データ分析技術

大阪大学情報科学研究科
教授　鬼塚真

取り組みの概要

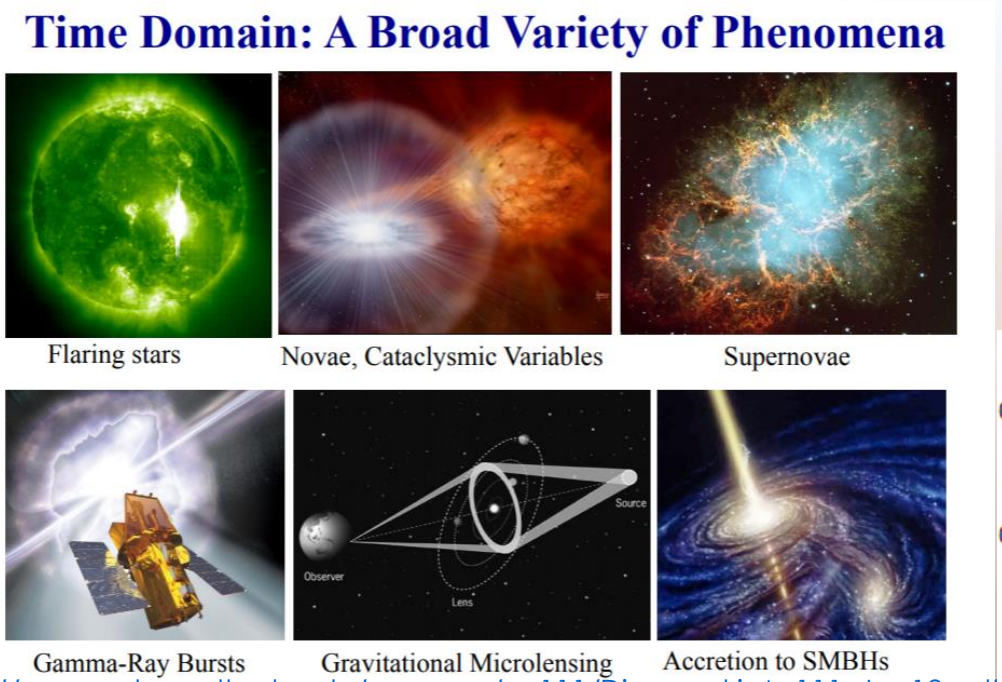
- **発見的データ分析(Exploratory data analysis)**
  - 概要: 有用性の高いデータを自動探索する技術
- **クエリワークロードの最適化**
  - 概要: クエリワークロードに対してスキーマあるいはクエリの最適化を行う技術. 特に cardinality estimation, 時間変化するワークロードに対する実体化ビュー推薦に取り組んでいる.

以降, 発見的データ分析とcardinality estimationについて説明

発見的データ分析の概要

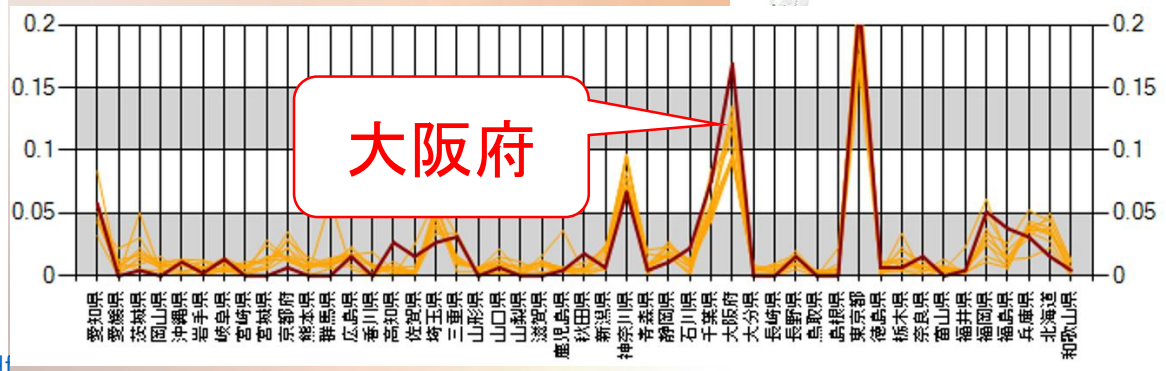
- 発見的データ分析とは，購買データや天体観測データなどの膨大なデータを対象として，通常とは異なる特徴的なデータを発見する技術
  - 国立天文台と協力し明るさが変化する**変動天体の発見**に取り組んでいる

応用例1: 変動天体の発見



応用例2: 特徴的なデータ発見

1人1日当たりゴミ排出量 第１位 大阪府
（環境省， 2011年）



https://www.astro.caltech.edu/~george/ay111/Djorgovski_Ay111_Jan12.pdf

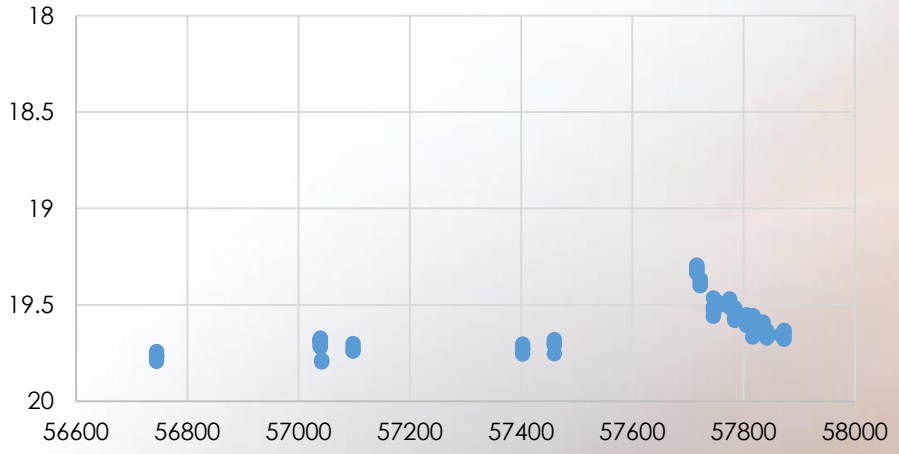
発見的データ分析の技術ポイント

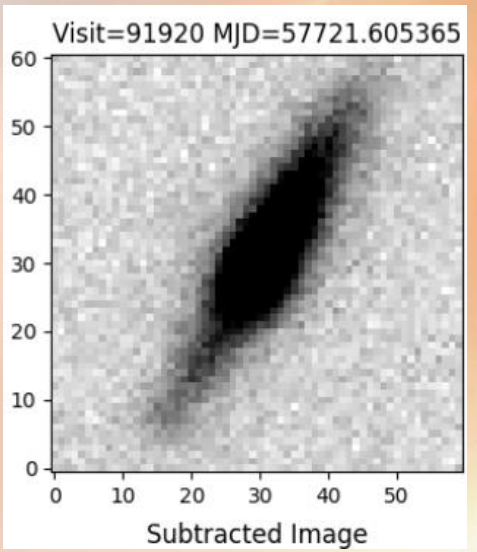
- **発見的データ分析の特徴**: 多様な観点からOLAPクエリを実行した結果群から，通常とは異なる特徴的な分析結果を高速に検出する
- **要素技術**:
  - **サンプリングベースの異常検知技術**: サンプルデータから全量データを利用した際に得られるOLAPクエリ結果を推定し(中心極限定理を活用)，異常度の高い結果を高速に特定する(10倍高速)[1,2]
  - **欠損値・計測誤差を含むデータに対する異常検知技術**: 天体観測などの計測データが含む大量の欠損値および計測誤差に対して，ロバストな異常検知技術を研究開発中
    - データを事前にクラスタ分解しクラスタ毎に近傍補完することで補完量を削減
    - 誤差の標準偏差を利用し異常検知のノイズ耐性を向上

天文台での応用例

- **到達点**
  - 検証用データから変動天体を11件発見したことを確認
  - 明るめの天体群を含むクラスタでは，上位10件中9件が正解（正解と分かっていた4件，新規で発見できた変動天体5件，1件は計測エラーによる異常値）

超新星と思われる例





取り組みの概要

- 発見的データ分析(Exploratory data analysis)
  - 概要: 有用性の高いデータを自動探索する技術
- **クエリワークロードの最適化**
  - 概要: クエリワークロードに対してスキーマあるいはクエリの最適化を行う技術. 特に cardinality estimation, 時間変化するワークロードに対する実体化ビュー推薦に取り組んでいる.

Cardinality estimation の概要

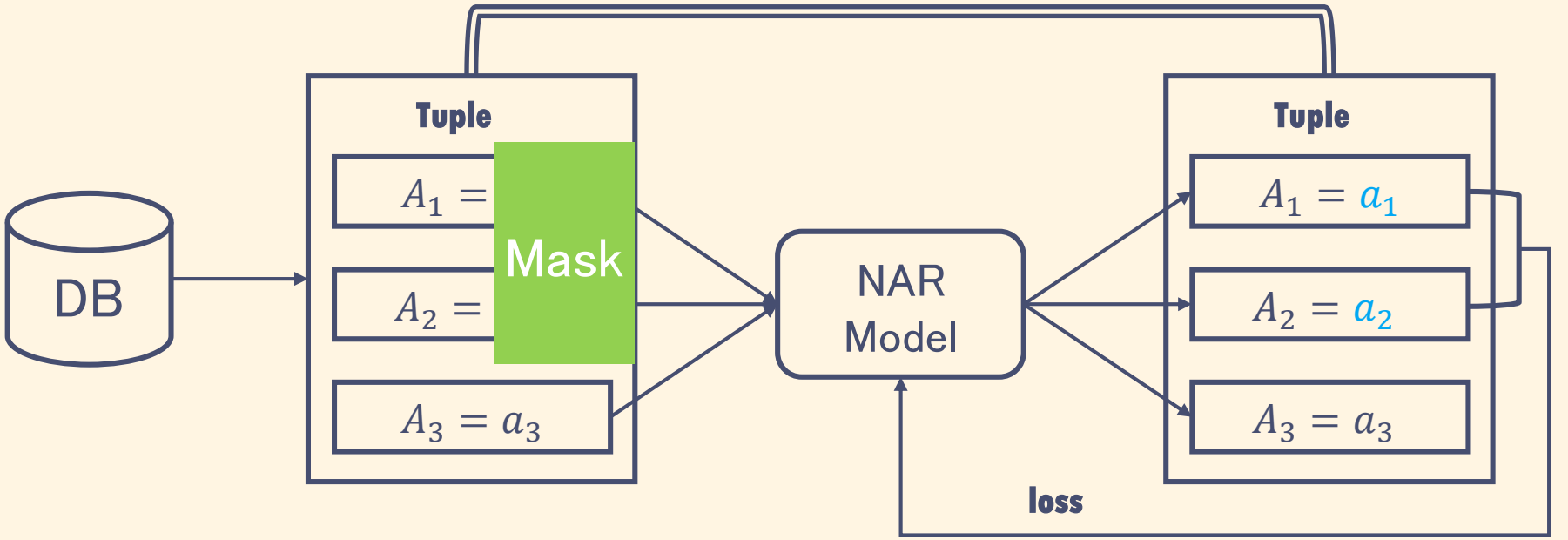
- Cardinality estimation は，与えられたクエリに対して何件のレコードがヒットするかを推定する技術
  - 最適なクエリプランの決定の際に有効
- 通常のDBMSにおけるCardinality estimation
  - 各カラム毎に値の分布をヒストグラムとして管理することで1カラムに対するCardinality を推定する．2つ以上のカラムに対する条件を含むクエリ（joinクエリ含む）に対しては，カラム間に依存性が無いことを仮定推定するため低精度
- 最近の研究動向
  - カラム間の依存関係を機械学習することで，高精度にCardinality を推定する

Cardinality estimation の技術ポイント

- **機械学習ベース**: データベースを入力にカラム間の依存関係を学習
- **Non-Autoregressive model** を適用した最初の研究
  - 従来技術の autoregressive model を利用する技術は学習におけるカラム順に依存してしまい，精度が安定しない問題があった
  - Non-autoregressive model では学習時のカラム順に非依存であるため，従来技術と比較して安定した精度を達成

提案手法:
Cardinality Estimation by Non-Autoregressive Model (NAR)

- 各属性の生成確率を出力するNARモデル (Density Estimation)
- マスクされた入力を当てる穴埋め問題として学習



実験結果

- データ: DMV[9]: 11.6M 件, 11 属性
- 評価クエリ: データに対してランダムに生成された2000クエリ
- 評価指標: q-error[1]: 実際の値から何倍離れているかを示す値(1がベスト)
- 評価結果: 既存技術より安定的に高性能を達成

手法	平均 q-error
提案手法 (Transformer-based)	1.098
Naru-ResMADE-0to10	1.076
Naru-ResMADE-10to0	210.935

安定して高性能

性能が不安定

大半のクエリは
q−error 3以内

log_{10}q-error のクエリ頻度

今後の予定

- **発見的データ分析(Exploratory data analysis)**
  - 並列分散化することで大規模データに対応予定(Spark 化)
- **クエリワークロードの最適化**
  - cardinality estimation: Tsurugi のDBMSエンジンとの連携, 天文台ワークロードへの適用
  - 時間変化するワークロードに対する実体化ビュー推薦: TPC-H ワークロードで性能検証

参考文献

[1] 松本 拓海, 山室 健, 小笠原 麻斗, 佐々木 勇和, 鬼塚 真: 探索的データ分析におけるフレームワークの効率化, DEIM 2018. https://db-event.jpn.org/deim2018/data/papers/122.pdf

[3] Takumi Matsumoto, Yuya Sasaki, Makoto Onizuka: Data Slice Search for Local Outlier View Detection: A Case Study in Fashion EC. EDBT/ICDT Workshops 2019. http://ceur-ws.org/Vol-2322/DARLIAP_12.pdf