Matrix Profile XXV: Introducing Novelets: A Primitive that Allows Online Detection of Emerging Behaviors in Time Series

Ryan Mercer Eamonn Keogh
University of California, Riverside
rmerc002@ucr.edu, eamonn@cs.ucr.edu

***Abstract*—While offline exploration of time series can be useful, time series analysis is almost unique in allowing the possibility of direct and immediate intervention. For example, if we are monitoring an industrial process and our algorithm predicts imminent failure, the algorithm could direct a controller to open a release valve or alert a response team. There now exist mature tools to monitor time series for *known* behaviors (template matching), previously unknown highly *conserved* behaviors (motifs) and *unexpected* behaviors (anomalies). In this work we claim that there is another useful primitive, *emerging* behaviors, that are worth monitoring for. We call such behaviors *Novelets*. We explain that Novelets are neither anomalies nor motifs but can be loosely thought of as initially *apparent* anomalies that are later realized to be motifs. We will show Novelets have a natural interpretation in many disciplines, including science, medicine, and industry. As we will further demonstrate, Novelet discovery can have many downstream uses, including prognostics and abnormal behavior detection. We will demonstrate the utility of our proposed primitive on a diverse set of domains.**

I. Introduction

In recent years the data mining community has generalized many of its batch algorithms to the more actionable online setting. In particular, there now exist mature tools to monitor time series for *known* behaviors (template matching) [12], previously unknown *repeated* behaviors (motifs) [21] and *unexpected* behaviors (anomalies or discords) [24]. However, there is almost no work to address the idea of *emerging* behaviors. Perhaps the reason this primitive has escaped the community's attention is that this idea may *seem* to be encapsulated in other primitives. However, none of these notions captures the general sense of emerging behaviors, which can perhaps be thought of as an anomaly/discord that later transitions to a motif. This view of emerging behaviors is more than just a mnemonic device, it offers a direction to create an emerging behavior discovery algorithm. This is because there is a widely used and mature framework to monitor for both motifs and discords, the Matrix Profile [21].

In Fig. 1 we show an example of an emerging behavior in a respiration dataset. Here most cycles look like this ⎍, a healthy breath cycle. In addition, there are some noisy cycles. However, there is also this unusual shape ⎍‿, that emerges after 40 seconds. Note that while this shape is neither the minimum nor maximum of a Matrix Profile (or left Matrix Profile), we can expand and adapt the Matrix Profile framework to produce an *Emergence Profile*, allowing an emerging behavior discovery algorithm. We call our approach *Novelets*, Newly Observed Variation while Excluding Learned and Established Time Series.

To make the actionability of Novelets clear, consider the following:

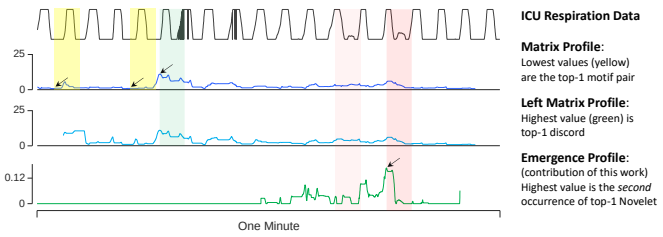

Fig. 1. A short respiration data snippet featuring normal, noisy, and emergent behaviors. Neither motifs nor discords capture either of the two instances of emergent behavior (highlighted in red). The Emergence Profile peaks at the *second* instance of the emergent behavior, as expected. The explanation of the cause and medical implications of this Novelet are available at [25].

- **Medicine**: Much of medical telemetry is replete with motifs (i.e., repeated heartbeats or respiration cycles), and with discords (i.e., sensors artifacts, movement artifacts). In the presence of such data, new patterns can emerge, either naturally or as the result of an intervention. For example, in an ICU a doctor may give her patient Digoxin. This drug can change the shape of the heartbeat, generally for the better, but sometimes in negative ways. Moreover, it can take 0.5 to 6 hours to produce an effect [17]. The doctor may wish to receive an alert the moment a new heart rhythm emerges.
- **Science**: The white-throated sparrow (*Zonotrichia albicollis*) is a bird that is found mostly in Canada, and pleasingly, its distinctive song rhythm is often transcribed as "*O-oh sweet Canada, Canada, Canada*". However, beginning in the early 2000s a new unexpected song variant emerged, roughly transcribed as "*O-oh sweet Cana, Cana, Cana*" [18]. Moreover, this newly emerged variant began to spread across the country, creating a flurry of academic interest [14]. In this case, the discovery of the emerging behavior was accidental [18], but one could imagine an algorithmic effort to discover novelty in bird songs.
- **Industry**: Industrial assets are often heavily monitored. For batch processes only a handful of patterns are normally observed, corresponding to different recipes being observed. Anomalous data is often associated with a spoiled batch. However, an emerging behavior is suggestive of an undocumented and unexpected change. For example, an operator may be unconsciously compensating for degrading infeed product by allowing a pasteurizing regime to run longer than normal [22]. The plant manager would like to be made aware of such emerging patterns.

In Section IV we consider several real-world examples that closely model these motivating examples. Note that unlike *anomalies* which are normally considered undesirable,

Novelets may be beneficial, neutral, or negative, depending on the context.

The rest of this paper is organized as follows. In Section II we explain the essential definitions, notation, and related work. With these developed intuitions, we explain exploratory data mining tasks in Section III. We evaluate the utility of Novelet discovery in Section IV, followed by model comparisons in Section V. Section VI delivers conclusions.

II. Definitions and Notation

Our data type of interest is *time series*.

Definition 1: A *time series* $\mathbf{T} = t_1, t_2, \ldots, t_n$ is a sequence of real-valued numbers.

For the task-at-hand we are not interested in any global properties of a time series but rather we are concerned with the shapes of small regions called *subsequences*.

Definition 2: A *subsequence* $\mathbf{T}_{i,m}$ is a contiguous subset of values from $\mathbf{T}$ starting at index i with length m.

We can measure the distance between any two time series of equal length using a distance measure. We will use the ubiquitous z-normalized Euclidean distance [21]. One minor modification to the Euclidean distance is that we clip it at $\sqrt{2m}$ because values above this are anti-correlated in the Pearson Correlation space [11]. This is done in order to make the greatest use of the normalized range. If we wish to measure the distance between a short time series and every subsequence from a long time series, we can produce a *distance profile*.

Definition 3: A *distance profile* $\mathbf{DP}_i^{(\mathrm{AB}),m}$ is the vector of distances between a query subsequence $\mathbf{T}_{i,m}^{(\mathrm{A})}$ and a reference time series $\mathbf{T}^{(\mathrm{B})}$.

The distance profile can be computed very efficiently using the MASS algorithm [12]. Fig. 2 illustrates these definitions on a running example of a synthetic time series containing three types of signals. This example is comprised mostly of a (slightly noisy) sine wave which acts as the background pattern. There is also a region of pure noise, and three examples of a modified sine wave that can be seen to "emerge" from the background pattern over time.

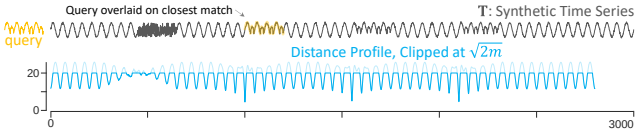

Fig. 2. *top*) A short synthetic time series containing three signal types: a sine wave, a modified sine wave, and noise. *bottom*) A segment closely matching the modified sine wave is used as a query to generate distance profile $\mathbf{DP}^{(\mathrm{AB}),200}$. Lower values indicate subsequence locations which are more similar to the query.

This conceptual element is helpful when constructing the *AB-join Matrix Profile* [21].

Definition 4: An *AB-join Matrix Profile* $\mathbf{MP}^{(\mathrm{AB}),m}$ between reference time series $\mathbf{T}^{(\mathrm{A})}$ and a query time series $\mathbf{T}^{(\mathrm{B})}$ is a vector of Euclidean distances between each subsequence $\mathbf{T}_{i,m}^{(\mathrm{A})}$ and its nearest neighbor $\mathbf{T}_{j,m}^{(\mathrm{B})}$. Formally,

$$\mathbf{MP}^{(\mathrm{AB}),m} = \left[\boldsymbol{min}\left(\mathbf{DP}_1^{(\mathrm{AB}),m}\right), \boldsymbol{min}\left(\mathbf{DP}_2^{(\mathrm{AB}),m}\right), \ldots, \boldsymbol{min}\left(\mathbf{DP}_{n-m+1}^{(\mathrm{AB}),m}\right)\right]$$

Note that in general, $\mathbf{MP}^{(\mathrm{AB}),m} \neq \mathbf{MP}^{(\mathrm{BA}),m}$. Even if they do happen to have equal lengths, they correspond to different reference time series.

We incorporate $\mathbf{MP}^{(\mathrm{AB}),m}$ into our running example in Fig. 3. The top motifs are a pair of sine waves which are outlined in $\mathbf{T}^{(\mathrm{A})}$ and $\mathbf{T}^{(\mathrm{B})}$.

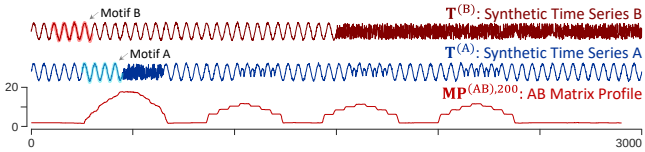

Fig. 3. *top*) Time series $\mathbf{T}^{(\mathrm{B})}$ is synthetically created to contain equal lengths of sine wave and noise. *center*) Time series $\mathbf{T}^{(\mathrm{A})}$, first introduced in Fig. 2, contains three signal types. *bottom*) $\mathbf{MP}^{(\mathrm{AB}),200}$ reveals behaviors common to both $\mathbf{T}^{(\mathrm{A})}$ and $\mathbf{T}^{(\mathrm{B})}$ as low values and behaviors unique to $\mathbf{T}^{(\mathrm{A})}$ as higher values.

Subsequence comparisons can also be adapted to subsequences *within* a single time series using the *self-join Matrix Profile*.

Definition 5: A *self-join Matrix Profile* $\mathbf{MP}^{(\mathrm{AA}),m}$ of a time series $\mathbf{T}^{(\mathrm{A})}$ is a specialization of the AB-join Matrix Profile with identical query and reference time series. An exclusion zone of length m is centered around each query index i for query subsequence $\mathbf{T}_{i,m}^{(\mathrm{A})}$ to suppress trivial nearest neighbors.

Fig. 4 shows $\mathbf{MP}^{(\mathrm{AA}),200}$ for our running example. Note that the embedded modified sine signal is neither a motif nor a discord.

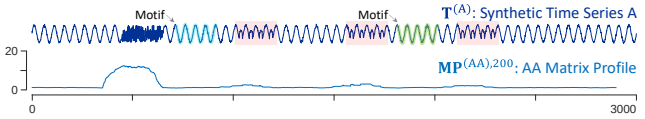

Fig. 4. *top*) The synthetic time series $\mathbf{T}^{(\mathrm{A})}$ and its corresponding $\mathbf{MP}^{(\mathrm{AA}),200}$ (*bottom*). The top motif pair are outlined. The three instances of the altered sine wave, which are highlighted in red, each have a corresponding mid-valued bump in $\mathbf{MP}^{(\mathrm{AA}),200}$, which excludes them from consideration as either motifs or discords.

In online applications where future subsequences have not yet arrived, it can be useful to reason about the similarity between newly arriving subsequences and all the subsequences observed up to that point by utilizing the *left self-join Matrix Profile*.

Definition 6: A *left self-join Matrix Profile* $\mathbf{LMP}^{(\mathrm{AA}),m}$ of a time series $\mathbf{T}^{(\mathrm{A})}$ is a specialization of the self-join Matrix Profile where each subsequence $\mathbf{T}_{i,m}^{(\mathrm{A})}$ is only compared to subsequences $\mathbf{T}_{j,m}^{(\mathrm{A})}$ where $j < i$.

The $\mathbf{LMP}^{(\mathrm{AA}),200}$ of the running example is shown in Fig. 5. The largest noticeable difference can be seen around the first instance of the modified sine. The difference in distances between the first and second instances offers a clue to discovering this elusive emerging behavior.

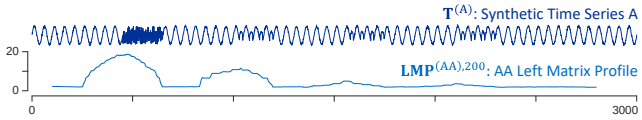

Fig. 5. *top*) Our running example time series. *bottom*) The left Matrix Profile $\mathbf{LMP}^{(AA),m}$ will highlight the first instance of a repeated behavior as an anomaly, but the first instance of the modified sine is still a moderate distance value, which excludes it from classification as a top discord compared to noise.

Note the similarity between $\mathbf{MP}^{(AB),200}$ from Fig. 3 and $\mathbf{LMP}^{(AA),200}$ from Fig. 5. They are very similar in most regions. This makes sense when the behaviors present in $\mathbf{T}^{(A)}$ and $\mathbf{T}^{(B)}$ are similar. Even noisy patches similarly have high distance within the two Matrix Profiles. Here is the key observation that informs our proposed algorithm:

Key Observation: $\mathbf{MP}^{(AB),m}$ and $\mathbf{LMP}^{(AA),m}$ only diverge in regions where there are emerging repeated behaviors in $\mathbf{T}^{(A)}$.

We are now in a position to introduce two new definitions. To do so, we will substitute the generic notations $\mathbf{T}^{(A)}$ and $\mathbf{T}^{(B)}$ for the specialized $\mathbf{T}^{(+)}$ and $\mathbf{T}^{(-)}$, which carry mild assumptions about their contents. In essence, we would like to identify emerging behaviors within $\mathbf{T}^{(+)}$ while suppressing those already known within $\mathbf{T}^{(-)}$.

Definition 7: An *Emergence Profile* $\mathbf{EP}^{m}$ is the element-wise difference between Matrix Profiles $\mathbf{MP}^{(+-),m}$ and $\mathbf{LMP}^{(++),m}$, where $\mathbf{MP}^{(+-),m}$ joins $\mathbf{T}^{(+)}$ with $\mathbf{T}^{(-)}$, and $\mathbf{LMP}^{(++),m}$ is the left self-join of $\mathbf{T}^{(+)}$.

$$\mathbf{EP}^{m} = \mathbf{MP}^{(+-),m} - \mathbf{LMP}^{(++),m}$$

The Emergence Profile is defined for any pair of time series which are longer than the subsequence length m. Moreover, it is still defined when the initial $\mathbf{T}^{(-)}$ is empty. In this case, $\mathbf{MP}^{(+-),m}$ is defined to be a constant vector of length $|\mathbf{T}^{(+)}| - m + 1$, set to the maximum distance value.

Fig. 6 shows a discovered emerging behavior, identified as a peak in $\mathbf{EP}^{200}$. The discovery confirms by visual inspection that the corresponding location within $\mathbf{T}^{(+)}$ is the *second* instance of the behavior of interest (highlighted in red).

This definition resembles that of the recently introduced Contrast Profile [11], so we follow its normalization step which maximizes the effective normalization range by dividing $\mathbf{EP}^{m}$ by $\sqrt{2m}$, then clipping negative values to zero. This $\sqrt{2m}$ factor makes the results unitless, and independent of both the sampling rate and subsequence length.

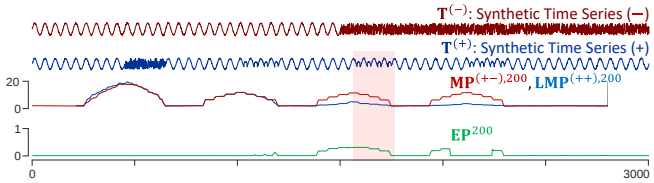

Fig. 6. *top*) Our running example time series. *middle*) By subtracting $\mathbf{LMP}^{(++),200}$ from $\mathbf{MP}^{(+-),200}$, the second instance of the emerging modified sine wave is clearly highlighted within $\mathbf{EP}^{200}$(*bottom*). Once a new behavior is learned, future instances are automatically suppressed.

The slight difference in definition leads to a desirable property. While the Contrast Profile must be completely updated as its time series grow, using the left self-join Matrix Profile means that the Emergence Profile is immutable as either $\mathbf{T}^{(+)}$ or $\mathbf{T}^{(-)}$ grows, a desirable property for online discovery of *Novelets.*

Definition 8: A *Novelet* $\mathbf{T}_{i,m}^{(+)}$ is the first instance of an emerging behavior. The second instance of an emerging behavior $\mathbf{T}_{j,m}^{(+)}$ produces a local maximum $\mathbf{EP}_{j}^{m}$ above some novelty threshold d, and its nearest neighbor in $\mathbf{LMP}^{(++)}$ is the Novelet $\mathbf{T}_{i,m}^{(+)}$.

It is important to note the following. While a Novelet is the *first* instance an emerging behavior, it is never possible to recognize it as such at the instant it was encountered. It is only possible to *retroactively* recognize it as a Novelet, when we see a later occurrence. Before that time, we cannot discount the possibility that it was a literally unique event.

The identified first and second instances of the modified sine from the running example are outlined in Fig. 7. The first two instances are enough to trigger a learning event, which results in the third instance being suppressed.

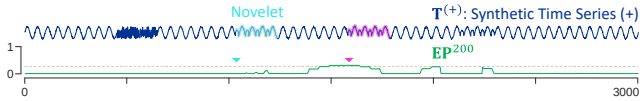

Fig. 7. *top*) The first and second instances of the emerging modified sine wave are outlined in $\mathbf{T}^{(+)}$. *bottom*) The third instance of the emerging behavior has been suppressed in $\mathbf{EP}^{m}$ by logically appending the first instance to $\mathbf{T}^{(-)}$ (not shown).

Whether discovered in batch or online mode, Novelets are discovered chronologically. Once discovered, a Novelet is logically appended to $\mathbf{T}^{(-)}$ so that future occurrences of the discovered behavior *are* suppressed in $\mathbf{EP}^{m}$.

Batch mode can generate $\mathbf{EP}^{m}$ with a brute force nested for loop algorithm requiring $O(m|\mathbf{T}^{(+)}|(|\mathbf{T}^{(-)}| + |\mathbf{T}^{(+)}|))$. In real world applications, the cumulative length of all learned Novelets is much shorter than the length of $\mathbf{T}^{(-)}$, so this contribution can be ignored. If we consider the subsequence length m, which may be in the thousands, the naïve time complexity is clearly untenable. By building our definitions on top of the Matrix Profile, we can leverage a large and diverse community of researchers (academic and industrial) that continue to accelerate the Matrix Profile [21][24], and we can apply Novelets in real time to diverse online applications.

A. *Differentiating Novelets from Related Concepts*

While Novelets have analogues in text and other domains, we believe that they have not been considered by the time series data mining community before. Here we take the time to explicitly state the difference between Novelets and other time series primitives/definitions. Where appropriate, we will use the familiar trick of using text strings as a proxy for time series, with Hamming distance replacing Euclidean distance.

- **Time series motifs** [21]: While Novelets may be time series motifs, most motifs are not Novelets. For example, consider ..`BeatzBeatyBeatxBeat`.. Here "`Beat`" is a strong motif

that represents a known healthy heartbeat. If we exclude "Beat" when considering the incoming stream ..BeatwBeatvBeatsPadakBeatqPatajBeatnBeat.., we notice the less well conserved pattern "Pada/Pata" that emerges late in the string, which may indicate an emerging arrhythmia pattern. This analogy explains why motif discovery algorithms may overlook the emergence of less conserved motifs which may eventually become Novelets.

- **Time series discords/anomalies** [21]: An anomaly is typically a shape that occurs once, whereas a Novelet is the first instance of a repeated shape. For example, when de Vlamingh became the first European to see a black swan in 1697, it was a stunning anomaly to him. Since at least Roman times, the phrase "black swan" was used to mean something that could not exist[1]. However, within days of the first sighting, black swans ceased to be anomalies, but simply became (temporarily) novel birds. As Fig. 1 shows, anomalies are generally unique, and not newly emerging behaviors. It is only after we spot a *second* occurrence of an "anomaly" that we can reconsider it as an emerging pattern. While there are hundreds of anomaly detection algorithms [2][4] in the literature, none has this ability to retrospectively recategorize an "anomaly" as a newly emerged behavior.
- **Time series chains** [23]: Time series chains track gradually changing behaviors, while Novelets track sufficiently new behaviors. A chain must be of at least length three, whereas a Novelet may be detected with just two occurrences. We can best think of chains as a slowly changing behavior, say a slowly degrading batch cycler or a slowly improving golf swing. In contrast, a Novelet is a completely novel behavior.
- **Time series segmentation/change detection:** In a trivial case where there are two regimes, the output of segmentations and Novelets may both point to the start of each regime. For example, when monitoring an infant's coaxed speech development, Novelets should identify the first instances of "MAMA" and "PAPA" in "MAMAzMAMAyMAMAxPAPAvPAPAtPAPA". As a practical matter, most time series segmentation algorithms only work well if the data is highly periodic, and they can see *many* periods per regime, say at least a few dozen walk cycles, followed by a few dozen run cycles. In contrast, Novelets only needs to see *two* examples of a new behavior.
- **Time series clustering**: Time series subsequence clustering typically groups many (but critically, not *all* [8]) subsequences per cluster in an attempt to simplify class representation. While Novelets identify subsequences which are sufficiently different, there is no intent in the core functionality to assign group IDs to any subsequences similar to the Novelets. The string "DogHogBinGinFrogTin" contains two clusters of animals ending with "og" and objects ending with "in". The Novelets with a new character threshold of two are "Dog" and "Bin". When applied to song lyrics, Novelets would identify when new rhyme schemes *began*, whereas clusters would identify all words belonging to a scheme.

To summarize, we believe that Novelets represent a truly new primitive that is not captured by any existing definition.

B. Related Work

The previous section implicitly covered much of the related work. In addition, we should also consider rare time series motif discovery [1] which attempts to identify infrequent repeating patterns across large spans of time. In the effort outlined in [1], exactness guarantees are abandoned due to computational limitations. Instead, repeated patterns are identified with a probability proportional to their frequency. While the ambition is admirable, the SAX approximations at the heart of this approach require careful parameter tuning and allow false dismissals. Additionally, there is no mechanism to ignore well understood (background) patterns in order to focus on emerging patterns.

C. The Novelet Assumptions

Recall the mild assumptions of Novelets: $\mathbf{T}^{(-)}$ contains previously observed patterns which will be suppressed and only emerging patterns exceeding the novelty threshold will be identified within $\mathbf{T}^{(+)}$. In practice, Novelets can still be run with a completely cold start if there exist no established behaviors in $\mathbf{T}^{(-)}$, although there will be an early brief surge of emerging behaviors that will be reported, as *anything* will be considered novel in comparison to *nothing*.

One subtlety arises when processing time series in batches. Suppose Novelets are discovered on Monday, then Monday's batch is used as $\mathbf{T}^{(-)}$ in Tuesday's batch for Novelet discovery. There may be a single occurrence of a pattern within Monday's batch that was not reported as a Novelet. If a second instance occurs in Tuesday's batch, it will be suppressed. This can be solved by using an amnestic sliding window, as described in Section II.*E*.

A pattern pair will only trigger Novelet discovery if they are sufficiently novel, but fortunately, setting the novelty threshold d is straightforward through a few heuristics. In practice, Novelet discovery has low sensitivity to the choice of d. A first sanity check is to run a self-join Matrix Profile on a representative training time series in order to determine if any motifs exist. The contribution of Novelets is to show where the first instance of an emerging motifs occurs, so if there are no motifs, then nothing will emerge. The next step is to identify whether there are any contrasting behaviors by running the Contrast Profile. With similar reasoning to the Matrix Profile, if there are no contrasting motifs, then a first contrasting motif cannot be identified.

Assuming that contrasting motifs exist, it is reasonable to set d to 80% of the contrast value of a compelling contrasting motif. While this process may require some trial and error, it is possible to quickly rediscover Novelets without recomputing the base $\mathbf{MP}^{(+-),m}$ and $\mathbf{LMP}^{(++),m}$, which account for the bulk of the time expense. Note that the order in which Novelets are discovered can influence which future shapes are judged as Novelets. Setting d too low may result in shapes which appear out of phase or unintuitively novel. An

[1] The Roman satirist Juvenal wrote in AD 82 of *rara avis in terris nigroque simillima cygno* ("*a rare bird in the lands, and very like a black swan*"), meaning that since a black swan did not exist, the proposed "rare bird" did not exist. Here "rare bird" was not literally a bird, it is just something that did not exist, like an honest politician.

alternative technique that eliminates this parameter is discussed in Section V.*A*.

As is true with other habituating processes, an emerging pattern may pass under the novelty threshold due to past learned behaviors. If a sufficiently long time series has been monitored, the frequency of emerging behaviors slows to a trickle. Novelets may be useful as a canary-in-the-coal-mine for local change detection, so it may be beneficial to incorporate pruning functionality which forgets Novelets which have not been observed recently.

D. General Novelet Observations

The Emergence Profile is subsequence length normalized so that values are bound between zero and one. A value of one for subsequence $\mathbf{T}^{(+)}_{i,m}$ means that the there exists a maximally conserved left nearest neighbor $\mathbf{T}^{(+)}_{j,m}$ while also a maximally dissimilar nearest neighbor $\mathbf{T}^{(-)}_{k,m}$. Here, maximally dissimilar corresponds to uncorrelated subsequences, such as a pair of random noise subsequences, rather than anti-correlated subsequences.

The intentional consequence of subsequence length normalization is that Emergence Profiles may be compared across subsequence lengths and sampling rates. As there may exist emerging patterns at multiple scales, we can compute the Emergence Profile for all subsequence lengths within a range in order to determine appropriate subsequence lengths to focus on. If subsequence length m is too small, then we may report many sequential Novelets, though in reality, they represent pieces of a larger repeated pattern. If m is too large, then we are considering additional context of an emerging pattern, which may reduce its novelty score and prevent triggering Novelet discovery.

To demonstrate this, consider a snippet from an ECG containing some Premature Ventricular Contractions (PVC). Even healthy heartbeats can be challenging to model due to the variability in their beat length, but as Fig. 8 shows, there is high stability in the Emergence Profile across subsequence lengths. By choosing a sufficiently high novelty threshold, we can inspect an Emergence Profile without the effects of past learned patterns. In Fig. 8.*left* we show the "Pan-Emergence Profile" for an ECG from the Long-Term ECG Database with three instances of PVC [6]. The Pan-Emergence Profile is simply the Emergence Profile computed for every subsequence length in a given range, then plotted as a surface.

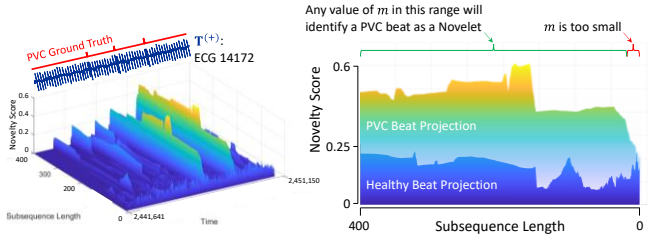

Fig. 8. *left*) A surface plot of the Emergence Profile which contains two high-valued ridges corresponding to the second and third PVC instances. *right*) All but the smallest subsequence lengths will result in detection of the PVC beats, suggesting that Novelets are highly robust to at least one of the two user-supplied parameters.

There are two regions with high Emergence Profile values which align with the second and third PVC instances. This is emphasized in Fig. 8.*right* where we see a projection of the healthy beats overlaid on a projection of the two PVC regions. This shows that all subsequence lengths in a large range have sufficiently high novelty threshold value. With a novelty threshold set to 0.25, only the PVC beats emerge as Novelets.

E. Online Novelets

The most compelling use case for Novelets is when used in an online fashion, allowing for the possibility of intervention to exploit or control the new behavior. There are a variety of definitions of "online", but we use the useful variant that incrementally updates $\mathbf{T}^{(-)}$ and $\mathbf{T}^{(+)}$ at real-world sampling rates such that computed values of $\mathbf{EP}^{m}$ are immutable.

The proposed immutability of $\mathbf{EP}^{m}$ arises from the usage of its subcomponents $\mathbf{LMP}^{(++),m}$ and $\mathbf{MP}^{(+-),m}$. First, a left Matrix Profile is naturally immutable as newly arriving subsequences are only compared with past subsequences. The AB-join Matrix Profile $\mathbf{MP}^{(+-),m}$ may appear to require updating the entire profile if $\mathbf{T}^{(-)}$ grows, but our insight is that $\mathbf{T}^{(-)}$ grows only as Novelets are discovered. Any behaviors appended to $\mathbf{T}^{(-)}$ results in suppression of that behavior. A discovered Novelet which influences past $\mathbf{MP}^{(+-),m}$ values would invalidate its own discovery. Therefore, it is admissible to omit updating past values of $\mathbf{MP}^{(+-),m}$.

Now that intuitions have been developed for how variables need to be updated, we can discuss the naïve algorithm that updates the OnlineNovelets in TABLE I. Following this, we show how the naïve update algorithm can be improved to support real-time discovery of Novelets.

TABLE I. THE ONLINENOVELETS ALGORITHM

Algorithm: OnlineNovelets($\mathbf{T}^{(-),old}$, $\mathbf{T}^{(+),old}$, $t^{(+)}$, $\mathbf{MP}^{(+-),m,old}$, $\mathbf{LMP}^{(++),m,old}$, $\mathbf{EP}^{m,old}$, m, d)

Input: negative time series $\mathbf{T}^{(-),old}$, positive time series $\mathbf{T}^{(+),old}$, a new positive time point $t^{(+)}$ following $\mathbf{T}^{(+),old}$, Matrix Profile $\mathbf{MP}^{(+-),m,old}$, Left Matrix Profile $\mathbf{LMP}^{(++),m,old}$, Emergence Profile $\mathbf{EP}^{m,old}$, subsequence length m, and novelty threshold d.

Output: The Emergence Profile $\mathbf{EP}^{m}$, the incrementally updated Matrix Profiles $\mathbf{MP}^{(+-),m}$ and $\mathbf{MP}^{(++),m}$, the current positive time series $\mathbf{T}^{(+)}$, the current negative time series $\mathbf{T}^{(-)}$, and a *novelet* if discovered.

```
1  novelet ← [ ]
2  T(+) = [T(+),old, t(+)]
3  last ← |T(+)| − m + 1      // index of newest subsequence of T(+)
4  NEW ← T(+)last,m           // newest subsequence in T(+) of length m
5  DP(+−),m ← MASS(NEW, T(−),old)      // Begin AB-join update
6  MP(+−),mlast ← Min(DP(+−),m)
7  DP(++),m ← MASS(NEW, T(+),old)  // Begin left self-join update
8  LMP(++),mlast, j ← Min(DP(++),m)   // Get min, argmin
9  EPmlast ← (MP(+−),mlast − LMP(++),mlast)/√2m   // Length norm to [0,1]
10 If EPmlast ≥ d
11     novelet ← T(+)j,m      // The first instance of NEW behavior
12     T(−) ← [T(−),old, novelet] // Suppress future occurrences
```

The algorithm begins in lines 1 and 2 by initializing an empty Novelet and appending the most recently arriving time point $t^{(+)}$ to $\mathbf{T}^{(+)}$. This newly arriving time point completes the newest valid subsequence *NEW*, located at index *last* in

lines 3 and 4. The query subsequence *NEW* and reference time series $\mathbf{T}^{(-),old}$ are input into MASS to generate a distance profile $\mathbf{DP}_{last}^{(+-),m}$ in line 5. For this and the following index assignments, it is assumed that indexing past the existing vector length results in a vector expansion. Here $\mathbf{DP}_{last}^{(+-),m}$ represents the distances between the latest subsequence from $\mathbf{T}^{(+)}$ and all known subsequences which should be suppressed from $\mathbf{T}^{(-)}$. In line 6 the distance to the closest matching subsequence is then appended to $\mathbf{MP}_{last}^{(+-),m}$. A similar process is completed for updating the Left Matrix Profile $\mathbf{LMP}_{last}^{(++),m}$ using the minimum distance of $\mathbf{DP}_{last}^{(++),m}$ in lines 7 and 8.

Now that the foundational Matrix Profile work is complete, line 9 subtracts the last distance value of $\mathbf{LMP}_{last}^{(++),m}$ from $\mathbf{MP}_{last}^{(+-),m}$ to produce the newest length-normalized value of the Emergence Profile $\mathbf{EP}_{last}^{m}$. This difference value is the novelty score. The algorithm returns the updated vectors here if the novelty threshold has not been reached. Though if *NEW* exceeds a novelty score of *d*, then in lines 10 to 12, *NEW's* nearest neighbor in $\mathbf{T}^{(+)}$ is returned as a Novelet and appended to $\mathbf{T}^{(-)}$ in order to be suppressed as a Novelet in future iterations.

In the simplified (for clarity of presentation) pseudo-code of TABLE I. , the time complexity of the OnlineNovelets algorithm is dominated by the O($n \log n$) MASS function, where n represents the larger of $|\mathbf{T}^{(+)}|$ and $|\mathbf{T}^{(-)}|$. After the arrival of b time points, the time complexity is O($bn \log n$). Eventually, when referencing some intermediate state, b is of equal magnitude to n, resulting in a time complexity of O($n^2 \log n$). By availing ourselves of the Matrix Profile, whose complexity is only O(n^2), we can approach a time complexity of O(bn) through processing newly arriving time points in sufficiently large batches of length b.

In practice, it can be difficult to gain intuition about an algorithm's computational demands without a concrete empirical grounding. To offer such a grounding we propose to consider the *maximum time horizon* (MTH): the answer to the question, "*How far back can the Emergence Profile refer before batch updates are slower than the sampling rate?*"

We consider the maximum time horizon of the two common sampling rates, run on an Intel® Core i7-9700 CPU at 3.00GHz with 32 GB of memory (full details at [25]).

- An Emergence profile initialized with ten periods in $\mathbf{T}^{(-)}$, with samples arriving at 100 Hz (a typical medical or human activity sampling rate), can reference back 4.4 hours when processing buffered input every minute.
- An Emergence profile initialized with ten periods in $\mathbf{T}^{(-)}$, with samples arriving at 1 Hz (a common industrial monitoring sampling rate), can reference back 3.2 years when processing buffered input every ten minutes.

The latter case is effectively *infinity*. Three years from now when we approach the MTH, Moore's law will allow us to reboot with a decade-long MTH. In the former case, we are close to being able to monitor a patient for an eight-hour operation recovery sleep. Moreover, we have a technique to gracefully degrade performance. As the dominating operations are the all-to-all pairwise comparisons, *time* will run out far before main *memory*. Since memory is of no consequence, it is possible to allow the OnlineNovelets algorithm to run until the MTH is reached, then transition to an *amnestic* mode where past time points are left out of future comparisons as new time points arrive. This slightly changes the meaning of Novelets to a behavior which emerges within a bounded time window. A behavior repeated with an offset greater than the maximum time horizon will then be considered an anomaly, though any previously discovered Novelets will continue to be suppressed. When past samples are no longer referenced, the past results may be written to disk in a parallel sub-process, allowing the OnlineNovelets algorithm to effectively run indefinitely.

A further investigation of theoretical MTH for various sampling rates is shown in Fig. 9.*left*. Real-time actionability is possible where the curves have a run-time gain (buffer-time divided by run-time) which is greater than or equal to 1.0. For example, the MTH for 1 kHz sampling is roughly 1.6 minutes ($|\mathbf{T}^{(+)}| = 10^5$). The subsequence length m and batch length for these curves were set to 100 and 1,000, respectively. In Fig. 9.*right*, the choice of batch length is tested when sampling at 100 Hz with $|\mathbf{T}^{(+)}| = 10^6$. We notice that the choice of 1,000 ($10m$) as batch length for Fig. 9.*left* was near the length of diminishing returns.

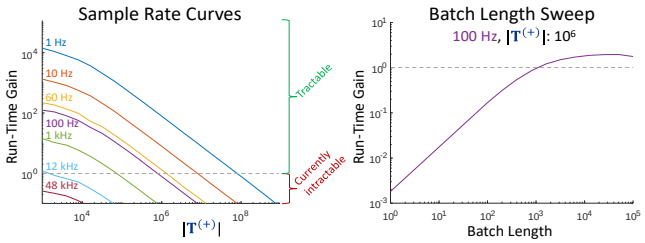

Fig. 9. *left*) For various sampling rates used in this work, we show the theoretical MTH as the point where the sampling rate curve descends below a real time gain of 1.0 (gray, dashed). *right*) We also explore the effect of buffer length on run-time gain.

III. Exploratory Data Mining

The discovery of emerging behaviors may be the end goal. Discovered Novelets can be reported to an end-user, and that may be the end of the analytical task. However, Novelets can potentially find patterns which have thus far gone unnoticed, and these patterns could be fed into any of dozens of downstream algorithms. Below we show some examples.

A. Electrooculogram

Time series telemetry from medical monitoring is often complex, leading to mostly feature-based or critical-limit monitoring. One such time series is an electrooculogram (EOG), which measures eye-related muscle movement. This is of significant interest, as eye movement can be a proxy for sleep-state and neurological disorders. We investigate the first record in the Sleep-EDF Database [6][7], which is roughly twenty-two hours long. When examining the top scoring Novelets with a three-second subsequence length, one Novelet was particularly interesting, as shown in Fig. 10.*top*.

The center of the Novelet subsequence is a familiar shape corresponding to an eye-blink. What is interesting about it is

that the leading and following sections are *also* part of blinks; by our now familiar analogy, ink**blink**blin. This is somewhat surprising, as blinks are normally thought to be random, one would not expect to see three in a row [9].

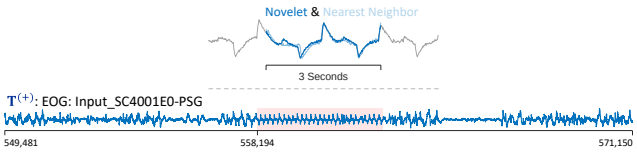

Fig. 10. *top*) One of the top ten EOG Novelets corresponds to an unexpectedly long sequence of eye-blinks. Context in gray. *bottom*) A three-minute snippet from a twenty-two-hour EOG recording. The highlighted region of interest is a highly periodic pattern in an otherwise visually noisy time series.

Intrigued by this result, we searched the rest of the data using this Novelet as the template (using MASS, Definition 3) and discovered the extraordinary region shown in Fig. 10.*bottom*. This region contains a series of twenty-six consecutive eye-blinks. We showed this data to Dr. Greg Mason who has more than forty years' experience in examining such data, and he found it to be astonishing. He had two theories that might explain this finding. The first is quotidian, perhaps a technician asked the patient to blink continuously to calibrate or test the recording apparatus. The second theory is that the patient has a rare neurological condition that produces a *fasciculation*. In either case, this demonstrates the utility of Novelets in the discovery of interesting and unexpected behaviors.

B. *Pedestrian Traffic*

Pedestrian traffic is a highly time-aligned behavior which often contains multi-scale periodicity. Motifs and discords are useful tools for uncovering the most conserved and most anomalous events within human behavior-based time series such as these multi-scale datasets [21].

Using Novelets, we have discovered that there are infrequent behaviors which are neither motifs nor discords. One such behavior, shown in Fig. 11.*top* occurs at Tin Alley-Swanston St (West) at the University of Melbourne [15], only on two Friday evenings at 4pm, specifically February 2, 2017 and April 21, 2017. Other than this day of week and time similarity, the Novelet and its nearest neighbor do not appear to share a standard weekly, semesterly, or yearly periodicity. The peculiarity of these two events is reinforced when examining the distance profile in Fig. 11.*bottom* between the three-year time series and the Novelet. There are clearly only two occurrences of this behavior.

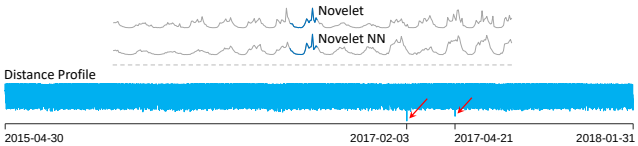

Fig. 11. *top*) The Novelet and its nearest neighbor. *bottom*) Distance profile between the three-year time series and the Novelet. The distance profile reveals one repetition of the Novelet.

Given the physical location of the sensor, and how the behaviors are non-periodic, yet both occur on Fridays at 4pm, they may be campus recruitment events or other pop-up events. While it is difficult to know for certain with only historic information, when running Novelets in real-time, it will be possible to flag such events as they occur in order to catalogue them for future reference.

In domains with a similar hourly sampling rate where behaviors are tracked at the 24-hour time scale, the maximum time horizon is essentially unlimited, where a thirty day buffer referencing the past thousand years takes less than a minute.

IV. Experimental Evaluation

To ensure that our experiments and figures are reproducible, we have built a website [25] that contains all the data/code used in this work. All experiments were conducted on an Intel® Core i7-9700 CPU at 3.00GHz with 32 GB of main memory, unless otherwise stated. We conduct experiments that demonstrate the following properties of Novelets:

- With high probability, Novelets can discover emerging behaviors, even in nosily "imperfect" data.
- As a corollary to the above, *zero* novelets will be returned if there is no true emerging behavior. Some algorithms always return *some* structure. For example, the Matrix Profile always returns the best motif, even in random data.
- Patterns in a time series which may be too subtle to be visually identified can still potentially be identified with Novelets. Of course, Novelets could still be useful if the patterns they discovered are *also* discoverable by human inspection. However, this seems to lack ambition. We would hope that Novelets could discover differences so slight that they would escape human attention.
- Novelets are robust to past false negatives such that if the first two instances of a behavior were not identified due to noise, either one is still a candidate in matching the k^{th} instance. This is an important property. The first occurrences of a new emerging pattern may be too tentative or noisy to be discovered, but we do not want processing them to spoil the chance of later discovering a slightly clearer version.

Our experiments are a mixture of somewhat anecdotal examples that show the diversity of problems that Novelets can be applied to, and more rigorous experiments designed to stress test our ideas. We begin with an example of the former.

A. *Bird Song Detection I*

Recall the motivating application in the introduction describing an observed change in the song of a white-throated sparrow. The song's rhythm changed from "*Canada*" to "*Cana*". We believe that Novelets in conjunction with existing wildlife monitoring efforts could detect instances of such a change in behavior.

To avoid constructing and potentially contriving a dataset to test this idea, we will simply use the audio of a ten-minute long video describing the phenomenon [10]. The video in question includes narration, other ambient sounds, and music, however we do not edit it in any way. We use a single 62.5 Hz Mel-Frequency Cepstral Coefficient (MFCC) as $\mathbf{T}^{(+)}$ and habituate to the narration in the first half of the video by

partitioning it as $\mathbf{T}^{(-)}$. When searching among two-second-long subsequences, four Novelets are identified: the song transcribed as "*Canada*", the song transcribed as "*Cana*", a solitary chirp, and silence preceding a higher amplitude sound. Fig. 12 shows the two Novelets corresponding to the two song compositions.

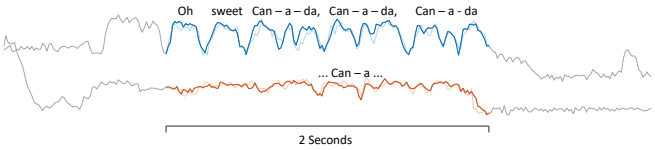

Fig. 12. The emergence of the two song compositions of the white-throated sparrow is detected using Novelets. Novelets and their second instances are shown in color with context shown in gray.

There is a quality difference between the two Novelets shown in Fig. 12. This is likely a result of two distinct birds singing at different pitches, which can appear quite different when decomposed as MFCC components. Nevertheless, this result is promising, given how little effort we needed to create this result, and the fact that we used no domain knowledge.

B. Bird Song Detection II

The previous section described an anecdotal example of the utility of our proposed idea. We will now more forcefully demonstrate the ability to identify a newly emerging bird song among a variety of other bird songs and forest sounds.

To demonstrate these properties in a principled manner, we have constructed two natural test sets with unambiguous ground truth, while also preserving real-world noise levels. Each test contains one hundred subtest iterations. Each subtest is composed of time series pairs, where $\mathbf{T}^{(-)}$ represents previously recorded forest sounds [16] known to lack the bird song of interest and $\mathbf{T}^{(+)}$ represents a similar distribution of forest sounds along with random insertions of four different instances of the bird song of interest. The bird songs of interest come from a white-crowned sparrow (*Zonotrichia leucophrys*) [19]. The bird songs are added to the forest audio with amplitudes that emulate a bird in the foreground in one test and in the background for the second test. The time series pairs are then decomposed into MFCC components, using default settings. The test sets were constructed so that the prior probability of correctly selecting any combination of two of the four songs of interest within a single subtest (i.e., the default rate) is less than 1%.

Identifying the background bird songs is non-trivial. The synthesized time series with four instances of background bird songs in Fig. 13.*top* shows the difficulty in identifying the target behaviors using visual inspection of amplitudes and shapes. The self-join Matrix Profile [21] is a standard tool for identifying motifs and discords within time series. The nearest neighbor distance of the identified Novelet is represented as a dashed gray line in Fig. 13.*center*. The nearest neighbor distance is a moderate MP value, indicating that it is neither a motif nor discord. The Emergence Profile successfully identifies the first and third instances of the target behavior, as confirmed by the alignments of ground truth and Novelet indicators in Fig. 13.*bottom*. The Emergence Profile of the test iteration shown in Fig. 13.*bottom* illustrates Novelets' robustness to false negatives. While Novelets fail to identify the target behavior within the first two instances, the third instance matched with the first instance.

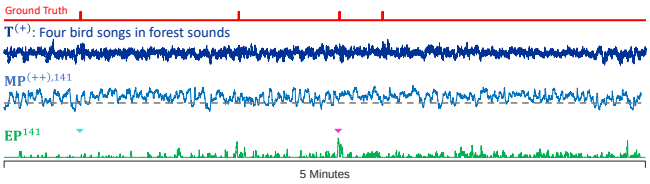

Fig. 13. *top*) Synthesized time series $\mathbf{T}^{(+)}$ with four embedded bird songs with locations indicated by ground truth annotations (red). *center*) Self-join Matrix Profile $\mathbf{MP}^{(++),141}$ with the nearest neighbor distance of the first instance indicted by a dashed line (gray). *bottom*) Emergence Profile $\mathbf{EP}^{141}$ with Novelet indicators at the first (cyan) and third instances (magenta) of the bird song.

The corresponding Novelet shown in Fig. 14 demonstrates that a Novelet may still be discovered with non-trivial noise levels between two behavior instances.

Fig. 14. The first and third instances of the embedded bird song, along with context (gray).

The test results demonstrate high utility of Novelets in the area of wildlife monitoring. Of the 100 background song subtests, the true positive rates for discovering the Novelet within the first two, three, and four instances are 13%, 33%, and 44% respectively. These are remarkably high considering the default rate of a single subtest is just 0.63%. These true positive rates drastically increase when simulating foreground bird calls. True positive rates with discovery within the first two, three, and four instances are 79%, 89%, and 89% respectively. While there may be multiple Novelets discovered within a subtest, for ease of analysis, only the first Novelet of each subtest was evaluated for accuracy.

Further analysis of the results demonstrates a valuable property of Novelets. Because Novelet discovery is triggered by a novelty threshold, it is possible that no Novelets will be returned. This means that a low true positive rate does not necessarily imply a high false positive rate. Fig. 15 demonstrates this with a large false negative margin for both background and foreground bird song tests.

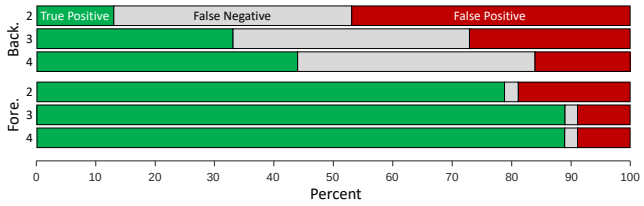

Fig. 15. *top*) True positive, false negative, and false positive rates shown for 100 iterations of identifying bird songs in the background within the first two, three, and four song instances. *bottom)* The test repeated for bird songs in the foreground.

Even at the lowest true positive rate of 13% when discovering a background bird song within the first two instances, Novelets demonstrates high repeatability considering the difficulty of the task. The prior probability of

positively identifying the first two instances in at least 13 of 100 iterations is 1.2×10^{-15}.

Finally, we consider scalability. Given an audio recording represented as MFCC at 60 Hz, a two second subsequence length, and twenty seconds of initialization, our algorithm will have an MTH of about twelve hours.

C. *Industrial Process Monitoring: Bearings*

Degradation of industrial components is often a gradual process where early prognosis can save large sums of money by reducing unplanned downtime. Typically, the sources of degradation such as wear, cracks, and corrosion are carefully modeled and tracked over a component's lifetime [13]. This requires detailed domain knowledge as the model parameters are unique to the application and setting. Our hypothesis is that *some* prognostic indications of wear may reveal themselves as newly emerging patterns. These novel patterns will not generally rise to the priority level of anomalies (which must be acted upon immediately) but can be reported to an engineer in say a daily or weekly report.

Case Western Reserve University's Bearing Data Center [3] provides datasets which capture component acceleration as bearings fail from seeded faults. The bearings are physically damaged in a variety of locations as a proxy for a range of failure sources. The dataset is split into four subsets: Normal and three types of failure. In this domain, *classification* accuracy is typically used to demonstrate the utility of an algorithm. Clearly Novelets are not a classifier. Here the question we will investigate is whether new repeated behaviors emerge as a process approaches a failure state.

In order to test this, we partition the first 5,000 samples of each time series as $\mathbf{T}^{(-)}$ in order to habituate to initial behaviors. With RPMs ranging between 1,720 to 1,797, the time series is analyzed with subsequences approximately the length of one rotation, which is about 400. As a manual training step, we observe that there are zero Novelets discovered within the Normal subset when the novelty threshold is set to 0.25.

Within the fifty-two failure configurations of Failure-Type-1 class, forty-five contained at least one Novelet, which is an error rate of 13.5%. Similarly, 53 of 60 failure configurations of Failure-Type-2 class contained at least one Novelet, which is an error rate of 11.7%. Finally, in the Failure-Type-3 class, 40 of 45 configurations contained at least one Novelet, which is an error rate of 11.1%.

The first two of ten discovered Novelets in a subtest of Failure-Type-1 are shown in Fig. 16. Each Novelet is overlaid on its second instance. Recall, the subsequence length corresponds to about one rotation of the bearing. Note how surprisingly well conserved each Novelet is. This is an interesting test because the question of whether a failing process is conserved for a long period, or quickly decays into chaos is not immediately obvious. It is likely that this is a domain dependent question, but in the case of bearing failure, our results strongly support the hypothesis that new behaviors emerge as failure approaches.

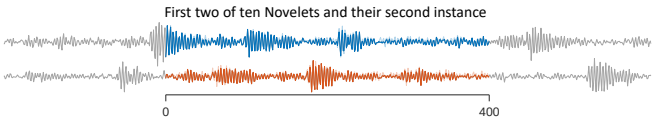

Fig. 16. Two discovered Novelets shown with context (gray) and overlaid on their second instance.

High speed industrial time series such as this tests the boundary of real-time actionability. The 12 KHz sampling rate here results in a computation time exceeding the arrival rate, though not by much. A 0.33 second buffer requires 1.26 seconds to process. As our current Emergence Profile is based on a vanilla Matrix Profile implementation, both future software optimizations, and current GPU-based Matrix Profile algorithms such as SCAMP [24] offer a path to achieve the needed 3x speed increase in the very near future.

D. *Industrial Process Monitoring: Buggy Machine*

The experiments in the previous section are satisfying in that they show promising results in a well-studied dataset. However, we do not have the ability to intervene in this system to test the limits of our ideas. To allow for more "hands-on" experiments, we constructed a simple apparatus for measuring the vibration of a motor.

Here we wanted to model a problem that we heard about anecdotally. Industrial motors that power fans are typically housed in enclosures to prevent the ingress of birds, bats and insects. Surprisingly, it is the latter that is most likely to cause damage. While insects are small, they are much more likely to be able to circumvent a screen barrier, especially if it has a small puncture[2]. Moreover, while the mass of an individual insect is small, the swarming behavior of many insects means that a fan with a damaged screen might ingest hundreds of insects in a single night. If enough insects are ingested into a fan, they can cause damage to the system in multiple ways.

We hypothesized that insects hitting a computer fan may produce Novelets when measured with accelerometers. As a proxy for real insects, we have chosen to use three weight classes of cotton balls. The small, medium, and large balls weigh about 50 mg, 240 mg, and 660 mg, respectively.

Fig. 17 summarizes our test, where we habituate on one minute of undisturbed fan motion, then drop each weight class six times from roughly three centimeters above the fan at twenty second intervals.

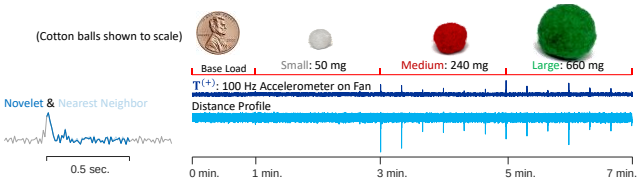

Fig. 17. *left*) The first Novelet discovered, shown with context in gray. *right*) Using the Novelet as a query on the accelerometer data produces a distance profile which detects all instances of the six medium and six large ball drops, confirming that the Novelet detected was indeed caused by a collision.

[2] This story is reminiscent of, but is distinct from, the famous story of the discovery of the first computer "bug" (a moth) by Dr. Grace Hopper in 1945.

A half-second Novelet discovers all the medium and large instances, but the small weight class is too light to register as a repeated acceleration shape. It is unlikely that a housefly (*Musca domestica*, 12 mg [20]) collision would register as a Novelet. However, these results suggest a larger insect collision *such as* a honeybee (*Bombus californicus*, 120 mg [20]) *would* likely register as a Novelet.

V. Model Comparison

A. Comparison to Segmentation

In some cases, Novelets may perform a function similar to semantic segmentation [5]. When a time series experiences a regime change, a Novelet may be discovered in the first few instances of a new behavior. Such a regime change may occur in patients experiencing onset of *Pulsus Paradoxus* where there is a physiological change in heart function. An existing example of this is found in [5], which we explore in Fig. 18.

Thus far, we have downplayed the importance of the novelty threshold due to claimed stability over a range of threshold values. Here, we reinforce this with a visually intuitive technique for identifying segmentation points without having an intuition for choice of threshold value. By sweeping over threshold values and plotting locations of Novelet indices, we see regions with high Novelet stability.

Each row in Fig. 18.*bottom* represents Novelet indices for a given threshold value. One strategy for reasoning about Fig. 18.*bottom* is to choose the index region with greatest frequency of discovered Novelets, which is 10,043. This is within just two heartbeats of the ground truth change point.

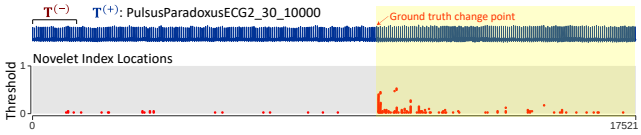

Fig. 18. *top*) A time series of Pulsus Paradoxus with a visually non-trivial change point located at index 10,000. *bottom*) A threshold parameter sweep of Novelet indices. There is a stable Novelet index within two subsequence lengths of the ground truth.

This example is impressive in that the semantic change is impossible to see with the naked eye, and is only known due to out-of-band data available to the attending physician.

VI. Conclusions

We have introduced Novelets, a primitive that allows online discovery of emerging behaviors. We have shown the actionability of Novelets in real-time monitoring of medical telemetry, wildlife, and industrial processes. We have also demonstrated our algorithm's low sensitivity to the two user-set parameters: subsequence length and novelty threshold, providing a low learning curve for the operator and high stability of results. By defining our ideas explicitly in terms of the Matrix Profile, we can leverage future work by the MP community. We have made all data and code available [25]. In future work we plan to investigate multidimensional Novelets in order to increase their actionability.

Acknowledgements

We acknowledge funding from NSF award IIS 2103976.

References

[1] N. Begum and E. Keogh, "Rare time series motif discovery from unbounded streams," Proc. VLDB. vol. 8:2, pp. 149–160, 2014.
[2] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A Review on Outlier/Anomaly Detection in Time Series Data," ACM Comput. Surv., vol. 54, no. 3, p. 56:1-56:33, Apr. 2021.
[3] "Case Western Reserve University Bearing Data Center" School of Engineering, Aug. 05, 2021. (accessed Apr. 19, 2022) https://engineering.case.edu/bearingdatacenter.
[4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Surv., vol. 41, no. 3, p. 15:1-15:58, Jul. 2009.
[5] S. Gharghabi, Y. Ding, C.-C. M. Yeh, K. Kamgar, L. Ulanova, and E. Keogh, "Matrix Profile VIII: Domain Agnostic Online Semantic Segmentation at Superhuman Performance Levels," in 2017 ICDM, Nov. 2017, pp. 117–126.
[6] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet," Circulation, vol. 101, no. 23, pp. e215–e220, Jun. 2000, doi: 10.1161/01.CIR.101.23.e215.
[7] B. Kemp, et. al., "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," IEEE Transactions on Biomedical Engineering, vol. 47, no. 9, pp. 1185–1194, Sep. 2000.
[8] E. Keogh and J. Lin, "Clustering of time-series subsequences is meaningless" *Knowl Inf Syst*, vol. 8, no. 2, pp. 154–177, Aug. 2005.
[9] R. W. Lawson, "Blinking and Sleep," Nature, vol. 165, no. 4185, Art. no. 4185, Jan. 1950, doi: 10.1038/165081b0.
[10] LesleytheBirdNerd, The White-throated Sparrow | Adorable Songster of the North, (Jun. 04, 2021). Accessed: May 02, 2022. [Online Video]. Available: https://www.youtube.com/watch?v=KsBj5nL0yUs
[11] R. Mercer, S. Alaee, A. Abdoli, S. Singh, A. Murillo, and E. Keogh, "Matrix Profile XXIII: Contrast Profile: A Novel Time Series Primitive that Allows Real World Classification," in 2021 ICDM, pp. 1240–45.
[12] A. Mueen et al., "The fastest similarity search algorithm for time series subsequences under Euclidean distance," 2015. (accessed Jan. 18, 2021). www.cs.unm.edu/~mueen/FastestSimilaritySearch.html.
[13] A. Muller, et. al "Formalisation of a new prognosis model for supporting proactive maintenance implementation" Reliability Engineering & System Safety, vol. 93, no. 2, pp. 234–253, Feb. 2008.
[14] K. A. Otter, A. Mckenna, S. E. LaZerte, and S. M. Ramsay, "Continent-wide Shifts in Song Dialects of White-Throated Sparrows," Current Biology, vol. 30, no. 16, pp. 3231-3235.e3, Aug. 2020.
[15] Pedestrian Counting System, "City of Melbourne - Pedestrian counting system, 2013. www.pedestrian.melbourne.vic.gov.au/#date=28-10-2021&time=8 (accessed Oct. 27, 2021).
[16] TheSilentWatcher, 4K Forest Birdsong 2 - Birds Sing in the Woods - No Loop Realtime Birdsong - Relaxing Nature Video, (Sep. 25, 2017). Accessed: May 02, 2022. [Online Video]. Available: https://www.youtube.com/watch?v=XxP8kxUn5bc
[17] P. Thornton, "Digoxin Uses, Dosage & Side Effects," Drugs.com, Dec. 06, 2021. www.drugs.com/digoxin.html (accessed Mar. 08, 2022).
[18] C. Wetzel, "Sparrows are singing a new song, in a rapid, unprecedented shift," Animals, Jul. 02, 2020. https://www.nationalgeographic.com/animals/article/new-sparrow-birdsong-replaces-old-tune (accessed Mar. 08, 2022).
[19] White-crowned Sparrow (audio recording). Retrieved May 5th 2022. Recordist Ian Cruickshank. https://xeno-canto.org/251101
[20] Wolfram|Alpha. https://www.wolframalpha.com (accessed May 10, 2022). with query [weight of Bombus californicus], and query [weight of Musca domestica].
[21] C. M. Yeh et al., "Matrix Profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets," in 2016 IEEE 16th ICDM. 2016, pp. 1317–1322.
[22] C. M. Yeh, Y. Zhu, H. A. Dau, A. Darvishzadeh, M. Noskov, and E. Keogh, "Online amnestic DTW to allow real-time golden batch monitoring," in ACM SIGKDD. 2019, pp. 2604–2612.
[23] Y. Zhu, M. Imamura, D. Nikovski, and E. Keogh, "Introducing time series chains: a new primitive for time series data mining," Knowl Inf Syst, vol. 60, no. 2, pp. 1135–1161, Aug. 2019.
[24] Z. Zimmerman et al., "Scaling Time Series Motif Discovery with GPUs: Breaking the Quintillion Pairwise Comparisons a Day Barrier," in Proc. ACM Symp. Cloud Comput. 2018.
[25] Novelets Supporting Website: https://sites.google.com/view/novelets