KARL: Fast Kernel Aggregation Queries

Tsz Nam Chan[1,2], Man Lung Yiu[2], and Leong Hou U[3]

[1]The University of Hong Kong, [2]Hong Kong Polytechnic University, [3]University of Macau

[1]tnchan2@hku.hk, [2]{cstnchan,csmlyiu}@comp.polyu.edu.hk and [3]ryanlhu@umac.mo

What is Kernel Aggregation Queries?

Kernel Aggregation Function

$$\mathcal{F}_P(\mathbf{q}) = \sum_{\mathbf{p_i} \in P} w_i \exp(-\gamma \cdot dist(\mathbf{q}, \mathbf{p_i})^2)$$

Type of weighting	Used in model
Type I: identical, positive w_i (most specific)	Kernel density
Type II: positive w_i (subsuming Type I)	1-class SVM
Type III: no restriction on w_i (subsuming Types I, II)	2-class SVM

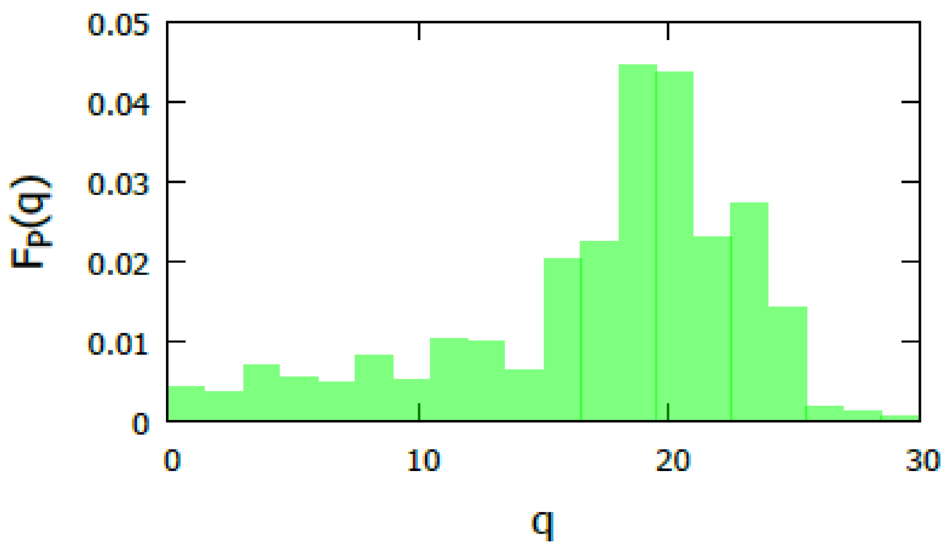
Approximate Kernel Aggregation Query (ε-KAQ)

- Input: query vector $\mathbf{q}$, dataset P, relative error ε
- Output: value $\hat{F}$

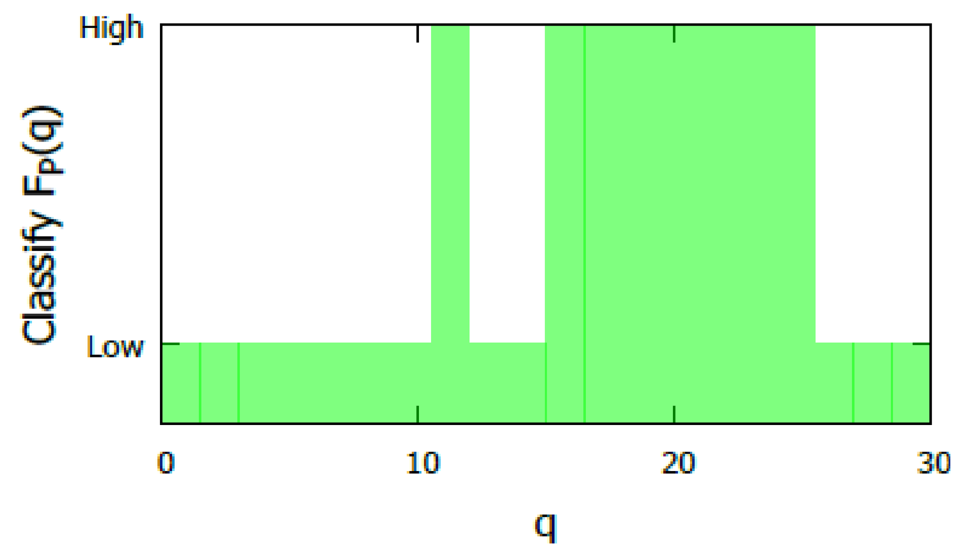
where: $(1-\varepsilon)\mathcal{F}_P(\mathbf{q}) \le \hat{F} \le (1+\varepsilon)\mathcal{F}_P(\mathbf{q})$

Threshold Kernel Aggregation Query (τ-KAQ)

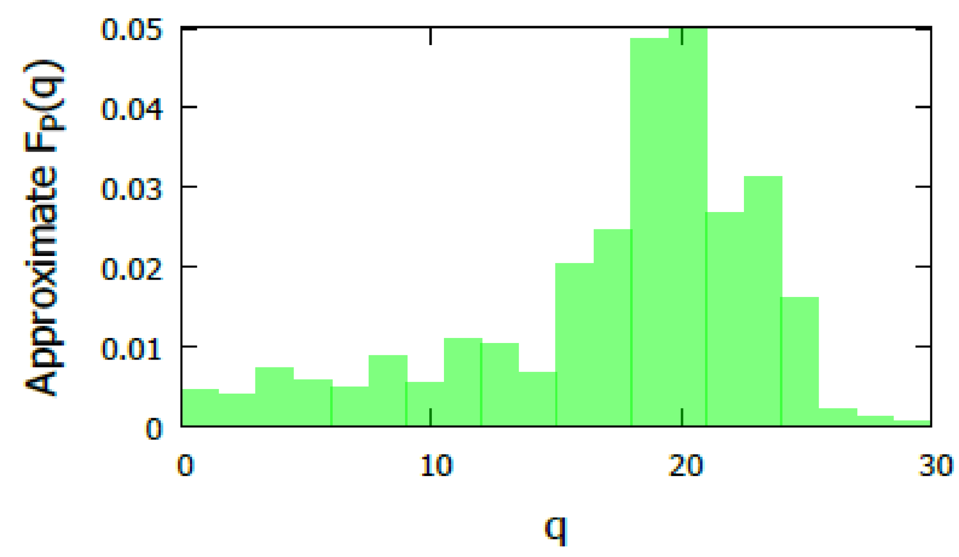
- Input: query vector $\mathbf{q}$, dataset P, threshold τ
- Output: 1 (if $\mathcal{F}_P(\mathbf{q}) \ge \tau$) or -1 (if $\mathcal{F}_P(\mathbf{q}) < \tau$)



KAQ



τ-KAQ, τ=0.01



ε-KAQ, ε=0.2

Applications of Kernel Aggregation Queries

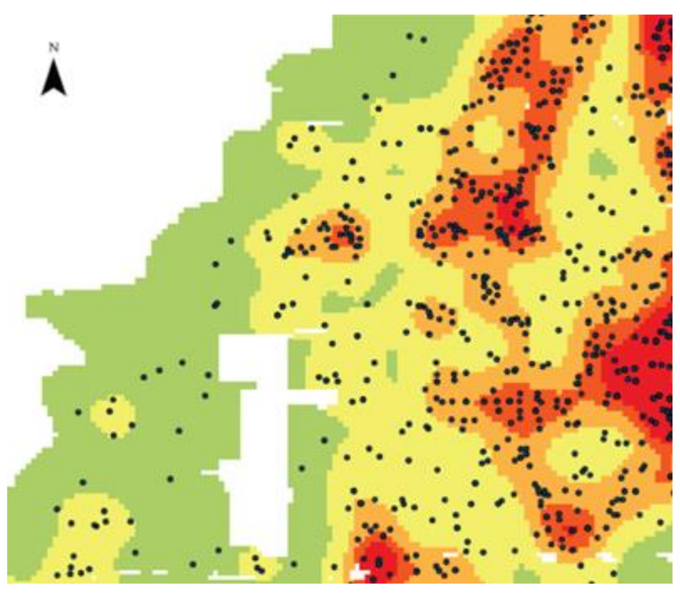
Kernel Density Estimation

Black dots (Crimes)

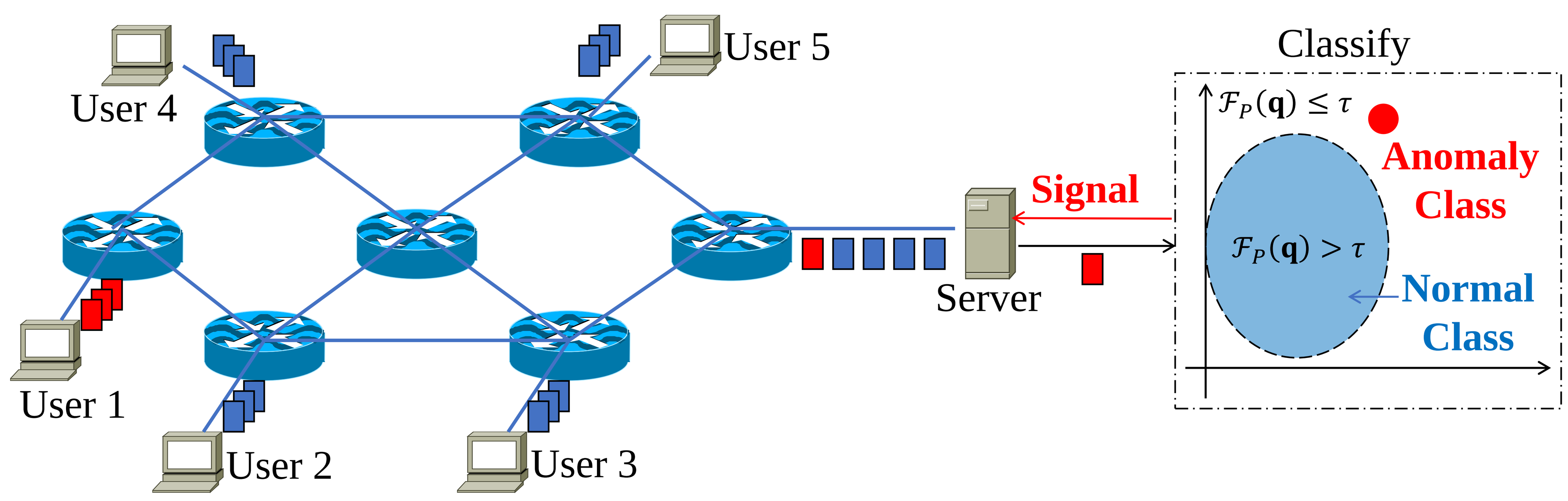
- Aggravated assault
- Robbery
- Commercial burglary
- Motor vehicle theft

Goal:

- Crime rates prediction



Kernel Support Vector Machine Classification



How to speed up?

$$\mathcal{F}_P(\mathbf{q}) = \underbrace{\sum_{\mathbf{p_i} \in P} w \exp(-\gamma \cdot dist(\mathbf{q}, \mathbf{p_i})^2)}_{O(|P| \times d) \text{ time}}$$

$$LB(\mathbf{q}) \le \mathcal{F}_P(\mathbf{q}) \le UB(\mathbf{q})$$

Much smaller than O(|P| × d) time

τ-KAQ (Stop Condition)

UB($\mathbf{q}$)

$\mathbf{q}$ is in class +1

LB($\mathbf{q}$)

+1

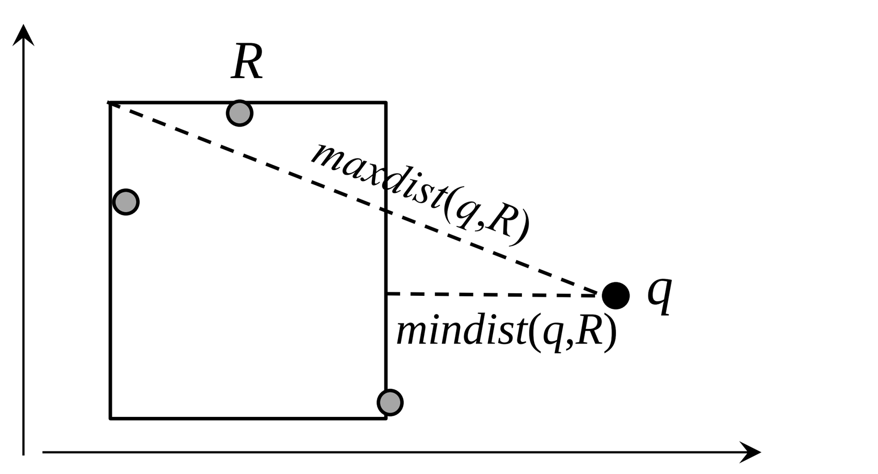
τ

UB($\mathbf{q}$) -1

$\mathbf{q}$ is in class -1

LB($\mathbf{q}$)

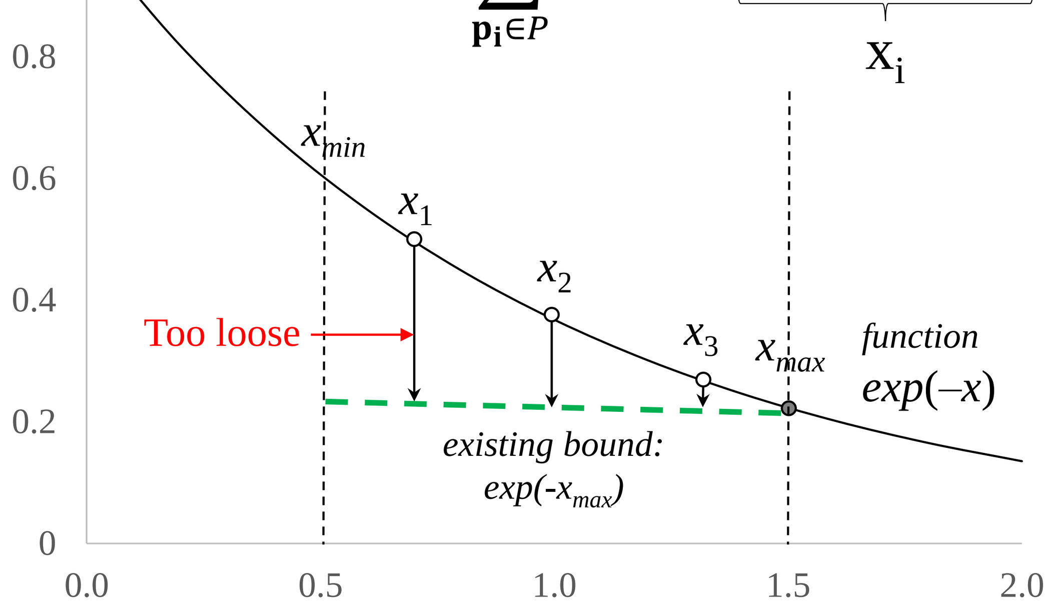
State-of-the-art Method and its Weakness



$$LB_R(\mathbf{q}) = w \cdot R.\text{count} \cdot \exp(-\gamma \cdot maxdist(\mathbf{q}, R)^2)$$
$$UB_R(\mathbf{q}) = w \cdot R.\text{count} \cdot \exp(-\gamma \cdot mindist(\mathbf{q}, R)^2)$$

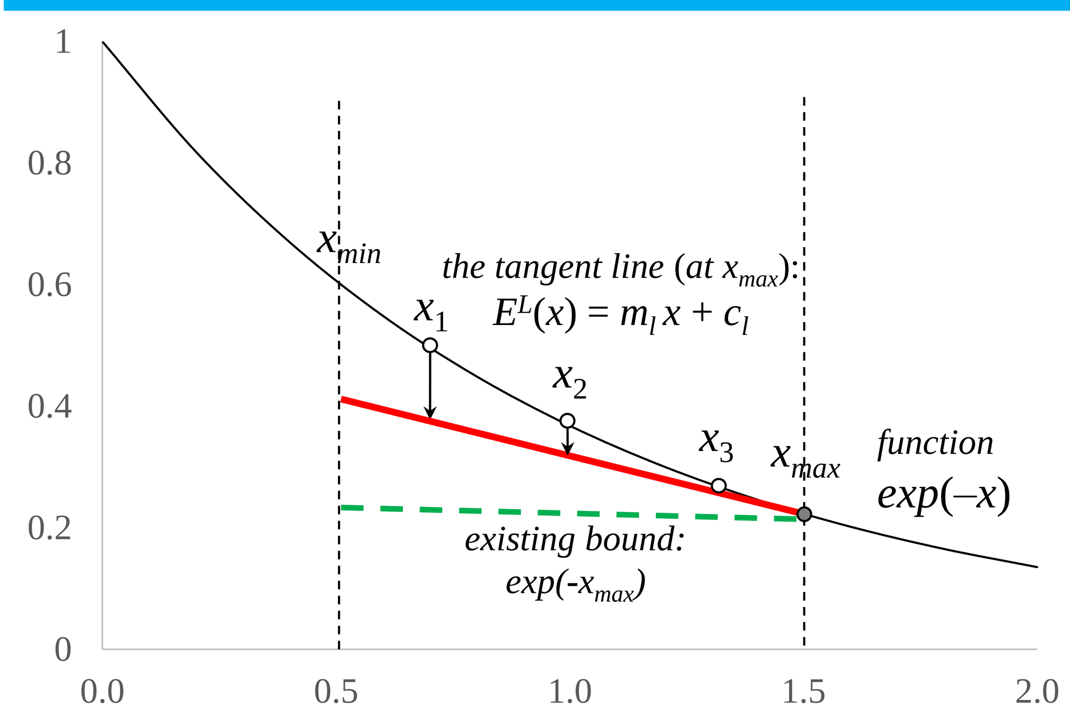
O(d) time

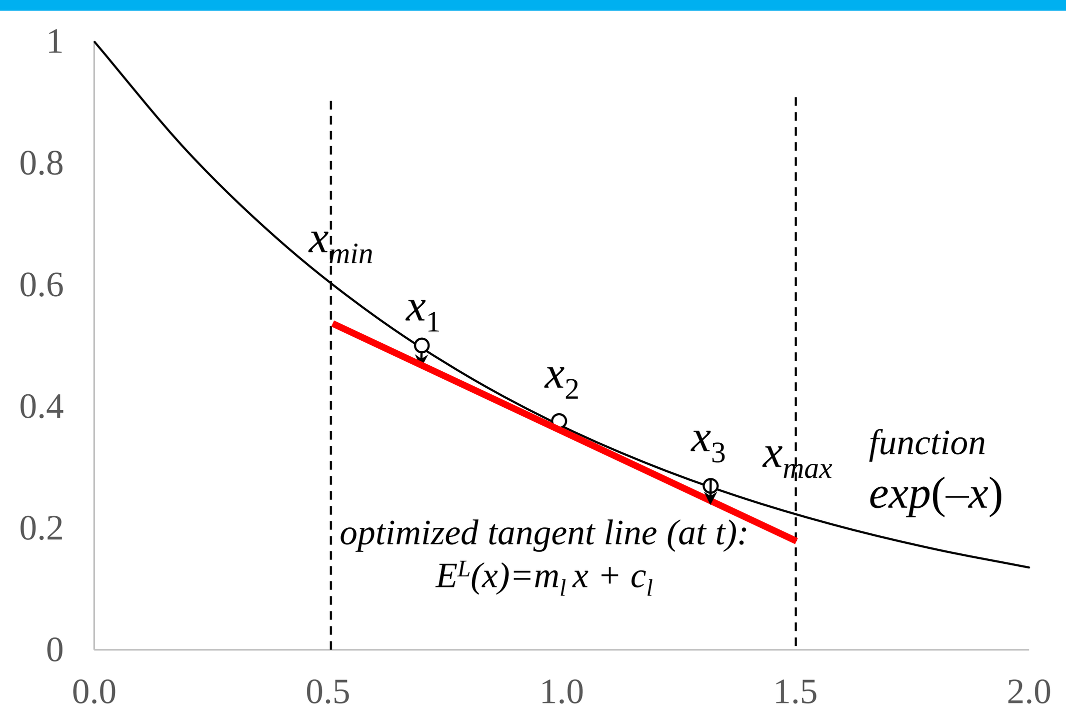
$$\mathcal{F}_P(\mathbf{q}) = \sum_{\mathbf{p_i} \in P} w \exp(-\gamma \cdot \underbrace{dist(\mathbf{q}, \mathbf{p_i})^2}_{x_i})$$



$$LB_R(\mathbf{q}) = w \cdot R.\text{count} \cdot \exp(-\gamma \cdot \underbrace{maxdist(\mathbf{q}, R)^2}_{x_{max}})$$

Our techniques

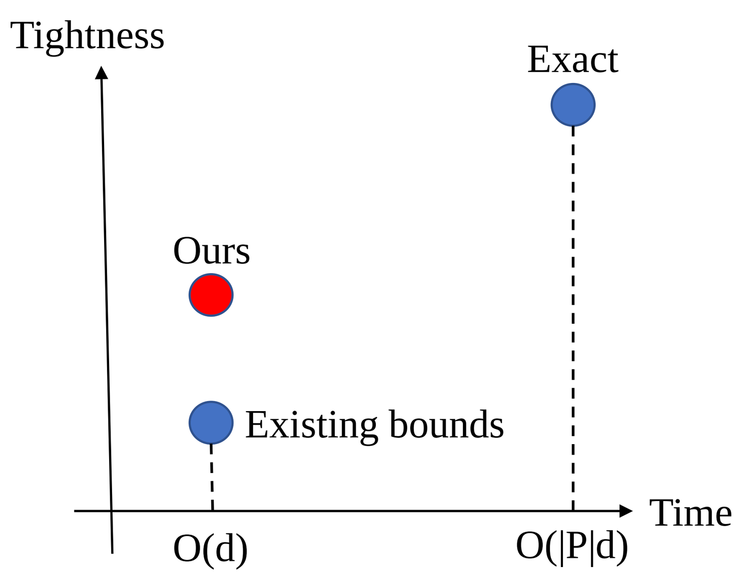




$$\mathcal{FL}_P(\mathbf{q}, Lin_{m,c}) = \sum_{\mathbf{p_i} \in P} w\,(m(\gamma dist(\mathbf{q}, \mathbf{p_i})^2) + c)$$

$$= wm\gamma(|P| \underbrace{\|\mathbf{q}\|^2}_{O(d)} - 2 \underbrace{\mathbf{q} \cdot \mathbf{a_P}}_{O(d)} + b_P) + wc|P|$$

where $\mathbf{a_P} = \sum_{\mathbf{p_i} \in P} \mathbf{p_i}$ and $b_P = \sum_{\mathbf{p_i} \in P} \|\mathbf{p_i}\|^2$



Experimental Results

Type	Datasets	SCAN	LIBSVM	Scikit	SOTA	KARL
I-ϵ	miniboone	36.1	n/a	36	16.5	**301**
	home	15.2	n/a	11.9	36.2	**187**
	susy	2.02	n/a	1.17	0.77	**13.2**
I-τ	miniboone	36.1	34	n/a	102	**510**
	home	15.2	14.1	n/a	93.2	**258**
	susy	2.02	1.86	n/a	3.58	**83.4**
II-τ	nsl-kdd	283	481	n/a	748	**20668**
	kdd99	260	520	n/a	1269	**11324**
	covtype	158	462	n/a	448	**6022**
III-τ	ijcnn1	903	1170	n/a	1119	**826928**
	a9a	162	610	n/a	546	**6885**
	covtype-b	13	38.4	n/a	33.9	**274**

SCAN ○ SOTA ♦ KARL △