CEFET/RJ

Data Management and Analysis of Spatial-Time Series

(Gerência e Análise de Séries Espaço-Temporais)

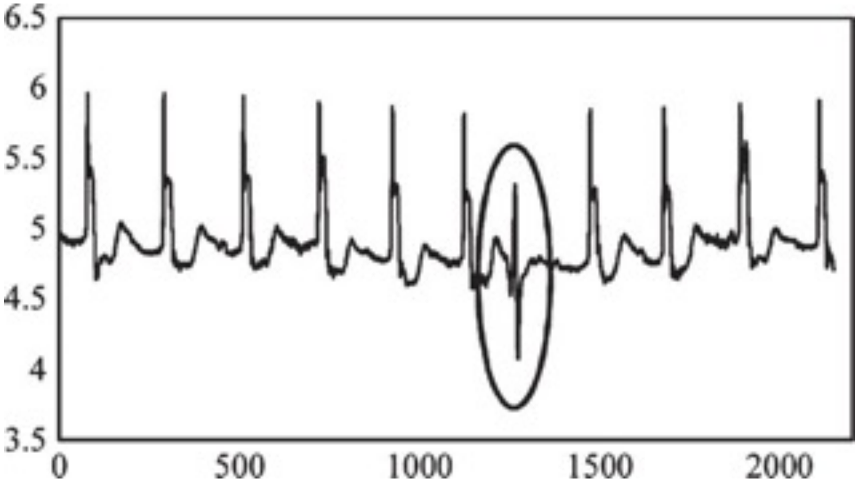
V EIC Workshop

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

Why the study of time series & spatial-time series is import?

Many phenomena are modeled in space-time

Anticipate decision-making regarding forthcoming events



Big Data, IoT, Deep Learning, HPC, and DISC

Knowledge Discovery in Time Series

- Big Data
  - Data deluge (volume and velocity)
  - Different data models (variability)
  - Science: astronomy, seismic
  - Business/Persons: IoT, flights
  - Government: smart cities, urban mobility
- Challenges for knowledge discovery
  - Data management
    - Data preprocessing
    - Workflows
  - Data analysis
    - Prediction / classification
    - Pattern identification

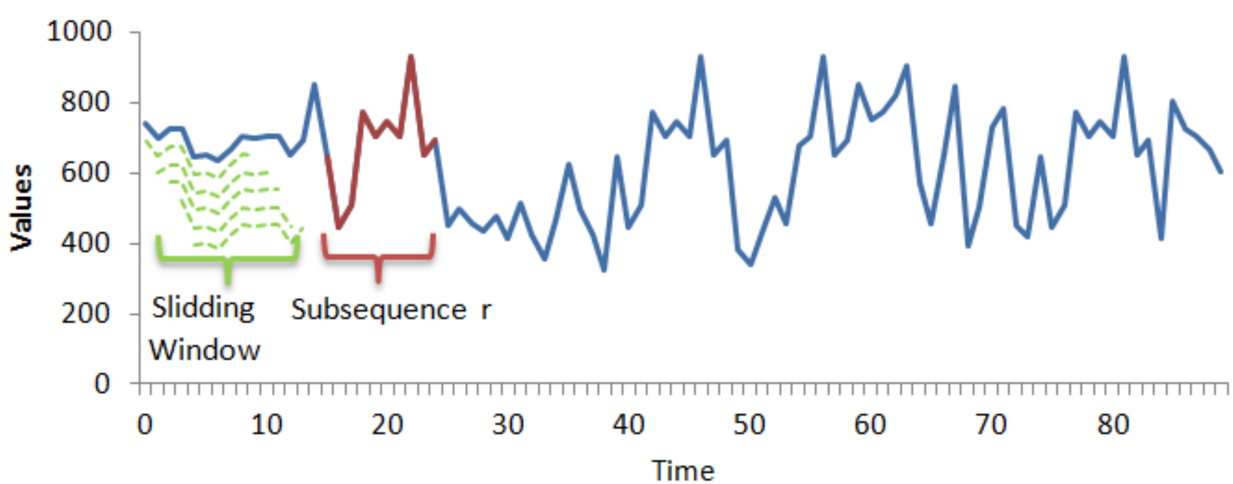
Time series definitions

Time Series: Let $t = <v_1, v_2, \cdots, v_n>$ be a time series, *i.e.*, a **sequence** of items, where $|t| = n$ is the number of items in t. A time index j is an integer value between 1 and n that is related to item v_j.

A **time interval** (or simply **interval**) $i = (i_s, i_e)$ is defined by a start time i_s and an end time i_e. The length of an interval i is given by: $|i| = i_e - i_s + 1$. Given a interval i, a sequence $s = <w_1, w_2, \cdots, w_k>$ is a **subsequence** of another sequence $t = <v_1, v_2, \cdots, v_n>$: $s = subseq(t, i)$ iff $i_s \geq 1 \wedge i_e \leq n$, $|i| = k$ and $\forall j \in [1..k], w_j = v_{i_s+j-1}$.

A **sliding window** is a function $sw(t, n)$ that produces a matrix W of size $(|t| - n + 1)$ by n that contains all sub sequences of size n for the time series t. Each line in W is a subsequence of t of size n.

Given $W = sw(t, n), \forall w_k \in W, w_k = subseq(t, (i_k, i_{k+n-1}))$

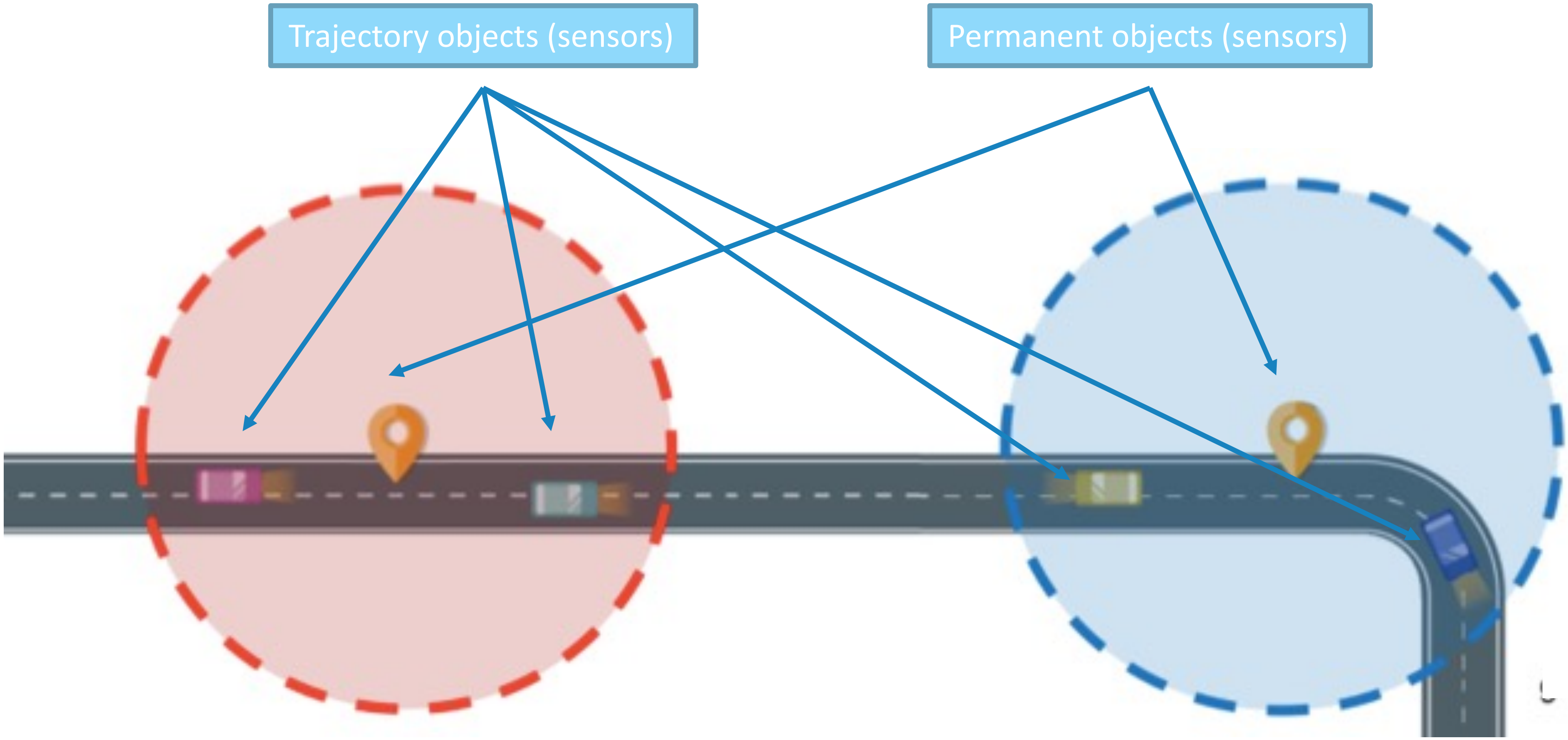


Spatial-time series

Let $P = \{p_1, p_2, \ldots, p_m\}$ be a set of positions, a **spatial-time series** d is a couple (p, t) where $p \in P$ is a position and t is the associated time series.

A **spatial-time series dataset** D is a set of spatial-time series $\{d_j\}$.

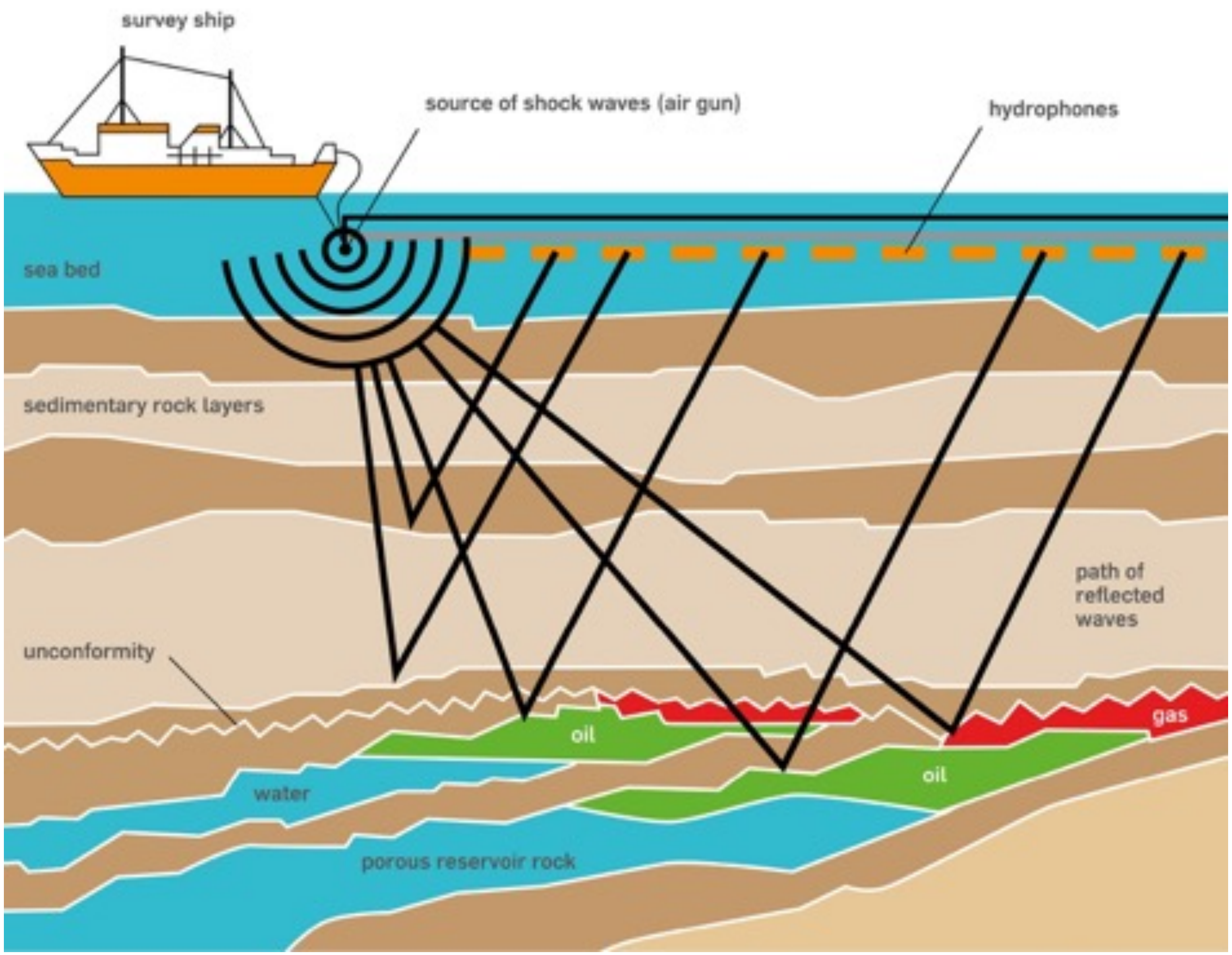
Given a $d = (p, t)$, if p varies according to time, d is a trajectory object, otherwise, d is a permanent object.



Knowledge discovery in Time Series (domain)

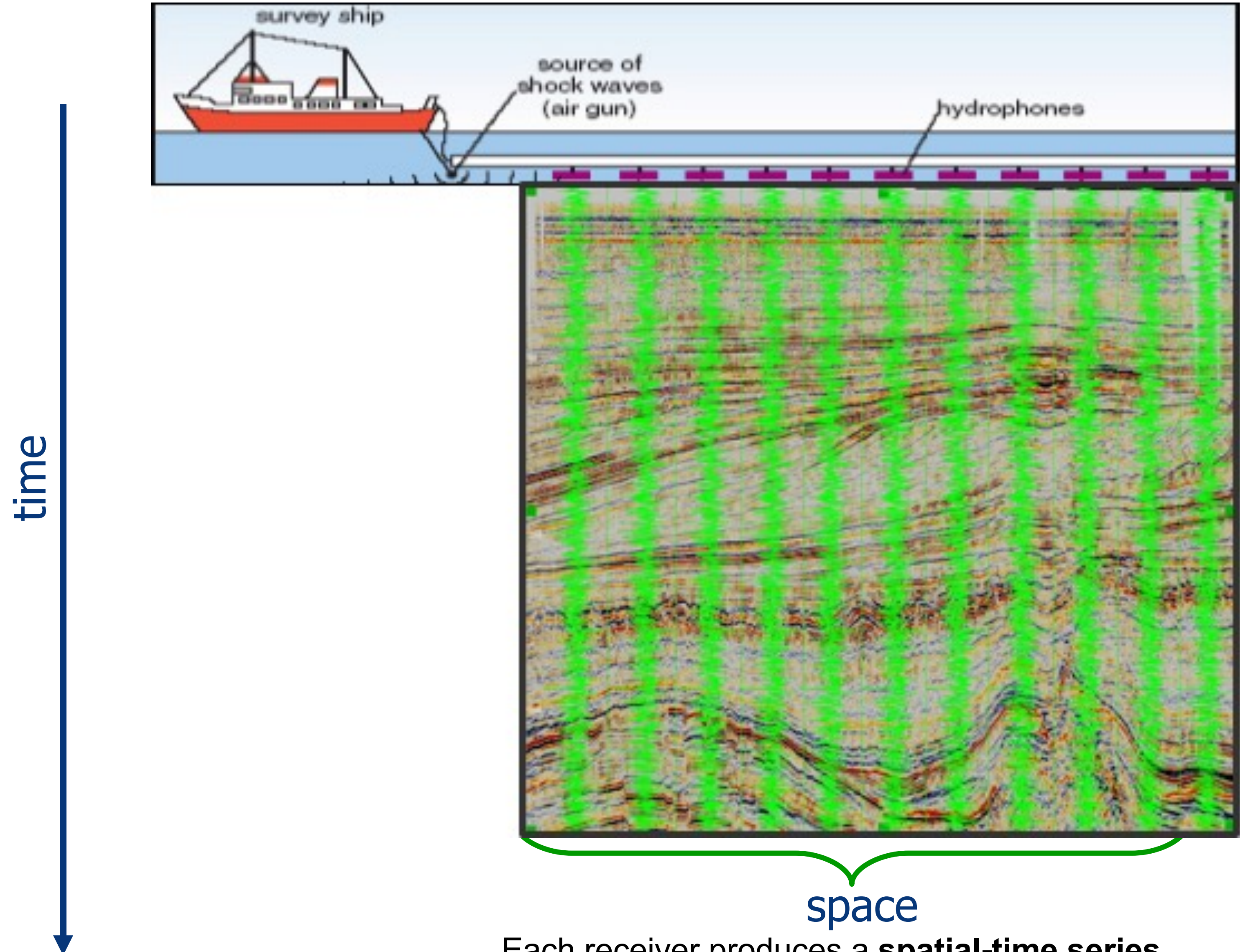
- Big Data
  - Data deluge (volume and velocity)
  - Different data models (variability)
  - Science: astronomy, Seismic
  - Business/Persons: IoT, Flights
  - Government: Smart cities, Urban mobility
- Challenges for Knowledge Discovery
  - Data management
    - Data Preprocessing
    - Workflows
  - Data analysis
    - Prediction / Classification
    - Pattern Identification

(Scientific)
Seismic Analysis Example



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

Seismic Traces Analysis



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

(Business/Industrial)
Analysis of Flight Delays

Analysis of delays in airports according to time

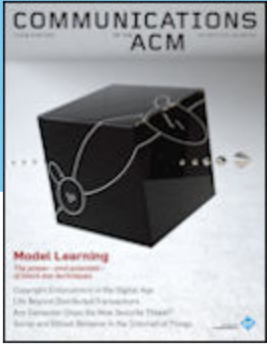
(Government)
Urban Mobility

Buses as trajectory sensors: Analysis of Trajectory Data
Buses stops as permanent object sensors
(Spatial-time aggregation of buses data according to buses stops)

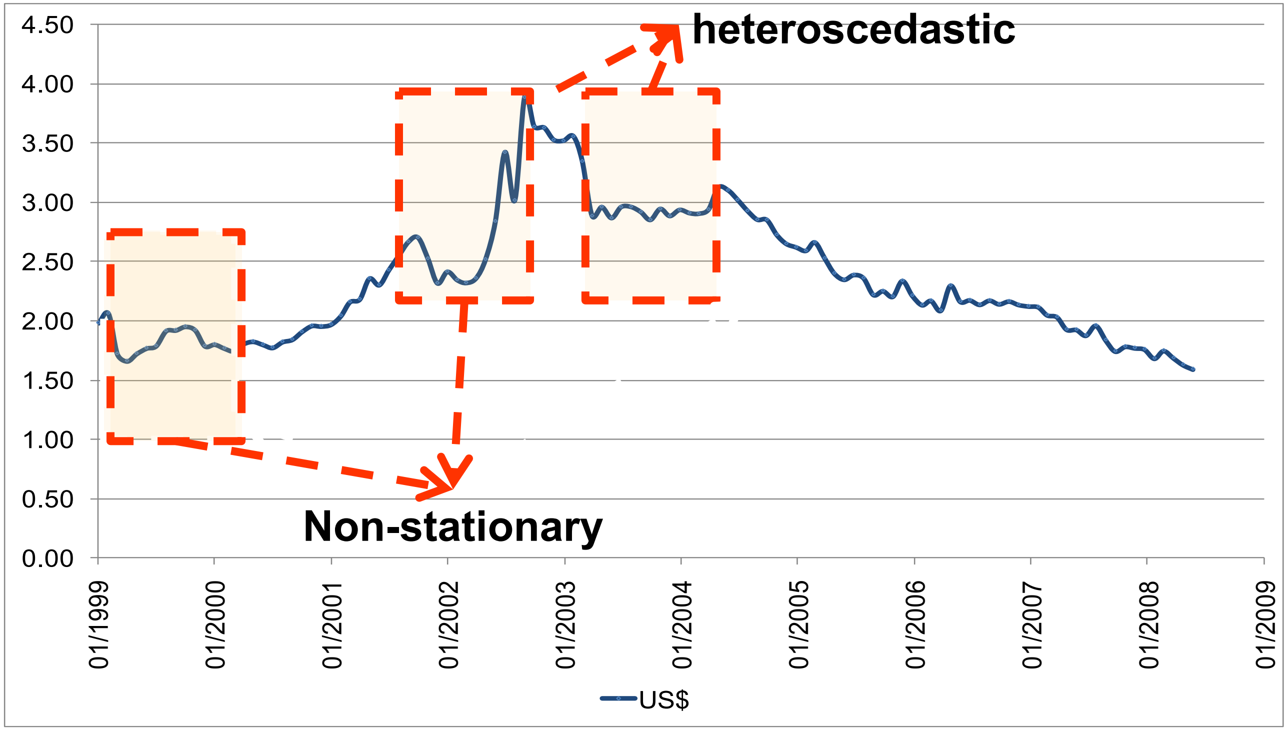
Knowledge Discovery in Time Series (data management)

- Big Data
  - Data deluge (volume and velocity)
  - Different data models (variability)
  - Science: astronomy, Seismic
  - Business/Persons: IoT, Flights
  - Government: Smart cities, Urban mobility
- Challenges for Knowledge Discovery
  - Data management
    - Data Preprocessing
    - Workflows
  - Data analysis
    - Prediction / Classification
    - Pattern Identification

Times Series Properties



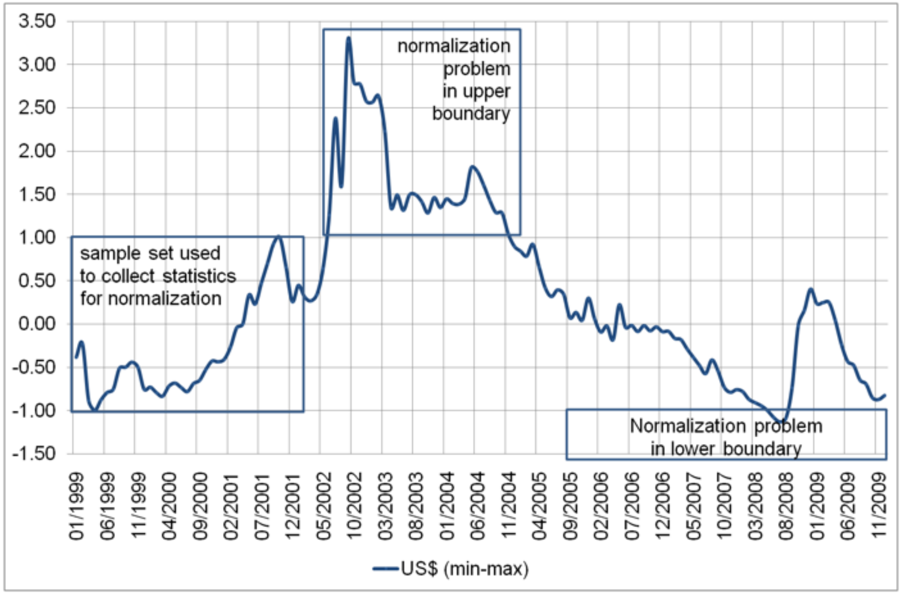
- Many of these real worlds phenomena are:
  - Non-Stationarity and Heteroscedastic



- Data preprocessing techniques: Normalization, Binning, Indexing, Sliding windows
- Machine Learning: Training, Quality of results

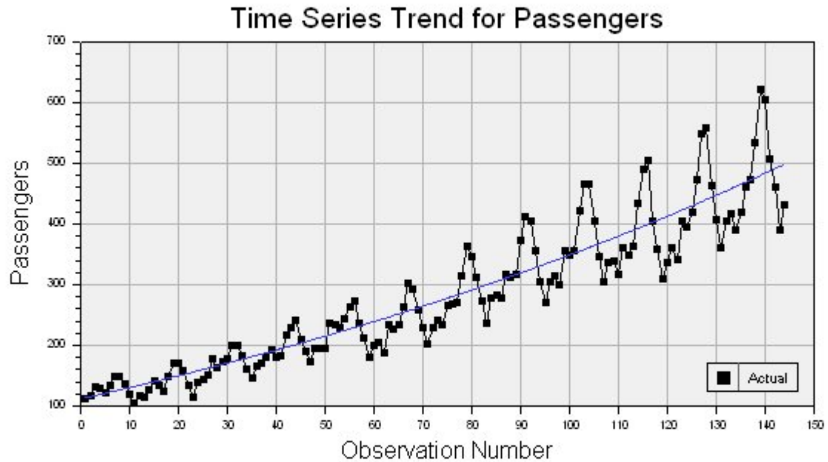
Non-Stationarity affects

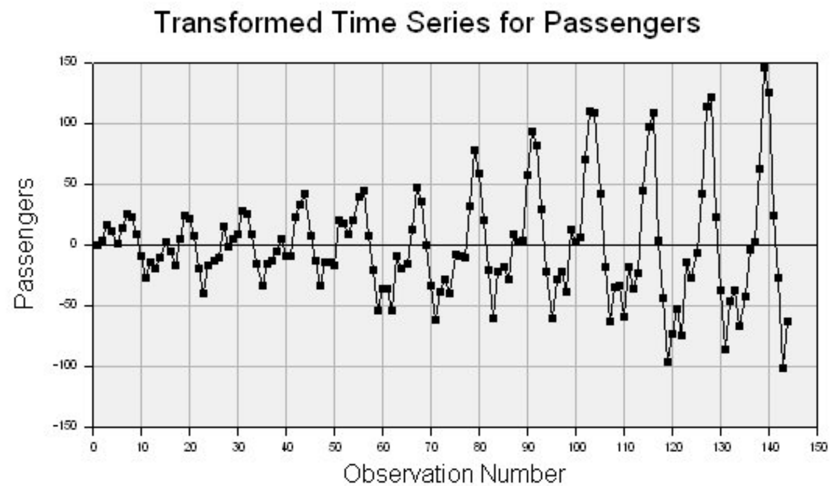
- Data preprocessing techniques
  - Normalization
  - Binning
  - Indexing
  - Sliding windows
- Machine Learning
  - Training
  - Quality of results



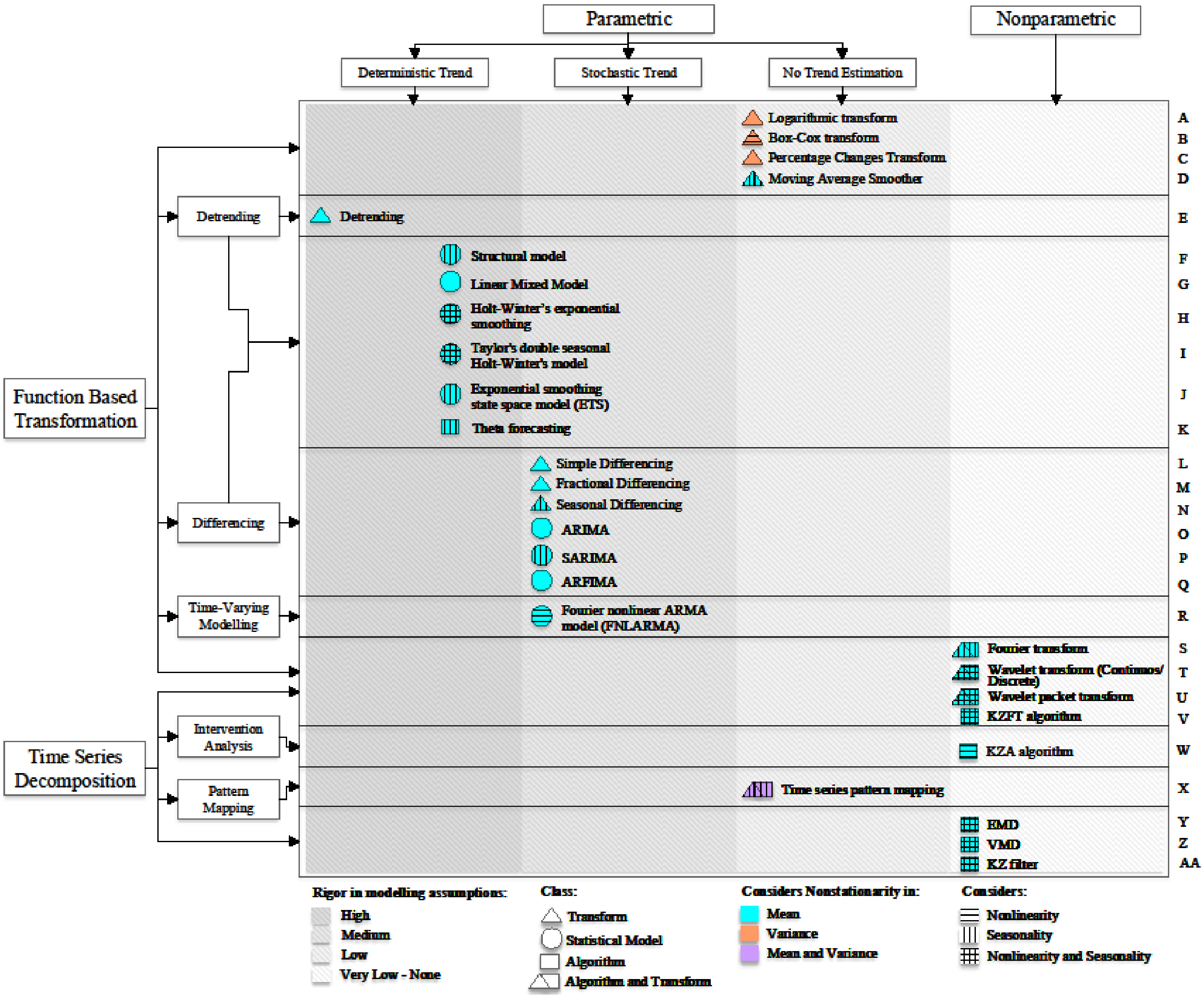
Non-Stationarity in Data Preprocessing: Statistical techniques

- Common approaches
  - Trend removal
  - Differentiation
  - ARIMA models
  - Log transformation
  - Fourier and Wavelet transforms
- Main Problems
  - Many of these techniques were mainly explored in linear models for time series prediction
  - Choosing these techniques is not easy





Review on non-stationary time-series



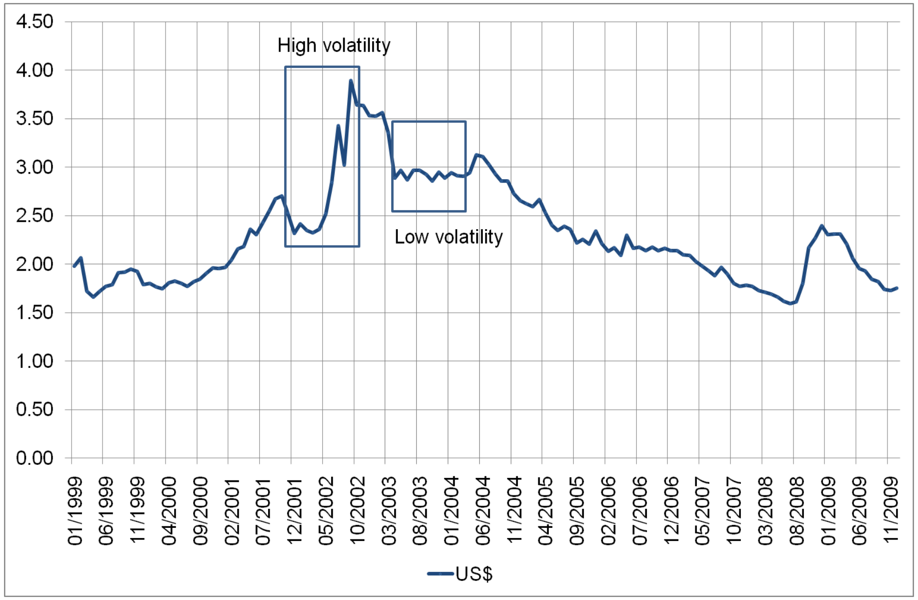
Main works that addresses non-stationary time-series

	Scientific	Socioeconomic/Financial	Industrial
2017	O, T – Nury et al.		
2016		H, O, Y – Wang et al. Q – Sadaei et al. T, Y, Z – Lahmiri	Z – Sun et al. I, J, K, O, Q – Girish and Tiwari T – Chiroma et al. X – Dudek H, O – Akpinar and Yumusak
2015		O, P, T – Joo and Kim	
2014	O – Ljung et al.	O – Claveria and Torra	A, B – Stefanakos and Schinas O, S – Shu et al.
2013	M – Maynard et al.	T – Gao et al. M – Alberiko Gil-Alana and Jiang	
2012	T – Percival and Mondal F – Chilès and Delfiner U – Atto and Berthoumieu	L – James and Murthy	
2011	M – Jara	T – Roshan et al.	T – An et al.
2010	A, B, E, F, L, S, V, W, AA – Yang and Zurbenko	T – Minu et al. D – Ogasawara et al. O, T – Stolojescu et al.	
2009		O, P – Brandão and Nova	
2008		R, T – Nachane and Clavel E, F, M, O, P – Mills and Markellos	
2007	M – Haldrup and Nielsen Q – Brockwell S – Morana	M – Caporale and Gil-Alana	
2006	Q, T – Ko and Vannucci D – F, Q, T – Palma Q, T – Ko and Vannucci T – Fryzlewicz and Nason	T – Fryzlewicz et al. L – Hendry A, B, D, E, M, N – Marrocu	
2005	L – Omtzigt and Paruolo M – Gil-Alana		O, T – Conejo et al.
2004		M, Q – Gil-Alana	
2003	B, Q – D'Elia and Piccolo	S, T – Los	M, O – Abraham and Balakrishna
2002	M – Dittmann and Granger		
2001	E – Maier and Dandy		
1996	M, Q – Baillie		
1992			L – Bhattacharya and Basu
1987			L – Sarma et al.
1981	B, O, P – Hipel E, O – Stensholt and Tjostheim		

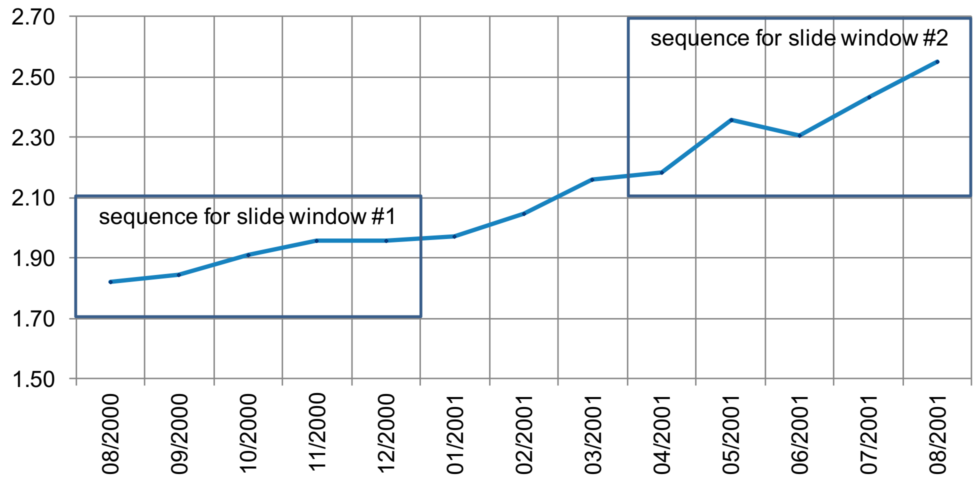
Function Based Transformation — Time Series Decomposition — Function Based Transformation and Time Series Decomposition

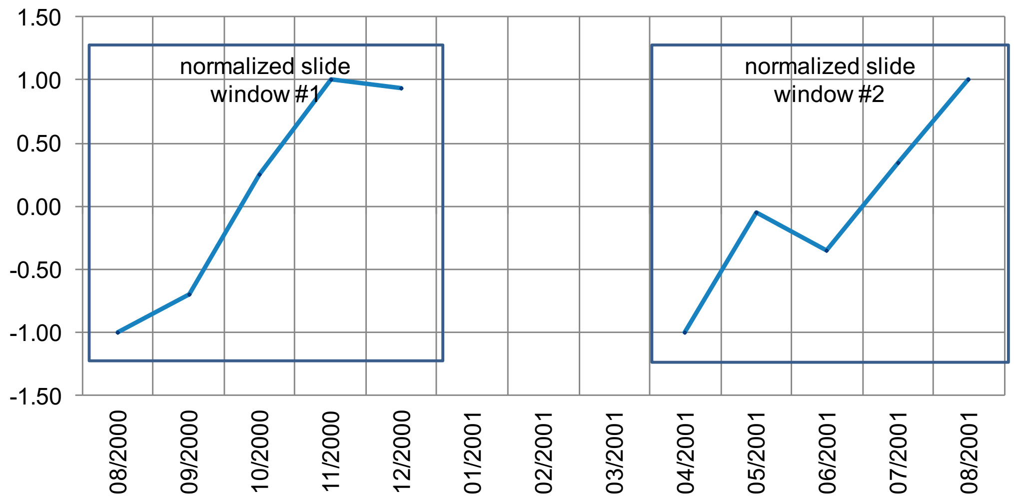
Non-Stationarity in Data Preprocessing: Techniques for Machine Learning

- Machine learning
  - Common Approaches
    - Incremental learning
    - Pseudo-stationary assumption
- Problems
  - Plasticity–stability dilemma
  - When combining the choice of preprocessing techniques with machine learning techniques, the problem becomes even more computational and data intensive



Normalization problem using sliding window

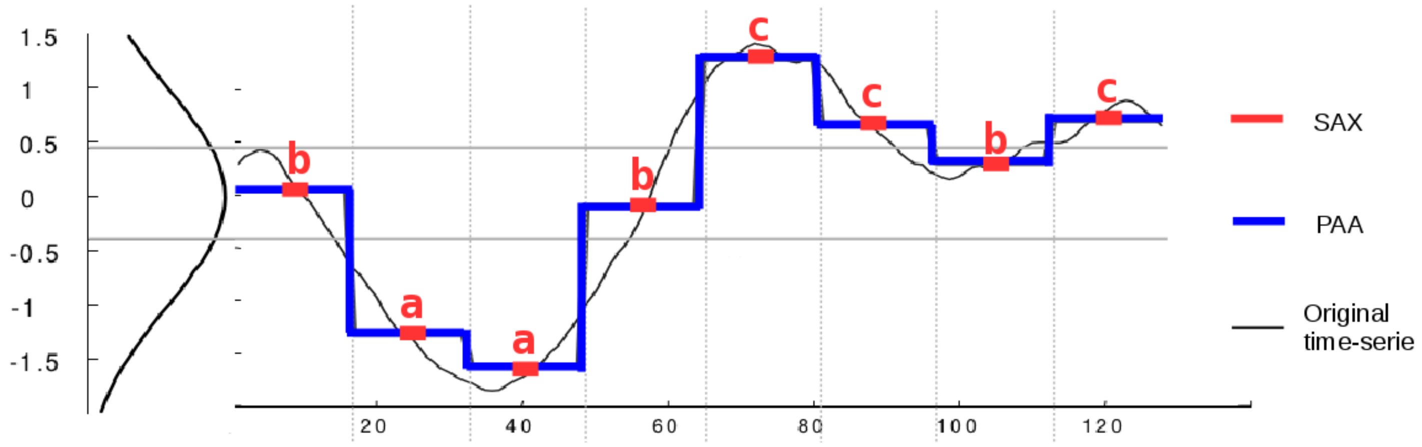




Monthly average exchange rate of U.S. Dollar to Brazilian Real
normalized by sliding window technique from aug/2000 to dec/2000 and from apr/2001 to aug/2001

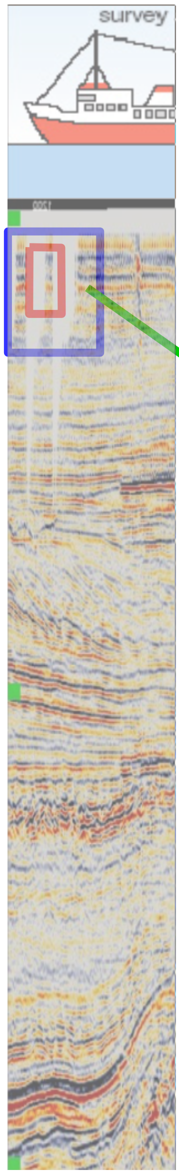
Data Indexing

- Time series contains continuous (non discrete) values
- Is not possible to find patterns performing an exact match between items of such sequences
- SAX indexing was applied to convert continuous values to a discrete symbolic representation



Binning would be change a range to a representative value

SAX Transformation



	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

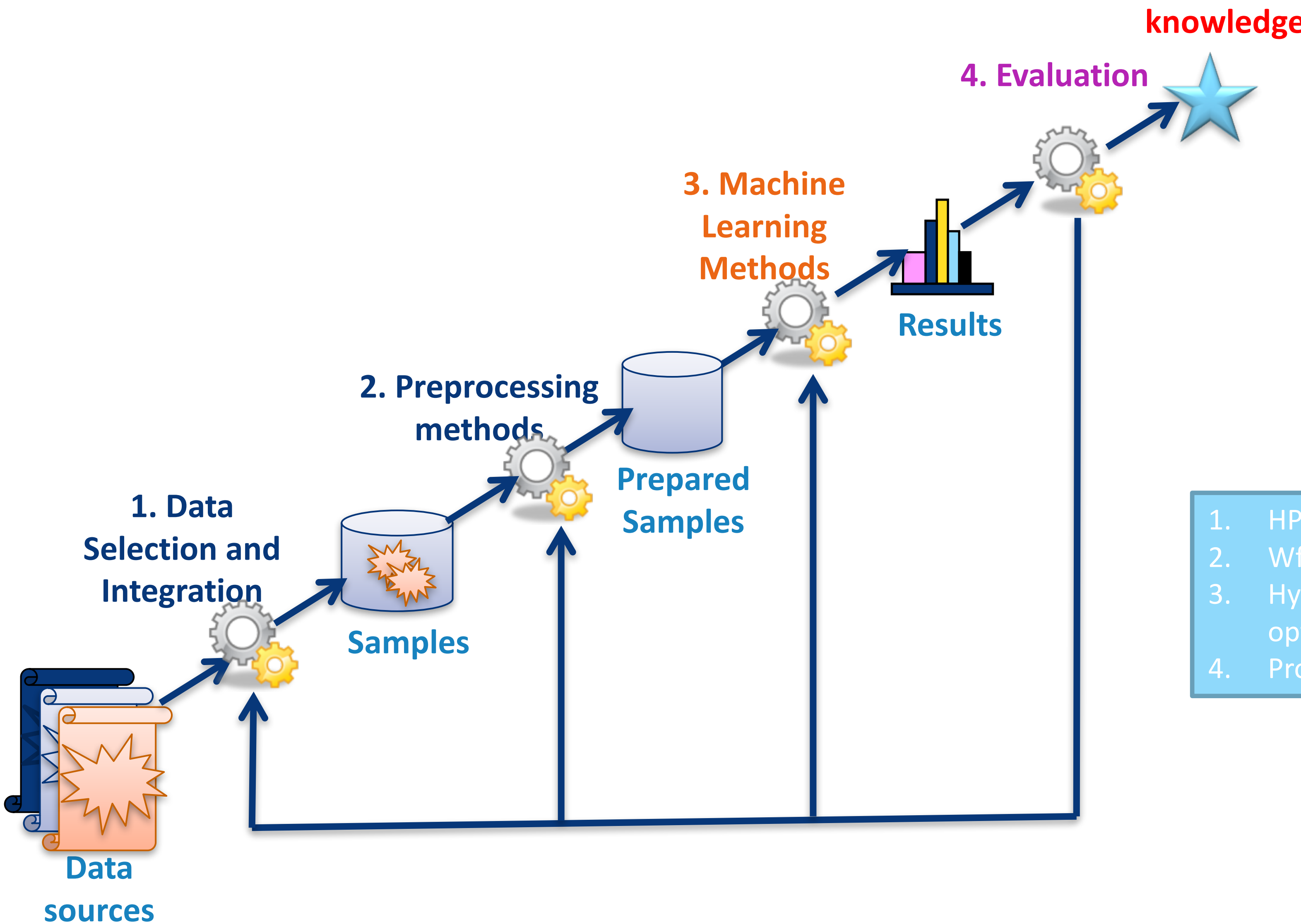
Portion of original seismic dataset

Alphabet [a-z]

	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

SAX converted data

Time Series Data Mining Process (Workflows)

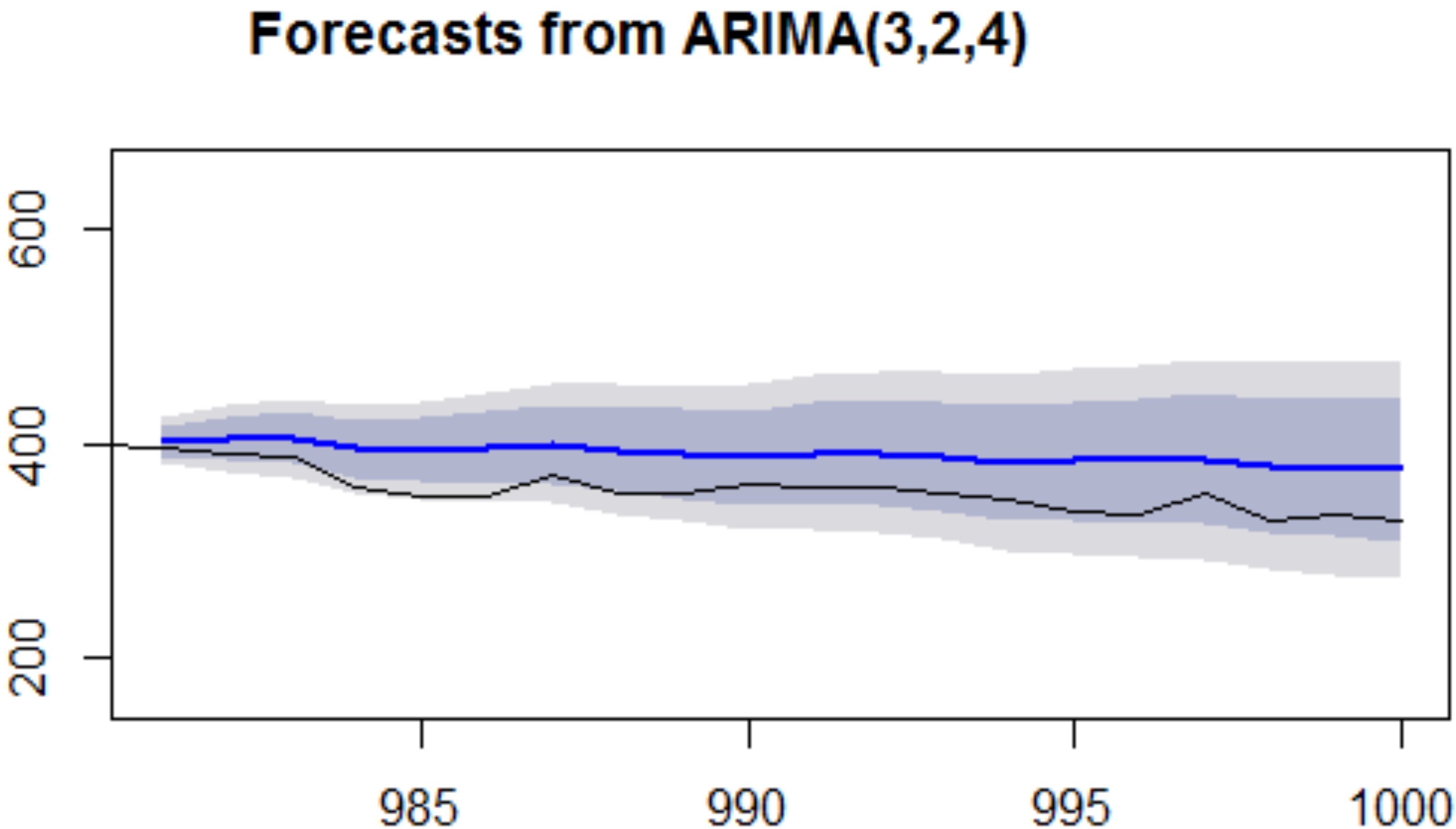


1. HPC and DISC
2. Wf. Algebra
3. Hyper-parameter optimization
4. Provenance

Knowledge Discovery in Big Data (domain)

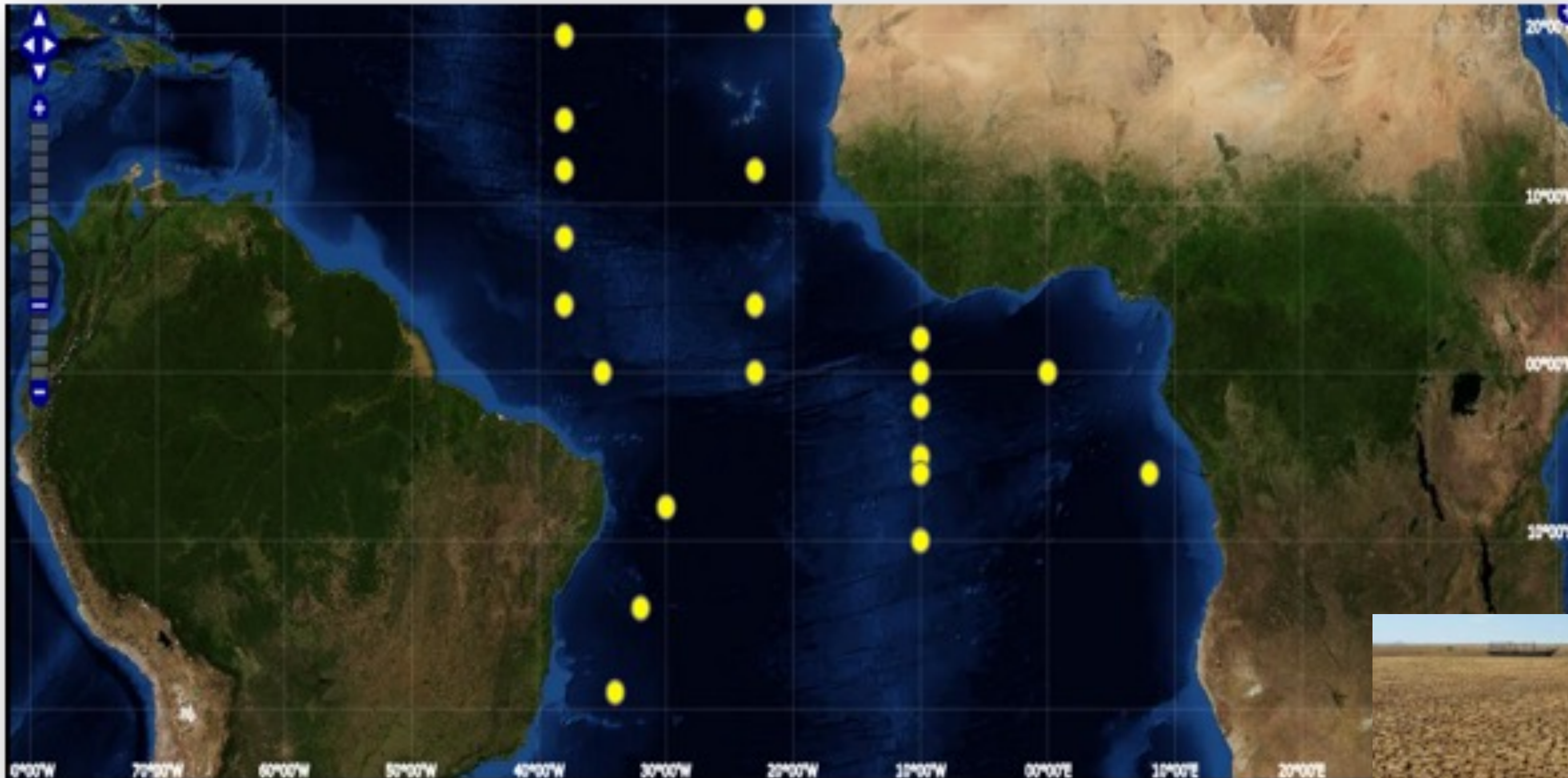
- Big Data
  - Data deluge (volume and velocity)
  - Different data models (variability)
  - Science: astronomy, Seismic
  - Business/Persons: IoT, Flights
  - Government: Smart cities, Urban mobility
- Challenges for Knowledge Discovery
  - Data management
    - Data Preprocessing
    - Workflows
  - Data analysis
    - Prediction / Classification
    - Pattern Identification

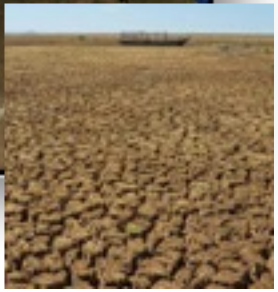
Time series prediction using linear models



Prediction of slice of CATS benchmark dataset

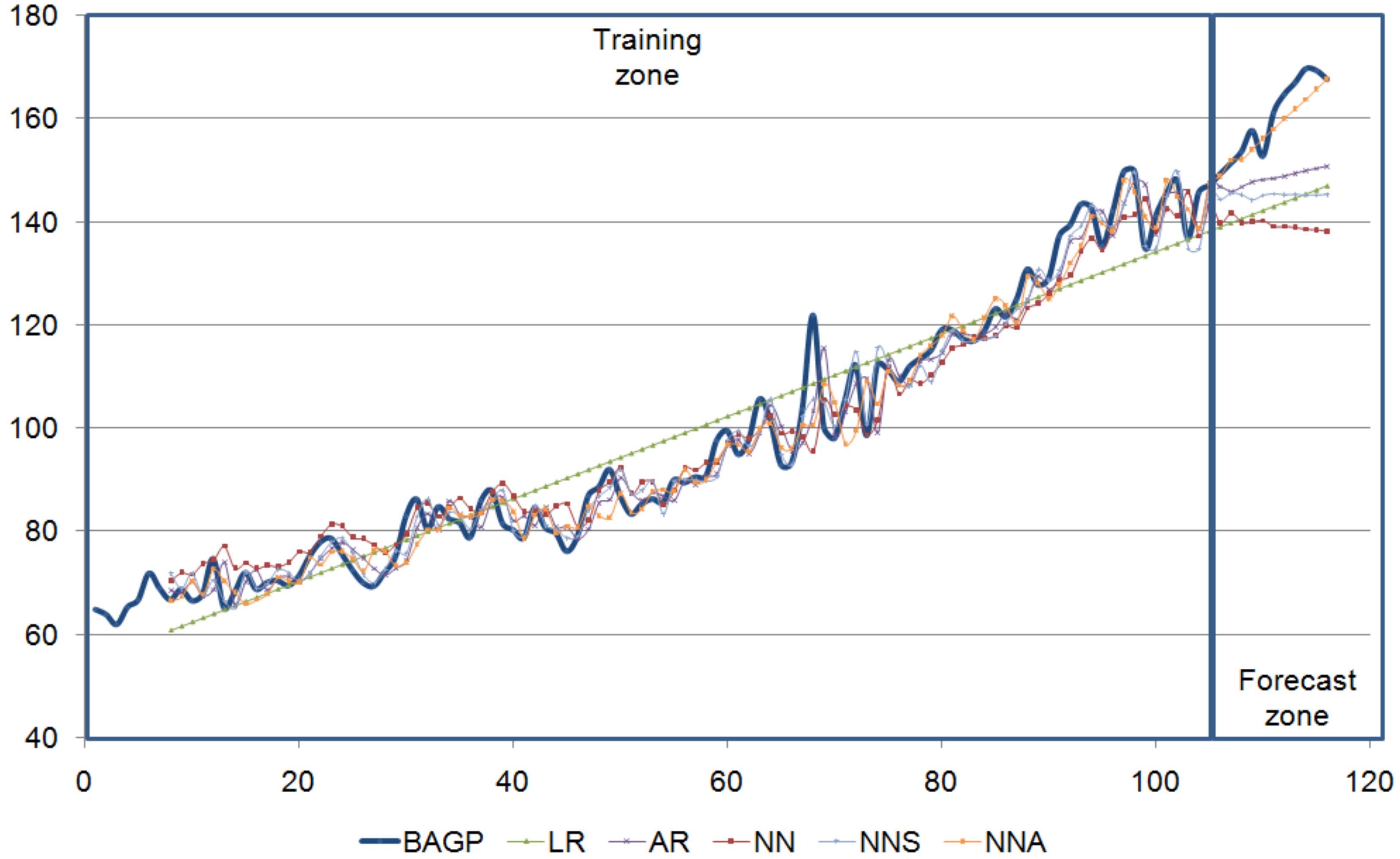
Prediction of sea surface temperature in South Atlantic Ocean





Spatial-time prediction

Time series prediction using machine learning

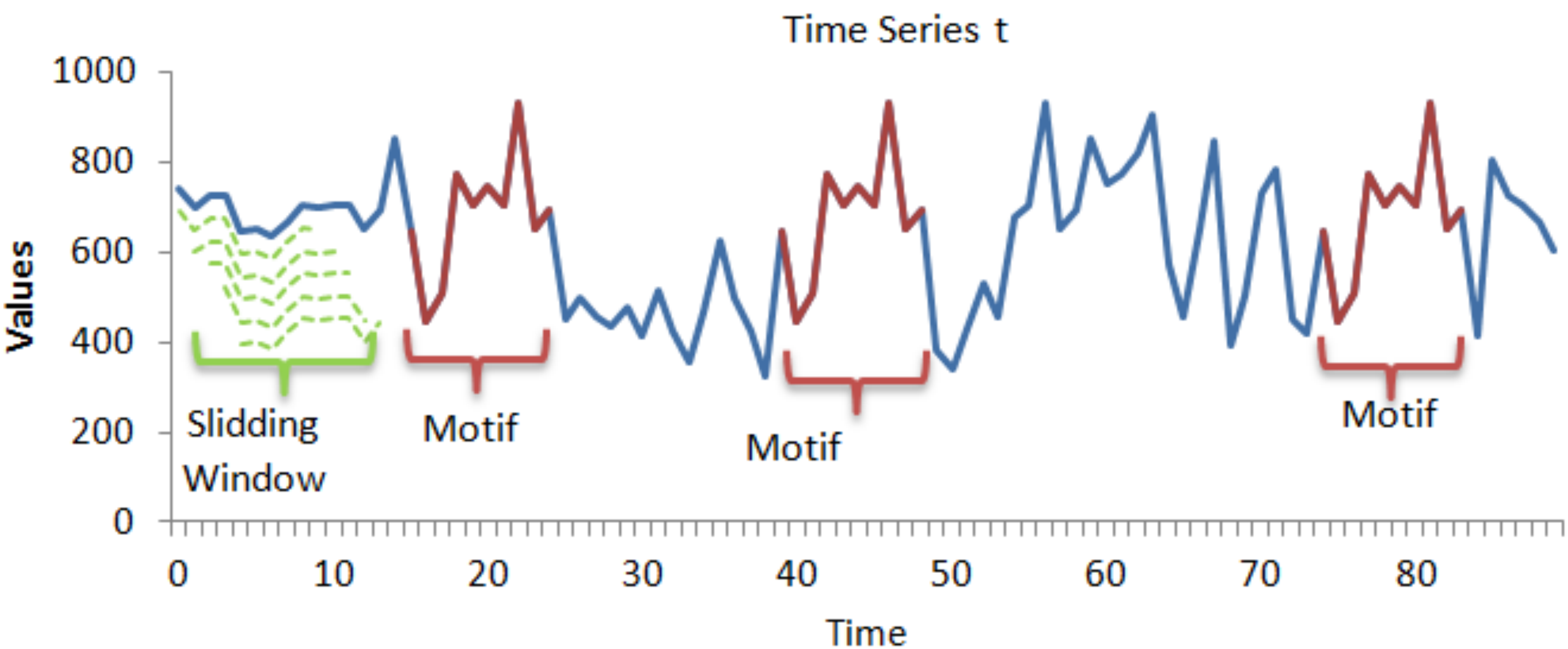


How to build good ML models for non-stationary time series?
Are conventional linear transformations adequate for ML?
How to address Lucas Theorem?

Motif in time series

A sequence $s = <w_1, w_2, \cdots, w_k>$ is **included** in time series $t = <v_1, v_2, \cdots, v_n>$ if there exist integers $i_1 < i_2 < \cdots < i_k$ such that $w_1 = v_{i_1}, w_2 = v_{i_2}, \cdots, w_k = v_{i_k}$.

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



What is a motif in spatial-time series?
How to find motifs in spatial-time series?
How to do it in non-stationarity?

What is next?

Research project in Management and Analysis of Spatial-Time Series

- Non-stationary resilient techniques in data preprocessing
- Novel algorithms for prediction/classification and pattern identification
  - Motif identification
  - Tight spatial-time sequence mining
- Explore spatial-time series applications
  - Frequent pattern mining, Classification/Prediction
- Explore data management and parallel processing for mining non-stationary time/spatial-time series
  - Algebraic-based workflows for spatial-time series data mining using Spark

Adaptive normalization

- Transformation
  - transforming the non-stationary time series into a stationary sliding window
- Outlier removal
- Normalization
- Data Mining:
  - Prediction/Classification
  - Pattern Identification

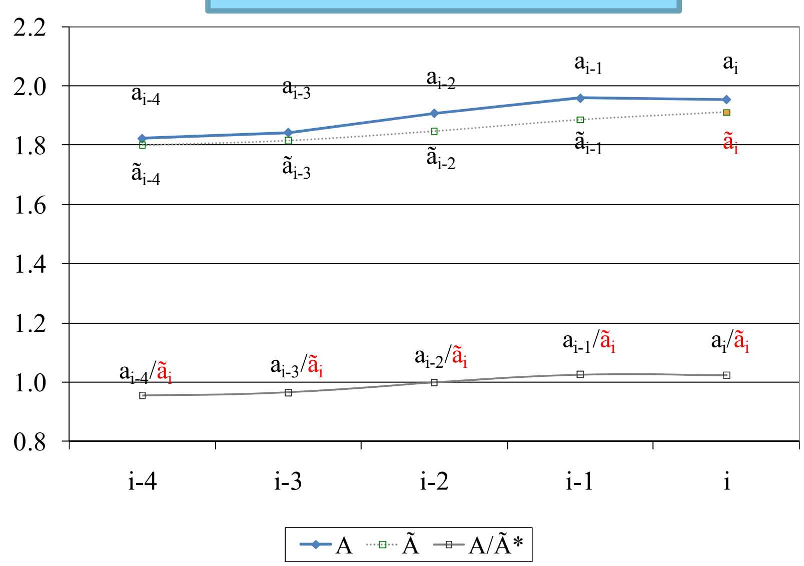
Adaptive Normalization
Phase 1: Transformation

i	US$/R$ S	EMA : $S^{(5)}$
1	1.734	1.721
2	1.720	1.729
3	1.707	1.734
4	1.708	1.742
5	1.735	1.745
6	1.746	1.747
7	1.744	1.752
8	1.759	1.752
9	1.751	1.760
10	1.749	-
11	1.763	-
12	1.753	-
13	1.774	-

Original time series S and its MA

i	$S[i]$ / $S^{(5)}[i]$	$S[i+1]$ / $S^{(5)}[i]$	$S[i+2]$ / $S^{(5)}[i]$	$S[i+3]$ / $S^{(5)}[i]$	$S[i+4]$ / $S^{(5)}[i]$	$S[i+5]$ / $S^{(5)}[i]$
1	1.008	1.000	0.992	0.993	1.008	1.015
2	0.995	0.987	0.988	1.003	1.010	1.009
3	0.984	0.985	1.000	1.007	1.006	1.014
4	0.980	0.996	1.002	1.001	1.010	1.005
5	0.994	1.000	0.999	1.008	1.003	1.002
6	1.000	0.999	1.007	1.003	1.001	1.009
7	0.995	1.004	0.999	0.998	1.006	1.001
8	1.004	0.999	0.998	1.006	1.000	**1.012**

Transformed slide window R



Adaptive Normalization
Phase 2: Outlier removal

- Method based on Boxplots:
  - values at least 1.5 x IQR below the first quartile or above the third quartile are considered outliers
- In Adaptive Normalization, any DSW that contains at least one outlier is discarded
- Q1 = 0.996, Q3 = 1.006, IQR = 0.10
- Q1 – 1.5 x IQR = 0.981Q3 + 1.5 x IQR = 1.021
- Discards DSW number 4

i	S[i] / $S^{(5)}[i]$	S[i+1] / $S^{(5)}[i]$	S[i+2] / $S^{(5)}[i]$	S[i+3] / $S^{(5)}[i]$	S[i+4] / $S^{(5)}[i]$	S[i+5] / $S^{(5)}[i]$
1	1.008	1.000	0.992	0.993	1.008	1.015
2	0.995	0.987	0.988	1.003	1.010	1.009
3	0.984	0.985	1.000	1.007	1.006	1.014
4	0.980	0.996	1.002	1.001	1.010	1.005
5	0.994	1.000	0.999	1.008	1.003	1.002
6	1.000	0.999	1.007	1.003	1.001	1.009
7	0.995	1.004	0.999	0.998	1.006	1.001
8	1.004	0.999	0.998	1.006	1.000	**1.012**

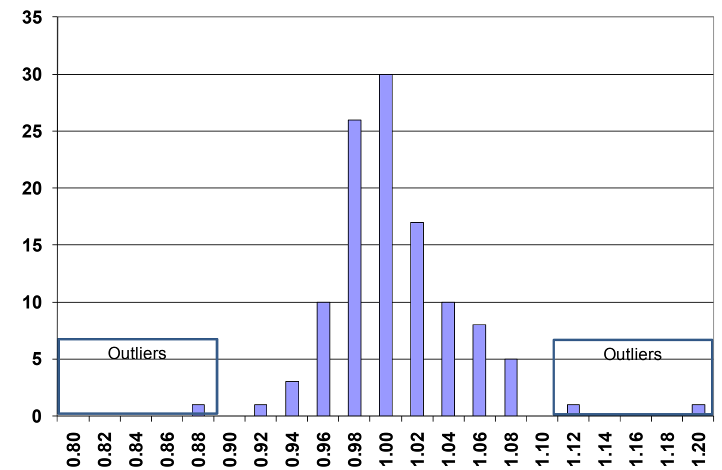


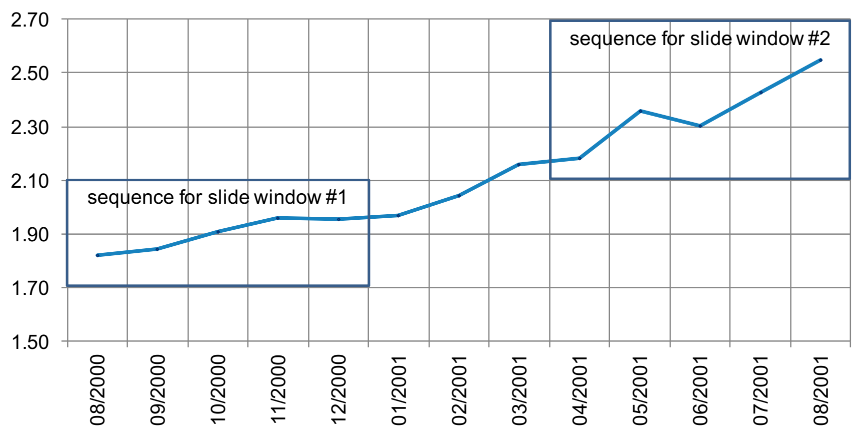
Figure 6. Outlier removal for U.S. Dollar to Brazilian Real Exchange Rate

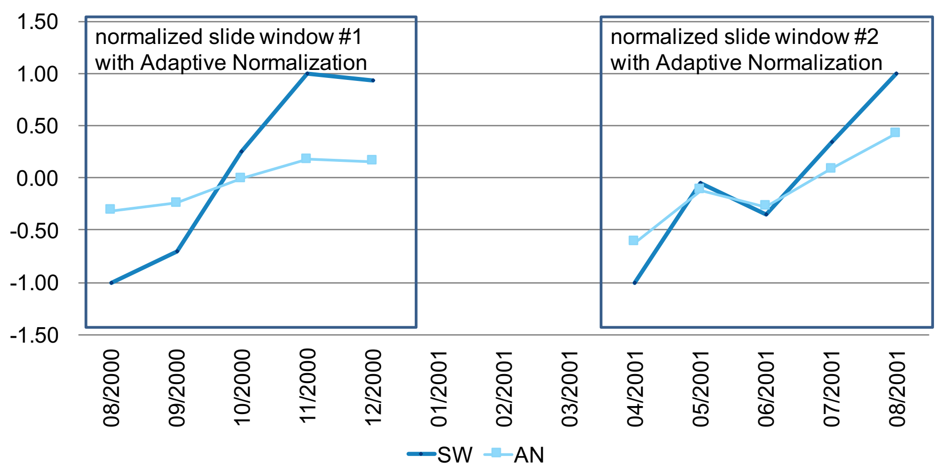
Adaptive Normalization
Phase 3: Normalization

- In the example:
  - Min-max normalization method to normalize the values of sequence in the range [−1, 1]
  - Min: 0.981
    Max(Min(R), (Q1 − 1.5 × IQR))
  - Max: 1.015
    Min(Max(R), (Q3 + 1.5 × IQR))

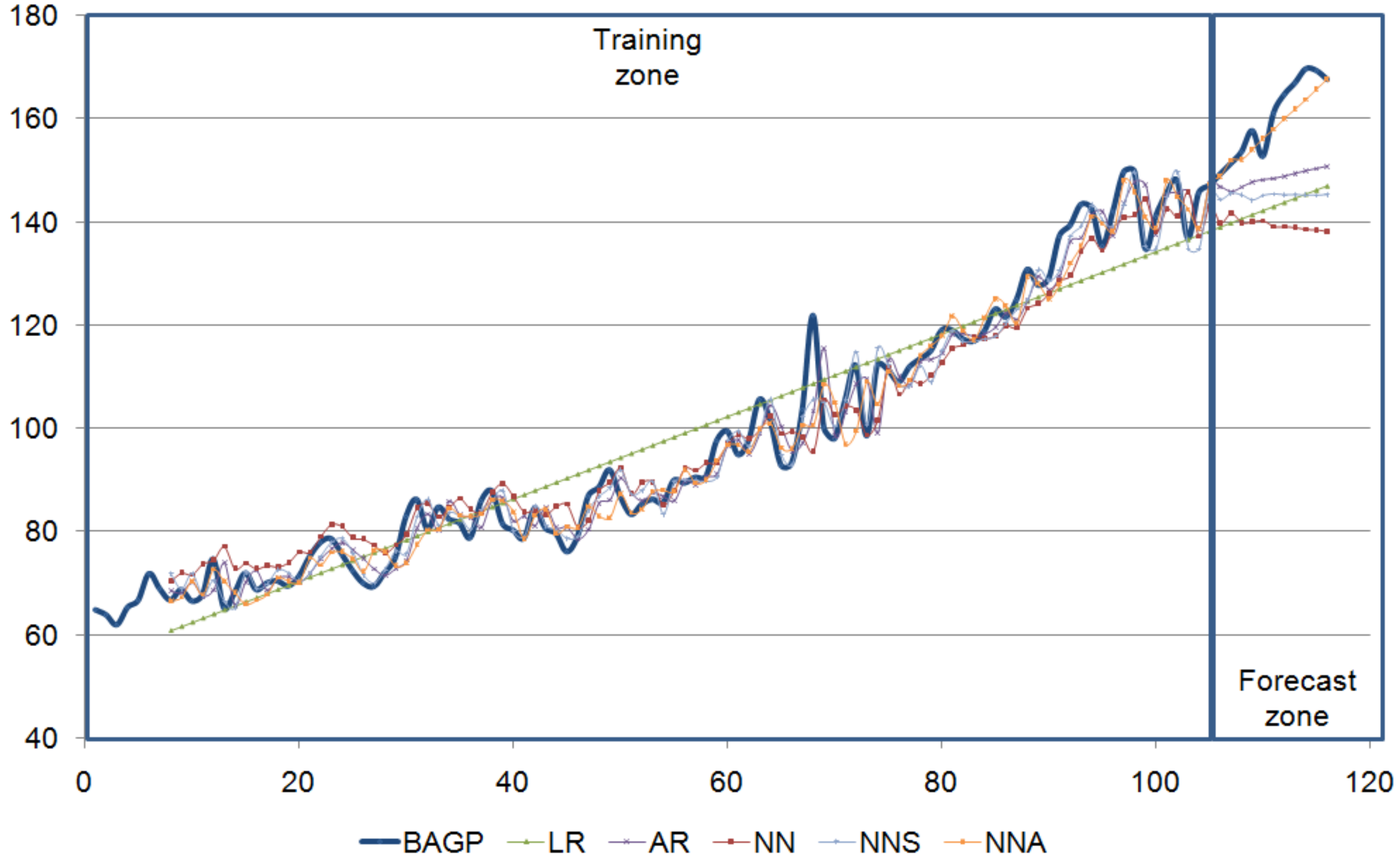
i	Normalized Sliding Window					
1	0,585	0,102	-0,347	-0,313	0,620	1,000
2	-0,187	-0,634	-0,599	0,329	0,707	0,638
3	-0,801	-0,766	0,159	0,536	0,468	0,982
-	-	-	-	-	-	-
5	-0,221	0,154	0,086	0,597	0,324	0,256
6	0,112	0,044	0,554	0,282	0,214	0,690
7	-0,142	0,366	0,095	0,027	0,502	0,163
8	0,355	0,084	0,016	0,491	0,152	**0,864**

Normalized sliding window in the range [-1,1]





Time series prediction using machine learning



Data transformations challenges for machine learning

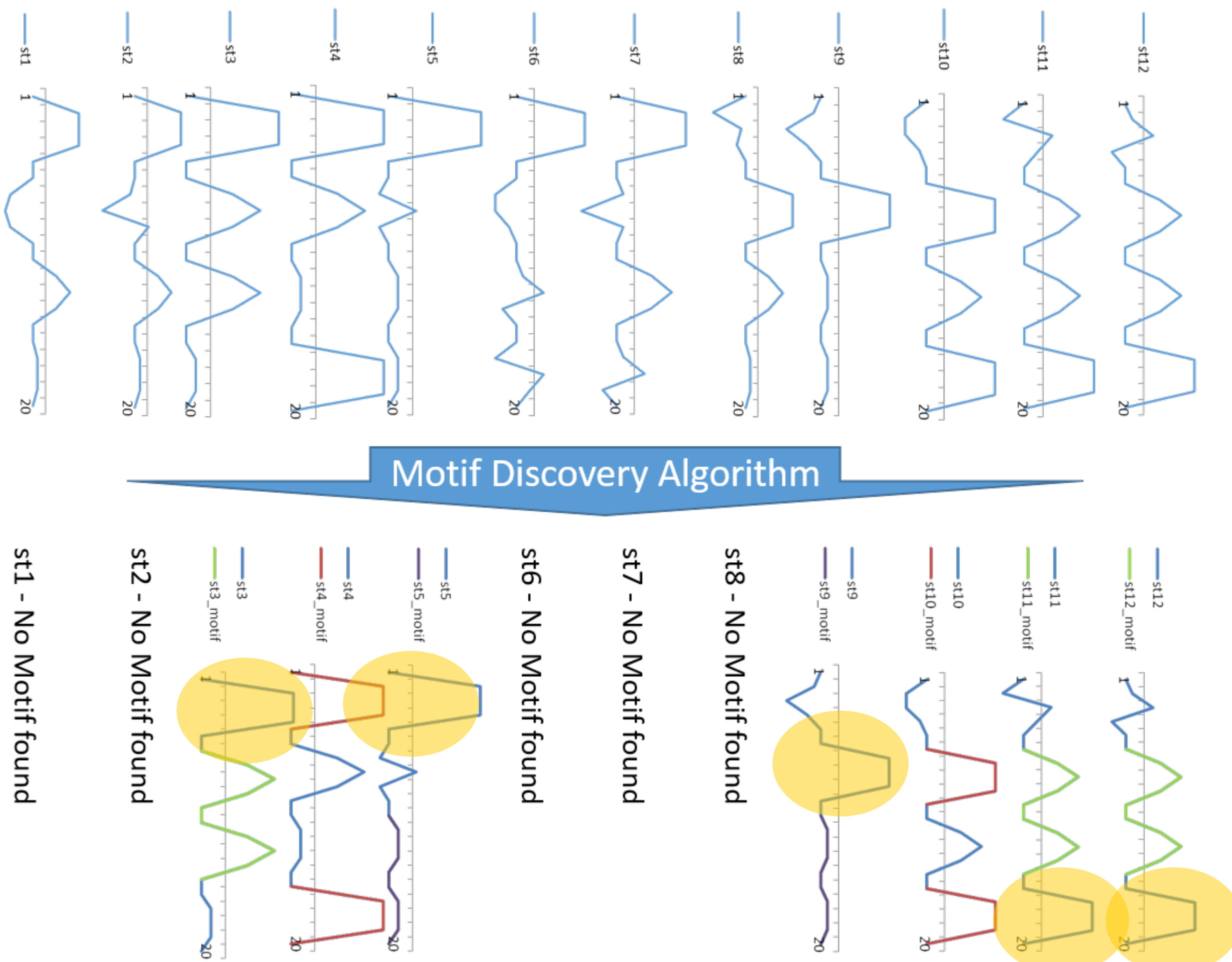
- Explore different inertia functions
  - Isaac Newton
- Explore new differentiation approaches
  - Solve division by zero problem
- Explore different machine learning algorithms
- Explore different mining tasks

Research Project In Management and Analysis of Spatial-Time Series

- Non-stationary resilient techniques in data preprocessing
- Novel algorithms for prediction/classification and pattern identification
  - Motif identification
  - Tight spatial-time sequence mining
- Explore spatial-time series applications
  - Frequent pattern mining, Classification/Prediction
- Explore data management and parallel processing for mining non-stationary Big Data
  - Algebraic-based spatial-time series data mining workflow using Spark

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.



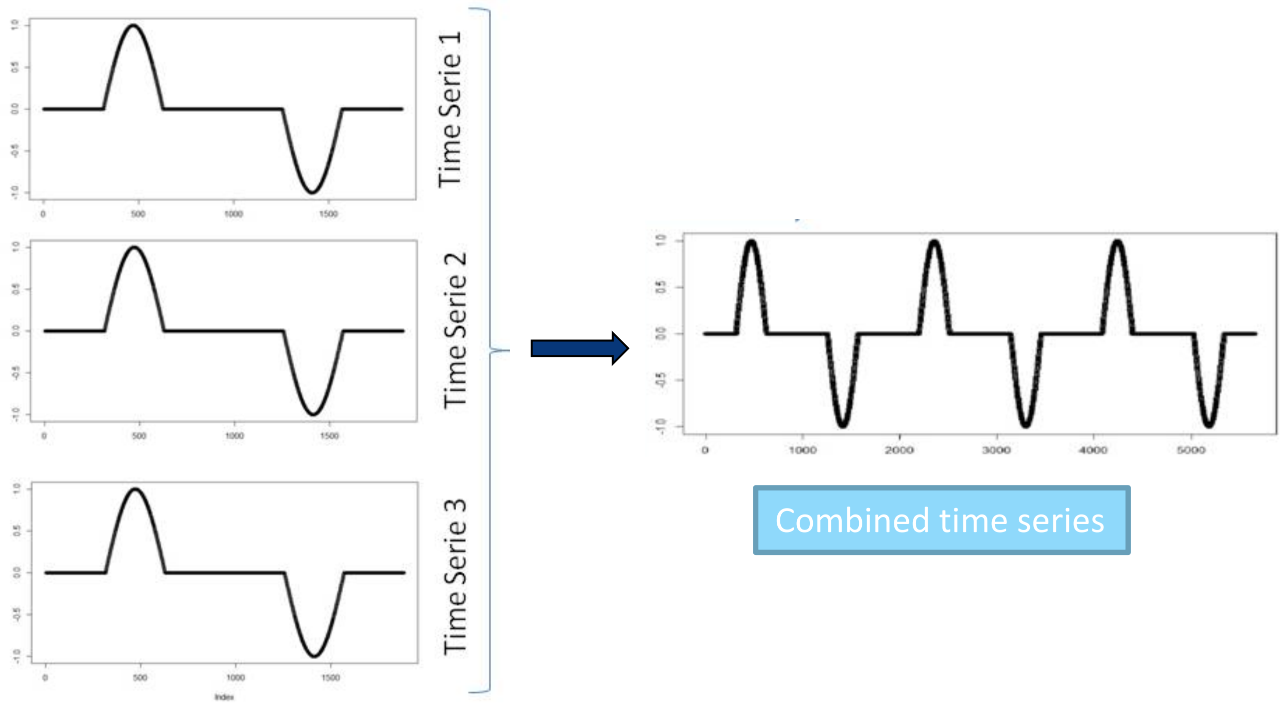
Traditional motif discovery algorithm applied in spatial-time series dataset. (i) red trapeziums and green triangles are identified motifs; (ii) blue trapeziums are not identified and not linked with red ones; (iii) blue triangles are not identified and not linked with green ones; (iv) purple shapes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

Spatial-Time Motif

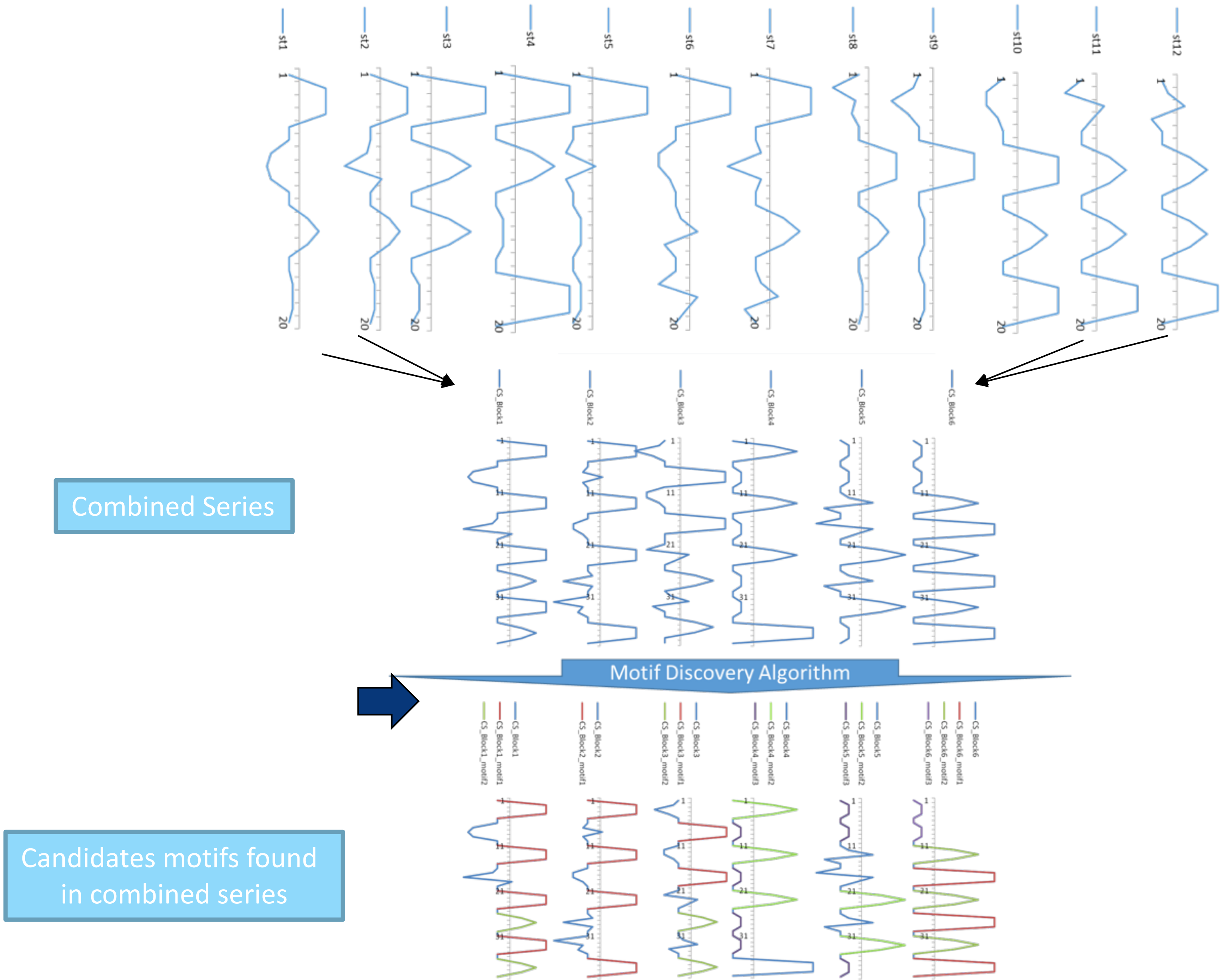
A **spatial range** (or simply **range**) $r = (p_s, p_e)$ is defined by a start position p_s and an end position p_e.

A **block** b is a couple (r, i) where r is a range $(r \in PR)$ and i is an interval $(i \in PI)$.

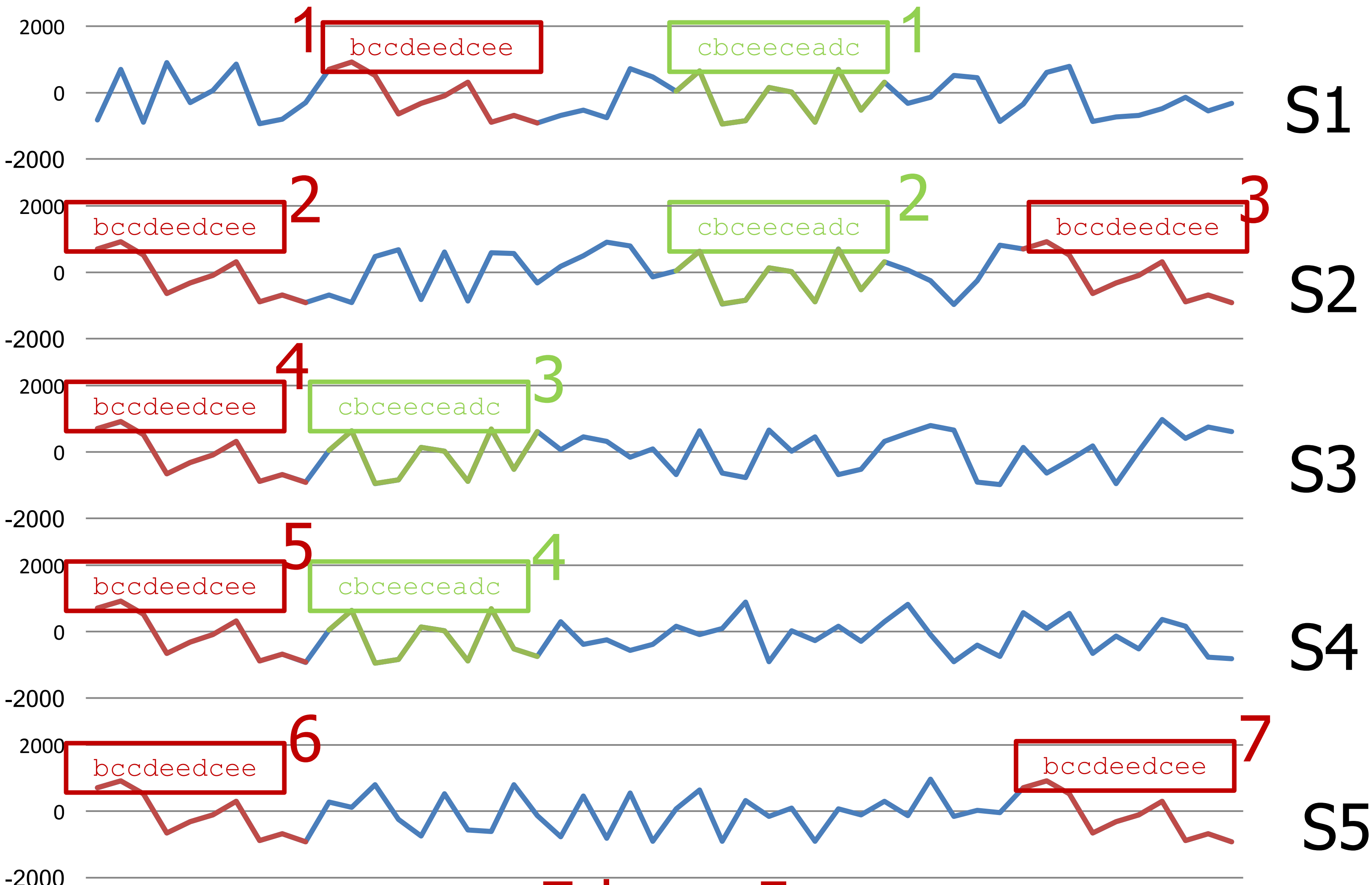
Let σ and κ be two thresholds, such that $\sigma \geq \kappa$. A sequence q is a **spatial-time motif** in a block $b \subset S$ iff q is included at list σ times $linear(b) \wedge support(q, b.r) > \kappa$.



Combined Series Approach



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	s	k	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

function STMotif(b, sw, w, a, bs, bt)

  $b_i \leftarrow partition(b, bs, bt)$

  for each $b_i \in b$ **do**

    $t \leftarrow combine(b_i)$

    $CSTM \leftarrow identify(t)$

    $STM \leftarrow STM \cup constraintST(CSTM)$

  end for

  $rankSTM = aggregate(STM)$

  return $rankSTM$

end function

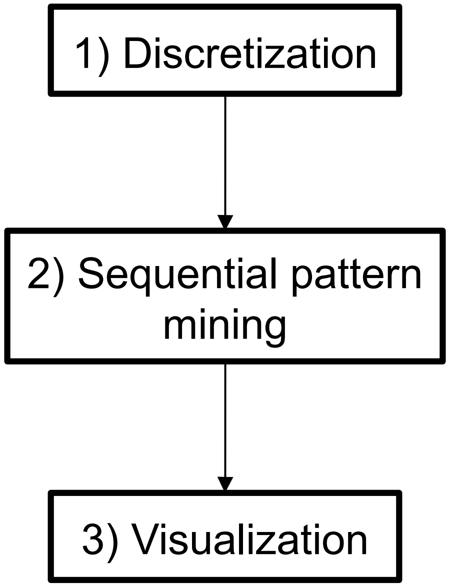
Research Project In Management and Analysis of Spatial-Time Series

- Non-stationary resilient techniques in data preprocessing
- Novel algorithms for prediction/classification and pattern identification
  - Motif identification
  - Tight spatial-time sequence mining
- Explore spatial-time series applications
  - Frequent pattern mining, Classification/Prediction
- Explore data management and parallel processing for mining non-stationary Big Data
  - Algebraic-based spatial-time series data mining workflow using Spark

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



A priori principle

Time Square

Pattern Identification in Space-Time Series

D / t	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
v_1	**a**	b	c	d	τ	θ	**i**	g	**a**	h
v_2	k	l	m	n	p	q	**u**	s	t	v
v_3	w	**e**	**e**	x	y	m	**a**	r	δ	α
v_4	h	**o**	**o**	g	**e**	ι	ε	**i**	χ	β
v_5	**i**	φ	κ	λ	**o**	z	ν	**u**	ζ	π
v_6	**u**	**a**	ρ	σ	τ	μ	c	d	f	**a**

(sr_1) (sr_3) (sr_4) (sr_2)

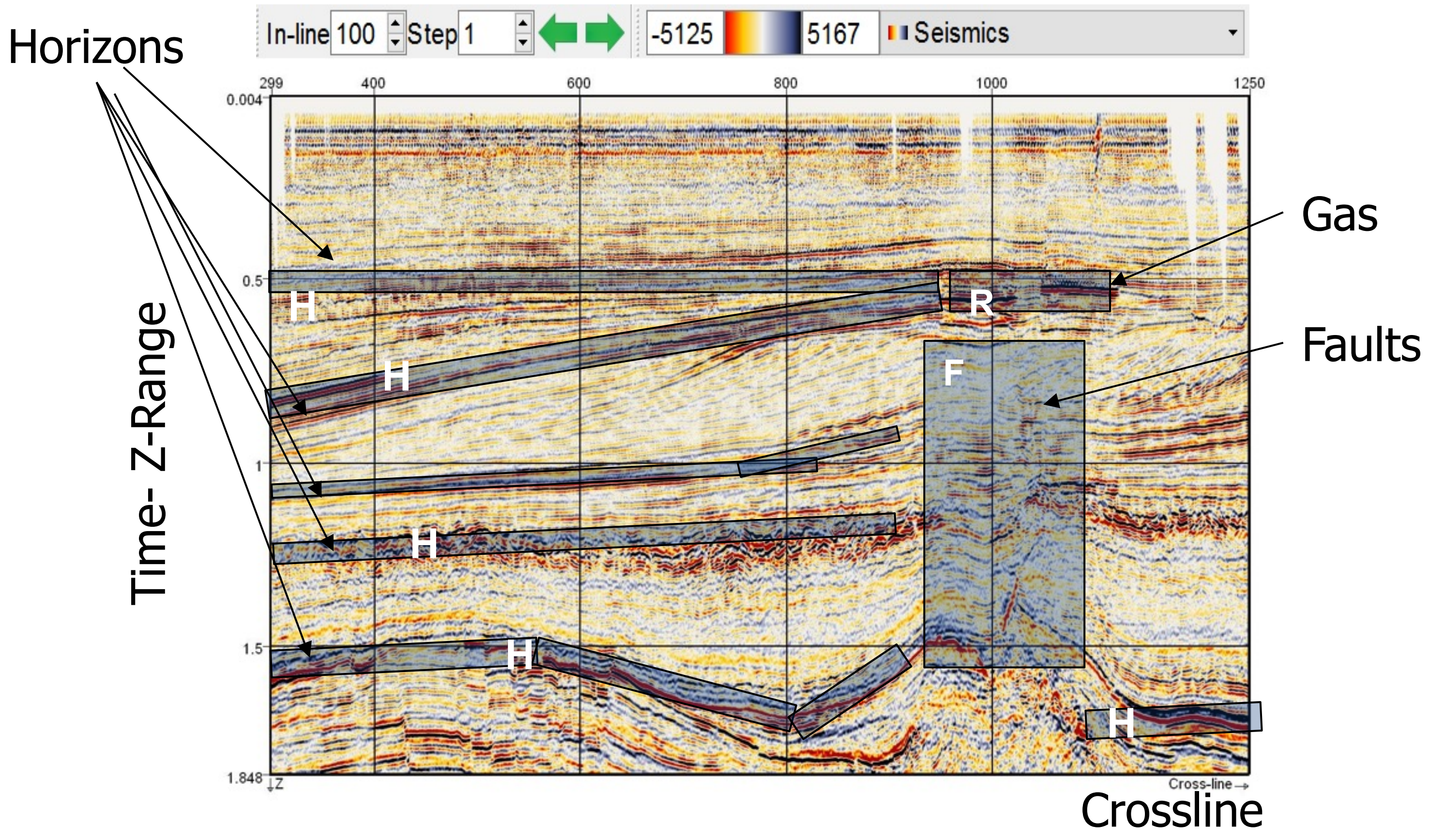
Spatial-time sequence miner

Algorithm 1 Spatio-Temporal Sequence Miner

function $STSM(D, \gamma, \delta)$
    $C_1 \leftarrow generateCandidates(D, nil)$
    $k \leftarrow 0$
    repeat
        $k \leftarrow k + 1$
        $SR_k \leftarrow solidRangedSequences(D, C_k, \gamma)$
        $C_{k+1} \leftarrow generateCandidates(D, SR_k)$
    until $C_{k+1} \neq \emptyset$
    for $(i \in \{1 \cdots k\})$ **do**
        $SB_i \leftarrow solidBlockedSequences(D, SR_i, \delta)$
    end for
    return $\{SB_1, \cdots, SB_k\}$
end function

Seismic Analysis

- 2D Slice of seismic dataset (inline 100)



Seismic Analysis – Results

- Motifs Analysis
  - Discovering spatial-time motifs in seismic datasets

Murillo Dutra master degree

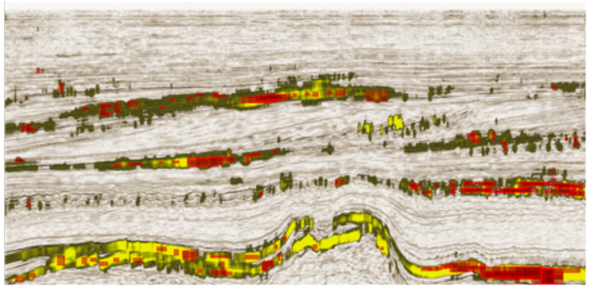


Figure 2: Identified solid-blocked sequence $<a, a, j, j>$ for *inline* 401, alphabet size 10, solid range threshold γ 80% and solid block threshold δ 20%. Its density was 206. Solid-blocked sequences are marked in red. The results follow the yellow pattern produced using the previously known *bright spots* for this dataset [3].

- Sequence Mining of Spatial-Time Series
  - Identification of solid spatial-time sequences

Riccardo Campisano master degree

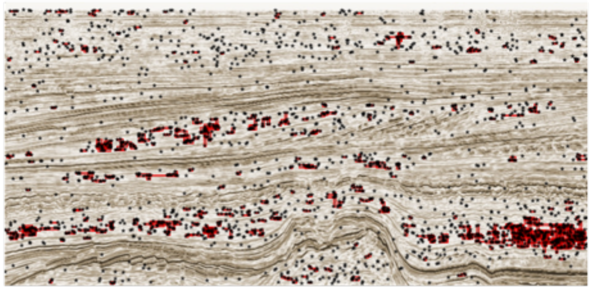


Figure 3: Comparison of quality between GSP and $STSM$ for sequence $<e, e, f>$ in *inline* 401 using alphabet size 10, with support of 80% for GSP and with solid range threshold (γ) of 80% and solid block threshold (δ) of 20% for $STSM$. Identified occurrences are marked as red when identified by $STSM$ and as black in GSP. Although occurrences from $STSM$ correspond to seismic horizons, many occurrences from GSP correspond to noise.

Research Project In Management and Analysis of Spatial-Time Series

- Non-stationary resilient techniques in data preprocessing
- Novel algorithms for prediction/classification and pattern identification
  - Motif identification
  - Tight spatial-time sequence mining
- Explore spatial-time series applications
  - Frequent pattern mining, Classification/Prediction
- Explore data management and parallel processing for mining non-stationary Big Data
  - Algebraic-based spatial-time series data mining workflow using Spark

Seismic Analysis – Research Opportunities

- 3D Analysis (x, y, and time)
  - Solid Cube Patterns
- Techniques for faults detection
  - Intuition that absence of solid patterns drives faults detection
- Techniques for shape detections
  - Combinations of motifs/solid patterns
- Comparison between motifs identification and sequence mining

Flight Delays

Brazilian Flights Dataset
Airports Meteorological Dataset

Flight Delays – Results

- Data warehouse
  - Brazilian National Flights
  - Meteorological condition
- Identification of frequent patterns that leads to delays

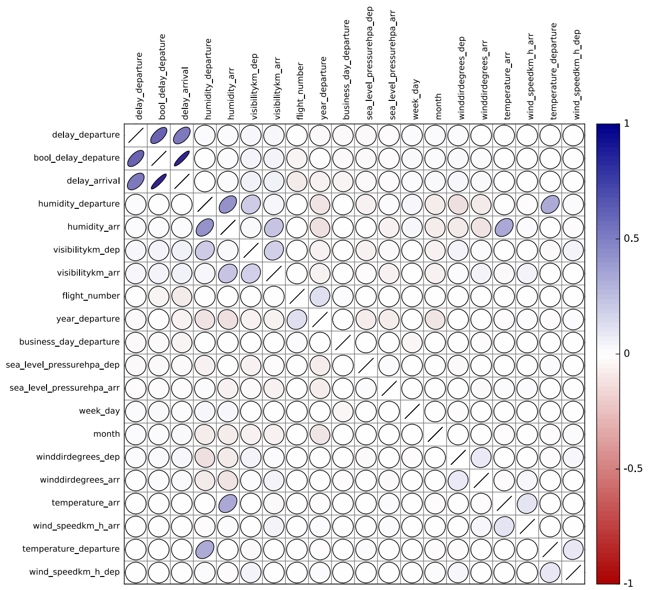
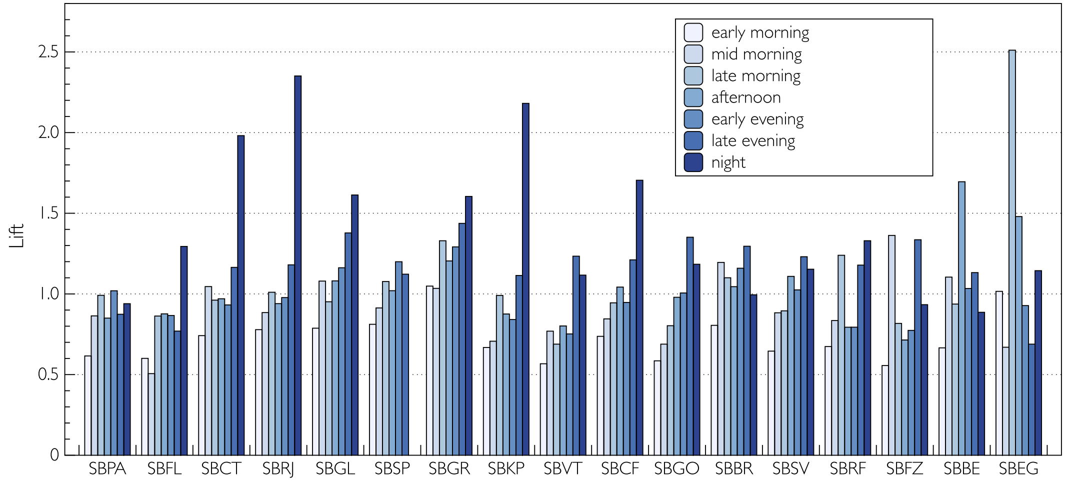


Fig. 3. Correlation matrix considering the Pearson coefficient between all the attributes of the Brazilian flight dataset.



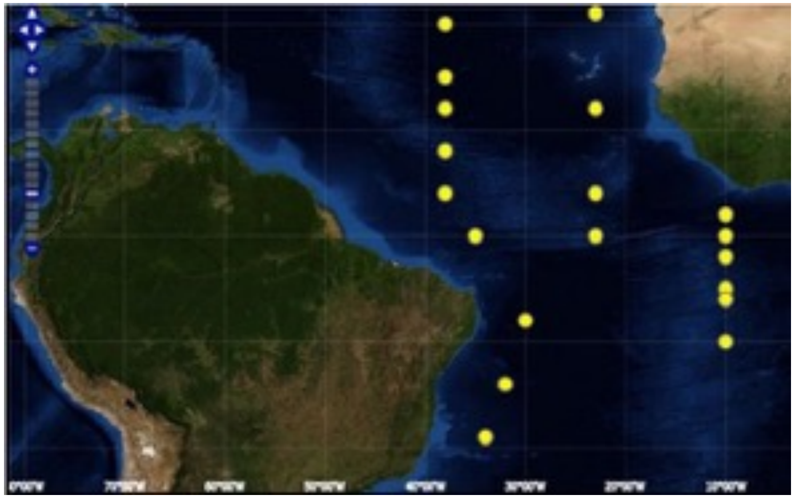
Alice Sternberg master degree

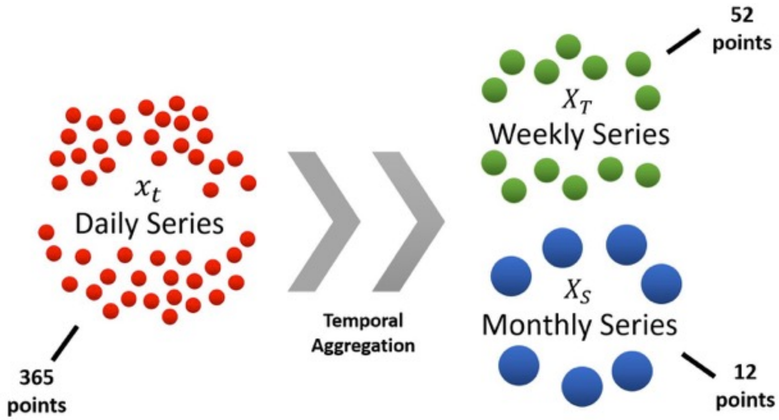
Fig. 7. Lift analysis of the rules containing the airport and the time of departure on the antecedent and a delay on the consequent – the airports are ordered from south to north.

Flight Delays – Research Opportunities

- Airport delays propagation
  - On going
- Flight delays propagation
  - On going
- Prediction of flight delays
  - On going*
- Replication of techniques using American datasets

- Long term prediction of sea surface temperature





Time-Series Prediction – Results

- Framework for analysis of prediction performance compared to linear models

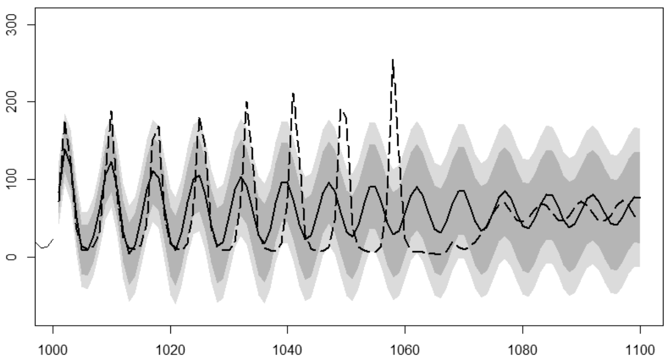


Fig. 2: ARMA predictions (solid line) for the time series A of the Santa Fe Competition. The actual time series values are represented by the dashed line.

TABLE III: Rankings of the top 25 results of the chosen competition datasets including results from TSPred R-package

Rank	Santa Fe				EUNITE		CATS			NN3		NN5	
	Dataset A		Dataset D							Dataset A		Dataset A	
	index	NMSE	index	NMSE[1]	Participant	MAPE [%]	Participant	E1	E2	Participant	Mean SMAPE	Participant	Mean SMAPE
1	W	0.02	ZH	0.08	Chih-Jen Lin	1.982	Sarkka*	408	346	Illies*	15.18%	Andrawis	20.40%
2	Sa	0.08	**TSPred$_{(ARIMA)}$**	**0.54**	Esp	2.149	Cai*	441	402	Adeodato*	16.17%	Vogel	20.50%
3	M	0.38	U	1.30	Brockmann	2.498	Kurogi*	502	418	Flores*	16.31%	D'yakonov	20.60%
4	L	0.45	**TSPred$_{(PR)}$**	**1.61**	**TSPred$_{(PR)}$**	**2.779**	Hu*	530	370	Chen*	16.55%	Rauch	21.70%
5	U	0.62	Z	4.80	Zivcak	2.873	Palacios-Gonzalez	577	395	D'yakonov	16.57%	Luna	21.80%
6	A	0.71	C	6.40	Kowalczyk	2.985	Maldonado*	644	542	Kamel*	16.92%	Wichard	22.10%
7	McL	0.77	W	7.10	Lewandowski	3.223	Simon*	653	351	Abou-Nasr	17.54%	Gao	22.30%
8	**TSPred$_{(ARIMA)}$**	**0.90**	S	17.00	Kowalczyk	3.264	Verdes*	660	442	Theodosiou*	17.55%	Puma-Villanueva	23.70%
9	**TSPred$_{(PR)}$**	**0.99**			Ortega	3.380	Chan*	676	677	**TSPred$_{(ARIMA)}$**	**17.79%**	Dang	25.30%
10	N	1.00			King	3.388	Wichard*	725	222	de Vos	18.24%	Pasero	25.30%
11	P	1.30			Lotfi	3.389	Beliaev*	928	762	Yan	18.58%	Adeodato	25.30%
12	Can	1.40			Guijarro	3.421	Kong	954	994	C49	18.72%	undisclosed	26.80%
13	K	1.50			Weizenegger	3.694	Wang	1037	402	Perfilieva*	18.81%	undisclosed	27.30%
14	Sw	1.50			**TSPred$_{(ARIMA)}$**	**3.820**	Cellier*	1050	278	Kurogi*	19.00%	**TSPred$_{(ARIMA)}$**	**27.80%**
15	Y	1.50			Boger	3.958	Crone*	1156	995	Beadle	19.14%	Tung	28.10%
16	Car	1.90			Bontempi	3.997	**TSPred$_{(ARIMA)}$**	**1173**	**917**	Lewicke	19.17%	undisclosed	33.10%
17					Pelikan	4.348	Acernese*	1247	1229	Sorjamaa*	19.60%	undisclosed	36.30%
18					Brockmann	4.373	Yen-Ping*	1425	894	Isa	20.00%	undisclosed	41.30%
19					Pelikan	4.437	**TSPred$_{(PR)}$**	**7387**	**6778**	C28	20.54%	**TSPred$_{(PR)}$**	**41.50%**
20					Riviecchio	4.502				Duclos-Gosselin	20.85%	undisclosed	45.40%
21					Brockmann	4.580				Papadaki*	22.70%	undisclosed	53.50%
22					Ivakhnenko	4.653				Hazarika	23.72%		
23					Brockmann	4.712				C17	24.09%		
24					Brockmann	5.087				Njimi*	24.90%		
25					Brockmann	5.425				Pucheta*	25.13%		

* et al.

[1] NMSE error for the 15 first predicted observations

Time-Series Prediction – Results

- Effect of temporal aggregation for long-term prediction of sea surface temperature

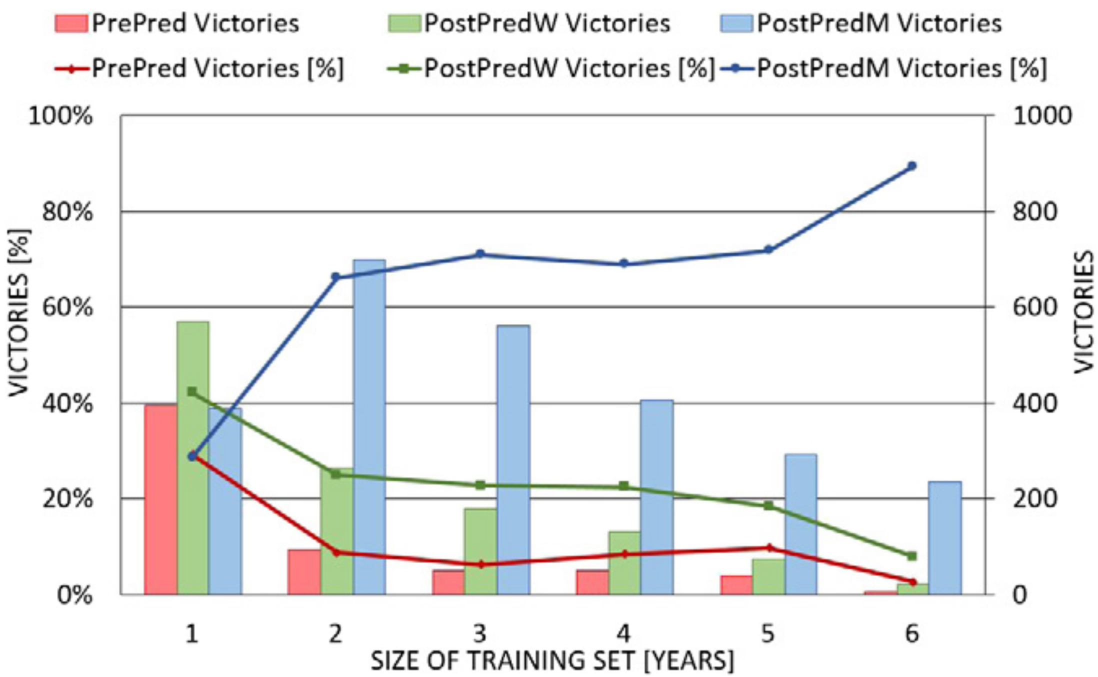


Fig. 8. Graphic of the victories of each prediction approach regarding their performances in generating up to twelve monthly aggregated forecasts.

Rebecca Salles
Scientific initiation

Time-Series Prediction – Research Opportunities

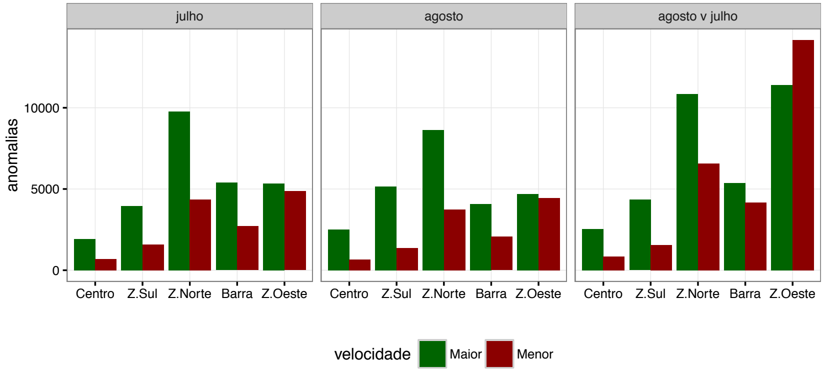
- Expansion of framework prediction for machine learning methods
  - On going
- Study of different preprocessing methods for supporting non-stationarity
  - On going
- Creation of novel methods for non-stationarity for machine learning methods

Urban Mobility

Approximately more than 4 million of observations per day
Bus as trajectory sensors
Spatial-Temporal Aggregation: Regions as virtual sensors

Urban Mobility – Results

- Data collection (done by UFF)
- Data Cleaning, Spatial-Time Aggregation
- Preliminary Analysis of Anomalies



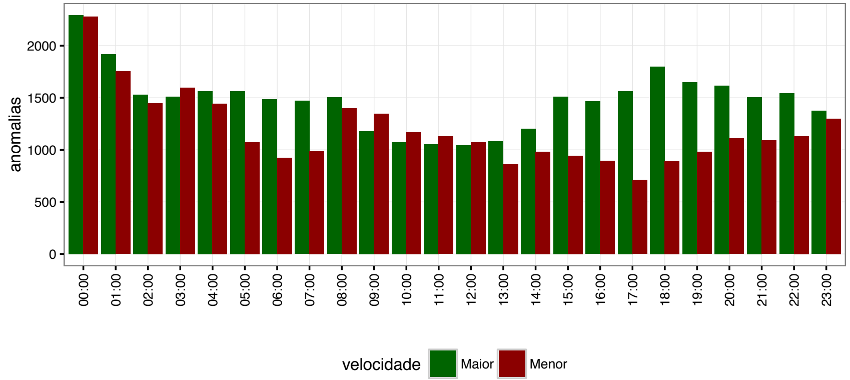
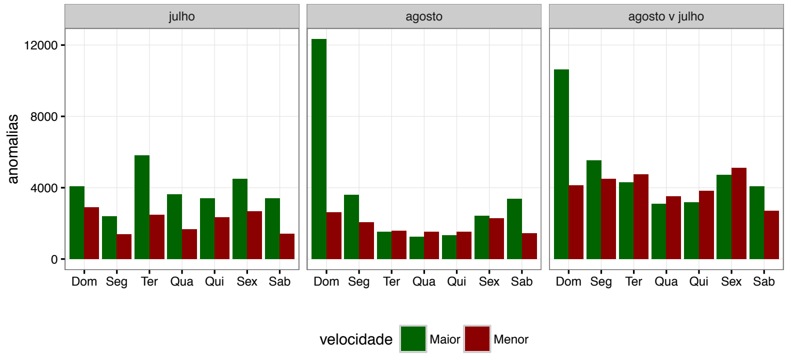


Figura 3. Anomalias identificadas por faixa de horário (ago v julho)

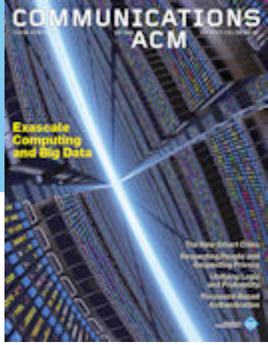


Ana Beatriz Cruz
Master degree

Urban Mobility – Research Opportunities

- Persistence and Querying
- Trajectory or Aggregated analysis
- Identification of Patterns, Anomalies, and Paradigm Change

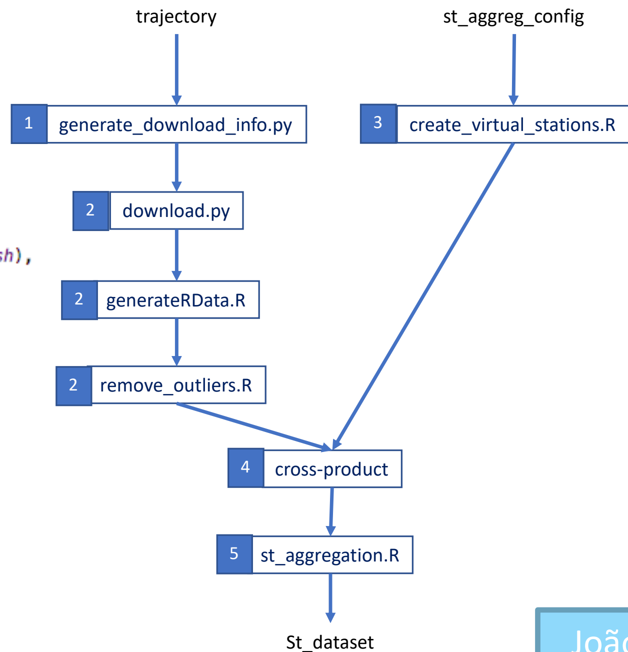
Research Project In Management and Analysis of Spatial-Time Series



- Non-stationary resilient techniques in data preprocessing
- Novel algorithms for prediction/classification and pattern identification
  - Motif identification
  - Tight spatial-time sequence mining
- Explore spatial-time series applications
  - Frequent pattern mining, Classification/Prediction
- Explore data management and parallel processing for mining non-stationary Big Data
  - Algebraic-based spatial-time series data mining workflow using Spark

Parallel and Distributed Execution Using Spark

```
val trajectory: Relation = Relation(Schema(key, initialTime, endTime),
  Tuple("copa-do-mundo-2014", "2014-06-01", "2014-07-31"))
val st_aggreg_config: Relation = Relation(Schema(radius, interval, busesMesh),
  Tuple("10", "10", "malha-2014.csv"))
w = Workflow("2014CupAggregation", () => {
  r1 = SplitMap(Activity("generate_download_info.py"), key, trajectory)
  r2 = Map(Activity("download.py"), r1)
  r3 = Map(Activity("generateRdata.R"), r2)
  r4 = Map(Activity("remove_outliers.R"), r3)
  r5 = Map(Activity("create_virtual_stations.R"), st_aggreg_config)
  r6 = Query(CrossProduct, r4, r5)
  result = Map(Activity("st_aggregation.R"), r6)
})
w.execute()
```



João Ferreira
Master degree

(a) (b)

Figura 2. Workflow para análise de tráfego durante a COPA de 2014 : a) Especificação do Workflow usando linguagem Scala; b) grafo mostrando as dependências entre as atividades

Recent Published Papers Related to the Project

- Cruz A. et al., 2017 - Detecção de anomalias no transporte rodoviário urbano. In SBBD.
- Ferreira J. et al, 2017 - Uma Proposta de Implementação de Álgebra de Workflows em Apache Spark no Apoio a Processos de Análise de Dados. In: BreSci
- Salles R. et al. 2017 - A Framework for Benchmarking Machine Learning Methods Using Linear Models for Univariate Time Series Prediction, IJCNN
- Marinho A. et al. 2017 - Deriving scientific workflows from algebraic experiment lines: A practical approach. Future Generation Computer Systems.
- Guedes G. et al. 2016 - Discovering top-k Non-Redundant Clusterings in Attributed Graphs. Neurocomputing.
- Sternberg A. et al., 2016 - An analysis of Brazilian flight delays based on frequent patterns. Transportation Research. Part E, Logistics and Transportation Review
- Salles R. et al, 2016 - Evaluating Temporal Aggregation for Predicting the Sea Surface Temperature of the Atlantic Ocean. Ecological Informatics.
- Machado E. et al, 2016 - Exploring machine learning methods for the Star/Galaxy Separation Problem. In: IJCNN
- Cruz A. et al, 2016 - Identificação de Motifs em Agregações de Séries Espaço-Temporais de Mobilidade Urbana. In: WTDBD/SBBD
- Campisano, R., Porto. F., Pacitti, E., Florent M., Ogasawara E., Spatial Sequential Pattern Mining for Seismic Data. In: SBBD
- Salles et al., 2015 - Evaluating Linear Models as a Baseline for Time Series Imputation. In: SBBD
- …
- Ogasawara, E. et al., 2010 Adaptive Normalization: A Novel Data Normalization Approach for Non-Stationary Time Series. In: IJCNN.

Main collaborators

Inria
informatics mathematics

National Laboratory Scientific Computing

COPPE UFRJ
Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia

Instituto de computação uff

Observatório Nacional

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

FGV

CEFET/RJ Team (12 active students)

Graduate students

D.Sc.
- Heraldo Borges

M.Sc.
- Ana Beatriz Cruz
- Carla Palmieri
- Fernanda Britto
- João Ferreira
- Lais Baroni
- Rebecca Salles

Undergraduate

Final Project
- Adílio Rosa
- Bernardo Monteiro
- Felipe Feder
- Pedro Castro
- Philipp Mendonça

Opportunities:
- Scientific initiation
- Final Projects
- D.Sc. (PPPRO)
- M.Sc. (PPCIC)

Graduated

M.Sc.
- Leonardo Mosqueira
- Murillo Dutra
- Riccardo Campisano

Graduate
- Lara Mello
- Luana Fragoso

Recent defenses

M.Sc.
- Amir Khatibi

Final Projects
- Arthur Rita
- Christopher Dantas
- Diego Vaz
- Iuri Bloch
- Josué Dias
- Leonardo Oliveira
- Luana Piani

CEFET/RJ Team

Dec/2016

Aug/2017