Spatio-temporal Range Searching Over Compressed Kinetic Sensor Data

Sorelle A. Friedler
Google
Joint work with David M. Mount

Motivation

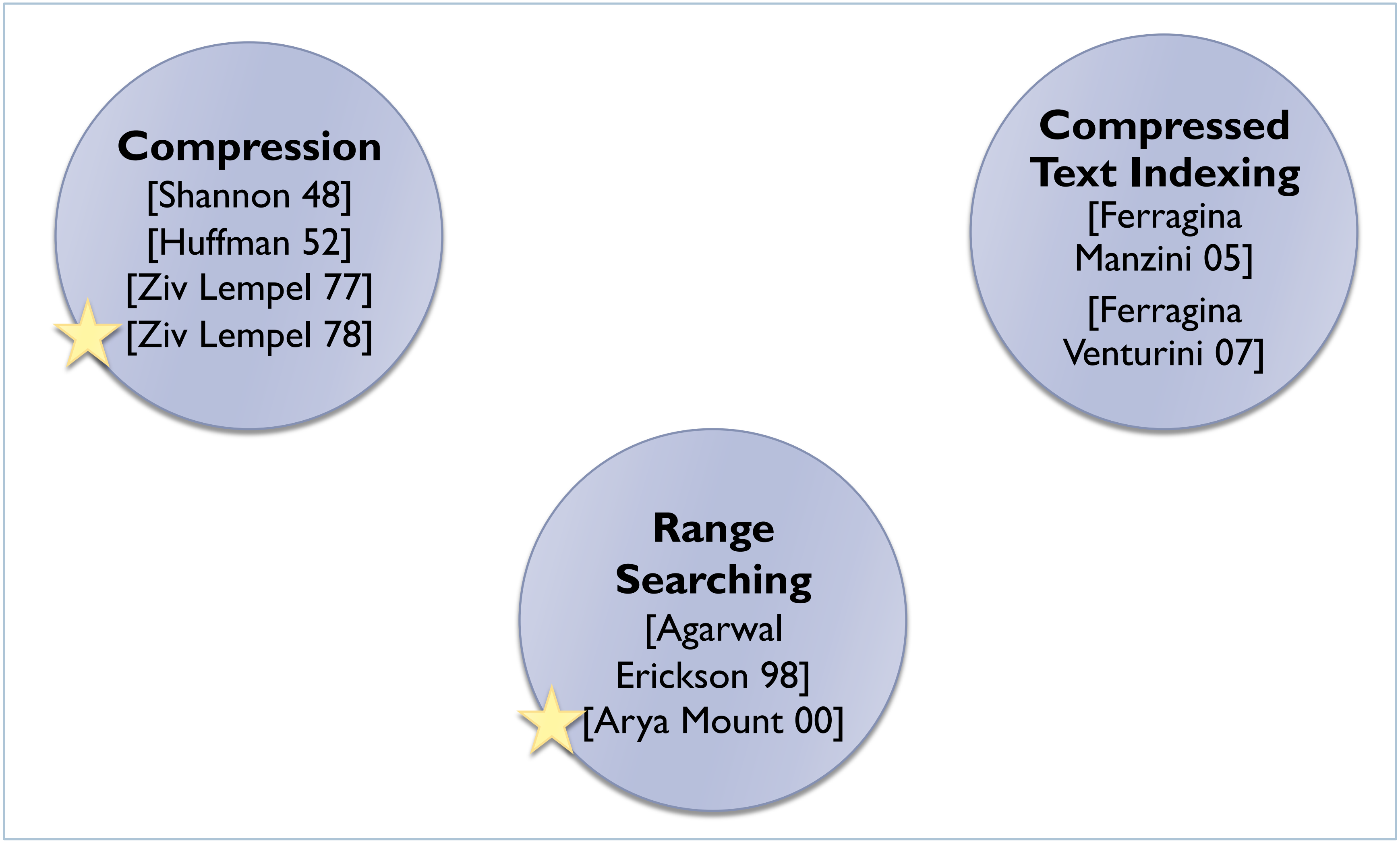
- Kinetic data: data generated by moving objects
- Sensors collect data
- Large amounts of data
- Collect and perform lossless compression

- Goal: Retrieve without decompressing
- Long Term: Analyze

Motivation

- Computer Science
  - Graphics: Image and video segmentation, animation
  - Databases: Maintenance over time
  - Sensor Networks: Data analysis
  - Cell phone users: Motion data analysis
    - 4.6 billion subscribers worldwide (in 2009)
    - 4.1 billion text messages per day in the US (in 2009)
- Biology
  - Mathematical ecology: Migratory paths, invasive species
  - Genomic data analysis: HIV strain analysis
- Engineering
  - Traffic patterns and identification

Related Work



Compression

[Shannon 48]
[Huffman 52]
[Ziv Lempel 77]
[Ziv Lempel 78]

Compressed Text Indexing

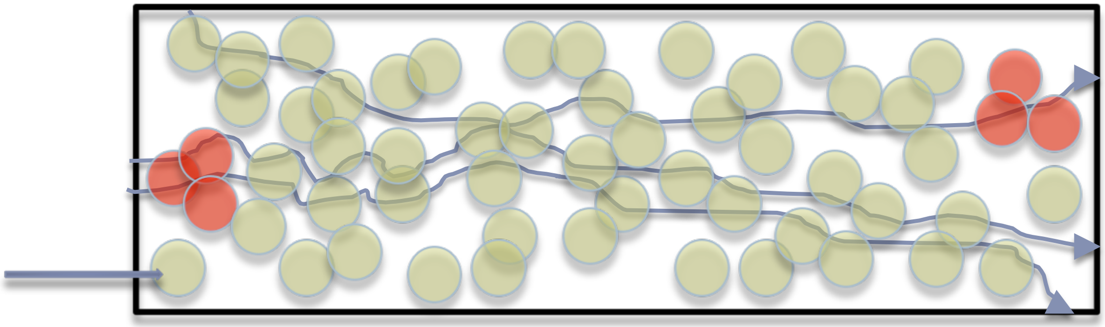
[Ferragina Manzini 05]
[Ferragina Venturini 07]

Range Searching

[Agarwal Erickson 98]
[Arya Mount 00]

Our Framework (Friedler Mount 09)

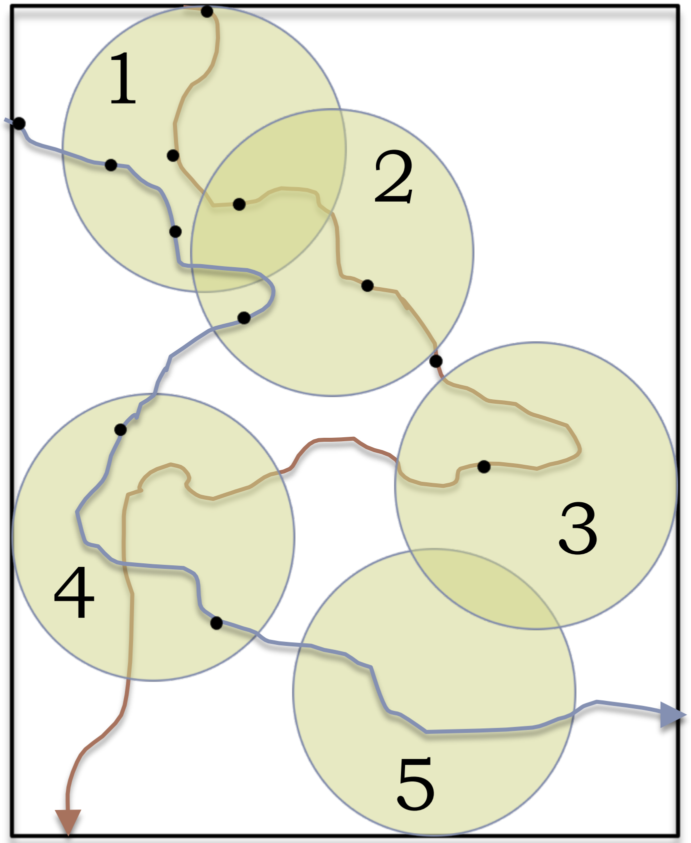
- Detection region around each sensor (stationary sensors)
- Point motion unrestricted
- No advance knowledge about motion
- Each sensor reports the count of points within its region at each synchronized time step
- *k*-local: Sensor outputs statistically dependent only on *k* nearest neighbors



sensor regions

Data Collection

Data based on underlying geometric motion



Sensor data streams

x_1	x_2	x_3	x_4	x_5
1	0	0	0	0
2	0	0	0	0
2	1	0	0	0
0	2	0	0	0
0	0	0	1	0
0	0	1	1	0

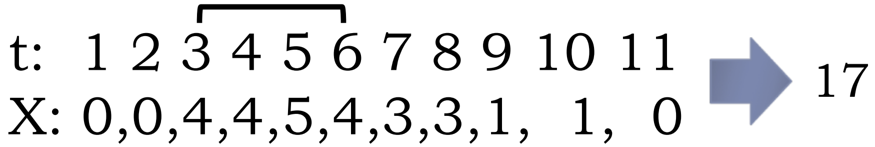
time ↓

Range Searching: Our Problem

Compress and preprocess the data so as to perform…

- Temporal range query: Given a time interval, return an aggregation of the counts over that time interval.

aggregation type: sum

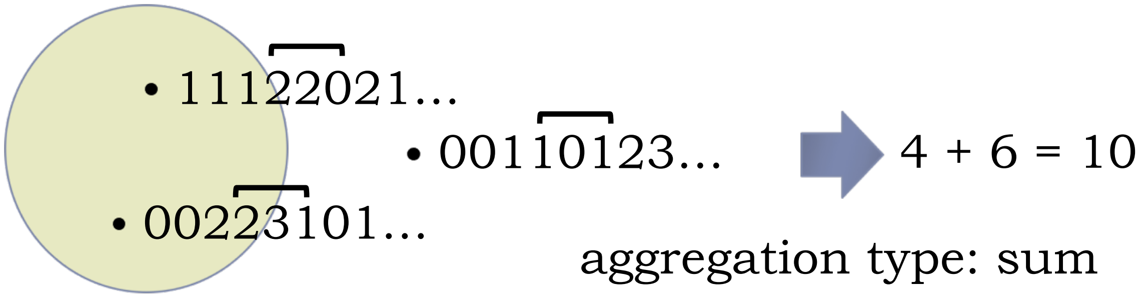


t: 1 2 3 4 5 6 7 8 9 10 11

X: 0,0,4,4,5,4,3,3,1, 1, 0

17

- Spatio-temporal range query: Given a time interval and spherical spatial region, return an aggregation of the counts over that time interval and within that region.



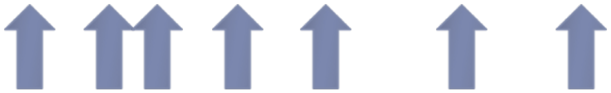
- 11122021…
- 00110123…
- 00223101…

4 + 6 = 10

aggregation type: sum

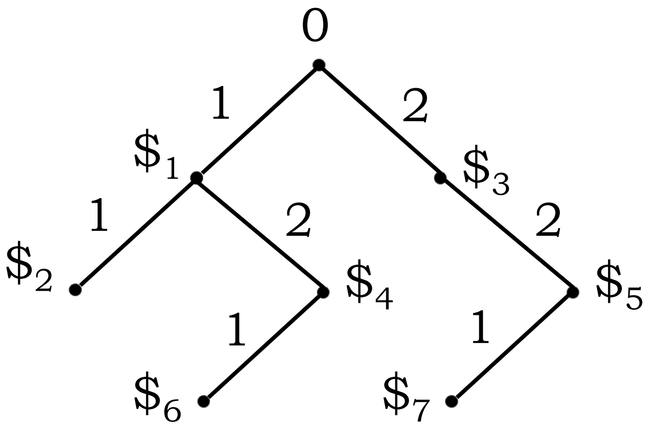
Lempel-Ziv Dictionary Compression [LZ78]

11121222121221



1 11 2 12 22 121 221

$\$_1$ $\$_2$ $\$_3$ $\$_4$ $\$_5$ $\$_6$ $\$_7$



Create a trie while scanning through a string. The compressed string contains pointers to this dictionary.

(LZ78 is an optimal entropy encoding algorithm.)

Temporal Range Searching

- Create trie with accompanying pointers
- Annotate trie with aggregate values and word start times
- Given a temporal range $[t_0, t_1]$ find the anchor points $\$^0$ and $\$^1$ such that $\$^0 \leq t_0$ and $\$^1 \geq t_1$ (binary search)
- Use stored prefixes, words, and subtraction of prefixes to find aggregates

Query Examples

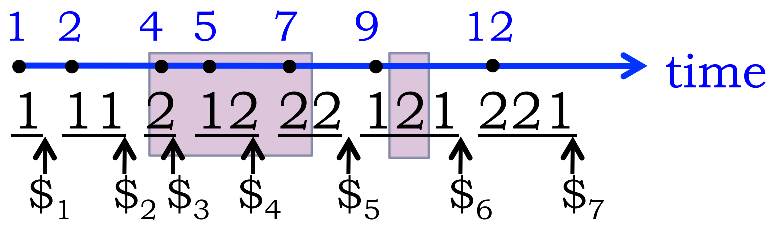
overlapping query: [4,7]

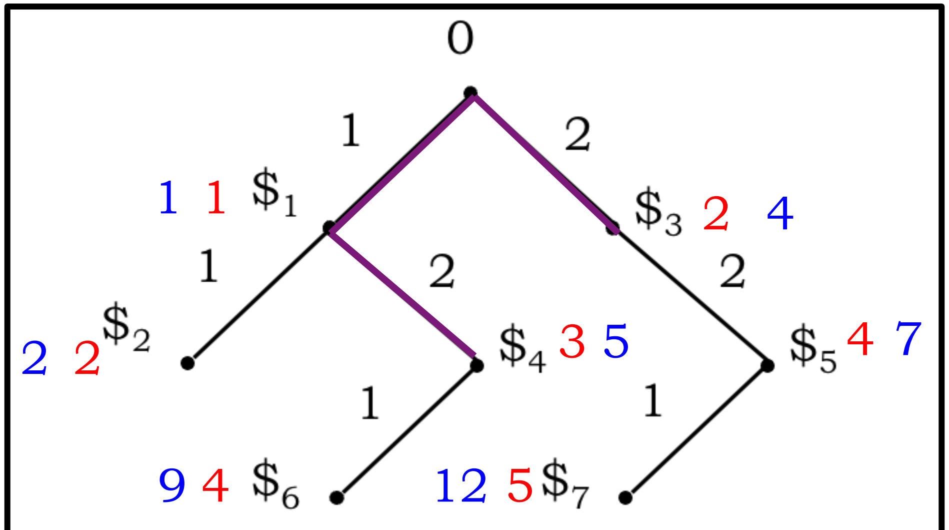
2 + 3 + 2 = 7

internal query: [10,10]

3 - 1 = 2

Encoded String

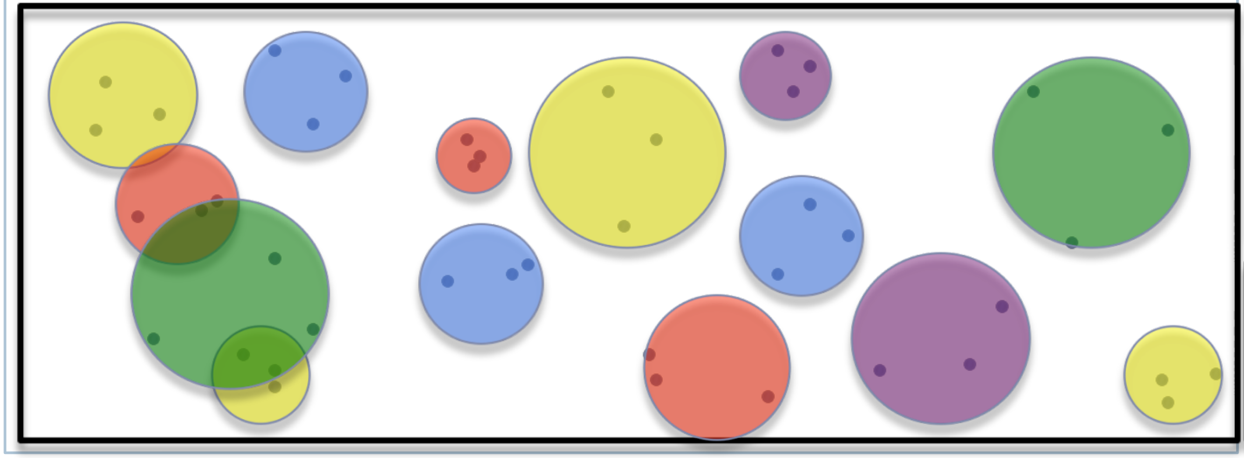


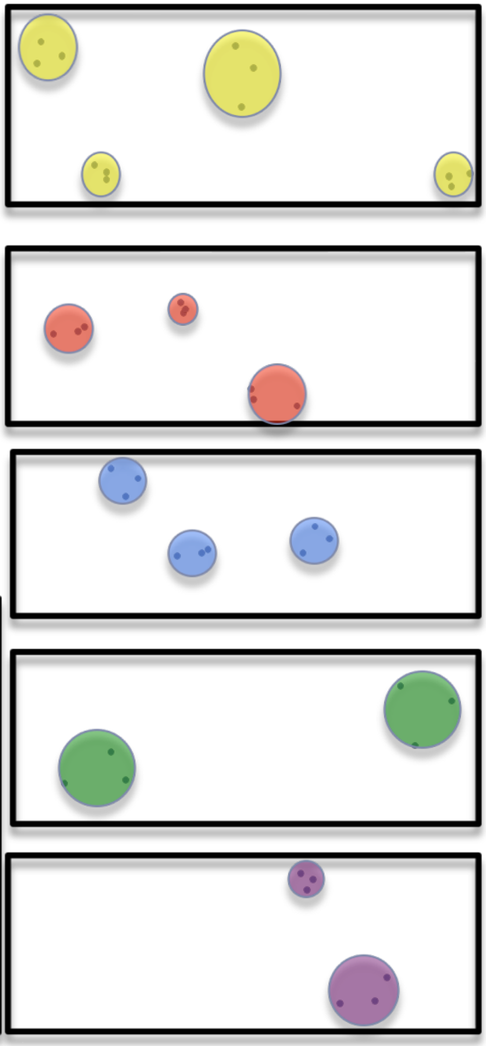


aggregate values word start times

Data Compression Algorithm: Partitioning Lemma (Friedler Mount 09)

- Lemma: There exists an integral constant *c* such that for all $k>0$ any point set can be partitioned into *c* partitions that are each *k*-clusterable.
  - $c = O(1 + 12^{O(1)})$ dependent on dimension

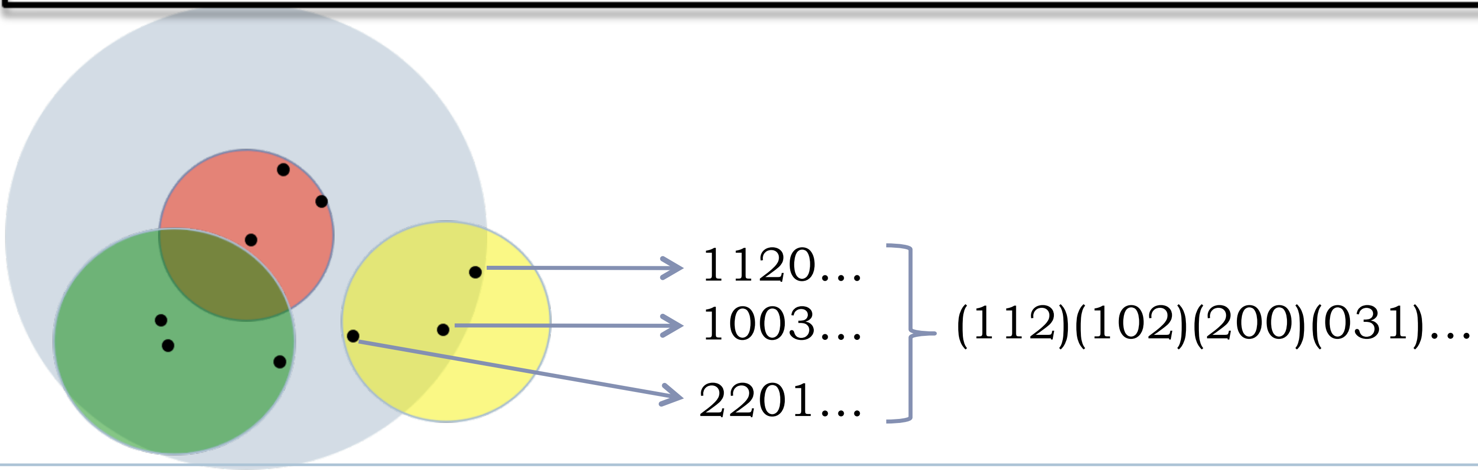




Data Compression Algorithm (Friedler Mount 09)

- Partition and cluster the sensors, then compress

```
for each partition $P_i$
    for each cluster in $P_i$
        combine the cluster's streams into one
        with longer characters and compress it
return the union of the compressed streams
```



1120...

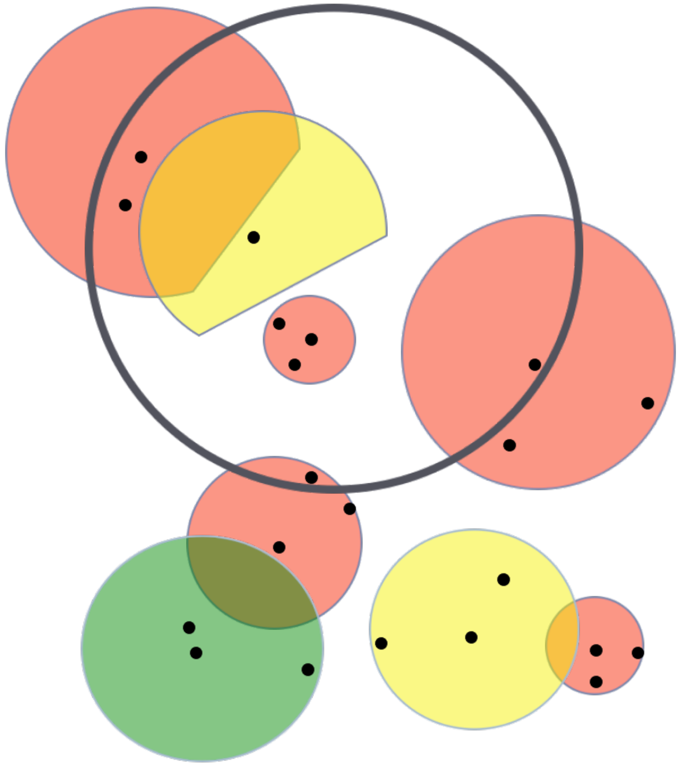
1003...

2201...

(112)(102)(200)(031)...

Sensor Clumps

- Recall: The sensors are partitioned, clustered, and compressed
- *Set of clumps*: A finite set of balls with a packing property limiting the number of intersections of any ball with a clump.
- Lemma: In a single partition, the nearest neighbor balls form a set of clumps that contain the sensor clusters

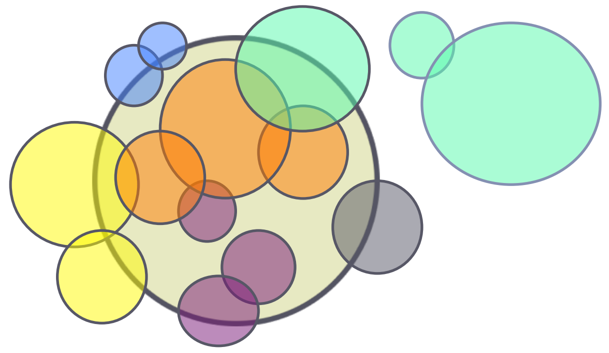


Range Searching Among Clumps

<u>Range Searching Among Clumps:</u>

Given any query range $\mathcal{R}$ report

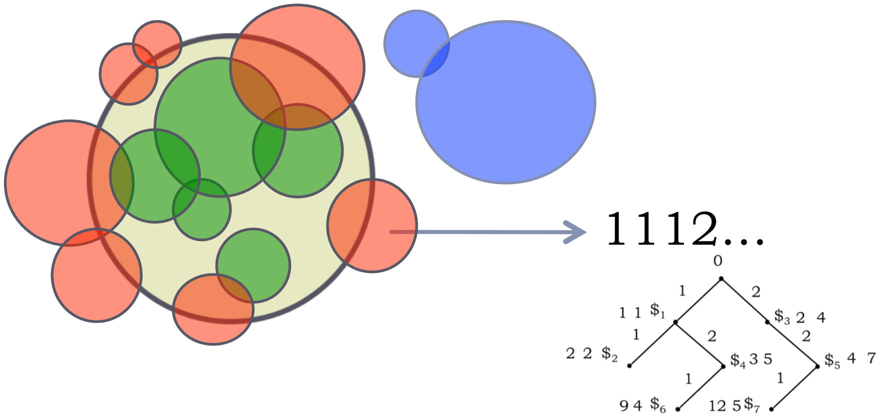
- a subset of clump subsets that form a disjoint cover of the clumps within $\mathcal{R}$
- the subset of clumps that $\mathcal{R}$ intersects



- <u>Lemma</u>: A quadtree variant based data structure can answer range searching queries among clumps.

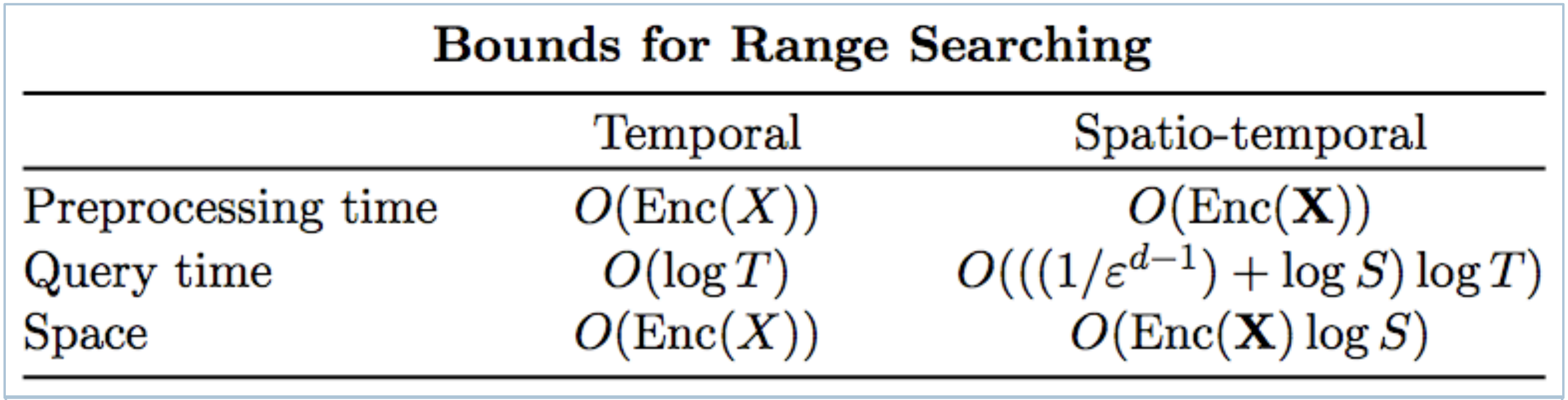
Spatio-temporal Range Searching

- Main Theorem: By adding an auxiliary data structure to answer temporal range queries to each node in the range searching among clumps solution we can answer spatio-temporal range queries.



- One range searching among clumps structure for each partition
- One temporal range structure for each clump and each internal quadtree node
- Get temporal sums for each clump and overlapped sensor
- Sum over all partitions

Results



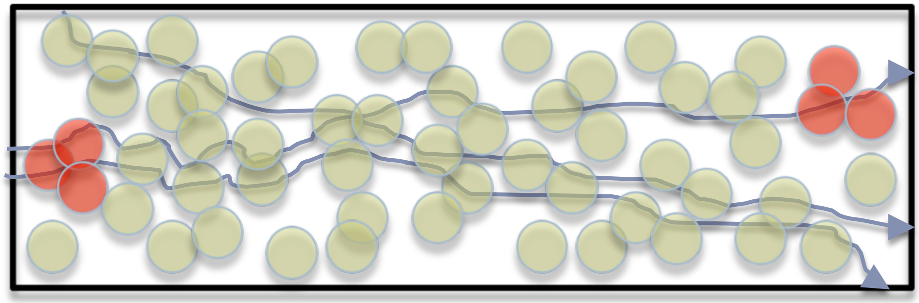
Bounds for Range Searching

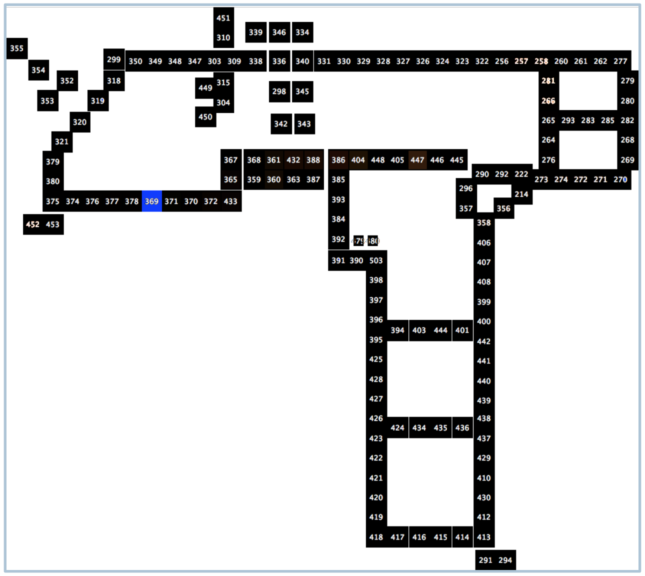
	Temporal	Spatio-temporal
Preprocessing time	$O(\mathrm{Enc}(X))$	$O(\mathrm{Enc}(\mathbf{X}))$
Query time	$O(\log T)$	$O(((1/\varepsilon^{d-1}) + \log S) \log T)$
Space	$O(\mathrm{Enc}(X))$	$O(\mathrm{Enc}(\mathbf{X}) \log S)$

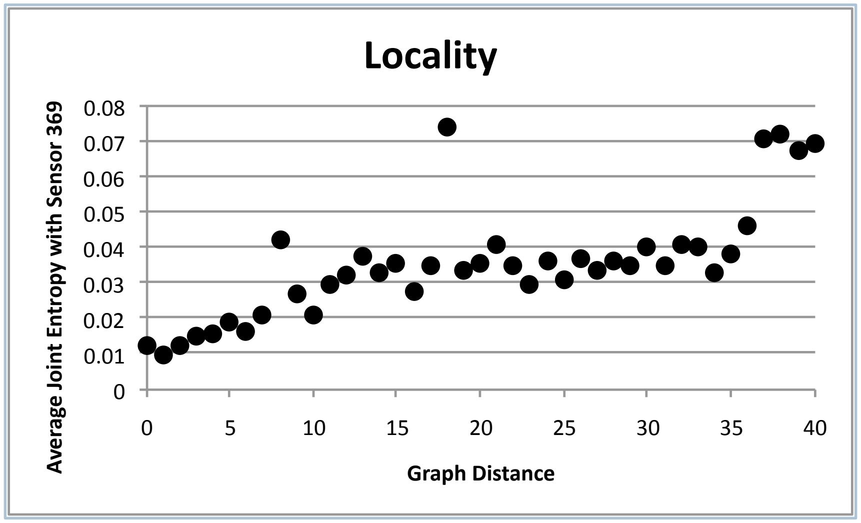
- X: The set of sensor system observations
- Enc(X): The encoded size (in bits) of the compressed data
- T: The total time over which data was collected
- S: The total number of sensors
- d: The dimension of the sensor space
- ε: An error parameter (for approximate range searching)

First range searching bounds over compressed data

Experimental Results: Locality

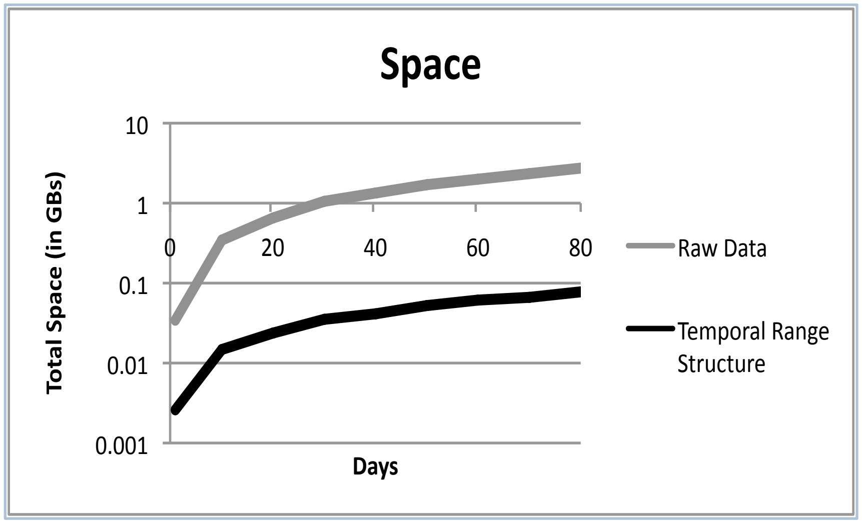


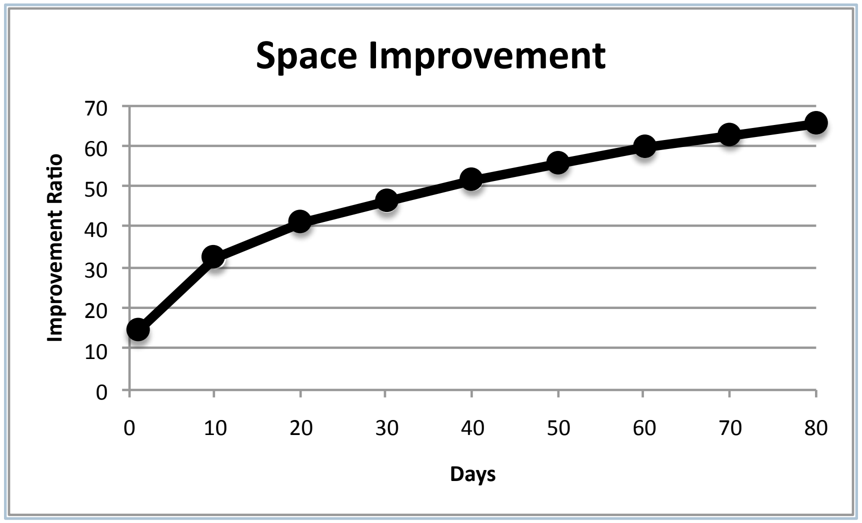




C. R. Wren, Y. A. Ivanov, D. Leigh, and J. Westbues.
The MERL motion detector dataset: 2007 workshop on massive datasets.
Technical Report TR 2007-069,
Mitsubishi Electronic Research Laboratories, Cambridge, MA, USA, August 2007.

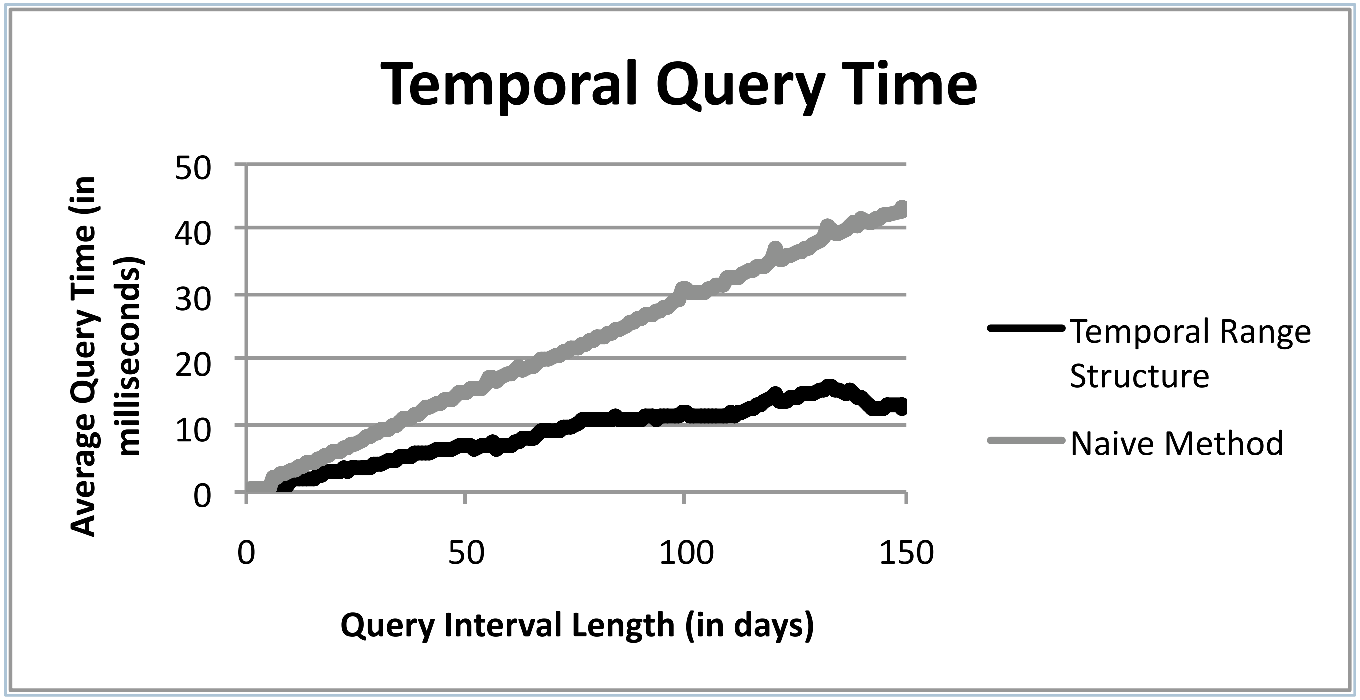
Experimental Results: Space





C. R. Wren, Y. A. Ivanov, D. Leigh, and J. Westbues.
The MERL motion detector dataset: 2007 workshop on massive datasets.
Technical Report TR 2007-069,
Mitsubishi Electronic Research Laboratories, Cambridge, MA, USA, August 2007.

Experimental Results: Time



C. R. Wren, Y. A. Ivanov, D. Leigh, and J. Westbues.
The MERL motion detector dataset: 2007 workshop on massive datasets.
Technical Report TR 2007-069,
Mitsubishi Electronic Research Laboratories, Cambridge, MA, USA, August 2007.

Open Problems

- I/O-efficiency
- Streaming Model
- Other range searching questions
  - halfspace range searching
- Statistical analysis over compressed data
  - clustering over space and time

Thank you!

Questions?