XAI-TS Workshop 2023

Towards explainable time series classification

Turin, September 18, 2023

Panagiotis Papapetrou, *Professor, Stockholm University*

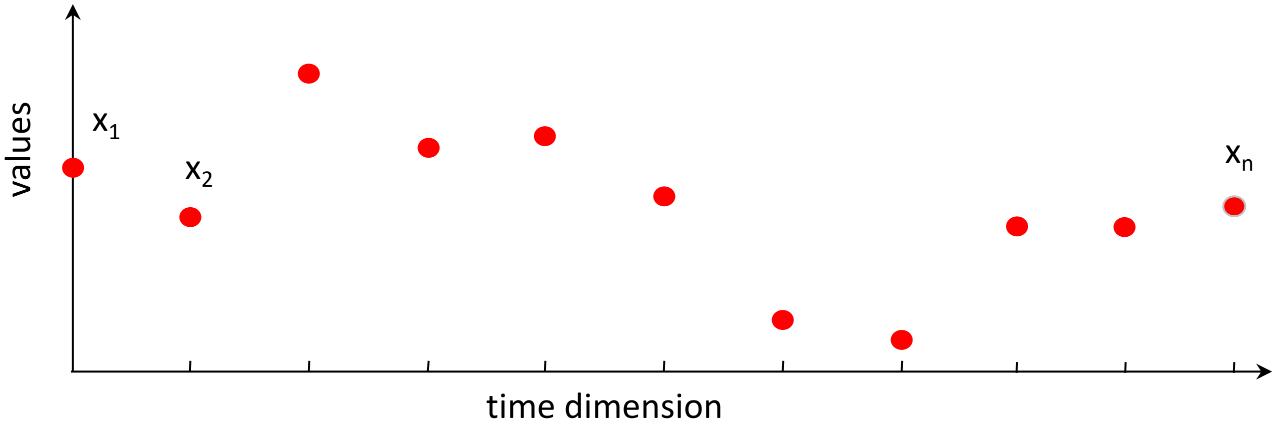
Stockholms universitet

Agenda

- Introduction
- Time series classification
- Explainable time series classification
- Time series counterfactuals
- Challenges and future directions

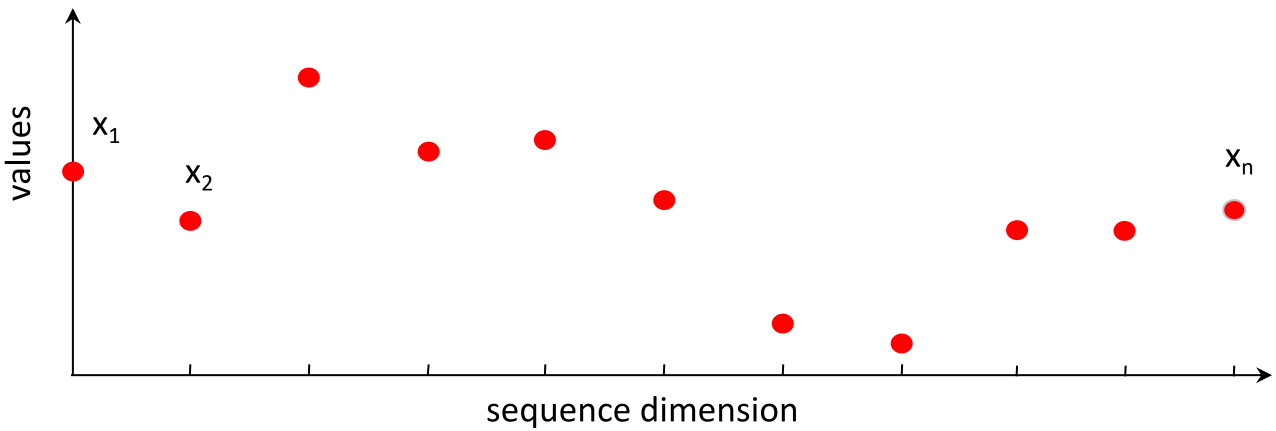
Time series

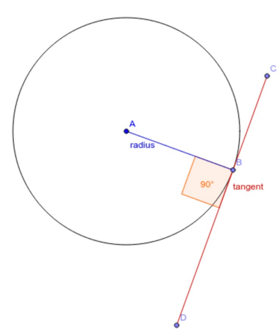
- Sequence of measurements ordered over time



Data series

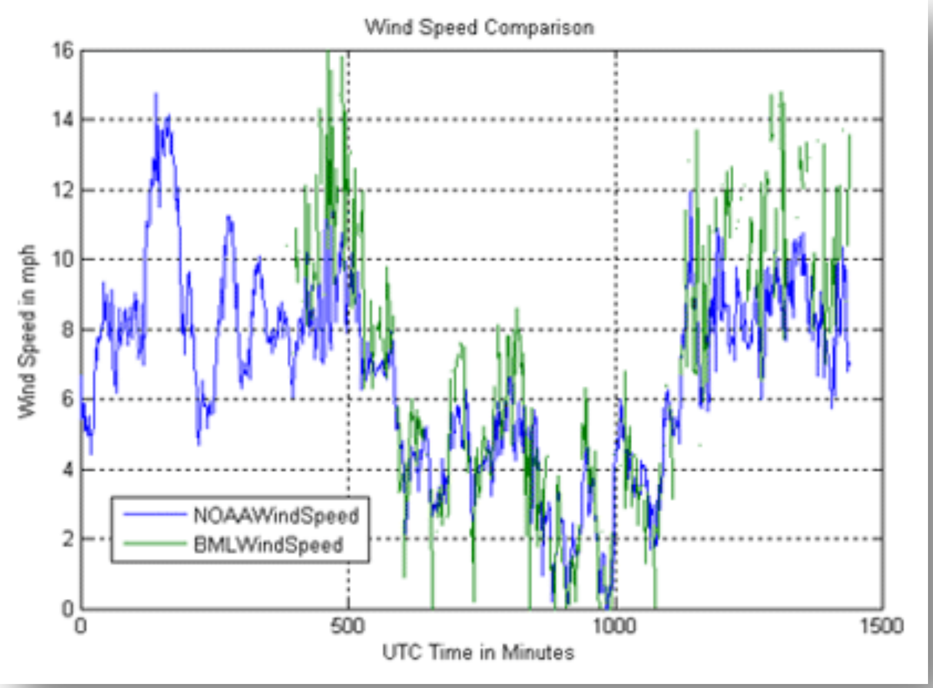
- Sequence of points ordered along some dimension





Data series

- Sequence of points ordered along some dimension



Wind speed
From ocean observing node project, http://bml.ucdavis.edu/boon/wind.html

Data series

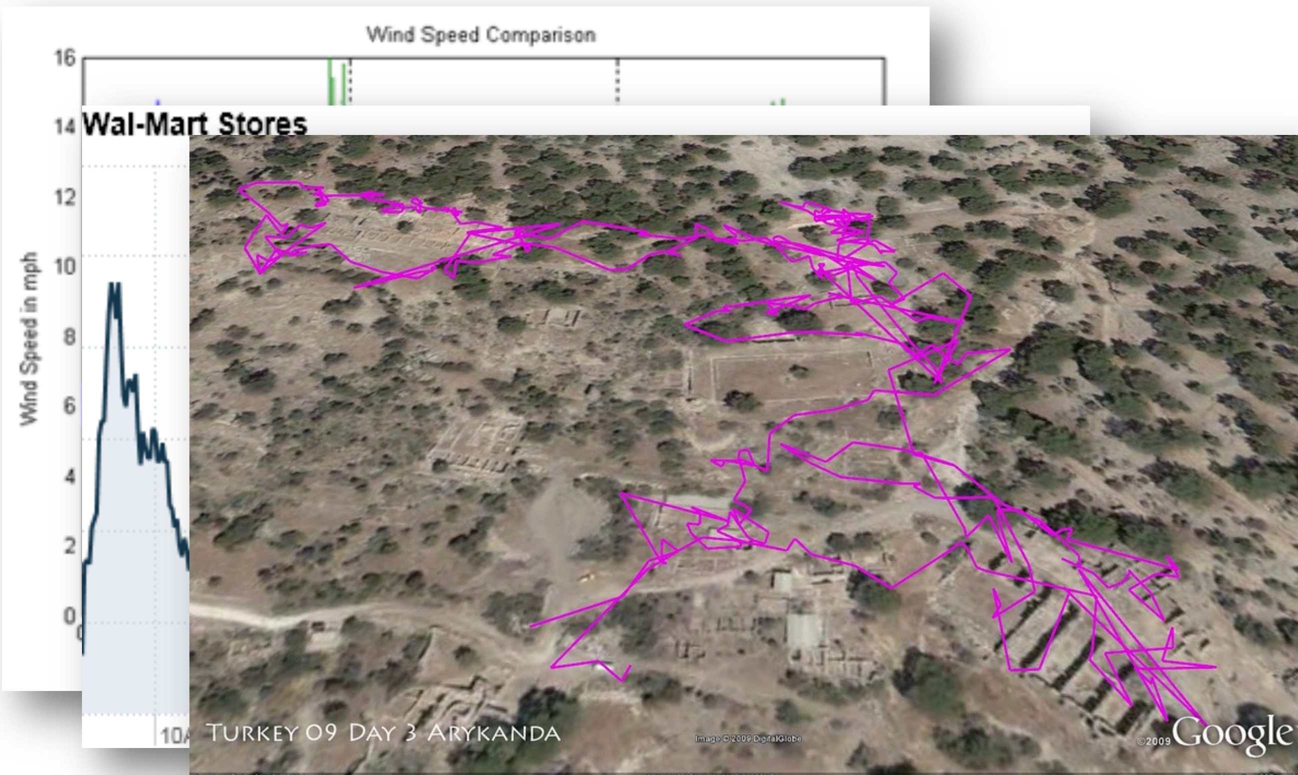
- Sequence of points ordered along some dimension

Historical stock quotes

From http://money.cnn.com/2012/04/23/markets/walmart_stock/index.htm

Data series

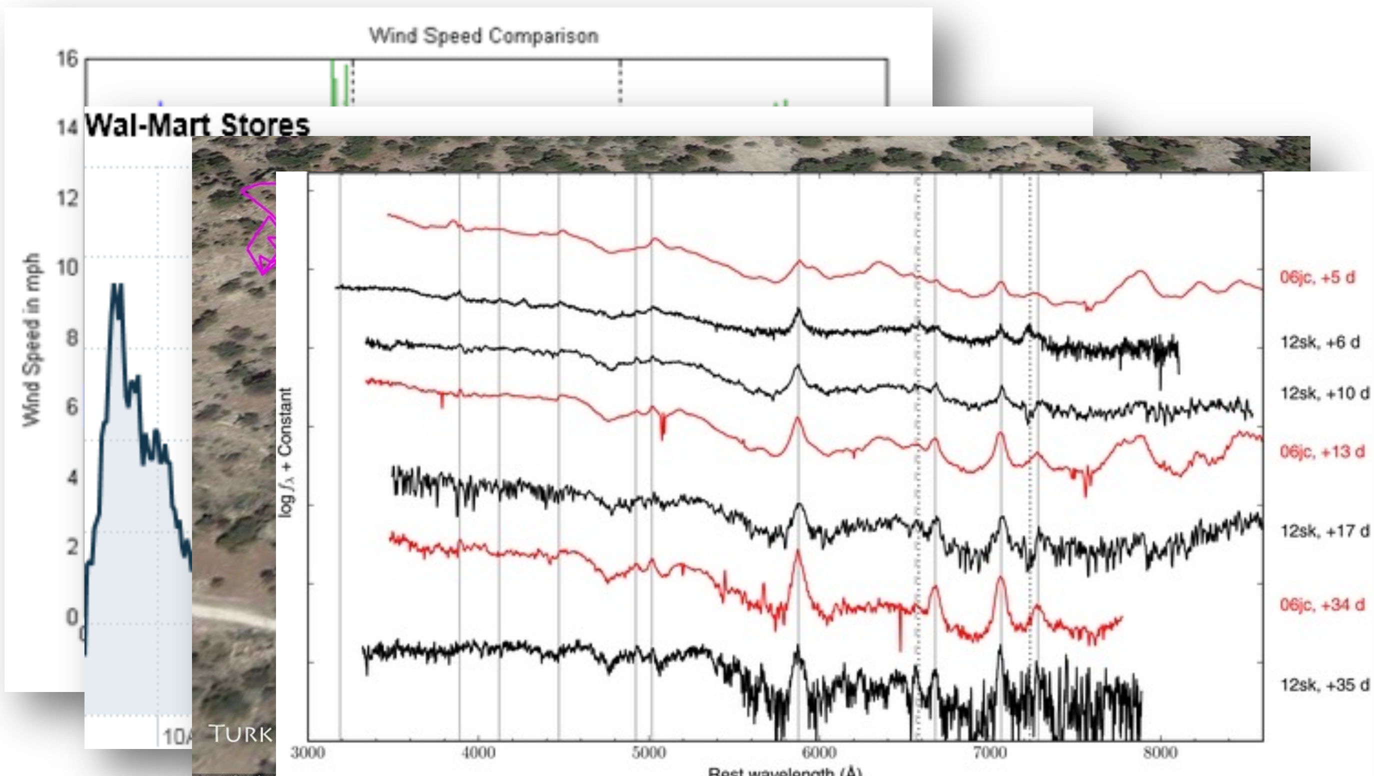
- Sequence of points ordered along some dimension



Trajectories from GPS logs
From http://www.flickr.com/photos/kitepuppet/3604115258

Data series

- Sequence of points ordered along some dimension

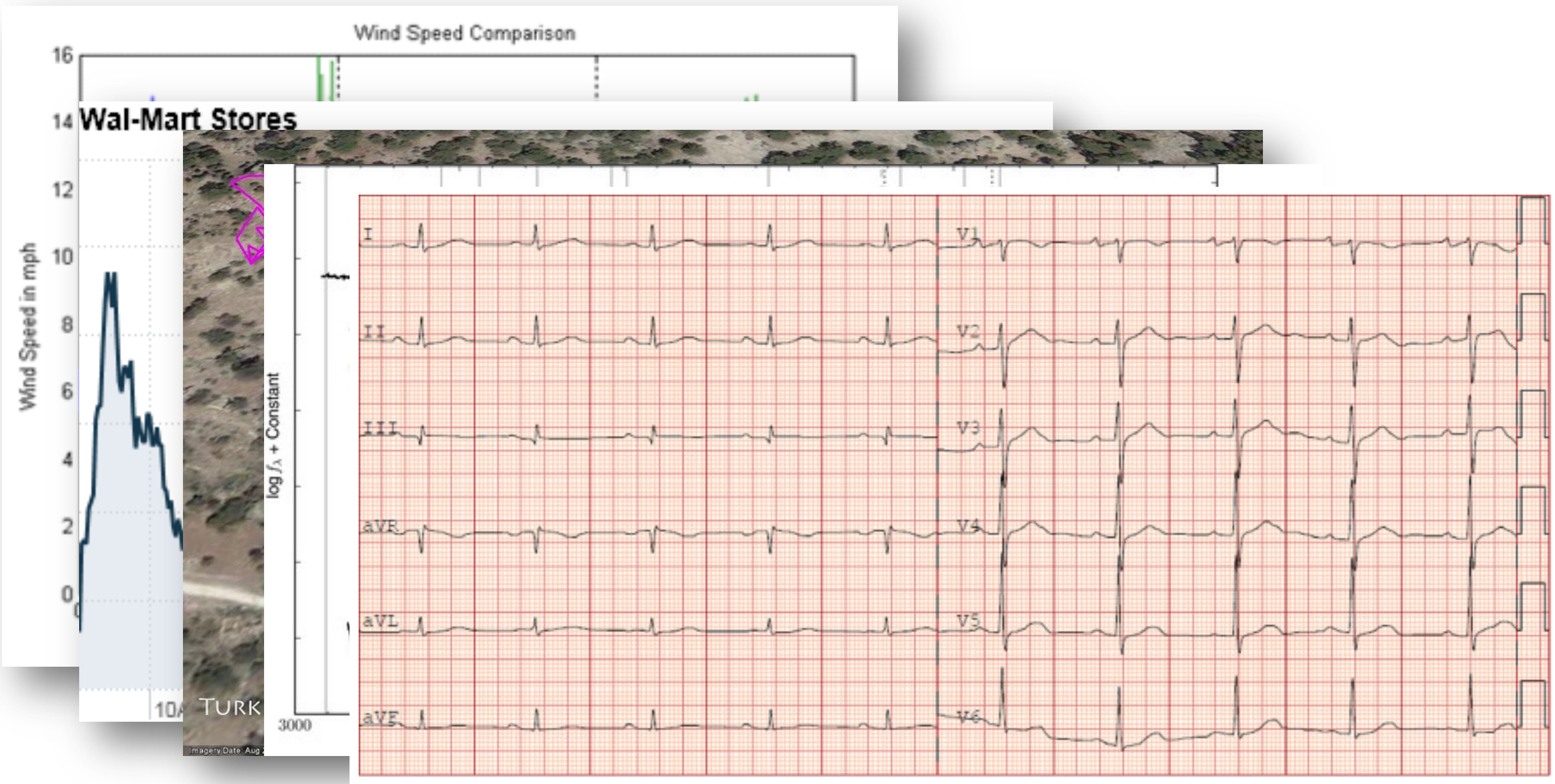


Spectroscopic sequence data (astronomy)

From Sanders et al., http://dx.doi.org/10.1088/0004-637X/769/1/39

Data series

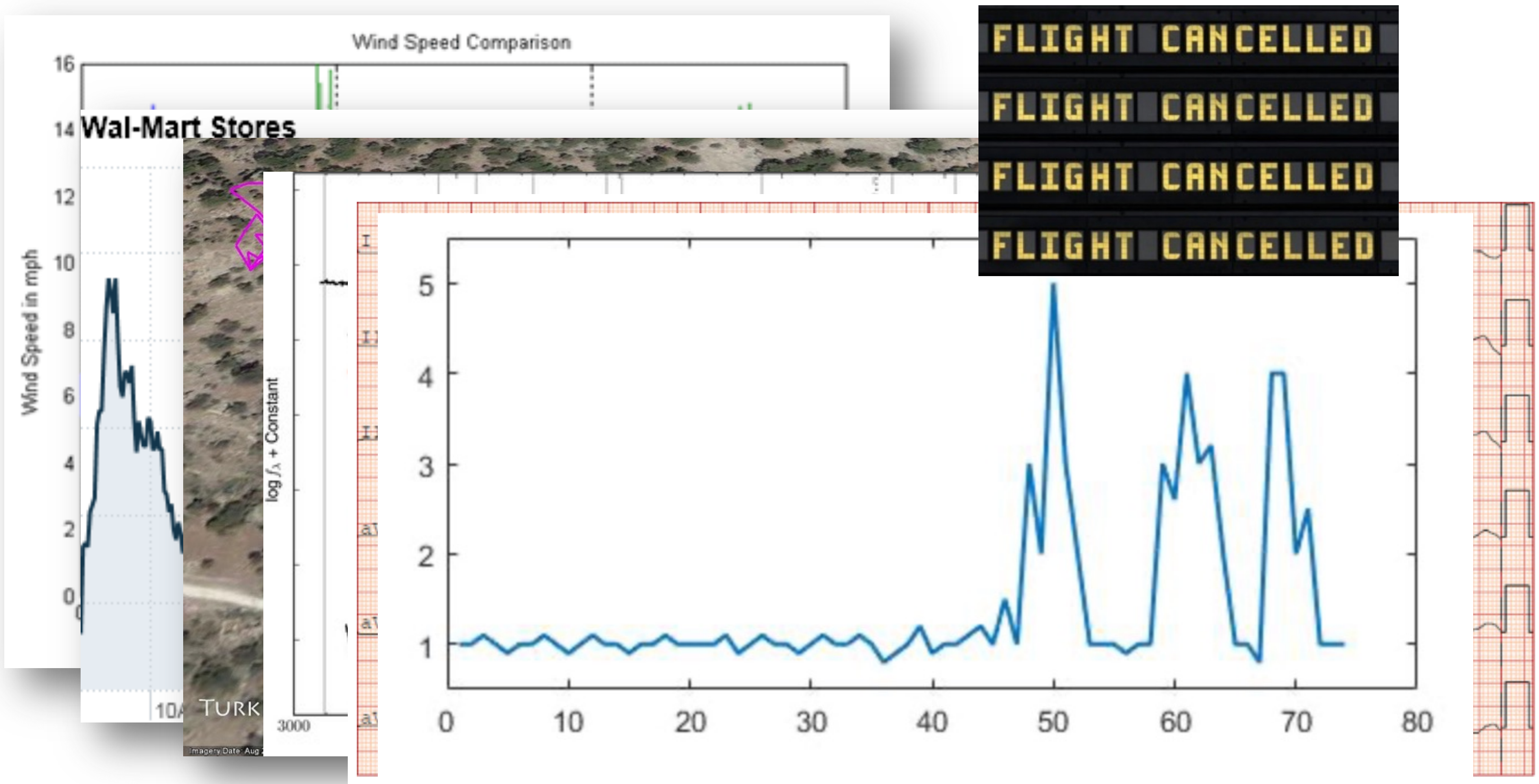
- Sequence of points ordered along some dimension



Electocardiograms (cardiology)
https://archive.physionet.org/physiobank/

Data series

- Sequence of points ordered along some dimension



Customer satisfaction/frustration

Data series analysis tasks

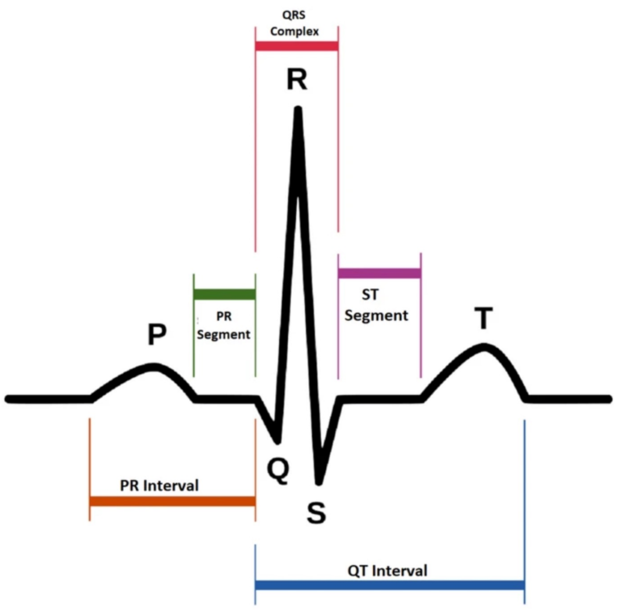
- Clustering
- Anomaly Detection
- Forecasting
- Classification
- Motif Discovery
- Similarity search

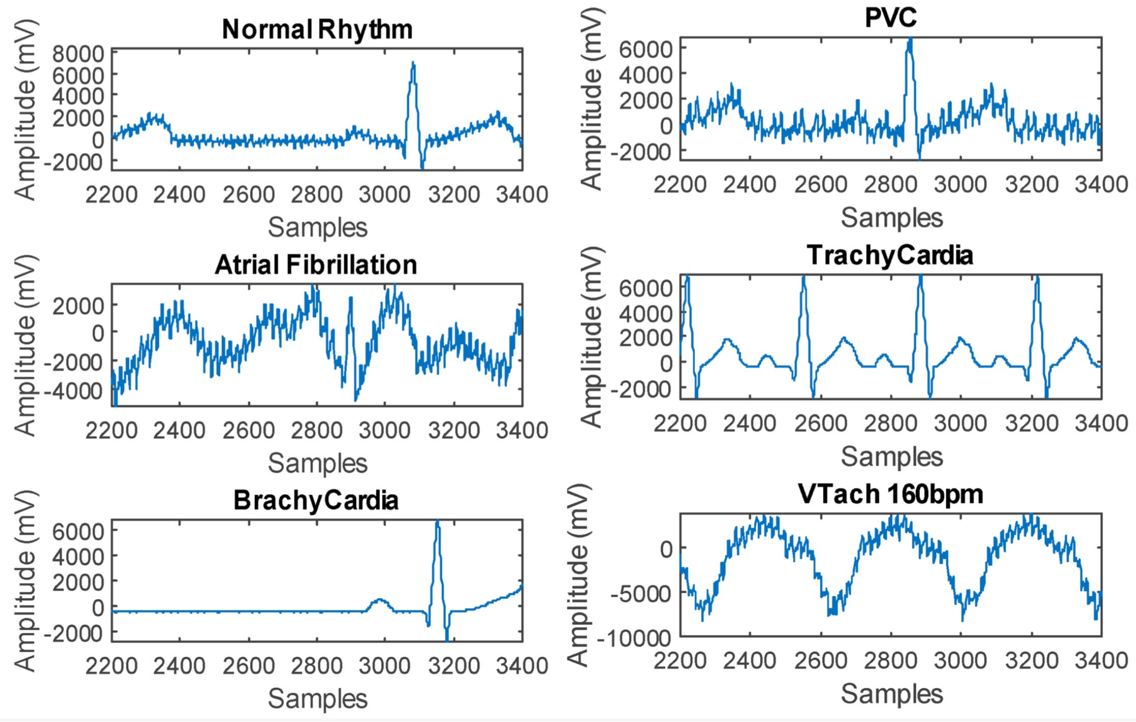
Agenda

- Introduction
- Time series classification
- Explainable time series classification
- Time series counterfactuals
- Challenges and future directions

Time series classification

Abnormal (binary)
atrial fribilation (multiclass)





Many time series classifiers

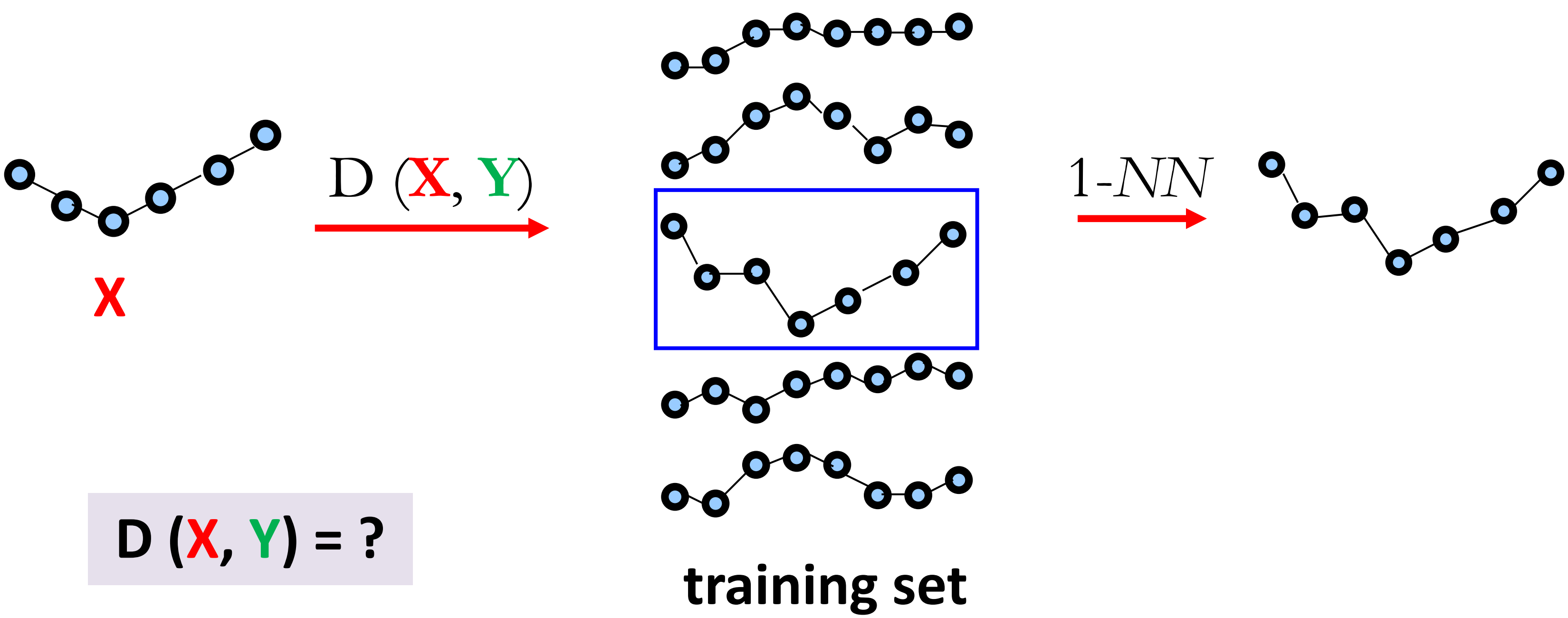
Feature-based

Distance-based

Deep learning-based

k-NN time series classification

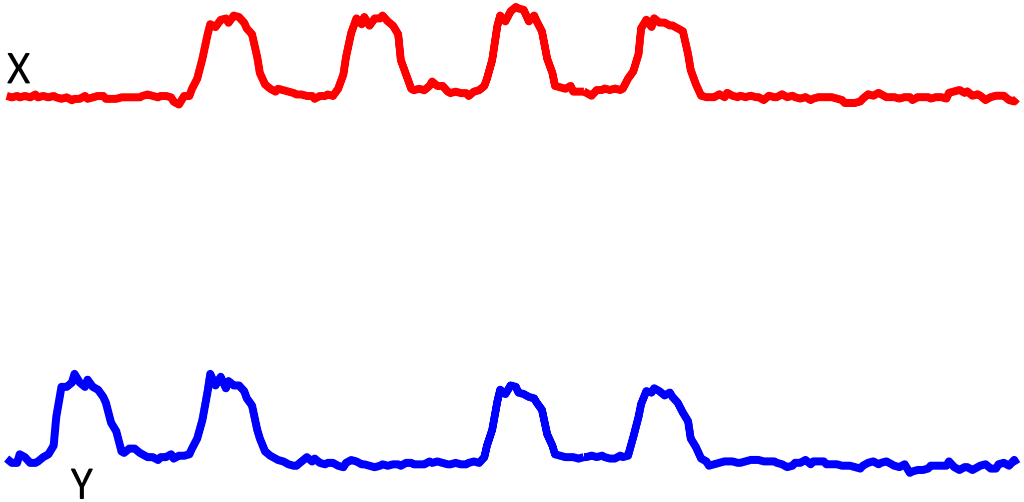
- Given a time series training set **Y** and a test time series **X**
- Find the best match of **X** in **Y**
- Assign the class of the *1-NN* to Q



D (X, Y) = ?

Euclidean and Dynamic Time Warping

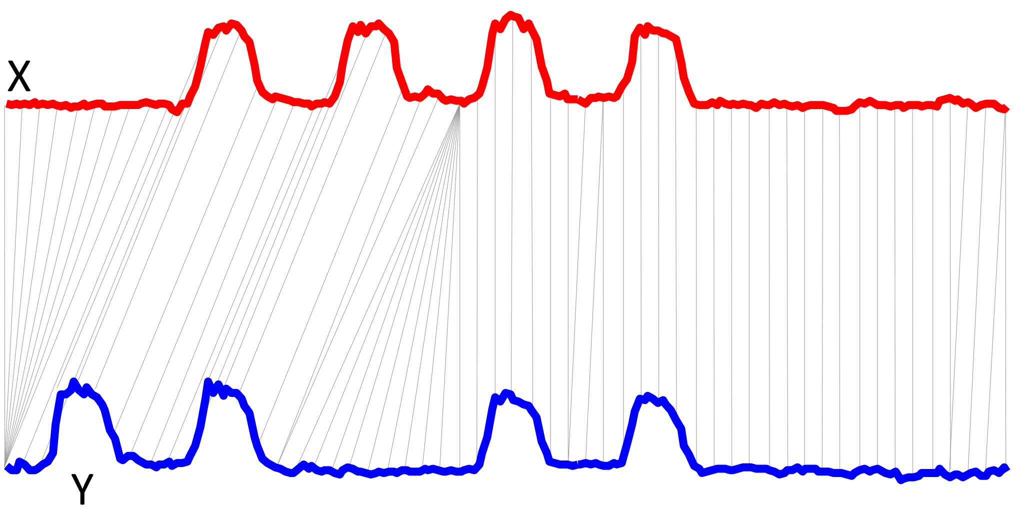
figures taken from Eamonn Keogh, University of California, Riverside



Euclidean Distance

Sequences are aligned "one to one".

$$D(X,Y) \equiv \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$



"Warped" Time Axis

Nonlinear alignments are possible.

Other time series distance measures

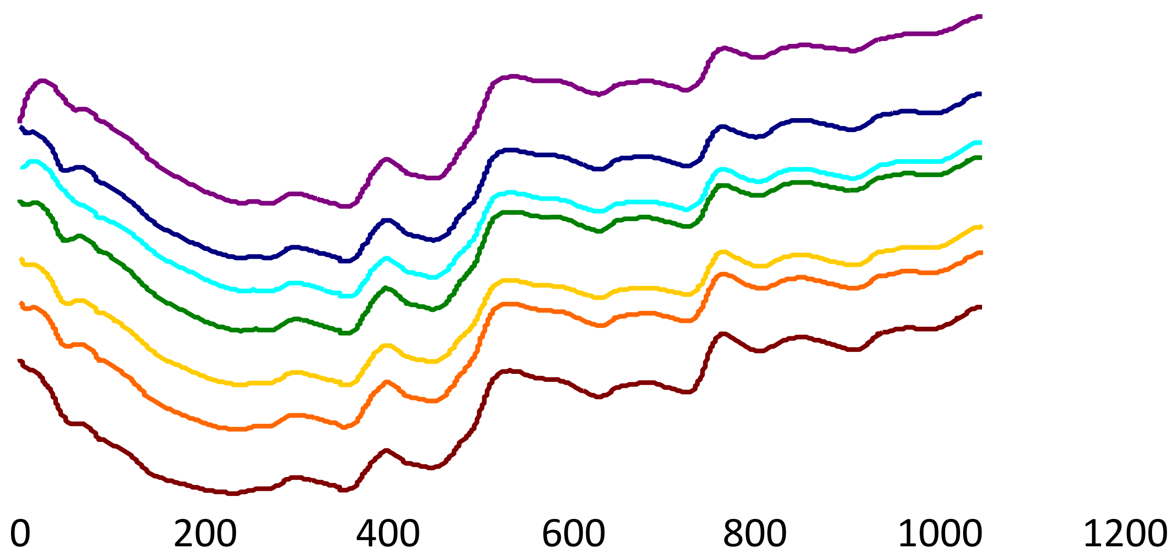
- **DDTW**: Derivative DTW
- **WDTW**: Weighted DTW
- **LCSS**: Longest Common Subsequence
- **MSM**: Move-Split-Merge
- **ERP**: Edit Distance with Real Penalty
- **TWE**: Time Warp Edit

https://hal.science/hal-03515496/document

Limitations of *k-NN* time series classifiers

figure taken from Eamonn Keogh, University of California, Riverside

- Given seven time series classes



- *k-NN* is unable to identify smaller patterns or shapes that are class discriminant

Many time series classifiers

Feature-based

Distance-based

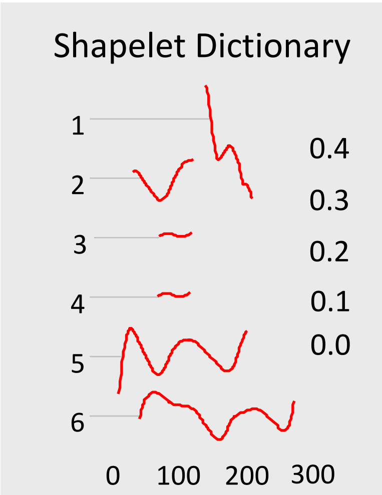
Deep learning-based

How about feature-based classification?

- Use **shapelets** as "attributes" or "features" for splitting a node in the decision tree

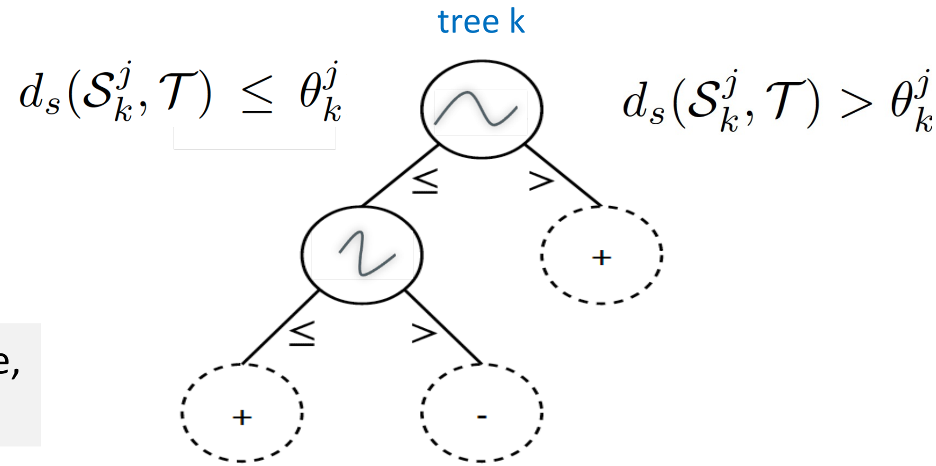
- **Shapelets:**
  - time series subsequence
  - *maximally representative* of a class
  - *discriminative* from other classes

The Shapelet Tree classifier

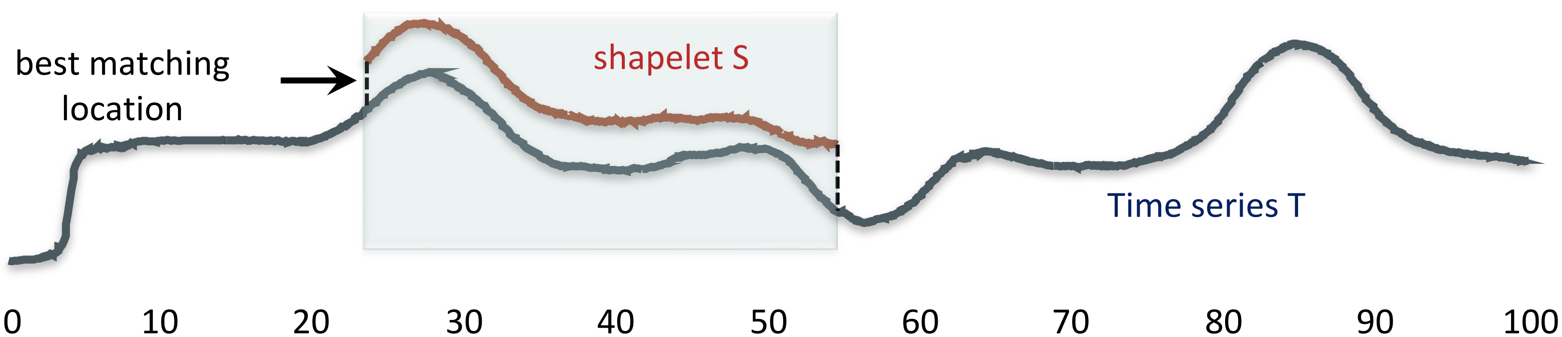


The tree contains several root-leaf paths

$$p_{k,j} = \{(x_1 \lesseqgtr \theta_1), (x_2 \lesseqgtr \theta_2), \ldots, (x_n \lesseqgtr \theta_n)\}$$

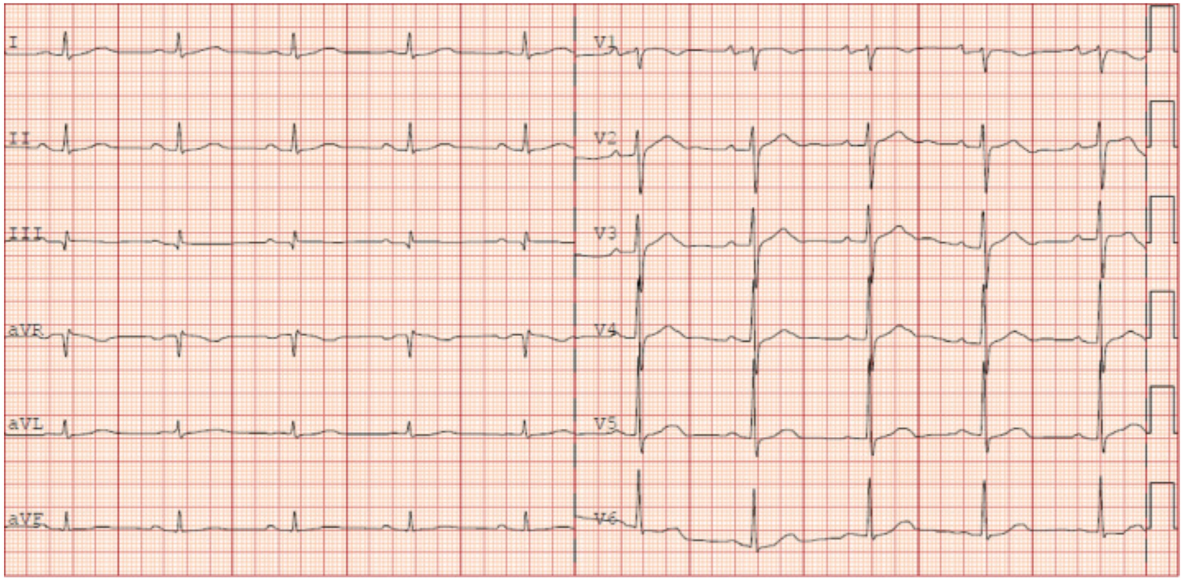


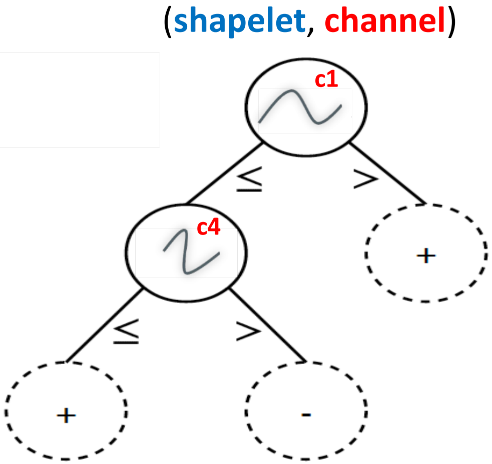
non-leaf node condition: Euclidean distance, lowest scoring subsequence match of S in T



Generalized Random Shapelet Forest (gRSF)

- A generalization of RSF for multivariate time series classification
- ***T*** random shapelet trees are built
  - each tree is built from a random sample (with replacement) of *time series channels* in the training set (channels are recorded in the decision nodes)
  - inspect ***r*** random shapelets at each node





Other shapelet-based approaches

- Transformations + k-NN
  - improved subsequence searching and matching, using online normalization, early abandoning, and re-ordering
  - dimensionality reduction using SAX
- Shapelet-based features
  - select the top k most informative shapelets as features
  - learn any suitable classifier (e.g., SVM, Random Forest) using the transformed dataset
- Synthetic shapelet generation
  - initialize using, e.g., K-means clustering
  - learn synthetic Shapelets

	s_1	s_2	$\dots$	s_k
d_1	0.3	3.3	$\dots$	0.1
d_2	0.2	3.2	$\dots$	3.8
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
d_n	3.1	0.9	$\dots$	9.6

Other feature-based classifiers

- **STC**: Shapelet Transform
- **BOSS**: Bag-of- SFA-Symbols
- **WEASEL**: Word eXtrAction for time SEries cLassification
- **MrSEQL**: Multiple Representation Sequence Learner

https://hal.science/hal-03515496/document

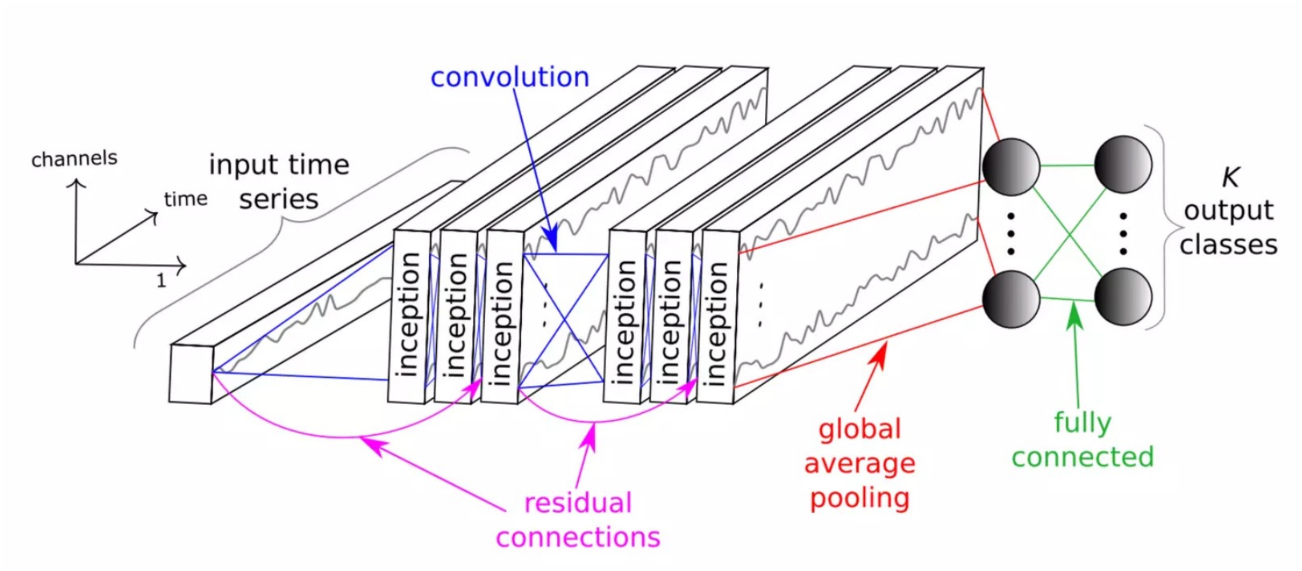
Many time series classifiers

Feature-based

Distance-based

Deep learning-based

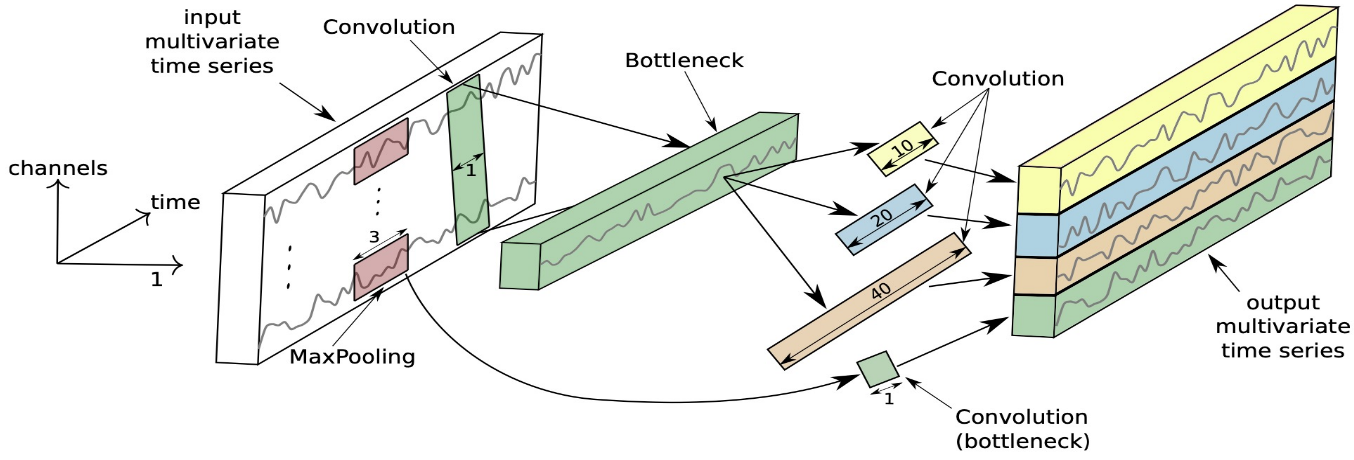
Inception Time [Fawaz 2020]



- The equivalent of **AlexNet** for time series
- An ensemble of five **deep learning models**
  - each created by cascading multiple inception modules
  - each having exactly the same architecture but with different randomly initialized weight values

https://arxiv.org/abs/1909.04939

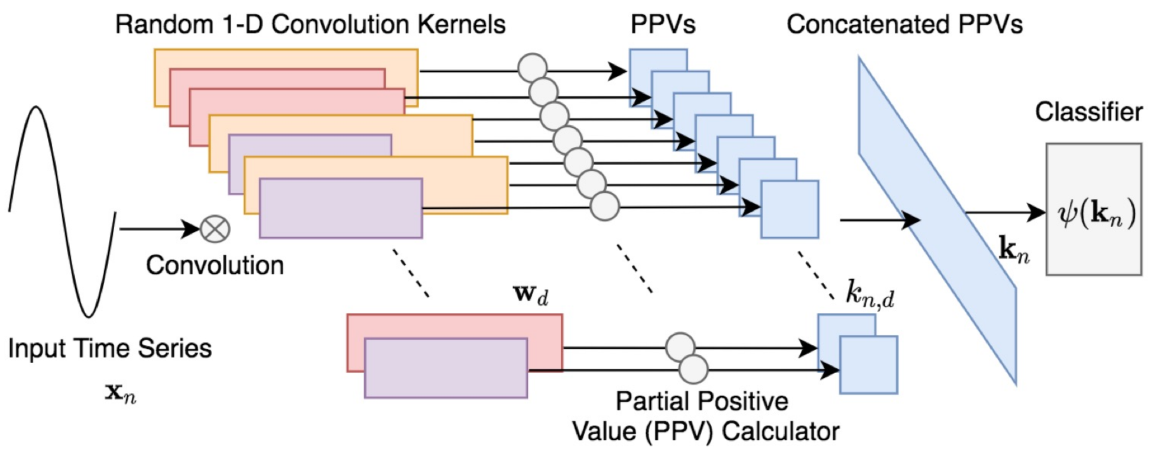
Inception Time [Fawaz 2020]



- **Core idea of an inception module:**
  - apply multiple filters simultaneously to an input time series
  - includes filters of varying lengths allowing the network to automatically extract relevant features from both long and short time series

https://arxiv.org/abs/1909.04939

ROCKET [Dempster et al. 2021]



In short…

- **ROCKET** initializes a bank of random convolution kernels (e.g., 10 000)
- The convolution of each kernel with an input time series produces a feature vector
- Each feature vector is represented by the proportion of positive values (**PPV**) and/or the maximum value (**max pooling**)
- The concatenation of PPV values from the kernels + the max pooling values is used as the input feature vector to train a Ridge regression classifier

https://arxiv.org/pdf/1910.13051.pdf https://github.com/angus924/rocket

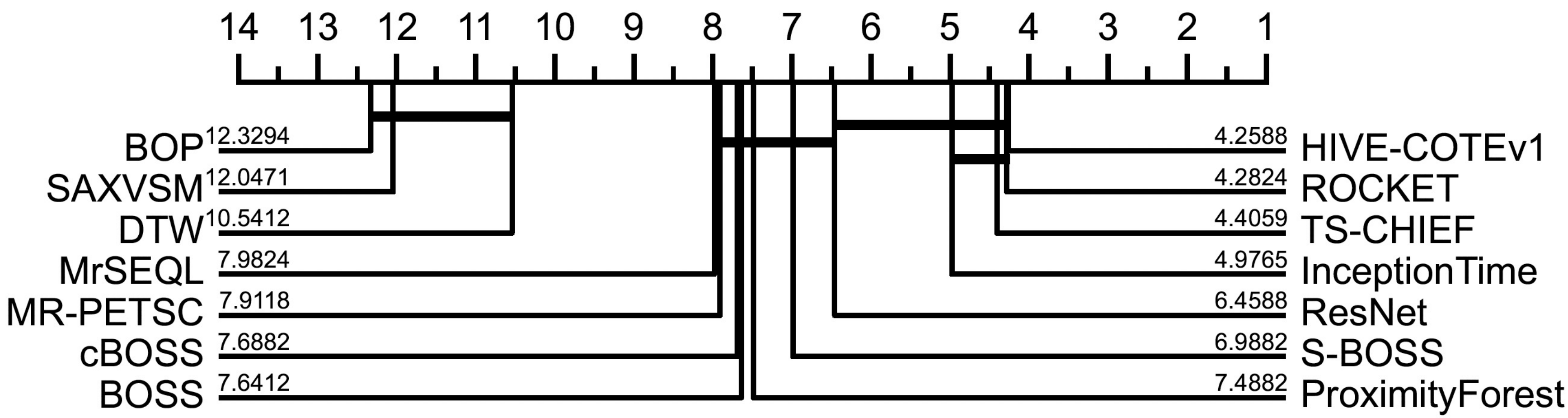
Other deep classifiers and ensembles

- **TapNet**: Time Series Attentional Prototype Network
- **ResNet** for time series classification
- **TS-CHIEF**: Time Series Combination of Heterogeneous and Integrated Embeddings Forest
- **HIVE-COTE**: Hierarchical Vote Collective of Transformation-based Ensembles
- **PETSC**: Pattern-Based Embedding for Time Series Classification
- **XEM**: An Explainable-by-Design Ensemble Method for Multivariate Time Series Classification

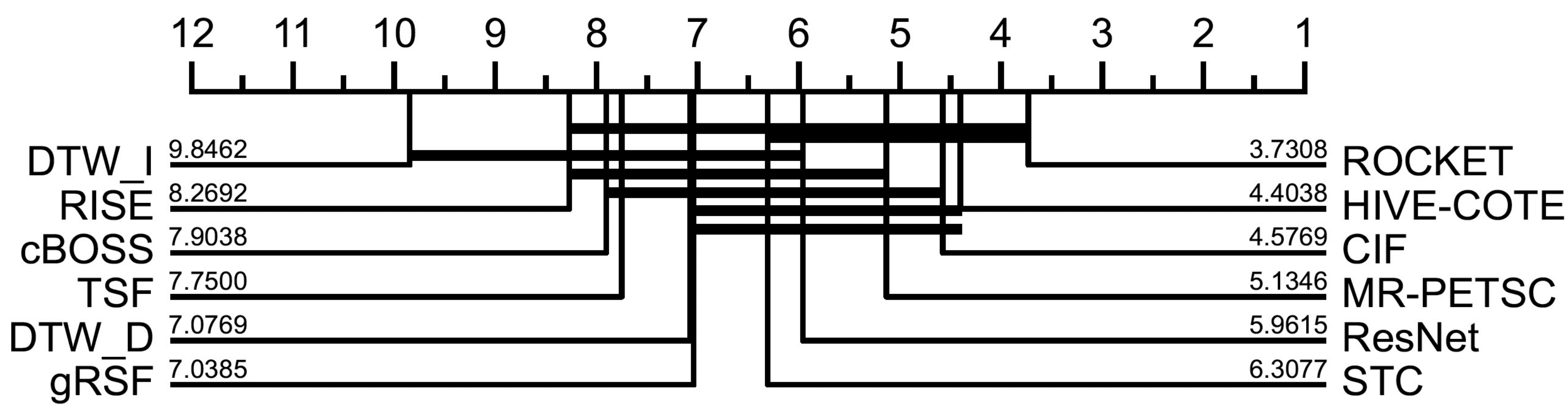
https://hal.science/hal-03515496/document

Overall winner?

Univariate time series classification



Multivariate time series classification



Agenda

Introduction

Time series classification

Explainable time series classification

Time series counterfactuals

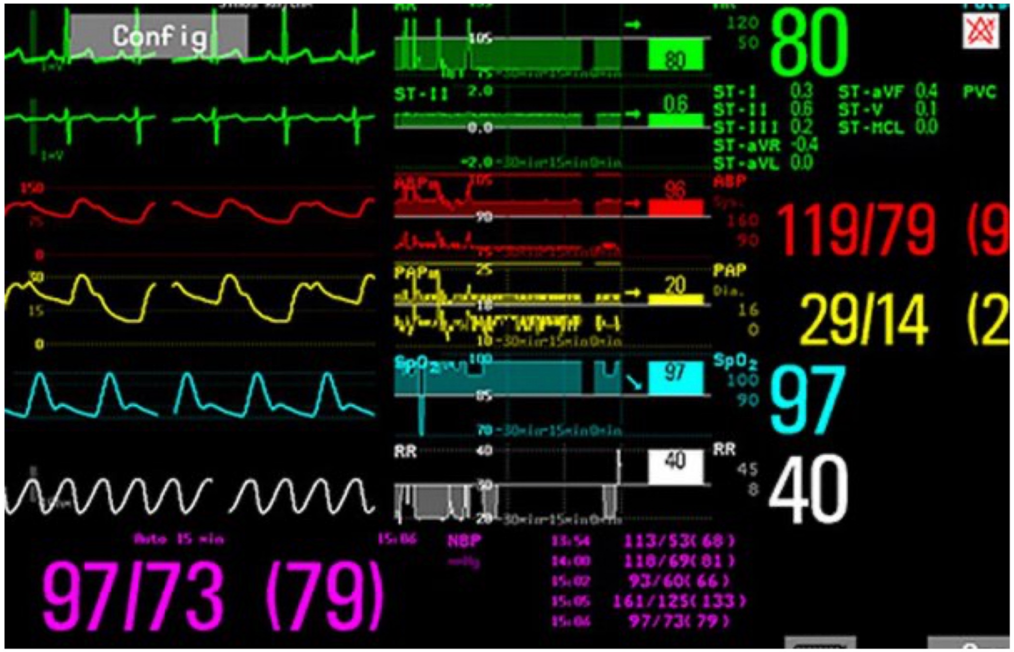
Challenges and future directions

Why explainability

- Interpretation/understanding of results
- Error discovery and management
- Bias avoidance
- Effectiveness improvement
- Trust

Proposition (J. Holmes 2023):
XAI-based systems need to start from modeling the underlying domain in order to obtain a true understanding of the context in which these systems will be used

Medical time series - in the ICU



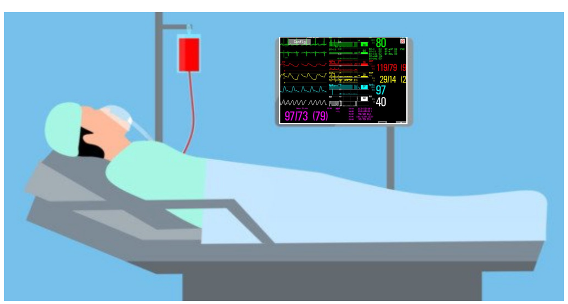
heart rate

systolic/diastolic blood pressure

pulmonary artery pressure

blood oxygen supply

respiration rate

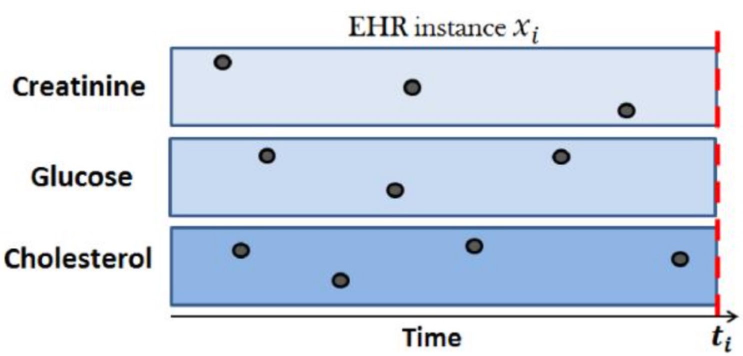


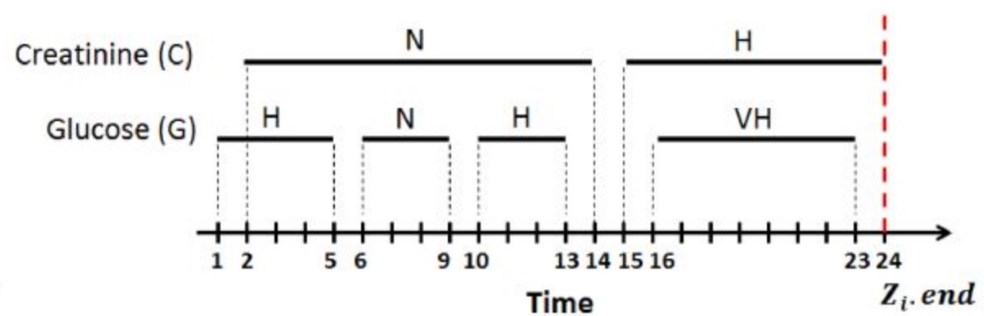
Over 100 variables are measured over time

Medical experts need to understand why…
…in order to be able to act timely

Temporal abstractions

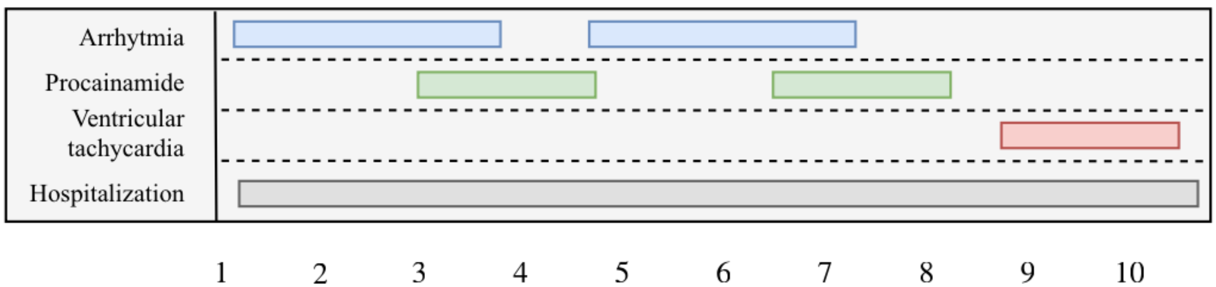
- Multiple temporal variables registered and evolving concurrently
- Each variable with multiple readings until a critical time point t_i, e.g., glucose, creatinine, cholesterol
- Class label: diagnosis/symptom detected at time t_i (event of interest)
- Main question: are all values over time really relevant?

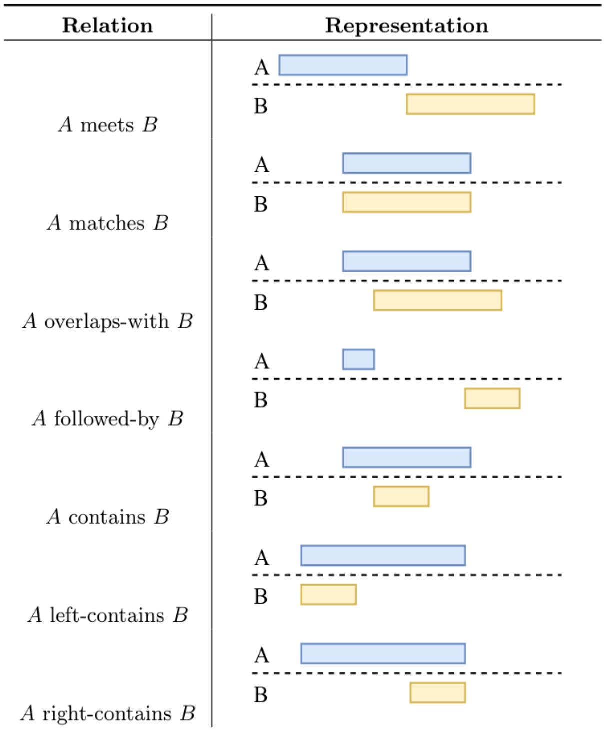




Temporal abstractions

- Trend abstraction:
  - e.g., decreasing, steady, increasing
- Value abstraction:
  - e.g., very low, low, normal, high, very high





Allen's temporal logic

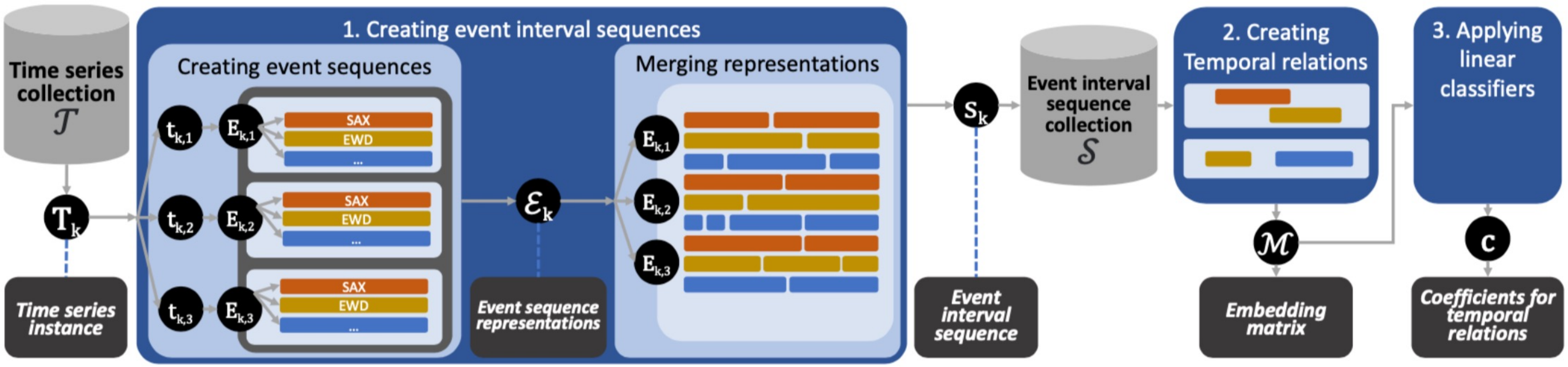
What is a temporal feature?

a sequence of ***"temporal relations"*** between two or more event intervals

What are the types of "temporal relations"?

Z-time [Lee et al. 2023]

- Employs temporal abstractions
- Builds temporal relations of event intervals to create interpretable features across multiple time series dimensions

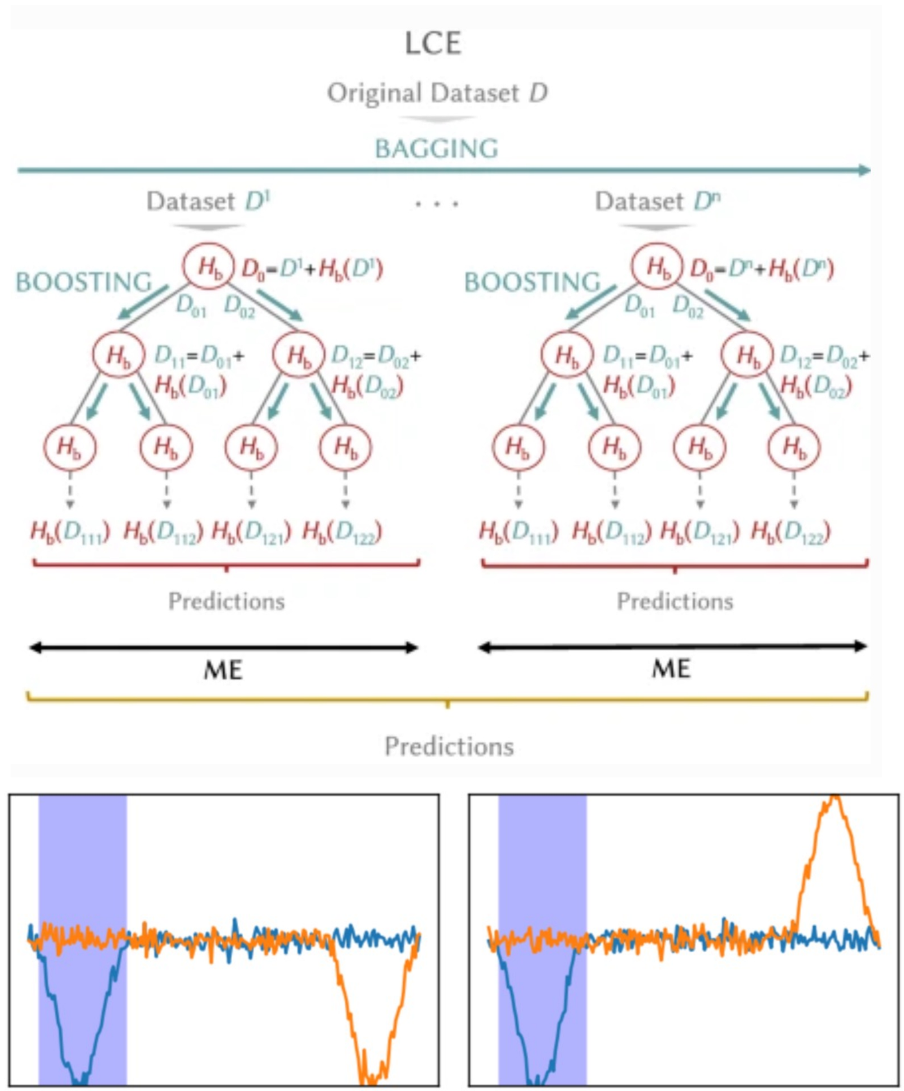


- Faster than the two interpretable competitors, XEM and MR-PETSC
- Handles missing data without applying interpolation

Z-Time: Efficient and Effective Interpretable Multivariate Time Series Classification, Lee et al. (session: time series II, 16:30-18:30)

XEM (Fauvel et al. 2022)

- Relies on an ensemble of eXtreme Gradient Boosting local cascade (LC) models
- The prediction is based on the subsequence that has the highest class probability, i.e., the subsequence on which LCE is the most confident
- XEM provides explainability-by-design through the identification of the time window used to classify the MTS

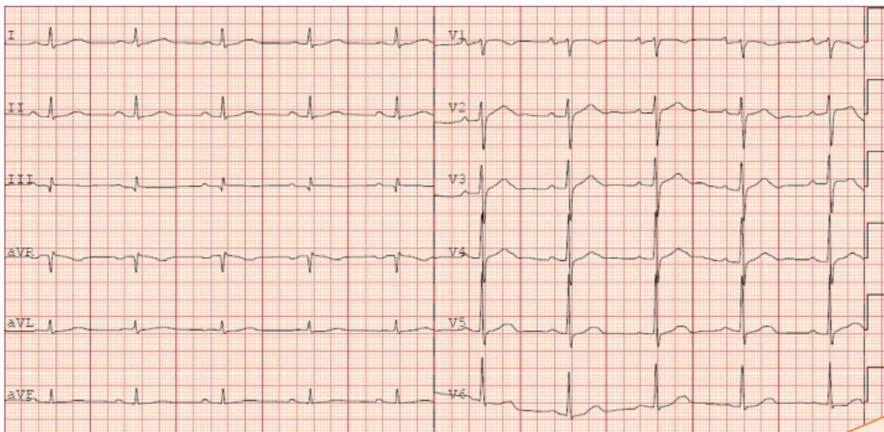


Agenda

- Introduction
- Time series classification
- Explainable time series classification
- Time series counterfactuals
- Challenges and future directions

Interpretable and actionable models

- It is desired to understand the predictions + outcomes without compromising predictive perf



Explaining: I can indicate the ECG segments and features that have affected my decision the most!

black box classifier

The patient will suffer a stroke in <u>2 days</u>!

Now what? Please tell me **why**?

Preventing: I can tell you what changes you need to make to the patient record, so that I can change my prediction ☺

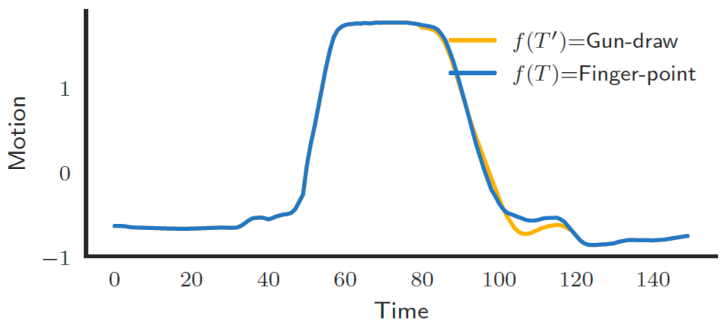
What is a counterfactual (CF)?

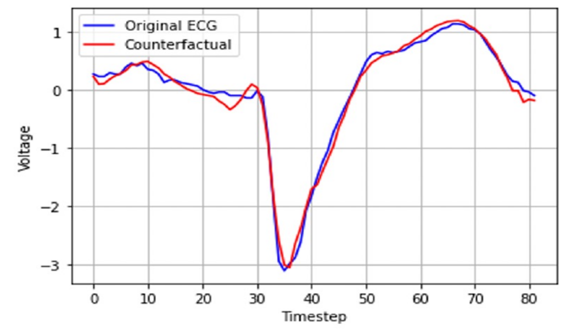
- Given a classifier f, an input instance x with predicted class label c, defined over a set of variables

- A counterfactual explanation **x’** can provide an answer to the following question:

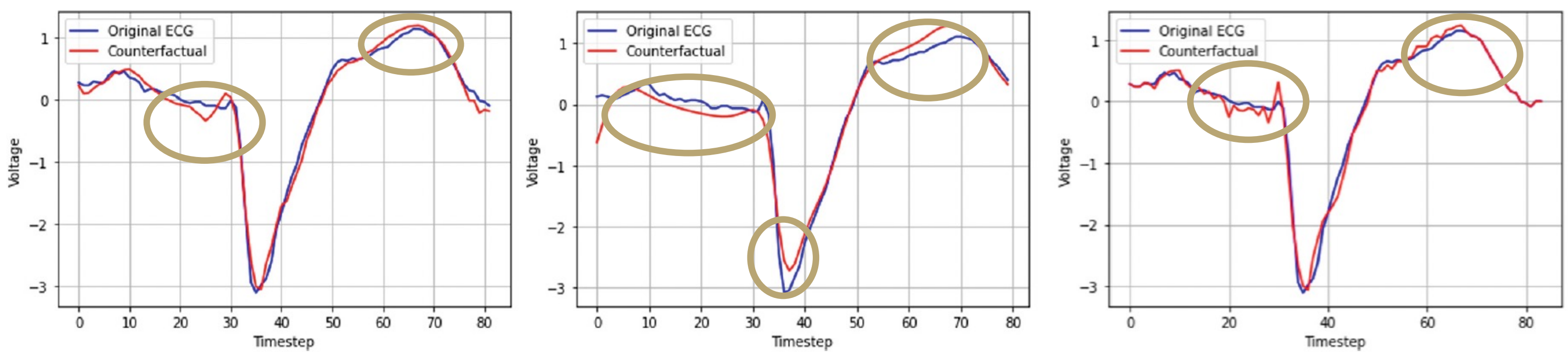
How should the configuration of the variables in x change
to obtain class label c’ instead of c ?

Time series counterfactuals



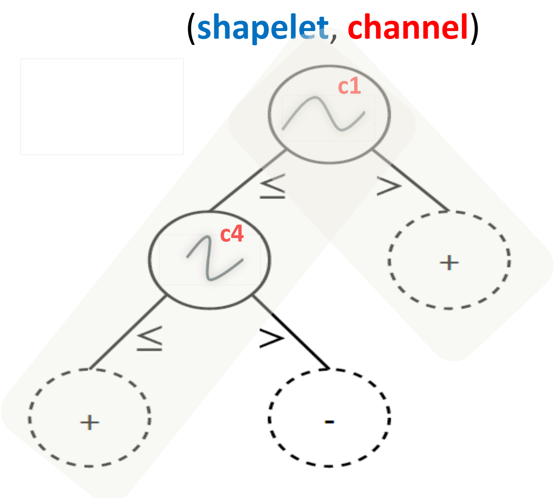


Goal: What is the minimum number of changes to apply to a time series T so that a given opaque classifier changes its prediction?



Time series counterfactuals for gRSF

- Focus on the trees that predict ***neg***
- For each tree ***T***, explore the positive paths, i.e., those that predict ***pos***
- Try to force those trees to predict ***pos*** by changing the shapelet features of ***T***

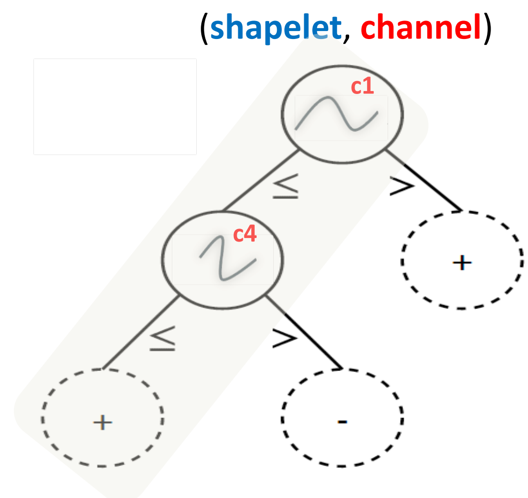


Given a non-leaf node (S^j_k, $\boldsymbol{\theta}^j_k$)

- Increase distance:
  - if S^j_k exists in ***T***, that is $d_s(\mathcal{S}_k^j, \mathcal{T}) \leq \theta_k^j$
  - and the current node condition demands otherwise
  - ✓ increase the distance of all matching instances of S^j_k, so that they all fall above the distance threshold $\boldsymbol{\theta}^j_k$

Time series counterfactuals for gRSF

- Focus on the trees that predict ***neg***
- For each tree ***T***, explore the positive paths, i.e., those that predict ***pos***
- Try to force those trees to predict ***pos*** by changing the shapelet features of ***T***

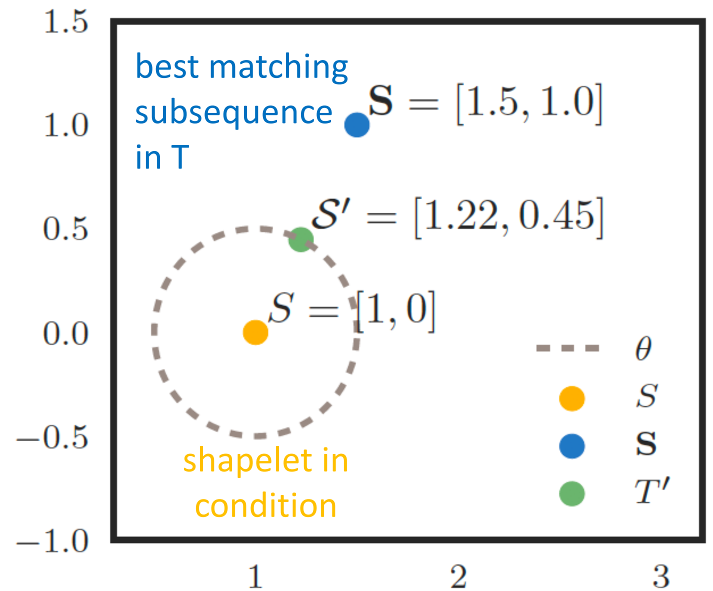


Given a non-leaf node (S^j_k, $\boldsymbol{\theta}^j_k$)

- Decrease distance:
  - if S^j_k does not exist in T, that is $d_s(\mathcal{S}_k^j, \mathcal{T}) > \theta_k^j$
  - and the current node condition demands otherwise
  - ✓ decrease the distance of the best matching instance of S^j_k, so that it falls below the distance threshold $\boldsymbol{\theta}^j_k$

How to transform the time series?

- Consider shapelet S as an m-dimensional point
- Define an m-sphere with S as its center and radius $\boldsymbol{\theta}$



- The transformed time series counterpart of S is given by the following equation:

$$\tau_S(\mathbf{S}, p_{ik}^j, \epsilon) = \mathcal{S}_k^j + \frac{\mathcal{S}_k^j - \mathbf{S}}{\|\mathcal{S}_k^j - \mathbf{S}\|_2}(\theta_k^j + (\epsilon\delta_{ik}^j))$$

Karlsson et al. Explainable time series tweaking via irreversible and reversible temporal transformations, ICDM 2018

Evaluation metrics?

proximity

Average cost of successful transformation, i.e.,

how costly is the transformation?

$$c_\mu(\tau, y') = \frac{1}{n}\sum_{i=1}^{n} c(\mathcal{T}_i, \tau(\mathcal{T}_i, y'))$$

sparsity

Compactness of transformation, i.e.,

how much of the time series is changed?

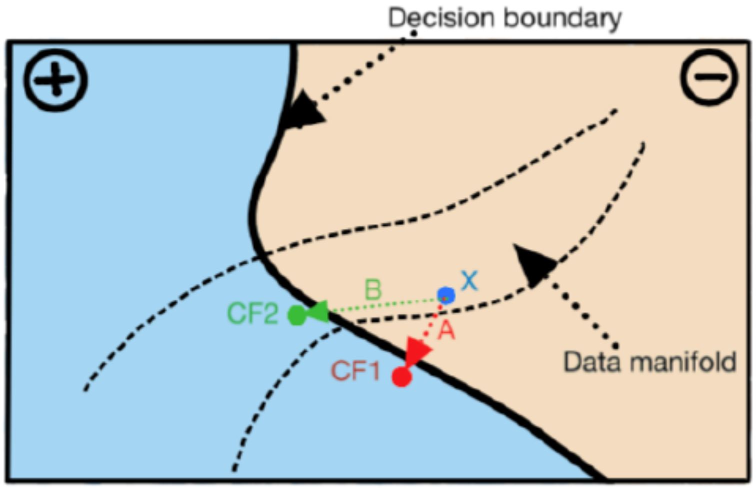
$$compact(\mathcal{T}, \mathcal{T}') = \frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{T}|} \mathit{diff}\,(T_i, T_i') \ ,$$

where

$$\mathit{diff}\,(T_i, T_i') = \begin{cases} 1, \text{ if } |T_i - T_i'| \leq e \\ 0, \text{ otherwise.} \end{cases}$$

Counterfactual quality

- It is not only sparsity and proximity that matter
- Counterfactuals should also be:
  - compliant with the original data distribution
  - should be expected to be observed

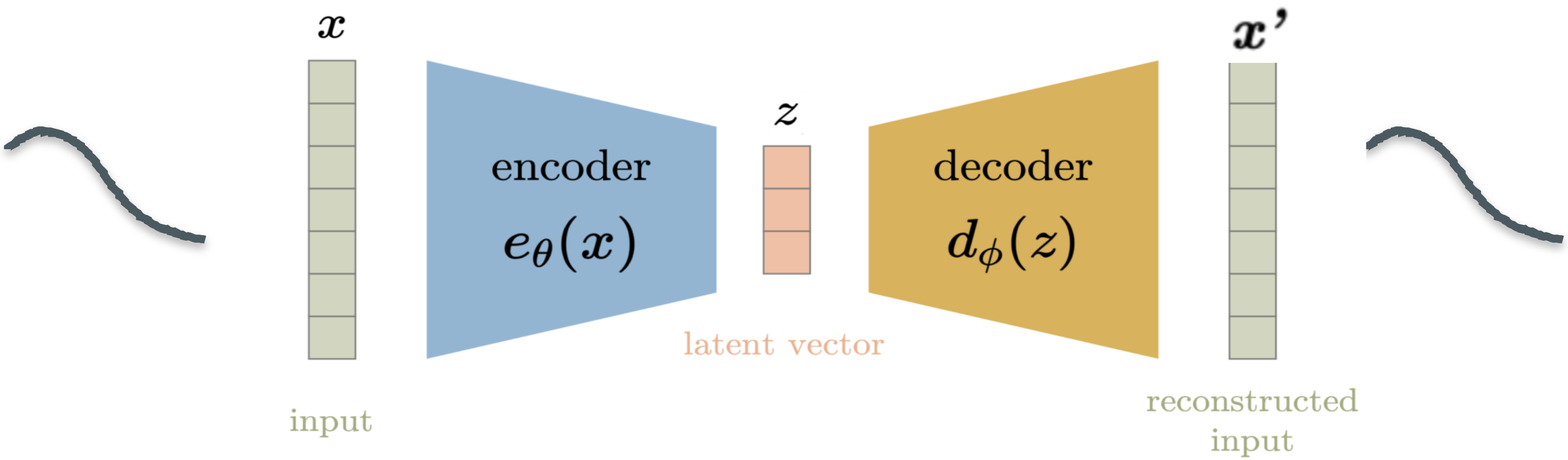


Several CF "goodness" measures:
- proximity
- validity
- sparsity
- faithfulness
- fairness
- …

- One direction: find a way to learn the data manifold / distribution per class

* Figure source: Verma, S., Dickerson, J., Hines, K.: Counterfactual Explanations for Machine Learning: A Review

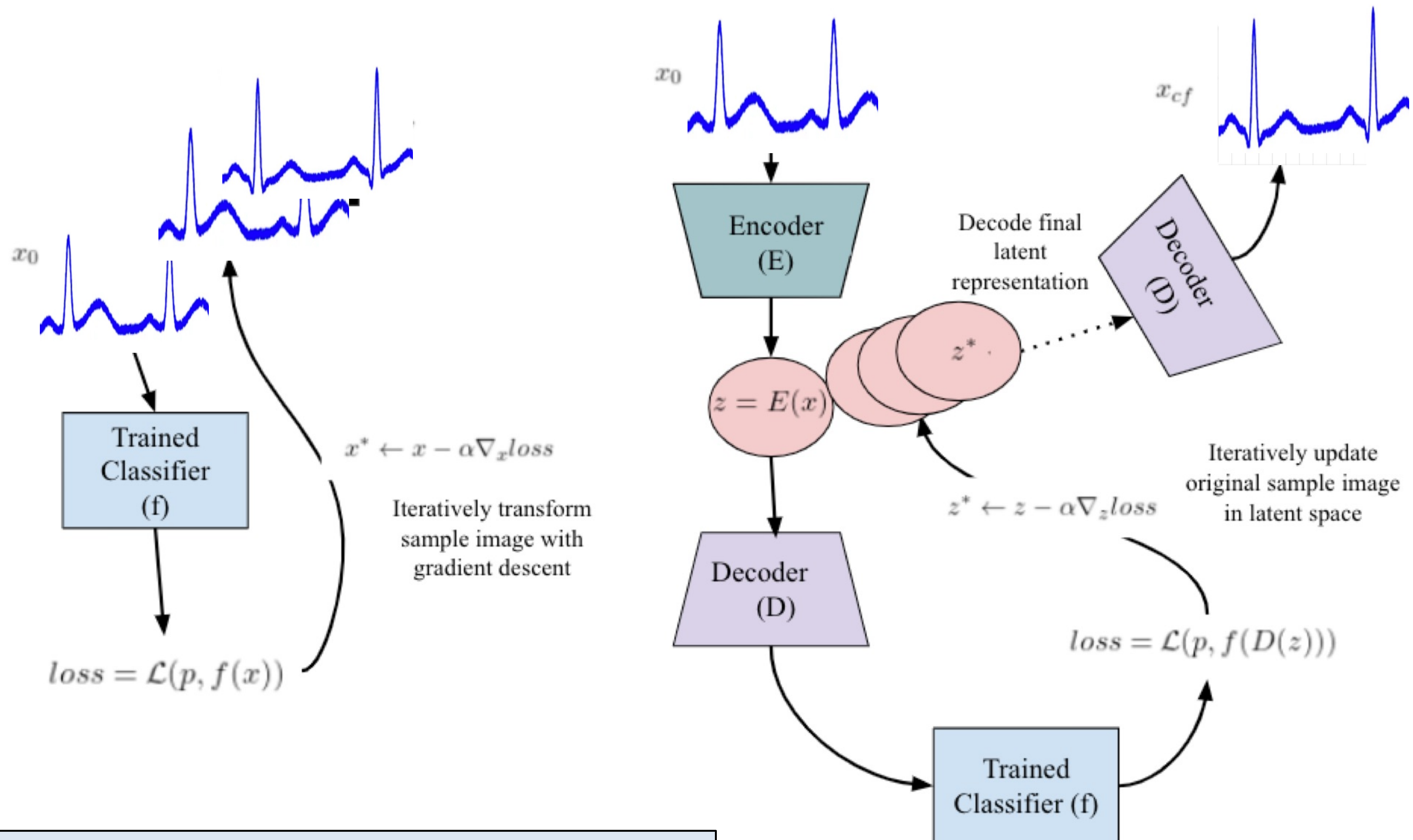
Autoencoders



$$loss = \|x - x'\|_2 = \|x - d_\phi(z)\|_2 = \|x - d_\phi(e_\theta(x))\|_2$$

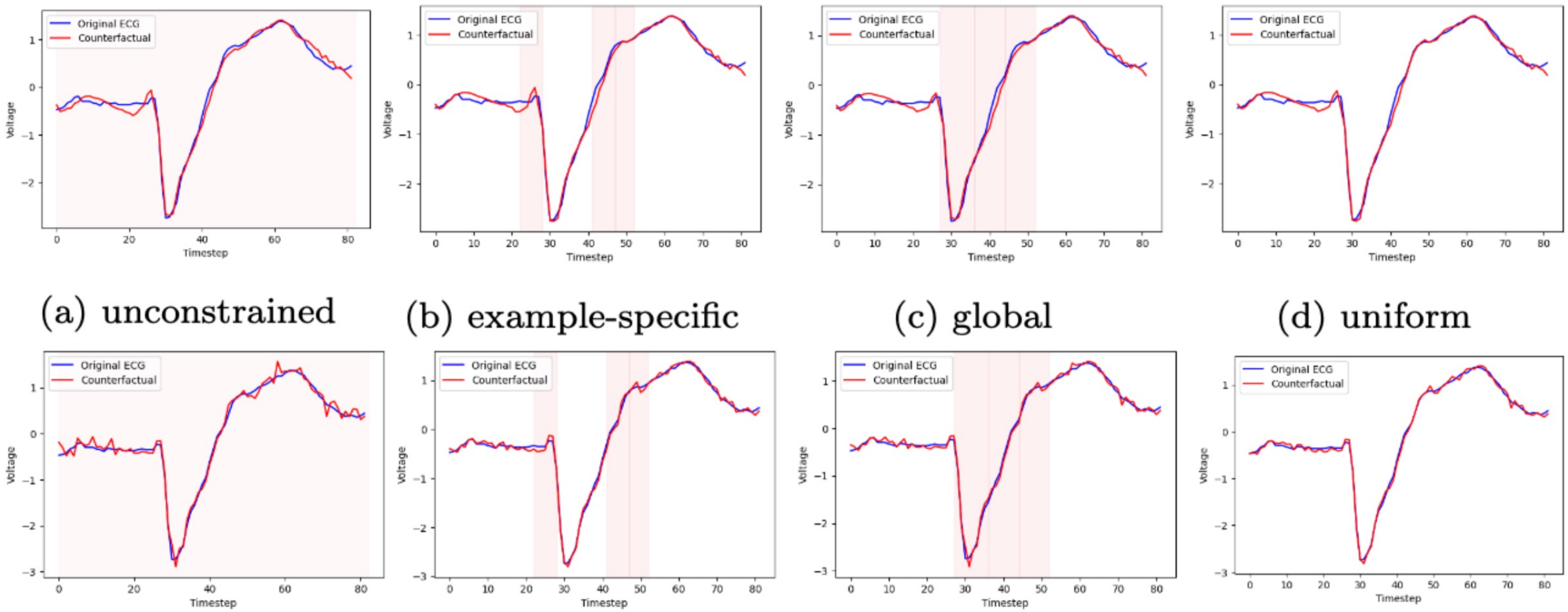
- Use an auto-encoder to find the generated counterfactual with the desired class (e.g., positive) outcome
- Perturb the encoded latent representation $z = e(x)$ through a gradient descent optimization approach iteratively to generate a new time series sample $x' = d(z)$ such that the output target $f(x') = ‘+’$

Latent space CFs



Balasubramanian et al. Latent-CF: A Simple Baseline for Reverse Counterfactual Explanations, Arxiv 2020

LatentCF for time series



(a) unconstrained (b) example-specific (c) global (d) uniform

(e) unconstrained (f) example-specific (g) global (h) uniform

Wang et al. Learning Time Series Counterfactuals via Latent Space Representations, Discovery Science 2022 and MACH (to Appear)

Agenda

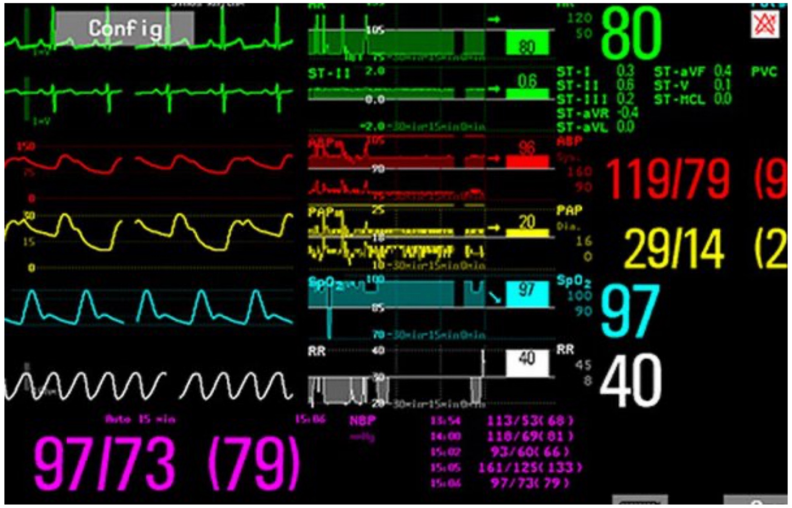
- Introduction
- Time series classification
- Explainable time series classification
- Time series counterfactuals
- Challenges and future directions

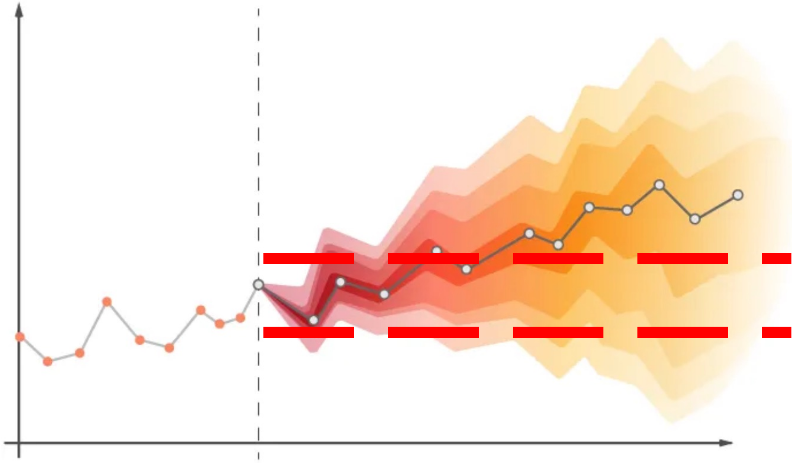
Challenges in XAI-TS

- Multimodal learning
- Sparsity in time series measurements
- Short time series
- Assessing explanations
- Actionable explanations
- Actionable time series forecasting

Counterfactuals for time series forecasting

- Monitor current patient vitals
- Forecast their progression
- Identify timely interventions
- Define forecasting counterfactuals



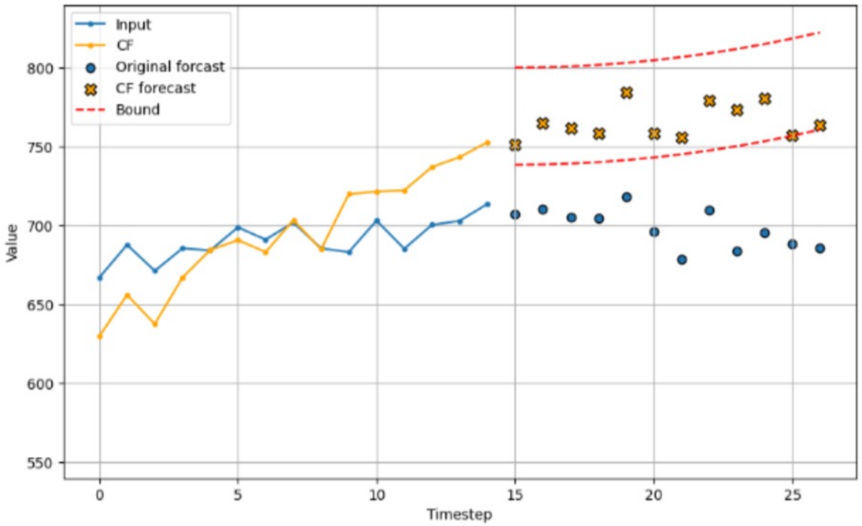


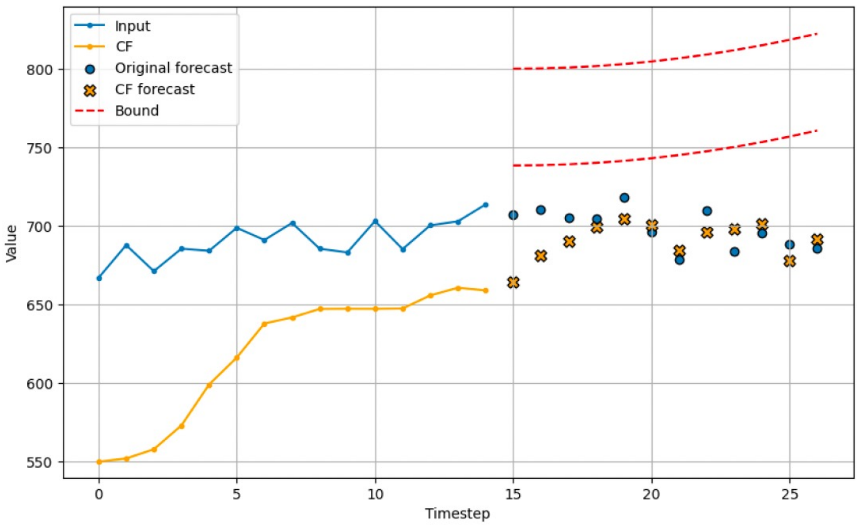
Maintain the prediction within a constaint band

Early interventions to prevent "violating" the band

Wang et al. Counterfactuals for time series forecasting, ICDM 2023

Counterfactuals for time series forecasting





Challenges:

- Defining proper constraints
- Defining proper and timely interventions
- Integrating external variables
- Multivariate forecasting

Wang et al. Counterfactuals for time series forecasting, ICDM 2023

Take-home messages

- **Understand** the domain you are explaining
- Consult with **domain experts**
- Ensure that your explanations are **compliant** with the **data domain**
- **Multivariate** and **multimodal** data is *challenging* but can be *critical*

Thank you!

Stockholms universitet

Panagiotis Papapetrou
Professor, Stockholm University
panagiotis@dsv.su.se