Tropical Cyclone Event Sequence Similarity Search via Dimensionality Reduction and Metric Learning

Shen-Shyang Ho, Wenqing Tang, W. Timothy Liu

Jet Propulsion Laboratory

University of Maryland, College Park

KDD-2010

This work was carried out at the Jet Propulsion Laboratory, California Institute of Technology. It was funded by the National Aeronautics and Space Administration (NASA) Advanced Information Systems Technology (AIST) Program under grant number AIST-08-0081.

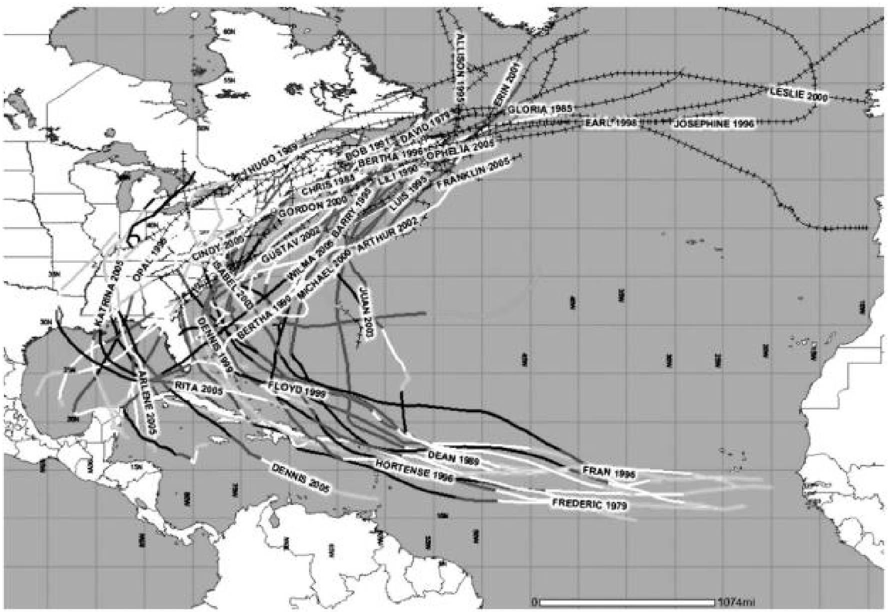
Outline

- Brief Motivation
- Methodology: Intuitions
- Some Experimental Results
- Open Problems

Brief Motivation (I)

Examples of “Conventional” similarity and objective query of interest to scientists or meteorologists:

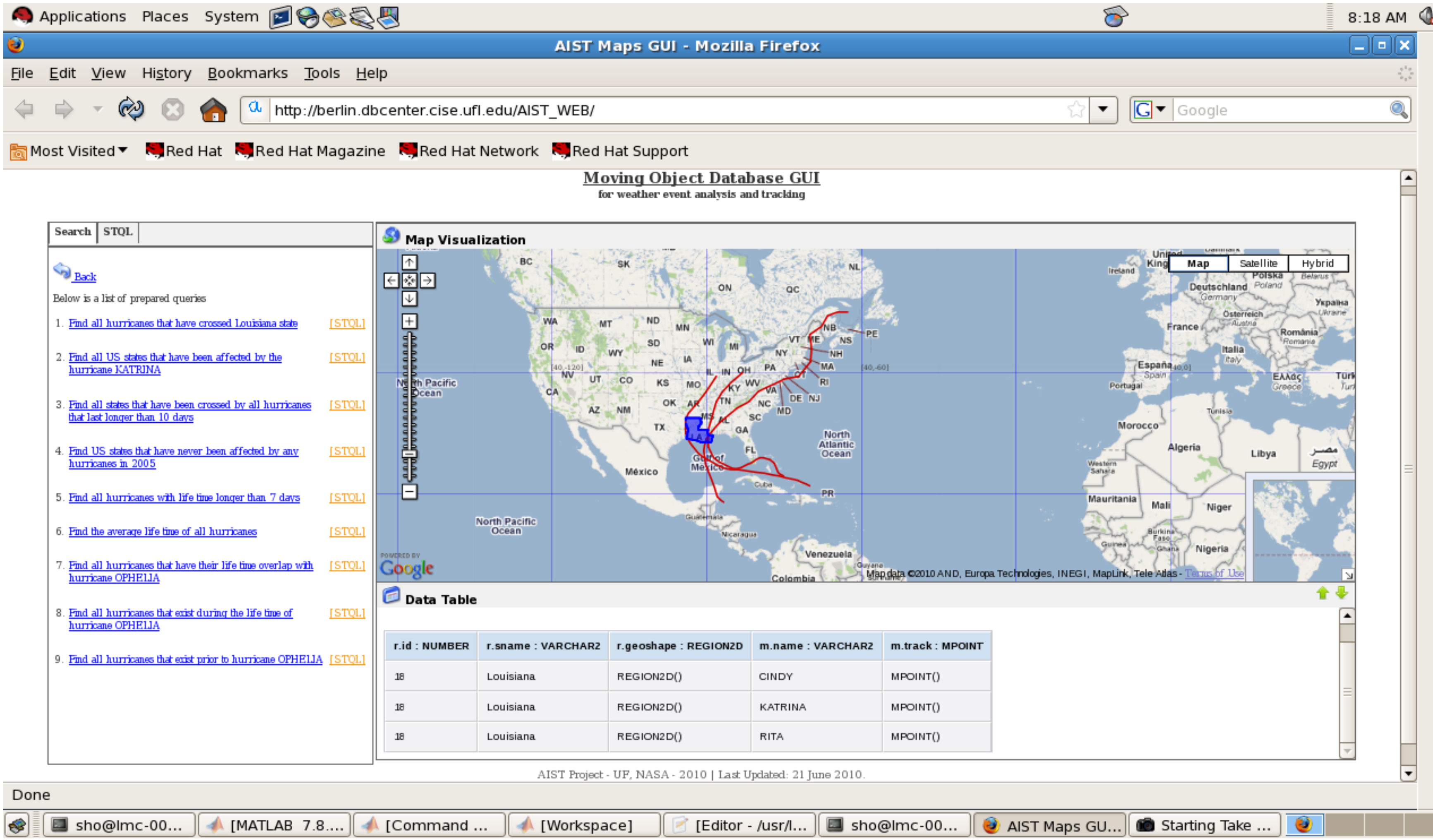
- Track Similarity: Find all hurricanes that crossed region ***A***, ***B***, ***C***, …
- Intensity Similarity: Find all hurricanes that had intensity at least *D* km/h.
- Origin Similarity: Find all hurricanes that evolved from region ***E***.



“Find all hurricanes that crossed Windsor, Ontario, and St. John's, Newfoundland from 1979 to 2005.”

Shawn M. Milrad, Eyad H. Atallah, and John R. Gyakum, ``Dynamical and Precipitation Structures of Poleward-Moving Tropical Cyclones in Eastern Canada, 1979--2005'', Monthly Weather Review, vol. 137, pp. 836-851, Mar. 2009

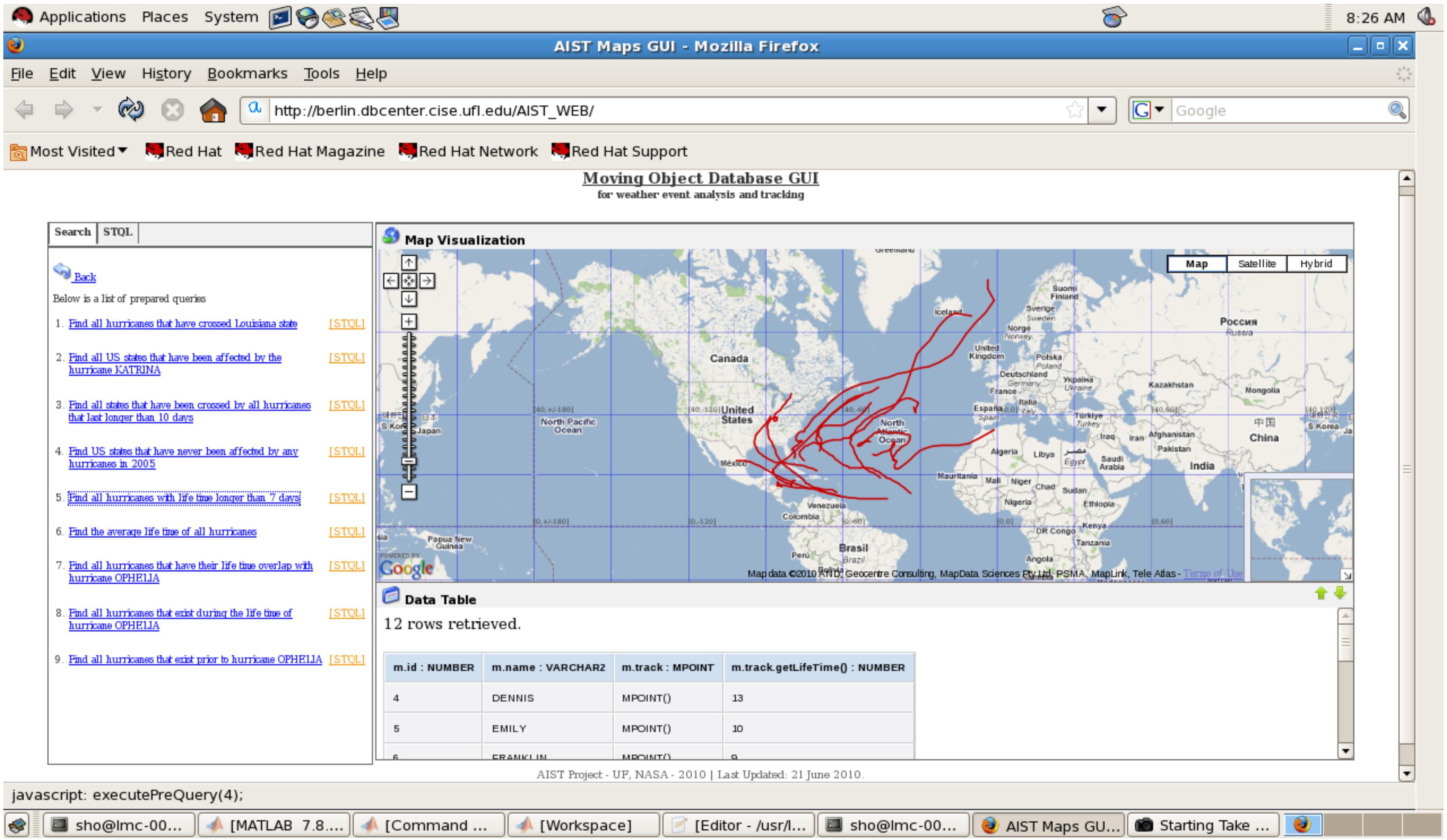
Brief Motivation (II)



Find all hurricanes that have crossed Louisiana state in 2005.

M. Schneider, S.-S. Ho, T. Chen, A. Khan, G Viswanathan, W. Tang, and W. T. Liu, Moving Objects Database Technology for Ad-Hoc Querying and Satellite Data Retrieval of Dynamic Atmospheric Events, 2010 Earth Science Technology Forum, June 22-24, Arlington, VA, 2010.

Brief Motivation (III)



Find all hurricanes that have life-time longer than 7 days in 2005

Brief Motivation (IV)

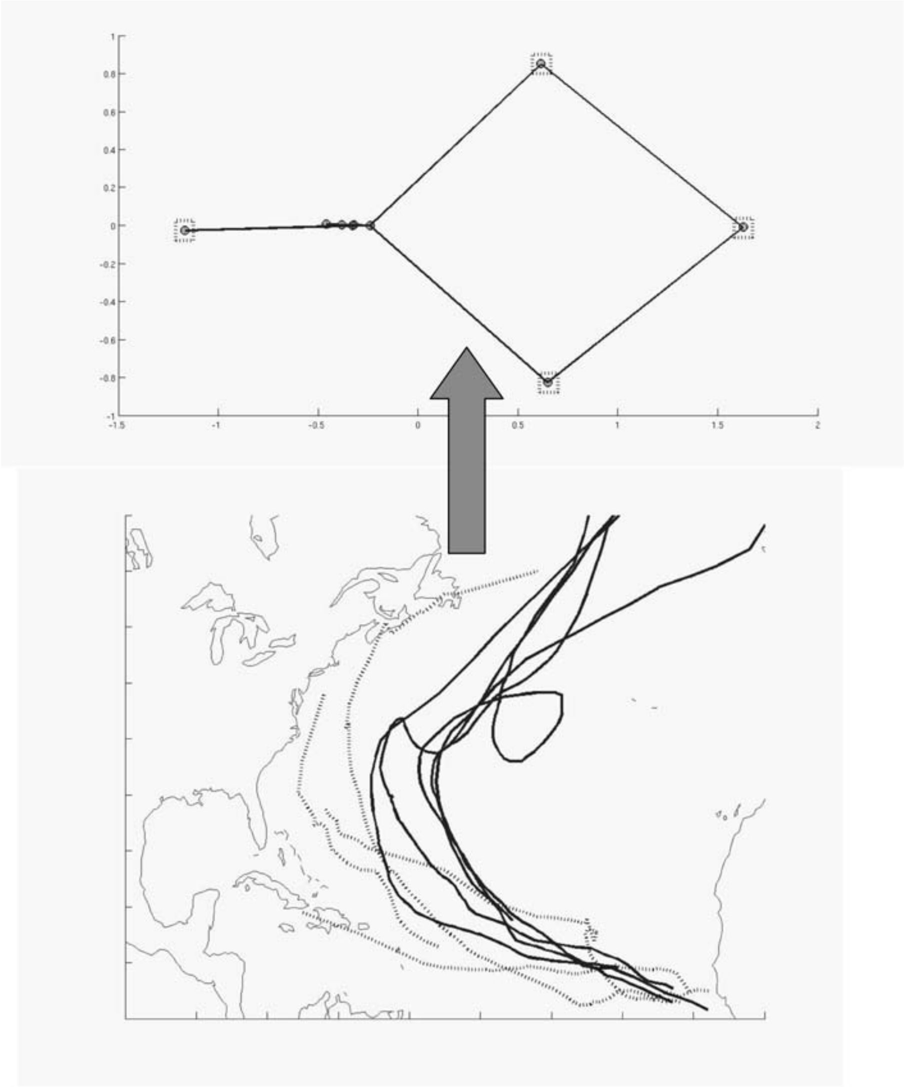
Query with instance-level constraints/user-defined "subjective" data constraints

Find all hurricanes that are *similar* to hurricane A, B, C, … and *dissimilar* to hurricane a, b, c, …

General Data Assumptions:

Multidimensional (Spatial, Temporal, Features/Attributes)
Arbitrary Length Sequences

Methodology (I)



Standard Data Mining Techniques Used:

- **Similarity Metric: Longest Common Subsequence (LCSS) – [Vlachos et al., ICDE, 2002] - generalize to multidimensional sequences**

$$S1(A,B,\delta,E) = \frac{LCSS_{\delta,E}(A,B)}{\min(|A|,|B|)}$$

$$LCSS_{\delta,E}(A,B) = \begin{cases} 0 : & \text{if } |A| = 0 \text{ or } |B| = 0 \\ 1 + LCSS_{\delta,E}(Head(A), Head(B)) : & c_k > 0, \forall c_k, \ |t_i - t_j| < \delta \text{ is satisfied} \\ \max(LCSS_{\delta,E}(Head(A), B), LCSS_{\delta,E}(A, Head(B))) : & \text{otherwise} \end{cases}$$

$$\begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix} = \begin{pmatrix} \epsilon_1 - |a_{1,t_i} - b_{1,t_j}| \\ \vdots \\ \epsilon_m - |a_{m,t_i} - b_{m,t_j}| \end{pmatrix}$$

- **Dimensionality Reduction: Isometric Feature Mapping (ISOMAP) – [Tenenbaum et al, Science, 2000)**

- **Metrics Learning: [Xing et al, NIPS, 2002]**

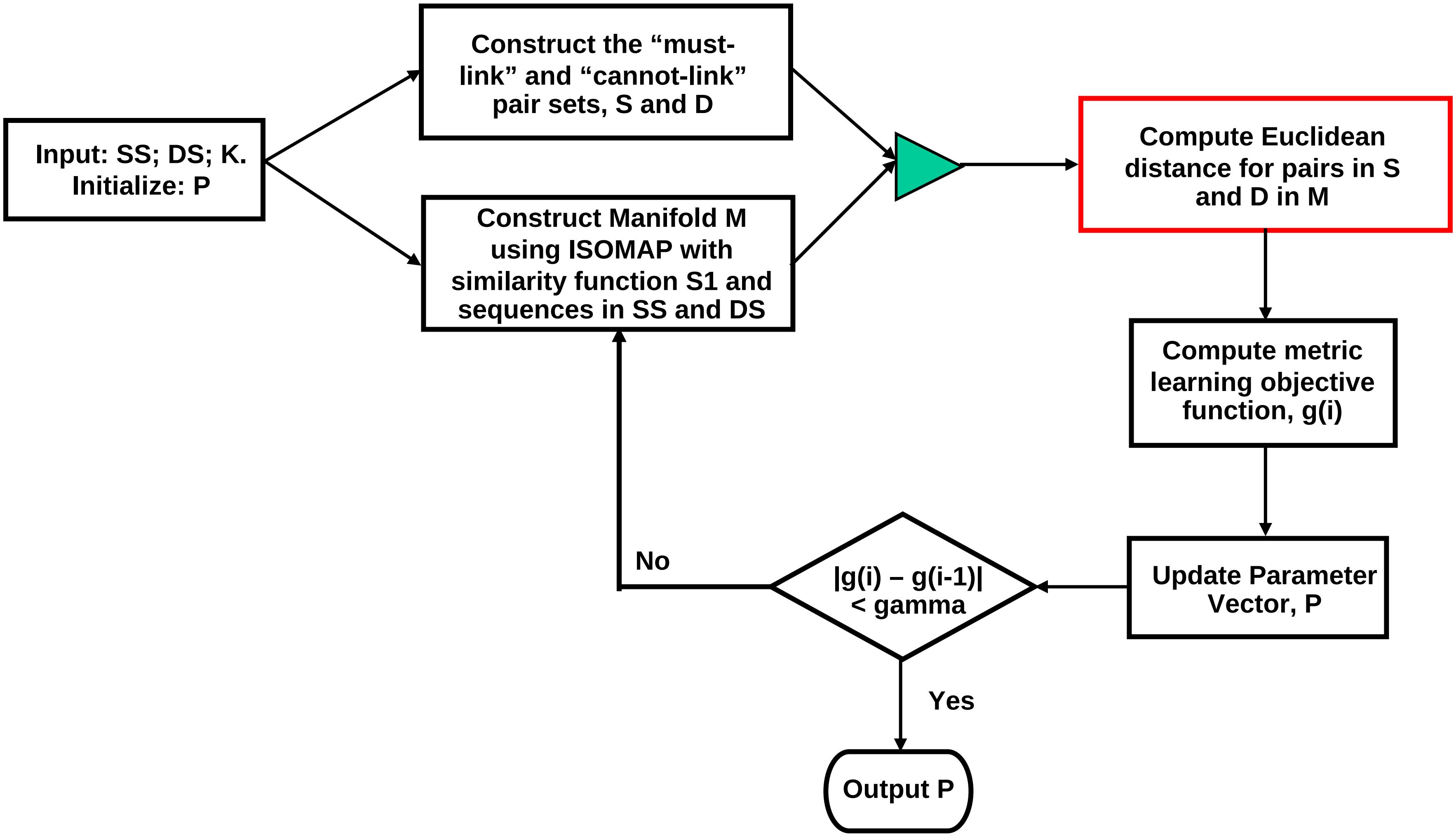
$$\min_{E,\delta} \quad \sum_{(x_i,x_j)\in S} ||f_{S1}(x_i) - f_{S1}(x_j)||^2$$

$$\text{such that} \quad \sum_{(x_i,x_j)\in D} ||f_{S1}(x_i) - f_{S1}(x_j)| \geq 1$$

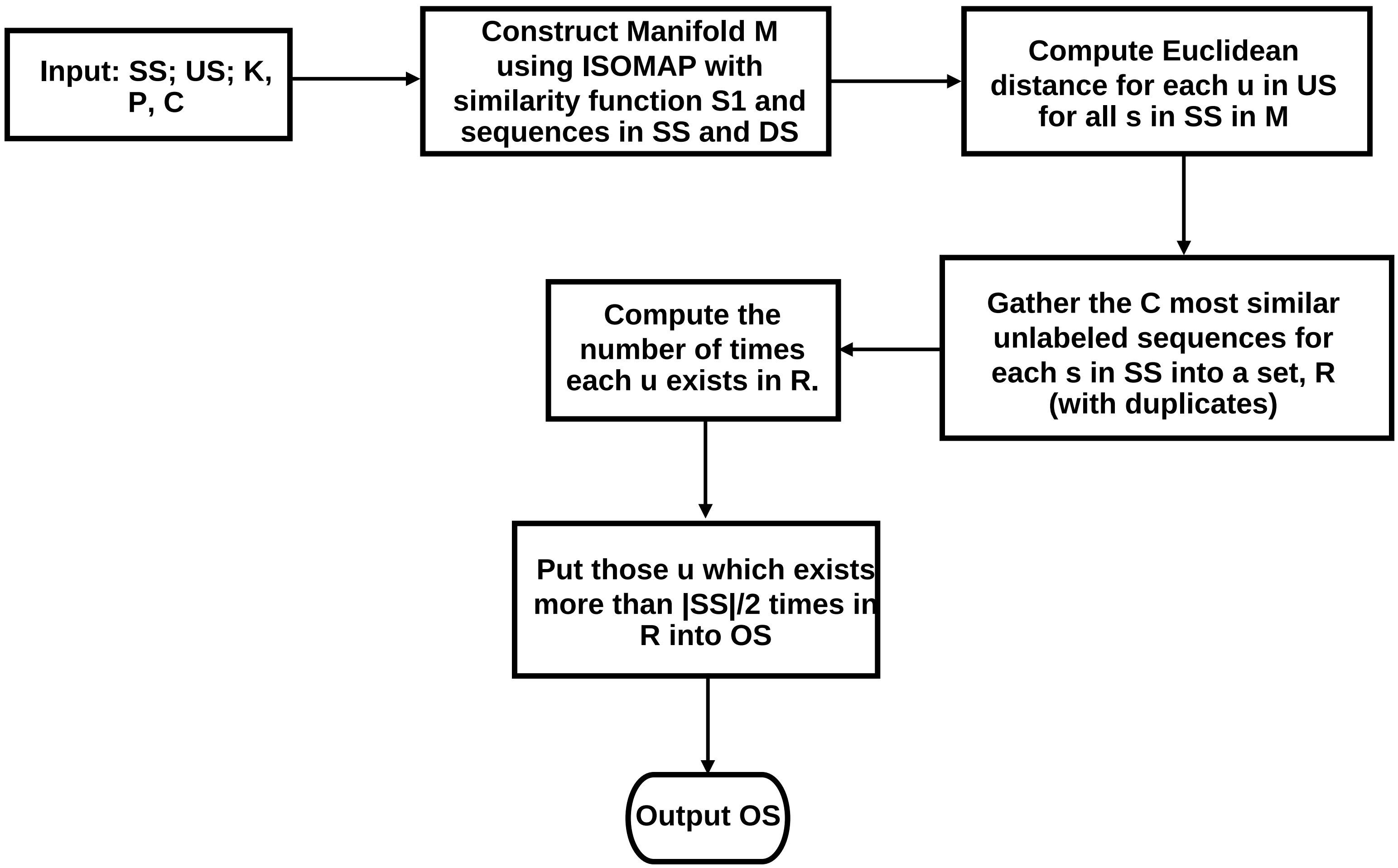
$$\text{and } P > 0 \qquad \text{where } P = (\epsilon_1, \epsilon_2, \ldots, \epsilon_m, \delta) \in (R^+)^m \times Z^+.$$

- **Similarity Search: Voting Approach**

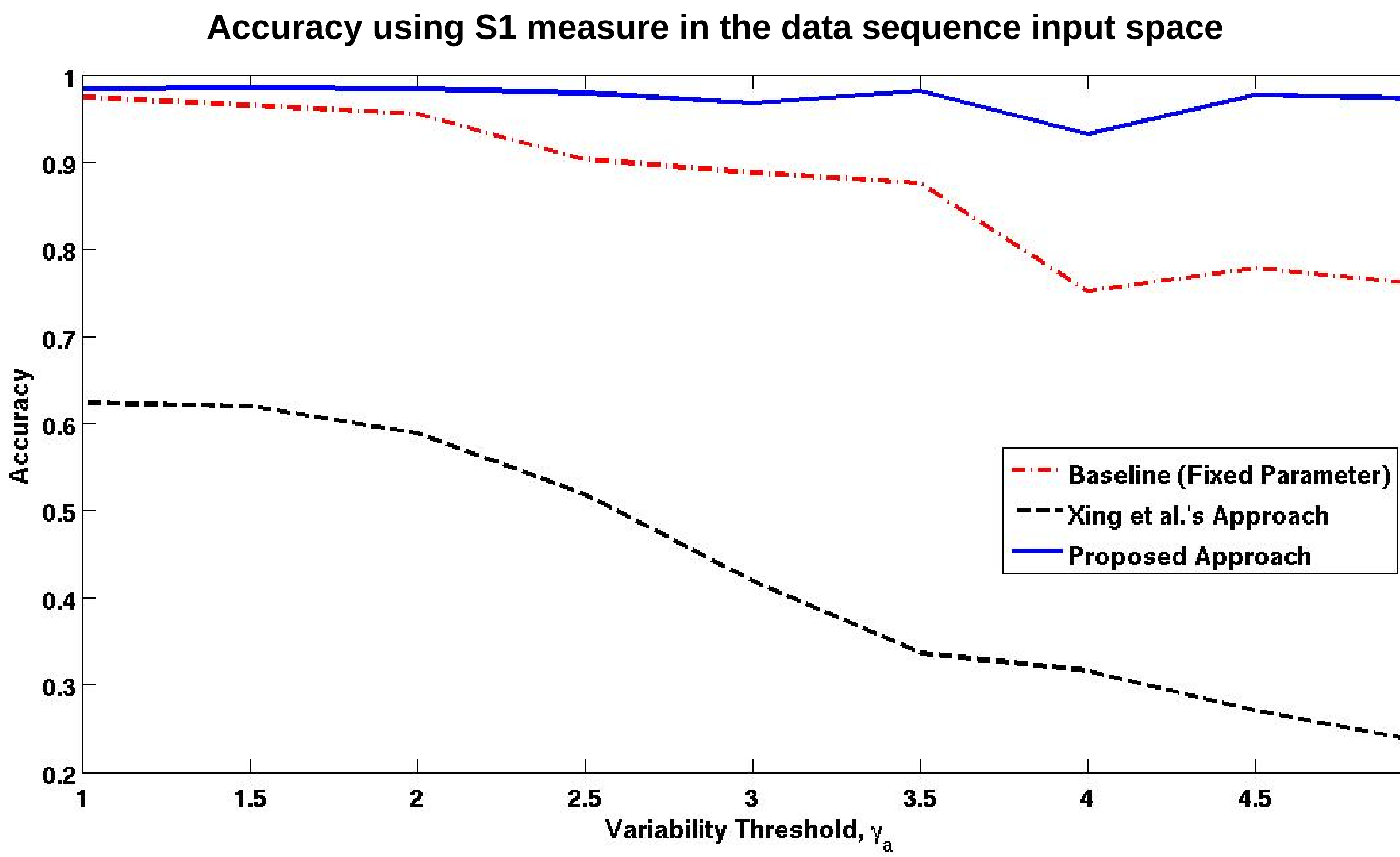
Methodology (III) – Learning LCSS Parameter P



Methodology (IV) – Similarity Search on Unlabeled Sequences U'

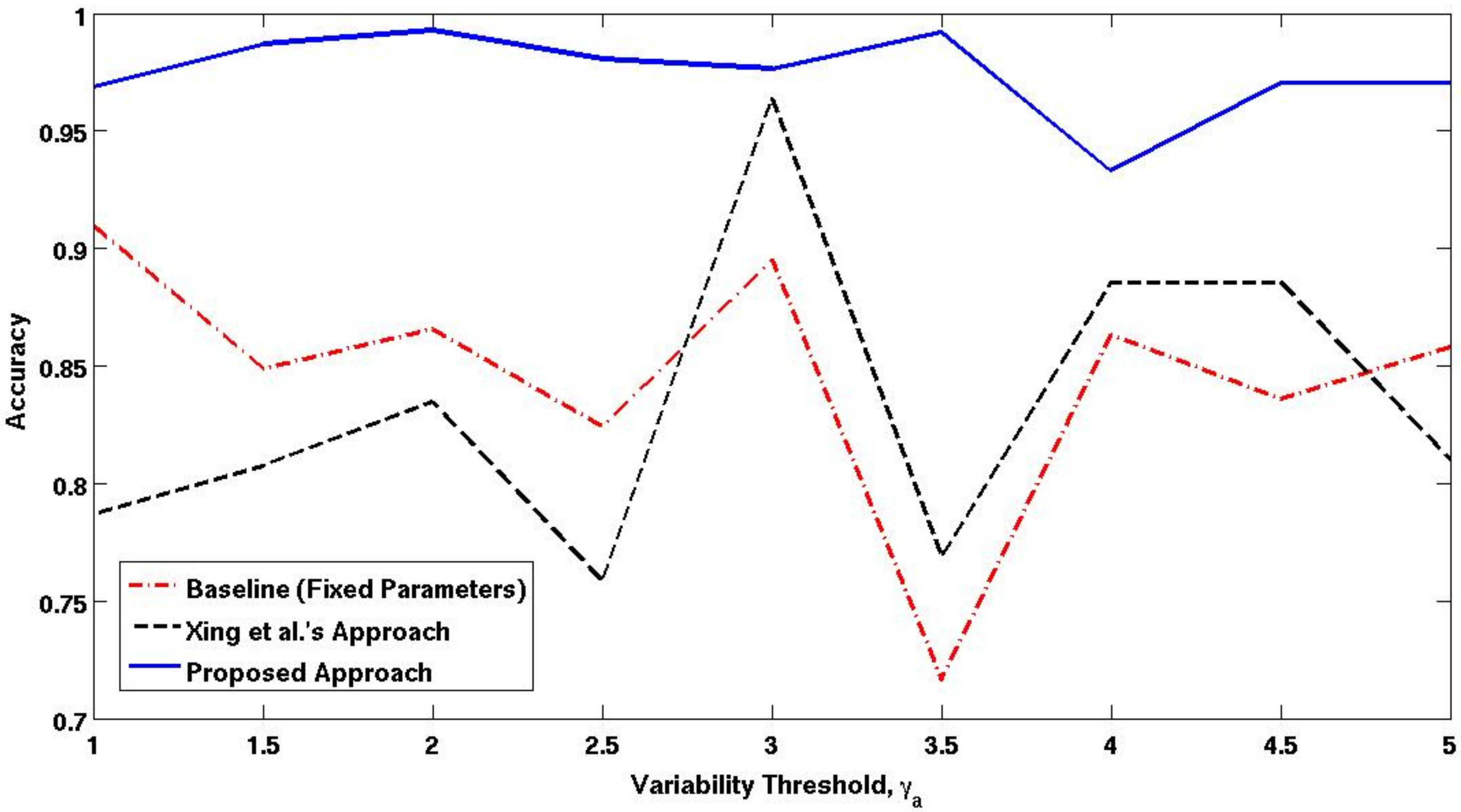


Experimental Results (I) – Learning P

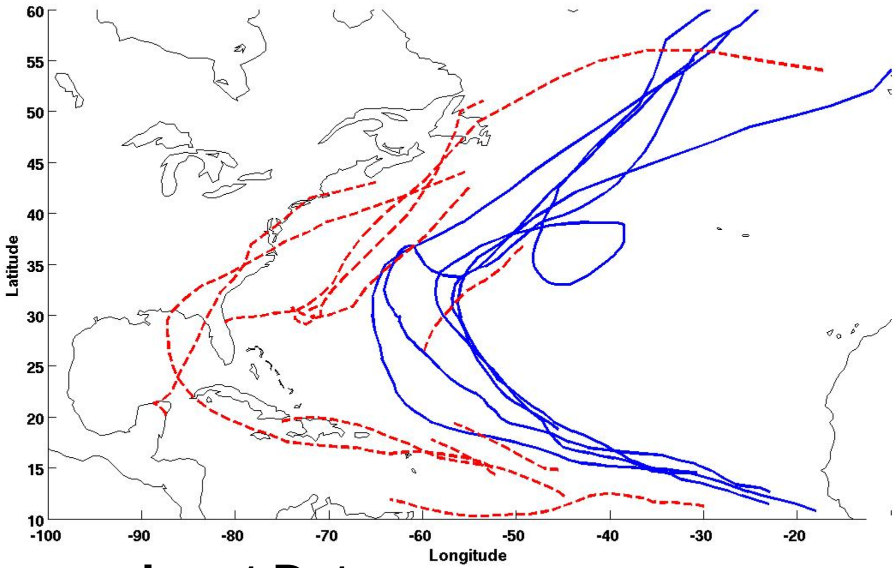


Experimental Results (II) – Learning P

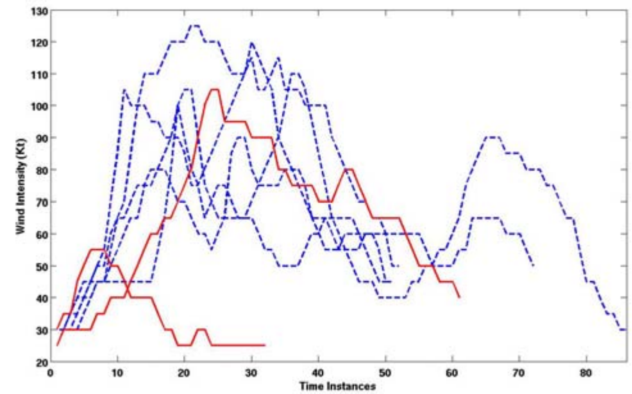
Accuracy using Euclidean distance in the low dimensional manifold

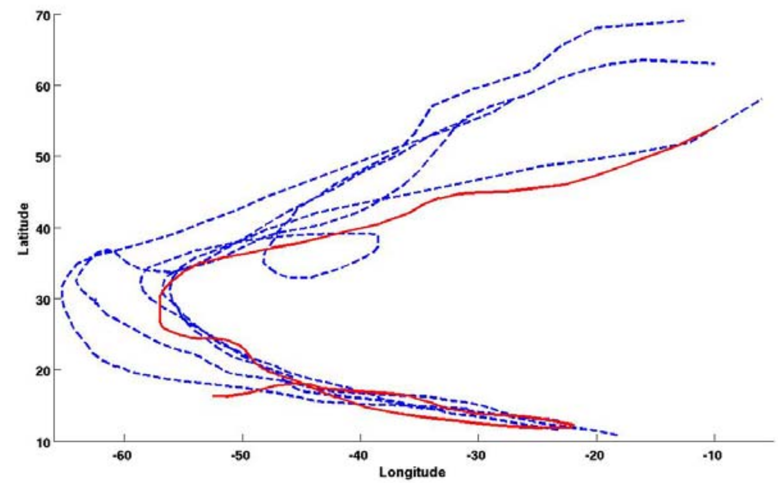


Experimental Results (III) – Similarity Search



Input Data





Output Result (Red)

Open Problems

- Theoretical justification for the setting and data assumptions.
- Which is the best similarity metric, dimensionality reduction approach, and metric learning approach for the framework?
- Local/partial sequence similarity search
- Other Applications …

See You at the Poster !!