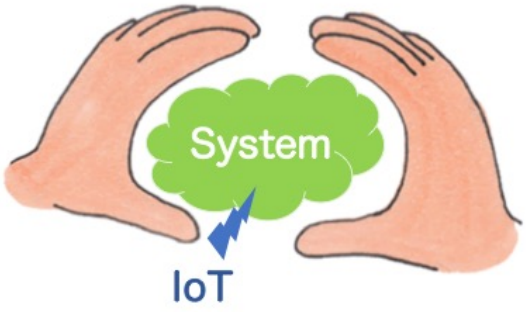
多次元データのオンライン要約とその応用に向けて

山本 泰生

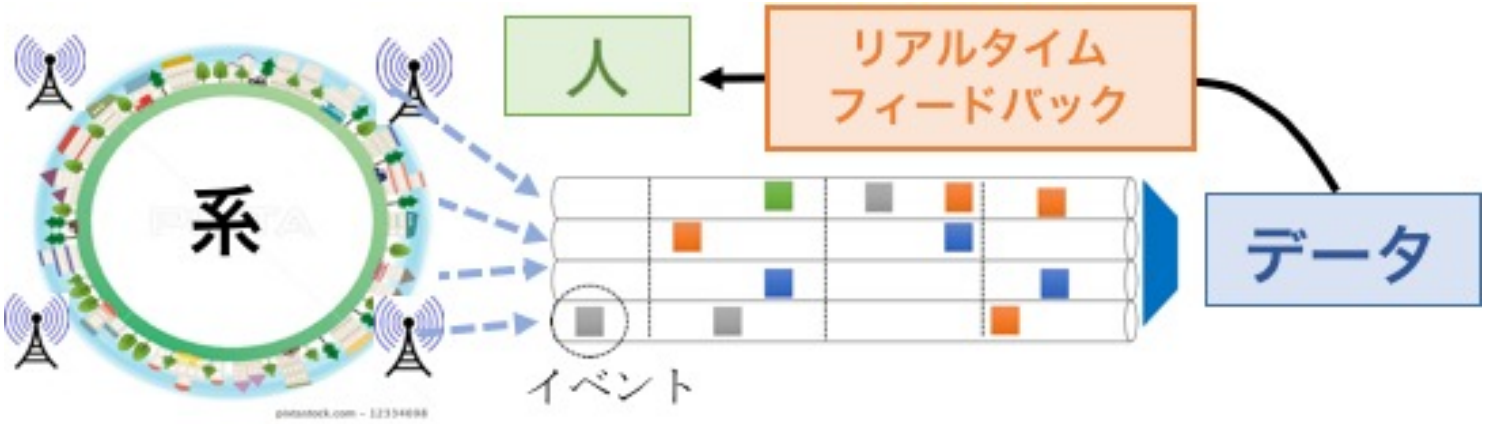
静岡大学情報学部・理研AIPセンター

木曽シュミットシンポジウム2019
2019年7月10日

なぜここに?



- 自己紹介
  - 専門: 人工知能，データマイニング，ビッグデータ処理
  - ストリーム型ビッグデータのオンラインデータマイニング
    JST さきがけ研究 (ビッグデータ基盤領域 2014年-2018年)
- データ駆動型の観測発見の研究
  - 系のデータから「面白そうな」観測をいち早く見つける

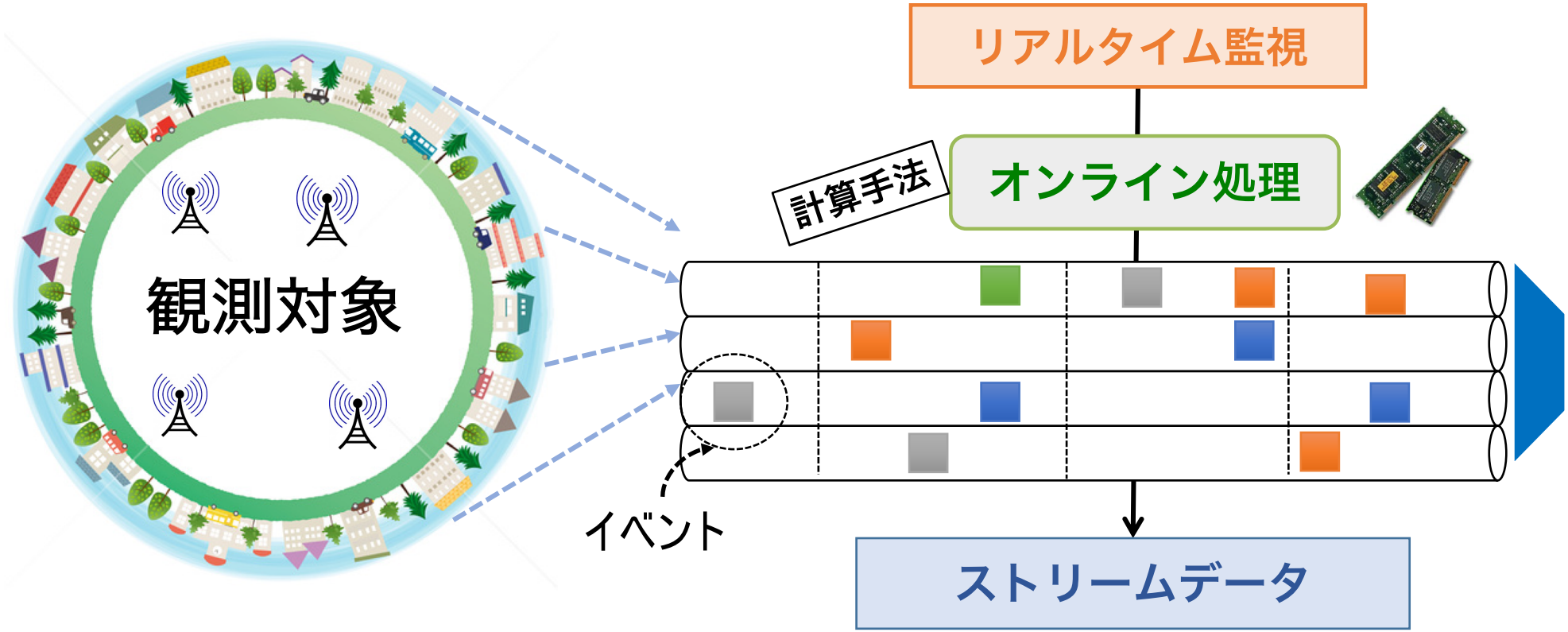


発表の流れ

- ストリームデータの要約 (サマリ) 法について
  - サマリとは
  - 確率的メンバーシップサマリ
  - 予備実験
- まとめ

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)**



ストリームデータの要約 (サマリ)

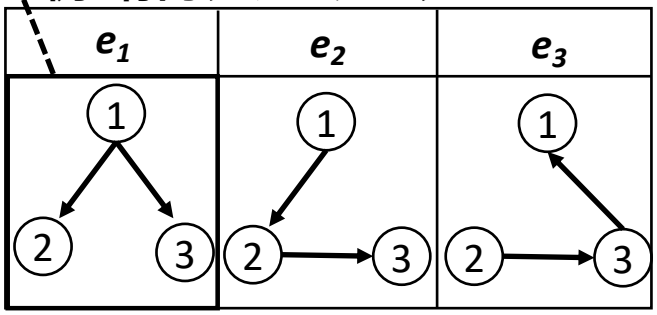
- **順序関係**に基づくストリームデータのモデリング
  - 解析対象のイベントの全体集合: E
  - E 上の順序関係: $\leq$
  - ストリームデータ: $V_n = \langle e_1, e_2, \ldots, e_n \rangle$ ただし $e_i \in E$
  - イベント e のサポート: $e_i \leq e$ を満たす V_n 中の e_i の個数
- **ストリームデータのサマリ S_n**
  - 任意のイベント e に対し，e の**出現回数**を与えるデータ構造

イベント

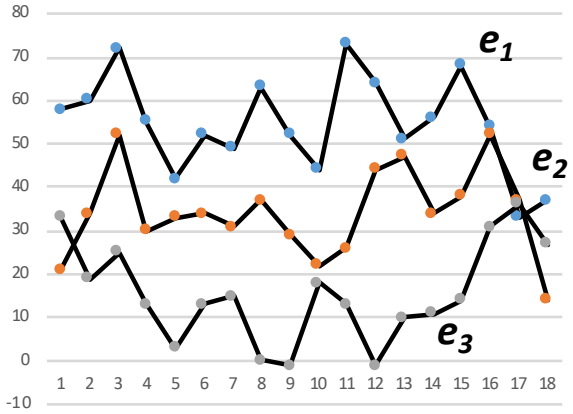
例. 気象データ

V_3	気温	降水量	風速	風向	日照
e_1	14	0	3	0	10
e_2	11	10	1	0	5
e_3	19	0	2	1	14

例. 有向グラフデータ



例. 時系列データ

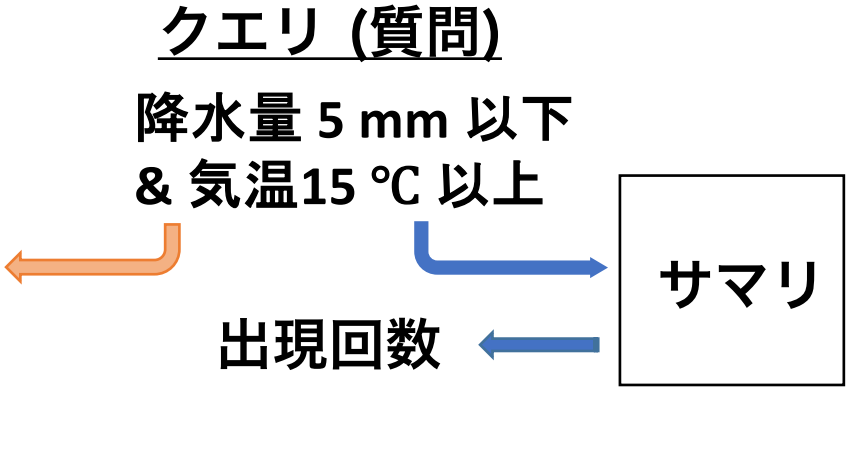


ストリームデータの要約 (サマリ)

- **順序関係**に基づくストリームデータのモデリング
  - 解析対象のイベントの全体集合: E
  - E 上の順序関係: $\leq$
  - ストリームデータ: $V_n = \langle e_1, e_2, \ldots, e_n \rangle$ ただし $e_i \in E$
  - イベント e のサポート: $e_i \leq e$ を満たす V_n 中の e_i の個数
- **ストリームデータのサマリ S_n**
  - 任意のイベント e に対し，e の**出現回数**を与えるデータ構造

例. 気象データ (関係データベース)

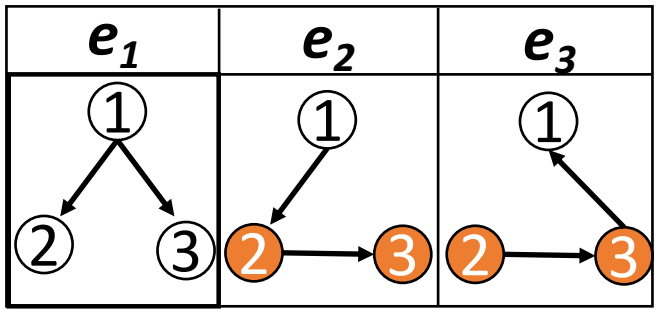
V_3	気温	降水量	風速	風向	日照
e_1	14	0	3	0	10
e_2	11	10	1	0	5
e_3	19	0	2	1	14



ストリームデータの要約 (サマリ)

- **順序関係**に基づくストリームデータのモデリング
  - 解析対象のイベントの全体集合: E
  - E 上の順序関係: $\leq$
  - ストリームデータ: $V_n = \langle e_1, e_2, \ldots, e_n \rangle$ ただし $e_i \in E$
  - イベント e のサポート: $e_i \leq e$ を満たす V_n 中の e_i の個数
- **ストリームデータのサマリ S_n**
  - 任意のイベント e に対し，e の**出現回数**を与えるデータ構造

例. 有向グラフデータ



クエリ (質問)

部分グラフ ② → ③
を含むイベント数は?

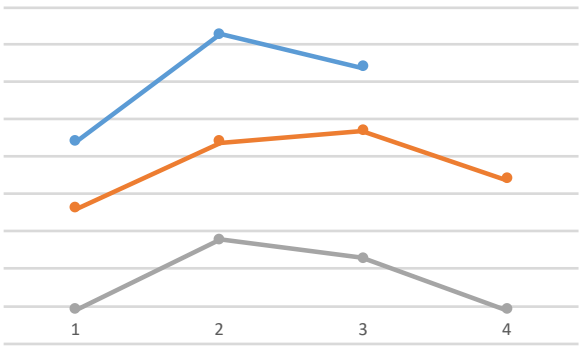
出現回数

サマリ

ストリームデータの要約 (サマリ)

- **順序関係**に基づくストリームデータのモデリング
  - 解析対象のイベントの全体集合: E
  - E 上の順序関係: $\leq$
  - ストリームデータ: $V_n = \langle e_1, e_2, \ldots, e_n \rangle$ ただし $e_i \in E$
  - イベント e のサポート: $e_i \leq e$ を満たす V_n 中の e_i の個数
- **ストリームデータのサマリ S_n**
  - 任意のイベント e に対し，e の**出現回数**を与えるデータ構造

例. 時系列データ



クエリ (質問)

10 ～ 20 上昇した後，5 ～ 10 だけ減少する波形を含むイベント数は?

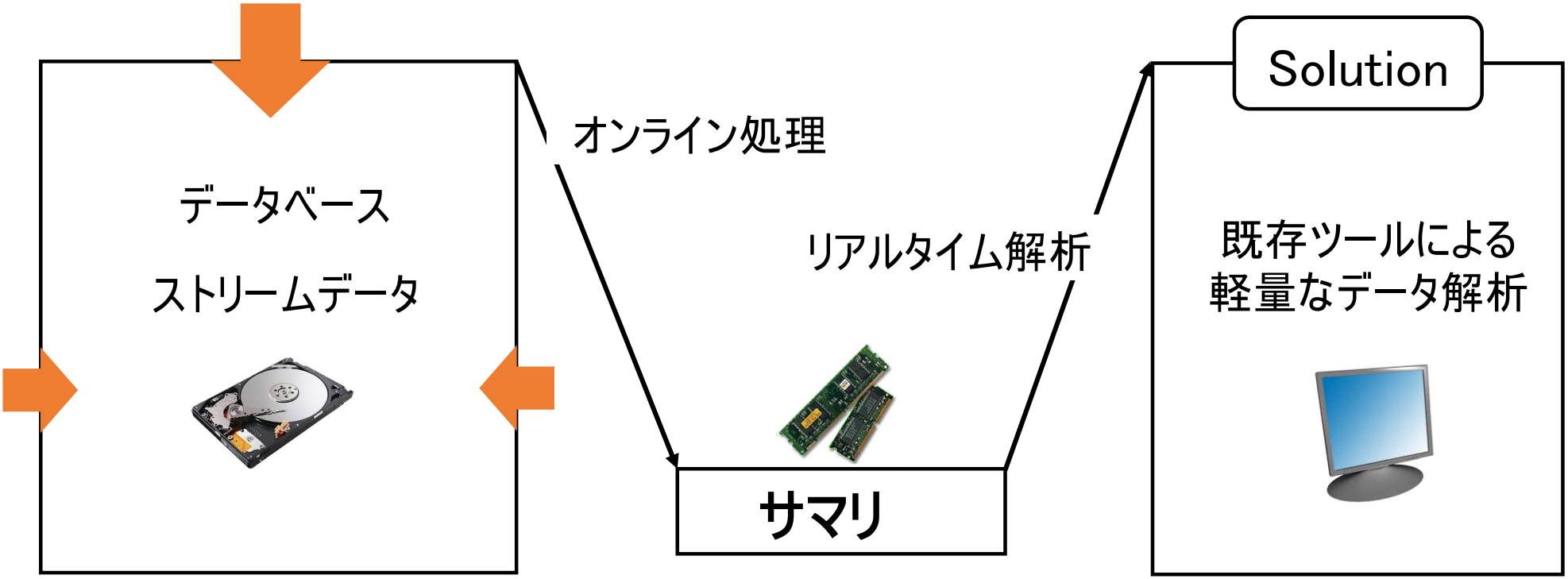
出現回数

サマリ

サマリ構築のインパクト

Point

以下の3つの条件を満たす **サマリ** を構築する
（1） **一度きり**のデータスキャン;
（2） **メモリ内に保持できる**;
（3） 見積値の**誤差保証**ができる！



サマリの位置付け

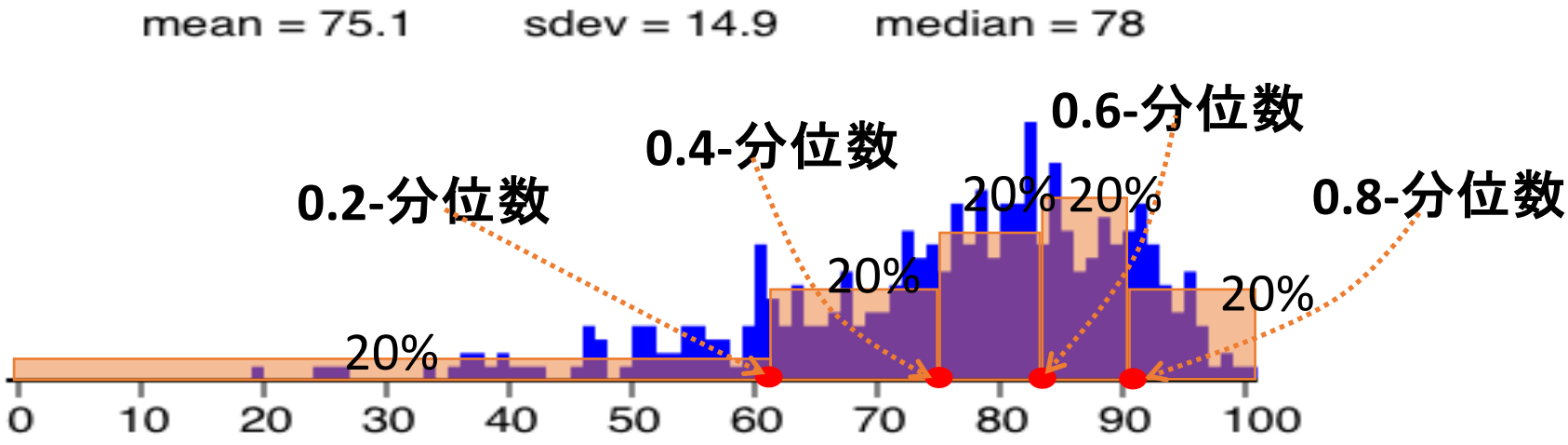
- 出現分布を特徴づけるノンパラメトリックなデータ表現

1次元データ (実数列) の場合: **分位数サマリ**

- 統計検定 (例. Q-Q プロット, Kolmogorov-Smirnov 検定)
- R, Excel, Google's Log Analysis のパッケージとして配布
- **劣線形アルゴリズム**の研究が2000年頃から活発

多次元データ: **半順序データのサマリ**

等深 (Equi-depth) ヒストグラム: 4分位数の例



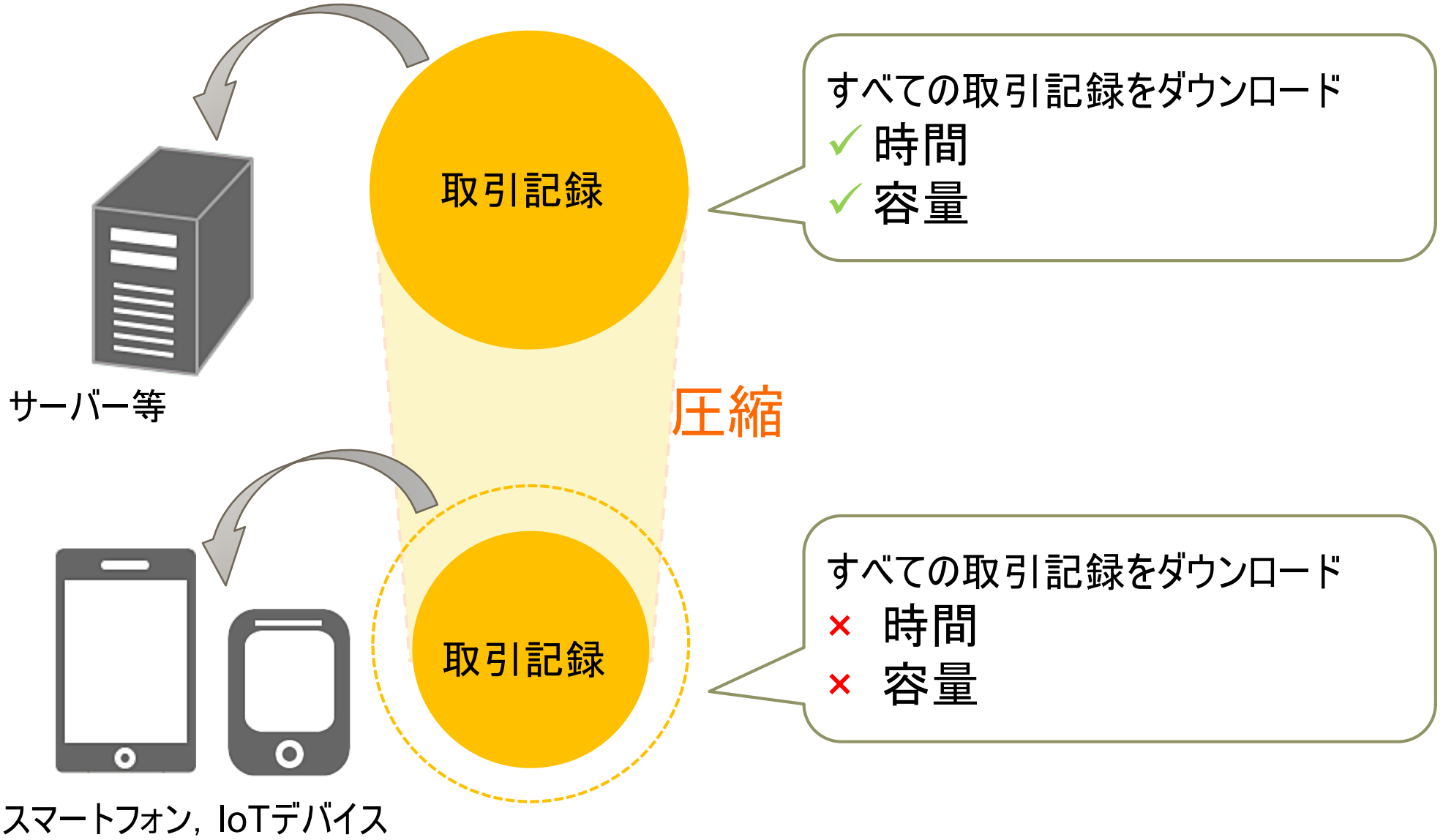
確率的メンバーシップサマリ

許容誤差の範囲内で

データが出現したかどうかを判定する

データ構造

ビットコインの運用

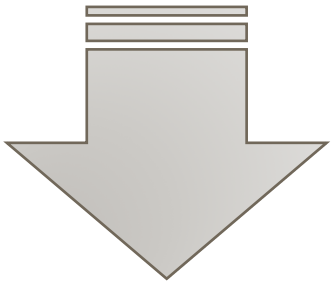

取引記録
すべての取引記録をダウンロード
✓ 時間
✓ 容量
サーバー等
圧縮
取引記録
すべての取引記録をダウンロード
× 時間
× 容量
スマートフォン，IoTデバイス

ビットコインでの使用例

すべての取引記録をダウンロードせずに取引検証を行いたい

- 二重支払い
- トランザクションがどのブロックに格納されているか

SPV (Simplified Payment Verification) : 簡単な取引検証
→取引記録の一部のみダウンロードする



ブルームフィルターを用いることで実現

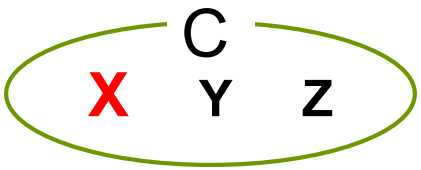
ブルームフィルター

確率的メンバーシップサマリの代表例

1970年にバートン・ブルーム (Burton H. Bloom) が考案

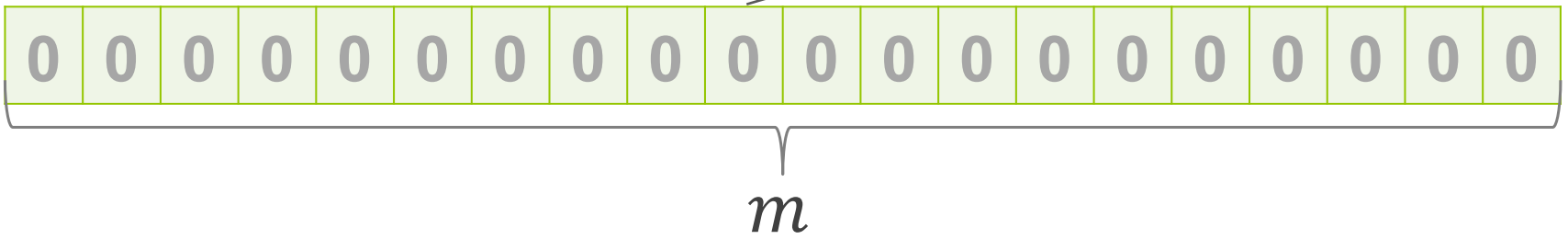
特徴

- k個のハッシュ関数と、mビットのメモリを使用する
- 探索の際の計算量が優れている→$O(k)$
- 偽陽性によるデータの誤判定の可能性がある



h_1, h_2, h_3

空のブルームフィルターは、
0で初期化される



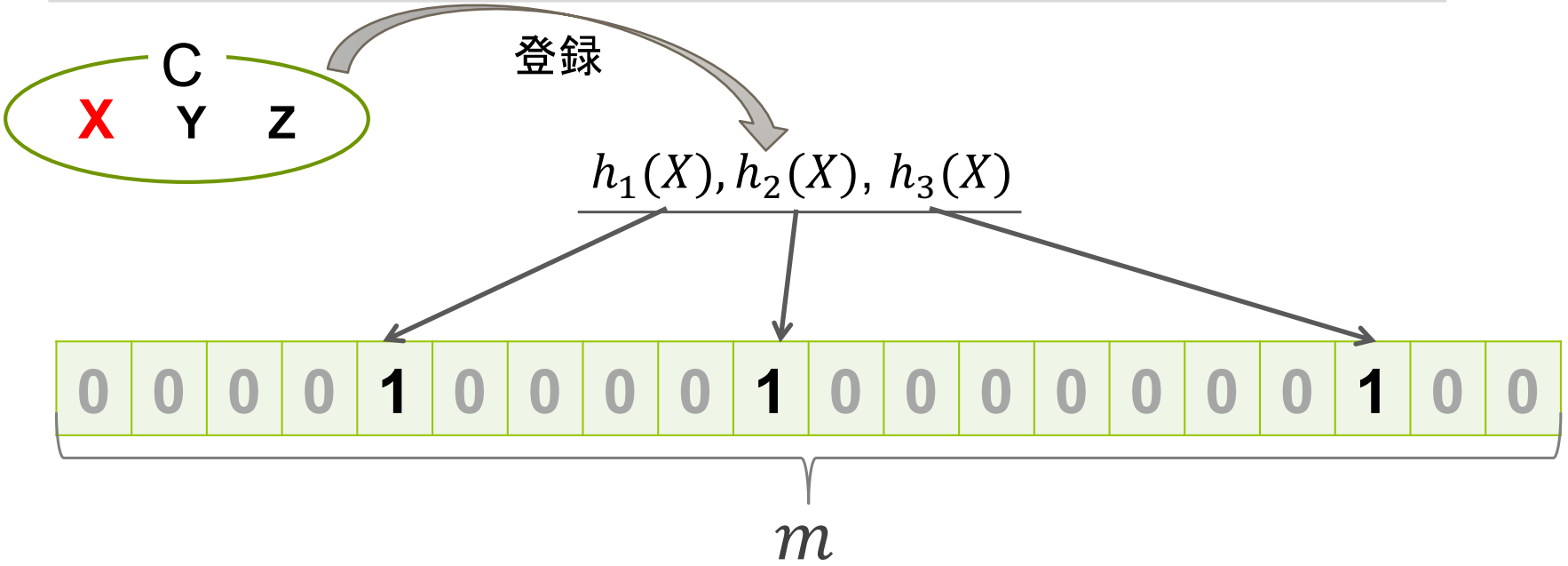
ブルームフィルター

確率的メンバーシップサマリの代表例

1970年にバートン・ブルーム(Burton H. Bloom)が考案

特徴

- k個のハッシュ関数と、mビットのメモリを使用する
- 探索の際の計算量が優れている→$O(k)$
- 偽陽性によるデータの誤判定の可能性がある



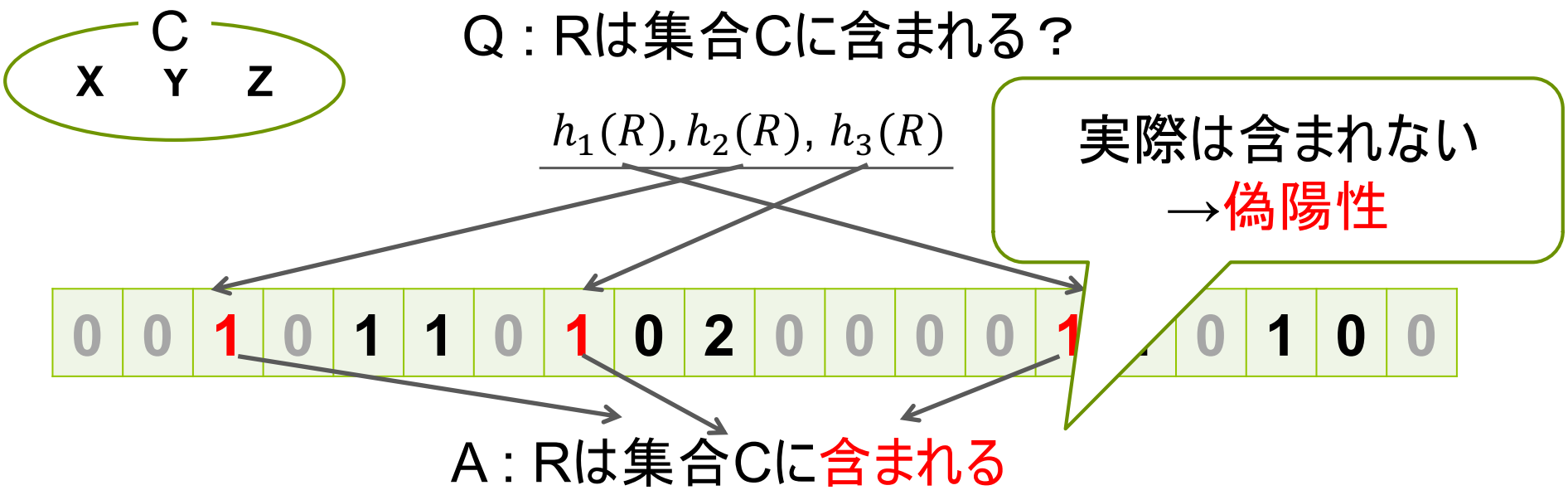
ブルームフィルター

確率的メンバーシップサマリの代表例

1970年にバートン・ブルーム(Burton H. Bloom)が考案

特徴

- k個のハッシュ関数と、mビットのメモリを使用する
- 探索の際の計算量が優れている→$O(k)$
- 偽陽性によるデータの誤判定の可能性がある



ブルームフィルターの偽陽性率

確率的メンバーシップサマリの代表例

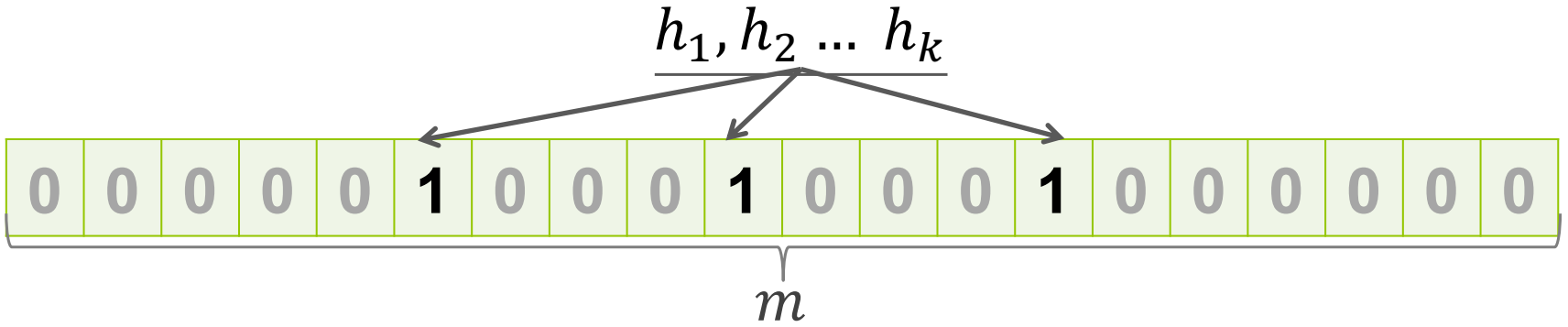
1970年にバートン・ブルーム(Burton H. Bloom)が考案

特徴

- k個のハッシュ関数と、mビットのメモリを使用する
- 探索の際の計算量が優れている→$O(k)$
- 偽陽性によるデータの誤判定の可能性がある

n個の要素を追加した状態で1である確率は $1-\left(1-\frac{1}{m}\right)^{kn}$

偽陽性となる確率の上界は$\left(1-\left(1-\frac{1}{m}\right)^{kn}\right)^{k} \simeq \left(1-e^{\frac{-kn}{m}}\right)^{k}$



ブルームフィルターの偽陽性率

確率的メンバーシップサマリの代表例

1970年に

特徴

- k個の
- 探索の
- 偽陽性

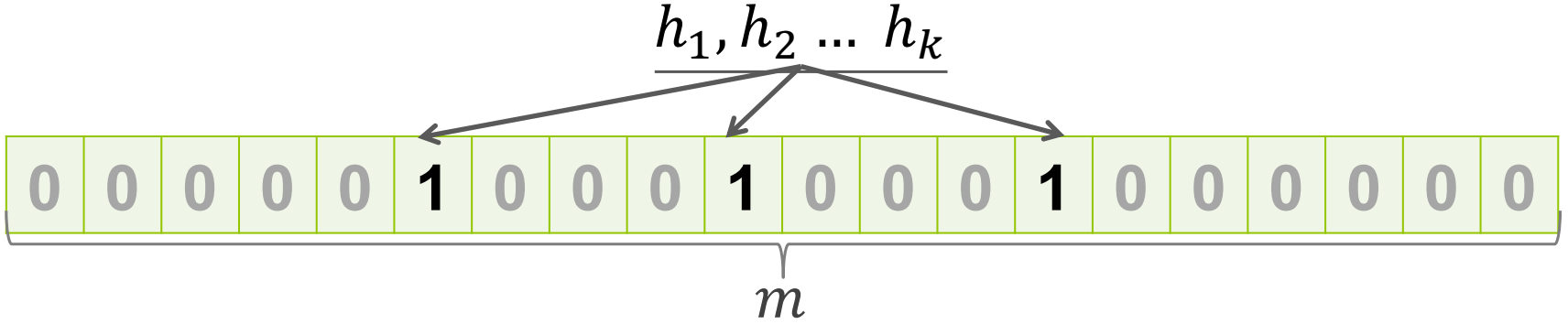
偽陽性率を小さくするためには

データ数 ＜ ハッシュサイズ

を満たさなければならない

偽陽性の発生確率を最小にするkの値は　$k = \frac{m}{n} \log 2$

この時の偽陽性の発生確率の上界は　$\frac{1}{2}^{k} \approx 0.6185^{\frac{m}{n}}$



モチベーション

- ブルームフィルタの空間計算量：$O(n)$

  → n の値は非常に大きい（ビッグデータ）

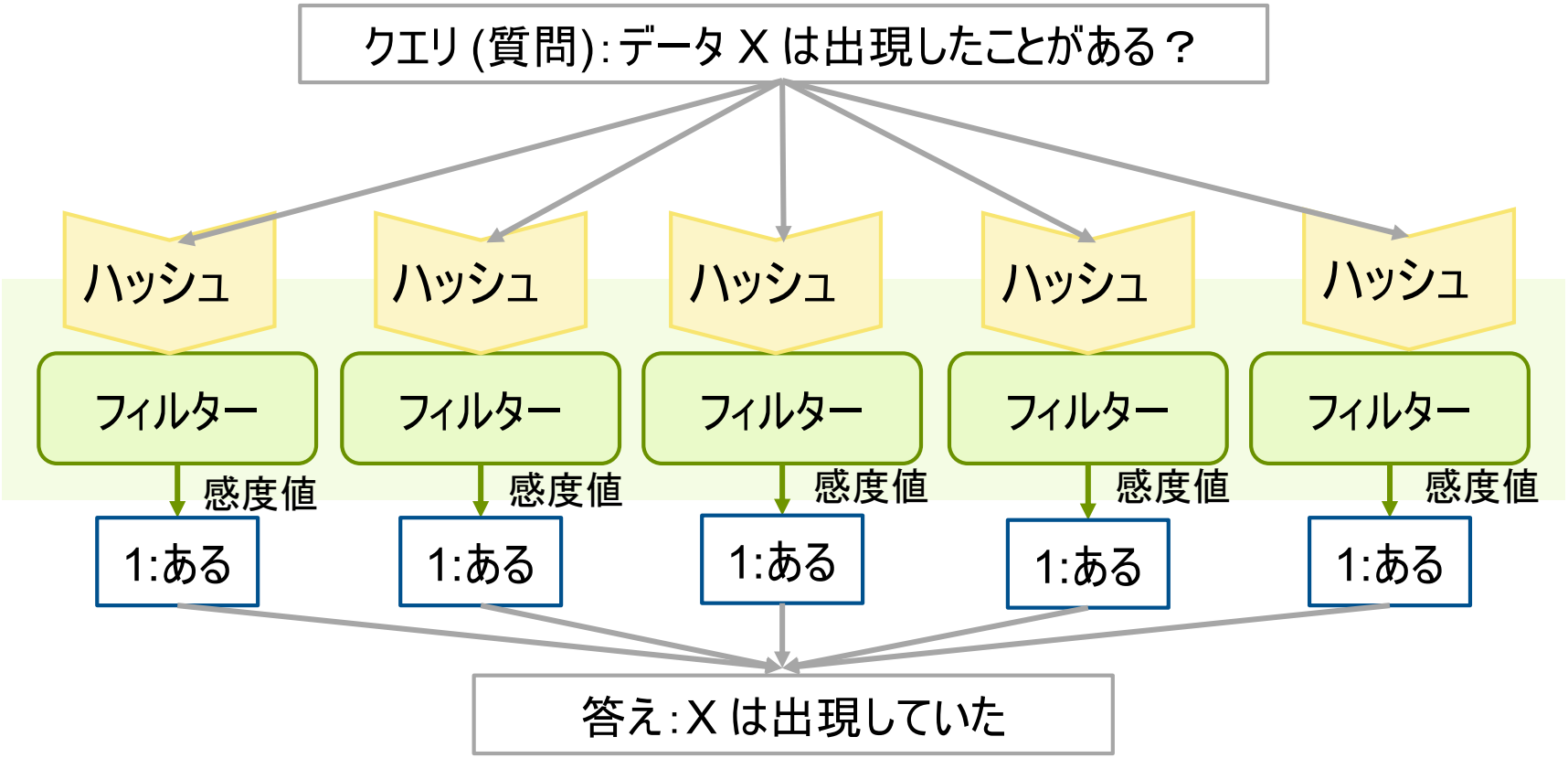
  → ハッシュサイズ M を大きくする必要がある

	ブルームフィルター	モチベーション
空間計算量:M	$M > n$	$M << n$

提案法

Compressed Hashing に基づく フィルター + アンサンブル法

- プロジェクションに基づく省メモリ型の近傍探索法 [CVMR2011]
- 理論的な空間計算量は O(log n)



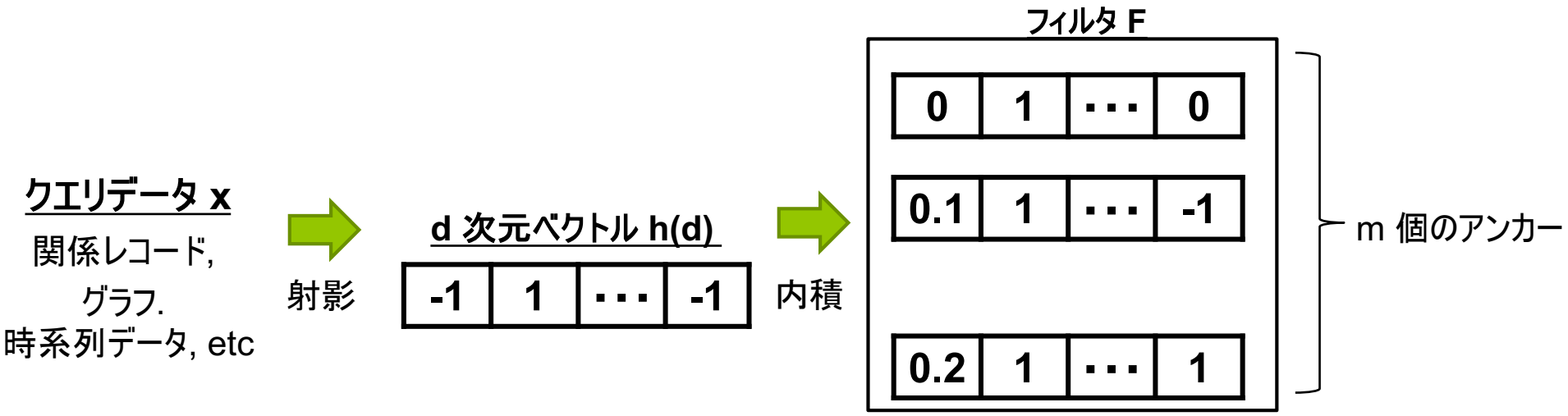
人工知能学会 SIG-KBS にて発表予定

Compressed Hashing に基づくフィルタ

- 一様かつ独立なハッシュ関数によるプロジェクション
- 射影ベクトルのクラスタリング (クラスタ中心をアンカーと呼ぶ)
- クエリデータのメンバーシップは下式の感度値から判定する

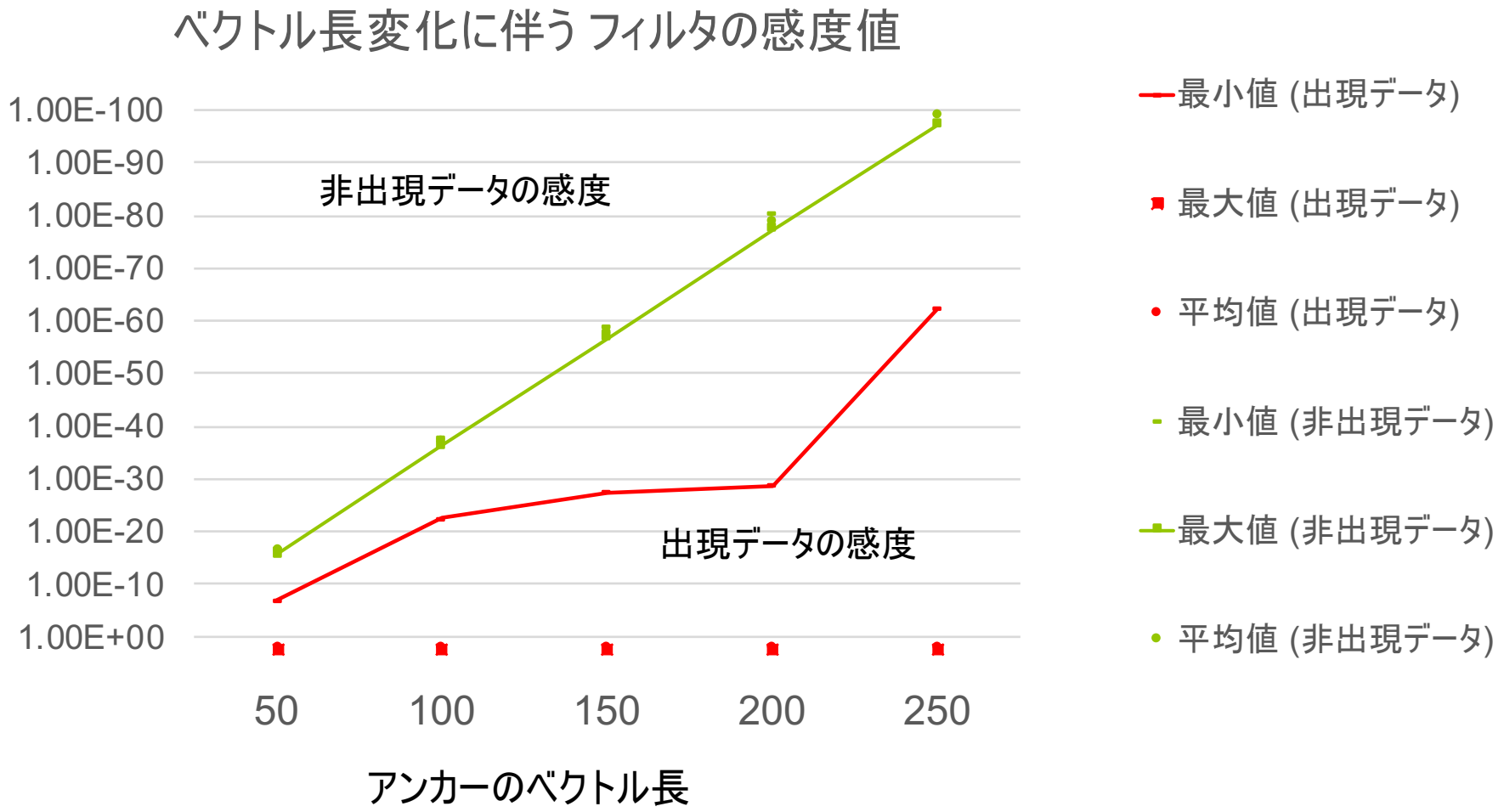
$$\mathrm{G}(x, F) = \frac{1}{m} \sum_{a \in F} \exp\left(\frac{-1}{h^2} \left(\frac{\langle h(x), a \rangle}{d} - 1\right)^2\right)$$

クエリデータ　フィルタ　アンカー　ベクトル長



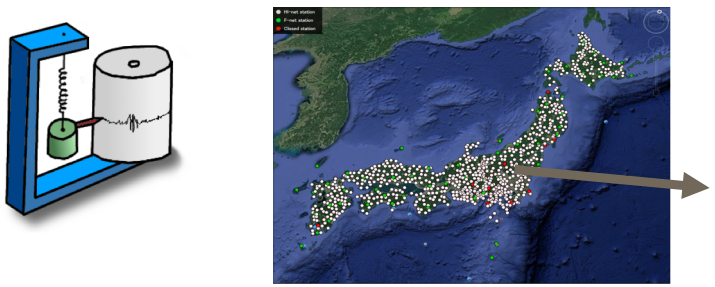
予備実験

- 時系列データ (18,000 点) を フィルタ に登録
- 出現データと非出現データに対する感度値差を計算

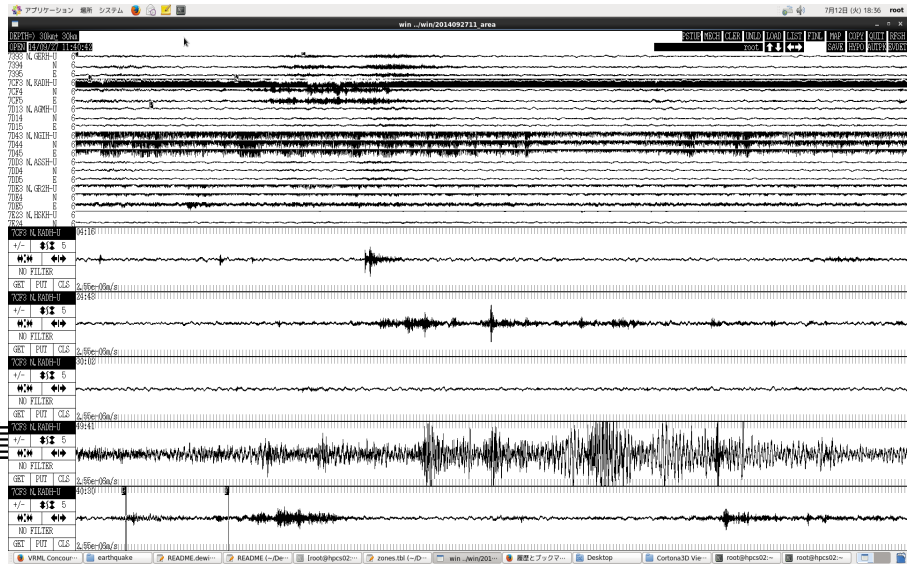


時系列データの傾向変化検知

- 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の傾向変化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 (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

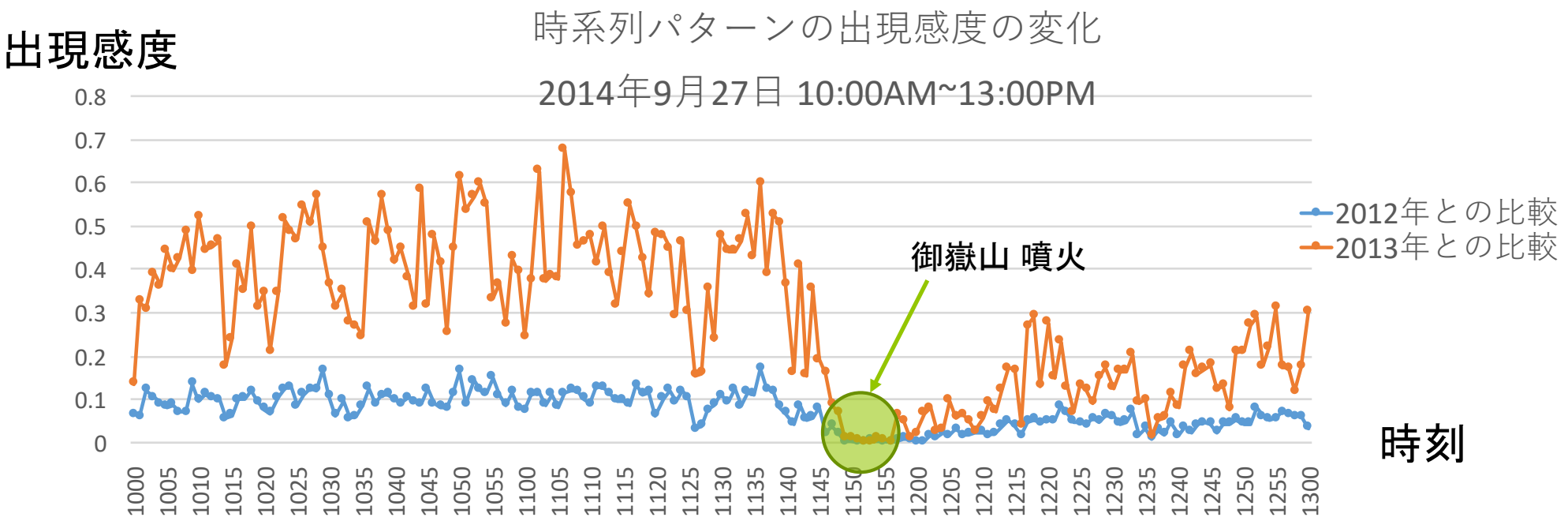


高感度地震観測網 (Hi-net)
1000 グリッド, 3チャネル (水平2軸+垂直
100Hz パケット通信: ３テラバイト/月



時系列データの傾向変化検知

- 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の傾向変化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 (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



発表のまとめ

- ストリームデータのオンライン要約法の紹介
  - 省メモリ型データサマリの構築
  - 確率的メンバーシップサマリ
- 時系列データの傾向変化のリアルタイム検知