Prominent Streak Discovery in Sequence Data

Xiao Jiang[1], *Chengkai Li*[2], *Ping Luo*[3], *Min Wang*[3], *Yong Yu*[1]

([1] Shanghai Jiao Tong University; [2] The University of Texas at Arlington; [3] HP Labs China)

1. Motivation

Prominent streaks stated in news articles:

- *This month the Chinese capital has experienced 10 days with a maximum temperature in around 35 degrees Celsius – the most for the month of July in a decade.*
- *The Nikkei 225 closed below 10000 for the 12th consecutive week, the longest such streak since June 2009.*
- *He (LeBron James) scored 35 or more points in nine consecutive games and joined Michael Jordan and Kobe Bryant as the only players since 1970 to accomplish the feat.*

2. Problem Formulation

In a sequence of values, a streak **<[l, r], v>** is the triple of left-end, right-end, and the minimum value in the interval. E.g.

3 1 7 7 2 5 4 6 7 3

<[6, 8], 4>

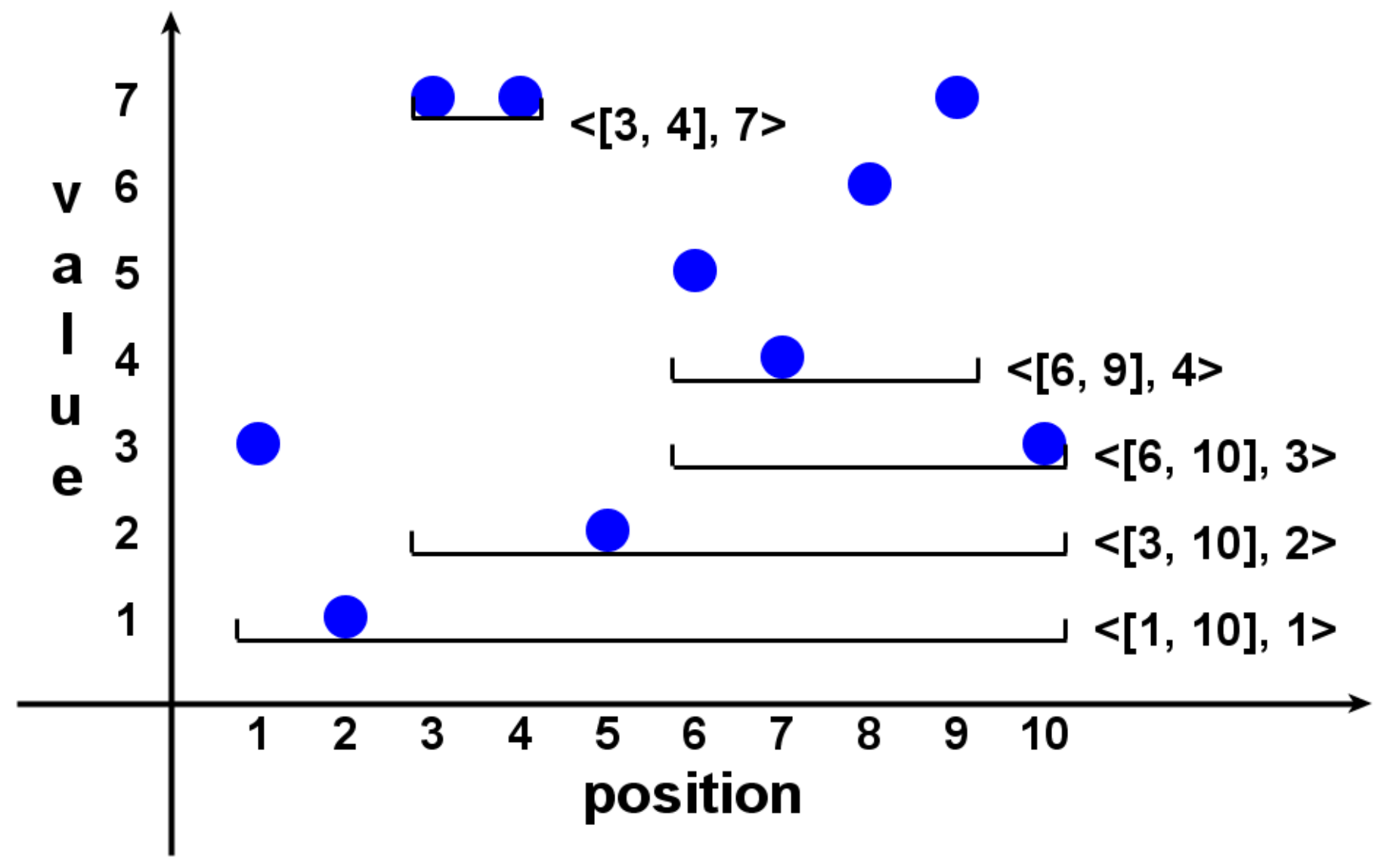
We define the **Dominance Relationship**:

Streak $s_1=\langle[l_1,r_1],v_1\rangle$ dominates streak $s_2=\langle[l_2,r_2],v_2\rangle$ iff

$$\begin{matrix} r_1-l_1 \geq r_2-l_2 \\ v_1 > v_2 \end{matrix} \quad \text{or} \quad \begin{matrix} r_1-l_1 > r_2-l_2 \\ v_1 \geq v_2 \end{matrix}$$

Prominent streaks are streaks that are not dominated by others.

Task 1: given a data sequence, compute the prominent streaks. E.g.

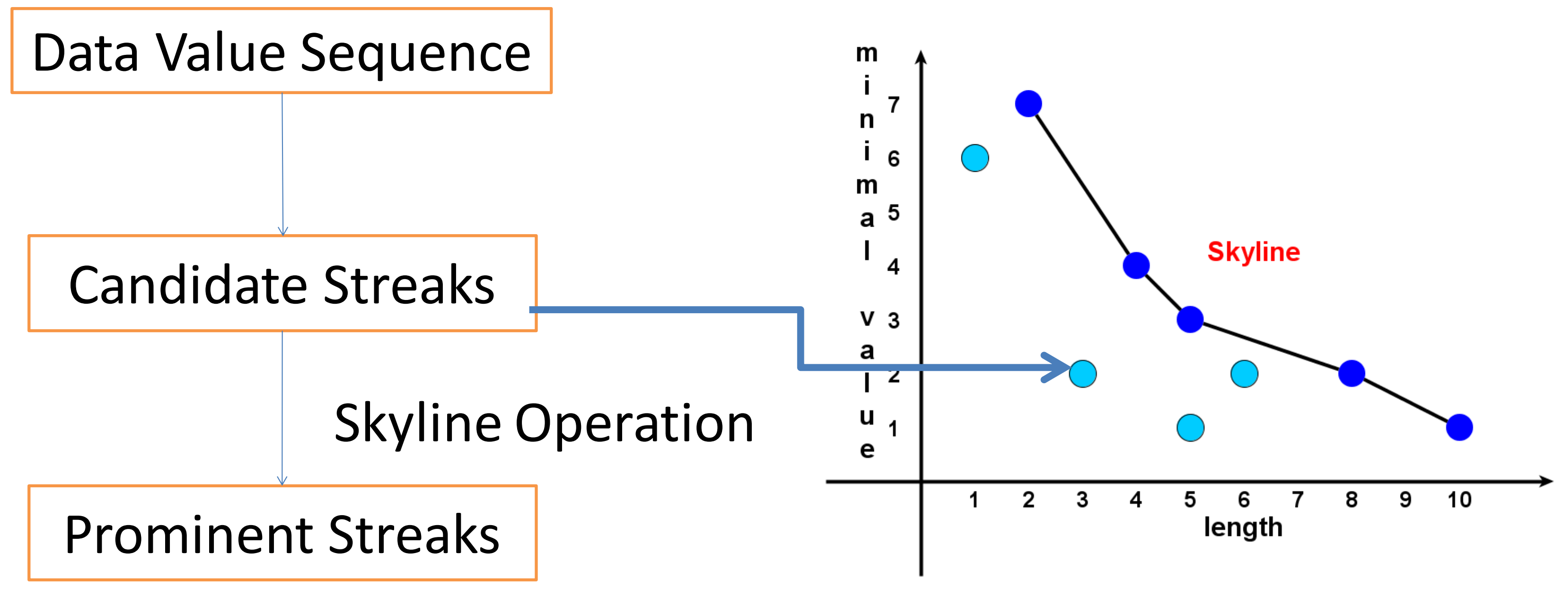


Monitoring:

Real-world data sequence often grows, with newly appended values.

Task 2: keep the prominent streaks up-to-date

3. Solution Framework



4. Local Prominent Streak

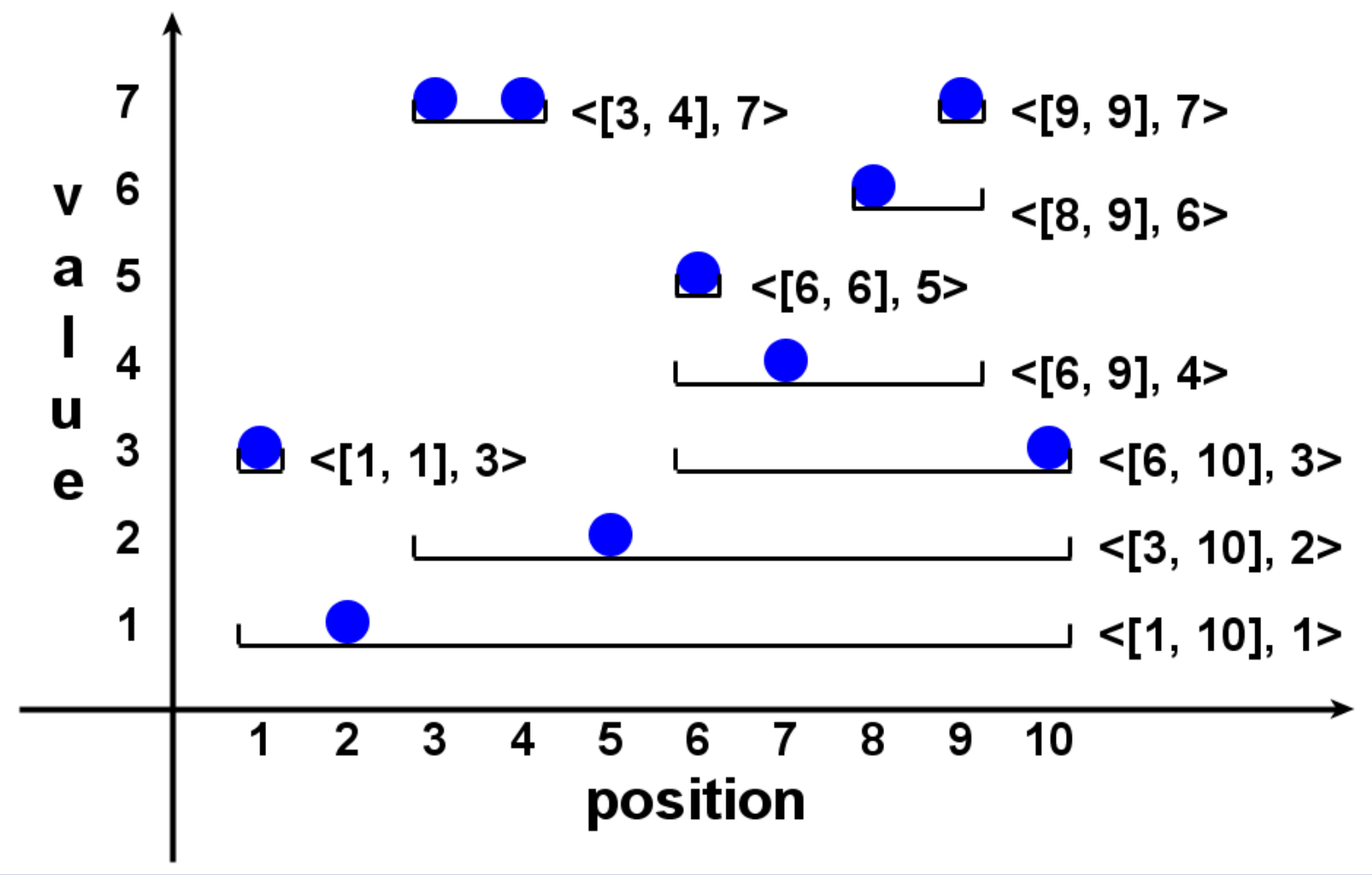
Streak $s_1=\langle[l_1,r_1],v_1\rangle$ dominates streak $s_2=\langle[l_2,r_2],v_2\rangle$ **locally** iff s1 dominates s2 and $[l_1,r_1]\supseteq[l_2,r_2]$

Local prominent streaks (LPS) are streaks that are not *locally dominated* by others.

Property 1: prominent streaks are also LPS

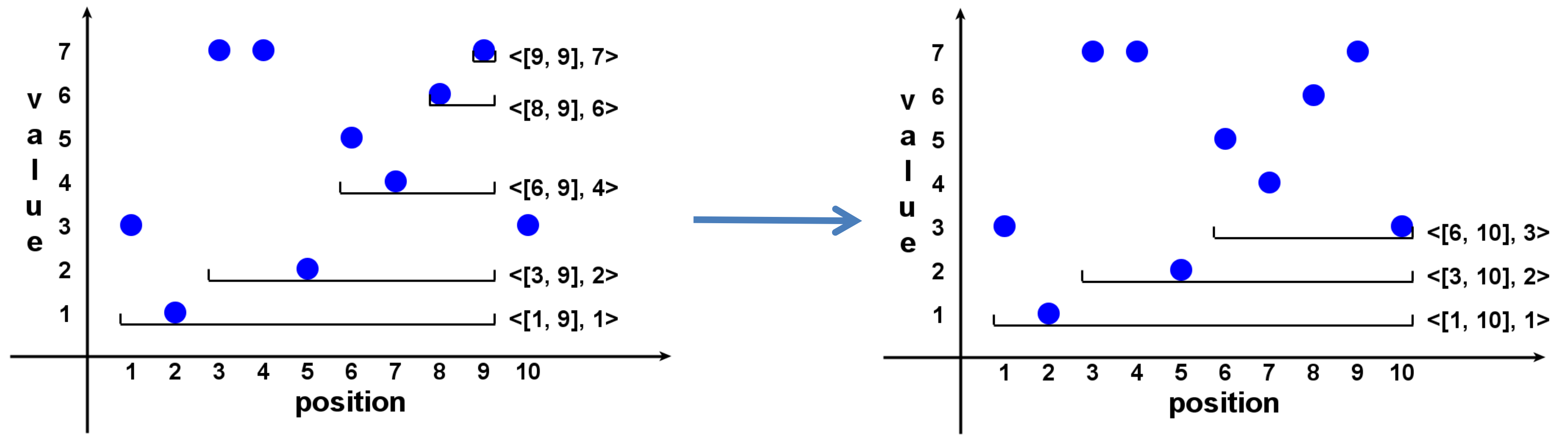
Property 2: the number of LPS is less than or equal to sequence length

Conclusion: LPS is a good set of candidate streaks

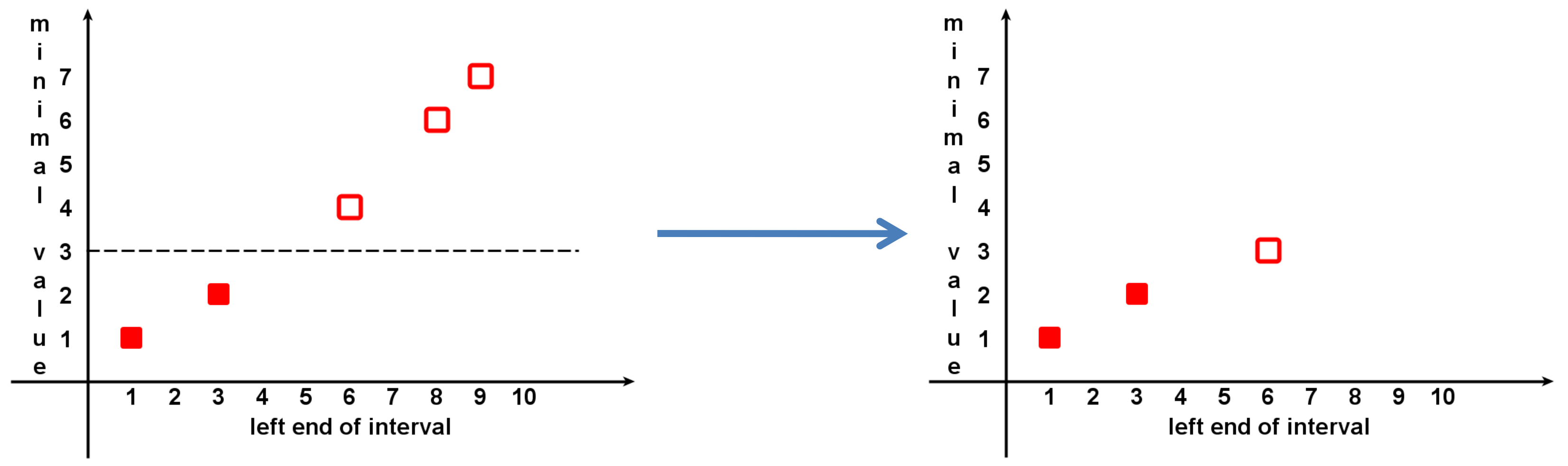


5. Linear LPS Method

1. Maintain a list of streaks when scanning the sequence rightward.
2. After scanning the *k*th value, right-ends of streaks in the list are all *k*.
3. When scanning *(k+1)*-th value, try to extend the streaks rightward.
   - 3.1 Streaks whose *v* is less than *(k+1)*-th value should be extended.
   - 3.2 Only the longest streak of the rest should be extended.
   - 3.3 The streaks whose *v* is greater than *(k+1)*-th value are LPS.
4. After scanning the last value, all the streaks in the list are LPS.



As the streaks share the same right-end, the minimum values are increasing if the streaks are listed in the increasing order of left-ends. Figure below: l-v plot for the above example.



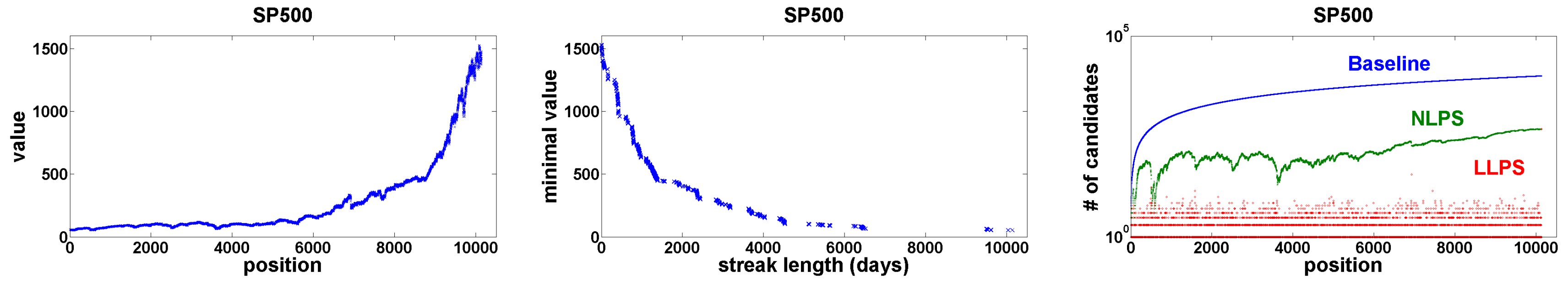
6. Experiments

Real-world datasets, sequence length from thousands to millions.

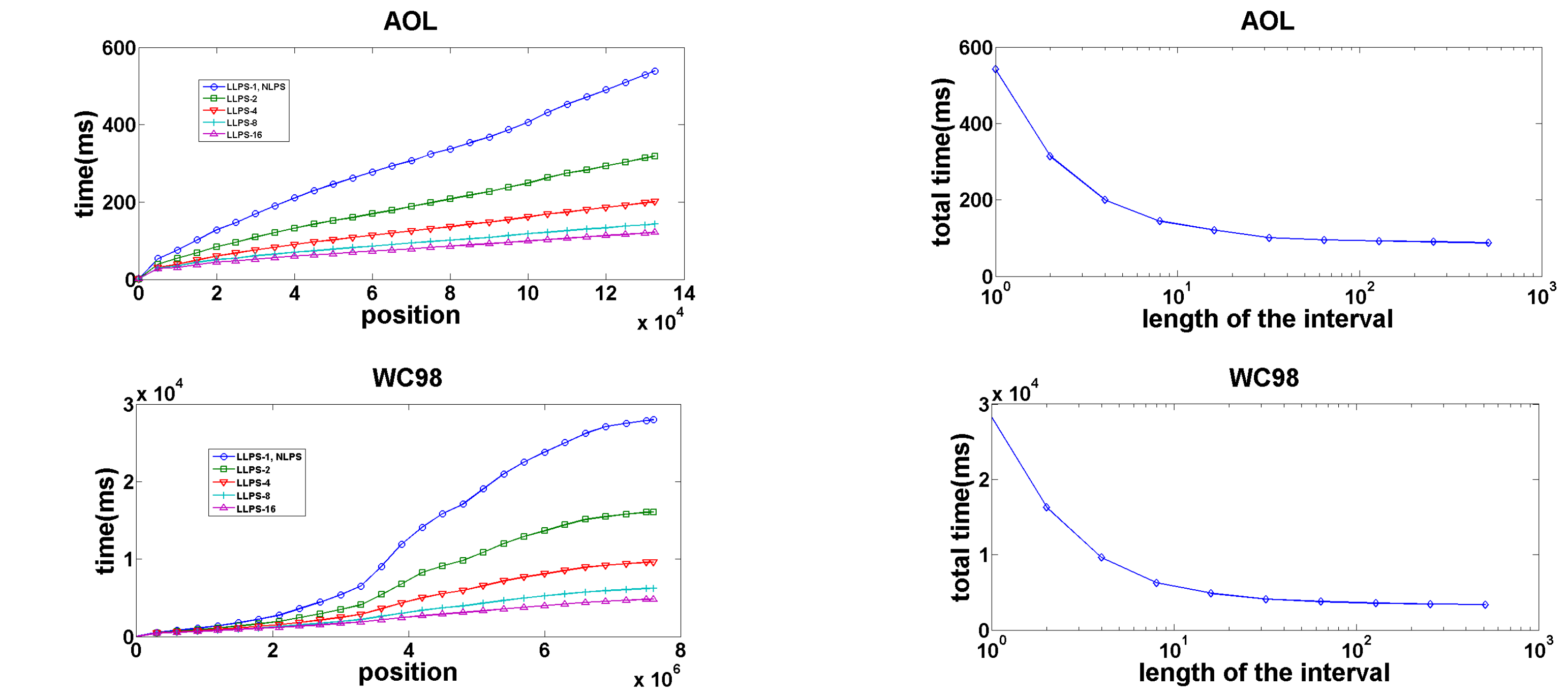
A variety of application scenarios, including meteorology, finance, network traffic.

Data sets	Length	#PS	Baseline(ms)	LLPS(ms)
Gold	1074	137	78	9
River	1400	93	78	3
Melb1	3650	55	390	12
Melb2	3650	58	387	15
Wiki1	4896	58	711	15
Wiki2	4896	51	711	15
Wiki3	4896	118	689	16
SP500	10136	497	4717	21
HPQ	12109	232	6099	18
IBM	12109	198	5079	22
AOL	132480	127	446622	78
WC98	7603201	286	>1 hour	3404

A closer look at SP500 (S&P 500 index, 06/1960-06/2000)



Monitoring prominent streaks of AOL and WC98:



Cumulative Execution Time at Various Positions, for Different Reporting frequencies

Total Execution Time by Reporting Frequencies

7. Interesting Prominent Streaks

Melbourne daily min/max temperature between 1981 and 1990 (Melb1 & Melb2)

- more than 2000 days with min temperature above zero
- 6 days: the longest streak above 35 degrees Celsius

Traffic count of Wikipedia page of Lady Gaga (Wiki2)

- more than half of the prominent streaks are around Sep. 12th (VMA 2010)
- at least 2000 traffic hourly lasting for almost 4 days