

# Road map

- Overview of Time Series Event Detection
- Data-Centric Al Initiatives
- Challenges



- Time series events are commonly instants or intervals in the time series where observations change in a manner that is considered important for analysis or decision-making processes
  - The interpretation of an event can vary significantly across different domains
  - They can be categorized into main types: anomalies, change points, and motifs



[1] V. Guralnik and J. Srivastava, "Event Detection from Time Series Data," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '99. New York, NY, USA: ACM, 1999, pp. 33–42. doi: <u>10.1145/312129.312190</u>.

- Process of identifying events
- Important for monitoring and surveillance
  - Industry, seismic, oil exploration, epidemiology, climate
- There are many studies, but
  - Focused on specific types of events
  - Lacking a holistic view of the problem



#### Taxonomy



 Anomalies are observations that do not conform to the typical ones at the time series [1]



- Change points are time intervals where there is a significant change in the statistical properties in a time series [1]
  - This can include changes in mean, variance, correlation, distribution
- They represent a transition between different states in a process that generates the time series [2]



[1] T. Górecki, L. Horváth, and P. Kokoszka, "Change point detection in heteroscedastic time series," Econometrics and Statistics, vol. 7. pp. 63–88, 2018. doi: 10.1016/j.ecosta.2017.07.005.

[2] C. Truong, L. Oudre, and N. Vayatis, "Selective review of offline change point detection methods," Signal Processing, vol. 167. 2020. doi: 10.1016/j.sigpro.2019.107299

# **Motifs**

- Time series motifs are sequences of significantly similar observations within a time series
  - It is an approximately repeated subsequence within a longer time series [1]



8

#### **Offline versus online detection**



[1] E. Ogasawara, R. Salles, F. Porto, and E. Pacitti, *Time Series Event Detection*. Springer, (to appear).

#### **Basic metrics for event detection**



# Data-Centric AI Initiatives

- Data-centric AI is an approach that emphasizes data preparation
  - Data Quality: accurate, complete, and representative data
  - Data Transformation: normalization, encoding categorical variables, and transforming features to improve model performance
  - Feature Engineering: new features to better capture the underlying patterns
  - Data labeling: Maintaining consistent data labels
  - Bias mitigation: Identifying and addressing biases in the data
  - Data augmentation: Using techniques to increase dataset size artificially



It might be a buzzword for data preprocessing

- Integrated normalization for sliding windows
- Compute a moving average for each sliding window
- Differentiate in each sliding window observation relative to its moving average
- Remove windows with outliers
- Scale each window between 0 and 1 with respect to the maximum and minimum differences of all windows



PERFORMANCE OF ALGORITHMS TO FORECAST THE MONTHLY AVERAGE EXCHANGE RATE OF U.S. DOLLAR TO BRAZILIAN REAL TIME SERIES

	RMSE		
Algorithm	1-step	12-step	
AR	0.082	0.545	
NN-MM	0.177	1.173	
NN-DS	0.094	1.444	
NN-ZS	0.126	0.814	
NN-SW	0.088	0.451	
NN-AN	0.062	0.345	

[1] E. Ogasawara, L. C. Martinez, D. De Oliveira, G. Zimbrão, G. L. Pappa, and M. Mattoso, "Adaptive Normalization: A novel data normalization approach for nonstationary time series," Proceedings of the International Joint Conference on Neural Networks. 2010. doi: 10.1109/IJCNN.2010.5596746.

- Provides inertia during time series analysis
  - Higher moving average, higher inertia
- It usually provides good step-ahead predictions using machine learning
- It enables outlier removal (could be used for anomaly detection)
- Limitations
  - Should establish the moving average

#### **FBIAD: Forward-Backward Inertia Anomaly Detection**

Use AN ideas for anomaly detection



fanio Lima Rebecca Salles CEHTRO CENTRO jaciolina/brano acito gle mbecca sales/bric acito gle Pedro Alpis Laciana Escobar CENTRO CONTROL CENTROL IN	Fubio Putto Rafaelli Coasisteo LNCC CERTORU gerio@haci.tr rafaeli.constele@cente.tjh Exister Pacitii Eduardo Opanesan Lik. & University of Monpeliar CERTORU
pedmatps/#incested-ej/tr lastamonicolme/#escentei-ej/tr Aforson-Than write event detection is related to multiple methods for detecting observations is a write with special restricted at a second second second second second second restricted at a second second second second second second or there are a former in the second of data second second second second second secon	edherpacili#mm.tr angunyars#inec.etg The literature presents many methods developed to dat events in time series. However, theoring and applying withit method for a proof time series to not a simple init. However, anguneerization of a development are dete
then error. Therefore, shipping to then error charge between result. Thes, bulkering errors associated with their charges is constant for intelly and corrict devidementaling Athmship likes are any sensitive to inter errors in a still public between the state of the state of the state of the state and the state of the state of the state of the state and historical between 40% is order to fill the pape in the floresters, the vert AL condition is possible, among lowers and historical between 40% is because 10% is the state of the state of the state of the state of the state of the state of the state of the state of the state of the state between its that are state. In condition is possible of the state between the state of the state of the state of the state between the state of the state of the state of the state of the state between the state of the st	retard to the initial assumptions regarding the behavior the time series and the matrixed properties intrinsic to data. It is a complex tank, particularly considering that is manuer of events observed in a time action is often abased in the series of the series of the series of the abased in the series of the series of the series of the series of the abased of the series of the series of the series of the series of the series of the series of the series of the series of the series. These areas is not series at the series of the series of the series of the series of the series of the series. There is a loss of codMMrg in and these techniques to control applications is both cases.
In the second s	We can also the probability of

[1] J. Lima, R. Salles, F. Porto, R. Coutinho, P. Alpis, L. Escobar, E. Pacitti, and E. Ogasawara, "Forward and Backward Inertial Anomaly Detector: A Novel Time Series Event Detection Method," Proceedings of the International Joint Conference on Neural Networks, vol. 2022-July. 2022. doi: 10.1109/IJCNN55064.2022.9892088. Dataset studied (Yahoo, Numenta and Gecco)

Method	Precision	Recall	F1	Accuracy
FBIAD	0.066	0.528	0.085	0.731
ARIMA	0.045	0.556	0.067	0.746
LSTM	0.041	0.534	0.062	0.735
ELM	0.041	0.517	0.063	0.726
Conv1D	0.036	0.519	0.055	0.724
SVM	0.030	0.542	0.049	0.732

#### **Inspecting Performance Comparison**



[1] J. Lima, R. Salles, F. Porto, R. Coutinho, P. Alpis, L. Escobar, E. Pacitti, and E. Ogasawara, "Forward and Backward Inertial Anomaly Detector: A Novel Time Series Event Detection Method," Proceedings of the International Joint Conference on Neural Networks, vol. 2022-July. 2022. doi: 10.1109/IJCNN55064.2022.9892088.

# Addressing moving average limitation using EMD

 Empirical Mode Decomposition (EMD) is a technique for decomposing non-linear and non-stationary series into a series of functions called Intrinsic Mode Functions (IMFs)



[1] Y. Lei, J. Lin, Z. He, and M. J. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," Mechanical Systems and Signal Processing, vol. 35, no. 1–2. pp. 108–126, 2013. doi: 10.1016/j.ymssp.2012.09.015.

#### REMD

- REMD: A hybrid method consisting of four steps
  - EMD decomposition
  - IMF aggregation
  - ARIMA adjustment
  - Anomaly detection: analysis of distribution error





- Datasets: Yahoo, Numenta, and Gecco
- REMD presents a much better performance than the secondplaced method
  - EMD-based method, when we use F1 as the main selection criterion

Method	Precision	Recall	F1
REMD	0.684	0.386	0.448
EMD	0.243	0.408	0.207
FBIAD	0.066	0.528	0.085
ARIMA	0.045	0.556	0.067
LSTM	0.041	0.534	0.062
ELM	0.041	0.517	0.063
Conv1D	0.036	0.519	0.055
SVM	0.030	0.542	0.049

#### Inspecting performance comparison



[1] J. Souza, E. Paixão, F. Fraga, L. Baroni, R. F. S. Alves, K. Belloze, J. Santos, E. Bezerra, F. Porto, and E. Ogasawara, "REMD: A Novel Hybrid Anomaly Detection Method Based on EMD and ARIMA," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2024-July. pp. 1–8, 2024.

# Time should count while evaluating events



[1] R. Salles, J. Lima, R. Coutinho, E. Pacitti, F. Masseglia, R. Akbarinia, C. Chen, J. Garibaldi, F. Porto, and E. Ogasawara, "SoftED: Metrics for Soft Evaluation of Time Series Event Detection." arXiv, Apr. 01, 2023. doi: 10.48550/arXiv.2304.00439.

# Still need improvements when it comes to streaming and online events

- The schizophrenic behavior of detectors in online detection
- Detection Probability:  $DP(x_i) = \frac{df(x_i)}{bf(x_i)}$ 
  - *df*: detection frequency
  - *bf*: batch frequency
- Detection Lag:  $Lag_i^s = fdb_i sb_i$ 
  - *fdb* (first detection batch)
  - sb (start batch)

João Eduardo Ferreira USP	Ismael Sa Potrobra	antine M	habeta	Guimarães Sigarira Potobras
bisaine refera	unastrepton	GLOBAT I	(all search all	ware balons but it.
Diego Carvalho	Fabio Porto	Rafaelli Ca	ofeites	Eduardo Ogasawara
CEPETARS	LNCC	CIPET	9K.1	CERTINE
d.carvatho@iese.org	Ipono@iscche r	alaelii.contisho	#celei@ht	solaran ang solo tal
Alterar-fullie next densities a critical tack with applications of the for correctly alternating the image by for correctly alternating the image by the correctly alternating the image between the correct sourcement and al- ternation of the image between the image densities of the appear iteration of the alternation of the appear iteration of the alternation of the alternation of the alternation of the alternation of the alternation of the alternation of the alternation of the alternation of the alternation of the alternation of the alternation of the alternation of the	in meaning the order is reverse trainer domains, its entition for counted options is a related to the events. How excitons the option of the option is size differences more role obtained option of the option obtained option of the option obtained option option obtained option option of the option option option is a size of the option option option option option option is a size of the option option of a large series it is a created option option option.	bebaster of on inform identifying To explo- become in and pring is and pring is and pring is and pring is and pring is described in detection fiere aspect for event de bitty and is bitty and is the results	rnathods through atom about the changes in the re these questi- tion areas, this ine action. The line of a detection in the the interde- tion interde- tion interde- res are imported in the interde- tion. Section II y tection. Section is the experime of the experime granthysis in c.	chest the streaming process" Ho behavior of methods counthus instanting lines water? ms and chemica stream waters of even detection workstone for or entropology to any result and to even any other streams and the stream of the stream of the proton countribution of the paper proton countribution of the paper behavior of the paper of the proceed pub- ter and section N presents of
identified using change point or a	nosuly delection method			
N 9 10 111 When offline they	rely on prior second to other	1000	1L B	ACADAGOND
historical data. When online, they by monitoring the streaming of pro-	have to detect occurrence cess data while it is being	This sect subsequence	e analysis conci	fundamental event detection at opts in time series.
generated (12, 9, 13); Many control visitors downall	richt.on.line desertion h	A. Even D	encasion	
enable undestrable situations to I	w addressed promptly, h	. Event de	incline can be	based on predicted values, who
this way, continuously monitoring :	dreaming time series raise	a methods in	clude a step for	predicting the next values in d
an important fact. An observation	this characterizes as even	t titte setes.	Events are det	ected when there is a discrepan
might not be detected when the s	diservation appears in th	e between the	e actual value o	f a new observation and prodicts
time series. Now observations might	a bu appended belins even	<ul> <li>values, Alte</li> </ul>	matively, event	s can be detected based on boan
And the second s	4, 15, 16]. Thin, if as even	aties, defini	ing upper and he	wer limits beyond which each ne
defension methods discover meth [1	and they have confirmed which of the	a conversion of		eked in an event [17, 18, 16, 7, 15
occurs in time 1 had is only detec	and by the contrast wave-cont			
access in time 1 hat is only doted method at time 1 + 8, the detection Considering the detection lag pro-	tag to A. biem, it is possible to rais	There are streaming to	e beo main me ime series stabi	stes for cellae event detection c and dynamic. In static mode, if
occurs in time 2 hat is only dote method at time 2 hat is only dote considering the dotection lag pre- the following research questions:	t lag in 8. doinn, it is possible to rais Now can one evaluate th	There an streaming to model deter	t two main me ine series stati- ts events in the	ides for celline event detection c and dynamic. In static mode, if steaming time series using a pr
decision mension discover tens () occurs in time 2 had is only dotec method at time 2 h 6, the detection Considering the detection lag pre- the following research questions: clapsed time between when an ob-	t lag in X. blem, it is possible to rais flow can tote evaluate th ervation in the time serie	There an streaming it model deter trained mod	t two main me ime series stati its events in the fet with historic	sdes flar crelike event detection c and dynamic. In static mode, it streaming time series using a pr at data. In dynamic axide, trainis

Online Event Detection in Streaming Time Series:

Novel Metrics and Practical Insights



[1] J. Lima, L. G. Tavares, E. Pacitti, J. E. Ferreira, I. Santos, I. G. Siqueira, D. Carvalho, F. Porto, R. Coutinho, and E. Ogasawara, "Online Event Detection in Streaming Time Series: Novel Metrics and Practical Insights," Proceedings of the International Joint Conference on Neural Networks, vol. 2024-July. pp. 1–8, 2024.

- Holistic view of the problem
  - Anomalies
  - Change points
  - Motif discovery
- Properties
  - Uniform Data Model
  - Rigid interface (algebraic)
  - Expansible
  - Based on experimental line
- Inspiration from Sci-Kit Learn
  - Fit()
  - Detection()
- More than 50 event detectors
- R Package available at CRAN



Harbinger is a framework for event detection in time series. It provides an integrated environment for time series anomaly detection, change points, and motif discovery. It provides a broad range of event detection methods and functions for plotting and evaluating event detections.

In the anomaly classes, methods are based on machine learning model deviation (Conv1D, ELM, MLP, LSTM, Random Regression Forest, SVM), machine learning classification model (Decision Tree, KNN, MLP, Naive Bayes, Random Forest, SVM), clustering (kmeans and DTW) and statistical methods (ARIMA, FBIAD, GARCH).

In the change points classes, methods are based on linear regression, ARIMA, ETS, GARCH, AMOC, ChowTest, BinSeg, GFT, PELT.

In the motifs classes, methods are based on Hash and Matrix Profile. There are specific methods for multivariate series. The evaluation of detections includes both traditional and soft computing.

Harbinger architecture is based on Experiment Lines and is built on top of the DAL Toolbox. Such an organization makes it easy to customize and add novel methods to the framework.

Installation

The latest version of Harbinger at CRAN is available at: <u>https://CRAN.R-project.org/package=harbinger</u>

You can install the stable version of Harbinger from CRAN with:

install.packages("harbinger")

You can install the development version of Harbinger from GitHub https://github.com/cefet-rj-dal/harbinger with:

# install.packages("devtools")
library(devtools)

# **CEFET/RJ** Team

#### D.Sc. students











Lucas Giusti\*



**Paulo Elias\*** (UFF)

#### M.Sc. students



Antônio Mello



**Arthur Garcia** 



**Cristiane Gea** 



**Edson Sobrinho** 



Fabiana Santos\*





Jéssica de Souza



**Josélia Rabelo** 



Luiz Oliveira



Michel Reis\*



**Ricardo Buçard\*** 



**Rodrigo Machado** 



Thiago Marques

- Multivariate Change-point Detection and Concept Drift
- Volatility Anomalies and Change Points
- Model Management Under Concept Drift
- Detection Metrics for Interval Events
- Data Augmentation for Time Series



# Challenges: Multivariate Change-point Detection and Concept Drift

- Virtual Concept Drift (data distribution)
- Real Concept Drift (model change)
- Virtual Concept Drift
  - Univariate time series (detection in isolation)
  - Multivariate time series

# Multivariate Change-point & Concept Drift



- Compute 1<sup>st</sup> principal component
- Analyze transformed time series



# **Consider a 3D time series**



# **PCA** does not capture change points



- Consider a multivariate time series  $X = \{x_1, x_2, x_3, x_4, x_5\}$
- Train multivariate to produce same output
- Extract encoded layer
- Detect change points in encoded time series



#### **Detection with autoencoder**



# Challenges: Model management under concept drift

- Assume we have model M<sub>0</sub>
- Novel concept appear and new model is built M<sub>1</sub>



- New batch introduces new concept
- New model is built *M*<sub>2</sub>



- New batch introduces new concept
- New model is built M<sub>3</sub>



- New batch introduces new concept
- Should we build new model  $(M_4)$  or can we reuse  $M_1$ ?
- Could we forget M<sub>0</sub>?



# Challenges: Detection Metrics

#### **Detection Metrics for Interval Events**

Traditional scoring methods, such as precision and recall, are

not sufficient to assess the performance of event detection OFTED: METRICS FOR SOFT EVALUATION OF TIME SERIES EV DETECTION They do not incorporate time and do not reward close detections. True positives are rewarded All other outcomes are equally penalized 30 20 TP • FP • FN traditional 10 0 20 10 30 0

[1] R. Salles, J. Lima, R. Coutinho, E. Pacitti, F. Masseglia, R. Akbarinia, C. Chen, J. Garibaldi, F. Porto, and E. Ogasawara, "SoftED: Metrics for Soft Evaluation of Time Series Event Detection." arXiv, Apr. 01, 2023. doi: 10.48550/arXiv.2304.00439.

# Challenges: Anomalies and Change Points related to Volatility

# **Detection of anomalies and change points related to volatility**

