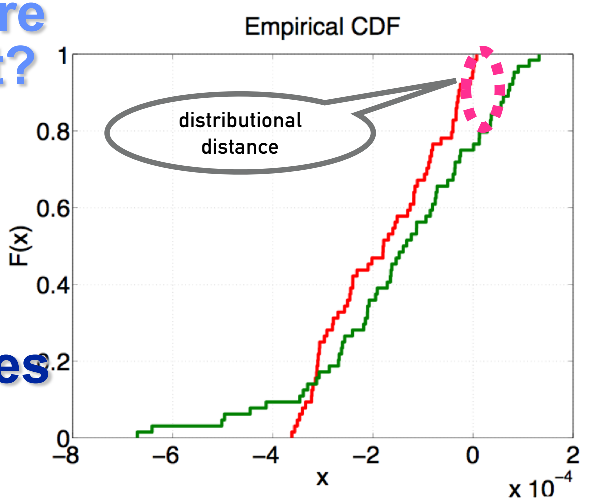
Similarity-based Compression with Multidimensional Pattern Matching

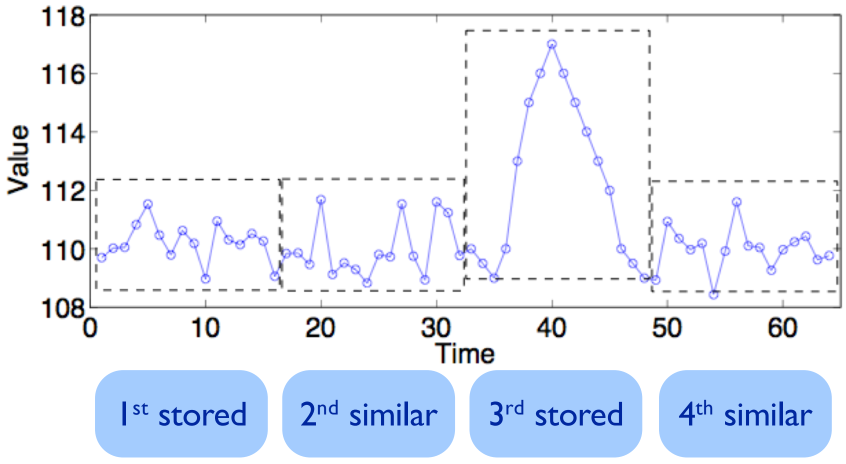
Olivia Del Guercio[1], Rafael Orozco[2]
Alex Sim[3], K John Wu[3]

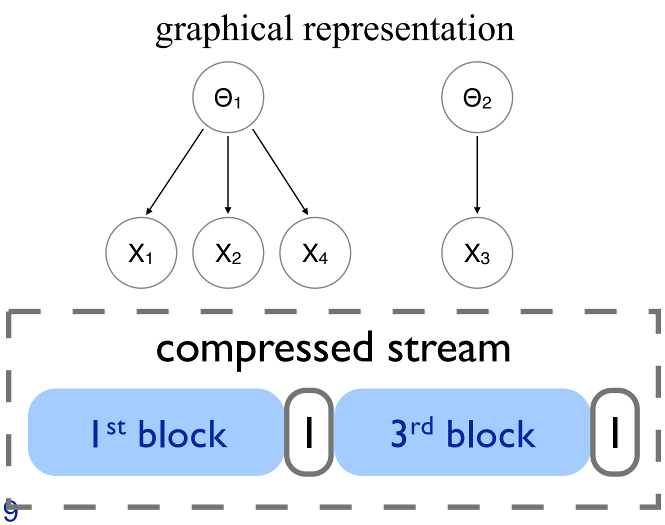
1. Scripps College
2. Bucknell University
3. Scientific Data Management Research Group
Computational Research Division
Lawrence Berkeley National Laboratory

Locally Exchangeable Measures – New Perspective on Data Compression

- **Question:** random-looking sequence of values are hard to compress, can we do something about it?
- **Answer:** IDEALEM (Implementation of Dynamic Extensible Adaptive Locally Exchangeable Measures) @ SSDBM2016, 2017, BigData2018, DCC2019
- **Dictionary compression with alternative measures of distance**
  - Based on Kolmogorov-Smirnov test (KS test)
  - Distributional distance/similarity of two random variables

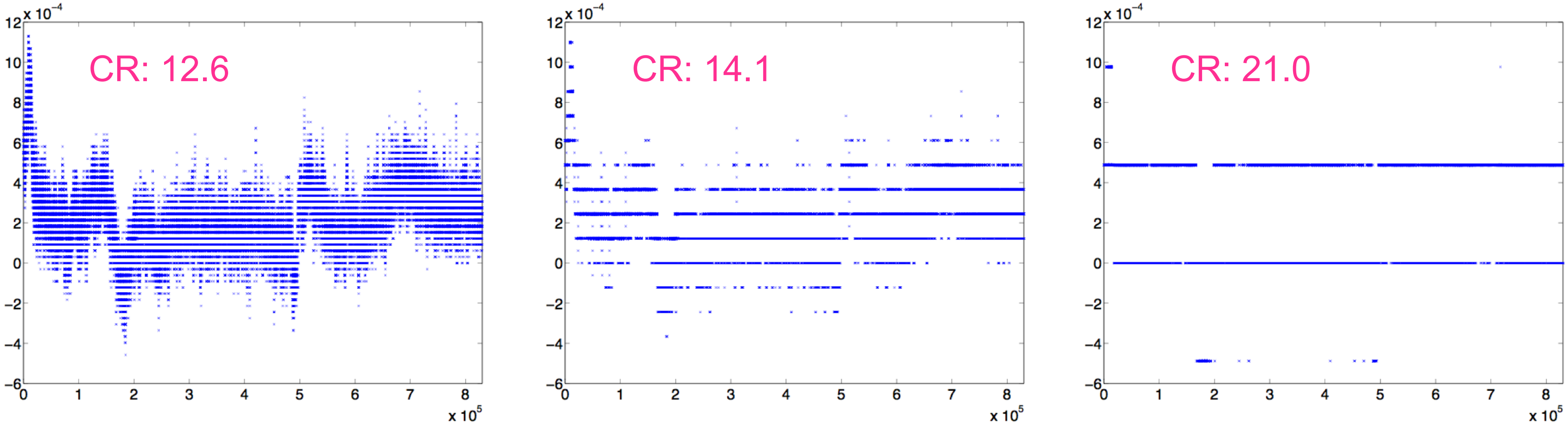




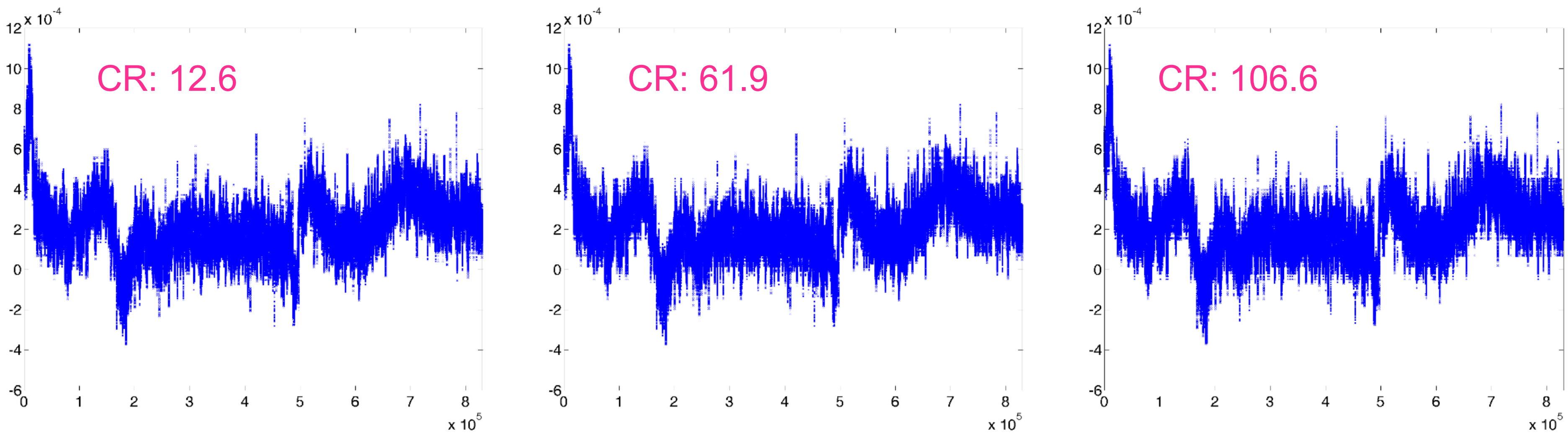


IDEALEM Achieves High Compression Ratio and High Reconstruction Quality on Power Grid Data

ZFP

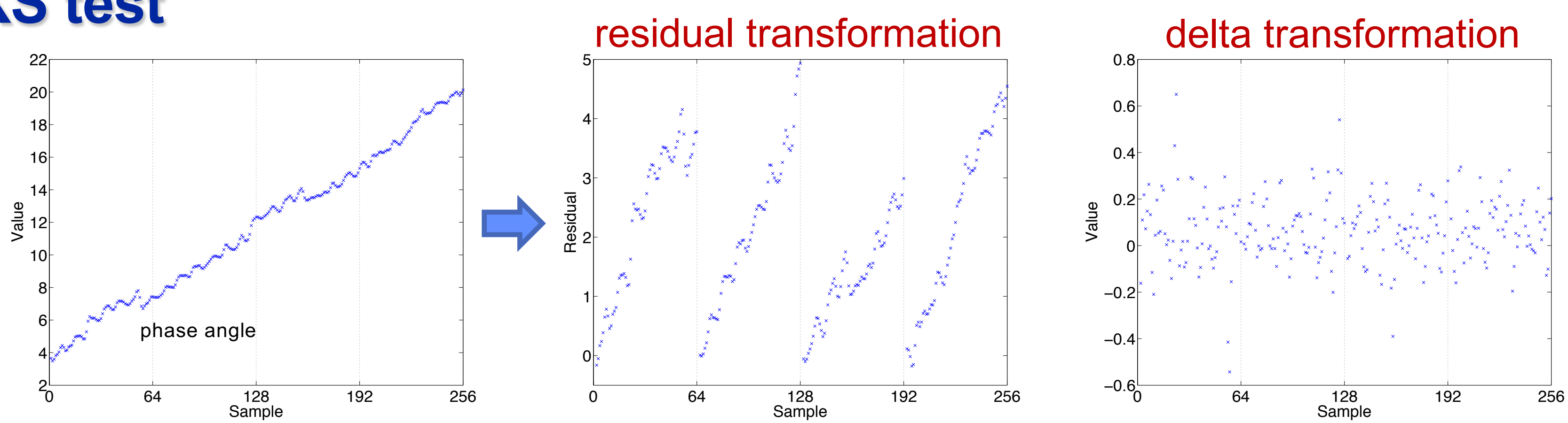


IDEALEM



IDEALEM Extension – Non-Stationary Data

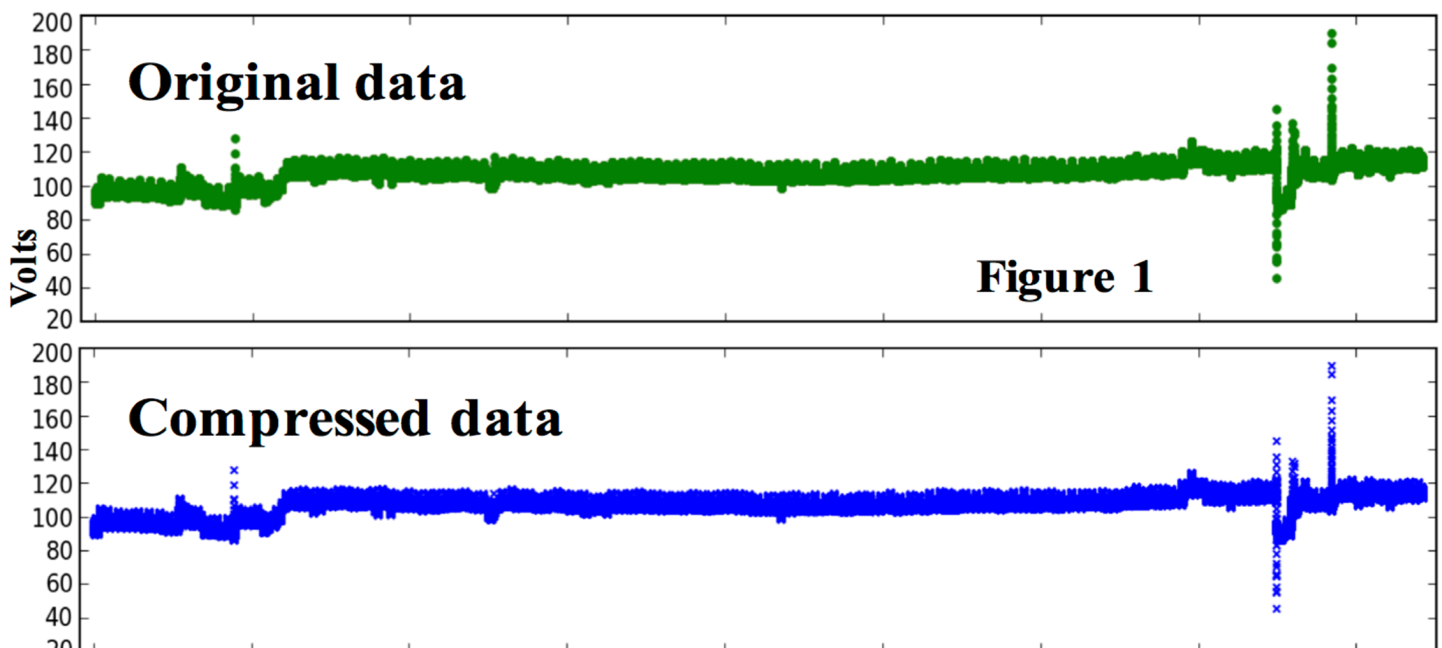
- Early IDEALEM was not so effective for non-stationary data such as phase angle of electricity data
- IDEALEM now offers two methods for transforming non-stationary data into locally stationary block to promote exchangeability/similarity
    - Residual transformation
    - Delta transformation
- These methods allow local variations to be compared through KS test



compressible using KS test!

New IDEALEM Extension – Multidimensional Data

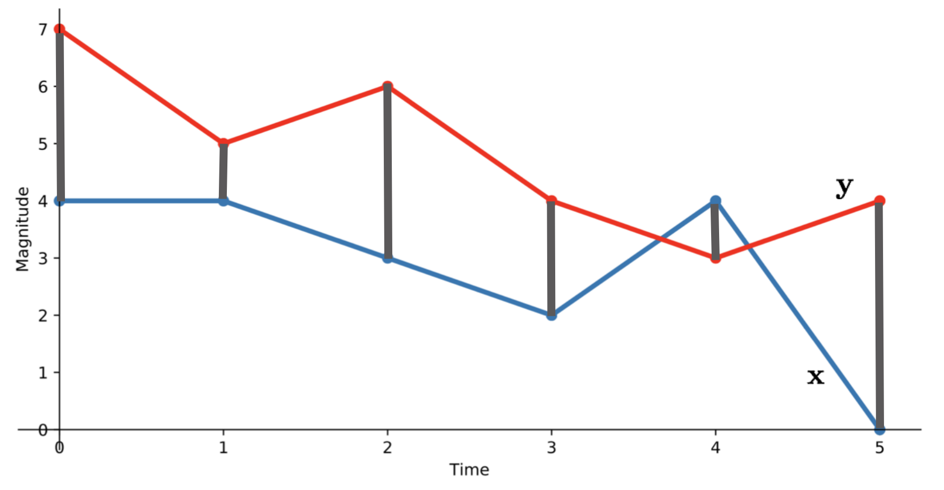
- Original IDEALEM algorithm only supports one dimensional data with K-S test.
- This paper is about extending the algorithm to support 2D and n-dimensional data with multidimensional similarity measures



Multidimensional Similarity Measures

- **Common measure between two time series**

$$d_{\mathrm{Euc}}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{M}(x_i - y_i)^2}$$



- **From many alternative similarity measures, we selected Dynamic Time Warp (DTW) and Minimum Jump Cost (MJC)**

Dynamic Time Warp (DTW)

- **DTW performs nonlinear "warping" on the sequences where differences in time are not penalized**
  - For time series of length n, it is necessary to do n^2 computations (through Dynamic Programming)

$$d_{\mathrm{DTW}}(\mathbf{x}, \mathbf{y}) = D_{M,N}$$

$$D_{i,j} = d_{\mathrm{Euc}}(x_i, y_i) + \min\{D_{i,j-1}, D_{i-1,j}, D_{i-1,j-1}\}$$

DTW example

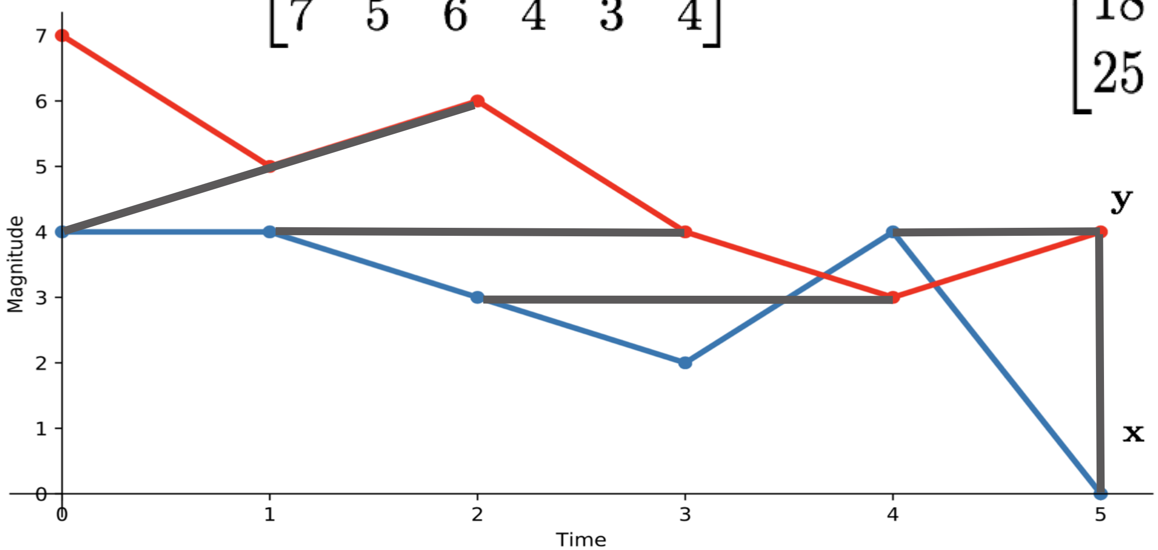
$$d_{\mathrm{DTW}}(\mathbf{x},\mathbf{y}) = D_{M,N}$$

$$D_{i,j} = d_{\mathrm{Euc}}(x_i, y_i) + \min\{D_{i,j-1}, D_{i-1,j}, D_{i-1,j-1}\}$$

x=(4, 4, 3, 2, 5, 0)
y=(7, 5, 6, 4, 3, 4)
$A_{i,j} = d_{\mathrm{Euc}}(x_i, y_j)$

$$A(\mathbf{x},\mathbf{y}) = \begin{bmatrix} 3 & 1 & 2 & 0 & 1 & 0 \\ 3 & 1 & 2 & 0 & 1 & 0 \\ 4 & 2 & 3 & 1 & 0 & 1 \\ 5 & 3 & 4 & 2 & 1 & 2 \\ 3 & 1 & 2 & 0 & 1 & 0 \\ 7 & 5 & 6 & 4 & 3 & 4 \end{bmatrix}$$

$$D(\mathbf{x},\mathbf{y}) = \begin{bmatrix} 3 & 4 & 6 & 6 & 7 & 7 \\ 6 & 4 & 6 & 6 & 7 & 7 \\ 10 & 6 & 7 & 7 & 6 & 7 \\ 15 & 9 & 10 & 12 & 7 & 8 \\ 18 & 10 & 11 & 10 & 8 & 7 \\ 25 & 15 & 16 & 14 & 11 & 11 \end{bmatrix}$$



$$d_{\mathrm{DTW}} = D_{6,6} = 11$$

Minimum Jump Cost (MJC)

- MJC works by accumulating the cost of jumping forward from one time series data point to the nearest data point in the other time series

$$d_{\mathrm{MJC}} = \sum_i c_{\min}^{(i)}$$

$$c_{\min}^{(i)} = \min\{c_{t_x}^{t_y}, c_{t_x}^{t_y+1}, c_{t_x}^{t_y+2}, \ldots\}$$

- Instead of calculating all n^2 distance values of between x and y, only the distance between points of index greater than the recursive starting point are calculated
- Expected to reduce runtime

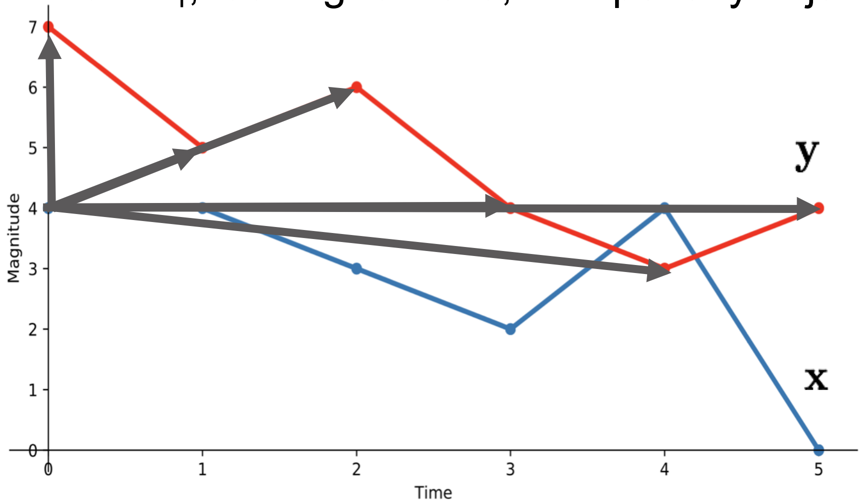
Minimum Jump Cost (MJC)

- **Minimum Jump Cost (MJC)**

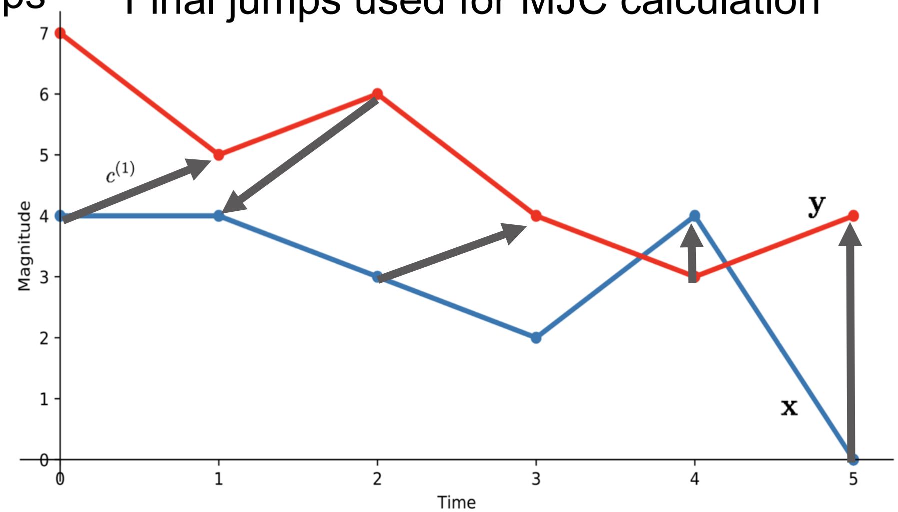
$$d_{\mathrm{MJC}} = \sum_i c_{\min}^{(i)}$$

$$c_{\min}^{(i)} = \min\{c_{t_x}^{t_y}, c_{t_x}^{t_y+1}, c_{t_x}^{t_y+2}, \ldots\}$$

From x_1, looking forward, with penalty in jumps



Final jumps used for MJC calculation



Summary of Compression Performance

- **Mean Squared Error (MSE)**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

- **Peak Signal to Noise Ratio (PSNR)**

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_x^2}{MSE}\right)$$

Similarity Measure	**Mean Squared Error (lower is better)**	**Execution Time (seconds)**
Minimum Jump Cost	**0.344**	**0.255** sec
Dynamic Time Warp	0.388	0.578 sec

Dictionary size comparison for CR = 100

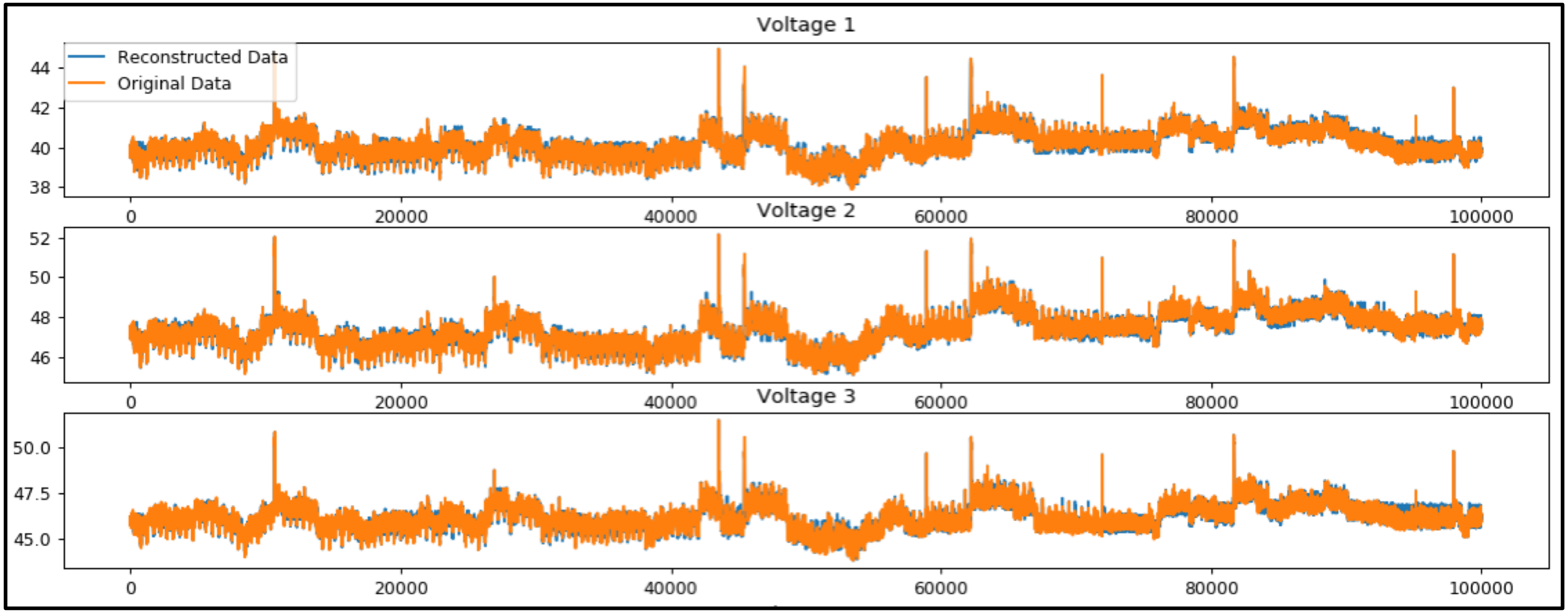
Dictionary Size		PSNR	Runtime	MSE
2	DTW	32.4	0.214	0.281
	MJC	32.6	0.183	0.262
20	DTW	34.4	0.624	0.215
	MJC	34.3	0.336	0.217
100	DTW	35.5	1.339	0.0580
	MJC	35.4	0.629	0.0586
255	DTW	34.8	1.318	0.0580
	MJC	35.4	0.686	0.0586

MJC has lower error at larger dictionary size

DTW takes 2x the time used by MJC

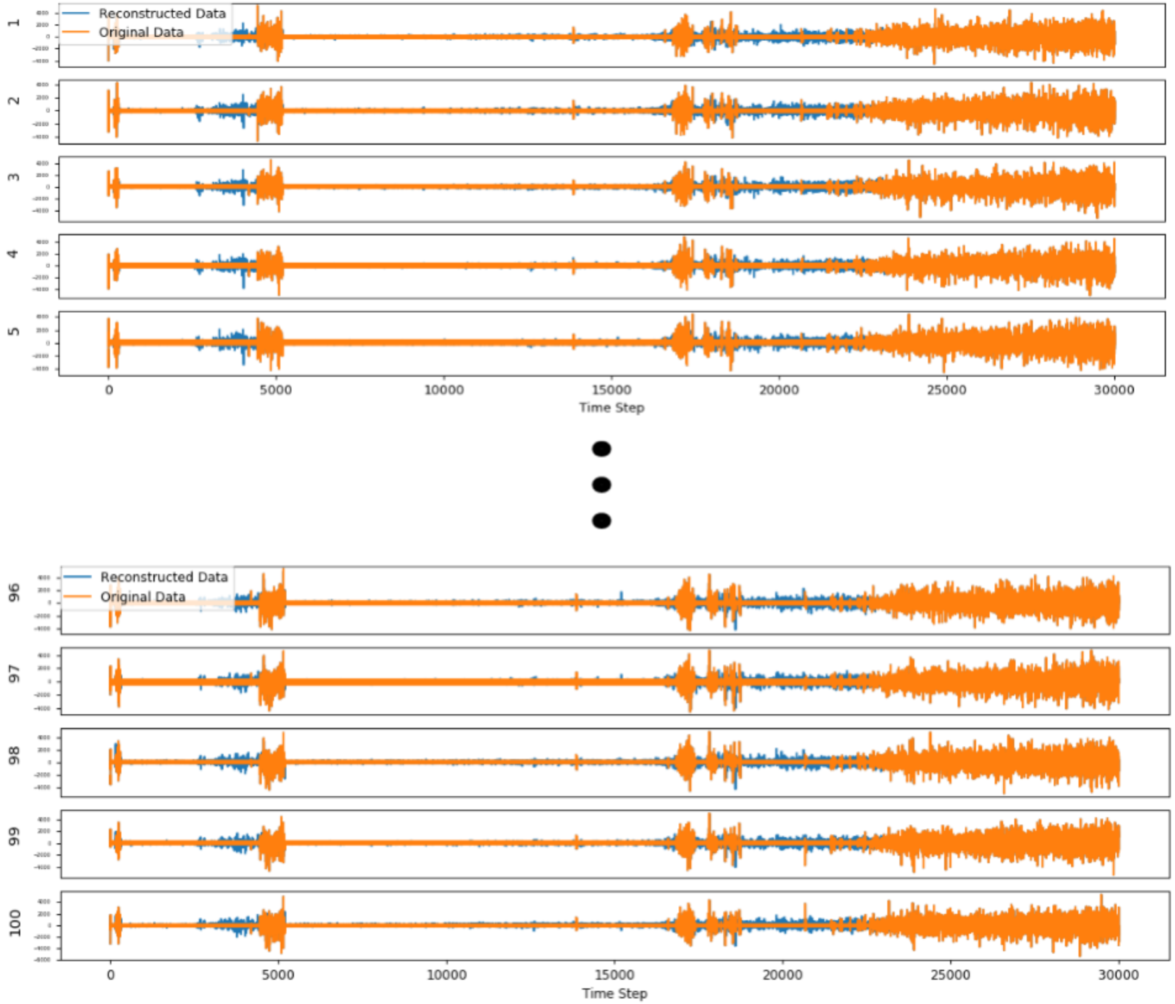
Results

- **Power Grid dataset on 3 dimensional input data**



Results

- **Distributed Acoustic Sensing dataset with 100 dimensions**



Results – image compression

- **ORIGINAL PHOTO (2560 x 1440 = 3,686,400 pixels) Each pixel has three dimensions of color (RGB) (image courtesy by Nina Fox)**

CR = 7.71

CR = 19.61

CR = 57.65

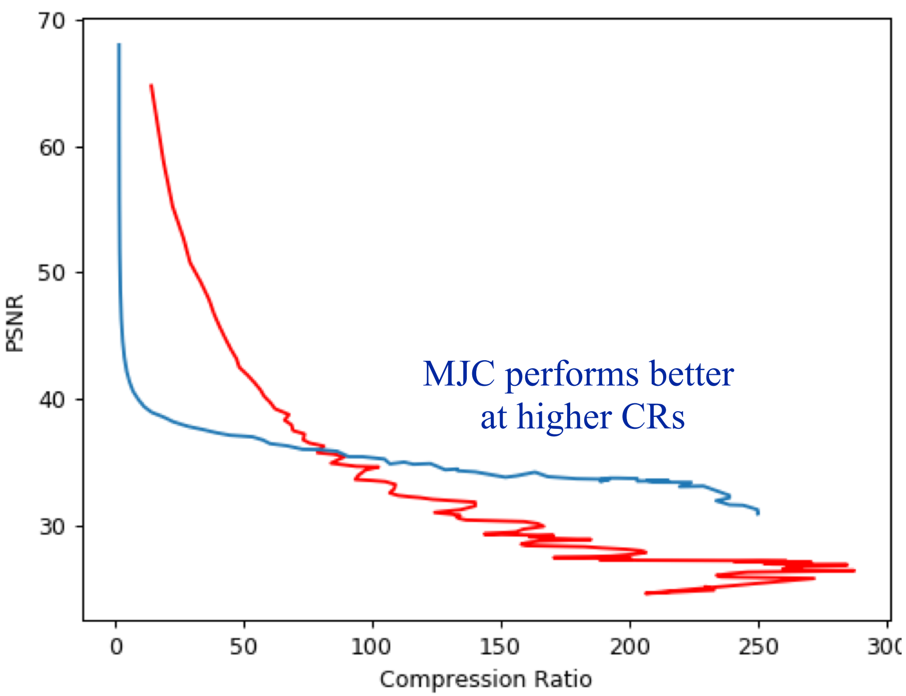
Results – video compression

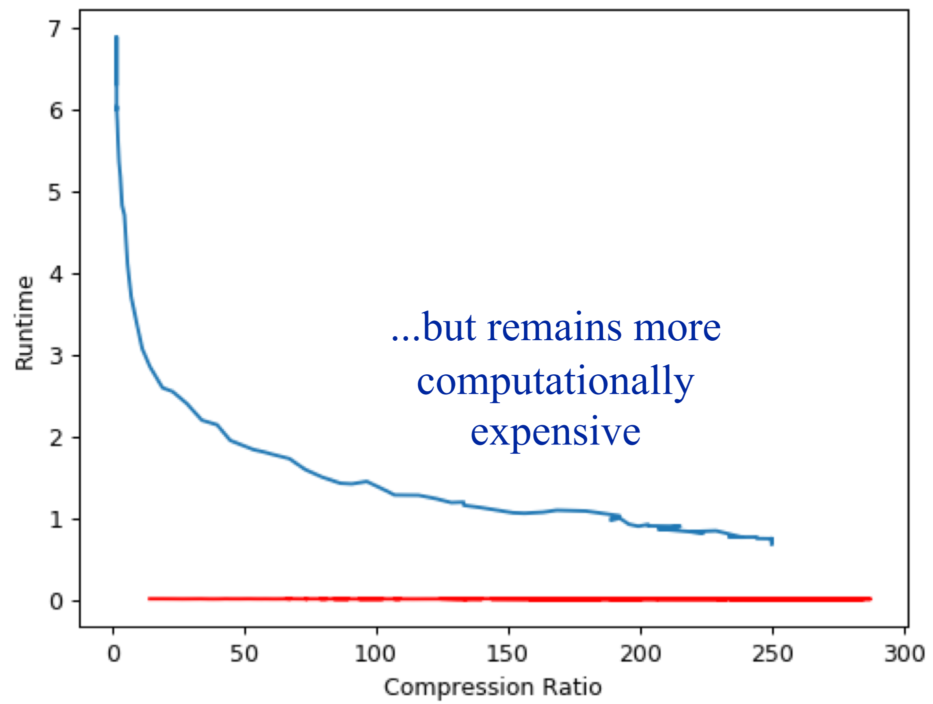
ORIGINAL VIDEO

300 frames x
300 height x
300 width x
3 colors

Comparison to SZ

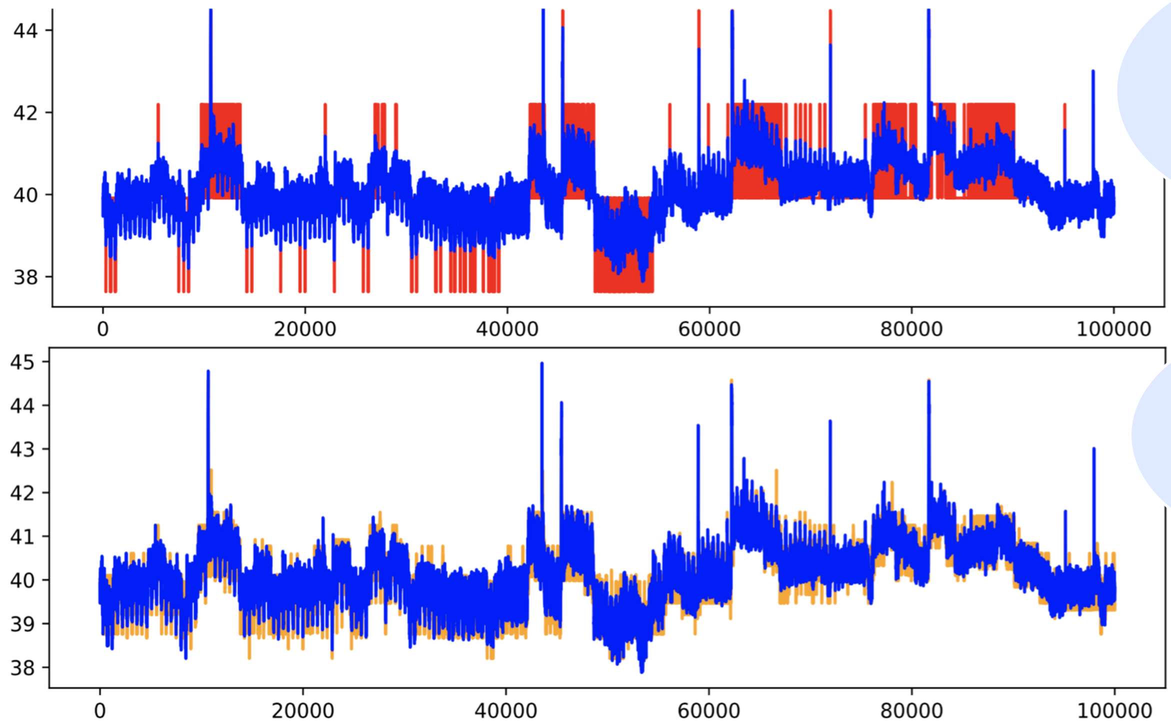
- **State of art compression technique developed by Cappello and colleagues from Argonne National Lab**
- **Work based on predictions made by nearby points**
  - **1D version was notable for incorporating multiple curve fitting**
  - **Multidimensional version uses a "layer" of points for prediction**
- **CR vs. PSNR (left) and Runtime (right) for SZ (red) and MJC (blue)**





SZ vs. MJC Reconstruction Results for Power Grid Data

- **SZ relies on nearby points to perform reconstruction, leading to inaccurate decompression in highly variable data**
- **IDEALEM: Better PSNR at the cost of compute time**



CR=100

SZ
0.009 (s)
PSNR 33.5

MJC
0.686 (s)
PSNR 35.4

Summary

- **IDEALEM: Similarity-based Compression with Multidimensional Pattern Matching**
  - An promising alternative to leading lossy compression algorithms
  - In addition to K-S test, MJC and DTW are used for multidimensional data
  - Applied to photos and videos, in additional to scientific multidimensional floating point data
- **Future work**
  - Understand MSE behavior
  - Study additional test statistics as a similarity measure
  - Study run-time optimization

IDEALEM

HTTPS://SDM.LBL.GOV/IDEALEM/

PAPERS:

HTTPS://SDM.LBL.GOV/MANA/

Contact E-Mail

K. John Wu <KWu@lbl.gov>