Tutorial 2

Finding Repeated Structure in Time Series: Algorithms and Applications

(we will start 5 min late to allow folks to find the room)

Abdullah Mueen
University of New Mexico, USA
Eamonn Keogh
University of California Riverside, USA

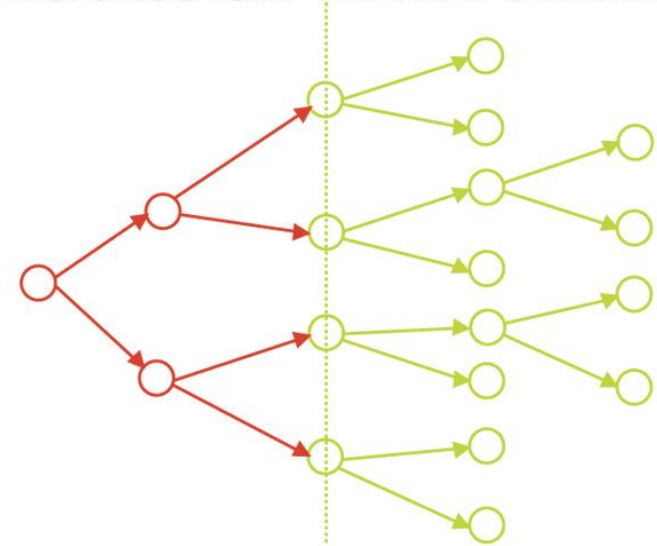
Slides available at
http://www.cs.unm.edu/~mueen/Tutorial/SDM2015Tutorial2.pdf

Funding by NSF
IIS-1161997 II

Vancouver

2015 SIAM International Conference on DATA MINING

April 30–May 2, 2015



Pinnacle Vancouver Harbourfront Hotel
Vancouver, British Columbia, Canada

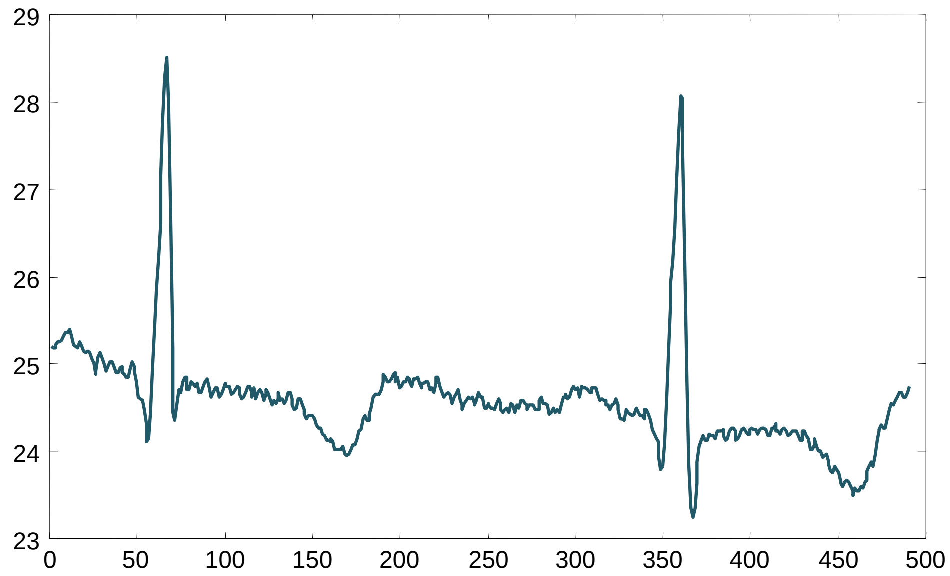
Tutorial Structure

- I will start with applications and talk about algorithms after that.
- There will be four Q&A segments. Please hold your question till the next segment.
- There is a feedback form available. Negative/positive, anonymous/known feedbacks are welcome.
- There will be a break at 5:00PM for 10 minutes.

25.1750
25.2250
25.2500
25.2500
25.2750
25.3250
25.3500
25.3500
25.4000
25.4000
25.3250
25.2250
25.2000
25.1750
..
..
24.6250
24.6750
24.6750
24.6250
24.6250
24.6250
24.6750
24.7500

What are Time Series?

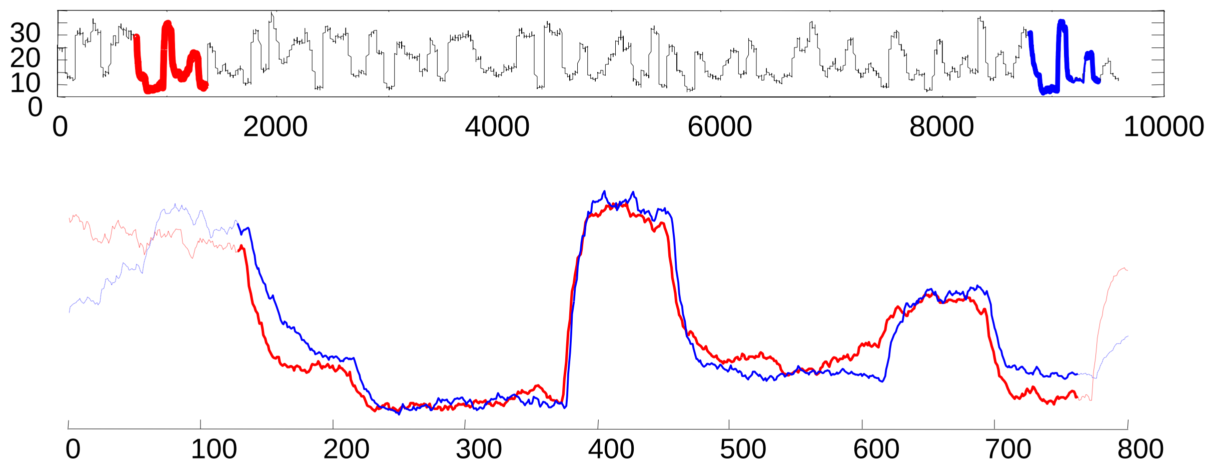
A time series is a collection of observations made sequentially in time.



Repeated Pattern (Motif)

时间序列中的重复模式

Find the subsequences having very high similarity to each other.



Finding Motifs in Time Series, Jessica Lin, Eamonn Keogh, Stefano Lonardi, Pranav Patel, KDD 2002

General Outline

- Applications (50 minutes)
  - As Subroutines in Data Mining
  - In Other Scientific Research
- Algorithms (100 minutes)
  - Uni-dimensional
  - Multi-dimensional
  - Open Problems

Applications Outline

- Applications
  - As Subroutines in Data Mining
    - Never Ending Learning
    - Time Series Clustering
    - Rule Discovery
    - Dictionary Building
  - In Other Scientific Research
    - Data center chiller management
    - Worm locomotion analysis
    - Physiological Prediction
    - Activity recognition
  - Motifs in Other Data-types
    - Audio
    - Shapes
    - Motion

Motifs allow us to learn, forever, without an explicit teacher…

If you have parallel texts, then over time you can learn a dictionary with high accuracy.

…And God said, “Let there be light”; and there was light. And God saw the light, that it was good. And God …

..Y dijo Dios: Sea la luz; y fue la luz. Y vio Dios que la luz *era* buena . Y llamó Dios a …

Motifs allow us to learn, forever, without an explicit teacher…

If you have parallel texts, then over time you can learn a dictionary with high accuracy.

…And God said, "Let there be light"; and there was light. And God saw the light, that it was good. And God …

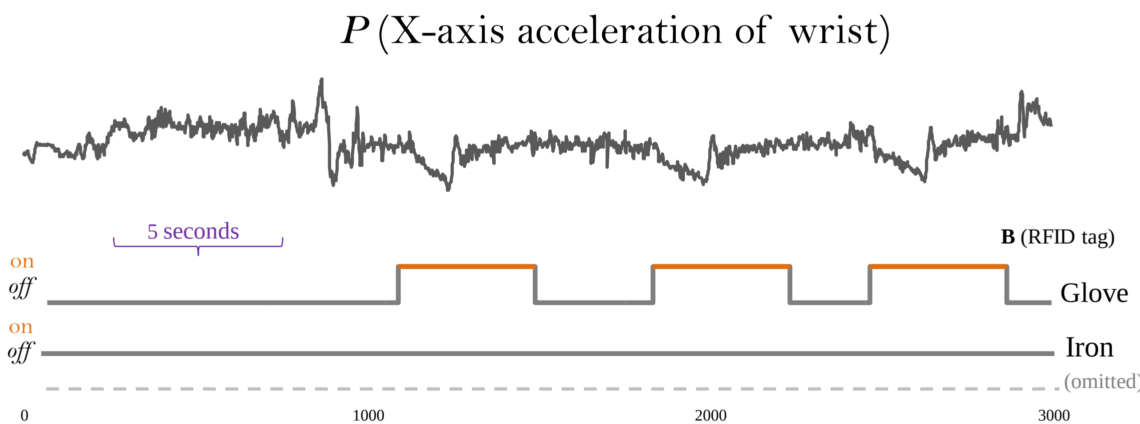
..Y dijo Dios: Sea la luz; y fue la luz. Y vio Dios que la luz *era* buena . Y llamó Dios a …

Note the mapping is non-linear, the learning algorithms in this domain are non-trivial.

Suppose however that the unknown "language" is not *discrete*, but *real-valued* time series? In this case, repeated pattern discovery can help*…

*Yuan Hao, Yanping Chen et al (2013). Towards Never-Ending Learning from Time Series Streams. SIGKDD 2013

Motifs allow us to learn, forever, without an explicit teacher…

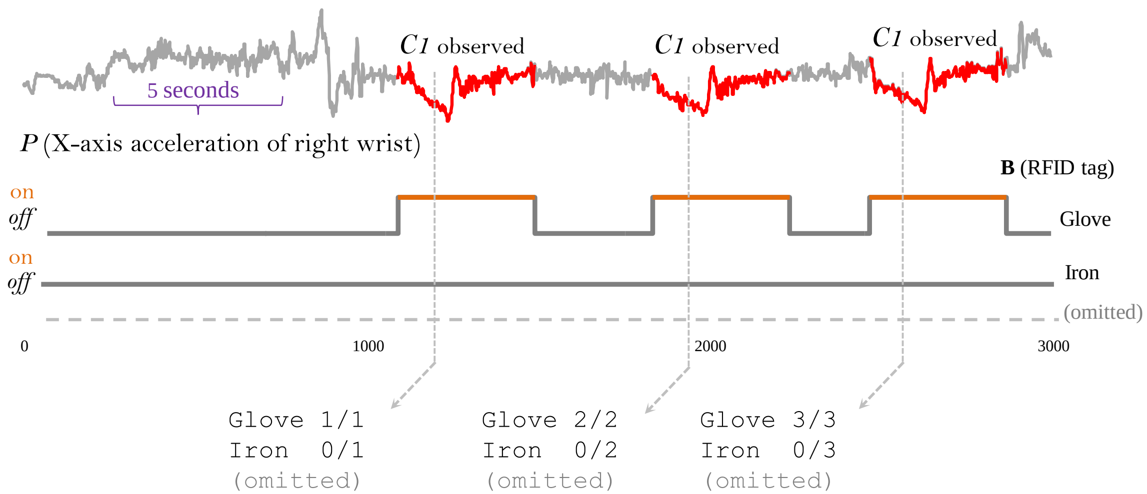


Real-valued “text”

Binary “text”

This dataset contains standard IADL housekeeping activities (vacuuming, ironing, dusting, brooming,, watering plants etc). We have a discrete (binary) “text” that notes if the hand is near a cleaning instrument, and a real-valued accelerometer “text”

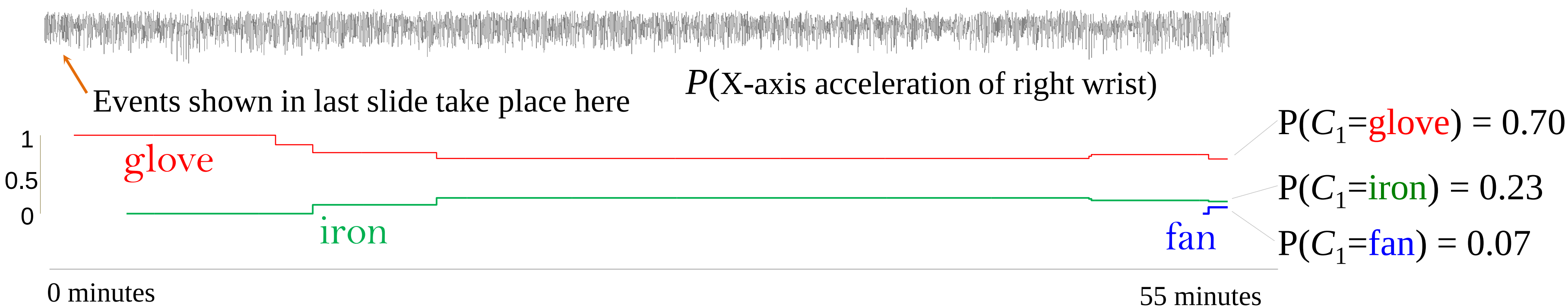
We can run motif discovery on the time series stream. If we find motifs, we can see if they correlate with the discrete streams…



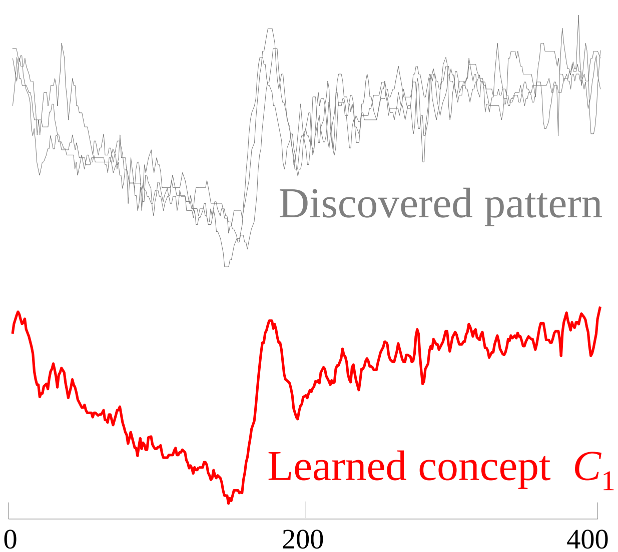
In this snippet, the motifs seem to correlate with the presence of a glove…

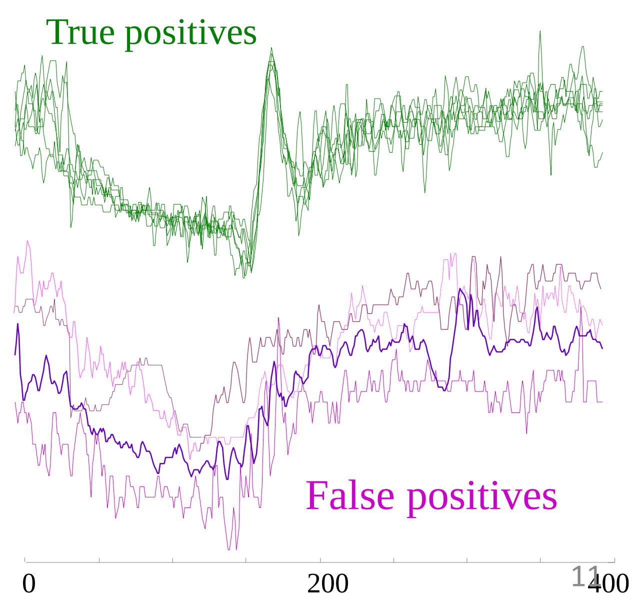
How well does this work?

Over a hour of activity, we learn to recognize a behavior in the time series that indicates the user is putting on a glove.

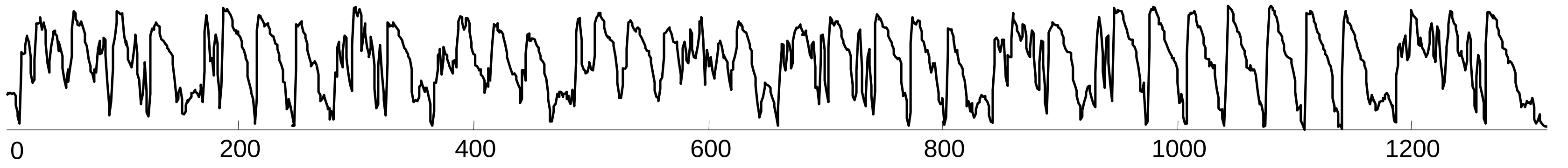


Note: There are false positives, but we considered only a single axis for simplicity.





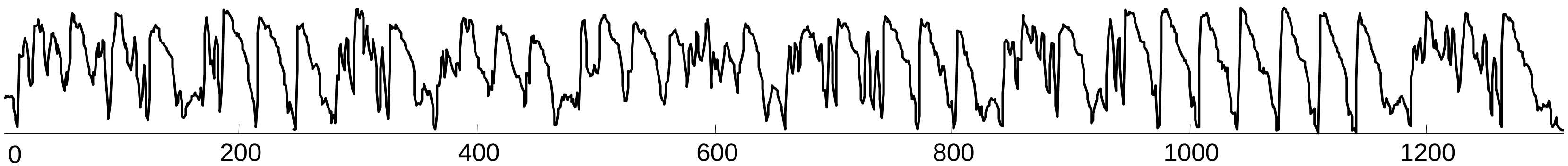
Motifs allow us to cluster subsequences of a time series…



And how would we evaluate our answer?

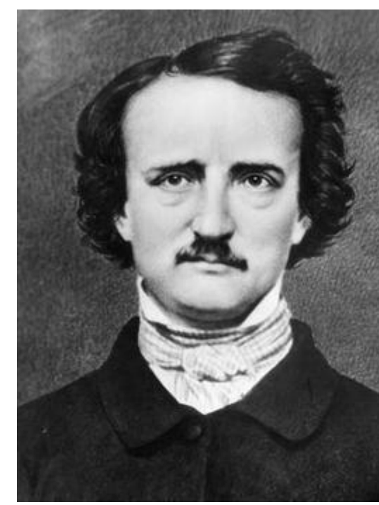
Thanawin Rakthanmanon, Eamonn Keogh, Stefano Lonardi, and Scott Evans (2011). Time Series Epenthesis: Clustering Time Series Streams Requires Ignoring Some Data. ICDM 2011

Motifs allow us to cluster subsequences of a time series…



And how would we evaluate our answer?

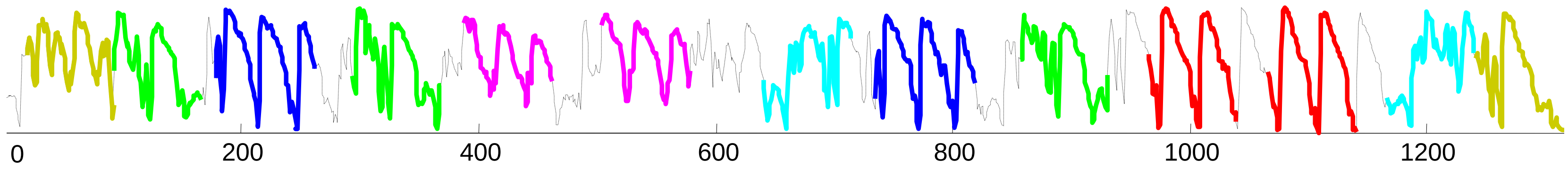
== Poem ==
In a sort of Runic rhyme,
To the throbbing of the bells--
Of the bells, bells, bells,
To the sobbing of the bells;
Keeping time, time, time,
As he knells, knells, knells,
In a happy Runic rhyme,
To the rolling of the bells,--
Of the bells, bells, bells--
To the tolling of the bells,
Of the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



Edgar Allan Poe

Data: last 30 seconds from 4-min poem, The Bells by Edgar Allan Poe

Motifs allow us to cluster subsequences of a time series…



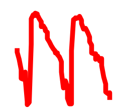
And how would we evaluate our answer?

== Poem ==

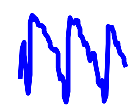
In a **sort of Runic rhyme,**
To **the throbbing of the bells**--
Of the bells, bells, bells,
To **the sobbing of the bells**;
Keeping **time, time, time**,
As he **knells, knells, knells**,
In a happy Runic rhyme,
To the rolling of the bells,--
Of the bells, bells, bells--
To **the tolling of the bells**,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

== Text in each clusters ==

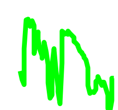
bells, bells, bells,
Bells, bells, bells,



Of the bells, bells, bells,
Of the bells, bells, bells--

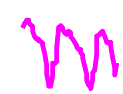


To **the throbbing of the bells**--
To **the sobbing of the bells**;
To **the tolling of the bells**,

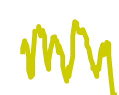


To the rolling of the bells,--
To the moaning and the groan-

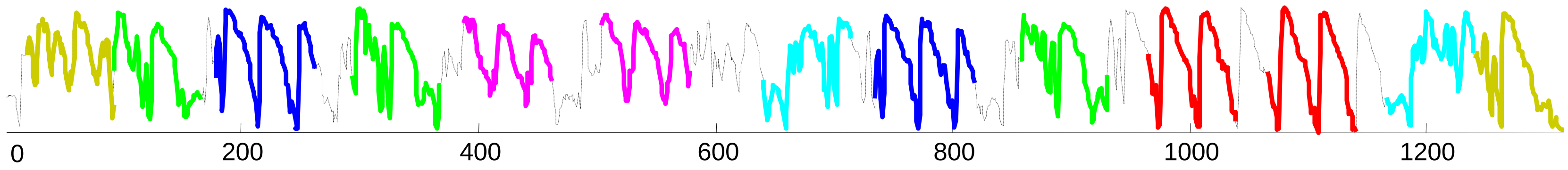
time, time, time,
knells, knells, knells,



sort of Runic rhyme,
groaning of the bells.



Motifs allow us to cluster subsequences of a time series…



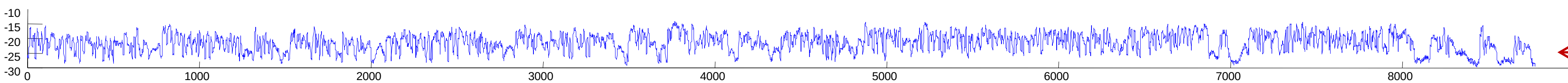
Key observations that make this possible:

- Time Series Motifs!
- We are willing to allow some data to be unexplained by the clustering
- We score the possible clustering's with MDL, this is parameter-free!
- Allowing the clusters to be of different lengths/sizes

Thanawin Rakthanmanon, Eamonn Keogh, Stefano Lonardi, and Scott Evans (2011). Time Series Epenthesis: Clustering Time Series Streams Requires Ignoring Some Data. ICDM 2011

Motifs are useful, but can we *predict* the future?

What happens next?



(unpublished work, email Dr. Keogh for preprint)

Prediction vs. Forecasting (informal definitions)

Forecasting is “always on”, it constantly predicts a value say, two minutes out (we are not doing this)

Prediction only make a prediction occasionally, when it is sure what will happen next

Why Predict the (short-term) Future?

If a robot can predict that is it about to fall, it may be able to..

- Prevent the fall
- Mitigate the damage of the fall

More importantly, if the robot can predict a *human's* actions

- The robot could catch the human!
- This would allow more natural human/robot interaction.
- Real time is not fast enough for interaction! We need to be a half second *before* real time.
-

- Other examples:
    - Predict a car crash, tighten seatbelts, apply brakes
    - Predict the next spoken word after '*data*' is '*mining*', then begin prefetching WebPages..
    - etc

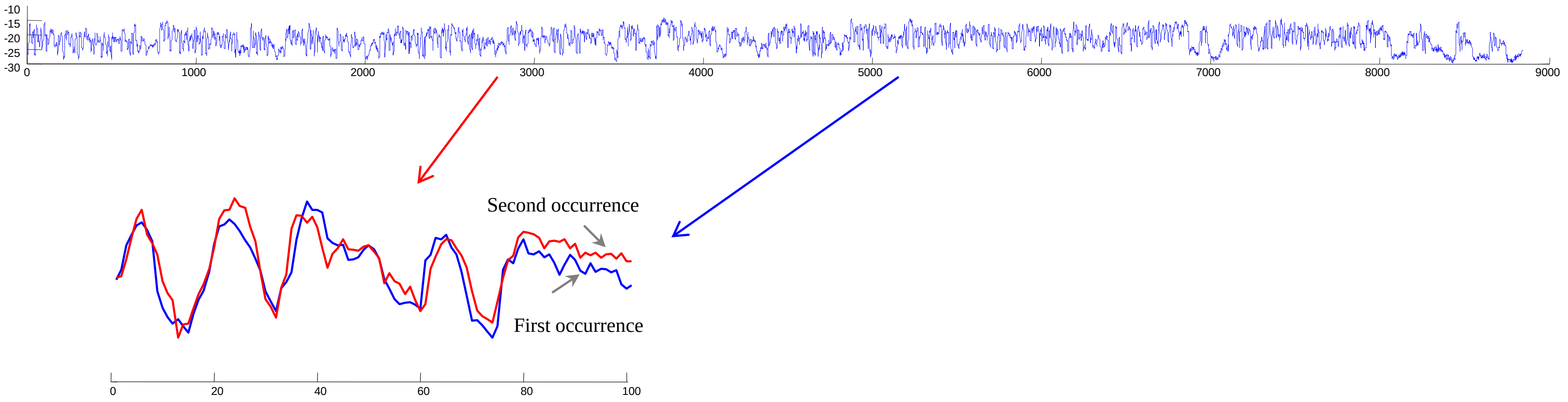
Previous attempts at this have largely failed…

However, we *can* do this, and time series motifs are the key tool

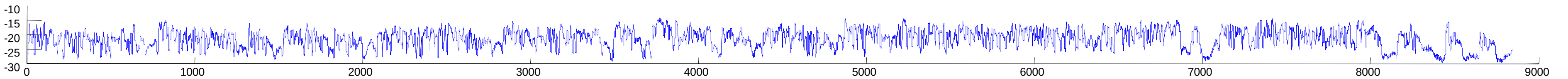
The rule discovery technique will use:

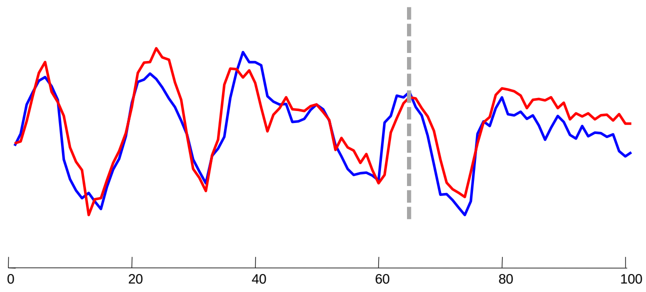
- Time Series Motifs
- MDL (minimum description length)
- Admissible speed-up techniques (not discussed here)

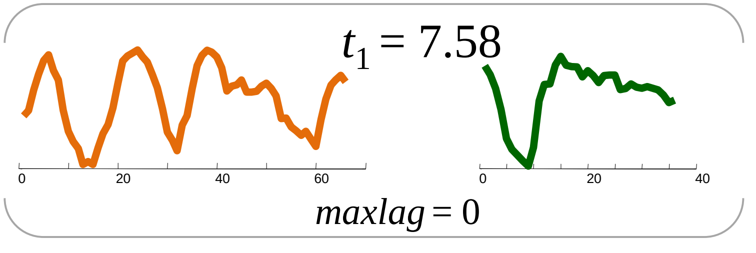
Let us start by finding motifs



We can convert the motifs to a rule





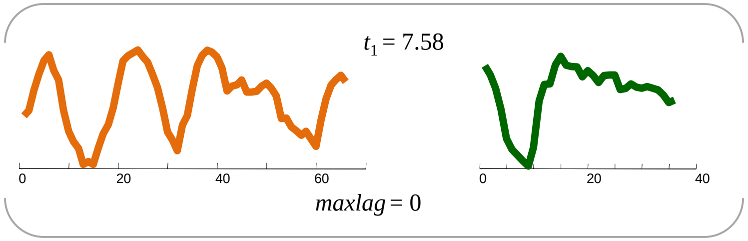


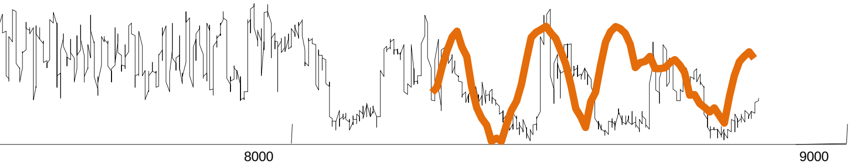
We can use the motif to make a rule…

IF we see **thisshape**, (antecedent)
THEN we see **thatshape**, (consequent)
within *maxlag* time

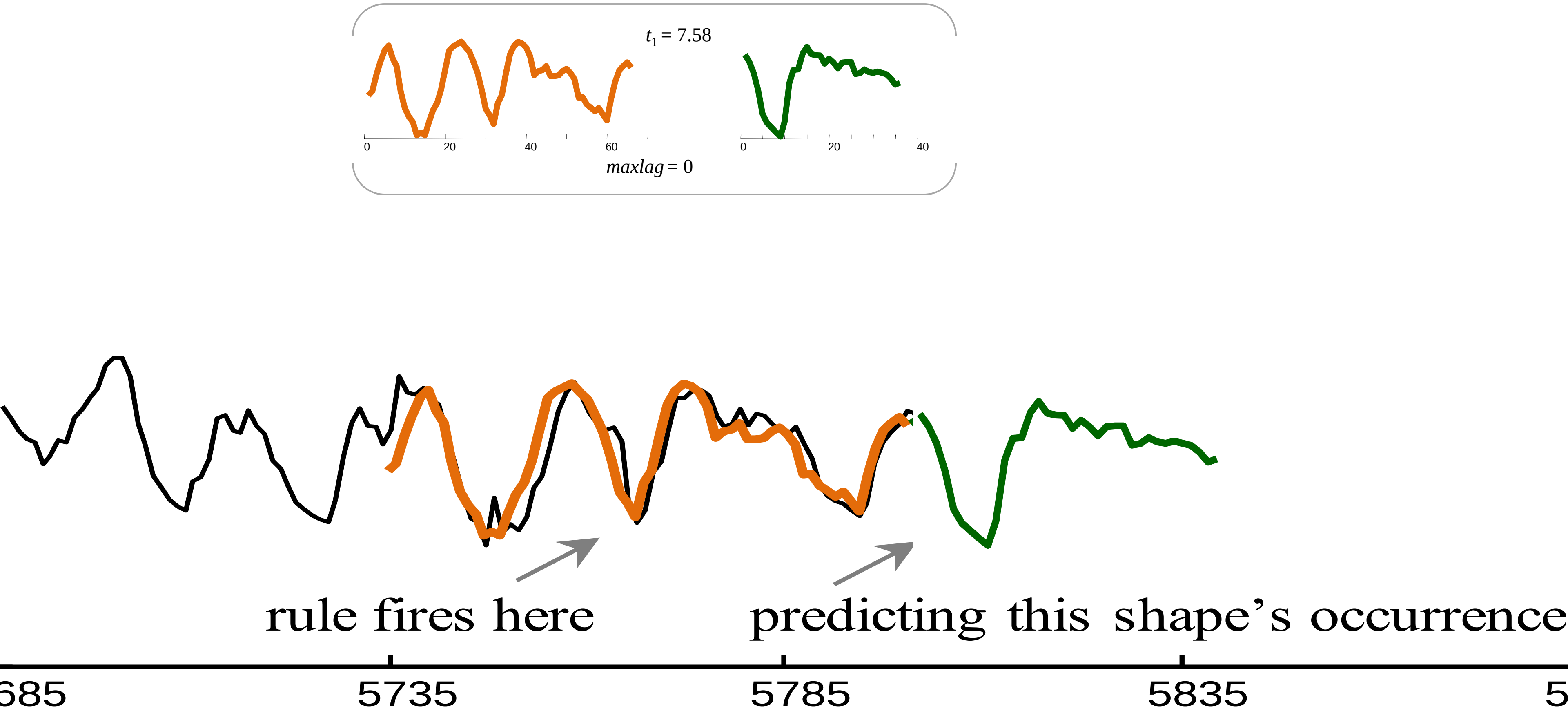
The Euclidean distance between **thisshape** and the observed window must be within a threshold t_1 = *7.58*

We can monitor streaming data with our rule..

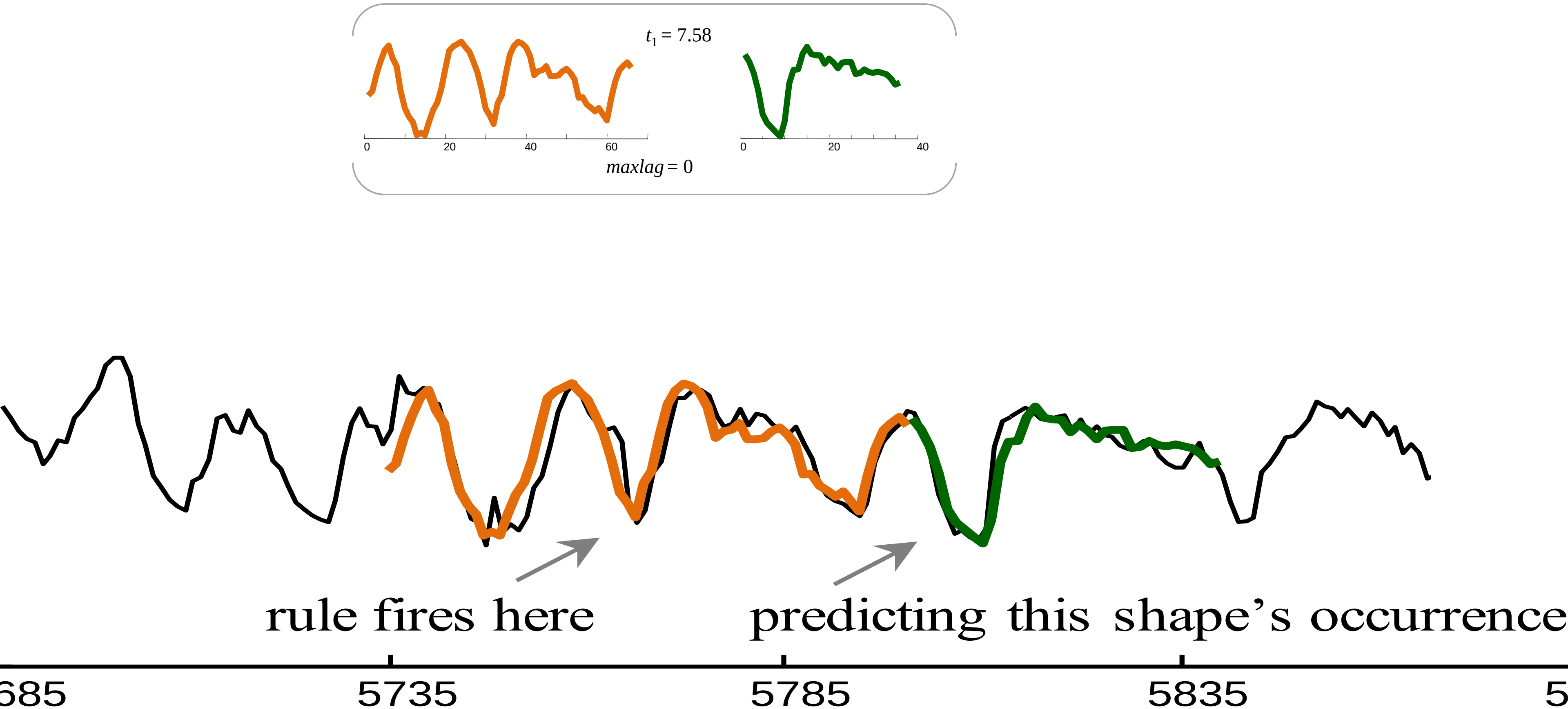




The rule gets invoked…

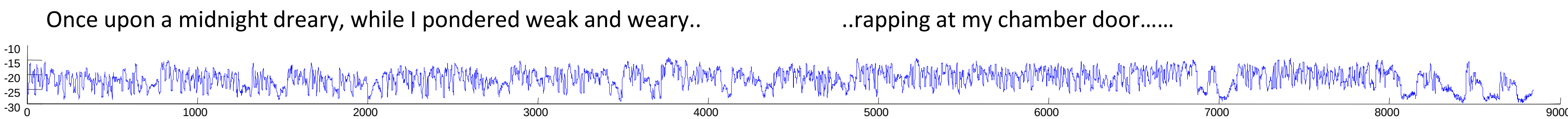


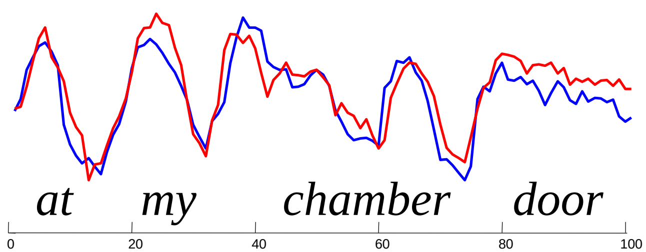
It seems to work!

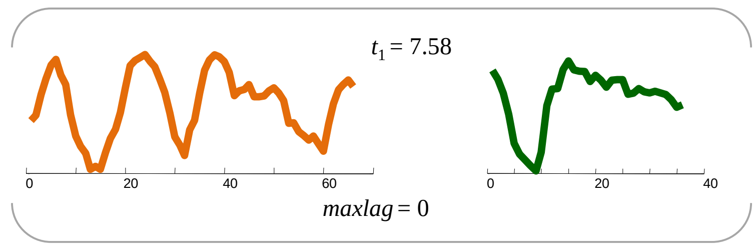


What is the ground truth?

The first verse of *The Raven* by Poe in MFCC space





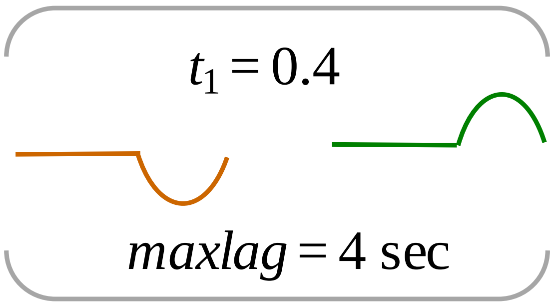


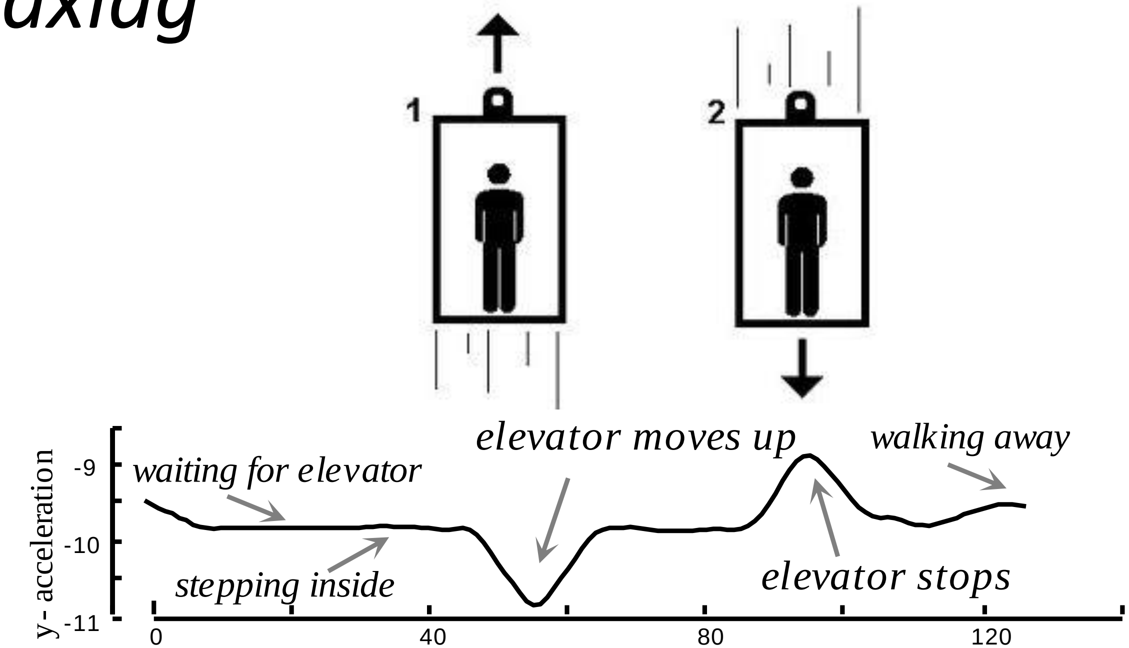
The phrase "*at my chamber door*" does appear 6 more times, and we do fire our rule correctly each time, and have no false positives.

What are we invariant to?

- Who is speaking? Somewhat, we can handle other males, but females are tricky.
- Rate of speech? To a large extent, yes.
- Foreign accents? Sore throat? etc

Why we need the *Maxlag* parameter



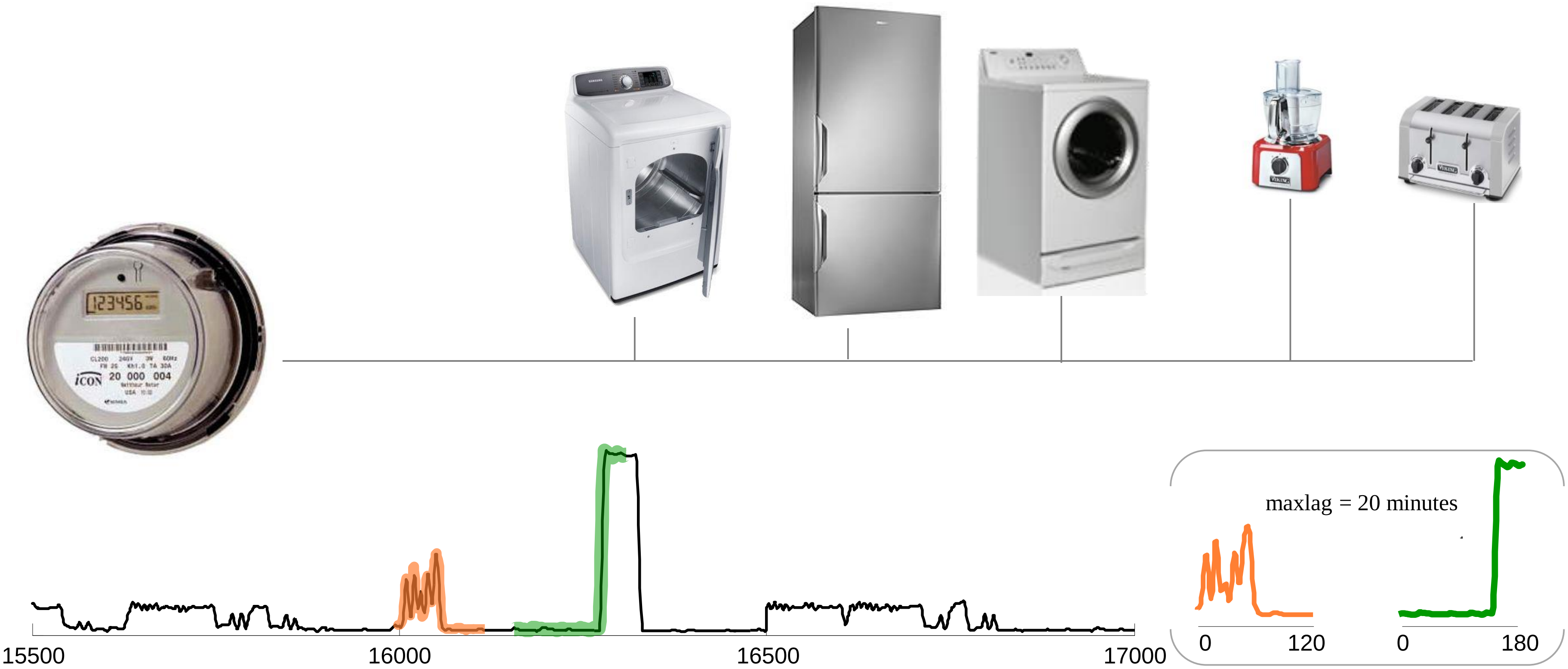


Here the *maxlag* depends on the number of floors we have in our building.

We can hand-edit this rule to generalize for short buildings to tall buildings

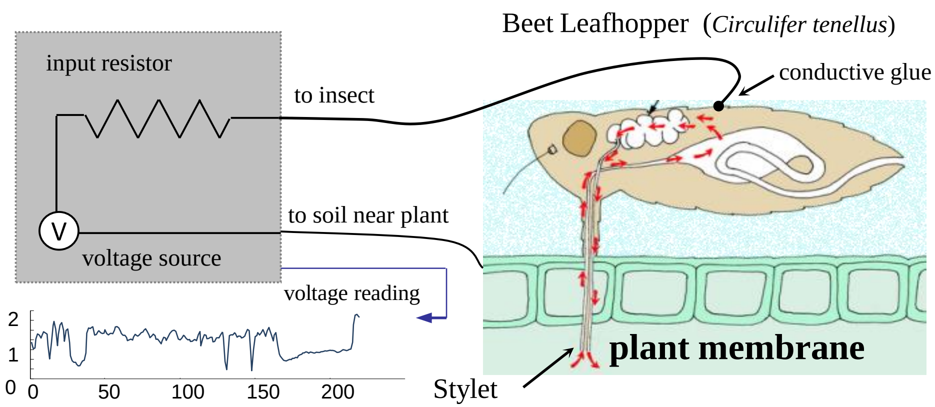
Can physicians edit medical rules to generalize from male to female…

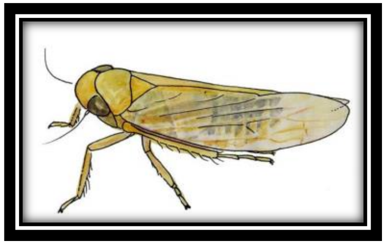
This works, *really*!



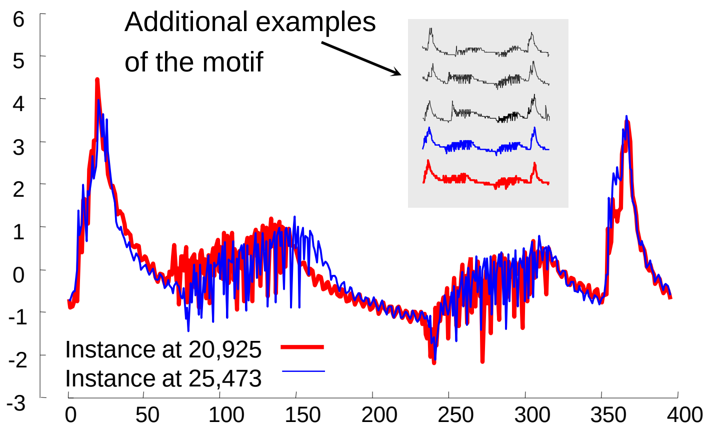
IF *we see a* ***Clothes Washer used***
THEN *we will see* ***Clothes Dryer used*** *within 20 minutes*

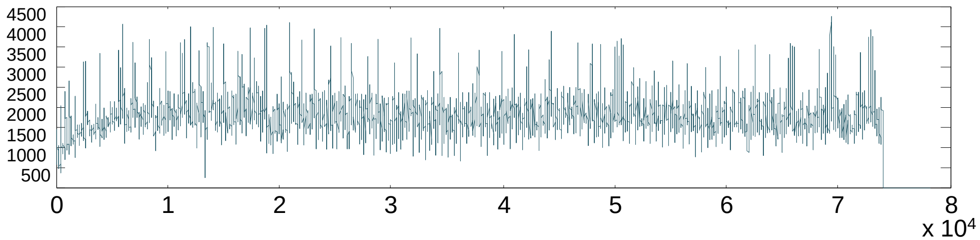
Insect Behavior Analysis





The **electrical penetration graph** or **EPG** is a system used by biologists to study the interaction of insects with plants.



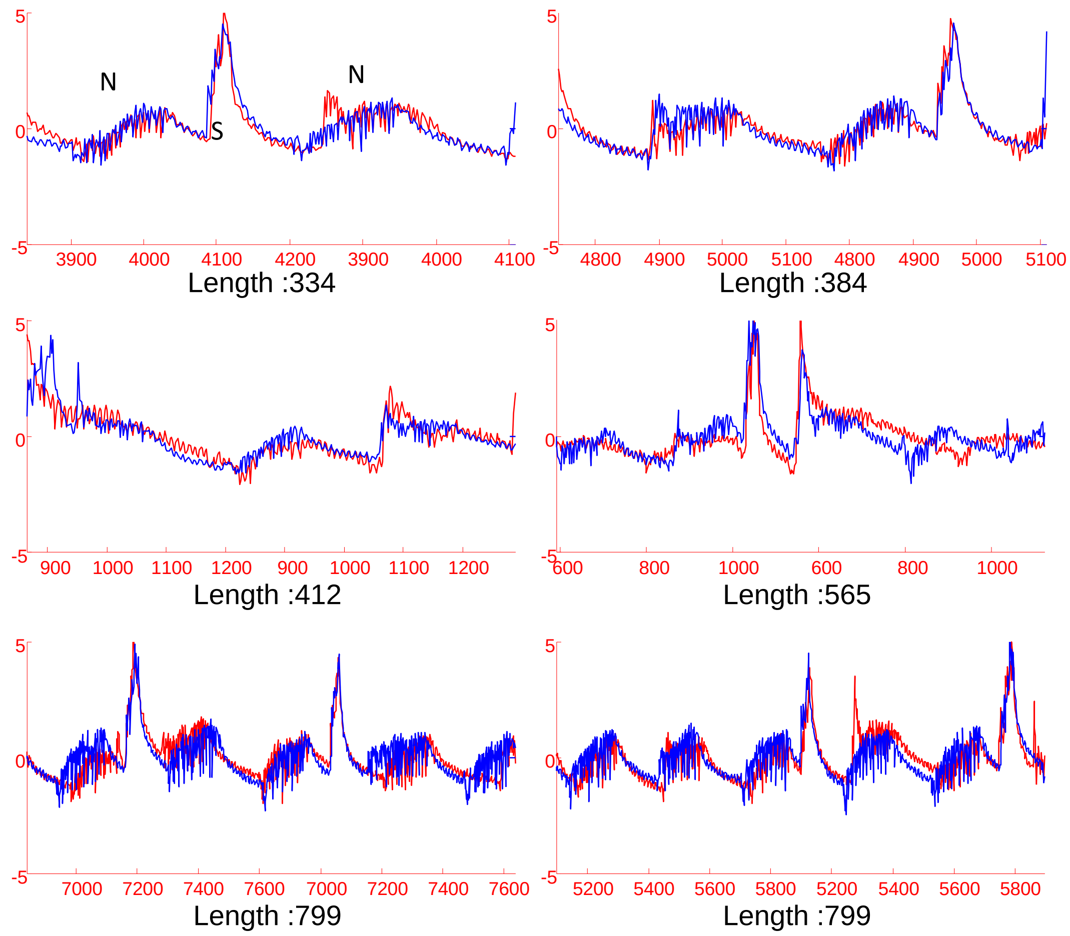


15 minutes of EPG recorded on Beet Leafhopper

As a bead of sticky secretion, which is by-product of sap feeding, is ejected, it temporarily forms a highly conductive bridge between the insect and the plant.

Exact Discovery of Time Series Motifs. A Mueen, et al. SDM, 2009.

Insect Behavior Analysis



More motifs reveal different feeding patterns of Beet Leafhopper.

Applications Outline

- Applications
  - As Subroutines in Data Mining
    - Never Ending Learning
    - Time Series Clustering
    - Rule Discovery
    - Dictionary Building
  - In Other Scientific Research
    - Data center chiller management
    - Worm locomotion analysis
    - Physiological Prediction
    - Activity recognition
  - Motifs in Other Data-types
    - Audio
    - Shapes
    - Motion

Sustainable Operation and Management of Data Center Chillers using Temporal Data Mining

HP Labs with Virginia Tech

“Our primary goal is to link the time series temperature data gathered from chiller units to high level sustainability characterizations… thus using **time series motifs** as a crucial intermediate representation to aid in data reduction.”

*“switching from **motif** 8 to **motif** 5 gives us a nearly $40,000 in annual savings!”* Patnaik et al. SIGKDD09

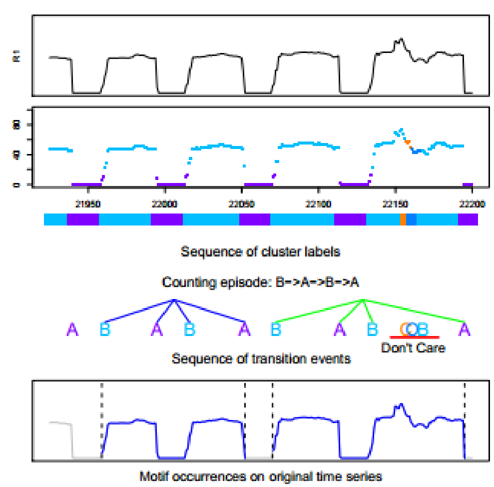
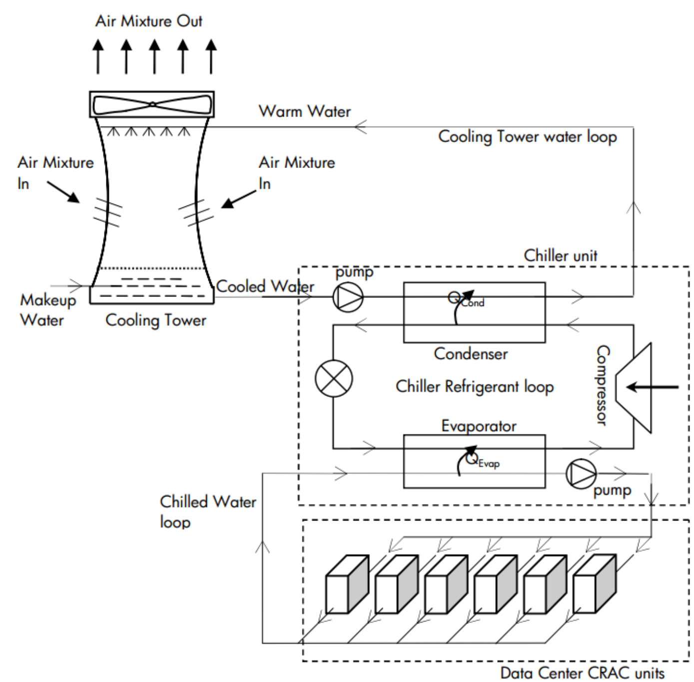


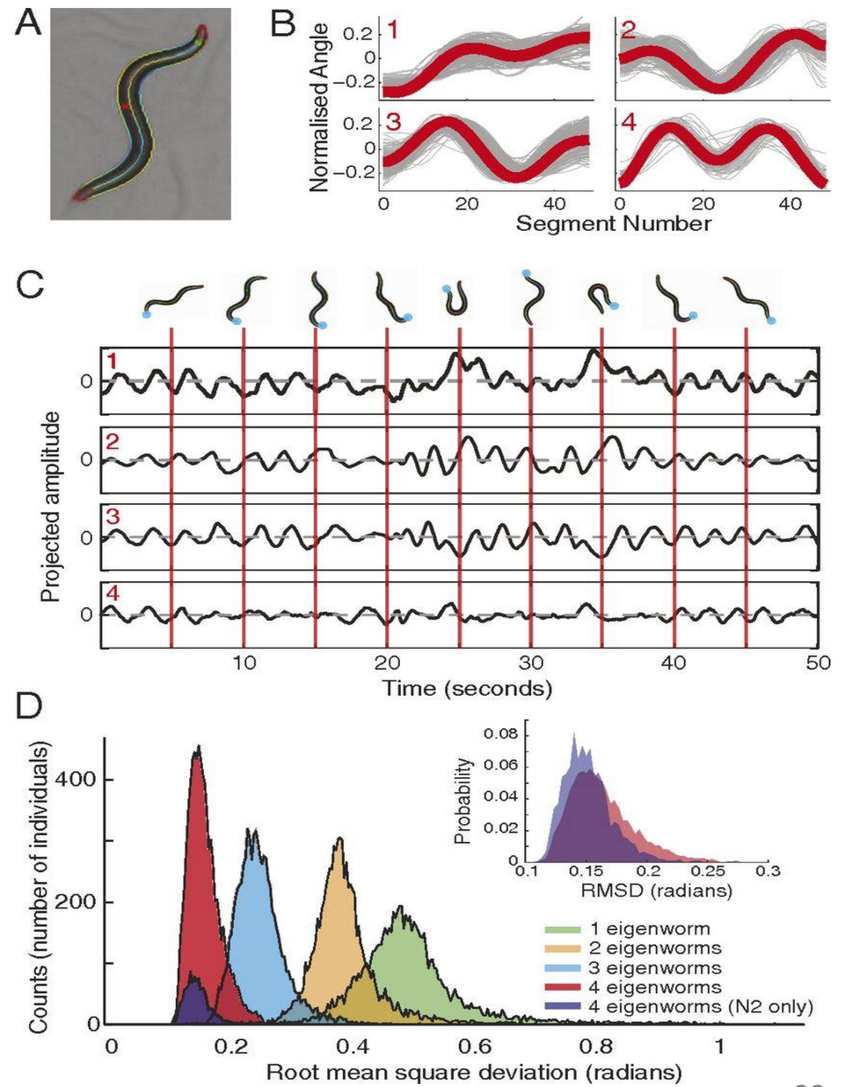
Figure 6: Illustration of motif mining in a single time-series using frequent episodes



A dictionary of behavioral motifs reveals clusters of genes affecting *C. elegans* locomotion

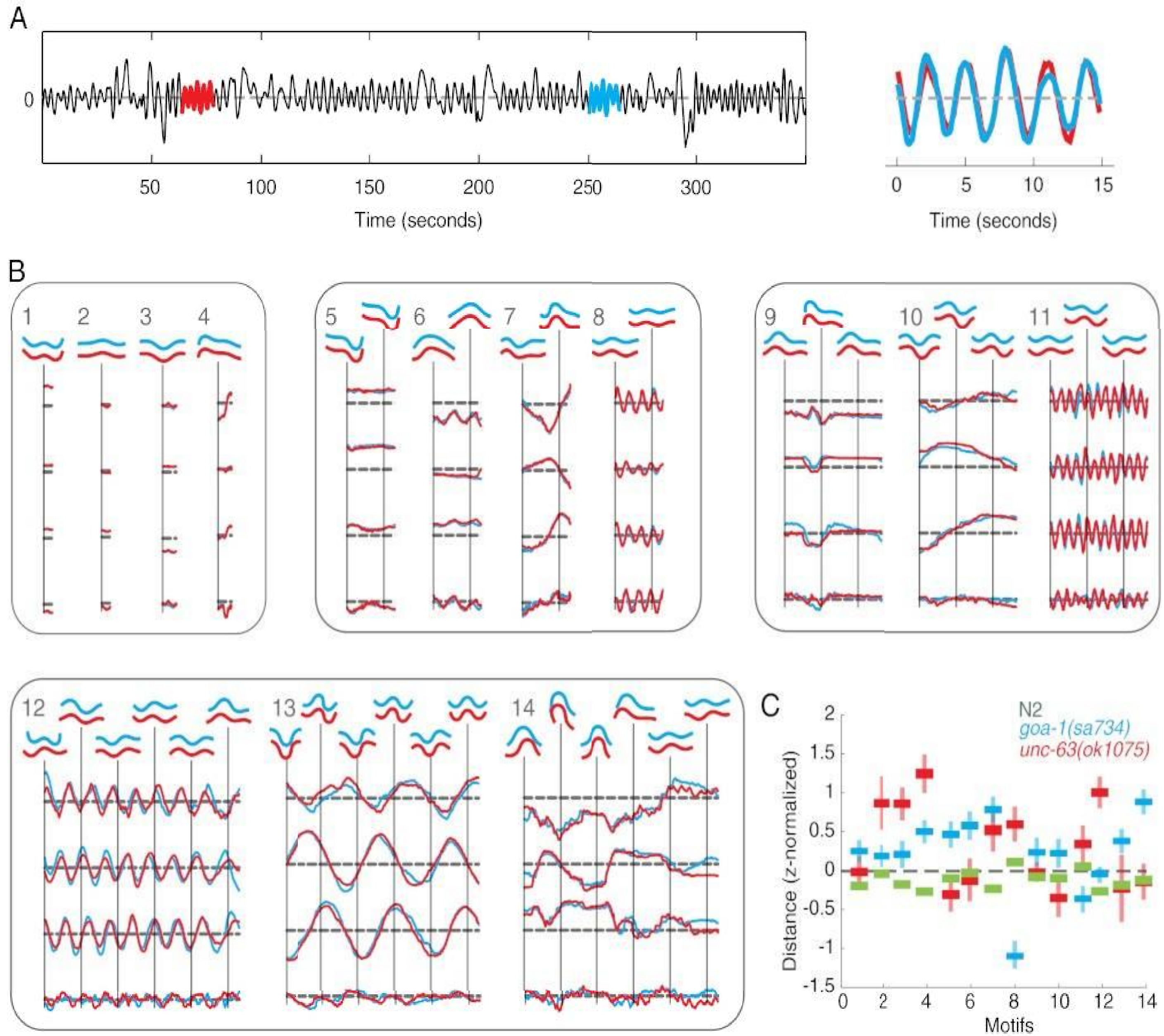
Laboratory of Molecular Biology, Cambridge, United Kingdom

Goal: Detect genotype by using the locomotion only. Convert postures to four dimensional time series.



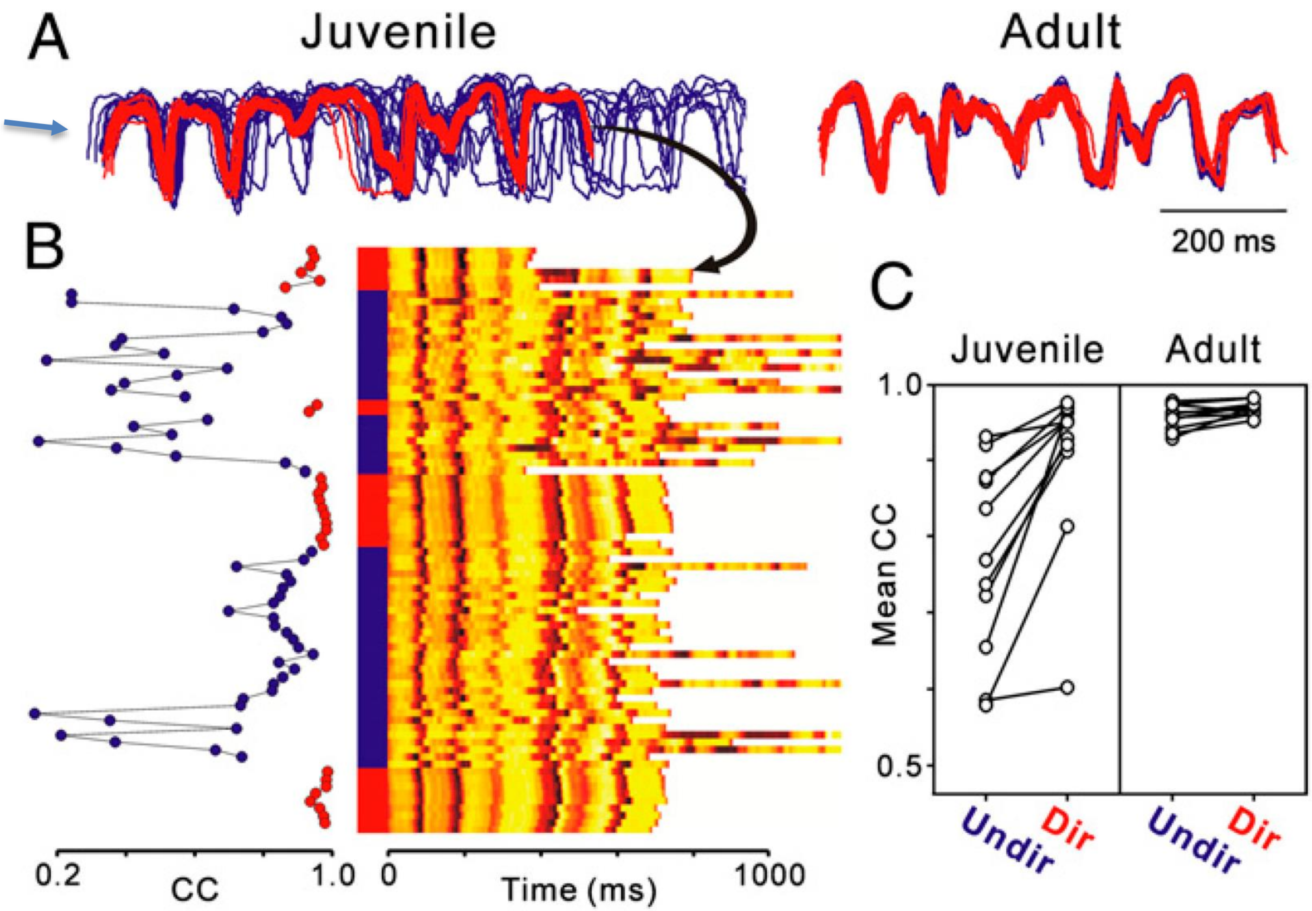
Brown et al. PNAS 2013; 110(2): 791–796.

A dictionary of behavioral motifs reveals clusters of genes affecting *C. elegans* locomotion



Variability in motif structure is lower in juvenile Directed than in Undirected and similar to that in adult song.

Amplitude Envelop



Social performance reveals unexpected vocal competency in young songbirds
Satoshi Kojima and Allison J. Doupe, PNAS 2011, 108 (4) 1687-1692.

Motif discovery in physiological datasets: A methodology for inferring predictive elements

University of Michigan and MIT

We evaluated our solution on a population of patients who experienced sudden cardiac death and attempted to discover electrocardiographic activity that may be associated with the endpoint of death. To assess the predictive patterns discovered, we compared likelihood scores for **time series motifs** in the sudden death population…

Motif Discovery in Physiological Datasets • 2:5

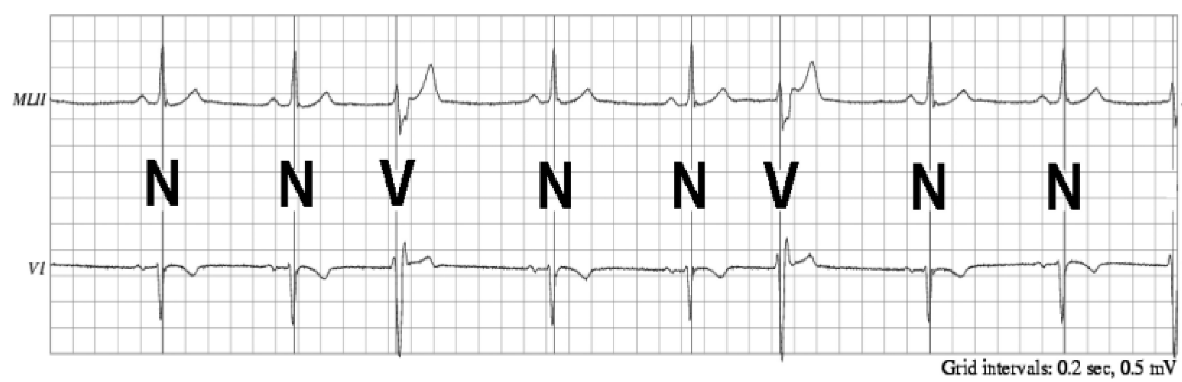


Fig. 3. Example symbolization of continuous ECG waveforms using clinical annotations (N = normal, V = premature ventricular contraction).

Zeeshan Syed et al. TKDD 2010

Constrained Motif Discovery in Time Series

Toyoaki Nishida, Kyoto University

*"we use **time series motifs** to find gesture patterns with applications to robot-human interactions"* Okada, Izukura and Nishida 2011

Constrained Motif Discovery in Time Series 25

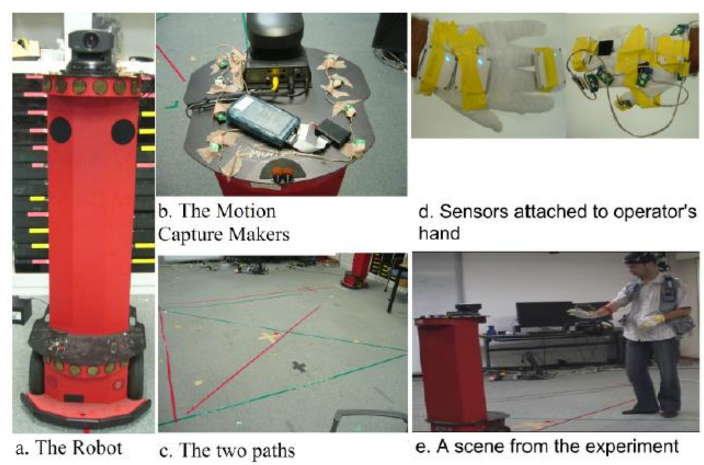
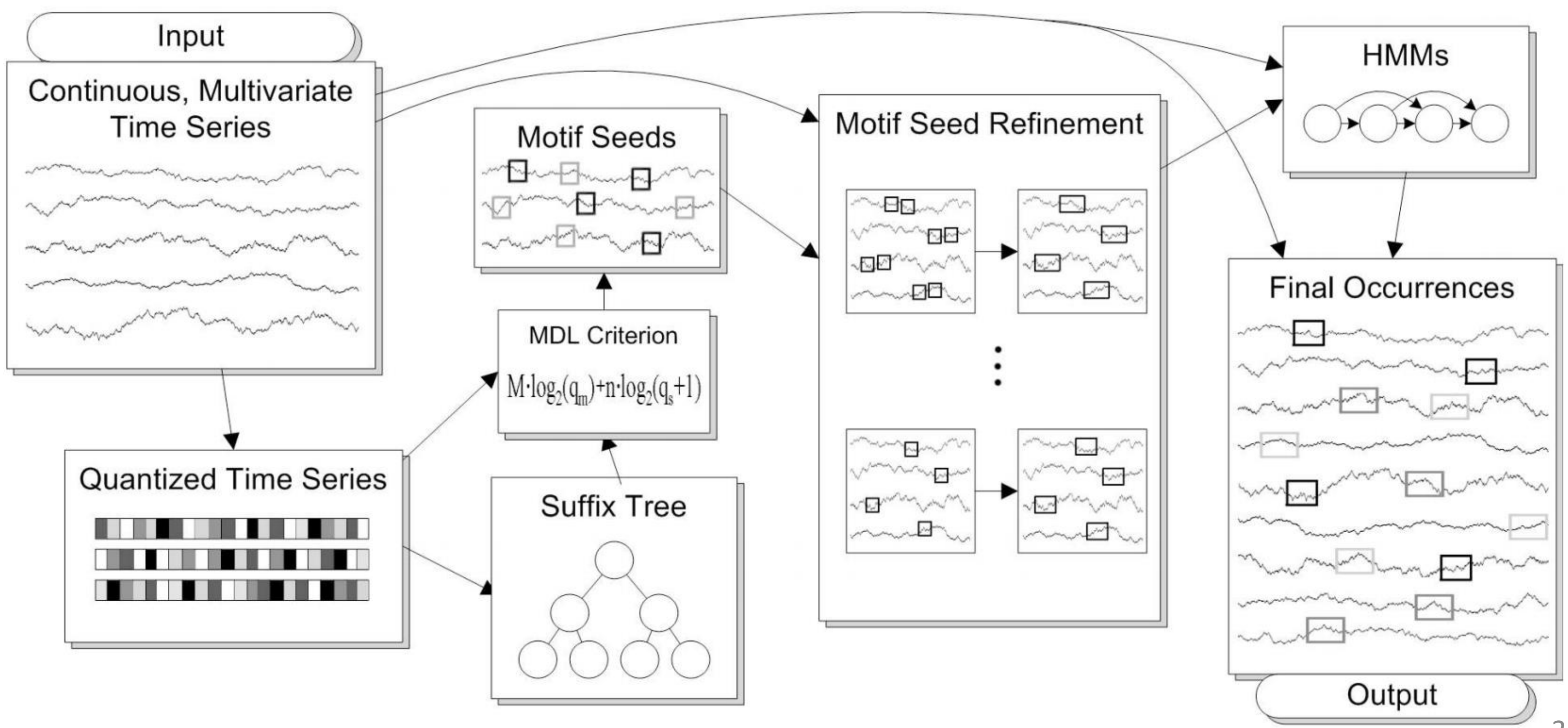


Fig. 8 The robot used in the experiment (a), the motion capture markers attached to it (b), the paths that were used (c), the sensors attached to the operator's hands (d), and a scene from the experiment (e)

Discovering Characteristic Actions from On-Body Sensor Data

David Minnen, Thad Starner, Irfan Essa, and Charles Isbell, Georgia Tech

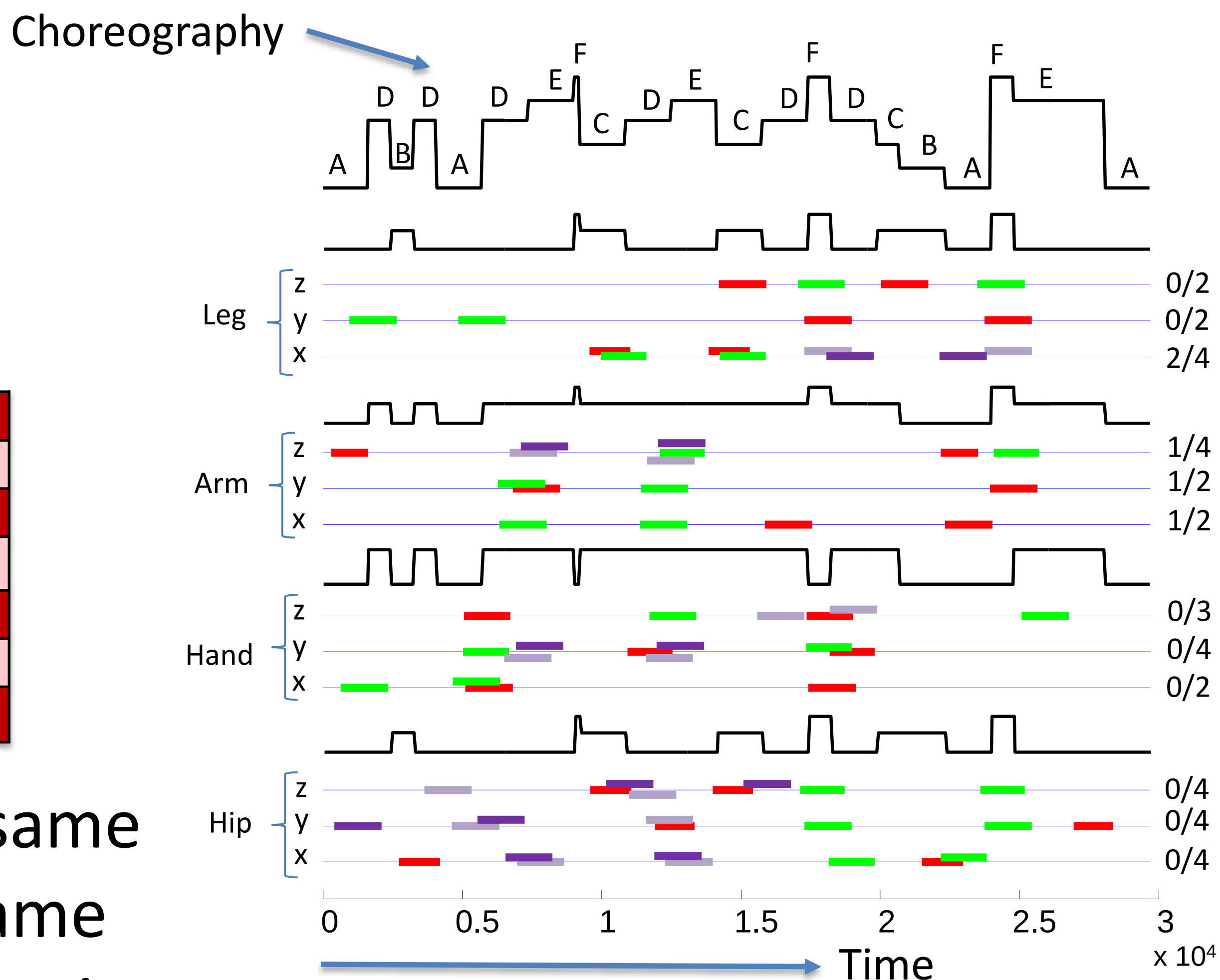
Our algorithm successfully discovers ***motifs*** that correspond to the ***real exercises*** with a recall rate of 96.3% and overall accuracy of 86.7% over six exercises and 864 occurrences.



Minnen et al. Symp. Wearable Computers 2006

Motifs can Spot Dance Moves…

Step	Action
A	Side steps with no arm movement
B	Rock steps sideways without arm movement
C	Rock steps sideways with arm movement
D	Side steps with arm movement
E	Side steps with arms up in the air
F	Standing still with head bopping



Motifs are from the same dance steps or the same transitions 86% of the time.

Data: H. Pohl et al. SMC 2010
Algorithm: Mueen, ICDM 2013

Applications Outline

- Applications
  - As Subroutines in Data Mining
    - Never Ending Learning
    - Time Series Clustering
    - Rule Discovery
    - Dictionary Building
  - In Other Scientific Research
    - Data center chiller management
    - Worm locomotion analysis
    - Physiological Prediction
    - Activity recognition
  - Motifs in Other Data-types
    - Audio
    - Shapes
    - Motion

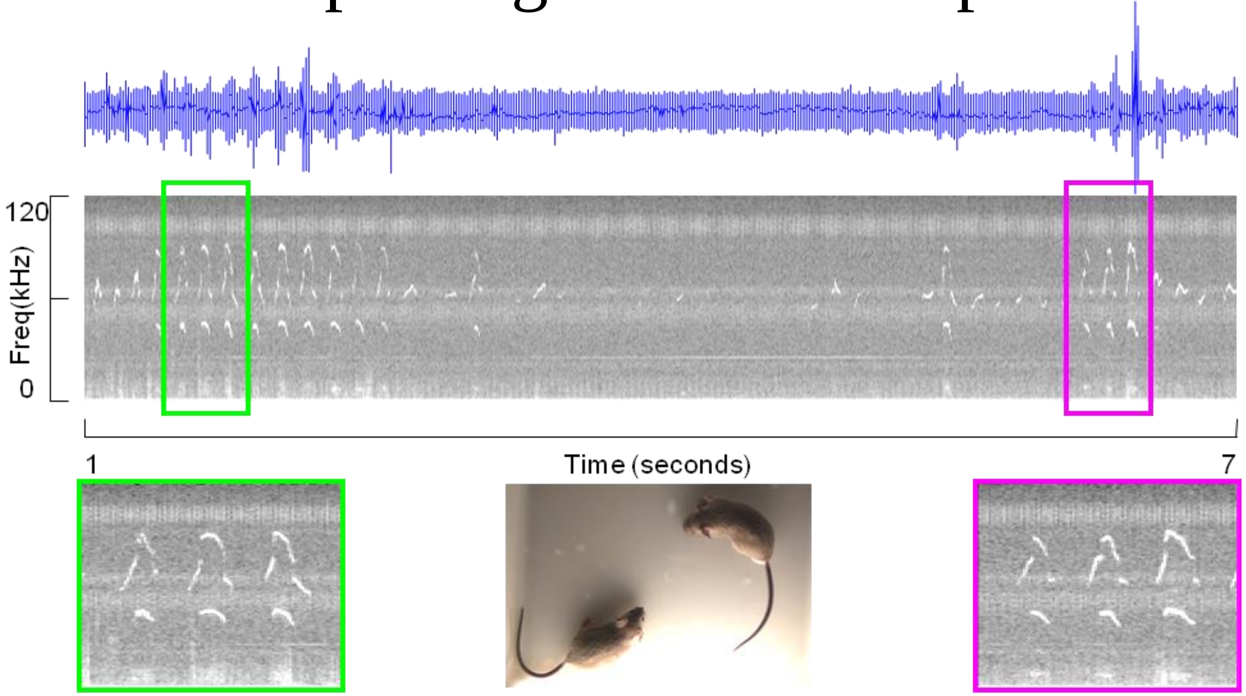
Motifs in Audio

Mice calls are inaudible and have significant noise

Manual inspection over temporal signal is impossible

Features like MFCC are not good for animal song

Just use the raw spectrogram to find repeated calls



Yuan Hao et al. Parameter-Free Audio Motif Discovery in Large Data Archives. ICDM 2013: 261-270

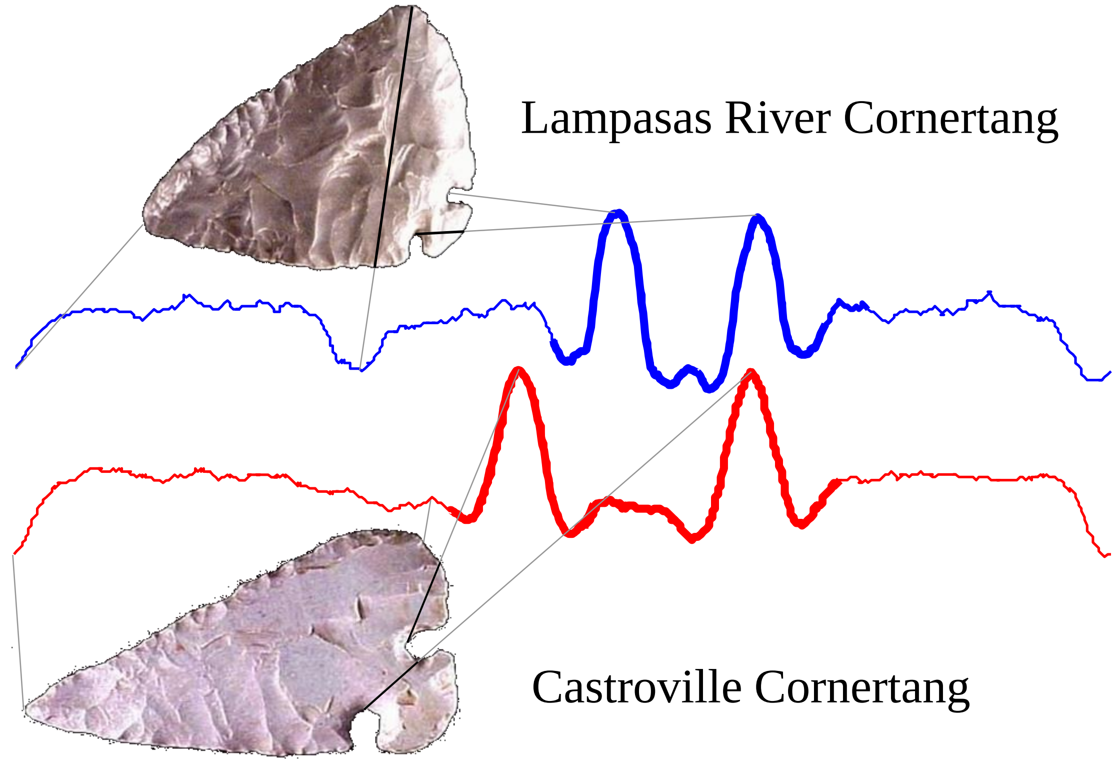
Motifs in Shapes

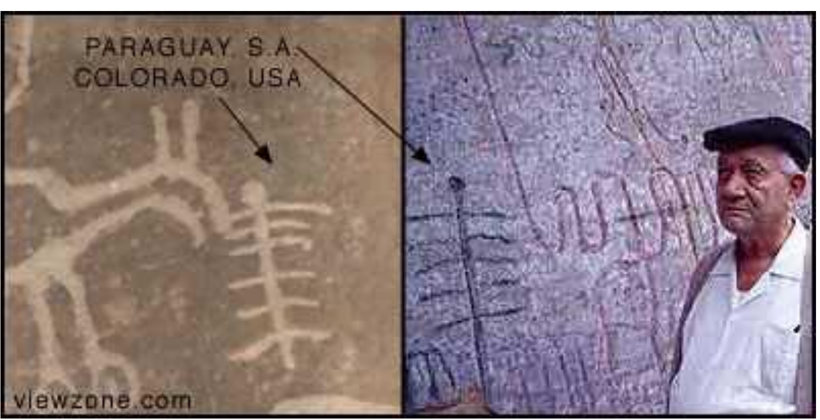
Projectile shapes

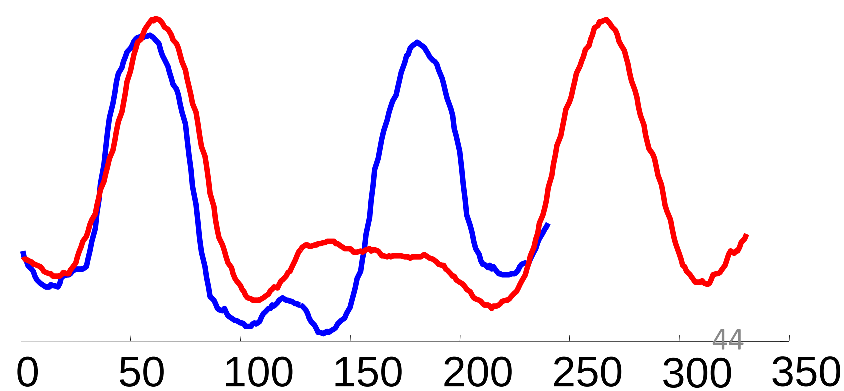
Algorithm detects a rare cornertang segment – an object that has long intrigued anthropologists.

Petroglyphs

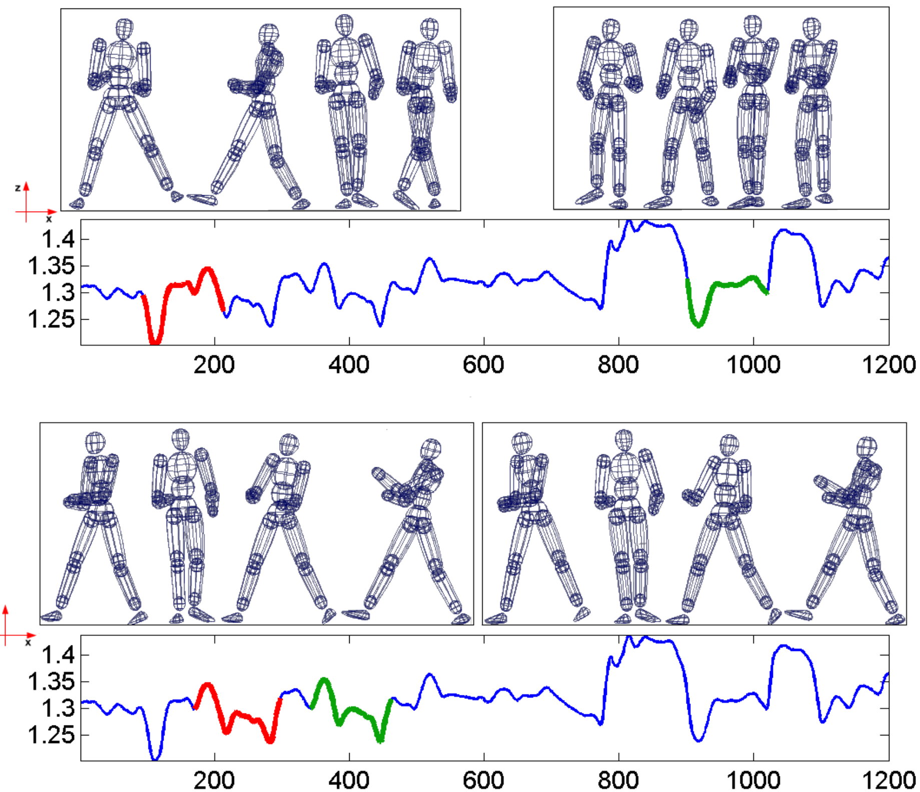
Algorithm detects similar petroglyphs drawn across continents and centuries

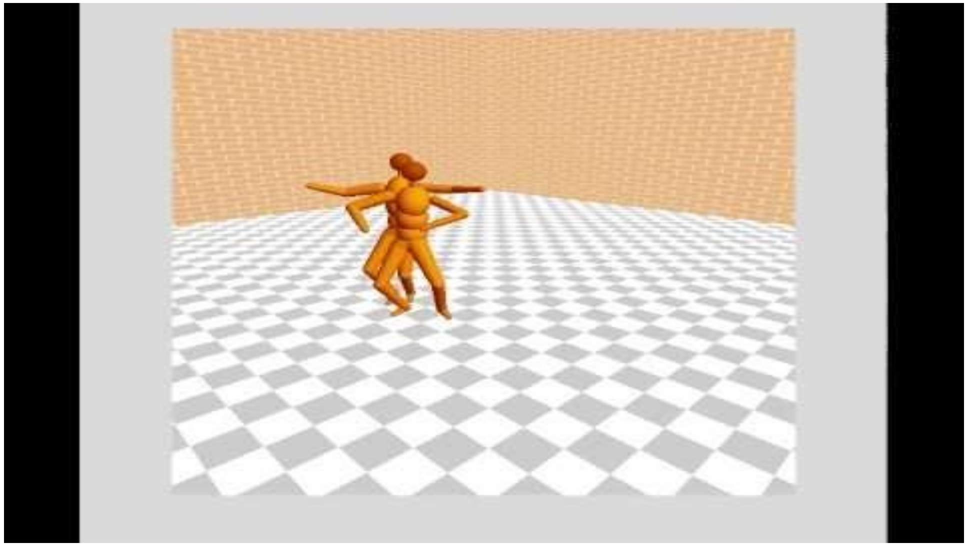






Motifs in Motion





Two motion can be stitched together by transiting from one motif to the other, a very useful technique for motion synthesis.

Mueen et al. A disk-aware algorithm for time series motif discovery. Data Min. Knowl. Discov. 2011.

Yankov et al. Detecting time series motifs under uniform scaling. KDD 2007

Time Series Motifs have 1,000 of Uses

- *..for discovering **motifs** in the **music** data is called the Mueen-Keogh (**MK**) algorithm*.. Cabredo et al. 2011
- *we apply the **MK motif** algorithm to time series retrieved from **seismic** signals…* Cassisi et al 2012
- *we take **motif** developed by Keogh in order to support a **medical** expert in discovering interesting knowledge.* Kitaguchi.
- *for the problem of estimation of Micro-drilled Hole Wall of PWBs we take the **Motif** method developed by Keogh...* Toshiki et al. (***fabrication***)
- *the most efficient **motif** provided a **power** savings of 41 This translates to an annual reduction of 287 tons of CO2*. Watson InterPACK09.
- *We use Keogh's **Motifs** for unsupervised discovery of abnormal **human** behavior in multi-dimensional time series data...* Vahdatpour SDM 2010.
- *variability of behavior, using **motifs**, provides more consistent groupings of households across different clustering algorithms…* Ian Dent 2014

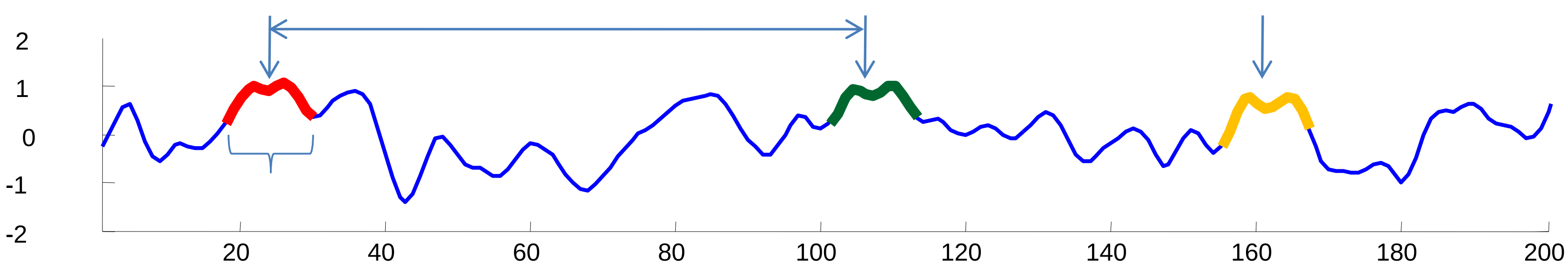
Questions and Comments

Algorithms Outline

- Algorithms
  - Definition, Distance Measures and Invariances
  - Exact Algorithms
    - Fixed Length
    - Enumeration of All length
    - K-motif Discovery
    - Online Maintenance
  - Approximate Algorithms
    - Random Projection Algorithm
  - Multi-dimensional Motif Discovery
  - Open Problems

Definition of Time Series Motifs

1. Length of the motif
2. Support of the motif
3. Similarity of the Pattern
4. Relative Position of the Pattern



Distance Measures

- The choices are
  - Euclidean Distance
  - Correlation
  - Dynamic Time Warping
  - Longest Common Subsequences
  - Uniformly Scaled Euclidean Distance
  - Sliding Nearest Neighbor Distance

Euclidean Distance Metric

Given two time series

$\mathbf{x} = x_1 \ldots x_n$

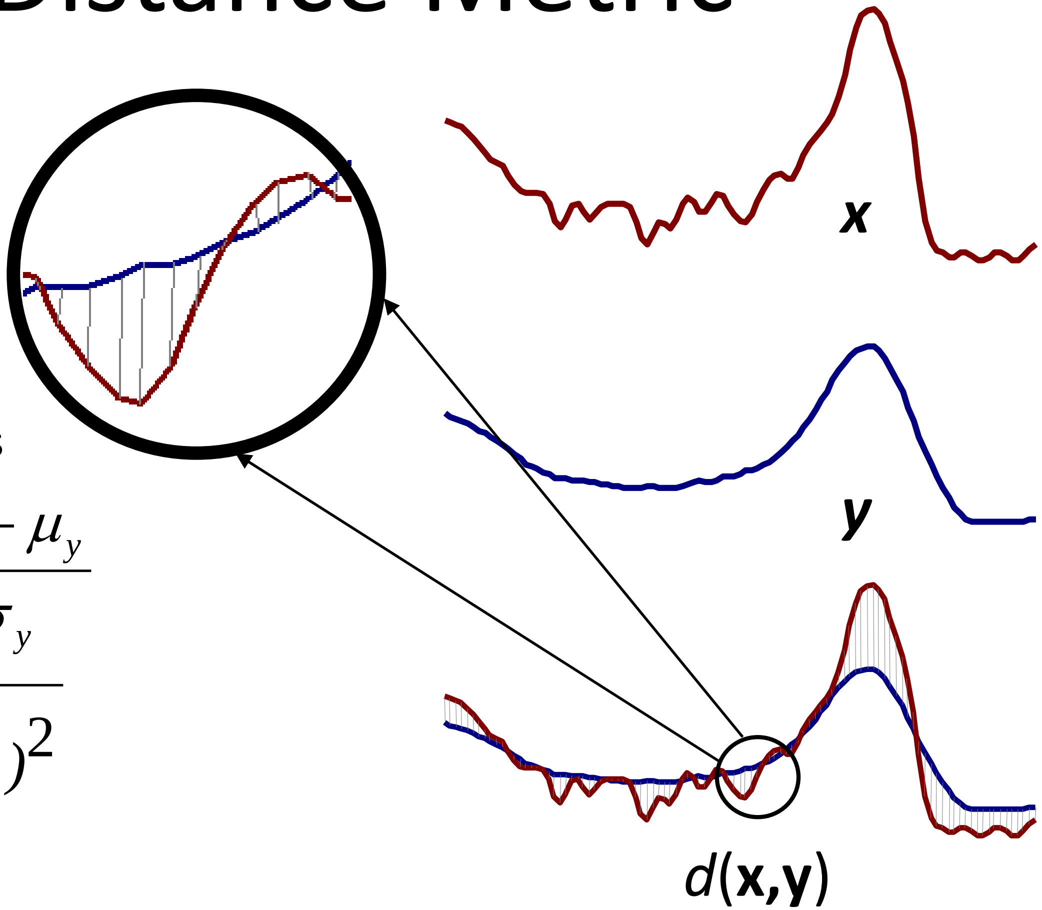
and

$\mathbf{y} = y_1 \ldots y_n$

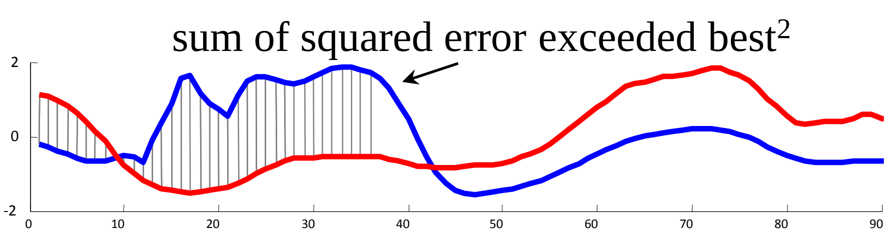
their z-Normalized Euclidean distance is defined as:

$$\hat{x}_i = \frac{x_i - \mu_x}{\sigma_x}, \hat{y}_i = \frac{y_i - \mu_y}{\sigma_y}$$

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (\hat{x}_i - \hat{y}_i)^2}$$



Early abandoning reduces number of operations when minimizing



Pearson's Correlation Coefficient

- Given two time series x and y of length m.
- Sufficient Statistics:

$$\sum_{i=1}^{m} x_i y_i \quad \sum_{i=1}^{m} x_i \quad \sum_{i=1}^{m} y_i \quad \sum_{i=1}^{m} x_i^2 \quad \sum_{i=1}^{m} y_i^2$$

- Correlation Coefficient:

$$corr(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{m} x_i y_i - m\mu_x \mu_y}{m\sigma_x \sigma_y}$$

Where $\mu_x = \frac{\sum_{i=1}^{m} x_i}{m}$ and $\sigma_x^2 = \frac{\sum_{i=1}^{m} x_i^2}{m} - \mu_x^2$

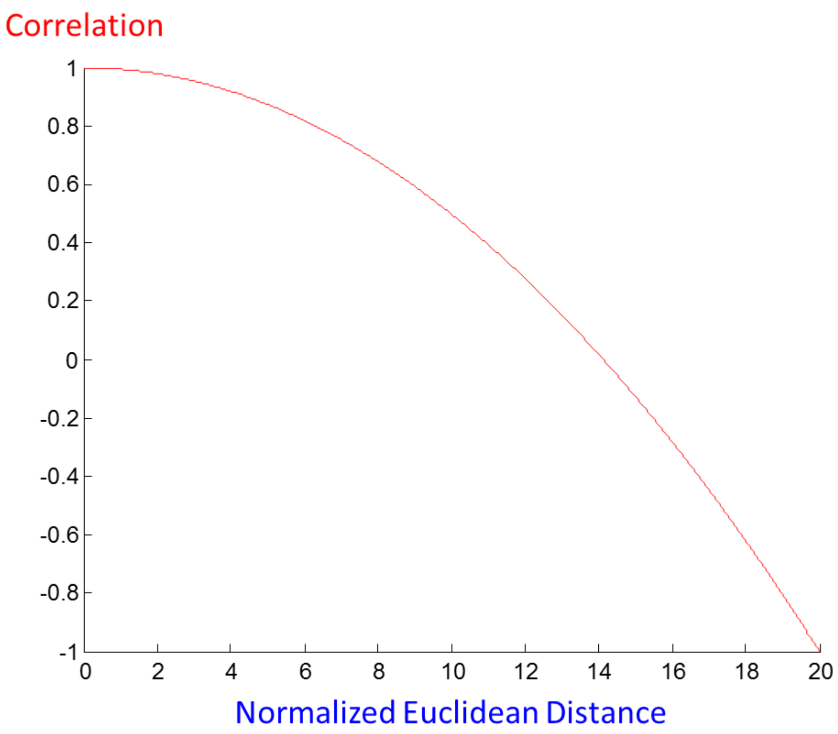
- Early abandoning is possible when maximizing
- Correlation is not a metric, therefore, use of triangular inequality needs special attention

Relationship with Euclidean Distance

$$d(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = \sqrt{2m(1 - corr(\boldsymbol{x}, \boldsymbol{y}))}$$

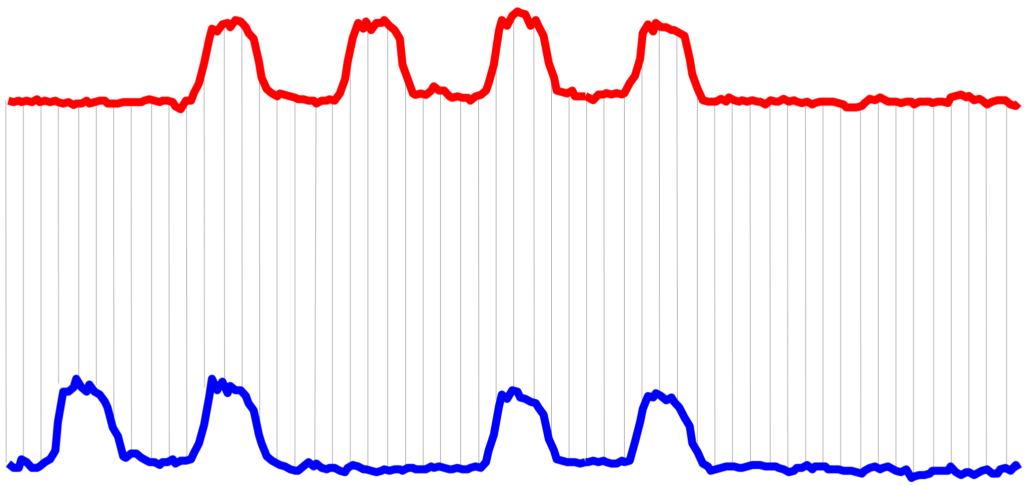
$$\hat{x}_i = \frac{x_i - \mu_x}{\sigma_x} \text{ and } \hat{y}_i = \frac{y_i - \mu_y}{\sigma_y}$$

$$d^2(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = \sum_{i=1}^{m} (\hat{x}_i - \hat{y}_i)^2$$



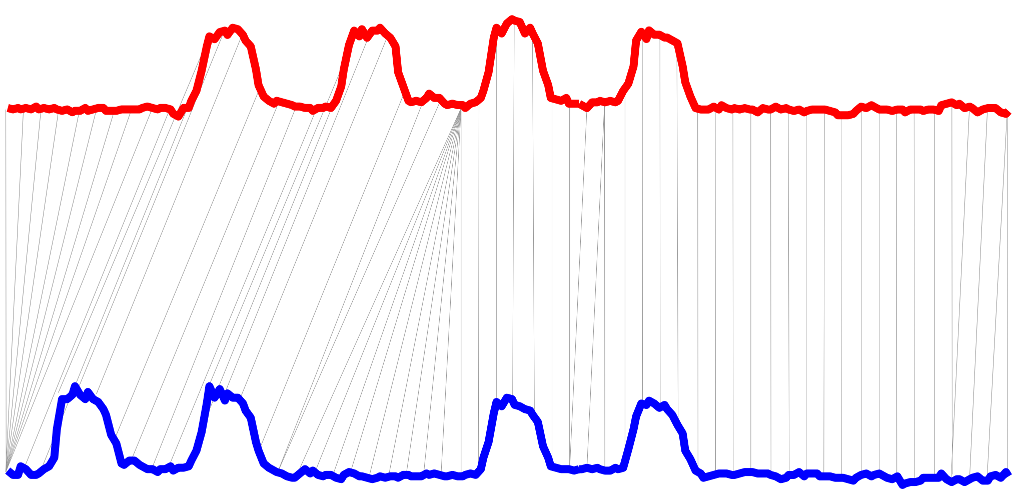
Minimizing z-normalized Euclidean distance and Maximizing Pearson's correlation coefficient are identical in effect for motif discovery.

Euclidean Vs Dynamic Time Warping



Euclidean Distance

Sequences are aligned “one to one”.



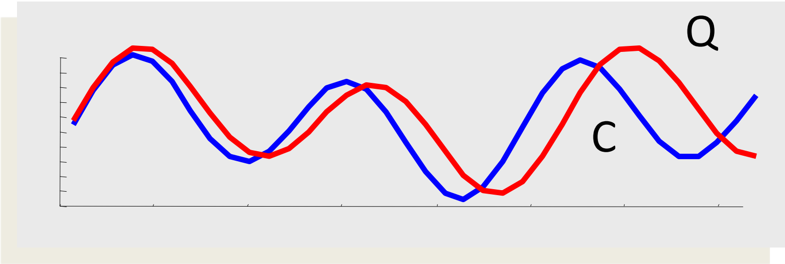
“Warped” Time Axis

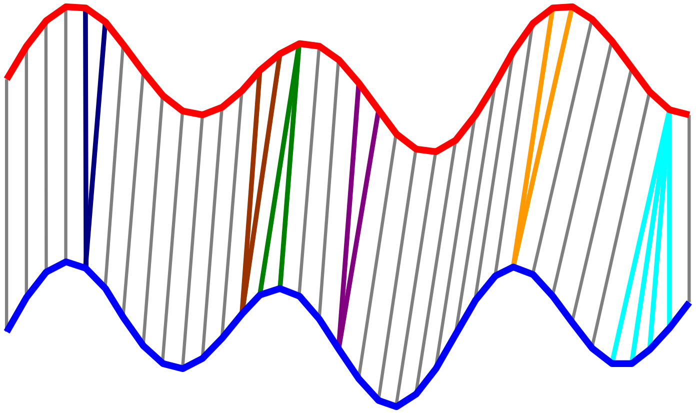
Nonlinear alignments are possible.

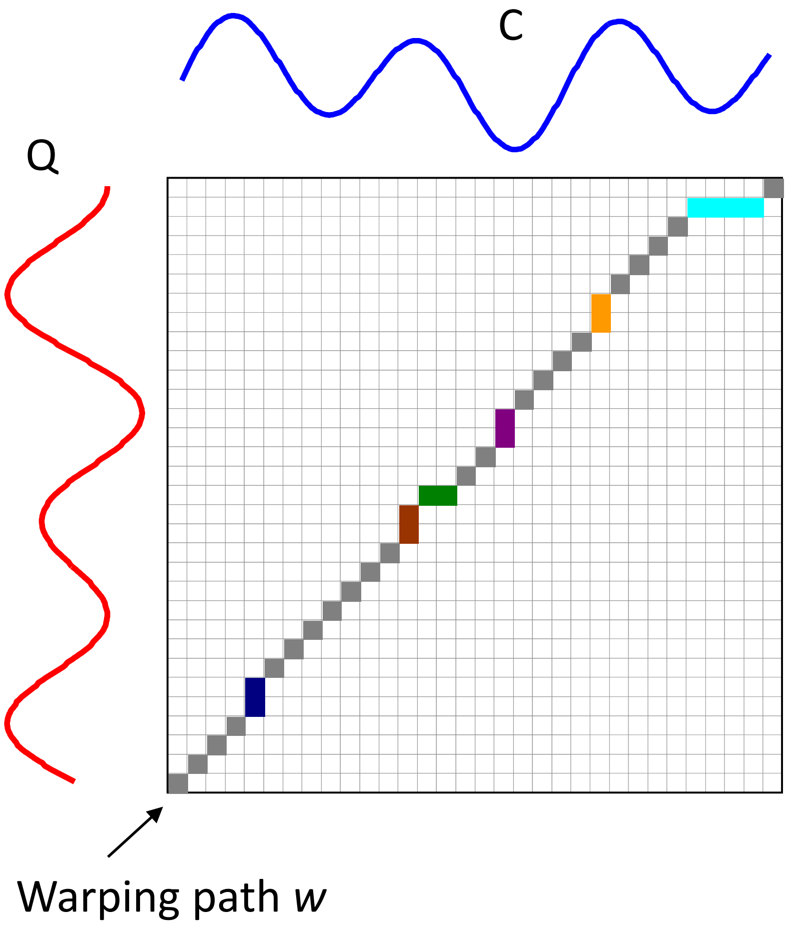
How is DTW Calculated?

$$DTW(Q,C) = \sqrt{D(m,n)}$$

$$D(i,j) = (q_i - c_j)^2 + \min\{D(i,j-1), D(i-1,j), D(i-1,j-1)\}$$







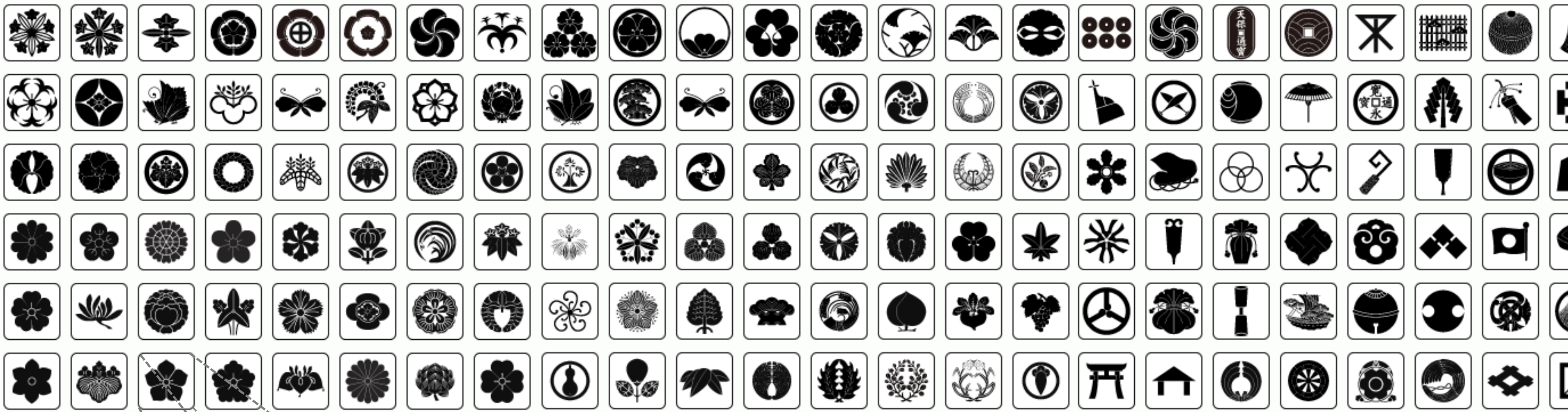
- Quadratic time complexity
- DTW is not a metric

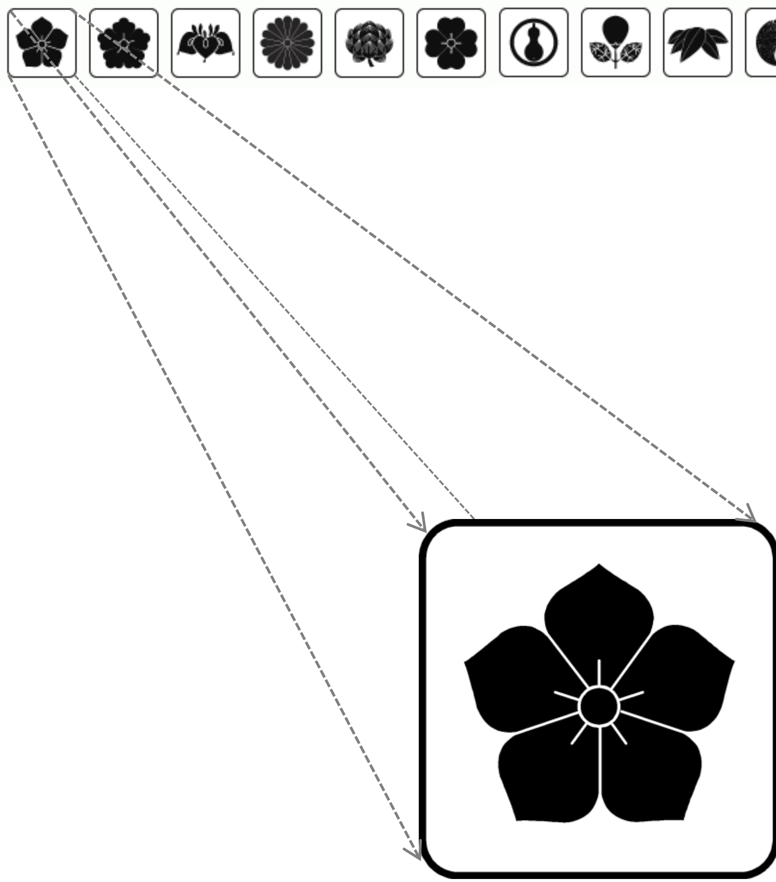
A four-slide digression, to make sure you understand what *invariances* are, and why they are important

Suppose we are walking in a cemetery in Japan.

We see an interesting grave marker, and we want to learn more about it.

We can take a photo of it and search a database….

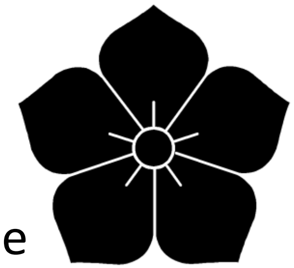


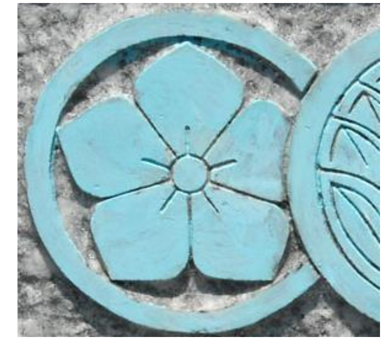


Campana and Keogh (2010). A Compression Based Distance Measure for Texture. SDM 2010.

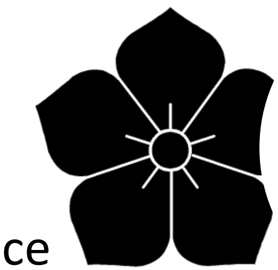
In order to do this, we must have a distance measure with the right *invariances*

Color invariance

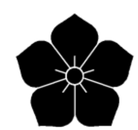




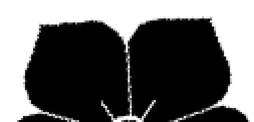
Occlusion invariance



Size invariance



Rotation invariance



Time Series Data has Unique Invariances

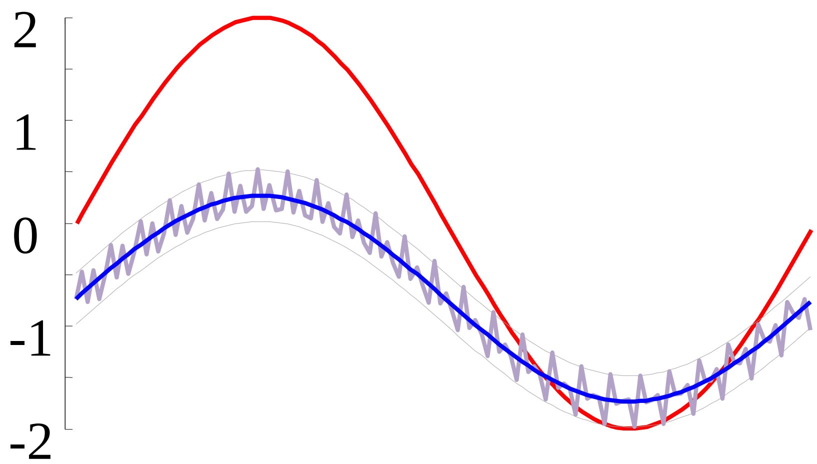
- These invariances are domain/problem dependent
- They include
  - Complexity invariance
  - Warping invariance
  - Uniform scaling invariance
  - Occlusion invariance
  - Rotation/phase invariance
  - *Offset invariance*
  - *Amplitude invariance*

Z-normalization of each subsequence removes these

- Sometimes you achieve the invariance in the distance measure, sometimes by preprocessing the data.
- In this work, we will just assume offset/amplitude invariance. See [a] for a visual tour of time series invariances.

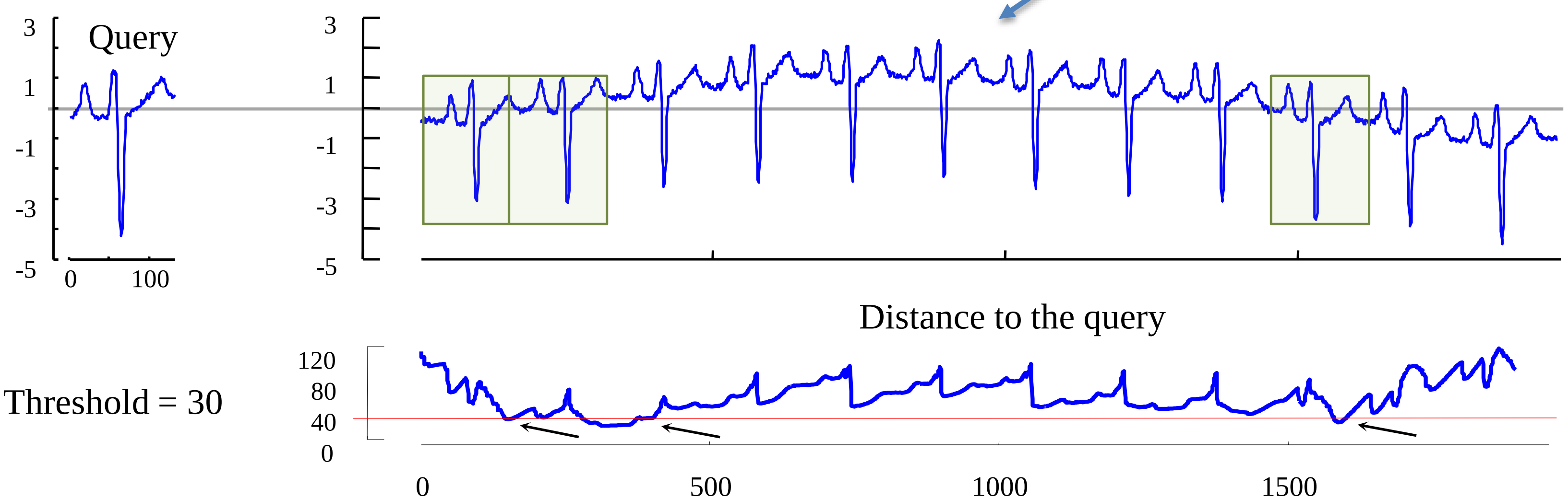
[a] Batista, et al (2011) A Complexity-Invariant Distance Measure for Time Series. SDM 2011

Z-Normalization ensures scale and offset invariances



$$\hat{x}_i = \frac{x - \mu_x}{\sigma_x}$$

Wandering Baseline



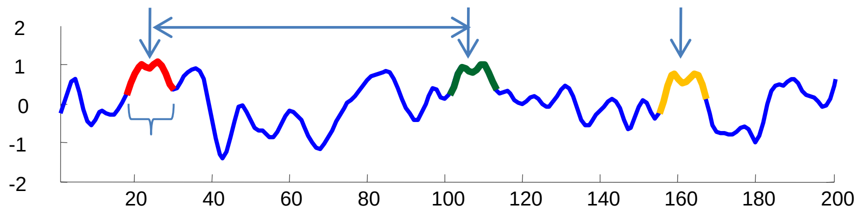
Without Normalization only 75% of the beats are missed

Algorithms Outline

- Algorithms
  - Definition, Distance Measures and Invariances
  - Exact Algorithms
    - Fixed Length
    - Enumeration of All length
    - K-motif Discovery
    - Online Maintenance
  - Approximate Algorithms
    - Random Projection Algorithm
  - Multi-dimensional Motif Discovery
  - Open Problems

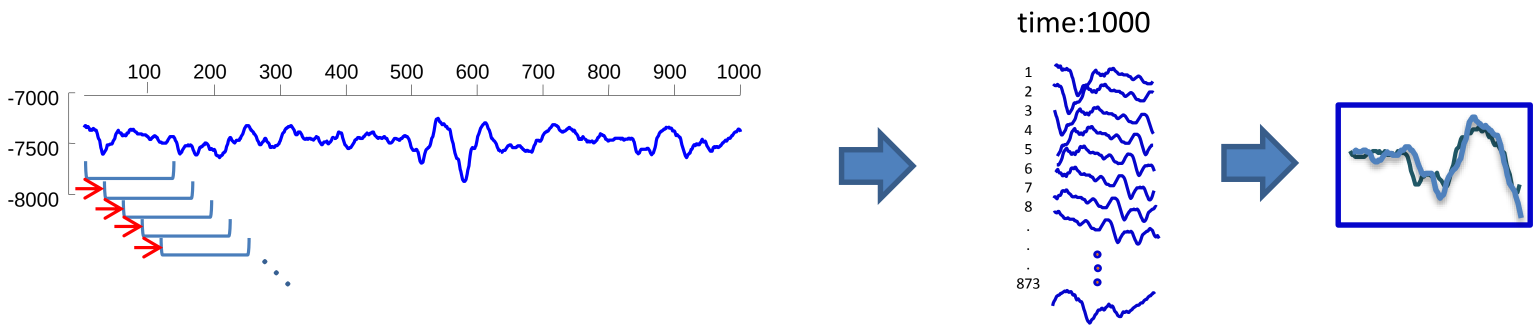
Simplest Definition of Time Series Motifs

Given a length, the most similar/least distant pair of non-overlapping subsequences



1. Length of the motif = **Given**
2. Support of the motif = **2**
3. Similarity of the Pattern = **Euclidean distance**
4. Relative Position of the Pattern = **non-overlapping**

Problem Formulation



The most similar pair of non-overlapping subsequences → The closest pair of points in high dimensional space

- ❖ Optimal algorithm in two dimension : $\Theta(n \log n)$
- ❖ For large dimensionality d, optimum algorithm is effectively $\Theta(n^2d)$

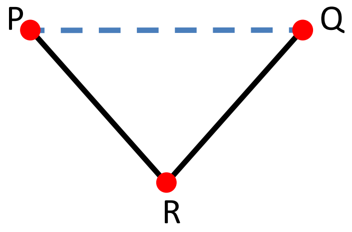
Lower Bound

- If P, Q and R are three points in a d-space

  $$d(P,Q)+d(Q,R) \geq d(P,R)$$

  $$\Rightarrow$$

  $$d(P,Q) \geq |d(Q,R) - d(P,R)|$$



- A third point R provides a very inexpensive lower bound on the true distance
- If the lower bound is larger than the existing best, skip d(P, Q)

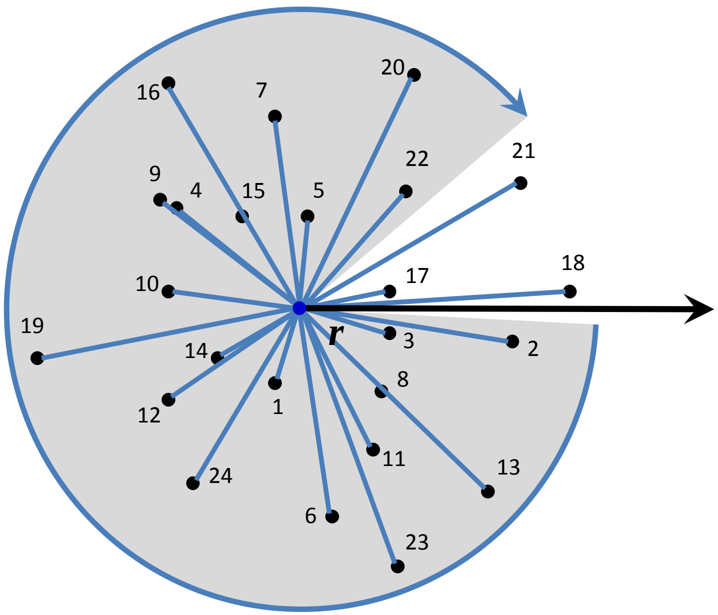
  $$d(P,Q) \geq |d(Q,R) - d(P,R)| \geq \text{BestPairDistance}$$

Circular Projection

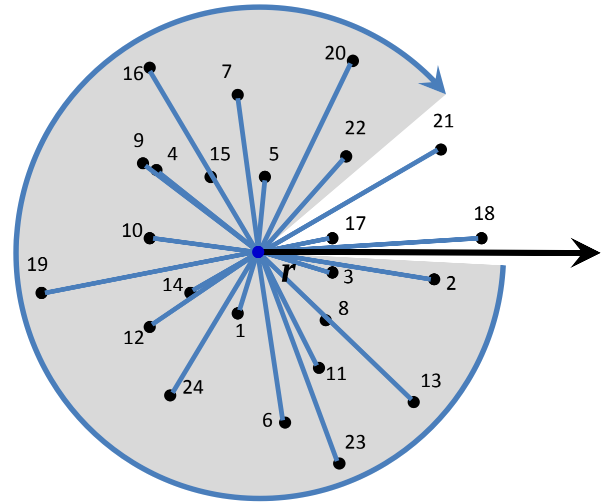
Pick a reference point *r*

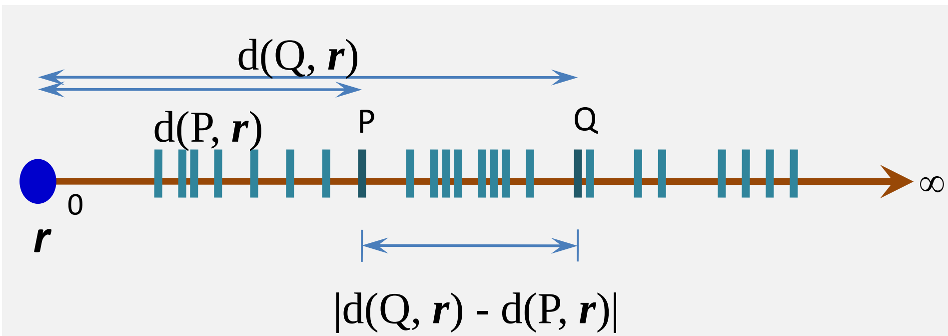
Circularly Project all points on a line passing through the reference point

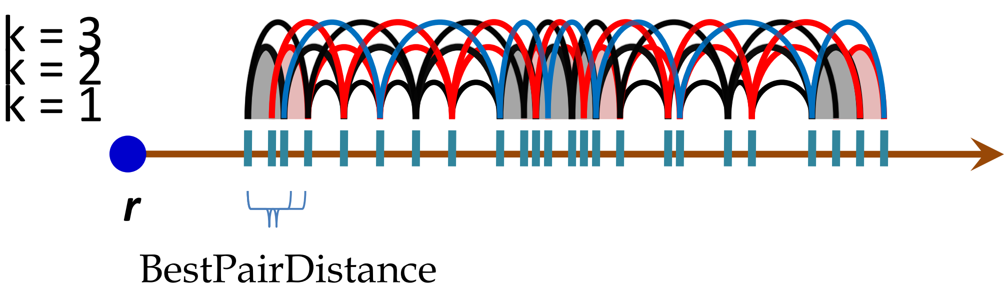
Equivalent to computing distance from *r* and then sorting the points according to *distance*



The Order Line







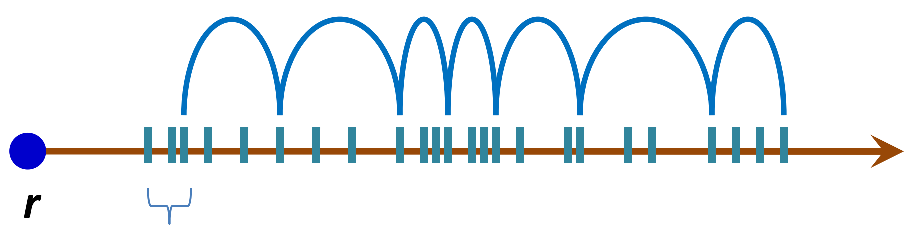
k=1:n-1

- Compare every pair having k-1 points in between

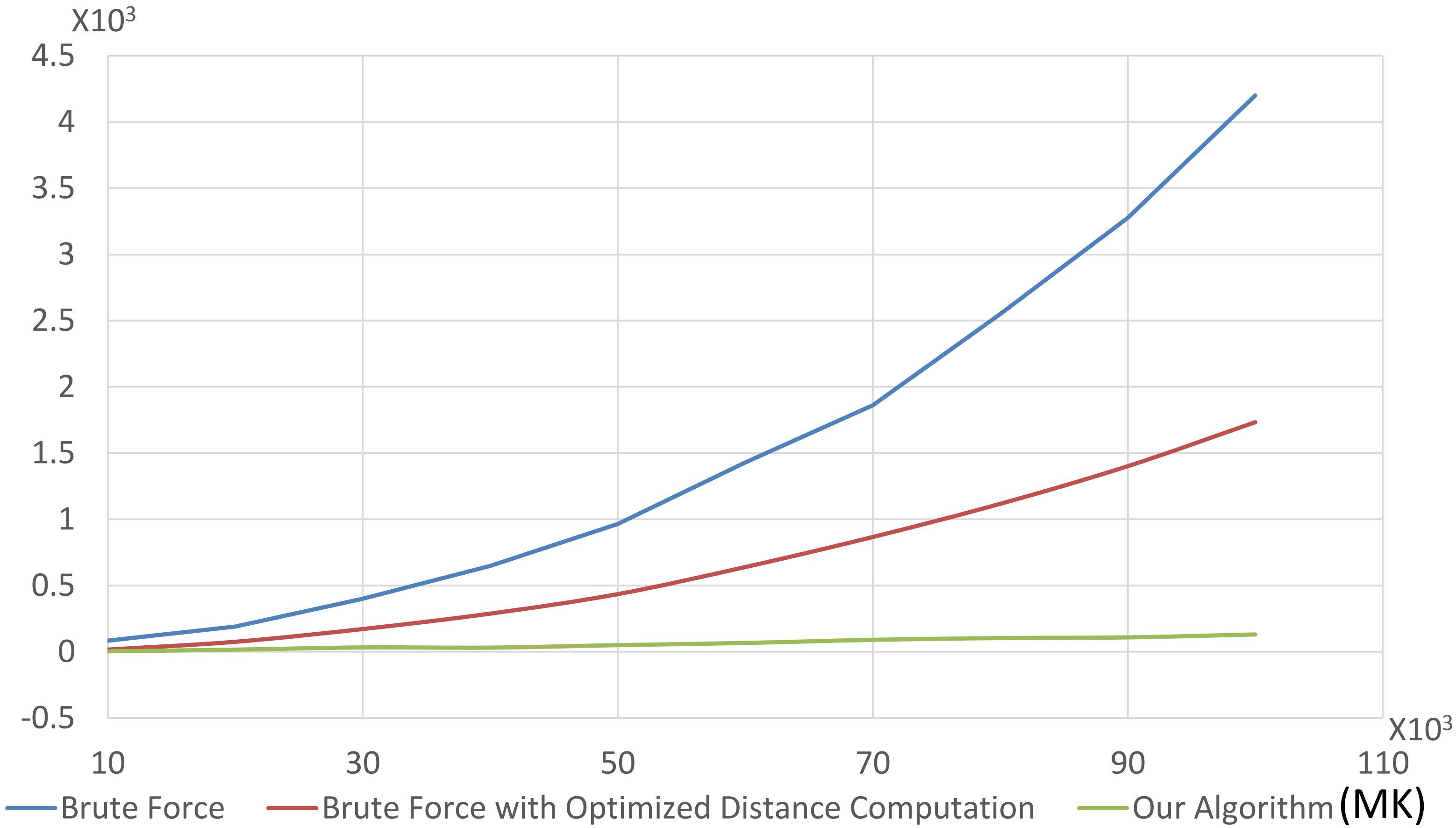
- Do k scans of the order line, starting with the 1st to kth point

Correctness

- If we search for all offset=1,2,…,n-1 then all possible pairs are considered.
  - n(n-1)/2 pairs
- **if** for any offset=k, none of the k scans needs an actual distance computation
  then for the rest of the offsets=k+1,…,n-1 no distance computation will be needed.



Performance



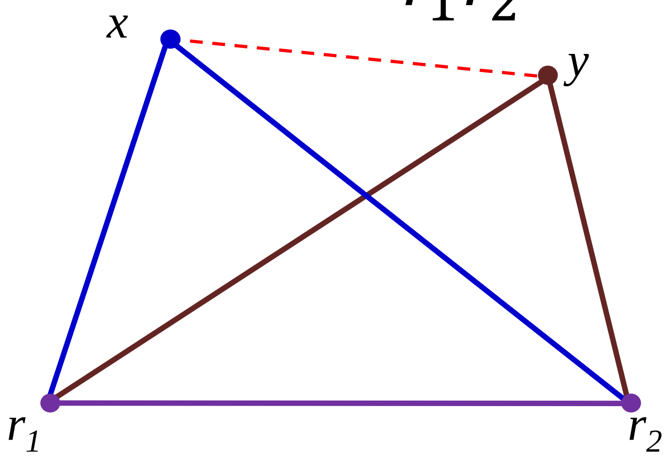
- Orders of Magnitude faster
- Exact in execution
- No sacrifice of the quality of the results

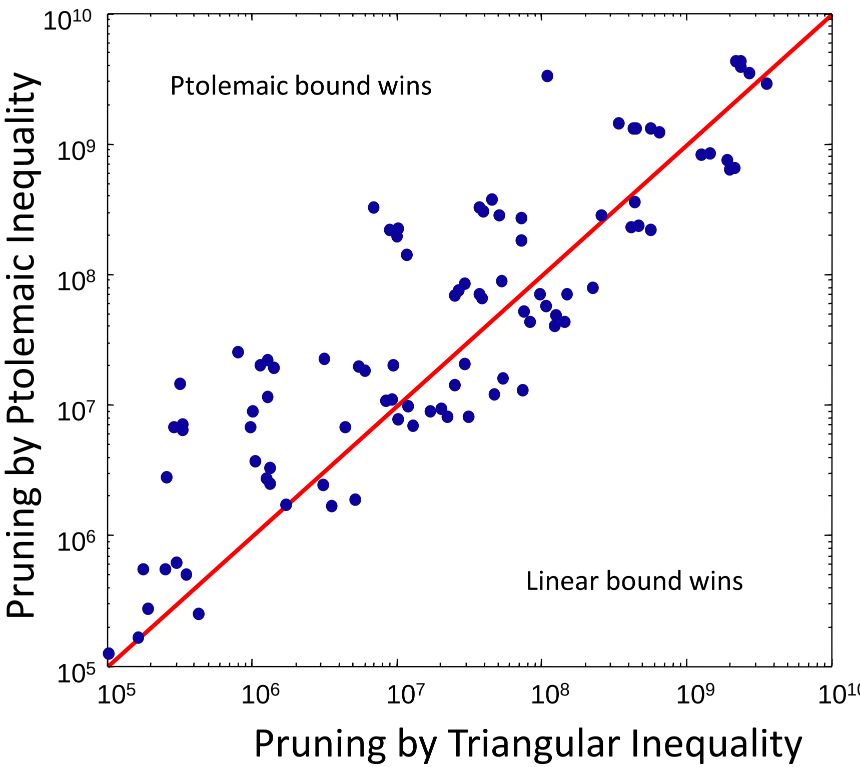
Multiple References

- Use multiple reference points for tighter lower bounds.

Ptolemaic bound

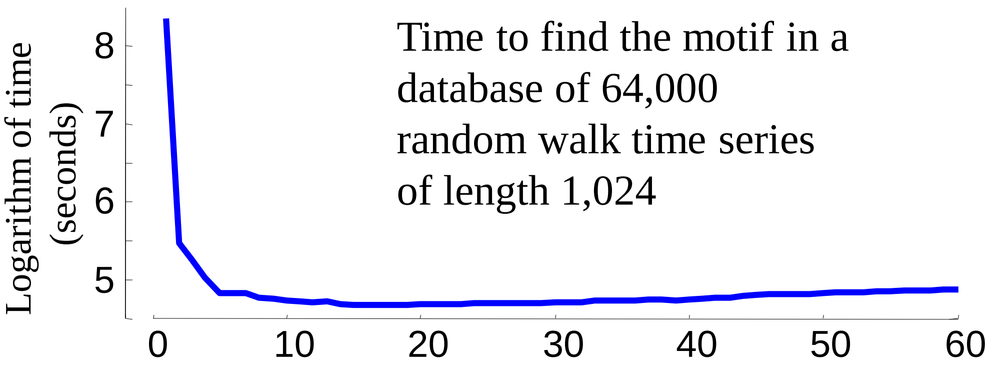
$$xy \geq \frac{|xr_1.yr_2 + xr_2.yr_1|}{r_1 r_2}$$



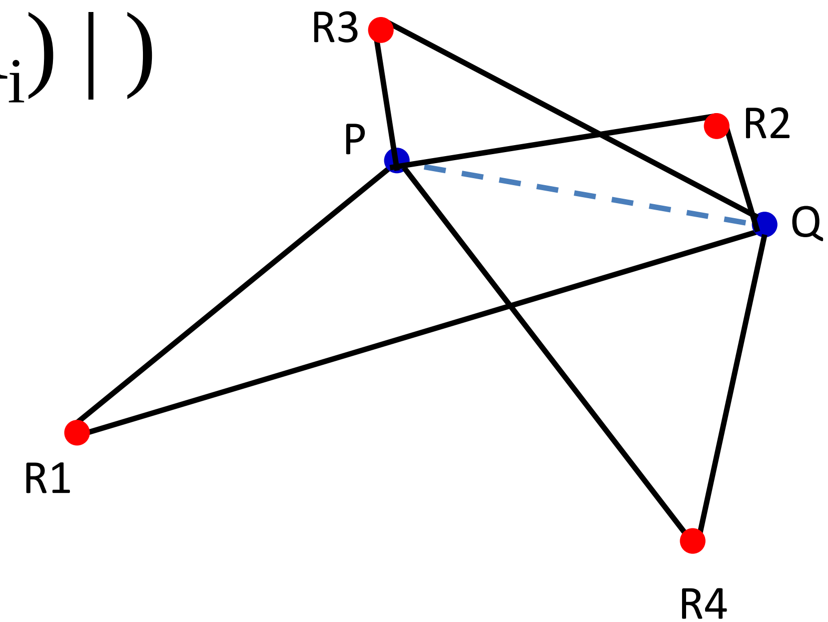


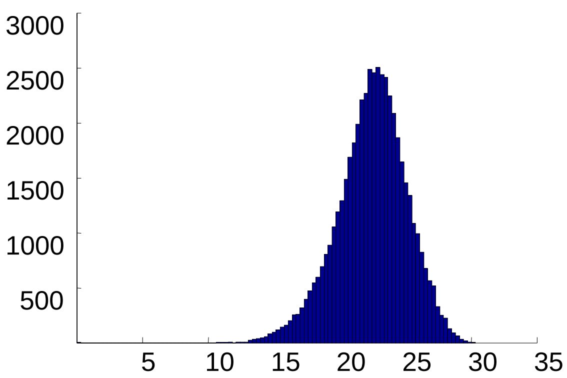
Pruning by Multiple References

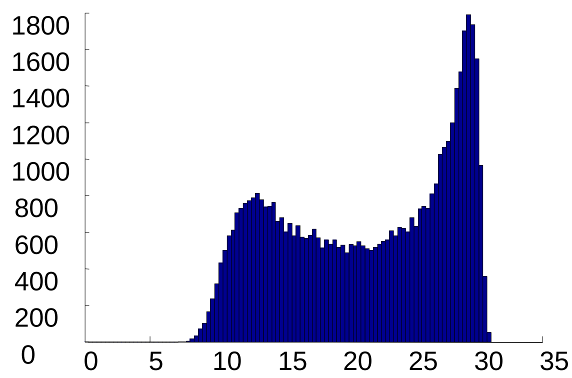
$$\max(\ |\ d(P,R_i) - d(Q,R_i)\ |\ )$$

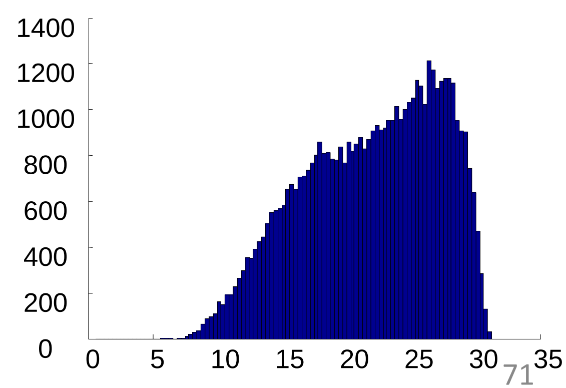


Time to find the motif in a database of 64,000 random walk time series of length 1,024









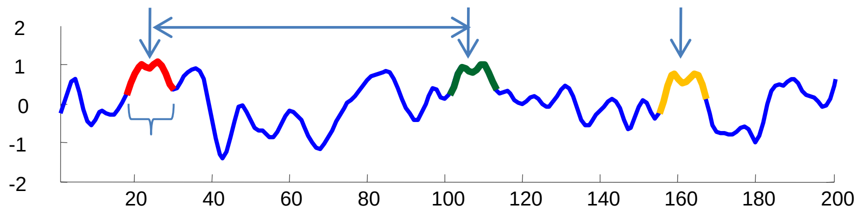
Algorithms Outline

- Algorithms
  - Definition, Distance Measures and Invariances
  - Exact Algorithms
    - Fixed Length
    - Enumeration of All length
    - K-motif Discovery
    - Online Maintenance
  - Approximate Algorithms
    - Random Projection Algorithm
  - Multi-dimensional Motif Discovery
  - Open Problems

Questions and Comments

~~Simplest~~ Definition of Time Series Motifs

The most similar/least distant pairs of non-overlapping subsequences at all lengths.



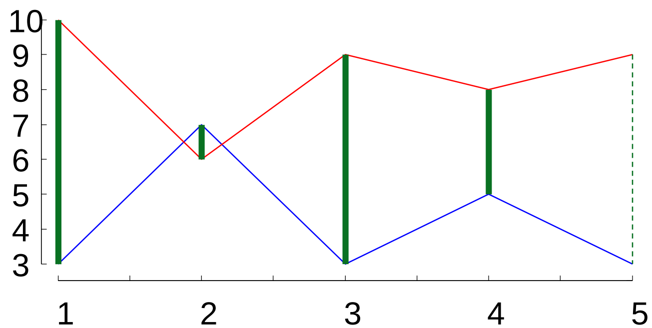
1. Length of the motif = ~~Given~~ **All**
2. Support of the motif = **2**
3. Similarity of the Pattern = **Euclidean distance**
4. Relative Position of the Pattern = **non-overlapping**

Goals: Enumerating Motifs

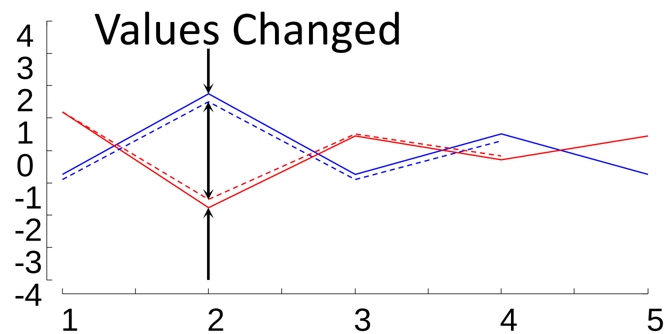
1. Remove the length parameter
2. Search for motifs in a range of lengths and report
   - **ALL** of the motifs of all of the lengths
3. Retain Scalability

Bound on Extension

1. Two time series $\boldsymbol{x}$ and $\boldsymbol{y}$ of length m
2. Their normalized Euclidean distance $d(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$
3. Find $d_{LB}(\hat{\boldsymbol{x}}_{+1}, \hat{\boldsymbol{y}}_{+1})$ if we increase the length of $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ by appending the next two numbers.



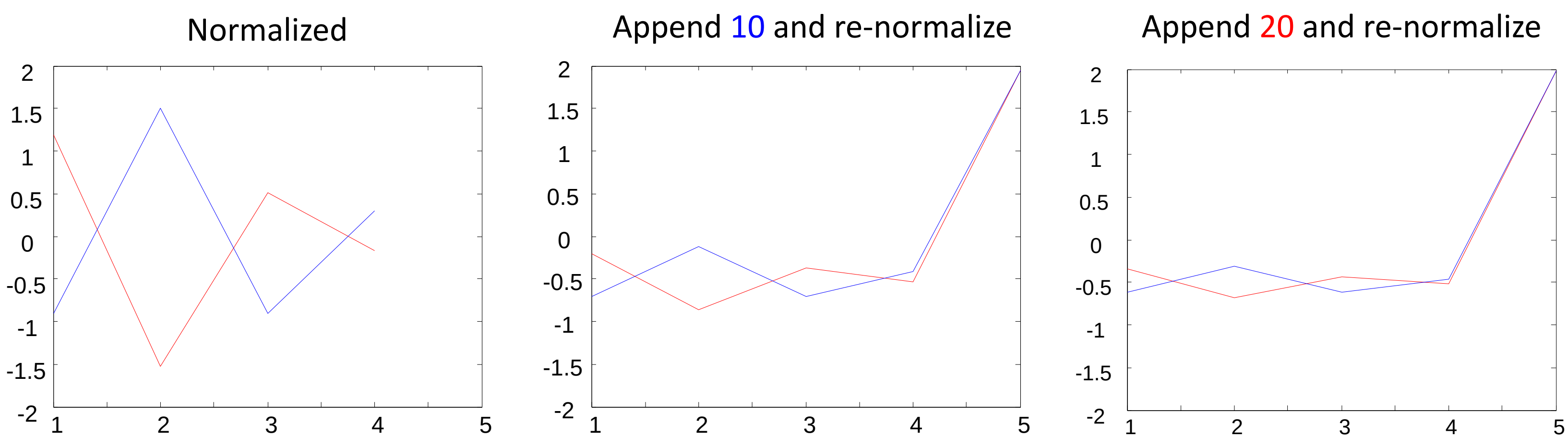
Without Normalization



With Normalization

Intuition

Area between blue and red is
the distance between the signals



Length 4

Length 5
Area shrinks

Length 5
Area shrinks further

If infinity is appended to both the signals, they will have zero area/distance.

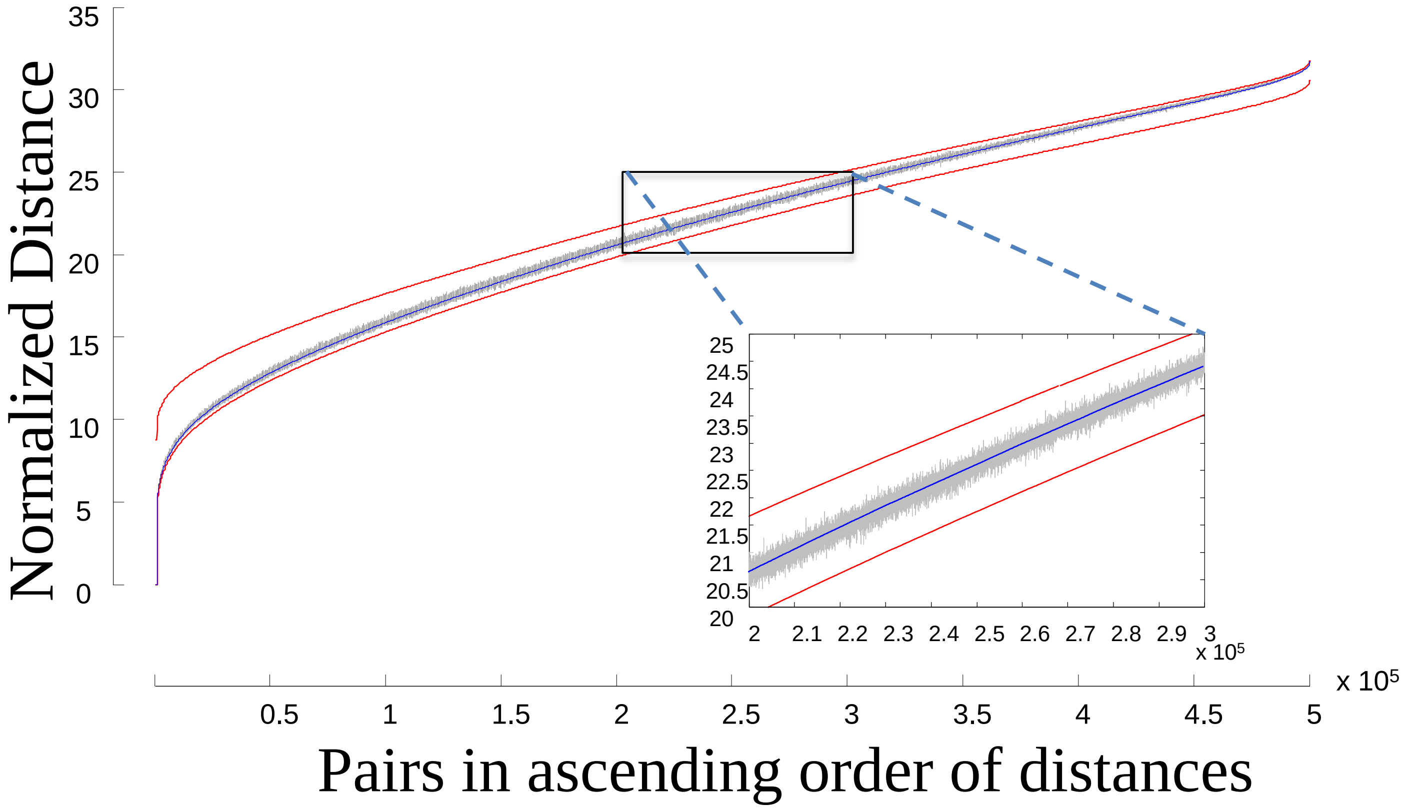
Bounding Euclidean Distance

$$d_{LB}^2(\widehat{\boldsymbol{x}}_{+1}, \widehat{\boldsymbol{y}}_{+1}) = \frac{1}{\sigma_m^2} d_m^2(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) < d_m^2(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})$$

Variances of $\widehat{\boldsymbol{x}}_{+1}$ and $\widehat{\boldsymbol{y}}_{+1}$, $\sigma_m^2 = \frac{m}{m+1} + \frac{m}{(m+1)^2} z^2$

$z =$ maximum normalized value in the database
A safe approximation $z = \max(abs(\widehat{\boldsymbol{x}}), abs(\widehat{\boldsymbol{y}}))$

Experimental Validation of the Bounds



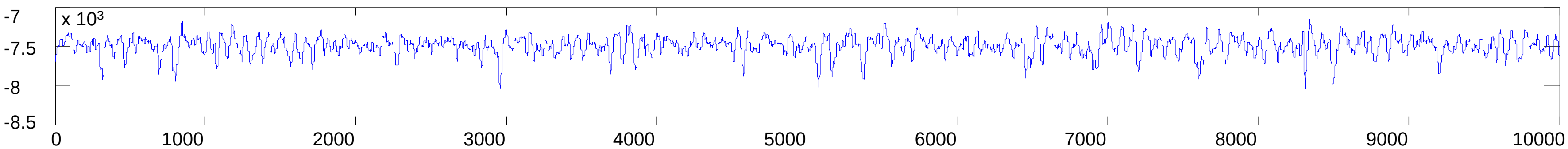
Blue: Distances of random signals of length 255

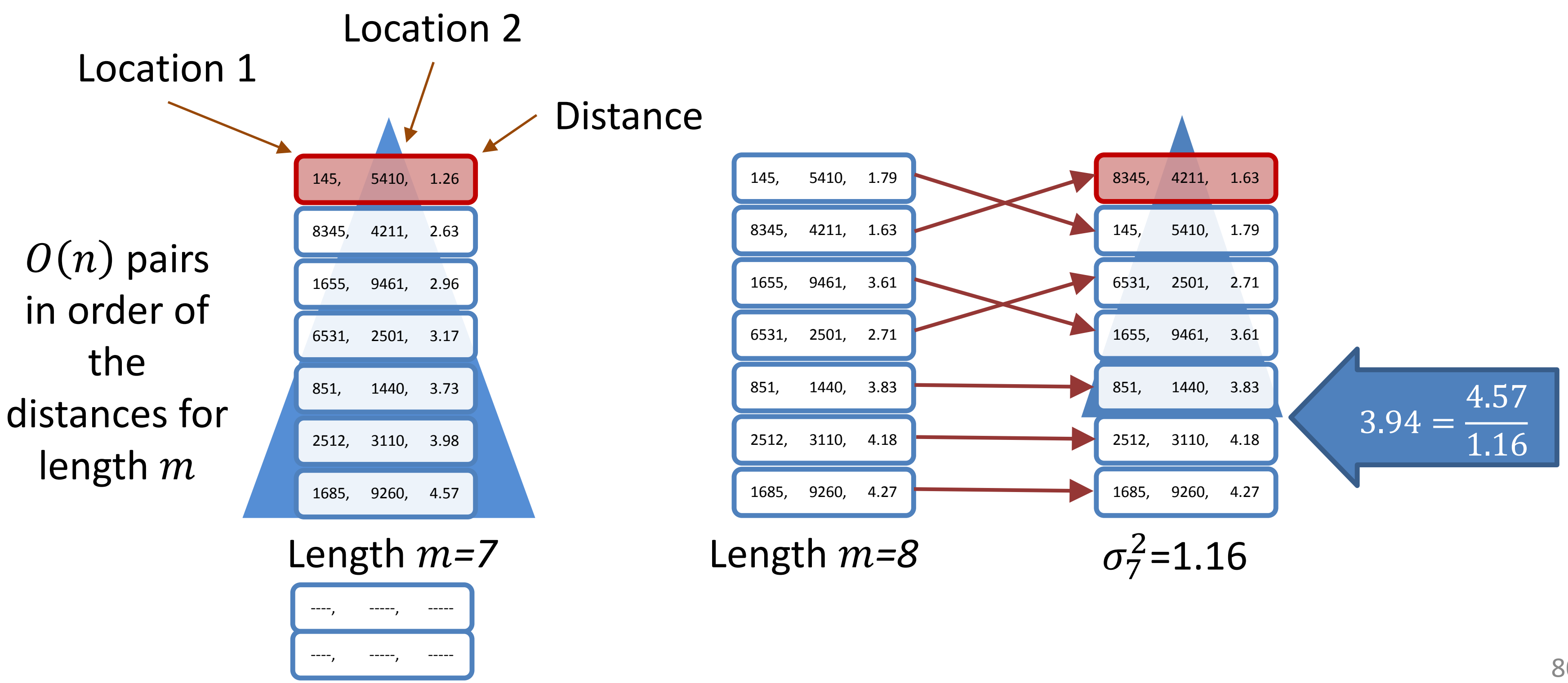
Gray: Distances of the signals when they are extended by one random sample

Red: The upper and lower bounds before observing the extensions

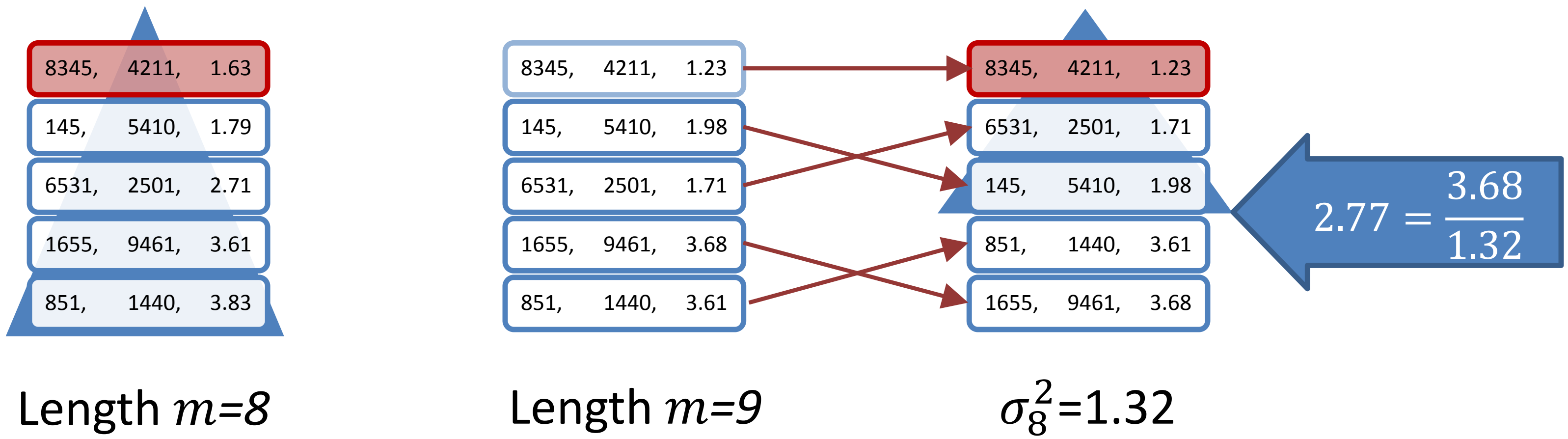
Intuition

n = 10000



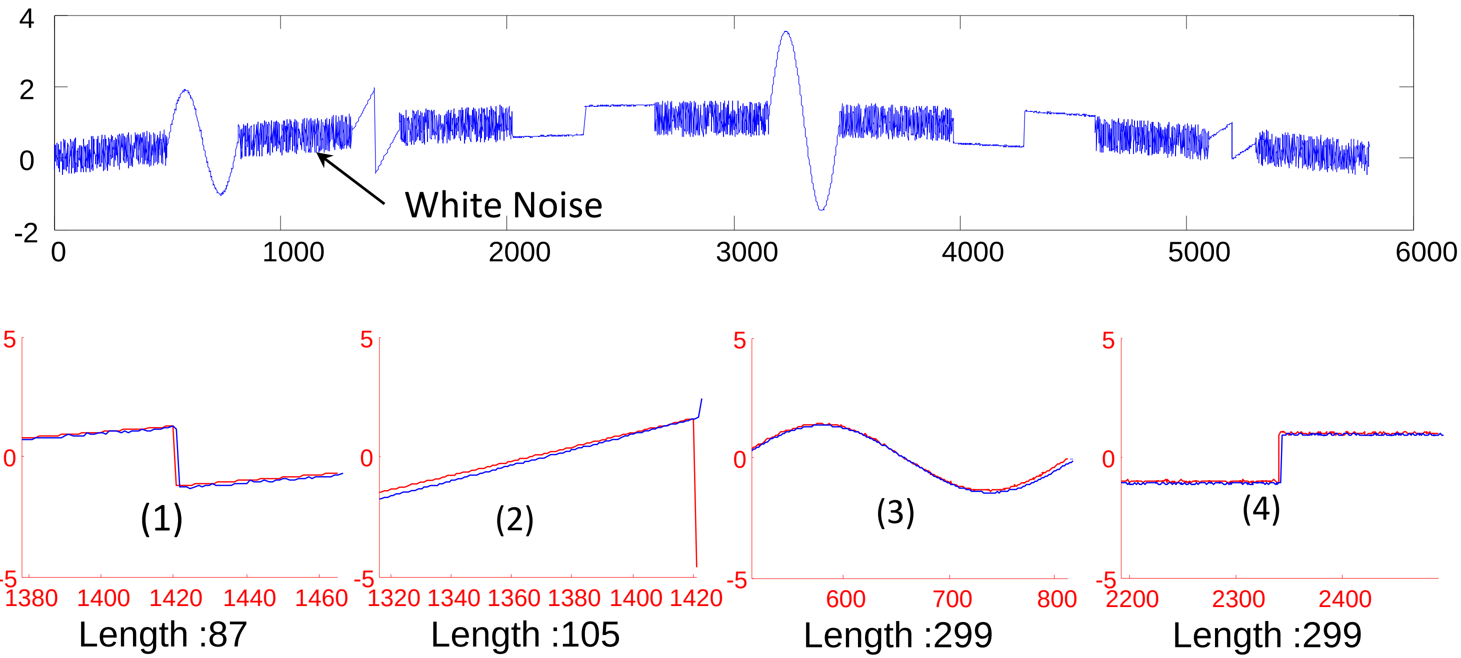


Intuition



- Once in every 10 lengths, the exact ordered list is required to be populated.
- This yields a 10x speed-up from running fixed-length motif discovery for all lengths.

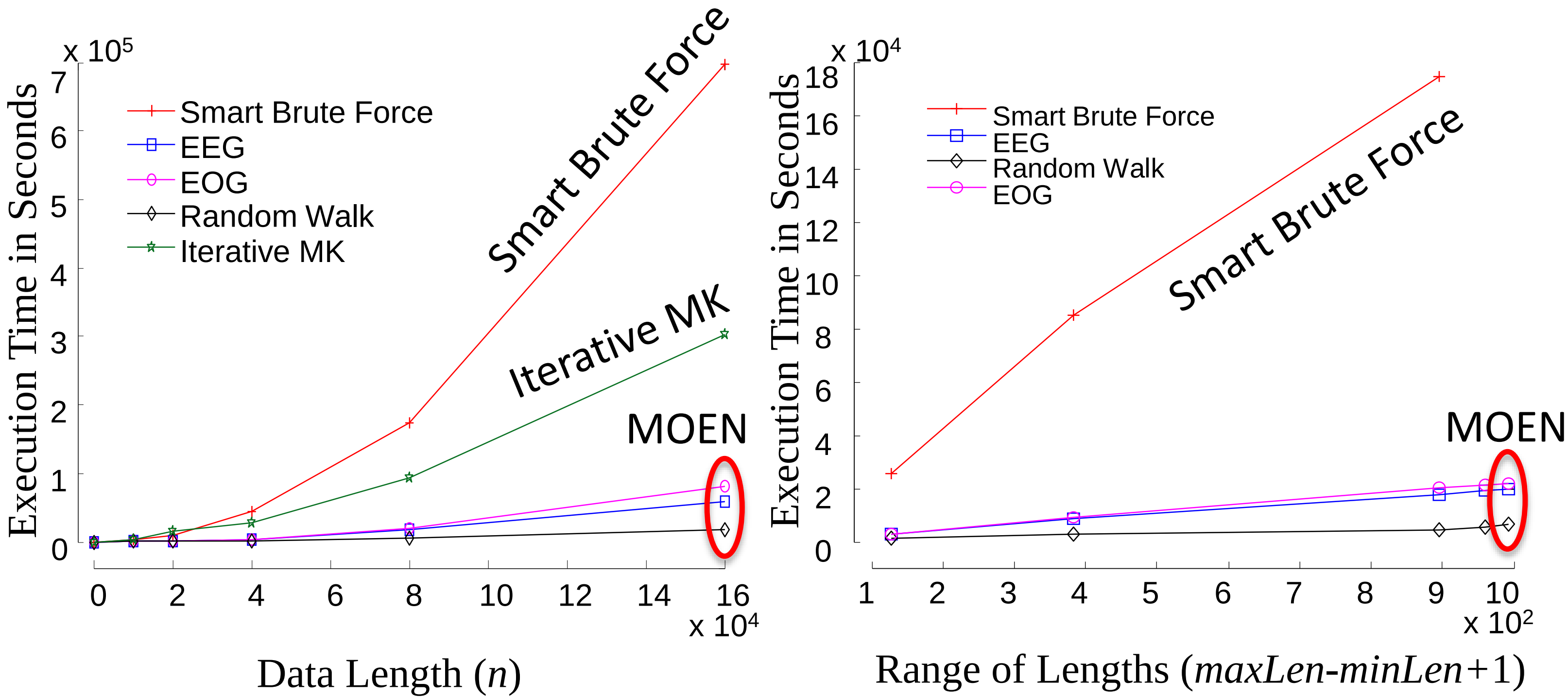
Sanity Check



- Three Patterns planted in a random signal with different scaling.
- The algorithm finds them appropriately.

http://www.cs.unm.edu/~mueen/Projects/MOEN/index.html

Experimental Results: Scalability

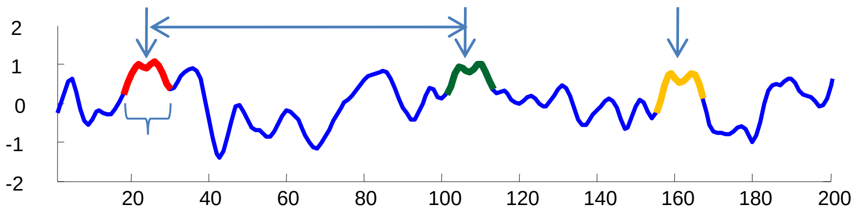


Algorithms Outline

- Algorithms
  - Definition, Distance Measures and Invariances
  - Exact Algorithms
    - Fixed Length
    - Enumeration of All length
    - K-motif Discovery
    - Online Maintenance
  - Approximate Algorithms
    - Random Projection Algorithm
  - Multi-dimensional Motif Discovery
  - Open Problems

~~Simplest~~ Definition of Time Series Motifs

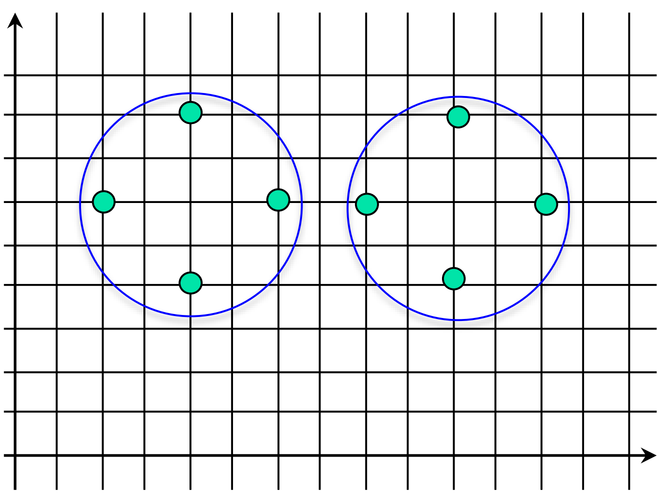
The non-overlapping subsequences at all lengths having k or more τ-matches.

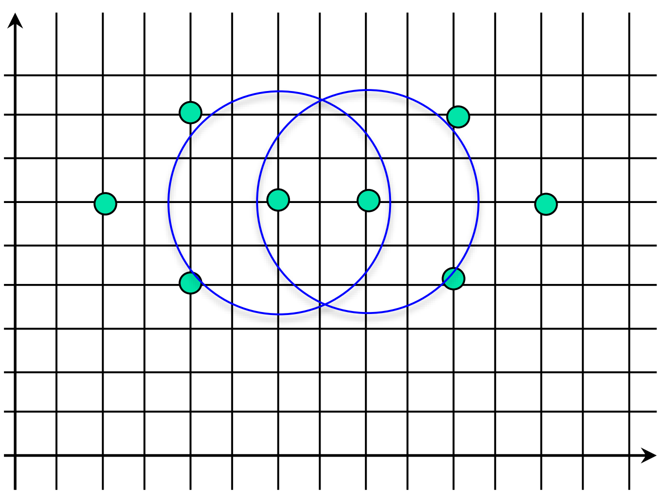


1. Length of the motif = ~~**Given**~~ **All**
2. Support of the motif = ~~**2**~~ **k and τ**
3. Similarity of the Pattern = **Euclidean distance**
4. Relative Position of the Pattern = **non-overlapping**

Optimal algorithm is hard

- Search for locations of the τ-balls that contain k subsequences
- NP-Hard
- Instead we search for a motif representative that has k subsequences within τ

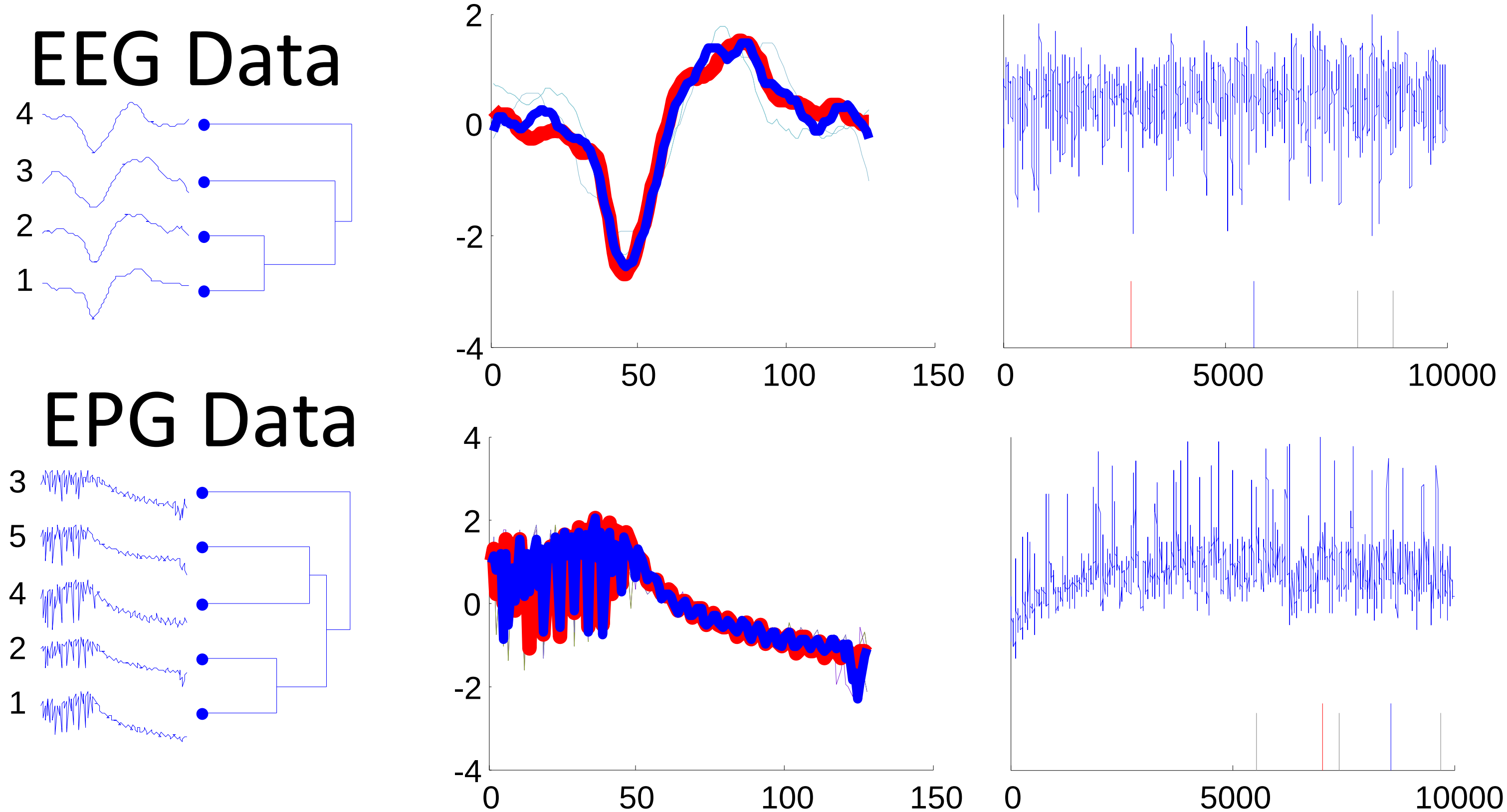




How do we find the motif representative?

- Simply take one of the two occurrences as the representative
- Take the average of the two
- Find all occurrences within a threshold of pair and train a HMM to capture the concept (Minnen'07)

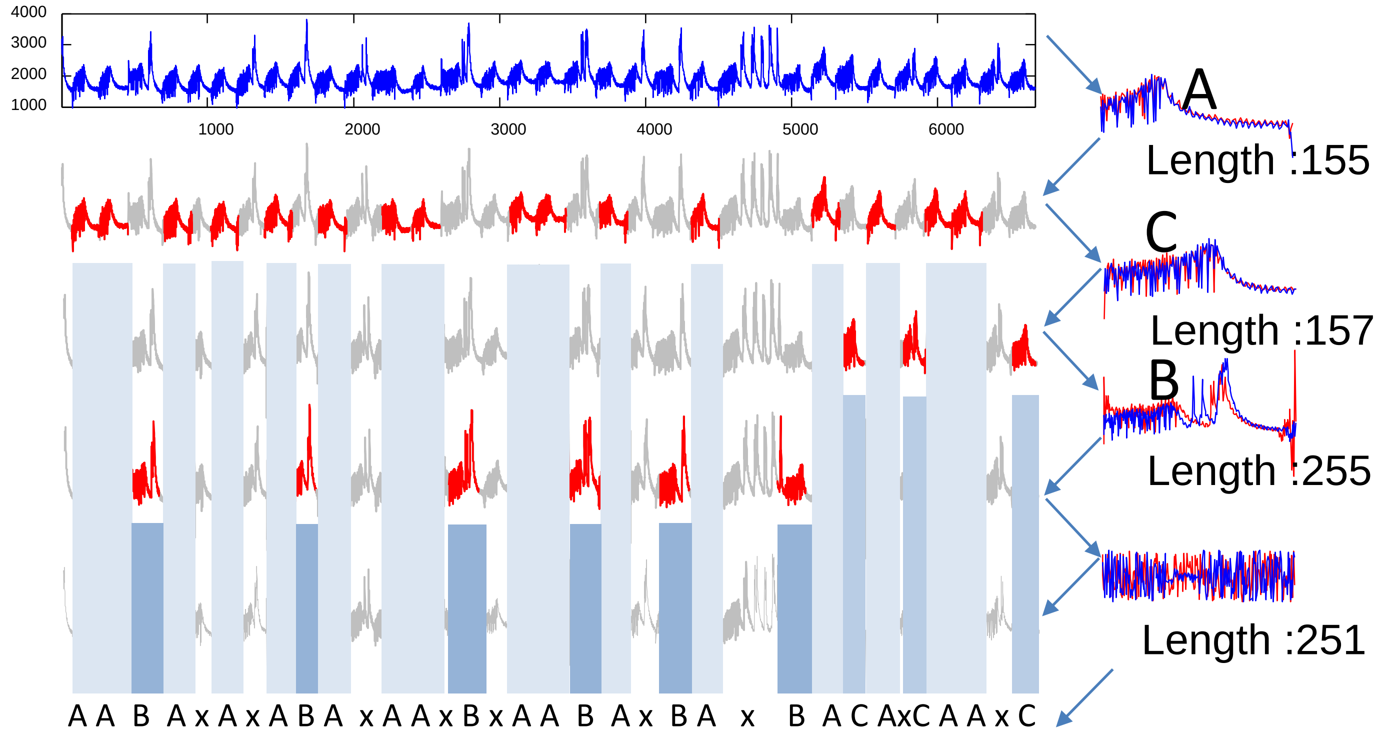
Using each one in the pair as the representative (k=4, τ=0.9)



It takes only k similarity searches to find other occurrences. The overall complexity remains the same.

Finding top-K motif

- Run MK for K times
- Replace occurrences by random noise between iterations

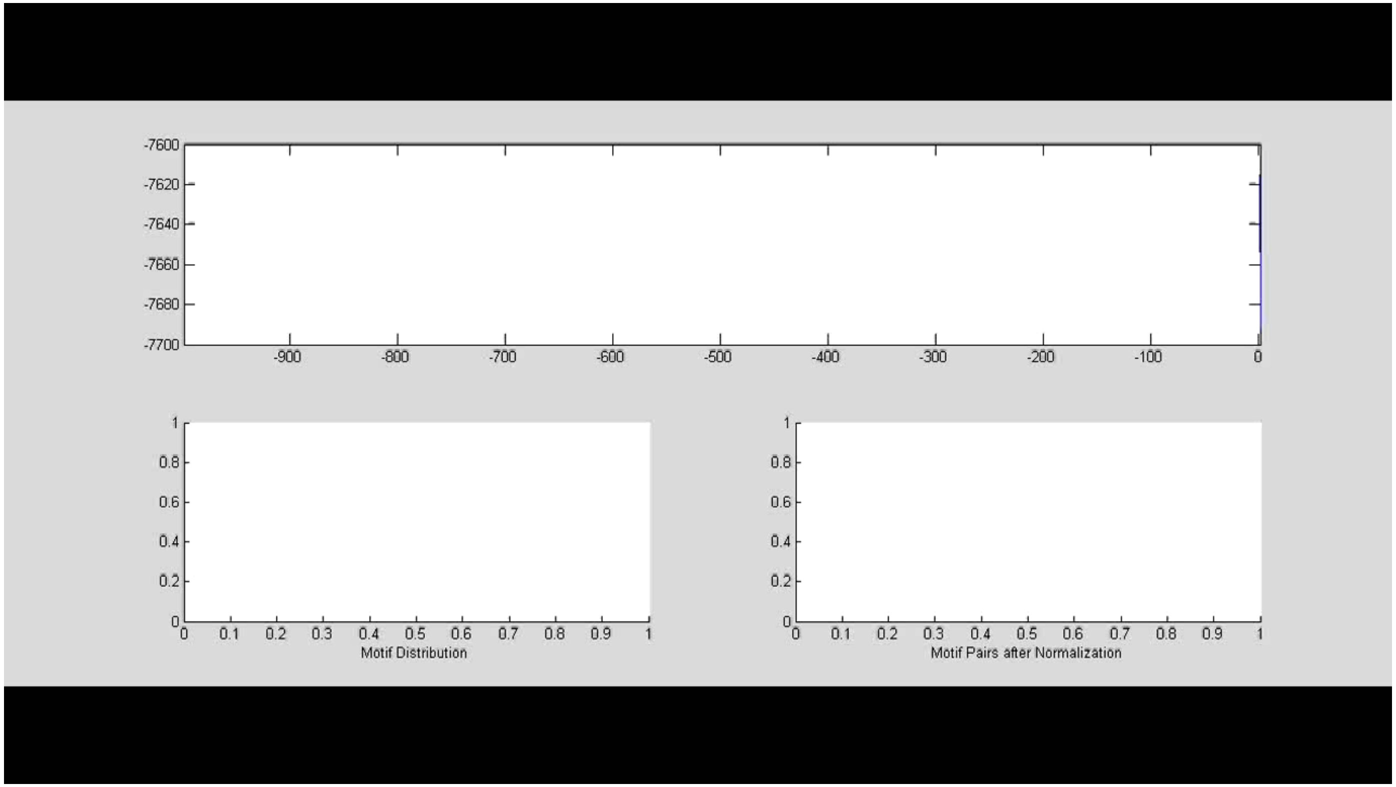


Algorithms Outline

- Algorithms
  - Definition, Distance Measures and Invariances
  - Exact Algorithms
    - Fixed Length
    - Enumeration of All length
    - K-motif Discovery
    - Online Maintenance
  - Approximate Algorithms
    - Random Projection Algorithm
  - Multi-dimensional Motif Discovery
  - Open Problems

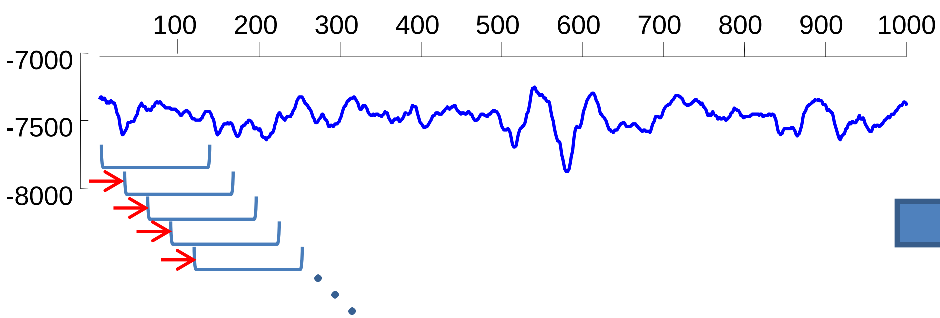
Online Time Series Motifs

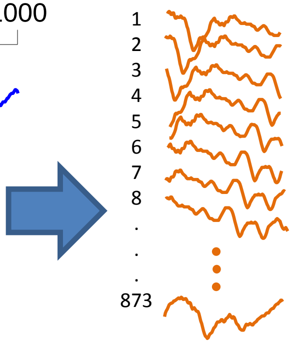
- Streaming time series
- Sliding window of the recent history
  - What minute long trace repeated in the last hour?



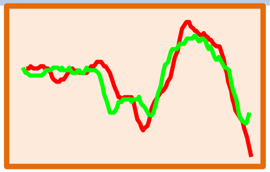
Problem Formulation

Discovery



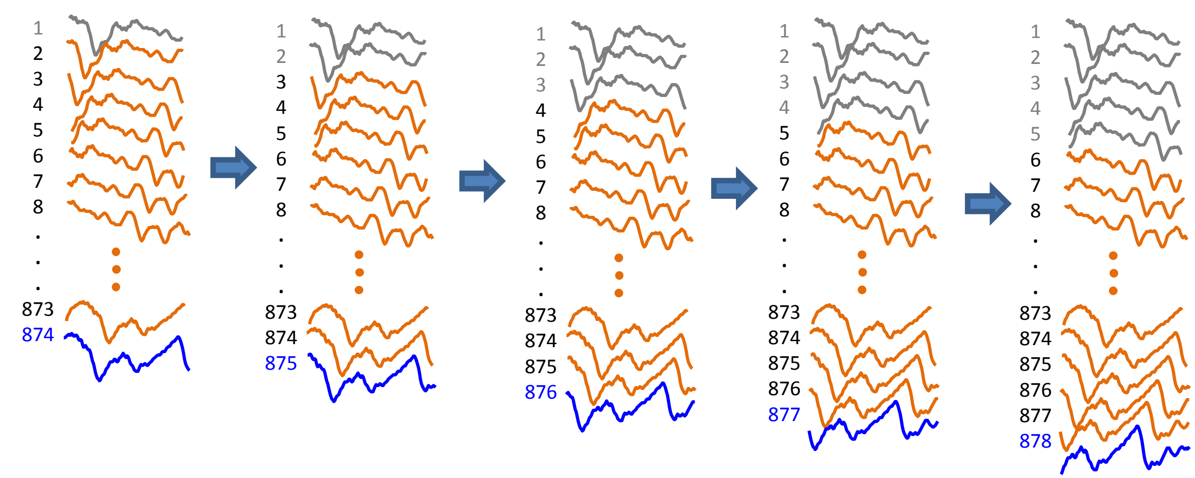


The most similar pair of non-overlapping subsequences

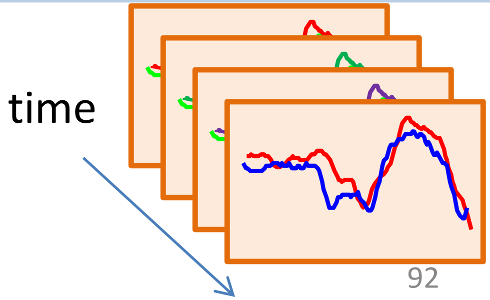


Maintenance

time:1001 time:1002 time:1003 time:1004 time:1005

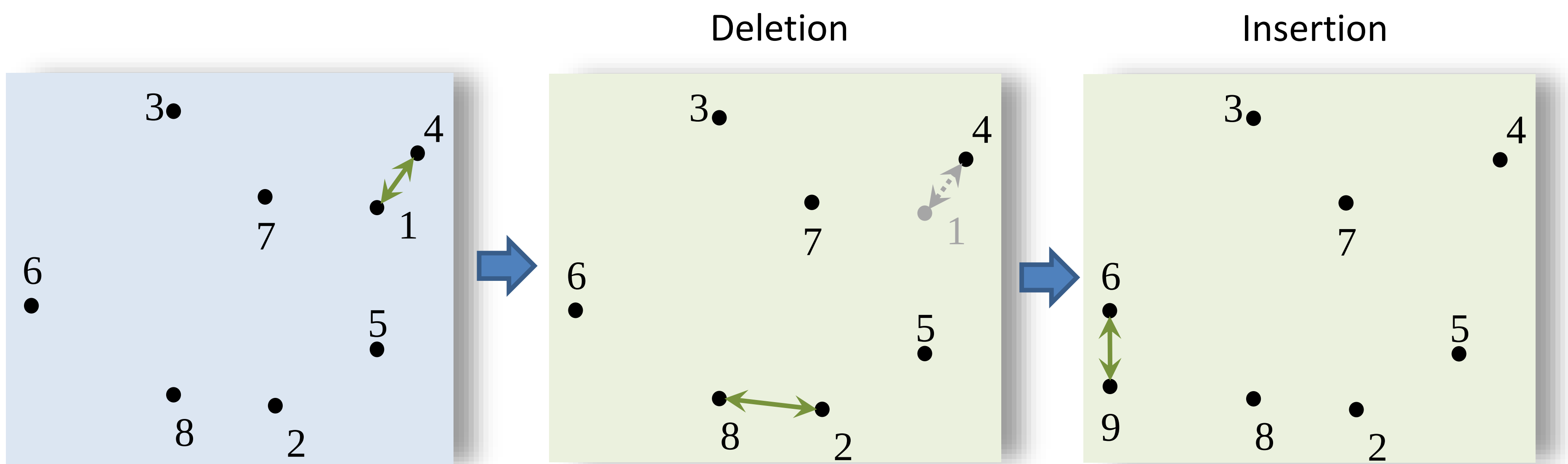


Update motif pair after every time tick



Challenges

- A subsequence is a high dimensional point
  - The dynamic closest pair of points problem
- Closest pair may change upon every update
- Naïve approach: Do quadratic comparisons.

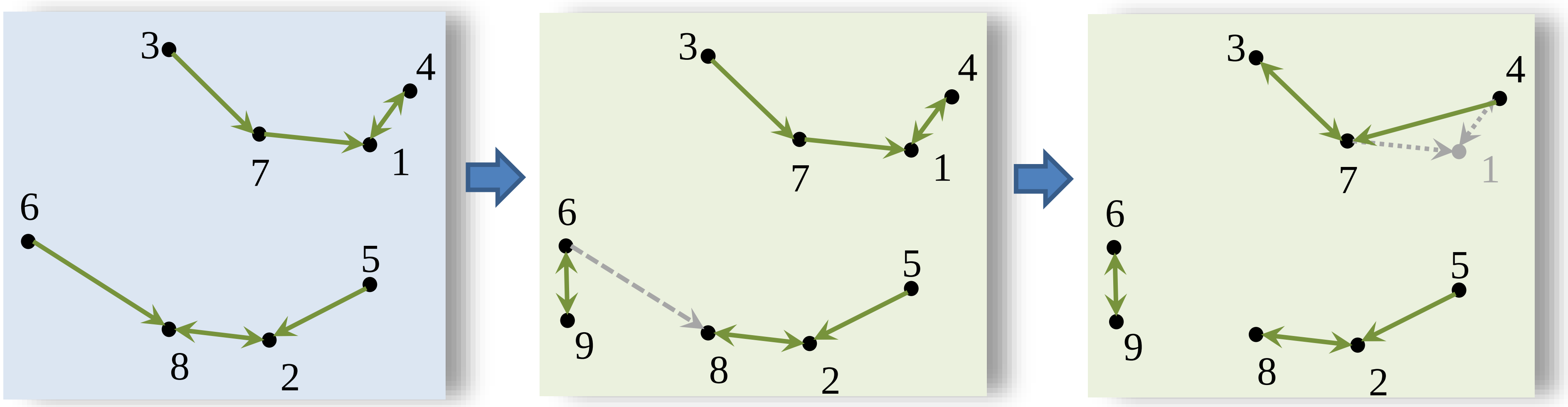


Related Work

- Goal: Algorithm with Linear update time
- Previous method for dynamic closest pair (Eppstein,00)
  - A matrix of all-pair distances is maintained
    - O(w^2) space required
  - Quad-tree is used to update the matrix
- Maintain a set of neighbors and reverse neighbors for all points
- We do it in O($w\sqrt{w}$) space

Maintaining Motif

- Smallest nearest neighbor → Closest pair
- Upon insertion
  - Find the nearest neighbor; Needs O(w) comparisons.
- Upon deletion
  - Find the next NN of all the reverse NN



Insertion

Deletion

Data Structure

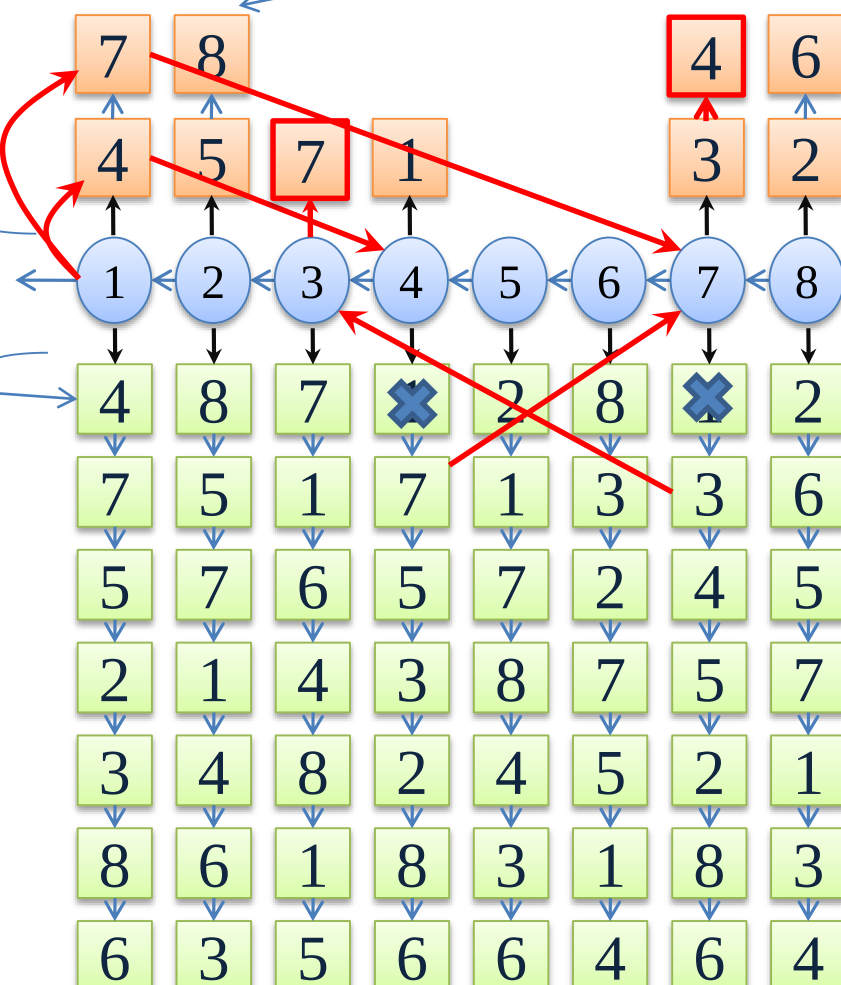
Total number of nodes in R-lists is *w*

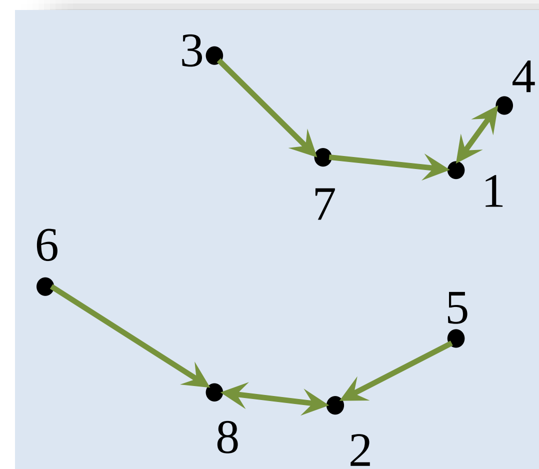
R-lists
Points that should be updated

FIFO Sliding window

(*ptr, distance*)

N-lists
Neighbors in order of distances

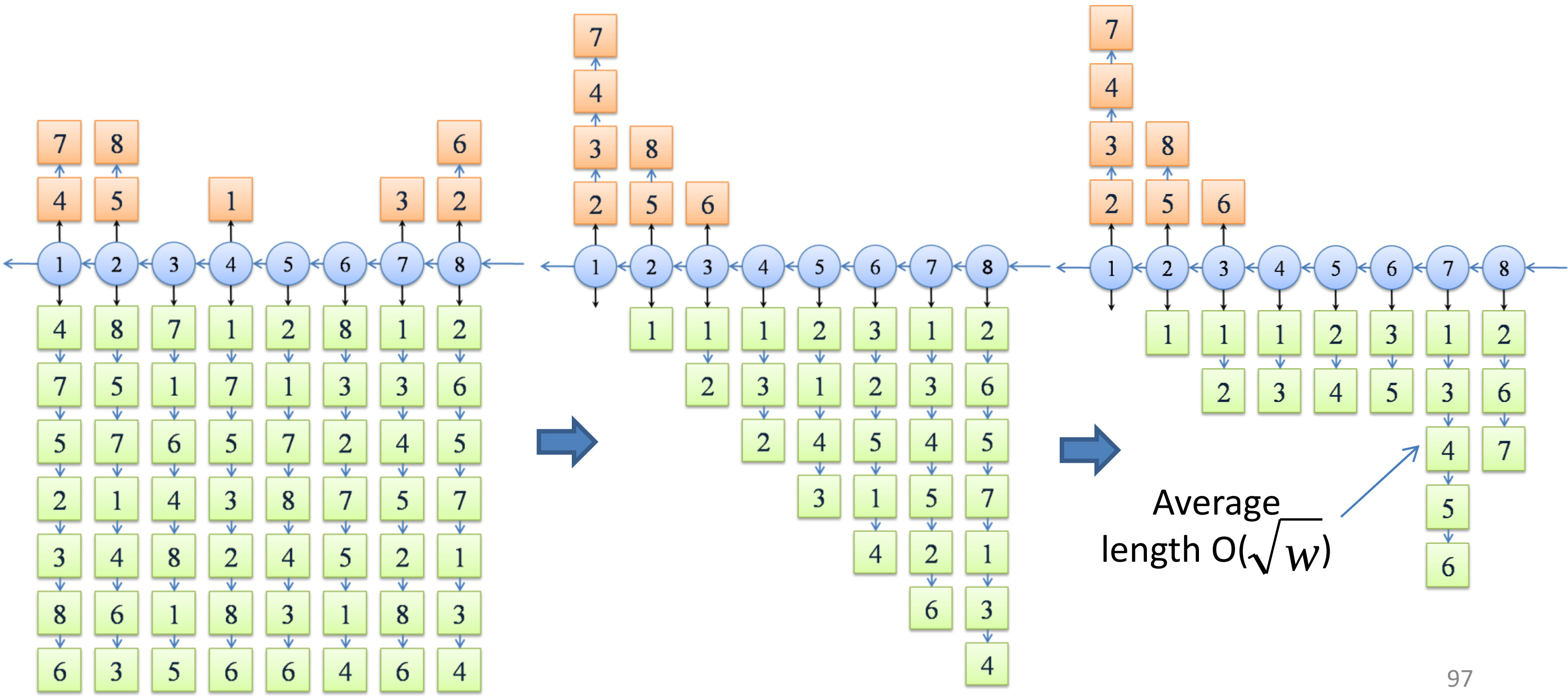




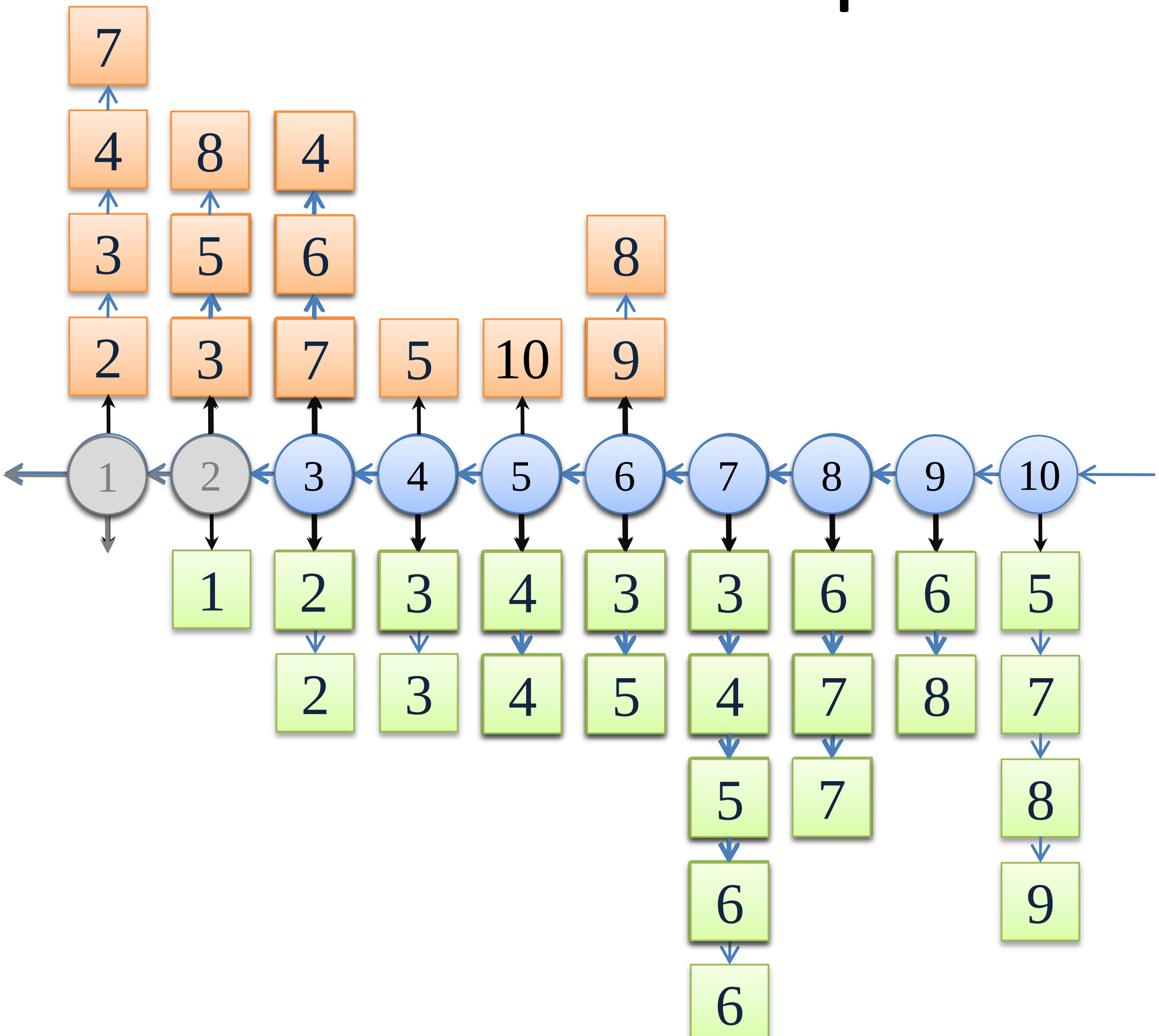
Observations

While inserting

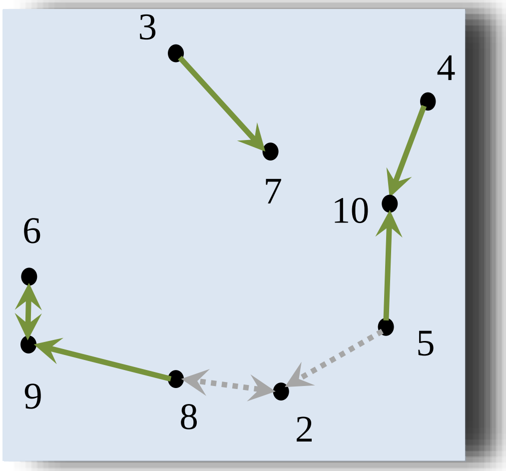
- ➢ Updating NN of old points is not necessary
- ➢ A point can be removed from the neighbor list if it violates the temporal order



Example

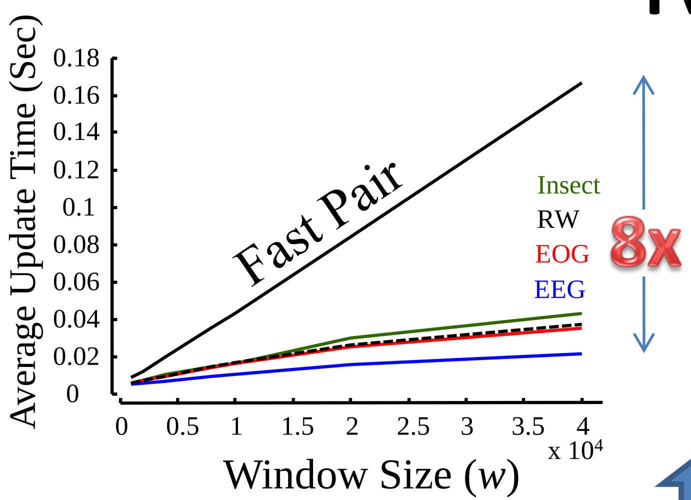


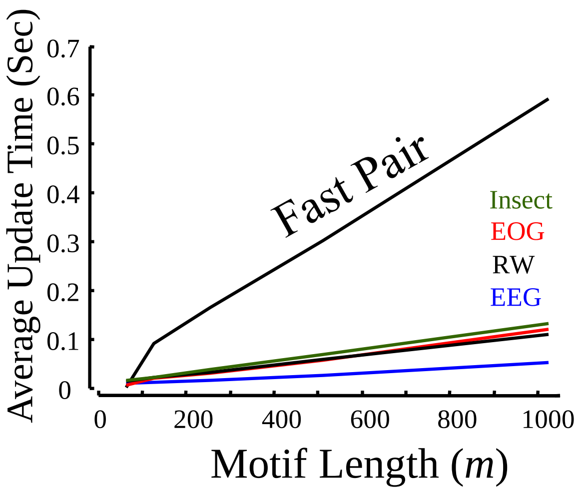
7
4
8
4
3
5
6
8
2
3
7
5
10
9
1
2
3
4
5
6
7
8
9
10
1
2
3
4
3
3
6
6
5
2
3
4
5
4
7
8
7
5
7
8
6
9
6



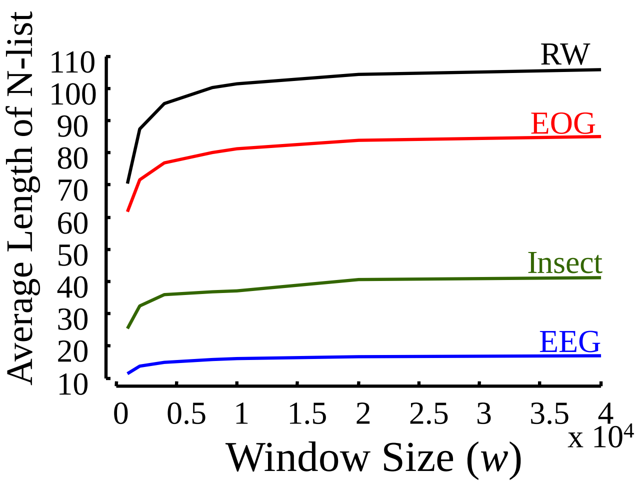
3
4
7
10
6
5
9
8
2

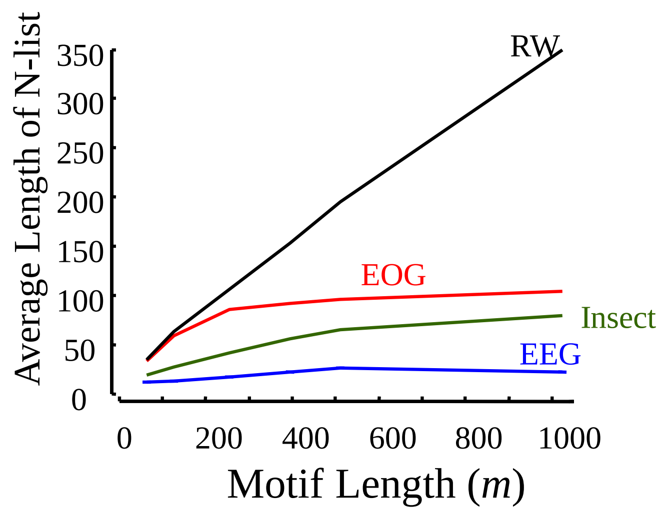
Results





- Up to **8x speedup** from general dynamic closest pair
- **Stable space cost** per point with increasing window size





Algorithms Outline

- Algorithms
  - Definition, Distance Measures and Invariances
  - Exact Algorithms
    - Fixed Length
    - Enumeration of All length
    - K-motif Discovery
    - Online Maintenance
  - Approximate Algorithms
    - Random Projection Algorithm
  - Multi-dimensional Motif Discovery
  - Open Problems

Questions and Comments

Finding Repeated Structure in Time Series: Algorithms and Applications

Break; We meet back in this room at 5:15PM

Abdullah Mueen
University of New Mexico, USA
Eamonn Keogh
University of California Riverside, USA

Slides available at: http://www.cs.unm.edu/~mueen/Tutorial/SDM2015Tutorial2.pdf

General Outline

- Applications (50 minutes)
  - As Subroutines in Data Mining
  - In Other Scientific Research
- Algorithms (100 minutes)
  - Uni-dimensional
  - Multi-dimensional

Algorithms Outline

- Algorithms
  – Definition, Distance Measures and Invariances
  – Exact Algorithms
    - Fixed Length
    - Enumeration of All length
    - K-motif Discovery
    - Online Maintenance
  – Approximate Algorithms
    - Random Projection Algorithm
  – Multi-dimensional Motif Discovery
  – Open Problems

How do we find approximate motif in a time series?

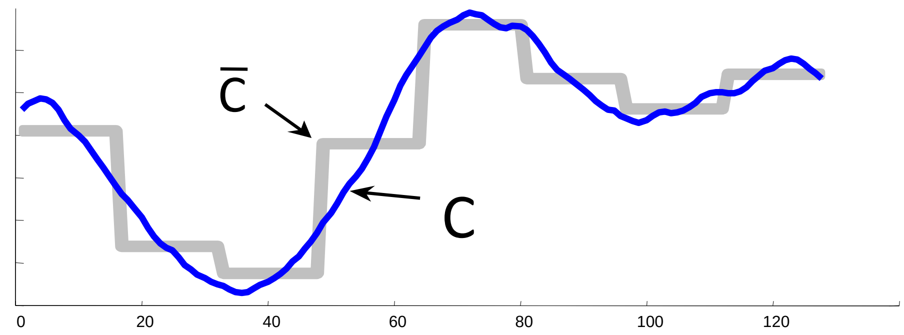
The obvious brute force search algorithm is just too slow…

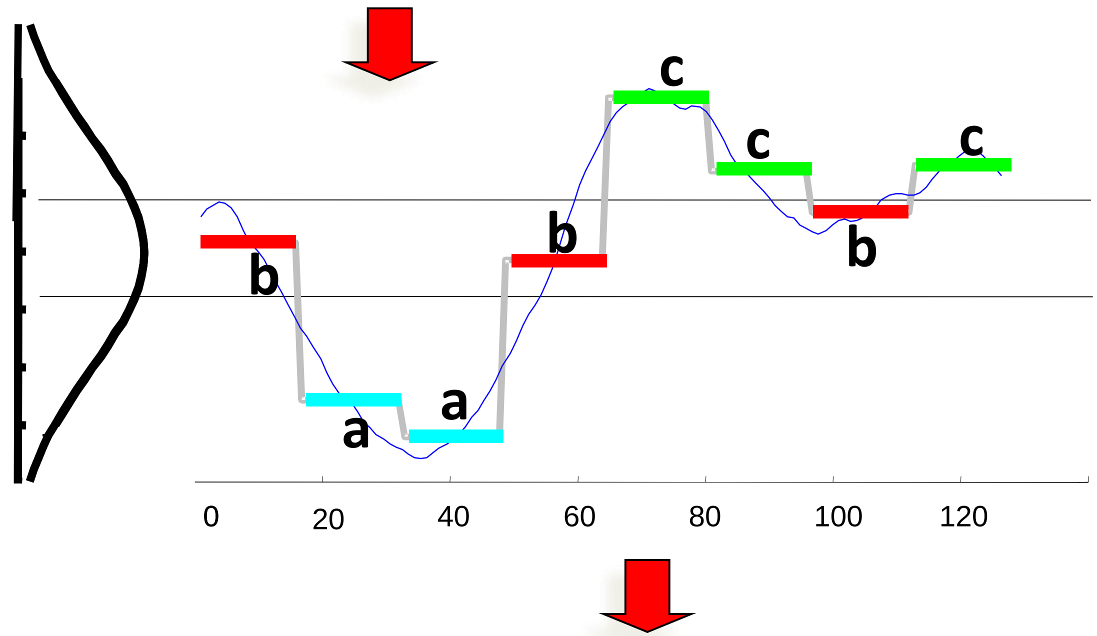
Our algorithm is based on a *hot* idea from bioinformatics, *random projection** and the fact that SAX allows us to lower bound discrete representations of time series.

*** J Buhler and M Tompa. *Finding motifs using random projections*. In RECOMB'01. 2001.**

Symbolic **A**ggregate Appro**X**imation

How do we obtain SAX?

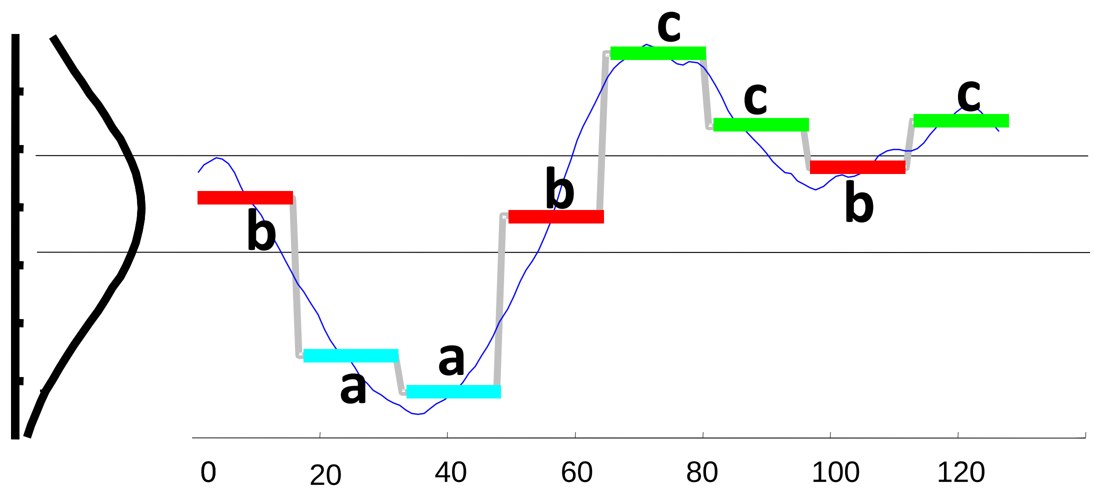




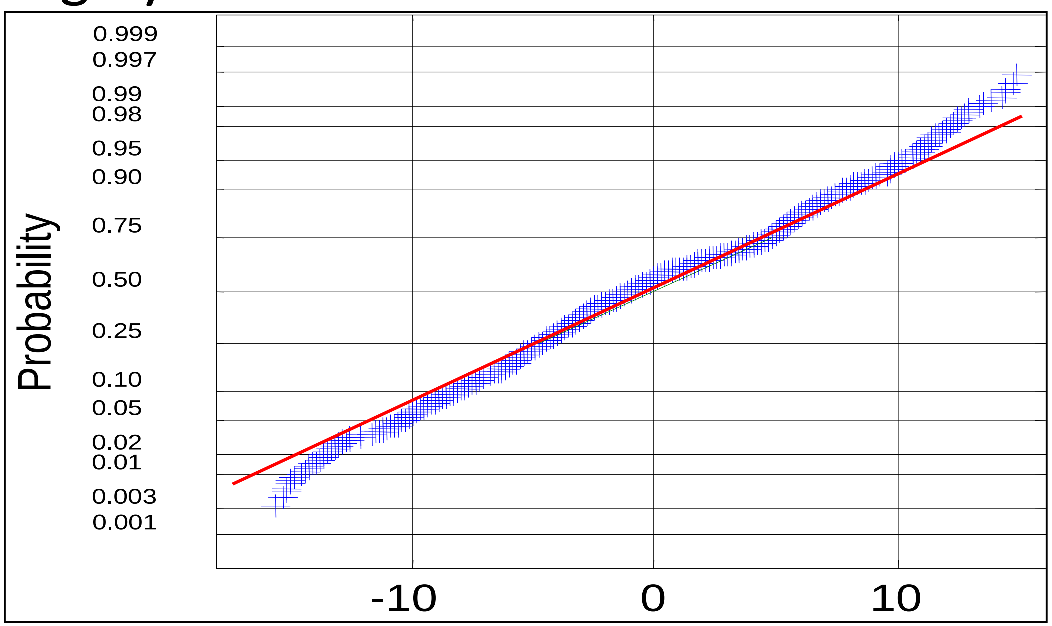
First convert the time series to PAA representation, then convert the PAA to symbols

It take linear time

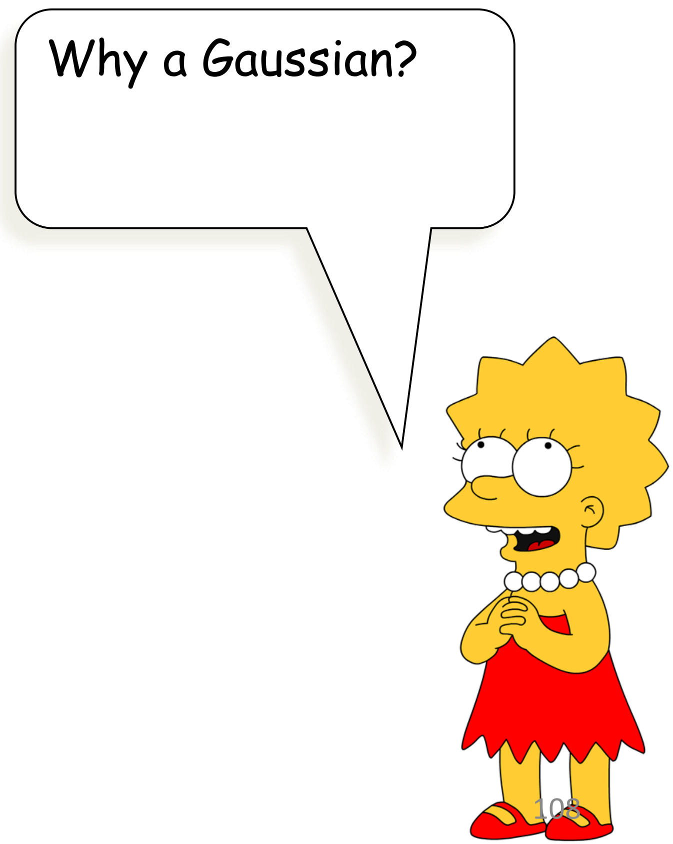
baabccbc



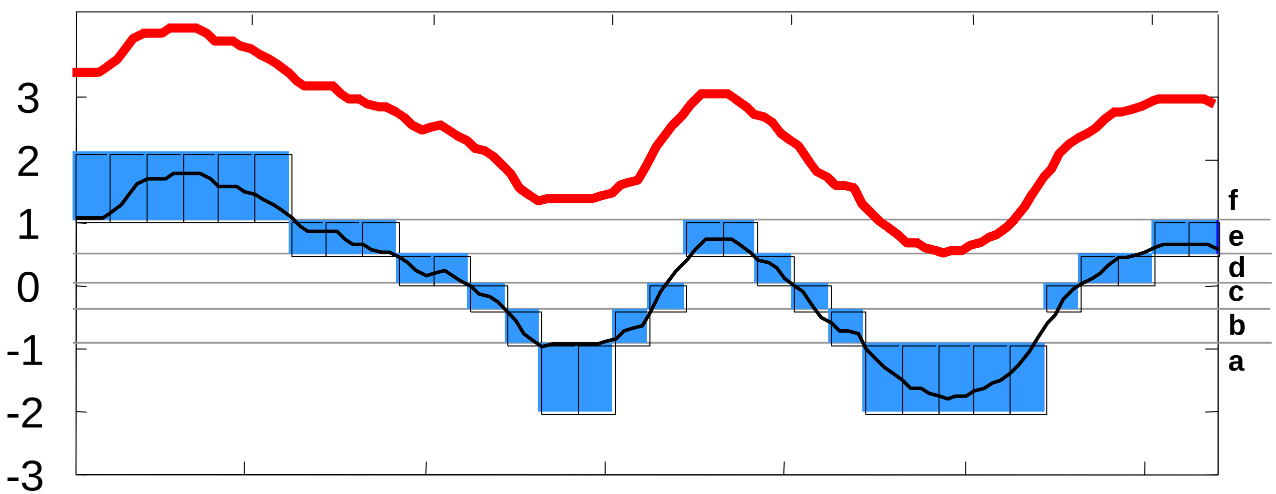
Time series subsequences tend to have a highly Gaussian distribution

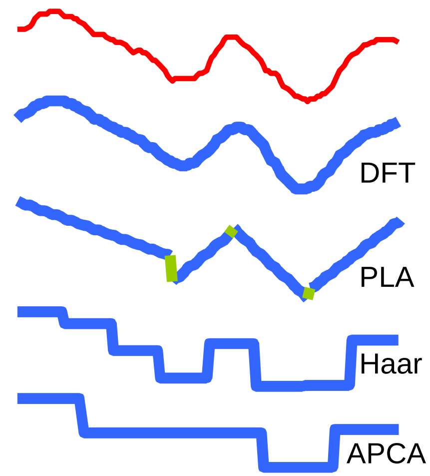


A normal probability plot of the (cumulative) distribution of values from subsequences of length 128.



Visual Comparison



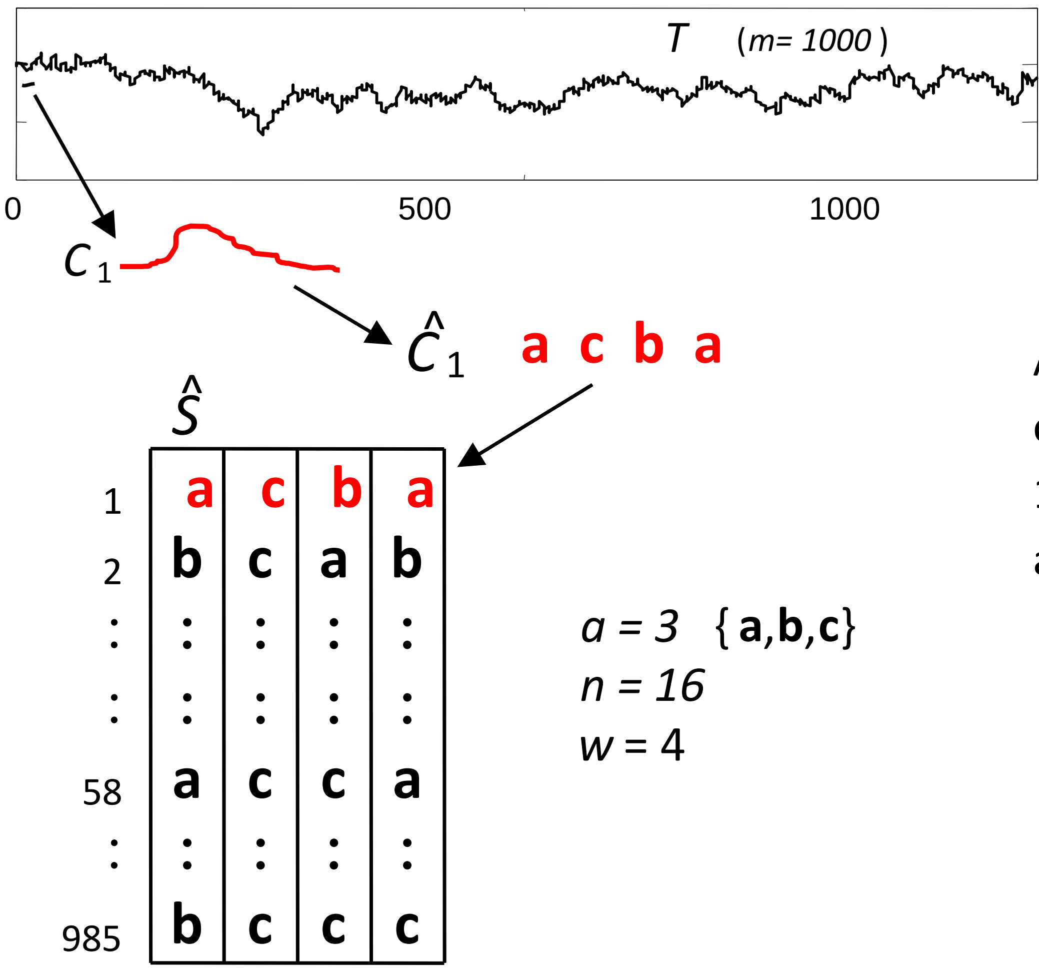


A raw time series of length 128 is transformed into the word "**ffffffeeeddcbaabceedcbaaaacddee**."

– We can use more symbols to represent the time series since each symbol requires fewer bits than real-numbers (float, double)

A simple worked example of approximate motif discovery algorithm

The next 3 slides

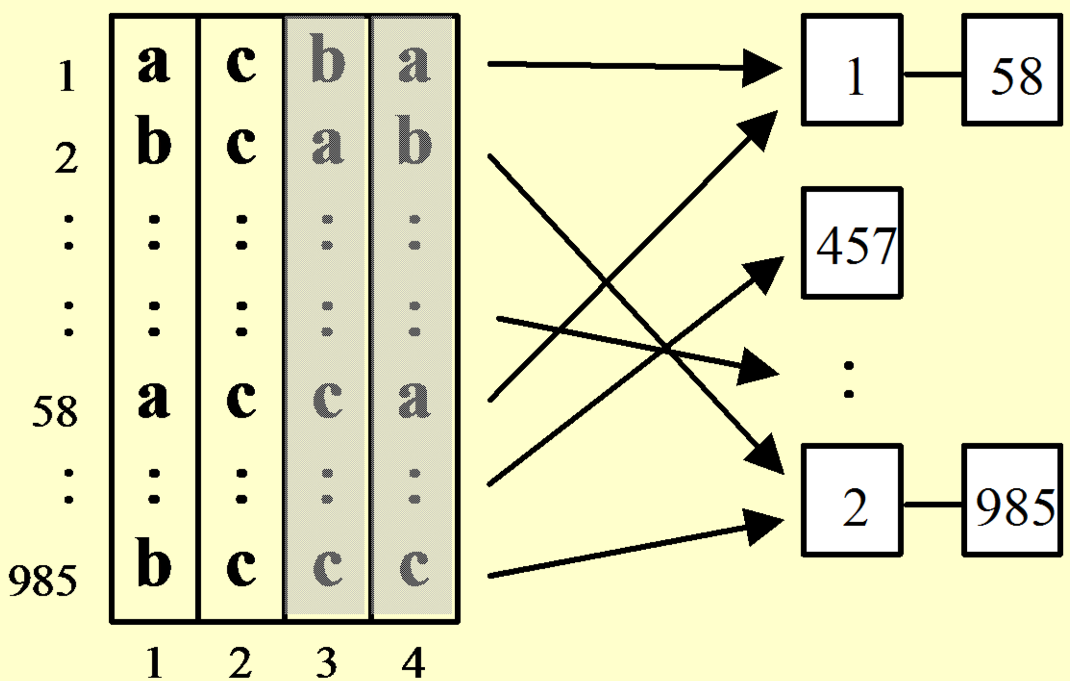


$a = 3$ {**a**,**b**,**c**}
$n = 16$
$w = 4$

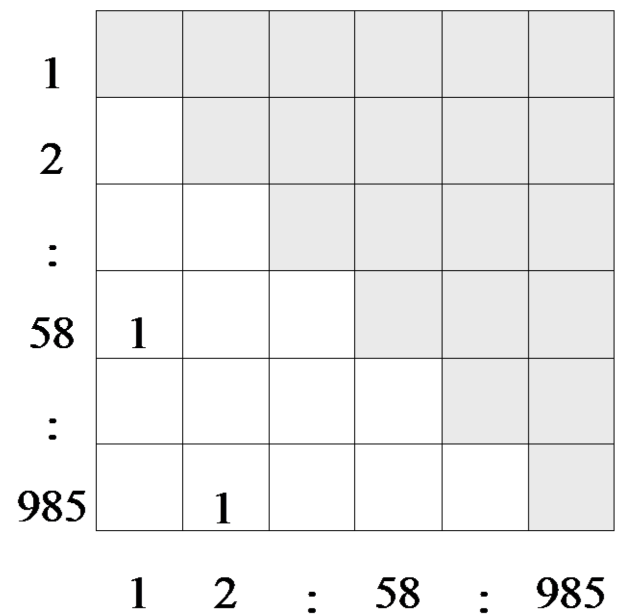
Assume that we have a time series T of length 1,000, and a motif of length 16, which occurs twice, at time T_1 and time T_{58}.

A simple worked example of approximate motif discovery algorithm

A mask {1,2} was randomly chosen, so the values in columns {1,2} were used to project matrix into buckets.

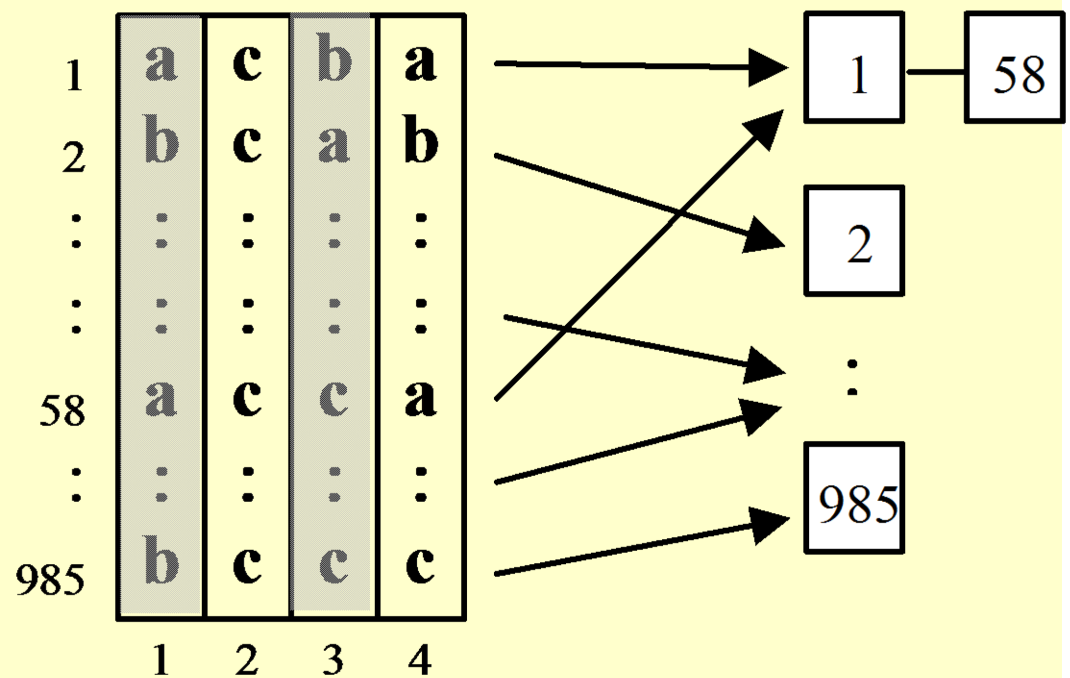


Collisions are recorded by incrementing the appropriate location in the collision matrix

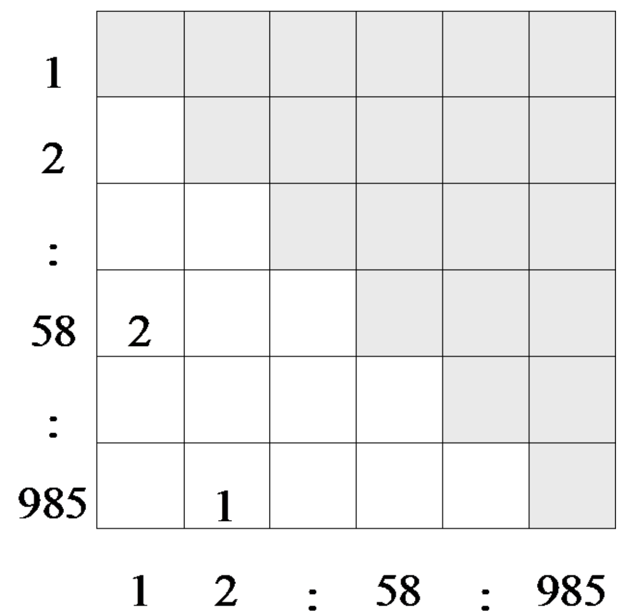


A simple worked example of approximate motif discovery algorithm

A mask {2,4} was randomly chosen, so the values in columns {2,4} were used to project matrix into buckets.



Once again, collisions are recorded by incrementing the appropriate location in the collision matrix



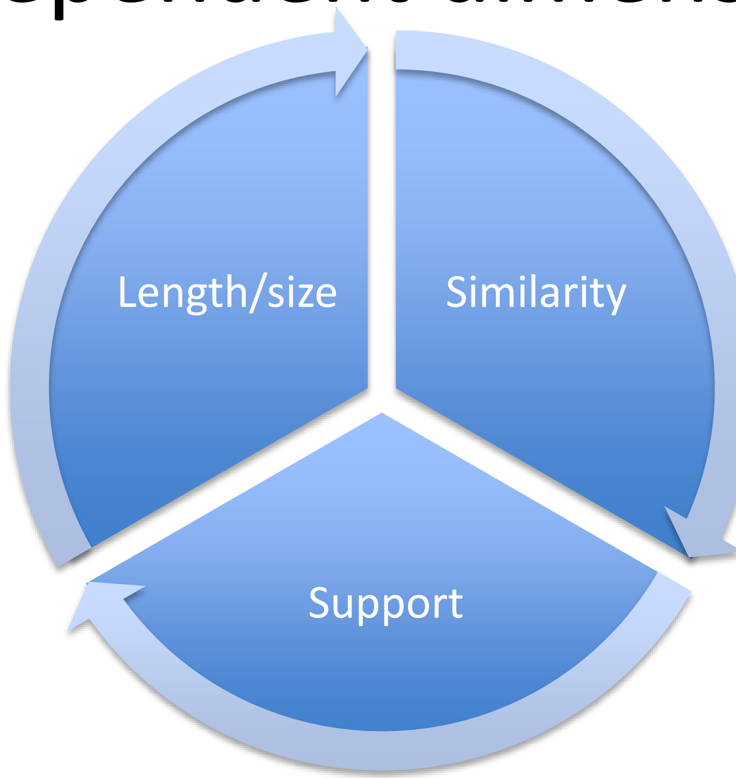
We can calculate the expected values in the matrix, assuming there are NO patterns…

$$E(k,a,w,d,t) = \binom{k}{2} \sum_{i=0}^{d} \left(1 - \frac{i}{w}\right)^{t} \binom{w}{i} \left(\frac{a-1}{a}\right)^{i} \left(\frac{1}{a}\right)^{w-i}$$

t is the length of the projected string. We conclude that if we have k random strings of size w, an entry of the similarity matrix will be hit on average

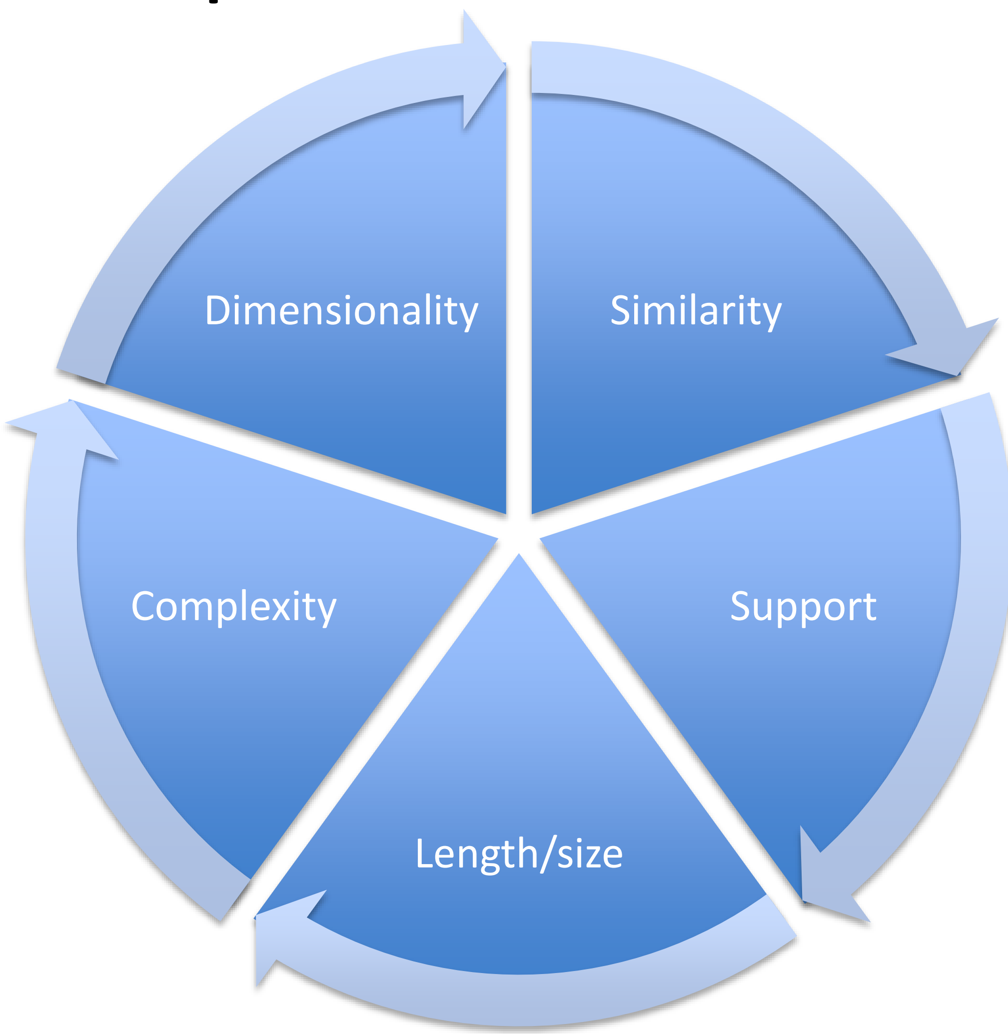
two randomly-generated words of size w over an alphabet of size a, the probability that they match with up to d errors

Motif significance involves several independent dimensions



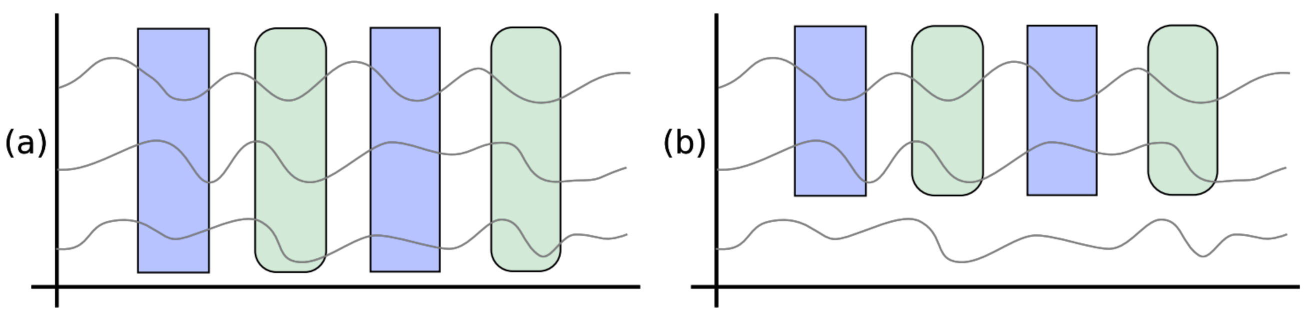
Assessing significance requires estimating a function $S{:}R^d \rightarrow R$ over these dimensions so we can rank the motifs

More invariances mean more independent dimensions



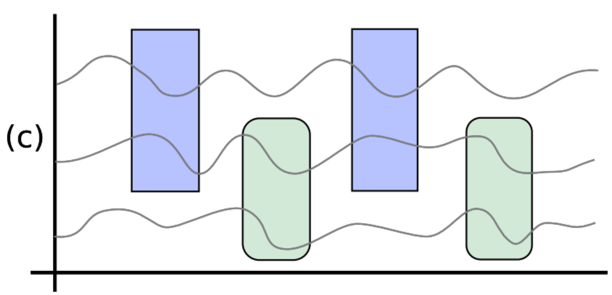
Multi-dimensional Motif

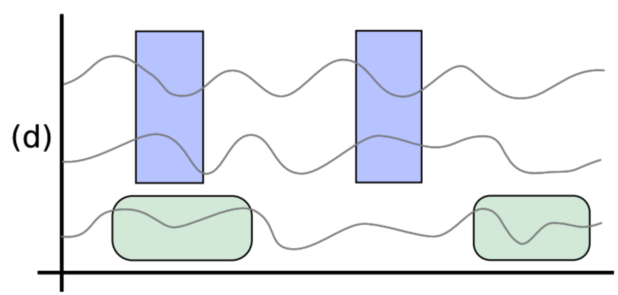
- Synchronous
  - Treat it as an even higher dimensional problem
  - Simple extensions of uni-dimensional algorithms work
  - To find sub-dimensional motifs, all possible sub-spaces have to be considered

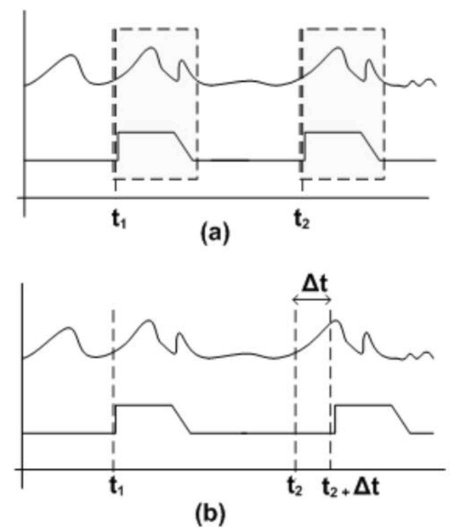


Multi-dimensional Motif

- Non-Synchronous
  - Lags among motifs are common
  - Subsets of dimensions can possibly construct a motif





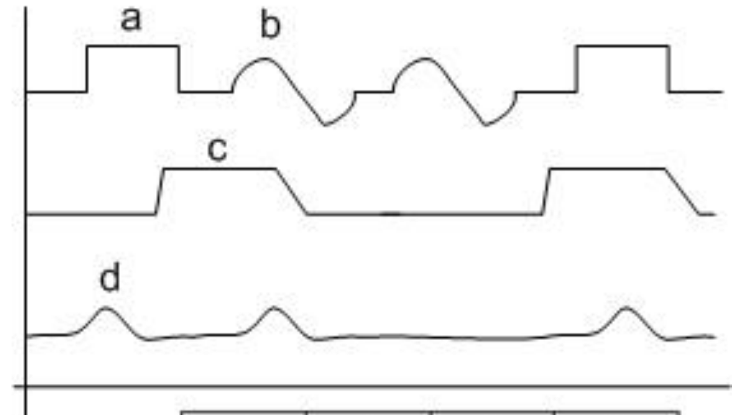


Alireza Vahdatpour, Navid Amini, Majid Sarrafzadeh: Toward Unsupervised Activity Discovery Using Multi-Dimensional Motif Detection in Time Series. IJCAI 2009: 1261-1266

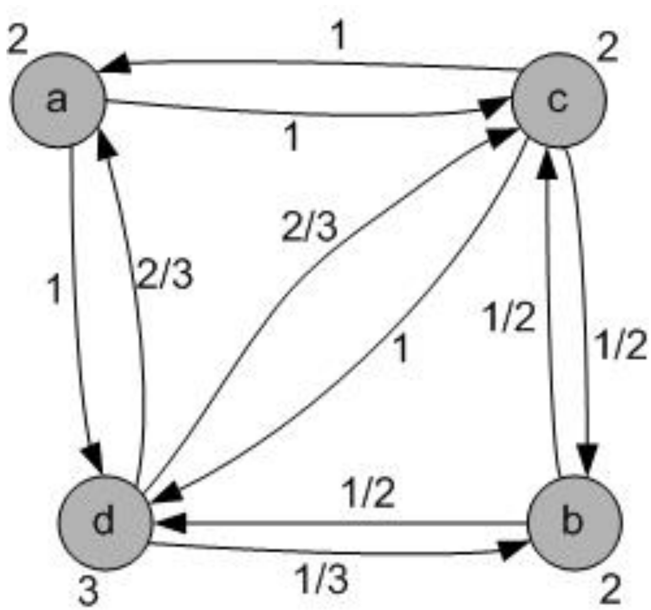
David Minnen, Charles Isbell, Irfan Essa, and Thad Starner. Detecting Subdimensional Motifs: An Efficient Algorithm for Generalized Multivariate Pattern Discovery. ICDM '07

Coincidence table

coincident(r_i, r_j) is the number of overlapping occurrences of motif *i* and motif *j*



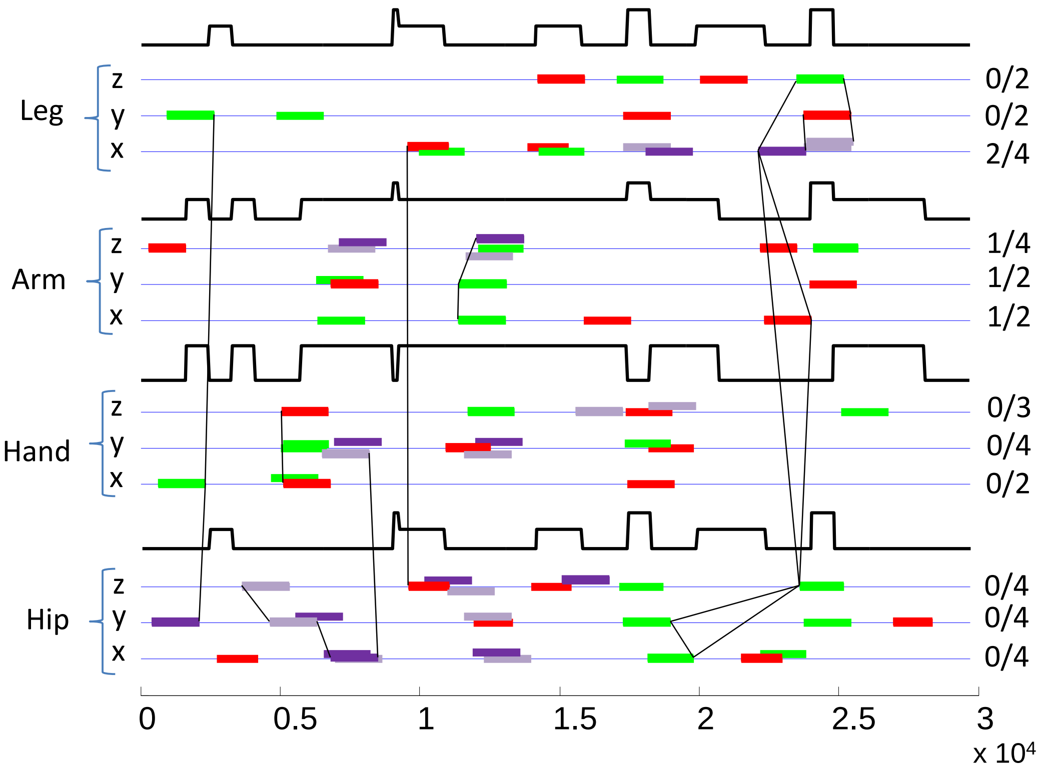
	a	b	c	d
a	1	0	1	1
b	0	1	0.5	0.5
c	1	0.5	1	1
d	0.66	0.33	0.66	1



$$w_{i,j} = coincident(r_i, r_j)/size_i$$

Single-dimensional motifs to graph

- Produce a co-occurrence graph
- Nodes are single dimensional motifs
- An edge between x and y denotes, x and y always co-occur within a time lag
- Cluster the graph using min-cut algorithms to find multi-dimensional motifs



Open Problems

- New Invariances:
  - P1: Find repeated patterns under warping distance.
  - P2: Finding motifs under Complexity invariance.
    - Uniform scaling (Yankov 06)
- Significance:
  - P3: Assessing significance of motifs without discretization.
    - Parameter-free
    - Data adaptive

Open Problems

- Algorithmic:
  - P4: ~~Optimal k-motif for a given threshold~~
  - P5: Exact multi-dimensional motif discovery
- Application:
  - P6: Finding hidden state machine from motifs
- States == clustering
- Rules between patterns only
- State machine is for rules among clusters
- Systems:
  - A suite with all the techniques added
- Parallel motif discovery using GPU

References

1. Abdullah Mueen, Eamonn J. Keogh: Online discovery and maintenance of time series motifs. KDD 2010: 1089-1098
2. Abdullah Mueen, Eamonn J. Keogh, Qiang Zhu, Sydney Cash, M. Brandon Westover: Exact Discovery of Time Series Motifs. SDM 2009: 473-484
3. Abdullah Mueen, Eamonn J. Keogh, Nima Bigdely Shamlo: Finding Time Series Motifs in Disk-Resident Data. ICDM 2009: 367-376
4. Abdullah Mueen: Enumeration of Time Series Motifs of All Lengths. ICDM 2013: 547-556
5. Yuan Hao, Mohammad Shokoohi-Yekta, George Papageorgiou, Eamonn J. Keogh: Parameter-Free Audio Motif Discovery in Large Data Archives. ICDM 2013: 261-270
6. Dragomir Yankov, Eamonn J. Keogh, Jose Medina, Bill Yuan-chi Chiu, Victor B. Zordan: Detecting time series motifs under uniform scaling. KDD 2007: 844-853
7. Xiaopeng Xi, Eamonn J. Keogh, Li Wei, Agenor Mafra-Neto: Finding Motifs in a Database of Shapes. SDM 2007: 249-260
8. Bill Yuan-chi Chiu, Eamonn J. Keogh, Stefano Lonardi: Probabilistic discovery of time series motifs. KDD 2003: 493-498
9. Pranav Patel, Eamonn J. Keogh, Jessica Lin, Stefano Lonardi: Mining Motifs in Massive Time Series Databases. ICDM 2002: 370-377
10. Alireza Vahdatpour, Navid Amini, Majid Sarrafzadeh: Toward Unsupervised Activity Discovery Using Multi-Dimensional Motif Detection in Time Series. IJCAI 2009: 1261-1266
11. Debprakash Patnaik, Manish Marwah, Ratnesh Sharma, and Naren Ramakrishnan. 2009. Sustainable operation and management of data center chillers using temporal data mining. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)
12. Yuan Li, Jessica Lin, Tim Oates: Visualizing Variable-Length Time Series Motifs. SDM 2012: 895-906
13. Tim Oates, Arnold P. Boedihardjo, Jessica Lin, Crystal Chen, Susan Frankenstein, and Sunil Gandhi. 2013. Motif discovery in spatial trajectories using grammar inference. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management(CIKM '13). ACM, New York, NY, USA, 1465-1468.
14. Sorrachai Yingchareonthawornchai, Haemwaan Sivaraks, Thanawin Rakthanmanon, Chotirat Ann Ratanamahatana: Efficient Proper Length Time Series Motif Discovery. ICDM 2013: 1265-1270
15. David Minnen, Charles Lee Isbell Jr., Irfan A. Essa, Thad Starner: Discovering Multivariate Motifs using Subsequence Density Estimation and Greedy Mixture Learning. AAAI 2007: 615-620

References

16. Philippe Beaudoin, Stelian Coros, Michiel van de Panne, and Pierre Poulin. 2008. Motion-motif graphs. In Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '08)
17. David Minnen, Charles L. Isbell, Irfan A. Essa, Thad Starner: Detecting Subdimensional Motifs: An Efficient Algorithm for Generalized Multivariate Pattern Discovery. ICDM 2007: 601-606
18. Yuan Hao, Yanping Chen, Jesin Zakaria, Bing Hu, Thanawin Rakthanmanon, Eamonn J. Keogh: Towards never-ending learning from time series streams. KDD 2013: 874-882
19. Gustavo E. A. P. A. Batista, Xiaoyue Wang, Eamonn J. Keogh: A Complexity-Invariant Distance Measure for Time Series. SDM 2011: 699-710
20. Thanawin Rakthanmanon, Eamonn J. Keogh, Stefano Lonardi, Scott Evans: Time Series Epenthesis: Clustering Time Series Streams Requires Ignoring Some Data. ICDM 2011: 547-556
21. Yasser F. O. Mohammad, Toyoaki Nishida:Constrained Motif Discovery in Time Series. New Generation Comput. 27(4): 319-346 (2009)
22. Nuno Castro, Paulo J. Azevedo:Time Series Motifs Statistical Significance. SDM 2011: 687-698
23. Bill Yuan-chi Chiu, Eamonn J. Keogh, Stefano Lonardi: Probabilistic discovery of time series motifs. KDD 2003: 493-498
24. Zeeshan Syed, Collin Stultz, Manolis Kellis, Piotr Indyk, John V. Guttag: Motif discovery in physiological datasets: A methodology for inferring predictive elements. TKDD 4(1) (2010)
25. Yasser F. O. Mohammad, Toyoaki Nishida: Exact Discovery of Length-Range Motifs. ACIIDS (2) 2014: 23-32

Questions and Comments

THANK YOU