

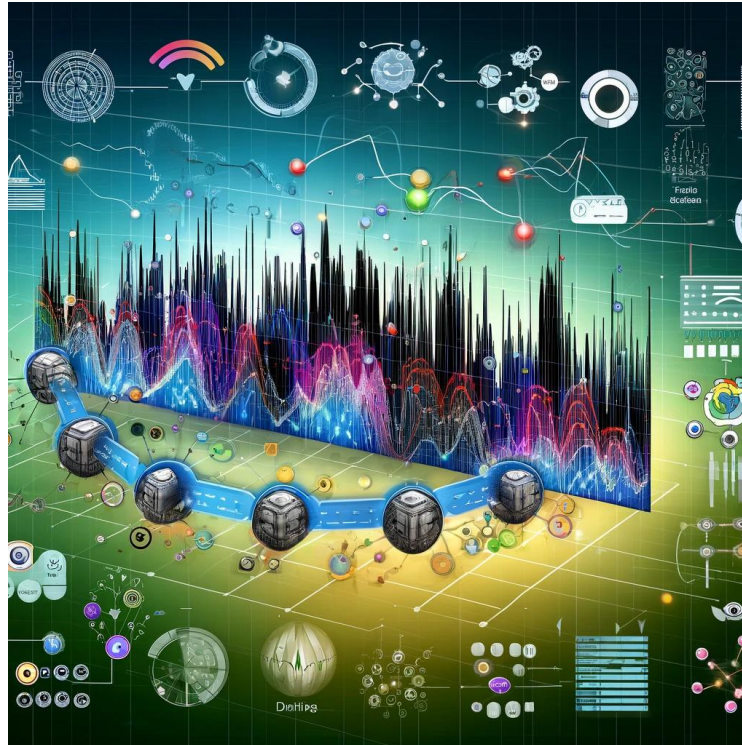


# DATA CENTRIC AI APPROACHES FOR TIME SERIES EVENT DETECTION

Eduardo Ogasawara  
eogasawara@ieee.org  
<https://eic.cefet-rj.br/~eogasawara>

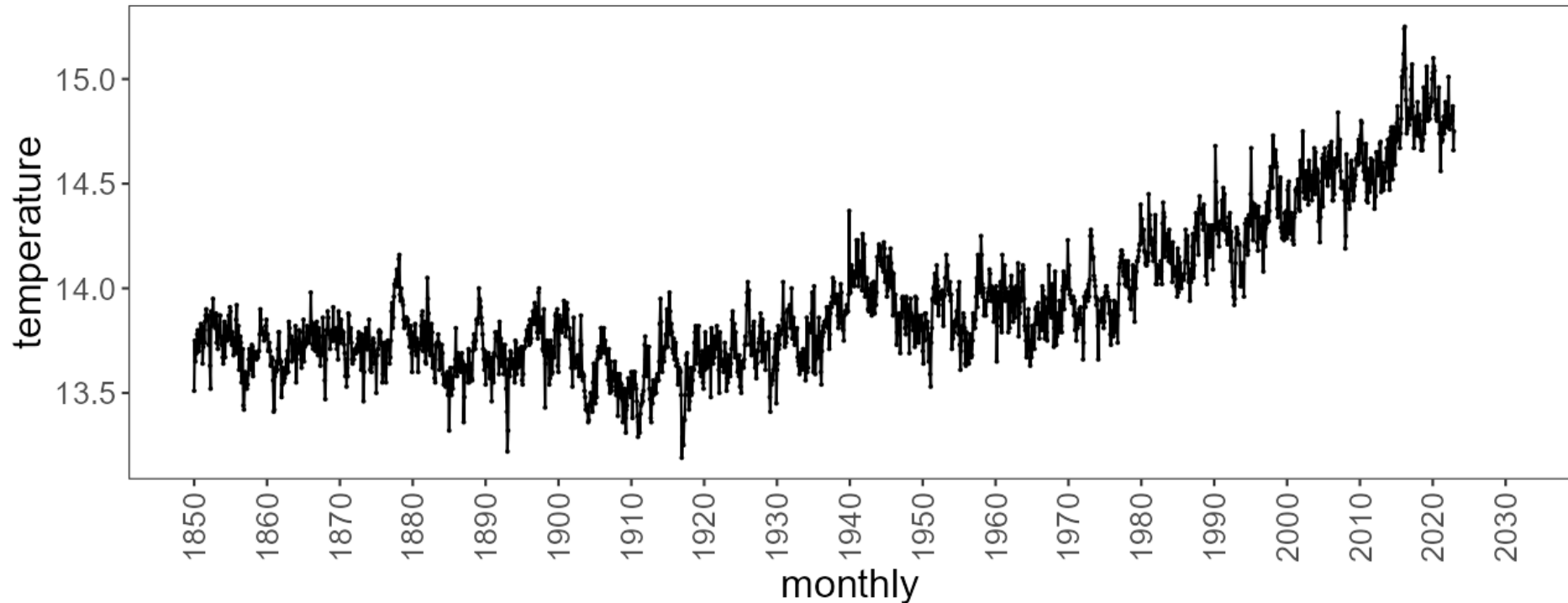
# Road map

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



## Time Series Events

- 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

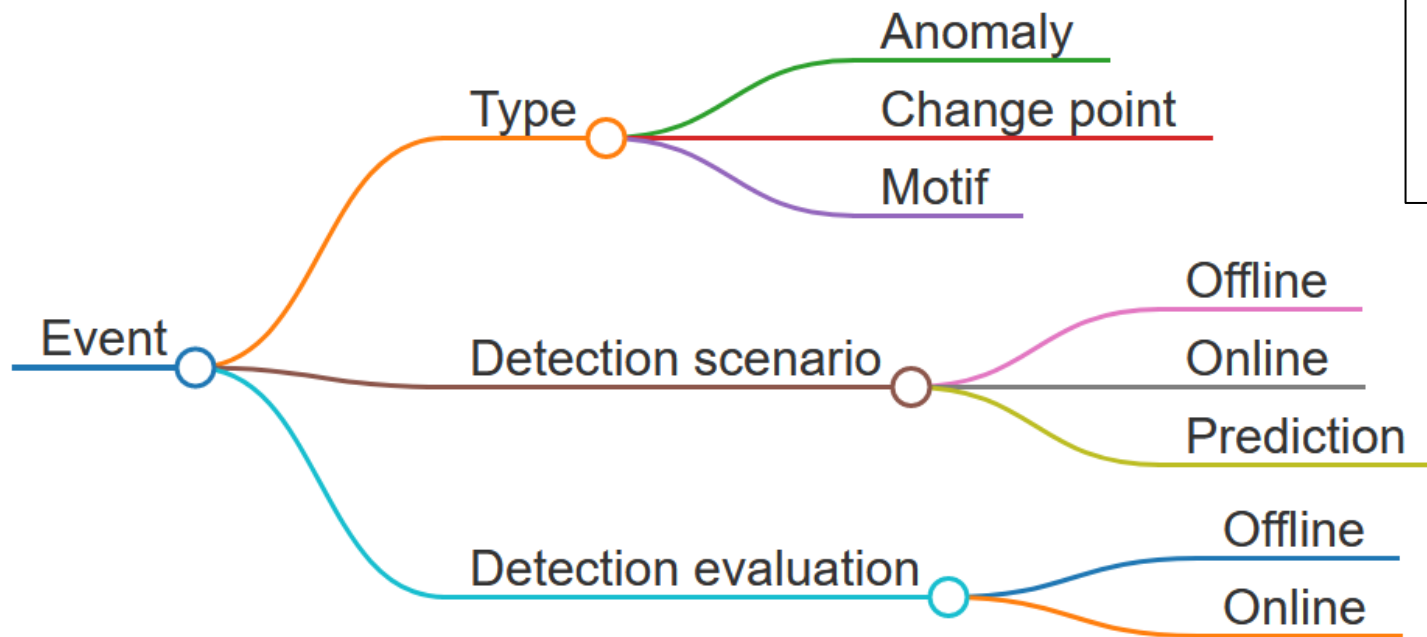


# ***Event Detection***

- 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



Eduardo Ogasawara, Rebecca Salles,  
Fábio Porto, Esther Pacitti

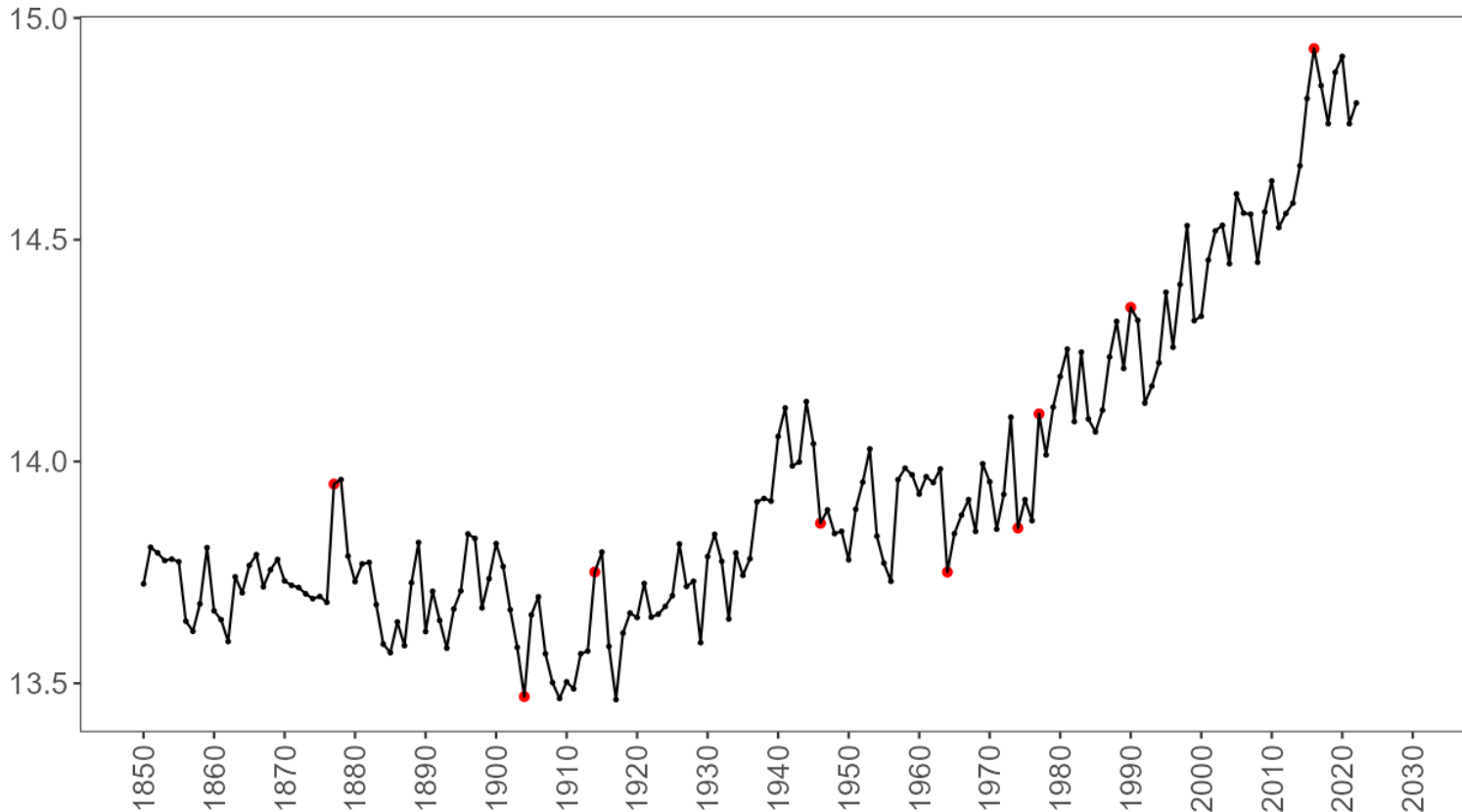
Event Detection in Time Series

May 26, 2024

Springer Nature

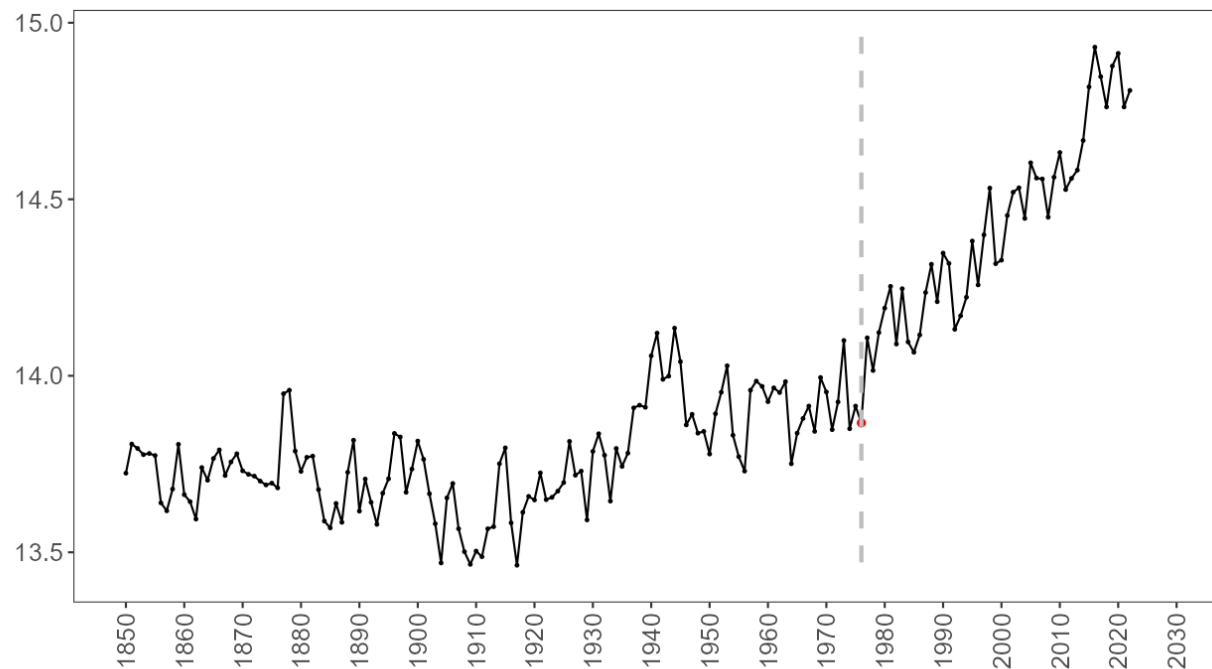
# Anomalies

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



# Change points

- 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]

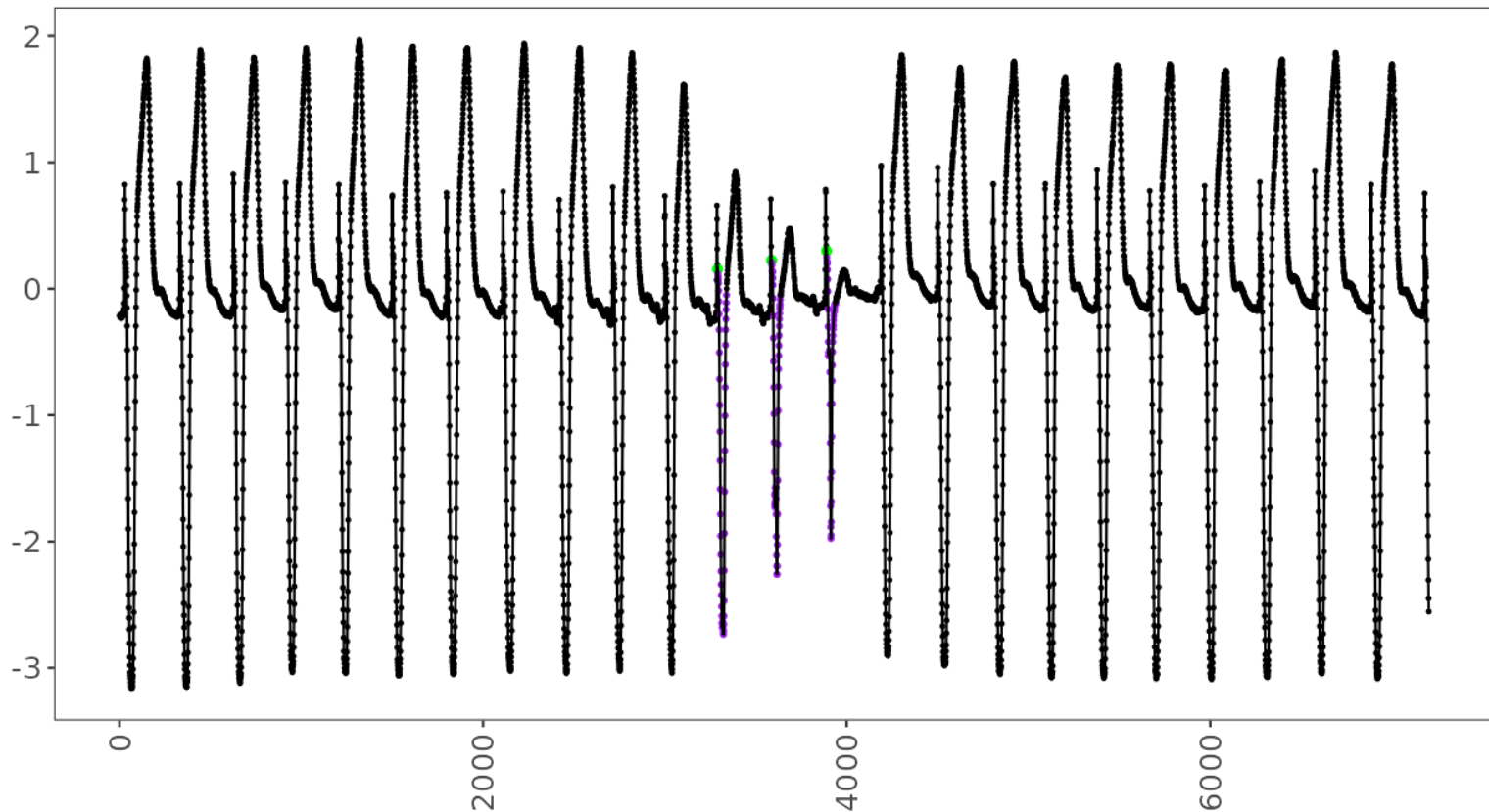


[1] T. Górecki, L. Horváth, and P. Kokoszka, "Change point detection in heteroscedastic time series," *Econometrics and Statistics*, vol. 7, pp. 63–88, 2018. doi: [10.1016/j.ecosta.2017.07.005](https://doi.org/10.1016/j.ecosta.2017.07.005).

[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](https://doi.org/10.1016/j.sigpro.2019.107299).

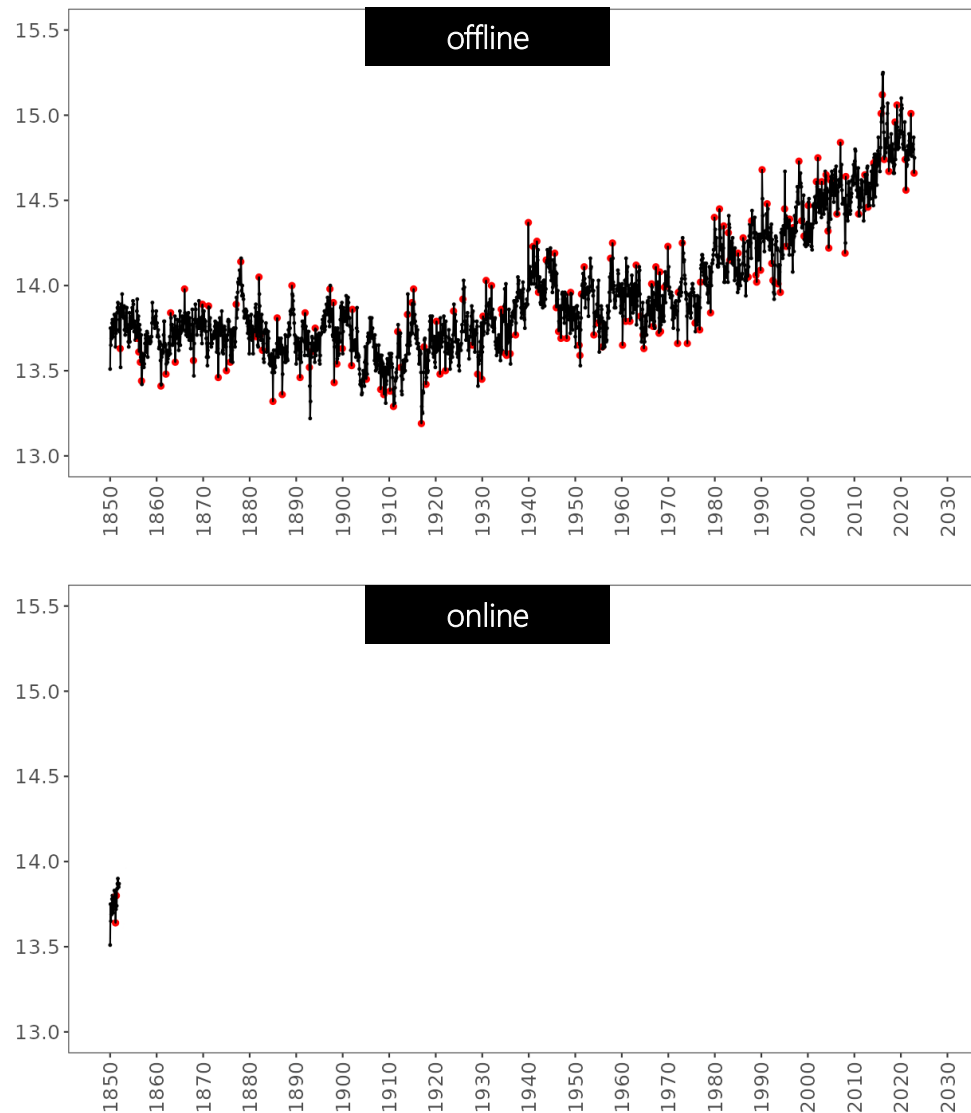
# Motifs

- 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]





# Offline versus online detection

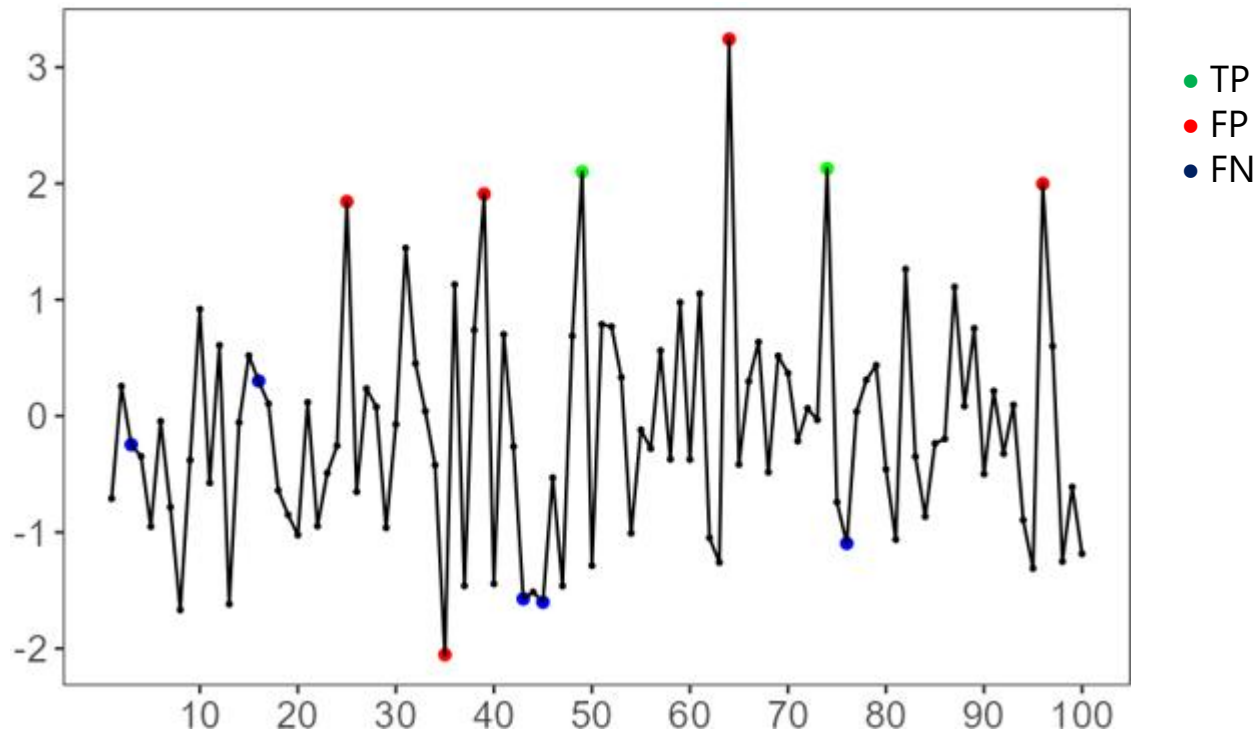


prediction



## Basic metrics for event detection

- $\text{precision} = \frac{TP}{TP+FP}$
- $\text{recall} = \frac{TP}{TP+FN}$
- $F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$



# Data-Centric AI Initiatives

# Data Centric AI

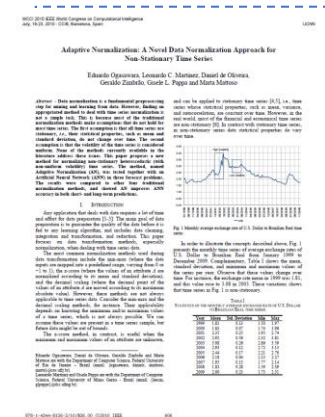
- Data-centric AI is an approach that emphasizes data preparation
  - Data Quality: accurate, complete, and representative data
  - Data Transformation: normalization, encoding categorical variables, and transforming features to improve model performance
  - Feature Engineering: new features to better 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*

# Adaptive normalization

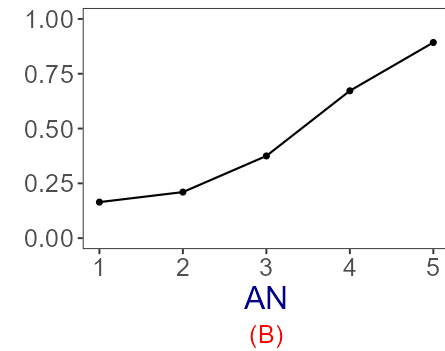
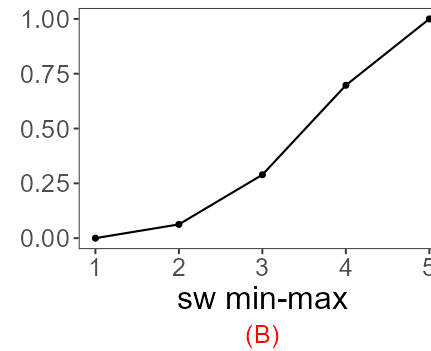
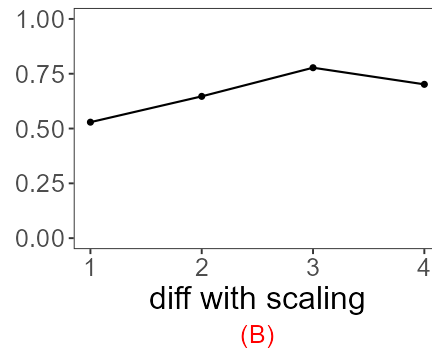
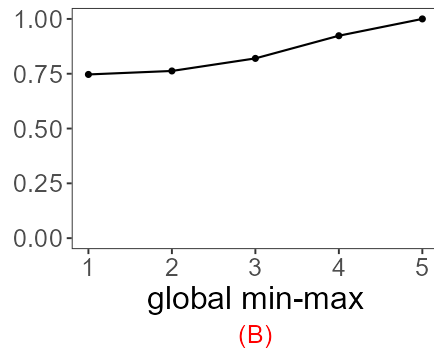
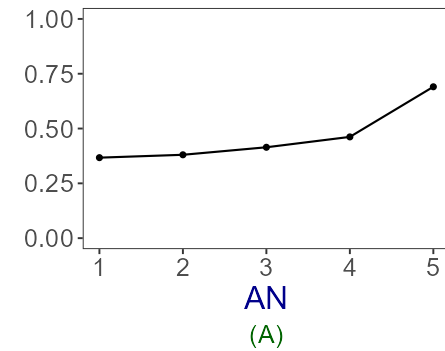
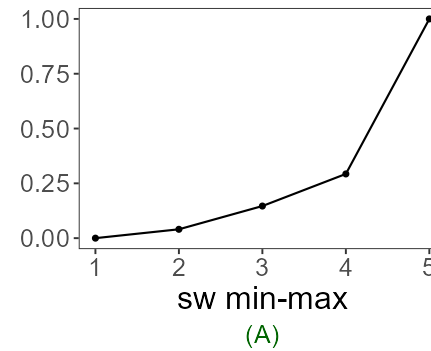
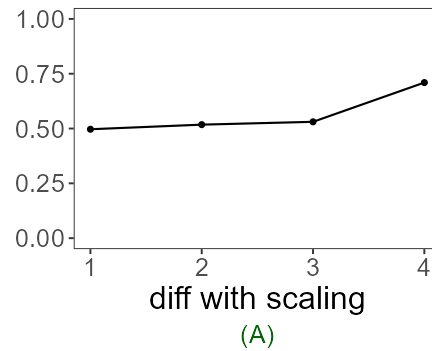
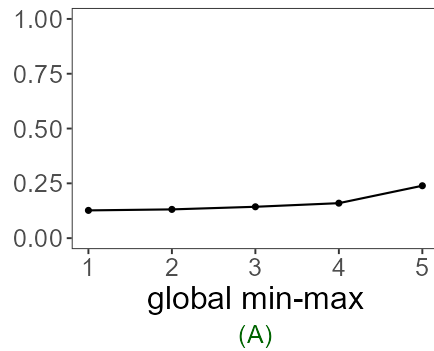
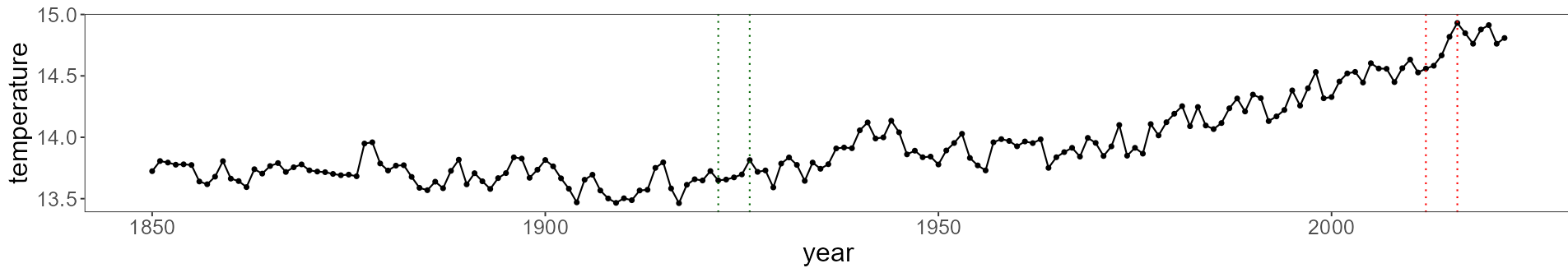
- Integrated normalization for sliding windows
- Compute a moving average for each sliding window
- Differentiate in each sliding window observation relative to its moving average
- Remove windows with outliers
- Scale each window between 0 and 1 with respect to the maximum and minimum differences of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



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

Algorithm	RMSE	
	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	<u>0.062</u>	<u>0.345</u>

# Inspecting comparison



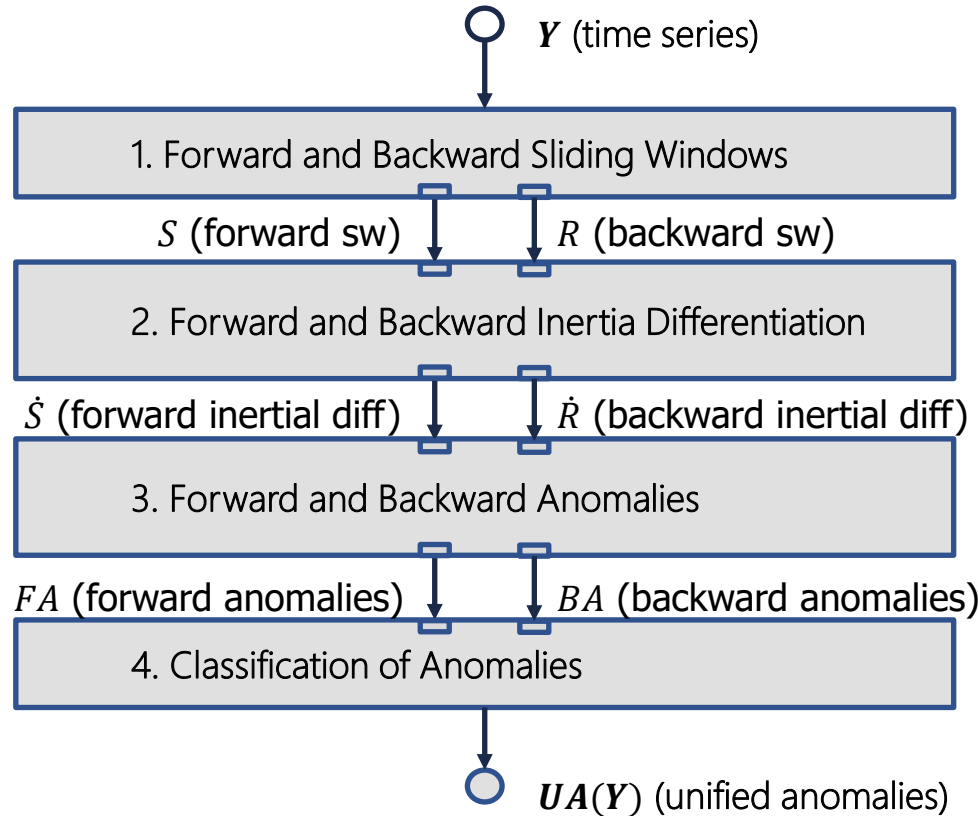
## ***AN Properties***

---

- 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



Forward and Backward Inertial Anomaly Detector:  
A Novel Time Series Event Detection Method

John Lima	Roberto Salles	Fabio Porto	Rafael Coutinho
UNICAMP	UNICAMP	UNICAMP	UNICAMP
john.lima@fci.unicamp.br	roberto.salles@fci.unicamp.br	fabio.porto@fci.unicamp.br	rafael.coutinho@fci.unicamp.br
UNICAMP	UNICAMP	UNICAMP	UNICAMP
john.lima@fci.unicamp.br	roberto.salles@fci.unicamp.br	fabio.porto@fci.unicamp.br	rafael.coutinho@fci.unicamp.br

The forward and backward inertial anomaly detector (FBIAD) is a novel time series event detection method. It is based on the idea of inertia, which is the resistance of an object to change its state of motion. In the context of time series, inertia is the resistance of a data point to change its value. The FBIAD is designed to detect anomalies in time series data by analyzing the inertia of each data point. It consists of four main steps: (1) Forward and backward sliding windows, (2) Forward and backward inertia differentiation, (3) Forward and backward anomalies, and (4) Classification of anomalies. The FBIAD is a novel time series event detection method that is based on the idea of inertia. It is designed to detect anomalies in time series data by analyzing the inertia of each data point. It consists of four main steps: (1) Forward and backward sliding windows, (2) Forward and backward inertia differentiation, (3) Forward and backward anomalies, and (4) Classification of anomalies.

1. FORWARD SLIDING WINDOWS  
In analyzing time series, it is often possible to observe a significant change in the behavior of a time series at a certain point or time interval. Such a behavior change usually characterizes the occurrence of an event. In this paper, we are interested in time series that can represent the occurrence of a phenomenon with specific and defined meaning to a certain domain of knowledge [1].  
Thus, the event detection problem is particularly relevant for applications based on sensor data streams. Examples of such applications include smart quality analysis [2], effective anomaly detection [3] and of detecting and predicting critical events [4] and of detecting and predicting critical events [5].  
From the point of view of time series, an anomaly is a data point that is significantly different from the rest of the data. In this paper, we will consider an anomaly to be a data point that is significantly different from the rest of the data. In this paper, we will consider an anomaly to be a data point that is significantly different from the rest of the data. In this paper, we will consider an anomaly to be a data point that is significantly different from the rest of the data.

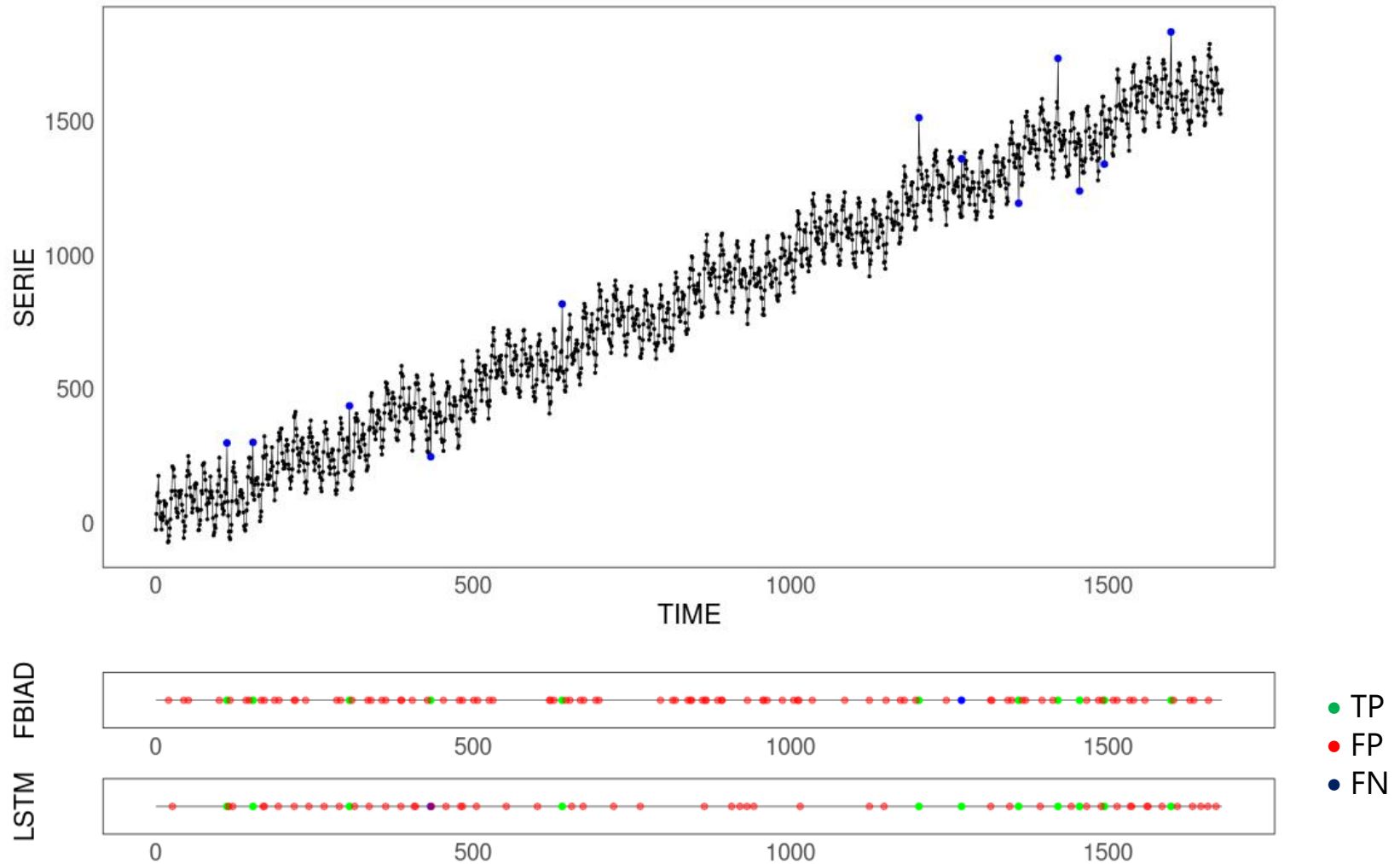


## ***Comparison***

- Dataset studied (Yahoo, Numenta and Gecco)

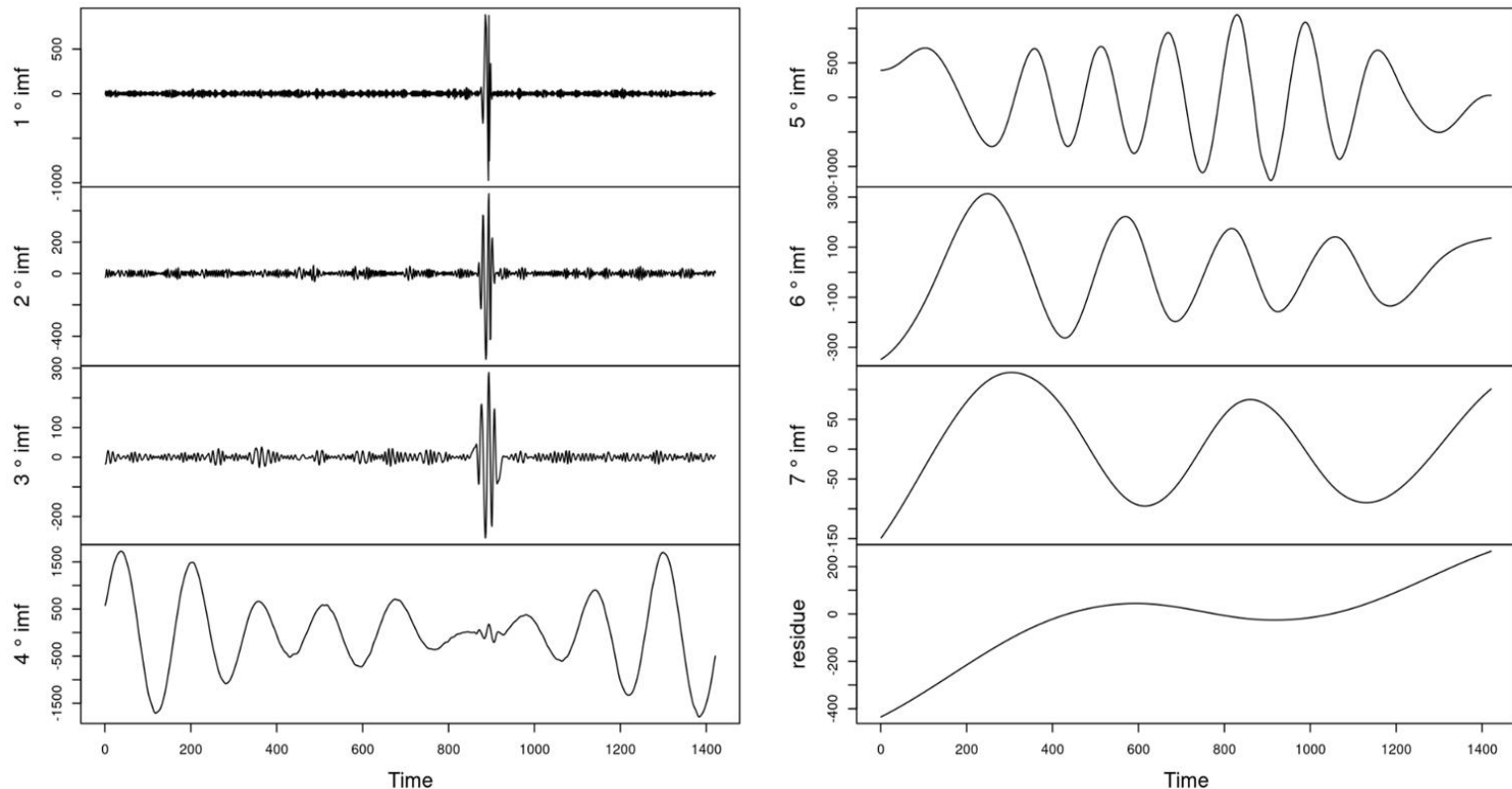
Method	Precision	Recall	F1	Accuracy
<b>FBIAD</b>	<b>0.066</b>	0.528	<b>0.085</b>	0.731
<b>ARIMA</b>	0.045	<b>0.556</b>	0.067	<b>0.746</b>
<b>LSTM</b>	0.041	0.534	0.062	0.735
<b>ELM</b>	0.041	0.517	0.063	0.726
<b>Conv1D</b>	0.036	0.519	0.055	0.724
<b>SVM</b>	0.030	0.542	0.049	0.732

# Inspecting Performance Comparison



# 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)



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

## REMD: A Novel Hybrid Anomaly Detection Method Based on EMD and ARIMA

Enrique Sánchez-Cabeza enrique.sanchez@unizar.es	Enrique Sánchez-Cabeza enrique.sanchez@unizar.es	Enrique Sánchez-Cabeza enrique.sanchez@unizar.es
Enrique Sánchez-Cabeza enrique.sanchez@unizar.es	Enrique Sánchez-Cabeza enrique.sanchez@unizar.es	Enrique Sánchez-Cabeza enrique.sanchez@unizar.es

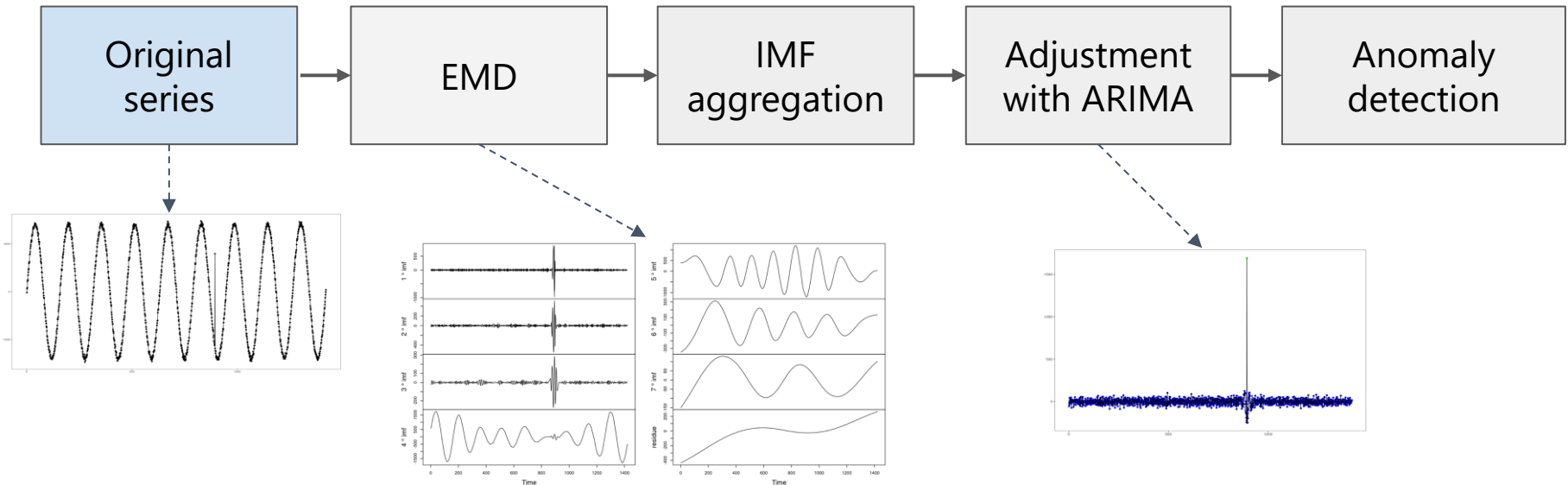
**Abstract.** Anomaly or outlier is a behavioral deviation from expected patterns and behavior. In this paper, we propose a hybrid method for anomaly detection based on EMD and ARIMA. The method consists of four steps: EMD decomposition, IMF aggregation, ARIMA adjustment, and anomaly detection. The method is evaluated using a real dataset and the results show that it is more effective than other methods.

**1. INTRODUCTION**

Time series analysis is a fundamental task in various fields, providing valuable insights into the behavior of data over time [1]. Through this analysis, it is possible to identify trends, seasonal patterns, and anomalies. Anomalies are data points that deviate significantly from the expected behavior of the data, and they can be caused by a variety of factors, such as errors, fraud, or system failures. Detecting anomalies is a challenging task, and it requires the use of sophisticated methods. In this paper, we propose a hybrid method for anomaly detection based on EMD and ARIMA. The method consists of four steps: EMD decomposition, IMF aggregation, ARIMA adjustment, and anomaly detection. The method is evaluated using a real dataset and the results show that it is more effective than other methods.

**2. RELATED WORK**

There are many methods for anomaly detection, and they can be classified into three main categories: statistical methods, machine learning methods, and deep learning methods. Statistical methods are based on the assumption that the data follows a certain distribution, and anomalies are data points that deviate from this distribution. Machine learning methods are based on the assumption that the data follows a certain pattern, and anomalies are data points that deviate from this pattern. Deep learning methods are based on the assumption that the data follows a certain complex pattern, and anomalies are data points that deviate from this pattern.

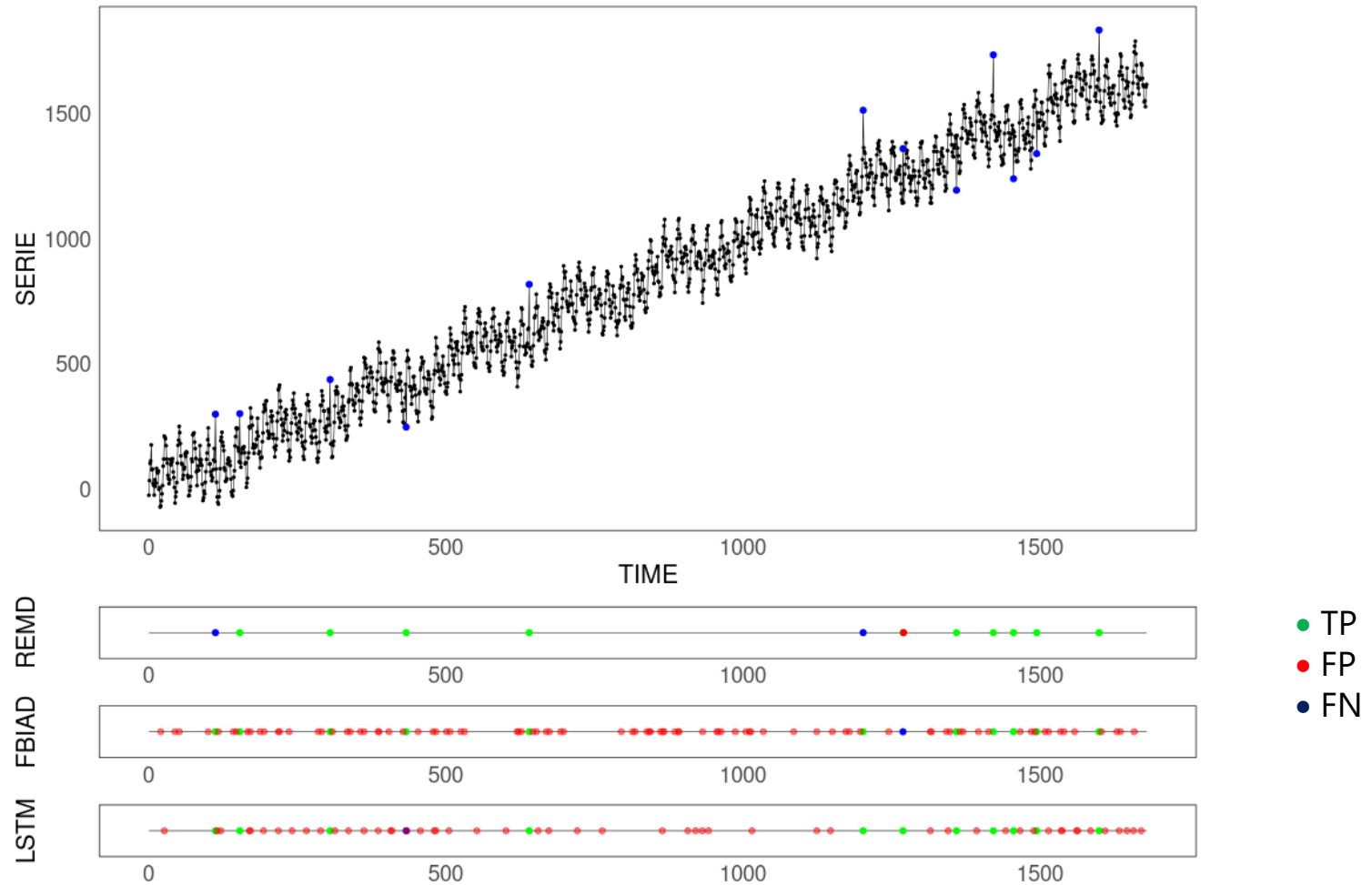


## Comparison

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

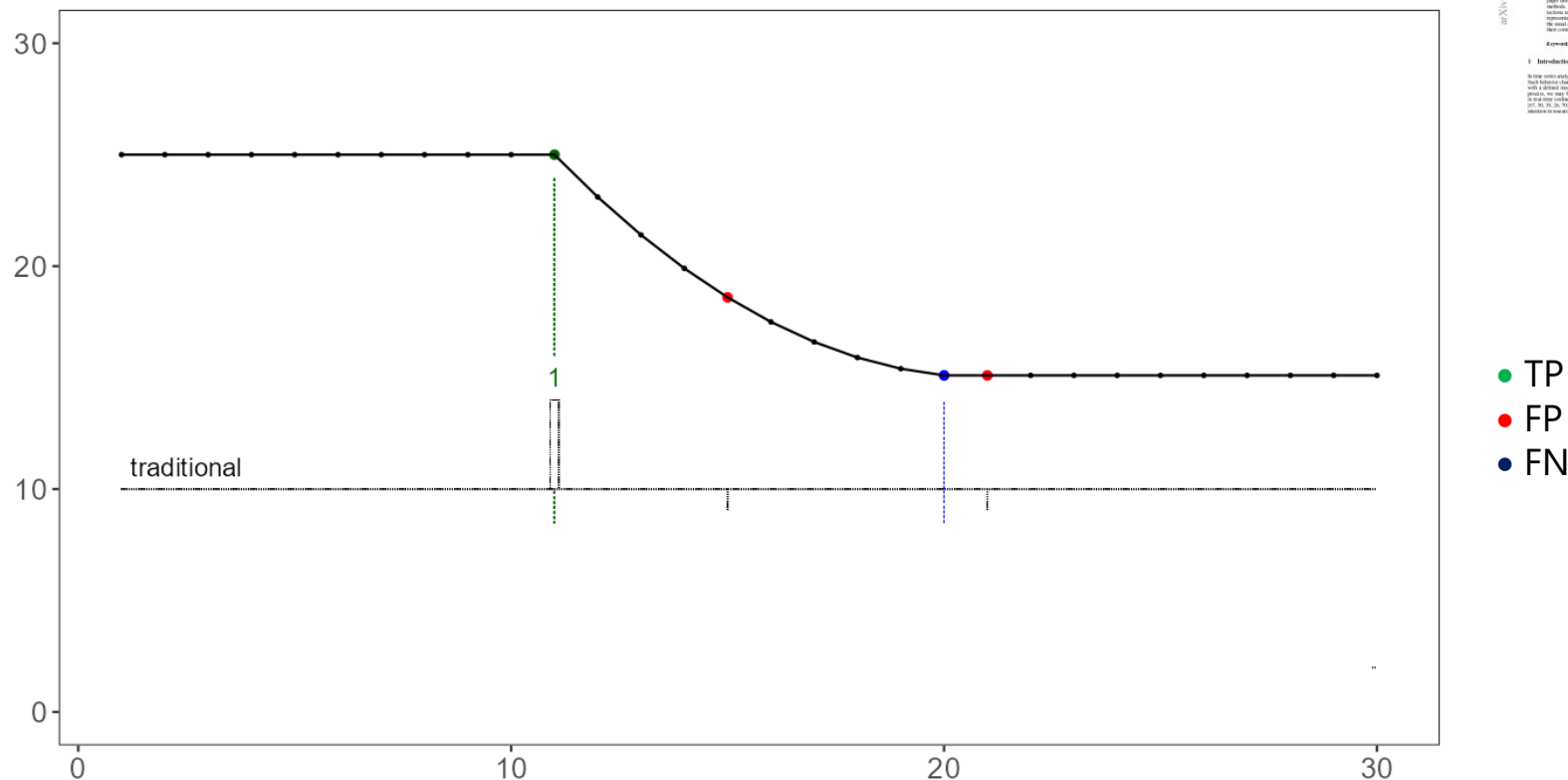
Method	Precision	Recall	F1
<b>REMD</b>	<b>0.684</b>	0.386	<b>0.448</b>
<b>EMD</b>	0.243	0.408	0.207
<b>FBIAD</b>	0.066	0.528	0.085
<b>ARIMA</b>	0.045	<b>0.556</b>	0.067
<b>LSTM</b>	0.041	0.534	0.062
<b>ELM</b>	0.041	0.517	0.063
<b>Conv1D</b>	0.036	0.519	0.055
<b>SVM</b>	0.030	0.542	0.049

# Inspecting performance comparison



## ***Time should count while evaluating events***

- Traditional scoring methods, such as precision and recall, are not sufficient to assess the performance of event detection
  - They do not incorporate time and do not reward close detections.
  - True positives are rewarded
  - All other outcomes are equally penalized



SQUARED METRICS FOR SHORT-LENGTHS IN TIME SERIES EVENT DETECTION			
<b>Reference Index</b> DOI: 10.3390/math12040439 arXiv:2004.00439v1 [stat.LG] 4 Apr 2020	<b>Article Index</b> DOI: 10.3390/math12040439 arXiv:2004.00439v1 [stat.LG] 4 Apr 2020	<b>Article's Content</b> DOI: 10.3390/math12040439 arXiv:2004.00439v1 [stat.LG] 4 Apr 2020	
<b>Article Title</b> SQUARED METRICS FOR SHORT-LENGTHS IN TIME SERIES EVENT DETECTION	<b>Article's Author</b> University of Montpellier & INRIA University of Montpellier & INRIA University of Montpellier & INRIA	<b>Article's Editor</b> University of Montpellier & INRIA University of Montpellier & INRIA University of Montpellier & INRIA	
<b>Article Type</b> Short Communication (stat.LG) (stat.LG) (stat.LG)	<b>Article's Editor</b> University of Montpellier & INRIA University of Montpellier & INRIA University of Montpellier & INRIA	<b>Article's Editor</b> University of Montpellier & INRIA University of Montpellier & INRIA University of Montpellier & INRIA	
<b>Article's Editor</b> University of Montpellier & INRIA University of Montpellier & INRIA University of Montpellier & INRIA	<b>Article's Editor</b> University of Montpellier & INRIA University of Montpellier & INRIA University of Montpellier & INRIA	<b>Article's Editor</b> University of Montpellier & INRIA University of Montpellier & INRIA University of Montpellier & INRIA	

# 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)

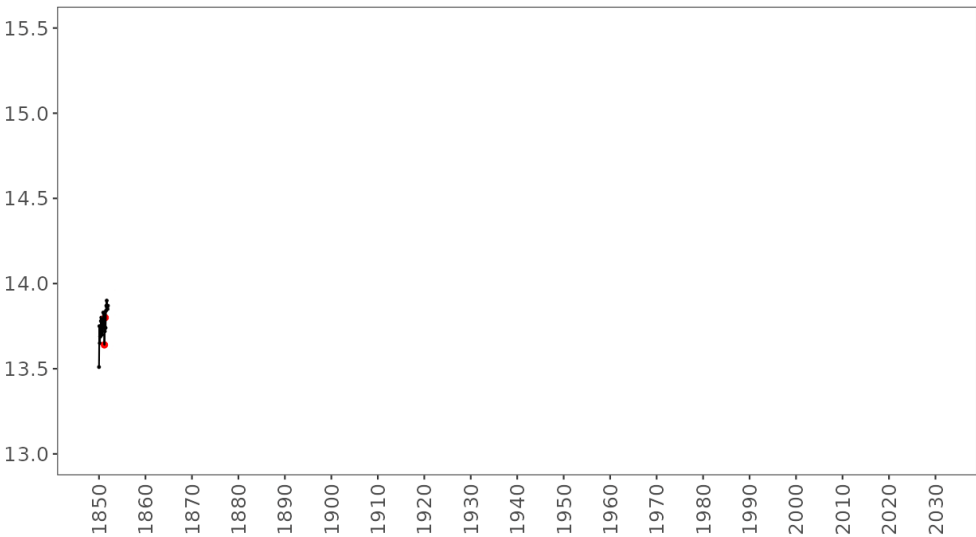
Online Event Detection in Streaming Time Series:  
Novel Metrics and Practical Insights

José Lima CITECH, A. Francisco joselima@ua.es	Luca Gualdi Santos CITECH luca.gualdi@ua.es	Edmar Pacifici INRIA & University of Strasbourg edmar.pacifici@inria.fr
Julio Eduardo Ferreira USP jferreira@usp.br	Isabel Gama Universidade Nova de Lisboa isabel.gama@fmf.unl.pt	Isabel Gama Universidade Nova de Lisboa isabel.gama@fmf.unl.pt
Diogo Carvalho CITECH, I. G. Siqueira diogo@ua.es	Rafael Porto UNIC rafael@ua.es	Edmar Gamas CITECH, I. G. Siqueira edmar.gamas@ua.es

Abstract—Online event detection in streaming time series is a critical task with applications across various domains. For instance, the right prediction of an event allows the system to take a decision and act accordingly. However, the current state-of-the-art methods are not able to handle the streaming nature of the data, leading to high detection lag and high false alarm rate. In this paper, we propose a novel metric, the Detection Probability (DP), to evaluate the performance of online detectors. The DP is defined as the ratio between the detection frequency (df) and the batch frequency (bf). The paper introduces novel metrics, the detection lag (Lag) and the detection lag standard deviation (LagSD), to evaluate the performance of online detectors. The paper also introduces a novel metric, the Detection Lag Standard Deviation (LagSD), to evaluate the performance of online detectors. The paper also introduces a novel metric, the Detection Lag Standard Deviation (LagSD), to evaluate the performance of online detectors.

1. INTRODUCTION  
Time series events are characterized by the occurrence of a significant change in the behavior of a time series at a certain point of time. In this paper, we propose a novel metric, the Detection Probability (DP), to evaluate the performance of online detectors. The DP is defined as the ratio between the detection frequency (df) and the batch frequency (bf). The paper introduces novel metrics, the detection lag (Lag) and the detection lag standard deviation (LagSD), to evaluate the performance of online detectors. The paper also introduces a novel metric, the Detection Lag Standard Deviation (LagSD), to evaluate the performance of online detectors.

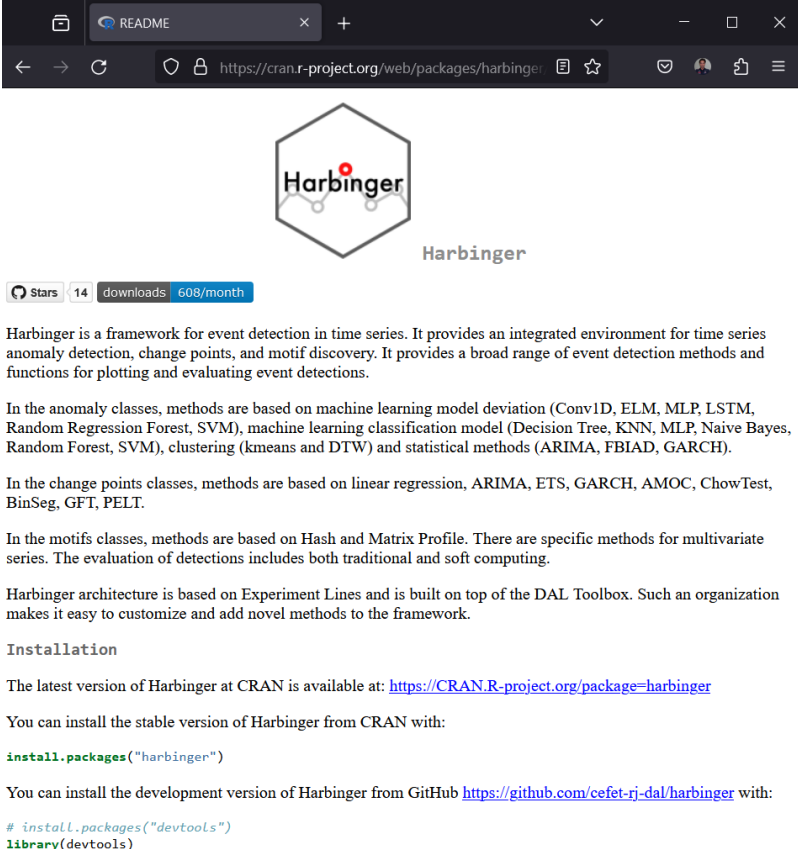
2. BACKGROUND  
A. Event Detection  
Event detection is a critical task in many domains, such as anomaly detection, intrusion detection, and fault detection. In this paper, we propose a novel metric, the Detection Probability (DP), to evaluate the performance of online detectors. The DP is defined as the ratio between the detection frequency (df) and the batch frequency (bf). The paper introduces novel metrics, the detection lag (Lag) and the detection lag standard deviation (LagSD), to evaluate the performance of online detectors. The paper also introduces a novel metric, the Detection Lag Standard Deviation (LagSD), to evaluate the performance of online detectors.





# Harbinger: Framework for Time Series Event Detection

- 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



The screenshot shows the CRAN page for the Harbinger R package. The browser address bar displays <https://cran.r-project.org/web/packages/harbinger/>. The page features the Harbinger logo, which is a hexagon with a red dot and the word "Harbinger" inside. Below the logo, the text "Harbinger" is displayed. The page also shows the number of stars (14) and downloads (608/month). The main content area contains a description of the framework, its capabilities, and installation instructions.

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: <https://CRAN.R-project.org/package=harbinger>

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

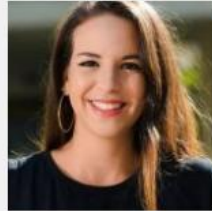
## D.Sc. students



Ellen Paixão



Janio Lima



Lais Baroni



Lucas Giusti\*



Paulo Elias\*  
(UFF)

## M.Sc. students



Antônio Mello



Arthur Garcia



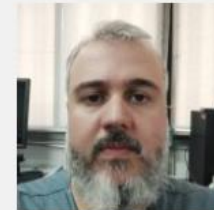
Cristiane Gea



Edson Sobrinho



Fabiana Santos\*



Frank Faisca



Jéssica de Souza



Josélia Rabelo



Luiz Oliveira



Michel Reis\*



Ricardo Buçard\*



Rodrigo Machado



Thiago Marques

# Biografia

- Doutor em Engenharia de Sistemas e Computação (COPPE/UFRJ) em 2011
- Professor no EIC - CEFET/RJ
  - Departamento de Ciência da Computação
  - Curso Técnico de Informática
- Programa de Pós-Graduação em Ciência da Computação (PPCIC)
- Programa de Pós-Graduação em Engenharia de Produção e Sistemas (PPPRO)
- Membro do Sênior da IEEE
- Membro da SBC e ACM
- Editor Associado da IEEE Latin America Transactions

<https://eic.cefet-rj.br/~eogasawara>





# DATA ANALYTICS LAB

<https://eic.cefet-rj.br/~dal>

