An Efficient Computational Framework for Big Maritime Traffic Data Preprocessing

Xiuju Fu (fuxj@ihpc.a-star.edu.sg)

by Xiuju Fu, Haiyan Xu, Vasundhara Jayaraman, Nasri Bin Oyhma, Liye Zhang, Zhe Xiao, Rick Goh

Computing Science Department, Institute of High Performance Computing

13 Nov 2019

Outline

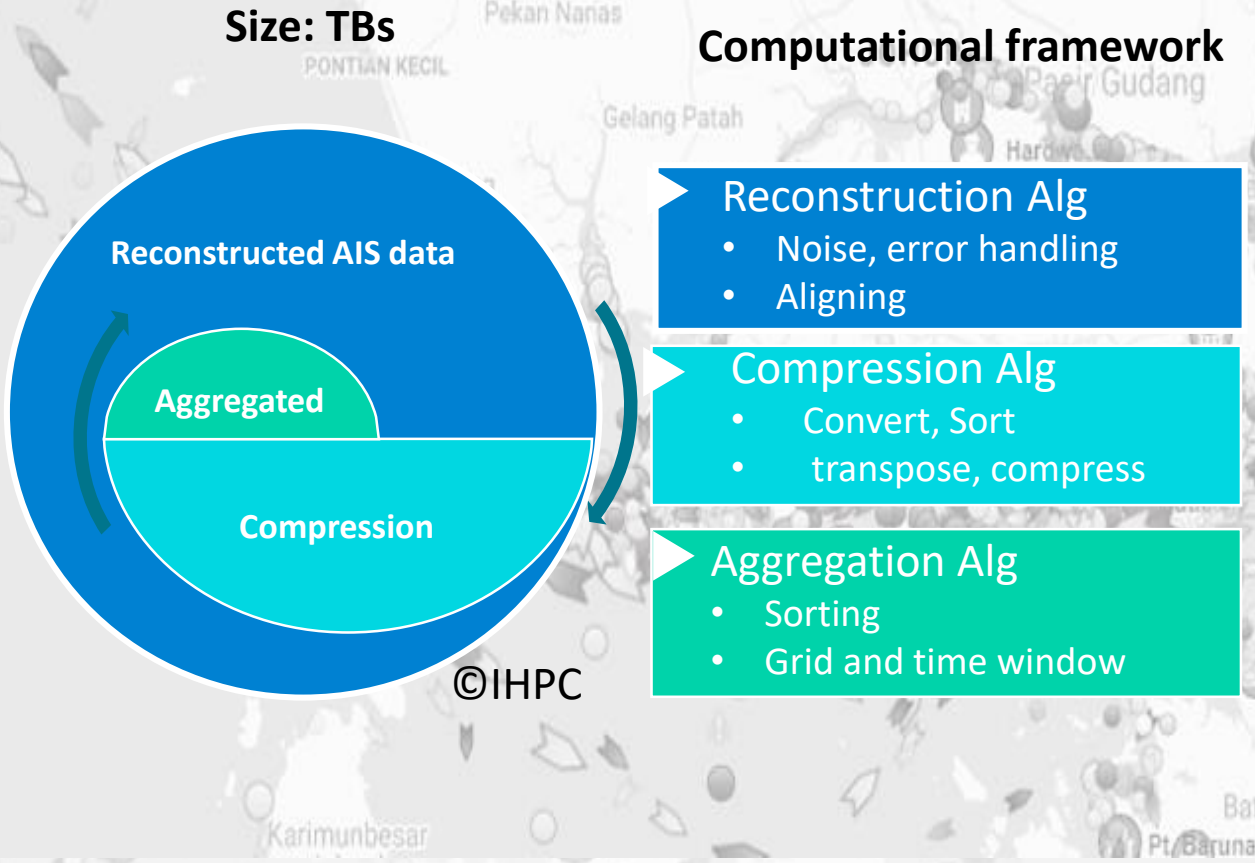
- Problem Statement
- Proposed framework
  - Reconstruction
  - Data-compression
  - Aggregation
- Results and Application Demonstration
- Conclusions

Problem Statement

- Large volumes of spatiotemporal data facilitates the understandings of objects' movement trajectories and activities.
  - Automatic Identification System (AIS) transmitter on vessels transmits location, speed, course, heading, destination information of the vessels every few seconds or minutes, and the size of AIS data generated is increasingly huge.
- The imperfect raw AIS data impedes maritime research and development
  - The presence of noises, errors, missing values and inconsistencies brings challenges for the application of AIS data analytics
- Challenging to store, transfer and load such a large volume of data into system memory for processing and analysis.
  - Study of historical and AIS data streams provides opportunities in understanding vessels' movement pattern for better maritime monitoring, prediction, optimization and management.
- Traditional algorithms are faced with difficulties in expansion and low access efficiency for increasingly large spatial-temporal data.

Proposed Framework

- A computational framework and model to efficiently construct, compress, transfer and acquire necessary information from large scale of spatiotemporal data
  - a reconstruction algorithm that deals with reducing noise, errors and misalignment
  - a lossless compression algorithm that compresses the spatiotemporal data into binary form for efficient storage, speedy loading and easy transfering across networks and systems within the organization;
  - an aggregation algorithm which derives movement, location and activity information of vessels grouped by grid, time window and/or other factors of interest from the compressed binary files, thereby improving data accessibility and reducing storage demand.
- The proposed method has been applied to extract and predict vessel movements



Institute of High Performance Computing
A*STAR

Missing and Noise Data in Raw AIS Data

- **Errors in the raw AIS database**

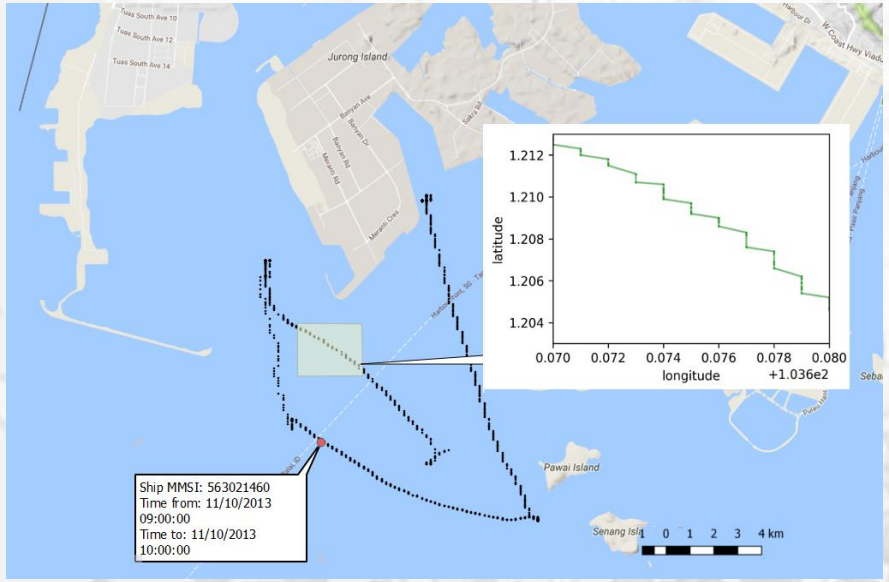


Figure Ship trajectory of AIS data

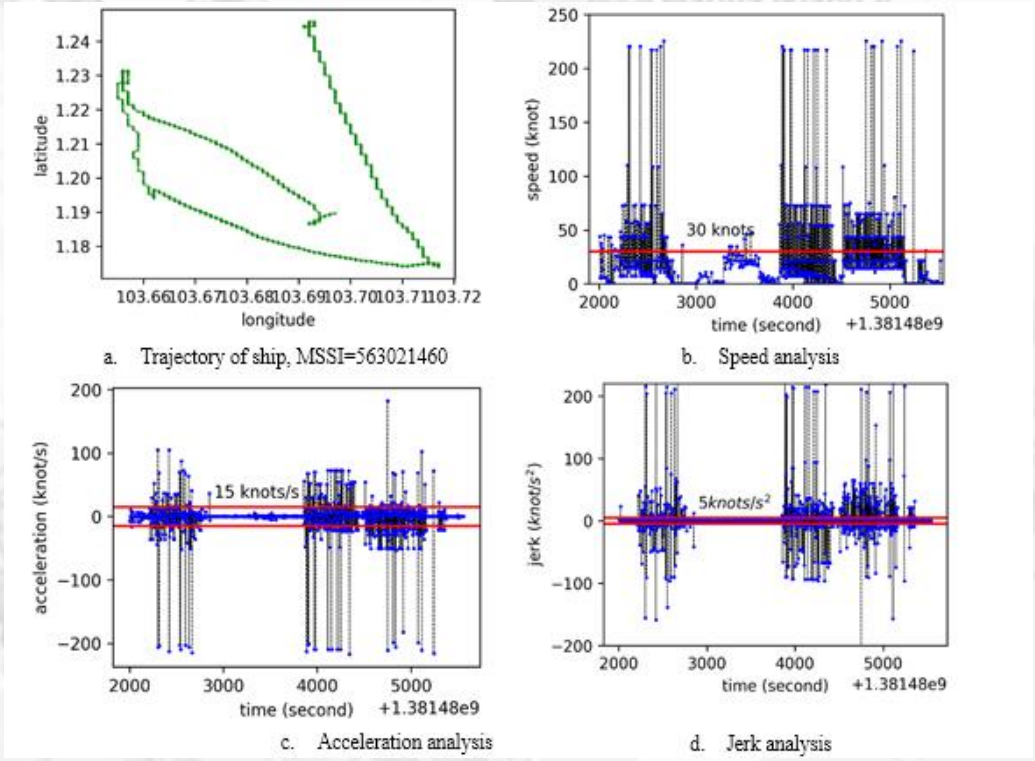


Figure Ship speed, acceleration and jerk analysis

Institute of High Performance Computing
A*STAR

1. Multi-regime Ship Trajectory Reconstruction Model

Step 1: Outliers removal
used vessel speed and ROT to identify the outliers;

Step 2: Ship navigational state estimation
hoteling, maneuvering and normal-speed sailing;

Step3: Vessel trajectory fitting

- *Hoteling*
- *Maneuvering: linear model*
- *Normal speed navigation*

$$s_x(t) = \sum_{j=-k}^{n} c_{x,k+2} M_{i,k+1}(t)$$

$$s_y(t) = \sum_{j=-k}^{n} c_{y,k+2} M_{i,k+1}(t)$$

of degree k with knots $a = t_0 < t_1 < \ldots < t_{n+1} = b$. Then, the trajectory reconstruction becomes an optimization problem:

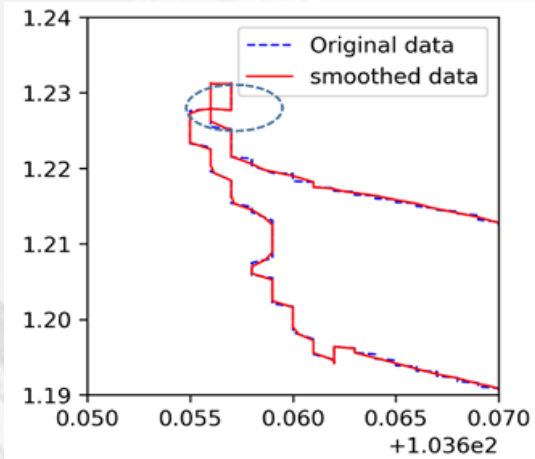
$$\min \sum_{j=1}^{n} \left[s_x^{(k)}(t_j + 0) - s_x^{(k)}(t_j - 0) \right]^2 + \left[s_y^{(k)}(t_j + 0) - s_y^{(k)}(t_j - 0) \right]^2$$

subject to

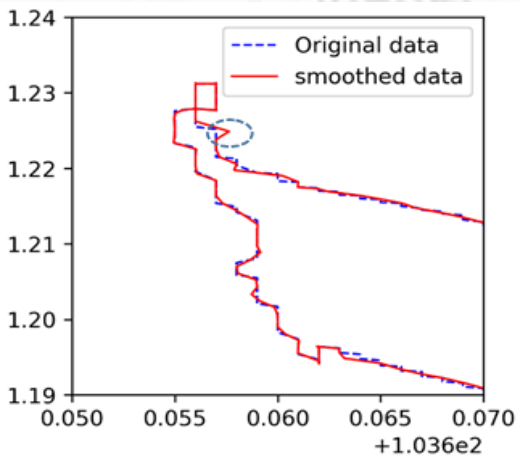
$$\sum_{j=1}^{m} w_j \left[x_j - s_x(t_j) \right]^2 + \left[y_j - s_y(t_j) \right]^2 \le S.$$

Institute of High Performance Computing
A*STAR

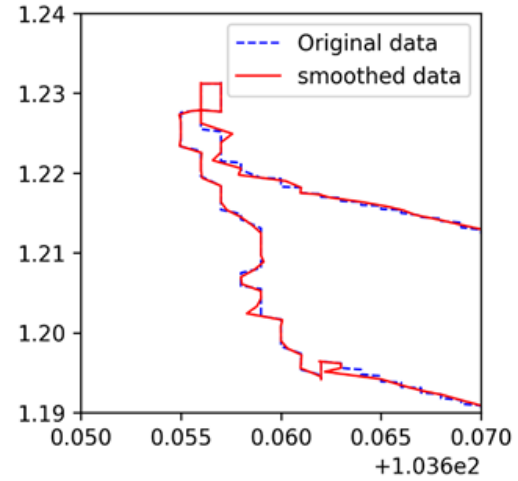
Model Evaluation – Reconstructed Trajectory Comparison



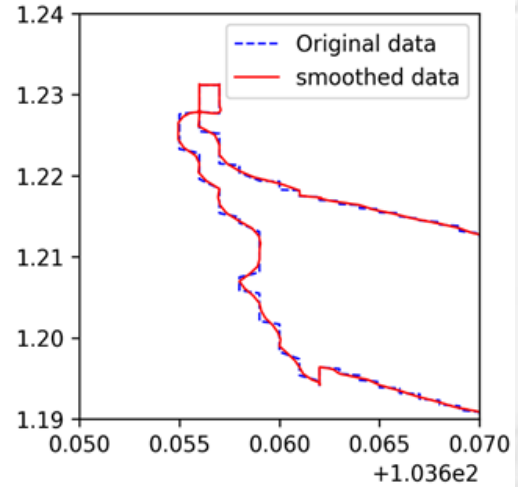
a. Linear interpolation model



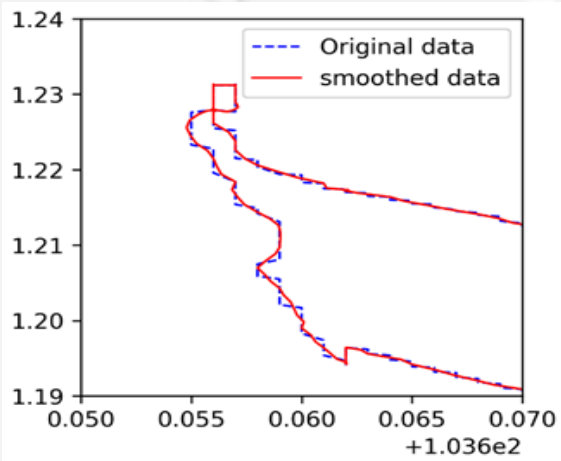
b. Polynomial regression model (n=5)



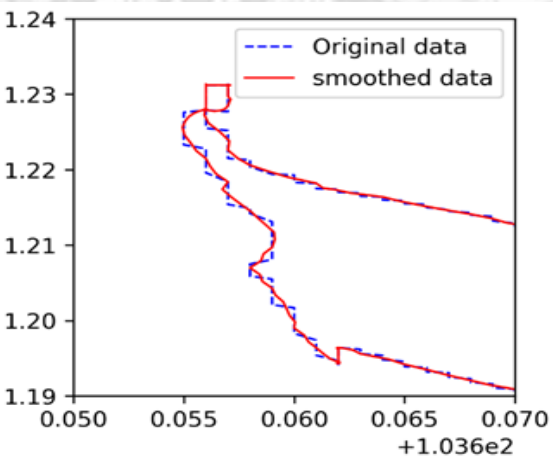
e. Weighted polynomial regression model (n=5)



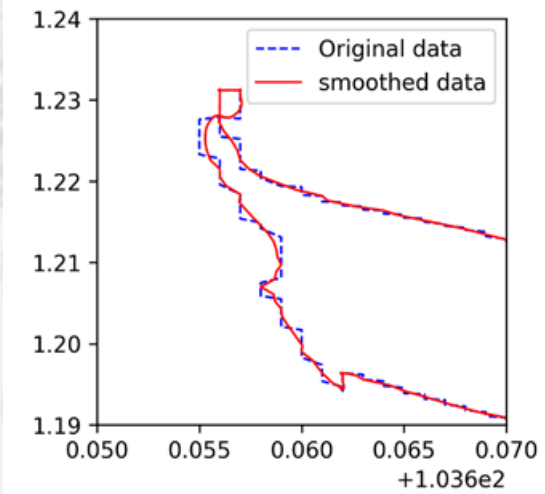
f. Weighted polynomial regression model (n=10)



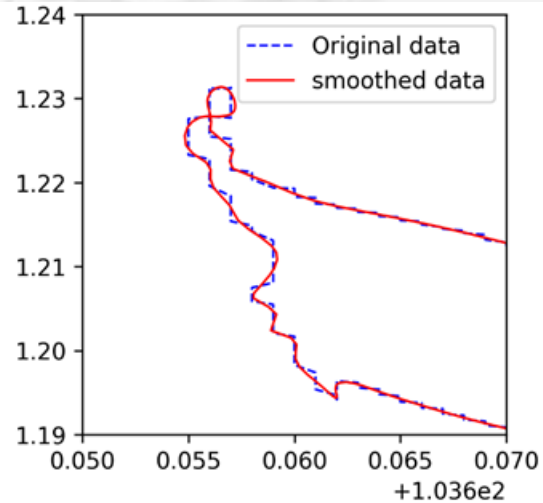
c. Polynomial regression model (n=10)



d. Polynomial regression model (n=15)



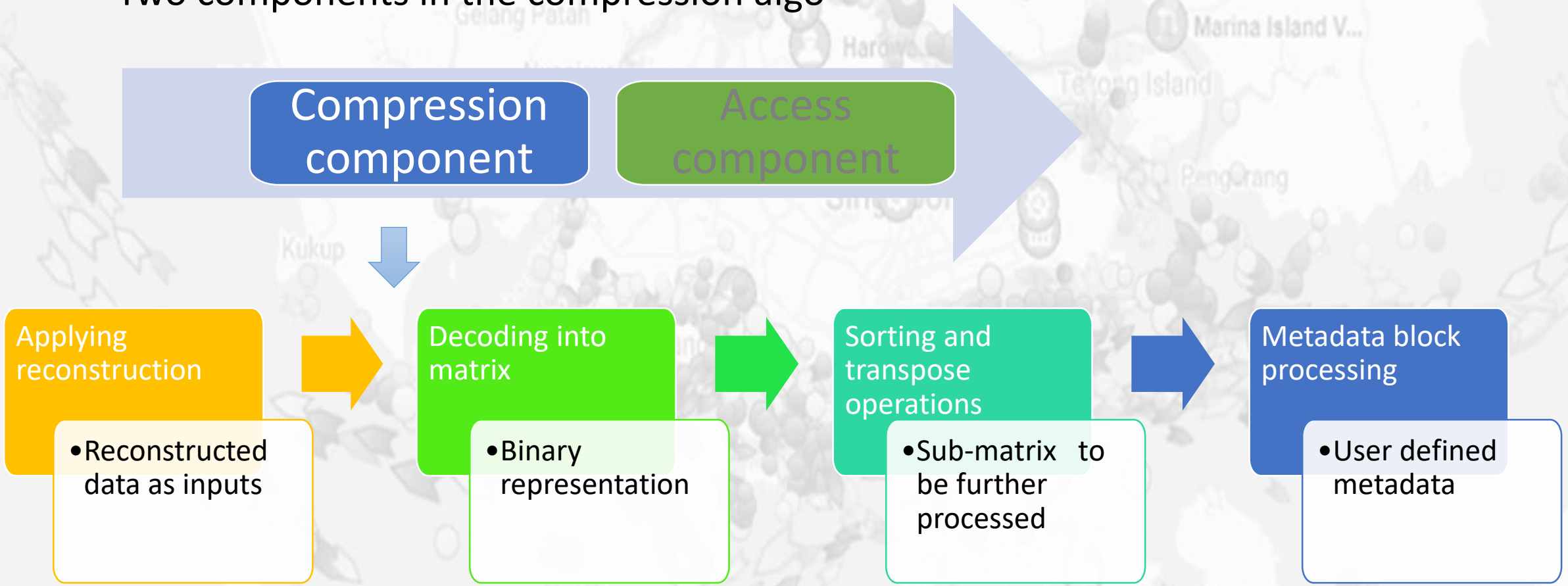
g. Weighted polynomial regression model (n=15)



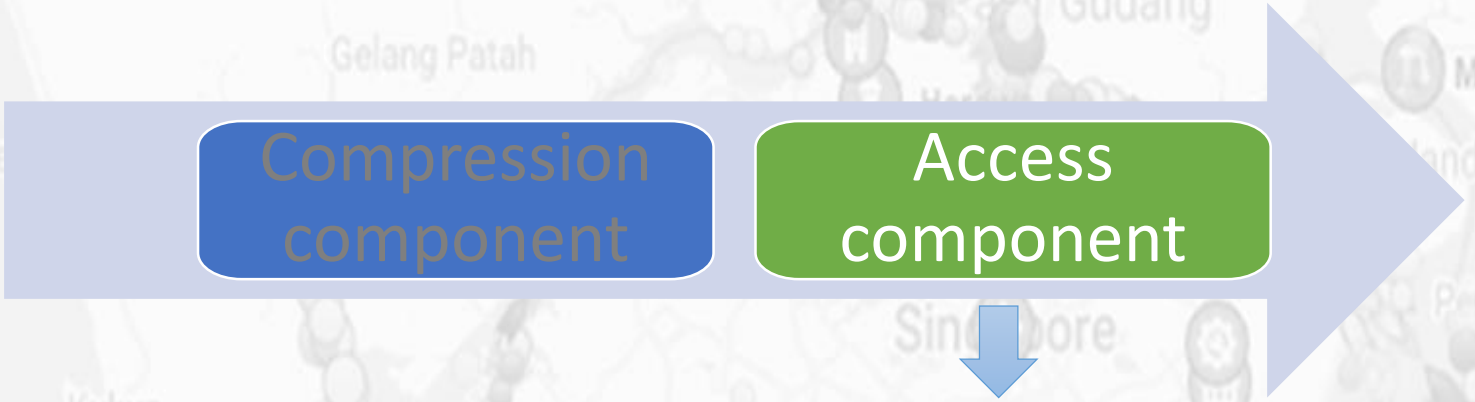
h. Multi-regime model

2. Compression Algorithm

- Two components in the compression algo



Compression Algorithm



Access management component to enable users to perform their analysis on the sub-matrix directly

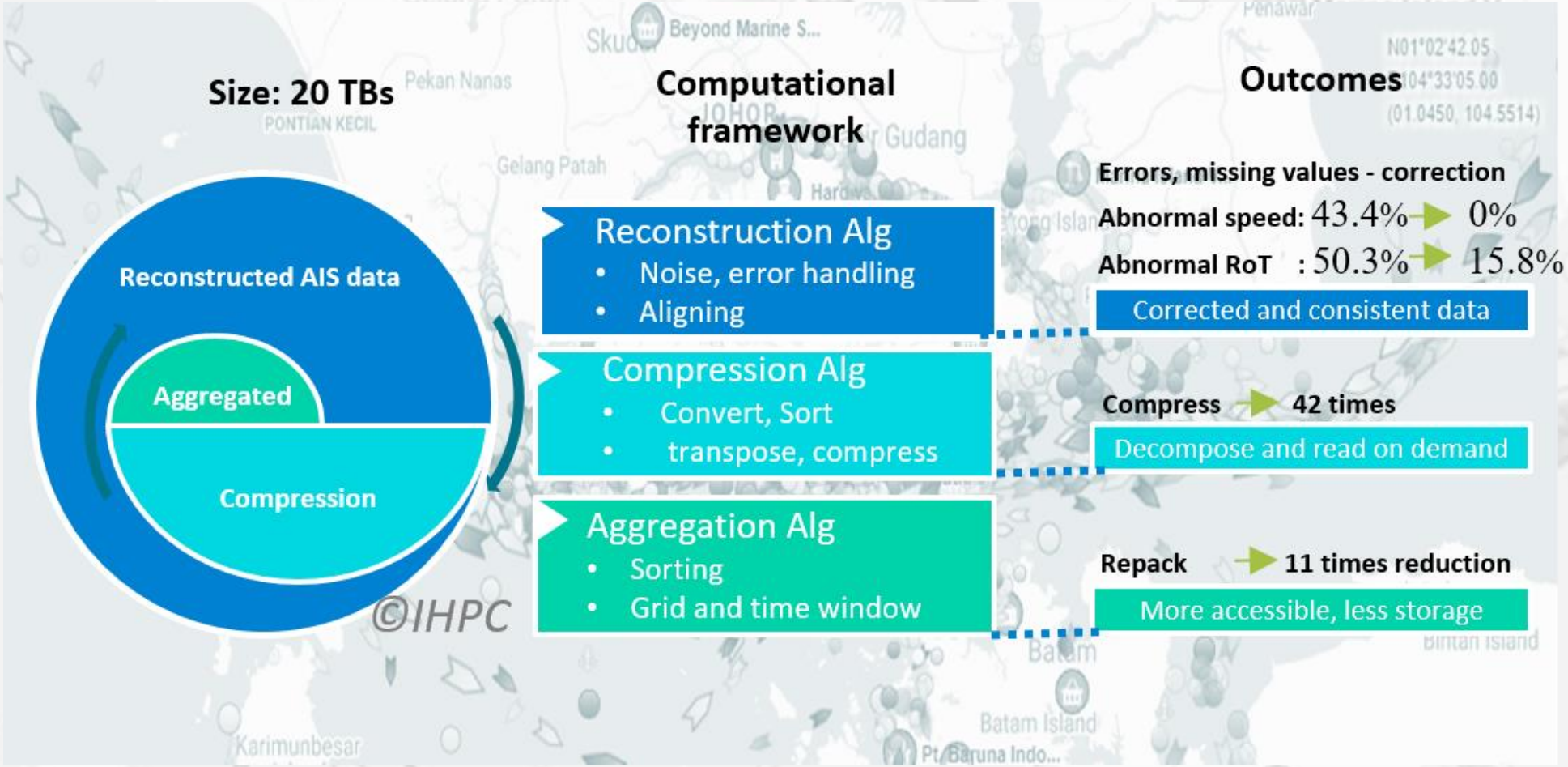
- Advantages
  - The batch partitioning and the sub-matrix decomposition offer rapid indexing into the relevant data and good cache locality.
  - The compressed sub-matrices are only decompressed when needed, reducing disk and space memory requirements.
  - Decompressed contents are cached so that subsequent accesses are fast.
  - The output files are self-contained archive files and can be replicated across a computing cluster for parallel flows.

3. Aggregation Algorithm

- Aggregation algorithm
  - For vessels within port entering any grid cell, record relevant features of the vessel movements within the cell.
  - Focus onto the S3 (SLOW SPEED/STATIONARY) state where the vessel is more likely to be performing loading/unloading, anchoring or bunkering activities.
- Advantages
  - Achieves a higher compression rate, while retaining the necessary information, provided that the aggregation algorithm is carefully designed (e.g., selection of attributes, grid cells' size, length of time window and interested derivatives).
  - Information for one vessel for its entire journey, which is stored in different AIS data files, can be combined and stored in one aggregated file.
  - Easy to index and search through to find the necessary information of each vessel at given conditions.

Results:Efficient Large Scale Spatiotemporal Data Processing

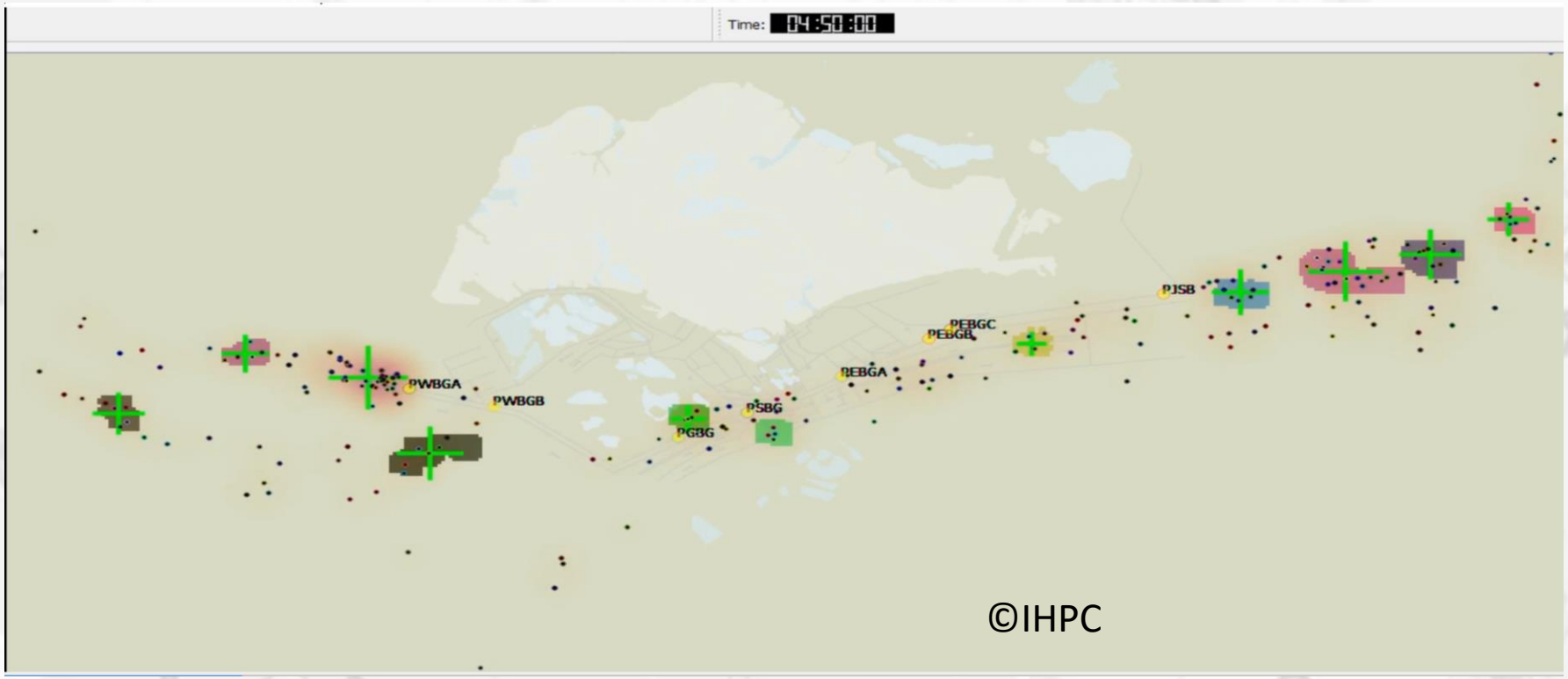
Big data handling for extracting insights



Descriptive Analytics: Maritime traffic management – hotspot detection

Motivation: Detect Traffic Hotspots (congested areas, risky areas) for proactive traffic management and JIT operation

Objective: Developing AI Algorithms for detecting and forecasting hotspots efficiently and accurately



Early morning period, hotspots observed at pilot boarding areas

Institute of High Performance Computing
A*STAR

Descriptive Analysis: Vessel Voyage Events Detection – Data Insights

Motivation: Decipher Vessel Voyage Events (berthing, anchorage, bunkering, slowing down) for future happenings to benefit operations and planning to improve efficiency.

Objective: Developing Efficient and Accurate Algorithms for efficiently detecting historical events & predicting.

95% of berths and 99.5% of berth stay duration were correctly identified

Conclusions

- Big spatio-temporal data preprocessing framework and model with the following features:
  - Efficient trajectory reconstruction for handling with errors, missing data and misalignments
  - The batch partitioning and the sub-matrix decomposition for compression offer rapid indexing into the relevant data and good cache locality for data analytics access, reducing disk space and memory requirements
    - The output files are self-contained archive files which can be stored in a database, and replicated across a computing cluster for parallel flows.
  - The final step of applying the aggregation algorithm achieves a further aggregation rate for big data, while retaining the necessary information, so as to improve accessibility, and reduce storage demand
- The proposed framework and model developed can also be applied for pre-processing other large scale spatiotemporal data