

KMVスケッチによる時系列データ要約 と突発検知応用に向けた予備検討

○西川 侑志, Thanapol Phungua-eng, 山本泰生

静岡大学 情報学部

木曾シュミットシンポジウム2024

(木曾福島保険センター)

2024年5月15日

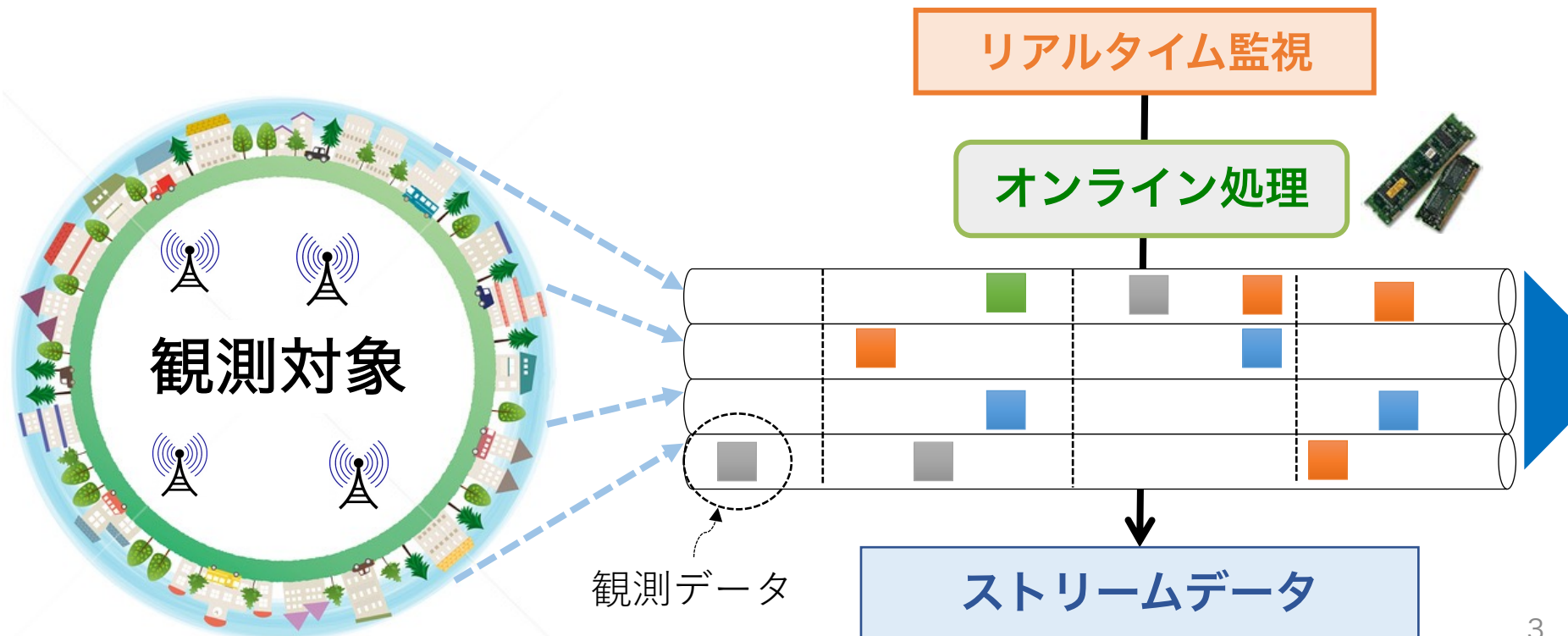
発表の概要

- 研究背景
 - ストリームデータ処理
 - データ要約
- KMVを用いた頻度サマリの構築
- 頻度サマリを用いた時系列データ要約
 - 問題設定
 - アプローチ
- ケーススタディ
 - 測光時系列データの頻度分布
 - エビデンスに基づく異常検知

ストリームデータの研究

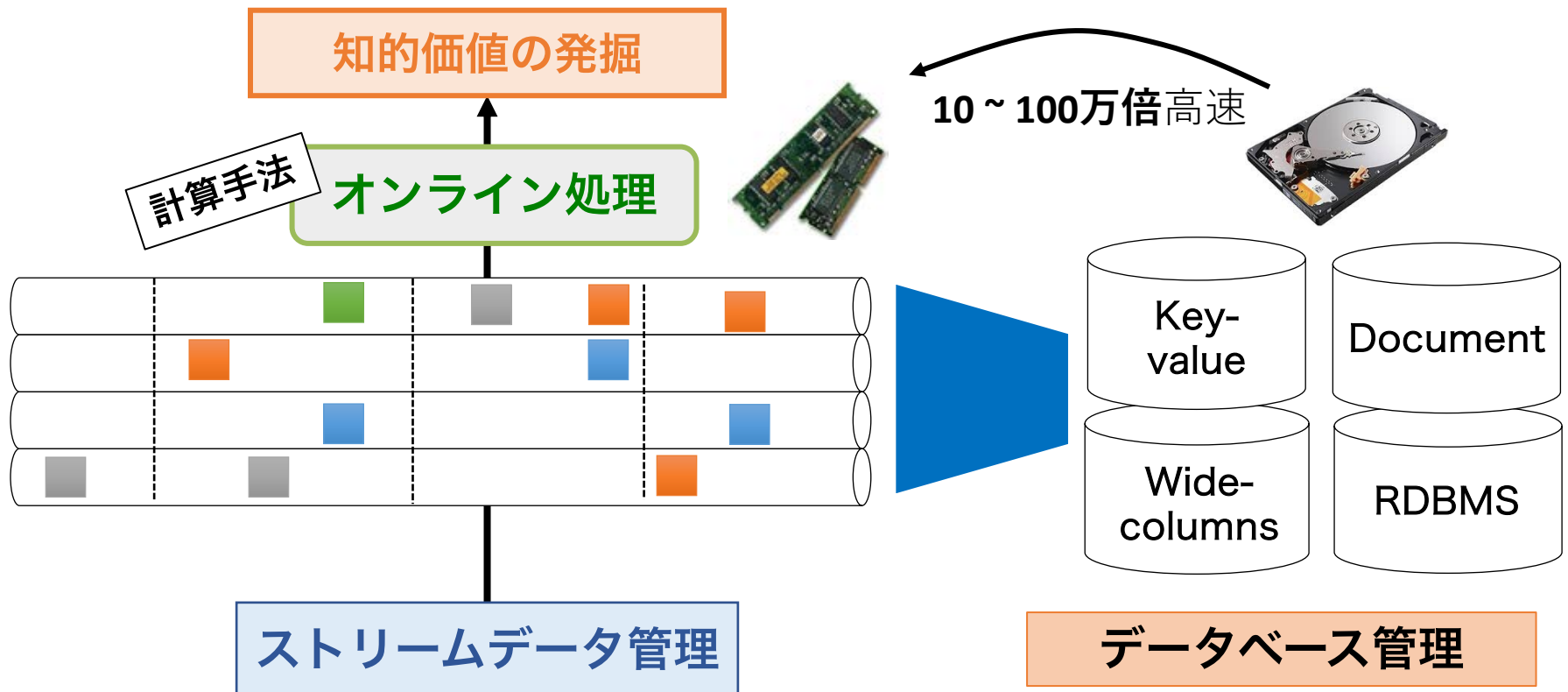
・ストリームデータとは？

- 高速に流れ続ける無限長のデータ列
- センサーノードから常時到着する観測データ
- 観測対象のリアルタイム分析 (**傾向の変化や異常の検出**)



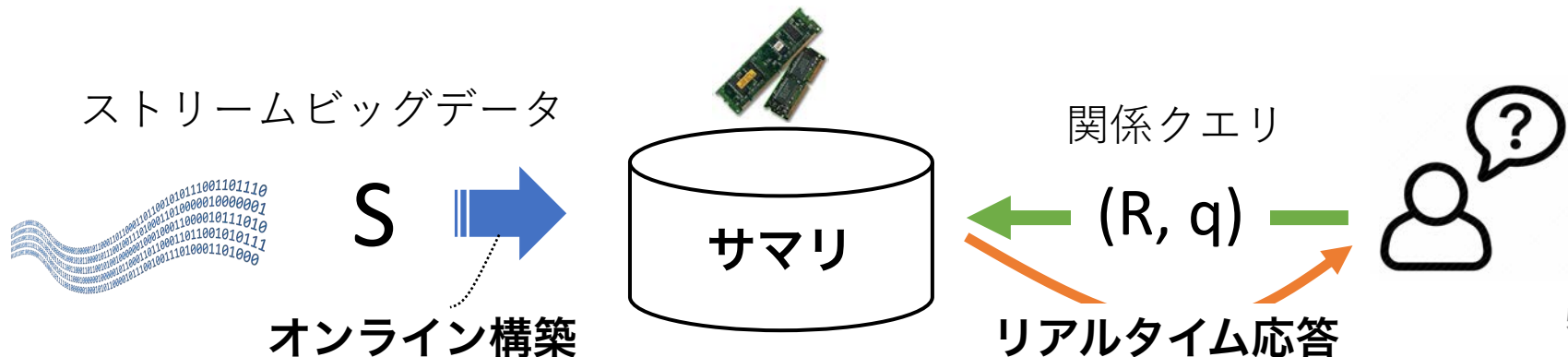
ストリームデータ分析

- ・ **ビッグデータ分野の重要課題** [IDC, 2015; 2018]
 - ハードディスク (+SSD) スキャンは原理的に困難
 - 到着データを “On-the-fly” でインメモリ処理する
 - 省メモリな**インメモリ管理技術**の開発が必要不可欠！



データ要約とは？

- **インメモリ処理**を可能にする**データ管理技術**
- サマリ: 特殊な関係クエリ (質問) に応答するデータ構造
 - 管理対象のデータ列: $S = \langle e_1, e_2, \dots, e_n \rangle$, n はデータ総数
 - 関係: R 🖱️ 問題に応じて設定
 - クエリ: q 🖱️ 何でもOK
- 2種類の関係クエリ (質問)
 - メンバーシップクエリ: q と関係 R を満たす e_i が S 中に存在した？
 - サポートクエリ: q と関係 R を満たす e_i が S 中に何回出現した？



発表の概要

- 研究背景
 - ストリームデータ処理
 - データ要約
- KMVを用いた頻度サマリの構築
- 頻度サマリを用いた時系列データ要約
 - 問題設定
 - アプローチ
- ケーススタディ
 - 測光時系列データの頻度分布
 - エビデンスに基づく異常検知

頻度サマリとは

- バイナリ (属性値が0もしくは1) データ集合の要約表現
- 頻度サマリのタスク
 - 入力: クエリ (属性の集合)
 - 出力: クエリの**頻度**

属性数 m

時刻 t	属性 a_1	属性 a_2	属性 a_3	...	属性 a_m
1	1	0	1	...	0
2	1	1	0	...	1
3	1	1	1	...	1
⋮	⋮	⋮	⋮	...	⋮
n	1	0	0	...	0

クエリの各属性値が1となるデータ数

例. $q = \{a_1, a_3\}$ のとき,
 $a_1 = 1$ and $a_3 = 1$ となるデータ数が
 q の頻度となる

クエリ q

線形探索の場合

時間計算量 $O(n)$

(データ数に比例して
実行時間が増加する)

頻度 $f(q)$

頻度サマリのアイデア

- バイナリデータの垂直配置
- 各属性の出現時刻集合をスケッチ化

バイナリデータ (a.k.a. トランザクション) の集合

KMVに基づく頻度サマリの構築 (1/2)

- 頻度サマリのアイデア (その1)

属性毎に値が1となるデータ出現時刻を保持する
(バイナリデータ集合の垂直配置)

元のデータ構造

時刻 t	属性 a_1	属性 a_2	属性 a_3	...	属性 a_m
1	1	0	1	...	0
2	0	1	0	...	1
3	1	0	1	...	1
...	...	...	...	...	...
n	0	0	0	...	0

垂直配置

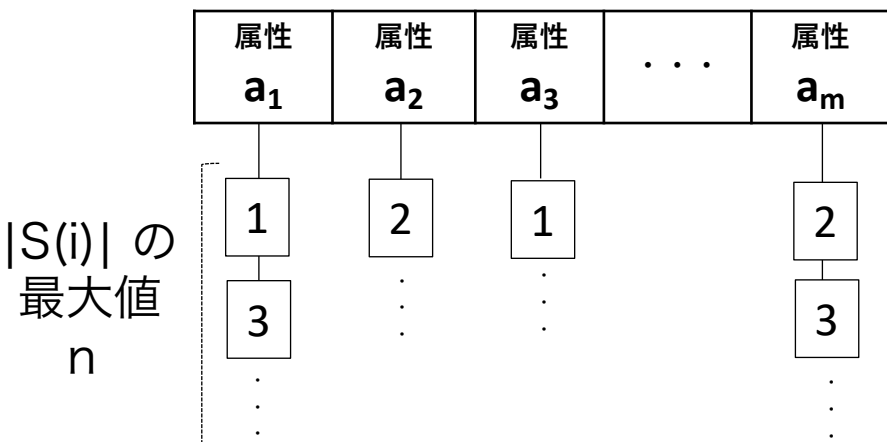


(非出現時刻は記録しない)

KMVに基づく頻度サマリの構築 (2/2)

- 頻度サマリのアイデア (その2)

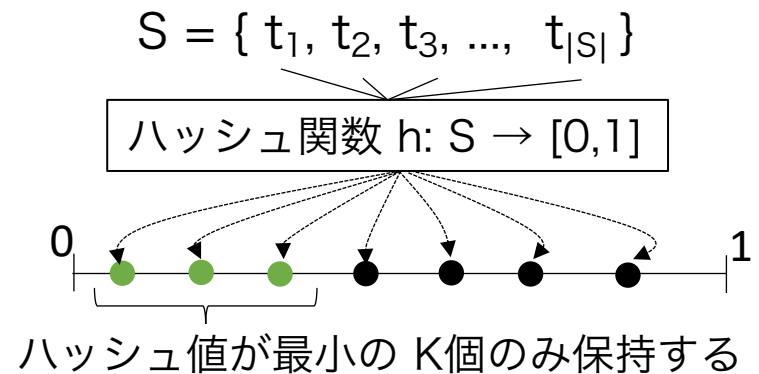
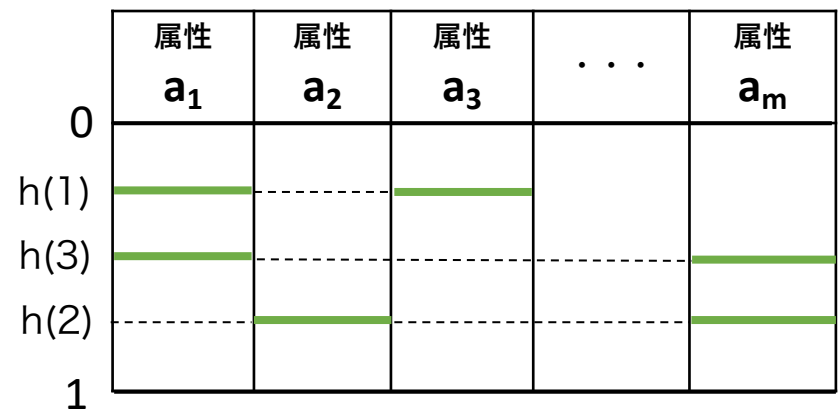
各属性の出現時刻集合を KMV を用いてスケッチ化



各属性 a_i の出現時刻集合を S_i と書く
 S_i には多数の時刻が保持される



K-Minimal Values (KMV)
を用いて 各 S_i を管理し、要約



KMVに基づく頻度推定

- S_i : 属性 a_i の出現時刻集合
- L_i : S_i の KMV (要素数K)
- U_i : L_i の最大値
- q : クエリ $\{a_{q,1}, a_{q,2}, \dots, a_{q,r}\}$
- $\hat{f}(q)$: q の頻度推定値

推定式 [Wang et al., 2021]

$$\hat{f}(q) \Leftrightarrow \frac{K_{\cap}}{\tilde{U}}, \text{ where}$$

$$K_{\cap} = |L_{q,1} \cap L_{q,2} \cap \dots \cap L_{q,r}|,$$

$$\tilde{U} = \max(U_{q,1}, U_{q,2}, \dots, U_{q,r})$$

$\hat{f}(q)$ の期待値は $f(q)$ と一致する

ハッシュ関数とKMVの要素数 (K) を増やせば誤差を平均的に小さくできる

元データ

t	a ₁	a ₂	a ₃	a ₄
1	1	0	1	1
2	0	1	0	0
3	1	1	1	1
4	1	1	0	0
5	1	0	1	0

頻度サマリ

a ₁	a ₂	a ₃	a ₄
0.4 0.3	0.2 0.3	0.4 0.3	0.4 0.3
0.5 0.7	0.5	0.7	

K=2 の場合

クエリ

$q = \{a_1, a_2\}$

$$\hat{f}(q) = \frac{1}{0.4} = 2.5$$

$$f(q) = 2$$

出力誤差0.5

ハッシュ値

h(1) = 0.4
h(2) = 0.2
h(3) = 0.3
h(4) = 0.5
h(5) = 0.7

発表の概要

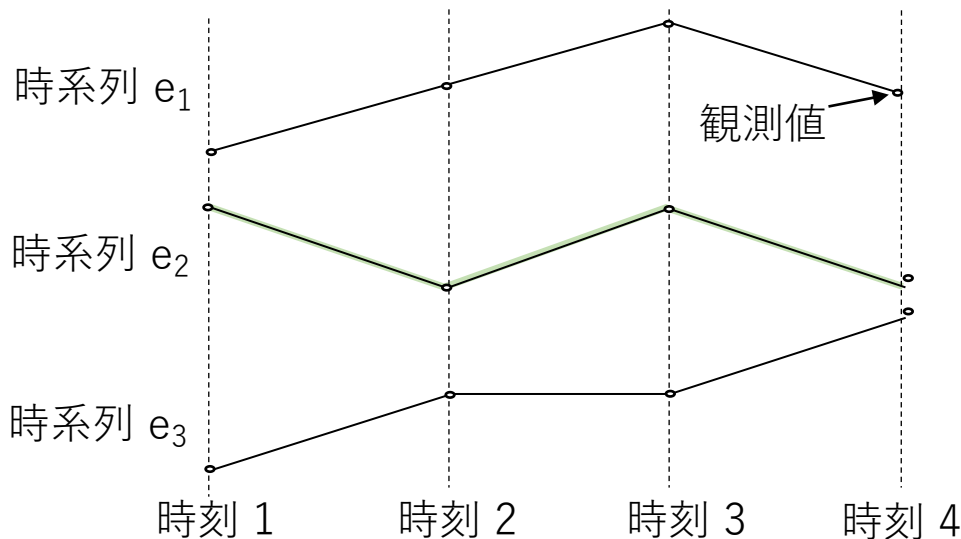
- 研究背景
 - ストリームデータ処理
 - データ要約
- KMVを用いた頻度サマリの構築
- 頻度サマリを用いた時系列データ要約
 - 問題設定
 - アプローチ
- ケーススタディ
 - 測光時系列データの頻度分布
 - エビデンスに基づく異常検知

時系列データ要約とは

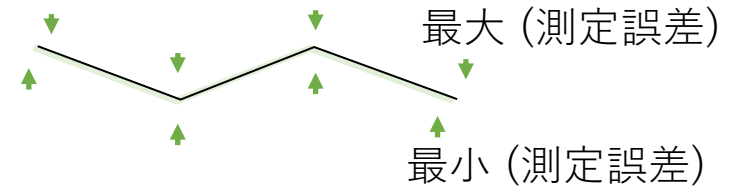
- 時系列データの集合を対象とする頻度サマリ
- 頻度サマリのタスク
 - 入力: クエリ (時系列) + 測定誤差
 - 出力: クエリの頻度、

クエリの各値 $\pm$ 測定誤差の範囲内にある時系列データの個数

時系列データの集合 $S = \langle e_1, e_2, e_3 \rangle$



クエリ (時系列) q + 測定誤差 r



$q \pm r$ に含まれる時系列は何回出現?

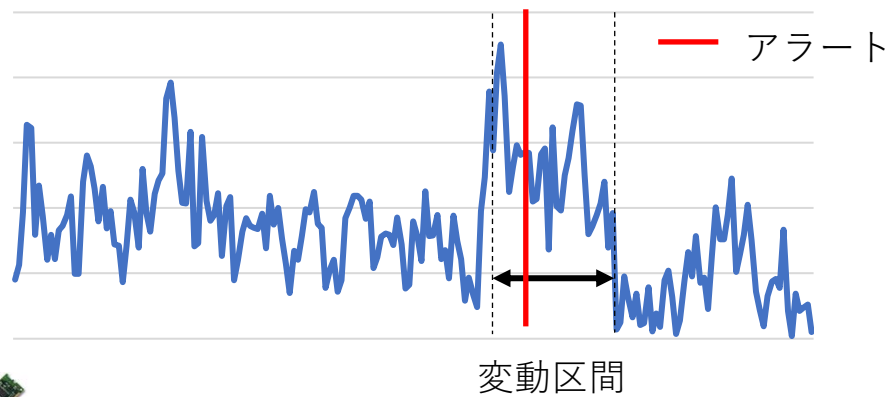
頻度サマリ

1回

定義 e は $q \pm r$ に含まれる $\Leftrightarrow$
 $\forall t (q(t) - r \leq e(t) \leq q(t) + r)$

- これまでに見たことのない**新種**パターンをリアルタイム検知する問題

実データ+ 変動パターン

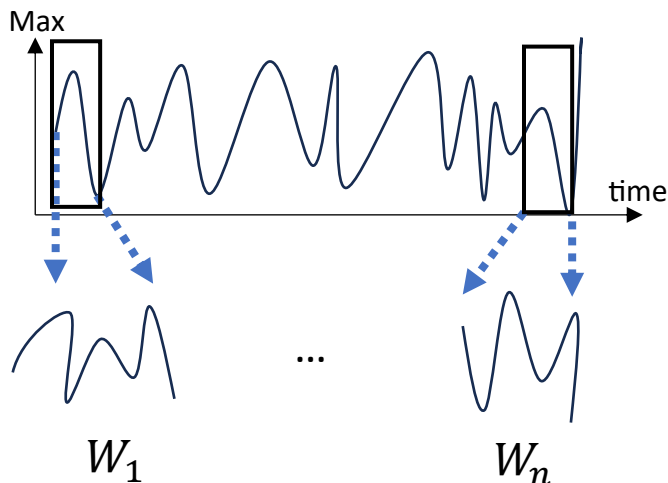


頻度サマリを用いた時系列データ要約

アプローチ

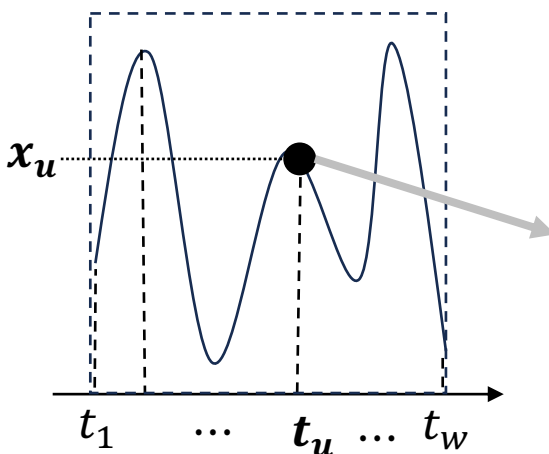
1. 時系列データをバイナリデータに変換
 - 時系列の各値を正規化 (例. Min-Max法, SAX)
 - 各値の上限と下限をバイナリとして符号化
2. バイナリデータの集合から頻度サマリを構築

正規化済み時系列



スライディングウィンドウ
による時系列切り出し

ウィンドウ



ウィンドウ幅 = w

x_u の上限を符号化

0	1	2	3	4	5	6	7	8
0	0	0	0	0	1	1	1	1

x_u の下限を符号化

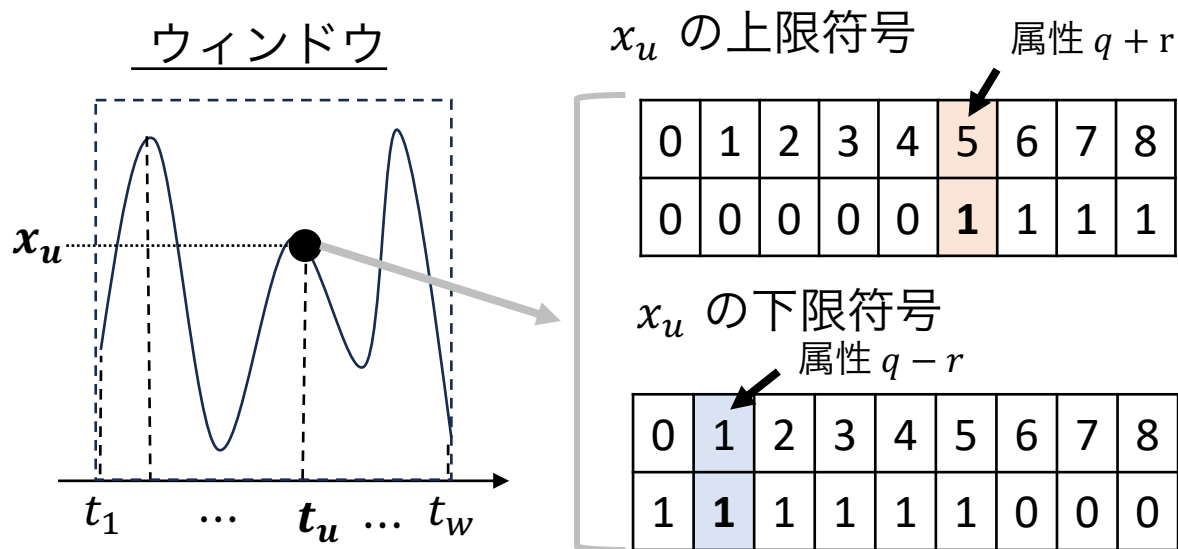
0	1	2	3	4	5	6	7	8
1	1	1	1	1	1	0	0	0

最大値 (Max) = 8
 $x_u = 5$ の場合 17

頻度サマリを用いた時系列データ要約

アプローチ

1. 時系列データをバイナリデータに変換
 - 時系列の各値を正規化 (例. Min-Max法, SAX)
 - 各値の上限と下限をバイナリとして符号化
2. バイナリデータの集合から頻度サマリを構築



上限符号の属性 $q+r$ と
下限符号の属性 $q-r$ の
値が共に1である $\leftrightarrow$

$$q-r \leq x_u \leq q+r$$

例. $q=3, r=2$ のとき

$$1 \leq x_u \leq 5$$

が成り立つ

最大値 (Max) = 8

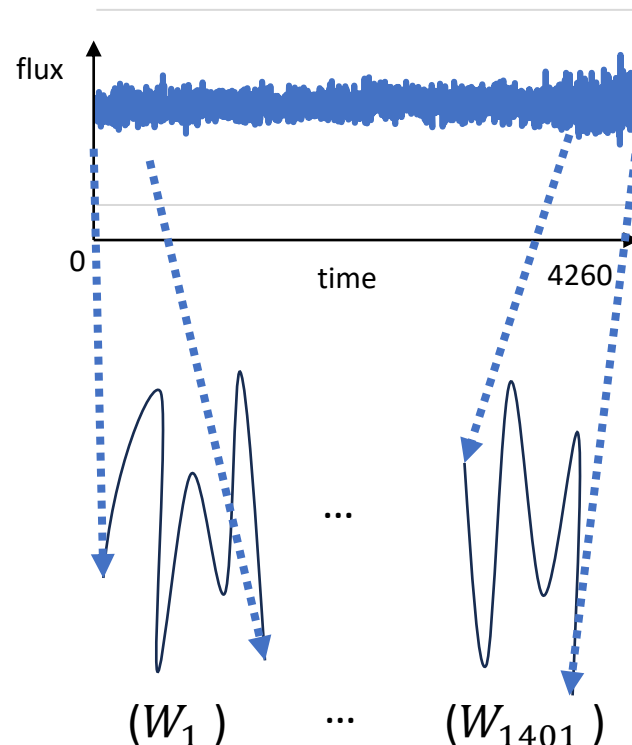
$x_u = 5$ の場合

発表の概要

- 研究背景
 - ストリームデータ処理
 - データ要約
- KMVを用いた頻度サマリの構築
- 頻度サマリを用いた時系列データ要約
 - 問題設定
 - アプローチ
- ケーススタディ
 - 測光時系列データの頻度分布
 - エビデンスに基づく異常検知

Tomo-eデータを用いた予備実験 (1/3)

- 対象ライトカーブ (檜山さんから提供)
 - 108天体*[1]
- ファイルごとに前処理
 - 正規化
 - Min-Max法 : [0~255]に設定
 - ウィンドウの切り出し法
 - ウィンドウサイズ : 60
 - スライド幅 : 3



全ウィンドウ数: $1401\text{windows} \times 108\text{天体} = \mathbf{151,308\text{windows}}$

Tomo-eデータを用いた予備実験 (1/3)

例

$W_1 = \{5032, 5144, 5674, 4867, 4336, \dots, 5232, 4962\}$

$W_2 = \{4867, 4336, 5144, 5313, 4605, \dots, 5213, 5122\}$

$\vdots$

$W_{1401} = \{5546, 4365, 5754, 5363, \dots, 5875, 5456\}$

各Windowは
シーケンシャル

60timestamps

各値を[0~255]*に正規化

大きな数値を扱いやすくする

Query(ウィンドウ)

測定誤差: **10%**

(*絶対誤差)

$W_u = \{75.5, 55.5, 95.5, 85.5, \dots, 65.5, 45.5, 55.5\}$

上限

↑ ↑ ↑ ↑ +25.5 ↑ ↑ ↑

Query: $W_q = \{50.0, 30.0, 70.0, 60.0, \dots, 40.0, 20.0, 30.0\}$

↓ ↓ ↓ ↓ -25.5 ↓ ↓ ↓

$W_l = \{24.5, 4.50, 44.5, 34.5, \dots, 14.5, 0.00, 4.50\}$

下限

Tomo-eデータを用いた予備実験 (1/3)

測定誤差: **10%**
(*絶対誤差)

$W_u = \{75.5, 55.5, 95.5, 85.5, \dots, 65.5, 45.5, 55.5\}$

最大

↑ ↑ ↑ ↑ +25.5 ↑ ↑ ↑

Query: $W_q = \{50.0, 30.0, 70.0, 60.0, \dots, 40.0, 20.0, 30.0\}$

↓ ↓ ↓ ↓ -25.5 ↓ ↓ ↓

$W_l = \{24.5, 4.50, 44.5, 34.5, \dots, 14.5, 0.00, 4.50\}$

最小

Query

ウィンドウの集合

$W_l(\text{誤差下限}) \leq \text{対象ウィンドウ} \leq W_u(\text{誤差上限})$

$W_1 \sim W_{150,308}$ のうちに測定誤差範囲内の W が何件発生しているのか？

Tomo-eデータを用いた予備実験 (1/3)

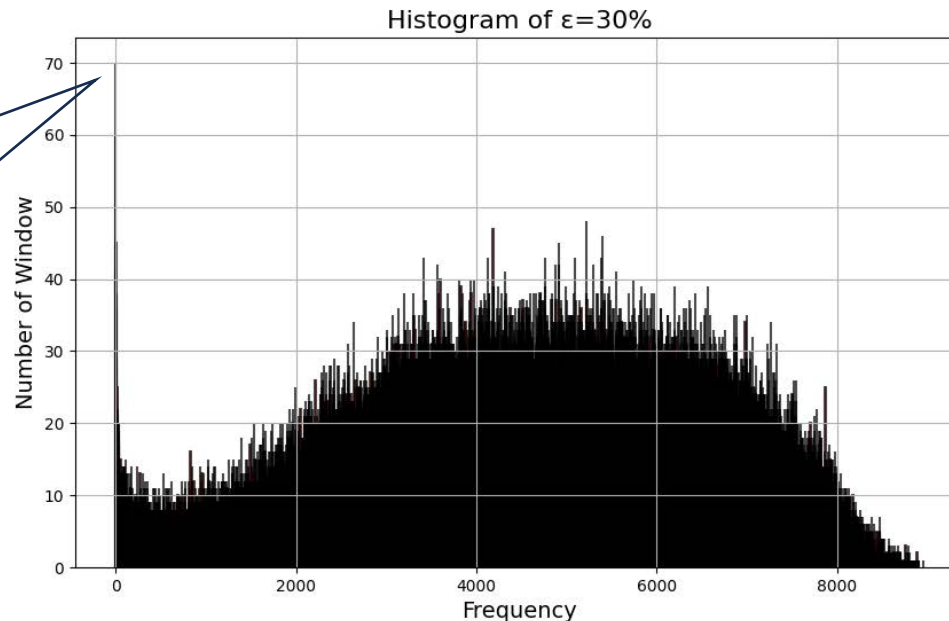
- 頻度分布: 各ウィンドウの頻度 (真値) 算出

全150,308ウィンドウの中で, 自身以外のウィンドウ(150,307ウィンドウ)に対して, 自身がクエリとして**頻度**を求める

どのウィンドウ
にもマッチしな
かった

70件

150,308件



測定誤差: 30%

このような頻度分布のデータセットに対して, 調べたいQueryを投げる
→頻度が低ければ, レアな現象(突発現象)

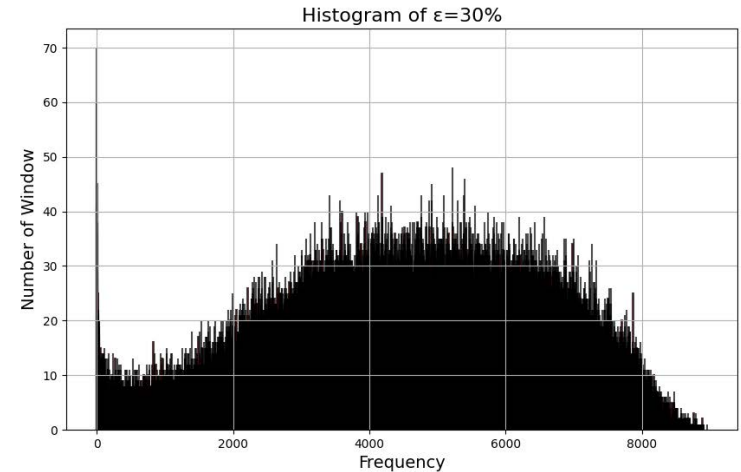
Tomo-eデータを用いた予備実験 (2/3)

- 予備実験における目的

- 真値による頻度分布をもとに, 突発信号をクエリとして頻度の算出

- 対象データ

- 実フレア (Tomo-e) *[2]
 - 逢澤さんから提供
- 人工フレア (Keppler) *[3]
 - Thanapolさんから提供



フレアのウィンドウをクエリとして, 頻度分布にて比較
→頻度が低ければ, レアな現象(突発現象)

[2] M. Aizawa, K. Kawana, K. Kashiya, R. Ohsawa, H. Kawahara, F. Naokawa, T. Tajiri, et al. Fast optical flares from M dwarfs detected by a one-second-cadence survey with Tomo-e Gozen. Publications of the Astronomical Society of Japan (PASJ), 74(5):1069–1094, 2022.

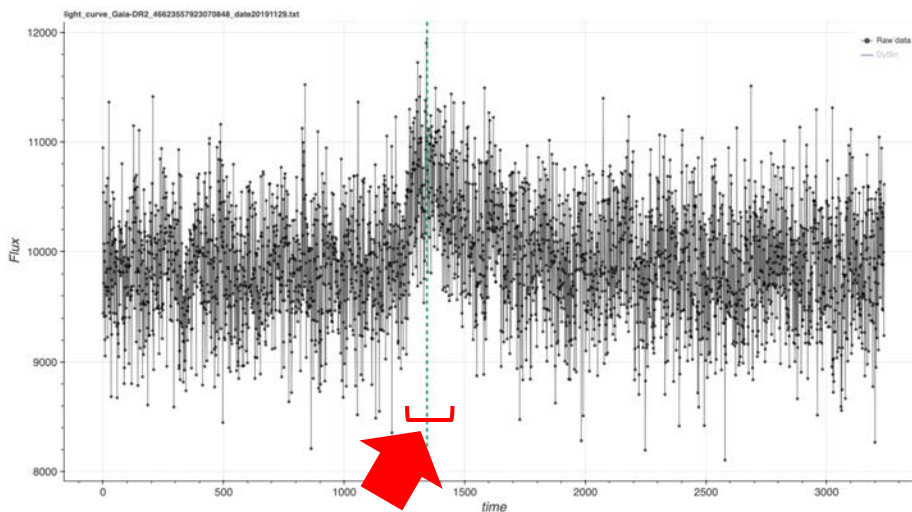
[3] T. Phungtua-eng et al., Elastic Data Binning for Transient Pattern Analysis in Time-Domain Astrophysics, *ACM/SIGAPP Symposium on Applied Computing*, 2022.12

Tomo-eデータを用いた予備実験 (2/3)

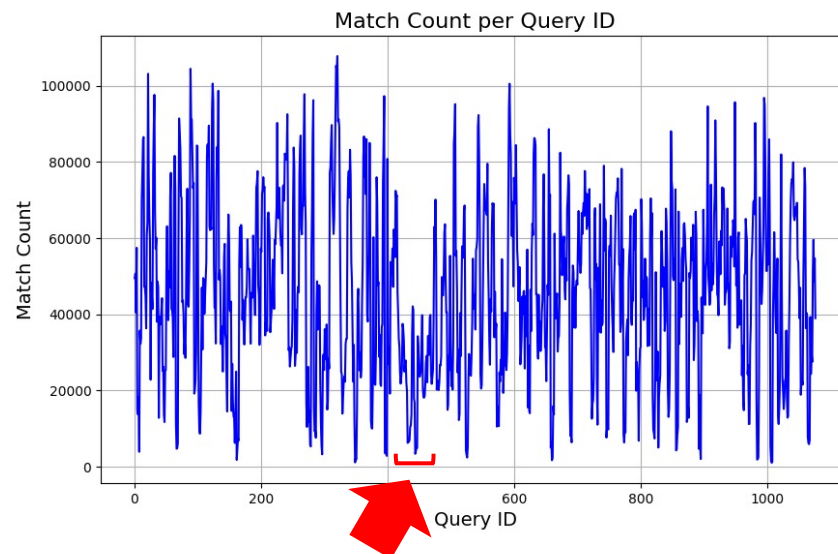
- ・ 実フレア (Tomo-e)

“light_curve_Gaia-DR2_46623557923070848_date20191129”

フレア以外で頻度が下がっている原因は, 上下が激しいデータだから？



フレア開始



区間として頻度が落ちている

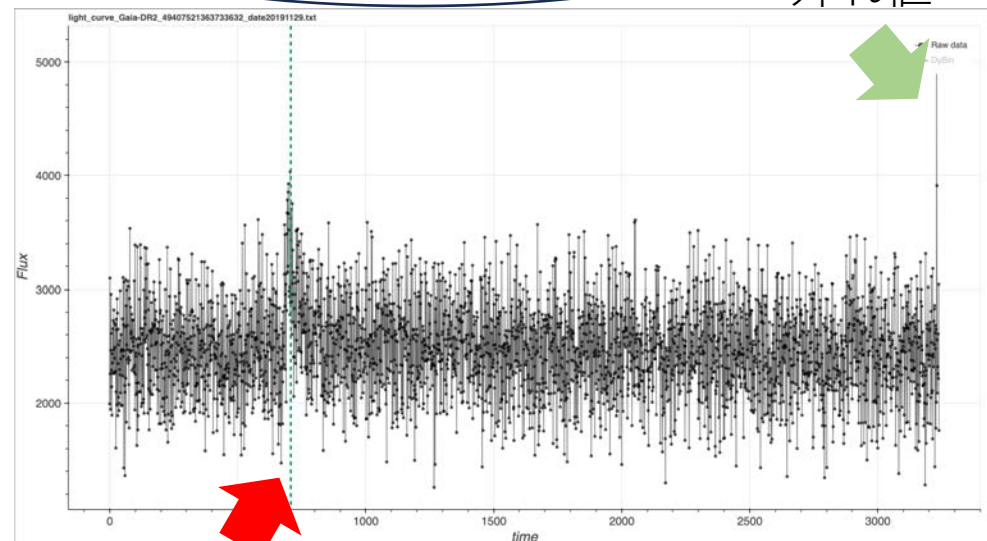
Tomo-eデータを用いた予備実験 (2/3)

- ・ 実フレア (Tomo-e)

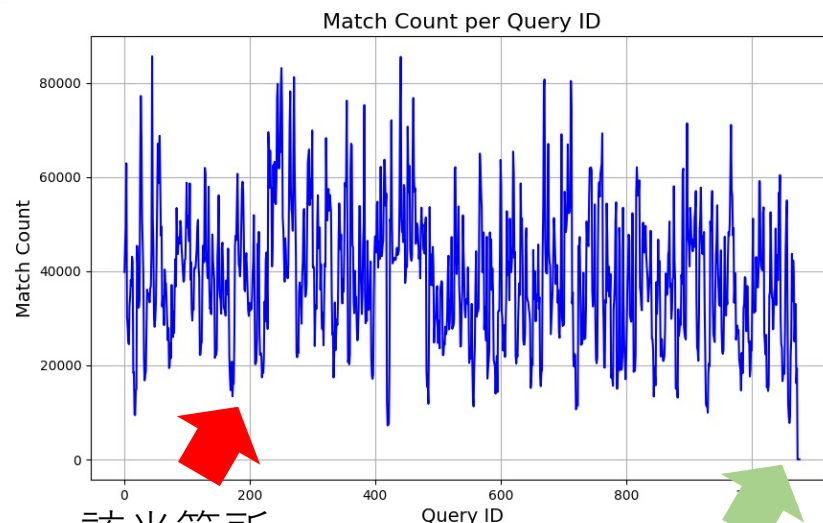
“light_curve_Gaia-DR2_49407521363733632_date20191129”

これが影響している

外れ値



フレア開始



該当箇所

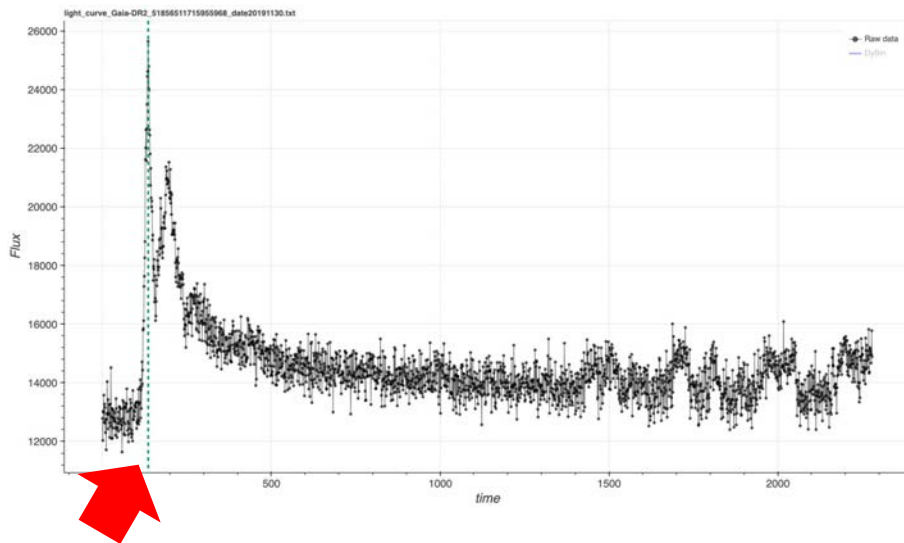
頻度0

Tomo-eデータを用いた予備実験 (2/3)

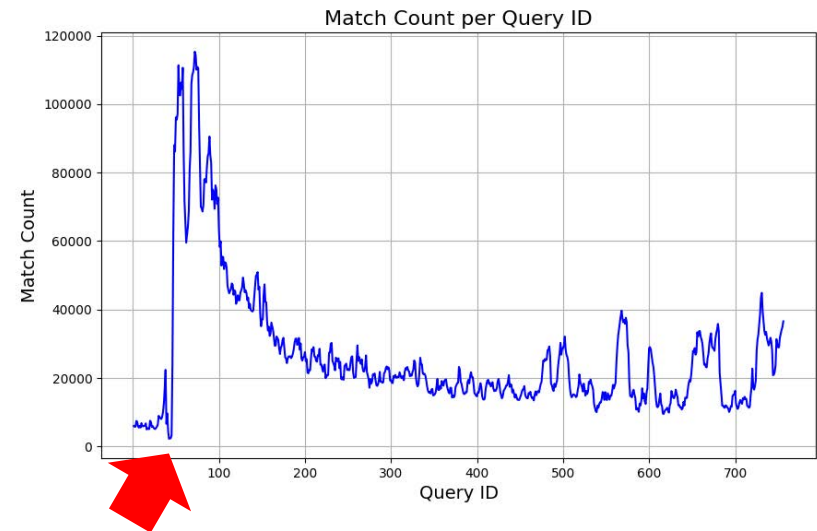
- 実フレア (Tomo-e)

“light_curve_Gaia-DR2_51856511715955968_date20191130”

似たようなグラフ結果



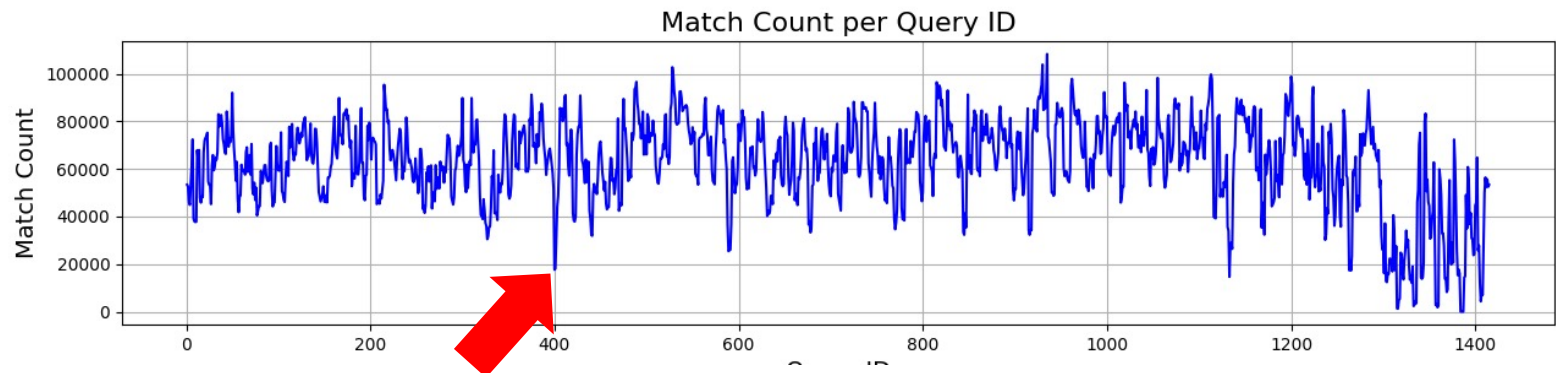
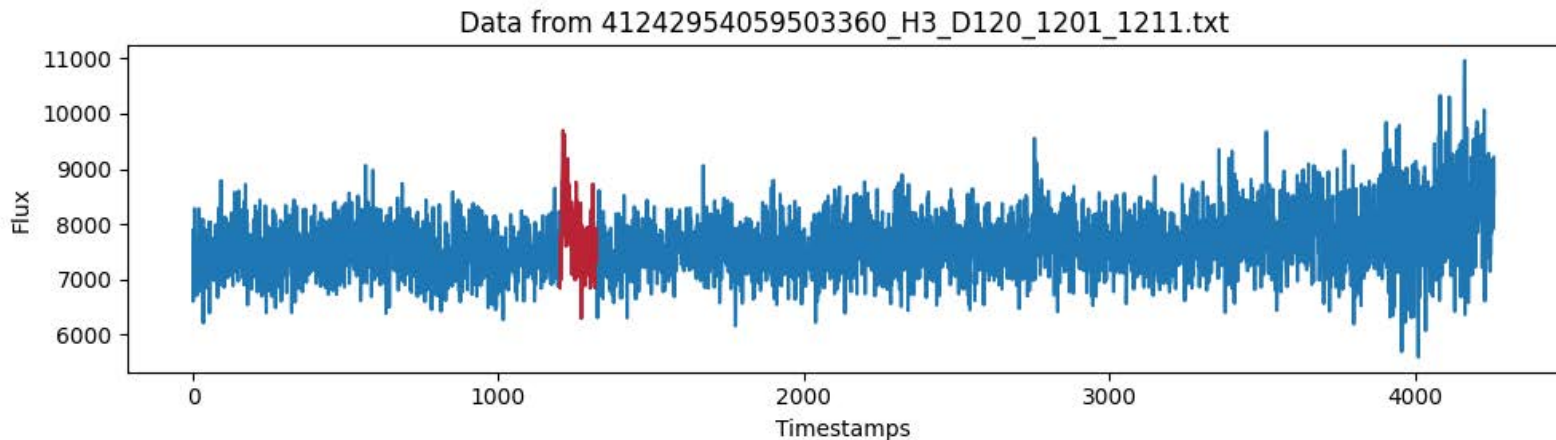
フレア開始



頻度が非常に低い

Tomo-eデータを用いた予備実験 (2/3)

- 人エフレア (Keppler)



頻度が落ちている

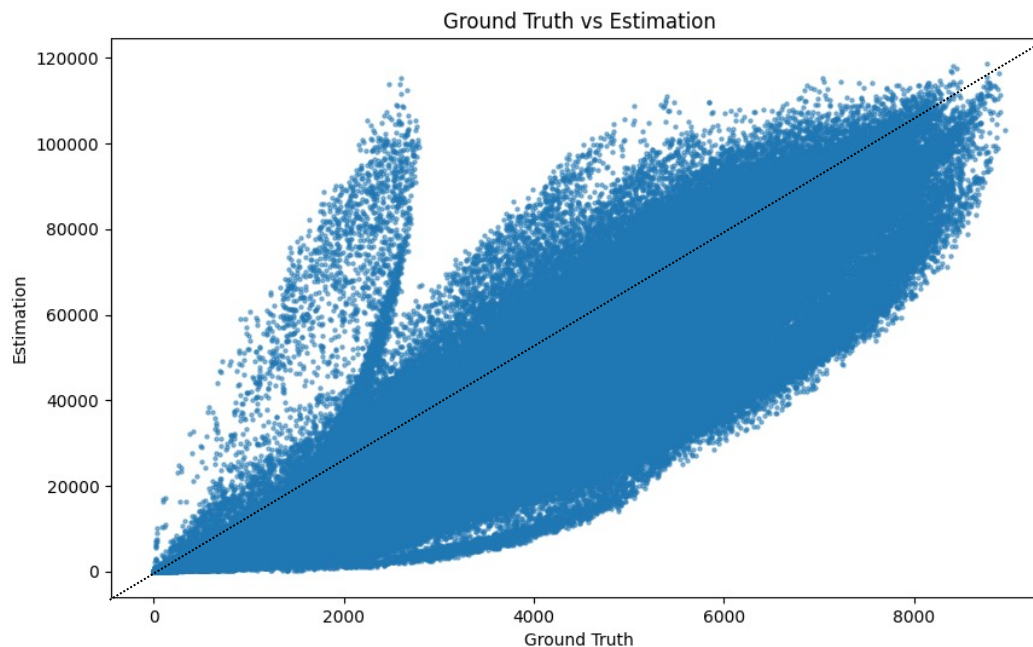
Tomo-eデータを用いた予備実験 (3/3)

- 真値と推定値の対角グラフ

KMV(データスキッチング)

自身以外の150,107ウィンドウのうち, K=5,000で頻度推定

測定誤差: 30%



線に近いほど
誤差**低**

ハッシュ関数とKMVの要素数 (K) を増やせば, 誤差を平均的に小さくできる

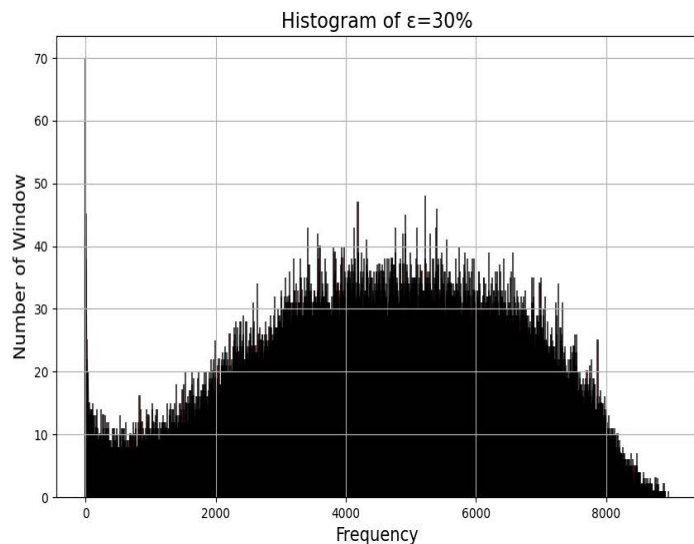
Tomo-eデータを用いた予備実験 (3/3)

- 真値と推定値の頻度分布

KMV(データスキッチング)

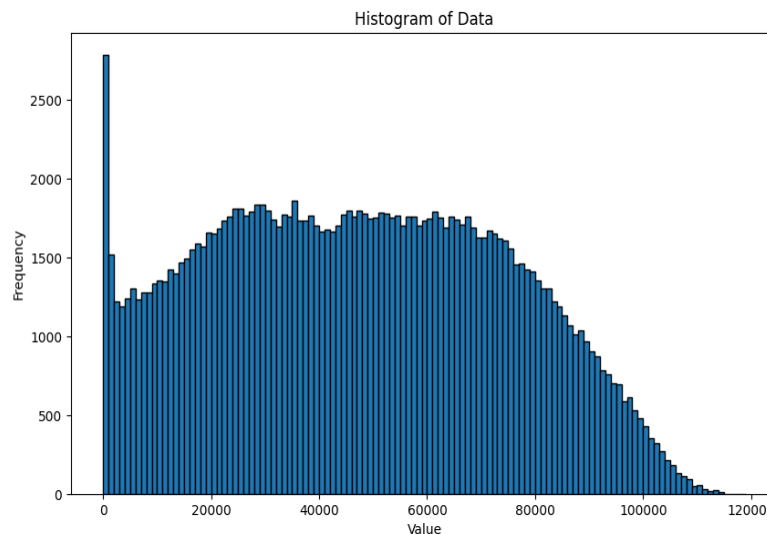
自身以外の150,107ウィンドウのうち, $K=5,000$ で頻度推定

真値



推定値

測定誤差: 30%



スケールを伸ばしたら, 似たような頻度分布
→ 少量のデータ($K=5000$)で**突発検知の可能性**

まとめと今後の課題

- まとめ

- KMVに基づく頻度サマリの紹介
- 時系列データ要約の提案
- ライトカーブデータを用いた予備実験
 - 突発クエリの頻度は低くなる
 - 真値と推定値の誤差の確認

- 今後の課題

- 推定値を用いた突発検知
- 出力誤差の理論値と実験結果の照合