# Error Metrics for Learning Reliable Manifolds from Streaming Data

Frank Schoeneman, Suchismit Mahapatra, Varun Chandola, Nils Napp, and Jaroslaw Zola

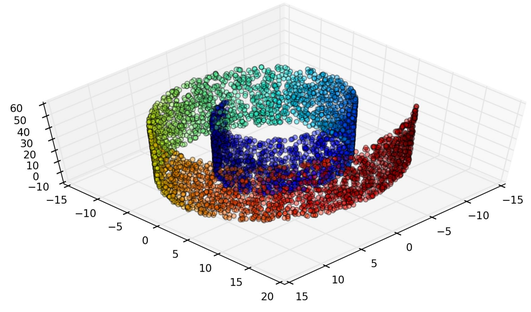
Department of Computer Science and Engineering
University at Buffalo, The State University of New York

## Motivation

- Progress in science and engineering depends more than ever on our ability to analyze huge amounts of data.
- Huge amounts of data is coming from high-performance high-fidelity numerical simulations, high-resolution scientific instruments or Internet of Things feeds.
- Real-world data is typically a result of complex non-linear processes, but can often times be described by a low-dimensional manifold.
- Manifold learning has numerous applications in scientific computing and bio-medical research, including fMRI analysis, clustering of oncology data, gene profiling, and many others.
- Isomap is a common method for manifold learning that has been widely-adopted in many application areas.

## Challenges

- Widely used manifold learning methods have been designed for off-line or batch processing.
- Standard methods are computationally expensive or impractical to apply to high-throughput data streams.
- Error in manifold learning is not yet completely understood, making error measurement on streaming data all the more complex.
- Until now, applying Isomap to data streams and understanding the collective error has not been well studied.



## Proposed Approach

- We formalize a notion of *collective error* in Isomap and describe strategies to quantify it using Procrustes Analysis.
- We assume that there exists a function $f\colon \mathbb{R}^d \rightarrow \mathbb{R}^D$ which maps $\boldsymbol{y_i} \in \mathbb{R}^d$ to $\boldsymbol{x_i} \in \mathbb{R}^D$. The goal of Isomap is to learn the inverse mapping, $f^{-1}(\cdot)$ that can be used to map high-dimensional $\boldsymbol{x_i}$ to low-dimensional $\boldsymbol{y_i}$ i.e. $\boldsymbol{y_i} = f^{-1}(\boldsymbol{x_i})$.
- Let $Y$ denote the data matrix containing the true low-dimensional mapping for the samples in $X$, i.e. $Y = [\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_n}]^T$ and $\hat{Y}$ denote the matrix with the approximate mapping, i.e. $\hat{Y} = \left[\hat{f}^{-1}(\boldsymbol{x_1}), \hat{f}^{-1}(\boldsymbol{x_2}), \cdots, \hat{f}^{-1}(\boldsymbol{x_n})\right]^T$.
- We then use Procrustes error to measure the difference between the true low-dimensional samples and the approximate samples.

$$\epsilon = d_{Proc}(Y, \hat{Y}) = \min_{R,t,s} \left\| sR\hat{Y} + t - Y \right\|_F$$

- We propose a new efficient algorithm call *Streaming Isomap (S-Isomap)* to incorporate streamed data into a stable Isomap manifold.

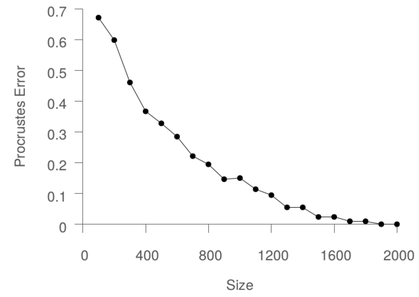
**Algorithm 1** STREAMING ISOMAP (S-ISOMAP)

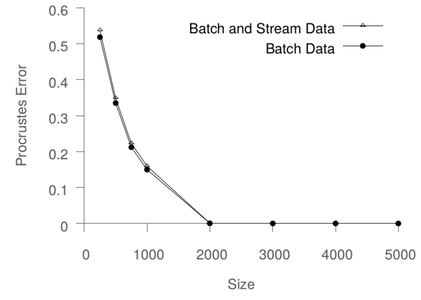
**Input:** $G_b$, $X_b$, $\mathbf{x}_s$, $k$
**Output:** $\mathbf{y_s}$

1: $\mathbf{kNN}, \mathbf{kDist} \leftarrow \mathrm{KNN}(\mathbf{x_s}, X_b, k)$ //Determine kNN
2:
3: **for** $1 \leq i \leq n$ **do**
4: $\quad g_i \leftarrow \min_{1 \leq j \leq k} \{\mathbf{kDist}_j + (G_b)_{kNN_j, i}\}$
5:
6: $\mathbf{g}_i^2 \leftarrow \mathbf{g_i}^2$
7: $\mathbf{G}_{ij}^2 \leftarrow \mathbf{G_{ij}}^2$
8:
9: $\mathbf{c} \leftarrow \frac{1}{2}(\bar{\mathbf{g}}^{\mathbf{2}} \cdot \mathbf{1}_n - \mathbf{g}^2 - \bar{\bar{G}} \cdot \mathbf{1}_n + \bar{\mathbf{G}})$
10: $\mathbf{p} \leftarrow (Y_b^\top Y_b)^{-1} Y_b^\top \mathbf{c}$ //least squares estimate
11: $\hat{Y} \leftarrow [Y_b; \mathbf{p}]$
12: $\mathbf{y}_s \leftarrow \mathbf{p} - \hat{Y}$
13: **return** $\mathbf{y}_s$

## Experiments

- We investigate the behavior of error in Isomap when applied to the Swiss Roll dataset, where we have low-dimensional ground truth available for comparison.
- A Reference-Sample method, for use where ground truth is not available, is shown to have similar behavior, making it a reliable method for error measurement on real-world data.
- Using these strategies to quantify error, we identify a **transition point** where a reliable manifold has been learned, and we can switch to our lightweight algorithm, S-Isomap.




## Reference

- F. Schoeneman, S. Mahapatra, V. Chandola, N. Napp, and J. Zola. Error Metrics for Learning Reliable Manifolds from Streaming Data. Under review for *SIAM International Conference on Data Mining*, 2017.