

Time Series: Defining a Search Engine

Philipp Beer

October 3, 2021

1 Thesis: Time Series Search Engine

1.1 Purpose

The purpose of this thesis is to explore the possibility of creating a time series series search engine.

1.2 Introduction

Time series is often described as "anything that is observed sequentially over time" which usually are observed at regular intervals of time [1]. They can be described as collection of observations that are considered together in chronological order rather than as individual values or a multiset of values. Their representation can be described as ordered pairs: $S = (s_1, s_2, \dots, s_n)$ where $s_n = (t_n, v_n)$. t_n can be a date, timestamp or any other element that defines order. v_1 represents the observation at that position in the time series.

Time series are utilized to analyze and gain insight from historic events/patterns with respect to the observational variable(s) and their interactions. A second area of application is forecasting. Here time series are utilized to predict the observations that occur in future under the assumption that the historic information can provide insight into the behavior of the observed variables.

Fu in their work [2] categorized time series research into (1) representation, (2) indexing, (3) similarity measure, (4) segmentation, (5) visualization and (6) mining. Research in these different fields started taking off in the second half of the 20th century. For example in [3] the authors worked on questions of representation via sampling the sampling of time series in 1969. All these different research areas always have to deal with the challenges that inhibit time series data. Generally datasets in this domain are large. Through this time series data incorporates the similar obstacles as high dimensional data, namely the "curse of dimensionality" [4] and requiring large computational efforts in order to generate insights. And as will be discussed in 1.2.1 there are applications fields where vast amounts of time series are generated and a comparison between them is required.

In this thesis we will focus on creating a algorithm allowing the fast and meaningful comparison of an input time series or template against a vast array time series. Within the research many different areas and approaches have been attempted (at list here). However, there is a tendency to apply the simplest methods possible to achieve the desired results. For time series those are mainly Euclidean Distance and Dynamic Time Warping. While those methods will be explored in section 1.3 it can be said that those methods are simple, easy to understand and produce mostly reliable results in their respective application domains. Good performance is not their strong suit. Therefore, our approach is targeted to achieving comparable capability in identifying similar series, while achieving it with a significant reduction in computational complexity.

1.2.1 Applications

Time series are encountered everywhere. Any metric that is captured over time can be utilized as time series. Granularity can be used as descriptor for the sampling rate of a series or more general how often measurements for a particular metric are taken. This granularity has a tendency to increase as well. As example consumer electronics that capture health and fitness data can be mentioned. Or sensors which are utilized in the automotive industry or heavy machinery where they are employed to capture information for predictive maintenance applications.

In the financial industry time series are a very fundamental component of decision making, like the development of stock prices over time or financial metrics of interest. The same is true for macro economic information or metrics concerning social structures in society, etc.

In the medical field time series are also ubiquitous. Whether they relate to patient data like blood pressure. The bio statistics field utilizes electro-graphical data like electrocardiography, electroencephalography and many others. In more aggregate medical analysis like toxicology analysis of drug treatments for vaccine approvals they are utilized and in many forms of risk management, for example, population level obesity levels.

In engineering fields the utilization is often times similar to the above but it also requires that information that is captured in time series is transferred between locations in an efficient manner. For example voice calls are required to be transferred between the participants in a fast manner and with minimized levels of noise in the data. Another interesting industrial example in the biomedical technology field is Neuralink which aims to implement a brain-machine-interface (BMI) utilizing mobile hardware like smartphones as the basis for its computation. Here a large amount of time series data is generated which requires quick processing to generate real-time information. Musk describes a recording system with 3072 electrodes generating time series data [5] that is used to capture the brain information and visualized in real-time [6].

Time series data is paramount to a wide variety of areas, relating to many different fields. Looking at the trajectory it seems likely that going forward more time series data on a higher granularity will be generated. This in turn increases the need to be able to process, analyze, compare and respond to the data with methods that are faster than today's standard options.

1.2.2 Organization of this thesis

The rest of this thesis is organized as follows:

1.2.3 TODO to be integrated

- refer to previous work on measures of similarity and outcome
- measure of similarity required
- challenges with time series (domains, granularity, length, outliers)
- area of signal processing interesting methods

1.3 Related work

Related work addressing the idea of time series search engine focuses on the system architecture and the data processing and pipelining aspect of this such an architecture [7]. However, in 2000 Keogh and Pazzani also applied a dimensionality reduction technique (Piecewise Constant Approximation) to execute fast search similarity search in large time series databases. Other papers address domain specific questions like the introduction of a "Time-series Subimage Search Engine for archived astronomical data" [9].

In order to be able to describe the closeness of time series or multiple time series to each a measure for similarity is required. In the literature various general measures and corresponding computation methods can be found. Wang et al. reviewed time series measures and categorized the similarity measures into 4 categories: (1) lock-step measures, (2) elastic measures, (3) threshold-based measures, and (4) pattern-based measures. Zhang et al. classify similarity measures in the categories: (1) time-rigid methods (Euclidean Distance), (2) time-flexible measures (dynamic time-warping), (3) feature-based measures (Fourier coefficients), and (4) model-based methods (auto-regression and moving average model) [11]. Lock-step measures include the L_p -norms (Manhattan and Euclidean Distance) as well as Dissimilarity Measure (DISSIM). Elastic measures include metrics like Dynamic Time Warping (DTW) and edit distance based measures like Longest Common Subsequence (LCSS), Edit Sequence on Real Sequence (EDR), Swale and Edit Distance with Real Penalty. An example for threshold-based measures are threshold query based similarity search (TQuEST). And Spatial Assembling Distance (SpADe) is an example for pattern-based measures. In another paper, Gharghabi et al. classify the space of similarity measures by the the most common measures into: (1) Euclidean Distance, (2) Dynamic Time Warping (DTW), (3) Least Common Subsequence (LCSS), and (4) K-Shape.

Dynamic Time Warping (DTW) is an elastic measure. It has been introduced by Berndt and Clifford in 1994 and its key advantage is the fact that comparison is applied on a one-to-many-basis allowing the comparison of regions from one series to regions of the other time series. This gives it the capability to warp peaks or valleys between different time steps of the two series as the resulting distance metric. As will be shown in section 1.4.2 this comes at the price of time complexity which renders it effectively useless in practice when applied to large scale data sets.

Other attempts are also made in introducing new distance metrics. Gharghabi et al. introduced a new metric called MPdist (Matrix Profile Distance) which is more robust than Euclidean Distance (ED) - more details can be found in section 1.4.1 - and Dynamic Time Warping (DTW) - more details can be found in section 1.4.2 - and computationally preferable. Interestingly, due to the use of subsequences in the comparison of two time series its time complexity ranges from $\mathcal{O}(n^2)$ in the worst case, to $\mathcal{O}(n)$ in the best case and with this can provide a significant advantage of prevalent methods like ED or DTW.

The other research area of interest for our task is time series representation. It concerns itself with the optimal combination of reduction of the data dimensionality but adequate capture of its particular properties. With these methods feats like minimizing noise, managing outliers can be achieved. For many activities this is also the basis for the reduction of time complexity in the resulting algorithms that analyze and compare the time series.

According to Li et al. the following methods are common methods for this task: (1) Discrete Fourier Transformation (DFT), (2) Singular Value Decomposition (SVD), (3) Discrete Wavelet Transformation (DWT), (4) Piecewise Aggregate Approximation (PAA), (5) Adaptive Piecewise Constant Approximation (APCA), (6) Chebyshev polynomials (CHEB), (7) Symbolic Aggregate approXima-

tion, and others [14]. In their paper, Pang et al. mention (1) Singular Value Decomposition (SVD), (2) Frequency-Domain transformation, (3) Piecewise Linear Representation (PLR), (4) model-based method, and (5) symbolic representation.

1.3.1 Dimensionality Reduction related to Singular Value Decomposition

Singular Value Decomposition is a fundamental matrix factorization technique with a plethora of applications and use cases. Its value comes from the capability of generating low rank approximations of data matrices that allow to represent the matrix values via the unitary matrices $\mathbf{U} \in \mathbb{C}^{n \times n}$ and $\mathbf{V} \in \mathbb{C}^{m \times m}$. The columns in \mathbf{U} and \mathbf{V} are orthonormal. The remaining matrix $\mathbf{\Sigma} \in \mathbb{R}^{n \times m}$, is a diagonal matrix with non-negative entries.

The power of the SVD is its ability to provide a low-dimensional approximation to high-dimensional data [16]. High dimensional data is often determined by a few dominant patterns which can be described by a low-dimensional attractor. Therefore, a prime application for the SVD is dimensionality reduction. It is complementary to the Fast Fourier Transform (FFT) which lays at the core of this work. Brunton and Kutz describe it as the generalization of the FFT.

Principal Component Analysis (PCA) is a very common application of the SVD. It was developed by Pearson in 1901. The main idea of PCA is to apply the SVD to a dataset centered around zero and subsequently computing the covariance of the centered dataset. Through the computation of the eigenvalues and their identifying the largest values the most important principal components are identified. Those are responsible for the largest variance in the dataset. And similar to the SVD their ranking and subsequent filtering can be used to focus on the most important components that allow to recreate majority of the of the variance in the dataset.

The Fast Fourier Transform (FFT) is based upon the Fourier Transform introduced by Joseph Fourier in early 19th century to analyze and analytically represent heat transfer in solid objects [18]. This transform is a fundamental component of modern computing and science in general. It has transformed how technology can be used in the in 20th century in areas such as image and audio compression and transfer. The concept will be introduced in more detail in section 1.4.3. Its core idea is to represent the data to be transformed as the coefficients of a basis of sine and cosine eigenfunctions. It is similar to the principles of the SVD with the notable difference that the basis are an infinite sum of sine and cosine functions. The ability to reduce to the transformed data to few key components is the same as in SVD and PCA.

1.3.2 Symbolic Aggregate approXimation

A dimensionality reduction technique that does not built on SVD and is geared directly towards time series is the Symbolic Aggregate approXimation (SAX) algorithm. Its core idea is to transform a time series into a set of strings via piecewise aggregate approximation (PAA) and a conversion of the results via a lookup table. Starting with PAA the reduction of a time series T of length n in vector $\bar{S} = \bar{s}_1, \bar{s}_2, \dots, \bar{s}_w$ of length w where $w < n$, can be achieved through the following computation:

$$\bar{s}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} s_j \quad (1)$$

This simply computes the mean of each of sub sequences determined through parameter w . An example from the M4 dataset can be seen in figure 1. For its application in SAX the time series

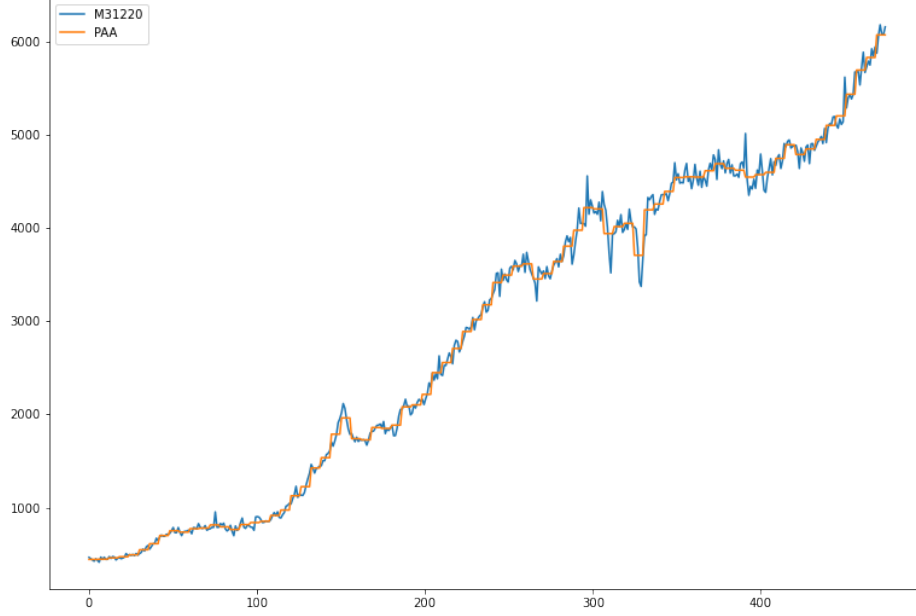


Figure 1: Piecewise Aggregate Approximation - M4 example: M31220 (window size - 6)

need to be standardized or mean normalized via:

$$z = \frac{x - \mu}{\sigma} \quad (2)$$

1.4 Relevant Concepts

1.4.1 Euclidean Distance

Euclidean Distance is the most widely used distance metric in the research of time series. It is either used as a metric on its own or as a metric used inside other methods to compute distances, for example, computation of distance of subsections of the data ([19]) or to compute the distance between various points of two time series (see section 1.4.2). Having two time series $S = \{s_1, s_2, \dots, s_n\}$ and $Q = \{q_1, q_2, \dots, q_n\}$ both of length n the Euclidean distance can be computed as:

$$D(S, Q) = \sqrt{\sum_{i=1}^n (S_i, Q_i)^2} \quad (3)$$

It is a measure that is easy to compute and comprehend and gives intuitive input for the distance of two time series. From the standpoint of time complexity the algorithm is applicable also to larger datasets with $\mathcal{O}(n)$. Its simplicity also creates some limitations. For example, to compute the euclidean distance between two series their length needs to be the same. Furthermore, it can be easily impacted in its results by the presence of outliers or increased levels of noise. It is not elastic with respect to the warping of information between two series in which effects that could indicate similarity happen even at slightly disparate steps.

Despite its shortcomings it is a prominent metric and widely used for distance calculations for short comings. Some of its limitations are addressed by more sophisticated metrics that utilize its computation as component.

1.4.2 Dynamic Time Warping

Berndt and Clifford introduced Dynamic Time Warping in 1994 finding the minimal alignment between two time series computed through a cost matrix and identifying the minimized path through the matrix starting from the final elements of each time series. This warps the points in time between the different series as shown in figure 2.

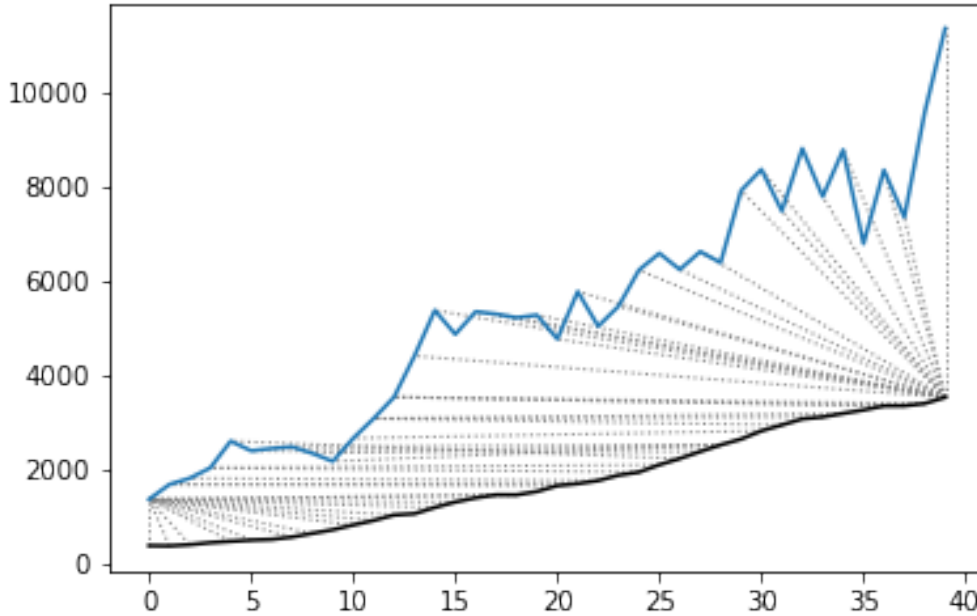


Figure 2: Dynamic Time Warping - M4 Example: Y5683 and Y5376

Two series $S = \{s_1, s_2, \dots, s_n\}$ of length n and $Q = \{q_1, q_2, \dots, q_m\}$ of length m are considered. For the series a n -by- m cost matrix M is constructed. Each element in the matrix represents the respective i^{th} and j^{th} element of each of the two series which contains the distance of those to points:

$$m_{ij} = D(s_i, q_j) \quad (4)$$

where often time euclidean distance is used as distance function $D(s_i, q_j) = (s_i - q_j)^2$. From the matrix a warping path P is chosen, $P = p_1, p_2, \dots, p_k, \dots, p_K$ where:

$$\max(m, n) \leq k < m + n - 1 \quad (5)$$

The warping path is constrained with bound with the following condition $p_1 = (1, 1)$ and $p_K = (m, n)$. That means that both first elements of each series, as well as, the last element of each series are bound to each other in the computation. The warping path also is continuous. This means that

from each chosen element p_k only the neighboring elements to the left, right and diagonally can be chosen for the continuation of the path: $p_k = (a, b)$ and $p_{k-1} = (a', b')$ with $a - a' \leq 1$ and $b - b' \leq 1$. The path elements p_k are also monotonous, meaning that $a - a' \geq 0$ and $b - b' \geq 0$. From the resulting matrix considering the mentioned constraints a cumulative distance $\gamma(i, j)$ is computed recursively:

$$\gamma(i, j) = D(s_i, q_j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \quad (6)$$

Therefore, the path can be obtained by the following definition:

$$DTW(S, Q) = \min\left\{\sum_{k=1}^P \delta(\omega_k)\right\} \quad (7)$$

Figure 3 provides an example for a warping path result.

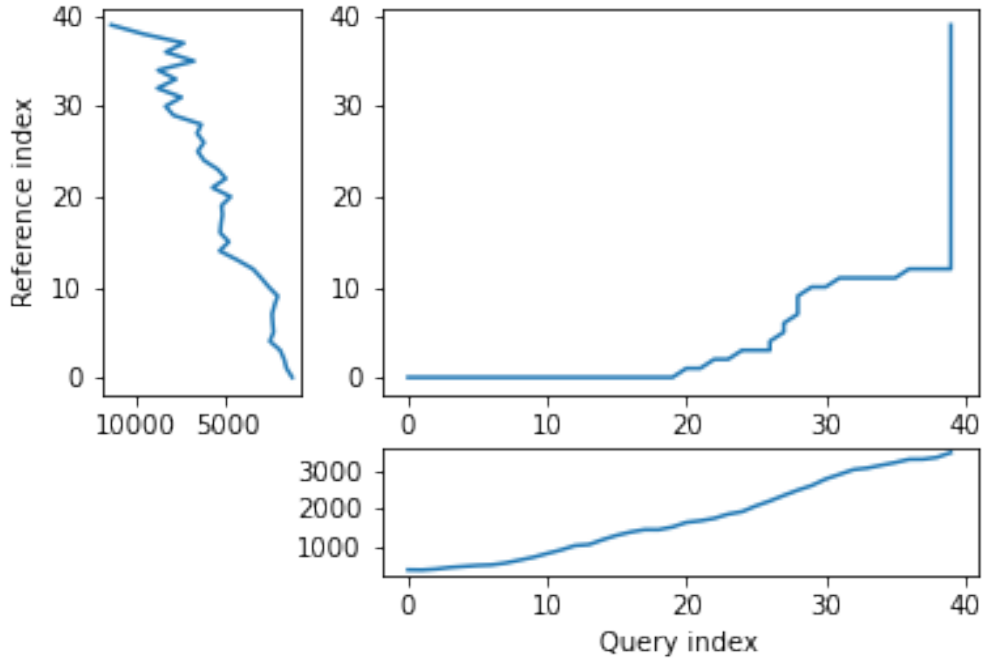


Figure 3: Warping path example - M4 data: Y5683 and Y5376

The challenge with the application of DTW is the time complexity of the algorithm $\mathcal{O}(m * n)$ due to the fact that the distance computation needs to be executed for each element of each series. Various methods for speed improvements have been introduced. The favorite principle was described by Ratanamahatana and Keogh. They introduced an adjustment window condition that where it is assumed that the optimal path does not drift very far from the diagonal of the cost matrix [20]. However, this does not change the fundamental nature of the algorithm and computing DTW for multiple time series against a database of time series will require days of computation time even on modern computer architectures.

In favor of DTW needs to be stated, that it is flexible with regards to the series used. The compared time series do not require to have the same length and can still be compared. This is a property

that is not available with Euclidean Distance. However, the user also needs to be aware of outliers in either data set which can lead to a clustering of the warping path or pathological matches around those extreme points in the series.

Therefore in practice, Dynamic Time Warping is not a method suitable for comparing a single time series against a large array of series when speed is an important criterion as well as the handling of outliers in the dataset.

- Similarity through decomposition
 - introduce time series decomposition (reference in [1])
 - trend and seasonality (mention assumptions about period)

1.4.3 Fast Fourier Transform

fast fourier transform

1.5 Time series representation

1.5.1 Challenges when building a time series

- length of series
- trend
- seasonality
- time complexity -> issue because of data size
- granularity or sampling rates
- noise
- data quality
- similarity is task dependent (level)
- usual need for preprocessing the time series data (denoising, detrending, amplitude scaling)
 - > any pre-processing does modify the series

1.5.2 Data Analysis

what does M4 data look like

1.5.3 Challenges

- How many frequencies to compare?
- priorities of frequencies (power spectrum)
- different length of time series (leading to different frequencies) - ranges solved with logs

1.6 Methodology

1.6.1 Used Data

The research in time series has been numerous and focused on various properties of them as well as finding methods to accurately predict them. Aside of forecasting all researched areas are measures of similarity and retrieval of time series.

- Forecasting In the arena of forecasting the M-competition organized by Prof. Makridakis played a big role in the development of forecasting methods shortly after their inception in 1979.

One of the aspects that has been correct up until the 5th installment of the M-competition is that statistical methods in forecasting have outperformed more complex machine learning methods. So learning algorithms did not benefit sufficiently from learning from multiple series to generate more accurate point predictions and prediction intervals compared to the statistics-based alternatives.

One interesting question in this area is whether clustering of time series that have similar properties and training algorithms per cluster of "similar" series can help simplify the learning process for machine learning methods and in consequence improve their performance in future competitions.

However, expressing similarity for time series is a challenging questions with respect to which metrics to utilize, time complexity as well as limiting assumptions that need to be made for time series.

1.6.2 Main contribution of the thesis

- transformation into Fourier-space
- transfer frequencies into frequency range band with increasing range width (using log scale)
- computation of frequency energy levels (sort and keep top 5) -> ask Prof. how to name this parameter
- conversion of ordered frequencies into frequency range band
- for each series to compare -> compare whether the frequency matches on the ordered positions -> provide exponential value per position -> match on more powerful frequencies is valued higher

1.6.3 additional computations

- utilization of FFT utilizes only frequency space (future work should consider comparison of energy levels per frequency)
- additional simple statistics computed (mean, std, quantiles)
- ts decomposition for trend estimation (requires parameter for period) -> then best line fit for slope of the time series
- computation of deltas for each series to search with statistics and slope of all other time series (review computational complexlity)

- ranking of matching series based highest frequency range match and ONE statistic

1.6.4 Preprocessing

- M4 data wide format vs. long format

1.6.5 Parallelization

- computation times
- scalability
- Samples for results only (stratification vs. non-stratification)
- Threads vs. Processes

1.6.6 Technology (check with Prof. if required)

R vs. Python vs. Mathematica, Matlab

1.6.7

- load
- transform to FFT vector space
- compare most important frequencies
- compare candidates
- select winner (which criteria)

1.7 Exploratory Data Study

- what do results look like

1.8 Formal Evaluation

- (maybe) improvement in forecasting approach
- find dataset with ground truth and compare DTW to this approach
- Distance metrics
- time complexity

1.9 Conclusion & future work

1.9.1 Successes

1.9.2 Failures

1.9.3 Flaws

- final computation

1.9.4 What is missing

- denoising of time series
- adjustment of number of frequencies used
-

1.10 Results & Discussion

1.11 References

- [1] R.J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2014. ISBN: 9780987507105. URL: <https://books.google.de/books?id=gDuRBAAQBAJ>.
- [2] Tak-chung Fu. “A review on time series data mining”. In: *Engineering Applications of Artificial Intelligence* 24.1 (Feb. 2011), pp. 164–181. ISSN: 0952-1976. DOI: 10.1016/j.engappai.2010.09.007. URL: <http://dx.doi.org/10.1016/j.engappai.2010.09.007>.
- [3] K.J. Åström. “On the choice of sampling rates in parametric identification of time series”. In: *Information Sciences* 1.3 (July 1969), pp. 273–278. ISSN: 0020-0255. DOI: 10.1016/S0020-0255(69)80013-7. URL: [http://dx.doi.org/10.1016/S0020-0255\(69\)80013-7](http://dx.doi.org/10.1016/S0020-0255(69)80013-7).
- [4] Yongqiang Tang, Yuan Xie, Xuebing Yang, et al. “Tensor Multi-Elastic Kernel Self-Paced Learning for Time Series Clustering”. In: *IEEE Transactions on Knowledge and Data Engineering* (2019), pp. 1–1. ISSN: 2326-3865. DOI: 10.1109/tkde.2019.2937027. URL: <http://dx.doi.org/10.1109/tkde.2019.2937027>.
- [5] Elon Musk. “An Integrated Brain-Machine Interface Platform With Thousands of Channels”. In: *Journal of Medical Internet Research* 21.10 (Oct. 2019), e16194. ISSN: 1438-8871. DOI: 10.2196/16194. URL: <http://dx.doi.org/10.2196/16194>.
- [6] Joshua H Siegle, Aarón Cuevas López, Yogi A Patel, et al. “Open Ephys: an open-source, plugin-based platform for multichannel electrophysiology”. In: *Journal of Neural Engineering* 14.4 (June 2017), p. 045003. ISSN: 1741-2552. DOI: 10.1088/1741-2552/aa5eea. URL: <http://dx.doi.org/10.1088/1741-2552/aa5eea>.
- [7] Chuanlei Zhang, Ji’an Luo, Shanwen Zhang, et al. “Introduction to time series search engine systems”. In: *2012 International Conference on Systems and Informatics (ICSAI2012)* (May 2012). DOI: 10.1109/icsai.2012.6223532. URL: <http://dx.doi.org/10.1109/ICSAI.2012.6223532>.
- [8] Eamonn J. Keogh and Michael J. Pazzani. “A Simple Dimensionality Reduction Technique for Fast Similarity Search in Large Time Series Databases”. In: *Lecture Notes in Computer Science* (2000), pp. 122–133. ISSN: 0302-9743. DOI: 10.1007/3-540-45571-x_14. URL: http://dx.doi.org/10.1007/3-540-45571-x_14.
- [9] Q. Kang, C. Yu, Y. Zhang, et al. “Astro-TS3: Time-series Subimage Search Engine for archived astronomical data”. In: *Astronomy and Computing* 34 (Jan. 2021), p. 100428. ISSN: 2213-1337. DOI: 10.1016/j.ascom.2020.100428. URL: <http://dx.doi.org/10.1016/j.ascom.2020.100428>.
- [10] Xiaoyue Wang, Abdullah Mueen, Hui Ding, et al. “Experimental comparison of representation methods and distance measures for time series data”. In: *Data Mining and Knowledge Discovery* 26.2 (Feb. 2012), pp. 275–309. ISSN: 1573-756X. DOI: 10.1007/s10618-012-0250-5. URL: <http://dx.doi.org/10.1007/s10618-012-0250-5>.
- [11] Zheng Zhang, Ping Tang, and Thomas Corpetti. “Time Adaptive Optimal Transport: A Framework of Time Series Similarity Measure”. In: *IEEE Access* 8 (2020), pp. 149764–149774. ISSN:

- 2169-3536. DOI: 10.1109/access.2020.3016529. URL: <http://dx.doi.org/10.1109/ACCESS.2020.3016529>.
- [12] Shaghayegh Gharghabi, Shima Imani, Anthony Bagnall, et al. “An ultra-fast time series distance measure to allow data mining in more complex real-world deployments”. In: *Data Mining and Knowledge Discovery* 34.4 (May 2020), pp. 1104–1135. ISSN: 1573-756X. DOI: 10.1007/s10618-020-00695-8. URL: <http://dx.doi.org/10.1007/s10618-020-00695-8>.
 - [13] Donald J. Berndt and James Clifford. “Using Dynamic Time Warping to Find Patterns in Time Series”. In: KDD Workshop. 1994, pp. 359–370.
 - [14] Duo Li, Yifei Zhao, and Yan Li. “Time-Series Representation and Clustering Approaches for Sharing Bike Usage Mining”. In: *IEEE Access* 7 (2019), pp. 177856–177863. ISSN: 2169-3536. DOI: 10.1109/access.2019.2958378. URL: <http://dx.doi.org/10.1109/ACCESS.2019.2958378>.
 - [15] Jingyue Pang, Datong Liu, Yu Peng, et al. “Intelligent pattern analysis and anomaly detection of satellite telemetry series with improved time series representation”. In: *Journal of Intelligent & Fuzzy Systems* 34.6 (June 2018), pp. 3785–3798. ISSN: 10641246, 18758967. DOI: 10.3233/JIFS-169551. URL: <https://doi.org/10.3233/JIFS-169551>.
 - [16] S.L. Brunton and J.N. Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019. ISBN: 9781108422093. URL: <https://books.google.de/books?id=CYaEDwAAQBAJ>.
 - [17] Karl Pearson. “On Lines and Planes of Closest Fit to Systems of Points in Space”. In: *Phil. Mag.* 6.2 (1901), pp. 559–572.
 - [18] J.B.J. Fourier and Firmin père & fils Didot. *Théorie analytique de la chaleur, par M. Fourier*. chez Firmin Didot, pere et fils, 1822.
 - [19] Christos Faloutsos, M. Ranganathan, and Yannis Manolopoulos. “Fast subsequence matching in time-series databases”. In: *Proceedings of the 1994 ACM SIGMOD international conference on Management of data - SIGMOD '94* (1994). DOI: 10.1145/191839.191925. URL: <http://dx.doi.org/10.1145/191839.191925>.
 - [20] Chotirat Ann Ratanamahatana and Eamonn Keogh. “Making Time-series Classification More Accurate Using Learned Constraints”. In: *Proceedings of the 2004 SIAM International Conference on Data Mining* (Apr. 2004). DOI: 10.1137/1.9781611972740.2. URL: <http://dx.doi.org/10.1137/1.9781611972740.2>.