

大規模データストリームの劣線形要約 とリアルタイム異常検知への応用

静岡大学 情報学部 情報科学科
准教授 山本 泰生

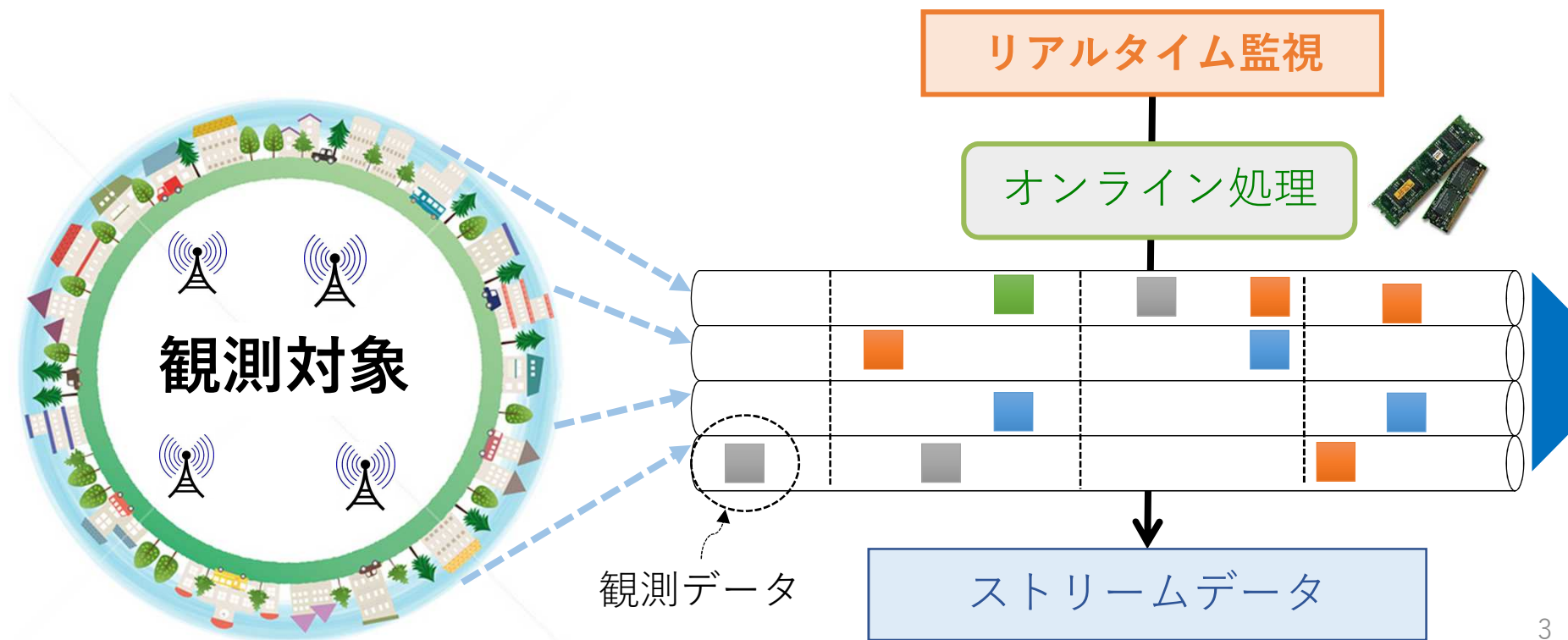
2024年2月8日

発表の概要

- 背景
 - ストリームデータ処理
 - データ要約
- 技術: 時系列データの要約
 - PSS-TS
- 応用: エビデンスに基づく異常検知
 - センシングデータを用いた行動認証
 - データ駆動の科学発見

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)



IoTとリアルタイム分析

- 2025年予想
 - ストリームデータが全体に占める割合: 25% [IDC, 2018]
 - 経済効果予測: **270~620兆円** [McKinsey, 2013]
 - 👉 ヘルスケア (41%), 製造 (33%), エネルギー (7%), 交通・インフラ・農業・小売 (15%)
- データ分析の3層構造 [M. Mohammadi et al., 2018]



IoTとリアルタイム分析

- 2025年予想

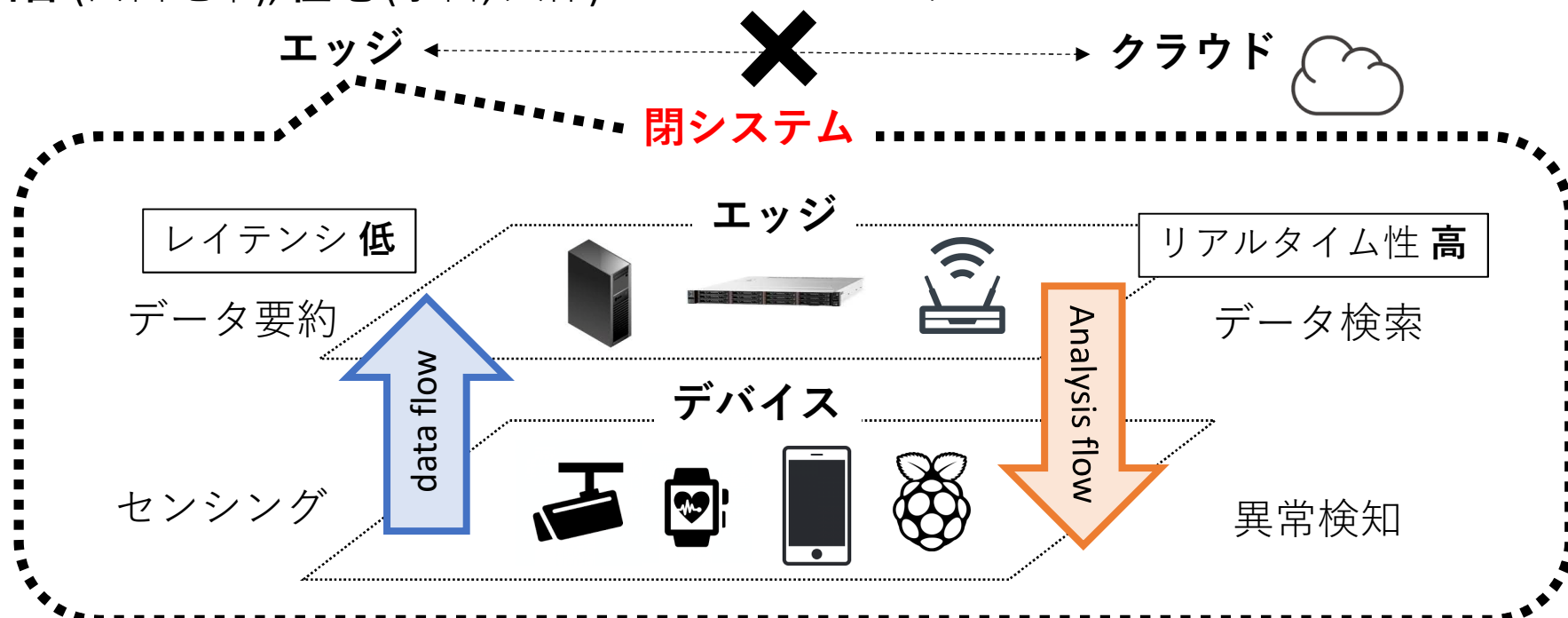
- ▶ ストリームデータが全体に占める割合: 25% [IDC, 2018]

- ▶ 経済効果予測: 270~620兆円 [McKinsey, 2013]

- 👉 ヘルスケア, 製造, エネルギー, 交通, インフラ, 農業, 小売

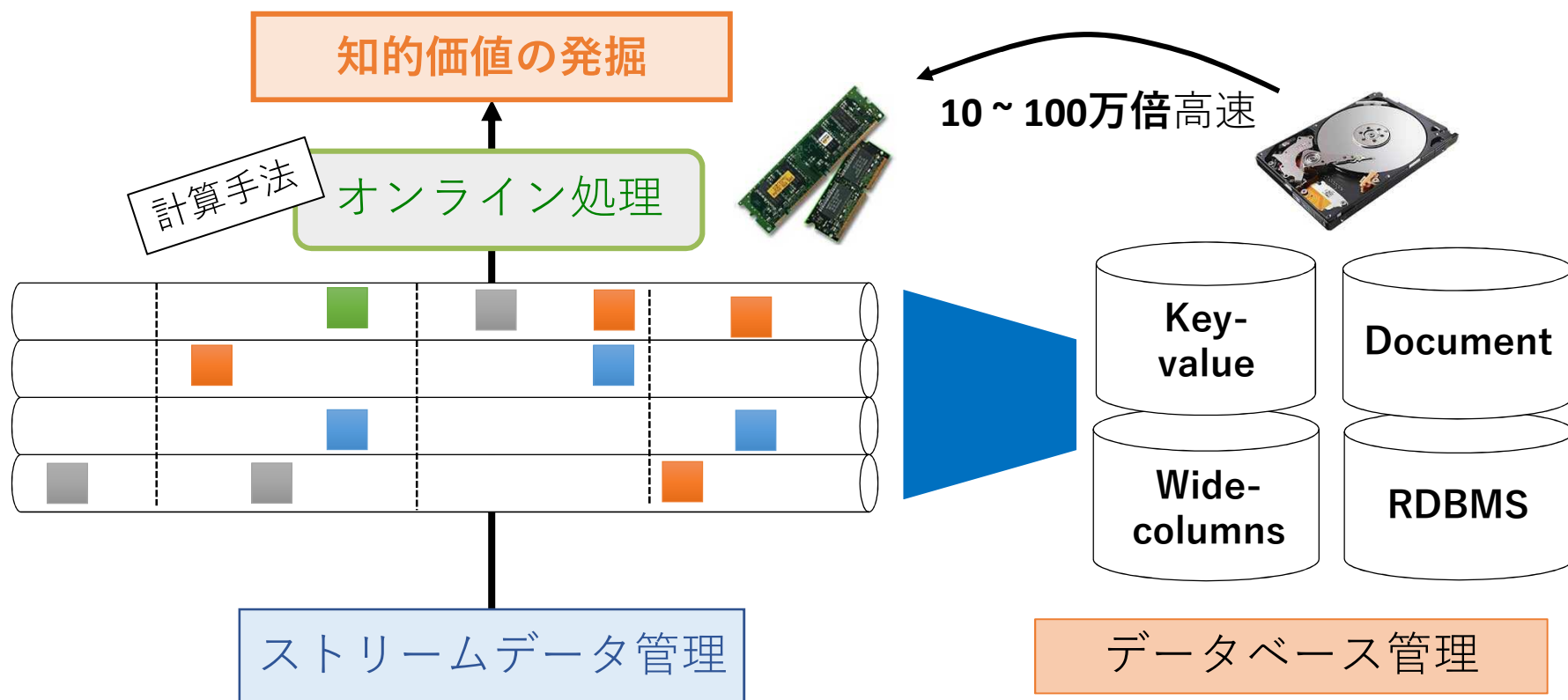
- データ分析の3層構造 [M. Mohammadi et al., 2018]

田舎 (山岳地帯), 極地 (宇宙, 人体) ネットワーク帯域 弱



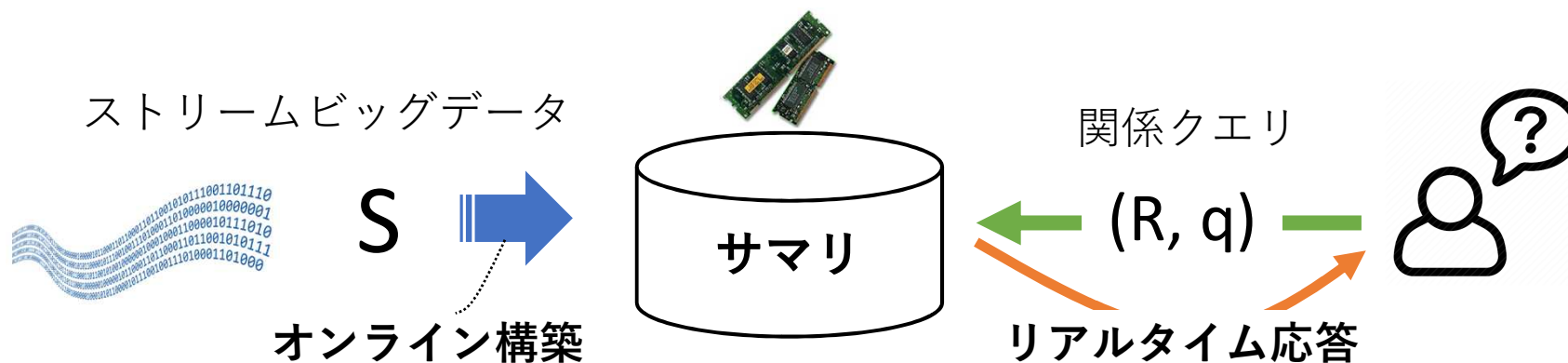
エッジ上のストリームデータ分析

- 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革新的なデータ管理技術
- 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種類のクエリ
 - 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?



関係データの例



「からから」 関係 R を次のように想定

任意の2日 e_i, e_j に対し e_i より e_j の方が「からから」 $R(e_i, e_j)$ 定義 $a_1^i \leq a_1^j$ かつ $a_2^i \geq a_2^j$
 気温が高く 降水量が低い

気象データ列

S	気温 a_1	降水量 a_2	風速 a_3	風向 a_4	日照 a_5
e_1	14	5	3	0	10
e_2	11	10	1	0	5
e_3	19	5	2	1	14
e_4	13	3	2	1	14

以下の2つのうち
成り立つものはどちら?

$R(e_2, e_4)$?

$R(e_4, e_1)$?

関係データの例 cont.

クエリ $q = (12, 5, -, -, -)$ のとき (ただし-は空)

メンバーシップクエリの意味

$R(e_i, q)$ となる e_i が S に存在するか?



S	気温 a_1	降水量 a_2	風速 a_3	風向 a_4	日照 a_5
e_1	14	5	3	0	10
e_2	11	10	1	0	5
e_3	19	5	2	1	14
e_4	13	3	2	1	14

気温 12 °C 以上 & 降水量 5 mm
以下となる日は過去にあった?



Yes

関係データの例 cont.



クエリ $q = (12, 5, -, -, -)$ のとき (ただし-は空)

サポートクエリの意味

$R(e_i, q)$ となる e_i が S に何回存在した?

S	気温 a_1	降水量 a_2	風速 a_3	風向 a_4	日照 a_5
e_1	14	5	3	0	10
e_2	11	10	1	0	5
e_3	19	5	2	1	14
e_4	13	3	2	1	14

気温 12 °C 以上 & 降水量 5 mm
以下だった過去の日数?



3日

サマリに関する研究動向 (1/2)

- 推定値に誤差を許容する近似サマリの研究が進む

q : クエリ, $f(q)$: q の真値, $\hat{f}(q)$: q の推定値とすると

右式を満たす: $\Pr(|f(q) - \hat{f}(q)| \leq \varepsilon) \geq 1 - \delta$

出力誤差

許容誤差

誤り発生確率

- メンバーシップサマリ: 空間計算量の下界は $\Omega(n)$

☞ データサイズ n 未満の近似サマリは存在しない

- サポートサマリ: 多くの劣線形近似サマリが発見されている
関係 R のタイプに応じてサポートサマリ (SS) を3つに分類

➤ 等価関係を扱う Equivalence SS (ESS)

クエリの出現頻度を答えるサマリ


➤ 全順序関係を扱う Linear order SS (LSS)

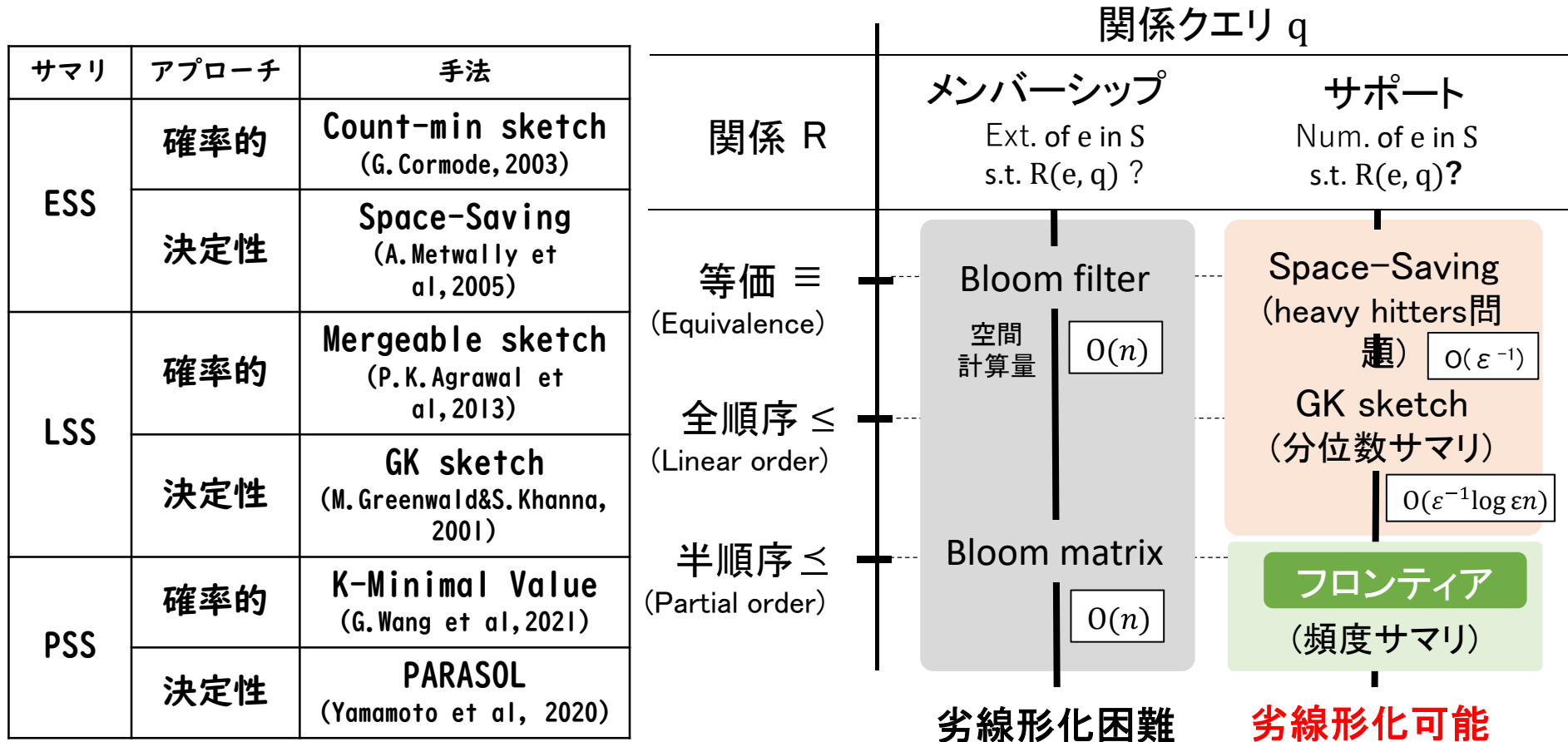
クエリのランキングを答えるサマリ

➤ 半順序関係を扱う Partial order SS (PSS)

クエリを含意するデータ数を答えるサマリ

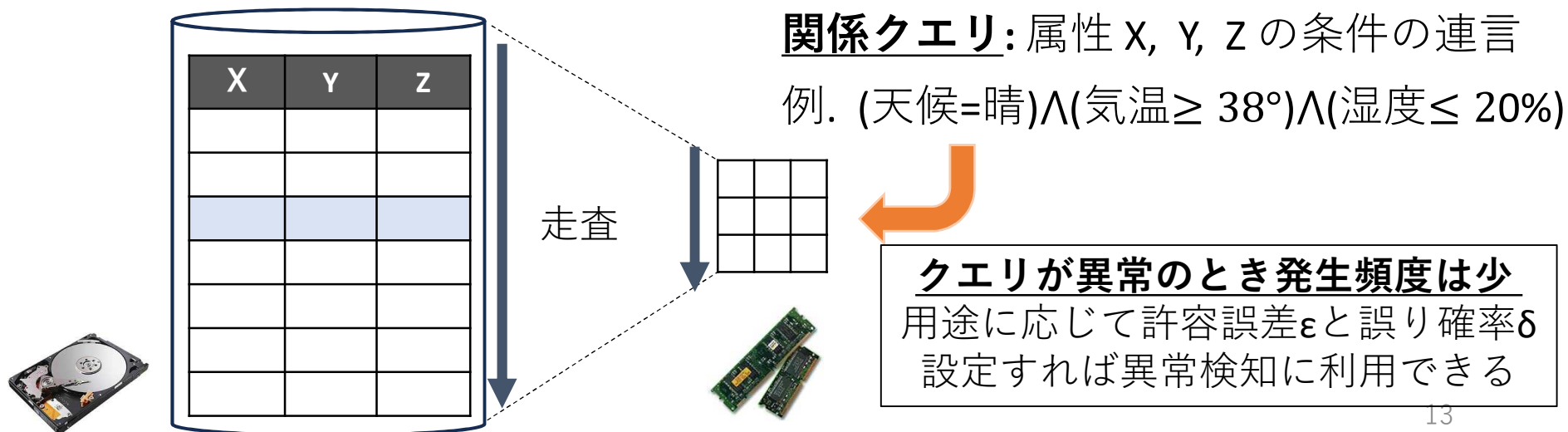
サマリに関する研究動向 (2/2)

- サマリ全体とサポートサマリの各クラスの現状
 - ESS と LSS: 確率的/決定性双方の優れたアルゴリズムが発見。機械学習やプライバシー保護の課題で多くの活用事例あり
 - PSS: 劣線形サマリを構築する決定性アルゴリズムは未探索。活用事例も少ない  応用領域はフロンティア



PSSのインパクト

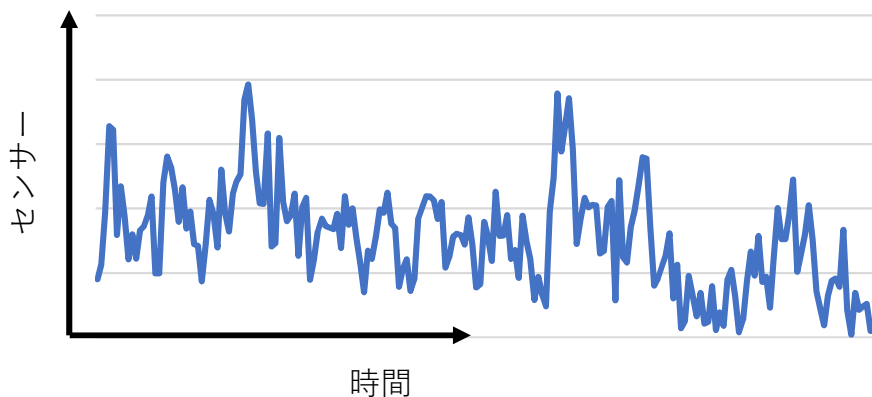
- データ量によらずストリーム全体をインメモリ管理する
超軽量なデータ構造を用いたリアルタイム検索
- 任意のクエリと関係を結ぶデータの発生頻度に応答する
関係表現を用いたリッチなクエリ設定
エビデンスを伴うクエリの異常度判定



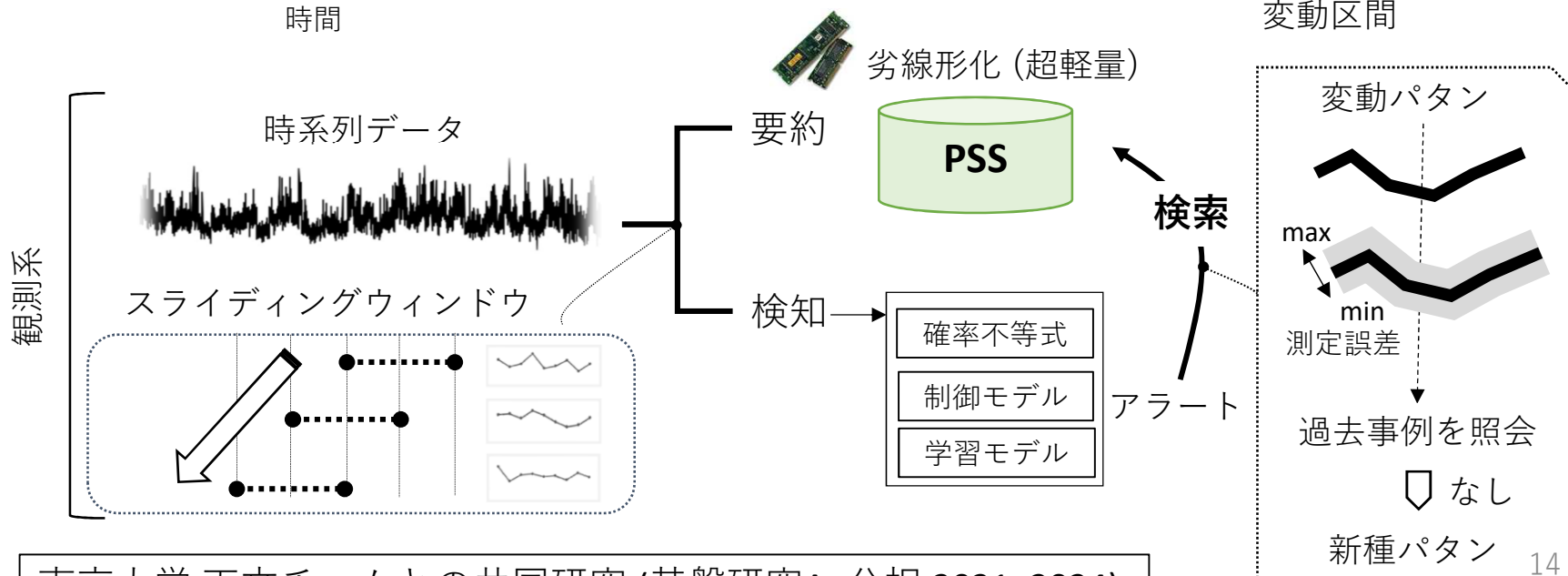
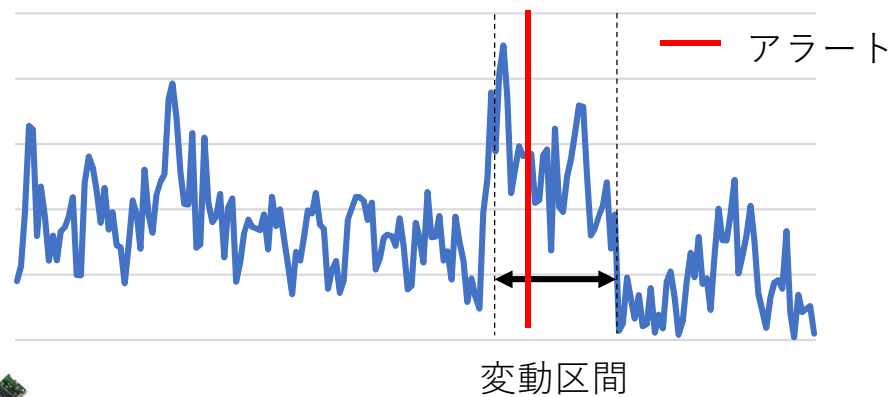
時系列データの場合

- 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 (時系列データ)



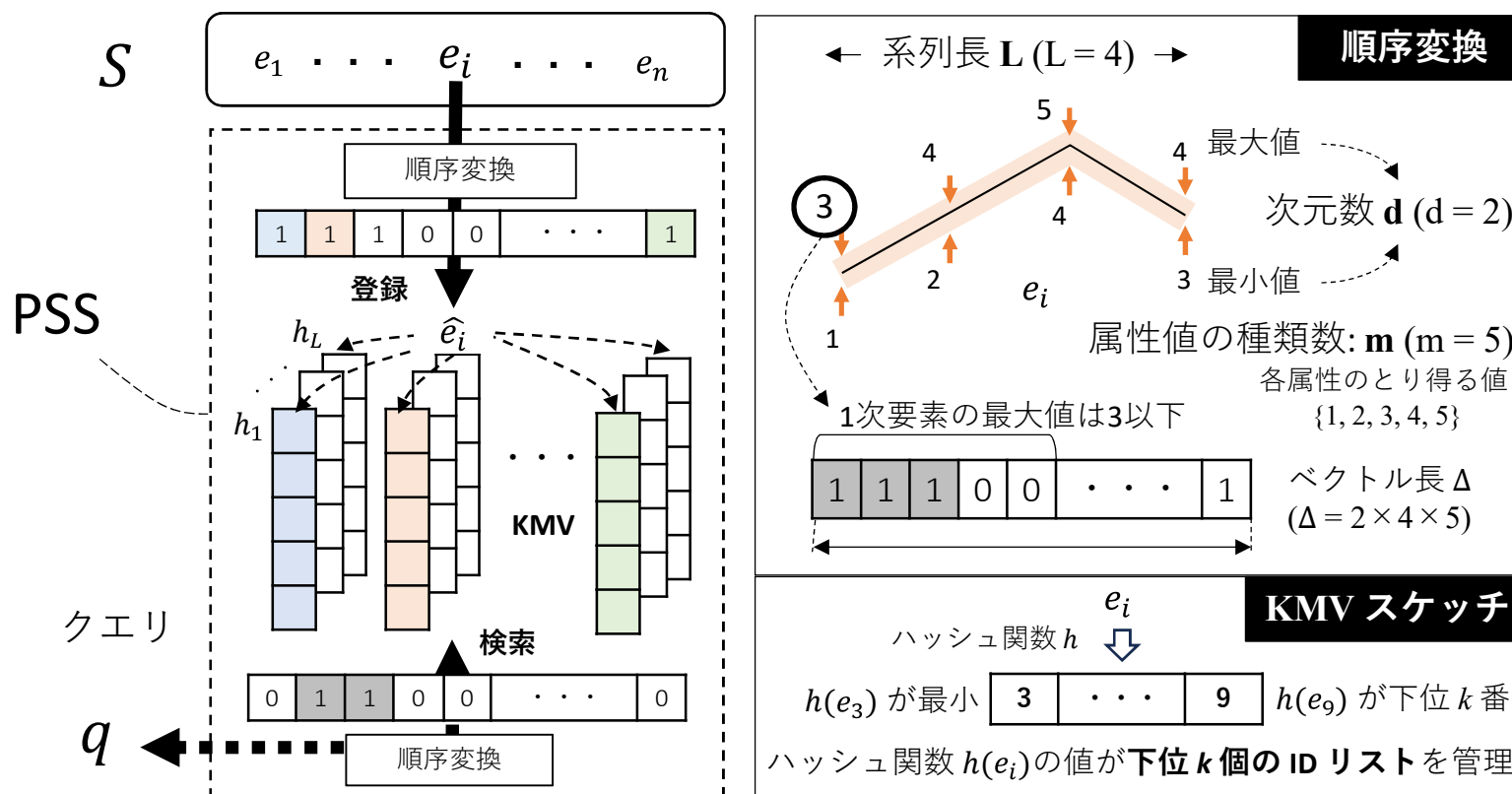
実データ+矩形波 (変動パターン)



東京大学 天文チームとの共同研究 (基盤研究A, 分担 2021~2024)

当研究室の取り組み（コア技術）

- 劣線形な頻度サマリ（PSS）アルゴリズムの開発
- **PSSを用いた時系列データ要約（PSS-TS）**
- 時系列データの時空間分解アルゴリズムの開発



時系列データ要約の応用可能性

1. 高速 & 超軽量なリアルタイム処理

2. オンライン解析

- エビデンスに基づくオンライン異常検知法
 現時刻の時系列が過去に出現したかどうかを判定
 ☞ 出現していなければ 異常 (新種) パターン
- ゼロトラスト認証の要素技術 *Sketch as FingerPrint*



HDD

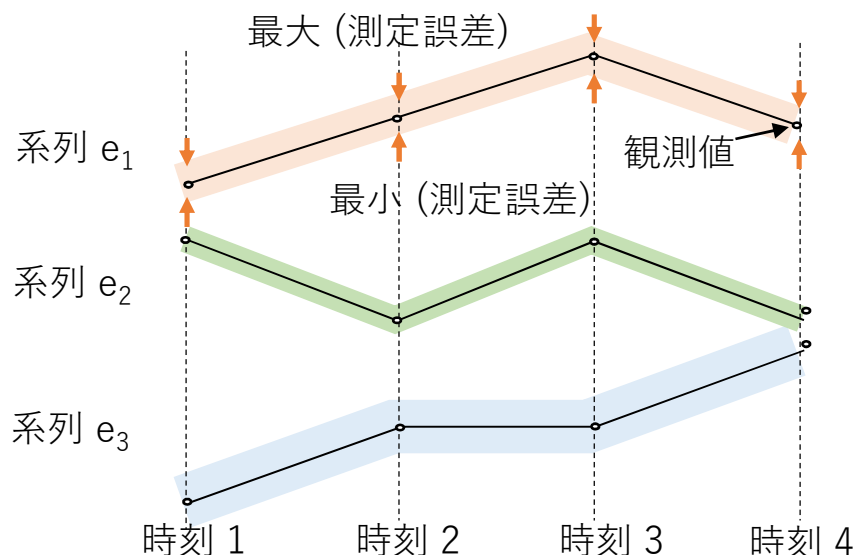


Memory

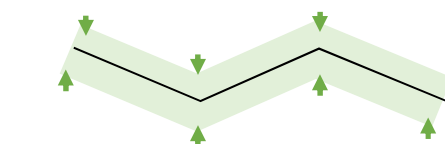


DPU

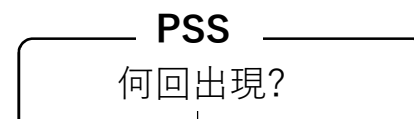
時系列データストリーム $S = \langle e_1, e_2, e_3 \rangle$



問い合わせ (クエリ) 時系列 q



q に含まれる系列は S に

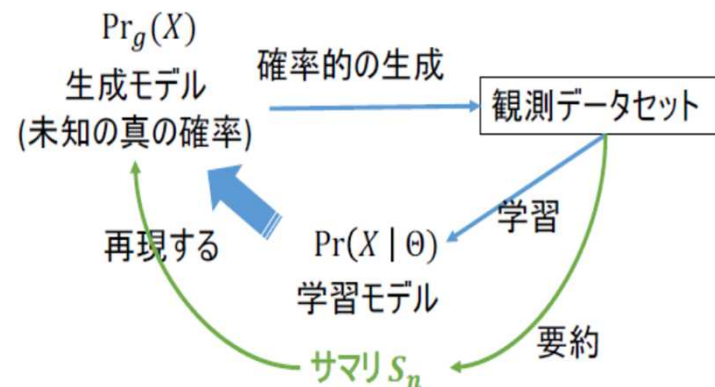


1回 ■ **レアな時系列**

定義: q と e_i の含意関係 $R(e_i, q)$

$$(\forall t e_i(t).min \geq q(t).min) \wedge (\forall t e_i(t).max \leq q(t).max)$$

既存技術と新技術の違い



• モデルに基づく異常検知

(learning-model-based anomaly detection)

- 事前の学習モデルが必要
 - 適切なモデル設計にコストがかかる (問題依存性が高い) 【問題点】
- モデルのダイナミクスが時間変化すると対応できない 【問題点】
- 学習モデルは高速・軽量 【ビッグデータ対応】

• エビデンスに基づく異常検知

(instance-based anomaly detection)

- 事前の学習モデルを必要としない
- **解析対象データに一切の分布を仮定しないノンパラメトリックな確率モデル**
 - 現象の発生メカニズムがわからないデータに利用可能
- **データセットシフト, ダイナミクスが時間変化する概念遷移に対応可**
- 既存サマリは一般に $O(n)$ で肥大化する 【問題点】 【ビッグデータ非対応】
 - **新技術により $O(\log n)$ に抑制可能**

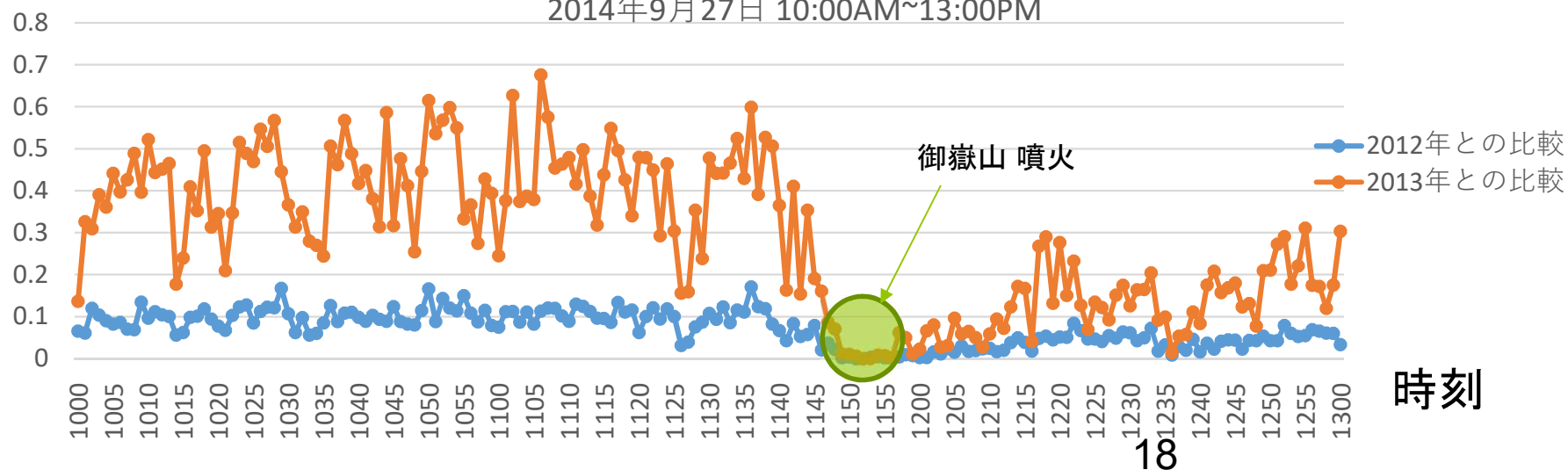
例. 地震データの異常イベント検知

- 以前に出現したことのありそうな時系列データかどうかで異常イベント(火山性噴火)をリアルタイム検知する
- 高感度地震観測網 開田観測点の地震計 (Hinet U成分) 60Hz データ
 - 2014年9月27日10:00AM – 13:00PM までの分単位の時系列データ
 - 2012年, 2013年それぞれの同時間帯の時系列パターンとの比較

出現感度

時系列パターンの出現感度の変化

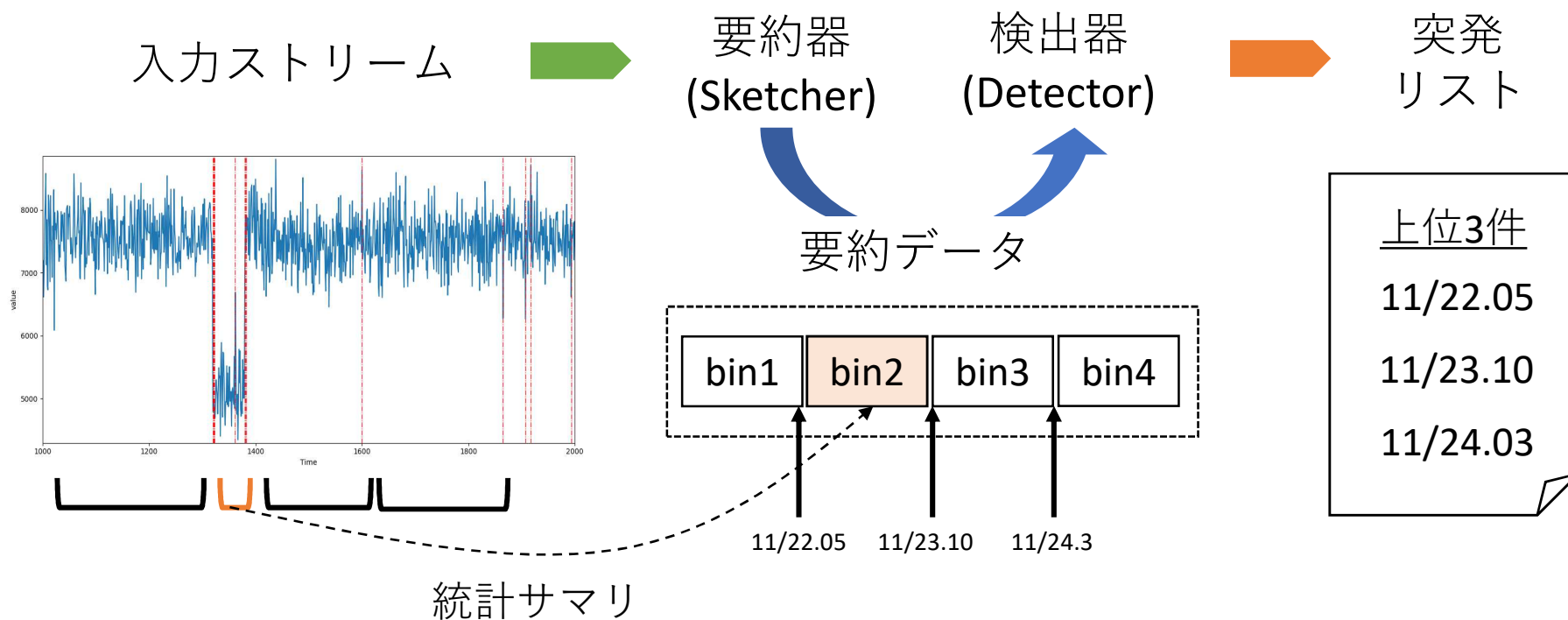
2014年9月27日 10:00AM~13:00PM



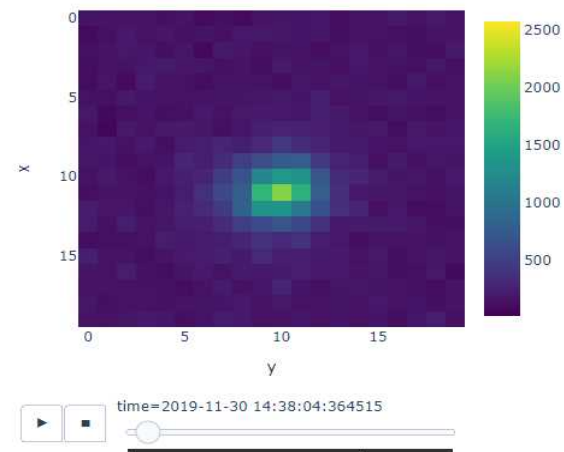
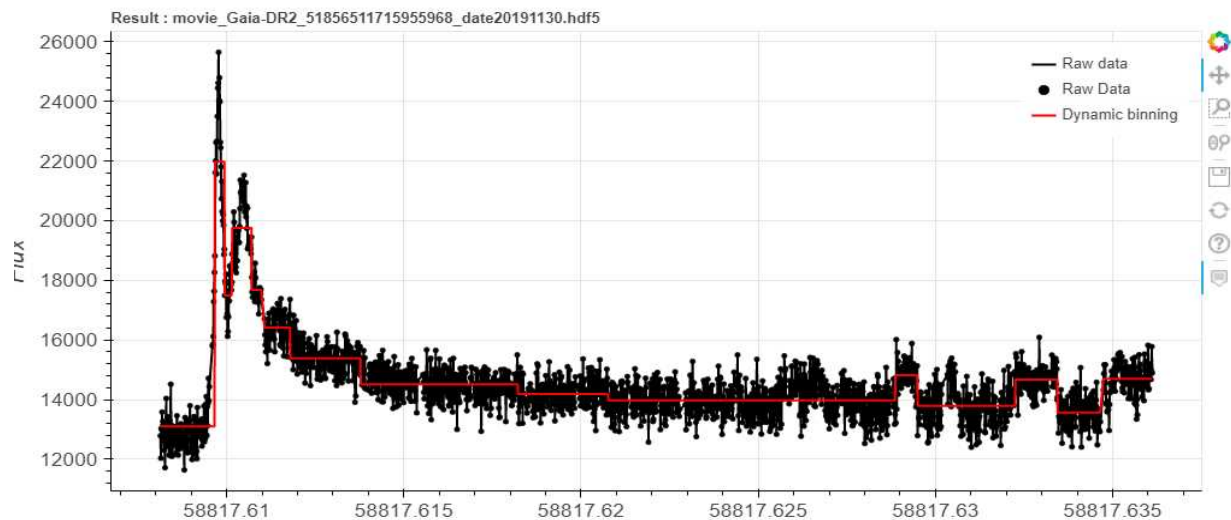
時刻

例. 測光データの突発イベント検出

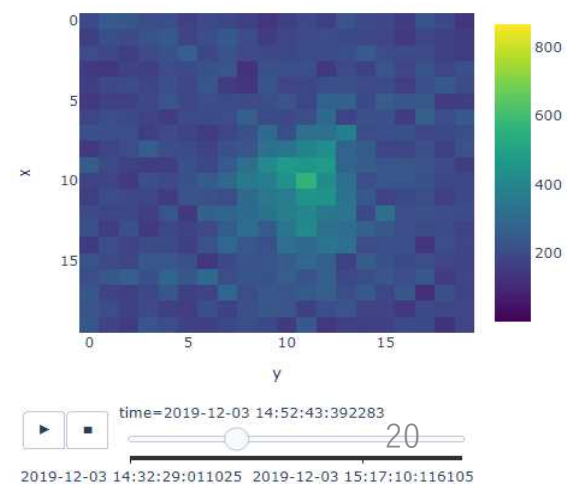
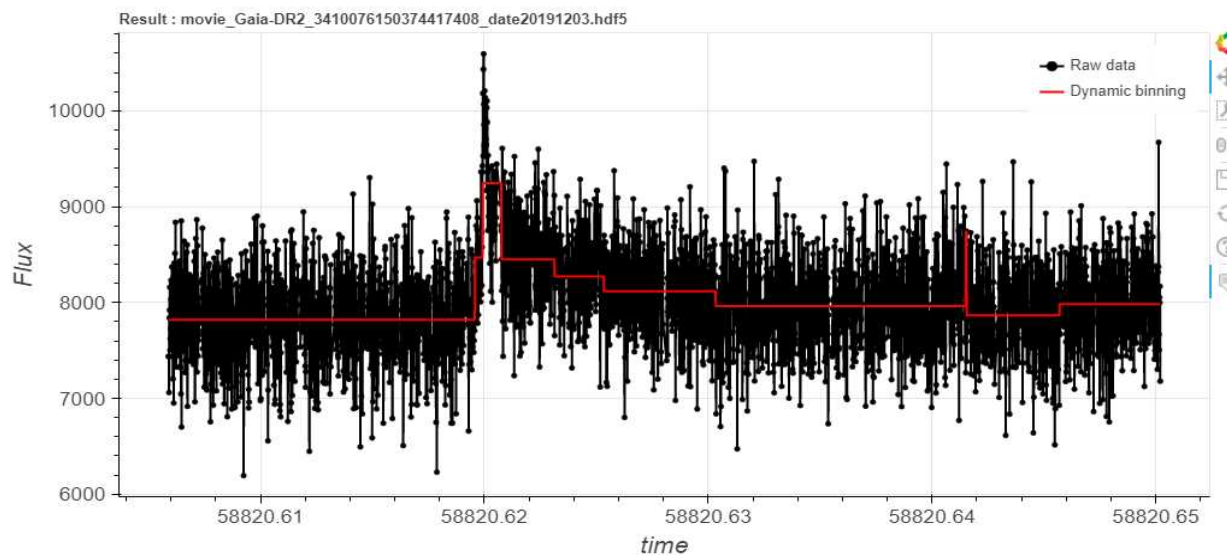
- データ要約 & 確率不等式に基づく検知
 - 同分布の時系列区間を一つのビン (bin) にまとめる
 - 各ビンの発生確率をもとに異常スコアを算出する



M dwarf データ上のフレア検出例



Movie result

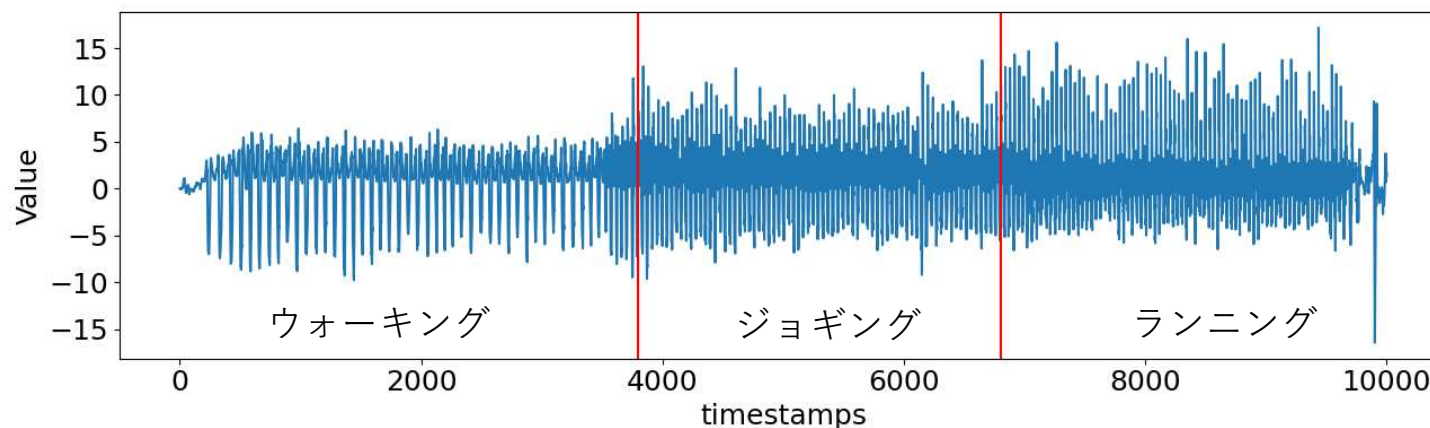


例. センシングデータを用いた行動認証

- スマートデバイスの各種センサーを利用
加速度, 角速度, 心拍, 皮膚温度, etc
- センシングデータを保存した PSS-TS
個人の ``行動指紋`` とみなせる
PSS-TSを用いたゼロトラスト認証の可能性
→ いつもと異なる非定常行動の識別
- 見守り・盗難防止といったセキュリティ分野へ利用



加速度データの例



発表のまとめ

- ストリームデータ処理とデータ要約（サマリ）
- 時系列データ要約の説明
 - **PSS-TS: 高速・超軽量なデータ構造**
- PSS-TSの応用可能性
 - インメモリ処理によるリアルタイム検索
 - エビデンスに基づく異常検知
 - ✓ 自然科学データを用いた科学発見
 - ✓ センシングデータを用いた行動認証の例

企業への期待

- アプリケーションを探索したい。
- リアルタイム解析を扱う領域での可用性検証を実施することで明確になると考えている。
- エッジ側でのデータ管理やリアルタイム分析に従事する企業との共同研究を希望。
- スマートデバイスのアプリ開発等でセキュリティ分野への展開を考えている企業には、本技術の導入は有効と思われる。

本技術に関する知的財産権

- 発明の名称 : データ操作プログラム、
データ操作システム、およびデータ操作
方法
- 出願番号 : 特願2022-17871
- 出願人 : 静岡大学
- 発明者 : 山本泰生、山本裕介

産学連携の経歴

期間 (年度)

- 2014-2017 JSTさきがけに採択
- 2017-2020 科研基盤研究(C)に採択
- 2020-2023 科研基盤研究(C)に採択
- 2023-2025 JST可能性検証に採択

【共同研究】

- 2020-2024 ヤマハ発動機
- 2021-2022 メンテック
- 2022-2023 デンソー
- 2023-2024 サッポロビール

お問い合わせ先

国立大学法人静岡大学

イノベーション社会連携推進機構

TEL 053-478-1710

FAX 053-478-1711

e-mail sangakucd@adb.shizuoka.ac.jp