Eötvös Loránd University (ELTE)
Faculty of Informatics (IK)
Pázmány Péter sétány 1/c
1117 Budapest, Hungary

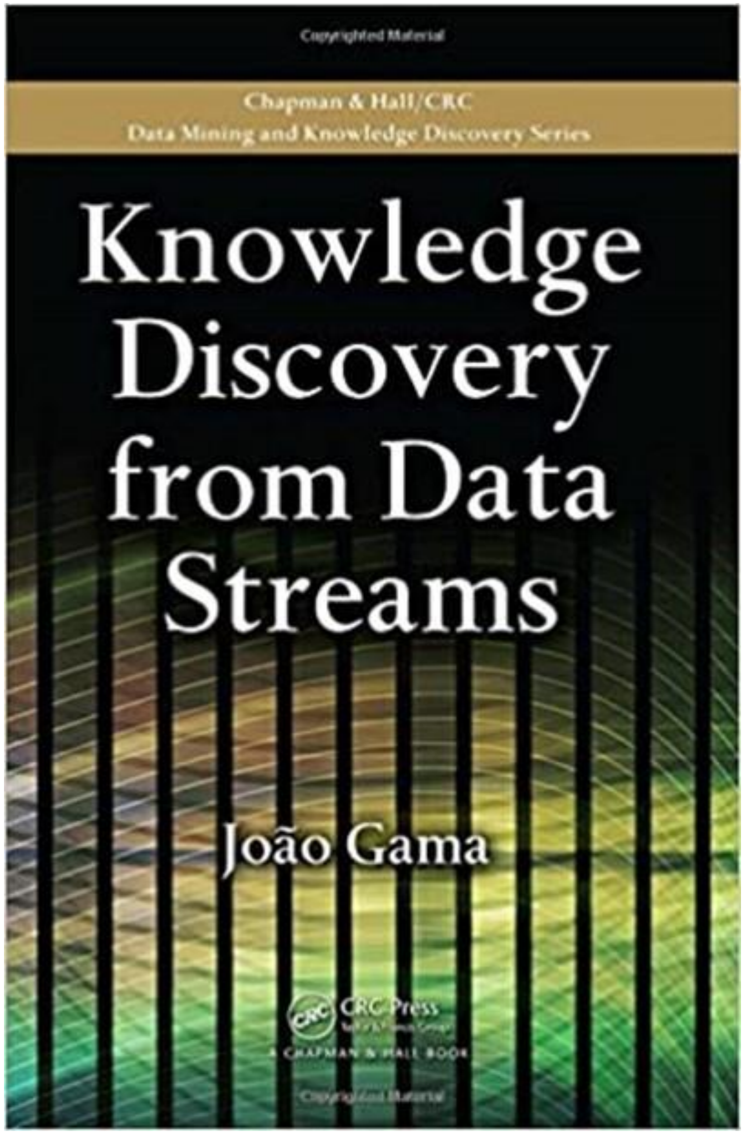
TIME SERIES ANALYSIS FOR DATA STREAMS

Stream mining (SM)

Imre Lendák, PhD, Associate Professor
Szegedi Gábor, PhD Student

2020
Budapest, Hungary

Overview & lecture topics



- Introduction
- Time series categories
- Trends & seasonality
- Time series similarity
- Clustering
- Classification
- Anomaly detection
- ~~Forecasting~~

- **Note:** we use a diverse range of other sources besides the KDDS book

Definitions

- **DEF: Time series** are sequences of measurements that follow non-random orders.
- Time series X notation:

$$x_1, x_2, \dots, x_{t-1}, x_t, \dots$$

- **DEF: Time series analysis** applies different data analysis techniques to model dependencies in the sequence of measurements
- Common components of time series analysis
  - **Trend** = represents a general systematic linear or (most often) nonlinear component that changes over time
  - **Seasonality** = represents re-occurring patterns appearing in systematic intervals over time
  - **Cycle** = the data exhibit rises and falls that are not of a fixed frequency.
  - **Noise** = a non-systematic component that is nor trend or seasonality within the data

Common use cases

- **Classification**, e.g. disease identification based on a one-off ECG diagram
- **Clustering**, e.g. identify novel patterns in large sets of medical measurements or financial data
- **Anomaly detection**, e.g. anomalous reading in an ECG data stream of a hospitalized patient → urgent reaction by the hospital staff
- **Forecasting**, e.g. predict future FOREX pair values based on historical data

- At least for classification, clustering and anomaly detection it is necessary to be able to (as) exactly (as possible) measure the **similarity** between time series

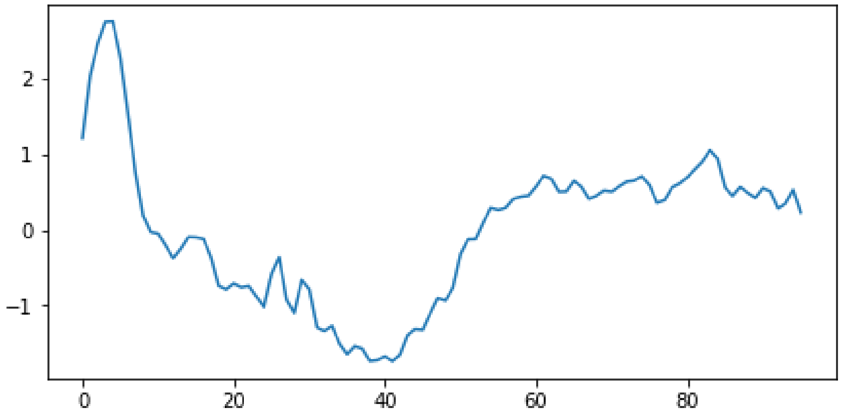
TIME SERIES CATEGORIES

Sample dimensionality categories

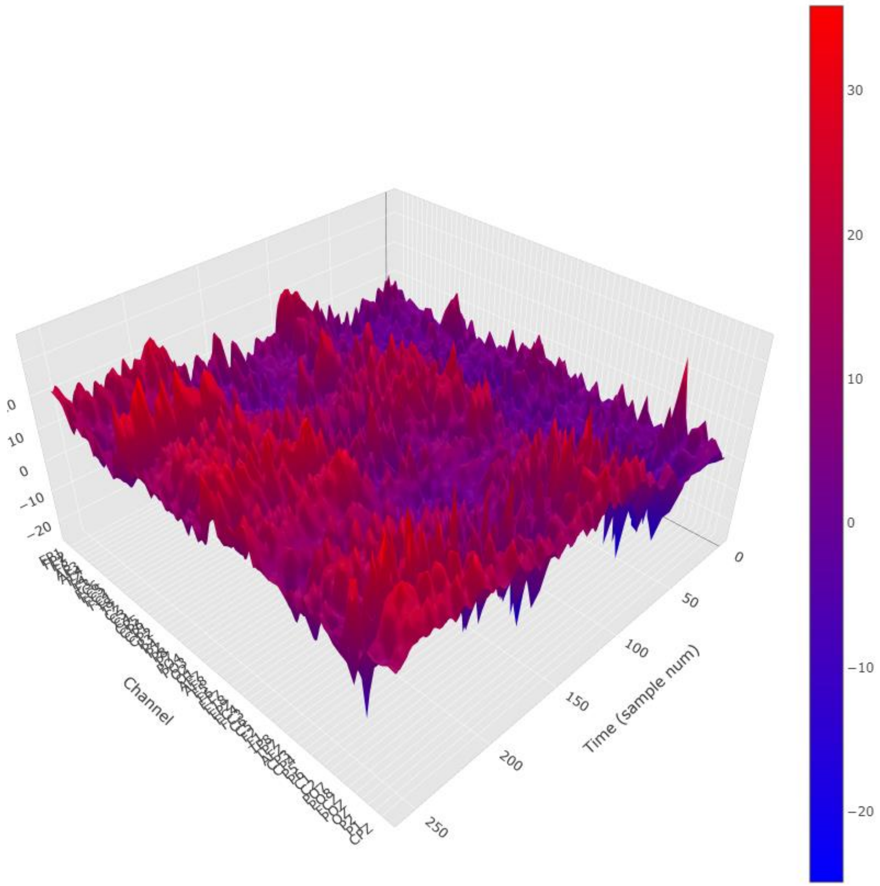
- We can categorize time series data based on what is the dimensionality of a single sample
- **Univariate:** A sample in the sequence is a single value.
  - Example: Daily changes in the average temperature.
- **Multivariate:** A sample in the sequence is a vector.
  - Example: Daily closing values of all the stocks on the NYSE stock market.
- **Complex:** A sample in the sequence is of higher dimension.
  - Example: A video feed.

Univariate time series

- A single variable as a function of time
  - E.g. a single load measurement in electric power systems, a flow meter in a water management system, stock price
- $TS = (x_1, \dots, x_n), x_i \in R$

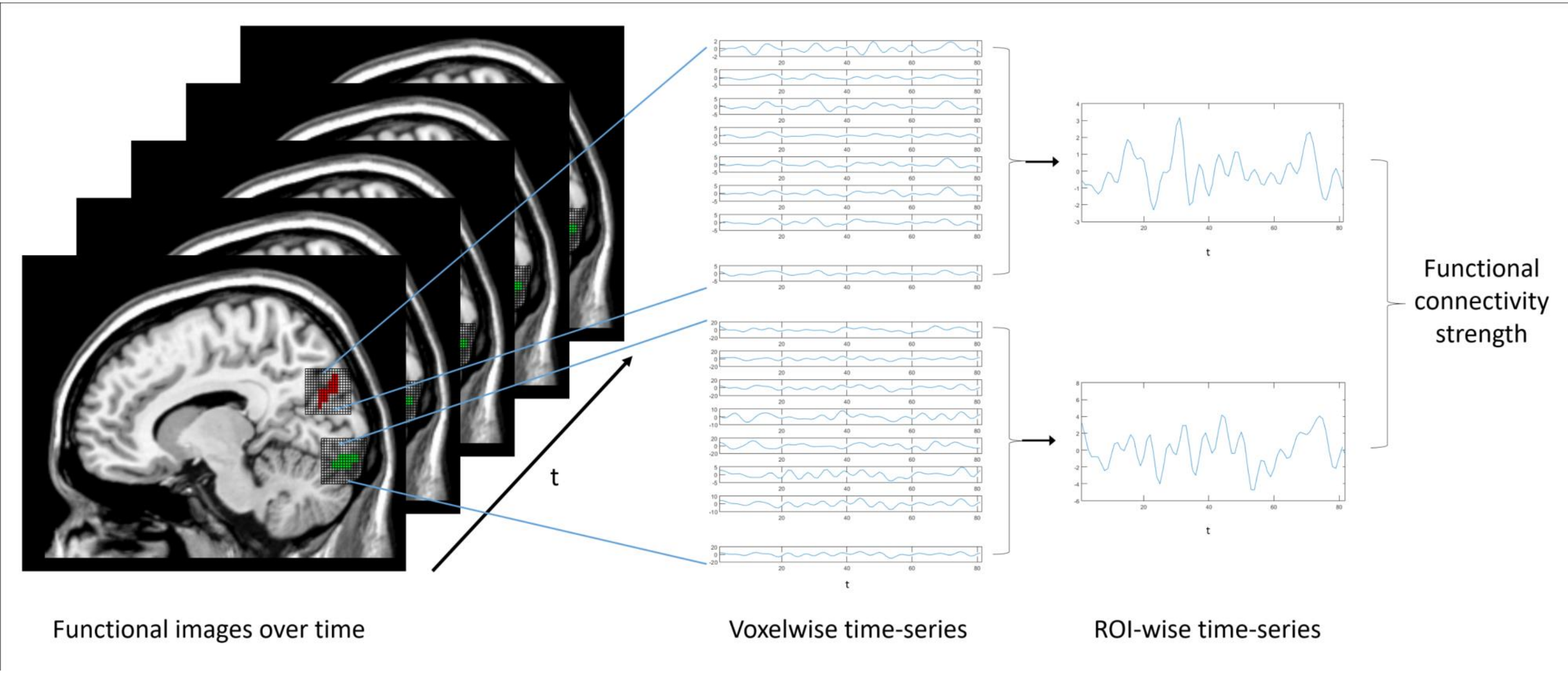


Multivariate time series



- Sequence of vectors
  - E.g. measurements describing weather conditions, ECG, EEG

Complex time series

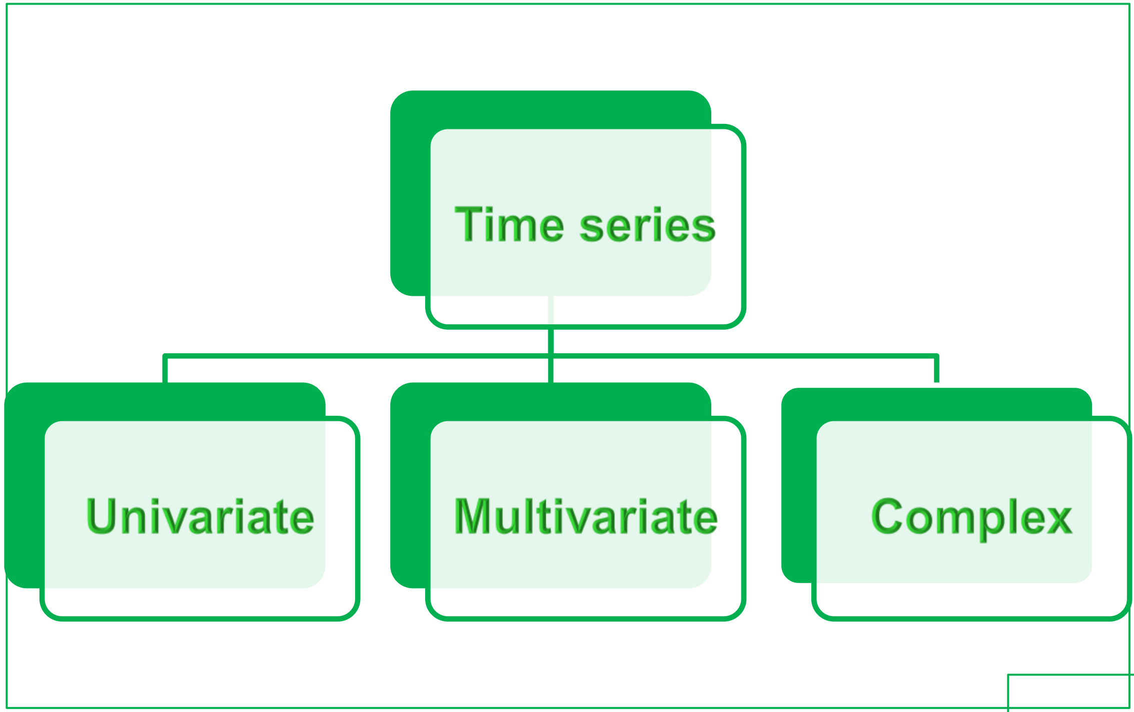


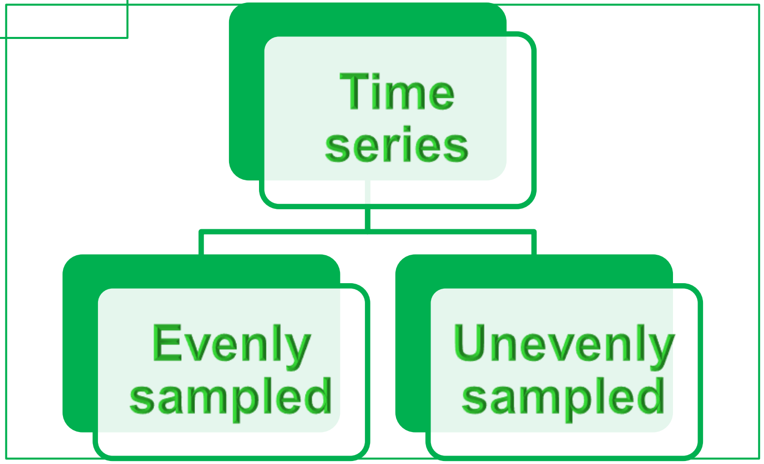
- E.g. functional magnetic resonance imaging (fMRI) data
- May be transformed to simpler time series for analysis

Sample frequency classification

- We can categorize time series data based on what is the frequency of samples (the time between 2 samples)
- **Evenly sampled:** Samples are distributed evenly during the time span of the series.
  - E.g. load measurement(s) in electric power systems sampled every 15 minutes for trading and forecasting
- **Unevenly sampled:** The frequency of samples is varying.
  - Each record is associated with a timestamp, but the frequency of samples is varying.
  - $TS = (t_1 : x_1, \dots, t_n : x_n)$
  - Note: observations x_i can be of any datatype
  - E.g. blood pressure of a patient (self-)measured twice a day, but at different times

Categorization summary





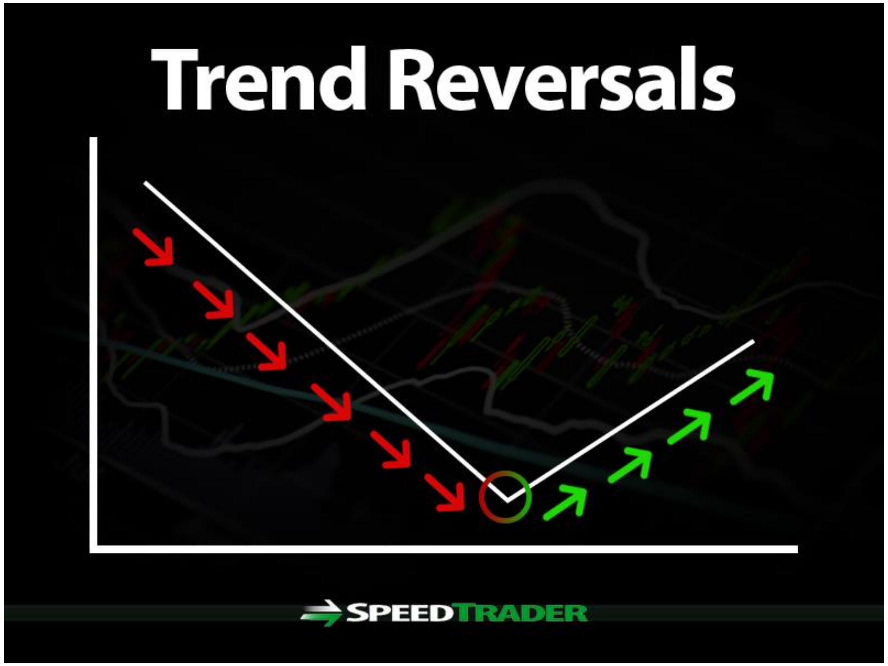
Trends

Trend primer

https://www.babypips.com/learn/forex/using-moving-averages

Trend intro

- **DEF:** A **trend** is a general systematic linear or (most often) nonlinear component that changes over time
- Trend-related challenges:
  - What is the mean of a time series with a trend? Or multiple trends?
  - What are the distributions of values?
  - Moving averages lag behind trends
- **DEF:** A **trend reversal** occurs when the direction of an existing trend changes (to the opposite)
  - Trend reversals are quite important in financial data analytics



speedtrader.com › methods-for-determining-trend-reversals

Trend analysis – 2

- **Moving averages** are used in trend detection → smooth out short-term fluctuations in the data → highlight longer-term trends or cycles

- **Averaging methods** → all items have the same relevance
- **Weighted averaging methods** → data points are associated with weights which depict their relevance

Moving averages

- **Moving average (MA)** = the mean of the previous n data points:

$$MA_t = MA_{t-1} - \frac{x_{t-n+1}}{n} + \frac{x_{t+1}}{n}$$

- **Cumulative moving average (CA)** = the average of all of the data up until the current data point

$$CA_t = CA_{t-1} - \frac{x_t - CA_{t-1}}{t}$$

Weighted moving averages

- **Weighted moving average** = different weights to different data points, usually the most recent data points are more "important"

$$WMA_t = \frac{nx_t + (n-1)x_{t-1} + \cdots + 2x_{t-n+2} + x_{t-n+1}}{n + (n-1) + \cdots + 2 + 1}$$

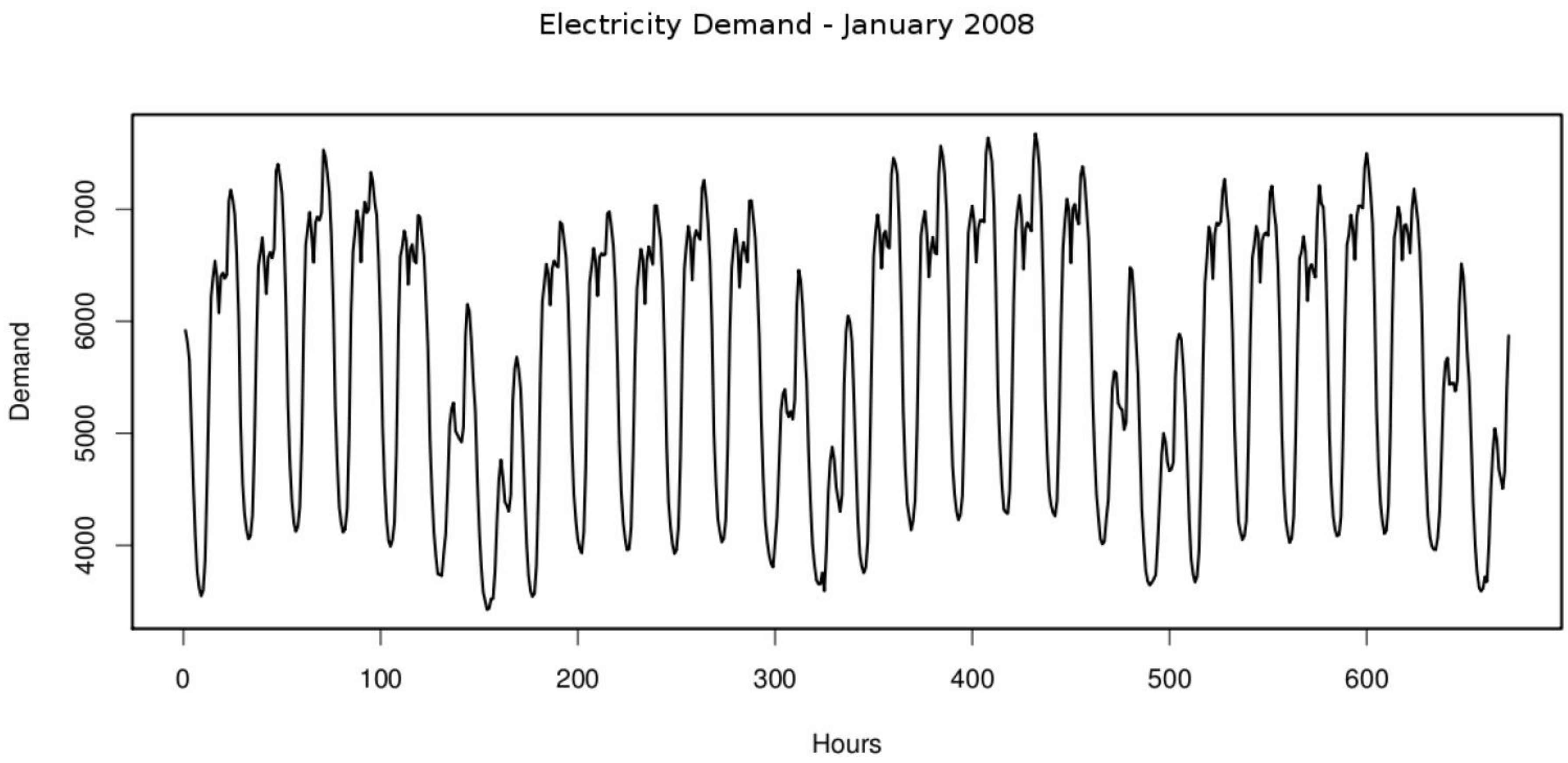
- **Exponential moving average** = the weighting for each older data point decreases exponentially, giving more importance to recent observations while still not discarding older observations entirely

$$EMA_t = \alpha \times x_t + (1-\alpha) \times EMA_{t-1}$$

  - **Note 1:** more weight to recent items
  - **Note 2:** choosing an adequate α is a difficult problem.

Seasonality

Seasonality primer



Gama J. Knowledge discovery in data streams. CRC Press. 2010.

Seasonality intro

- **DEF:** Seasonality is a time series characteristic which signifies regular and predictable changes
  - E.g. different (electricity) load patterns occurring yearly (or different time periods, e.g. weeks)
- A simple way to remove the seasonal component is differencing
- Seasonality is caused by various external and internal factors affecting the system under observation and producing the time series
  - Weather conditions → less travel during icy periods
  - Vacation periods → lower electricity consumption if people travel (not during covid)
  - Other sources? Discuss!

Seasonality and autocorrelation

- **DEF:** **Correlation** is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate)

$$corr_{x,y} = \frac{cov_{x,y}}{\sigma_x \sigma_y} = \frac{E\left[(x - \mu_x)(y - \mu_y)\right]}{\sigma_x \sigma_y},$$

$$x, y\text{: } random\ variables$$

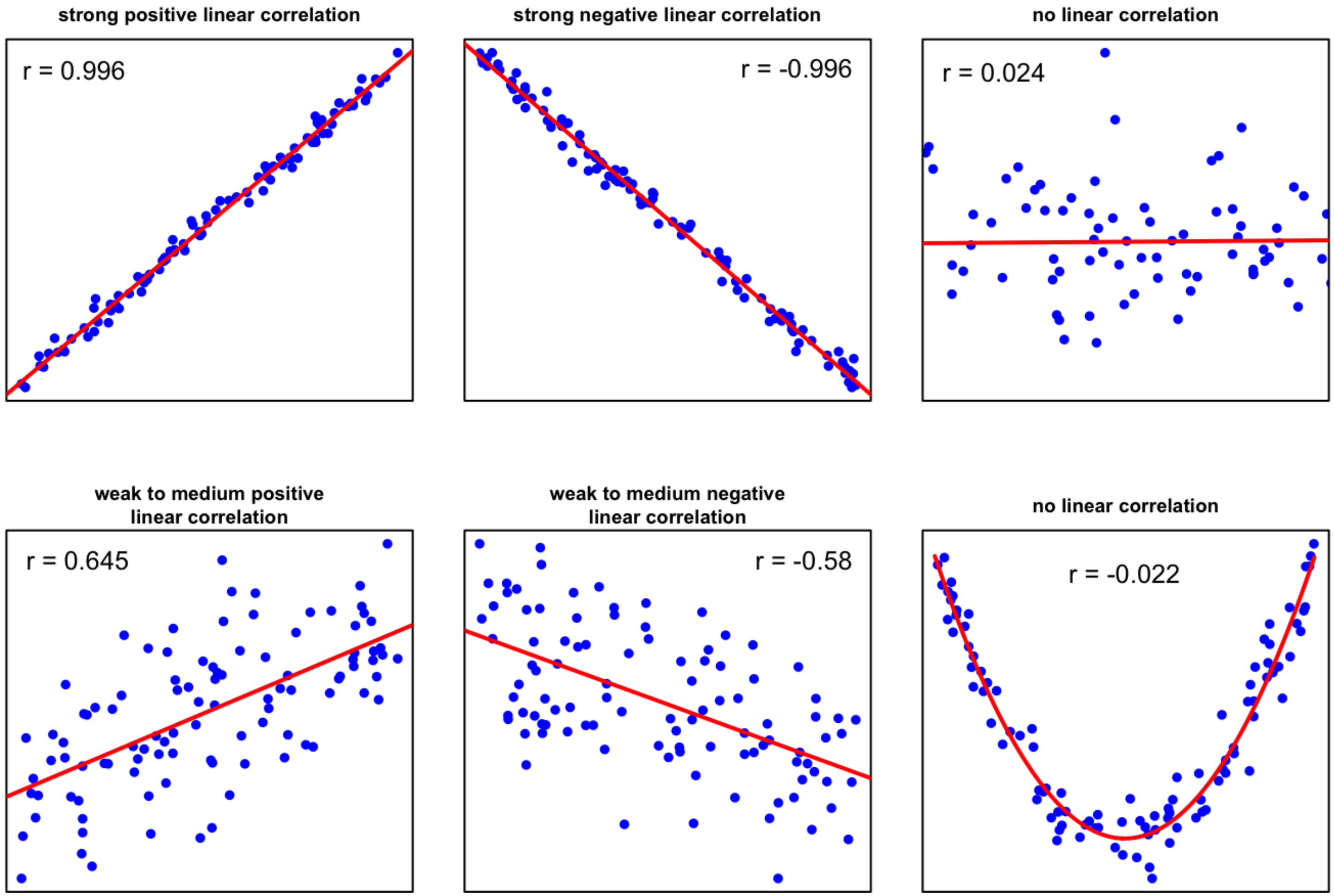
$$\mu_x, \mu_y\text{: } expected\ values$$

$$\sigma_x, \sigma_y\text{ : } standard\ deviations$$

- **DEF:** **Autocorrelation** is the correlation of a signal with a delayed copy of itself as a function of delay
  - Autocorrelation is the cross-correlation of a time-series with itself

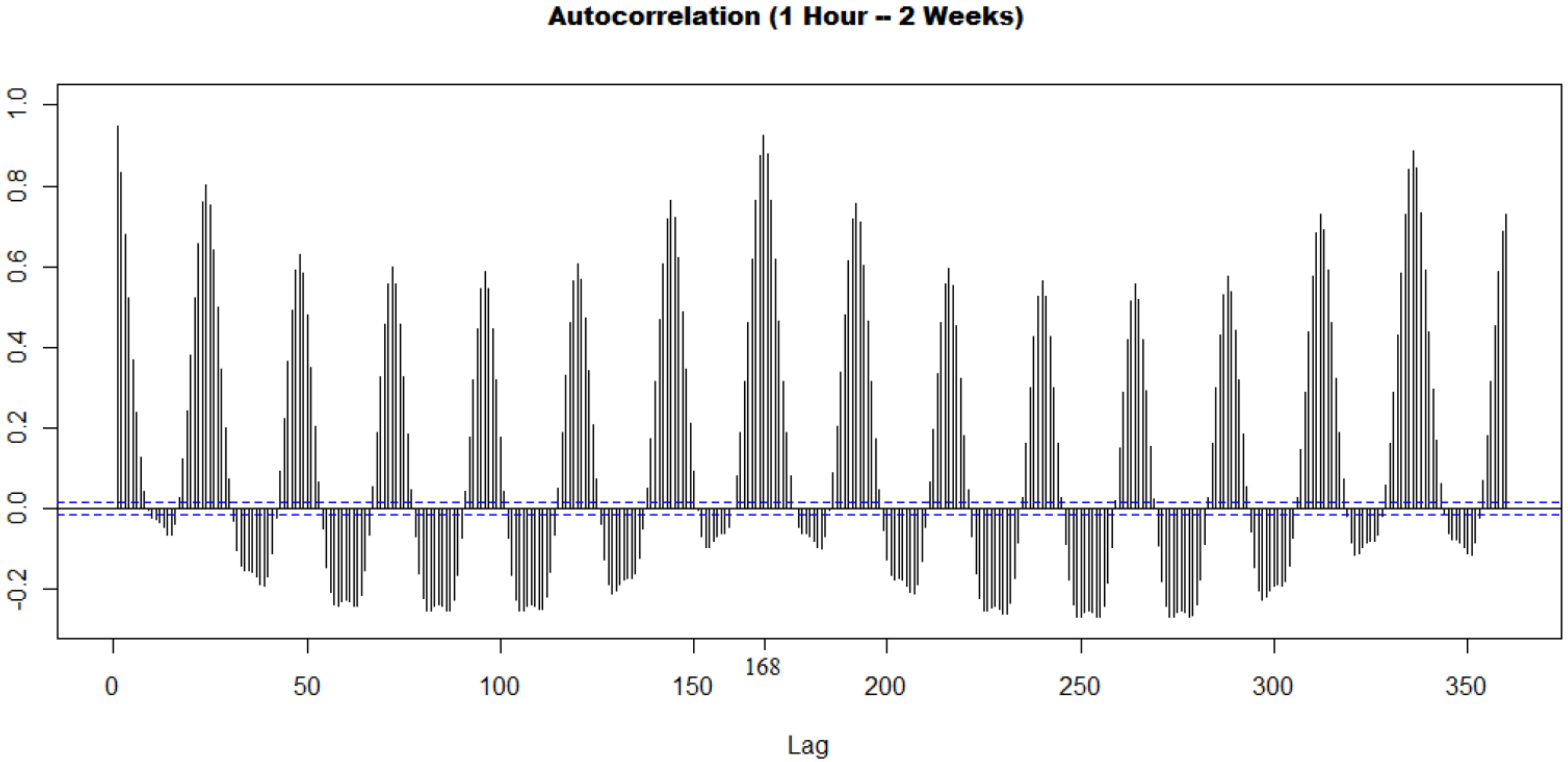
$$r(x, l) = \frac{\sum_{i=1}^{n-l}(x_i - \bar{x})(x_{i+l} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Correlation types



https://www.geo.fu-berlin.de/en/v/soga/Basics-of-statistics/Descriptive-Statistics/Measures-of-Relation-Between-Variables/Correlation/index.html

Power system data correlogram



Gama J. Knowledge discovery in data streams. CRC Press. 2010.

Seasonality and autocovariance

- DEF: Covariation is a measure of the joint variability of two random variables

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1},$$

$$x, y\text{: } random\ variables$$

$$\bar{x}, \bar{y}\text{: } means\ of\ x\ and\ y$$

$$N\text{: } number\ of\ values$$

  - Positive covariance → greater values of variable x correspond to greater values of variable y
  - Negative covariance → greater values of variable x correspond to smaller values of variable y

- **Note:** Both autocorrelation and autocovariance are useful statistics to detect periodic signals

SIMILARITY

Motivation

- **Similarity measures** are necessary for most time series analysis types
  - Assess distance between time series → form clusters
  - Measure distance from classes → assign to classes
  - Lack of (any) similarity → might signify an anomaly
- **Similarity criteria** in time series analysis can be based on
  - Raw data similarity
  - Time series feature similarity
  - Similarity between the underlying (data) generating processes (i.e. model)

Euclidean distance

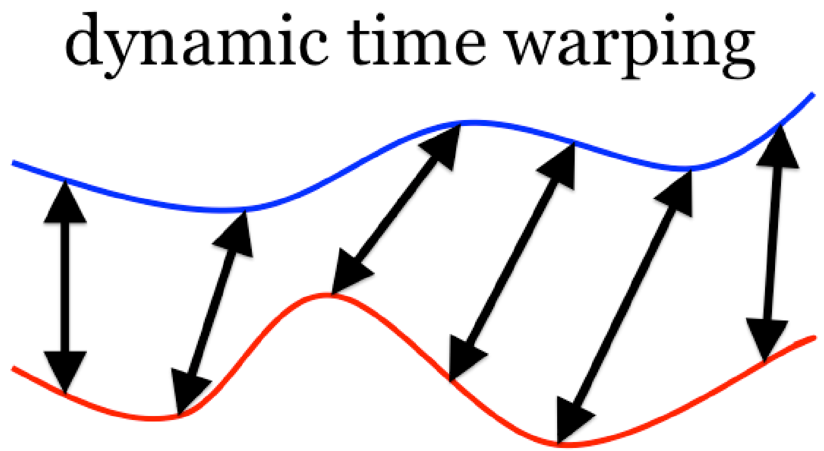
- **DEF:** The Euclidean distance between two time series is the square root of sum of the squared distances between each pair of points between 2 time series
  - Time series alignment is necessary

$$D(C,Q) = \sqrt{\sum_{i=1}^{n} (q_i - c_i)^2}$$

- Satisfies the 4 properties of distance:
  - Identity: $D(Q,Q) = 0$
  - Non-negative: $D(C,Q) \geq 0$
  - Symmetric: $D(C,Q) = D(Q,C)$
  - Satisfies the triangular inequality: $D(Q,C) + D(C,T) \geq D(Q,T)$

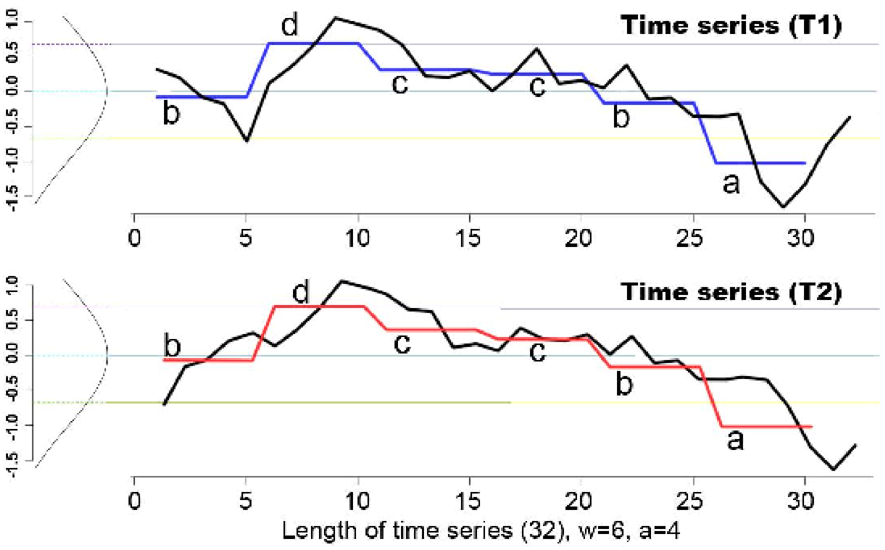
Dynamic Time Warping (DTW)

- Dynamic Time Warping (DTW) is a distance measure for comparing two temporal sequences, which may vary in speed → no alignment needed
- Time complexity: $O(N^2)$
- **Note:** does not allow time scaling of segments
- **Trivia:** A well-known use case is speech recognition with different speaking speeds



https://www.mathworks.com/matlabcentral/fileexchange/43156-dynamic-time-warping-dtw

Symbolic Aggregate Approximation (SAX)

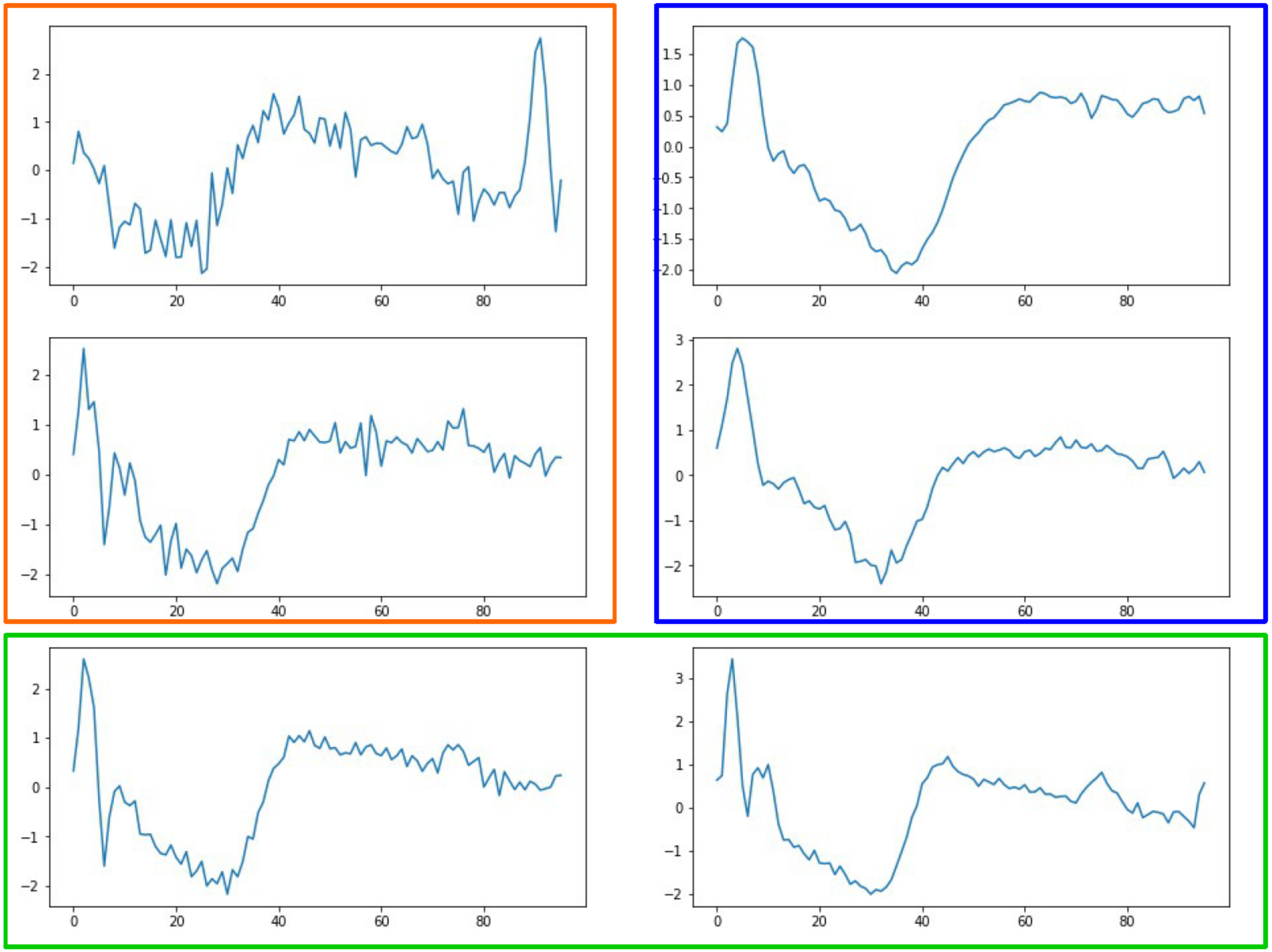


https://www.semanticscholar.org/paper/An-improved-symbolic-aggregate-approximation-based-Zan-Yamana/bf8267be7a70b1f9df982155f12c5786451c7756

- DEF: Symbolic Aggregate Approximation (SAX) transforms a time series into a string of characters
- Complexity: $O(N)$
- Steps:
  - Piecewise Aggregate Approximation (PAA)
  - Symbolic Discretization
  - Distance Measure
- SAX use cases:
  - Motfis = previously unknown frequent patterns
  - Discords = the most unusual time series sub-sequence

CLUSTERING

Clustering primer

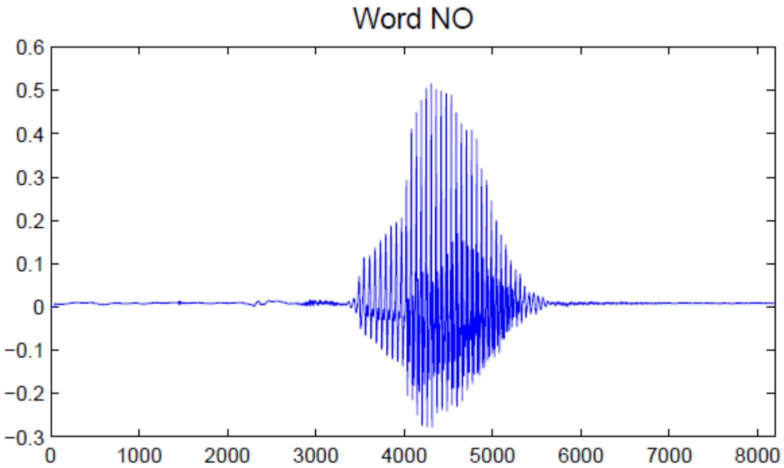


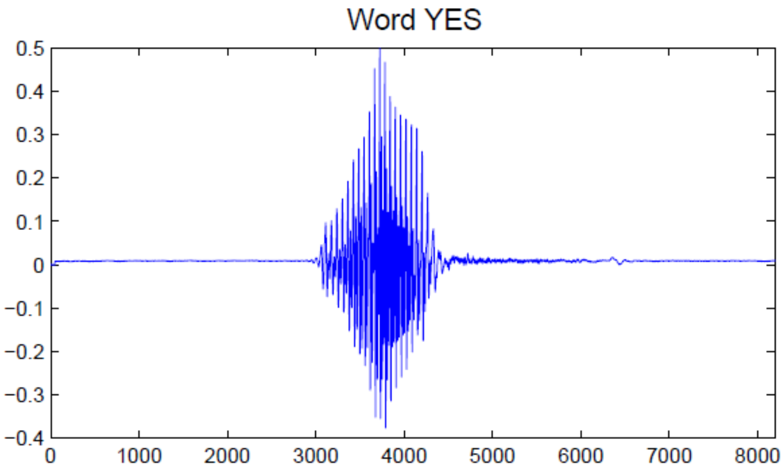
http://www.biointelligence.hu/pdf/timeseriestutorial.pdf

Problem definition and approaches

- **DEF:** In time **series clustering problems** multiple time series (or slices of a single time series) are analyzed with the goal to group them or their subsets into different clusters
- Latest application domains: finance, medicine, seismology, meteorology, etc.
- Approaches
  - Raw data clustering → direct
  - Clustering by features → indirect, based on features
  - Model-based clustering → indirect, based on a model
- Key time series clustering reference: Andrés M. Alonso’s slides from 2019 (unless otherwise stated)

Raw data classification





- DEF: Raw data clustering measures the **element-wise distance** between two (or more) time series

$$D(x_i, x_j) = d(x_i - x_j)$$

- The series need to be **perfectly aligned** → hard to achieve in real-life use cases

- Other raw data approaches: autocorrelation, extreme value

Feature-based clustering

- **DEF: Feature-based clustering** relies on derived statistical features of the time series
  - Assumption: a finite set of statistical measures can be used to capture the global nature of the time series
- **Common** time series features: mean, standard deviation, skewness, periodicity
- **Less common** features: kurtosis, energy, entropy
  - TSFEL: Time Series Feature Extraction Library (60 features)
- Feature-based clustering **advantages**:
  - Reduced dimensionality of the original time series
  - Lower sensitivity to missing data
  - Ability to handle different lengths of time series

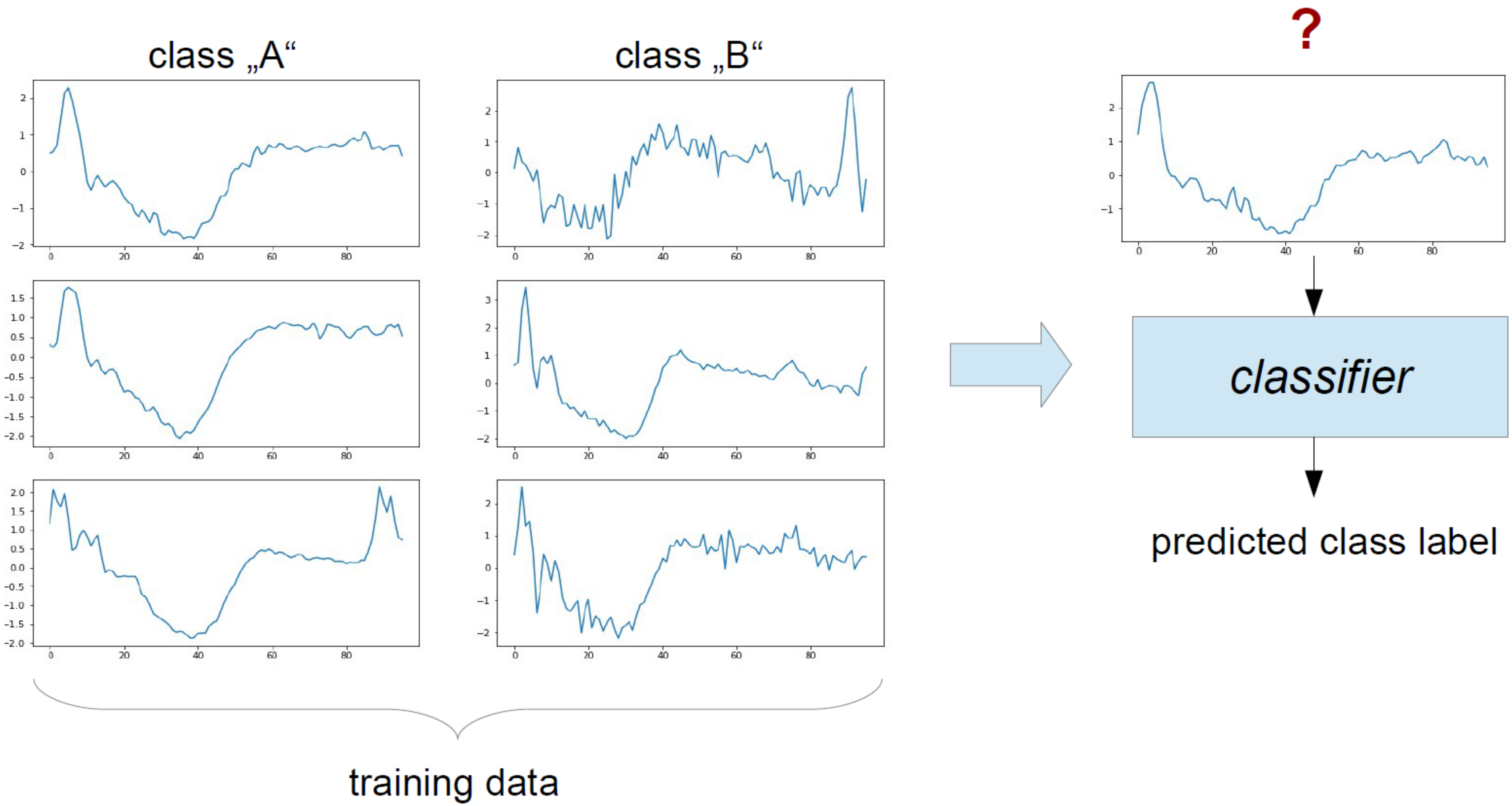
https://tsfel.readthedocs.io/en/latest/descriptions/feature_list.html

Model-based clustering

- **DEF: Model-based clustering** assumes that the data were generated by a model and tries to **recover the original model** from the data
- The model recovered from the data defines the clusters
  - In a K-means approach the model is a set of centroids which (are supposed to had) generated the data
- Advantages:
  - Low computational cost (if the model-matching is ‘cheap’)
- Disadvantages:
  - It might be challenging to derive a correct model
- Source: Stanford NLP Group

CLASSIFICATION

Classification primer



Buza K. Time Series Classification and its Applications, 8th International Conference on Web Intelligence, Mining and Semantics. June 25 – 27 2018, Novi Sad, Serbia.

Classification techniques

- **Similarity-based classification**, e.g. nearest neighbor, hubness-aware classifiers
  - Classification based on characteristic local patterns, e.g. motif-based, shapelet-based
- **Feature-based classification**
  - Feature extraction + a standard classifier such as SVM, Naive Bayes, decision tree...
  - Possible features: min, max, avg, std, etc.
- **Other** techniques:
  - Hidden Markov Models
  - Deep learning with CNN

Buza K. Time Series Classification and its Applications, 8th International Conference on Web Intelligence, Mining and Semantics. June 25 – 27 2018, Novi Sad, Serbia.

Evaluation

Evaluation protocol

- Simulate real-life applications and data as much as possible → why train a classifier it will not be used?!
- Independent test set
- Cross-validation

Evaluation metrics

- Accuracy, AUC, precision, recall, F-measure, AUPR
- Standard deviation, statistical significance tests
- Note: Be careful when evaluating any solution on unbalanced data

Stream mining challenge

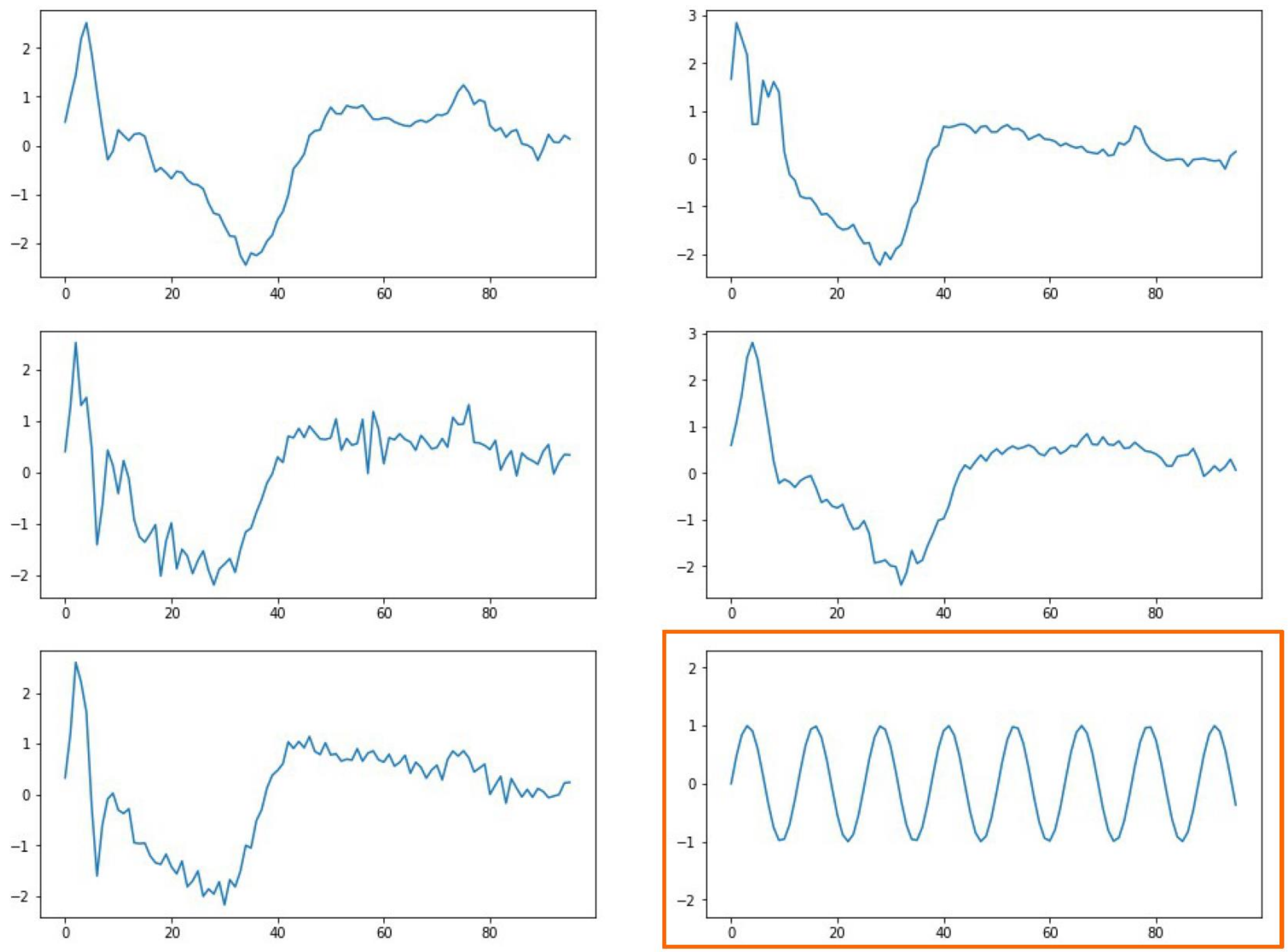
Buza K. Time Series Classification and its Applications, 8th International Conference on Web Intelligence, Mining and Semantics. June 25 – 27 2018, Novi Sad, Serbia.

Additional references

- Buza, Schmidt-Thieme (2009): **Motif-based** classification of time series with Bayesian networks and SVMs, Advances in Data Analysis, Data Handling and Business Intelligence. Springer, Berlin, Heidelberg, pp. 105-114
- Hills et al. (2014): Classification of time series by **shapelet** transformation, Data Mining and Knowledge Discovery, 28(4), pp. 851-881

ANOMALY DETECTION

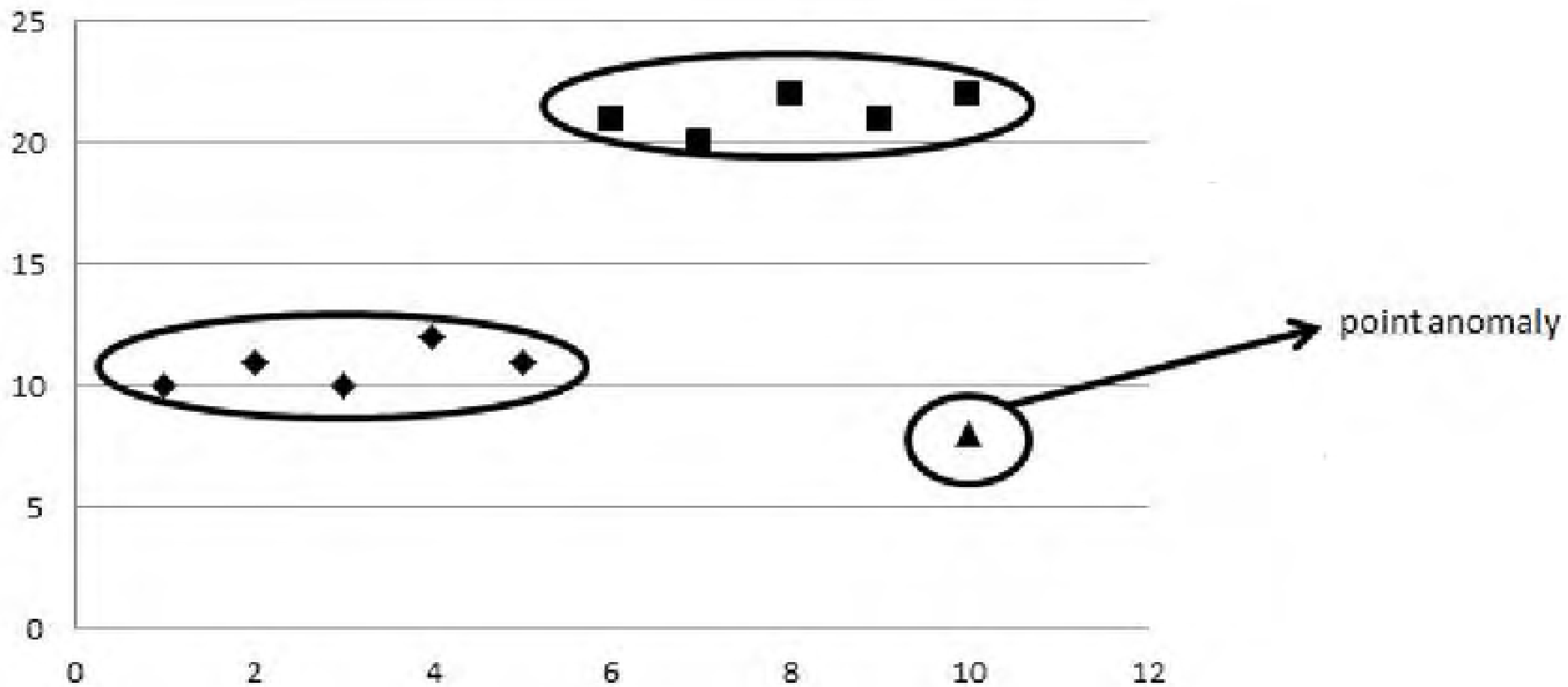
Anomaly detection primer



Buza K. Time Series Classification and its Applications, 8th International Conference on Web Intelligence, Mining and Semantics. June 25 – 27 2018, Novi Sad, Serbia.

Type #1: Point anomalies

- DEF: In a **point anomaly** an individual data instance is anomalous with respect to its surroundings



Baddar, S. W. A. H., Merlo, A., & Migliardi, M. (2014). Anomaly Detection in Computer Networks: A State-of-the-Art Review. *JoWUA*, *5*(4), 29-64.

Type #2: Contextual anomalies

- **DEF:** In **contextual anomalies** a data instance is anomalous in specific context, but otherwise might be

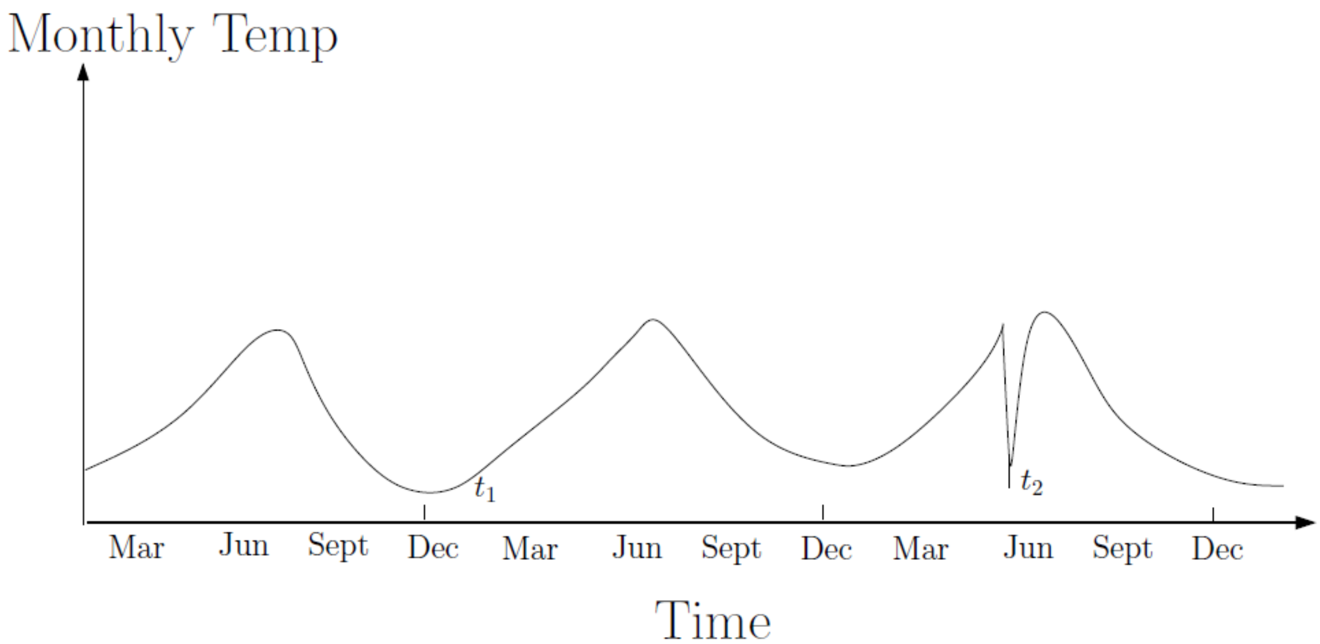


Fig. 3. Contextual anomaly t_2 in a temperature time series. Note that the temperature at time t_1 is same as that at time t_2 but occurs in a different context and hence is not considered as an anomaly.

Chandola V., Banerjee A., Kumar V., “Anomaly Detection: A Survey”, Technical Report TR 07-017, 2007

Type #3: Collective anomalies

- DEF: **Collective anomalies** are collections of data instances anomalous in relation to the entire data set

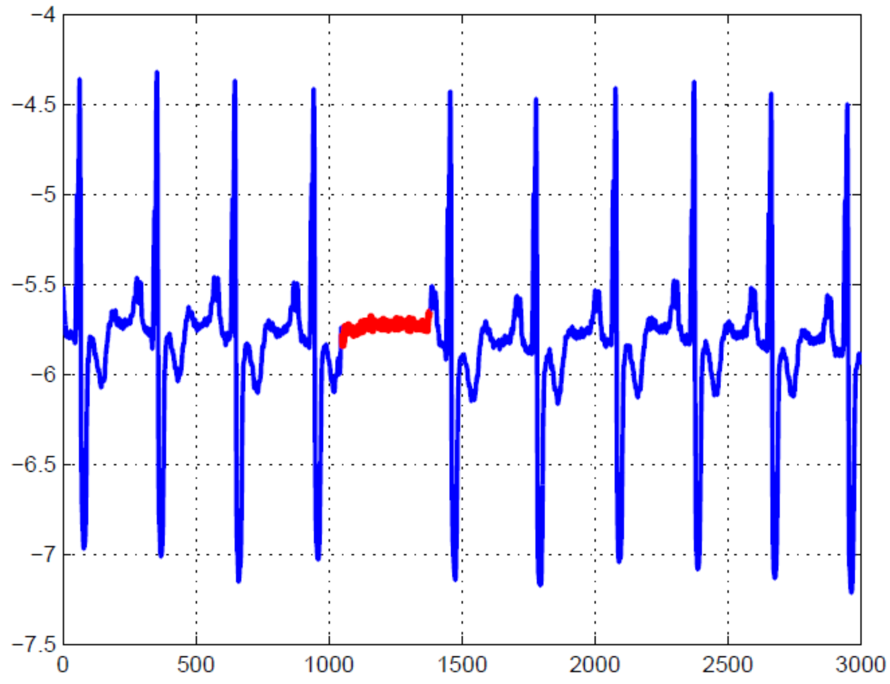


Fig. 4. Collective anomaly corresponding to an *Atrial Premature Contraction* in an human electrocardiogram output.

Chandola V., Banerjee A., Kumar V., “Anomaly Detection: A Survey”, Technical Report TR 07-017, 2007

Anomaly detection techniques

- **Seasonal and Trend decomposition using Loess (STL)** → split time series into (season, trend, residue)
  - The residue element contains the anomalies
- **Classification** → applicable if there is labeled data → no class means outlier/anomaly
- **Auto Regressive Integrated Moving Average (ARIMA)** → predict future points → detect discrepancies
  - Several points in the past used to forecast next point + noise
- **Long short-term memory (LSTM)**
  - Malhotra, Pankaj; Vig, Lovekesh; Shroff, Gautam; Agarwal, Puneet (April 2015). "Long Short Term Memory Networks for Anomaly Detection in Time Series". ESANN 2015.
- + many other methods

Summary

- Introduction
- Time series categories
- Trends & seasonality
- Similarity
- Clustering
- Classification
- Anomaly detection

Common references

- Aggarwal, C. C. (2015). Data mining: the textbook. Springer.
  - Note: chapter “Mining time series data”
- Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering – a decade review. Information Systems, 53, 16-38.
- Buza K. Time Series Classification and its Applications, 8th International Conference on Web Intelligence, Mining and Semantics. June 25 – 27 2018, Novi Sad, Serbia. http://www.biointelligence.hu/pdf/timeseriestutorial.pdf
- Esling, P., & Agon, C. (2012). Time-series data mining. ACM Computing Surveys (CSUR), 45(1), 1-34.
- Gama J. Knowledge discovery in data streams. CRC Press. 2010.
- Holan, S. H., & Ravishanker, N. (2018). Time series clustering and classification via frequency domain methods. Wiley Interdisciplinary Reviews: Computational Statistics, 10(6), e1444.
- Kotsakos, D., Trajcevski, G., Gunopulos, D., & Aggarwal, C. C. (2013). Time-Series Data Clustering.
- Maharaj, E. A., D'Urso, P., & Caiado, J. (2019). Time series clustering and classification. CRC Press.

Thank you for your attention!