Time Series Epenthesis:

Clustering Time Series Streams Requires Ignoring Some Data

Thanawin Rakthanmanon

Eamonn Keogh

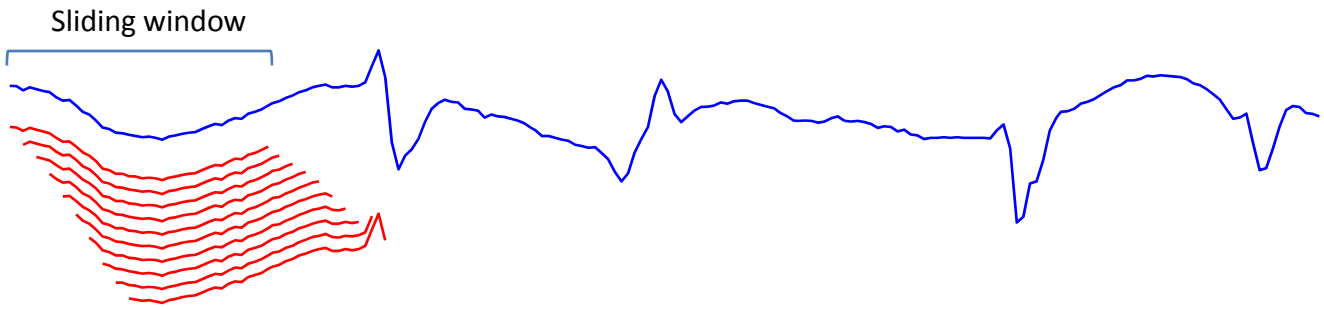
Stefano Lonardi

Scott Evans

UNIVERSITY OF CALIFORNIA
UCRIVERSIDE

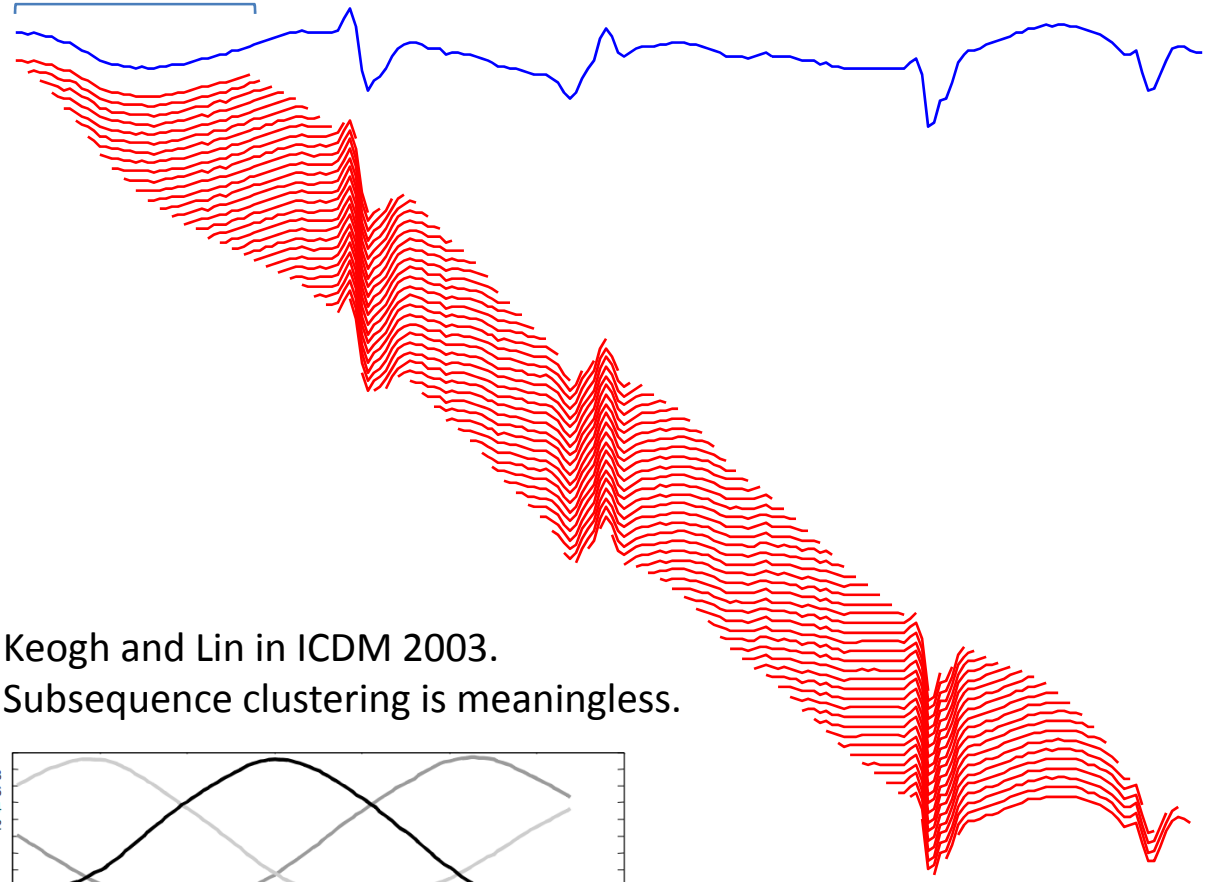
Subsequence Clustering Problem

- Given a time series, individual subsequences are extracted with a sliding window.
- Main task is to cluster those subsequences.

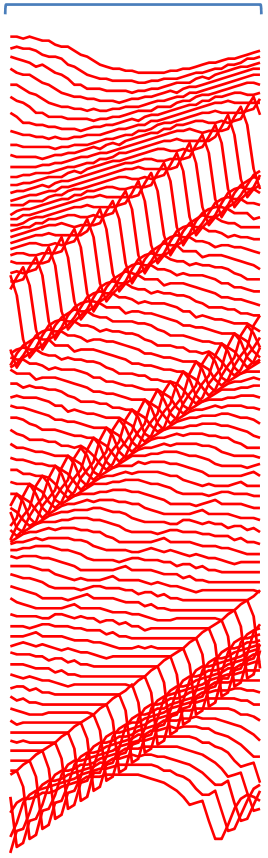


Subsequence Clustering Problem

Sliding window



All subsequences



Keogh and Lin in ICDM 2003.
Subsequence clustering is meaningless.

Centers of 3 clusters

Average subsequence

All data also contains ..

Transitions (the connections between words)

- Some transitions has good meaning and worth to be discovered
  - The connection inside a group of words
- Some transitions has *no meaning/structure*
  - **ASL:** hand movement between two words
  - **Speech:** (un)expected sound like um.., ah.., er..
  - **Motion Capture:** unexpected movement
  - **Hand Writing:** size of space between words

How to Deal with them?

Possible approaches are

- Learn it!
  - Separate noise/unexpected data from the dataset.
- Use a very clean dataset
  - dataset contains only atomic words.
- Simple approach *(our choice)*
  - Just ignore some data.
  - Hope that we will ignore unimportant data.

Concepts in Our Algorithm

Our clustering algorithm ..

- is a hierarchical clustering
- is parameter-lite
  - approx. length of subsequence (size of sliding window)
- ignores some data
  - the algorithm considers only non-overlapped data
- uses MDL-based distance, *bitsave*
- terminates if ..
  - no choice can save any bit (*bitsave* ≤ 0)
  - all data has been used

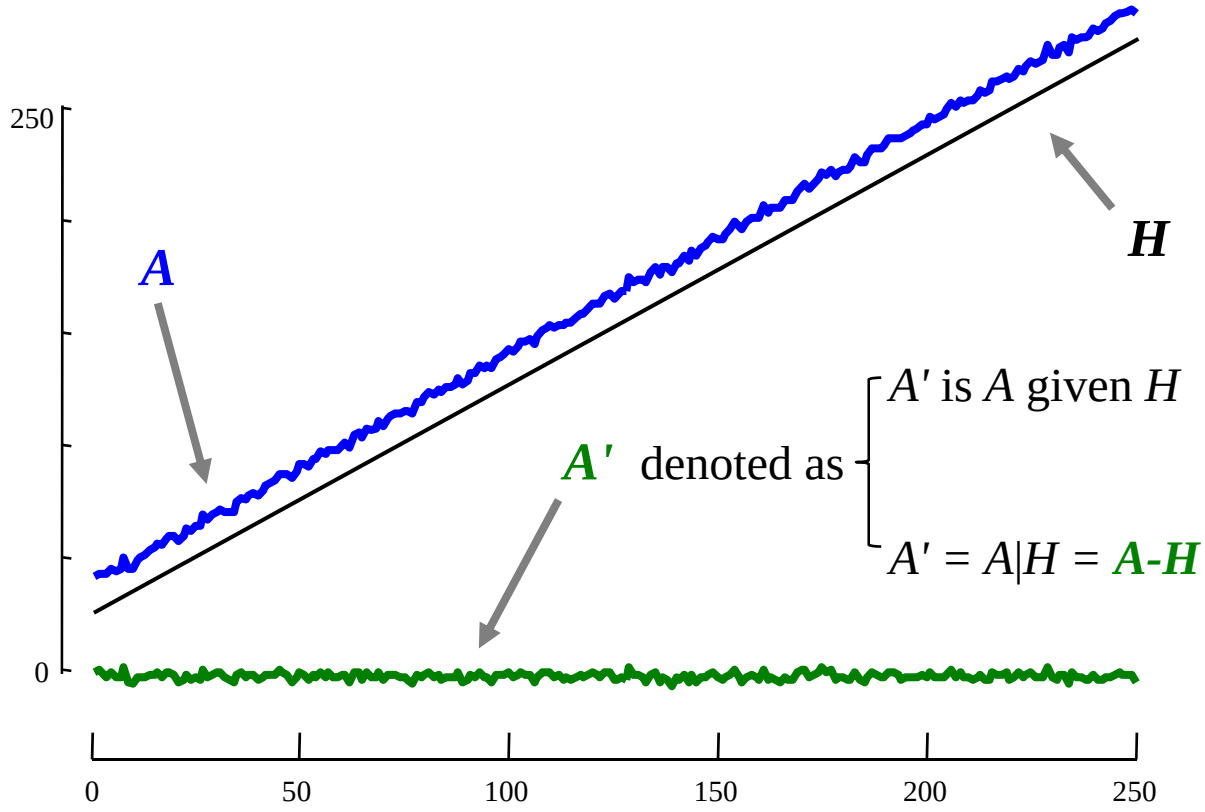
Minimum Description Length (MDL)

- The *shortest* code to output the data by Jorma Rissanen in 1978
- Intractable complexity (Kolmogorov complexity)

Basic concepts of MDL which we use:

- The *better* choice uses the *smaller* number of bits to represent the data
  - Compare between different operators
  - Compare between different lengths

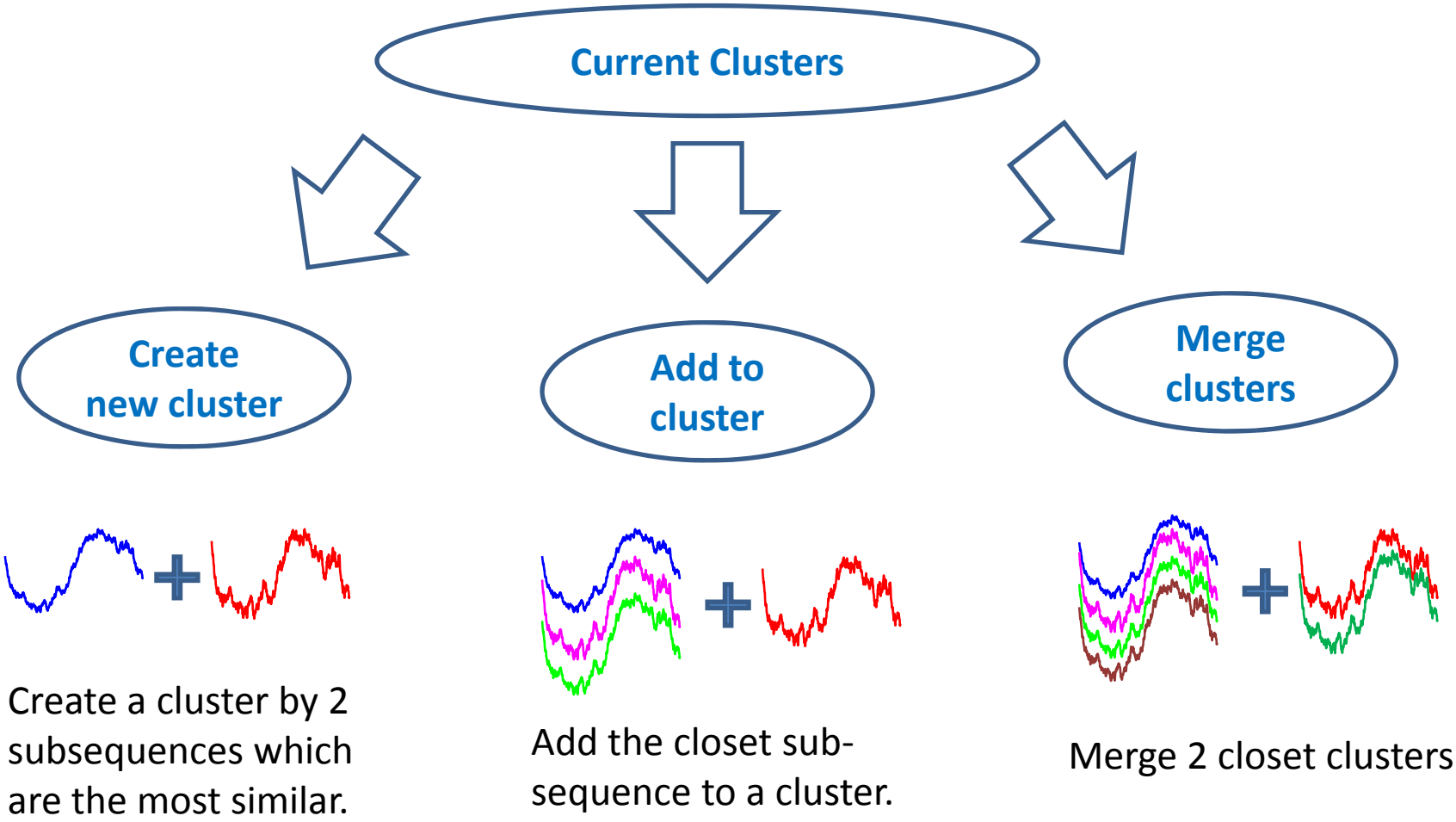
How to use Description Length?



If $DL(A) > DL(A') + DL(H)$, we will store A as A' and H

$DL(A)$ is the number of bits to store A

Clustering Algorithm



What is the best choice?

$$bitsave = DL(Before) - DL(After)$$

1) Create

$bitsave = DL(A) + DL(B) - DLC(\mathbf{C'})$

- a new cluster **C'** from subsequences *A* and *B*

2) Add

$bitsave = DL(A) + DLC(\mathbf{C}) - DLC(\mathbf{C'})$

- a subsequence *A* to an existing cluster **C**
- **C'** is the cluster **C** after including subsequence *A*.

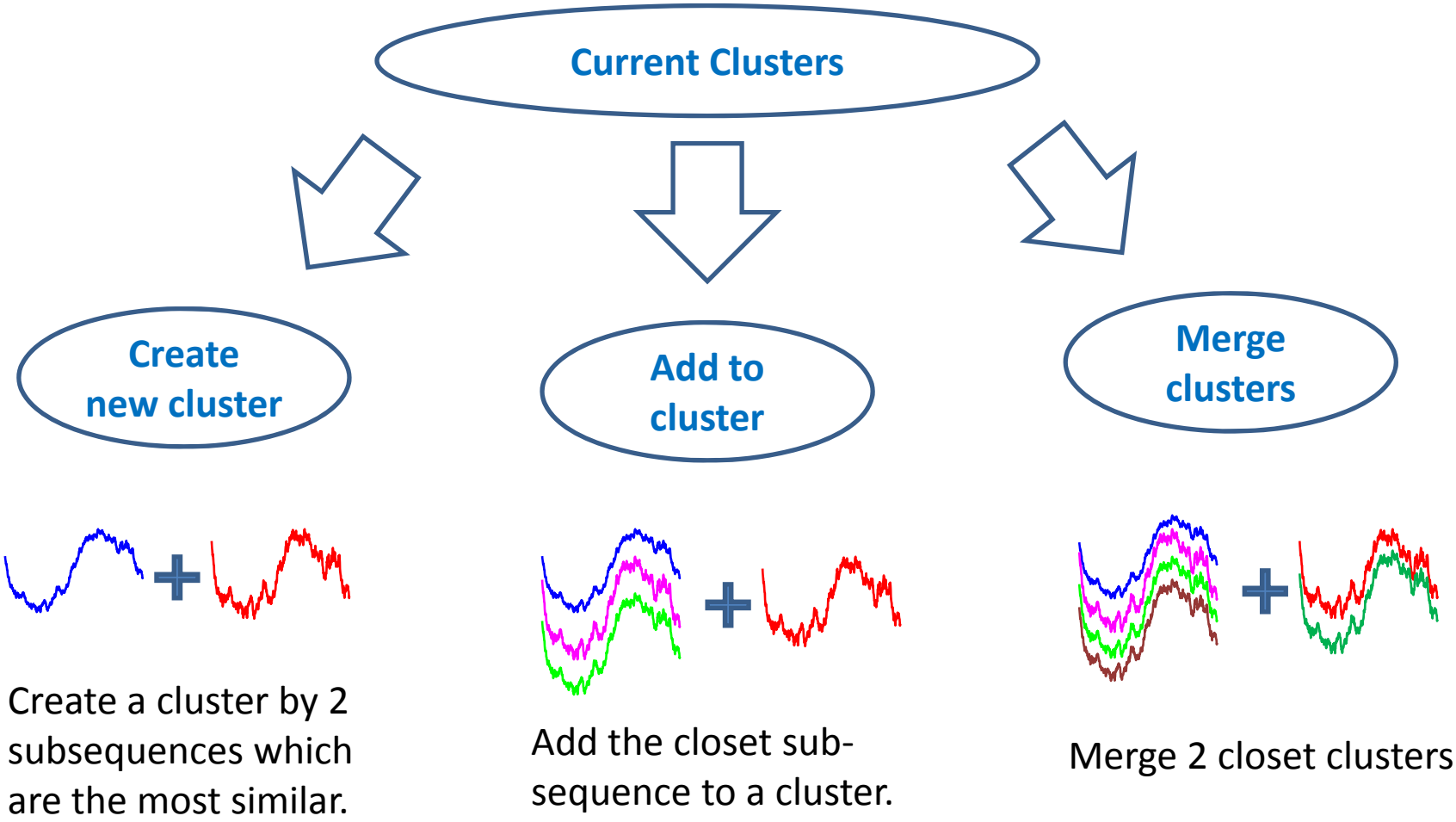
3) Merge

$bitsave = DLC(\mathbf{C}_1) + DLC(\mathbf{C}_2) - DLC(\mathbf{C'})$

- cluster $\mathbf{C}_1$ and $\mathbf{C}_2$ merge to a new cluster **C'**.

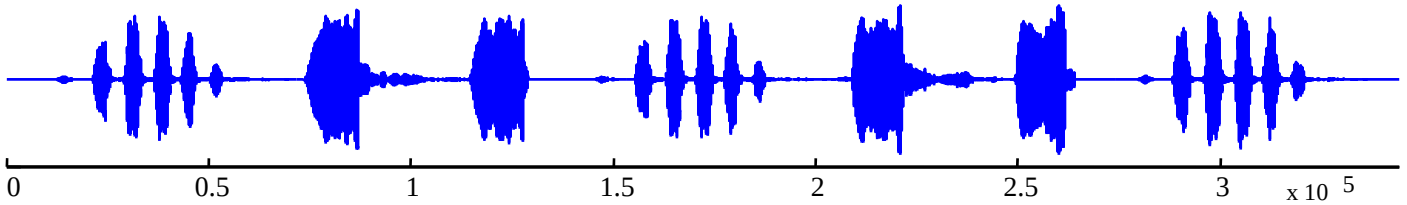
The bigger save, the better choice.

Clustering Algorithm

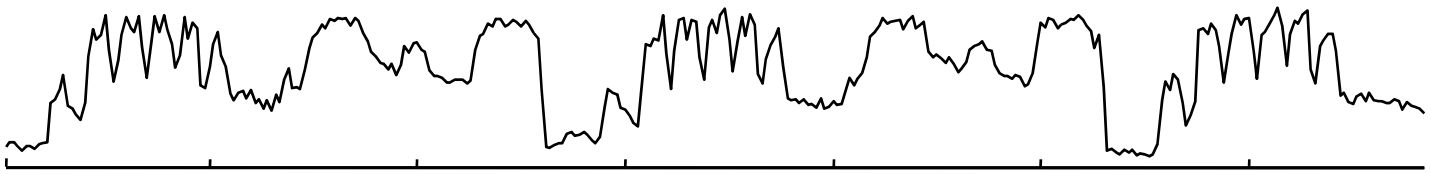


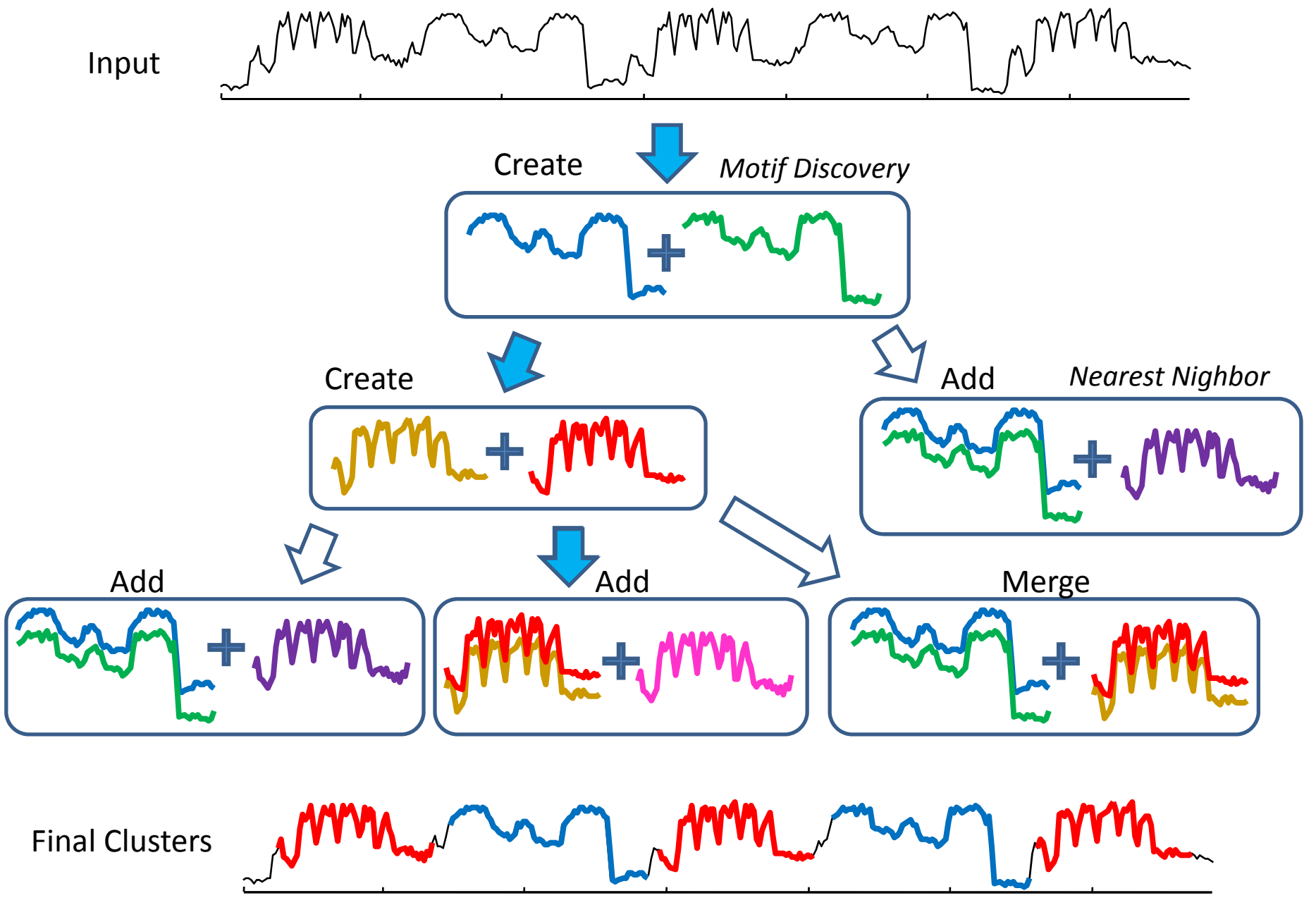
Bird Calls

Two interwoven calls from the *Elf Owl,* and *Pied-billed Grebe.*

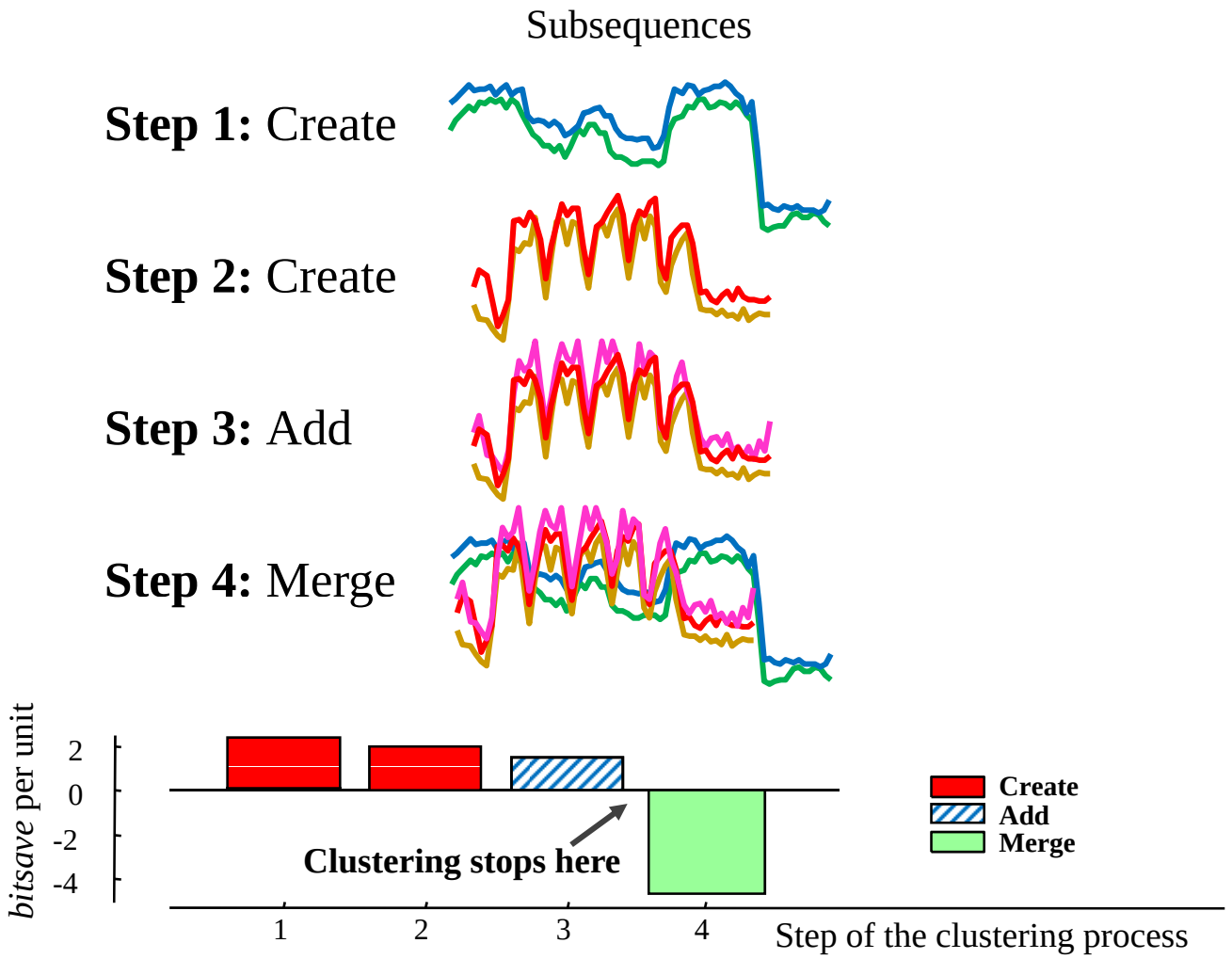


A time series extracted by using MFCC technique.



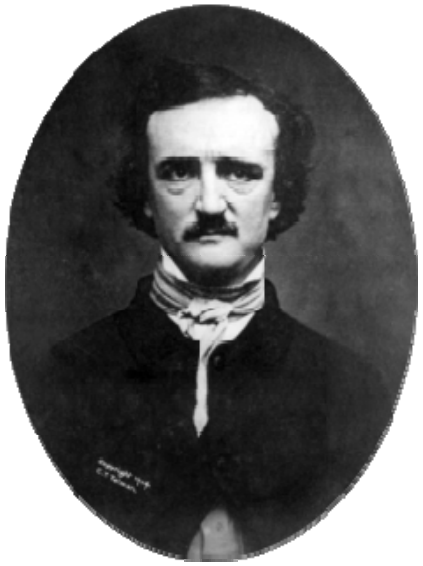

Input
Create
Motif Discovery
Create
Add
Nearest Nighbor
Add
Add
Merge
Final Clusters

Bird Calls: Clustering Result

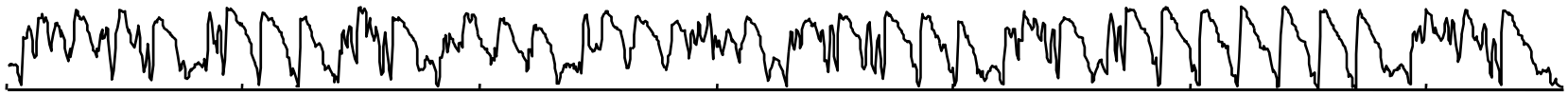


Poem *The Bells*

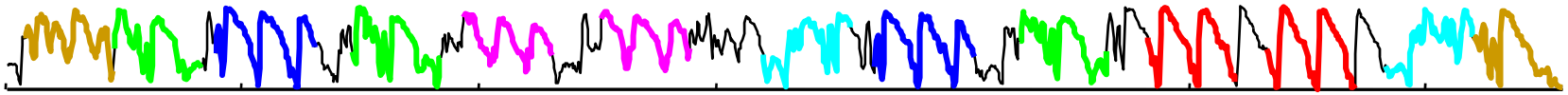
In a sort of Runic rhyme,
To the throbbing of the bells--
Of the bells, bells, bells,
To the sobbing of the bells;
Keeping time, time, time,
As he knells, knells, knells,
In a happy Runic rhyme,
To the rolling of the bells,--
Of the bells, bells, bells--
To the tolling of the bells,
Of the bells, bells, bells, bells,
Bells, bells, bells,--
To the moaning and the groan-
ing of the bells.



Edgar Allen Poe
1809-1849
(Wikipedia)



The Bells: Clustering Result



== Original Order ==

In a sort of Runic rhyme,
To the throbbing of the bells--
Of the bells, bells, bells,
To the sobbing of the bells;
Keeping time, time, time,
As he knells, knells, knells,
In a happy Runic rhyme,
To the rolling of the bells,--
Of the bells, bells, bells--
To the tolling of the bells,
Of the bells, bells, bells, bells,
Bells, bells, bells,--
To the moaning and the groaning of the bells.

== Group by Clusters ==

bells, bells, bells,
Bells, bells, bells,

Of the bells, bells, bells,
Of the bells, bells, bells—

To the throbbing of the bells--
To the sobbing of the bells;
To the tolling of the bells,

To the rolling of the bells,--
To the moaning and the groan-

time, time, time,
knells, knells, knells,

sort of Runic rhyme,
groaning of the bells.

Summary

- Clustering time series algorithm using MDL.
- Some data must be ignored or not appeared in any cluster.
- MDL is used to ..
  - select the best choice among different operators.
  - select the best choice among the different lengths.
- Final clusters can contain subsequences of different length.
- To speed up, Euclidean is used instead of MDL in core modules, e.g., motif discovery.

Thank you for your attention

QUESTION?

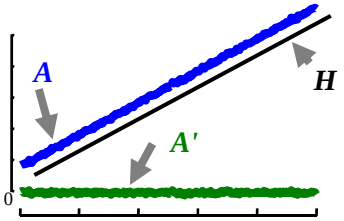
Supplementary

How to calculate *DL*?

A is a subsequence.

- *DL*(*A*) = entropy(*A*)
  - Similar result if use Shanon-Fano or Huffman coding.

H is a hypothesis, which can be any subsequence .



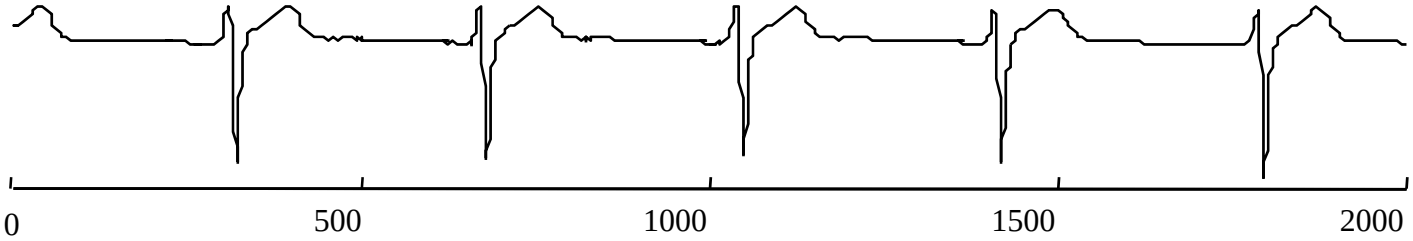
**DL*(*A*) = *DL*(*H*) + *DL*(*A* - *H*)

Cluster **C** contains subsequence *A and B*

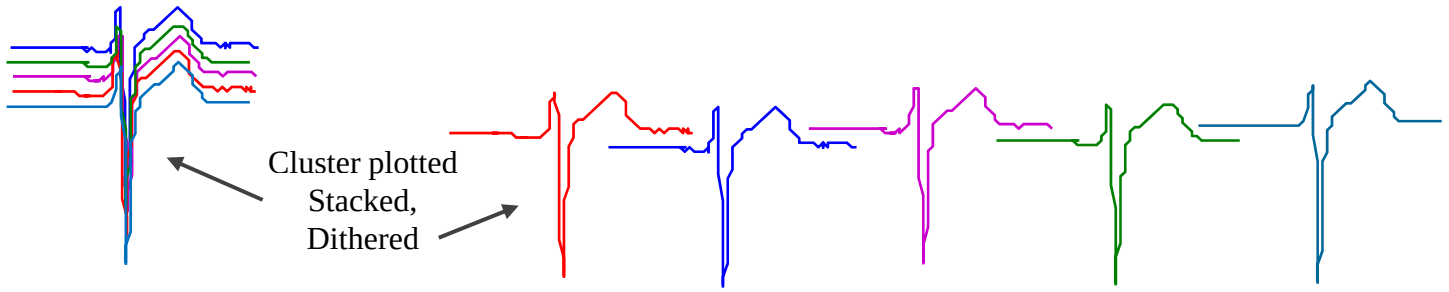
- *DLC*(**C**) = *DL*(*center*) + min(*DL*(*A-center*), *DL*(*B-center*))

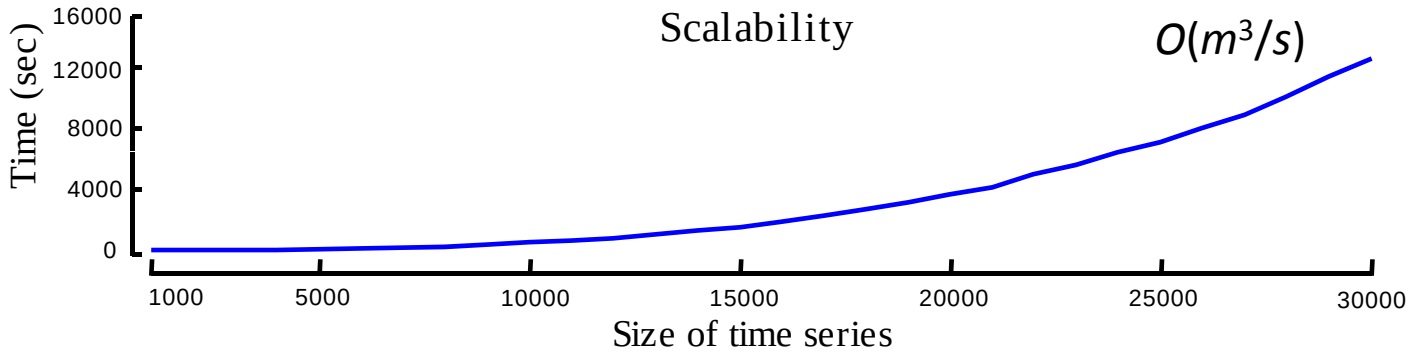
Running Time

Koshi-ECG time series

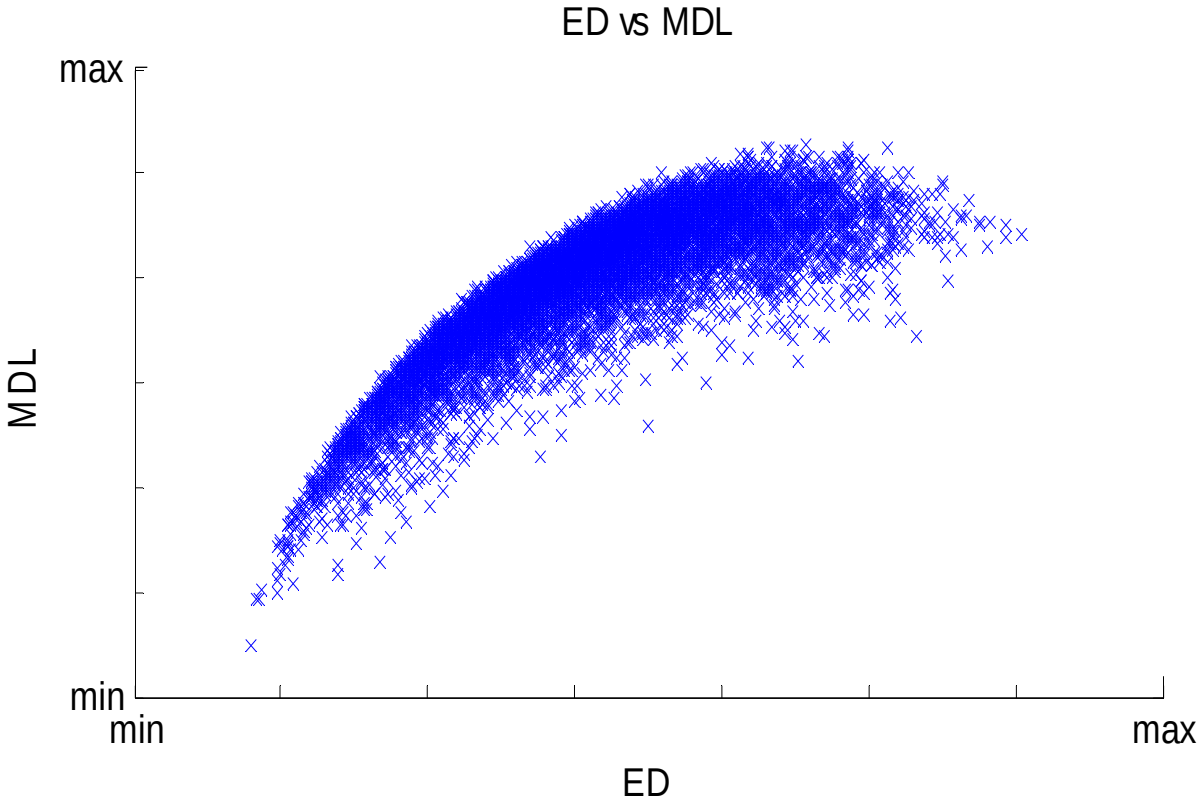


motif length
s = 350





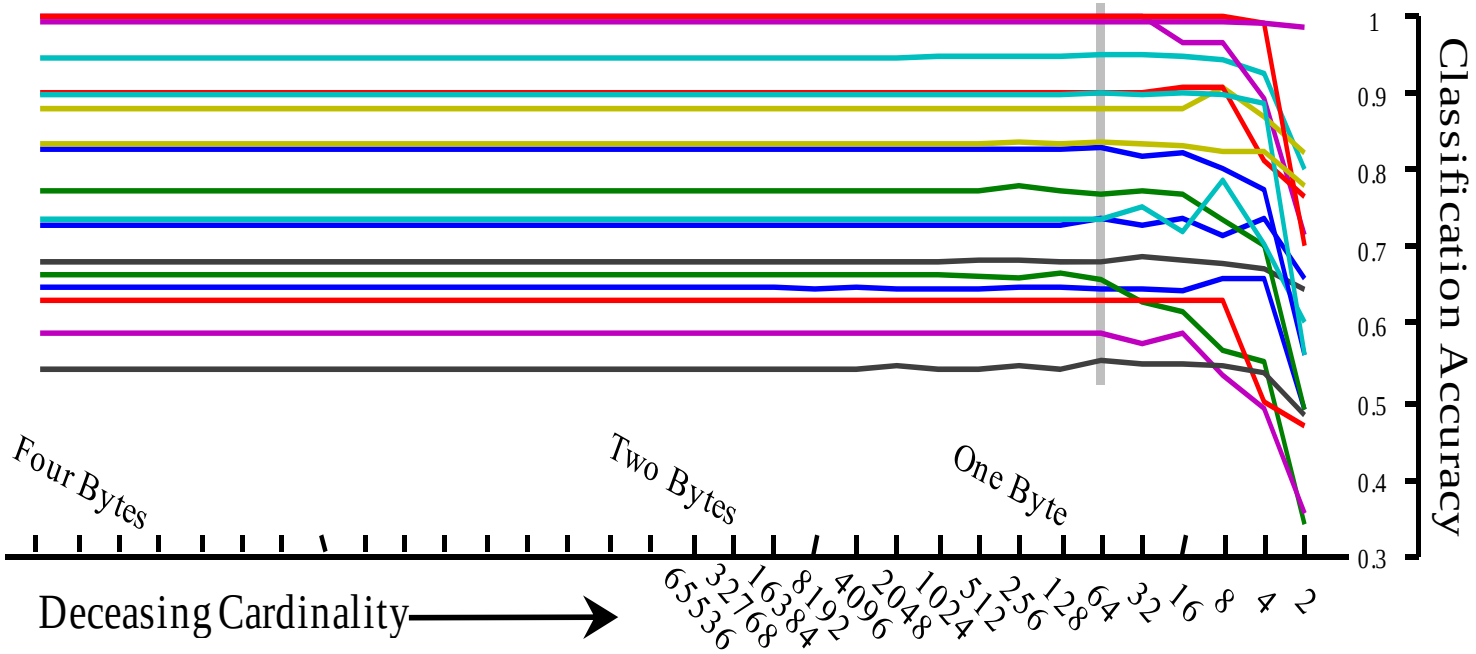
ED vs MDL in Random Walk



ED calculated in original continuous space

MDL calculated in discrete space (64 cardinality)

Discretization vs Accuracy



- Classification Accuracy of 18 data sets.
- The reduction from original continuous space to different discretization does not hurt much, at least in classification accuracy.