CERN openlab
Minor Review Meeting

21st April 2009

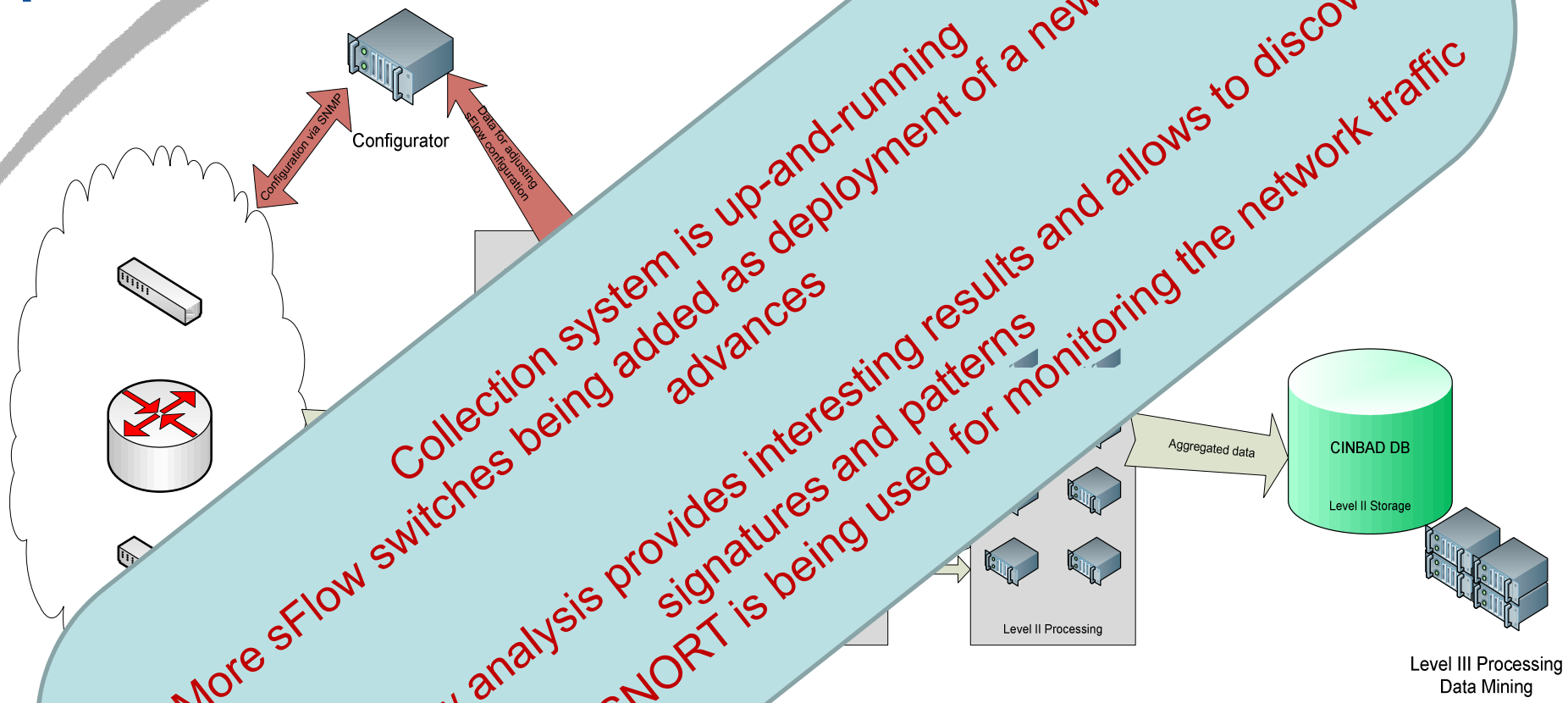
Milosz Marian Hulboj - CERN/Procurve

Ryszard Erazm Jurga - CERN/Procurve

Agenda

- Motivation for the Time Series Data Mining
- Time Series – short introduction
- SAX representation of the Time Series data:
  - Algorithm
  - Example of applications:
    - Time Series Bitmaps
    - Motif Detection
- Conclusions and Future Plans

Collector [illegible]tion



Collection system is up-and-running

More sFlow switches being added as deployment of a new firmware advances

Statistical flow analysis provides interesting results and allows to discover signatures and patterns

Modified version of SNORT is being used for monitoring the network traffic

Results are promising, but…

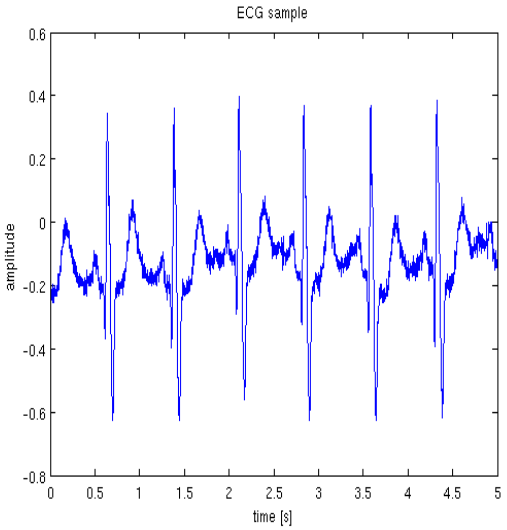
Some of the methods require a significant amount of manual work

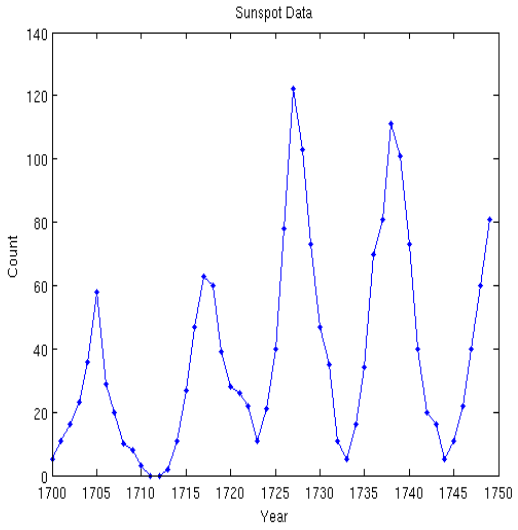
That is why we want to look at the Time Series Data Mining techniques…

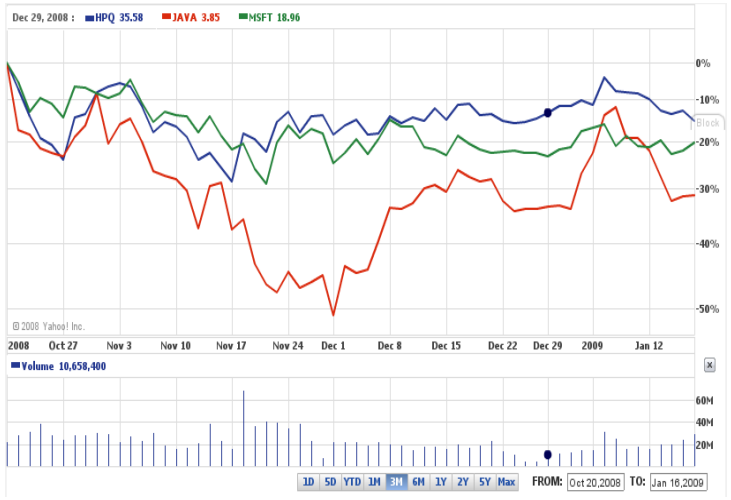
Which hopefully will allow to increase the automatisation

Time series and why do we care?

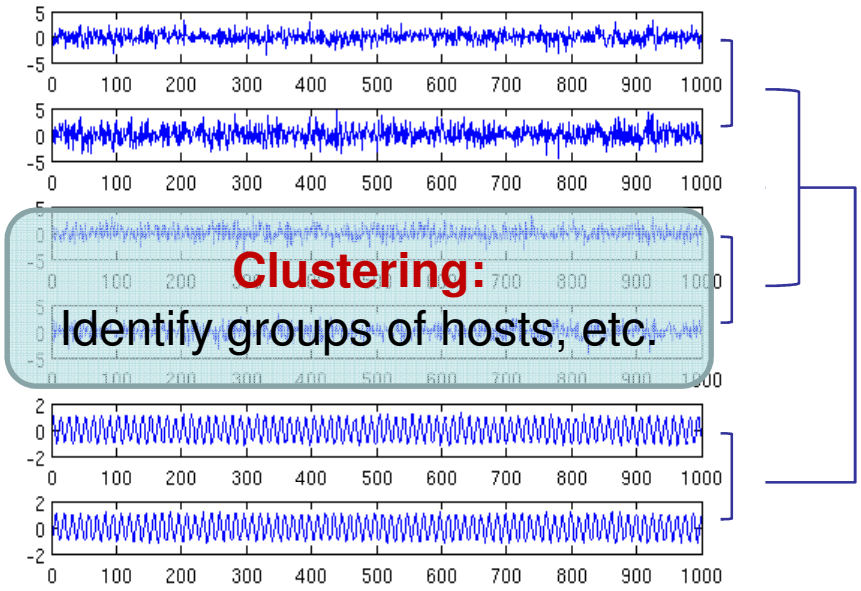
- What are the Time Series?
  - A **time series** is a sequence of data points, measured at successive times
  - Time series are ubiquitous, more and more data is being measured and collected
- Examples:



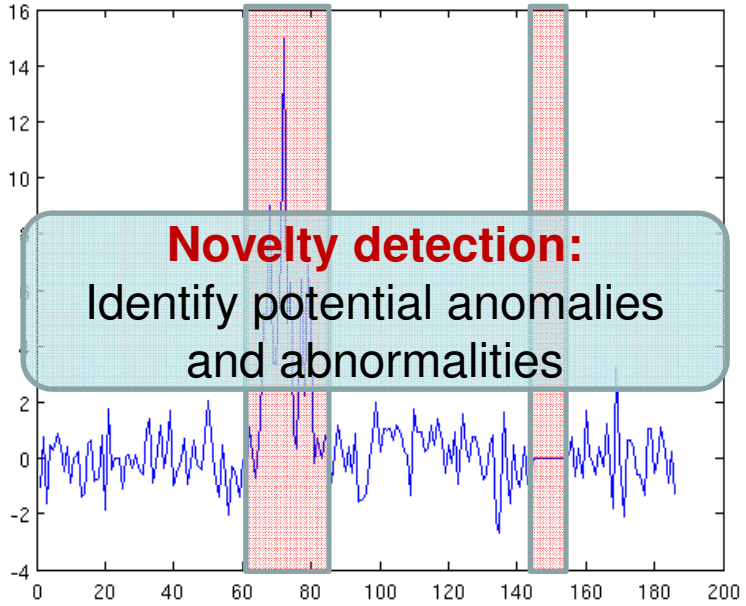




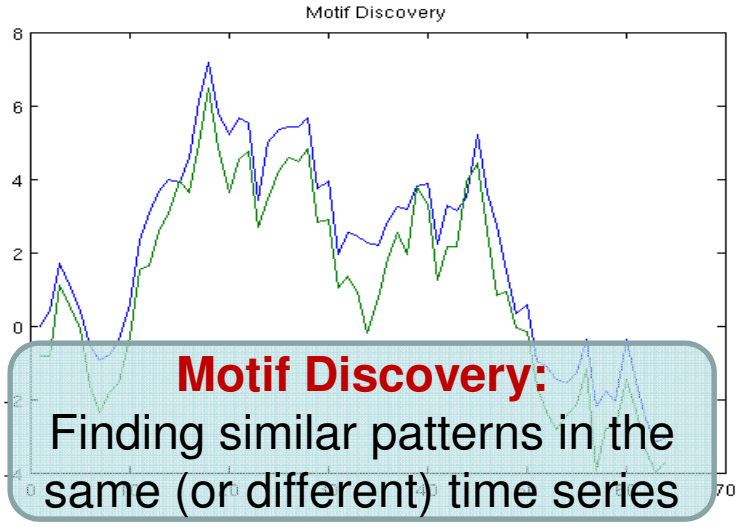
What can we find in time series data?



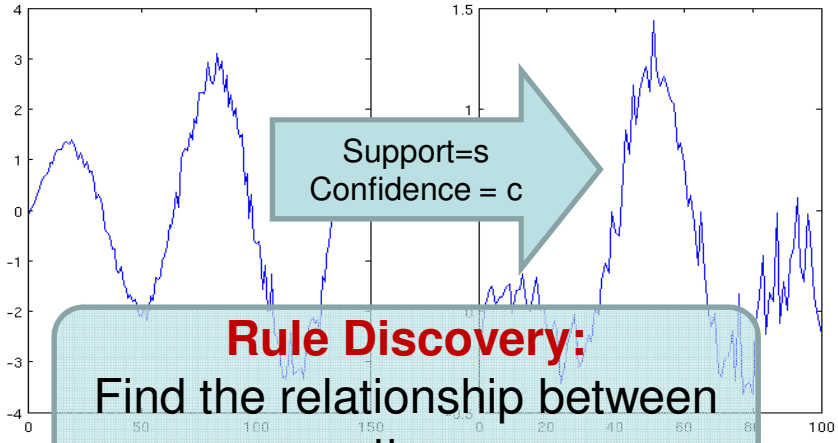
Clustering:
Identify groups of hosts, etc.



Novelty detection:
Identify potential anomalies and abnormalities



Motif Discovery:
Finding similar patterns in the same (or different) time series



Rule Discovery:
Find the relationship between patterns

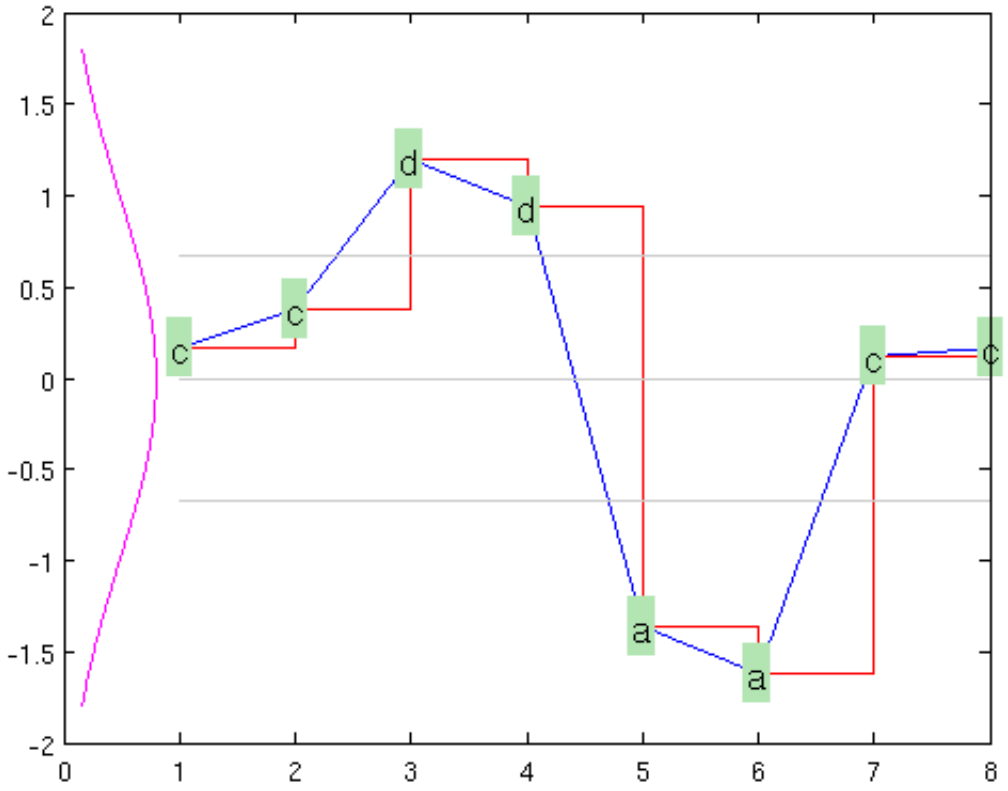
Time series mining

- Much ongoing research in the area of time series analysis, for example:
  - Financial data analysis (we all want to be rich…)
  - Bioinformatics, genomics (i.e. DNA analysis)
  - Medicine (i.e. attempts to build brain-computer interface)
  - …
  - Network traffic analysis (i.e. detecting traffic volume anomalies)
- Look at the current state of the time series data mining
- Develop methods useful for the CINBAD project

Time Series Representations

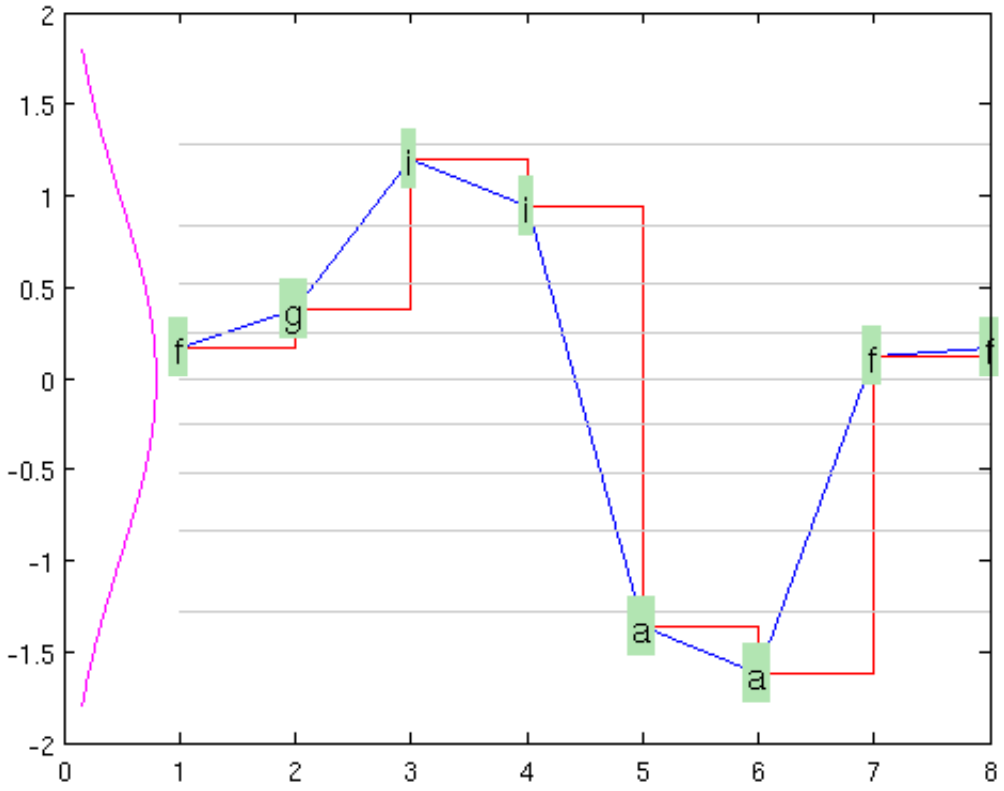
- Representation of data is the key to effective and scalable techniques:
  - Huge amounts of live, streaming data
  - Limited amount of storage
  - Many algorithms require discrete data
- Symbolic Aggregate approXimation
  - Discretisation with meaningful distance measure
  - Dimensionality reduction
  - Output suited for data mining procedures
  - Simple implementation and nearly real-time operation

SAX Example (I)



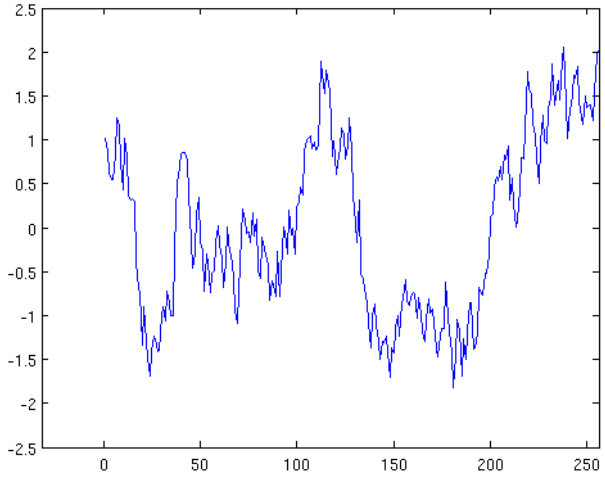
ccddaacc

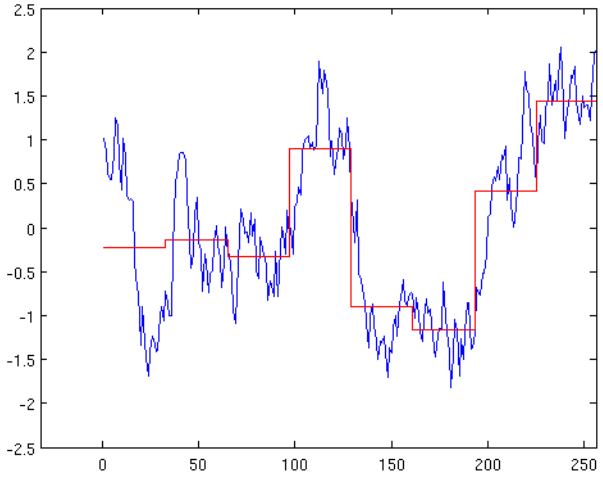
SAX Example (II)

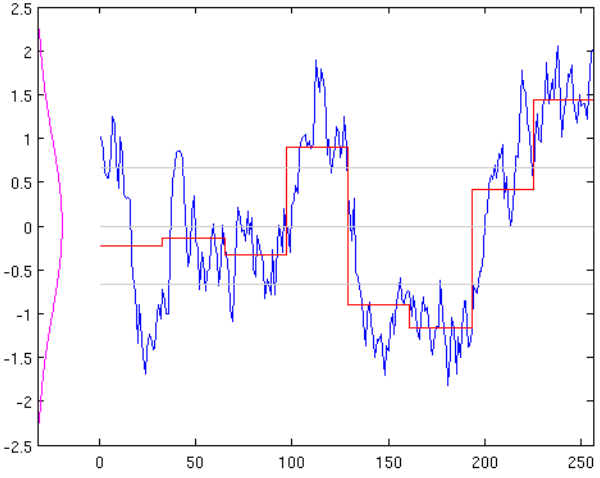


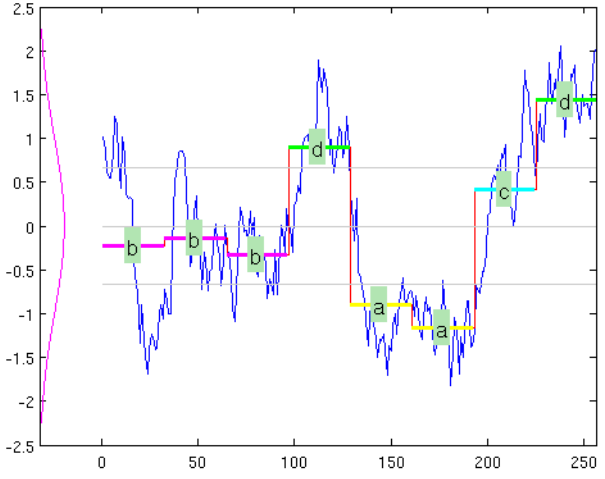
fgiiaaff

SAX Example (III)





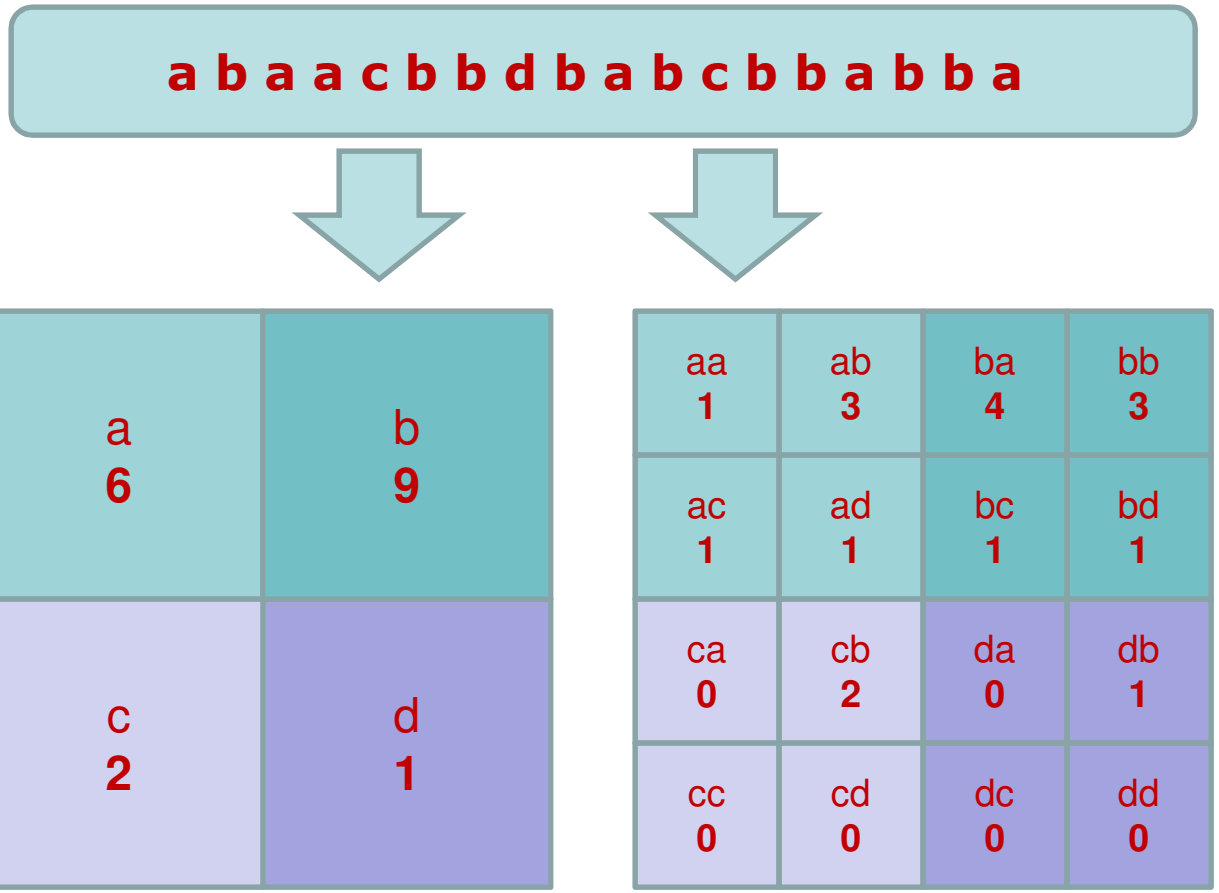




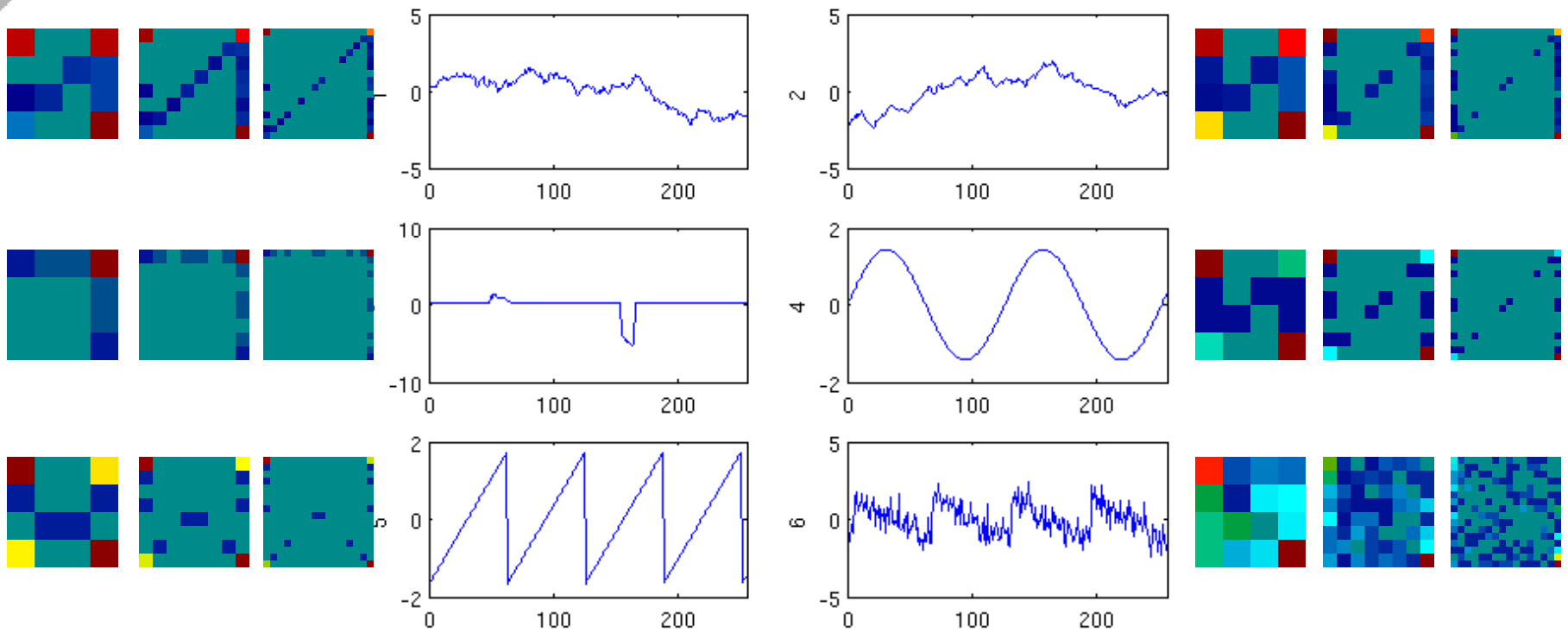
SAX Applications

- Transforming real-valued data into symbolic representation for data-mining algorithms:
  - Text mining and bioinformatics methods
  - Suffix trees/tries, hashing, etc.
  - Increasing speed of real-valued algorithms
    - Dimensionality reduction + easy distance calculation
- Time Series Bitmaps
- Motif Detection with Random Projection
- VizTree analysis

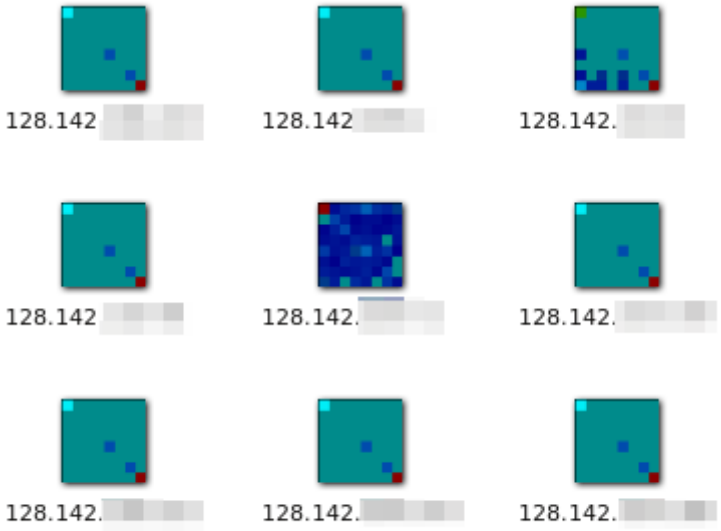
Time Series Bitmaps (I)



Time Series Bitmaps (II)

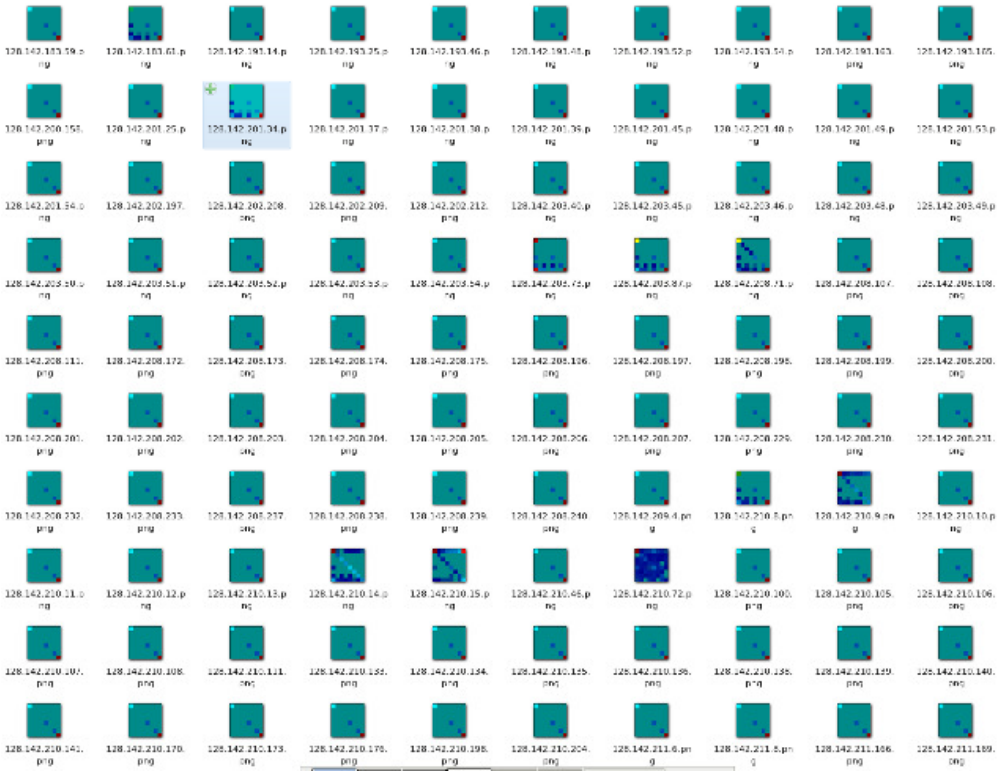


Time Series Bitmaps (III)

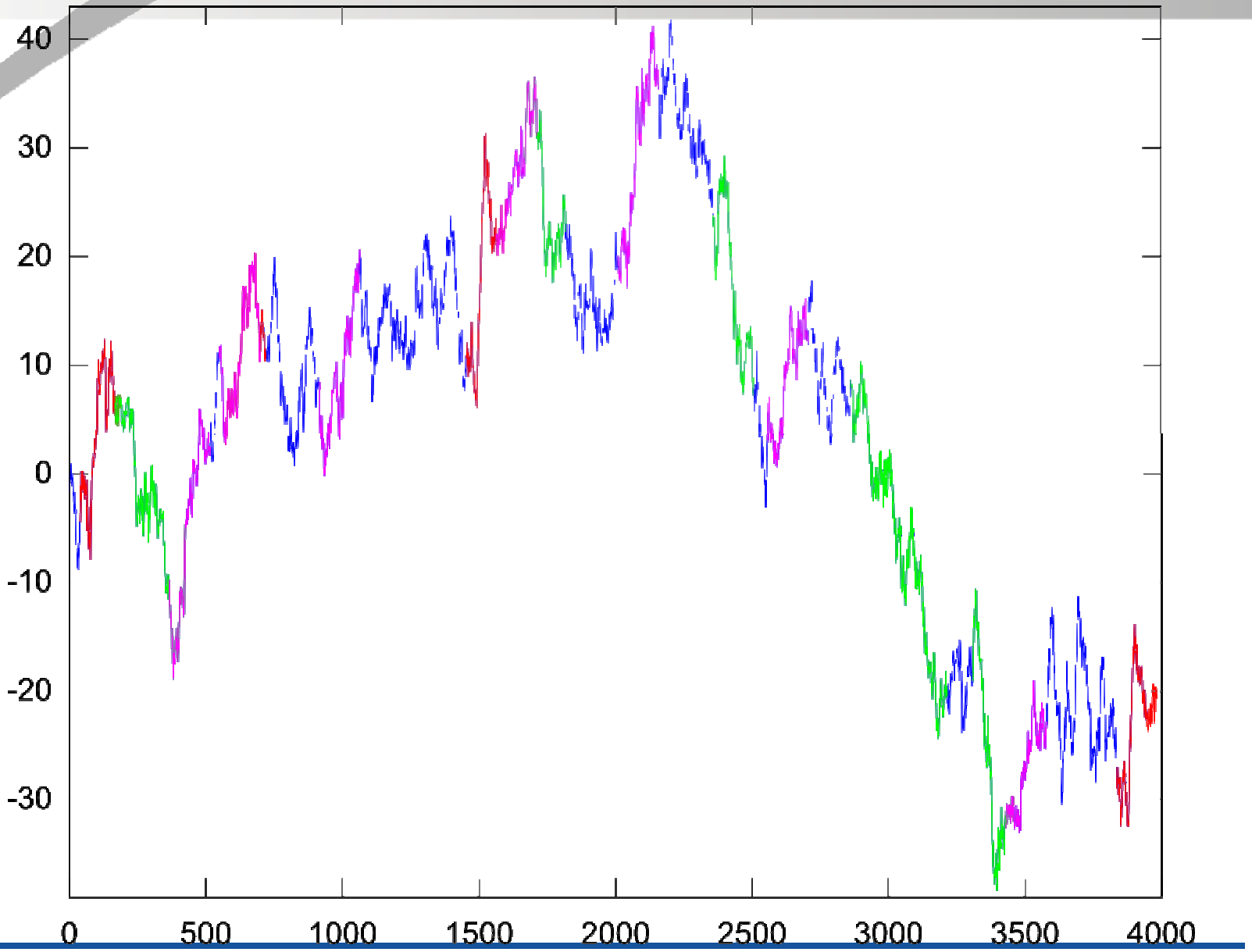


Looks nice…

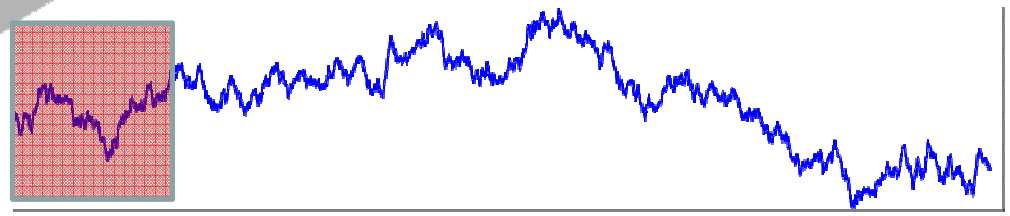
But there are some issues…

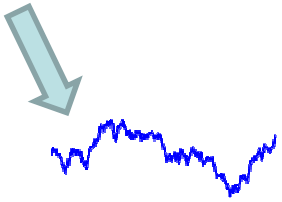


Motif Detection (I)



Motif Detection (II)





b d b a

1	b	d	b	a
2	c	a	c	d
..				
..				
150	b	d	c	a
..				
3947	c	a	c	c

SAX Representation

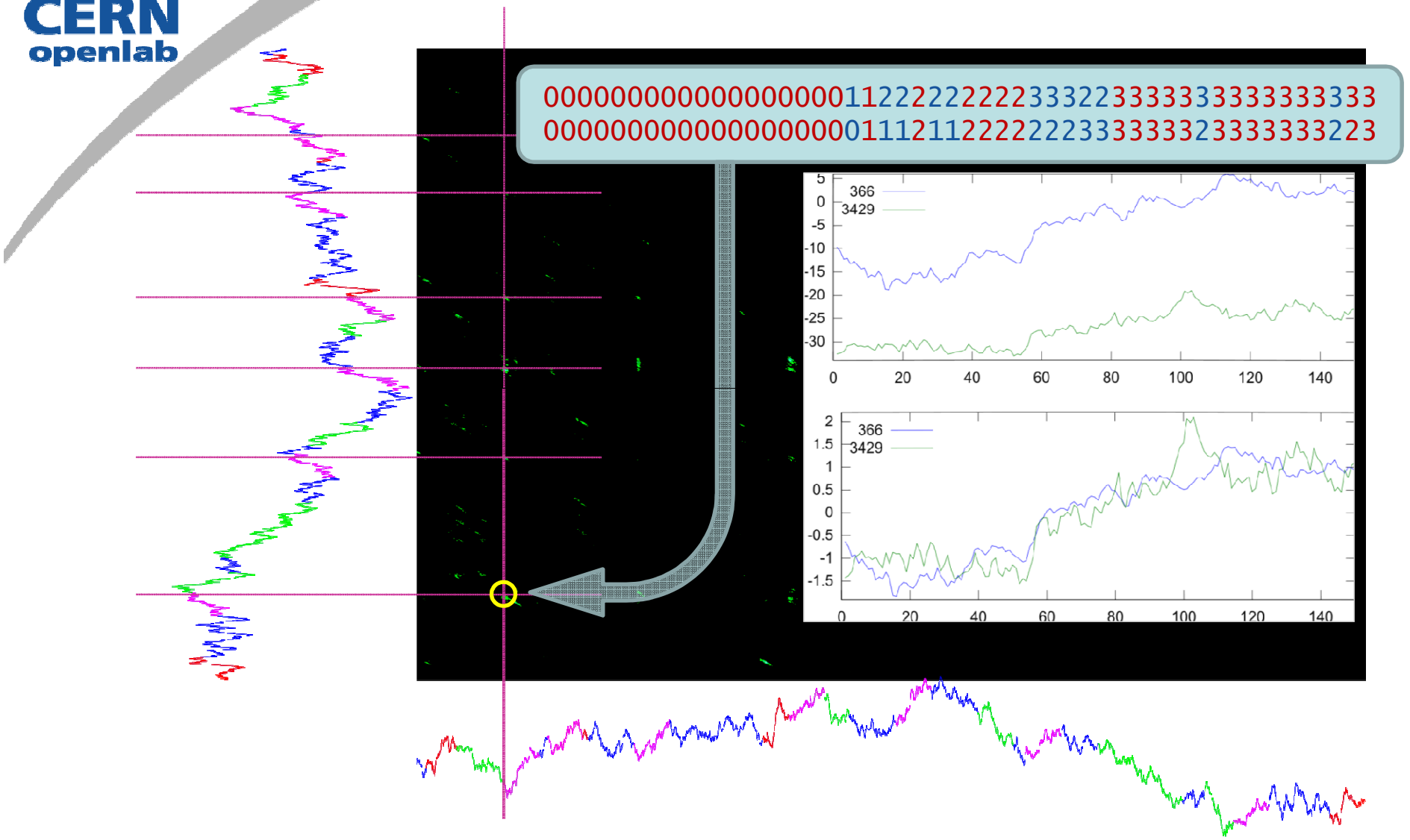
Random Projection

1	b		b	
2	c		c	
..				
..				
150	b		c	
..				
3947	c		c	

Collision Matrix

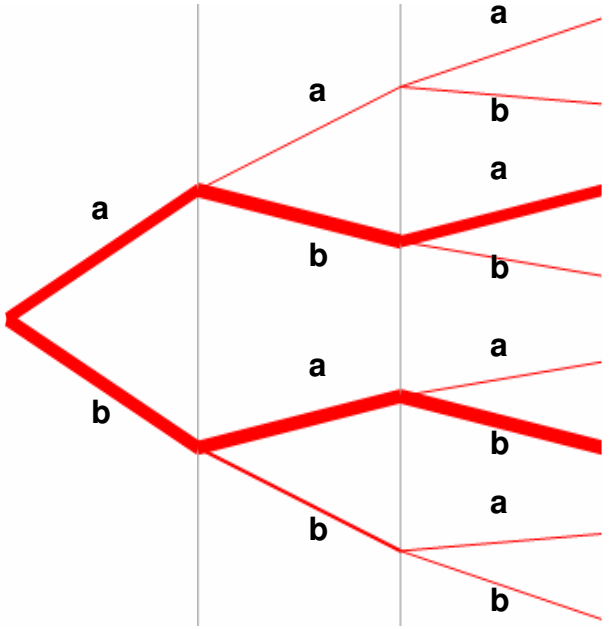
1					
2					
..					
..					
150	1				
3947		2			
	1	2	..	..	3947

Motif Detection (III)



VizTree

- Tool for graphical analysis of time series
- Simple and straightforward to use
- Helps identify motifs and anomalies
- Allows to compare two time series



window size = 3
# of symbols = 3
Alphabet size = 2

Picture from **VizTree** presentation by ***Huyen Dao and Chris Ackermann***
VizTree demo at: http://cs.gmu.edu/~jessica/viztree/viztree_demo.htm

Conclusions

- Still far away from parameter-free technique:
  - Sliding window size, PAA aggregation size, alphabet size, …
- SAX seems a promising way to pre-process the time series
- We are investigating other recent time-series developments
- We want to prepare a technical report summarising our findings.