Evaluation – Comparison with Prior Arts

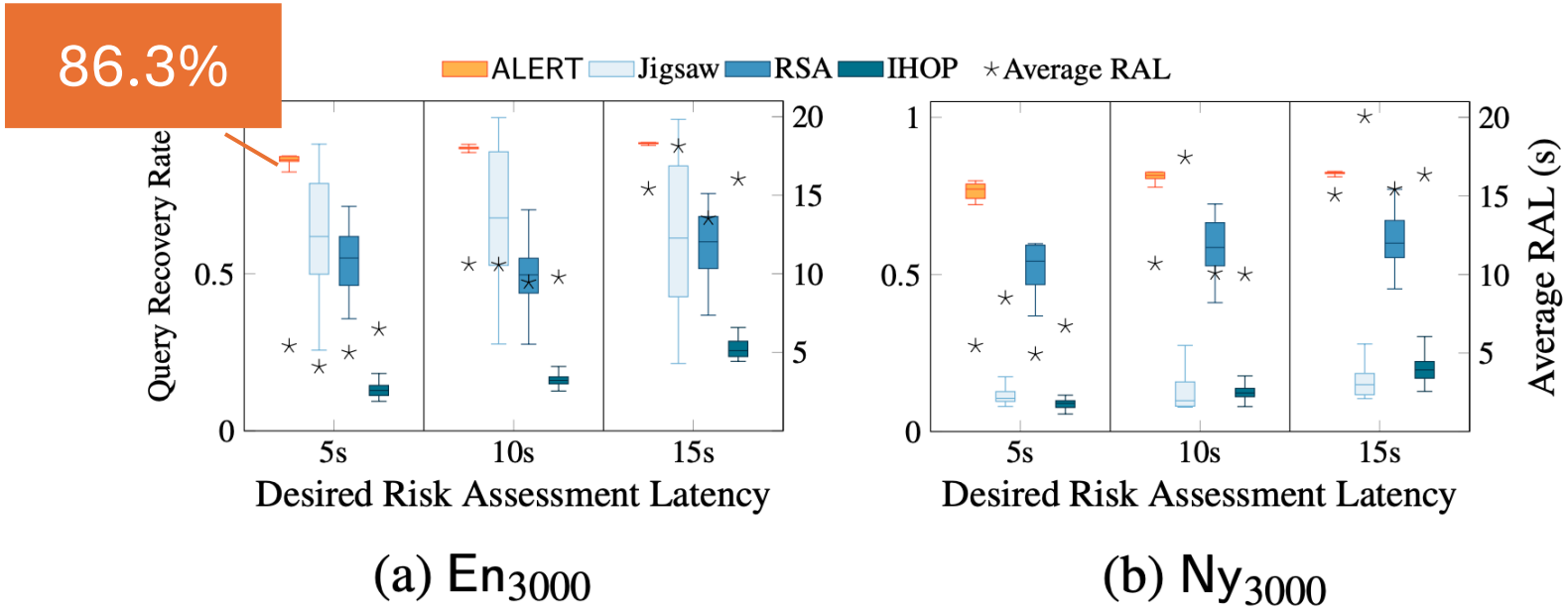
- Effectiveness **without time constraints** (sampled data, α=50%)
  - Illustrate 14.5X speed-up with only 5.2% recovery loss on En_{3000}. Even Higher accuracy on Ny_{3000}



[1] Nie Hao et al. Query Recovery from Easy to Hard: Jigsaw Attack against SSE. USENIX Security 2024.
[2] Simon Oya et al. IHOP: Improved Statistical Query Recovery against Searchable Symmetric Encryption through Quadratic Optimization. USENIX Security 2022.
[3] Marc Damie, Florian Hahn, and Andreas Peter. A highly accurate query-recovery attack against searchable encryption using non-indexed documents. USENIX Security 2021.

Evaluation – Comparison with Prior Arts

- Effectiveness **under low latency scenario** (sampled data, α=50%)
  - Achieve recovery rates of 86.3% with 5.4s runtime, also demonstrate superior recovery stability (1.5% interquartile range)

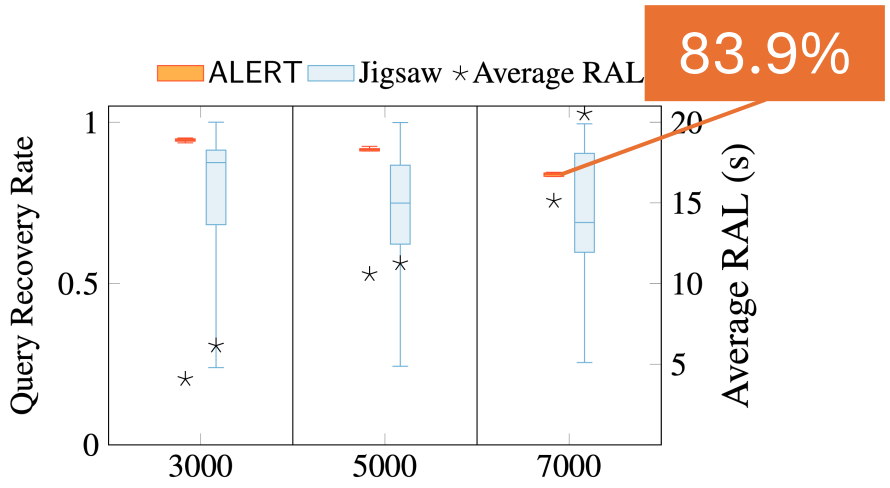


(a) En_{3000} (b) Ny_{3000}

[1] Nie Hao et al. Query Recovery from Easy to Hard: Jigsaw Attack against SSE. USENIX Security 2024.
[2] Simon Oya et al. IHOP: Improved Statistical Query Recovery against Searchable Symmetric Encryption through Quadratic Optimization. USENIX Security 2022.
[3] Marc Damie, Florian Hahn, and Andreas Peter. A highly accurate query-recovery attack against searchable encryption using non-indexed documents. USENIX Security 2021.

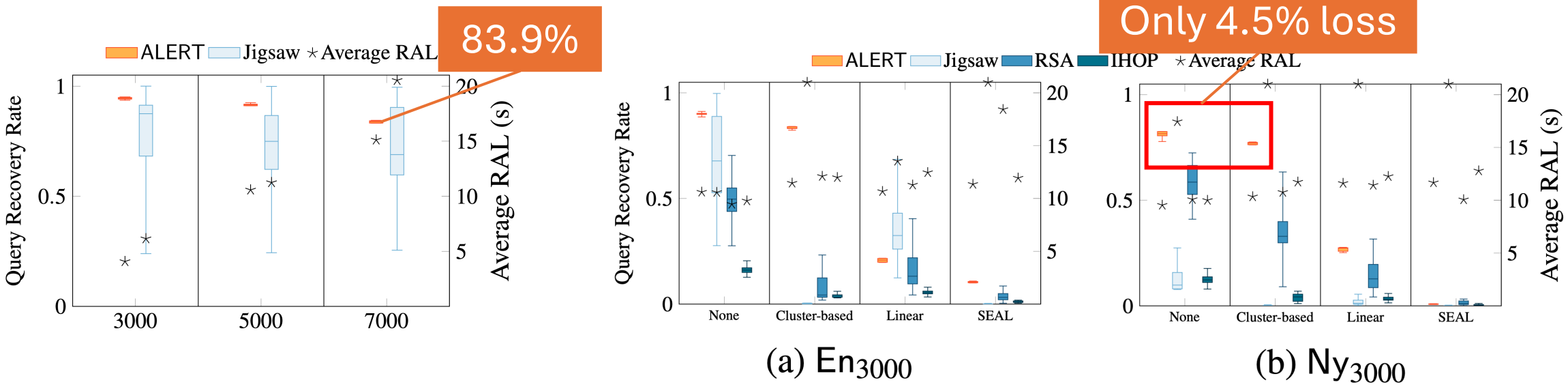
Evaluation – Comparison with Prior Arts

- Effectiveness under large keyword universe sizes
  - Under large keywords universe (Over 7000), only ALERT remains stable performance



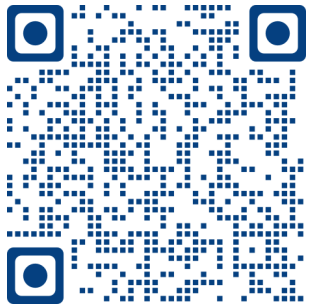
Evaluation – Comparison with Prior Arts

- Effectiveness under large keyword universe sizes
  - Under large keywords universe (Over 7000), only ALERT remains stable performance
- Effectiveness under access-pattern countermeasures
  - Only 4.5% recovery rate loss under cluster-based padding, also maintain performance under linear padding and SEAL



USENIX Security Symposium 2025

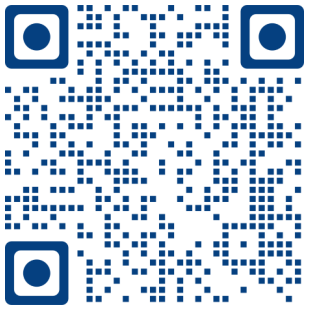
ALERT: Machine Learning-Enhanced Risk Estimation for Databases Supporting Encrypted Queries



Source Code

Opensource Code: https://doi.org/10.5281/zenodo.14726862

Any comments are welcome: longxiang.wang@my.cityu.edu.hk



Personal Webpage