Time Series Chains: A Novel Tool for Time Series Data Mining

Yan Zhu[1], Makoto Imamura[2], Daniel Nikovski[3], Eamonn Keogh[1]
[1]University of California, Riverside
[2]Tokai University
[3]Mitsubishi Electric Research Laboratories
yzhu015@ucr.edu, imamura@tsc.u-tokai.ac.jp, nikovski@merl.com, eamonn@cs.ucr.edu

Abstract

Since their introduction over a decade ago, time series motifs have become a fundamental tool for time series analytics, finding diverse uses in dozens of domains. In this work we introduce Time Series Chains, which are related to, but distinct from, time series motifs. Informally, time series chains are a temporally ordered set of subsequence patterns, such that each pattern is similar to the pattern that preceded it, but the first and last patterns are arbitrarily dissimilar. In the discrete space, this is similar to extracting the text chain "hit, hot, dot, dog" from a paragraph. The first and last words have nothing in common, yet they are connected by a chain of words with a small mutual difference. Time Series Chains can capture the evolution of systems, and help predict the future. As such, they potentially have implications for prognostics. In this work, we introduce a robust definition of time series chains, and a scalable algorithm that allows us to discover them in massive datasets.

1 Introduction

Time series motifs are approximately repeating subsequences embedded in a longer time series. Since their formulation in 2002 [Patel *et al*., 2002] they have emerged as one of the most important primitives in time series data mining. Motif discovery has been used as a sub-routine in higher-level analytics, including classification, clustering, visualization [Hao *et al*., 2012], and rule-discovery [Shokoohi-Yekta *et al*., 2015]. Moreover, motif discovery has been applied to domains as diverse as severe weather prediction, robotics, medicine [Syed *et al*., 2010] and seismology [Zhu *et al*., 2016].

In retrospect, it is easy to see why time series motifs are so useful. If a pattern is repeated (or *conserved*), there must be a latent system that occasionally produces the conserved behavior. For example, this system may be an overcaffeinated heart, sporadically introducing a motif pattern containing an extra beat (Atrial Premature Contraction [Lovallo *et al*., 2004]), or the system may be an earthquake fault, infrequently producing highly repeated seismograph traces because the local geology produces unique wave reflection/refractions [Zhu *et al*., 2016]. Time series motifs are a commonly used technique to gain insight into such latent systems, in essence, they can be seen as "*generalizing the notion of a regulatory motif to operate robustly on non-genomic data*" [Syed *et al*., 2010].

In this work, we expand the notion of time series motifs to the new primitive of *time series chains* (or just *chains*). Time series chains may be informally considered motifs that evolve or drift in some direction over time. Figure 1 illustrates the difference between time series motifs and time series chains (we defer formal definitions until Section 2).

Both motifs and chains have the property that each subsequence is relatively close to its nearest neighbor. However, the motif set also has a relatively small diameter. In contrast, the set of points in a chain has a diameter that is much larger than the mean of each member's distance to its nearest neighbor. Moreover, the chain has the property of *directionality*. For example, in Figure 1.*left*, if a tenth member was added to the motif set, its location will also be somewhere near the platonic ideal, but independent of the previous subsequences. In contrast, in Figure 1.*right*, the location of the tenth member of the chain would be somewhere just left of item nine.

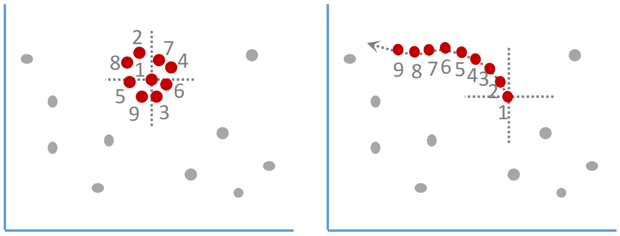

Figure 1. Visualizing time series subsequences as points in high-dimensional space. *left*) A time series motif can be seen as a collection of points that approximate a platonic ideal, represented here as the crosshairs. *right*) In contrast, a time series chain may be seen as an evolving trail of points in the space. Here the crosshairs represent the first link in the chain, the *anchor*.

This paper is an abridged version of the paper "Matrix Profile VII: Time Series Chains: A New Primitive for Time Series Data Mining" that appeared in 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, 2017, pp. 695-704.

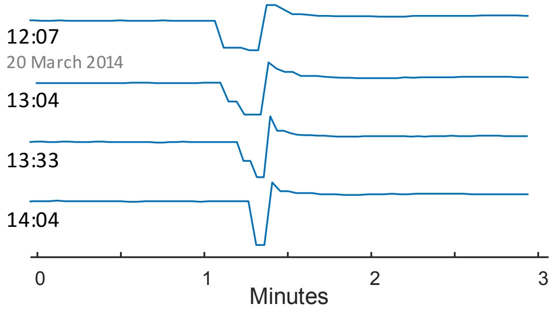

Figure 2. A time series chain discovered in an electrical power demand dataset monitoring domestic freezer usage [Murray *et al*., 2015]. Note that through the early afternoon, the valley becomes narrower and the peak that follows it becomes sharper.

While we can clearly define *chains*, it may not be obvious that such constructs exist in the real-world. In fact, as we preview in Figure 2, time series chains appear to be near ubiquitous in many domains, so long as the data trace is sufficiently long. As we will show later in Section 3, once given the formal definition and the computational ability to find time series chains, we begin to find them everywhere, in datasets from ten seconds, to ten years in length.

2 Formal Definitions of Time Series Chains

Before we formally introduce Time Series Chains, recall our guiding principle from the last section. We want something very like the definition of time series motifs [Patel *et al*., 2002; Yeh *et al*., 2016; Zhu *et al*., 2016], but with the additional property of *directionality*. For example, given a choice between the following:

{ ape → abe → ape → ape → abe → ape }

{ ape → apt → opt → oat → mat → man }

The latter is strongly preferred because the pattern is in some sense "evolving" or "drifting". We can now see this intuition in the *real-valued* space of interest. The definition below captures this spirit in the continuous case.

Definition 1: A time series chain of time series T is an ordered set of subsequences: $TSC=\{T_{C1,m}, T_{C2,m}, \dots T_{Ck,m}\}$ $(C1\leq C2\leq \dots \leq Ck)$, such that for any $1 \leq i \leq k\text{-}1$, we have $RNN(T_{Ci,m}) = T_{C(i+1),m}$, and $LNN(T_{C(i+1),m}) = T_{Ci,m}$. We denote k the *length* of the time series chain.

Here $T_{Ci,m}$ is a subsequence of time series T, which is of length m and starts from position Ci. $RNN(x)$ denotes the right nearest neighbor of x in T, $LNN(x)$ denotes the left nearest neighbor of x in T. To help the reader better understand this definition, let us consider the following time series:

47, 32, 1, 22, 2, 58, 3, 36, 4, -5, 5, 40

Assume that the subsequence length is 1, and the distance between two subsequences is simply the absolute difference between them (to be clear, we are making these simple and pathological assumptions here just for the purposes of elucidation; we are actually targeting much longer subsequence lengths and using z-normalized Euclidean distance in our applications). In Figure 3, we use arrows to link every subsequence in the time series with its left and right nearest neighbors.

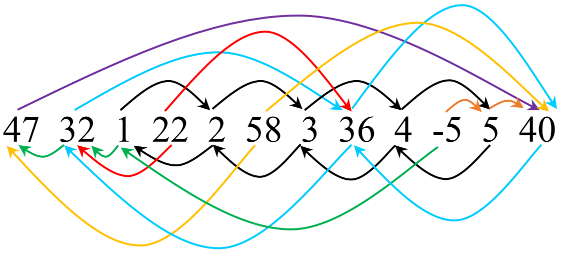

Figure 3. Visualizing the left/right nearest neighbors: every arrow above the time series points from a number to its right nearest neighbor; every arrow below the time series points from a number to its left nearest neighbor.

We call an arrow pointing from a number to its right nearest neighbor (arrows shown *above* the time series) a *forward arrow* (i.e. $x \rightarrow y$ means $RNN(x)=y$), and an arrow pointing from a number to its left nearest neighbor (arrows shown *below* the time series) a *backward arrow* (i.e. $x \leftarrow y$ means $LNN(y)=x$). Definition 1 indicates that every pair of consecutive subsequences in a chain must be connected by both a forward arrow and a backward arrow. The diligent reader may quickly discover the longest time series chain in our toy example:

47, 32, 1, 22, 2, 58, 3, 36, 4, -5, 5, 40 (*Raw data*)
1 ⇌ 2 ⇌ 3 ⇌ 4 ⇌ 5 (*Extracted chain*)

We can see that this chain shows a gradual increasing trend of the data. Note that in this one-dimensional example, the elements of the chain can only drift by increasing or decreasing. In the more general case, the elements can drift in arbitrarily complex ways. Our definition is also capable of discovering complex drifting patterns in high-dimensional space. For example, the reader can easily verify that the two-dimensional chain in Figure 1.*right,* a curvy evolving pattern, is captured by our definition. We defer real-world examples in much higher dimensional spaces to Section 3.

We are interested in two types of time series chains: anchored and unanchored chains.

Definition 2: An *anchored* time series chain of time series T starting from subsequence $T_{j,m}$ is an ordered set of **subsequences**: $TSC_{j,m} = \{T_{C1,m}, T_{C2,m}, \dots T_{Ck,m}\}$ $(C1\leq C2\leq \dots \leq Ck,\ C1=j)$, such that for any $1\leq i \leq k\text{-}1$, we have $RNN(T_{Ci,m})=T_{C(i+1),m}$, and $LNN(T_{C(i+1),m})=T_{Ci,m}$; for $T_{Ck,m}$, we have either $T_{Ck,m}$ is the last subsequence in T, *or* $LNN(RNN\ (T_{Ck,m}))\neq T_{Ck,m}$.

We can "grow" an anchored chain step-by-step as follows. Consider Figure 3 as an example. If we start from 1, we find $RNN(1)=2$ and $LNN(2)=1$, so 2 can be added to the chain; since $RNN(2)=3$ and $LNN(3)=2$, 3 can also be added; this process continues until we reach 5. As $RNN(5)=40$ and $LNN(40)\neq 5$, the chain terminates, and finally we find the chain 1⇌2⇌3⇌4⇌5 as the longest chain starting from 1.

We believe that of all the anchored chains in time series T, the longest one should reflect the most general trend within the data. We call this chain the unanchored time series chain:

Definition 3: An *unanchored* time series chain of time series T is the longest time series chain within T.

Note that there can be more than one unanchored time series chain of time series T with the same maximum length. In case of such ties, we report the chain with minimum average distance between consecutive components.

2.1 Finding the Unanchored Time Series Chain

According to definitions 1-3, the algorithm to find the unanchored Time Series Chain can be divided into three steps. First, we need to find the left and right nearest neighbor of all the subsequences in the time series. We use an algorithm called LRSTOMP, a variant of the STOMP algorithm [Zhu *et al.*, 2016] to compute such information. The time complexity of LRSTOMP is $O(n^2)$ and the space complexity is $O(n)$ (here n is the length of the time series), the same as STOMP [Zhu *et al.*, 2016]. Next, we find all the anchored chains within the time series based on Definition 2 (i.e., we "grow" chains from every subsequence in the time series). The time and space complexity of this step are both $O(n)$. Finally, the unanchored chain is simply the longest anchored chain discovered. We refer interested readers to [Supporting Webpage, 2017] for a detailed discussion of the algorithm.

3 Experimental Evaluation

We note in passing that all the experimental results in this paper are reproducible. To ensure this, we have created a website to archive all the datasets and code in perpetuity [Supporting Webpage, 2017].

We provide four case studies in which we applied our chain discovery algorithm to various datasets. These case studies will help the reader gain an appreciation for the utility of chain discovery. These datasets are designed to span the diverse types of data encountered in time series data mining, some are stationary, some have trends, some are smooth, some are noisy, the shortest is ten seconds long, the longest is ten years, etc.

Note that in this section, we are only showing the application of the unanchored time series chain. Unless otherwise stated, in the rest of this section, we use the term "time series chain" to represent unanchored time series chain in Definition 3, rather than Definition 1.

3.1 Case Study: Hemodynamics

In November 2016, we briefed Dr. John Michael Criley, Professor Emeritus at the David Geffen School of Medicine at UCLA, and Dr. Gregory Mason of UCLA Medical Center, a noted expert on cardiac hemodynamics, on the capabilities of time series chain discovery. They suggested more than a dozen possible uses for it in various clinical and research scenarios in medicine. Here we consider one example they are interested in.

Syncope is the loss of consciousness caused by a fall in blood pressure. The tilt-table test (see Figure 4.*top.left*) is a simple, noninvasive, and informative test first described in 1986 as a diagnostic tool for patients with syncope of unknown origin [Heldt *et al.*, 2003]. Beyond diagnosing the condition, the test may reveal the cause, neurological disorder, metabolic disorder, mechanical heart disease, cardiac arrhythmias, etc. [Moya, 2009].

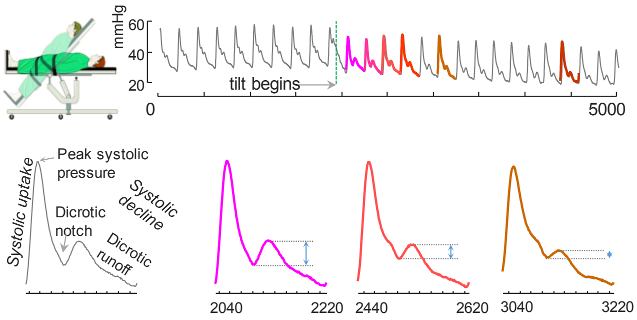

Figure 4. *left-to-right*, *top-to-bottom*) A patient lying on a medical tilt table has his arterial blood pressure monitored. Nomenclature for a standard beat. The chain discovered in this dataset shows a decreasing height for the dicrotic notch.

In brief, the clinician will want to contrast any evolving patterns in the patient's *arterial* blood pressure (ABP) that are a response to changes in positon induced by a tilt table, with evolving patterns that are not associated with changes of posture. As hinted at in Figure 4, time series chains are an ideal way to find and summarize such patterns. Here we set m=200, as this is the typical length of an ABP signal (Figure 4.*bottom.left*).

Figure 4 shows just a snippet of the time series searched. We encourage the reader to see the full dataset/results at [Supporting Webpage, 2017]. Nevertheless, even this snippet is visually compelling. It shows that as the table is tilted, the height of the dicrotic notch steadily decreases. Per Dr. Mason, the change in orientation "*dramatically increases central venous filling and subsequent left ventricular end-diastolic* volume*, for several heart beats. Left ventricular stroke volume and effective cardiac output increase transiently*, (likely due to) *relative hyperemia, which is well-described during recovery from transient vascular occlusion*".

3.2 Case Study: Penguin Behavior

In this case study, we decided to explore a dataset for which we have no expertise, to see if we could find time series chains, which we could then show to an expert for independent evaluation of meaning and significance (if any).

To this end, we consider telemetry collected from a Magellanic penguin (*Spheniscus magellanicus*). The dataset was collected by attaching a small multi-channel data-logging device to the bird. The full data consists of 1,048,575 data points recorded at 40 Hz (about 7.2 hours). While a suite of measurements was recorded, for simplicity we focus on the X-Axis acceleration (the direction of travel for a swimming bird). In Figure 5 we show the snippet of the data in which we found a chain, with m=28. This is about 0.7 seconds, and the approximate period of the data.

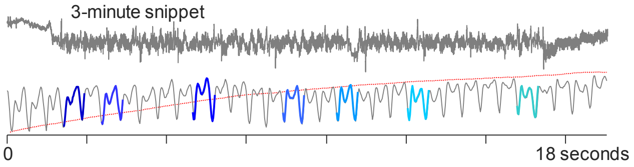

Figure 5. *top*) A random three-minute snippet of X-Axis acceleration of a Magellanic penguin (from a total of 7.2 hours). *bottom*) An eighteen-second long section containing the time series chain. In the background, the red time series records the depth, starting at sea-level and leveling off at 6.1 meters.

In fact, this chain *does* have a simple interpretation. Adult Magellanic penguins regularly dive to depths of up to 50m to hunt prey, and may spend as long as fifteen minutes under water. One of our sensors measures pressure, which we showed in Figure 5.*bottom* as a fine/red line. This shows that the chain begins just after the bird begins its dive, and ends as it reached its maximum depth of 6.1m. Magellanic penguins have typical body densities for a bird at sea-level, but just before diving they take a very deep breath that makes them exceptionally buoyant [Ponganis *et al*., 2015]. This positive buoyancy is difficult to overcome near the surface, but at depth, the compression of water pressure cancels it, giving them a comfortable neutral buoyancy [Ponganis *et al*., 2015; Williams *et al*., 2011]. In order to get down to their hunting ground below sea level it is clear that "(for penguins) *locomotory muscle workload, varies significantly at the beginning of dives*" [Williams *et al*., 2011]. The snippet of time series shown in Figure 5 does not suggest much of a change in *stroke-rate*, however penguins are able to vary the thrust of their flapping by twisting their wings [Williams *et al*., 2011]. The chain we discovered shows this dramatic sprint downwards leveling off to a comfortable cruise. Fortunately, our data contains about a dozen major dives, allowing us to confirm our hypothesis about the meaning of this chain on more data.

Note that our chain does not include every stroke in the dive. Our data is undersampled (only 40Hz for a bird that can swim at 36kph) and this data is recorded in the wild, the bird may have changed directions to avoid flotsam or fellow penguins. However, this is a great strength of our algorithm: we do not need "perfect" data to find chains; we can find chains in real-world datasets. Also, from Figure 5.*bottom* we can see that *m*=28 is longer than the actual period of the data; our algorithm is not sensitive to this and still discovered a meaningful chain.

3.3 Case Study: Human Gait

In the experiments in the previous section we could be sure of the validity of the discovered chains, because we had access to some ground truth. In this section and the next, we show examples of chains we discovered in datasets for which we do not have an obvious way to empirically verify. This demonstrates one use for chains, finding patterns that are interesting but speculative, and may warrant further investigation.

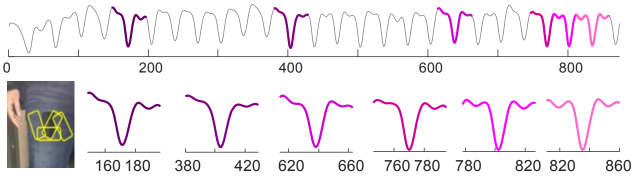

Figure 6. *top*) A 30-second snippet of data from an accelerometer on a mobile phone. The phone was placed in the user's front pocket (*inset*). *bottom*) The extracted chain shows an evolution to a stable and symmetric gait.

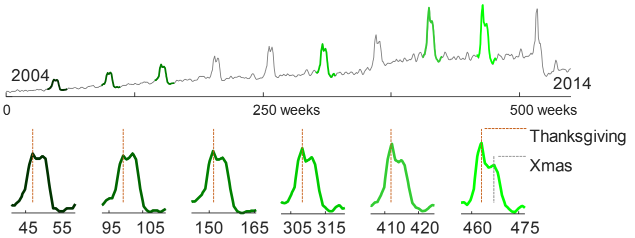

Figure 7. *top*) Ten years of query volume for the keyword *Kohl's*. *bottom*) The z-normalized links of the time series chain discovered in the data hints at the growing importance of "*Cyber Monday*".

We first consider a snippet of a gait dataset recorded to test a hypothesis about biometric identification [Hoang *et al*., 2015]. The dataset is shown in Figure 6.*top*. We set *m* = 50 here, as this is the approximate length of a period of the data. As hinted at in Figure 6.*inset* (taken from the original paper), the authors of the study were interested in "*the instability of the mobile in terms of its orientation and position when it is put freely in the pocket*" [Hoang *et al*., 2015]. Given the experimental setup, we suspected that the gait pattern might start out as being unpredictable as the phone jostled about in the user's pocket, eventually settling down as the phone settled into place. This is exactly what we see in Figure 6.*top*. Note that the first few links are far apart and asymmetric, but the last few links are close together, and almost perfectly symmetric.

3.4 Case Study: Web Query Volume

We next consider a dataset that is noisy, under-sampled and has a growing trend. We examined a decade-long *GoogleTrend* query volume for the keyword *Kohl's*, an American retail chain (data courtesy of [Matsubara *et al*., 2015]). As shown in Figure 7, the time series features a significant "bump" around the end-of-years holidays, unsurprising for a store known as a destination for gift buyers. Here we set *m* = 76 (the approximate length of a "bump"). Note that *m* does *not* need to be precisely set. If we set *m* = 114 (50% longer), we can still obtain the same basic chain (though each link is 50% longer, see [Supporting Webpage, 2017] for a visual comparison).

The discovered chain shows that over the decade, the bump transitions from a smooth bump covering the period between Thanksgiving and Xmas, to a more sharply focused bump centered on Thanksgiving. This seems to reflect the growing importance of *Cyber Monday*, a marketing term for the Monday after Thanksgiving. The phrase was created by marketing companies to persuade people to shop online. The term made its debut on November 28th, 2005 in a press release entitled "*Cyber Monday Quickly Becoming One of the Biggest Online Shopping Days of the Year*" [Smith, 2010]. Note that this date coincides with the first glimpse of the sharping peak in our chain.

Here we seem to "miss" a few links in the chain. However, note that the data is noisy and coarsely sampled, and the "missed" bumps are too distorted to conform with the general evolving trend. This noisy example again illustrates the robustness of our technique. As before, we note that we do not need "perfect" data to find meaningful chains. Even if some links are badly distorted, the discovered chain will still be able to include all the other evolving patterns.

4 Conclusions and Future Work

We introduced *time series chains*, a new primitive for time series data mining. We have shown that chains can be efficiently and robustly discovered from noisy and complex datasets, to provide useful insights. In future work we plan to consider a more theoretical treatment of the properties of chains, and adapt/apply them to online problems, including prognostics and concept drift.

Acknowledgments

We gratefully acknowledge funding from MERL and from NSF IIS-1161997 II and NSF IIS-1510741.

References

[Hao *et al*., 2012] Ming C. Hao, Manish Marwah, Halldór Janetzko, Umeshwar Dayal, Daniel A. Keim, Debprakash Patnaik, Naren Ramakrishnan and Ratnesh K. Sharma, Visual exploration of frequent patterns in multivariate time series. *Information Visualization,* 11.1(2012): 71-83.

[Heldt *et al*., 2003] T. Heldt, M. B. Oefinger, M. Hoshiyama, and R. G. Mark, Circulatory response to passive and active changes in posture. In *Computers in Cardiology* (2003): 263-266. IEEE.

[Hoang *et al*., 2015] Thang Hoang, Deokjai Choi, and Thuc Nguyen, On the instability of sensor orientation in gait verification on mobile phone. In *12th IEEE International Joint Conference on e-Business and Telecommunications (ICETE),* 4(2015): 148-159.

[Lovallo *et al*., 2004] William R. Lovallo, Michael F. Wilson, Andrea S. Vincent, Bong Hee Sung, Barbara S. McKey, and Thomas L. Whitsett, Blood pressure response to caffeine shows incomplete tolerance after short-term regular consumption. *Hypertension*, *43.4* (2004): 760-765.

[Matsubara *et al*., 2015] Yasuko Matsubara, Yasushi Sakurai, and Christos Faloutsos, The web as a jungle: Non-linear dynamical systems for co-evolving online activities. In *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 721-731.

[Moya, 2009] Angel Moya, Tilt testing and neurally mediated syncope: too many protocols for one condition or specific protocols for different situations?. Eur Heart J. 30.18 (2009): 2174-2176.

[Murray *et al*., 2015] D. Murray, J. Liao, L. Stankovic, V. Stankovic, R. Hauxwell-Baldwin, C. Wilson, M. Coleman, T. Kane, S. Firth, A data management platform for personalised real-time energy feedback. *EEDAL,* 2015.

[Patel *et al*., 2002] Pranav Patel, Eamonn Keogh, Jessica Lin, and Stefano Lonardi, Mining motifs in massive time series databases. In *Proceedings of the 2002 IEEE International Conference on Data Mining.* pp 370-377.

[Ponganis *et al*., 2015] P.J. Ponganis, J. St Leger and M. Scadeng, Penguin lungs and air sacs: implications for baroprotection, oxygen stores and buoyancy. *Journal of Experimental Biology*. (2015): 720-730.

[Shokoohi-Yekta *et al*., 2015] Mohammad Shokoohi-Yekta, Yanping Chen, Bilson Campana, Bing Hu, Jesin Zakaria, and Eamonn Keogh, Discovery of meaningful rules in time series. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining,* 2015, pp. 1085-1094.

[Smith, 2010] Jordan Smith, The Accidentally-on-Purpose History of Cyber Monday. 2010. URL retrieved February 5th 2017: www.esquire.com/news-politics/news/a23870/cyber-monday-online-shopping-4021548/

[Syed *et al*., 2010] Zeeshan Syed, Collin Stultz, Manolis Kellis, Piotr Indyk, and John Guttag, Motif discovery in physiological datasets: a methodology for inferring predictive elements. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4.1(2010): 2.

[Williams *et al*., 2011] Cassondra L. Williams, Katsufumi Sato, Kozue Shiomi, and Paul J. Ponganis, Muscle energy stores and stroke rates of emperor penguins: implications for muscle metabolism and dive performance. *Physiological and Biochemical Zoology* 85.2(2011): 120-133.

[Yeh *et al*., 2016] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh, Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View that Includes Motifs, Discords and Shapelets. In *Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 1317-1322.

[Zhu *et al*., 2016] Yan Zhu, Zachary Zimmerman, Nader Shakibay Senobari, Chin-Chia Michael Yeh, Gareth Funning, Abdullah Mueen, Philip Brisk, and Eamonn Keogh, Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins. In *Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 739-748.

[Supporting Webpage, 2017] Supporting Webpage: https://sites.google.com/site/timeserieschain/