

Time Series Motifs

Statistical Significance



NUNO C. CASTRO
PAULO J. AZEVEDO

May 18th, 2011

Roadmap

I. Introduction

- I. Data Mining
- II. Time Series
- III. Motivation
- IV. Examples

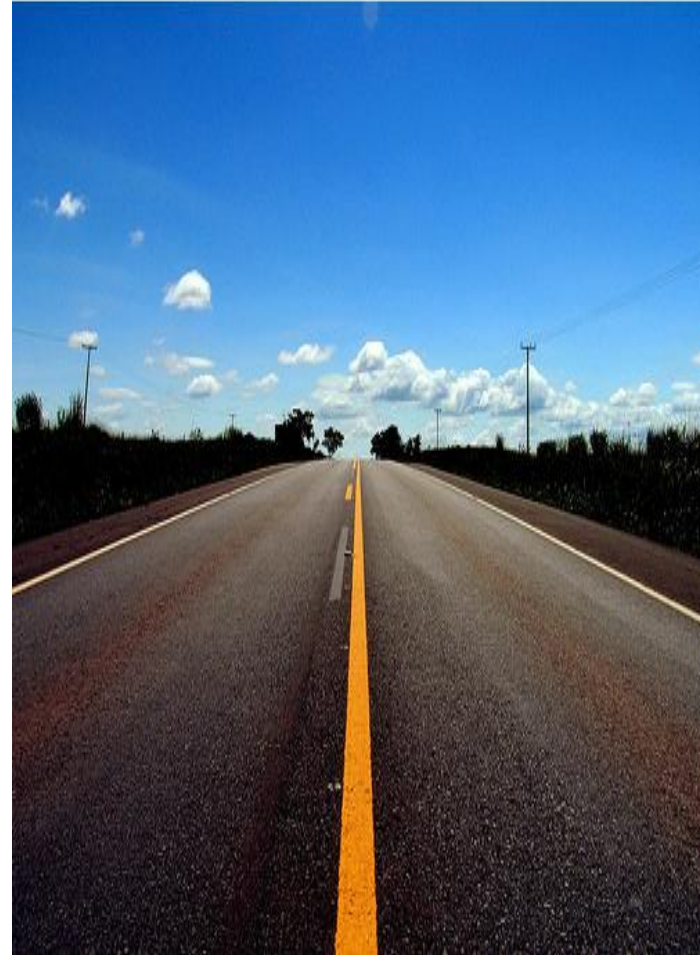
II. Motif Discovery

- I. Motif definition and motivation

III. Motif Statistical Significance

- I. Introduction
- II. Approach
- III. Experimental Analysis

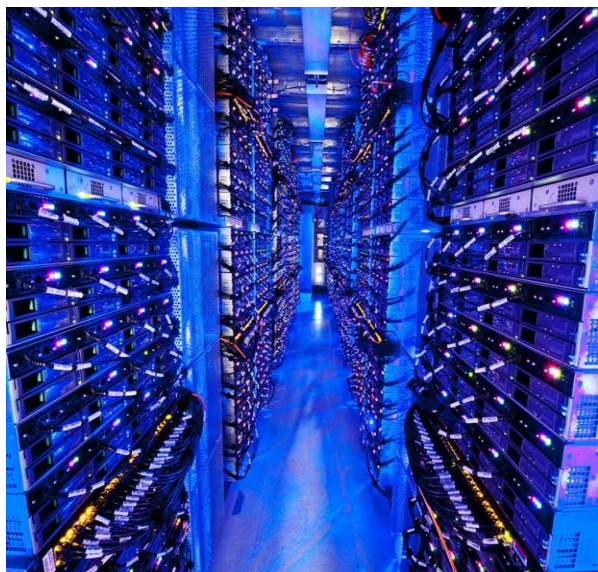
IV. Conclusions



Data Mining

- The extraction of **nontrivial**, **implicit** and **useful** knowledge from the data

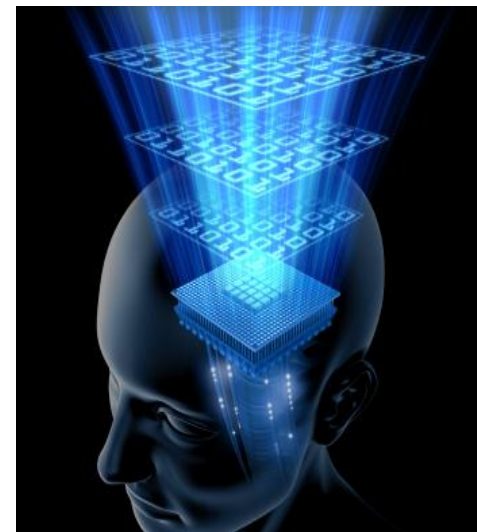
Data



Data Mining

- Artificial Intelligence
- Computer Science
- Statistics
- Information Retrieval

Knowledge



Data Mining goals

- To **find “structure”** in the large amount of information available from different sources
- To **organize** the data
- To **identify patterns** that translate into new understandings and viable predictions
- To **discover relationships** between data and phenomena that ordinary operations and routine analysis would otherwise overlook

Time Series

- People measure things:

- Oil price
- Sócrates popularity
- Blood pressure, etc.

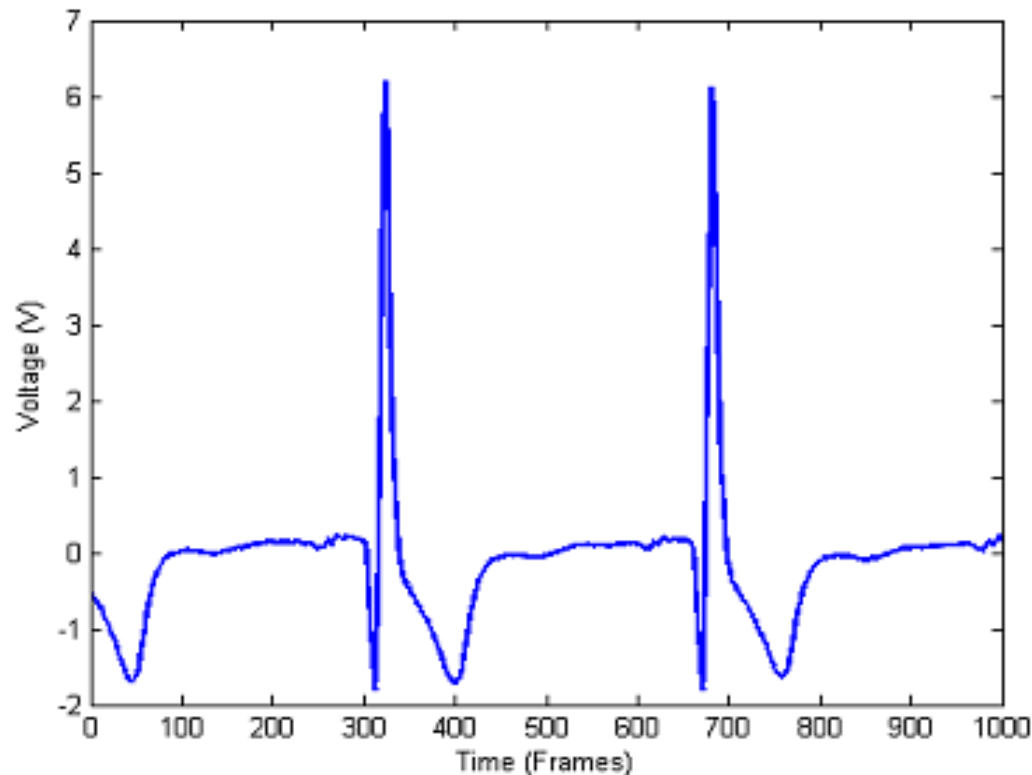


and things change over time, creating a time series

Time Series definition

- A (numeric) *time series* is a sequence of observations of a numeric property over time

-1,25
-1,00
0,01
0,05
...
5,45
0,00
...



Motivation to Work in Time Series

- Time series are **ubiquitous**
- Most of the information (data) produced in a variety of areas are time series
 - e.g. about 50% of all newspaper graphics are time series
- Other types of data can be converted to time series

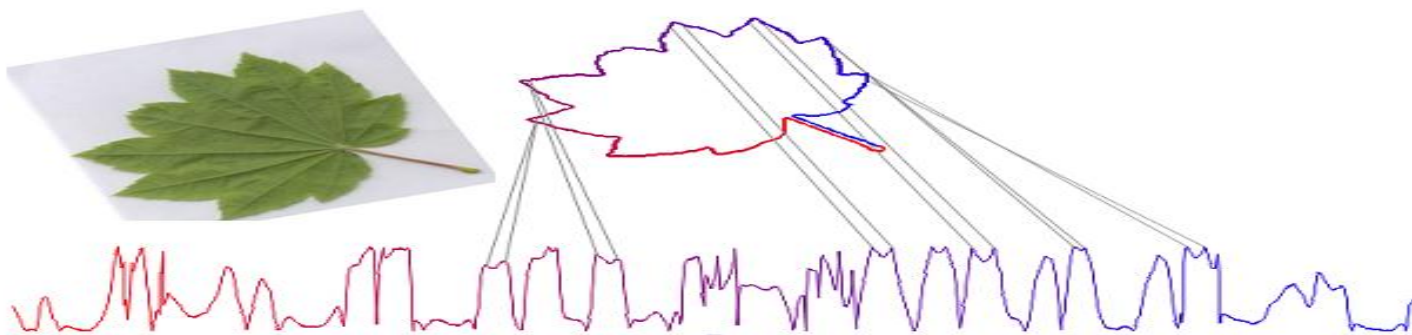
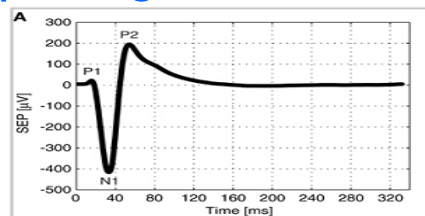
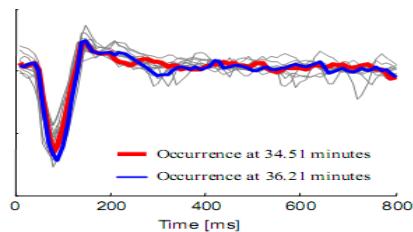


Image from E. J. Keogh. A decade of progress in indexing and mining large time series databases. In VLDB, page 1268, 2006.

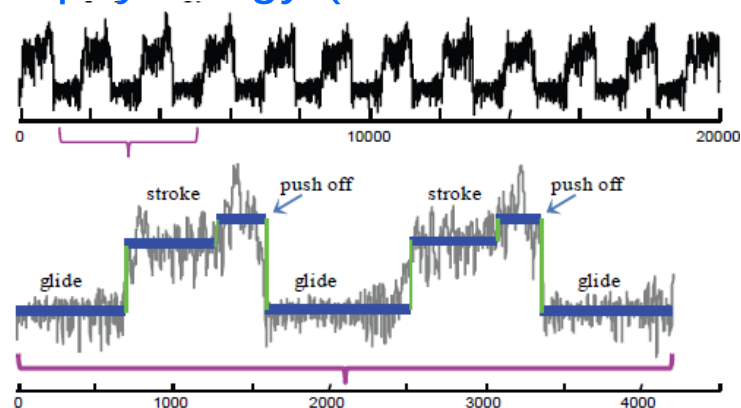
Time Series Examples

Images from a variety of papers by E. J. Keogh. Available at: www.cs.ucr.edu/~eamonn

electroencephalogram



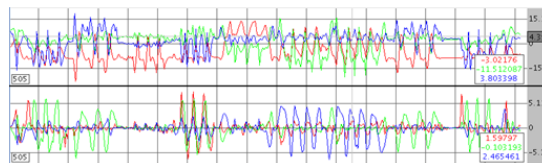
physiology (muscle activation)



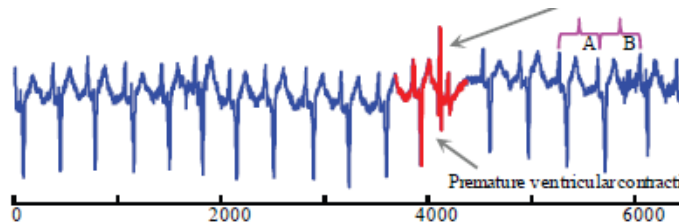
historical archives



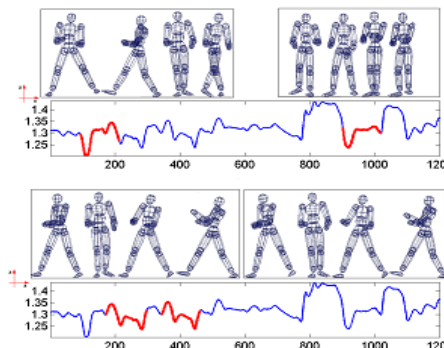
sensors



ECG



motion data



I – Introduction

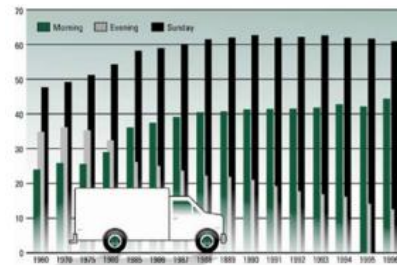
Time Series Examples (cont.)

Image from E. J. Keogh. A decade of progress in indexing and mining large time series databases. In VLDB, page 1268, 2006.

stocks



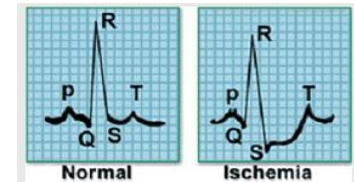
sales



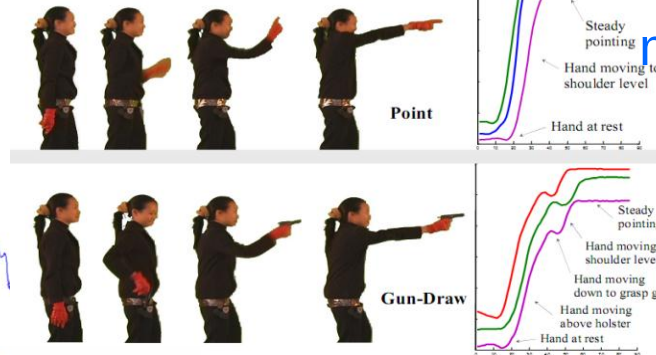
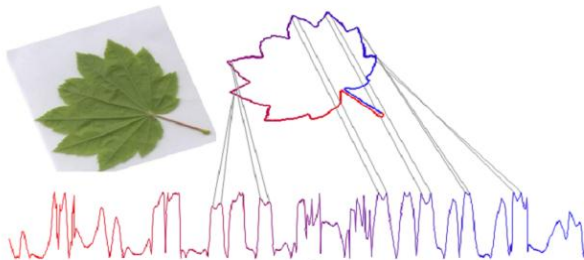
goods consumption



animal ECG



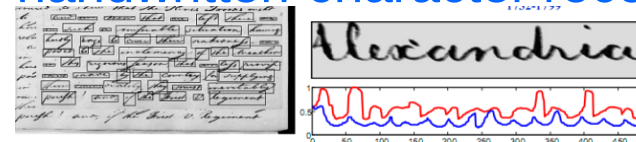
images



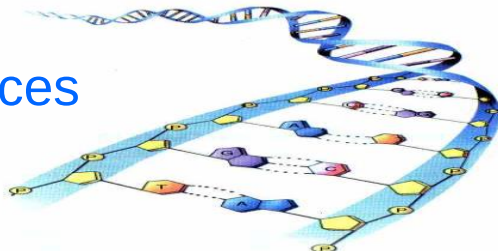
motion capture



handwritten character recognition



DNA sequences



Time Series data characteristics

- Analysis is hard, as we are typically dealing with **massive** data-sets:
 - One hour EEG: 1 GB of data
 - Typical weblog: 5 GB / week
 - MACHO database: 5 TB (growing 3 GB a day)
 - Stanford Linear Accelerator database: 500 TB
- **Quadratic** complexity algorithms are insufficient
- The data also present some **distortions** (noise, scaling effects, etc.) that make the analysis more difficult

Time Series Data Mining Tasks

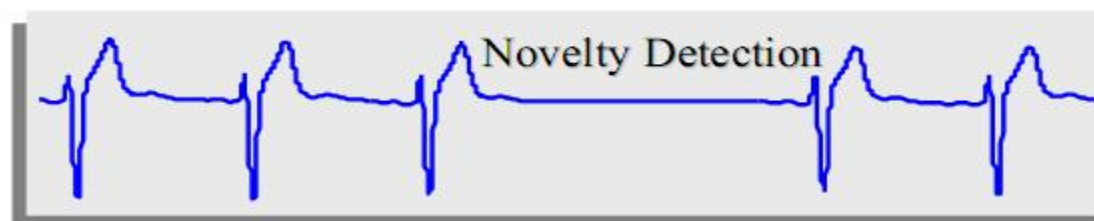
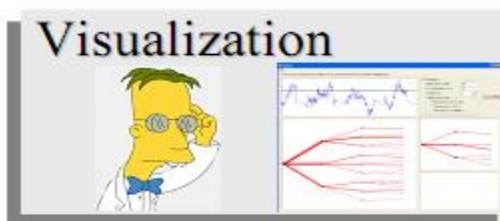
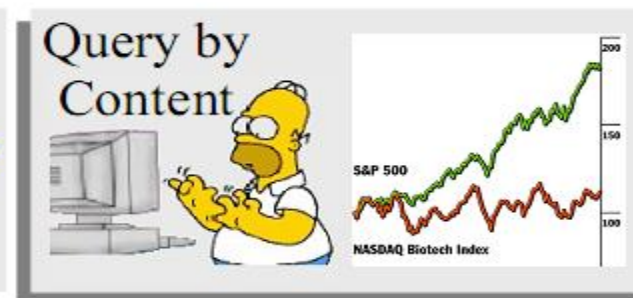
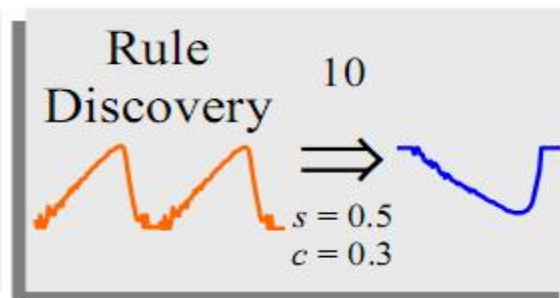
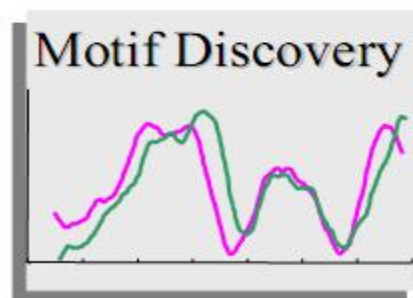
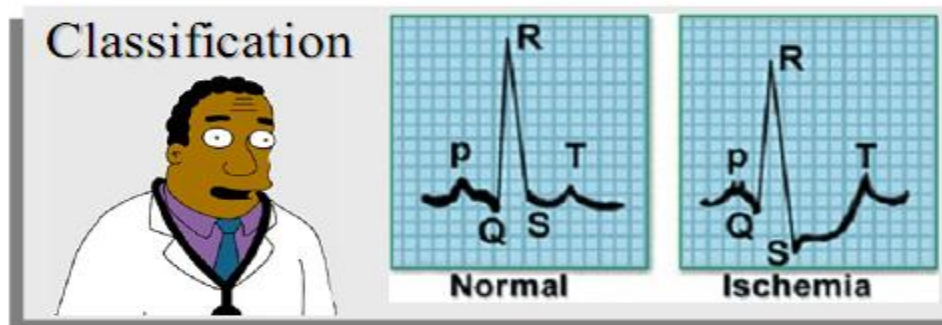
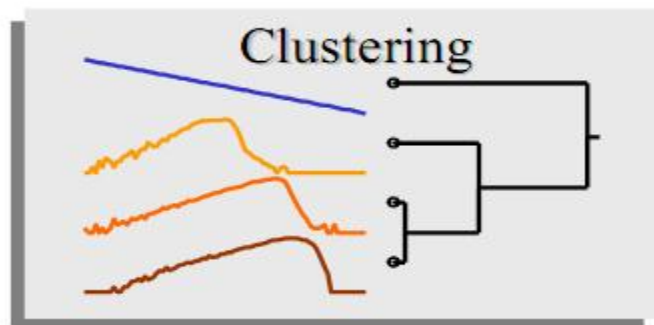
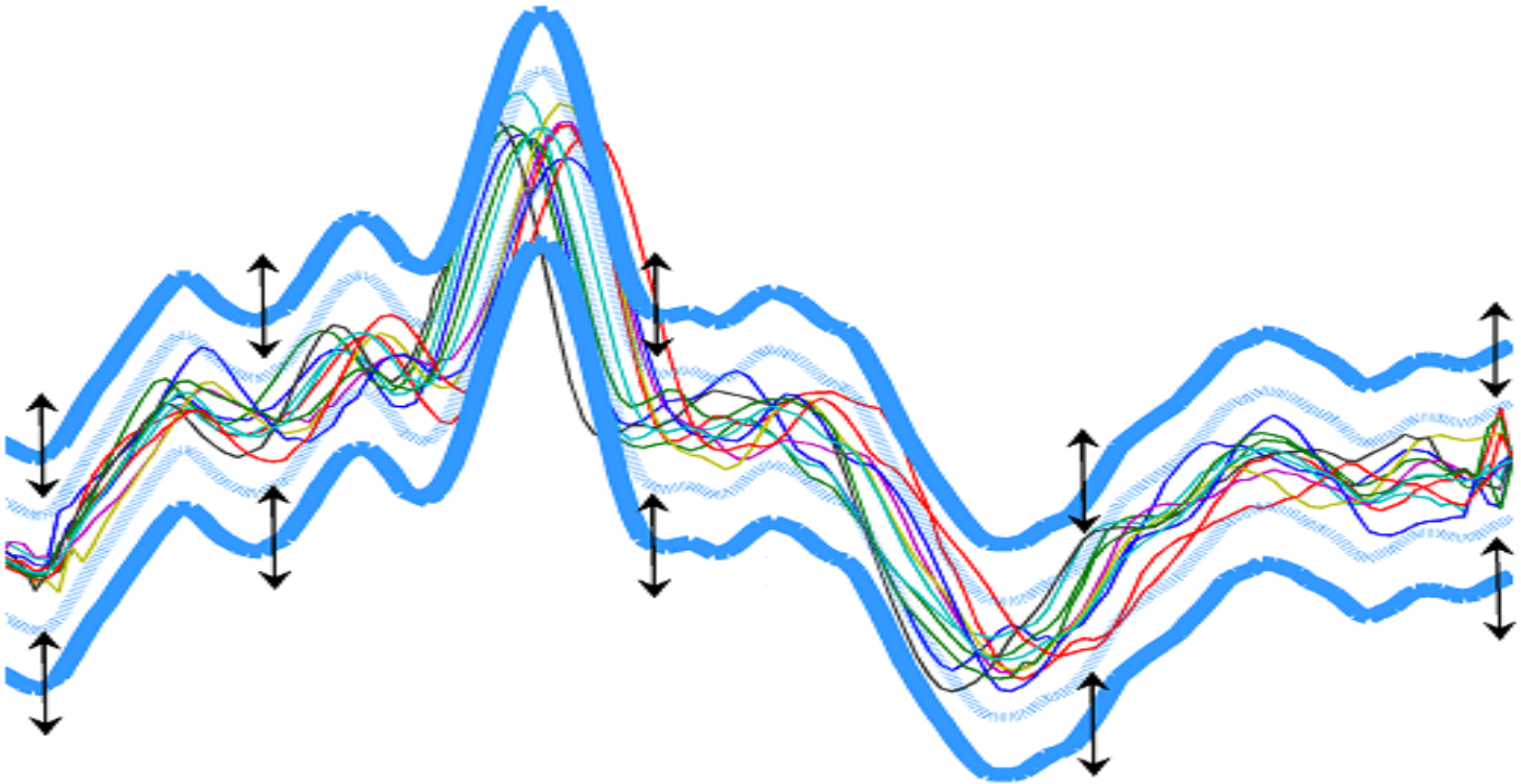


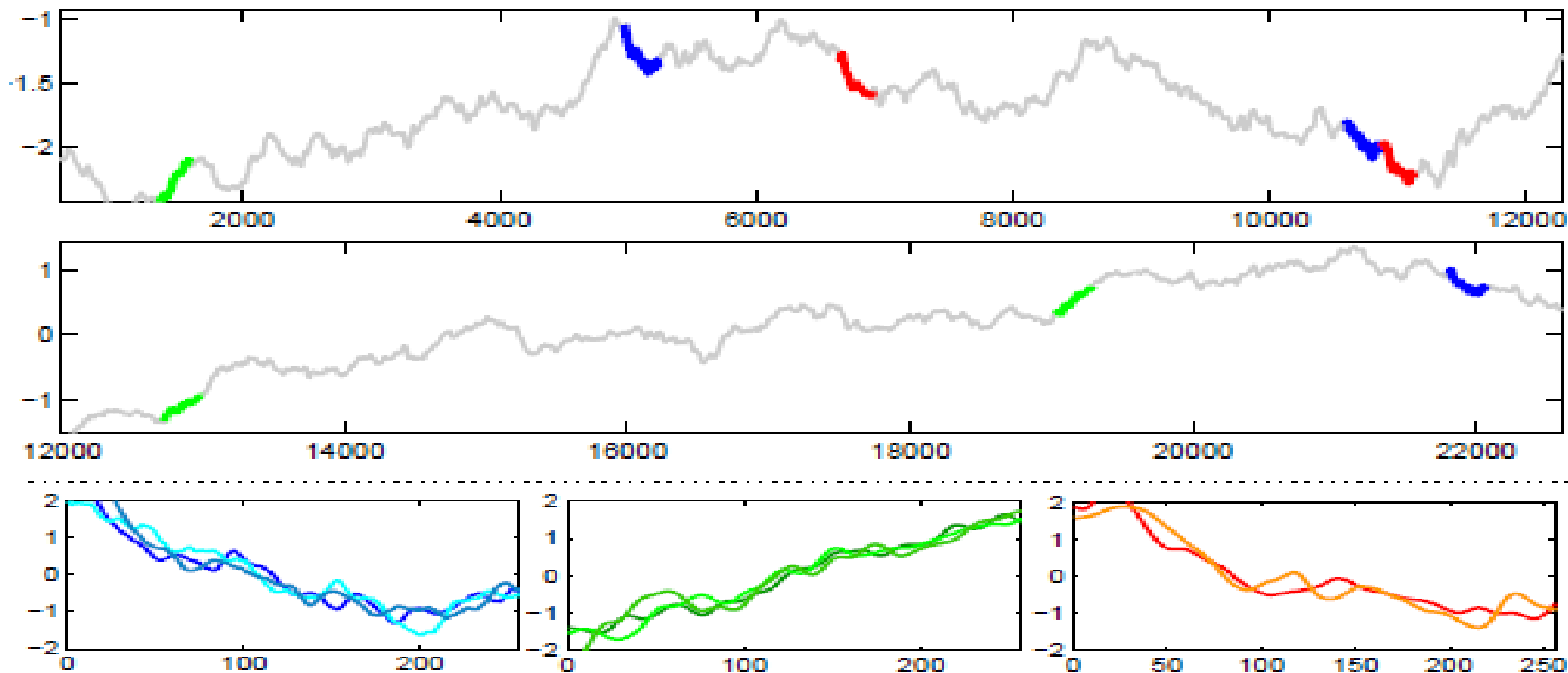
Image from E. J. Keogh. *A decade of progress in indexing and mining large time series databases*. In VLDB, page 1268, 2006.

II – Motif Discovery



Motif Definition

- **Motifs**, also known as “recurrent patterns”, “frequent patterns”, “repeated subsequences”, or typical shapes” are **previously unknown patterns in time series**

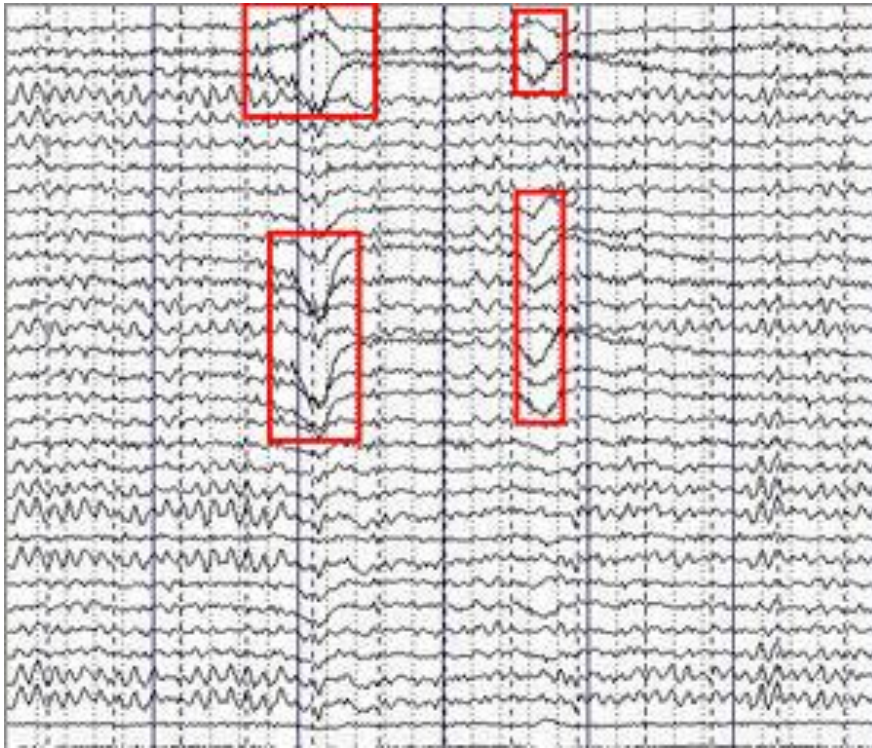


Motivation

- Finding motifs is an important task:
 - Describe the time series at hand
 - Help summarize/represent the database
 - Provide useful insight to the domain expert

Motif Example

- Patterns that precede a seizure in EEG



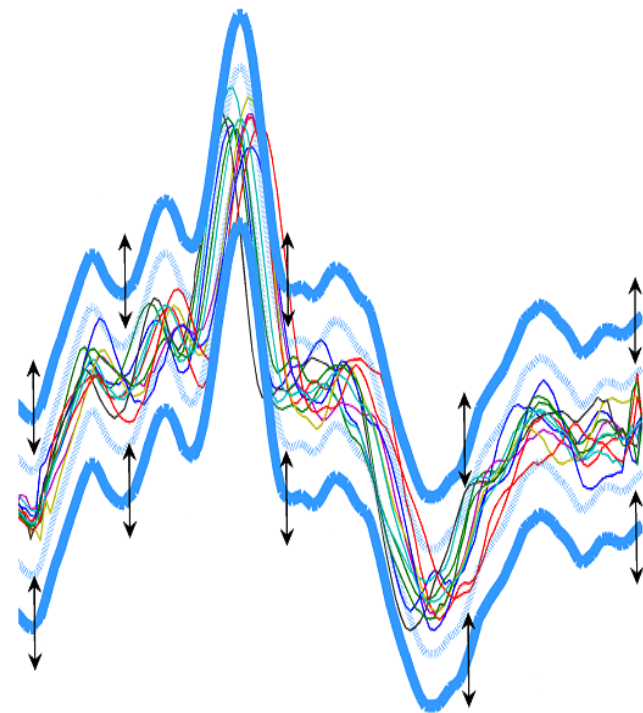
Motif Example (cont.)

- Bursts in telecommunication traffic



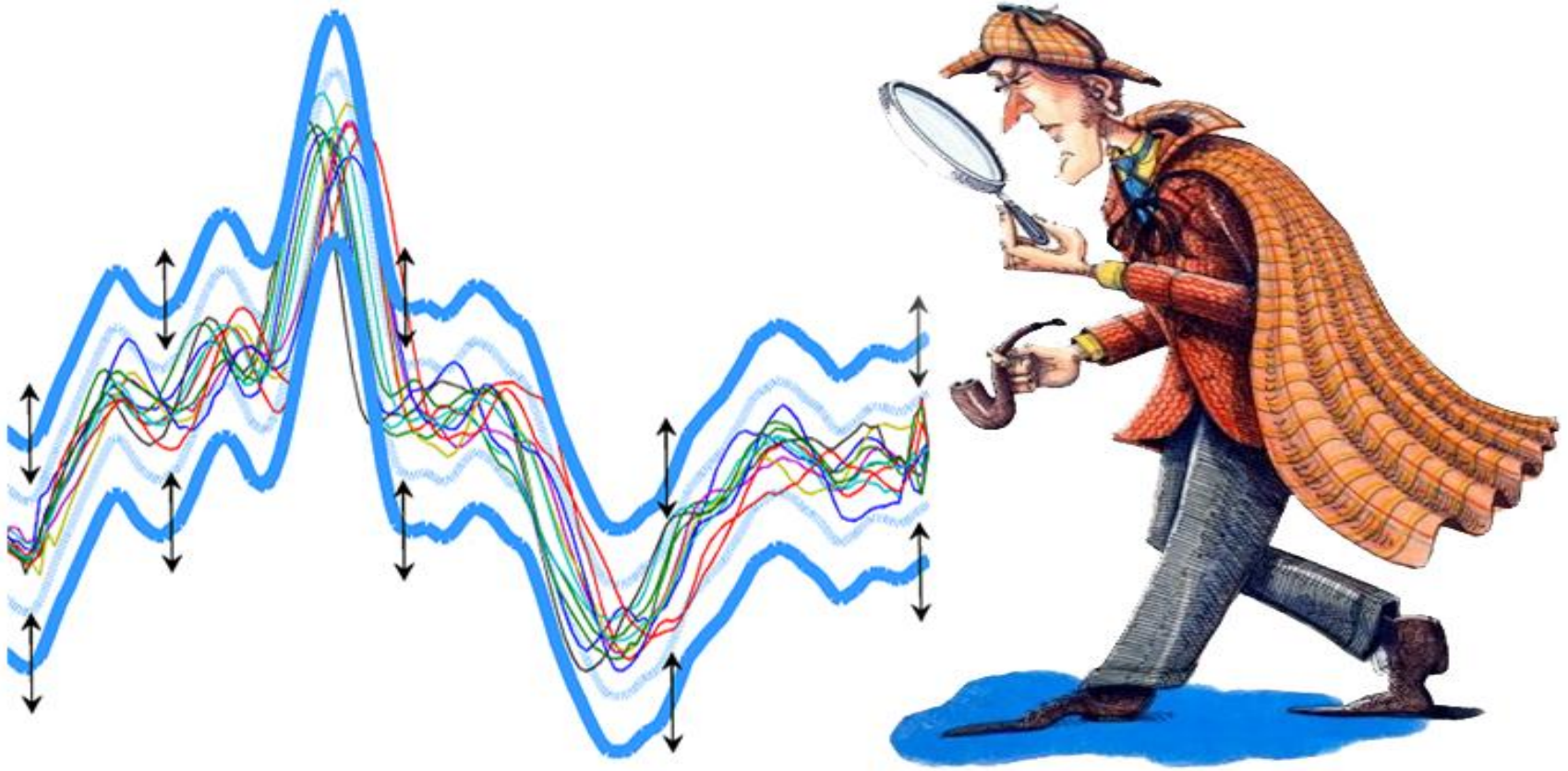
Our previous work

- We have proposed a motif discovery algorithm:
 - Multiresolution Motif Discovery in Time Series (**MrMotif**)*
 - Time efficient:
 - One single sequential disk scan
 - Clever representation technique (*iSAX*)
 - Use of constant access time structures
 - Memory efficient:
 - Combine our approach with the Space-Saving algorithm
 - Adjustable amount of memory to use



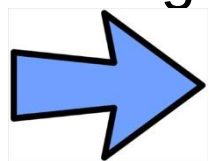
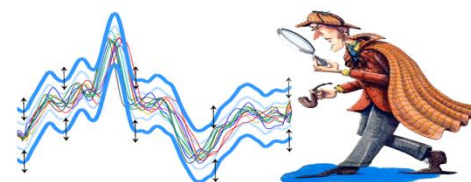
*Nuno Castro and Paulo J. Azevedo, *Multiresolution Motif Discovery in Time Series*, in Proceedings of the SIAM International Conference on Data Mining (SDM 2010), Columbus, Ohio, USA., pp. 665-676.

III – Motifs Statistical Significance



Problem

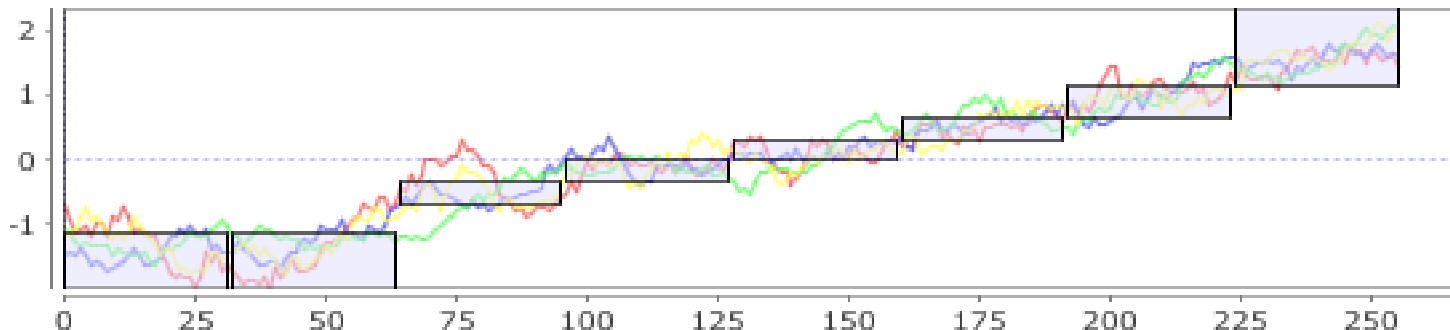
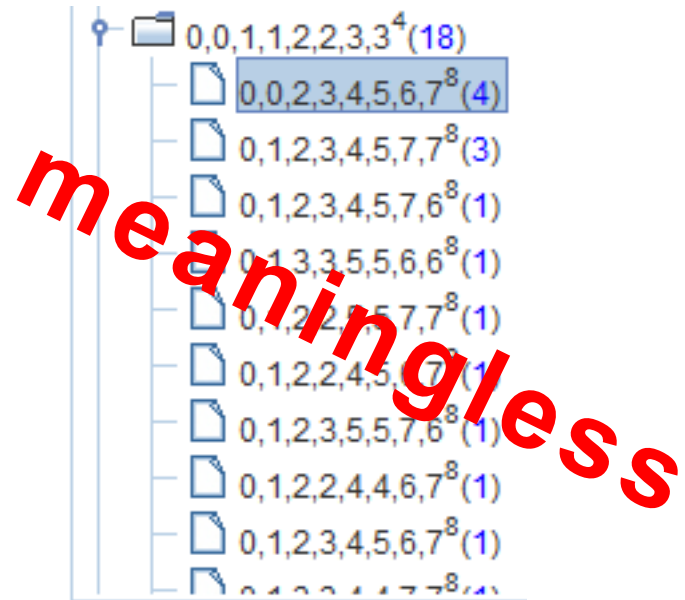
- A large number of proposals recently introduced on “how to efficiently **mine** motifs”
- Very **few** works on how to **evaluate** the motifs
- Motifs are typically evaluated by *humans*
 - Subjective
 - Slow
- **Unfeasible** for real-world datasets (**Terabytes** of data)
 - A large number of patterns are returned by motif mining algorithms



Automatic evaluation measures are necessary.

Example

- Randomly generated dataset with 65536 time series of length 256
- 65 motifs were discovered
- Most frequent motif: 4 repetitions
- Average motif count: 2.17



Solution

- Statistical tests are widely used in data mining
 - In bioinformatics, to detect DNA segments with unexpected frequency
 - In networks mining, to find significant subgraphs
 - In itemsets mining, to discard redundant rules
- They aim to answer the question:
 - “Can this pattern occur so many times just by chance?”
- We intend to compare a motif’s **expected** and **observed** count using statistical tests

Contribution

- To present an approach to assess the statistical significance of time series motifs:

calculate each **motif's p-value**

Our approach

- Motifs are extracted from the database
- Motif's expected count is calculated
- Statistical hypothesis tests are applied to assess each motif's p-value



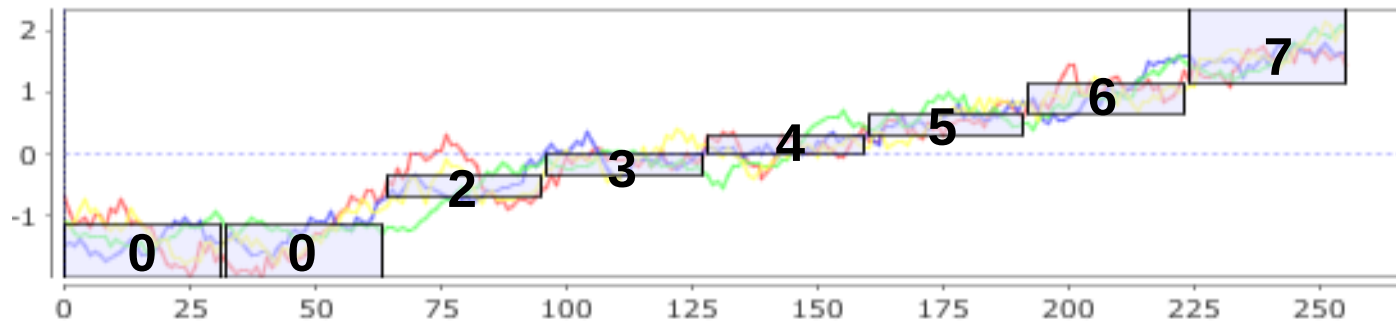
III – Motif Statistical Significance – Approach

- Motifs are extracted from the database
- Motif's expected count is calculated
- Statistical hypothesis tests are applied to assess each motif's p-value



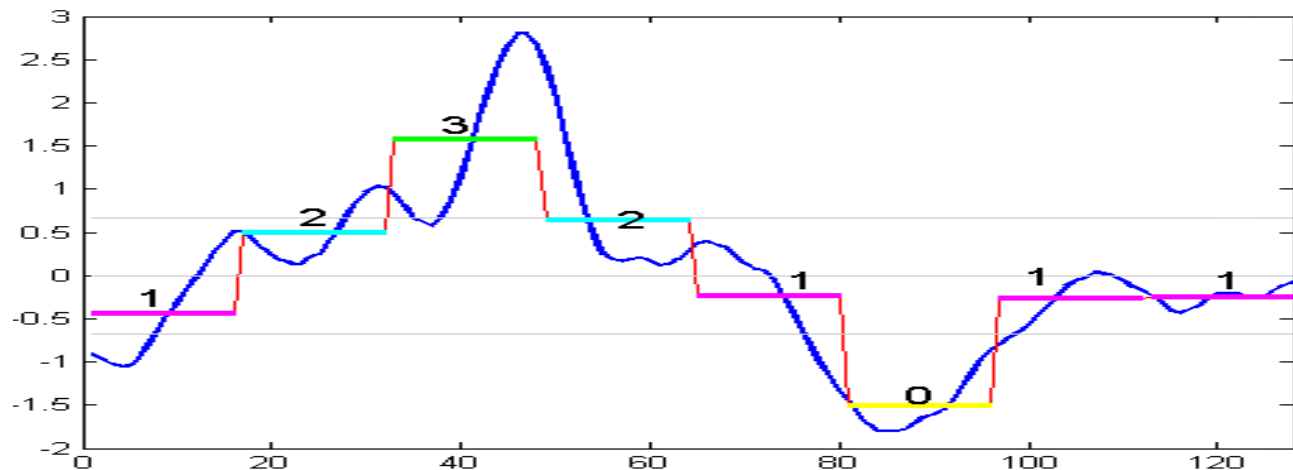
Extracting motifs

- In order to leverage existing work from the bioinformatics, we are interested in **symbolic** motifs
- A symbolic motif is the representation of a motif using symbols (integers, letters)
- For example, the motif $\{0, 0, 2, 3, 4, 5, 6, 7\}$:



Symbolic Aggregate Approximation (iSAX)

- State of the art time series representation technique
- Widely used in time series data mining
- Converts a time series to a sequence of symbols (word)
 - Given a *resolution* (alphabet size) and *word size*

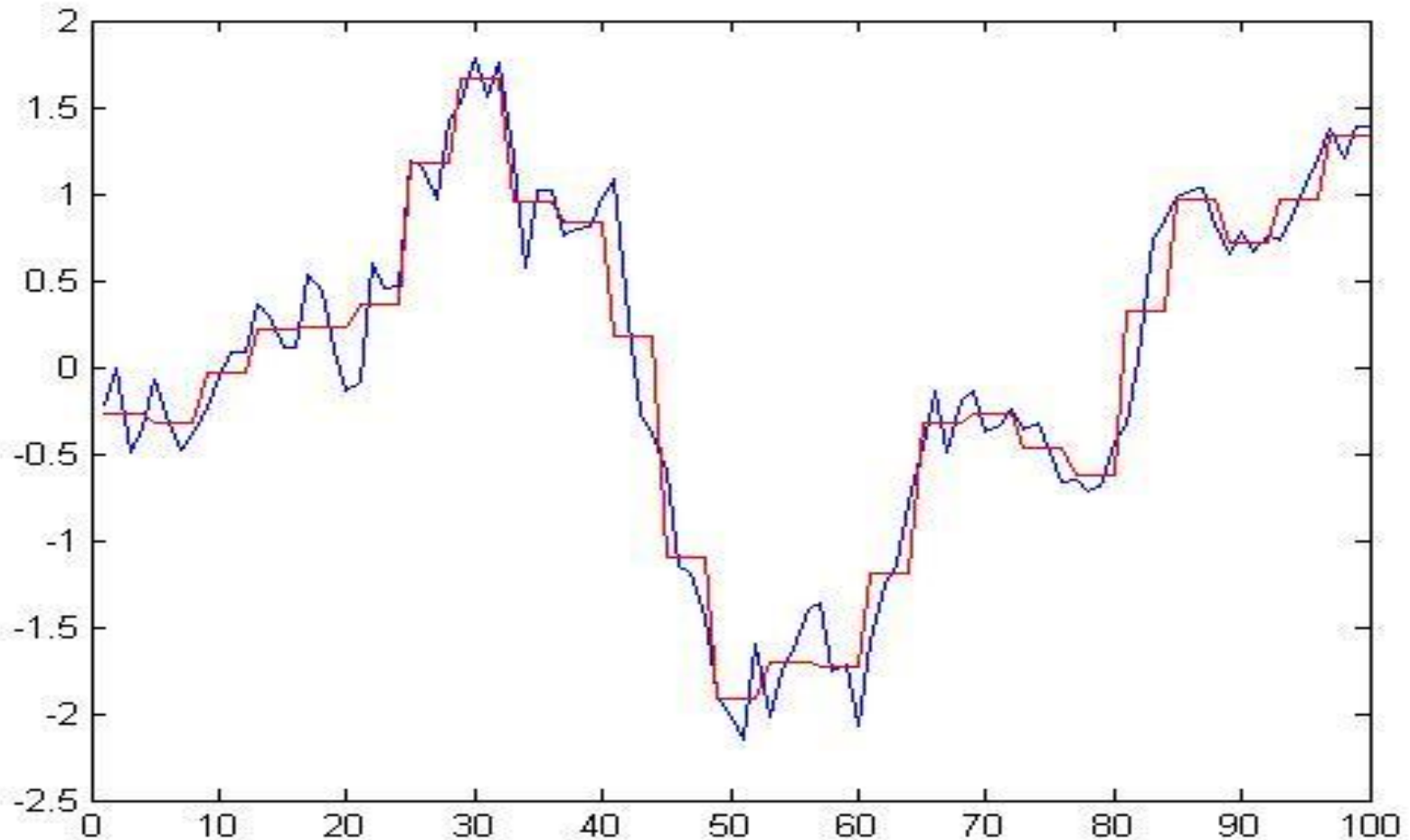


Time series is represented by the iSAX word:

1, 2, 3, 2, 1, 0, 1, 1, 1

* Shieh, J. and Keogh, E., *iSAX: indexing and mining terabyte sized time series*, in Proceedings of the 14th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (2008), pp. 623-631.

iSAX example



Obtained using MATLAB code made available by Eamonn Keogh at www.cs.ucr.edu/~eamonn

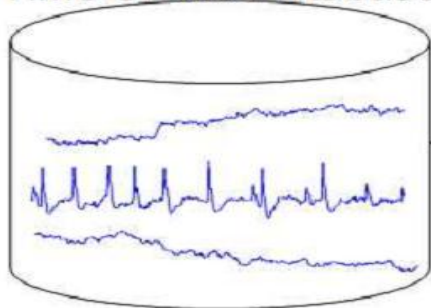
Nuno Castro and Paulo Azevedo

18/05/2011

Extracting motifs (cont.)

- Frequent motifs are extracted using a motif discovery algorithm and symbolized using iSAX*

Time Series Database



MrMotif

Frequent Motifs

Motif	Count
{ 2 , 5 , 7 , 5 , 3 , 0 , 3 , 3 }	54
{ 4 , 7 , 0 , 0 , 0 , 1 , 5 , 5 }	32
{ 0 , 0 , 0 , 4 , 5 , 2 , 0 , 0 }	25
...	

III – Motif Statistical Significance – Approach

- Motifs are extracted from the database
- Motif's expected count is calculated
- Statistical hypothesis tests are applied to assess each motif's p-value



Expected counts

- Frequency by its own does not guarantee that motifs are significant
- A better approach is to consider the difference between the motif **expected count** and its **observed count**
- The expected count is the number of repetitions of a motif we should expect in random sequences that are similar to our database

Expected counts (cont.)

- We use Markov Chain Models to estimate a motif's probability of occurrence
- For a motif, we consider its **subword** count
- For example, the motif “baccdfah”:

$$\text{M6} \quad \left| \quad \mu = \frac{N(\text{baccdfa}) N(\text{accdfah})}{3n N(\text{accdfa})}\right.$$

- Expected count: $\hat{N}_m(w) = n \mu$

III – Motif Statistical Significance – Approach

- Motifs are extracted from the database
- Motif's expected count is calculated
- Statistical hypothesis tests are applied to assess each motif's p-value



Statistical Significance

- We intend to calculate the motifs p-values:
 - P-value is the probability of the **motif count** to be *at least as large* as the **observed count**, just by chance.
 - We assume the motif count in time series is Binomial, therefore

$$\mathbb{P}(\mathcal{B}(n, \mu) \geq N^{obs}(w)) = 1 - \sum_{k=0}^{N(w)-1} \binom{n}{k} \mu^k (1 - \mu)^{n-k}$$

- If $P \leq \alpha$, we say the pattern is accepted as significant
 - α calculated using the Holm method
- Otherwise, pattern is rejected

Multiple hypothesis testing problem

- The significance level (α) is typically fixed to **0.05**
- Since we apply a test for **each** distinct motif, in a dataset with **100000** motifs we expect to have **5000** significant motifs by chance alone
- The higher the number of simultaneously executed tests, the higher the chance to find at least one that **incorrectly** rejects the null hypothesis

Multiple hypothesis testing problem

- Bonferroni adjustment
 - $\alpha' = \alpha / n$
 - e.g. $\alpha' = 0.05 / 65 = 0,00077$
 - too strict
- Holm procedure
 - all p-values are sorted increasingly from p_1 until p_n
 - the first one to reject $p_j \leq \alpha / (n-j+1)$ becomes α'

Experimental Analysis

- We test our approach on data from a **wide** range of applications and **sizes**
- **52** publicly available datasets from a variety of sources are used
- The **MrMotif** algorithm is used to extract **symbolic** motifs from the time series database
- The significance level (α) is automatically calculated using the **Holm** procedure

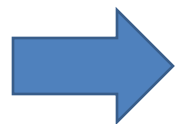
III – Motif Statistical Significance – Experimental Analysis

Results

	sequence length	distinct motifs	nr. significant motifs	adjusted cutoff	% accepted
Dataset	n	N_d	NSM	α'	%
ERP	47616	2628	95	1.97E-05	3.61
eog	67493	5882	95	8.64E-06	1.62
rateeg	576694	100438	95	4.98E-07	0.09
lightcurves	5327	376	70	0.000163	18.62
cl2	4310	54	36	0.002632	66.67
sasa	81280	754	29	6.89E-05	3.85
koskiecg	2394	360	24	0.000148	6.67
mallat	803	30	18	0.003846	60.00
motor	420	60	7	0.000926	11.67
stocks	18000	1394	7	3.6E-05	0.50
arrowheads	1231	161	5	0.000318	3.11
pen	510	46	4	0.001163	8.70
burstin	1310	221	4	0.000229	1.81
powerdata	1838	295	4	0.000171	1.36
shapemixed	160	14	2	0.003846	14.29
10000	10000	754	2	6.64E-05	0.27
TEK	180	51	1	0.00098	1.96
eegheartrate	373	85	1	0.000588	1.18
leaf	442	72	1	0.000694	1.39
network	1121	36	1	0.001389	2.78
insect	1471	77	1	0.000649	1.30
chaotic	109	4	0	0.0125	0
random	1718	65	0	0.000769	0
fortune	500	9	0	0.005556	0
logistic	2000	181	0	0.000276	0
packet	2332	187	0	0.000267	0
tide	2906	6	0	0.008333	0
eeg	62700	2767	0	1.81E-05	0

Pruning power

- Our approach **prunes** most of the false discoveries
- For some datasets, **all** frequent motifs were **discarded**
- Using statistical tests in time series motif discovery can act as a **filter**, pruning *meaningless* motifs



This seems to support the **need** for statistical tests in time series motif discovery.

Number of parameters

- Pruning the prohibitively large output of pattern discovery algorithms is typically done by **support** or (top) **K** parameters
- Unintuitive parameters
- Can only be optimized by experimentation
 - May be unfeasible for some datasets to re-run the algorithm with a new parameter setting

 Using our approach **avoids** the use of unintuitive parameters, since the adjusted cutoff value (α') is automatically derived

Motif ranking

- Motifs can be **ranked** according to their statistical significance, i.e. p-value
- To be able to rank motifs is important: a ranking yields a smooth way to select the most representative and relevant motifs
- For example, for the domain expert it is better to manually analyze 5 motifs, than 754
- In some cases, when the number of motifs makes the manual analysis very difficult, p-value based rankings may become a requirement

Motif ranking (cont.)

Datasets	Motif	Motif count	Motif Probability	Expected	p-value
		$N(w)$	μ		
<i>sasa</i>	gggfcbbb	17	3.9E-05	3.172479	4.77E-08
	hggdcbbb	8	8.79E-06	0.7143	8.93E-07
	bbbbgggg	14	3.37E-05	2.735099	1.19E-06
	bbbeggfg	10	1.67E-05	1.354194	1.68E-06
	abbdgggg	7	7.16E-06	0.58183	2.7E-06
<i>eog</i>	aacefggg	31	8.79E-05	5.932245	3.69E-13
	caacfggh	11	6.36E-06	0.429089	1.54E-12
	babbeggh	12	8.78E-06	0.592607	2.27E-12
	dbdgggfa	11	7.38E-06	0.497955	7.41E-12
	gabdeggd	12	1.03E-05	0.695669	1.2E-11
<i>cl2</i>	heddddbe	74	0.00193	8.319006	3.98E-13
	hecddcdf	37	0.001998	8.613394	7.54E-13
	hedcdccd	645	0.049903	215.0832	9.33E-13
	hedddcce	80	0.006069	26.1573	1.06E-12
	hedddccd	64	0.004855	20.92584	1.23E-12
<i>koskiecg</i>	gdddddbg	40	0.002734	6.544641	2.37E-12
	dddddbfh	34	0.00299	7.157086	2.88E-12
	hedddddb	43	0.006027	14.42812	7.89E-10
	dddddbgh	22	0.001817	4.350855	1.49E-09
	dbggdddd	45	0.00719	17.21198	1.55E-08
<i>mallat</i>	dgbcdche	90	0.03608	28.97219	6E-13
	cgbcdche	97	0.041707	33.49079	6.16E-13
	dgbbdche	92	0.038283	30.74089	6.57E-13
	dgbcdcege	59	0.024542	19.70757	7.29E-13
	dhbcdcege	137	0.056988	45.76165	7.92E-13

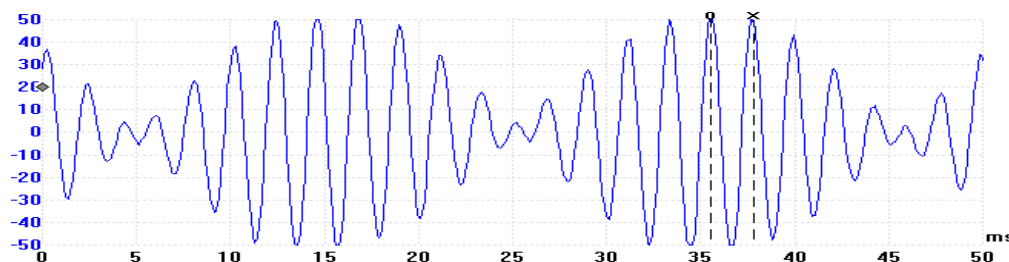
IV – Conclusions

- We proposed an approach to compute the p-values of time series motifs
- A motif is accepted if it passes a statistical hypothesis test
 - i.e. $p\text{-value} \leq \text{significance level}$.

Conclusions (cont.)

- Our approach:
 - Significantly **reduces** the number of returned patterns
 - Avoids the use of unintuitive support or top-K **parameters**
 - Allows to **rank** motifs according to their significance
 - Provides researchers and practitioners with an important technique to **evaluate** the degree of **relevance** of each pattern
- We aim to **highlight** the **importance** of motif evaluation, since we believe it is **crucial** to make motif mining an useful task in practice

Thank you for your attention!



- Contact: castro@di.uminho.pt
- Paper web site (executable, source code and datasets):

www.di.uminho.pt/~castro/stat

Future work

- Extend work to other statistical tests
- Integrate the approach in the motif discovery process (currently applied as post-processing)
- Use other approaches (e.g. FDR) to deal with the multiple hypothesis problem