Rodica Neamtu web: *http://web.cs.wpi.edu/ rneamtu/* rneamtu@wpi.edu

Research Statement

1 Previous and Current Research

My current research addresses the questions "**How can we get insights into time series collections, based on using similarity models to help answer complex questions? How can we expand and efficiently use similarity models for data discovery?** ".

My *current research* work towards answering these questions is anchored on introducing novel frameworks for designing complex similarity distances, and incorporating them in efficient and accurate tools for data discovery. My *previous research* tackled the critical problem of automating data integration from a variety of public websites by abstracting key features of multi-dimensional tables and interpreting them in the context of a knowledge-centered Unified Spatial Temporal Model. The classification-driven extractors I developed are trained to identify and classify entities from both structured and unstructured parts of spreadsheets. *Together, these two broad areas* contribute to creating an end-to-end solution to integrate data acquired from heterogeneous public resources and transform it into a unified model upon which newly designed, powerful, yet highly efficient analytics techniques are applied.

My *future research* will focus on creating tools to offer both descriptive and predictive data mining capabilities. These tools will provide the ability to explain the data and extract interesting properties and interrelationships. Domains such as astronomy, finance, e-commerce and genome sequencing, are currently collecting a staggering amount of data, a significant part of which is in the form of data series. To make sense of it, scientists need to interactively explore these time series, by formulating hypotheses and progressively refining them. My research focuses on adding expressive exploratory mechanisms to big time series collec-

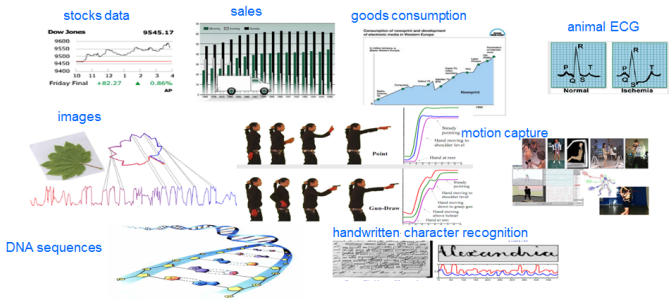

Figure 1: Time series in diverse domains. Image from E. Keogh. A decade of progress in indexing and mining large time series databases. (VLDB 2006.)

tions. To this end, I developed interactive tools that allow analysts to explore similarity and find best match sequences and patterns in very large datasets using different similarity distances. I also introduced novel techniques for visualizing high-cardinality query results. Such results are crucial to answer complex economic, financial, medical and societal questions. For example, a doctor can find the patterns immediately preceding a heart attack in a patient by identifying similar existing patterns in a multi-terabyte ECG dataset, or an analyst can find stocks having a similar growth compared with the Apple stock over a time period.

Finding similar trends and patterns among time series data is critical for many applications ranging from financial planning to policy making, as shown in Figure 1. A successful

data discovery system must be able to efficiently mine large time series collections of heterogeneous types, from multiple sources while allowing flexible interpretations provided by different parameters. This challenge raises some fundamental questions:
How do we *automatically integrate data* from heterogeneous data sources?
How do we *discover and extract* important insights from data?
How do we *perform "unified" analytics"* and enable users best interpret the data?
How do we capitalize on the data insights and use them for *predictive tasks*?
My existing research provides answers to the first three questions, while my future research will refine these answers and find solutions for the last question. I will highlight my work in the following areas:
(1) Automated integration of spatial temporal data using identification and classification.
(2) Interactive exploration of large time series datasets including the introduction of new framework for designing similarity distances and incorporating them in data reduction models.

(1) Automated integration of spatial-temporal data using identification and classification. Public web data sources include the Tax Policy Center[1] which contains information related to tax policies, rates and trends, the Census Bureau [2], reporting information about demographics, the National Science Foundation[3], the Bureau of Economic Analysis[4]. They represent valuable public knowledge ready to be leveraged for policy decision making and economic forecasting. The extraction and integration of such data is challenging and time consuming. Yet, the appetite for leveraging new data sources appears endless, so automation is critical to the success of building and growing rich economic indexes [6,7].
My Data Integration through Object Modeling framework (DIOM)[4,5] tackles the critical problem of automating data integration from a variety of public websites by abstracting key features of multi-dimensional tables and interpreting them in the context of a knowledge centered Unified Spatial Temporal Model. The classification-driven extractors are trained to identify and classify entities from both structured and unstructured parts of spreadsheets. The unstructured part contained in titles, headers and footers reveals critical information, so-called *implicit knowledge*, crucial to the correct interpretation of data. This "implicit knowledge" is used to automatically extract, integrate and transform data from heterogeneous public data sources by leveraging a spatial temporal model conceptualizing on the main entity types present in a large class of datasets.

(2) Interactive exploration of large time series datasets. My research work focuses on the detection of relationships between and among large time series data sets by tackling the challenge of inherent high cardinality of data and the complexity of the process of mining it. The need for flexible interpretations of similarity through parameter tuning as well as recommendations for similarity distances and thresholds is also addressed. Figure 2 offers a high level view of my generalized model for exploring time series similarity. Within this scope, my research focuses on four areas namely, interactive exploration of time series similarity, generalized similarity models, comparative analysis of the impact of various distances on similarity and interactive visual analytics.

- **Interactive exploration of time series similarity.** I introduced a novel paradigm

[1] www.taxpolicycenter.org/
[2] www.census.gov
[3] www.nsf.gov/
[4] www.bea.gov/

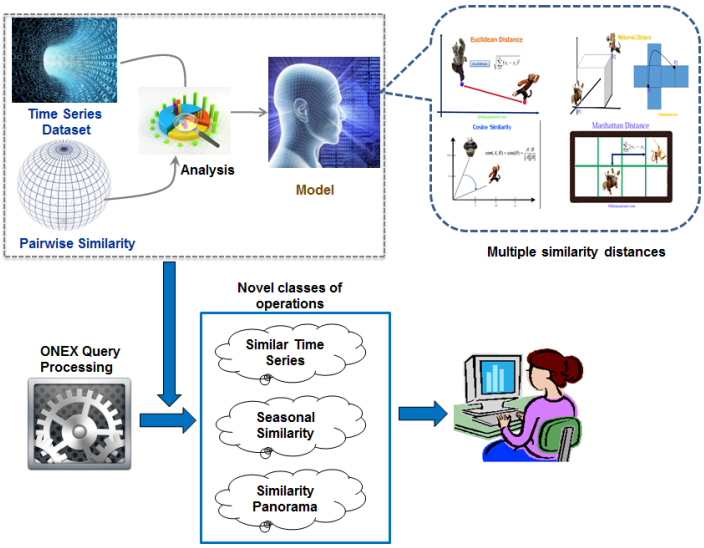

Figure 2: Generalized model for exploring time series similarity

called Online Exploration of Time Series (ONEX [1]) that employs a powerful one-time pre-processing step to compress the raw data into a compact knowledge base encoding critical similarity relationships among time series. This ONEX framework takes advantage of the computationally inexpensive Euclidean Distance for the construction of the ONEX base, yet the online explorer supports powerful time-warping using DTW[5] to facilitate the comparison of sequences of different lengths and flexible alignment within a few seconds of response time. My unique ONEX solution overcomes the prohibitive computational costs associated with this complex distance by using it over the surprisingly compact ONEX base instead of the raw data. ONEX emerges as a truly interactive time series exploration system. This unique approach based on the combination of two similarity distances leads to improvements in accuracy of up to 20% and up to 4 times shorter time responses compared to the fastest known state-of-the-art method.
ONEX renders the exploration of large time series datasets more practical and helps analysts better understand the similarity of time lines by supporting rich classes of operations. The ONEX query processor implements strategies for efficiently answering

Figure 3: Examples of answers that ONEX can provide

complex classes of questions from diverse domains. These classes include traditional similarity exploration, finding similarity patterns and offering guidance and parameter tuning. *For example,* as shown in Figure 3, using ONEX, a financial analyst can retrieve the stock similar to the stock fluctuations of the Apple Stock for a specific time period. Or, looking for repeating patterns, a doctor can find all 30 minutes long subsequences of

[5]Berndt et all, Using Dynamic Time Warping to Find Patterns in Time Series, In KDD workshop,1994

a patient ECG having similar shapes.

- **Generalized similarity models.** While analysts prefer domain-specific distance measures for exploring similarities among time series, these tend to be "point-to-point" distances. The point-wise nature limits their ability to perform comparisons among sequences of different lengths and alignments. Analysts thus instead must utilize "elastic"

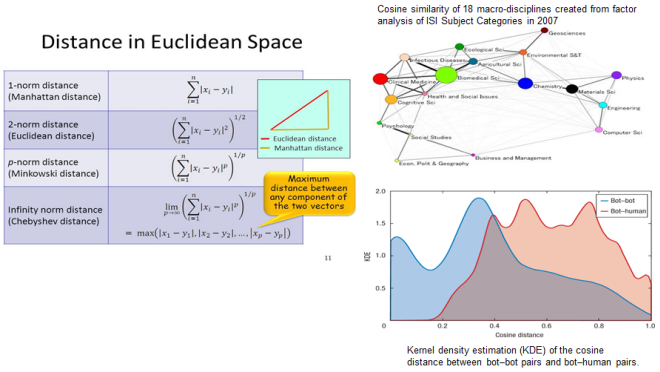

Figure 4: General and domain-specific distances that can be warped by GDTW framework

distances like Dynamic Time Warping (DTW) that enable flexible comparisons among such sequences. However, this is at the cost of elastic distances not incorporating the most suitable distances for their specific applications. To tackle this shortcoming, we introduced the first conceptual framework called Generalized Dynamic Time Warping (GDTW) [2] that supports warping of a large array of domain-specific similarity distances. While the classic DTW and its prior extensions utilize the Euclidean Distance for warping, this is *the first work* to generalize the ubiquitous DTW distance and "extend" its warping capabilities to a diversity of point-to-point distances. The GDTW framework is shown to support distances based on averages, max, min, fractions, square roots, (like the ones displayed in Figure 4) and their combinations – covering a wide range of popular functions for many different domains. Better yet, our time-warping framework efficiently computes these new warping paths by adapting the dynamic programming principles from DTW to this new context. Through extensive evaluation studies on numerous public datasets, we empirically showed that these generalized time warping distances produce interesting results and the ability to get more flexible similarity interpretations.

- **Efficient knowledge discovery in time series datasets powered by multiple distances.** We extended the work for interactive exploration of time series with a new paradigm called General Exploration of Time Series using Multiple Distances, or in short GENEX [3]. GENEX provides deep insights into time series datasets by revealing new data relationships in the rich context of using combinations of distances. This work ties together my previous two research efforts by combining specific distances with their "time-warped" counterparts, offering a novel mechanism for exploring time series similarity in specific application domains.

- **Interactive visual analytics.** We designed tools for interactive visual analytics to help analysts get insights into their datasets, as well as parameter tuning guidance contributing to better understanding and interpretations of similarity. Our new analytic interactive dashboard bridges the gap between the growing disparity between the volume of

time series data produced and the current capacity of domain experts to understand this data. Such visual analytics enable users to explore and interact with time series data sets and offer guidance and refinement of similarity parameters. Users can interactively construct rich classes of comparative queries to find insights in large time series data sets. Diverse visualizations further support interpretation of the results of matches.

2 Future Research Goals

My expertise in time series similarity exploration provides the solid foundation upon which to build new theoretical frameworks and practical solutions. Vital breakthroughs in data discovery are needed to understand the complexity of Big Data and explore hidden correlations to get insights that can lead to the much needed answers in diverse application domains. New generalized solutions are needed to offer more flexibility in interpreting Big Data and use it to make informed decisions. I will capitalize on my PhD dissertation work and my previous research experience to further data discovery in large complex, heterogeneous datasets.

I hope to develop a world-class reputation in data discovery as applied to diverse applications including economy, health, education and the environment. My background and life experiences lead me towards problems with large societal impacts like improving human health and achieving a deeper understanding of the impact of decisions on the economical and social health. Currently the research community is preoccupied with many different aspects of data discovery, including similarity and correlation exploration, and predicting future values and trends. I plan on expanding my research from the similarity exploration to finding and interpreting data correlations and use that to predict future values and trends.

I will outline below some future opportunities that I am excited to pursue:

(1) *Interactive mining of medical data.* I will continue and expand my current research on time series similarity to include motif discovery and rule discovery, while also devising and implementing new techniques to improve the performance of my existing similarity exploration techniques. For example, I plan on devising algorithms for speeding up the computations of general warped distances by exploiting lower bounds applicable over large classes of distances. Such performance improvements can render my ONEX-MD system a viable tool for mining large datasets in medicine. Generally, a typical dataset for a functional MRI scan can take up to several Gigabytes per person, providing the measurements on sub-regions of the brain at a spatial resolution of 1 to 5 mm per voxel, and a temporal resolution of one scan per seconds or so. In addition to fMRI, there are also other types of medical images (e.g., EEG, PET, MRI, DTI, CT, CAT, MEG) providing multiple views of the patient. Moreover, doctors may also have measures on thousands of other bio-markers of the patient, such as blood markers, anti-bodies, virus-levels, RNAs etc. Clinicians are interested in using all these measures (imaging and bio-markers) to map the brain as well as the blood system to detect the effects of stroke, brain injury, or diseases such as Alzheimer's and ADHD. I am interested in efficiently investigating these time series data provided by these heterogeneous sources of medical data, figuring out the relationships among different components and doing it efficiently, so doctors can benefit from these results by getting real-time answers to their questions.

(2) *Data discovery in large public heterogeneous datasets.* I will engage in data discovery using correlation measures to devise tools for evaluating and interpreting both positive and negative correlations. This area of research will be the stepping stone to creating more complex analytics, capable of answering not just domain-specific questions, but questions involving complex data from multiple domains. Such tools should be able to predict the impact that cer-

tain decisions will have on the economic health of an organization or state. The intellectual merit of my research stems from providing answers to complex questions, like "What is the predicted impact of introducing a new tax in MA?". I plan on developing mechanisms to analyze large time series data from diverse domains based on newly designed measures. These measures enable analysts to find key features likely to predict the impact of changes in various factors influencing political and economic decisions. Our tools will provide the ability to explain the data and extract interesting properties and interrelationships. I will also construct a set of models to infer the behavior of a new data set or the predicted impact of changes on the data. Classification-based comparisons across different domains will identify key features by constructing a concise summary of the stored data as well as data distribution information, such as variance. This can be the foundation for predicting the most plausible values of some missing data or value distribution of certain attributes.

In summary, I will actively seek the support of agencies such as NSF and NIH, as well other organizations interested in Big Data anlytics and data mining. I plan on engaging in collaborative research with researchers both in and outside of my area of expertise. Furthermore, I recognize the importance of recruiting and mentoring graduate and undergraduate students to make my research plans a success. In particular, I am committed to inspiring and helping my students conduct research. I plan to develop positive working relationships with students and faculty and engage in interdisciplinary research aimed at solving high impact problems.

3 References

[1] Rodica Neamtu, Ramoza Ahsan, Elke Rundensteiner, Gabor Sarkozy. Interactive Time Series Exploration Powered by the Marriage of Similarity Distances. In Proceedings of VLDB, Very Large Databases 2017, Vol. 10, No. 3, Endowment 2150-8097/16/11.
[2] Rodica Neamtu, Ramoza Ahsan, Gabor Sarkozy, Elke Rundensteiner. Generalized Dynamic Time Warping: Unleashing the Warping Power Hidden in Point-to-Point Distances. In submission for Proceedings of ACM SIGMOD 2017.
[3] Rodica Neamtu, Ramoza Ahsan, Gabor Sarkozy, Elke Rundensteiner. Efficient knowledge discovery in time series datasets powered by multiple distances. In submission for Proceedings of SIGKDD 2017.
[4] Rodica Neamtu, Ramoza Ahsan and Elke Rundensteiner The impact of Big Data on making evidence-based decisions. Book chapter. In Frontiers in Data Science. September 30, 2017 Forthcoming by CRC Press Reference - 450 Pages - 50 B/W Illustrations ISBN 9781498799324 - CAT K30579 Series: Chapman and Hall CRC Big Data Series.
[5] R Ahsan, R Neamtu, and E Rundensteiner. Towards spreadsheet integration using entity identification driven by a spatial-temporal model. In Proceedings of the 31st Annual ACM Symposium on Applied Computing, pages 1083âĂŞ1085. ACM, 2016.
[6] R Ahsan, R Neamtu, and E Rundensteiner. Using entity identification and classification for automated integration of spatial-temporal data. International Journal of Design Nature and Ecodynamics, 11(3):186–197, 2016.
[7] R Neamtu et al. "Taming Big Data: Integrating diverse public data sources for economic competitiveness analytics." Proceedings of the First International Workshop on Bringing the Value of Big Data to Users (Data4U 2014). ACM, 2014.
[8] R Ahsan, R Neamtu, et al. METIS: Massachusetts economy and technology index system. Proceedings of ACM SIGMOD 2014.