Data sketching for transient detection in Tomo-e light curve data

Phungtua-eng Thanapol*, 山本 裕介, 鈴木 健太, 西川 侑志, 山本 泰生

静岡大学・情報学部

木曽シュミットシンポジウム2022

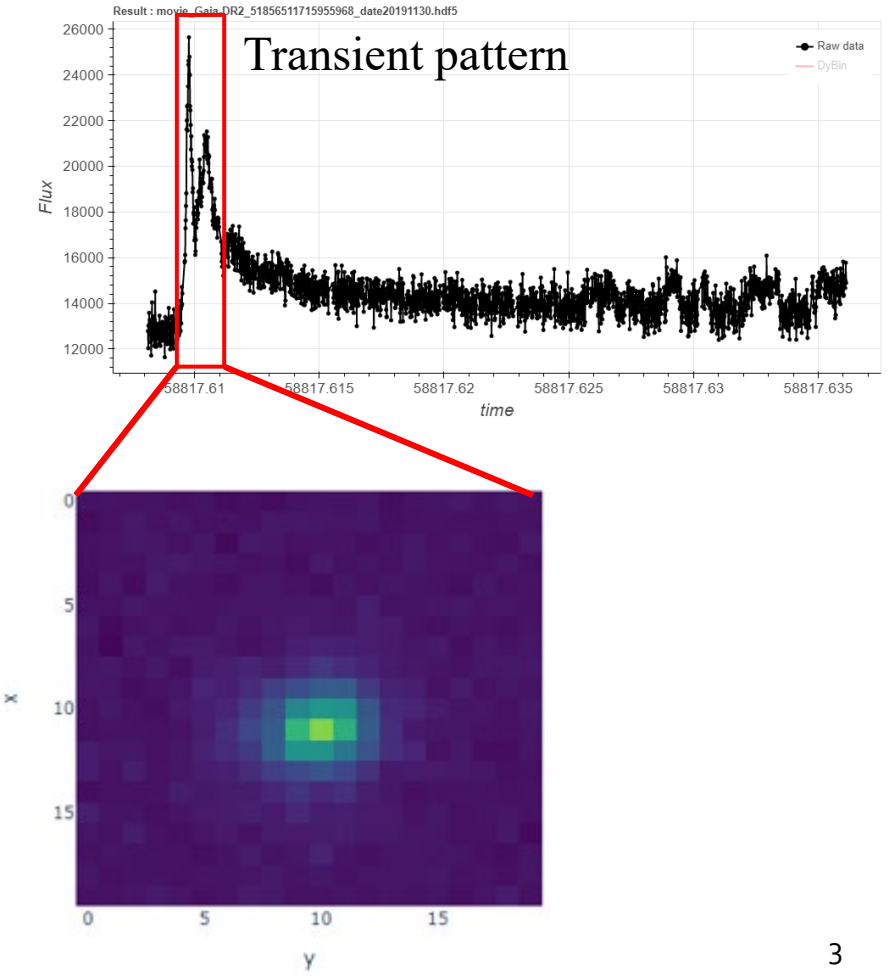
2022.07.06

Outline

- Introduction
- Problem definition
- Dynamic binning
- Experiments on the light curves
- Conclusion

What are *unknown transient patterns* ?

- Transient: short-term phenomenon suddenly and intensively happens in an astronomical object.
- Known (Unknown) transient: transient that has been observed in astronomy (ex: Flare, Blazar, etc.).
- Transient pattern: time-series signal appearing in light curve data, associated with the transient.

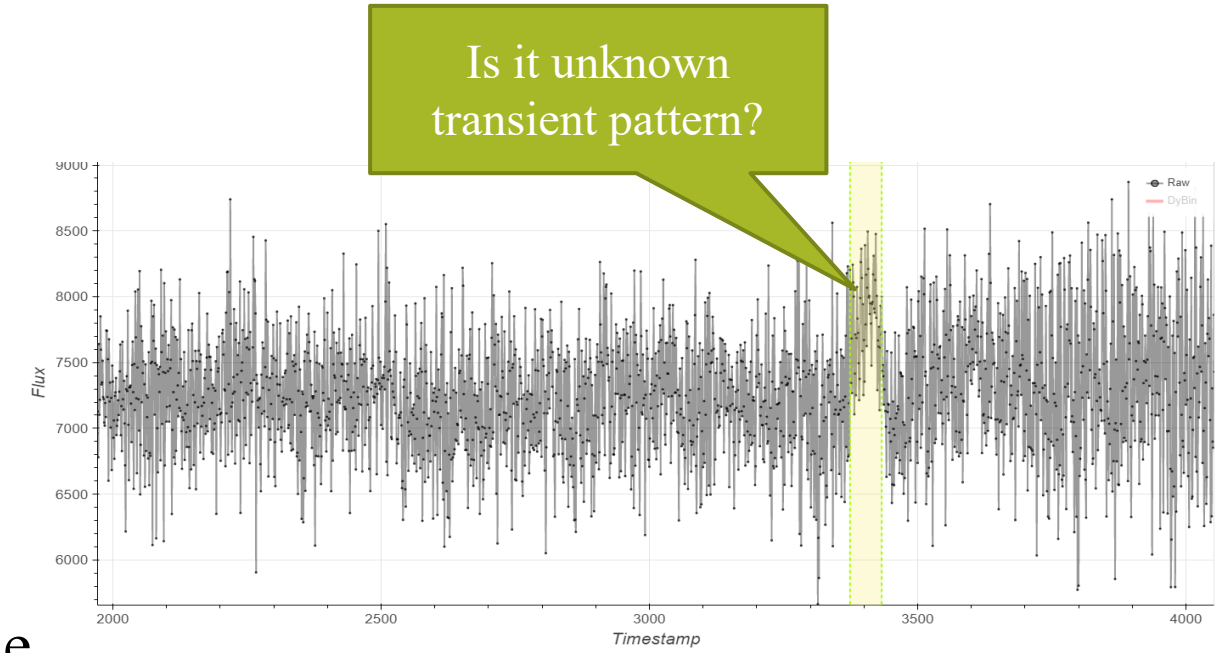


Light curves (LCs)

- LCs is a graph of light intensity of the astronomical object within the region over a certain time period.

- LCs usually contain unwanted external factors:
  - Atmospheric turbulence
  - Measurement error from hardware.

- They may lead to erroneous analysis conclusions.
- The general technique to avoid this issue is *data sketching*



Example of LCs

Data sketching

Q: What is data sketching?

- Summarizes or approximates data into a new representation of fewer lengths.

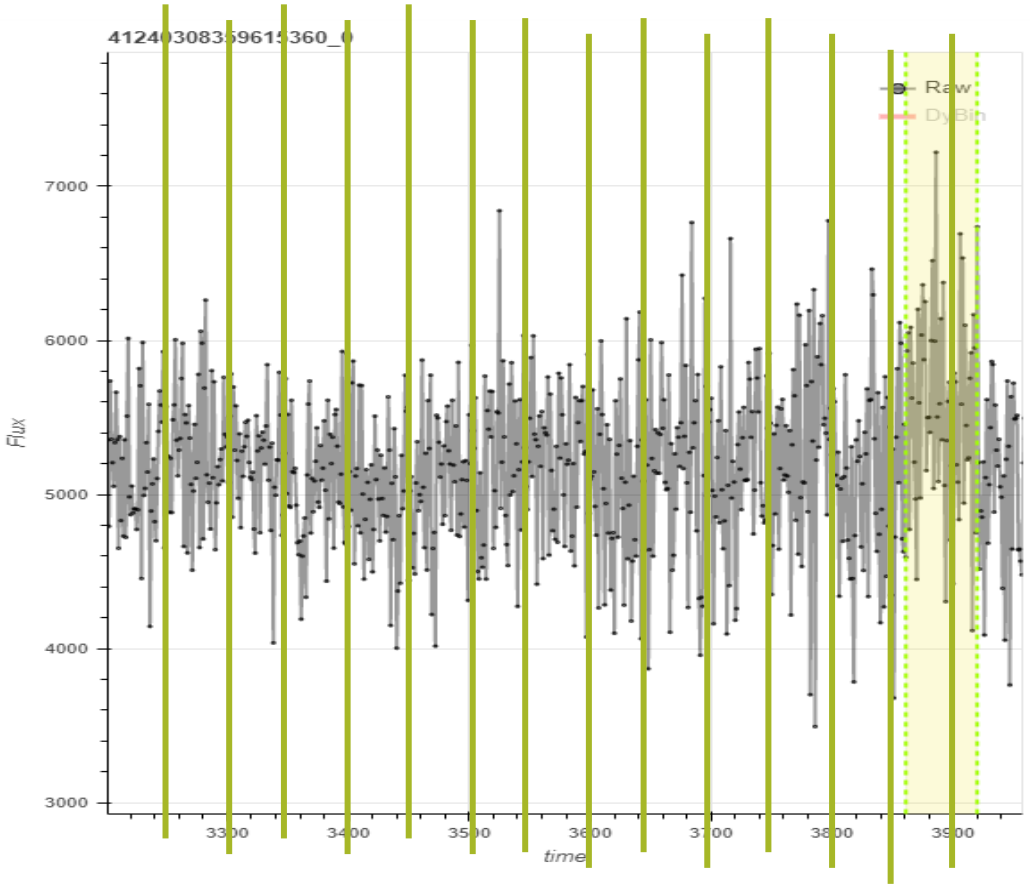
Q: What is a new representation?

- Presents the relevant feature of subsequences of LCs into approximate value, divided into bins.
- We call it *data binning*.

How to compute data binning

- Input : the LCs $\{X_1,X_2,X_3\ldots, X_t\}$, where *t* is timestamp.

- Step 1: the dimensionality reduction is divided into equal-sized bins.

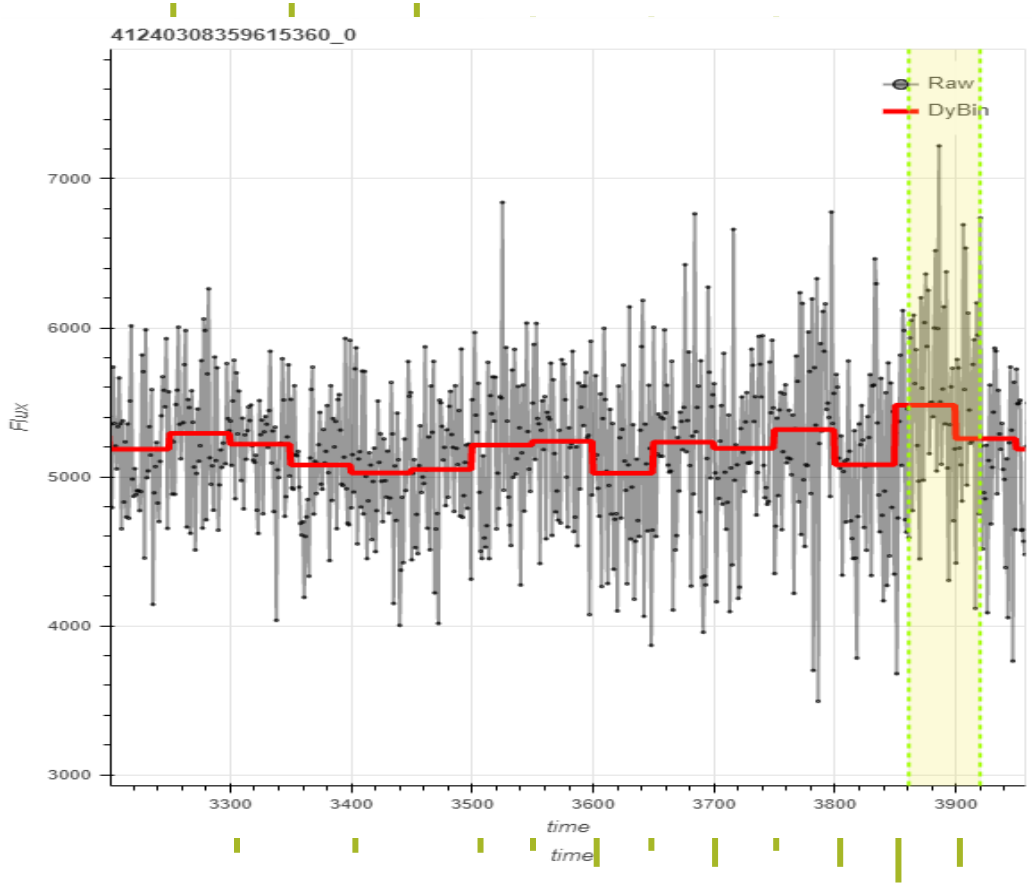


LCs : $\{X_1,X_2,X_3\ldots, X_t\}$

How to compute data binning

- Step 2: computes mean value within bins for representation.

- Output: $W_{1,t} = \{Bin_1, Bin_2, \dots, Bin_w\}$, where w is the latest bin and $w \leq t$.



$W_{1,t} = \{Bin_1, Bin_2, \dots, Bin_{16}\}$

Data binning

Q: What does data binning provide?

- Extracts relevant characteristics of transient patterns
- Reduces noise while preserving the characteristics of the original data

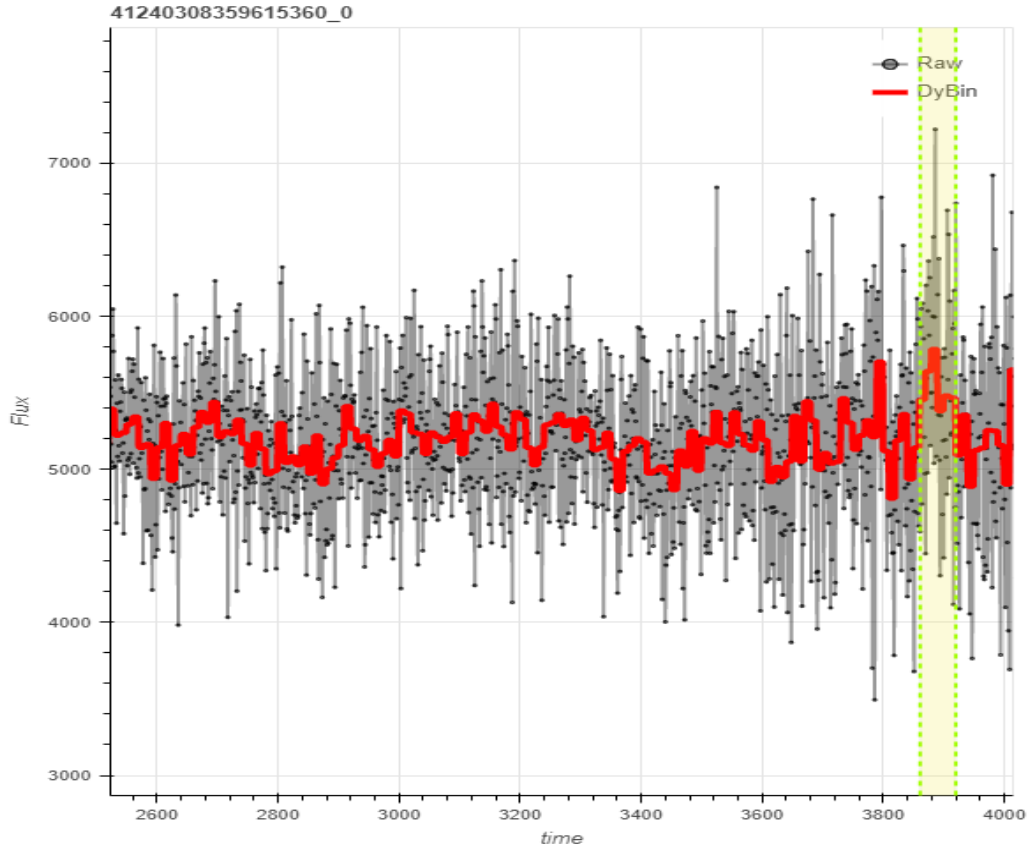
Q: What is required is an input parameter that produces a new representation by data binning?

- Bin sizes and window sizes for division into bins by compression.

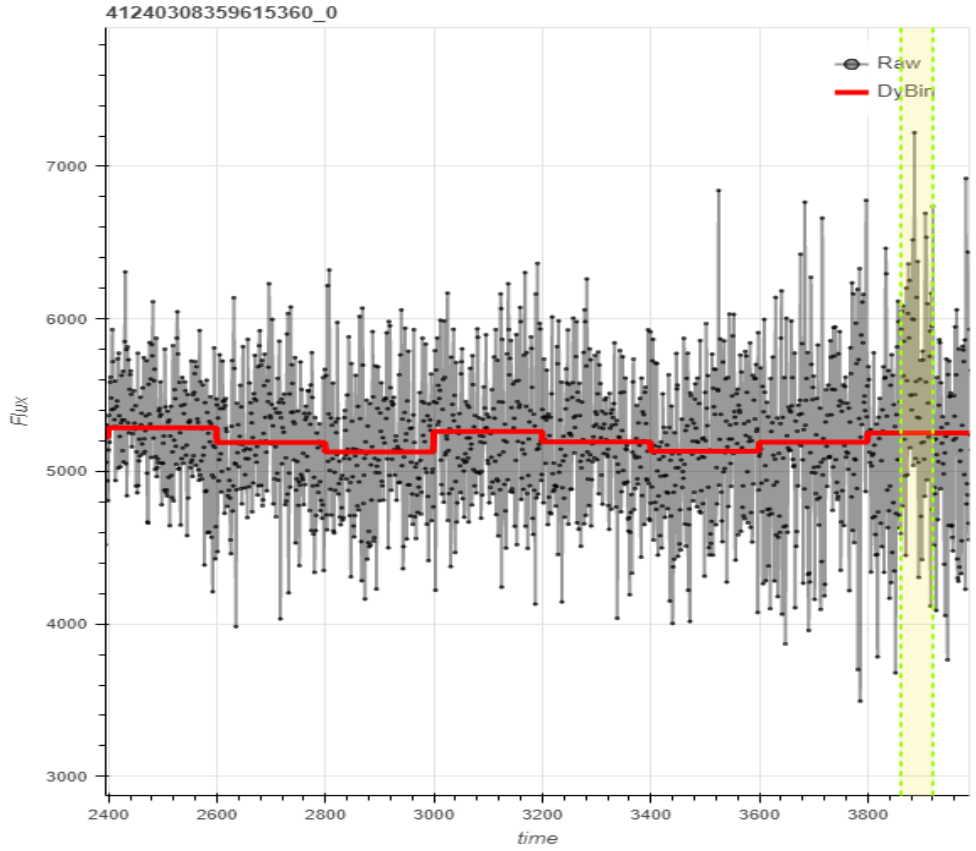
Q: Can we simply choose an arbitrary bin sizes and window sizes with ad hoc input?

- **The answer is NO.**

Arbitrary bin sizes



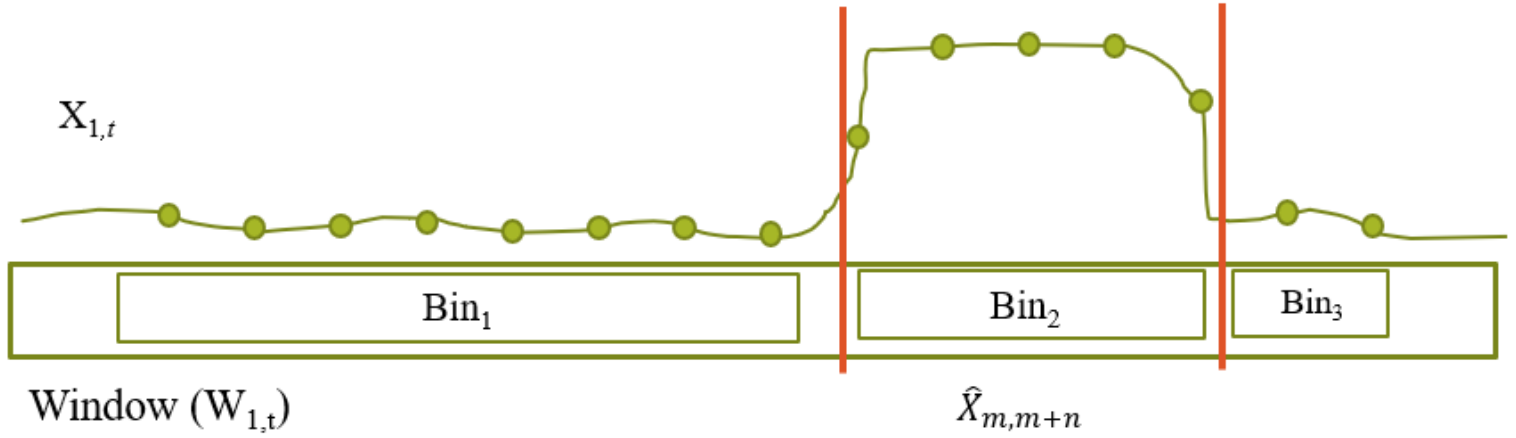
Sub-sequence length is too small



Sub-sequence length is too large

Preliminaries

- Given a time series data of (LCs) $X_{1,t}=\{x_1,x_2,x_3\ldots, x_t\}$, is a sequence of measurements of light intensity within the time interval $[1,t]$.
- Given a $\hat{X}_{m,m+n}$ is unknown transient pattern that is sub-sequence of LCs, where m and $m+n$ are the starting point and endpoint of the transient pattern.
- Output is to split the LCs into w continuous bins $W_{1,t}= \{Bin_1, Bin_2, \ldots , Bin_w\}$, and are collected into a window.



Problem definitions (Cont.)

Problem definitions

- We find the periods of stability and anomaly behavior to divide LCs into sequences of bins.
- Bins in the window are dynamic bin sizes and represent each period's features.
- It aims to produce the representation of LCs using only w bins, while w is minimized.
- It aims to produce the representation of $\hat{X}_{m,m+n}$ using only 1 bin, and that bin corresponds to features of $\hat{X}_{m,m+n}$.

Dynamic binning

- The dynamic binning to adjust the bin size was published in [1].
- Auto-adjusts bin sizes.
- Merges some bins that are likely to be similar to each other.
- Our proposed method includes two statistical hypothesis tests.
    1. T-test for equal means of two bins.
    2. F-test for equal variances of two bins.

Algorithm 1: $DynamicBin$

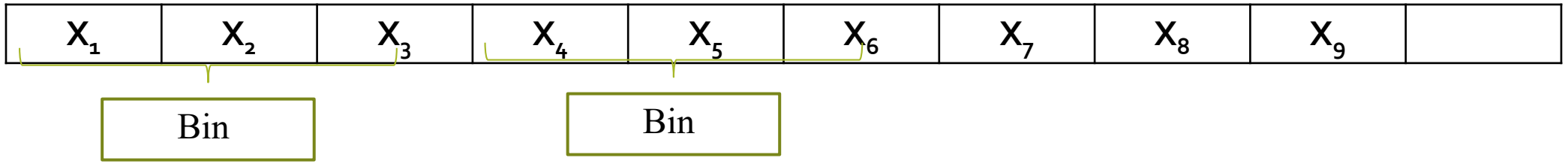
input: W: window, α: significance level

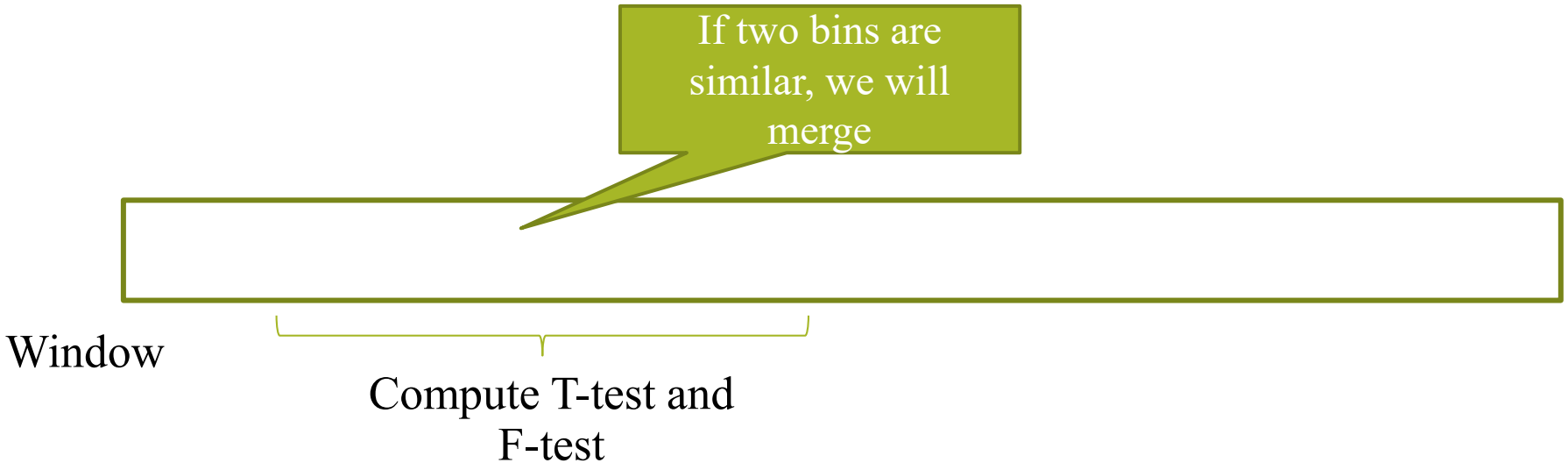
Compute T-test for equal means of Bin_l and Bin_{l-1}
if $|T_{test}| \leq T_{1-\frac{\alpha}{2}}$ **then**
  Compute F-test for equal maximum possible variances of Bin_l and Bin_{l-1}
  if $F_{test} \leq F_{1-\frac{\alpha}{2}}$ **then**
    Merge(Bin_l, Bin_{l-1})
  else
    if *W is full* **then**
      Search the bin Bin_n whose T-test value is minimum in W
      Merge(Bin_n, Bin_{n-1}).
    end
  end
end

Algorithm 1 : Dynamic binning

[1] T. Phungtua-Eng, Y. Yamamoto and S. Sako, "Dynamic Binning for the Unknown Transient Patterns Analysis in Astronomical Time Series," 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 5988-5990, doi: 10.1109/BigData52589.2021.9671917.

How to compute dynamic binning





T-test and F-test

$$\mathrm{T-test} = \frac{|\mu_i - \mu_{i+1}|}{\sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_{i+1}^2}{n_{i+1}}}}$$

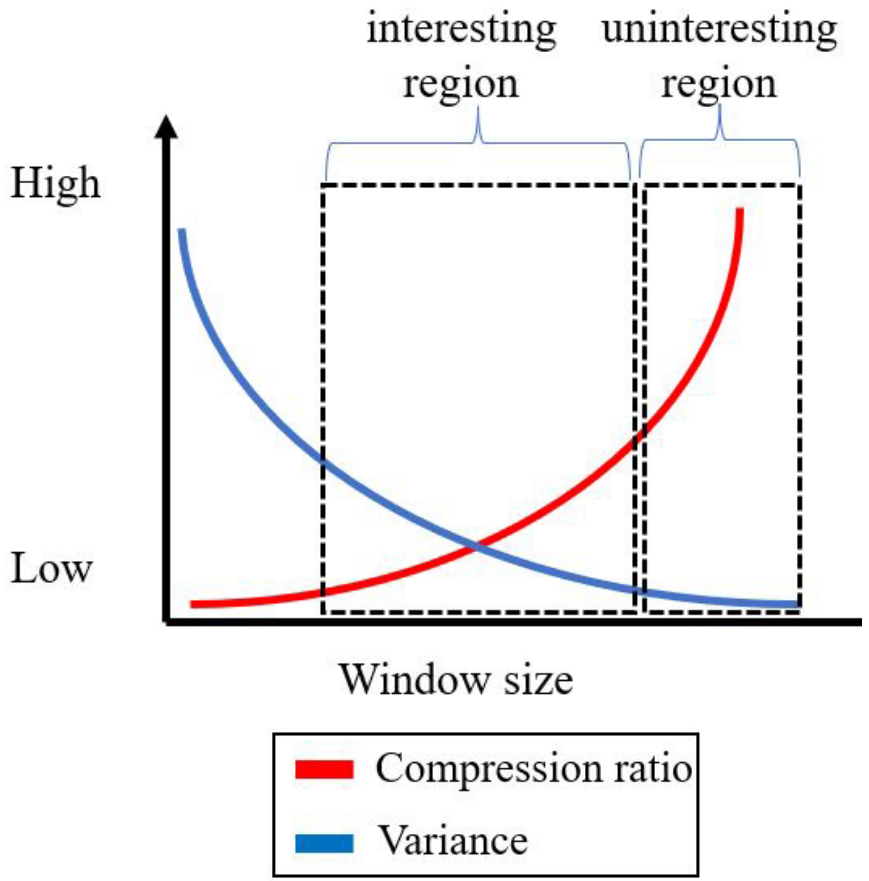
$$\mathrm{F-test} = \frac{\sigma_i^2}{\sigma_{i+1}^2}$$

- i and $i+1$ are index of the i-th and (i+1)-th Bin in the window.
- n is the number of instances that correspond to sub-sequence of LCs.
- μ is the mean, σ^2 is the variance of bin.

- The greater the value of the T-test, the greater the evidence two bins are not similar.
- The greater the value of the F-test, means two bins are significantly different distribution.

Compression ratio

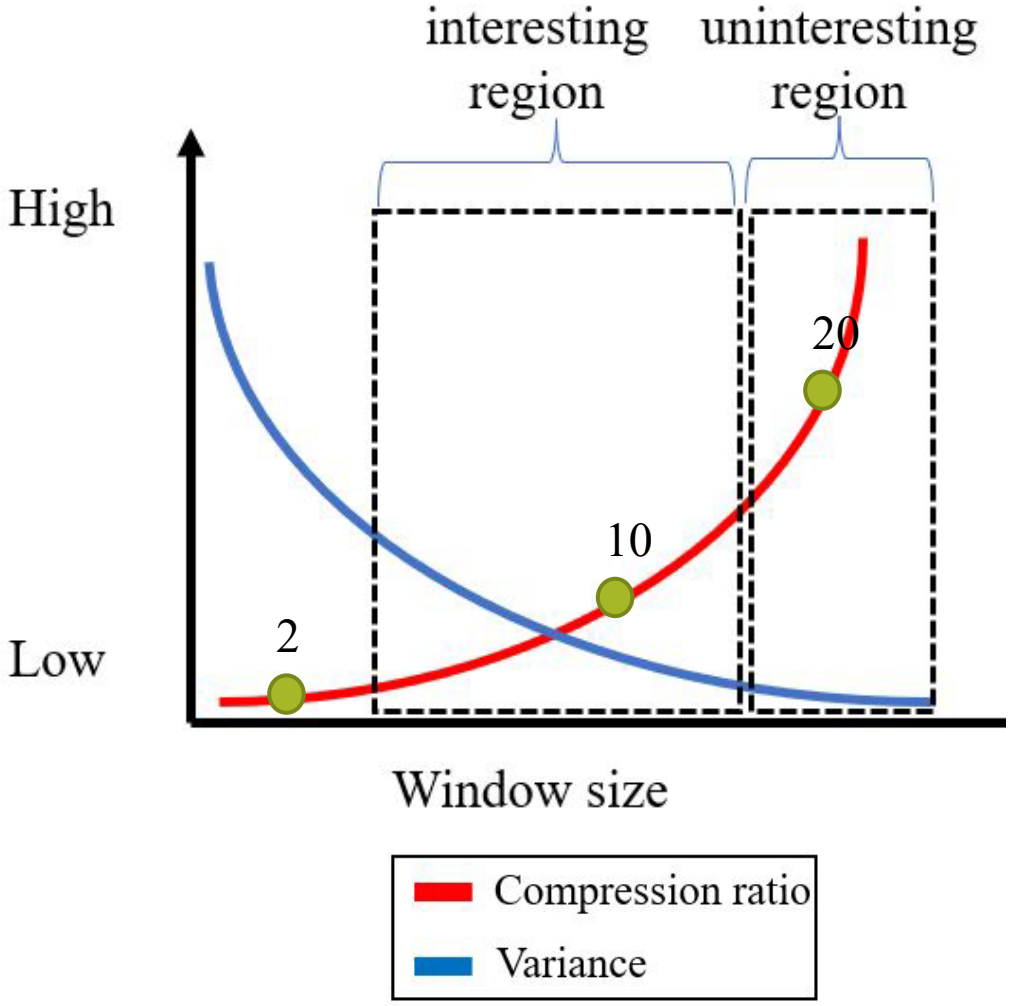
- The compression ratio is the ratio between the length of original data and length of the output by compression.
- The variance and compression ratio do have opposite behaviors.
- *Compression ratio* is an essential measurement for the judgment of window size.



Compression ratio (cont.)

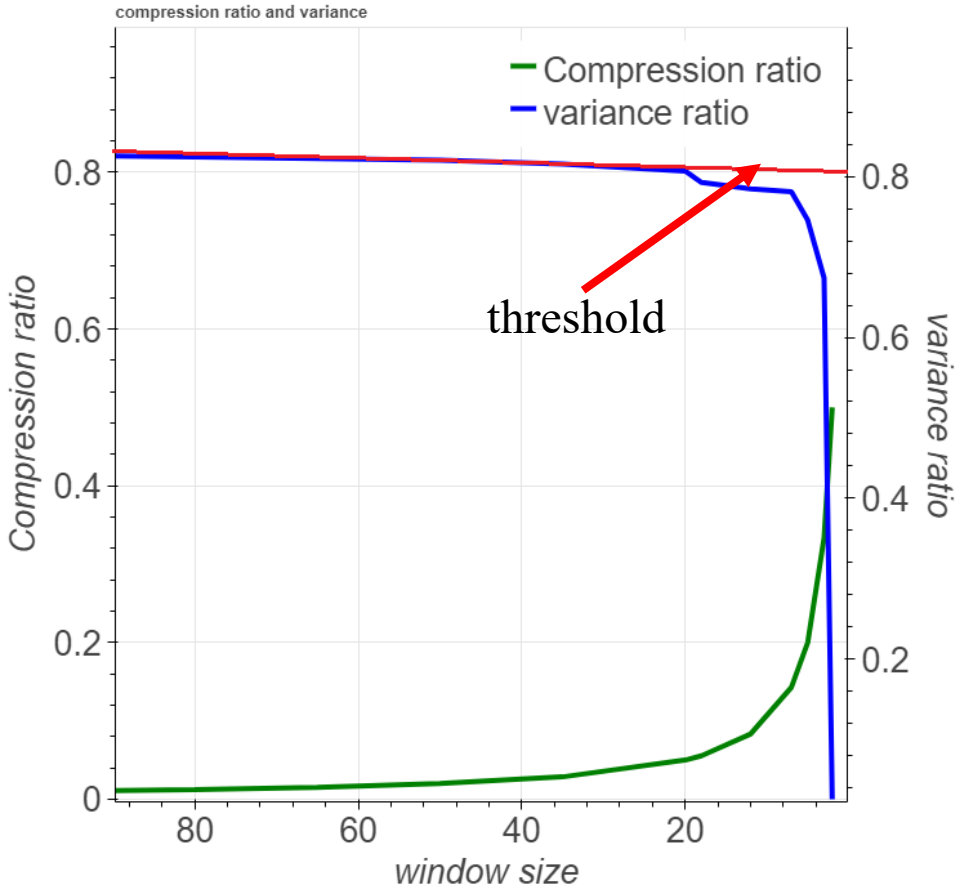
- For example:
  - Input data : $X_{(1,100)} = \{x_1, x_2,\ldots,x_{100}\}$; u = 100 instances.
    - Output is 50 bins > Compression ratio = 2
    - Output is 10 bins > Compression ratio = 10
    - Output is 2 bins > Compression ratio = 20

- A small value of compression ratio means a lot of noise in the input.
- A large value of compression ratio means a low redundancies.



Adjusting window size

- Gives a threshold for considering window size [2].
- When the variance ratio far from the threshold is an uninteresting region.
- In this case, window size is to more than ten bins.

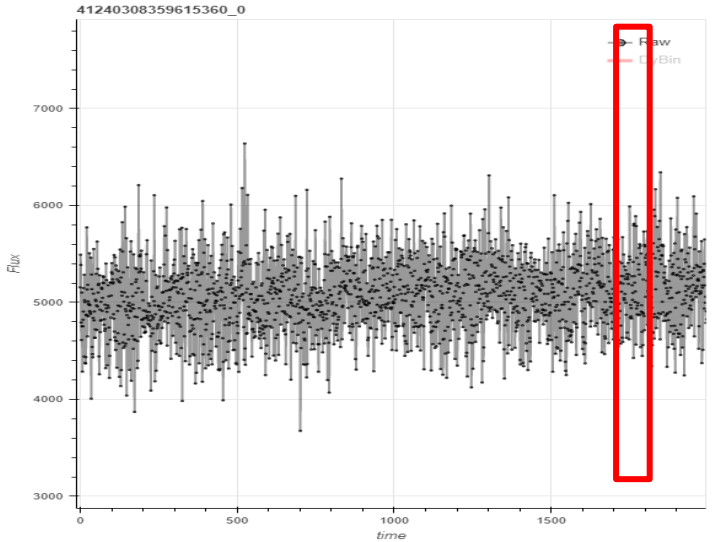


Trade-off between
variance ratio and compression ratio

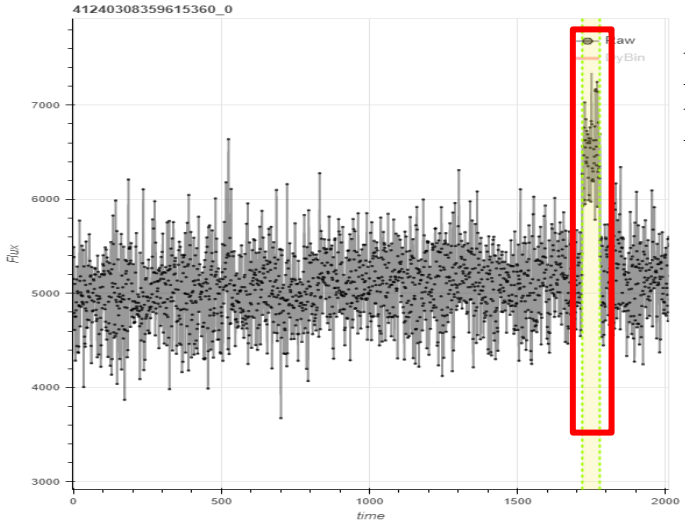
[2] T. Phungtua-Eng, Y. Yamamoto, “A novel framework of non-parametric for adjusting the window size”, 研究報告知能システム（ICS), 2022-ICS-206(5)

Experiments on the LCs

- We inject **artifact one square transient pattern** for each file.
  - The duration varies from time ranging in [1mins, 2mins, 3mins, and 6mins].
  - The power of the transient pattern is [1σ, 3σ], where σ is the standard deviation of the original data for each file.



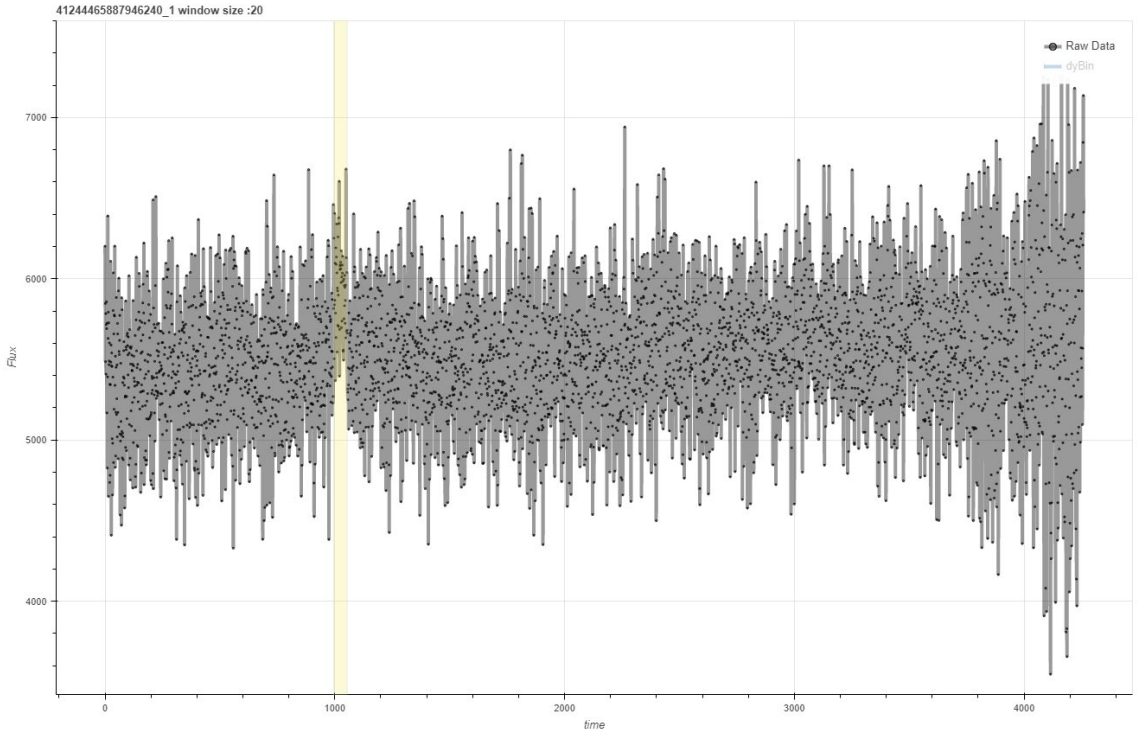
Original LCs



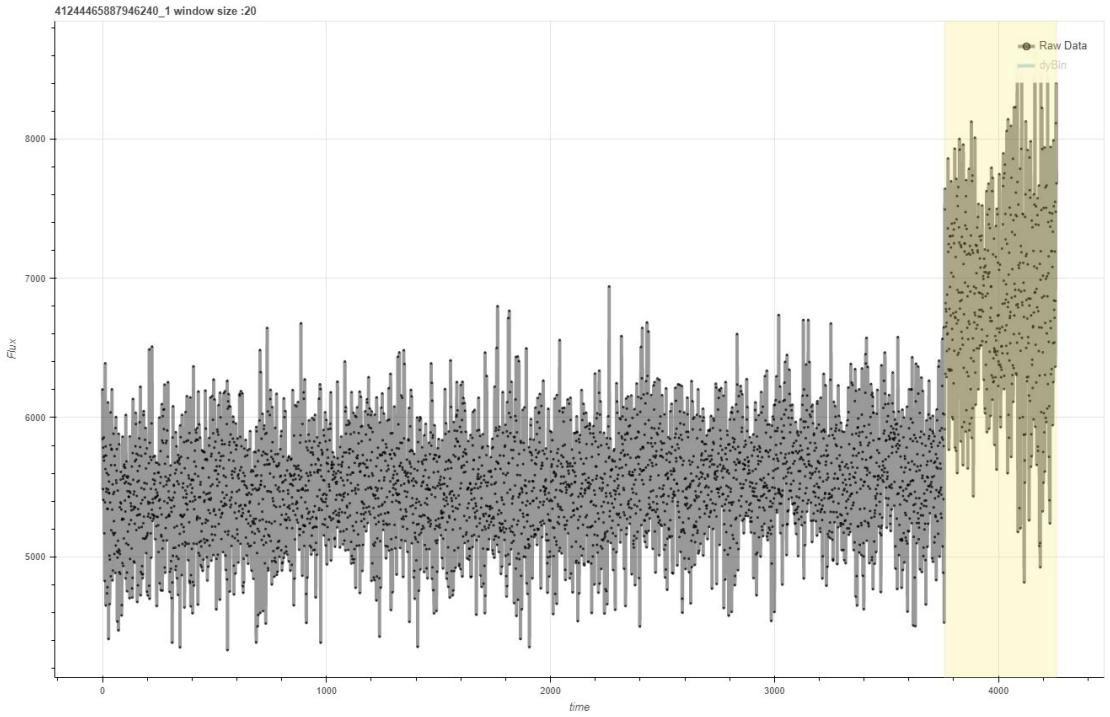
After transient pattern injection

Power : 3σ
Duration : 1mins

Experiments on the LCs (Cont.)



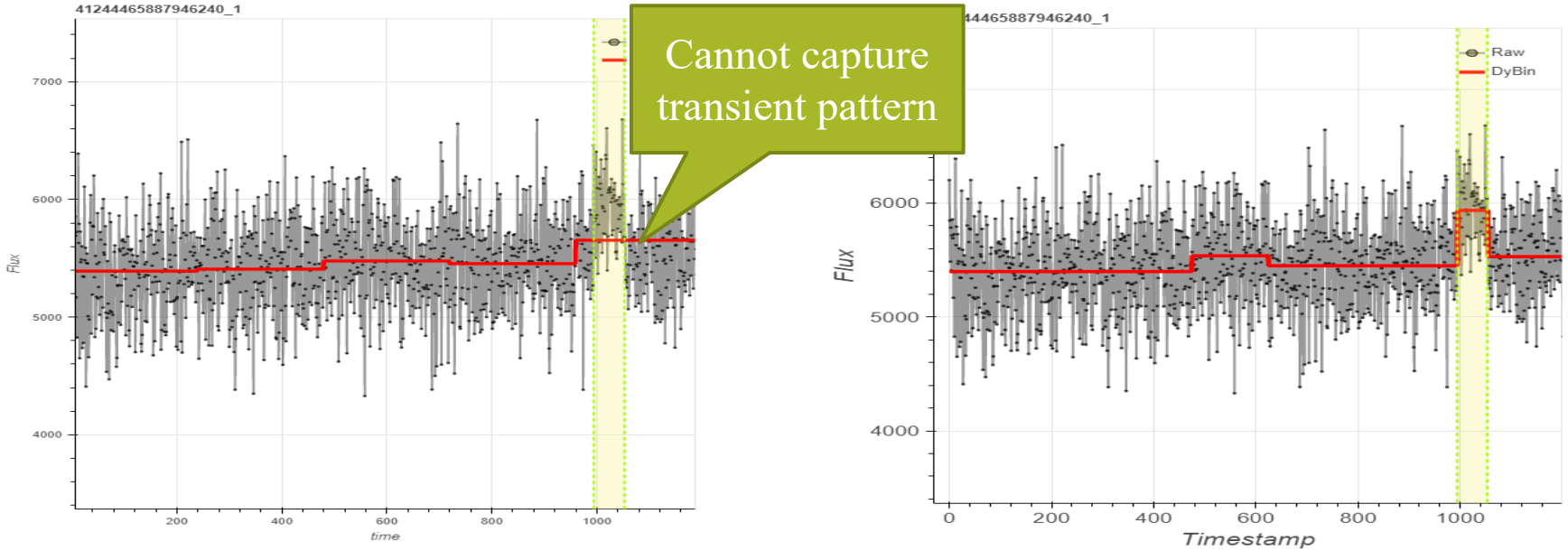
Lowest and shortest duration



Highest power and longest duration

Result of sketching

- We compare our dynamic binning against fixed bin size.



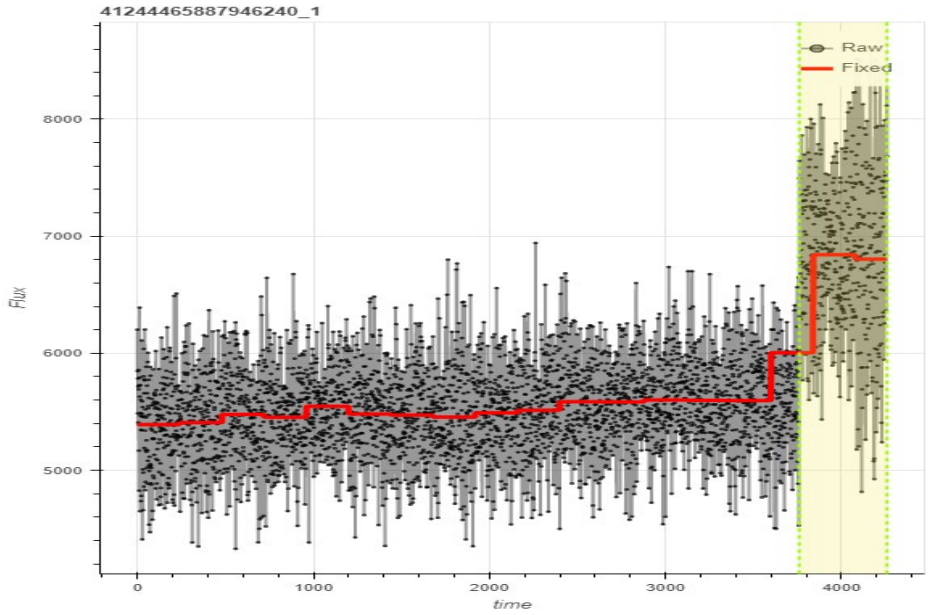
Fixed bin size

Dynamic binning

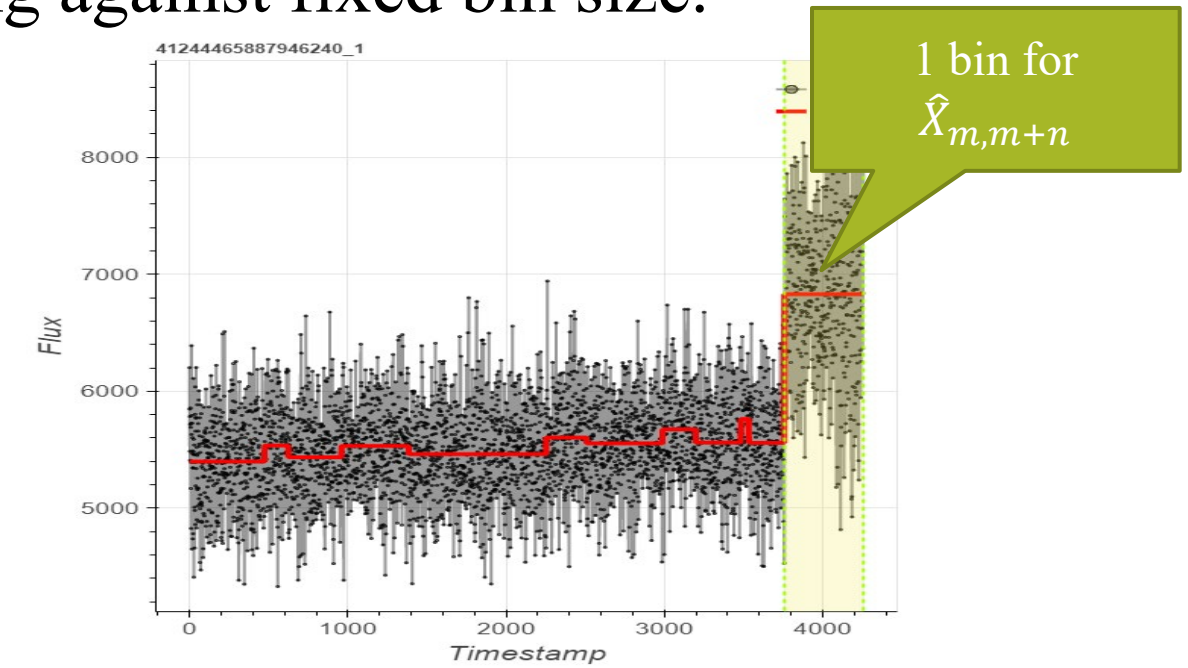
Comparison results dybin and fixed bin size with lowest and shortest duration

Result of sketching (Cont.)

- We compare our dynamic binning against fixed bin size.



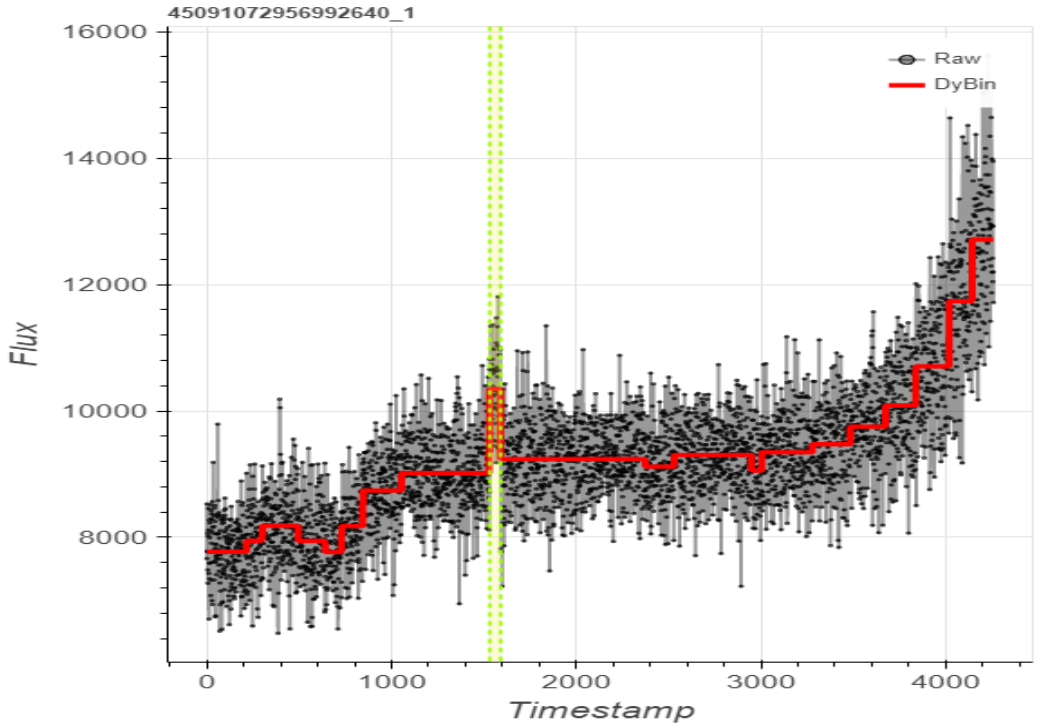
Fixed bin size

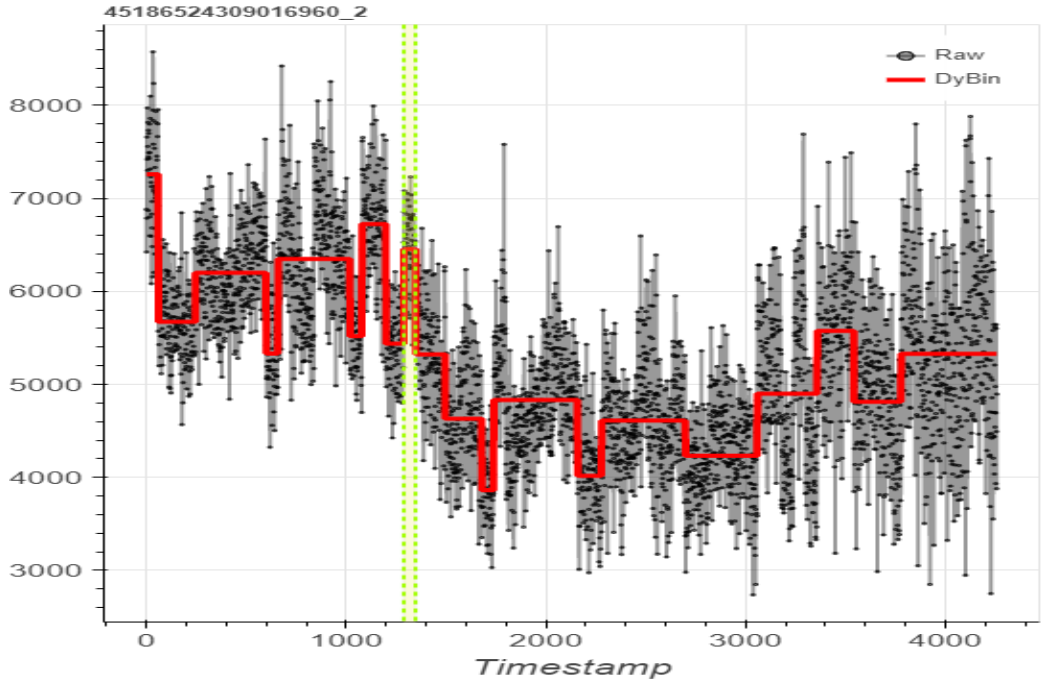


Dynamic binning

Comparison results dybin and fixed bin size with highest power and longest duration

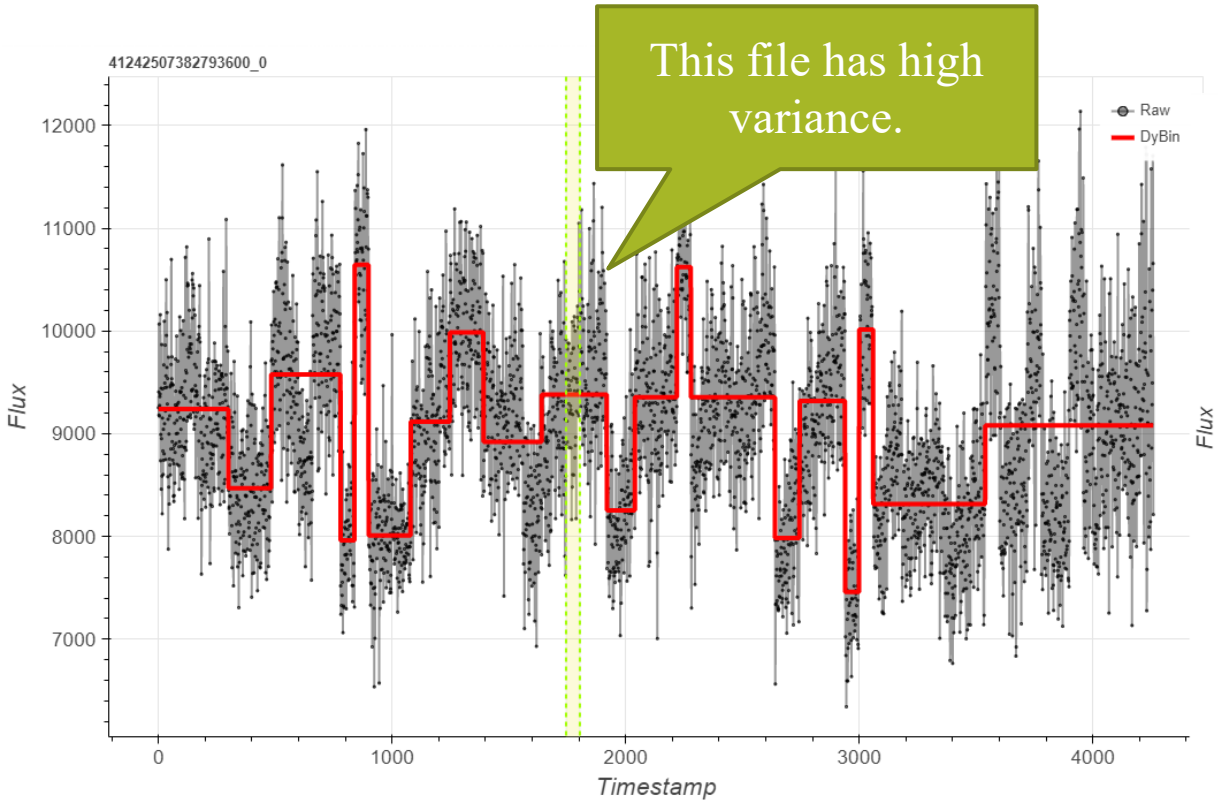
Result of sketching (Cont.)

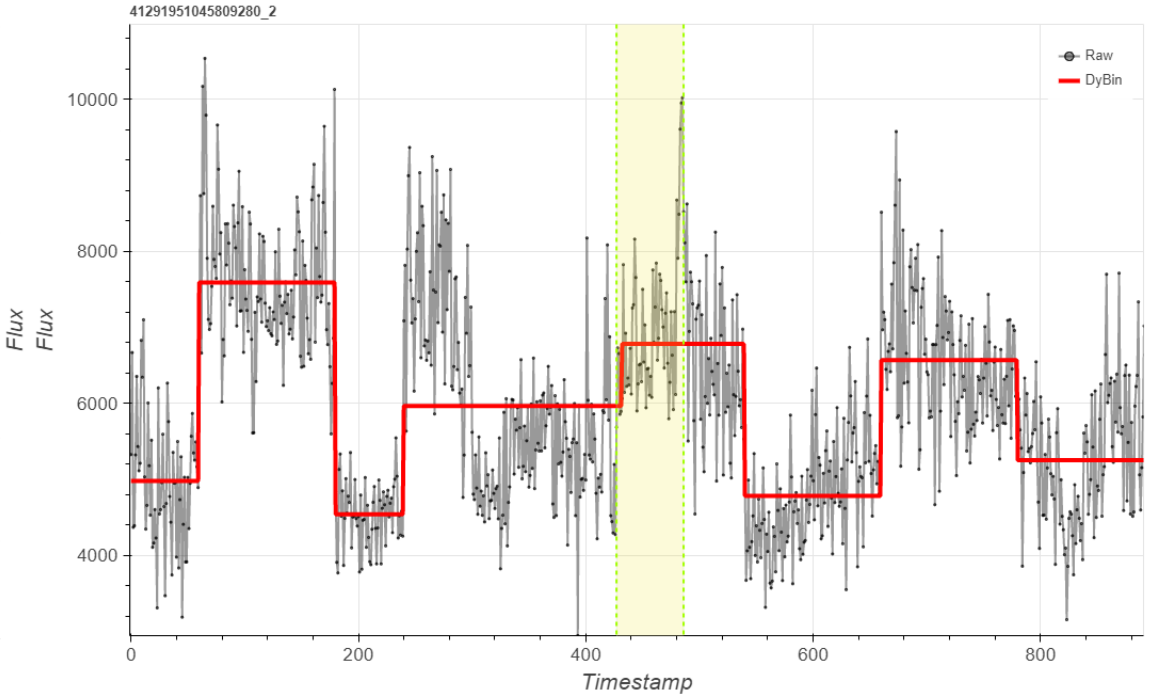




The results of dynamic binning with various scenarios

Result of sketching (Cont.)





Unsuccessful cases

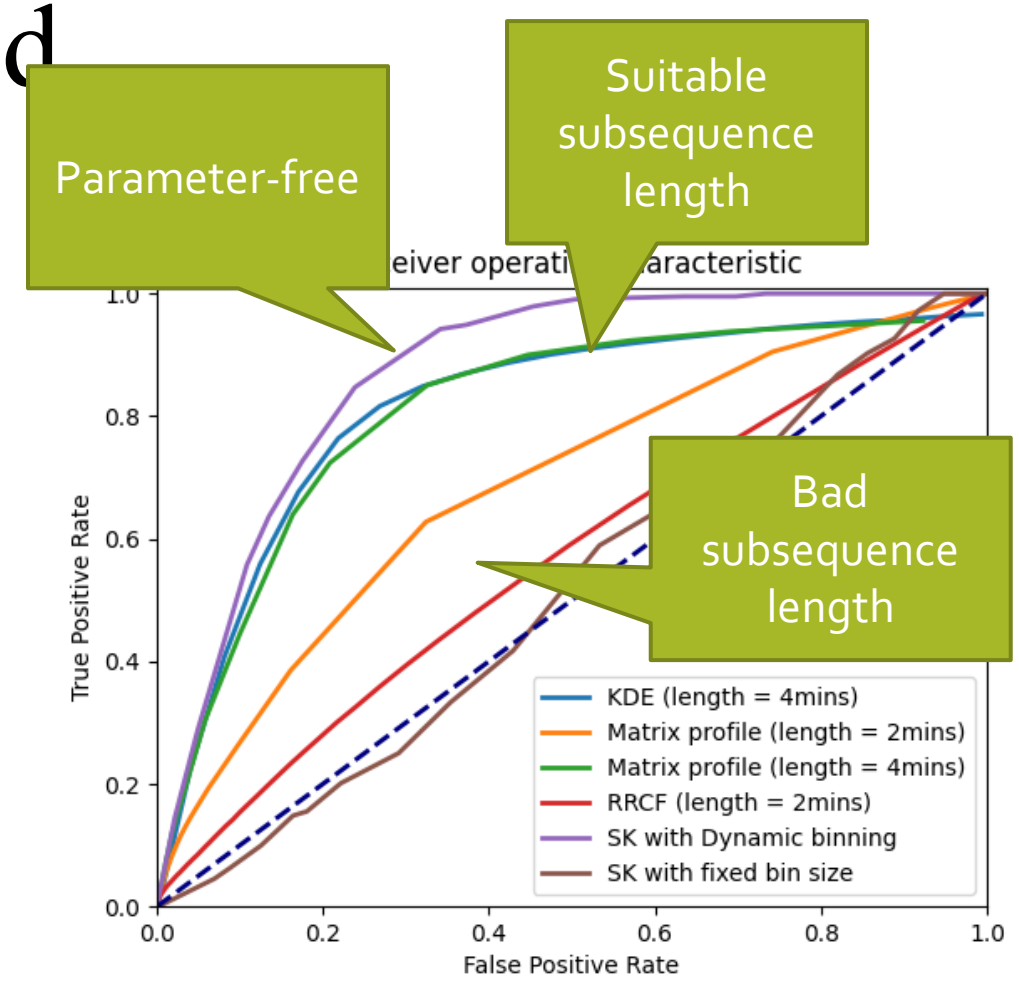
Transient pattern detection

- We apply dynamic binning with detection methods.
- We evaluate dynamic binning against state-of-the-art methods for comparison.
  - SK-method with fixed bin size [3]
  - SK-method with dynamic binning
  - Matrix Profile algorithm (MP) [4]
  - Robust Random Cut Forest (RRCF)[5]
  - Kernel Density Estimation (KDE) [6]

[3] G. Shevlyakov and M. Kan, "Stream Data Preprocessing: Outlier Detection Based on the Chebyshev Inequality with Applications," 26th Conference of Open Innovations Association (FRUCT), 2020, pp. 402-407
[4] C. M. Yeh et al., "Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets," IEEE 16th International Conference on Data Mining (ICDM), 2016, pp. 1317-1322
[5] S. Guha, N. Mishra, G. Roy, & O. Schrijvers, ”Robust random cut forest-based anomaly detection on streams”, 33rd International conference on machine learning, 2016, pp. 2712-2721.
[6] Latecki, L.J., Lazarevic, A., Pokrajac, D. (2007). Outlier Detection with Kernel Density Functions. In: Perner, P. (eds) Machine Learning and Data Mining in Pattern Recognition. MLDM 2007.

ROC of detection method

- We obtained good overall results of dynamic binning using the SK method to detect square transient patterns without put bin sizes and window size.
- MP, KDE, and RRCF based on fixed length.
- We suppose the good representation that may improve performance of RRCF, KDE ,and MP.



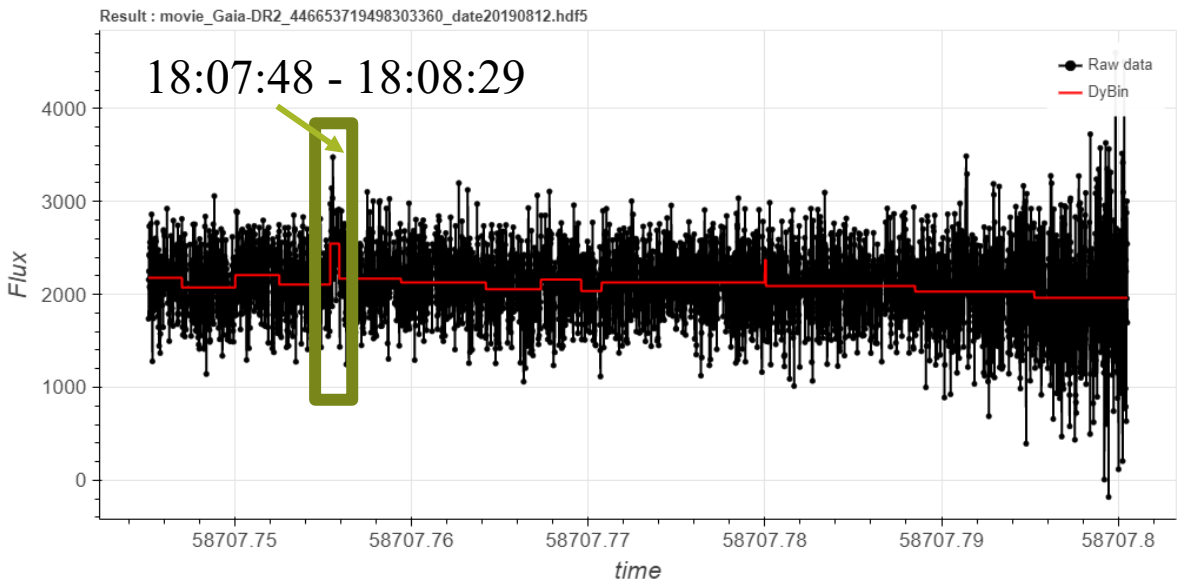
ROC curve of 5 methods

Evaluation with real flares

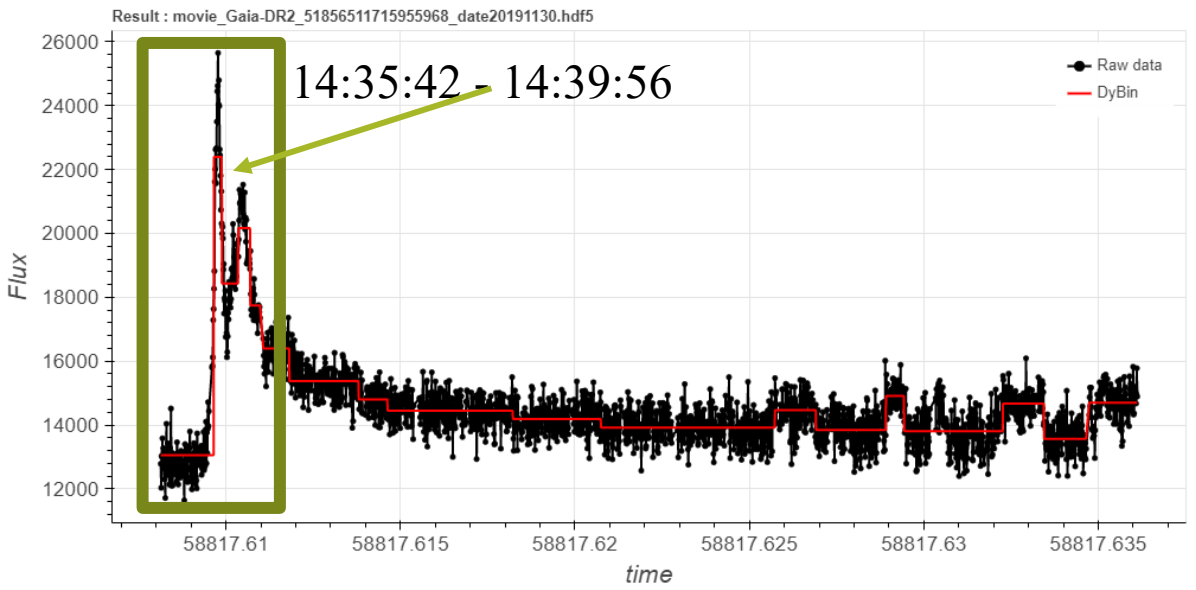
- We evaluate dynamic binning using SK-method with real dataset.
- This dataset contains 10,850 files and 18 files that are real flares.
  - The dataset provided is from Kashiyama-san and Aizawa-san.
- We have two part for evaluation.
  1. Results from our sketching with 18 flares.
  2. Result detection with dynamic binning using SK-method

Example of our sketching with real flares.

- Each bin in the window can represent each period's features.



Low-powered signal



High-powered signal

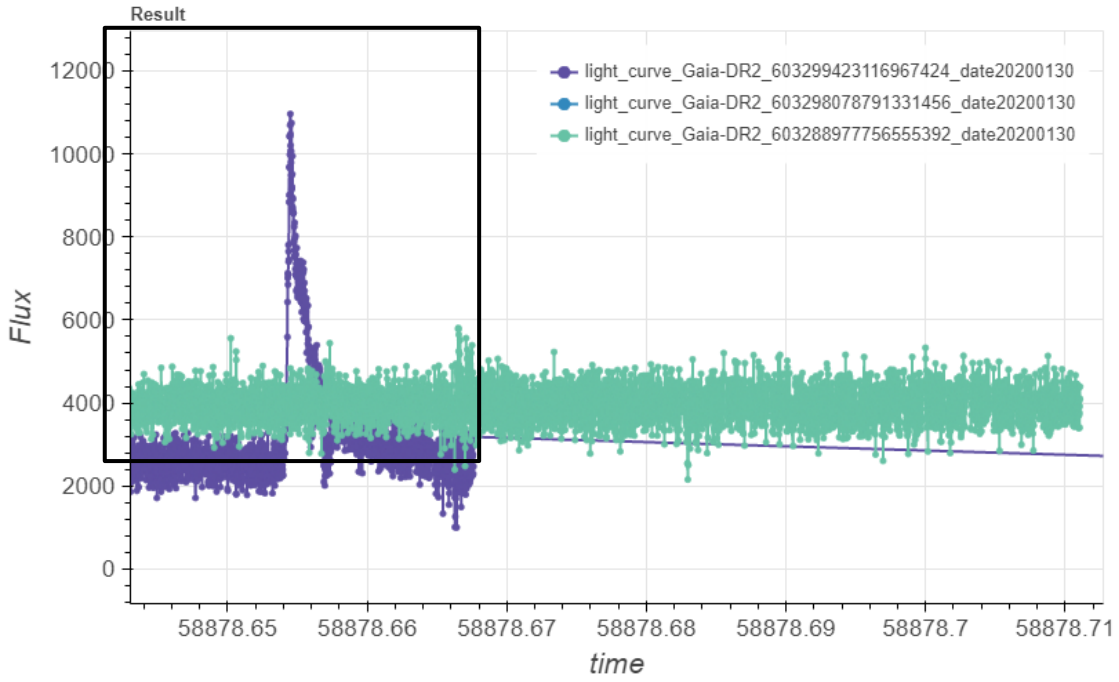
Result detection with dynamic binning using SK-method

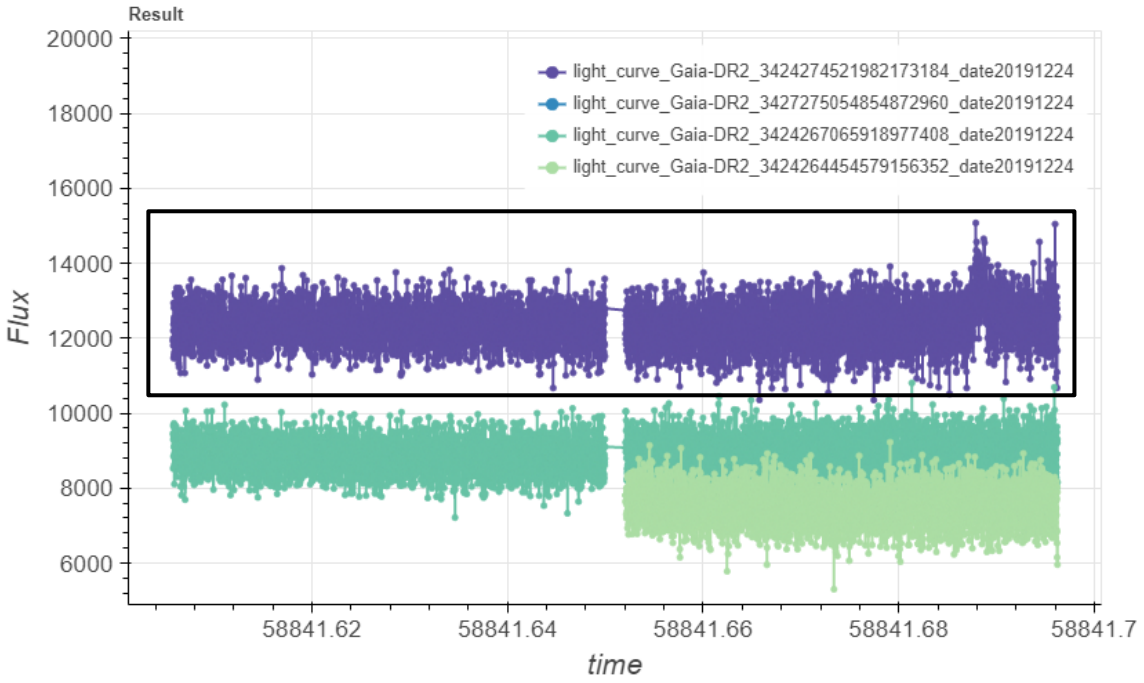
- Input : 10,850 files
- Output : Transient patterns list sorted by the metrics of SK-method.
- Our processes:
  - Sketches each file with **dynamic binning**.
  - Apply SK-method to detect candidate transient patterns [3].
  - Removes simultaneous events by Matrix profile [4].
  - List files whose high value is in Top-100.

[3] G. Shevlyakov and M. Kan, "Stream Data Preprocessing: Outlier Detection Based on the Chebyshev Inequality with Applications," 2020 26th Conference of Open Innovations Association (FRUCT), 2020, pp. 402-407
[4] C. M. Yeh et al., "Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets," 2016 IEEE 16th International Conference on Data Mining (ICDM), 2016, pp. 1317-1322

Detection result and discussion

- We found 8 flares from 18 in Top-100.





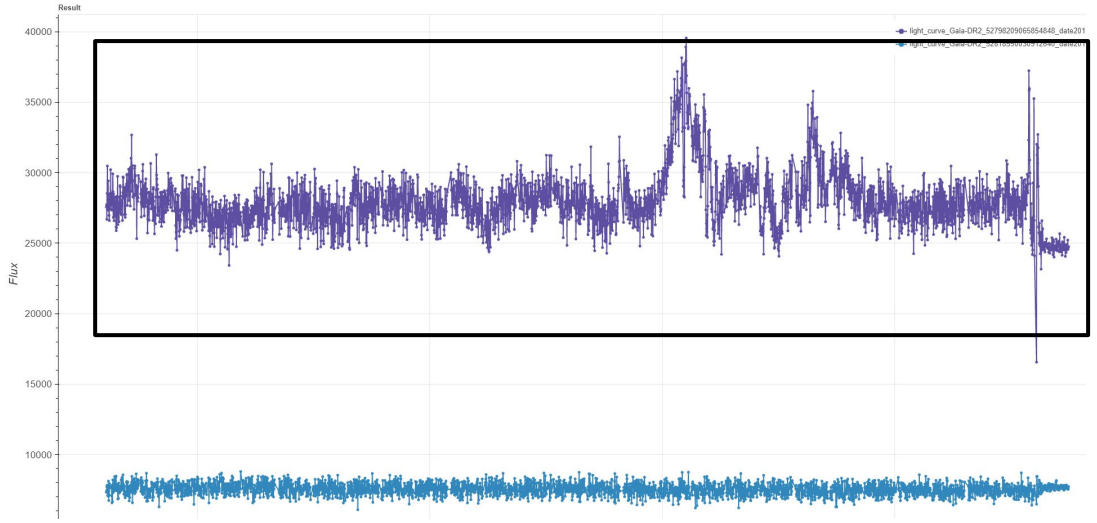
Example of real flares in Top-100

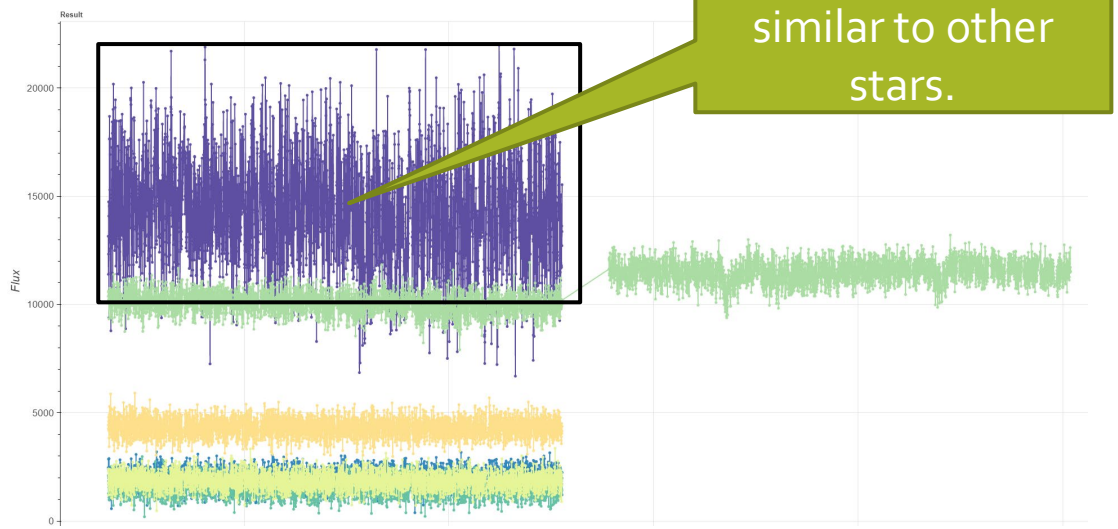
NO	File Name	Ranking
1	light_curve_Gaia-DR2_6049428794672028l6_date20200201	7
2	light_curve_Gaia-DR2_3398180156118506240_date20191224	16
3	light_curve_Gaia-DR2_6032994231169674 24_date20200130	26
4	light_curve_Gaia-DR2_3424781736143842432_date20191224	27
5	light_curve_Gaia-DR2_6027122839080747 52_date20200130	44
6	light_curve_Gaia-DR2_46623557923070848_date20191129	53
7	light_curve_Gaia-DR2_657563345604584064_date20200201	58
8	light_curve_Gaia-DR2_51856511715955968_date20191130	82
9	light_curve_Gaia-DR2_3424274521982173184_date20191224	106
10	light_curve_Gaia-DR2_631840133632923520_date20200201	249
11	light_curve_Gaia-DR2_3410076150374417408_date20191203	271
12	light_curve_Gaia-DR2_49407521363733632_date20191129	299
13	light_curve_Gaia-DR2_459557313084377600_date20190812	364
14	light_curve_Gaia-DR2_3311459749888116736_date20191129	406
15	light_curve_Gaia-DR2_2977433546209292672_date20191106	528
16	light_curve_Gaia-DR2_612157157509901568_date20200201	538
17	light_curve_Gaia-DR2_603188200643885696_date20200124	594
18	light_curve_Gaia-DR2_446653719498303360_date20190812	598

We can find all of them in Top-600.

Detection result and discussion (Cont.)

- Current problem is simultaneous event detection part.
  - Some files are unique when compare with near-star files.
  - We plan to evaluate with variety aperture sizes.

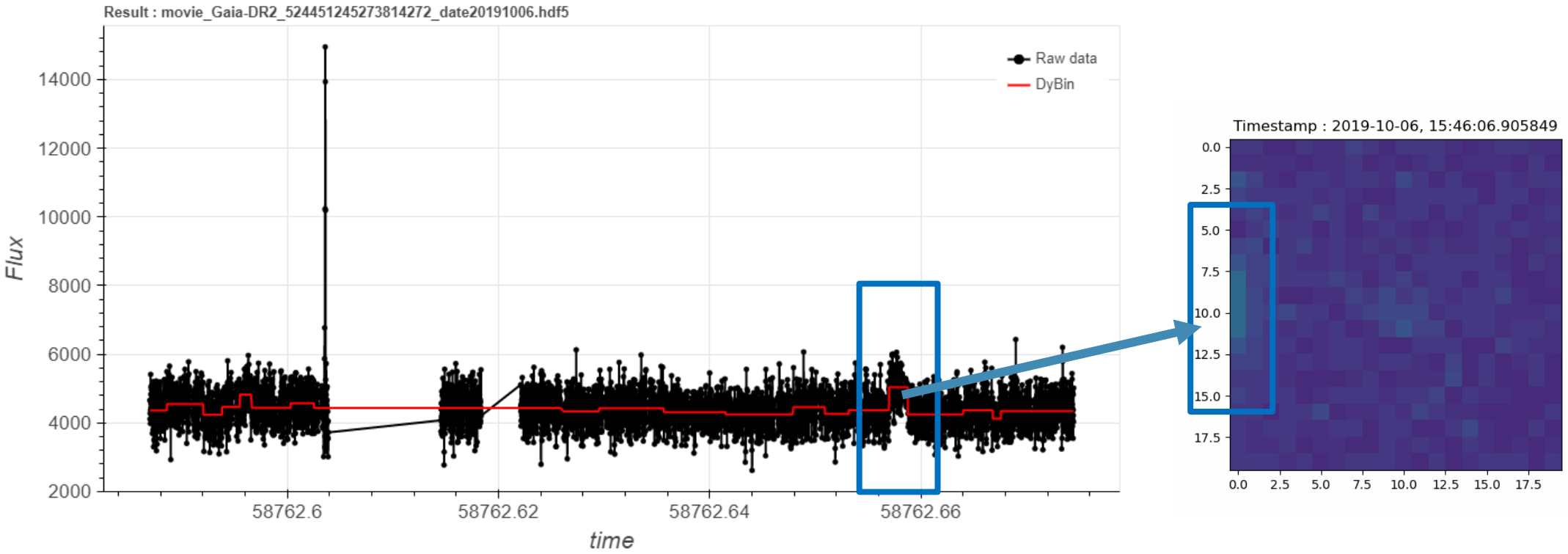




Files are not real flares but are unique behavior.

Detection result and discussion (Cont.)

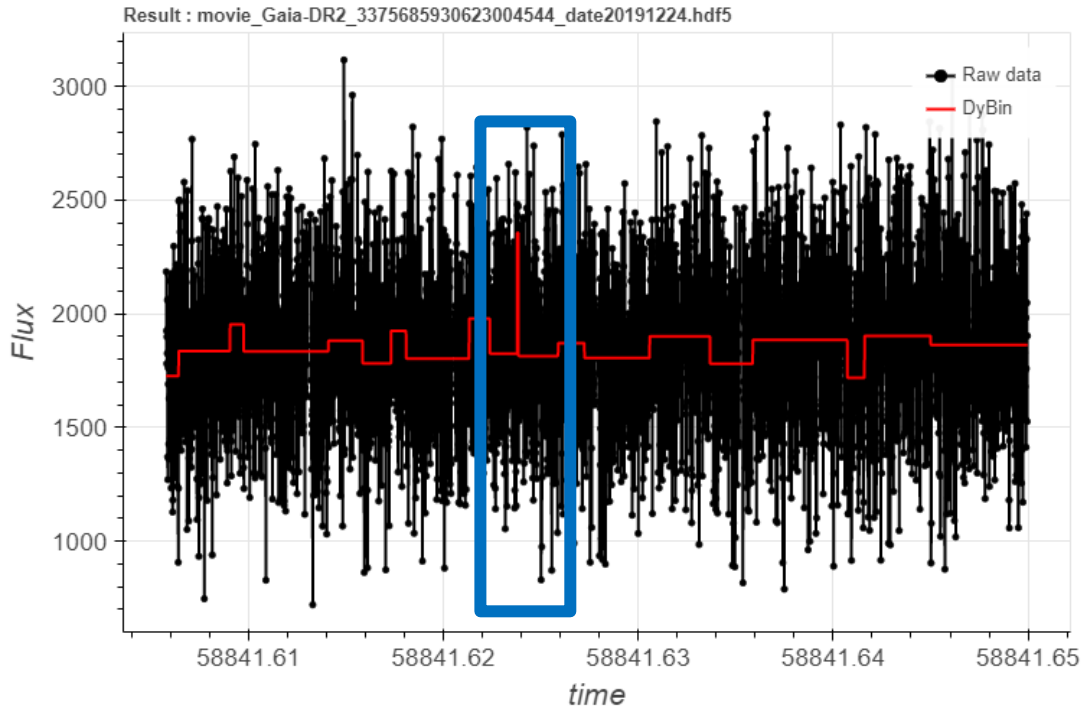
- We plan to improve simultaneous events removing for this issue.

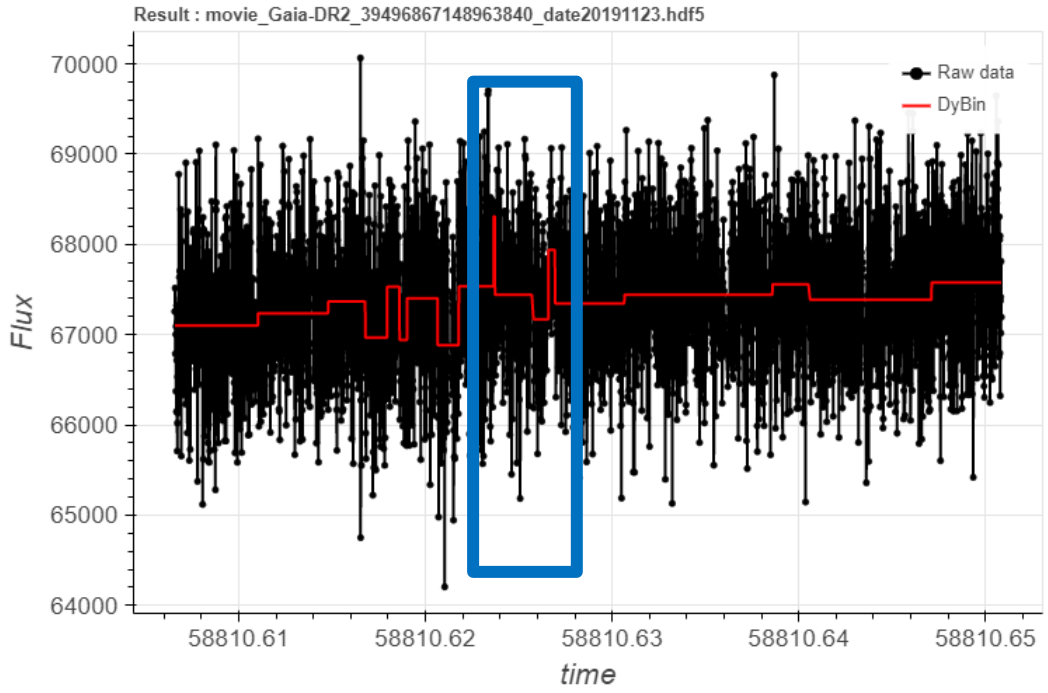


Example of false positive

Detection result and discussion (Cont.)

- The second problem is that the bin is too small and cannot be merged with neighboring bins.
- Our algorithm starts with a small bin and merges into a large bin (bottom-up approach).
  - Our further work is top-down approach and hybrid approach.





Conclusion

- Window sizes and bin sizes can freely parameter without cornering significant differences results.
- We can extract relevant characteristics of transient patterns.
- **Further plan**
  - Investigation for applying dynamic binning with other detection methods.
  - Is there a way to address simultaneous event detection?
  - Decreasing the computational cost for the minimal window size searching.

Conclusion (Cont.)

- We are currently developing a web application using this algorithm.

User interface in our web application