PCDaS
Ciência de Dados aplicada à Saúde

ICICT Instituto de Comunicação e Informação Científica e Tecnológica em Saúde

CEFET/RJ

DAL

OUTBREAKS DETECTION

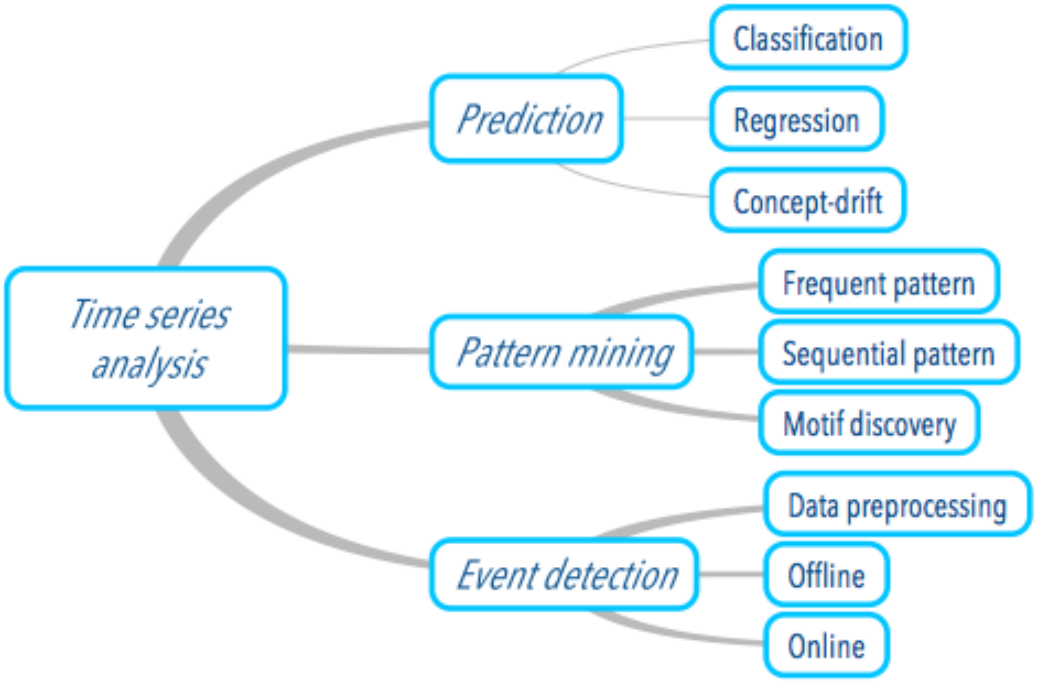
Eduardo Ogasawara
eogasawara@ieee.org
https://eic.cefet-rj.br/~eogasawara

Biograph

- D.Sc. in Systems and Computer Engineering (COPPE/UFRJ) in 2011
- Professor at EIC - CEFET/RJ
  - Computer Science Department
  - Computer Science Technical High School
- Graduate Program in Computer Science (PPCIC)
- Graduate Program in Eng. Production and Systems (PPPRO)
- Member of IEEE, SBC, ACM, and INNS
- Associated Editor of IEEE Latin America Transactions

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

Research Themes



Let’s start

Disease Outbreaks

- A disease outbreak is the occurrence of disease cases in excess of normal expectancy
  - The number of cases varies according to the disease-causing agent, and the size and type of previous and existing exposure to the agent

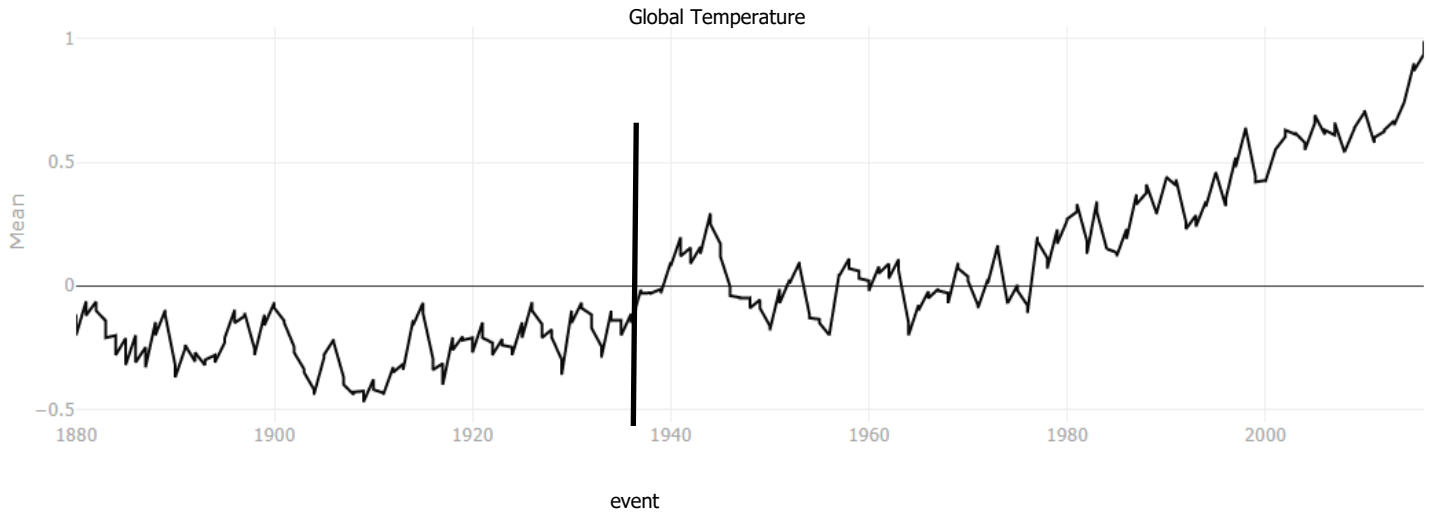
- Disease outbreaks are usually caused by an infection, transmitted through person-to-person contact, animal-to-person contact, or from the environment or other media
  - Outbreaks may also occur following exposure to chemicals or to radioactive materials

Source: WHO,2022

https://www.who.int/teams/environment-climate-change-and-health/emergencies/disease-outbreaks

Events

- A point or an interval where a significant change in the time series behavior occurs
- Events may appear as anomalies, change points, or frequent patterns (motifs)



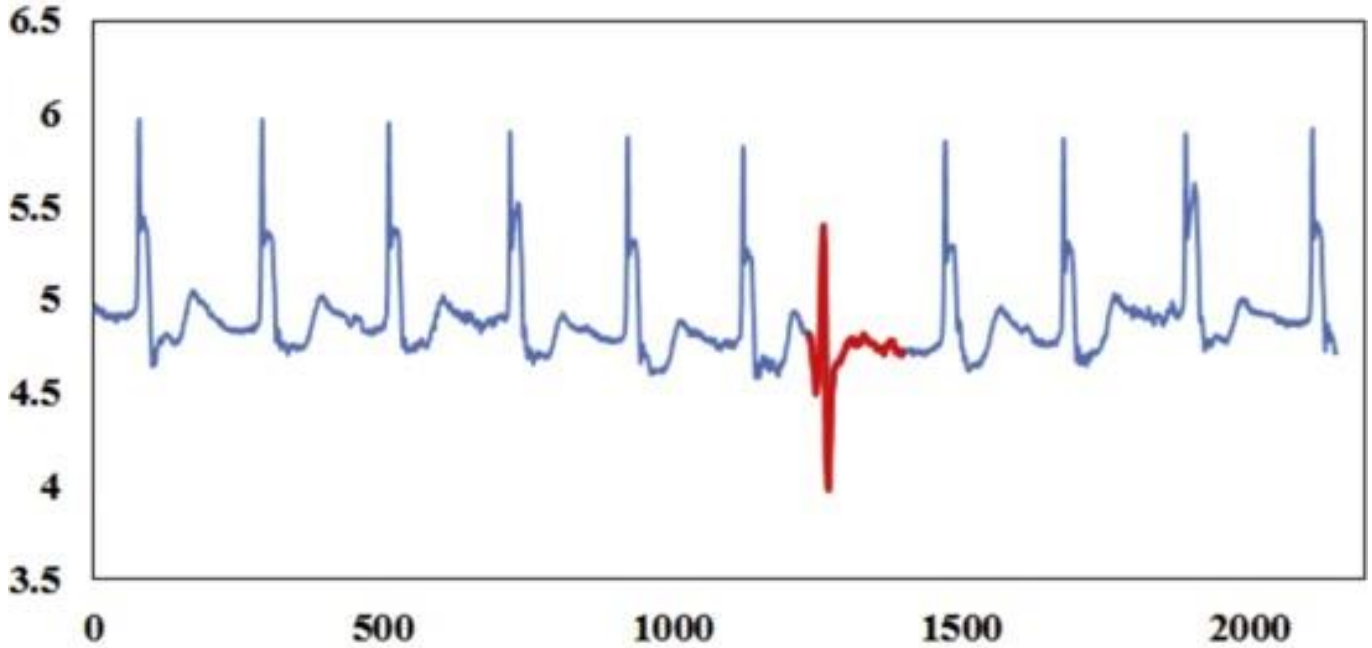
[1] V. Guralnik and J. Srivastava, 1999, Event Detection from Time Series Data, In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 33–42

Image source: World Global Temperature, https://datahub.io/core/global-temp

Anomalies

- A pattern or observation that do not conform to expected behavior [1]
- It can be categorized as punctual, contextual or collective

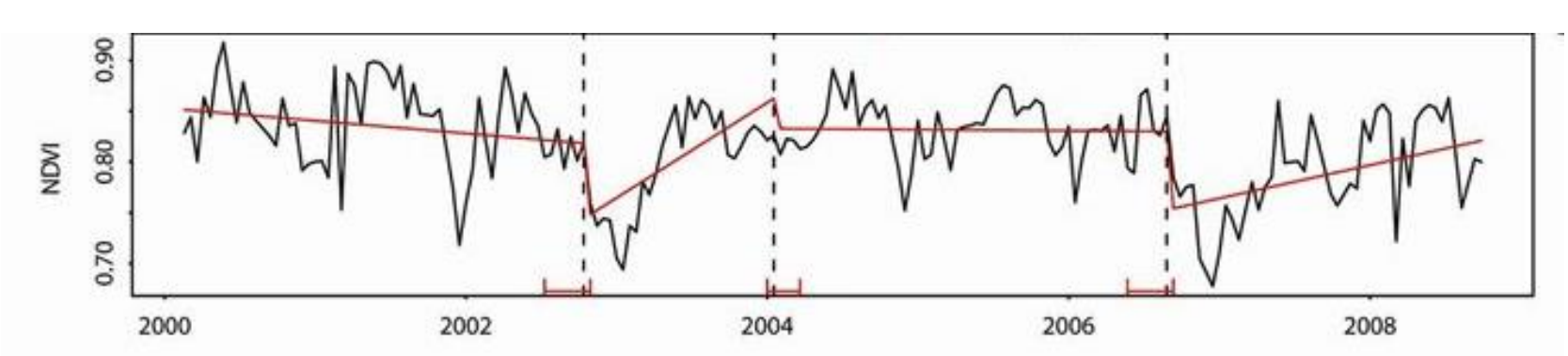
(*)



[1]] V. Chandola, A. Banerjee, e V. Kumar, 2009, Anomaly detection: A survey, ACM Computing Surveys, v. 41, n. 3
(*) In this example, it can also be classified as a discord

Change Points

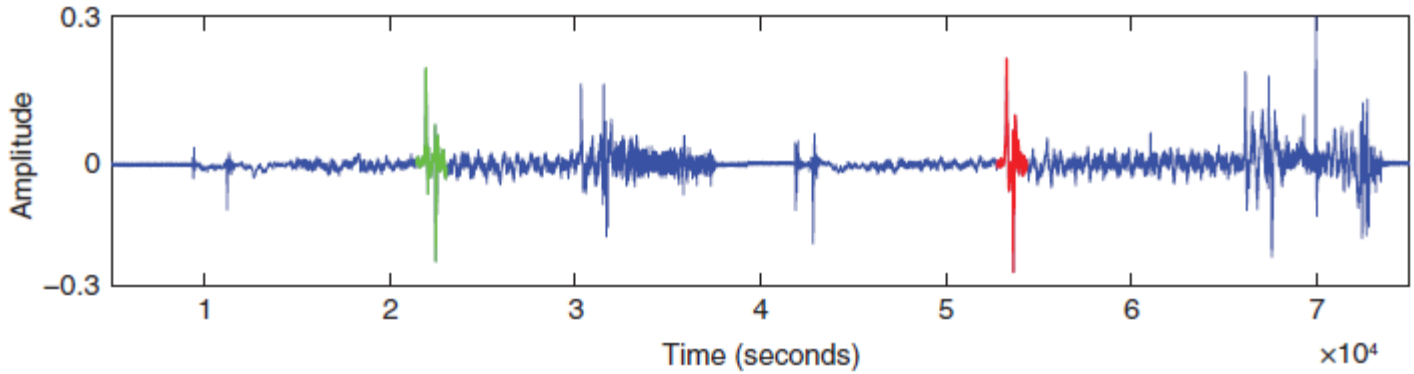
- Points (or time intervals) that mark significant change in time series behavior [1]
- They separate different states in the process that generates the time series



[1] J.-I. Takeuchi e K. Yamanishi, 2006, A unifying framework for detecting outliers and change points from time series, IEEE Transactions on Knowledge and Data Engineering, v. 18, n. 4, p. 482–492.
Image source: https://www.wur.nl/en/Research-Results/Chair-groups/Environmental-Sciences/Laboratory-of-Geo-information-Science-and-Remote-Sensing/Research/Integrated-land-monitoring/Change_detection_and_monitoring.htm

Motifs

- A pattern (unknown) that occurs a significant number of times in time series [1,2,3]



How to do it in non-stationarity time series?

[1] P. Patel, E. Keogh, J. Lin, and S. Lonardi, "Mining motifs in massive time series databases," in Proceedings - IEEE International Conference on Data Mining, ICDM, 2002, pp. 370–377
[2] A. Mueen, "Time series motif discovery: Dimensions and applications," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 4, no. 2, pp. 152–159, 2014
[3] S. Torkamani and V. Lohweg, "Survey on time series motif discovery," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 7, no. 2, 2017.

Event detection

- An event can represent a phenomenon with a specific meaning defined in a certain domain
- Event detection is the process of finding events
- It is a basic function in surveillance and monitoring systems
- Example of applications:

Electronic IT security

Healthcare

Medical diagnostics and monitoring

Industrial monitoring & Damage detection

Sensor Networks

[1] V. Chandola, A. Banerjee, e V. Kumar, 2009, Anomaly detection: A survey, ACM Computing Surveys, v. 41, n. 3
[2] M. Gupta, J. Gao, C.C. Aggarwal, e J. Han, 2014, Outlier Detection for Temporal Data: A Survey, IEEE Transactions on Knowledge and Data Engineering, v. 26, n. 9, p. 2250–2267.
[3] H. Wang, M.J. Bah, e M. Hammad, 2019, Progress in Outlier Detection Techniques: A Survey, IEEE Access, v. 7, p. 107964–108000.

Importance of event detection

[1] Marco A.F. Pimentel, David A. Clifton, Lei Clifton, Lionel Tarassenko, A review of novelty detection, Signal Processing, Volume 99, 2014, Pages 215-249, ISSN 0165-1684, https://doi.org/10.1016/j.sigpro.2013.12.026.

Event detection initiatives

Anomaly detection	Finding unexpected behavior (deviations)
Change point detection	Finding change points It is related to finding drifts in time series
Motif detection	Identifying frequent patterns in time series

[1] Marco A.F. Pimentel, David A. Clifton, Lei Clifton, Lionel Tarassenko, A review of novelty detection, Signal Processing, Volume 99, 2014, Pages 215-249, ISSN 0165-1684, https://doi.org/10.1016/j.sigpro.2013.12.026.

Anomaly detection (distribution analysis)

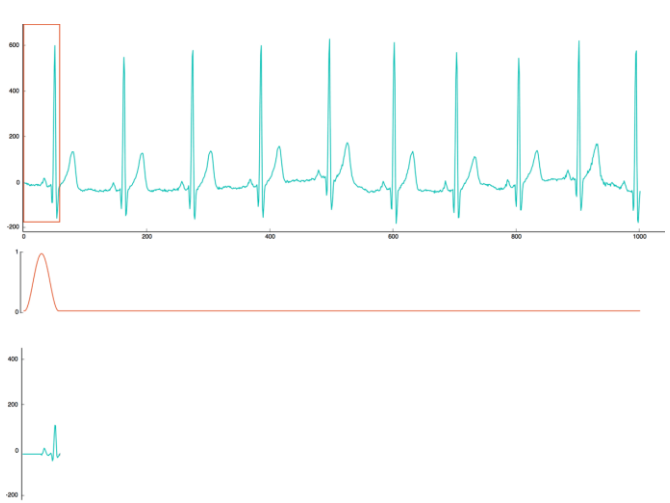
- Statistical analysis
  - Differentiation (backshift operator)
  - Residuals from moving average
  - Residuals from filters (Kalman)
- Model adjustment
  - Residuals from decomposed signal
  - Residuals from linear models (regression)
  - Residuals from autoregressive models (ARIMA)
  - Residuals from volatility models (GARCH)
  - Residuals from machine learning models
- Clustering of subsequences
  - Distribution analysis over difference between subsequences and centroids
  - DBScan
- Time series decomposition
  - Trend
  - Seasonal
  - Fourier transform
  - Wavelets
  - IMF - intrinsic mode function
  - Hilbert-Huang transform

[1] V. Chandola, A. Banerjee, e V. Kumar, 2009, Anomaly detection: A survey, *ACM Computing Surveys*, v. 41, n. 3

Change point detection

- Seminal change point [1]
- Change Finder [2]

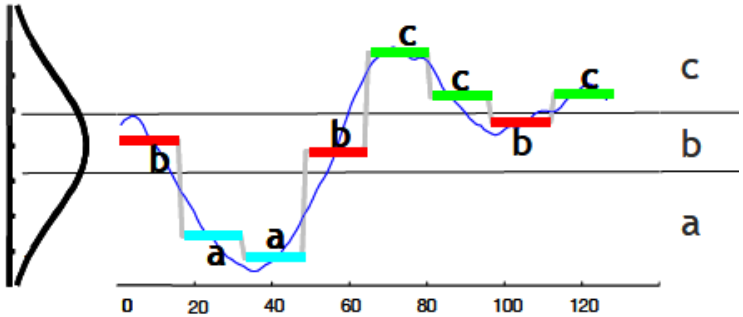
Windowed approach

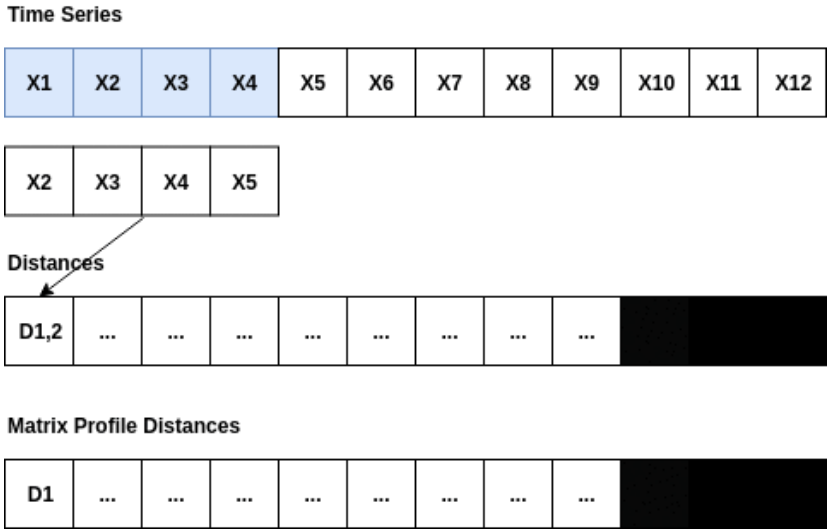


[1] J.-I. Takeuchi and K. Yamanishi, 2006, A unifying framework for detecting outliers and change points from time series, IEEE Transactions on Knowledge and Data Engineering, v. 18, n. 4, p. 482–492.
[2] V. Guralnik and J. Srivastava, 1999, Event Detection from Time Series Data, In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 33–42
Image source: https://towardsdatascience.com/real-time-time-series-anomaly-detection-981cf1e1ca13

Motif discovery

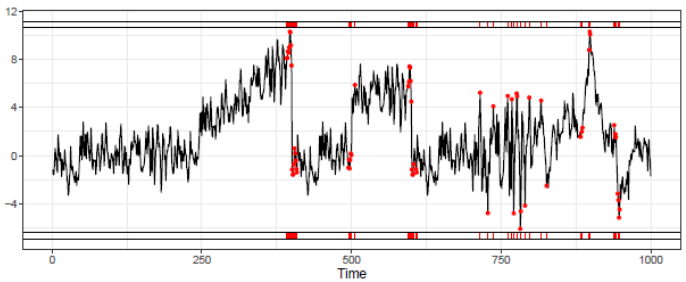
- Indexing
  - Discretization
  - SAX [1]
- Brute force
- Hash-based (random projection) [2]
- Matrix profile [3]



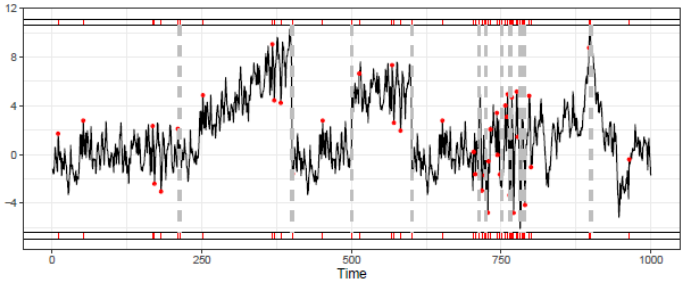


[1] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: a novel symbolic representation of time series," Data Mining and Knowledge Discovery, vol. 15, no. 2, pp. 107–144, 2007
[2] A. Mueen, "Time series motif discovery: Dimensions and applications," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 4, no. 2, pp. 152–159, 2014
[3] M. Linardi, Y. Zhu, T. Palpanas, and E. Keogh, 2020, Matrix profile goes MAD: variable-length motif and discord discovery in data series, *Data Mining and Knowledge Discovery*, v. 34, n. 4, p. 1022–1071.
(*) https://towardsdatascience.com/introduction-to-matrix-profiles-5568f3375d90

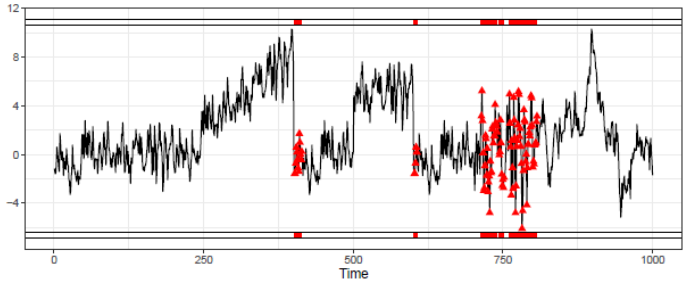
The many faces of event detection



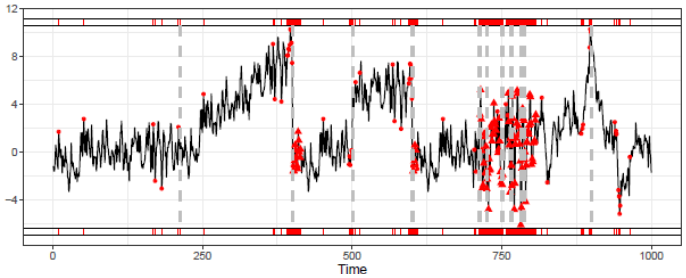
Method A: trend anomalies



Method B: trend anomalies & change points



Method C: volatility anomalies



Methods A,B & C:
trend anomalies, volatility anomalies and change points

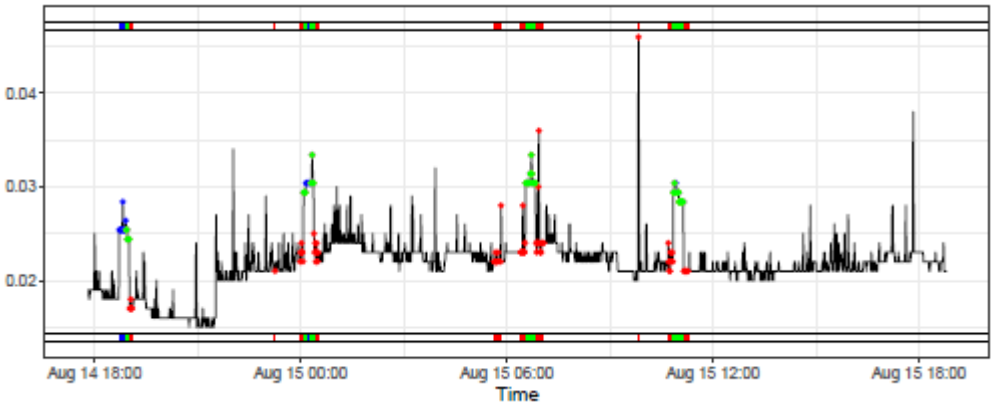
[1] R. Salles, L. Escobar, L. Baroni, R. Zorrilla, A. Ziviani, V. Kreischer, F. Delicato, P. Pires, L. Maia, et al., 2020, Um framework para integração e análise de métodos de detecção de eventos em séries temporais, In: *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*

Metrics for event detection

- Classifier Accuracy: percentage of test set tuples that are correctly classified
  - $accurary = \frac{TP+TN}{All}$
  - $precision = \frac{TP}{TP+FP}$
  - $recall = \frac{TP}{TP+FN}$
  - $F_1 = \frac{2 \cdot precision \cdot recall}{precision+recall}$
  - *ROC Curve*

Confusion Matrix (CM)

Predicted Actual	$\hat{E}$	$\neg\hat{E}$
E	TP	FN
$\neg$E	FP	TN



Online event detection

- Handles streaming time series

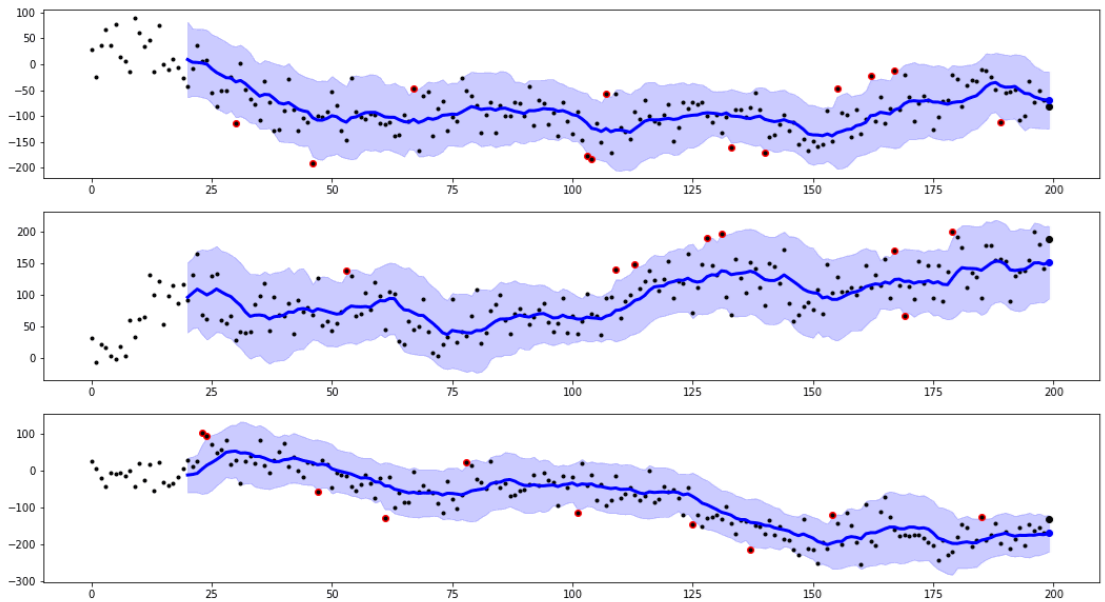
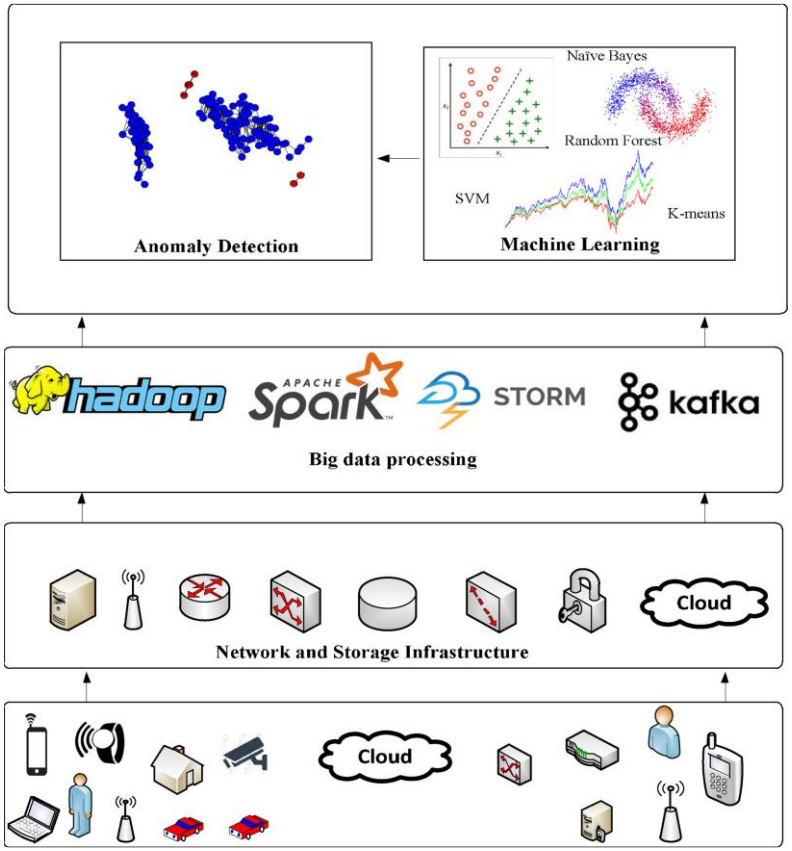


Image source: https://towardsdatascience.com/real-time-time-series-anomaly-detection-981cf1e1ca13

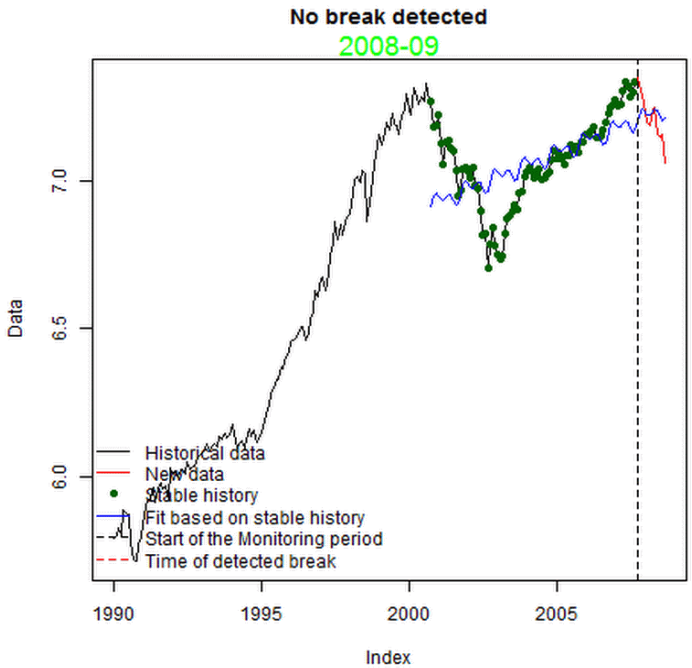
Online event detection infrastructure

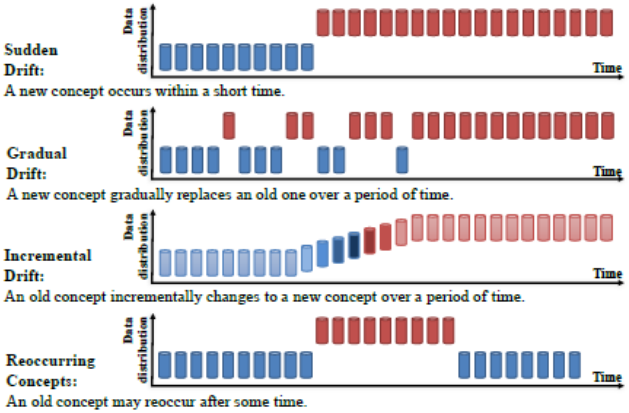


[1] R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, M. Imran, Real-time big data processing for anomaly detection: A Survey, International Journal of Information Management, Volume 45, 2019, Pages 289-307, ISSN 0268-4012

Online change-point detection

- Detection occurs incrementally





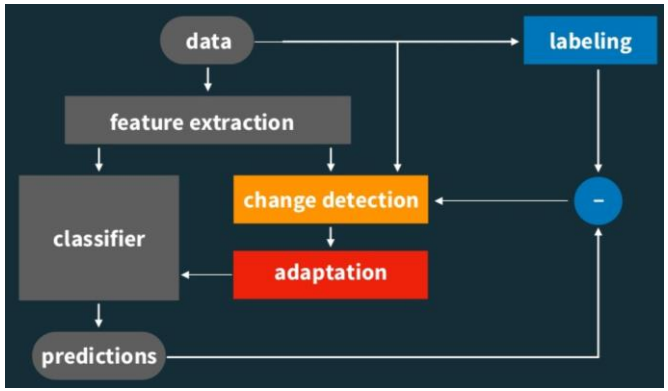


Image source: https://towardsdatascience.com/real-time-time-series-anomaly-detection-981cf1e1ca13

[1] G. Ditzler, M. Roveri, C. Alippi, e R. Polikar, 2015, Learning in Nonstationary Environments: A Survey, *IEEE Computational Intelligence Magazine*, v. 10, n. 4, p. 12–25.
[2] https://youtu.be/woRmeGvOaz4

Online event detection challenges (when to adapt)

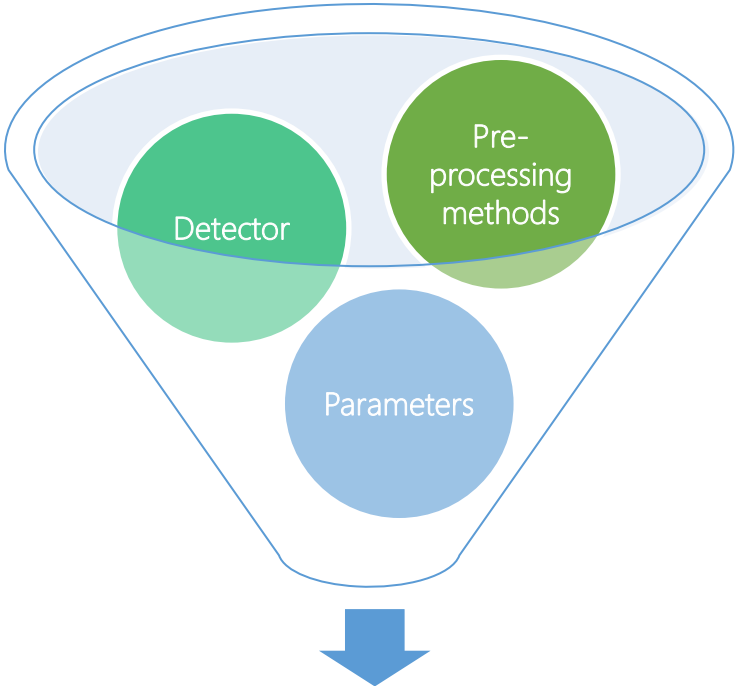
Adaptability:

- Incremental learning
- Plasticity and Stability Dilemma

[1] S.O. Haykin, 2008, *Neural Networks and Learning Machines*. 3 ed. New York, Prentice Hall.
[2] Grossberg, S., 1988. Neural Networks and Natural Intelligence, Cambridge, MA: MIT Press.
[3] G. Ditzler, M. Roveri, C. Alippi, e R. Polikar, 2015, Learning in Nonstationary Environments: A Survey, *IEEE Computational Intelligence Magazine*, v. 10, n. 4, p. 12–25.

Online event detection challenges (too many methods)

Myriad of event detection methods (detectors)



Detection performance

Choice of appropriate detectors/parameters for event detection is a challenge

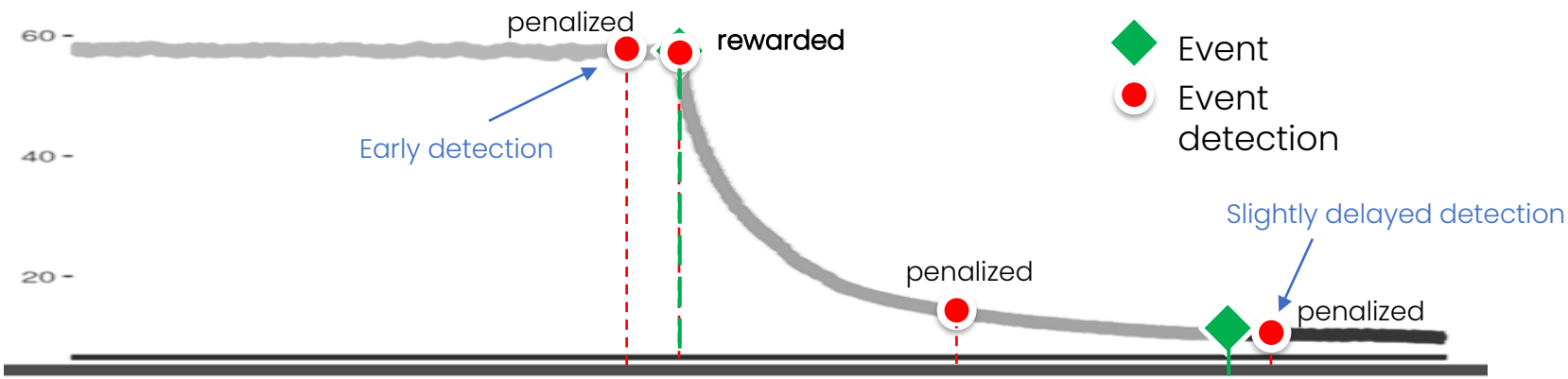
Directly related to initial assumptions about the behavior and statistical properties of data

The nature of the events observed is often unknown

Detectors specialized in a type of event may disregard the occurrence of another, or even misidentify them

Online event detection challenges (metrics)

- Traditional scoring methods, such as precision and recall, don't suffice for evaluating online event detection performance.
  - They do not incorporate time and do not reward early detection.
  - True positives are rewarded. All other results are "harshly" and equally punished.



[1] Lavin, A., & Ahmad, S. (2015, December). Evaluating Real-Time Anomaly Detection Algorithms--The Numenta Anomaly Benchmark. In 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA) (pp. 38-44). IEEE.
[2] Singh, N., & Olinsky, C. (2017, May). Demystifying Numenta anomaly benchmark. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 1570-1577). IEEE.

What type of disease outbreaks?

- Doenças Relacionadas ao Saneamento Ambiental Inadequado (DRSAI)
- Internações Hospitalares por Doenças Imunopreveníveis: CID-10 por Doenças Imunopreveníveis
- Taxa de Internação por infecção respiratória aguda de menores de cinco anos de idade (IRA5)

Data Analytics Lab Team

Doutorado

Lais Baroni (CEFET/RJ)

Leonardo Carvalho (CEFET/RJ)

Rebecca Salles (CEFET/RJ)

Mestrado

Antônio Mello (CEFET/RJ)

Arthur Garcia (CEFET/RJ)

Cristiane Gea (CEFET/RJ)

Diego Salles (CEFET/RJ)

Flávia Rocha (CEFET/RJ)

Janio Lima (CEFET/RJ)

Jéssica de Souza (CEFET/RJ)

https://eic.cefet-rj.br/~dal/equipe/

Other researches

- An Analysis of Malaria in the Brazilian Legal Amazon Using Divergent Association Rules
- Estimation of COVID-19 Under-Reporting in the Brazilian States Through SARI
- Neonatal mortality rates in Brazilian municipalities: from 1996 to 2017

New Generation Computing
https://doi.org/10.1007/s00354-021-00125-3

Estimation of COVID-19 Under-Reporting in the Brazilian States Through SARI

Balthazar Paixão[1] · Lais Baroni[1] · Marcel Pedroso[2] · Rebecca Salles[1] · Luciana Escobar[1] · Carlos de Sousa[2] · Raphael de Freitas Saldanha[2] · Jorge Soares[1] · Rafaelli Coutinho[1] · Fabio Porto[3] · Eduardo Ogasawara[1]

Received: 19 December 2020 / Accepted: 4 March 2021
© Ohmsha, Ltd. and Springer Japan KK, part of Springer Nature 2021

Abstract Due to its impact, COVID-19 has been stressing the academy to search for curing, mitigating, or controlling it. It is believed that under-reporting is a relevant factor in determining the actual mortality rate and, if not considered, can cause significant misinformation. Therefore, this work aims to estimate the under-reporting of cases and deaths of COVID-19 in Brazilian states using data from the InfoGripe. InfoGripe targets notifications of Severe Acute Respiratory Infection (SARI). The methodology is based on the combination of data analytics (event detection methods) and time series modeling (inertia and novelty concepts) over hospitalized SARI cases. The estimate of real cases of the disease, called novelty, is calculated by comparing the difference in SARI cases in 2020 (after COVID-19) with the total expected cases in recent years (2016–2019). The expected cases are derived from a seasonal exponential moving average. The results show that under-reporting rates vary significantly between states and that there are no general patterns for states in the same region in Brazil. The states of Minas Gerais and Mato Grosso have the highest rates of under-reporting of cases. The rate of under-reporting of deaths is high in the Rio Grande do Sul and the Minas Gerais. This work can be highlighted for the combination of data analytics and time series modeling. Our calculation of under-reporting rates based on SARI is conservative and better characterized by deaths than for cases.

Keywords COVID-19 · Under-reporting · SARI · Time series modeling · Data analytics

✉ Lais Baroni
lais.baroni@fiocr.cefet-rj.br

[1] Federal Center for Technological Education of Rio de Janeiro, CEFET/RJ, Rio de Janeiro, Brazil
[2] Oswaldo Cruz Foundation, Fiocruz, Rio de Janeiro, Brazil
[3] National Laboratory of Scientific Computing, LNCC, Rio de Janeiro, Brazil

Published online: 14 March 2021

Springer

Novidades

Inscreva-se em: https://eic.cefet-rj.br/~eogasawara/youtube