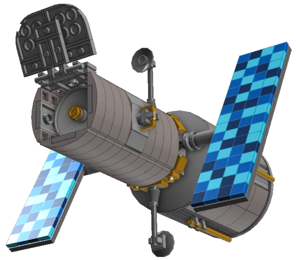
Comparing Synopsis Techniques for Approximate Spatial Data Analysis

A. B. Siddique, Ahmed Eldawy, Vagelis Hristidis
University of California, Riverside

Spatial Data is Ubiquitous…

- The amount of spatial data is exponentially increasing

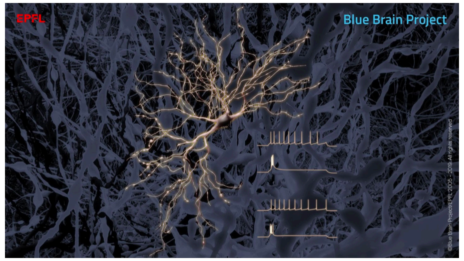




150 GB Weekly

50GB daily

60-80% Geo-referenced



Brain Simulation

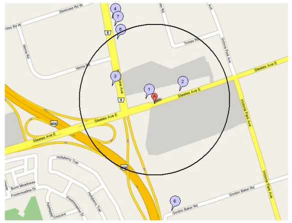
↑Crop Production

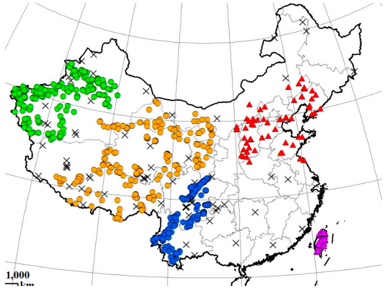
Event Detection

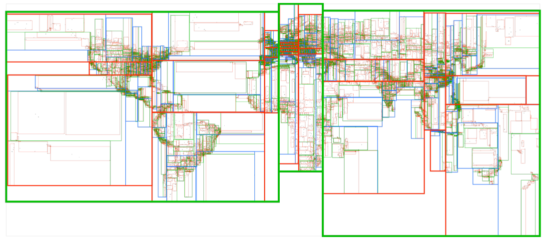
Geo- Advertisement

Spatial Data Synopsis

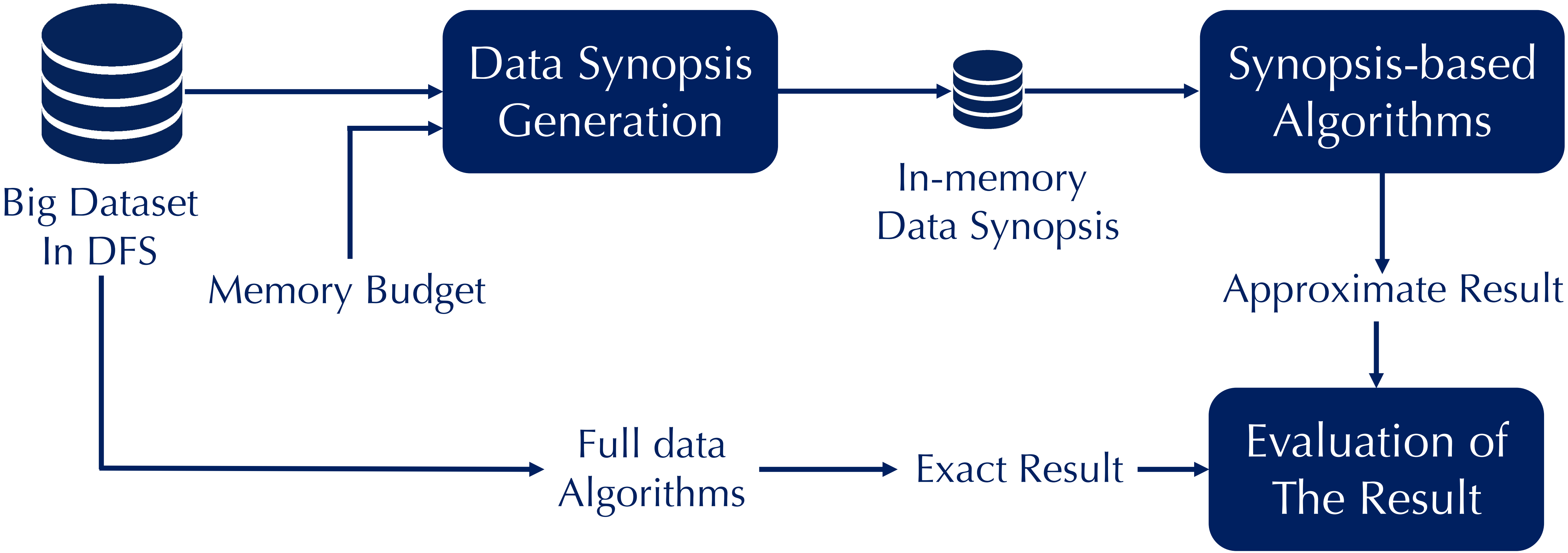
- Fast response algorithms are needed
- Data Synopsis:
  - Fast approximate results
  - Speed and accuracy trade-off
- Many spatial data synopsis
- Many common spatial operations
- **Given a problem-setting, when to use which?**
  - No readily-available answer



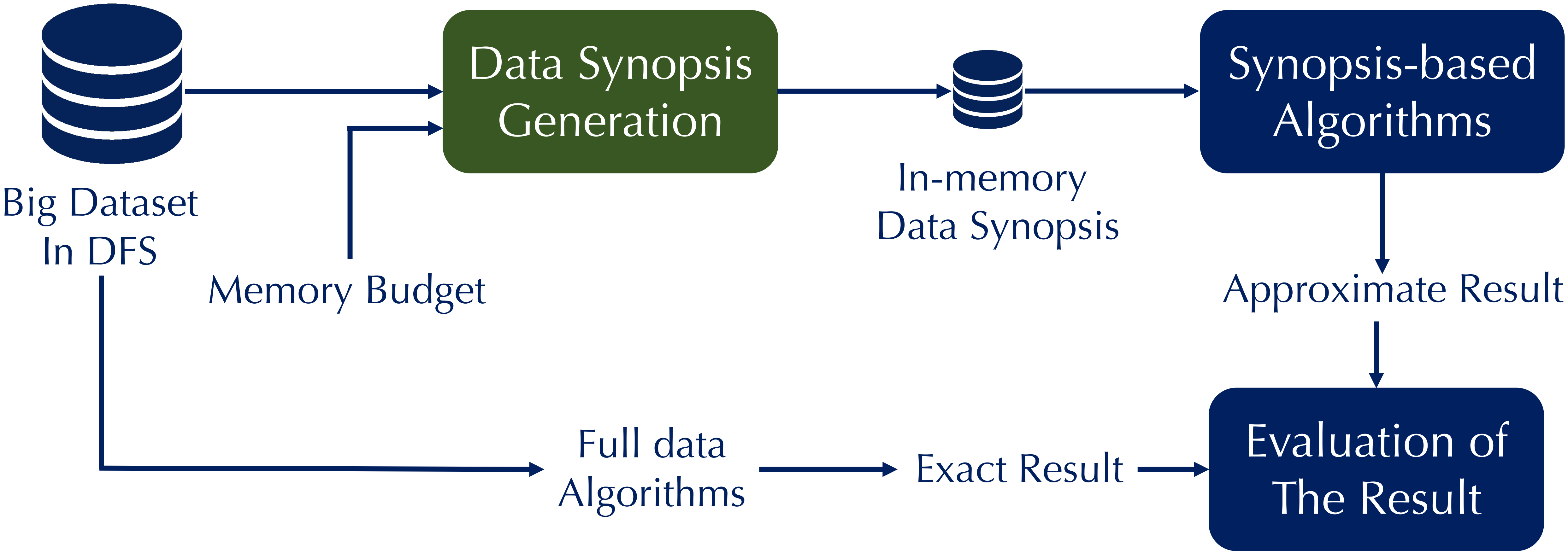




A Unified Environment



Outline



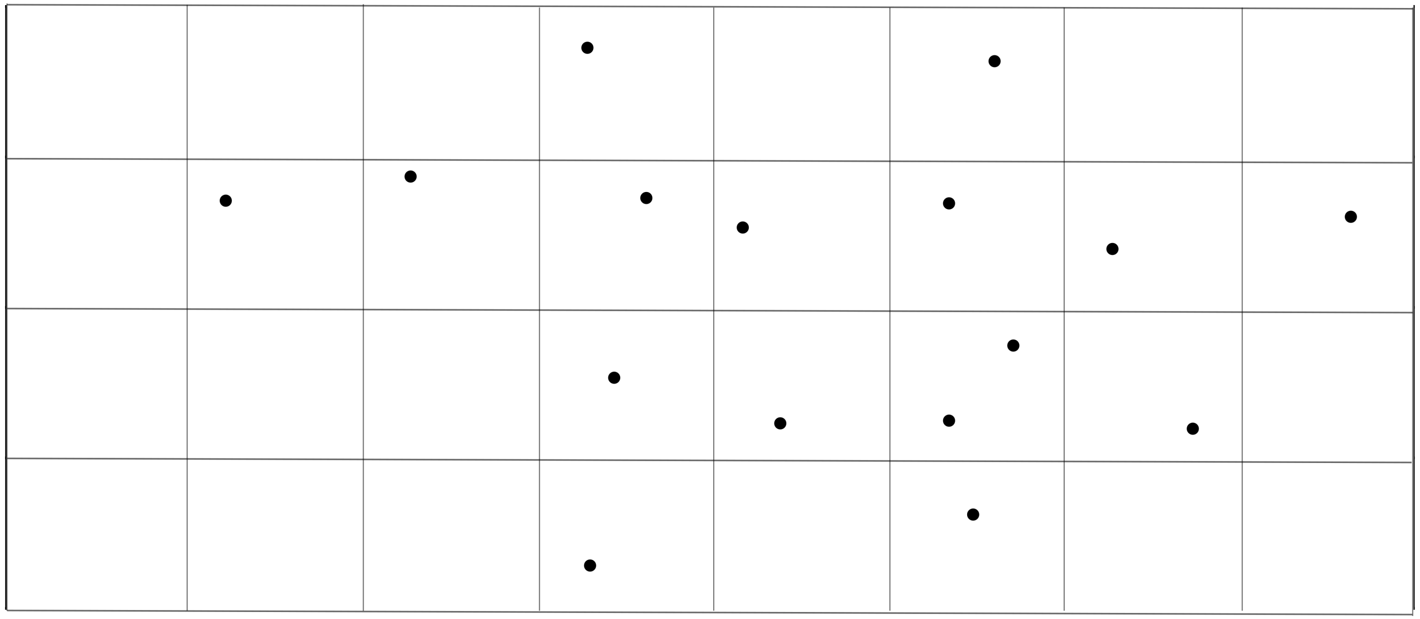
Data Synopsis Generation
Synopsis-based Algorithms
Big Dataset In DFS
Memory Budget
In-memory Data Synopsis
Approximate Result
Full data Algorithms
Exact Result
Evaluation of The Result

Synopsis Generation – Random Sample

- **Inputs:** Big dataset, Memory Budget, Synopsis algorithm
- **Output:** Synopsis of the required size

Synopsis Generation – Stratified Sample

- **Inputs:** Big dataset, Memory Budget, Synopsis algorithm
- **Output:** Synopsis of the required size



Synopsis Generation – Uniform Histogram

- **Inputs:** Big dataset, Memory Budget, Synopsis algorithm
- **Output:** Synopsis of the required size

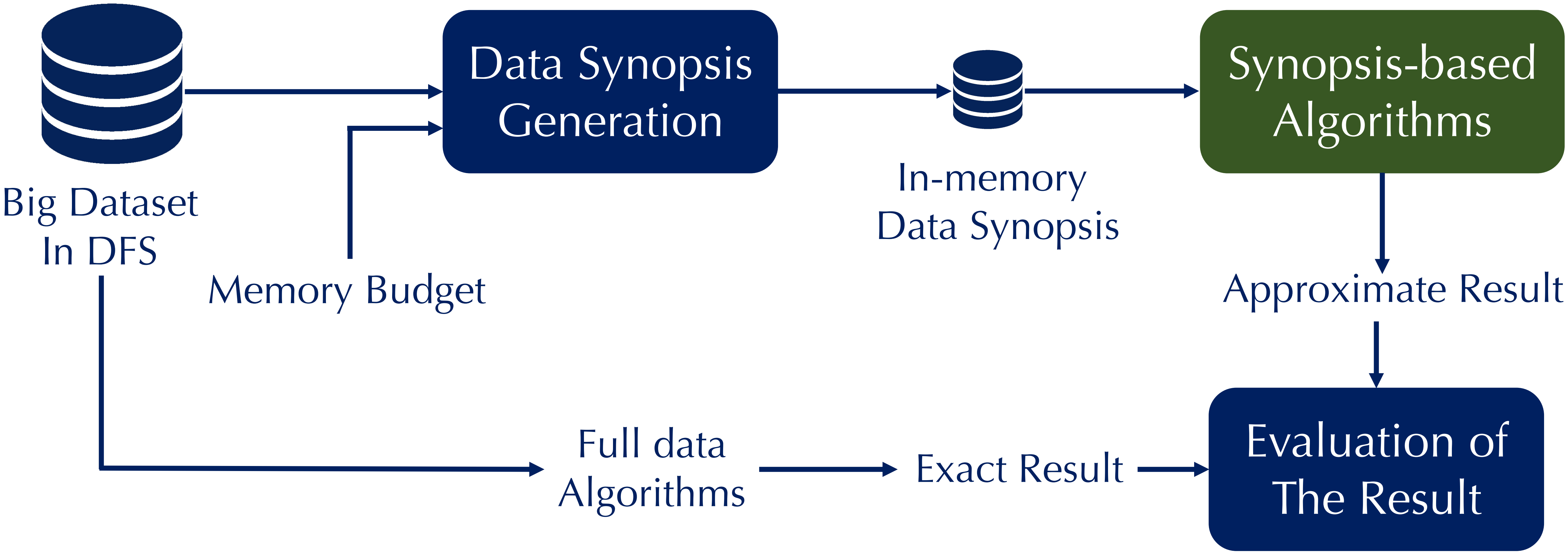
63	64	51	121	130	65	12	39
58	46	74	184	287	355	301	49
11	16	44	192	268	374	130	0
0	0	2	65	41	46	17	0

Synopsis Generation – Non-Uniform Histogram

- **Inputs:** Big dataset, Memory Budget, Synopsis algorithm
- **Output:** Synopsis of the required size

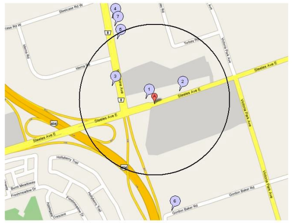
121	49	34	55	49	77	186
179	157	140	17 4	159	115	178
137	152	164	23 7	194	248	300

Outline

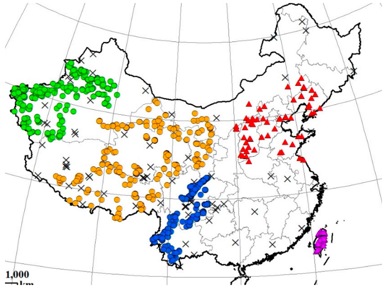

Data Synopsis Generation
Synopsis-based Algorithms
Big Dataset In DFS
Memory Budget
In-memory Data Synopsis
Approximate Result
Full data Algorithms
Exact Result
Evaluation of The Result

Spatial Operations

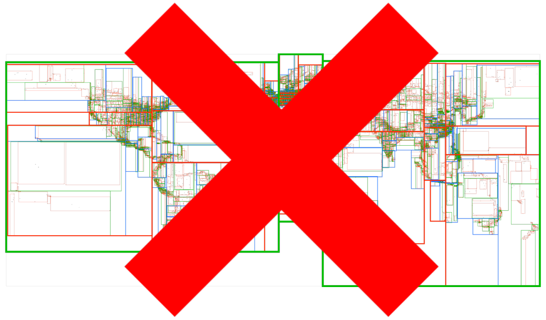
- Spatial Operations studied in this work



Selectivity Estimation



K-Means Clustering



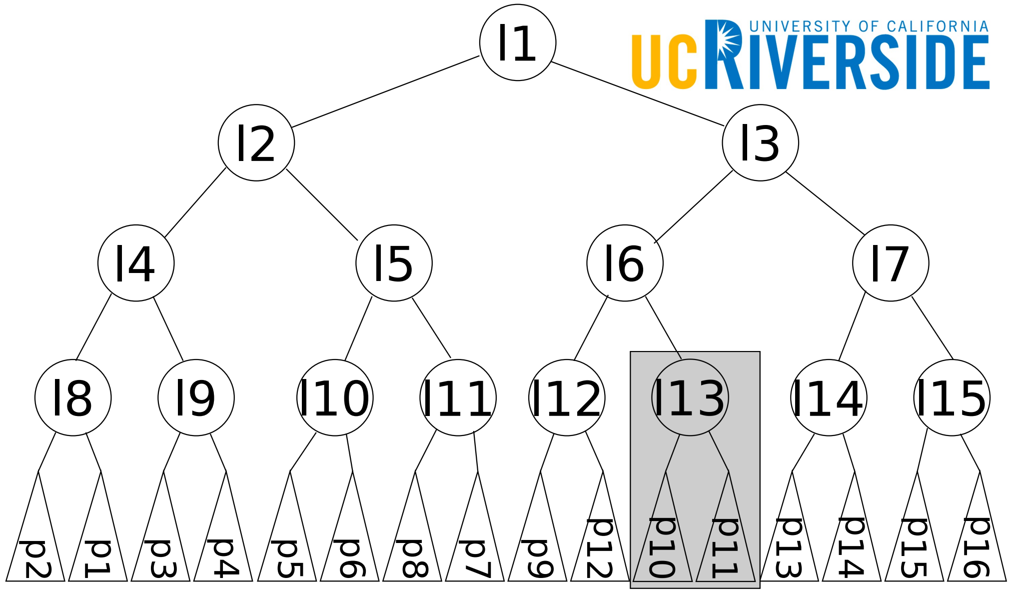
Spatial Partitioning

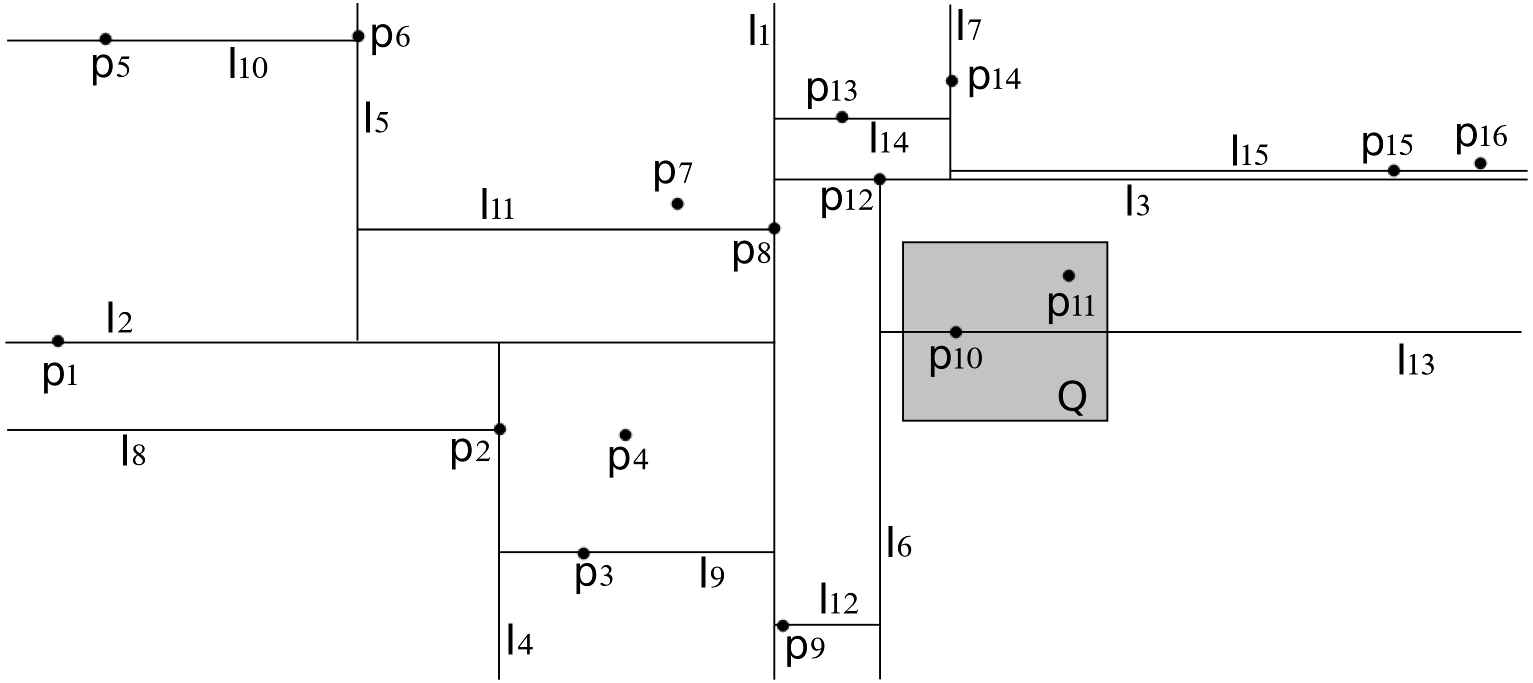
Selectivity Estimation (SE)

- Estimates the number of records in a given query range
- **Inputs:** Synopsis, Range Query
- **Output:** Number of records

SE – Based on Sample

- **Inputs:** Synopsis, Range Query
- **Output:** Number of records





SE – Based on Histogram

- **Inputs:** Synopsis, Range Query
- **Output:** Number of records

Horizontal Prefix-sum

Vertical Prefix-sum

63	127	178	299	429	494	506	545
121	231	356	661	1078	1498	1811	1899
132	258	427	924	1609	2403	2846	2934
132	258	429	991	1717	2557	3017	3105

+2403
-494
-427
+178
=1606

SE – Based on Histogram

- **Inputs:** Synopsis, Range Query
- **Output:** Number of records

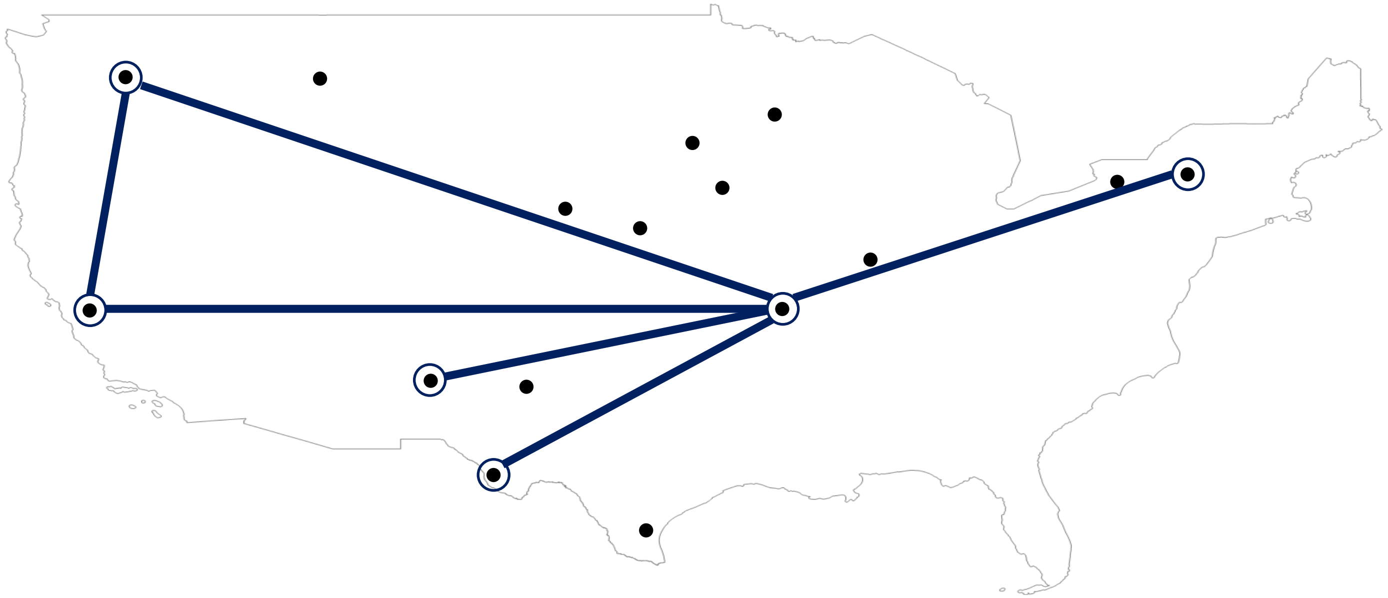
63	127	178	299	429	494	506	545
121	231	356	661	1078	1498	1811	1899
132	258	427	924	1609	2403	2846	2934
132	258	429	991	1717	2557	3017	3105

K-Means Clustering (KC)

- An unsupervised learning problem, which tries to group objects having some kind of similarity into one cluster
- **Inputs:** Synopsis, Number of clusters
- **Output:** K-cluster centroids

KC – Based on Sample

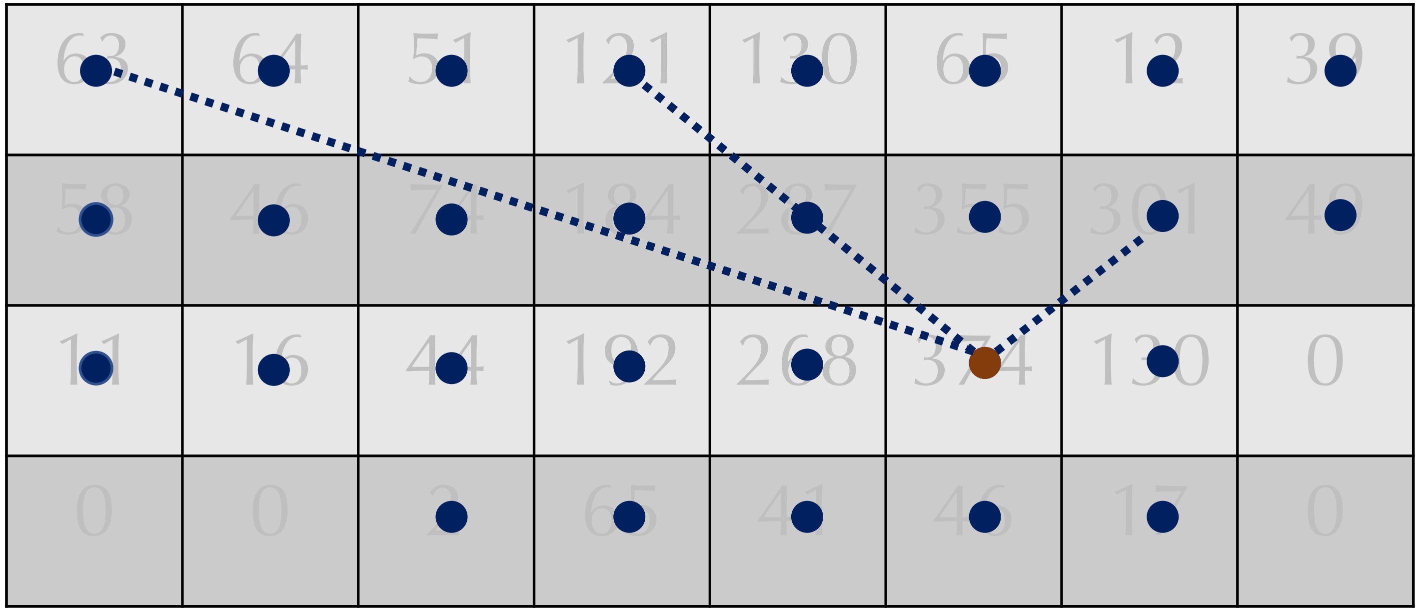
- **Inputs:** Synopsis, Number of clusters
- **Output:** K-cluster centroids



$$d(x, C)^2$$

KC – Based on Histogram

- **Inputs:** Synopsis, Number of clusters
- **Output:** K-cluster centroids

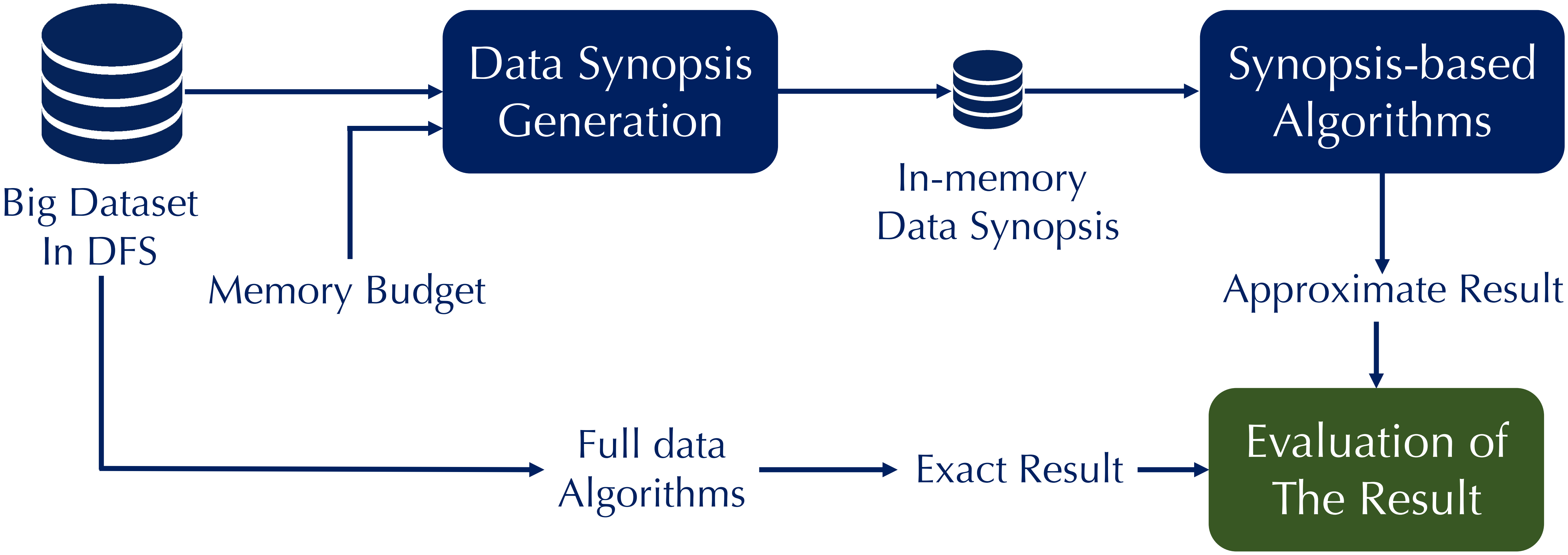


$$f \times d(x, C)^2$$

Full Dataset Algorithms

- State-of-the-art parallel algorithms for every operation
- **Selectivity Estimation**
  - Apache Spark's Filter and count
- **K-Means Clustering**
  - Apache Spark MLlib's Scalable K-Means++

Outline



Data Synopsis Generation
Synopsis-based Algorithms
Big Dataset In DFS
Memory Budget
In-memory Data Synopsis
Approximate Result
Full data Algorithms
Exact Result
Evaluation of The Result

Experimental Setup

Datasets

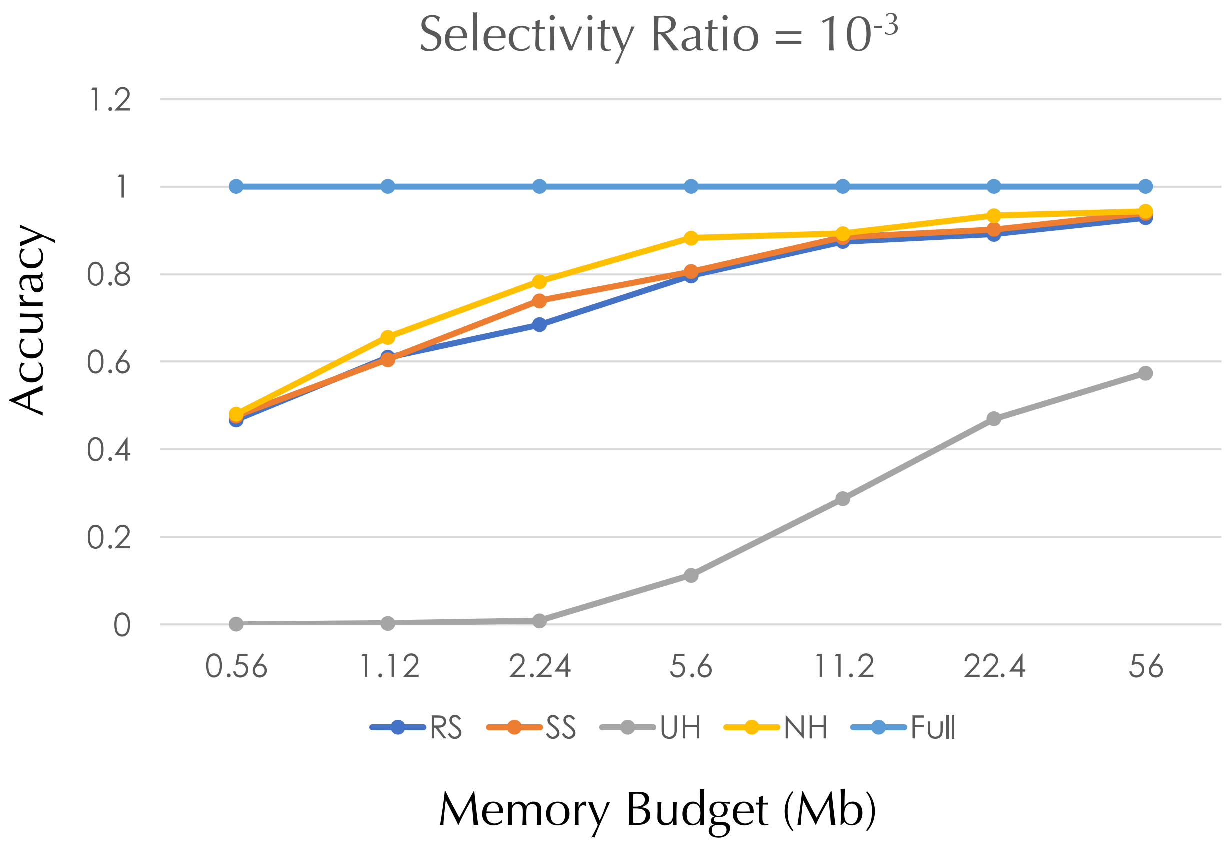
Name	Size	Records	Description
all-nodes	96 GB	2.7 billion	Points
edges	23 GB	70 million	Polygons
all-objects	92 GB	263 million	Mixed
Synthetic	51 GB	250 million	Rectangles

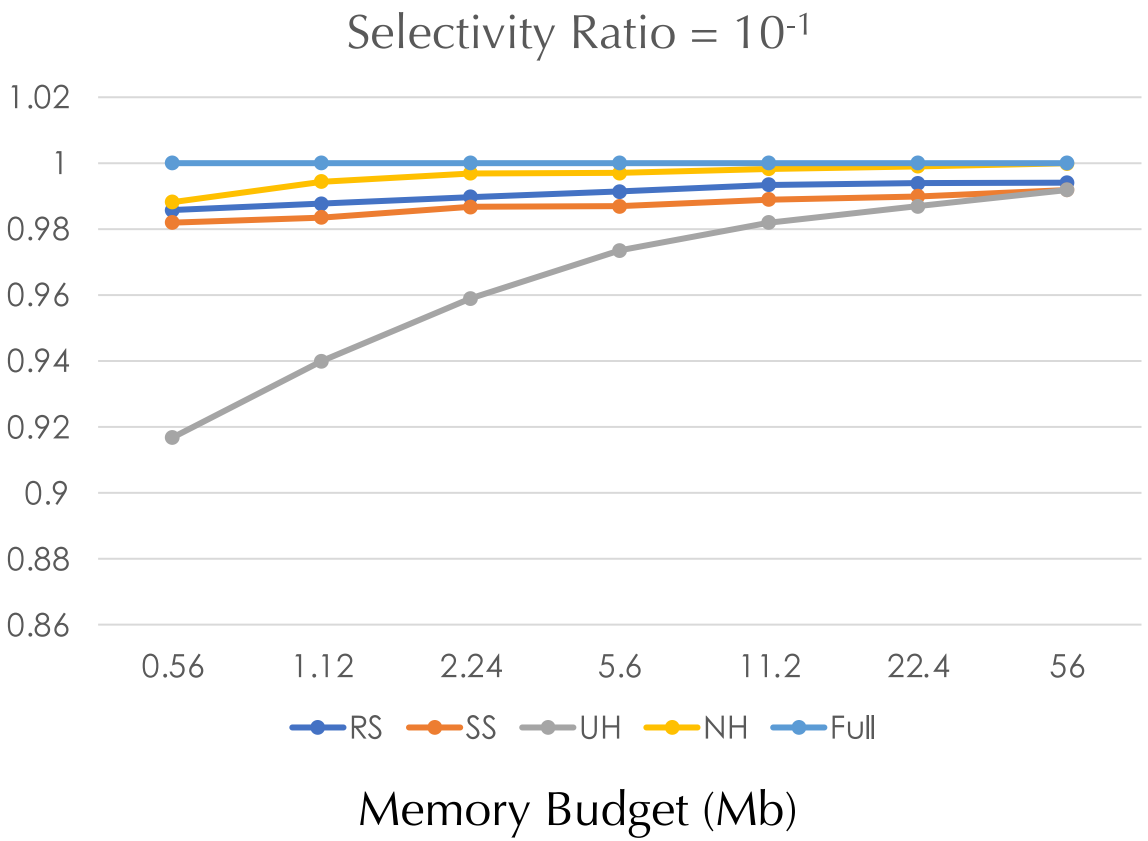
Performance Metric

Problem	Quality	Performance
SE	Absolute Relative Accuracy	Query Response Time
KC	Sum of Squared Error	Clustering Time

Selectivity Estimation: Quality Measure

Absolute Relative Accuracy

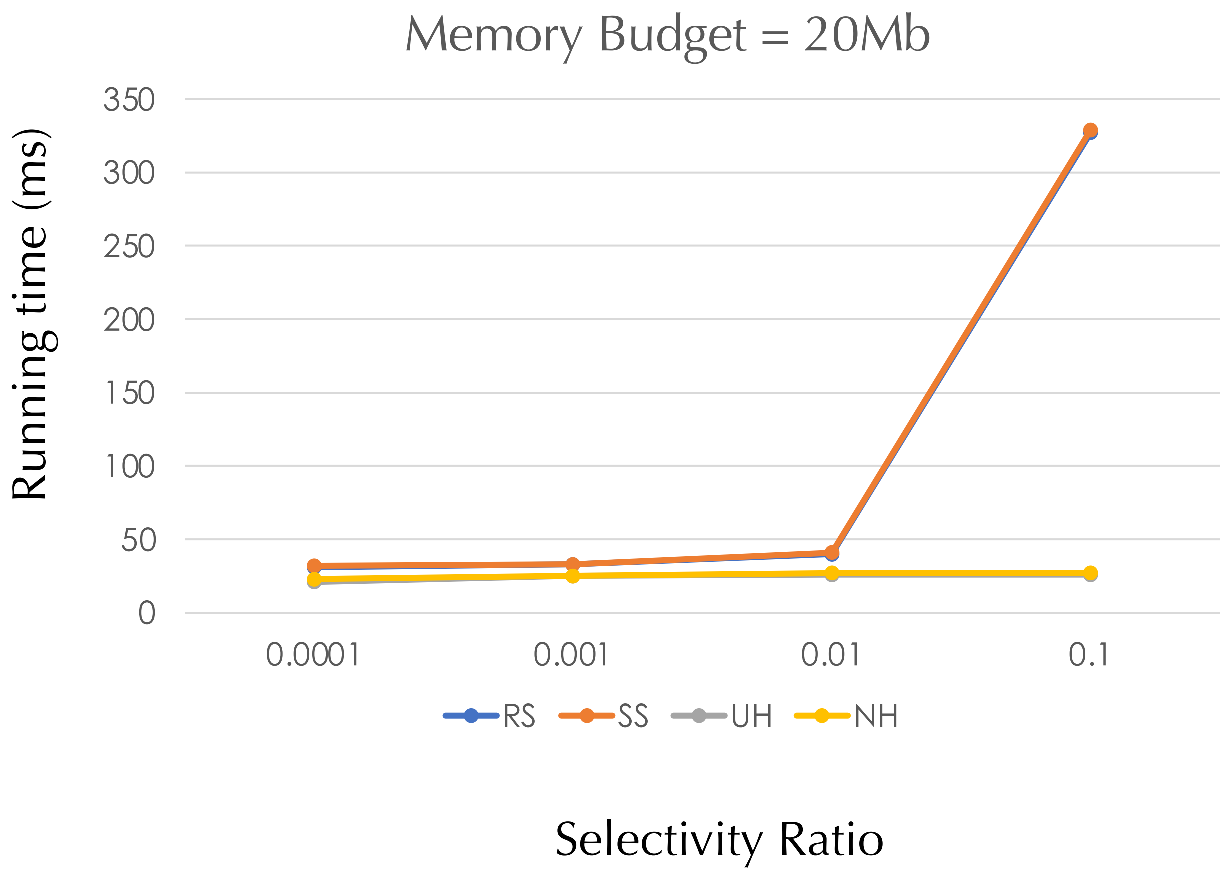


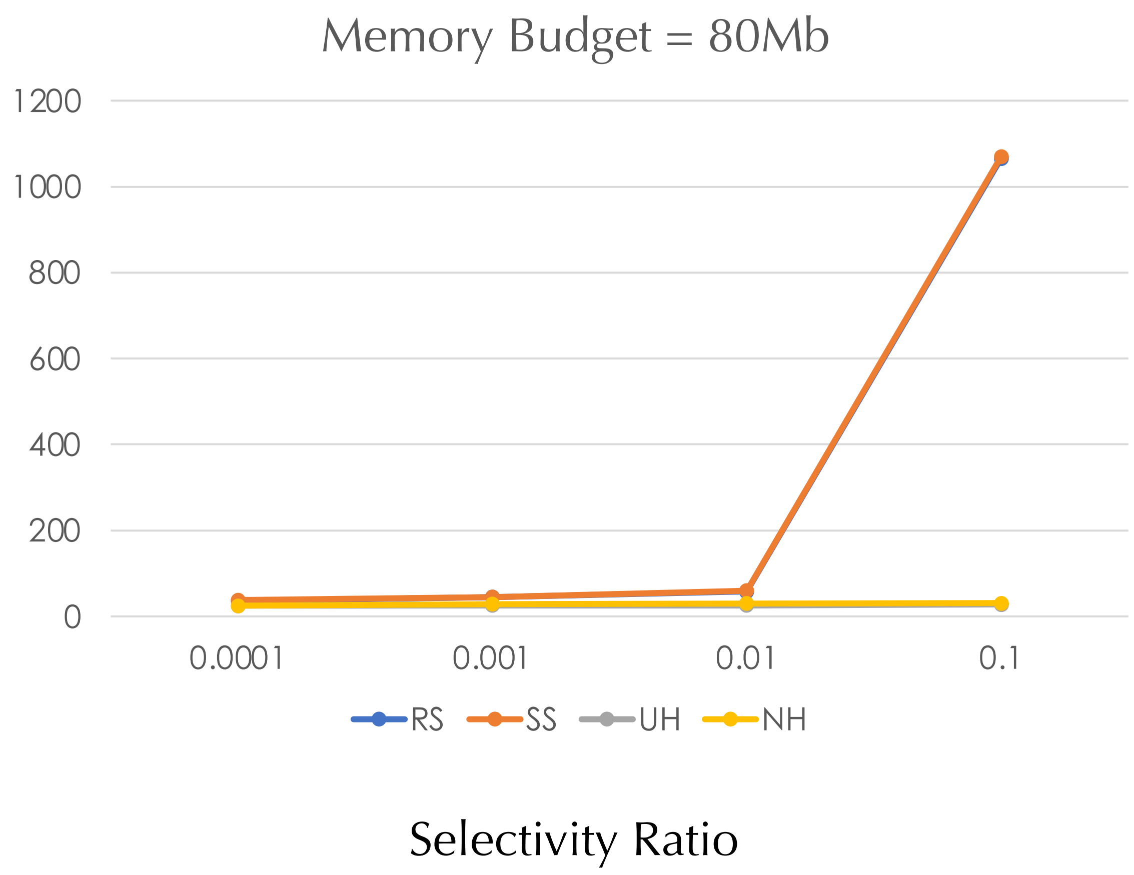


Edges dataset

Selectivity Estimation: Performance Measure

Query Response Time



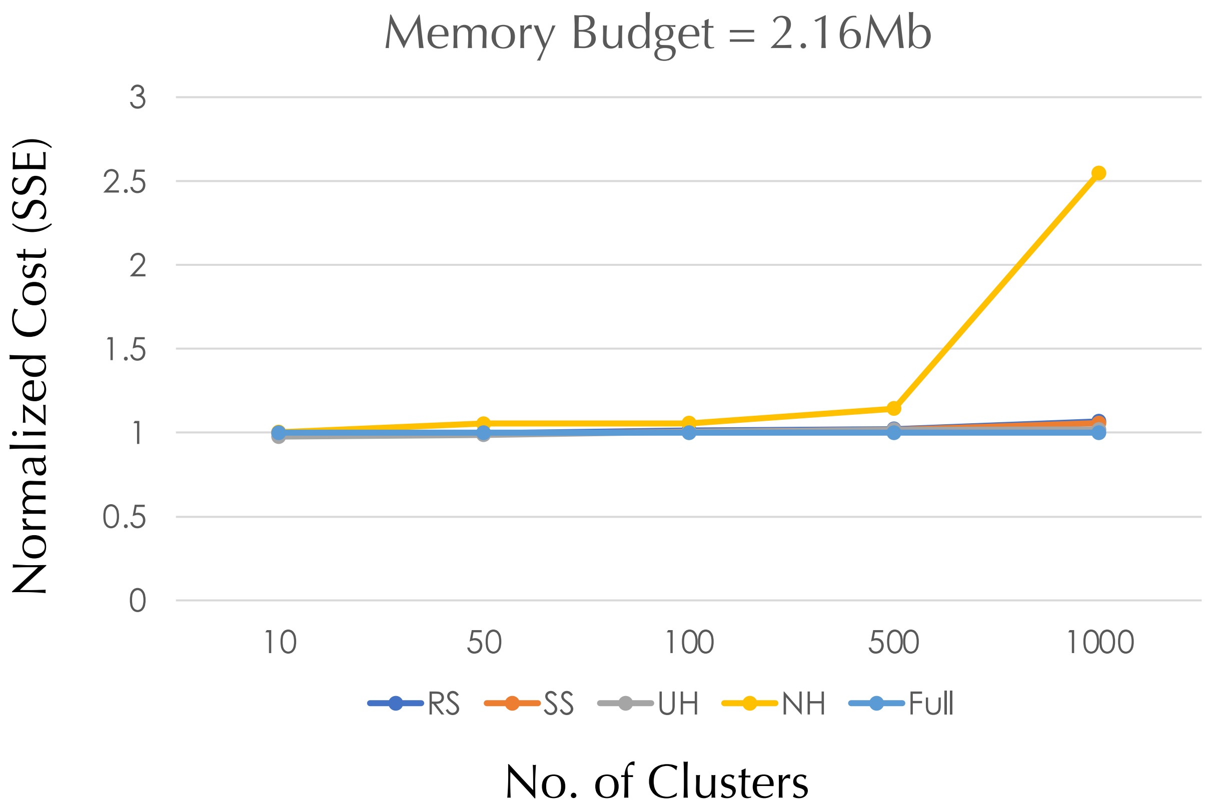


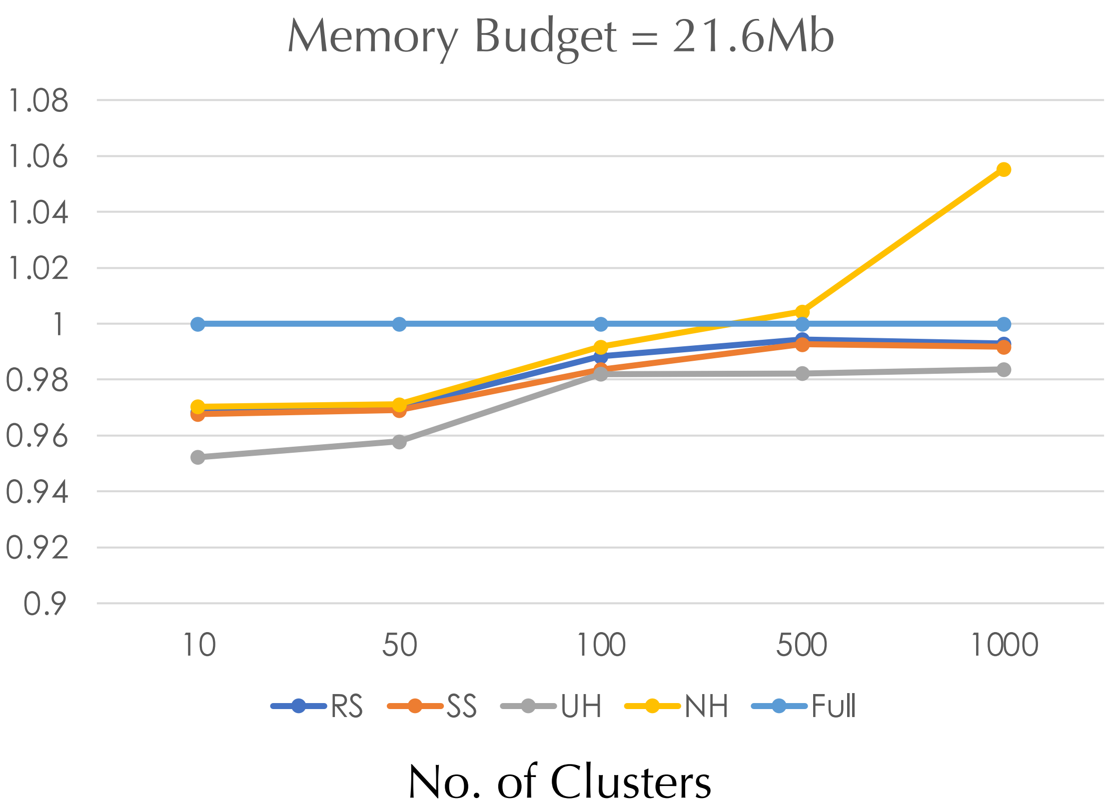
Synthetic dataset

Full Dataset >55 seconds

K-Means Clustering: Quality Measure

Sum of Squared Error

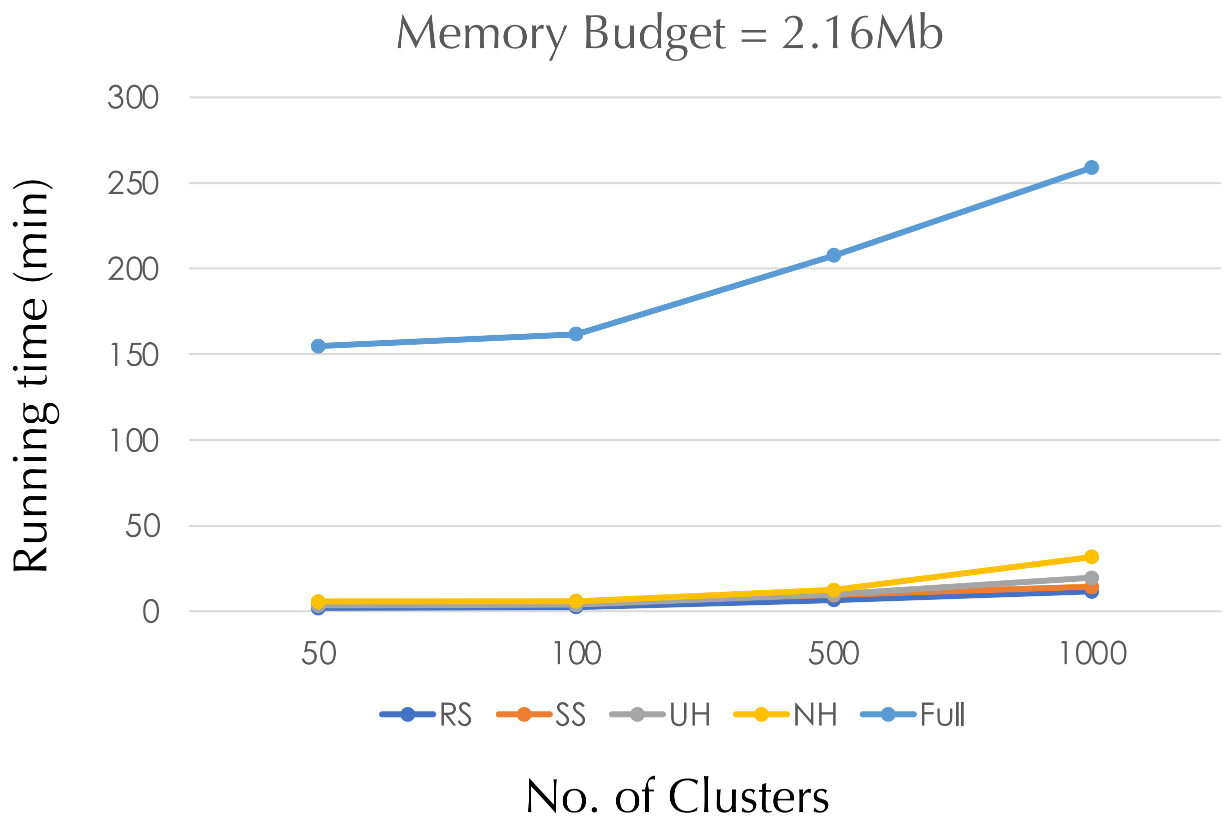


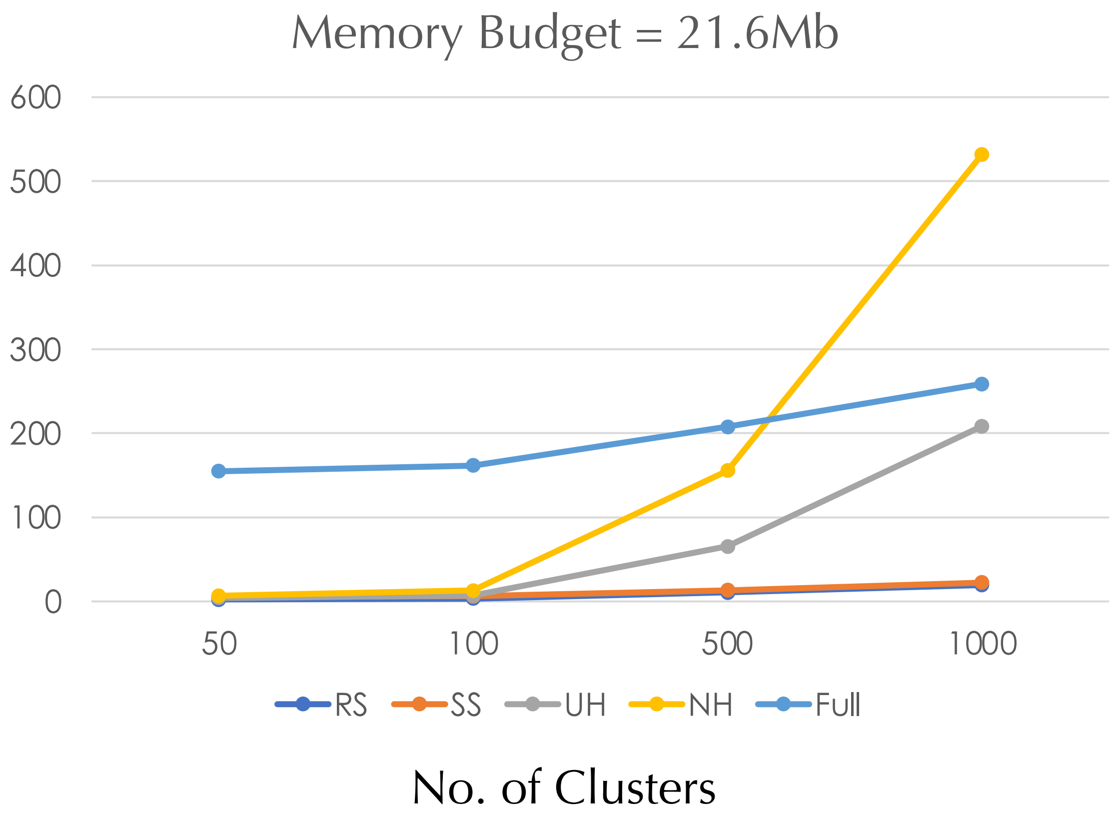


All-nodes dataset

K-Means Clustering: Performance Measure

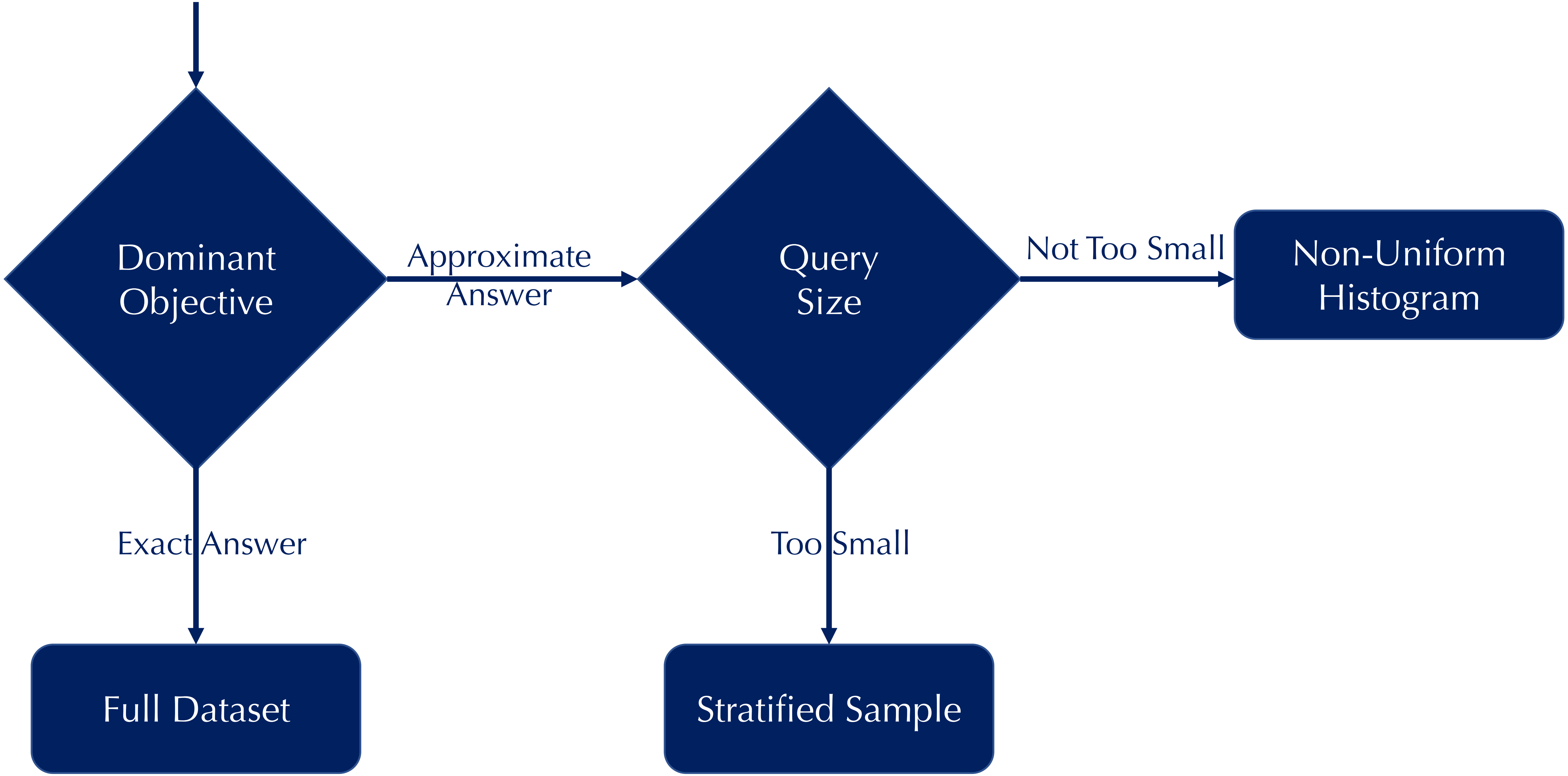
Clustering Time



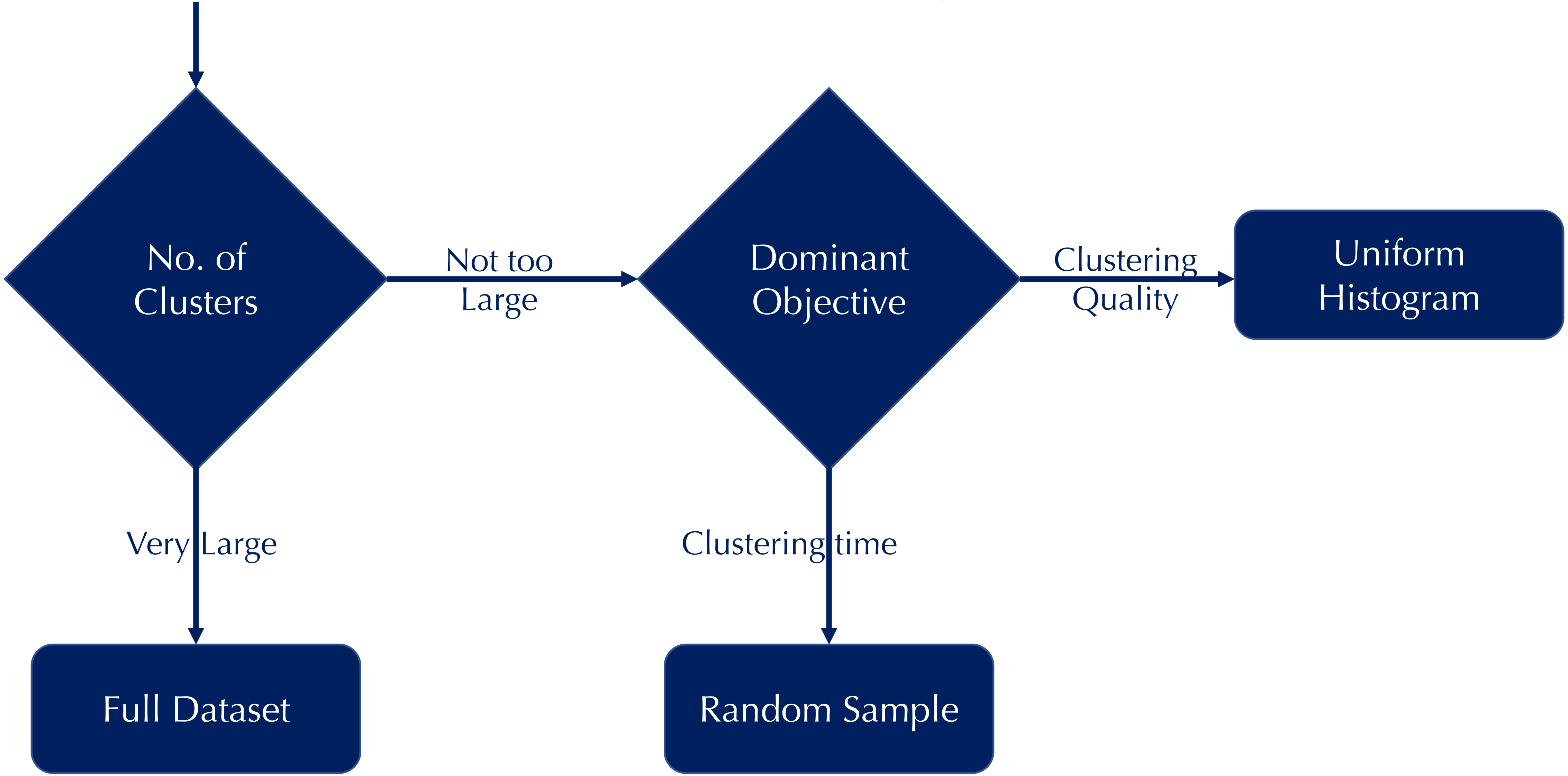


All-nodes dataset

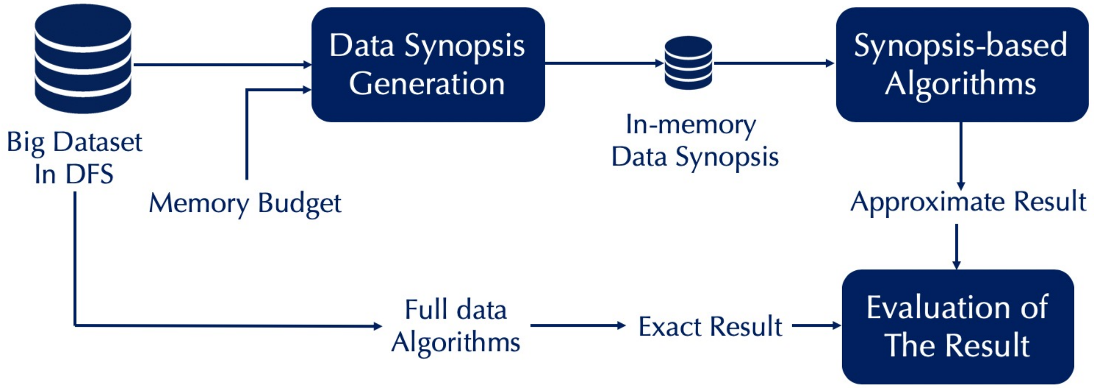
Guidelines: Selectivity Estimation

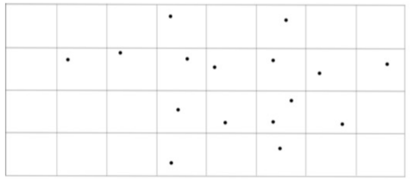


Guidelines: K-Means Clustering

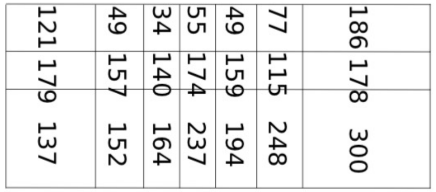


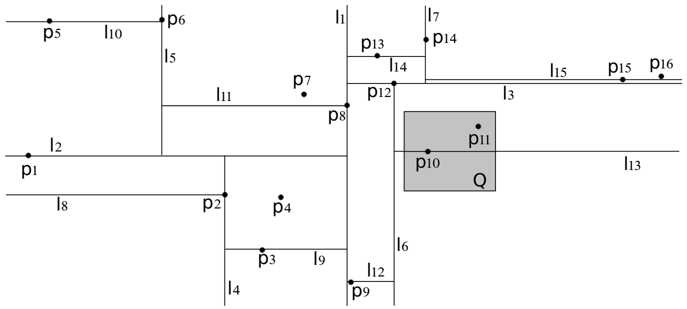
Summary

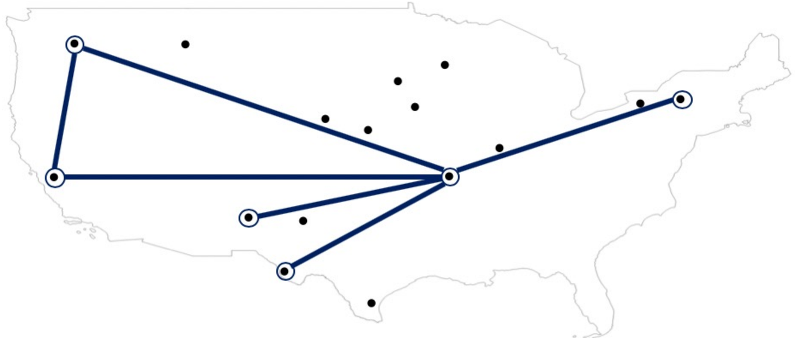


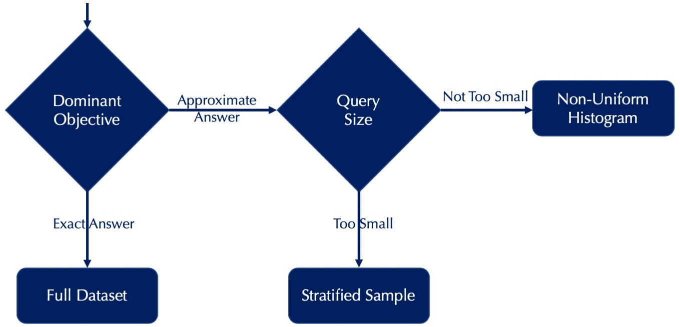


63	64	51	121	130	65	12	39
58	46	74	184	287	355	301	49
11	16	44	192	268	374	130	0
0	0	2	65	41	46	17	0





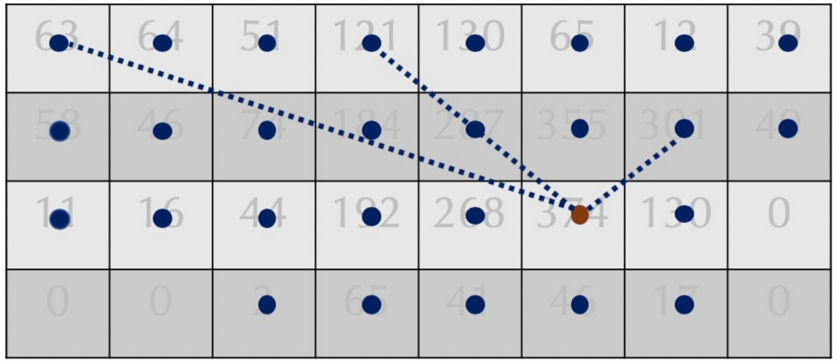


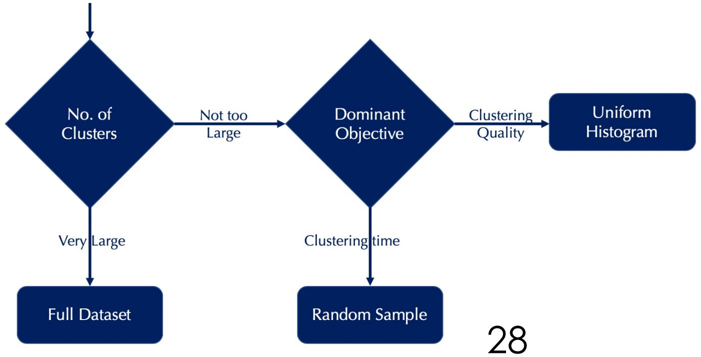


Horizontal Prefix-sum

Vertical Prefix-sum

63	127	178	299	429	494	506	545
121	231	356	661	1078	1498	1811	1899
132	258	427	924	1609	2403	2846	2934
132	258	429	991	1717	2557	3017	3105

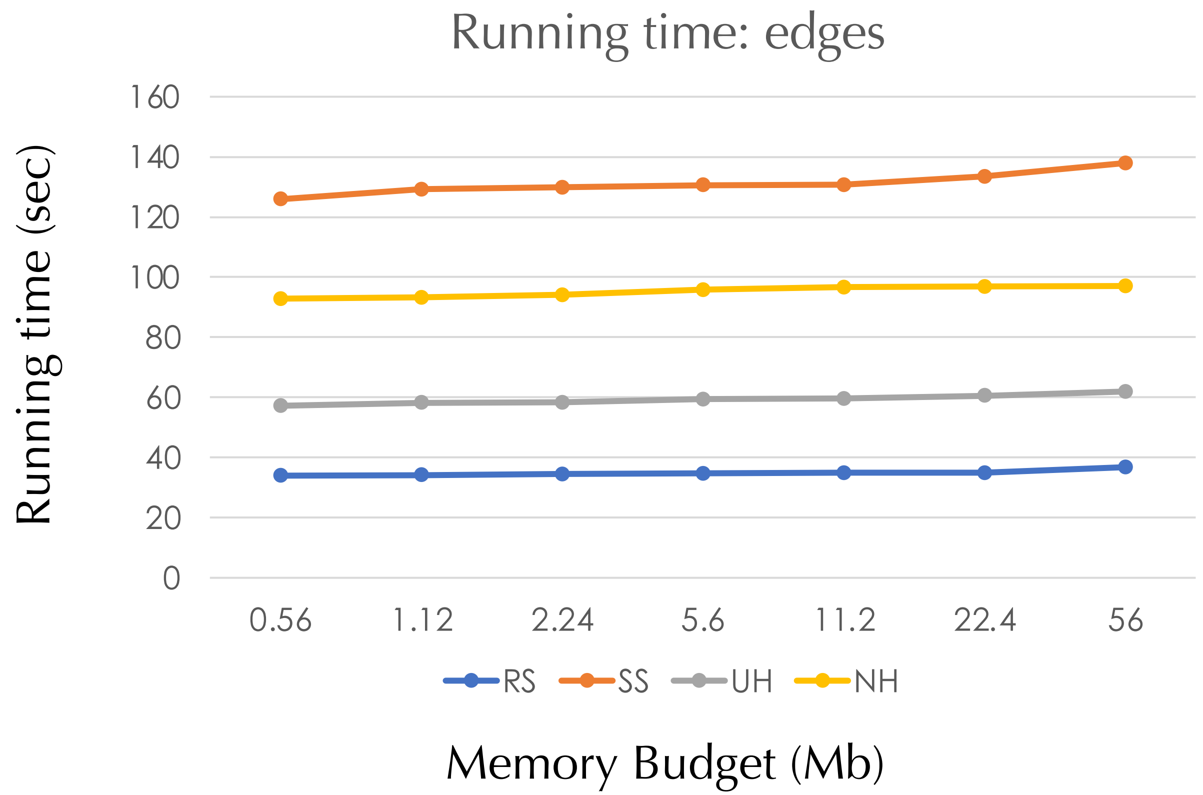


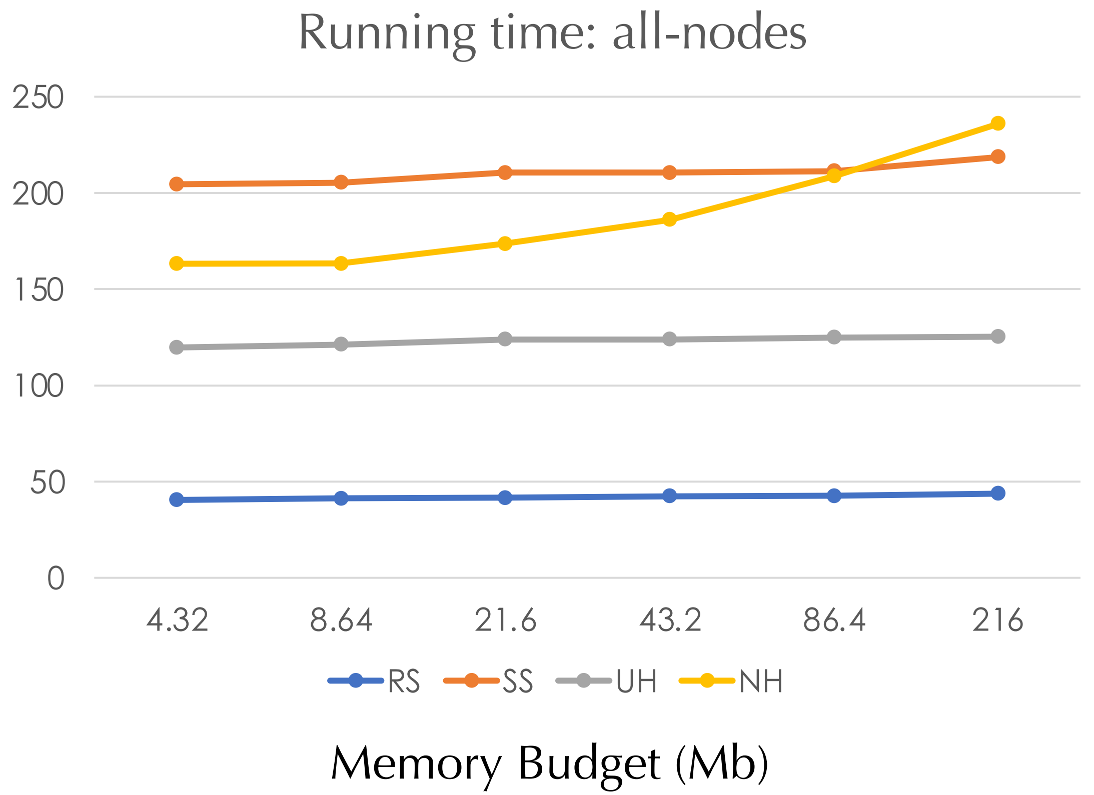


Acknowledgments

This work is supported in part by the National Science Foundation (NSF) under grants IIS-1838222, IIS-1619463, and IIS-1447826.

Synopses Performance



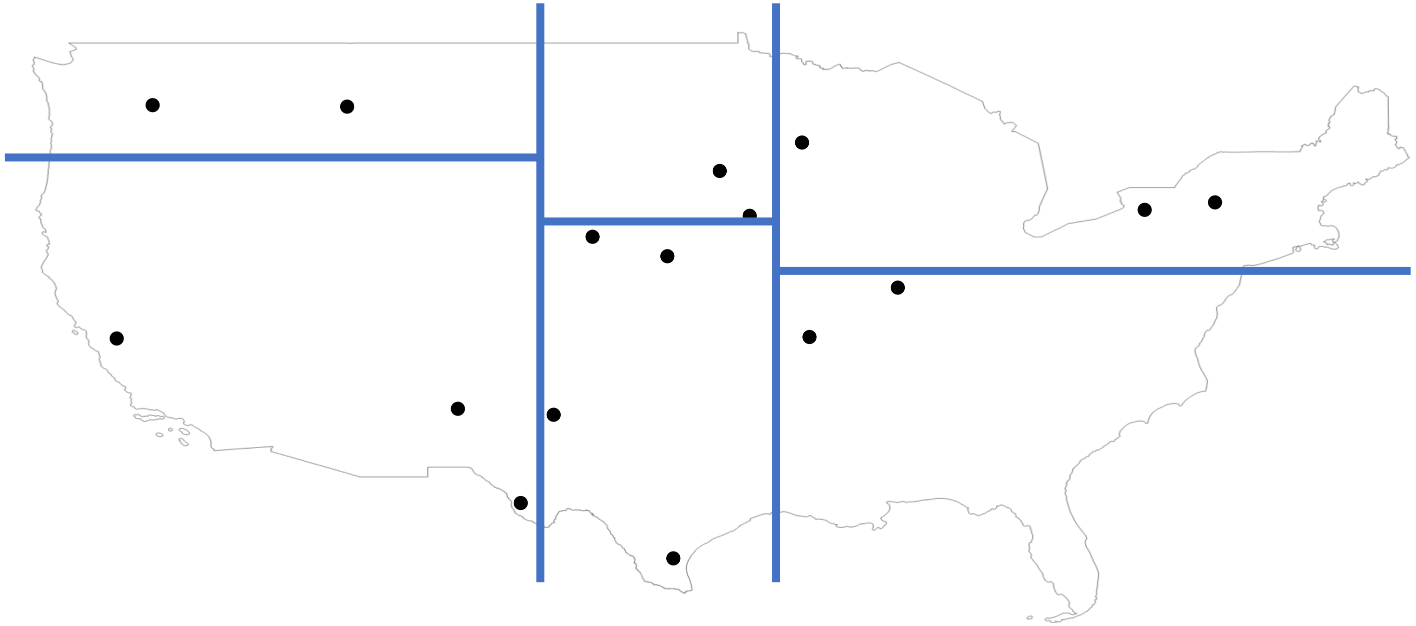


SPATIAL PARTITIONING (SP)

- Partitions into smaller-sized subsets
  - Balancing the sizes of the partitions
  - Maintaining the spatial locality
- **Inputs:** Synopsis, Number of partitions
- **Output:** A set of MBRs

SP - Based on Sample

- **Inputs:** Synopsis, Number of partitions
- **Output:** A set of MBRs



SP - Based on Histogram

- **Inputs:** Synopsis, Number of partitions
- **Output:** A set of MBRs

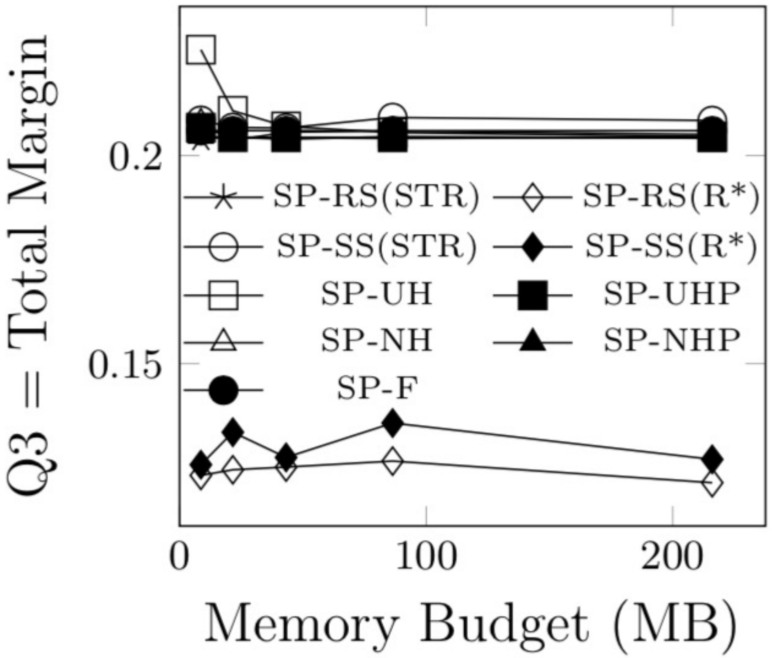
63	64	51	121	130	65	12	39
58	46	74	184	287	355	301	49
11	16	44	192	268	374	130	0
0	0	2	65	41	46	17	0

SP - Based on Histogram Partial cells

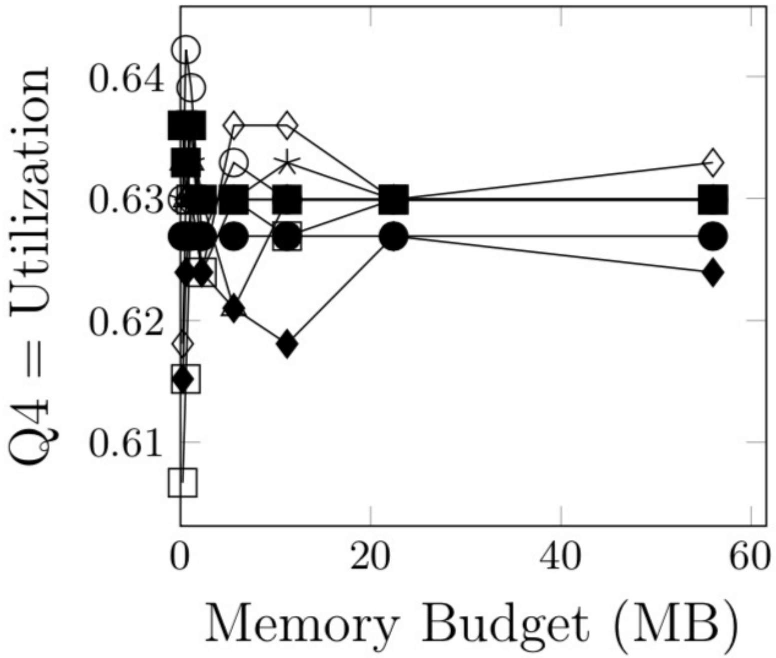
- **Inputs:** Synopsis, Number of partitions
- **Output:** A set of MBRs

63	64	51	121	130	65	12	39
58	46	74	184	287	355	301	49
11	16	44	192	268	374	130	0
0	0	2	65	41	46	17	0

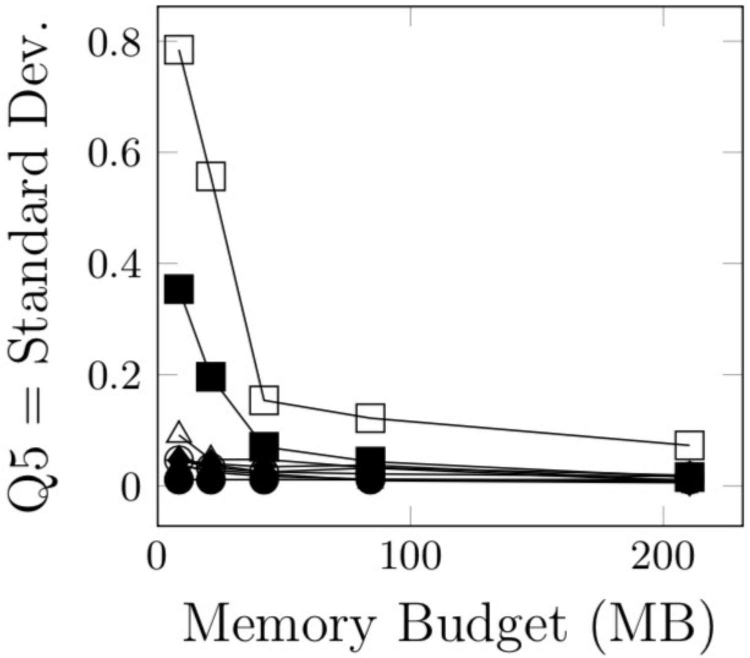
SP: Quality Measures



Q3 for all-nodes

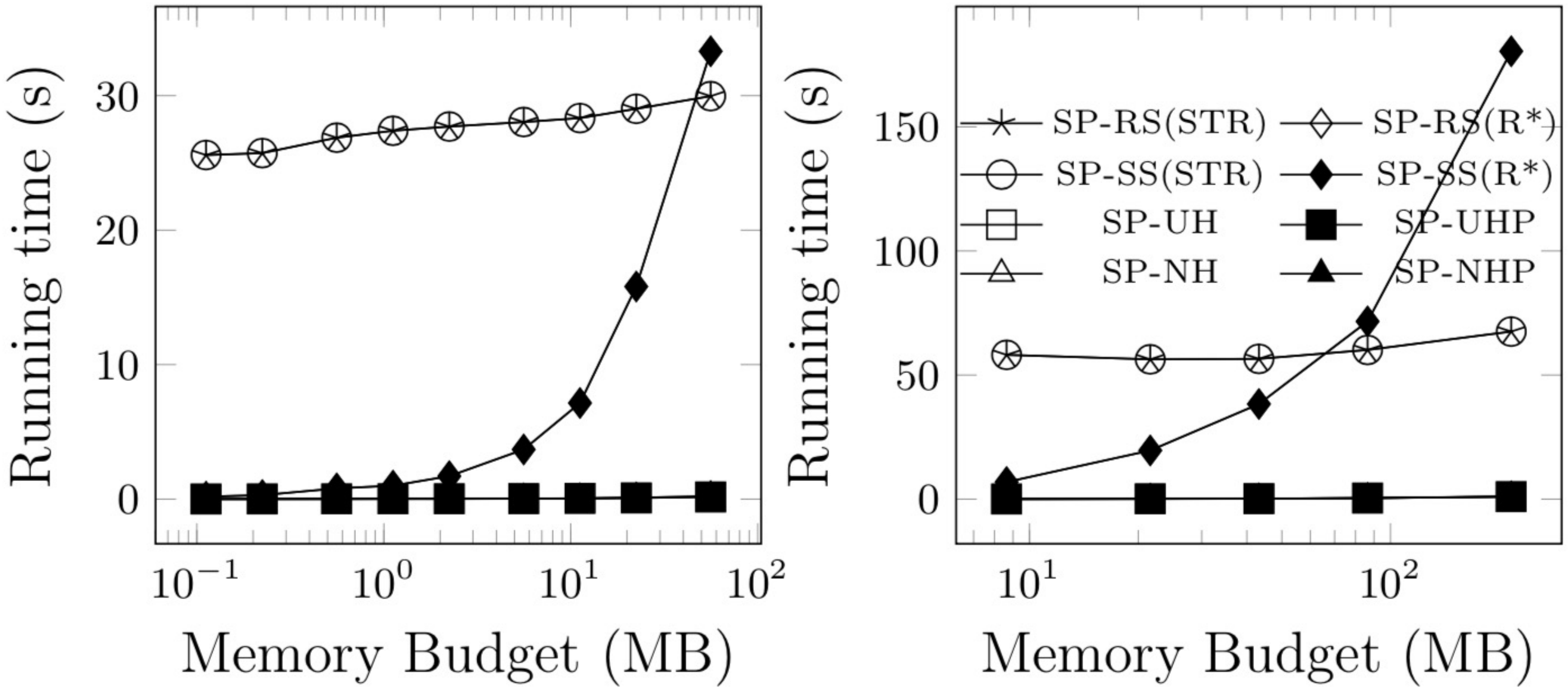


Q4 for edges



Q5 for all-objects

SP: Performance Measures



(a) SP-F $= 2275s$ for `edges` (b) SP-F $= 2993s$ for `all-nodes`

Summary of Results and Guidelines: SP