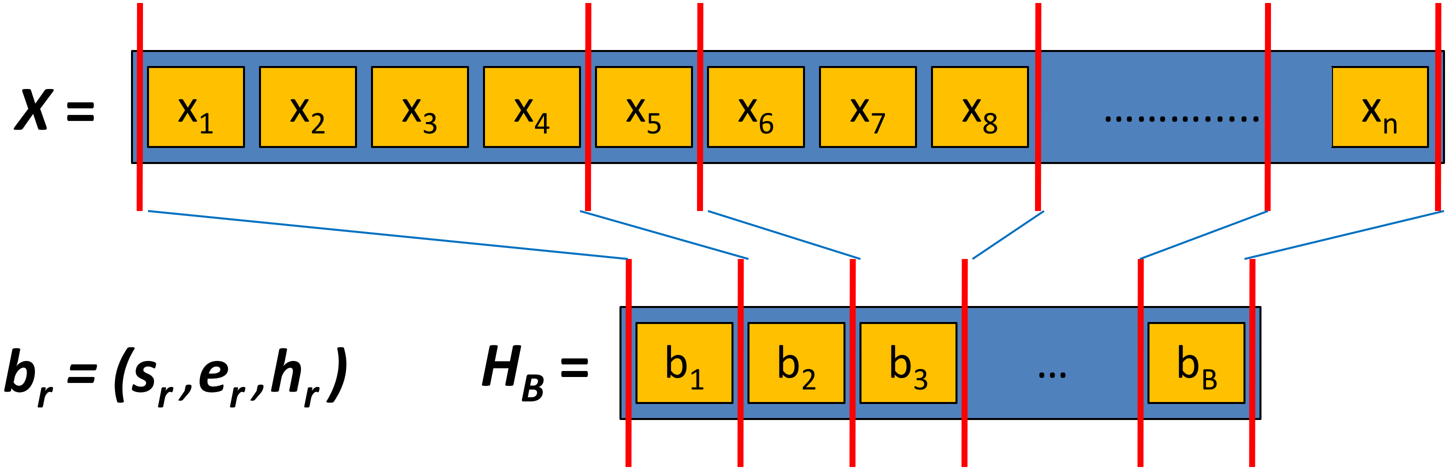
Local Search in Histogram Construction

Felix Halim, Panagiotis Karras, Roland Yap

NUS
National University
of Singapore

Problem Statement (1/2)

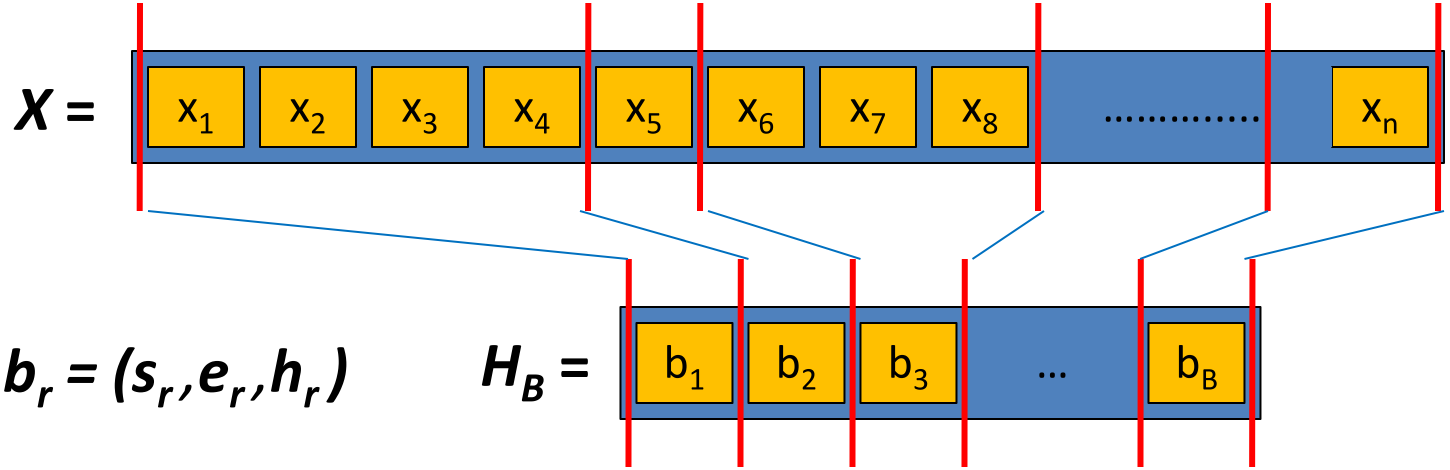
- Given a finite data sequence $\boldsymbol{X = x_1,\ldots,x_n}$
- Create and store a compact representation $\boldsymbol{H_B}$ of $\boldsymbol{X}$ using at most $\boldsymbol{B}$ storage space
- Minimize the total error of $\boldsymbol{E_X(H_B)}$



Problem Statement (2/2)

- Minimize the total error, **Ex**
- ***Min $E_X(H_B)$*** = $\sum_r \textsc{SqError}(b_r)$ for a for *r = 1 .. B*

$$\textsc{SqError}(b_r) = \sum_{i=s_r}^{e_r} (x_i - h_r)^2$$



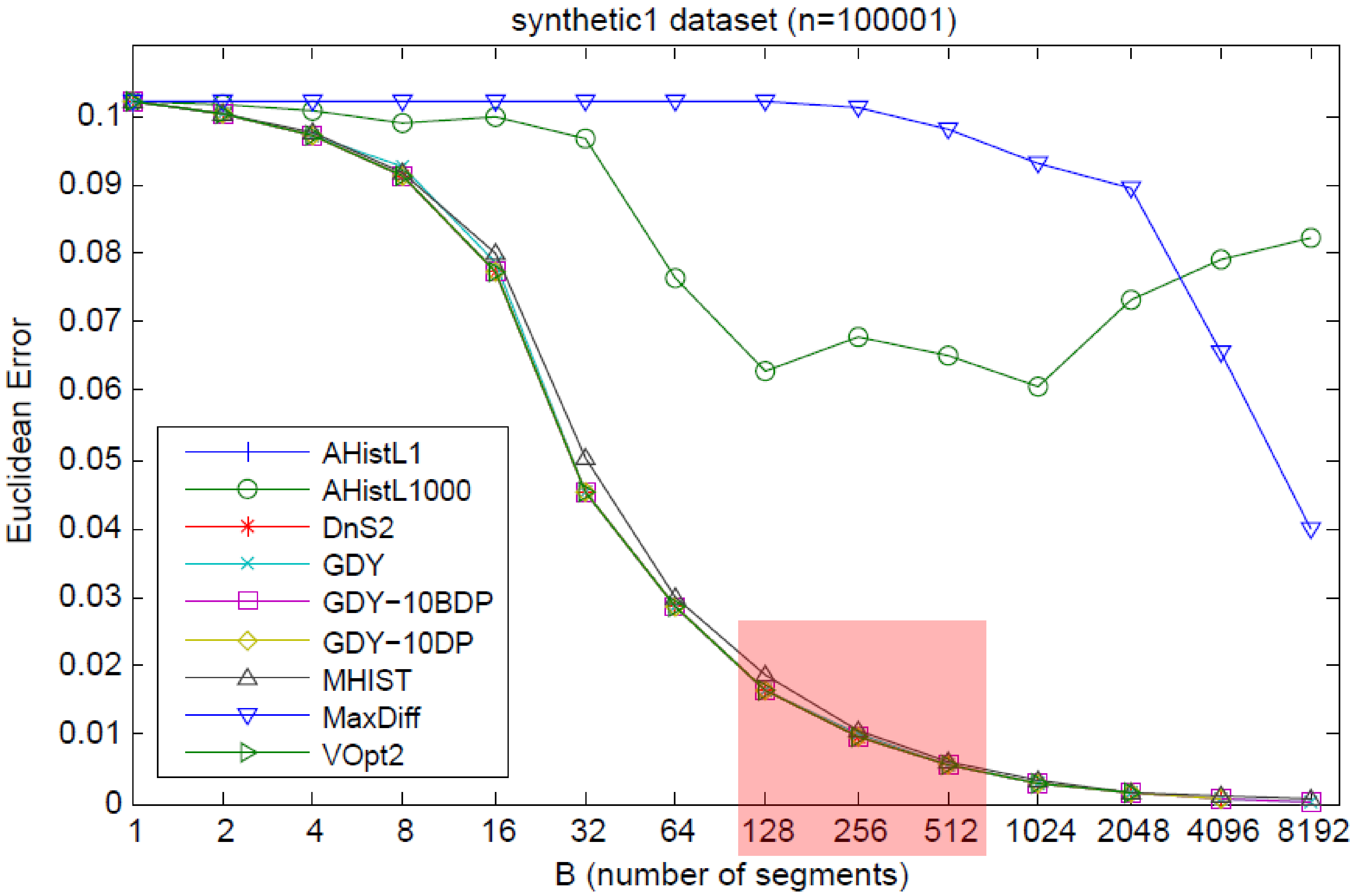
Applications

- Database Systems
- Decision Support Systems
- Bio-Informatics
- Information Retrieval

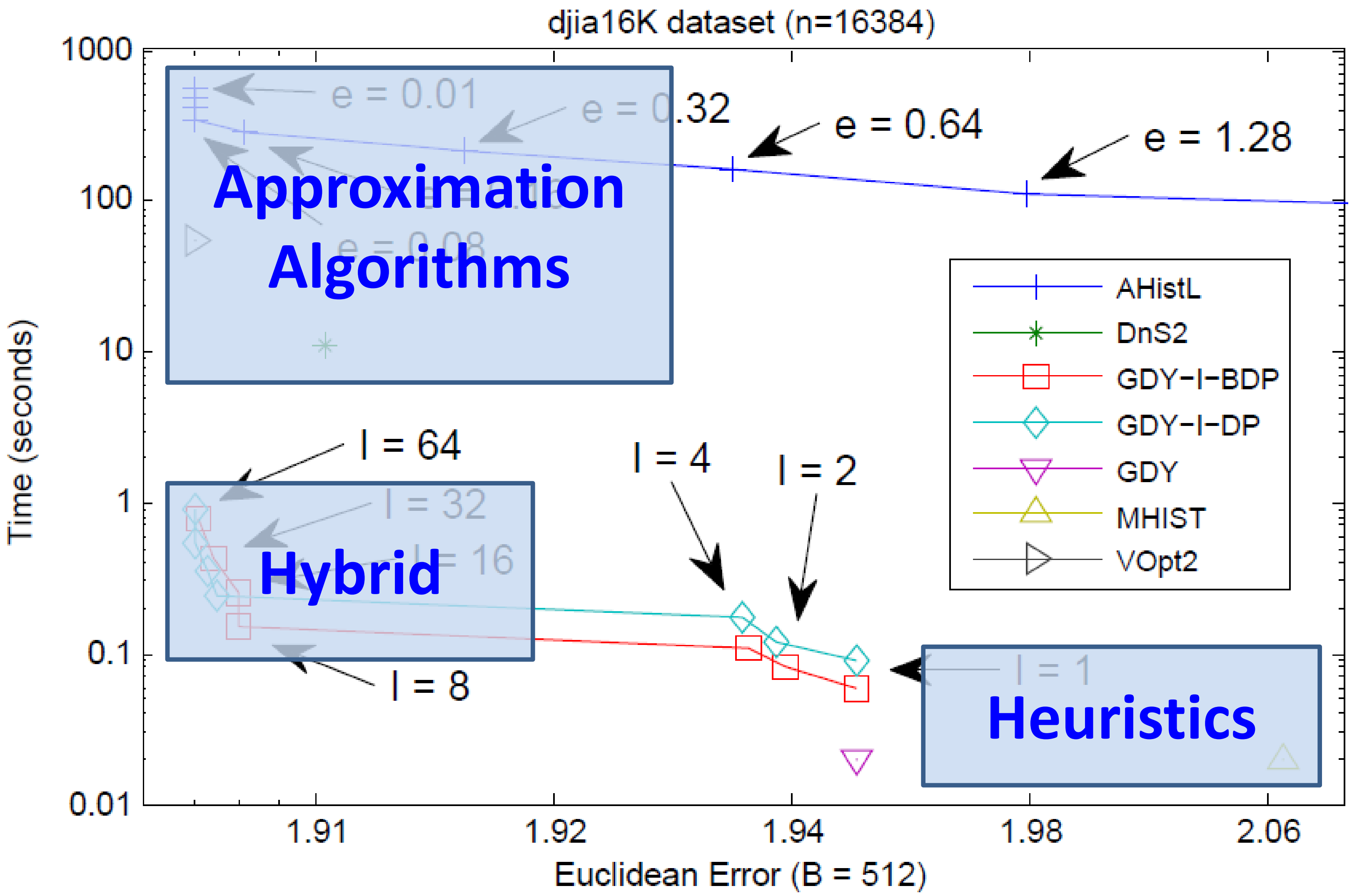
Results on Histogram Construction

Category	Name	Complexity
Optimal	V-Optimal	n^2 B
Heuristics	MHIST	B * (n + log B)
	MaxDiff	n * log B
Approximations	AHistL-Δ	n + B^3(log n + e^{-2}) log n
	DnS	$n^{4/3}$ $B^{5/3}$
Hybrid (CIKM 09)	**GDY-DP**	**n B (for B less than √n)**
	GDY-BDP	**n B**

Effective B range = [128 .. 512] for large n



Effectiveness - Tradeoff



Approximation Algorithms

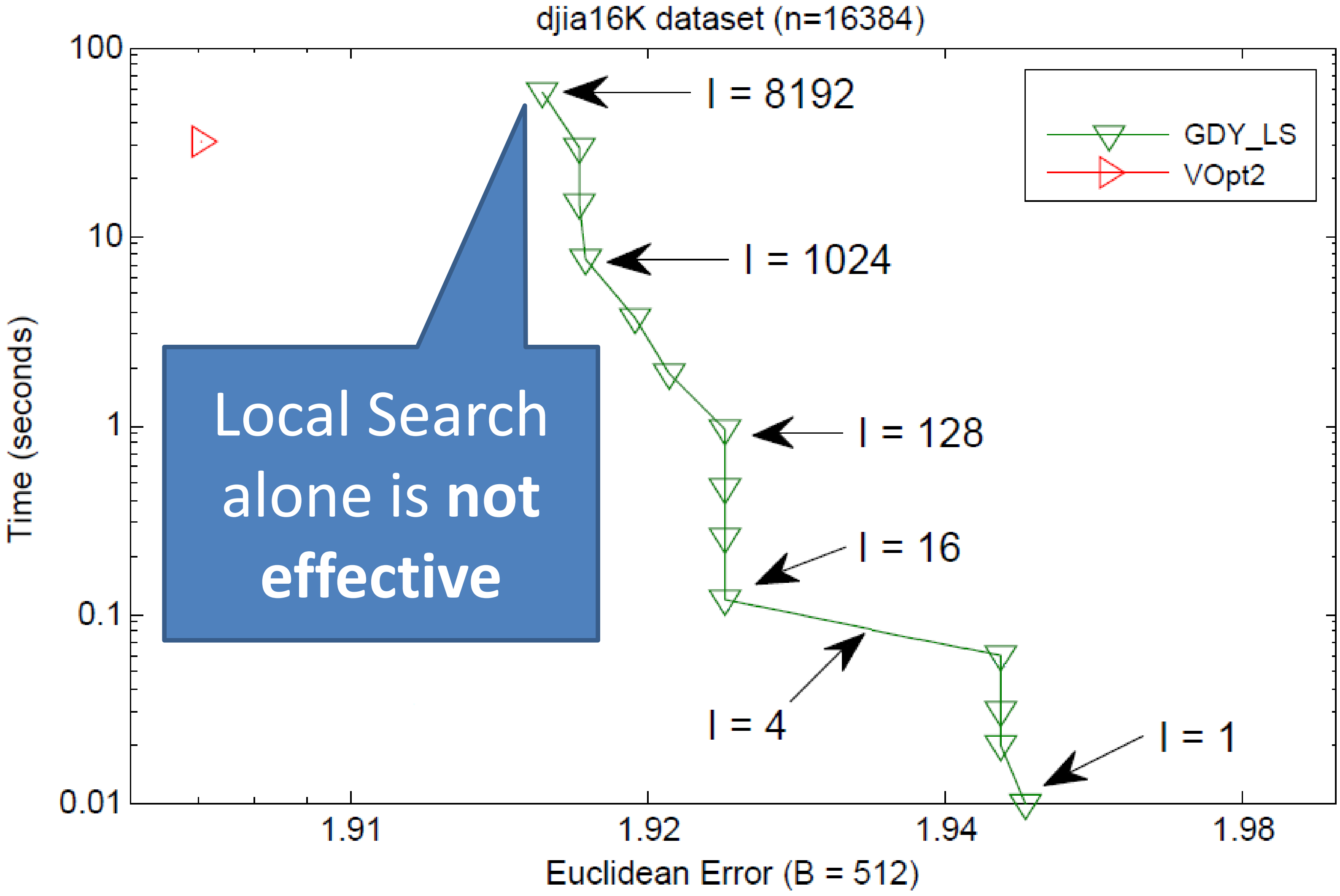
Hybrid

Heuristics

Hybrid

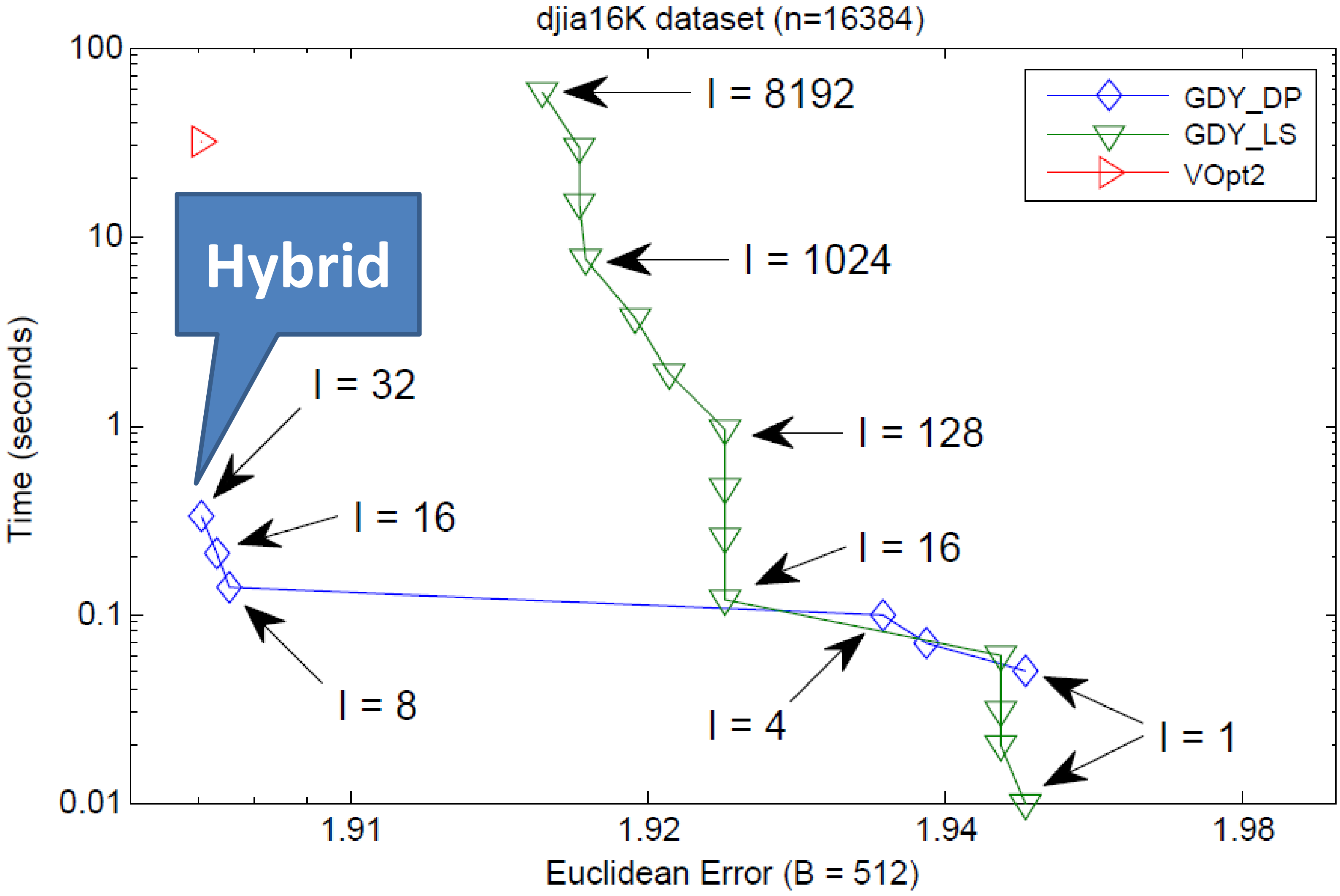
- Local Search
  - Moves from solution to solution
  - Used for collecting a good-diversified samples
    - AHistL and DnS fails to provide good sampling
- Optimal Algorithm
  - Dynamic Programming
    - Used to take the best out of the samples
  - Served as a ***catalyst*** for the Local Search

Optimal Algorithm as a ***Catalyst***



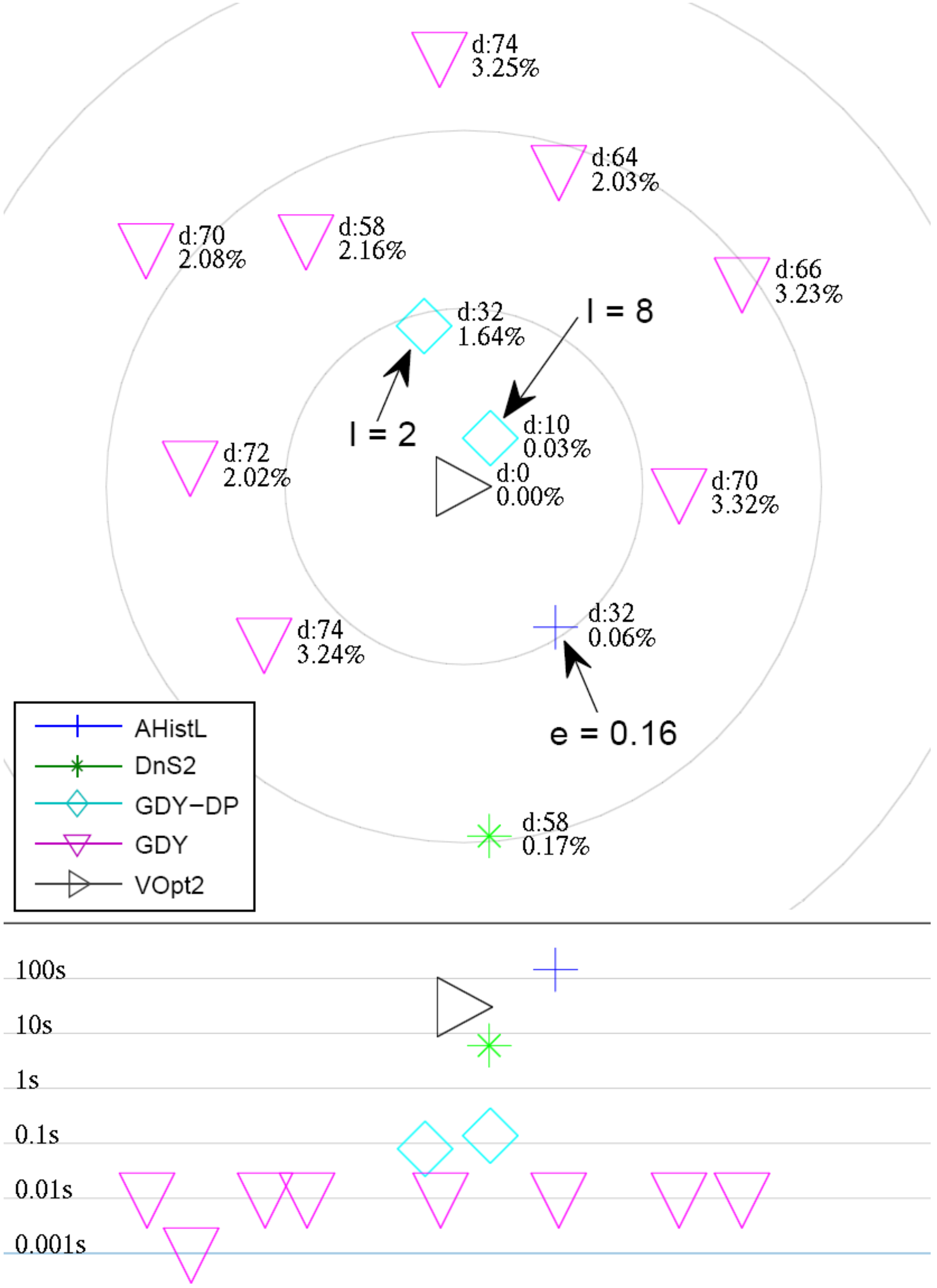
Local Search alone is **not effective**

Optimal Algorithm as a ***Catalyst***



Visualization of the Search

- AHistL
  - ➢ **0.06%** quality
  - ➢ **32** misplaced
- DnS
  - ➢ **0.17%** quality
  - ➢ **58** misplaced
- GDY-DP
  - ➢ **0.03%** quality
  - ➢ **10 misplaced**



Conclusion

- We *advanced* state of the art
  - Despite of long history of Histogram Construction
- Stand-alone algorithms
  - Heuristics
    - Good performance but poor quality
  - Approximation algorithms
    - Sacrificing performance for error guarantees
  - They are not effective / efficient enough

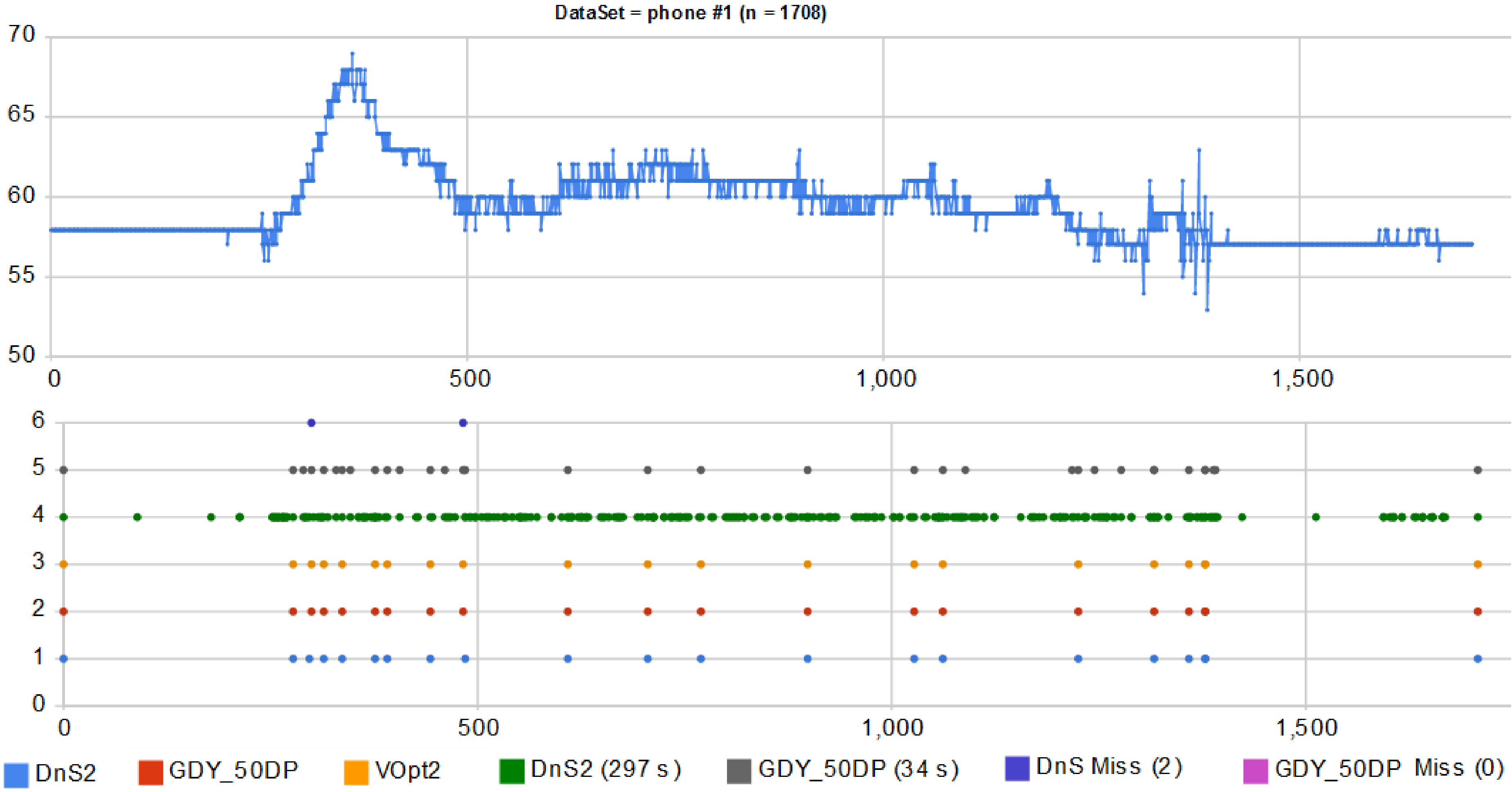
Conclusion

- Hybrid
  - Local Search (LS)
    - Used for collecting diversified samples
    - The better the LS, the better the samples
  - Optimal / Better algorithms
    - Used to select best of the samples
    - Served as a ***catalyst*** for the Local Search

Thank You

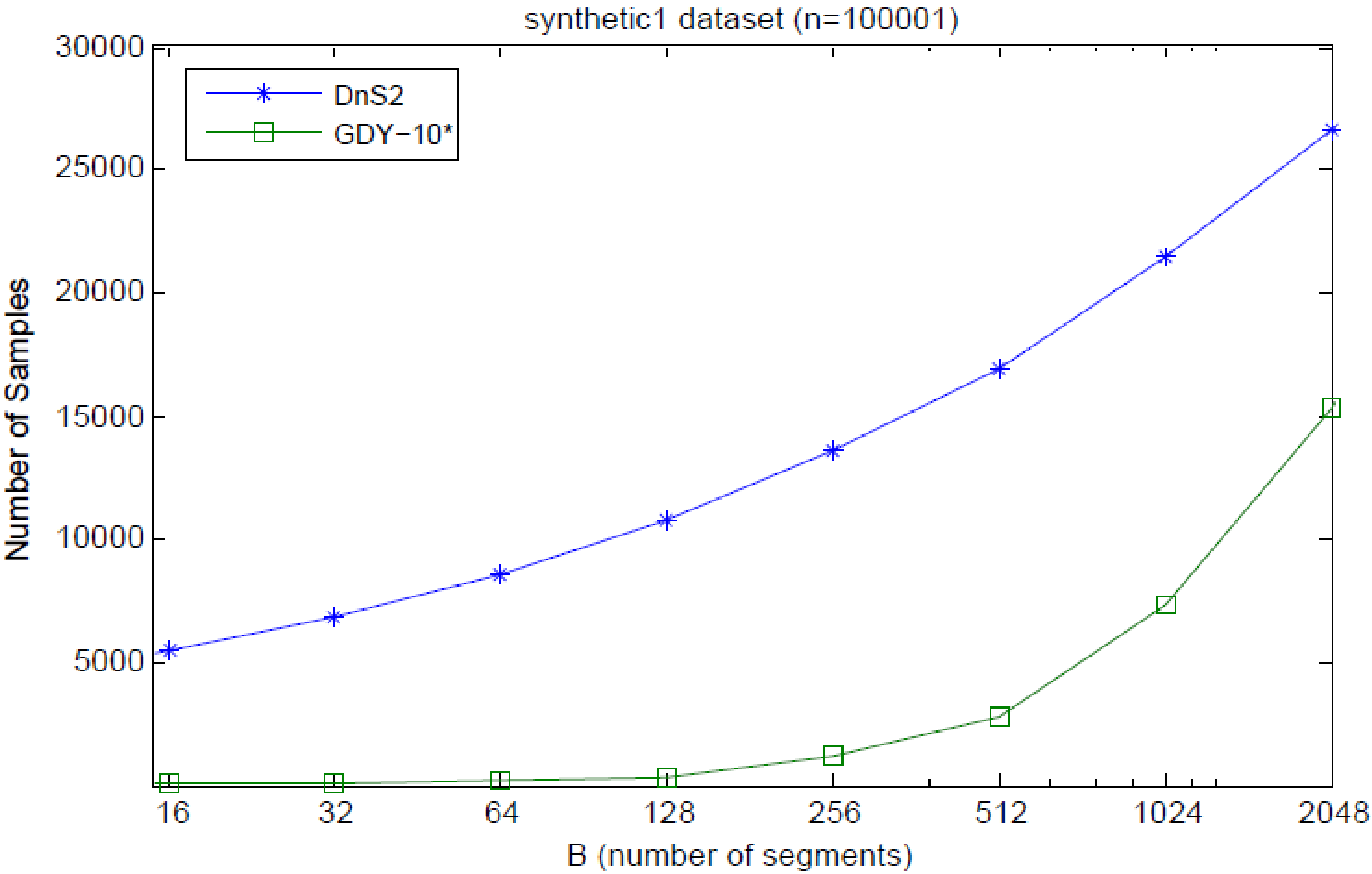
- Questions and Answers

LS Sampling Effectiveness



http://felix-halim.net/histogram

Sampling Quantity


synthetic1 dataset (n=100001)
DnS2
GDY−10*
Number of Samples
30000
25000
20000
15000
10000
5000
16
32
64
128
256
512
1024
2048
B (number of segments)