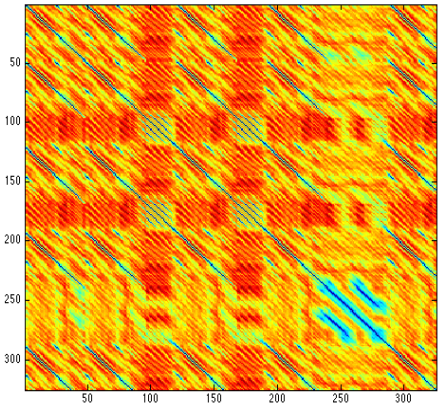
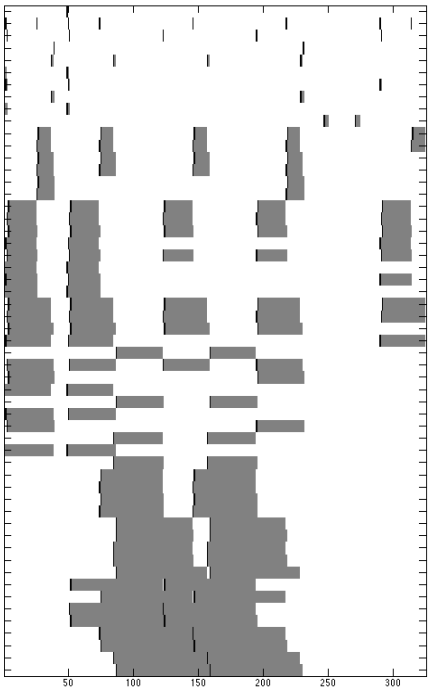
Research Statement - Katherine M. Kinnaird

My research is in applied mathematics, at the intersection of network theory and machine learning. My work is motivated by the following question: *How can high-dimensional and noisy data be represented by a low-dimensional object that encodes relevant and size-appropriate information?* Addressing this question draws naturally on theory from numerical linear algebra, statistical learning, complex networks, algebraic combinatorics, graph theory, and machine learning. Specifically, I work on addressing this question for *sequential data*; that is, data comprised of a sequence of observations where the order of the observations is as important as the individual observations themselves [1].

Figure 1: Visualizations of Representations for Score of Chopin's Mazurka Op. 6, No. 1



(a) Self-Dissimilarity Matrix



(b) Aligned Hierarchies

From GPS data on our phones to recordings of songs played on our radios, we interact with and compare sequential data streams every day. With the sheer volume of data created everyday, finding quick and computationally feasible methods for comparing data, including sequential data, is crucial to unraveling meaning from data and for data-based recommendations and predictions. We, as humans, base our comparisons of sequential data streams on the patterns found within them. For example, when listening to a song, we may note a chord progression found in the chorus is also found in the verse and the bridge, and also that this chord progression is not part of another song. This natural human comparison based on patterns found in sequential data streams is the motivation for my methodology for comparing sequential data.

My research program builds a methodology for comparing high-dimensional sequential data that can be broadly applied to many questions from a range of fields. My work has three interconnected components: a theoretical component, an application of my theoretical work to questions about music, and an open-access software component implementing the theoretical work. First, I build a low-dimensional representation for a data stream that mimics how humans notice, interpret, and understand patterns in the world around them. This representation is an algebraic object that can be embedded into a space with a natural distance function. With this framework, I have developed a representation with an associated metric that can compare sequential data

streams of the same size as well as representations that compare sequential data streams based on identically sized sections of the data streams. Second, noting that music is a natural example of sequential data, I address questions about songs, referred to as tasks in the interdisciplinary field of Music Information Retrieval (MIR), by applying my method of comparing sequential data streams using their representations. Lastly, I am building an open-access suite of software tools that implements my methodology for comparing high-dimensional sequential data. This software seeks to be accessible to any researcher working with high-dimensional data and will include examples from a range of domains, presented in the context of questions posed by researchers in those fields.

The significance of my work lays in the fact that, given a self-dissimilarity matrix presentation of a sequential data stream, the construction of my low-dimensional representation is domain neutral; that is, I do not use assumptions from any specific field to form these representations. Therefore, comparison techniques using these representations can be applied to self-dissimilarity matrices arising from any high-dimensional sequential data, such as medical data or gene sequencing data [1], as well as to recurrence plots [7]. Additionally, having proven that these representations can be embedded into a space with a natural metric, I have an objective method for comparing sequential data streams, even those based on cultural artifacts such as recordings of songs. The domain neutrality of my work does not exempt me from engaging with data domains. In fact, I developed a mathematical framework for a musically meaningful interpretation of a threshold commonly used in MIR [5].

Theoretical Work - Aligned Hierarchies

Comparing sequential data streams with the same number of observations begins by building *aligned hierarchies*, a low-dimensional object that synthesizes all possible hierarchical decompositions of repeated patterns, for each sequential data stream [4, 6]. To do this, I use self-dissimilarity matrices created from *overlapping feature vectors* (vectors of statistics derived from overlapping musical intervals) of sequential data streams, like songs. Figure 1(a) is a visualization of a self-dissimilarity matrix created from overlapping feature vectors extracted from the score of Chopin's Mazurka Op. 6, No. 1; the information that I seek to extract is contained in the dark blue diagonals. Using this matrix, I create the aligned hierarchies for each song by identifying repeated structures at a variety of scales; for an example, see Figure 1(b). I use these aligned hierarchies to create a network representing my set of songs and use this network to perform analysis of the data set.

To build the aligned hierarchies, a novel representation for sequential data streams, I find all repeated elements at a variety of scales, by using a mix of thresholding, searching, and inclusion/exclusion techniques. I then organize the found patterns into one algebraic object, denoting where each pattern occurs in the sequential data stream and which instances of structure are repeats of each other. We can visualize the aligned hierarchies as a two-dimensional object with the found patterns arranged along the observation ordering of the sequential data stream in rows with the largest patterns at the bottom. Figure 1(b) is a visualization of the aligned hierarchies for the score of Chopin's Mazurka Op. 6, No. 1. The gray blocks in the visualization represent the full duration of the repetition, and the black line at the beginning of each gray block denotes where the repetition starts in that score. The y-axis of Figure 1(b) denotes the type of repeated structure.

Comparing Sequential Data Streams of the Identical Lengths

The aligned hierarchies can be embedded as elements of a product space comprised of discrete quotient spaces each endowed with the same metric, as I demonstrated in [4, 6]. An element of this product space is a sequence of binary matrices, with one matrix for each positive integer. The k^{th} element in this sequence is a matrix encoding the repeated structure of exactly k units in the sequential data stream. Each component space of the product space is the quotient space arising from a row permutation equivalence relationship on $M_{m\times n}(\mathbb{Z}_2)$, because the aligned hierarchies group repetitive structure such that permuting rows encoding structure of the same length does not fundamentally change the representations. The metric on each quotient space is the minimum of the entry-wise 1-norm between one element and all row permutations of a second one [4, 6]. I proved that this matrix function is indeed a metric, by creating a leveled graph whose vertices on level i are the orbits with entry-wise 1-norm equal to i and whose edges connect vertices between level i and $(i+1)$ when an element of an orbit on level $(i+1)$ can be obtained by adding a 1 to an element in an orbit on level i. The metric on this product space of identical quotient spaces is the sum of the distances for each quotient space and provides a measure of total dissimilarity for pairs of aligned hierarchies.

Comparisons based on aligned hierarchies are total comparisons of one whole sequential data stream to another. Such comparisons can address rigid comparison tasks like the *fingerprinting task* in music information retrieval, which seeks to find all copies of the same recording of the same performance [2]. However, these total comparisons will fail to address more nuanced comparisons, like the cover song task which seeks to find all performances of the same piece of music, regardless of the performer [2, 3, 8, 9].

Comparing Sequential Data Streams of the Differing Lengths

Sequential data streams are rarely the same length; therefore, it is essential to develop techniques that can compare sequential data streams of differing lengths. Currently, there are a number of pre-processing techniques that alter the data streams to be all the same size. These techniques range from truncating data streams or padding them with zeros to *resampling*, a process that alters the widths of each of the observations either by smoothing several observations together or by selecting one observation from a region based on some criterion (for an example, see [3]). Such pre-processing techniques do create sequential data streams of the same size, but each can introduce or lose important artifacts in the data stream. Furthermore, there is little theoretical work examining the effects of such pre-processing techniques.

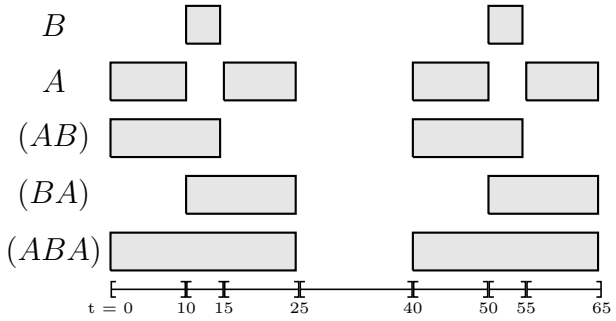
In contrast to these methods, I build mathematical objects that address the challenge of comparing sequential data streams without altering the width of the data first. Furthermore, noting that humans can compare sequential data streams based on elements of a sequential data stream, I have developed post-processing techniques that are applied to the aligned hierarchies and lead to comparisons based on comparing sections of sequential data streams. These post-processing techniques create low-dimensional representations of sequential data streams that are embeddable into a space with a natural metric [4].

One such post-processing technique isolates different repeated structures in a given sequential data stream and then finds the individual aligned hierarchies for each isolated re-

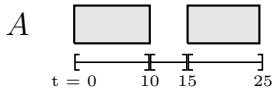
peated structure. The collection of these aligned hierarchies for each sequential data stream is called the *aligned sub-hierarchies* and, like the aligned hierarchies, the aligned sub-hierarchies can be embedded into a space with a natural metric.

For example, consider a sequential data stream with the aligned hierarchies shown in Figure 2(a). In this aligned hierarchies, the (ABA) structure is the only one with a smaller structure repeated within it and thus the only structure with its own aligned hierarchies. So the aligned sub-hierarchies for the sequential data stream with the aligned hierarchies shown in Figure 2(a) is comprised of only the aligned hierarchies for the (ABA) structure.

Figure 2: Visualization of aligned hierarchies and aligned sub-hierarchies for sequential data stream with segmentation ABACABA



(a) Aligned Hierarchies



(b) Aligned sub-Hierarchies

Song Comparisons via Aligned Hierarchies

Using the aligned hierarchies for songs, I address MIR tasks for data based on Mazurkas by Chopin. Specifically, I have applied my techniques to two song comparison tasks: the fingerprint task and the cover song task.

My experiments on data based on human-coded digital representations of scores have provided a proof of concept for the aligned hierarchies [4]. These experiments demonstrate that without any post-processing, I can address the fingerprint task using the aligned hierarchies as the basis of my song comparison and that with minimal post-processing, I can use the aligned sub-hierarchies to address the cover song task.

Thus far, my experiments addressing the cover song task performed on data based on the recordings of the performances of scores have been less successful than those performed on the data based on scores; these results provided insight into the sensitivity of the comparisons made using the aligned sub-hierarchies. As the experiments performed on the audio data suggest, the results of my approach are sensitive to the pre-processing techniques applied to the original data [4].

Open Access Software - Implementing Aligned Hierarchies

In the third part of my research, I am building open-access software that implements my theoretical work and provides tangible examples from a variety of domains. This suite of software will have implementations in a variety of programming languages and will be accessible to researchers in domains beyond my expertise. In creating this software, I will forge new collaborations with researchers beyond MIR and machine learning, with the intention that these collaborations will yield results in their domains of expertise while also forming examples to be part of the software packages. This kind of open-access software with examples from a variety of domains will encourage my methodology to be adopted in other fields that address questions about high-dimensional sequential data.

My three-part research program is based on a domain neutral methodology for comparing high-dimensional sequential data. My work is multidisciplinary, currently using theory and techniques from mathematics, statistics, machine learning, and music information retrieval and with the intention of including additional domains through productive interdisciplinary collaborations. I also seek to provide accessible and credible software for researchers in any domain addressing questions about high-dimensional sequential data.

Beyond MIR - Atypical Collaborations Beyond STEM

During my time in the Culture Analytics Long Program at the Institute for Pure and Applied Mathematics, I began collaborating with Prof. John Laudun of the Department of English at the University of Louisiana at Lafayette. Our project examines the influence of ideas among TED talks. Working with the transcripts of all the talks given at the main TED event, our work draws on ideas from topic modeling and network science as well as concepts from digital humanities. Currently we are exploring the popularity of topics over time as well as based on the speaker's gender.

This collaboration is emblematic of the multidisciplinary collaborations that I seek to build both in my research and in my classrooms. By working with those beyond the typical science and social science departments, I am on the cutting edge of research in cultural analytics and am better equipped to design research programs that support truly multidisciplinary work.

Research with Undergraduates - Training the Next Generation

Training the next generation of researchers in the skills required to engage in interdisciplinary research, like mine, and data science research is a necessity. This training can take two forms. In the first, more traditional approach, students could engage in faculty-directed research with the goal of publishing novel work. My diverse research program, with components in pure mathematics, applications to MIR, and software development, could support research projects with students interested in this more traditional undergraduate research setting. For example, an upper-level pure mathematics student could work on a project investigating further mathematical properties of the aligned hierarchies and their derivatives. Another example could be that student with interests in sequential data and biology could develop an authentic example of my theoretical work being applied to biological data that would be included in the software implementation of my theoretical work.

As an alternative to the traditional undergraduate research model, students could work in a collaborative setting that models the research process. In this second approach, students engage in student-directed research projects that further develop their research skills. During my postdoctoral work at Macalester College, I started the Data Science TRAIn Lab, a program that trained students to excel in data science research by encouraging students to (T)ry, (R)ead, (A)sk, and (In)corporate. As its name implies, this lab focused on the research process while simultaneously exposing students to the diversity of concepts, theories, and techniques found in data science and machine learning.

References

[1] C.M. Bishop, *Pattern recognition and machine learning*, Springer Science+Business, New York, NY, 2006.

[2] M. Casey, C. Rhodes, and M. Slaney, *Analysis of minimum distances in high-dimensional musical spaces*, IEEE Transactions on Audio, Speech, and Language Processing **16** (2008), no. 5, 1015 – 1028.

[3] P. Grosche, J. Serrà, M. Müller, and J.Ll. Arcos, *Structure-based audio fingerprinting for music retrieval*, 13th International Society for Music Information Retrieval Conference (2012).

[4] K. M. Kinnaird, *Aligned hierarchies for sequential data*, Ph.D. thesis, Dartmouth College, 2014.

[5] ———, *Musically meaningful thresholds for chroma-based audio features*, Pre-print (2015).

[6] ———, *Aligned hierarchies: A multi-scale structure based representation for music-based data streams*, 17th International Society for Music Information Retrieval Conference (2016).

[7] N. Marwan, M.C. Romano, M. Thiel, and J. Kurths, *Recurrence plots for the analysis of complex systems*, Physics Reports **438** (2007), no. 5-6, 237–329.

[8] J. Serrà, *Identification of versions of the same musical composition: Audio content-based approaches and post-processing steps*, Lambert Academic Publishing, Saabrücken, Germany, 2011.

[9] J. Serrà, M. Zanin, and P. Herrera, *Cover song networks: Analysis and accuracy increase*, International Journal of Complex Systems in Science **1** (2011), 55–59.