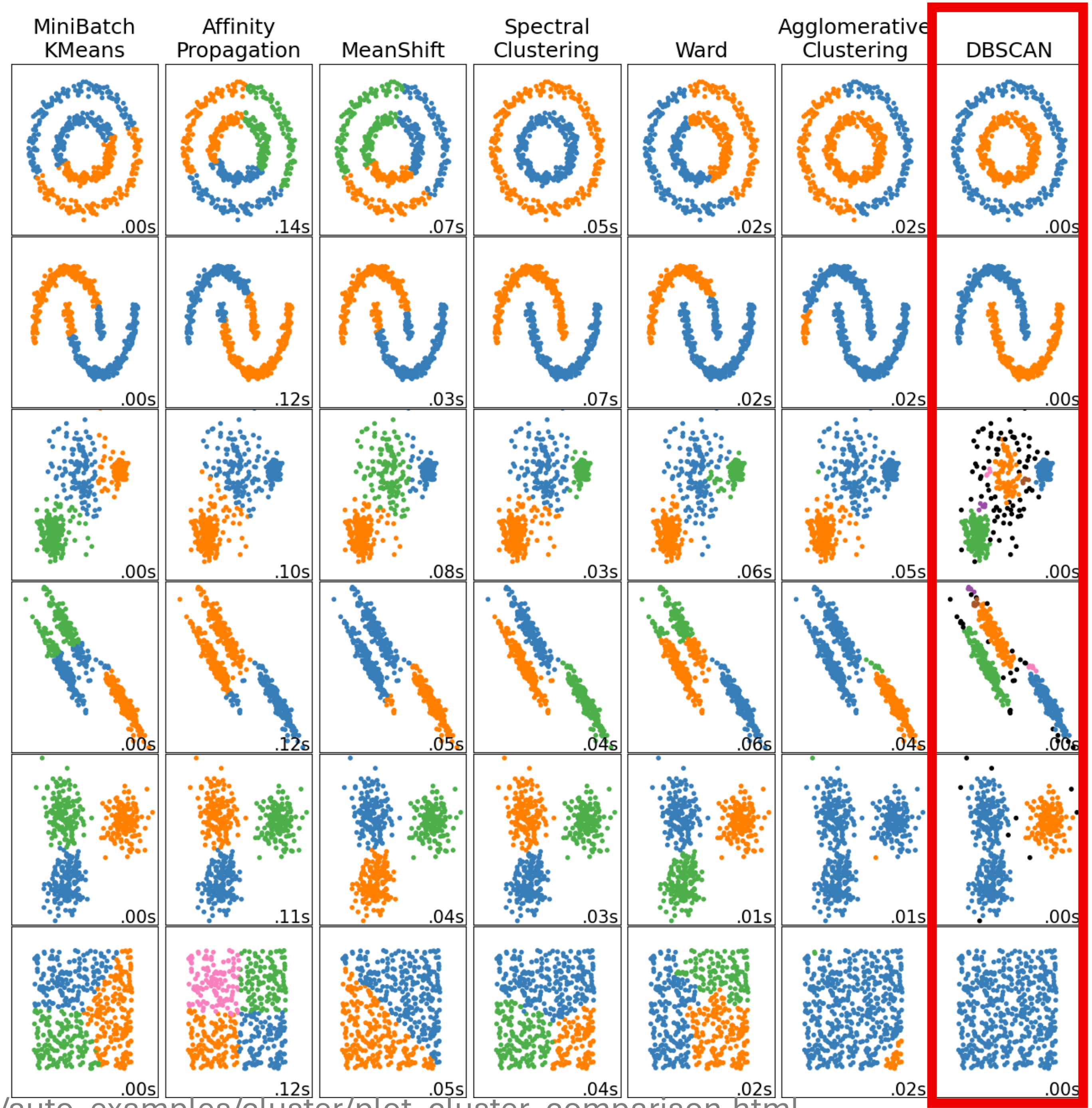
Unimodal Strategies in Density-Based Clustering

Oron Nir[1,2], Jay Tenenbaum[2], and Ariel Shamir[1]

[1] CANVAS Lab, Reichman University

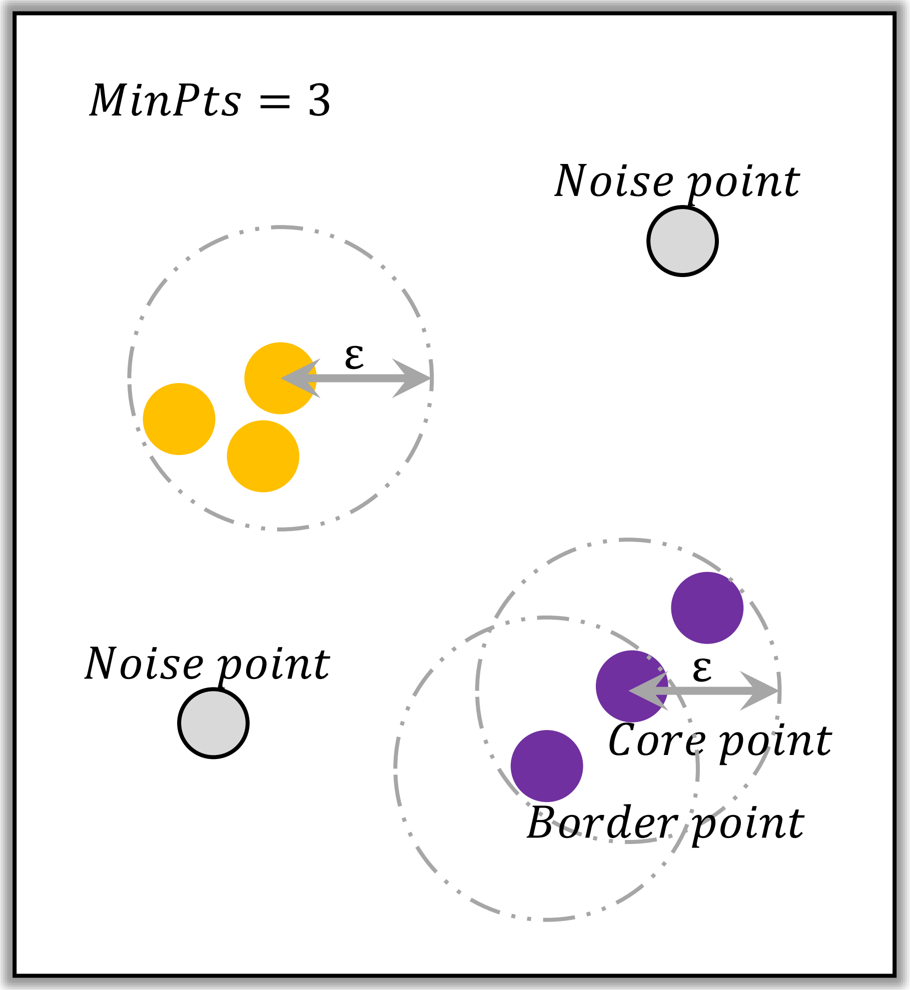
[2] Microsoft



https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

DBSCAN

- *Dataset:* $X \in \mathbb{R}^{N \times D}$
- *DBSCAN Parameters:*
  - d *– a distance metric*
  - $\varepsilon \in \mathbb{R}_{>0}$ *- the radius*
  - $MinPts \in \mathbb{N}_{\geq 2}$
- *Core and Border Points:*
  - *A* ***core point*** p *has at least* $MinPts$ *neighbors*
  - *A* ***border point*** b *is:*
    - $d(b, p) \leq \varepsilon$ *for some core point* p
    - *not a core point*
  - *A* ***noise point*** *is neither a core nor a border point*



Ester et al. *A density-based algorithm for discovering clusters in large spatial databases with noise*. In KDD, 1996.

Prior works on ε tuning

All prior works use synthetic, small, and low dimensional datasets…

1996	2007	2013	2021	2022	2023
DBSCAN	**VDBSCAN**	**AutoEps**	**AEDBSCAN**	**AMD-DBSCAN**	**SS-DBSCAN**
Ester et al.	Liu et al.	Gaonkar and Sawant	Mistry et al.	Wang et al.	Monko & Kimura

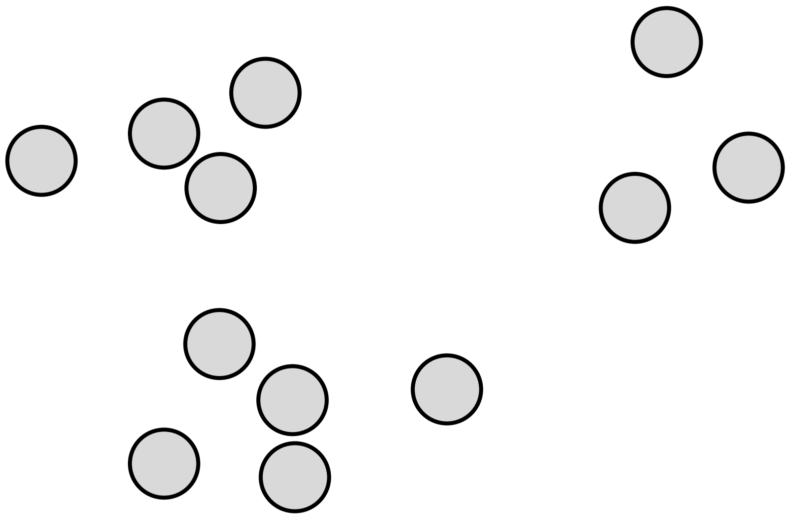
Our contribution

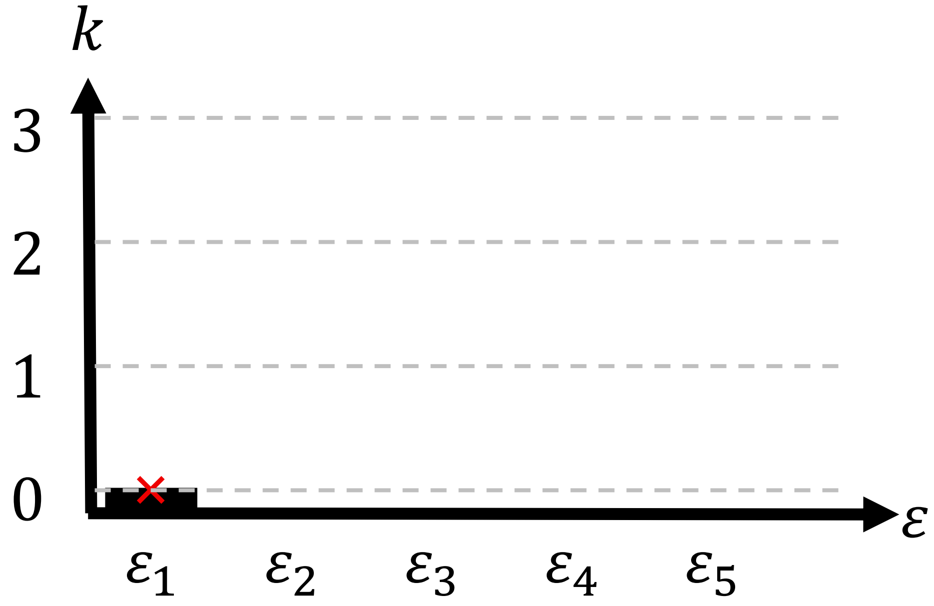
- We identify a key property of the relation between the number of clusters (k) and the radius (ε)
  - ➢it is unimodal
- We find that the mode (ε^*) yields a good solution
  - ➢and support it theoretically and empirically
- We devise an efficient algorithm to tune ε
  - ➢using the Ternary Search algorithm

Bajwa et al. *Ternary search algorithm: Improvement of binary search*. In Indiacom, 2015. IEEE

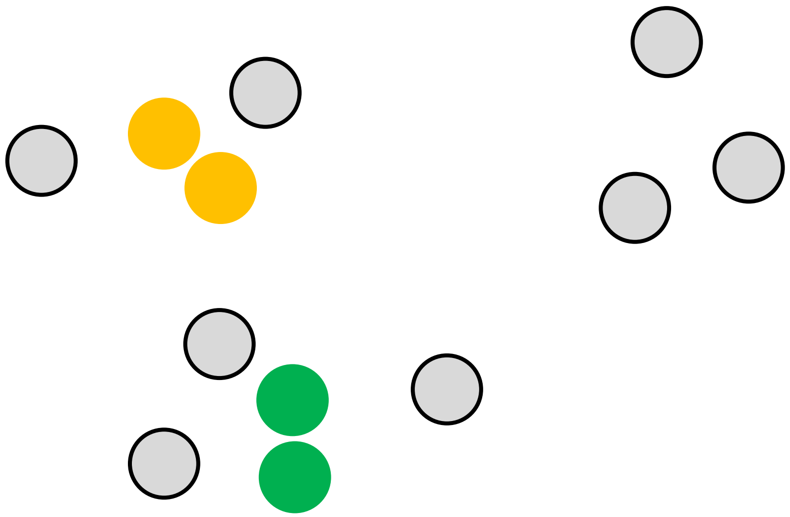
The Unimodality Property

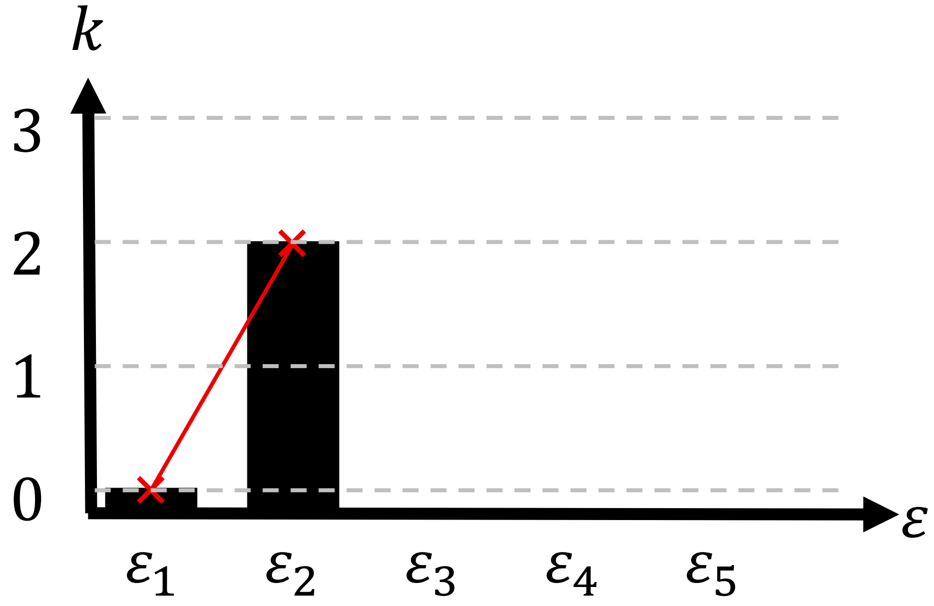
Low ε: k=0



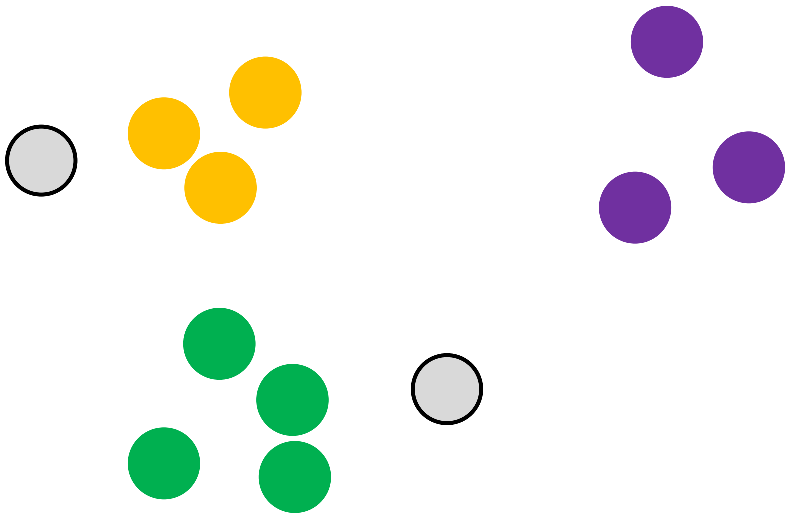


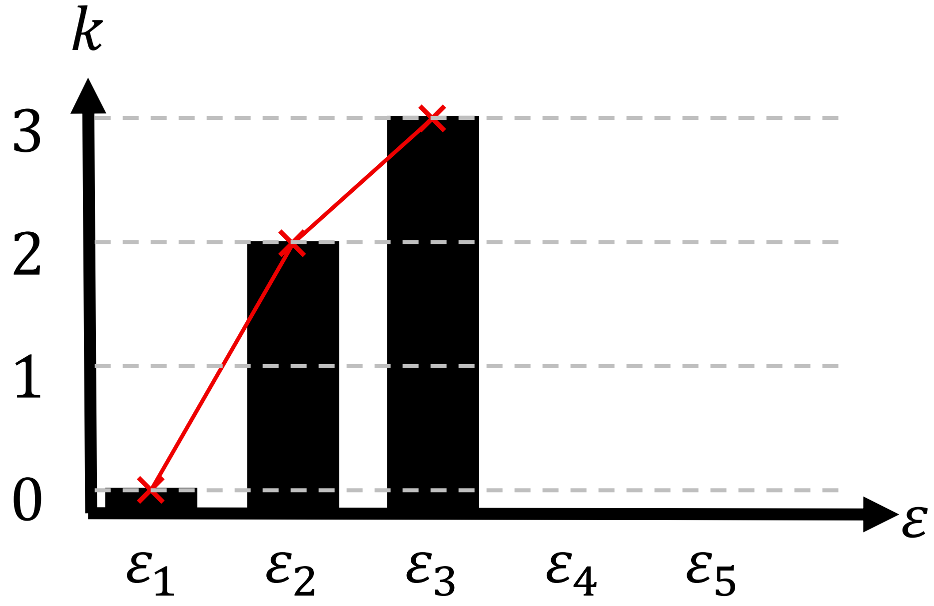
Increasing ε: k=2



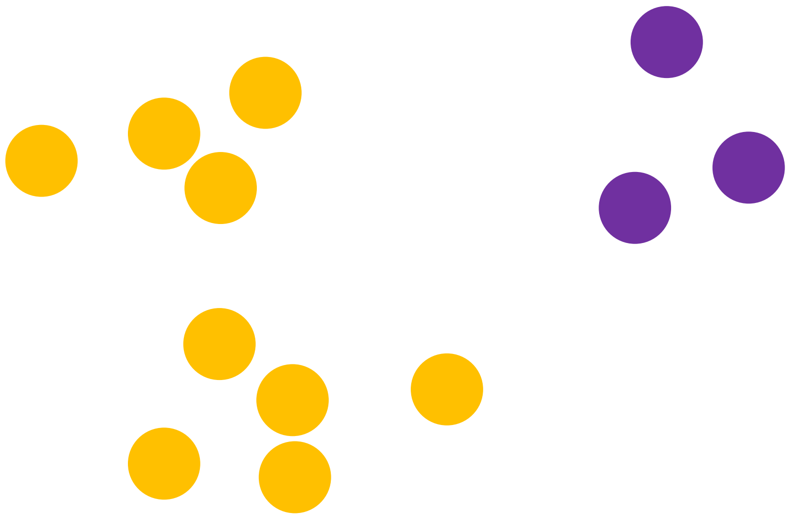


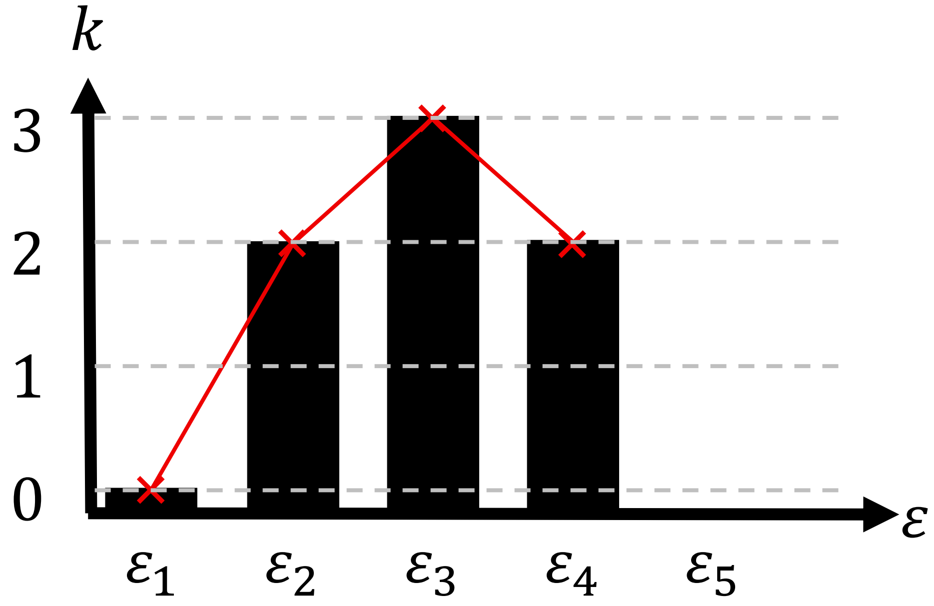
Good ε: k=3



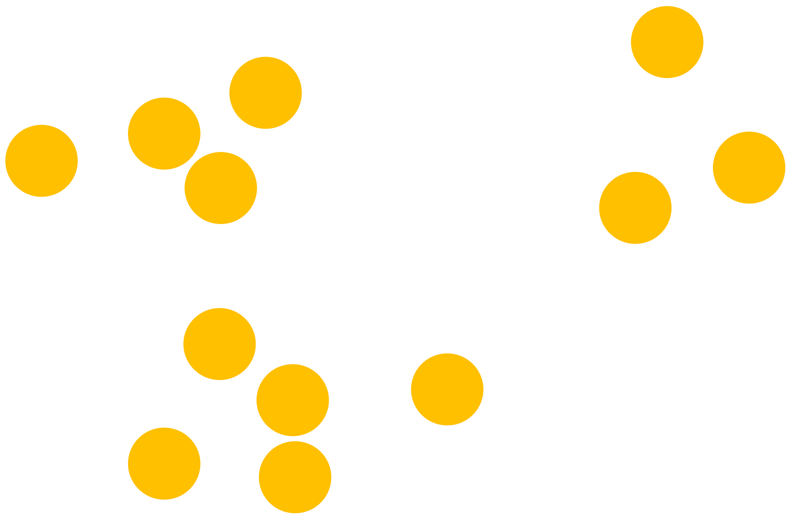


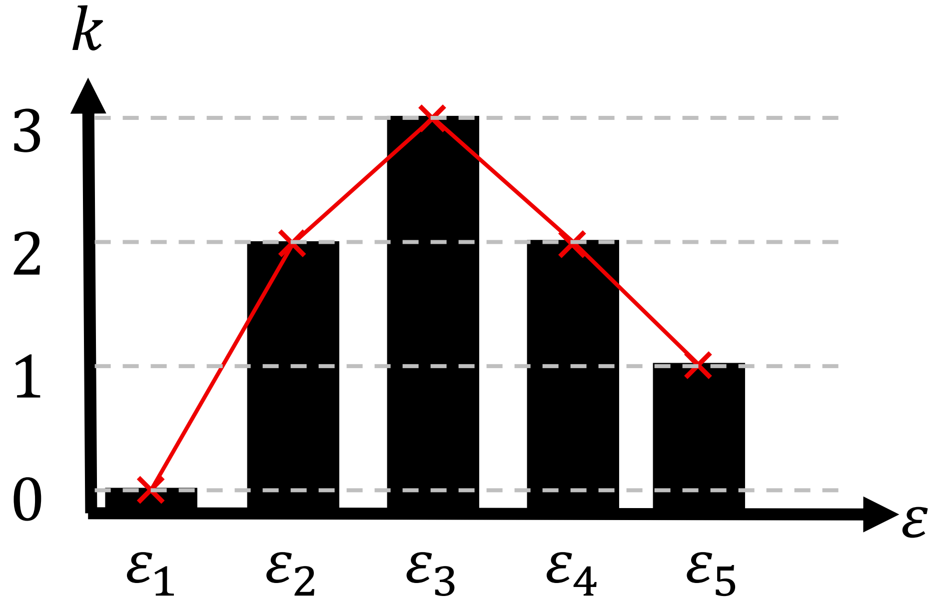
Oversized ε: k=2



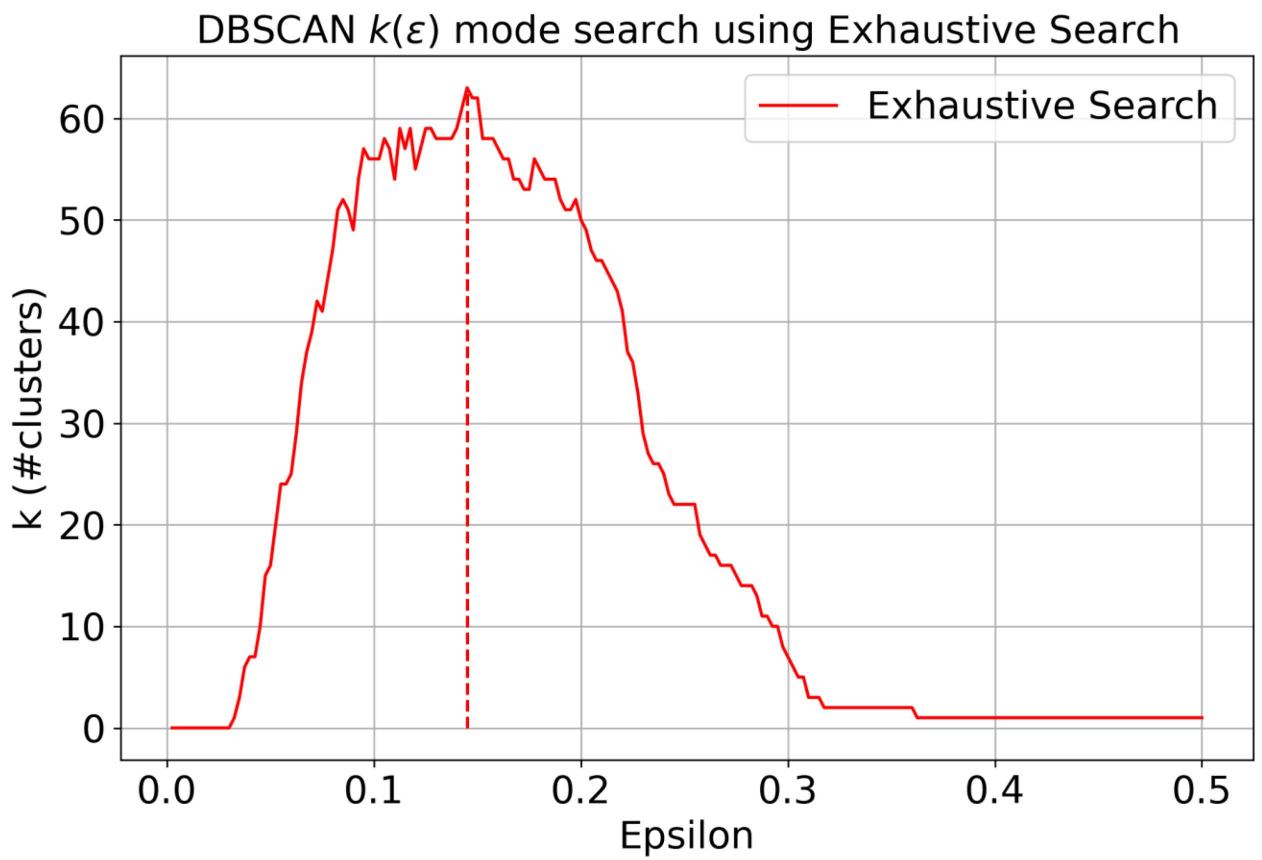


Way too large ε: k=1





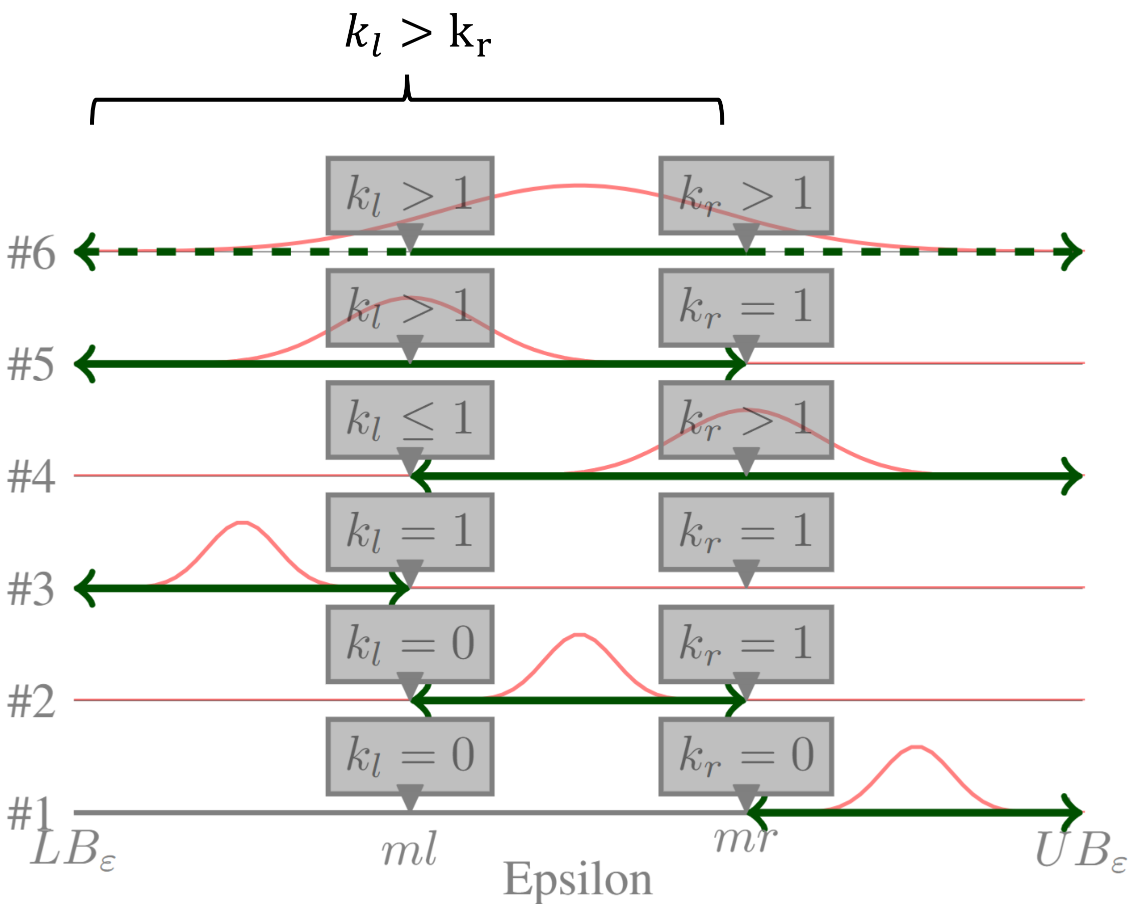
The Unimodality Property



Method

Method: Ternary Search

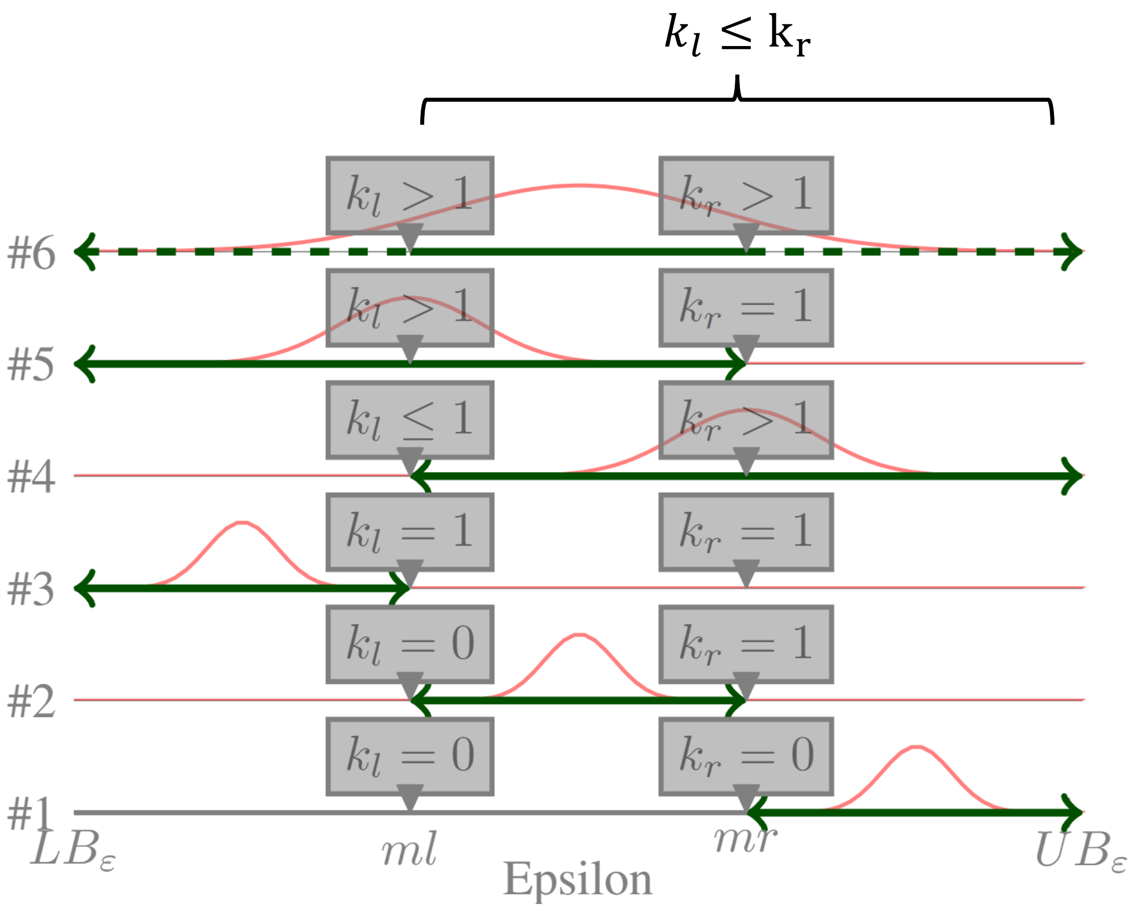
Our task is to efficiently find the mode of k(ε),



Bajwa et al. *Ternary search algorithm: Improvement of binary search*. In Indiacom, 2015. IEEE

Method: Ternary Search

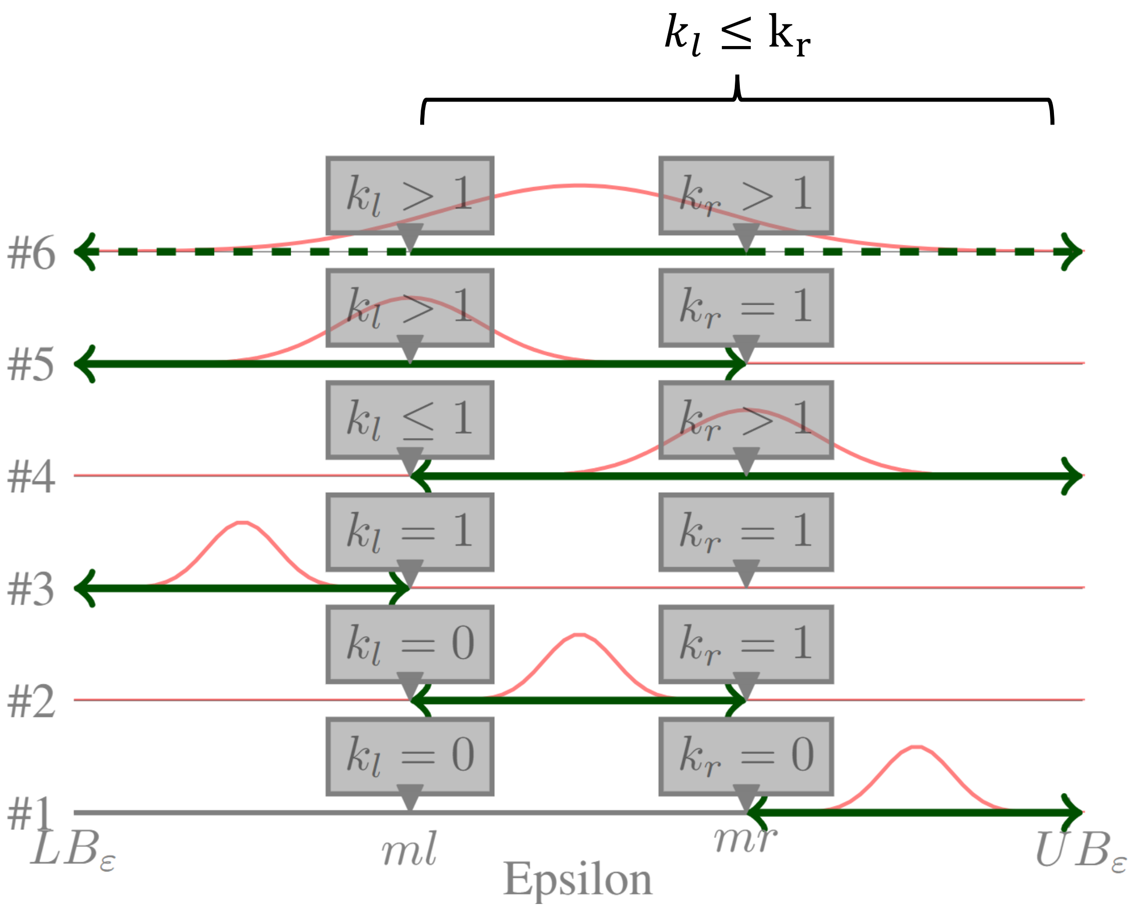
Our task is to efficiently find the mode of k(ε),

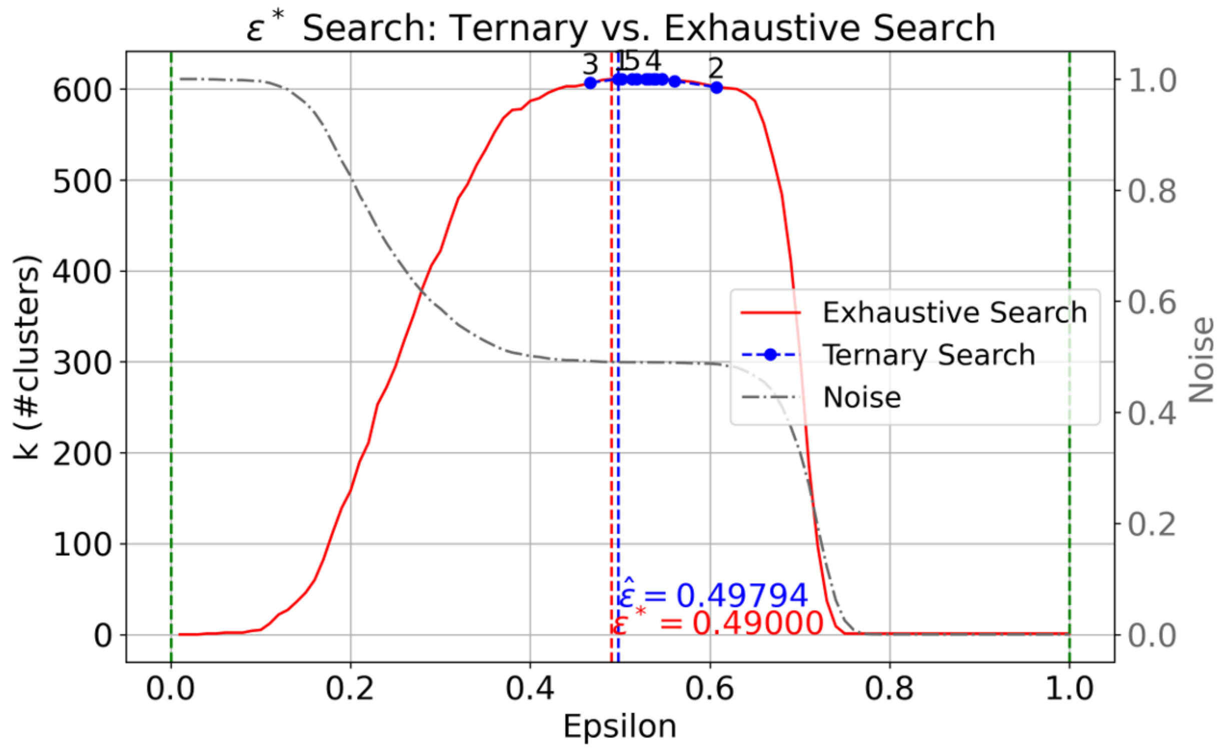


Bajwa et al. *Ternary search algorithm: Improvement of binary search*. In Indiacom, 2015. IEEE

Method: Ternary Search

Our task is to efficiently find the mode of k(ε),



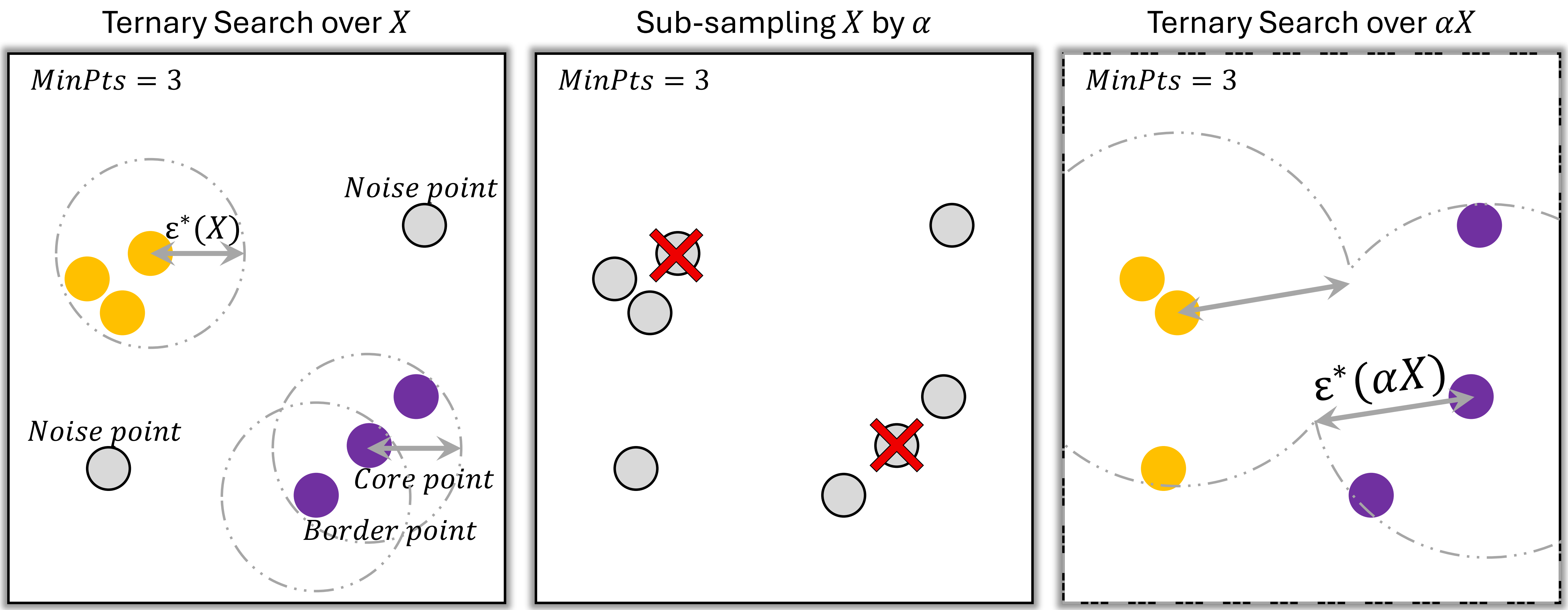


Bajwa et al. *Ternary search algorithm: Improvement of binary search*. In Indiacom, 2015. IEEE

Algorithm 1 $TS(X, LB, UB, MinPts, itr)$

for i=0 to itr **do**
  $m_l \leftarrow \frac{2LB+UB}{3}$
  $m_r \leftarrow \frac{LB+2UB}{3}$
  $\mathcal{C}_l \leftarrow \textbf{DBSCAN}(X, m_l, MinPts)$
  $\mathcal{C}_r \leftarrow \textbf{DBSCAN}(X, m_r, MinPts)$
  $k_l \leftarrow |\{c \in \mathcal{C}_l\}|$
  $k_r \leftarrow |\{c \in \mathcal{C}_r\}|$
  $\langle LB, UB \rangle \leftarrow TSConditions(LB, UB, m_l, m_r, k_l, k_r)$
end for
return $\frac{m_l+m_r}{2}$

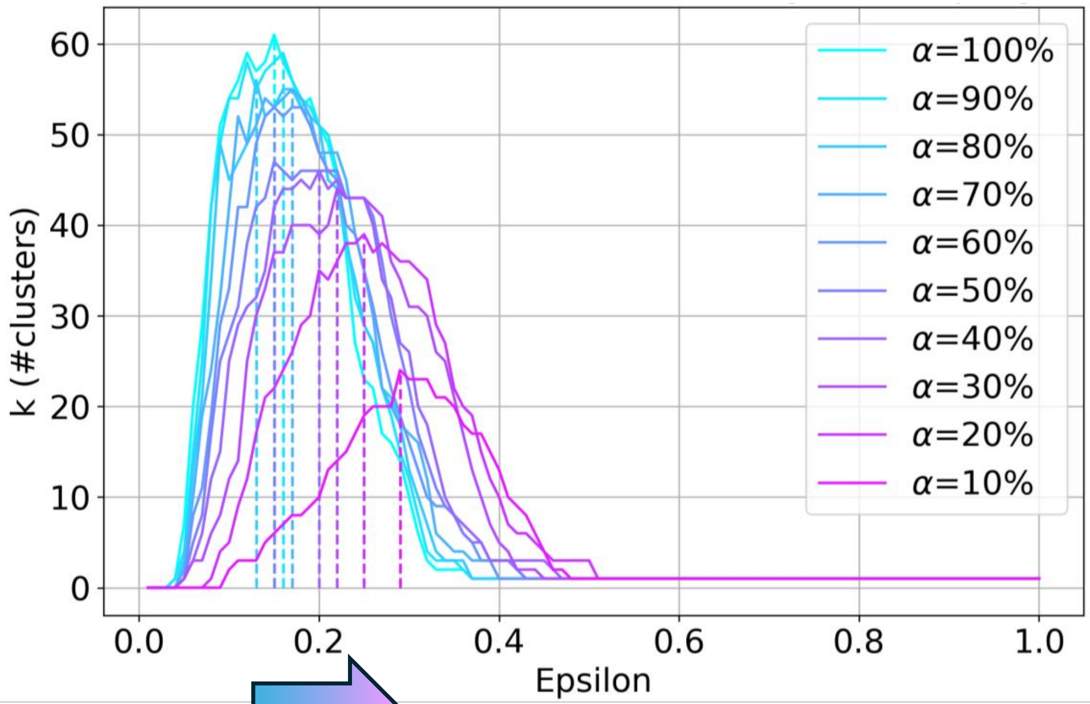
Upper Bound Motivation



Sub-sampling yield larger radii

Upper Bound

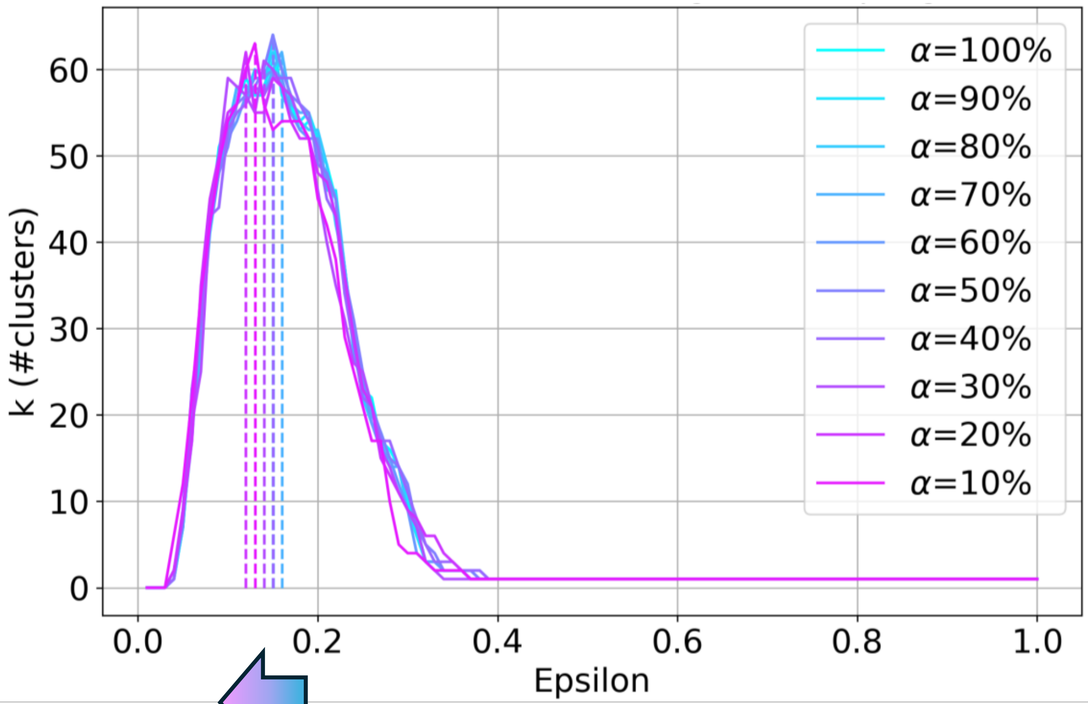
By sub-sampling αN points



Lower Bound

By sub-sampling αD dimensions

In lower dimensions points are closer and ε^* is underestimated



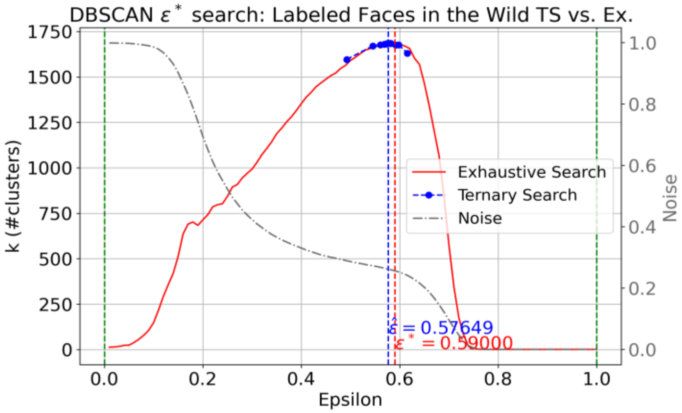
Results

Evaluation over Classification Datasets

Face Recognition

Labeled Faces in the Wild (LFW)

N=13,233; #Classes = 1,680

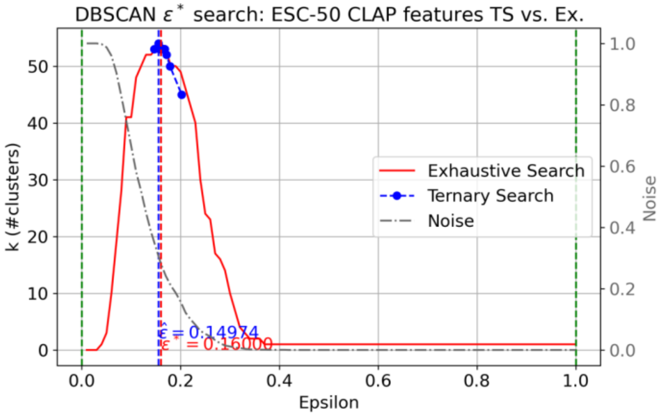


Audio Effect Classification

ESC-50

N=1,024; #Classes = 50

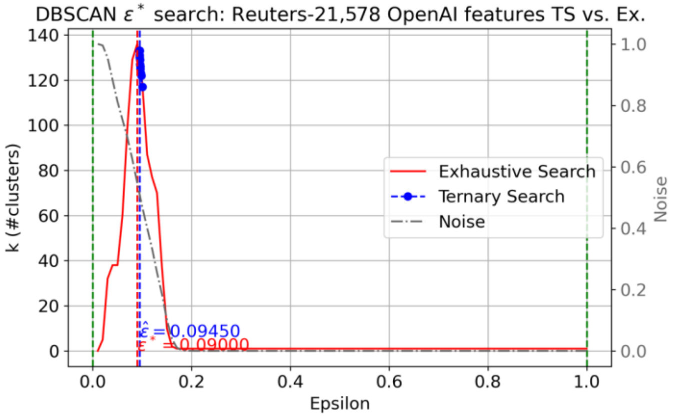
No.	Major Categories				
	Animals	Natural Soundscapes and Water Sounds	Human, Non-Speech Sounds	Interior/Domestic Sounds	Exterior/Urban Noises
1	Dog	Rain	Crying baby	Door knock	Helicopter
2	Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
3	Pig	Crackling fire	Clapping	Keyboard typing	Siren
4	Cow	Crickets	Breathing	Door, wood creaks	Car horn
5	Frog	Chirping birds	Coughing	Can opening	Engine
6	Cat	Water drops	Footsteps	Washing machine	Train
7	Hen	Wind	Laughing	Vacuum cleaner	Church bells
8	Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
9	Sheep	Toilet flush	Snoring	Clock tick	Fireworks
10	Crow	Thunderstorm	Drinking, sipping	Glass breaking	Hand saw



Document Classification

Reuters document classification

N=21,578; #Classes = 135

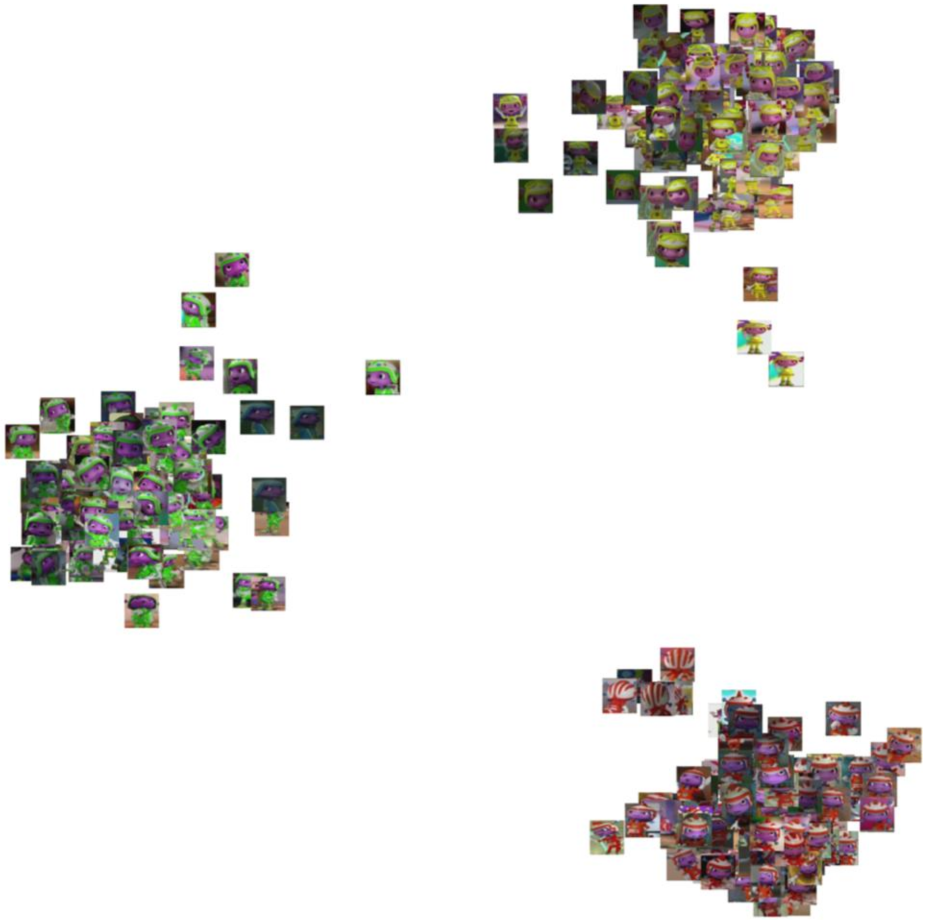


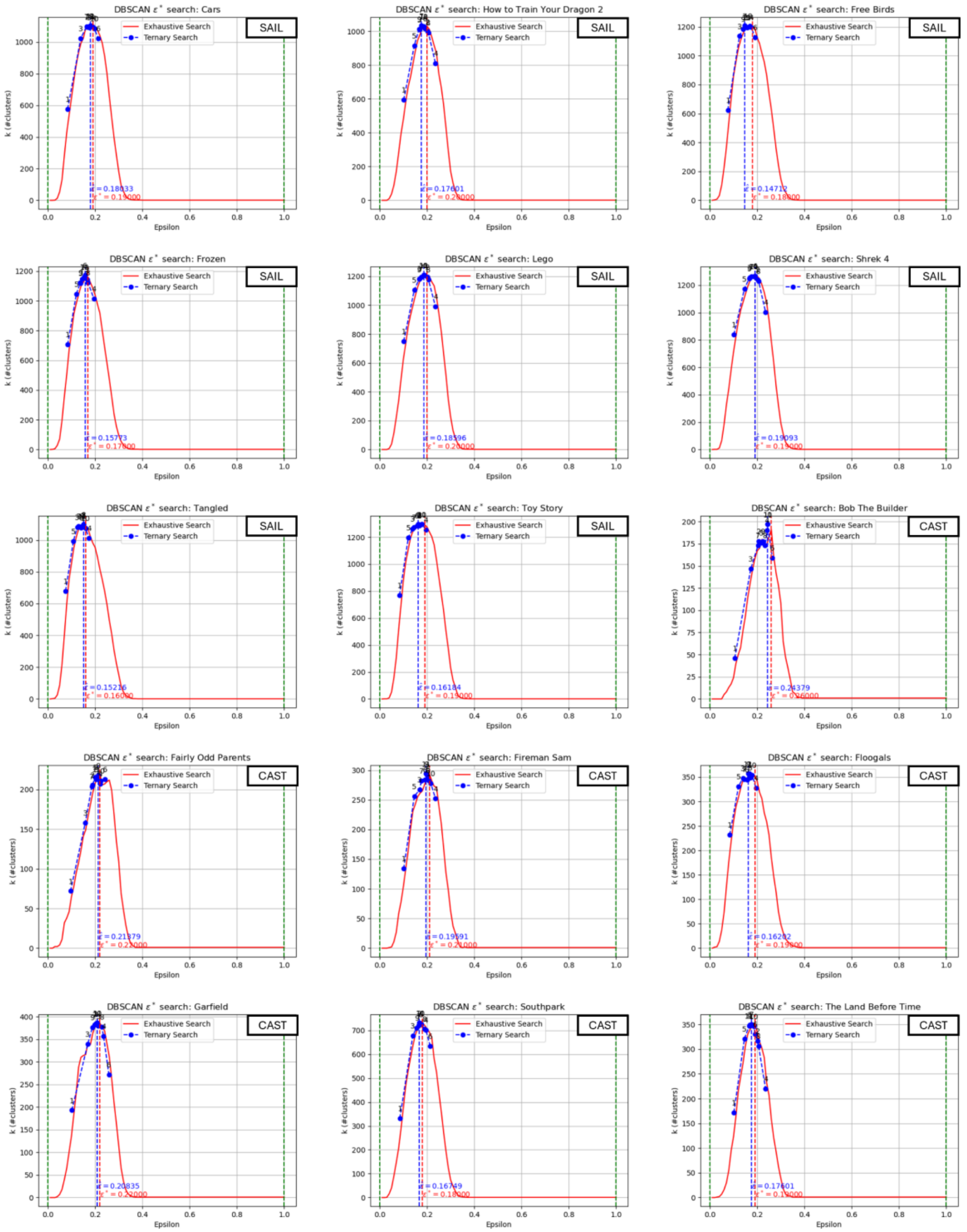
Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*. In Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition.

Karol J. Piczak. *ESC: Dataset for Environmental Sound Classification*. In ACM MM, Brisbane, Australia, (2015).

David Lewis. *Reuters-21578 text categorization test collection*. Distribution 1.0, AT&T Labs-Research, 1, (1997).

Results: Unimodality in Animation





Nir, O., Rapoport, G., & Shamir, A. *CAST*. In EG (2022).

Results: Unimodality across 24 domains

1. Varying #classes or labels: $[50, 1680]$
2. Large datasets of up to $\mathrm{N} = 60{,}000$ points
3. High dimensional with up to $D = 2{,}048$
4. 24 domains e.g., NLP, CV, Audio, Animation
5. DIP Test was found insignificant for all, i.e., for all datasets a unimodal distribution was observed(!)

Dataset	Labels	N	Embed.	D	Task	p_{val}^{DIP}	Dataset	Labels	N	Embed.	D	Task	p_{val}^{DIP}
LFW	1,680	13,233	DNet	256	Face	>99.9%	AMCDv5	N/A	13,406	CAST	2,048	Anim	14.9%
ImNet1k	1,000	50,000	CLIP	512	OD	>99.9%	AMCDv6	N/A	14,372	CAST	2,048	Anim	6.4%
ImNet1k	1,000	50,000	Hiera	1,000	OD	33.8%	AMCDv7	N/A	14,460	CAST	2,048	Anim	8.4%
CIFAR	100	60,000	CLIP	512	OD	8.9%	AMCDv8	N/A	14,748	CAST	2,048	Anim	14.8%
CIFAR	100	60,000	Hiera	1,000	OD	>99.9%	CASTv1	N/A	2,648	CAST	2,048	Anim	6.4%
Reuters	135	21,578	ADA2	1,536	Doc	99.8%	CASTv2	N/A	4,215	CAST	2,048	Anim	79.1%
ESC-50	50	1,024	CLAP	1,024	Audio	41.3%	CASTv3	N/A	4,633	CAST	2,048	Anim	14.8%
FACE	N/A	45,207	DNet	256	Face	>99.9%	CASTv4	N/A	4,163	CAST	2,048	Anim	99.4%
AMCDv1	N/A	15,395	CAST	2,048	Anim	29.3%	CASTv5	N/A	4,959	CAST	2,048	Anim	14.8%
AMCDv2	N/A	13,102	CAST	2,048	Anim	52.5%	CASTv6	N/A	5,639	CAST	2,048	Anim	99.4%
AMCDv3	N/A	14,676	CAST	2,048	Anim	79.0%	CASTv7	N/A	4,795	CAST	2,048	Anim	52.5%
AMCDv4	N/A	14,676	CAST	2,048	Anim	29.4%	Urban8k	N/A	8,732	CLAP	1,024	Audio	99.6%

J. A. Hartigan and P. M. Hartigan. *The dip test of unimodality*. The Annals of Statistics, 13(1):70–84, 1985.

Results: Clustering Quality

Method	Reuters (k=135)					LFW (k=1,680)					ESC (k=50)				
	$NMI\uparrow$	$ARI\uparrow$	$\hat{k}$	$Noise\downarrow$	T[s]↓	$NMI\uparrow$	$ARI\uparrow$	$\hat{k}$	$Noise\downarrow$	T[s]↓	$NMI\uparrow$	$ARI\uparrow$	$\hat{k}$	$Noise\downarrow$	T[s]↓
KMeans+Elbow	58.5%	19.9%	41	**0.0%**	1,917	78.0%	78.1%	773	**0.2%**	17,315	95.1%	83.3%	43	**0.0%**	306
HDBSCAN	62.0%	2.4%	1,247	61.4%	240	72.1%	36.3%	393	56.8%	105	86.2%	44.6%	**52**	17.3%	8
VDBSCAN	55.4%	0.3%	2,296	27.5%	246	92.3%	12.0%	2,661	38.0%	84	78.7%	20.9%	447	23.3%	10
OPTICS	61.3%	20.5%	37	97.1%	505	64.1%	24.5%	390	65.8%	202	56.3%	3.8%	53	59.7%	16
SS-DBSCAN	0.0%	0.0%	1	22.9%	230	13.5%	4.6%	2	52.7%	252	85.9%	46.6%	43	16.0%	9
AMD-DBSCAN	41.1%	28.3%	**134**	6.2%	69	71.7%	20.4%	281	34.0%	**29**	83.4%	31.5%	93	18.3%	13
AEDBSCAN	49.8%	4.5%	974	24.6%	144	91.8%	23.6%	1,944	48.7%	53	83.9%	46.7%	230	20.8%	7
AutoEps	66.4%	67.1%	646	67.5%	2,377	11.9%	1.7%	56	57.1%	82	91.8%	77.3%	144	36.0%	24
TS (ours)	**77.5%**	**93.9%**	138	55.2%	152	**99.0%**	**96.8%**	1,697	30.5%	60	**97.4%**	**90.3%**	57	32.3%	5
TSE (ours)	**77.8%**	92.9%	150	38.0%	**24**	**99.0%**	**96.7%**	**1,694**	30.4%	41	96.7%	85.2%	**48**	14.7%	**2**

Conclusions

- Observe the **Unimodality property** in density-based clustering
  - ➢Back it theoretically and empirically with the DIP test
- A **Ternary Search-based method** to automatically set epsilon
  - ➢Experiment over NLP, Vision, and Audio classification datasets
- **Our code** is available on GitHub

Q&A

Unimodal Strategies in Density-Based Clustering

Oron Nir[1,2], Jay Tenenbaum[2], and Ariel Shamir[1]

[1] CANVAS Lab, Reichman University

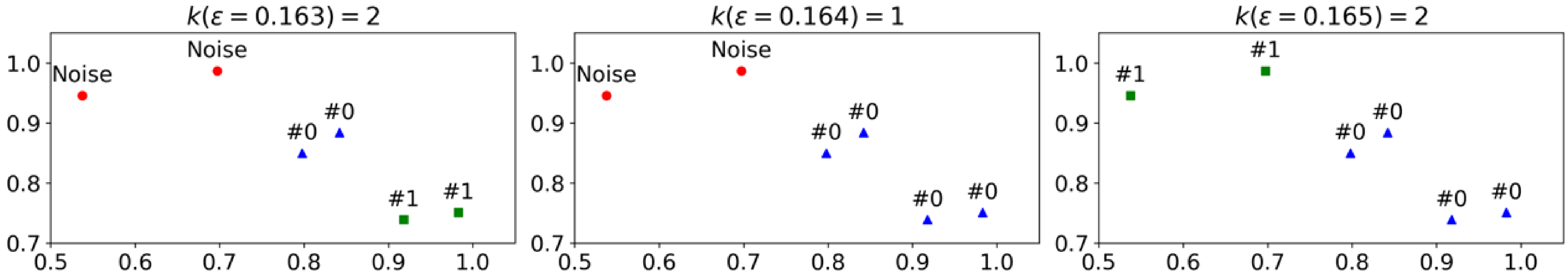
[2] Microsoft

Our code is available on GitHub:

CANVAS LAB

Microsoft

The Unimodality Property – A counter example



Theory

- Theorem 1: For uniform 1D data $X \sim U[0,1]$, $\mathrm{E}_{\mathrm{X}}[k(\varepsilon)]$ is unimodal.
  - Tools: order statistics $x_1 \leq \cdots \leq x_n$, and spacings analysis $s_i = x_i - x_{i-1}$.
- Theorems 2/3: WHP $\operatorname*{argmax}_{\epsilon} \mathrm{k}(\varepsilon) \in \left[\frac{1}{2}\sqrt[d]{\frac{MinPts}{n}}, \sqrt{d} \cdot \frac{1}{2}\sqrt[d]{\frac{MinPts}{n}}\right]$
  - Tools: divide $[0,1]^D$ to cubes, Hoeffding over #points in each cube.

Algorithm 1 $TS(X, LB, UB, MinPts, itr)$

for i=0 to itr **do**
  $m_l \leftarrow \frac{2LB+UB}{3}$
  $m_r \leftarrow \frac{LB+2UB}{3}$
  $\mathcal{C}_l \leftarrow \textbf{DBSCAN}(X, m_l, MinPts)$
  $\mathcal{C}_r \leftarrow \textbf{DBSCAN}(X, m_r, MinPts)$
  $\langle LB, UB \rangle \leftarrow Cond(LB, UB, \ldots$
    $m_l, m_r, \mathbf{K}(\mathcal{C}_l), \mathbf{K}(\mathcal{C}_r))$
end for
return $\frac{m_l+m_r}{2}$

Algorithm 2 $Cond(LB, UB, m_l, m_r, k_l, k_r)$

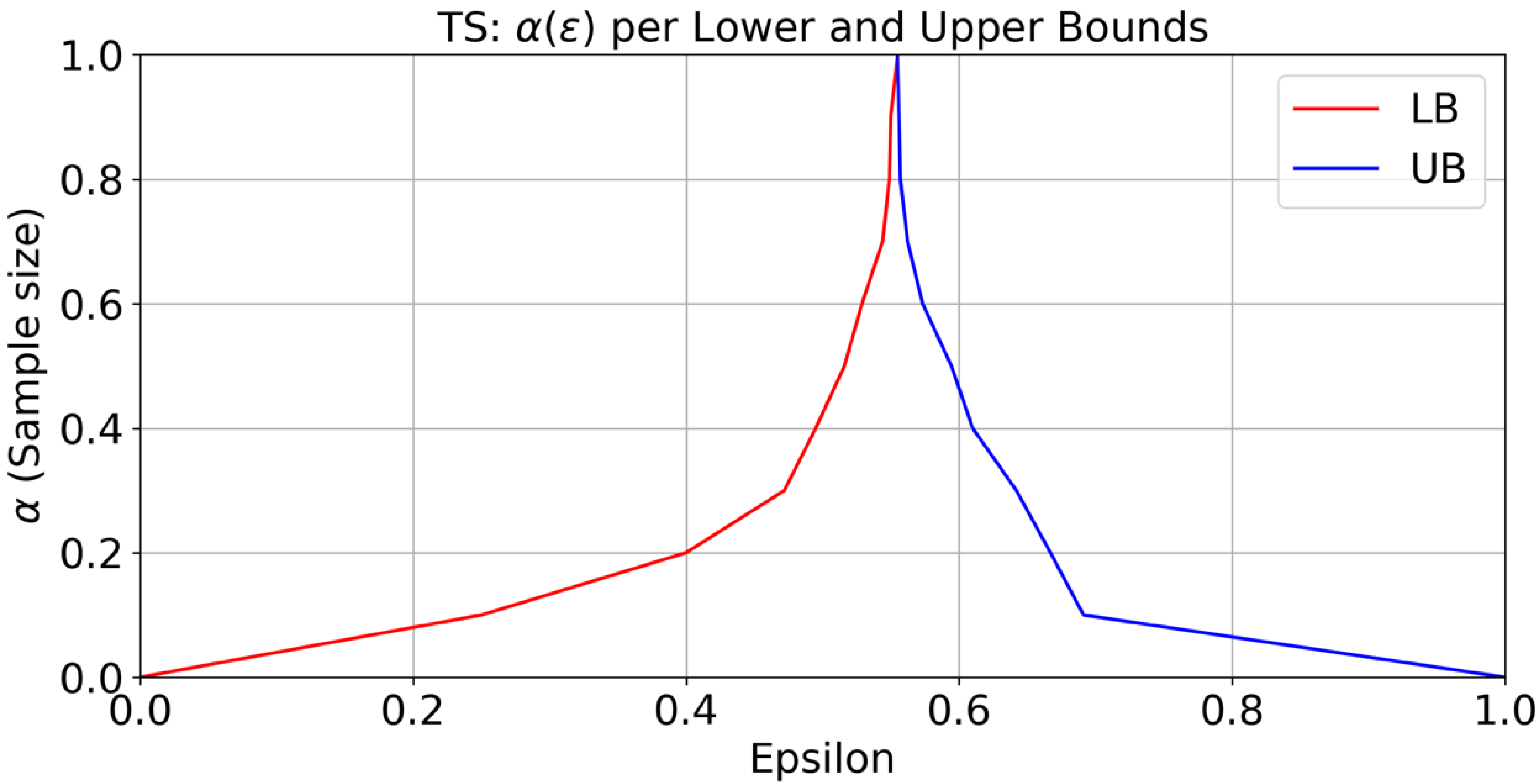
if $k_l == 1$ **and** $k_r == 1$ **then**
  return $\langle LB, m_l \rangle$
else if $k_l == 0$ **and** $k_r == 1$ **then**
  return $\langle m_l, m_r \rangle$
else if $k_l == 0$ **and** $k_r == 0$ **then**
  return $\langle m_r, UB \rangle$
else if $k_l > k_r$ **then**
  return $\langle LB, m_r \rangle$
else
  return $\langle m_l, UB \rangle$
end if

Method: TS Clustering

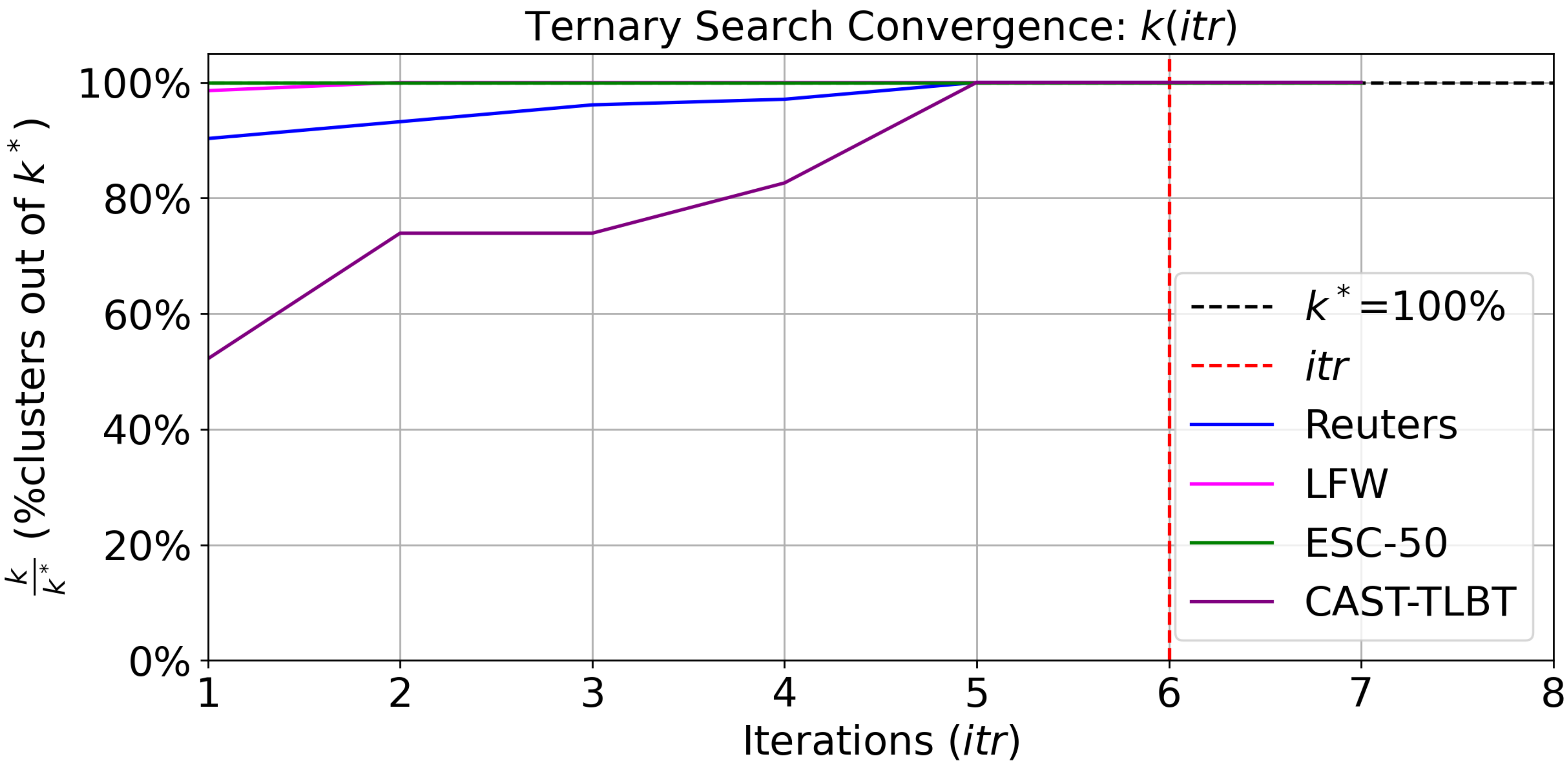
Algorithm 3 TSClustering(X, MinPts, itr)

1: $UB^0 \leftarrow max_{i \in \{2,\ldots,N\}} d(X_1, X_i)$
2: $LB^0 \leftarrow 0$
3: $\mathcal{R} \leftarrow$ sample $\lceil \alpha N \rceil$ points from X
4: $\mathcal{T} \leftarrow$ sample $\lceil \alpha D \rceil$ dimensions from X
5: $UB \leftarrow TS(X_{\mathcal{R},1:D}, LB^0, UB^0, MinPts, itr)$
6: $LB \leftarrow TS(X_{1:N,\mathcal{T}}, LB^0, UB, MinPts, itr)$
7: $\varepsilon^* \leftarrow TS(X, LB, UB, MinPts, itr)$
8: **return** $DBSCAN(X, \varepsilon^*, MinPts)$

Ablations: α sampling ratio vs. $[LB, UB]$



Ablations: Setting itr and convergence to k^*



Ternary Search Estimator (TSE)

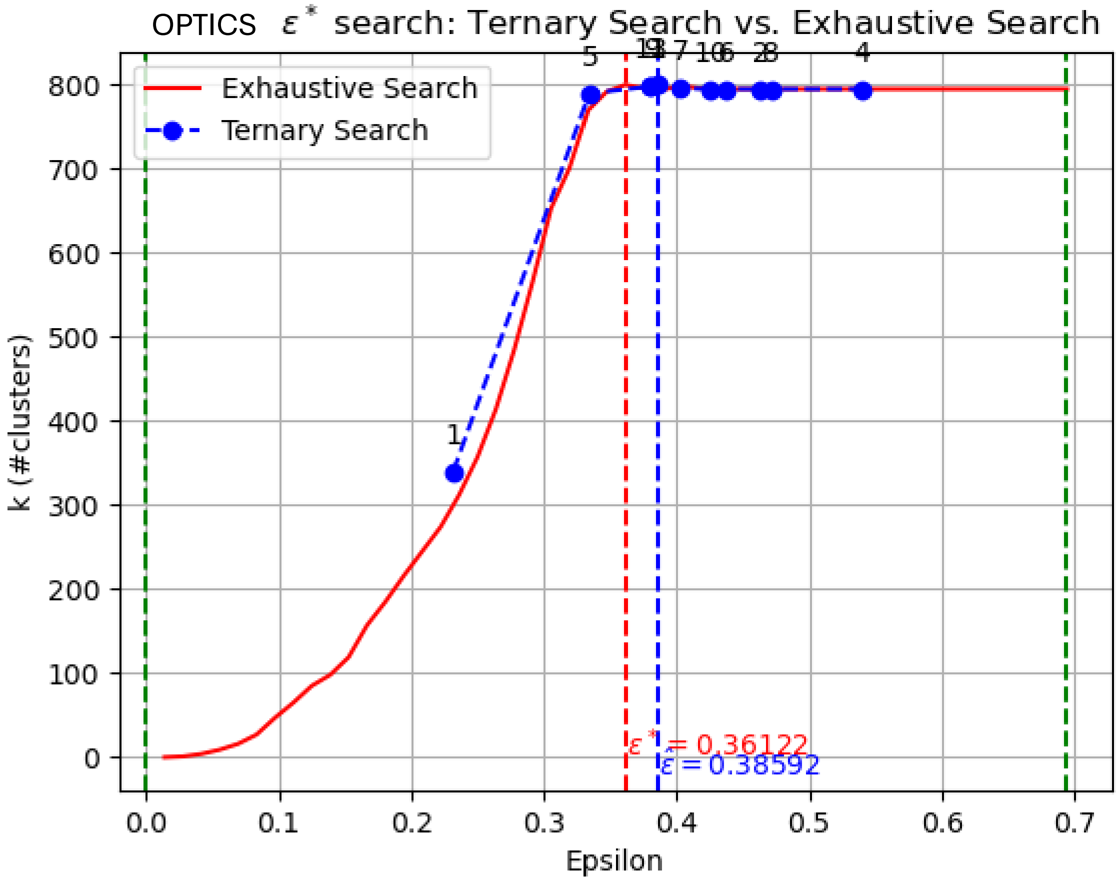
- We propose an estimator (TSE) for ε^* obtained by sampling an α fraction of the data and dimensions simultaneously
- The intuition is that the opposite influences of sampling the data and dimensions on ε^* should roughly cancel out
- We average this estimator over $m = 30$ iterations

$$\varepsilon^* = \frac{1}{m}\sum_{i=1}^{m} TS_i(X_{\mathcal{R},\mathcal{T}}, LB^0, UB, MinPts, itr)$$

Limitations: OPTICS

OPTICS takes the max-ε and supports multi-resolution clustering for clusters of different densities.

Resulting a non-unimodal relation.



NMI and ARI

Aspect	**Normalized Mutual Information (NMI)**	**Adjusted Rand Index (ARI)**
Definition	Based on **information theory**; measures mutual dependence between two clusterings normalized by entropy.	Based on **pairwise counting**; measures agreement of sample pairs assigned to clusters, adjusted for chance.
Formula	$NMI(U,V) = \frac{2MI(U,V)}{H(U)\ +H(V)}$	$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$
Range	$[0,1]$	$[-1,1]$
Interpretation	0 → independent clusterings; 1 → perfect agreement.	0 → random labeling; 1 → perfect agreement; negative → worse than random.
Sensitivity	Captures **distributional similarity**; tolerant of label permutation.	Captures **pairwise assignments**; highly sensitive to exact pair matching.
Bias	Not adjusted for chance; tends to give higher scores for unbalanced partitions.	Adjusted for chance; penalizes trivial solutions (e.g., all points in one cluster).
Use Cases	Comparing **overall information overlap** between partitions (e.g., topic clustering, community detection).	Comparing **fine-grained consistency** of partitions, especially where pairwise relations matter (e.g., ReID, MOT tracking, clustering evaluation).