Motifs identification in Spatial-Time Series

CEFET/RJ

Eduardo Ogasawara

http://eic.cefet-rj.br/~eogasawara

Collaborators

CEFET/RJ: **Murillo Dutra, Riccardo Campisano (MSc.students)**
LNCC: **Fabio Porto**
INRIA: **Florent Masseglia, Esther Pacitti**

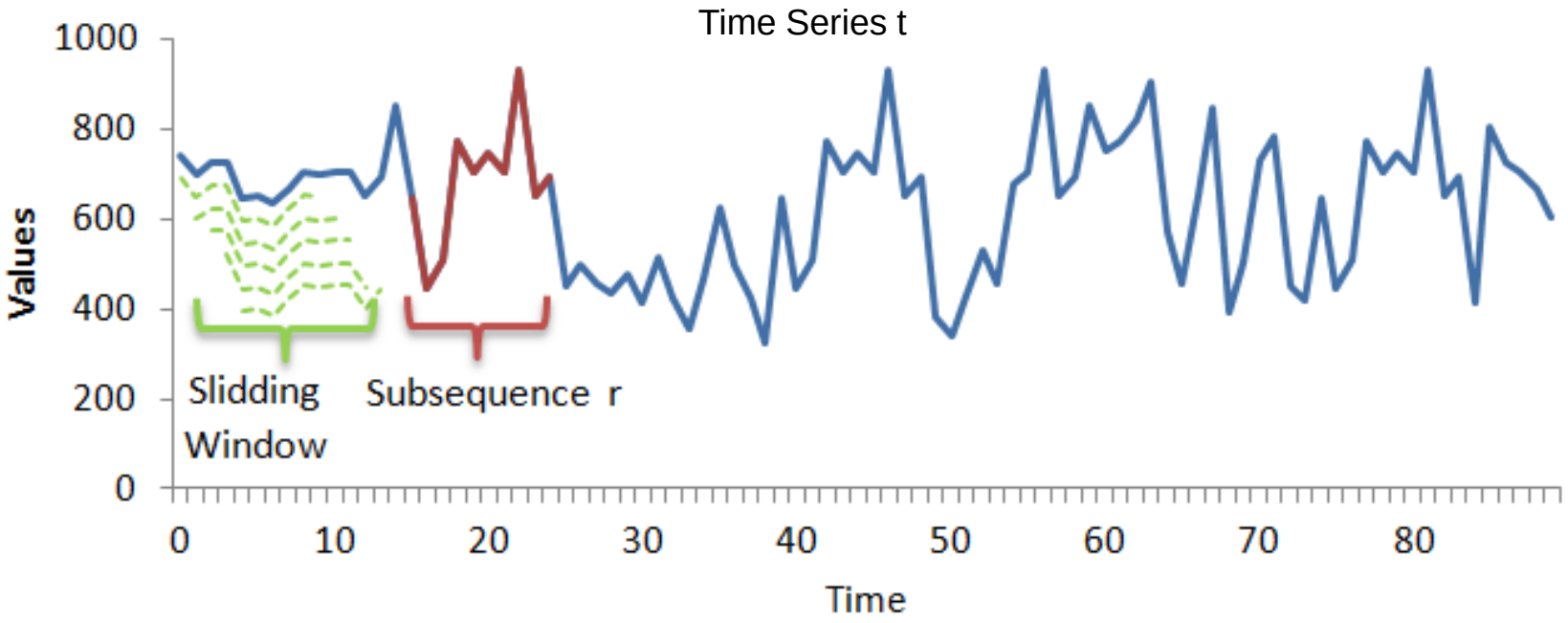
CEFET/RJ

National Laboratory Scientific Computing

informatics mathematics

Time series data mining background

- **Time series** can be defined as collection of observations of a phenomenon along a time-line
- **Subsequence** is a sample of time series
- **Sliding windows** is a set formed by all possible subsequences of a time



Time Series and Sequences

Definition 1. *A **time series** t is an ordered sequence of values in time [1], where each t_i is a value, $|t| = m$ is the number of elements in t, and t_m is the most recent value in t.*

$$t =< t_1, t_2, \ldots, t_m >, t_i \in \mathbb{R}$$

Definition 2. *The p-th **sub sequence** [2] of size n in a time series t, represented as $t^{p,n}$, is an ordered sequence of values $< t_p, t_{p+1}, \ldots, t_{p+n-1} >$, where $|t^{p,n}| = n$ and $1 \leq p \leq |t| - n$.*

$$t^{p,n} = subseq(t, p, n)$$

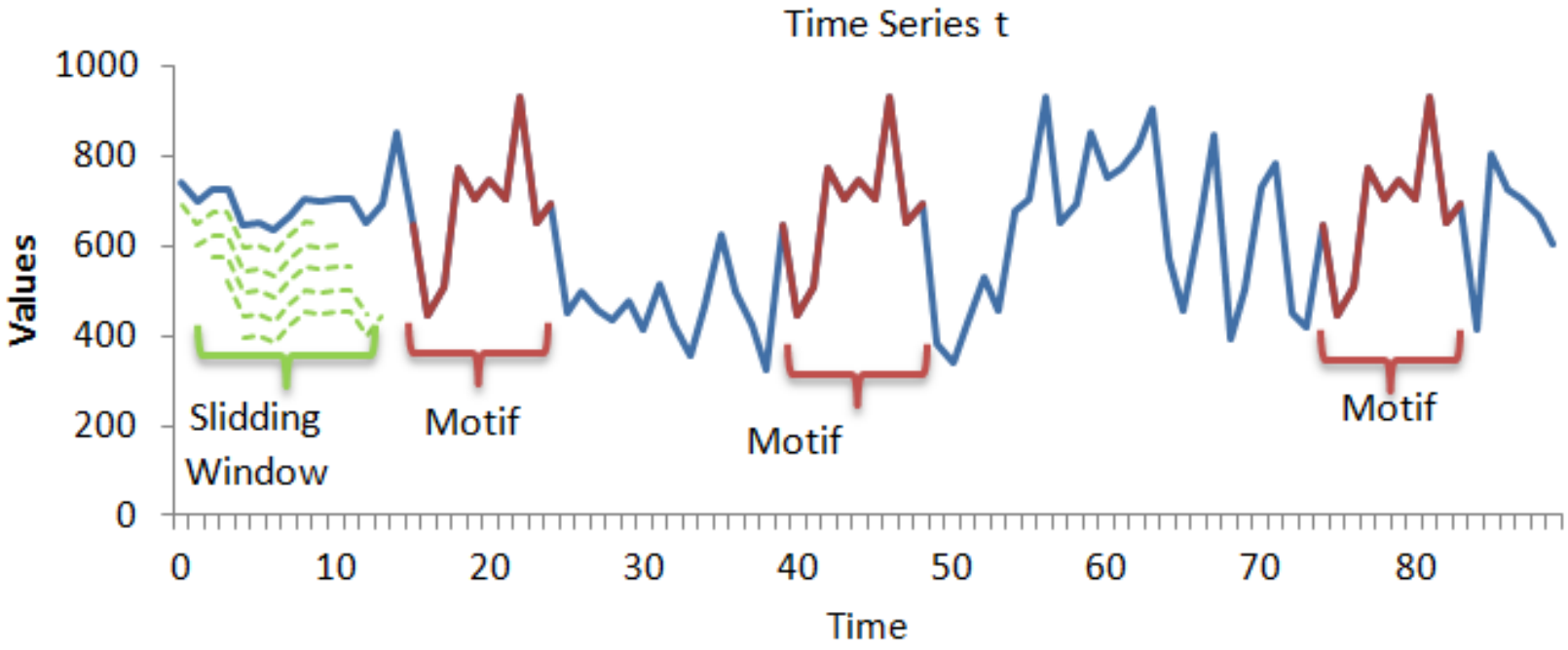
Sliding Window

Definition 3. *A **sliding window** [3] is a function* $sw(t, n)$ *with arguments* t *and* n *that produces a matrix* W *of size* $(|t| - n + 1)$ *by* n *that contains all sub sequences of size* n *of time series* t*. Each line in* W *is a sub sequence of* t *of size* n*. Given* $W = sw(t, n)$, $\forall w_i \in W$, $w_i = t^{i,n}$.

Definition 4. *Let* $q =< q_1, q_2, \ldots, q_n >$ *and* $t =< t_1, t_2, \ldots, t_m >$ *be two time series, such that* $|q| = n$, $|t| = m$, *and* $m > n$. q *is **included** in* t $(q \prec t)$ *iff* $\exists\, w_i \in W, W = sw(t, n) \mid q = w_i$.

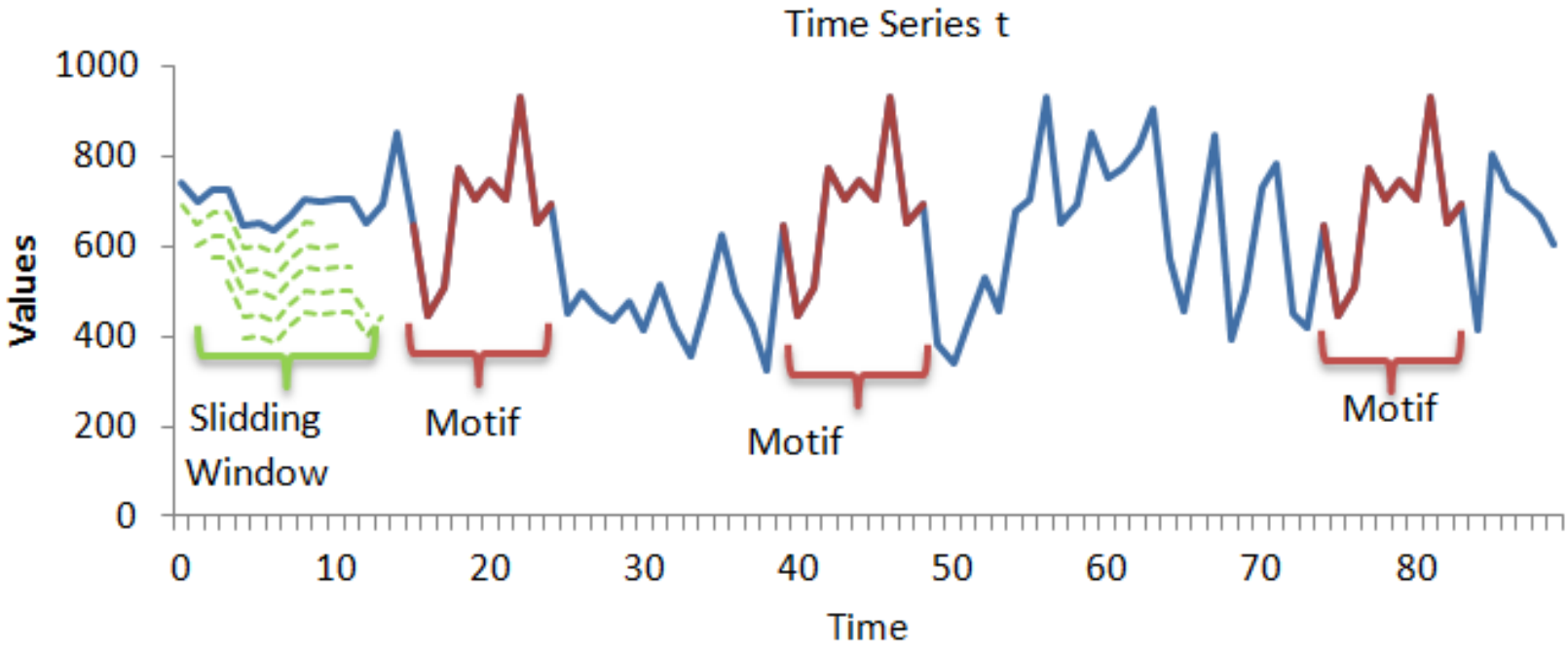
Motif

- **Motif** is a previously unknown subsequence of a time series with relevant number of occurrences in time series data

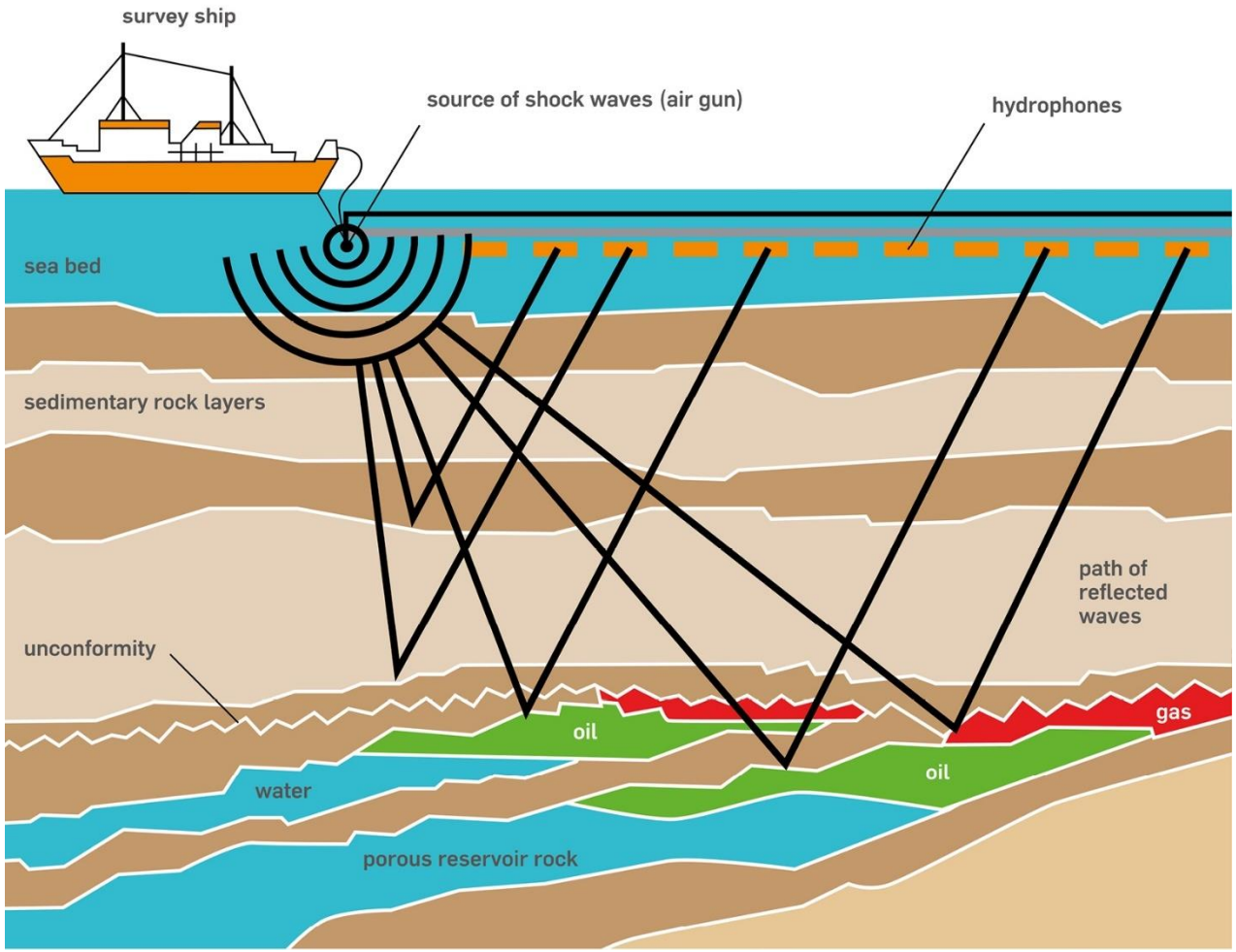


Motif in Time Series

Definition 5. *Given two time series q and t, q is a **motif** [4] with support σ, iff q is included in t at least σ times. Formally, given time series q and t such that $W = sw(t, |q|)$, $motif(q, t, \sigma) \leftrightarrow \exists\, R \subseteq W$, such that $\forall w_i \in R$, $w_i = q \wedge |R| \geq \sigma$.*

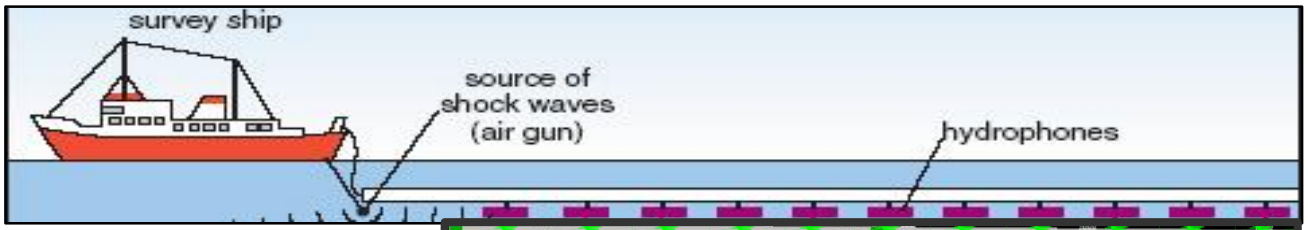


(Scientific)
Seismic Analysis Example



Source: https://krisenergy.com/company/about-oil-and-gas/exploration/

Seismic Traces Analysis



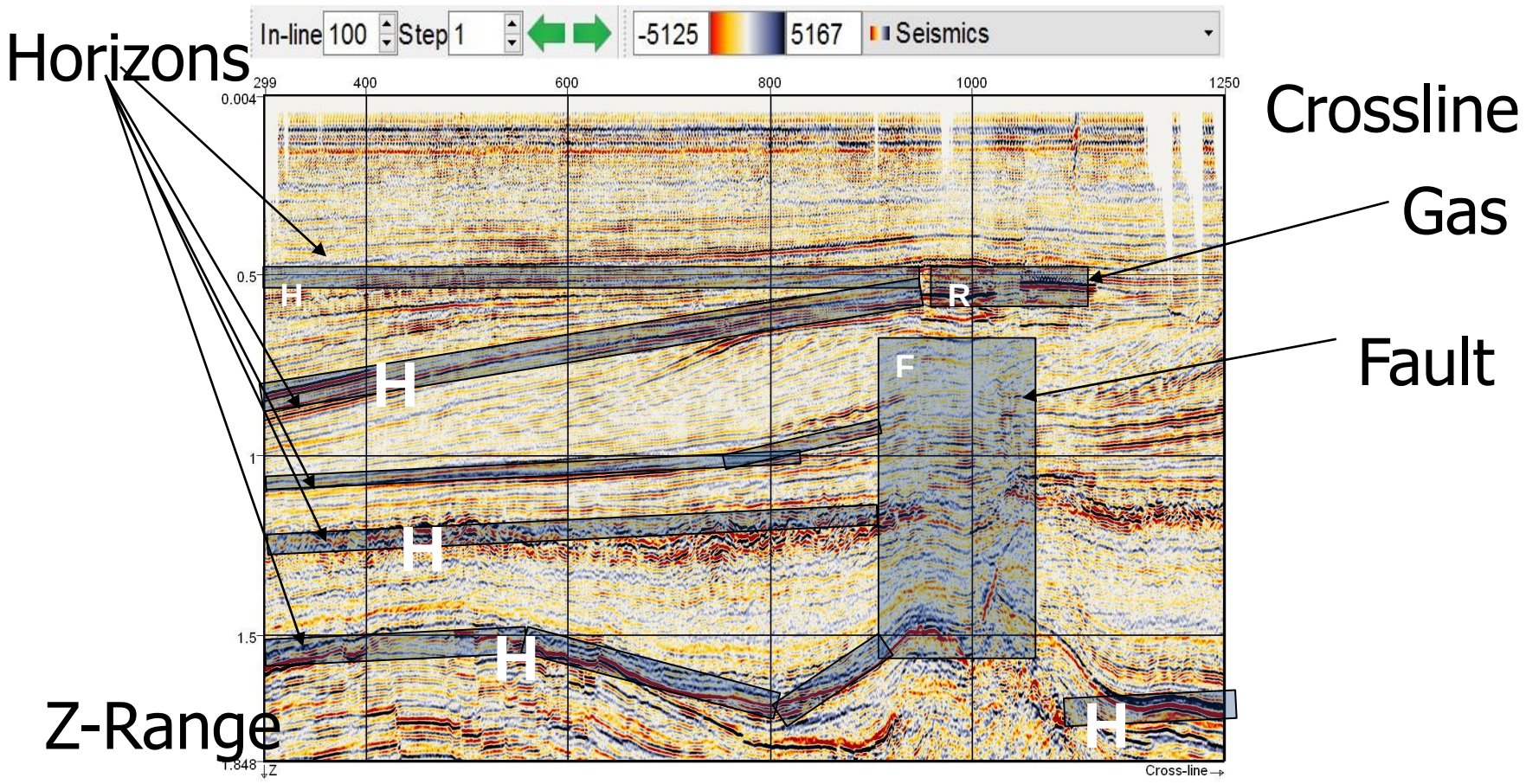
time

space

Each receiver produces a **spatial-time series** related to a specific position of the surface

Seismic Interpretation

- 2D slice of seismic dataset (Inline 401)



(Business/Industrial)
Analysis of Delays in Brazilian Air System

Analysis of delays in airports according to time

(Government)
Buses Stops Analysis

Buses as sensors: Analysis of Trajectory Data
Spatial-time aggregation of buses according to buses stops
Buses Stops as Spatial-Time derived sensors

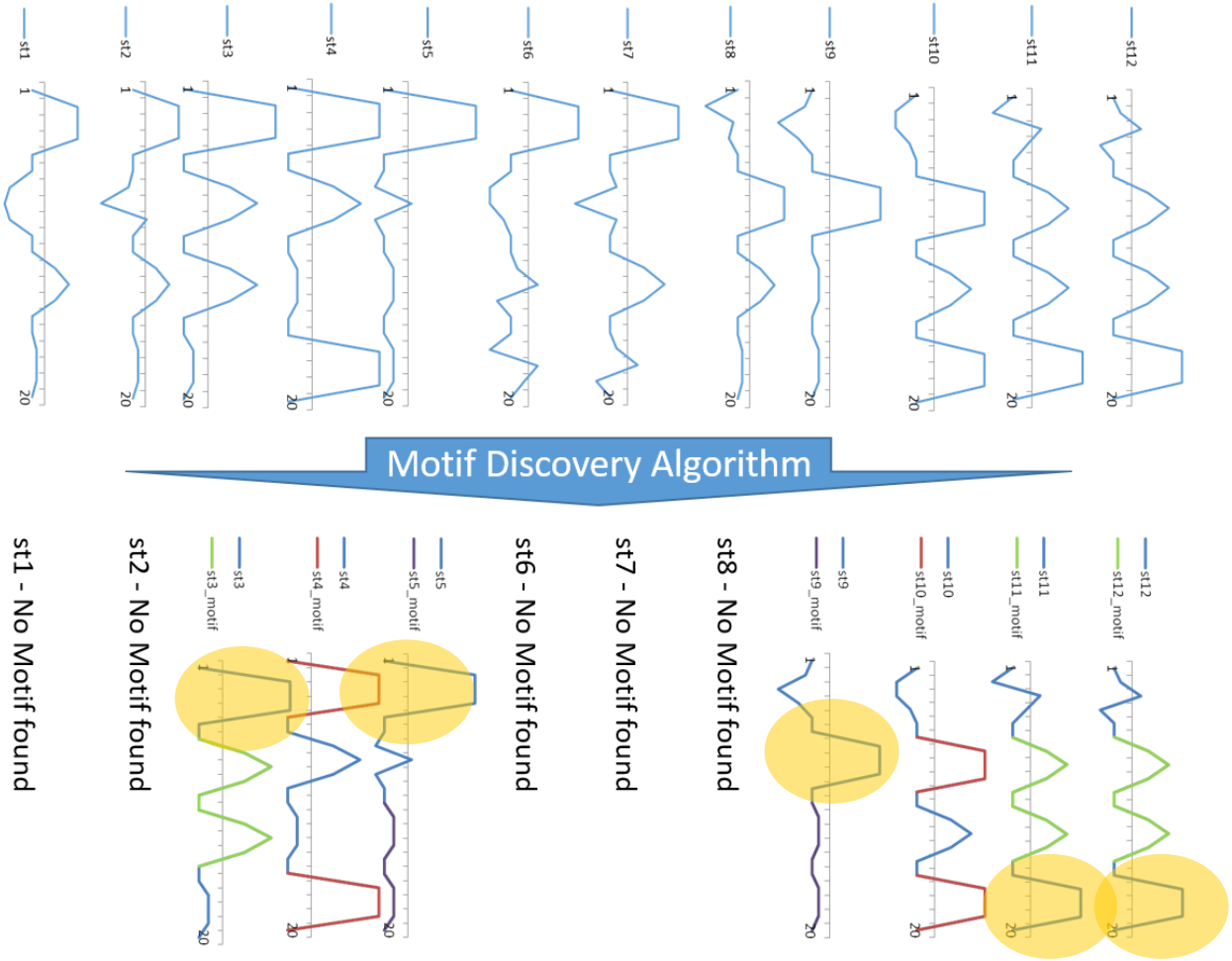
Spatial-Time Series

Definition 6. *A **spatial-time series** s is an abstraction for a time series with an associated position in space. Formally, a spatial-time series s is a time series composed of coordinates x and y and time series $t =< t_1, t_2, \ldots, t_m >$. $s.x$ and $s.y$ are coordinates of s, and $s.t$ is time series for s.*

Definition 7. *A **spatial-time series dataset** (for short, dataset) S is a set of spatial time series $\{s_z\}$. We define $t_{max}(S)$ as the maximum number of observations for all spatial series s_z inside dataset S. Formally, $t_{max}(S) = max(\{|s_z.t|\}), \forall s_z \in S$.*

Problem definition

- **Discover motifs in spatial-time series**
  - Running motif discovery algorithm in single time series:
    - In some cases, no motif is found.
    - Similar shapes in the neighbors are not identified.



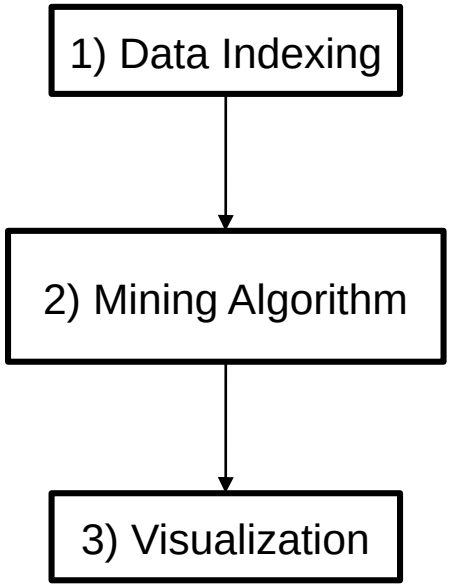
Traditional motif discovery algorithm applied in spatial-time series dataset. (i) red trapeziums and green triangles are identified motifs; (ii) blue trapeziums are not identified and not linked with red ones; (iii) blue triangles are not identified and not linked with green ones; (iv) purple shapes are not identified motifs

Spatial-Time Motif

Definition 10. *Let σ and κ be two support values such that $\sigma \geq \kappa$. A time series q is a **spatial-time motif** in a parallelepiped $pt_{i,j}^{k,n} \in S$ iff q is included at least σ times in $pt_{i,j}^{k,n}$ and $\forall s_z.t^{kn,n} \in \overline{pt}_{i,j}^{k,n}$, q is included in $s_z.t^{kn,n}$, $\overline{pt}_{i,j}^{k,n} \subseteq pt_{i,j}^{k,n}$ and $|\overline{pt}_{i,j}^{k,n}| \geq \kappa$.*

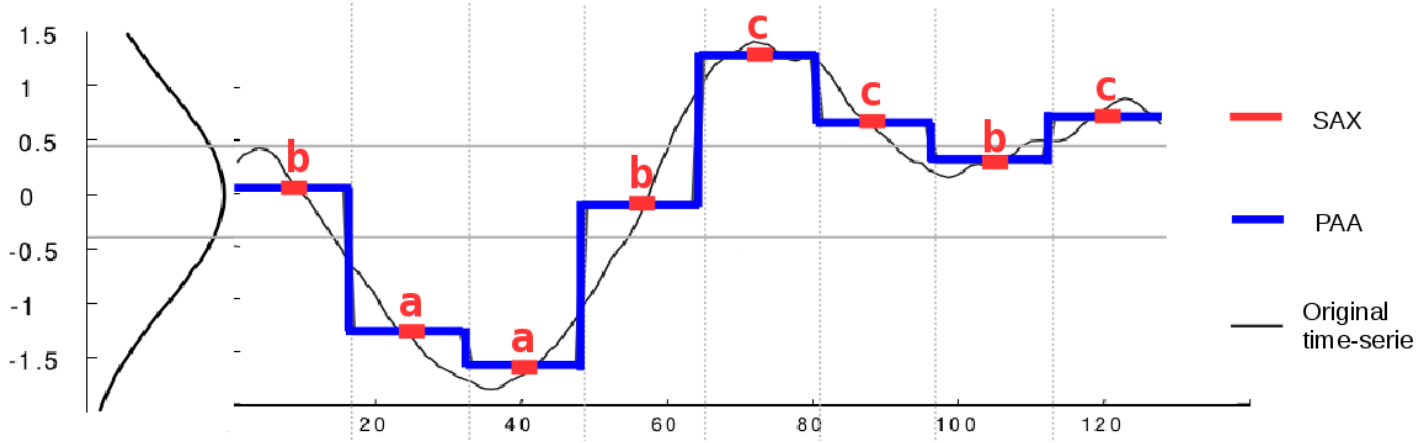
Data Mining Process

- Data Indexing
- Mining Algorithms
  - Combined Series
  - Sequence Mining
- Data Analytics & Visualization



Data Indexing

- Time series contains continuous (non discrete) values
- Is not possible to find patterns performing an exact match between items of such sequences
- SAX indexation [Lin et al., 2003] was applied to convert continuous values to a discrete symbolic representation



SAX Transformation

Translated figure
of *Inline* 100

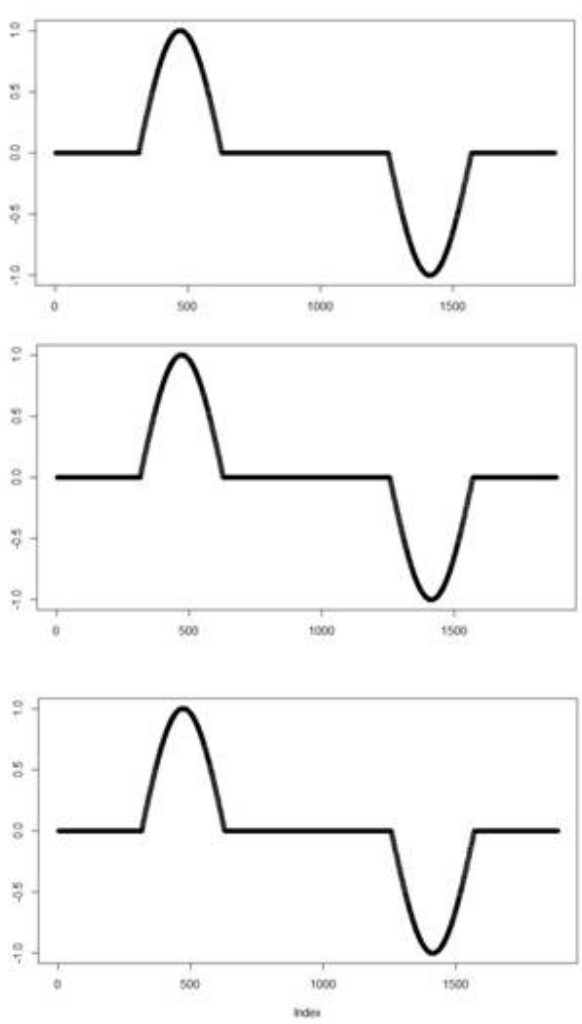
	X13	X14	X15	X16	X17	X18
15	180	106	283	648	482	-926
16	-662	-1468	-1762	-981	-107	-51
17	0	0	0	0	0	0
18	814	775	263	-986	-2138	-2763
19	-604	-1261	-1793	-1722	-965	-227
20	0	0	0	0	0	0
21	0	0	0	0	0	0
22	0	0	0	0	0	0
23	0	0	0	0	0	0
24	0	0	0	0	0	0
25	-1486	-2471	-2398	-1414	-441	-196
26	0	0	0	0	0	0
27	929	1141	508	-1203	-2278	-2824
28	-167	-1250	-2378	-2343	-1496	-705
29	0	0	0	0	0	0
30	347	265	132	-582	-1577	-2569
31	-632	-1556	-2231	-1993	-1207	-589
32	0	0	0	0	0	0
33	1213	1785	1485	-620	-3000	-4203
34	-882	-2066	-2936	-2947	-2220	-1214

	X13	X14	X15	X16	X17	X18	X19	X20
15	n	n	o	p	o	j	e	k
16	k	h	g	j	m	m	m	m
17	m	m	m	m	m	m	m	m
18	q	q	o	j	f	d	b	c
19	k	i	g	g	j	m	n	n
20	m	m	m	m	m	m	m	m
21	m	m	m	m	m	m	m	m
22	m	m	m	m	m	m	m	m
23	m	m	m	m	m	m	m	m
24	m	m	m	m	m	m	m	m
25	h	e	e	h	l	m	l	l
26	m	m	m	m	m	m	m	m
27	q	r	p	i	e	d	a	a
28	m	i	e	e	h	k	n	n
29	m	m	m	m	m	m	m	m
30	o	o	n	k	h	e	b	d
31	k	h	f	f	i	k	m	n
32	m	m	m	m	m	m	m	m
33	r	t	s	k	d	b	a	b
34	j	f	d	d	f	i	m	n

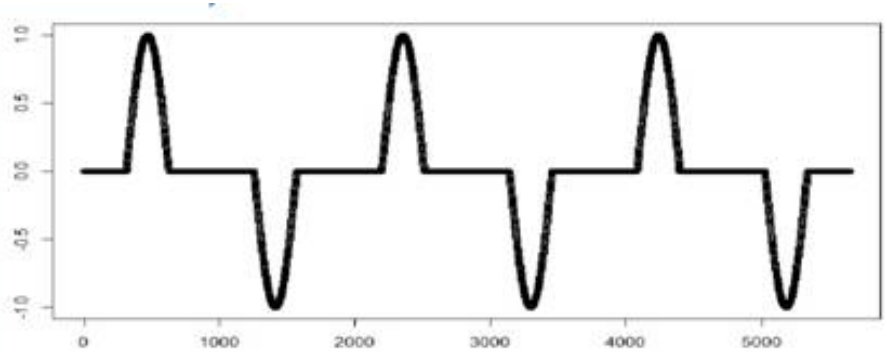
Portion of original seismic dataset

SAX converted data

Combined Spatial-Time Series



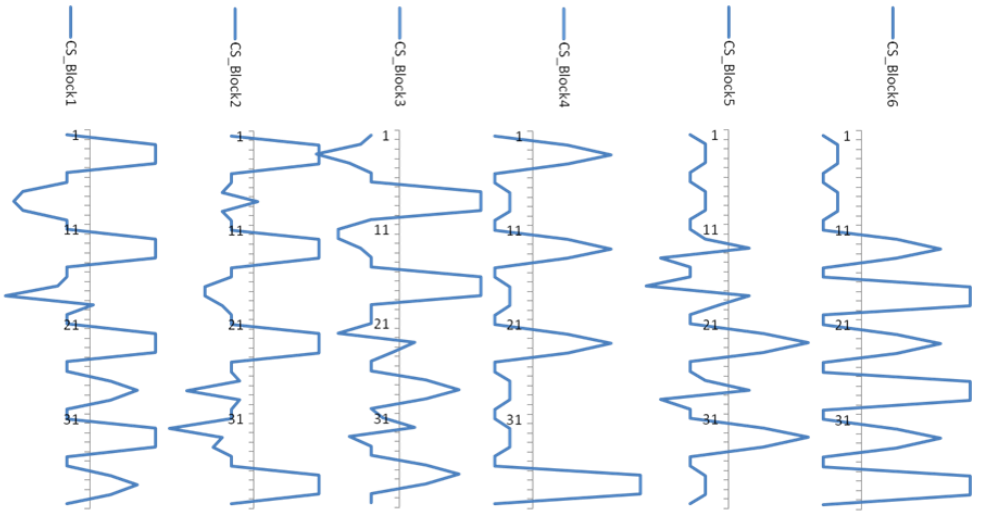
Combined time series



Combined Series Approach

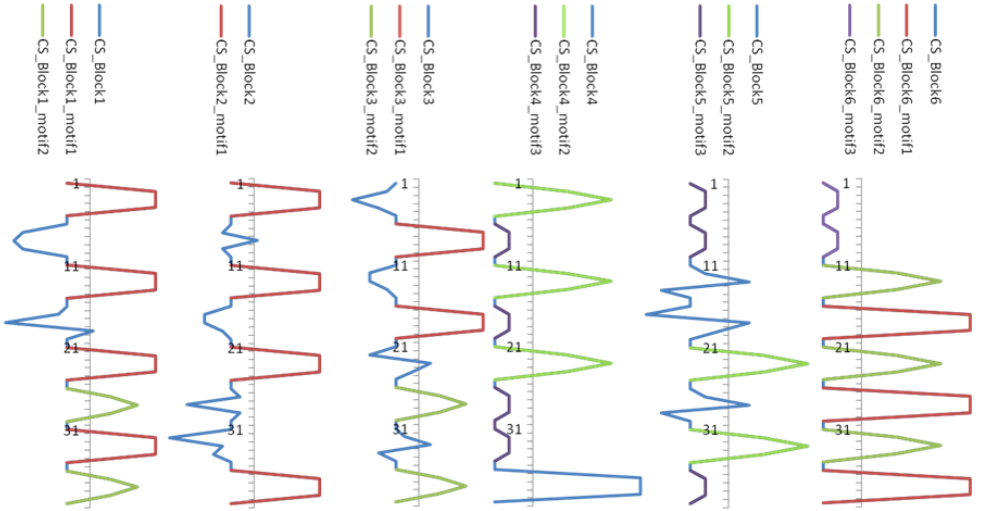
- Run the motif discovery algorithm for single time series

Combined Series

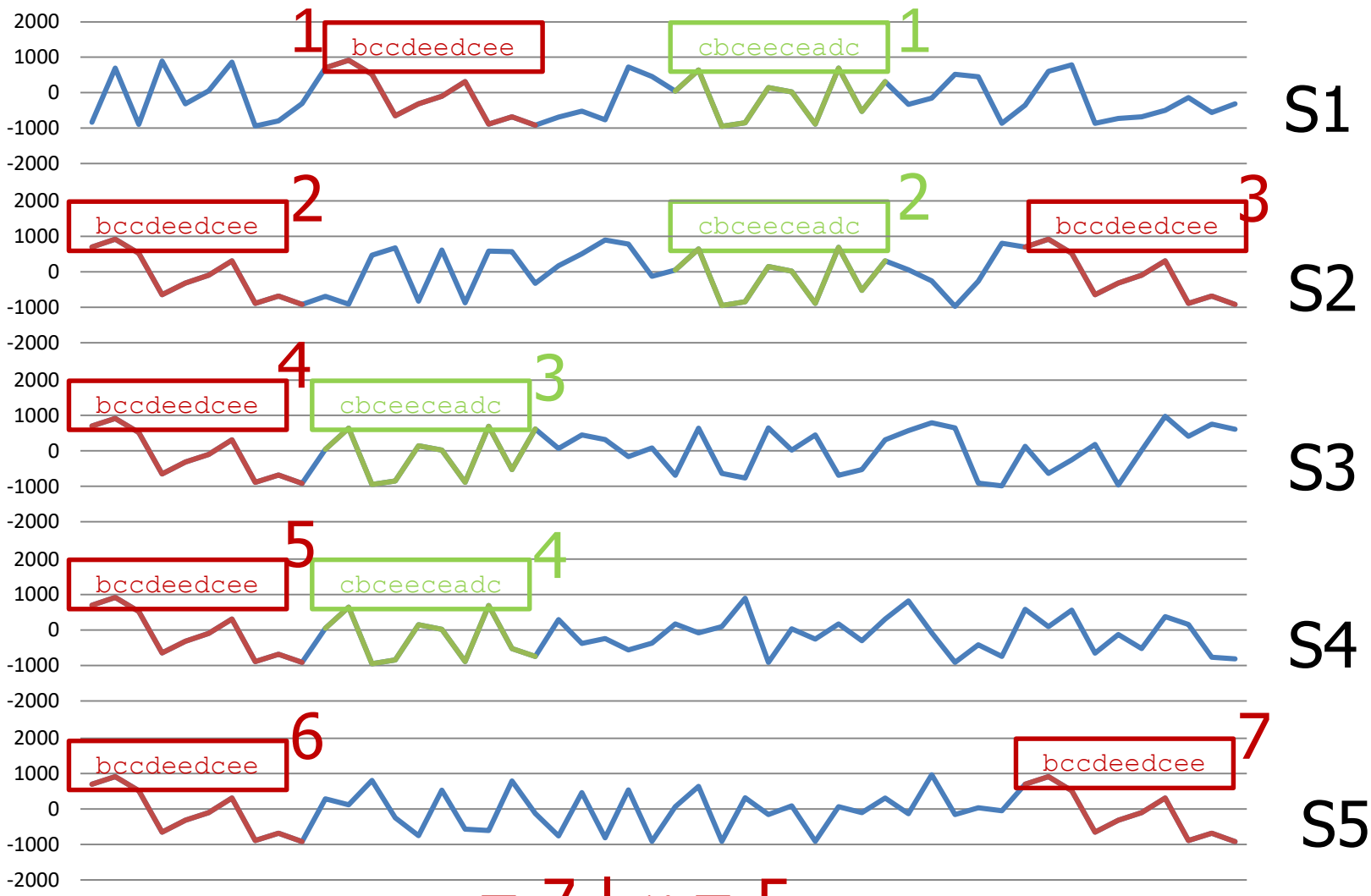


Motif Discovery Algorithm

Candidates motifs found in combined series



Identified Motifs in Original Spatial-Time Series



$\sigma = 7 \mid \kappa = 5$

$\sigma = 4 \mid \kappa = 4$

Spatial-Time Motif Ranking

- Rank identified spatial-time motifs

Motif	Word	σ	κ	Spatial-Time Motif
Motif 1	bccdeedcee	7	5	Yes
Motif 2	cbceeceadc	4	4	No

σ: total motif occurrences in block

κ: number of series that occurs the identified motif

Restriction Parameters:

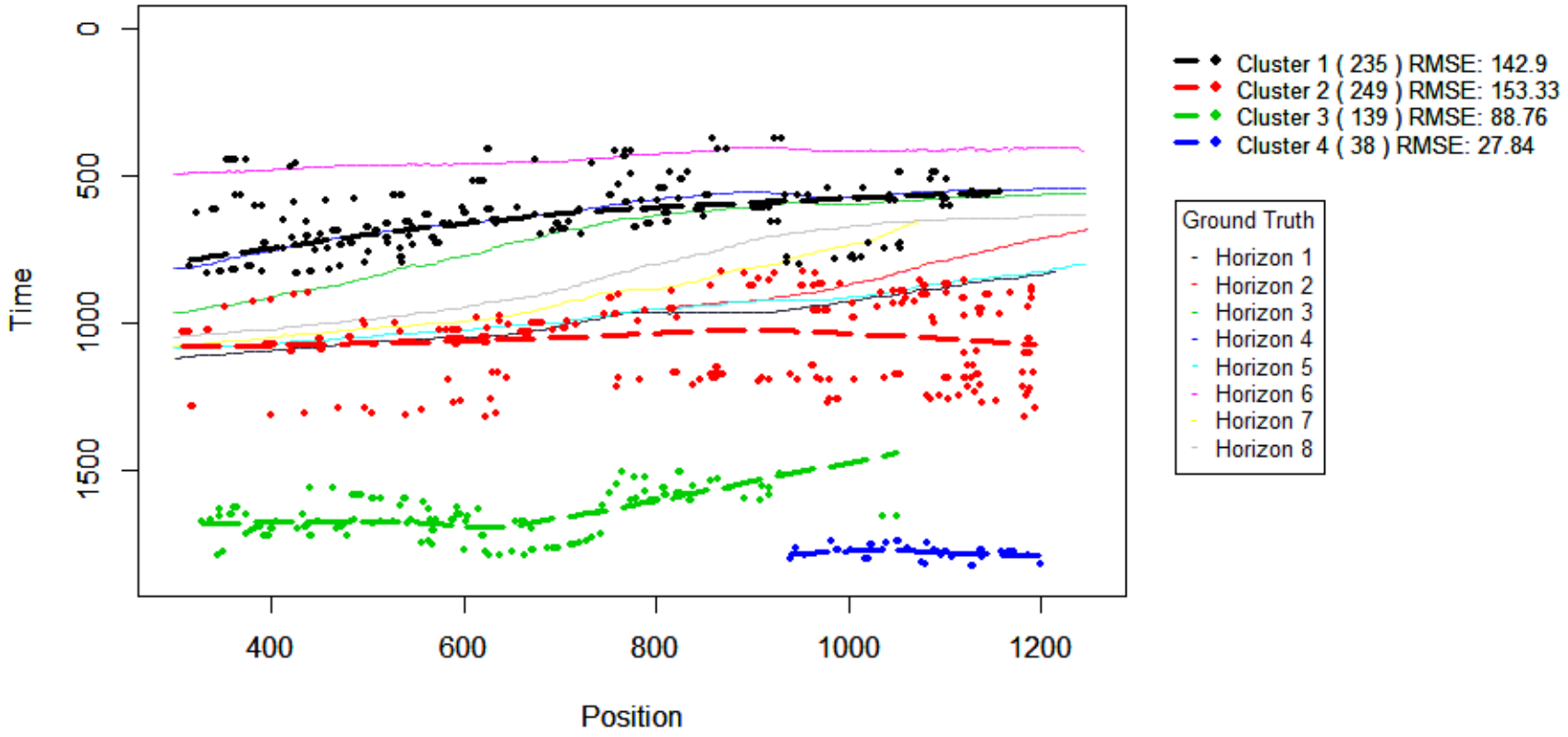
$\sigma \geq 5$

$\kappa \geq 3$

Algorithm

1: **function** STMOTIF(b, sw, w, a, bs, bt)
2: $b_i \leftarrow partition(b, bs, bt)$
3: **for each** $b_i \in b$ **do**
4: $t \leftarrow combine(b_i)$
5: $CSTM \leftarrow identify(t)$
6: $STM \leftarrow STM \cup constraintST(CSTM)$
7: **end for**
8: $rankSTM = aggregate(STM)$
9: return $rankSTM$
10: **end function**

Identified Motif x Ground-truth horizons

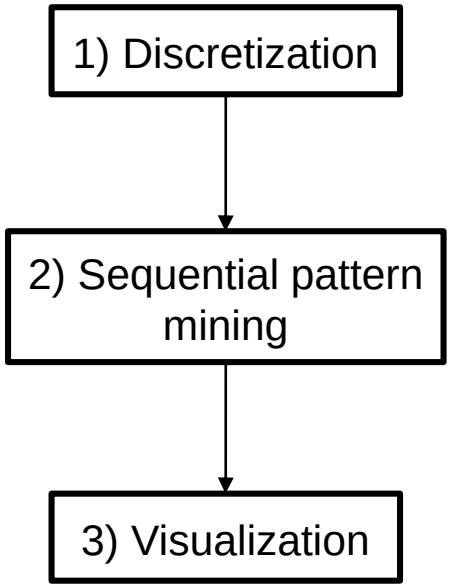


- **Black dashed line** is very close to **blue continuous line** (ground truth horizon).
- **Red dashed line,** is very close to **black continuous line** (ground truth horizon).
- **Green dashed line** and **Blue dashed lines** indicates that may exist other regions that were not previously mapped in ground-truth.

Approach 2: Sequence Mining

- Sequence pattern mining is used successfully to obtain insight from large volume of transactional databases.

- Scope of this work is the use of such technique to discover sequential patterns on seismic spatial-time series:
  - indexing technique used to discretize the input
  - adapted algorithm implemented to retrieve discovered patterns positions
  - results are presented over original seismic trace images to better evaluate the quality of results



Generating candidates

- The principle of Apriori applied to the candidate generation is still valid for sequences that are frequent in a specific range:
  - a sequence is frequent in a spatial-time dataset only if all its sub-sequences are frequent in it

- Verifying that, the candidates can be generated similarly to the algorithm defined in [Agrawal et al., 1995]:
  - <a,b,c,x> and <y,a,b,c> can be joined to obtain the candidates <y,a,b,c,x>

Sequential pattern mining

- A sequential pattern algorithm was adapted to performs sequence mining in spatial-time series dataset.
- It uses of the Apriori Principle: if a set of items is frequent, any of its subset is frequent too.
- itemsets of size k-1 → itemsets of size k

Algorithm 1 Sequential pattern mining on Spatial-Time Series

```
1:  function FINDFREQUENTSEQUENCES(d_seq, min_sup, max_stretch)
2:      freq_items ← getFrequentItems(d_seq, min_sup)
3:      freq_k_seq ← convertToSequences(freq_items)
4:      while count(freq_k_seq) > 0 do
5:          all_freq_seq ← all_freq_seq ∪ freq_k_seq
6:          cand_k_seq ← joinSequences(freq_k_seq)
7:          freq_k_seq ← pruneCandidates(
8:              cand_k_seq, d_seq, min_sup, max_stretch)
9:      end while
10:     return all_freq_seq
11: end function
```

Visualization

- For each detected frequent sequence the algorithm provide all the positions where the sequence was encountered.

- With this associated positions is possible to visually represents the match positions of each sequence and this allow a supervised evaluation of the quality of the results.

SBBD Paper 2016

alphabet-size: 25

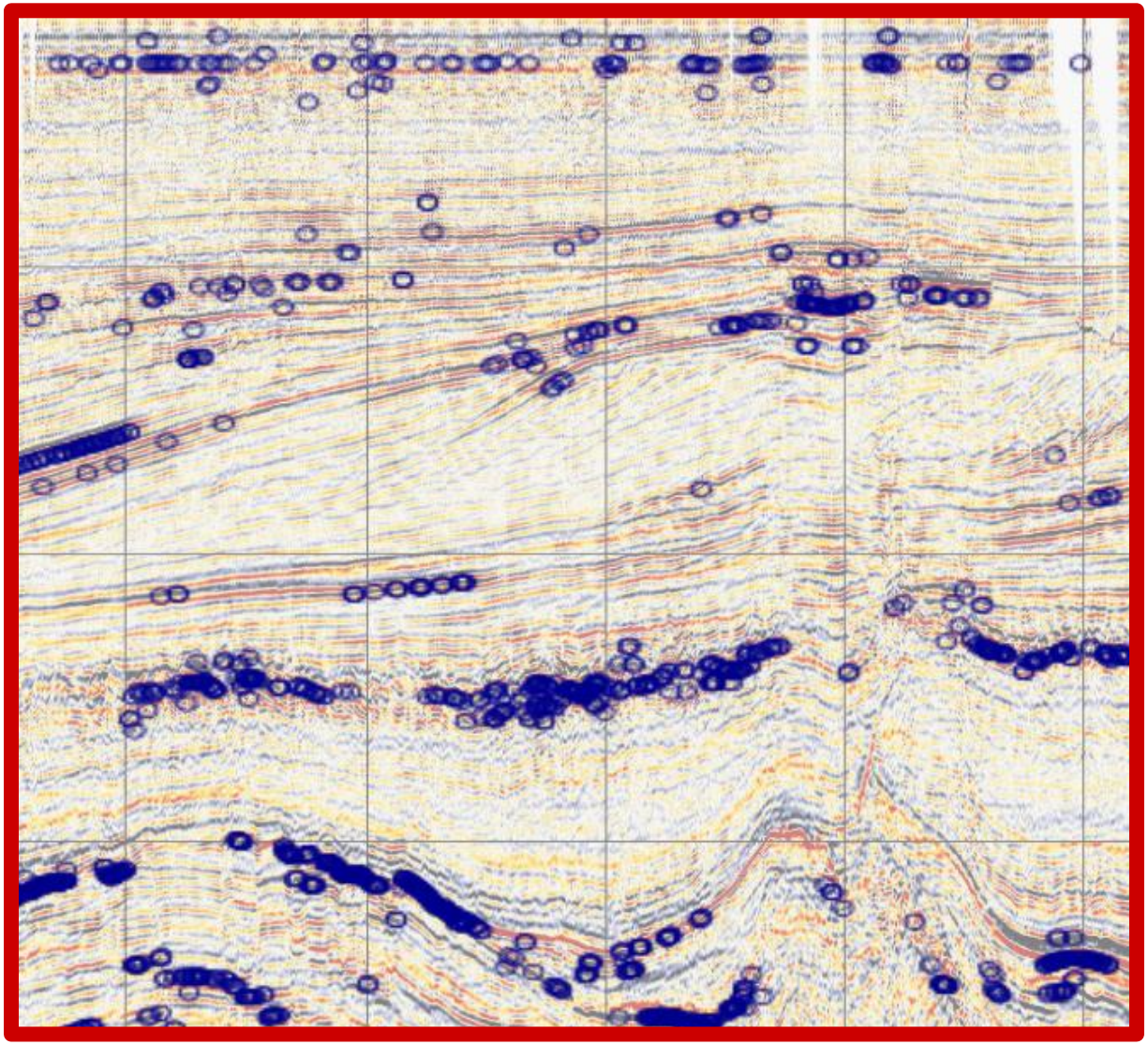
min-support: 70%

max-stretch: 2

Sequence:

<a,a,y,y,>

Several horizon segments detected



Inline 100

Motifs identification in Spatial-Time Series Applications & Methods

CEFET/RJ

Eduardo Ogasawara

http://eic.cefet-rj.br/~eogasawara