Category: Aerospace Engineering **Poster ID**: 67

Presenter: Hui Zhang, Graduate Student
Aerospace Engineering, U-M Ann Arbor

Co-Authors: Carlos E. S. Cesnik, Aerospace Engineering, U-M Ann Arbor

Title: Big Data for Damage Characterization of Aerospace Structures

Abstract: The purpose of the research is to develop efficient damage detection algorithms for effective characterization of various damage features inside aerospace structures, based on a large data library of damage scenarios generated by the nonlinear guided wave simulation tool UM/LISA. UM/LISA is a very efficient and versatile tool that is developed in A2SRL laboratory at the U-M's CoE for guided wave simulations in structures, considering features such as non-reflective boundary, piezoelectric coupled multi-physics field, strain-based damping model, nonlinear contact dynamics, and the capability for multiple GPU platform. The numeric simulation not only helps us understand guided wave interactions with structural damages, it also plays a vital role in the damage characterization algorithms, through building a large database of simulation cases with virtually all possible damage scenarios, and then correlating test signal with the database to obtain the information about the damage. Matching pursuit algorithm for extracting time and frequency centers from the signals is adopted to localize and estimate possible damage on typical aerospace structures, and big data analysis is performed on the damage library to determine most obvious signal features. The overall damage characterization framework for typical aerospace structures will be given, and the damage detection algorithm computes a special distribution of matching merit metric to illustrate regions of high likelihood of having damage while also providing the corresponding damage size estimation. The work will look into a practical case study in an effort to demonstrate the effectiveness of the algorithms for detecting complex engineering structures. Finally, the robustness of the algorithm will be briefly addressed.

Category: Astronomy Research **Poster ID**: 78

Presenter: Chris Dessert, Graduate Student
Physics, U-M Ann Arbor

Co-Authors: Ben Safdi, Physics, U-M Ann Arbor

Title: Dark Matter Decay in Andromeda and Virgo

Abstract: One of the most promising methods to search for dark matter (DM) is indirect detection of DM decay products, including photons, in the centers of galaxies. We present preliminary results on the DM lifetime limit for decays to b quarks using 413 weeks of Fermi Large Area Telescope (Fermi-LAT) Pass 8 gamma-ray data. The dataset contains photon counts with energies ranging from 200 MeV to 2 TeV, along with the spatial distribution. We can compare this distribution with the expected distribution of photon flux due to DM decays given the known spatial distribution of DM. We preform a stacked analysis of two nearby galaxies, Andromeda and Virgo, on the University of Michigan (UM) Flux cluster using python, cython, and Jupyter to obtain these limits. The procedure is as follows: we preform a Poissonian template fit with open-source code package NPTFit to model sources of flux which do not arise from DM decays and attribute the potential remaining flux as due to DM decays, yielding likelihood profiles for the decay intensity. We constrain the DM lifetime for masses between 20 GeV to 20 TeV. In the future, we will use a galaxy group catalog to preform the analysis across the full sky. We would like to additionally incorporate data up to larger redshifts, first with a cosmological simulation to understand uncertainties in the analysis and then with the Dark Energy Survey (DES) dataset pending release.

Category: Biological Sciences **Poster ID**: 62

Presenter: Umang Varma, PhD Candidate
Mathematics, U-M Ann Arbor

Co-Authors: Alexander H Vargo, Mathematics, U-M Ann Arbor; Qianyi Ma, Human Genetics, U-M Ann Arbor; Sue Hammoud, Human Genetics, U-M Ann Arbor; Jun Z Li, Human Genetics, U-M Ann Arbor; Anna C Gilbert, Mathematics, U-M Ann Arbor

Title: Marker Selection: Adapting Methods from 1-bit Compressed Sensing and Mutual Information

Abstract: Single cell RNA sequencing (scRNA-seq) gives biologists a unique insight into the heterogeneity of cells, even within the same tissue. The analysis of scRNA-seq data poses novel biological and algorithmic challenges, however. One such problem is that of marker selection: after the cells have been clustered into biological types, determine a relatively small subset of genes that is the most informative about the given clustering. In the machine learning literature, this problem is known as feature selection.

The focus of this work is on two methods for marker selection in scRNA-seq data that leverage over a decade of mathematical analysis and algorithmic development in sparsity. In one of these methods, we adopt a 1-bit compressed sensing algorithm (1CS) that was introduced in [1] for MS data. In order to select markers, the 1CS algorithm finds optimal hyperplanes that separate the given clusters of cells and depend only on a small number of genes. The second method is based on the mutual information (MI) framework developed in [2]. The MI algorithm greedily builds a set of markers out of a set of statistically significant genes that maximizes information about the target clusters and minimizes redundancy between markers.

Unlike differential expression-based methods, the methods we discuss choose a set of markers without requiring arbitrary cutoffs for the associated statistics and mitigate the problem of merging the lists created for each cluster. We hope that these tools will be widely applicable as multiclass feature selection becomes more useful in the analysis of scRNA-seq data.

References:

[1] Conrad T et al. (2017) BMC Bioinformatics, 18:160.
[2] Peng H, Long F, and Ding C (2005) IEEE Trans on Pattern Analysis and Machine Intelligence, 27(8):1226-1238.

Category: Biological Sciences **Poster ID**: 65

Presenter: Ricardo D'Oliveira Albanus, PhD Candidate
Computational Medicine & Bioinformatics, U-M Ann Arbor

Co-Authors: John Hensley, Computational Medicine & Bioinformatics, U-M Ann Arbor; Yasuhiro Kyono, Human Genetics, U-M Ann Arbor; Jacob Kitzman, Human Genetics, U-M Ann Arbor; Stephen C.J. Parker, Computational Medicine & Bioinformatics, Human Genetics, U-M Ann Arbor

Title: Bits of Information in the Human Epigenome Identify General Principles of Gene Regulation and Disease Predisposition

Abstract: Every cell in an individual human has a nearly identical genetic code, yet this code is interpreted in a cell-specific manner leading to diverse tissues and organs. This common >3 billion bases of genetic information would stretch out to approximately 2 meters, but must fit within a just few micrometers of a cell, requiring the genome to be compacted by about six orders of magnitude. This dense packaging is facilitated by a constellation of proteins that simultaneously compact the genetic material and allow for the relevant genes and regulatory circuits to be accessible in a cell-specific manner. The compact form of DNA and proteins is referred to as chromatin and represents the epigenome. In order to better understand in vivo chromatin landscapes, we develop a new method that uses chromatin accessibility data to predict where transcription factors (TFs), the protein effectors of chromatin regulation, bind genome-wide, and show using experimental TF binding data as cross-validation, that it performs better than competing methods. We additionally develop a new information theory approach to calculate the information encoded in chromatin. We show that the chromatin information landscape is non-uniform across the genome and varies by cell type. High chromatin information domains (CIDs) are significantly enriched in cell-identity regions of the genome, while low CIDs are significantly enriched in general regions. Furthermore, we observe that genetic variations associated with gene regulation are more enriched in low CIDs compared to high CIDs, suggesting that high CIDs could exert chromatin buffering by maintaining a more stable local chromatin organization. Additionally, high CIDs are significantly overrepresented in evolutionarily conserved and disease associated regions, indicating that they have lower tolerance to genetic variation. Collectively, our work provides insights on how epigenomic information is encoded in the human genome and how it influences health and disease.

Category: Biological Sciences **Poster ID**: 69

Presenter: Qianyi Ma, Research Area Specialist
Human Genetics, U-M Ann Arbor

Co-Authors: Christopher D. Green, Human Genetics, U-M Ann Arbor; Jun Z. Li, Human Genetics, Computational Medicine & Bioinformatics, U-M Ann Arbor; S. Sue Hammoud, Human Genetics, Obstetrics & Gynecology, U-M Ann Arbor

Title: A High-Resolution Atlas of Spermatogenesis Using Single-cell RNA Sequencing

Abstract: Spermatogenesis is a well-organized and tightly regulated biological process composed of three distinct biological activities: 1) Continuous stem cell self-renewal and the expansion of progenitor cells by mitosis, 2) the production of haploid cells from diploid progenitor cells by meiosis, and 3) the orderly differentiation of haploid cells into spermatozoa (spermiogenesis). The transition between these developmental stages is precisely timed, and is dependent upon both germ cell intrinsic programs and extrinsic factors secreted by supporting cells. In past studies, the spatial organization of these cells across the seminiferous tubule (ST) has been determined histologically, and their functional contribution to spermatogenesis has been examined through the purification of predefined cell subpopulations, followed by gene expression profiling using microarrays or bulk RNA-seq. However, our understanding of the functional heterogeneity of cell populations is limited, and the developmental transitions are poorly understood. To overcome these limitations, we have obtained gene expression profiles of 33,180 high-quality single cells of mouse testis in 24 experimental datasets. Unbiased clustering defined reproducible and high-confidence view of cellular heterogeneity of ST, revealing 10 major cell types, corresponding to interstitial cells (Leydig, Endothelial, Myoid, Macrophages and a newly recognized cell type), Sertoli cells, and known germ cell populations. Unsupervised ordering of the germ cell populations reflected the differentiation trajectory from spermatogonia to spermatocytes, and to round and mature sperm. Taken together, these findings provide a molecular portrait of major cell types in the ST at an unprecedented resolution. The cell type markers developed from this resource will be valuable for studying cell-cell interactions and their spatiotemporal patterns during different stages of spermatogenesis. Also importantly – these datasets are being used to improve the computational methods currently used in the field.

Category: Business and Marketing **Poster ID**: 36

Presenter: Longxiu Tian, Graduate Student
Marketing, U-M Ann Arbor

Co-Authors: Fred M. Feinberg, Marketing, Statistics, U-M Ann Arbor

Title: Bayesian Imputation for Freemium Subscription Service Choice: Analysis of an Online Dating Pricing Experiment

Abstract: Freemium subscription services allow customers to “try before they buy”, relying on a paying subscriber base for ongoing revenues. To entice free users to upgrade, firms typically offer an array of options as a price menu, rewarding longer commitments with lower unit prices. Key to their success is setting component prices to both encourage upgrading overall and to redistribute upgraders to longer-term contracts. Extant customer data suffer from low menu pricing update frequency and component plan prices changing in lockstep. We avail of a formal pricing experiment for an online dating site that orthogonalizes the midpoint and steepness of price menus. To correct for differential customer self-selection into upgrading, which is a function of the menu prices themselves, we develop a novel selectivity model relating binary selection (upgrading) and multinomial choice (among three available plans). The model is estimated using Riemann Manifold Hamiltonian Monte Carlo, for efficient recovery of highly nonlinear parameters and full covariance matrices. Model estimates confirm a number of expected pricing effects (e.g., higher prices discourage upgrading and dissuade choice of particular plans), but also an unanticipated decoy effect: raising the price of the 3-month contract actually helps the 6-month contract most desired by firm management.

Poster Abstracts

Category: Climate Research & Natural Disasters **Poster ID**: 1

Presenter: Kevin Fries, Graduate Student
Civil & Environmental Engineering, U-M Ann Arbor

Co-Authors: Branko Kerkez, Civil & Environmental Engineering, U-M Ann Arbor

Title: Real Time Flood Forecasting Through Data Fusion

Abstract: As computational power grows, hydrologists' ability to observe, model, and predict events is increasing at an ever-growing rate. The recent operationalization of the National Water Model is just one example of our capacity. But as our capacity grows, so do the size of the watersheds we model. As a result, the negligible uncertainties in smaller-scale models are no longer negligible. There is now a need for new tools and frameworks that can assess model performance and improve our models through data integration or assimilation. This study focuses on that state of Iowa and how effective the NWM could be in predicting floods at high resolution. In this paper we present a study across 182 gages in Iowa where we learned state-space representations of the stream stage based on NWM flow predictions. Of these 182 sites, approximately one-third have shown to be promising for flood forecasting. That is, the NWM does not predict storms when none occur (false positive) or miss storms when they do occur (false negative). What makes these sites viable are further characterized using principal component analysis and random forests. This type of approach can lend itself to being informative for identifying where large scale models may benefit from further observations, such as new stream gages. Further, this type of approach can help municipalities identify where they may want to place sensors and couple the observations with a large scale model so as to implement early-warning flood systems for their communities.

Category: Climate Research & Natural Disasters **Poster ID**: 12

Presenter: Sara Shashaani, Postdoctoral Fellow
Industrial & Operations Engineering, U-M Ann Arbor

Co-Authors: Seth Guikema, Industrial Operations & Engineering, U-M Ann Arbor

Title: Hurricane Power Outage Prediction with Out of Bag Feature Selection Approaches

Abstract: Predicting hurricane power outages facilitates natural hazards response decisions. However the spatial data is largely zero-inflated with a sizable number of explanatory variables, and finding statistical models that can provide reliable predictions remains to be a challenge. We study a feature selection approach by using out of bag performance of the predictors for such datasets. A combinatorial optimization model is developed to seek contributing variables to the accuracy of the prediction model. We found promising results experimenting on actual central gulf coast outage data.

Category: Climate Research & Natural Disasters **Poster ID**: 22

Presenter: Manish Verma, Data Science Consultant
Consulting for Statistics, Computing & Analytics Research, U-M Ann Arbor

Co-Authors:

Title: Coastal Cities and Hurricanes in a Changing Climate: A Geospatial Perspective

Abstract: Hurricanes cause loss of life, damage property and infrastructure, and can have lasting negative impact on local economy and society. Hurricane Harvey that made landfall in Texas in 2017 dumped more than 40 inches of rainfall in four to five days causing widespread flooding. It displaced more than 30000 people and caused billions of dollars of damage to property. Nearly 50,000 houses were affected by the flood and more than 10000 people had to be rescued. Although we cannot attribute unprecedented rains of individual hurricanes to climate change, overall climate models predict that hurricanes are going to become more powerful as sea surface temperature increases. Coastal cities thus have to integrate realities of climate change in their growth and expansion plans.

In this presentation, I will combine a variety of remotely sensed data, GIS, and spatial analysis to examine the patterns of urban growth and how it intersected with the hydrology of the greater Houston area. The presentation will examine if urban growth was partly responsible for the damage caused by the unprecedented rains from Harvey. The goal will be to identify spatial elements that coastal cities can integrate as they develop adaptation strategies.

Category: Climate Research & Natural Disasters **Poster ID**: 64

Presenter: Xianglei Huang, Associate Professor
Climate & Space Sciences & Engineering, U-M Ann Arbor

Co-Authors: Xiuhong Chen, Climate & Space Engineering, U-M Ann Arbor; Haoming Shen, Electrical Engineering & Computer Science, U-M Ann Arbor

Title: Solar Forecasting Based on the Suppor Vector Regression

Abstract: We here present some initial results of our project of using machine-learning algorithm to perform intra-day solar forecast. The performance of the algorithm is compared with a more physical-based forecast and the advantages and disadvantages of both methods will be delineated. Thoughts for further improving the SVR approach will be also discussed.

Category: Climate Research & Natural Disasters **Poster ID**: 71

Presenter: Matthew Bartos, Graduate Student
Civil & Environmental Engineering, U-M Ann Arbor

Co-Authors: Alexander Ritchie, Electrical & Computer Engineering, U-Michigan Ann Arbor

Title: A Graph Partitioning Approach for Controller Placement in Dendritic Networks

Abstract: In this study, we develop an algorithm for controller placement in directed graphs, with specific applications for active control of stormwater infrastructure. In the context of watersheds, flash floods may occur when large volumes of runoff arrive synchronously at a given location in the drainage network. If valve controllers are placed at strategic locations, flood waves can be mitigated by "de-synchronizing" the tributary flows through coordinated delays and releases of water. Controller placement can thus be formulated as a graph-cutting problem that seeks to remove subgraphs that contribute most to synchronous flows at a target location. We address this problem through two major contributions. First, we develop a new graph shift operator that describes the translation of a graph signal in continuous time by characterizing the rate of change of the signal at each vertex. The graph shift operator is used to determine the steady-state travel time of a nonlinear kinematic flood wave through a dendritic graph representing a stream network. Our second major contribution is the development of a controller placement algorithm that minimizes synchronous flows at a target location in the network. Our partitioning algorithm removes subgraphs that share a common distribution of edge distances with respect to a target location (using the steady-state travel time of a kinematic wave as the topological distance). The results of this controller placement approach are compared to configurations produced by traditional graph-cutting approaches, and the performance of each method is evaluated using a hydraulic model (EPA Stormwater Management Model). While formulated specifically for the problem of flood control, the results of this study can be used to determine controller placement strategies for other problems involving complex networks, such as contaminant transport in pipe networks, or control of traffic flows.

Category: Data Science Education **Poster ID**: 33

Presenter: Arya Farahi, Graduate Student
Physics, U-M Ann Arbor

Co-Authors: Jonathan Stroud, Computer Science Engineering, U-M Ann Arbor

Title: MDST - A Model for Data Science Education through Public Service

Abstract: The Michigan Data Science Team (MDST) is a student organization at the University of Michigan where students from all disciplines collaborate on impactful data science projects. Many MDST projects focus on public service, which allows students to exercise their skills to aid local communities. MDST's success in establishing community partnerships serves as a model for using public service projects to advance data science education.

Category: Data Science Methodology **Poster ID**: 2

Presenter: Bryan Goldsmith, Assistant Professor
Chemical Engineering, U-M Ann Arbor

Co-Authors: Dr. Mario Boley, Dr. Jilles Vreeken, Professor Matthias Scheffler, Dr. Luca M. Ghiringhelli

Title: Finding Patterns, Correlations, and Descriptors in Materials-science Data using Subgroup Discovery

Abstract: Data analytics applied to materials-science data often focuses on the inference of a global prediction model for some physical or chemical property of interest for a given class of materials, such as band gaps or molecule binding energies. However, the underlying mechanism for some target property could differ for different materials within a large pool of materials-science data. Consequently, a global model fitted to the entire dataset may be difficult to interpret and may well hide or incorrectly describe the actuating physical mechanisms. In these situations, local models would be advantageous to global models. Subgroup discovery (SGD) is presented here as a data-mining approach to find interpretable local models of a target property in materials-science data. We first demonstrate that SGD can identify physically meaningful models that classify the crystal structures of 82 octet binary semiconductors as either rocksalt or zincblende. The SGD framework is subsequently applied to 24 400 configurations of neutral gas-phase gold clusters with 5 to 14 atoms to discern general patterns between geometrical and physicochemical properties.

Category: Data Science Methodology **Poster ID**: 9

Presenter: Bing She, Postdoctoral Research Associate
Spatial Data Center & China Data Center; Institute for Social Research, U-M Ann Arbor

Co-Authors: Shuming Bao, Spatial Data Center & China Data Center, U-M Ann Arbor

Title: A Spatial Approach for Knowledge-based Discovery of Research Data

Abstract: Finding the right dataset for analysis is crucial. Researchers have to draw on extensive literature reviews and their own prior research experiences to identify candidate datasets. Despite many existing tools to promote effective data use, data search remains largely a manual process which requires researchers to choose certain search keywords or browse datasets by certain subject categories pre-defined by archivists. A fundamental challenge is how to better harness the rich data from multiple spatiotemporal constrained datasets that cut across existing disciplinary boundaries to inform the development of new theory and hypothesis testing. Literature-based discovery aims to assist researchers in generating meaningful hypotheses by mining the implicit relationships between terms in the literature. However, the previous studies assume homogenous influence mechanisms across regions. In addition, most connectivity matrices, which are among one of the core components for building the influence chains, were built on the static data structure. This work envisions a spatial approach for knowledge-based discovery related to these limitations. This presentation will discuss an integrated approach for identifying the optimized influence chains between terms based on different criteria and discovering the linkages among survey data across disciplines with ad-hoc spatial/temporal constraints. This will enable the generation of a broad range of novel research questions and scientific hypotheses in a much more efficient and flexible way than previous approaches.

Category: Data Science Methodology **Poster ID**: 14

Presenter: Mikhail Yurochkin, Graduate Student
Statistics, U-M Ann Arbor

Co-Authors: Mikhail Yurochkin, Statistics, U-M Ann Arbor; Aritra Guha, Statistics,U-M Ann Arbor; XuanLong Nguyen,Statistics, U-M Ann Arbor

Title: Conic Scan-and-Cover Algorithms for Nonparametric Topic Modeling

Abstract: We propose new algorithms for topic modeling when the number of topics is unknown. Our approach relies on an analysis of the concentration of mass and angular geometry of the topic simplex, a convex polytope constructed by taking the convex hull of vertices representing the latent topics. Our algorithms are shown in practice to have accuracy comparable to a Gibbs sampler in terms of topic estimation, which requires the number of topics be given. Moreover, they are one of the fastest among several state of the art parametric techniques. Statistical consistency of our estimator is established under some conditions.

Category: Data Science Methodology **Poster ID**: 27

Presenter: David Hong, Graduate Student
Electrical Engineering & Computer Science, U-M Ann Arbor

Co-Authors: Jeffrey Fessler, Electrical Engineering & Computer Science, U-M Ann Arbor; Laura Balzano, Electrical Engineering & Computer Science, U-M Ann Arbor

Title: Optimally Weighted PCA for High-Dimensional Heteroscedastic Data

Abstract: Principal Component Analysis (PCA) is a classical method for reducing the dimensionality of data by projecting them onto a low-dimensional subspace that captures most of their variation. It is a fundamental tool with numerous data analysis applications. However, conventional PCA treats all data uniformly and does not exploit any knowledge of the relative quality of each sample. If the noise variance of each sample is known, a more natural approach is to give noisier samples less weight, i.e., to use a weighted PCA. Common choices include binary weights (i.e., throw away noisier samples) and inverse noise variance weights (i.e., maximum likelihood or whitening). This work analyzes the statistical performance of weighted PCA for high-dimensional data drawn from a low-dimensional subspace and degraded by heteroscedastic noise (i.e., noise with non-uniform variance across samples). We show the surprising fact that the common weighting choices produce sub-optimal asymptotic estimation of the underlying low-dimensional subspace. In particular, we provide a simple expression for asymptotic recovery as a function of data properties (e.g., noise variances, samples-to-dimension ratio, etc.) and weights; this analysis enables us to find the optimal weights as a function of the data properties. The expression also quantifies the performance lost by using sub-optimal weights: sometimes it is practically optimal to throw away noisier samples and other times doing so significantly degrades performance.

Category: Data Science Methodology **Poster ID**: 28

Presenter: Yutong Wang, Graduate student
Electrical & Computer Engineering, U-M Ann Arbor

Co-Authors: Laura Balzano, Electrical & Computer Engineering, U-M Ann Arbor; Clayton Scott, Electrical & Computer Engineering, U-M Ann Arbor; Venkatesh Saligrama, Electrical & Computer Engineering, Boston University

Title: A Convex Clustering Formulation Using the Similarity Matrix

Abstract: We consider the problem of clustering given data drawn from a finite K-mixture in high dimension. Our approach uses the similarity matrix of the data and an optimization based on the Ky-Fan K-norm of the similarity matrix. We show that under certain separation assumptions on the mixture centroids, the optimization when given the true similarity matrix returns the true clustering. In practice, we do not have access to the true similarity matrix. However, we demonstrate in synthetic experiments that our method far outperforms Kmeans even when we only have a single empirical estimate of the similarity matrix.

Category: Data Science Methodology **Poster ID**: 35

Presenter: Yumu Liu, Graduate student
Statistics, U-M Ann Arbor

Co-Authors: Ji Zhu, Statistics, U-M Ann Arbor

Title: Network Community Detection via the Degree-Corrected Block Model with Node Covariates

Abstract: One fundamental problem in network analysis is community detection, which aims to cluster nodes into groups that share similar link patterns. Node covariates are often available together with the network in many nowadays applications, and it is desirable to leverage the node covariates information in community detection. In this project, we consider incorporating node covariates via the exible degree-corrected block model. Specically, we allow the community memberships to depend on node covariates, while the link probabilities are determined by both node community memberships and degree parameters. In terms of model estimation, we have developed two algorithms, one using the variational inference and the other based on the pseudo-likelihood. We found in simulation studies that the pseudo-likelihood approach, though scales well when the number of nodes increases, does not cluster as accurately as the variational inference approach when either the network becomes sparse or the within- vs between-community dierence becomes small. Further, we show under mild conditions, the community memberships and the covariate parameters can be estimated consistently. We have also applied the proposed model to a lawyer friendship network.

Category: Data Science Methodology **Poster ID**: 37

Presenter: MeiXing Dong, Graduate Student
Computer Science & Engineering, U-M Ann Arbor

Co-Authors: Rada Mihalcea, Computer Science & Engineering, U-M Ann Arbor

Title: Expanding Sparse Text with Induced Domain-Specific Lexicons and Embeddings

Abstract: We address the problem of expanding sparse textual content to increase the accuracy of data-driven prediction tasks. Using an alumni donor dataset and two social media datasets, we perform several comparative experiments and analyses, and show that our methods to automatically enhance sparse textual data significantly improve predictive performance. Methods used include lexicon induction and word embeddings.

Category: Data Science Methodology **Poster ID**: 38

Presenter: Micha Fischer, Graduate Student
Survey Methodology, Institute for Social Research, U-M Ann Arbor

Co-Authors: Felicitas Mittereder, Institute for Social Research, Program in Survey Methodology, U-M Ann Arbor

Title: Using Missing Data to Impute Missing Data

Abstract: In survey data, to impute missing values in a variable of interest we usually assume complete covariates. Often this is not the case and we first impute the missing values within the covariates (i.e. pre-imputation step). This method can be problematic in two ways: first, if the imputation models in the pre-imputation-step are not correctly specified, we might introduce imputation bias to survey estimates. Second, we lose the information that the respondent chose not to answer the question. Assuming MAR and MNAR, this information by itself might be important for the imputation model. Thus, the item missing pattern of respondents can be informative for the outcome variables. By including missingness as its own category we could improve imputation accuracy and therefore the estimators for survey data. Tree-based methods (e.g. random forest) can incorporate this additional information and account for complex interactions in the covariates at the same time. Using random forests in a simulation study, we investigate in which situations can this approach be more precise and accurate compared to the usual approaches.

Category: Data Science Methodology **Poster ID**: 39

Presenter: Boang Liu, PhD Candidate
Statistics, U-M Ann Arbor

Co-Authors: Ji Zhu, Statistics, U-M Ann Arbor

Title: Network Augmented Classification

Abstract: In classical classification, a data point is classified given a vector of its individual covariates. Often, additional network information describing the connectivity relationships between data points are also available, which in principle can be used to improve the classification performance. In this work, we develop a general statistical framework for network augmented classification. Under this framework, we derive the optimal Bayes classifiers for two general families of distributions incorporating both covariates and networks, one being generative and the other being discriminative. Further, we establish consistency results for plug-in classifiers with respect to the optimal classifier under the generative family. We also apply the general approaches to two specific models and propose two effective classification methods for practical use. The proposed methods have been evaluated using both simulation studies and real-world data examples, and the results are promising.

Category: Data Science Methodology **Poster ID**: 40

Presenter: Xuefei Zhang, Graduate Student
Statistics, U-M Ann Arbor

Co-Authors: Ji Zhu, Statistics, U-M Ann Arbor

Title: Prediction on Network-linked Data by Matrix Variate Models

Abstract: Classical prediction problems usually assume training data are independent samples. However in real-world data, individuals are often connected by a network and interact in complex ways, therefore the independence assumption may not hold. Incorporating network information in modeling is expected to improve the prediction performance, as it provides additional information about relationships among individuals. In this proposal, we first focus on predicting a continuous response variable using both covariates and network information. Specifically, we propose to use a matrix variate model, that allows two-way dependence among data points and among variables, to model the distribution of variables associated with nodes in a network. Network information is naturally incorporated into the matrix variate distribution, and the relationship between the response variable and predictors can be derived under such model. We have developed an efficient EM algorithm for parameter estimation. We demonstrate the performance of our method using simulation studies, for both in-sample prediction and semi-supervised learning settings, and the preliminary results are encouraging. We have also extended the proposed framework to classification problems when the variable of interest is categorical.

Category: Data Science Methodology **Poster ID**: 47

Presenter: Murali Mani, Associate Professor
Computer Science, U-M Flint

Co-Authors:

Title: Effective Big Data Visualization

Abstract: In the last several years, big data analytics has found an increasing role in our everyday lives. Data visualization has long been accepted as an integral part of data analytics. However, data visualization systems are not equipped to handle the complexities typically found in big data. Our work examines effective ways of visualizing big data, while also realizing that most visualization processes are interactive. During an interactive visualization session, an analyst issues several visualization requests, each of which builds on prior visualizations. In our approach, we integrate a distributed data processing system that can effectively process big data with a visualization system that can provide effective interactive visualization but for smaller amounts of data. The analyst's current request is used to infer contextual information about the analyst such as their expertise and tolerance for delay. This information is used to carefully determine additional data that can be sent to the visualization system for decreasing the response time for future requests, thus providing a better experience for the analyst and increasing their productivity.

Category: Data Science Methodology **Poster ID**: 48

Presenter: Neophytos Charalambides, Graduate Student
Electrical Engineering & Computer Science, U-M Ann Arbor

Co-Authors: Anthony Thomas Della Pella, Mathematics, U-M Ann Arbor

Title: Community Detection Through Polynomials

Abstract: Community detection is of great interest to social scientists, psychologists, economists, as well as engineers. We study this problem through random graphs by using partition functions which originate from physics. It is known that exact computation of these polynomials (for a specific class of partition function) answer the question as to whether or not “cliques“ exist in graphs. We seek a generalization to random graphs of an algorithm given for approximation of the clique partition function. The algorithm will be implemented and subsequently simulated on a dataset of multiple taxi journeys in Chicago, where trips are considered based on pickup and drop-off locations. The data set is structured to inherently include communities, so the obtained empirical results will serve as a control group for further numerical experimentation towards verifying the practical computability of the algorithm in the more general setting.

Category: Data Science Methodology **Poster ID**: 66

Presenter: Zhiyuan Lu, Graduate Student
Statistics, U-M Ann Arbor

Co-Authors: Moulinath Banerjee, Statistics, U-M Ann Arbor; George Michailidis, Informatics Institute, University of Florida

Title: Intelligent Sampling for Multiple Change-points in Exceedingly Long Time Series with Rate Guarantees

Abstract: Change point estimation (offline version) is traditionally performed by optimizing over the data set of interest, by considering each data point as the true location parameter and computing a data fit criterion. Subsequently, the data point that minimizes the criterion is declared as the change point estimate. For estimating multiple change points, the procedures are analogous in spirit, but significantly more involved in execution. Since change-points are local discontinuities, only data points close to the actual change point provide useful information for estimation, while data points far away are superfluous, to the point where using only a few points close to the true parameter is just as precise as using the full data set. Leveraging this ``locality principle", we introduce a two-stage procedure for the problem at hand, which in the 1st stage uses a sparse subsample to obtain pilot estimates, and in the 2nd stage refines these estimates by sampling densely in appropriately defined neighborhoods around them. We establish that this method achieves the same rate of convergence and even virtually the same asymptotic distribution as the analysis of the full data, while reducing computational complexity to $\sqrt{N}$ time (N being the length of data set), as opposed to at least $O(N)$ time for all current procedures, making it promising for the analysis on exceedingly long data sets with adequately spaced out change points. The main results are established under a signal plus noise model with independent and identically distributed error terms, but extensions to dependent data settings, as well as multiple stage (>2) procedures (which deliver even bigger computational gains without losing precision in truly humongous time series) are also provided. The performance of our procedure -- which is coined ``intelligent sampling" -- is illustrated on both synthetic and real Internet data streams.

Category: Data Science Methodology **Poster ID**: 72

Presenter: Elizabeth Hou, Graduate Student
Electrical Engineering & Computer Science, U-M Ann Arbor

Co-Authors: Yasin Yılmaz; Alfred O. Hero,Electrical Engineering & Computer Science, Biomedical Engineering, Statistics, Data Science Initiative, U-M Ann Arbor

Title: Sparse Network Tomography for Anomaly Detection

Abstract: We are interested in detecting anomalous activity in sparse traffic networks where the network is not directly observed. Given knowledge of what the node-to-node traffic in a network should be, any activity that differs significantly from this baseline would be considered anomalous. We propose two statistical estimators for the actual rates of the network traffic matrix. The first is a Bayesian hierarchical model, which uses the EM algorithm to simultaneously approximate the unobserved individual traffic between the nodes and estimate the rate matrix, and the second is a minimum relative entropy distribution formed by projecting a prior distribution centered around the baseline onto a feasible set bounded by the data. We show that by warm-starting the expectation-maximization (EM) algorithm at the mode of the minimum relative entropy distribution, we achieve better performance and require much fewer EM iterations. Additionally, the probabilistic framework of the Bayesian hierarchical model allows us to naturally perform statistical goodness-of-fit tests to detect significant deviations from the baseline network. We apply our models to both simulated and real datasets to demonstrate their superior performance over existing alternatives.

Category: Data Science Methodology **Poster ID**: 73

Presenter: Morteza Noshad, Graduate Student
Electrical Engineering & Computer Science, U-M Ann Arbor

Co-Authors:

Title: Optimal Estimation of Information Measures and their Applications

Abstract: We propose a direct estimation method for information measures such as Renyi and f-divergence and mutual informations, based on a new graph theoretical interpretation. Suppose that we are given two sample sets X and Y, respectively with N and M samples, where $\eta := M/N$ is a constant value. Considering the k-nearest neighbor (k-NN) graph of Y in the joint data set (X,Y), we show that the average powered ratio of the number of X points to the number of Y points among all k-NN points is proportional to Renyi divergence of X and Y densities. A similar method can also be used to estimate measures of divergence and correlation. We derive bias and variance rates, and show that for density functions with continuous and bounded derivatives of up to the order d, our estimator achieves the parametric MSE rate of O(1/N). Our estimators are more computationally tractable than other competing estimators, which makes them appealing in many practical applications such as change detection, learning graphical models, finding the error of classification, feature selection, etc.

Category: Data Science Methodology **Poster ID**: 80

Presenter: Salimeh Yasaie Sekeh, Postdoctoral Research Fellow
Electrical Engineering & Computer Science, U-M Ann Arbor

Co-Authors: Brandon Oselio, Electrical Engineering & Computer Science, Statistics, U-M Ann Arbor; Alfred O. Hero, Electrical Engineering & Computer Science, Biomedical Engineering, Statistics, Data Science Initiative, U-M Ann Arbor

Title: A Dimension-Independent Discriminant Between Distributions

Abstract: Henze-Penrose divergence is a non-parametric divergence measure that can be used to estimate a bound on the Bayes error in a binary classification problem. In this paper, we show that a cross-match statistic based on optimal weighted matching can be used to directly estimate Henze-Penrose divergence. Unlike an earlier approach based on the Friedman-Rafsky minimal spanning tree statistic, the proposed method is dimension-independent. The new approach is evaluated using simulation and applied to real datasets to obtain Bayes error estimates.

Category: Data Security **Poster ID**: 5

Presenter: Eric Boyd, Director, Research Networks
Information & Technology Services, U-M Ann Arbor

Co-Authors: Ferdinando Fioretto, Industrial & Operations Engineering, U-M Ann Arbor

Title: PrivaScope: Protect Privacy while Enabling Data Analytics

Abstract: The University of Michigan is committed to sharing operational data with researchers and administrative units to further the research mission of the university, as long as doing so does not compromise individual privacy, is supportable with reasonable effort, and does not compromise operational efficiency or security.

In order to meet the imperative of sustainably and securely supporting research while preserving privacy, the university is developing PrivaScope (a contraction of “Privacy” and “Microscope”), a software framework to enable research use of operational data while maintaining individual privacy. It is important to note that this data exists for operational reasons and is not being created solely for research use. PrivaScope is being submitted to the university under the aegis of a data repository IRB. Privacy preservation is central to this effort, not ancillary.

PrivaScope will help realize many practical outcomes:

- Protection of the individual privacy in every area of research by implementing state-of-the-art technology: advanced differential privacy for data analysis

- Effective and efficient transportation for all community members

- Easy and efficient access to healthy food choices

- Individual effectiveness either through quick access to study space by providing study area heat maps, or through finding correlation of the extracurricular activities and academic achievement

Further goal of this initiative is to disseminate acquired knowledge and deployed digital tools by open-sourcing data transformation software.

Category: Data Security **Poster ID**: 58

Presenter: Xinyan Zhao, Graduate Student
School of Information, U-M Ann Arbor

Co-Authors: V.G. Vinod Vydiswaran, Learning Health Sciences, U-M Ann Arbor

Title: HyDeXT: A Hybrid De-identification and Extraction Tool for Health Text

Abstract: HyDeXT, a Hybrid De-identification and eXTraction tool. It combines the labels derived from a machine learning-based sequential tagger and a pattern-based labeler, to effectively de-identify eighteen categories of Protect Health Information (PHI).

Methods: HyDeXT consists of a text processing pipeline with the following four stages: (i) Cleaning and pre-processing. (ii) Pattern-based and Gazetteer-based labeling. (iii) Machine learning-based sequence tagger. (iv) The final combiner.

The machine learning-based sequence tagger is a linear-chain Conditional Random Field(CRF) model defined over five classes of features; namely (a) Word form features (b) Part-of-speech features (c) Neighborhood features (d) Regex features, and (e) IsInX features: features that check if a word matches elements in a list or gazette for certain closed form PHI fields.

Result: The HyDeXT tool was evaluated on the 2016 i2b2 Research Domain Criteria dataset. We compared our system with the current state-of-the-art medical de-identification system, MIST. The micro-averaged accuracy of MIST on the cleaned text version (after stage (i)) is 0.851. HyDeXT achieves an overall accuracy of 0.898, a relative improvement of 5.5%.

Impact: HyDeXT is able to identify eighteen PHI fields using a machine learning-based sequential tagger combined with a pattern-based labeler, which results in a highly efficient and effective PHI de-identification system; (2) the individual stages of the pipeline can be utilized separately, in addition to allowing for a plug-and-play architecture to support further improvements; and (3) HyDeXT provides a reliable name recognition and extraction on unseen health data, including social media data. The tool has been used to successfully identify person names and other PHI-like fields from Facebook posts. In future, we plan to further develop HyDeXT as a health information extraction tool.

Category: Energy Research **Poster ID**: 60

Presenter: Jingxing Wang, PhD Candidate
Industrial & Operations Engineering, U-M Ann Arbor

Co-Authors: Abdullah Alshelahi, Industrial & Operations Engineering, U-M Ann Arbor; Mingdi You, Industrial & Operations Engineering, U-M Ann Arbor; Eunshin Byon, Industrial & Operations Engineering, Civil & Environmental Engineering, U-M Ann Arbor; Romesh Saigal, Ind

Title: Integrated Prediction of Wind-Farm Power Output

Abstract: The market share of wind energy in the electricity market has been increasing significantly during the recent decade. However, wind farm power output is volatile because it is highly affected by stochastic weather conditions. This feature brings lots of challenges to manage the power production process. This study aims to develop a new method for predicting non-stationary wind farm outputs by capturing highly volatile wind speed characteristics, capturing the time-dependent wind-to-power relationship, and quantifying the variability of both wind speeds and power outputs in an integrative way.

We present a new integrated prediction methodology for wind power production under the assumption that the underlying dynamics follow the inhomogeneous geometric Brownian motion. Dual Kalman filtering is used to estimate the time-dependent parameters in the wind speed model. An adaptive learning method to capture the time-varying wind-to-power relationship, that is, a dynamic power curve. Therefore, the wind farm power out can also be modeled as an inhomogeneous geometric Brownian motion, whose parameters comes from the wind speed model and the power curve. As a result, the wind farm power output follows a closed-form log-normal distribution. Prediction of power output is made to minimize the weighted sum of underestimation and overestimation, which are quantified by real options theory.

Our method provides promising results. After tested on two data sets from inland commercial wind farms, our method achieves over 20% lower prediction errors than alternative methods.

Category: Energy Research **Poster ID**: 79

Presenter: Kristen Schell, Postdoctoral Research Fellow
Industrial & Operations Engineering, U-M Ann Arbor

Co-Authors: Seth Guikema, Industrial & Operations Engineering, U-M Ann Arbor; Pierre Pinson, University of Denmark; Brent McRoberts, Texas A&M University

Title: Wind Drought Synchronicity with Peak Demand

Abstract: Although accurate resource assessment is still a major challenge for wind farm investors, the technology has seen wide global deployment and remains an attractive proposition. Decreasing cost, compounded with policy incentives to achieve high renewable energy penetration, ensure wind will play a large part in electric grids of the future. While experience with incorporating large amounts of variable renewable energy generation into the electricity system has grown in recent years - especially in locations like California and Texas - achieving political mandates for deep decarbonization will require better informed power plant siting, to ensure adequacy of supply. Using 1.3 terabytes (TB) of wind speed and electricity system load data, we identify locations with a wind resource that is better matched to electricity demand through a cross-spectral analysis of the wind speed and electricity load time series. Applied to the wind resource across the electricity system of Texas (ERCOT), we show that 92% of wind farms located in ERCOT, as of 2012, are not optimally sited to provide the best power output at peak demand. Further, only 22% of the wind resource across ERCOT aligns with peak demand in the system. Future wind farms must be sited according to their ability to contribute to system resource adequacy, or rate-payers risk paying hundreds of millions in capacity payments for curtailed wind power, and investors risk stranded assets.

Category: Engineering Research **Poster ID**: 8

Presenter: Zhe Du, Graduate Student
Electrical & Computer Engineering, U-M Ann Arbor

Co-Authors: Laura Balzano, Electrical & Computer Engineering, U-M Ann Arbor; Necmiye Ozay, Electrical & Computer Engineering, U-M Ann Arbor

Title: Robust Online Switched System Identification

Abstract: Our work focuses on the problem of online switched system identification. The system we are interested in is Switched AutoRegressive eXogenous (SARX) system, which is a hybrid system consists of several subsystems with different parameters. At one time, only one subsystem is dominating and generating the output, and the dominant system could switch over time arbitrarily. Our goal is to estimate the parameters for subsystems as well as the subsystem switching sequences in an online fashion.

Existing methods solving this online problem often suffer from a disadvantage that well learned estimate for a subsystem may easily be affected by data from other subsystems. Focusing on this, we proposed a two-step algorithm with several candidate estimates to learn each subsystem parameters. Every time we receive new data, we first assign this data to one candidate based on a robust criterion that incorporates both residual error and a confidence level. Then, we use Randomized Normalized Least Mean Squares (RNLMS) algorithm to update the parameter estimate of chosen candidate. As for the theoretical guarantees, we showed the local convergence of our algorithm. Specifically, when the candidate estimates are well initialized, we can guarantee every data assignment will be made correctly and estimate can converge. Though theories only guarantee convergence with good initialization, simulation results show even with random initialization, our algorithm still has good performance.

Finally, we showed that our algorithm outperforms existing method in both synthetic data as well as real world data.

Category: Engineering Research **Poster ID**: 10

Presenter: Arvind Prasadan, Graduate Student
Electrical Engineering & Computer Science, U-M Ann Arbor

Co-Authors: Raj Nadakuditi, Electrical Engineering & Computer Science, U-M Ann Arbor

Title: Hilbert Dynamic Mode Decomposition

Abstract: Dynamic Mode Decomposition (DMD) is a technique for analyzing the behavior of a dynamical system. DMD decomposes the output from such a system into a collection of static modes and corresponding frequencies governing the time-evolution of the system. DMD is widely applied in fields as diverse as fluid dynamics, biology, and economics. However, when applied to real-valued data, DMD is problematic. In particular, we show that recovery of the frequencies can be impossible. Via connections to random matrix theory, we explore the behavior on real- and complex-valued data, and derive a new, adjusted algorithm that resolves these issues: Hilbert DMD (hDMD). We validate our theoretical results via numerical simulations, and demonstrate the superiority of hDMD over DMD.

Category: Healthcare Research **Poster ID**: 13

Presenter: Omar Iftikhar, Graduate Student
Computer & Information Science, U-M Dearborn

Co-Authors: Omid Dehzangi, Computer & Information Science, U-M Dearborn; Omar Iftikhar, Computer & Information Science,U-M Dearborn; Bhavani Anantapur Bache, Computer & Information Science,U-M Dearborn; Ying Li, Orthopaedic Surgery, Michigan Medicine, U-M Ann Arbor;

Title: Context-Aware Sensor Solution for Remote Monitoring of Adolescent Idiopathic Scoliosis Brace Treatment

Abstract: Examining the effectiveness and compliance of scoliosis patients wearing back braces is an on-going challenge that many physicians currently face. Scoliosis is a medical condition that is caused by an irregular curve in an individual's spine. Adolescent Idiopathic Scoliosis(AIS) occurs in individuals 11 years and older. Brace treatment with a thoracolumbosacral orthosis (TLSO) is indicated for skeletally immature patients with AIS with a curve measuring 25-45o. The brace prevents further worsening of the curvature in an attempt to avoid a surgical procedure. This makes it highly significant to monitor compliance of brace-wear and its effectiveness. To successfully monitor compliance with brace treatment, a wearable multi-modal sensor solution is embedded into the patient's brace. The custom-designed hardware consists of a sensor board, a force sensor, an accelerometer and a gyroscope. In this project, we implement a novel data-mining method to identify patient activities and evaluate the effectiveness of the brace treatment pervasively. Our aim is to design a context-aware remote monitoring system for ubiquitous evaluation and enhancement of brace treatment compliance of AIS patients. Data from the sensors is filtered and processed using custom-made analyzing algorithms, a 'weekly report card' will be generated. The report card contains highly valuable information with regards to the duration of brace-wear for each day. This report card is highly beneficial to both, the physician, and the patient to confirm whether the treatment is being properly adhered to. We investigated an experimental scenario in which, the patient performs a series of pre-defined activities at home during day- long segments of brace wear. Our experimental results demonstrated that we achieved an overall accuracy of a 100% for activity detection in a semi-supervised experimental setup. It was also observed that the duration of brace wear increased from 20% to 80% in 4 weeks.

Category: Healthcare Research **Poster ID**: 16

Presenter: Mojtaba Taherisadr, Research Scientist
Computer & Information Science, U-M Dearborn

Co-Authors: Omid Dehzangi, Computer & Information Science, U-M Dearborn

Title: Shoe-based Human Activity Recognition: Multi-Pressure Sensor Characterization and Selection using Coalitional Game Theory

Abstract: Human activity monitoring and detection is of great significance in applications such as healthcare and sport. Methods for activity detection include camera based and motion sensors techniques. Motion sensors suffer from obtrusiveness which requires wearing sensor additionally, and also they are not capable of distinguishing between sitting and standing. Camera based methods are limited to the place in which camera has been installed, and also suffers from privacy issues. In this study we propose a new wearable sensor platform in order to detect the human activity efficiently and unobtrusively. Our proposed system is based on the pressure sensors embedded and distributed on the insole surface. Data has been collected for different activities including Sitting, standing, walking, running, jumping, and cycling. We first extract a comprehensive set of features including various categories such as temporal, spectral, and statistical features. Then, we implement a game-theoretic based algorithm to consider all combination of the sensors to evaluate the performance of the sensors individually, and also when they are combined with the other sensors. Results show that choosing a proper subset of channels is able to improve the results, which, in turn, leads into more computationally efficient system. Furthermore, selected subset of channels are mostly from the same region of the insole which means that region is the most informative part of the insole platform.

Category: Healthcare Research **Poster ID**: 17

Presenter: Mojtaba Taherisadr, Research Scientist
Computer & Information Science, U-M Dearborn

Co-Authors: Omid Dehzangi, Computer & Information Science, U-M Dearborn

Title: Human Gait Identification Using Two Dimensional Multi-resolution Analysis

Abstract: With the recent advances in wearable technologies like as inertial sensors, they have become widely used in various daily living applications particularly, in wearable gait analysis due to several factors, such as being easy-to-use and low-cost. Many methods have been suggested to extract various heuristic and high-level features of motion sensor data to identify discriminative gait signatures to distinguish the target individual from others. In this study, we propose a novel approach for human gait identification using spectro-temporal two dimensional (2D) expansion of human gait cycles, and then implement a 2D multi-resolution analysis technique to decompose the expanded 2D space to different levels of resolution. We propose a systematic methodology for processing non-stationary motion signals for the purpose of human gait identification with 3 major components: 1) Gait cycle extraction, 2) 2D spectro-temporal representation, and 3) Multi-resolution analysis. We collect raw motion data from five inertial sensors placed at chest, lower-back, right hand wrist, right knee and right ankle. We pre-process the raw recordings via motion signal processing and then we propose an effective heuristic segmentation method to extract gait cycle from the processed data. Spectro-temporal features are extracted by merging key instantaneous spectral descriptors in a gait cycle which characterize the non-stationarities in the each gait cycle inertial data in two dimensions (2D). The 2D time-frequency representation of the gait cycle extracted from inertial sensor data from a population of 10 subjects are decomposed to several levels of resolution. Based on the extracted features from each level of decomposed space, identification task is accomplished. Based on our experimental results, 93.36% subject identification accuracy was achieved.

Category: Healthcare Research **Poster ID**: 18

Presenter: Mojtaba Taherisadr, Research Scientist
Computer & Information Science, U-M Dearborn

Co-Authors: Omid Dehzangi, Computer & Information Science, U-M Dearborn

Title: IMU-based Gait Recognition using Convolutional Neural Networks and Multi-Sensor Fusion

Abstract: Several methods have been suggested to extract various heuristic and high-level features from gait motion data to identify discriminative gait signatures and distinguish the target individual from others. However, the manual and hand crafted feature extraction is error prone and subjective. Furthermore, the motion data collected from inertial sensors have complex structure and the detachment between manual feature extraction module and the predictive learning models might limit the generalization capabilities. In this paper, we propose a novel approach for human gait identification using time-frequency (TF) expansion of human gait cycles in order to capture joint 2 dimensional (2D) spectral and temporal patterns of gait cycles. Then, we design a deep convolutional neural network (DCNN) learning to extract discriminative features from the 2D expanded gait cycles and jointly optimize the identification model and the spectro-temporal features in a discriminative fashion. We collect raw motion data from five inertial sensors placed at the chest, lower-back, right hand wrist, right knee, and right ankle of each human subject synchronously in order to investigate the impact of sensor location on the gait identification performance. We then present two methods for early (input level) and late (decision score level) multi-sensor fusion to improve the gait identification generalization performance. We specifically propose the minimum error score fusion (MESF) method that discriminatively learns the linear fusion weights of individual DCNN scores at the decision level by minimizing the error rate on the training data in an iterative manner. 10 subjects participated in this study and hence, the problem is a 10-class identification task. Based on our experimental results, 91% subject identification accuracy was achieved using the best individual IMU and 2DTF-DCNN. We then investigated our proposed early and late sensor fusion approaches, which improved the gait identification accuracy of the system to 93.36% and 97.06%, respectively.

Category: Healthcare Research **Poster ID**: 19

Presenter: Yi Luo, Research Lab Specialist
Radiation Oncology, U-M Ann Arbor

Co-Authors: Daniel L McShan, Radiation Oncology, U-M Ann Arbor; Randall K Ten Haken, Radiation Oncology, U-M Ann Arbor; Issam El Naqa, Radiation Oncology, U-M Ann Arbor

Title: A Data-Driven Bayesian Network Approach and Its Application to Personalized Radiotherapy in Lung Cancer

Abstract: Purpose/Objective(s): The computational complexity arises exponentially with the number of nodes in a Bayesian network (BN) structure learning. The purpose of our research is to develop a practical and robust data-driven BN approach to predict tumor local control (LC) in lung cancer for personalized radiation treatment.

Methods: First, we selected relevant biophysical predictors influencing LC before and during the treatment through an extended Markov blanket (EMB) method. From these, the most robust BN structures for LC prediction were found via a wrapper-based approach. Our dataset has 98 NSCLC patients, and each had 288 extracted feature. While 68 patients were used to find full BN models for LC prediction, 30 patients were reserved for independent testing of the developed BNs. A nested cross-validation (NCV) was developed to evaluate the performance of the two-step BN approach, and an ensemble BN model was generated to evaluate the sensitivity of the two-step BN approach by comparing its similarity with the corresponding full BN model.

Results: The full BN model for LC prediction before or during the treatment yields an area under the ROC curve (AUC) of 0.81 or 0.85 based on cross-validation for BN structure learning. Independent testing of the full BN before or during the treatment for the reserved patients still yielded an AUC of 0.77 or 0.79. The two-step BN approach before and during the treatment yields an AUC of 0.75 and 0.80, respectively, per NCV. The full BN model and the ensemble BN model are highly related based on the Mantel statistic, and their correlation is statistically significant with p value= 0.001.

Conclusions: The BN approach is a stable approach for LC prediction before and during radiotherapy. The BN predictions can be improved from incorporating during treatment information, which could be an important component of decision support systems for personalized radiotherapy.

Category: Healthcare Research **Poster ID**: 20

Presenter: Huan-Hsin Tseng, Postdoctoral Fellow
Radiation Oncology, U-M Ann Arbor

Co-Authors: Jen-Tzung Chien, Electrical & Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan; Issam El Naqa, Radiation Oncology, U-M Ann Arbor; Hsin-Yi Lin, Mathematics, University of Maryland, College Park, Maryland

Title: Power-Law Distribution in Alpha-Divergence Manifold Learning

Abstract:

Purpose: Stochastic neighbor embedding (SNE) aims to transform the observations in high-dimensional space into a low-dimensional space which preserves neighborhood identities. However, data pairs in the latent space are forced to be compressed due to the loss of dimensions. We developed a new dimensionality reduction technique in SNE using power-law distribution to further relax the crowding problem and study its nonlinear embedding behavior under variation of the alpha-divergence probability measure.

Methods: The equation of motion for SNE is first analyzed using general mathematical formulation, studying the relationship and the corresponding analogy in physics mechanics. By establishing the relations, we are able to construct a strong-force type SNE transformation via introducing the power-law distribution, which enables latent variables (particles) to attract/repel neighboring particles of the same/distinct class. We further adopt alpha-divergence for probability measure rather than the usual Kullback-Leibler distance to test the role of divergence in SNE representations.

Results: Owing to the fact of strong forces of the p-SNE, better clustering effects are achieved and separation between different groups within the high-dimensional data structure is preserved in the lower-dimension embedding. The experiments of p-SNE are demonstrated on 3 image datasets: MNIST, COIL-20, and Olivetti faces for comparison. Different exponents (beta) of the power-law distribution are found to provide different sensitivity for detecting data structure. Larger beta yields sensitive results, while smaller beta shows the opposite. In the 3 learning tasks, better Davies-Bouldin indices (DBI) of p-SNE are observed compared to the traditional SNE and t-SNE, where DBI indicates the clustering effect of an unsupervised learning.

Conclusion: The novel p-SNE is proposed to provide a better dimension reduction for further mitigating the crowding problem due to the dimension compression and viewing data structure in the lower-dimension space, which also yields an avenue for different realizations on manifold learning.

Category: Healthcare Research **Poster ID**: 21

Presenter: Huan-Hsin Tseng, Postdoctoral Fellow
Radiation Oncology, U-M Ann Arbor

Co-Authors: Yi Luo, Radiation Oncology, U-M Ann Arbor; Sunan Cui, Radiation Oncology, U-M Ann Arbor; Jen-Tzung Chien, Electrical & Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan; Randall K Ten Haken, Radiation Oncology, U-M Ann Arbor; Issam El Naqa, Radiation Oncology, U-M Ann Arbor

Title: Deep Reinforcement Learning for Automated Radiation Adaptation in Lung Cancer

Abstract:

Purpose: Investigate deep reinforcement learning (DRL) applicability in radiation-based treatment plans via learning from historical data. We aim to develop an automated radiation dose adaptation framework for maximizing tumor local control at reduced rates of certain side effect such as radiation pneumonitis (RP) in the treatment of non-small cell lung cancer (NSCLC).

Methods: In a retrospective population of 114 NSCLC patients who received radiotherapy, a 3-component neural network framework was developed for reinforcement learning of dose fractionation adaptation. First, a generative adversarial network (GAN) was employed to learn from the large number of characteristics of the limited number of patients. Secondly, an artificial environment (AE) was reconstructed using deep neural network fed with the GAN data to estimate the transition probabilities between phases of the patient treatment courses. Thirdly, Deep Q-Network (DQN) of DRL was applied to the AE to choose the optimal dose in a response-adapted treatment setting. This machine learning approach was bench marked against clinical decisions actually made in an adaptive dose escalation previously used for 34 of the patients (based on avid PET imaging signal in the tumor and limited by an RP rate of 17%.

Results: Taking our adaptive dose escalation protocol as a blueprint for the GAN+AE+DQN architecture, we obtained an automated dose adaption estimate for use ~2/3 through the treatment. By letting the DQN freely control adaptive dose per fraction ranging from 1~5 Gy, it automatically favored dose escalation/de-escalation between 1.5~3.8 Gy, a range similar to one used in the protocol. Moreover, the DQN also suggested similar (but generally lower in magnitude) individual adaptive fraction doses as those used in the protocol with an RMS error=0.5 Gy.

Conclusion: We demonstrated that automated dose adaptation by DRL is promising and would yield similar results given by clinicians. However, further validation on larger datasets would still be required.

Category: Healthcare Research **Poster ID**: 23

Presenter: Sanjeeva Wijeyesakere, Postdoctoral Fellow
Cheminformatics Group, The Dow Chemical Company

Co-Authors: Dan Wilson, Dow Cheminformatics, The Dow Chemical Company; Amanda Parks, The Dow Chemical Company; Tyler Auernhammer, The Dow Chemical Company; Sue Marty, Toxicology & Environmental Research & Consulting, The Dow Chemical Company

Title: Using Public Data to Develop Open-access Computational KNIME Workflows that Identify Cholinergic Compounds

Abstract: A key goal of high-throughput data mining in the life sciences is to align compounds with mechanistic targets. As a proof-of-concept, we sought to identify the likelihood of any unknown compound interacting with the cholinergic nervous system by using both two-dimensional (2D) structural scaffolding and/or three-dimensional (3D) docking. The sympathetic and parasympathetic cholinergic nervous systems encompass the foundation of signaling at the mammalian neuro-muscular junction where neurotransmission is mediated via interaction of acetylcholine (ACh) with the nicotinic (nAChR) and muscarinic (mAChR) ACh receptors followed by its hydrolysis by acetylcholinesterase (AChE). We hypothesized that the cholinergic potential of almost any novel compound could be computationally predicted using exhaustive data mining based on the presence of structural motifs (scaffolds) shared across known cholinergic agents. We developed a KNIME workflow to capture and cluster virtually all (~19,000) compounds known to target the cholinergic system and identified 453 scaffolds that explain their structural diversity. Using the Tanimoto similarity of these scaffolds as a covariate within a random forest machine-learning algorithm, we can predict the likelihood a novel compound will target the cholinergic system with high sensitivity (99.3%) and accuracy (94%). We can predict the likelihood of a compound interacting separately with the nAChR, mAChR or AChE with similarly sensitivity and accuracy (>96%). We demonstrate added power by combining 3D docking with 2D scaffolding in predicting the interaction of compounds with target receptors (nAChR and mAChR) where non-covalent interactions drive the binding of small molecules.

Category: Healthcare Research **Poster ID**: 24

Presenter: Omar Iftikhar, Graduate Student
Computer & Information Science, U-M Dearborn

Co-Authors: Omid Dehzangi, Computer & Information Science, U-M Dearborn; Omar Iftikhar, Computer & Information Science, U-M Dearborn; Bhavani Anantapur Bache, Computer & Information Science, U-M Dearborn

Title: Smart Insole Based Human Activity Recognition Model

Abstract: Activity monitoring is important for remote health monitoring of patients. Vision based human activity monitoring are deployed in many of applications. However, such systems are not privacy preserving. Several motion-based activity monitoring systems have been developed, where wearable motion sensors are used to identify human activities. These sensors need to be worn continuously in order to monitor the activities. In this paper, we designed and developed an unobtrusive way of monitoring and identifying activities. We acquire data from high density pressure sensors that are embedded at 13 different locations of insoles. The subjects wear a pair of the 'smart insoles' and continue on with their daily lifestyle. The insole data is analyzed to identify different activities including sitting, standing, walking, running, cycling and jumping. We present two types of analysis including subject independent and subject dependent sensor based analysis. In the subject independent analysis, we identity the sensors which are important in identifying different activities. In the subject dependent analysis, we analyze the impact of different sensors on subjects. We also evaluated the impact of different sensors on different subjects. Machine learning and classification techniques are used to generate a reliable and robust activity detection model. Even though this project related to patient-specific activity detection, this can be expanded to applications such as sports and fitness. The methodology and data analysis remain consistent. Our aim is to use the 'smart insoles' to provide a highly accurate method of activity detection.

Category: Healthcare Research **Poster ID**: 25

Presenter: Jerome Cheng, Assistant Professor
Pathology Informatics, U-M Ann Arbor

Co-Authors: Ulysses Balis, Pathology Informatics, U-M Ann Arbor; David McClintock, Pathology Informatics, U-M Ann Arbor

Title: Use of A High-Dimensional Stochastic Classifier (VIPER) in Tandem with Deep Learning Approaches to Detect and Quantify Crescentic Glomerulonephritis

Abstract: Ulysses G .J. Balis*, Jerome Y. Cheng*, David McClintock
*These authors contributed equally to the work

Background: Crescentic glomerulonephritis is a severe form of glomerulonephritis, defined by crescent formation present in greater than 50% of the glomeruli. Traditionally, diagnoses are arrived upon by histopathological interpretation by pathologists, who manually count the percentage of glomeruli involved by crescentic changes. In this report, we present an automated method by which accurate counting of glomeruli is carried out by a highly automated machine learning computational pipeline. This approach is novel in that it exceeds conventional convolutional neural network approached by use of a high-dimensional stochastic spatial pre-filter (VIPER – Validated Idempotent Pattern Extraction & Recognition) , which benefits from subject matter expertise directing the initial classification archetypes.

Design: Sections of whole slide images (WSIs) were manually annotated for areas containing characteristic crescentic glomerular changes, generating multiple candidate VIPER-based feature vectors. Multiple heatmaps representing the goodness of fit of each vector as applied to the total surface area of the field of view, were created using the VIPER algorithm. This data, in tandem with ground truth maps, was utilized to generate a machine learning model optimized for the detection of both normal and diseased glomeruli. The same vector set in tandem with the realized machine learning model was then applied to different cases and images, and in so doing, generating a second set of heatmaps. These were applied to the model to predict the regions comprising glomeruli exhibiting crescentic injury.

Results: The combined VIPER/ deep learning pipeline was able to identify the far majority of glomeruli exhibiting crescentic injury in the cross-validation set of WSI images.

Conclusion: Machine learning tools, in tandem with spatial, high-dimensional stochastic filters, such as VIPER, are a highly effective methodology for the detection and classification of glomerular injury, such as crescentic glomerulonephritis.

Category: Healthcare Research **Poster ID**: 26

Presenter: Hanbo Sun, Graduate Student
Statistics, U-M Ann Arbor

Co-Authors: Ivo Dinov, Health Behavior & Biological Sciences, Computational Medicine &Bioinformatics, U-M Ann Arbor

Title: BrainPrint: Computable Volume Phenotypes

Abstract: Machine learning methods applied to neuroimages for disease diagnosis and prediction have been extensive researched in recent years. However, two critical problems exist. First, though many recent publications have studied brain relevant phenotypes, Alzheimer's Disease, brain lesion and brain age for example, which showed neuroimaging and convolutional neural network are useful tools for computer assisted intervention for brain disease diagnosis and phenotypes prediction, less work has been done on predicting phenotypes that intuitively less relevant to brain volume, such as Autism spectrum disorder(ASD), intelligence quotient (IQ), etc. Another vital part is feature representation. To our best knowledge, the previous methods in the literatures either used hand-crafted features, which are awkward and potentially problematic, or extracted and fused features based on pre-training. Self-taught manners usually outperform manual methods. However, simpler but computationally efficient feature representation method is expected.

To address these two problems, we proposed a 7 layers Convolutional Neural Network (ChpCNN) architecture to predict ASD on 3D MRIs, on which we achieved overall 64% and Autism disease group 69% accuracy. Further, by using ensemble method, both the overall accuracy and sensitivity improved ~2%. We compared the performances with that of multiple machine learning methods and VGG16. Also, compared it with various ML methods and LeNet on 2D axial view slices. Ensemble ChiCNN outperforms all benchmarks. Finally, we applied 3D wavelet transform and denoised wavelet coefficients to obtain 3D sparse brain feature representation, which we named as BrainPrint. Comparable performance was achieved by BrainPrint.

Future, we hope to understand more from BrainPrint. Based on that, our ambition is to pre-train multiple models for various phenotypes so that other researches or clinical sites, that require computer intervention diagnosis, can use the pre-trained models directly or update models based on transfer learning by adding small scale additional data points.

Category: Healthcare Research **Poster ID**: 29

Presenter: Zhenke Wu, Assistant Professor
Biostatistics, School of Public Health; MIDAS, U-M Ann Arbor

Co-Authors:

Title: Estimating AutoAntibody Signatures to Detect Autoimmune Disease Patient Subsets

Abstract: Autoimmune diseases are characterized by highly specific immune responses against molecules in self-tissues. Different autoimmune diseases are characterized by distinct immune responses, making autoantibodies useful for diagnosis and prediction. In many diseases, the targets of autoantibodies are incompletely defined. Although the technologies for autoantibody discovery have advanced dramatically over the past decade, each of these techniques generates hundreds of possibilities, which are onerous and expensive to validate. We set out to establish a method to greatly simplify autoantibody discovery, using a pre-filtering step to define subgroups with similar specificities based on migration of radiolabeled, immunoprecipitated proteins on sodium dodecyl sulfate (SDS) gels and autoradiography [Gel Electrophoresis and band detection on Autoradiograms (GEA)]. Human recognition of patterns is not optimal when the patterns are complex or scattered across many samples. Multiple sources of errors - including irrelevant intensity differences and warping of gels - have challenged automation of pattern discovery from autoradiograms. In this paper, we address these limitations using a Bayesian hierarchical model with shrinkage priors for pattern alignment and spatial dewarping. The Bayesian model combines information from multiple gel sets and corrects spatial warping for coherent estimation of autoantibody signatures defined by presence or absence of a grid of landmark proteins. We show the pre-processing creates more clearly separated clusters and improves the accuracy of autoantibody subset detection via hierarchical clustering. Finally, we demonstrate the utility of the proposed methods with GEA data from scleroderma patients.

Category: Healthcare Research **Poster ID**: 41

Presenter: Daniel; Prahlad, Tejas Zeiberg, Undergraduate Students
Computer Science & Engineering, U-M Ann Arbor

Co-Authors: Jenna Wiens, Computer Science Engineering, U-M Ann Arbor; Michael Sjoding, Internal Medicine, U-M Ann Arbor

Title: Data-Driven Tools for Patient Risk Stratification for ARDS

Abstract: The Acute Respiratory Distress Syndrome (ARDS) develops in the lungs of critically ill patients and prevents effective gas transport, leading to severely low oxygen levels in these patients. Mortality rates associated with ARDS range from 26-58%. This high mortality rate is due in part to the fact that an estimated 70% of cases are diagnosed late or not at all. We hypothesize that accurate and early patient risk stratification for ARDS will lead to more timely diagnosis/treatment and better outcomes. Current risk stratification models for ARDS (e.g., the Lung Injury Prediction Score, LIPS) are based on a small number of risk factors that are not easily abstracted from the electronic health record (EHR). In contrast, we propose a risk stratification model that utilizes clinical features based on routinely available longitudinal EHR data. Trained on data from 307 ICU patients, the proposed model uses data from the first six hours of a patient's stay to predict whether the patient will develop ARDS at any subsequent point during their hospitalization. Applied to a held-out dataset of 76 patients, our model achieves good discriminative performance (AUC of 0.81 [95% C.I.: 0.59-0.93]). Moreover, we outperform the LIPS model, which requires significantly more effort to extract input variables, (AUC of 0.73 [95% C.I.: 0.53-0.88]). The EHR represents a rich source of medical data useful for developing an accurate ARDS risk stratification model, which has the potential to help clinicians provide better care to ARDS patients if ultimately deployed in a clinical environment.

Category: Healthcare Research **Poster ID**: 46

Presenter: Neida Sechague, Research Assistant
Public Health, U-M Flint

Co-Authors: Rie Suzuki, Public Health, U-M Flint; Kimberly R. Warden, Hamilton Community Health Network, Inc.

Title: The Prevalence of Pregnancy Loss in Flint, Michigan

Abstract: Poverty is associated with pregnancy loss and may hinder aspects of prenatal care usage. Although health promotion programs targeting maternal and child health are available, few studies have investigated types of pregnancy loss in women living in Flint, Michigan. The purpose of the study was to identify the relationship between the timing of prenatal care use and pregnancy loss. Data were derived from a Flint community health center's electronic medical records from 2010–2016 for adults aged 17–45. The dependent variable was pregnancy loss, indicating either miscarriages (fetal deaths occurring at < 20 weeks gestation) or stillbirths (fetal deaths occurring at > 20 weeks gestation). Descriptive statistics and logic regression analyses were conducted. Of 182 women who experienced pregnancy loss, a majority were African American (70%), 20 years and older (89%), obese (78%), non-smokers (98%), lived in Flint (70%), had Medicaid (77%), used the Main or North Point clinics (68%), visited during the first trimester (75%), and experienced miscarriages (59%). The logistic regression analysis revealed that initial prenatal care use during second trimesters was associated with the frequency of stillbirths (AOR = 6.11). The interaction effects indicated that race and the timing of prenatal care use did not have an impact on the types of pregnancy loss. To summarize, delayed prenatal care use affected the prevalence of types of pregnancy loss. Further studies are needed to identify possible major barriers of first trimester visits to prevent pregnancy loss among African American obese adult women living on the north side of Flint, Michigan.

Category: Healthcare Research **Poster ID**: 49

Presenter: Katherine Hoffman, Graduate Student
Biostatistics, U-M Ann Arbor

Co-Authors: Peter XK Song, Biostatistics, U-M Ann Arbor; Rajiv Saran, Nephrology, U-M Ann Arbor; Jennifer Bragg-Gresham, Internal Medicine, U-M Ann Arbor

Title: Learning Impact of Air Pollution on Kidney Disease Epidemiology in US and East China Using Big Public Health Data

Abstract: The incidence rate of end stage renal disease (ESRD) and the mortality rate among ESRD patients are high in both China and the US. Air quality in China has been worsening over the past few decades and often reaches or exceeds hazardous level in many regions. Understanding the adverse effect of the exposure level to the air pollution on the ESRD patients' mortality and morbidity is of great importance but has been insufficiently studied in both China and the US. ESRD patients are subject to high risk of morbidity and co-mortality and are more susceptible to adverse outcomes attributable to particle pollution exposure. The significant difference of exposure to suspended particulate matter and hazardous gases in China (high) and US (low) enables us to compare and assess risk of exposure to air pollution on kidney disease incidence and distribution. Thus, it allows us to develop prevention measures for control of disease and other risk factors relating to well-being. Currently, daily means of particulate matter and hazardous gas levels from air monitors by county are being collected from the Environmental Protection Agency's Air Quality DataMart. Meanwhile in Shanghai, similar regional data collection is ongoing. The next steps for the US data will include developing a model-based kriging for particle pollution exposure. Cox regression adjusting for various county-level socioeconomic covariates will be run using ESRD survival rates obtained from the United States Renal Data System. The results of the geospatial associations, in combination with the results from Shanghai's parallel analysis, are expected to help facilitate the development of prevention measures and related public health policies in both the US and China.

Category: Healthcare Research **Poster ID**: 51

Presenter: Mojtaba Taherisadr, Research Scientist
Computer & Information Science, U-M Dearborn

Co-Authors: Omid Dehzangi, Computer & Information Science, U-M Dearborn

Title: Location Detection Automation for Preclinical Stereotactic Neurosurgery Procedure Using Convolutional Neural Network and Template Matching

Abstract: Locating desired area for any further actions in stereotactic neurosurgeries is time consuming and prone to human error. Thus it requires automation to accurately and efficiently locate the region of interest (ROI) and desired spot (e.g. injection site). Distortions such as change in shape due to camera lens, different lighting conditions, different poses, and presence of partial occlusions, horizontal and vertical shifts are the inherent challenges involved with image processing-based automation techniques. In this study we propose a general framework based on two complementary areas including deep convolutional neural network (DCNN) and template matching to efficiently automate the localization process. We first detect the ROI using DCNN, and then localize and specify the injection site by applying previously designed custom template. In order to evaluate the performance of the proposed method, we implement an empirical study to detect the ROI in rodents, and then localize the injection site for needle insertion, or implementation of electrode/cannula/optic fiber.

Poster Abstracts

Category: Healthcare Research **Poster ID**: 54

Presenter: Sunan Cui, Graduate Student
Applied Physics, U-M Ann Arbor

Co-Authors: Randall K Ten Haken, Radiation Oncology, U-M Ann Arbor; Issam El Naqa, Radiation Oncology, U-M Ann Arbor

Title: Prediction of Toxicity in Radiotherapy with Mixture of Sequential and Non-sequential Data by Deep Neural Network

Abstract:

Purpose: A major challenge in application of deep neural network to predict radiotherapy outcome is the limited sample size, therefore, to take advantage of the correlation among sequential data, reducing the size of samples needed to train the neural network, we investigated two architectures which take a mixture of sequential and non-sequential data as input and output prediction of radiation-induced lung damage.

Method: We investigated the proposed architectures on a heterogeneous dataset contains 106 patients treated with radiotherapy. Features include 3 micro-RNAs, 1 SNP (single nucleotide polymorphism), 4 cytokines and 2 radiomics from PET (positron emission tomography), among which cytokines and image information were measured over several times during the fractionated treatment. The first architecture used gated recurrent unit (GRU) to take sequential data as input, then, the output was concatenated with non-sequential data, and fed into following fully-connected layers. Both GRU and fully-connected layer were implemented with “dropout” to mitigate overfitting. The second architecture processed sequential data with 2-dimension locally-connected layer, whose length of filter and stride were set properly to ensure desired grouping. This layer is similar to 2D convolutional layer except the weights are not shared. Similarly, the output of the 2D locally-connected layer was merged with non-sequential data and then fed into following-connected layers. Linear activation function was implemented in 2D locally-connected layer.

Result: We varied the parameters of GRU (dropout rates, dimensionality of outputs), the architecture yielded the best AUC (the area under ROC curve) of 0.727 on 5-fold cross-validation. The second architecture achieved AUC of 0.76 on 5-fold cross-validation.

Conclusion: Architectures considering correlation among feature can potentially reduce the size of parameters in the neural network which is especially desirable in a small-size sample. Compared with GRU, locally-connected layer yielded better results in proposed architecture to predict toxicities in radiotherapy.

Category: Healthcare Research **Poster ID**: 55

Presenter: Muhamed Farooq, Graduate Student Research Assistant
Computer & Information Science, U-M Dearborn

Co-Authors:

Title: Portable Brain Computer Interface for the Intensive Care Unit Patient Communication using Subject-Dependent SSVEP Identification

Abstract: A major predicament for Intensive Care Unit (ICU) patients is inconsistent and ineffective communication means. Patients rated most communication sessions as difficult and unsuccessful. This, in turn, can cause distress, unrecognized pain, anxiety, and fear. Therefore, ICU needs should be communicated quickly and effectively. As such, we designed a portable BCI system for ICU communications optimized to operate effectively in an ICU environment. The system utilizes a wearable EEG cap coupled with an Android app designed on a mobile device that serves as visual stimuli and data processing module. Furthermore, to overcome the challenges that BCI systems face today in real-world scenarios, we propose a novel subject-specific Gaussian Mixture Model (GMM)-based training and adaptation algorithm. First, we incorporate subject-specific information in the training phase of the SSVEP identification model using GMM-based training and adaptation. We evaluate subject-specific models against other subjects. Subsequently, from the GMM discriminative scores, we generate the transformed vectors, which are passed to our predictive model. Finally, the adapted mixture mean scores of the subject-specific GMMs are utilized to generate the high dimensional super vectors. Our experimental results demonstrate that the proposed system achieved 98.7% average identification accuracy, which is capable of providing effective and consistent communication for ICU patients.

Category: Healthcare Research **Poster ID**: 57

Presenter: Yaya Zhai, Graduate Student
Computational Medicine & Bioinformatics, U-M Ann Arbor

Co-Authors: Alfred O. Hero, Electrical Engineering & Computer Science, Data Science Initiative, U-M Ann Arbor

Title: Temporal Gene Expression Network and Regulatory Mechanism During Acute Respiratory Viral Infection

Abstract: Acute respiratory viral infection is one of the most prevalent diseases and poses considerable economic burdens to the society. But to date, there is no system biological study of acute respiratory viral infection in subject level yet. Here we constructed and analyzed a temporal gene co-expression network and explored the regulatory mechanism during the development of infection.

Category: Healthcare Research **Poster ID**: 61

Presenter: Teal Guidici, PhD Candidate
Statistics, U-M Ann Arbor

Co-Authors: George Michailidis, Informatics Institute, University of Florida

Title: Detecting Differentially Expressed Metabolic Pathways with Adjustments for Macronutrient Intake

Abstract: Differential expression testing and set enrichment analysis are commonly used to summarize the results of high throughput biological experiments, to generate biologically meaningful hypothesis for further analysis. and to aid in the planning of validation experiments. Conventional approaches to differential expression testing and set enrichment analysis do not usually account for individual variation in relevant background features, in many cases due to lack of pertinent data. These features are especially relevant in the context of metabolomics, where blood metabolite levels can react sensitively and quickly to changes in nutrient intake. In this project we introduce a network based method for detecting differentially expressed metabolites and metabolic pathways, while adjusting for individual variation in the consumption of relevant macronutrients through the integration of nutrition intake data. We test our method on metabolomic and nutrition intake data from a controlled feeding study featuring two distinct diets (a high polyunsaturated fat diet and a high carbohydrate diet)

Category: Healthcare Research **Poster ID**: 75

Presenter: Chao Gao, Graduate student
Biostatistics, U-M Ann Arbor

Co-Authors:

Title: Model-based and Model-free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease

Abstract: Parkinson's disease (PD) is an age-related neurodegenerative disorder affecting over 10 million people worldwide. In PD patients, loss of dopaminergic neurons in the Substantia Nigra are observed. PD itself is not fatal, however, people with PD suffer from disrupted balance, tremor, rigidity, slowness in movement, etc. Patients are at high risk of falling due to the motor impairment or freeze of gait. Serious falls, especially among elderly patients, can lead to a variety of injuries including head injuries and bone fractures. Complications from these injuries may further lead to hospitalization or long-term rehabilitation and death at worst.

In this study, our goal is to investigate the clinical, demographic and neuroimaging information of PD patents from two different sources (University of Michigan and Tel Aviv University) using machine learning techniques and construct predictive models classifying fallers and non-fallers among PD patients. Through feature selection, we successfully identified important predictors for patient's fall, such as gait speed, Hoehn and Yahr scale as well as postural instability and gait difficulty-related measurements. The methods we applied for modeling includes logistic regression, random forests, support vector machines, XGboost, etc. As evaluated by 5-fold cross validation and external validation, the obtained classification models achieve from 70 to around 80% overall accuracy, depending on the dataset used.

In the future, we look forward to expanding our study on larger dataset and diving deeper in to neuroimaging data (e.g. MRI), with the hope to recognize patterns associated with falling of PD patients in the brain. The ultimate goal is building reliable statistical models that can be applied to provide PD patients with more individualized health care based on their personal and medical information.

Category: Healthcare Research **Poster ID**: 76

Presenter: Xuefei Zhang, Graduate student
Statistics, U-M Ann Arbor

Co-Authors: Tanner Caverly,Learning Health Sciences, Internal Medicine, U-M Ann Arbor; Akbar Waljee, Internal Medicine, U-M Ann Arbor; Ji Zhu, Statistics, U-M Ann Arbor

Title: The reliability of eliciting detailed tobacco history in the clinic and the clinical implications of imperfection

Abstract:

Background: Determining whether a person is eligible for lung cancer screening with low-dose computed tomography has important clinical consequences. Because this determination requires eliciting a person's detailed smoking history, understanding the reliability of this history when elicited in time-pressured clinical settings is critical.

Methods: We conducted a retrospective cohort study of 10,449 patients who had a detailed smoking history documented in their medical record at least twice between 10/2013-7/2017, across 8 academic Veterans Affairs (VA) health systems. We examined 2 specific tobacco-history measures -- a person's smoking duration (in years) and smoking intensity (average packs/day while smoking). First, we fit a linear random effects model to quantify the reliability of both measures (smoking-duration and smoking-intensity) using the intraclass correlation coefficient (ICC). Next, we used simulation to examine how the observed reliability of both measures moderates the accuracy of determining screening-eligibility, and how different interventions could improve this accuracy.

Results: We found that eliciting smoking duration in real-world clinical settings had good reliability (ICC=0.67), while eliciting smoking intensity had fair reliability (ICC=0.55). This imperfect reliability led to false-positive and false-negative determinations of screening-eligibility in our simulations: 11.4% false-positive rate (10.9%-12.0%) and 12.24% false-negative rate (11.8%-12.8%). Taking the average of repeated smoking-history assessments would improve accuracy. Repeating the assessment at 5 time-points would decrease false-positive rates to 9.1% and false-negative rates to 4.3%. If the reliability of these real-world assessments could be improved to match that observed in standardized national surveys, the false-positive rate of a single assessment would decrease to 8.1% and false-negatives to 8.8%.

Conclusions: Measuring detailed smoking history in real-world practice is inherently imperfect, and this leads to substantial misclassification of patients when determining lung cancer screening eligibility. Interventions to improve the reliability of this history can modestly decrease misclassifications, but may be difficult to achieve in practice.

Category: Law **Poster ID**: 3

Presenter: Daniel W. LinnaJr, Visiting Professor of Law
Law School, U-M Ann Arbor

Co-Authors:

Title: Measuring Legal-Service Delivery Innovation & Technology Adoption

Abstract: From www.LegalTechInnovation.com : This study is part of a pilot project to create an index of legal-service delivery innovation. ... The problem to be solved is the lack of access to legal services. Experts estimate that approximately 80 percent of the impoverished and 50 percent of the middle class lack access to legal services. Even corporations say that they do not get what they need from their lawyers. There are many opportunities for lawyers to better serve their clients and the public and a growing number of lawyers are seizing these opportunities.

This index is intended to serve both consumers and producers of legal services. My hope is that clients, including legal departments, will consult this index and engage in deep discussions with their lawyers about how to improve legal-service delivery. Those discussions should include not only how to improve efficiency, but also how to improve quality and obtain better substantive outcomes.

This index is also intended to be a resource for the producers of legal services, from lawyers in legal departments and law firms to technologists, project managers, data analysts, and other professionals across the industry, including legal startups, legal process outsourcers, and alternative legal service providers. Our discussions about legal innovation and technology tend to include a lot of generalizations. One purpose of the index is to get more specific about innovation and technology, including the disciplines and tools driving it forward and the substantive legal areas where we see activity.

This index should also be a resource for law schools and law students. It will help law schools better understand the evolution of the legal landscape, which will help them better prepare their students for the future. Law students can use this index to learn more about how the profession is changing and the knowledge and skills that they should develop for long-term success. The index also aims to provide law students information about the law firms recruiting them as well as a framework for assessing each law firm's strategies for the future. Again, I caution that this index is simply an initial attempt to measure indicators of innovation and various weaknesses have been acknowledged. That said, the index and this initial information provides a starting point for very important discussions.

Finally, this index should be a resource for improving access to legal services. The various sectors of the legal industry–from legal aid and courts to legal departments and law firms–face challenges that are similar in many ways. These sectors can learn a lot from each other. They could work together on certain problems for the benefit of all. Law schools could play a role coordinating that collaboration, serving as laboratories for innovation and research and development. This provides the opportunity for law firms and legal departments to do well by doing good, partnering with law schools and legal aid organizations to improve legal-service delivery and access to legal services for everyone.

Category: Learning Analytics **Poster ID**: 15

Presenter: SungJin Nam, Graduate Student
School of Information, U-M Ann Arbor

Co-Authors: Kevyn Collins-Thompson, School of Information, U-M Ann Arbor

Title: Predicting Short- and Long-Term Vocabulary Learning via Semantic Features of Partial Word Knowledge

Abstract: We show how the novel use of a semantic representation based on Osgood's semantic differential scales can lead to effective features in predicting short- and long-term learning in students using a vocabulary learning system.

Previous studies in measuring students' intermediate knowledge states during vocabulary acquisition did not provide much information on which semantic knowledge students gained during word learning practice. Moreover, these studies relied on human ratings to evaluate the students' responses. To solve this problem, we propose a semantic representation for words based on Osgood's semantic decomposition of vocabulary.

Our method provides a representation of a word in ten different semantic scales. For example, each student response is converted to cosine similarity score of word vectors for the response and difference between each semantic scale's end (e.g., good and bad, or big and small) from Word2Vec. Target words, which are difficult words that presented to students from the system as a learning task, are also evaluated in the same way. A comparison between two representations is expected to capture student's partial knowledge for the target word.

To demonstrate our method can effectively represent students' knowledge in vocabulary acquisition, we build models for predicting the student's short-term vocabulary acquisition and long-term retention. We compare the effectiveness of our Osgood-based semantic representation to that provided by Word2Vec neural word embedding, and find that prediction models using features based on Osgood scale-based scores perform better than the baseline and are comparable in accuracy to those using Word2Vec score-based models.

By using more interpretable Osgood-based scales, our study results can help with better understanding of students' ongoing learning states and design personalized learning systems that can address an individual's weak points in vocabulary acquisition.

Category: Learning Analytics **Poster ID**: 43

Presenter: Shibamouli Lahiri, Graduate Student
Computer Science & Engineering, U-M Ann Arbor

Co-Authors: Carmen Banea, Computer Science & Engineering, U-M Ann Arbor; Rada Mihalcea, Computer Science & Engineering, U-M Ann Arbor

Title: Matching Graduate Applicants with Faculty Members

Abstract: Every year, millions of students apply to universities for admission to graduate programs (Master's and Ph.D.). The applications are individually evaluated and forwarded to appropriate faculty members. Considering human subjectivity and processing latency, this is a highly tedious and time-consuming job that has to be performed every year. In this paper, we propose several information retrieval models aimed at partially or fully automating the task. Applicants are represented by their statements of purpose (SOP), and faculty members are represented by the papers they authored. We extract keywords from papers and SOPs using a state-of-the-art keyword extractor. A detailed exploratory analysis of keywords yields several insights into the contents of SOPs and papers. We report results on several information retrieval models employing keywords and bag-of-words content modeling, with the former offering significantly better results. While we are able to correctly retrieve research areas for a given statement of purpose (F-score of 57.7% at rank 2 and 61.8% at rank 3), the task of matching applicants and faculty members is more difficult, and we are able to achieve an F-measure of 21% at rank 2 and 24% at rank 3, when making a selection among 73 faculty members.

Category: Learning analytics **Poster ID**: 44

Presenter: August Evrard, Professor
Physics, Astronomy, Digital Innovation Greenhouse, Office of Academic Innovation, U-M Ann Arbor

Co-Authors: Kyle Shultz, Digital Innocation Greenhouse, U-M Ann Arbor; Dhayaa Anbajagane, Physics, U-M Ann Arbor; Raihan Haque, U-M Ann Arbor; Kerby Sheddon, Statistics, CSCAR, U-M Ann Arbor

Title: Gender Bias in Introductory STEM Grades: Clues from Problem Roulette

Abstract: At the University of Michigan, multi-year samples show that men historically outperform women in final grade earned in large introductory STEM courses, with quarter-letter grade mean differences typical. Use of multiple choice examinations for discriminatory assessment is a common feature of such courses. In this work, we present evidence indicating that the multiple choice format itself is not the primary factor driving the gender gap in grades. Analyzing millions of student-problem interactions on the optional Problem Roulette preparation service, we either do not detect or find greatly reduced gender differences in correct response rates in six Physics, Chemistry and Statistics courses. We also present comparisons of study behaviors for men and women, finding modest differences that are generally smaller than differences in study behavior by academic subject.

Category: Learning analytics **Poster ID**: 45

Presenter: Nia Dowell, Postdoctoral Research Fellow
School of Information, U-M Ann Arbor

Co-Authors: Christopher Brooks, School of Information, U-M Ann Arbor

Title: Modeling Temporal Changes in the MOOC Learner Population

Abstract: Recent innovations in educational technology brought with it the introduction of massive open online courses, commonly referred to as MOOCs. Since the rise of their popularity in 2012, there has been an emerging trend in higher education for the adoption of MOOCs. However, despite this interest in learning at scale, there has been limited work investigating how MOOC participants have changed over time. In this study, we explore the temporal changes in MOOC learners' language and discourse characteristics. In particular, we demonstrate that there is a clear trend within a course for language in discussion forums to be of both more on-topic and reflective of deep learning in subsequent offerings of a course. We measure this in two ways, and demonstrate this trend through several repeated analyses of different courses in different domains. While not all courses show an increase beyond statistical significance, the majority do, providing evidence that MOOC learner populations are changing as the educational phenomena matures. The findings have important implications for both research and instructional design in scaled learning environments.

Category: Materials Science **Poster ID**: 50

Presenter: Jie Shen, Professor
College of Engineering & Computer Science, U-M Dearborn

Co-Authors: David Yoon, Computer & Information Science, U-M Dearborn; Jinghui Mao, Computer & Information Science, U-M Dearborn; Yujie Liu, Computer & Information Science, U-M Dearborn; Qiuwei He, Computer & Information Science, U-M Dearborn; Hao Chen, Computer & Information Science, U-M Dearborn

Title: Understanding Materials Aging and Degradation via Data Sampling, Modeling and Simulation

Abstract: Environmental or functional aging and degradation of materials are an important natural phenomenon that has a profound impact on the safety or integrity of components, structures, and systems. Understanding and predicting the behavior of aging and degradation will provide a fundamental insight into how to control and mitigate the impacts. The Virtual Engineering Lab at the University of Michigan-Dearborn has conducted a series of studies on the materials degradation behavior via data sampling, modelling and simulation. As a leader in the digital diagnosis of materials degradation, we have made the following contributions: (1) Establishment of a multi-resolution transformation rule of material defects for the first time, (2) Design of an accurate digital diagnosis method for material damage, (3) Reconstruction of defects in material domains from X-ray CT data, (4) Parallel computation of materials damage, and (5) Extensive application potentials in various industries.

Poster Abstracts

Category: Science and Society **Poster ID**: 6

Presenter: Muzammil M. Hussain, Assistant Professor
Communication Studies, Center for Political Studies, Institute for Social Research, U-M Ann Arbor

Co-Authors: Fan Liang, Communication Studies, U-M Ann Arbor; Vishnupriya Das, Communication Studies, U-M Ann Arbor; Nadiya Kostyuk, Political Science, U-M Ann Arbor

Title: Constructing a Harmonious Data-Driven Society: China's Social Credit System as a State Surveillance Infrastructure

Abstract: Big data technologies have been adopted by both public and private sector actors to develop and expand surveillance capacities. This study traces the institutional process and political-economic interests of public and private stakeholders advancing the construction of China's national "Social Credit System" (SCS), currently on-track for full deployment by 2020 upon 1.4 billion citizens. The official goal of the SCS is to centralize multiple existing public and private platforms into a single big data-enabled surveillance infrastructure that is designed to manage, monitor and predict the trustworthiness of each actor in China. By conducting an inductive grounded case study of the SCS's historical context and developmental trajectory, this study identifies the emergent institutional features constituting a state surveillance infrastructure: state-sponsored big data-enabled surveillance efforts that are increasingly instrumentalized by state powers to govern and regulate the political, economic, and social dominions of society. By taking explicit interest in the behind-the-scenes organizational forces and processes shaping the construction of the SCS, we propose that the 'state surveillance infrastructure' framework allows for nuanced comparative assessments of similar state-backed surveillance platforms and practices rapidly being integrated for public administration by state powers across the world.

Category: Science and society **Poster ID**: 52

Presenter: Junqi Min, Graduate Student
Management Studies, U-M Dearborn

Co-Authors: Omar Bah, Management Studies, U-M Dearborn

Title: Analysis of Dearborn Fire Department Network

Abstract: When it comes to response time of 911 calls, time is life. This project analyzes real data and proposes solutions for the Dearborn Fire Department who currently provides service to approximately 110,000 residents covering 27.3 square miles. Specifically, the project focuses on locating a second fire station to serve both district 2 and 5 (the west side of Dearborn), districts that currently have longer response time on average. To start with, we compute distances for subsequent analyses as response time is largely contingent on distance. We then apply SITATION (a software application for location analysis) and various algorithms to identify a theoretically optimal solution. In addition, we develop multiple approaches to work around limitations associated with SITATION. Optimal solutions by different approaches are compared to a pre-determined location, and their implications to the response time are examined. We wrap up the project with final recommendations and takeaway points for the Dearborn Fire Department.

Category: Science and society **Poster ID**: 59

Presenter: Natsuko Nicholls, Research Manager
Institute for Research on Innovation & Science; Survey Research Center, Institute for Social Research, U-M Ann Arbor

Co-Authors: Jason Owen-Smith, Sociology, Institute for Social Research, U-M Ann Arbor; Xiyun Xiang, School of Information, U-M Ann Arbor; Yi Xiao, School of Information, U-M Ann Arbor; Oshin Nayak, School of Information, U-M Ann Arbor

Title: Nexus Between Research and Public Engagement

Abstract: This poster presentation addresses the public impact of research, focusing on the way that scientific knowledge is translated into society through researchers' public engagement and services. In estimating the public value of research, a single dataset or database will rarely suffice. Linking records from multiple data sources is a promising method and we demonstrate benefits of linking our own dataset IRIS data (administrative data on award and research activities) to other data sources, including a large dataset of federal advisory committee members who advised the President and the Executive branch on various policy issues between 1997 and 2016.

In our poster presentation, we focus on both technical and theoretical components. Record linkage is to bring together information from two records that we believe relate to the same entity, for instance, the same individual (linking asset). To bring together the records for the same entity across datasets, however, often comes with challenges that lie in data quality, data preprocessing (e.g., standardization and parsing), and record linkage techniques (e.g., deterministic vs. probabilistic record linkage methods, string comparator metrics, blocking strategies as stratification, etc.). We discuss our record linkage method, describing how we solved these technical and algorithmic challenges.

Our presentation also discusses the finding that funded research are significantly related to federal advisory committee service. With a data visualization tool (e.g., alluvial charts), we showcase our finding that helps to make academic public engagement (via federal advisory committee service data) visible and measurable in the context of university research activities. This presentation focuses on cases of top leading research universities.

Category: Social Science **Poster ID**: 4

Presenter: Jonathan Kummerfeld, Postdoctoral Research Fellow
Computer Science & Engineering, U-M Ann Arbor

Co-Authors:

Title: Automatic Disentanglement of Multi-Party Conversations Online

Abstract: In comment threads, Slack channels, Google Hangouts and other multi-party communication settings there are often multiple threads of conversation mixed together. This makes analysis for data science over such resources difficult. To address this, we have developed new neural network models for discourse disentanglement: determining for every message which other message(s) it is a response to. Previous work has used rules to extract threads of conversation, but these miss the internal structure of each thread, and often make mistakes. By annotating a new dataset of 20,000 messages of IRC logs with discourse structure, and training our new models of discourse, we have been able to extract a more detailed and accurate set of dialogue structures.

Category: Social Science **Poster ID**: 34

Presenter: Charles Crabtree, PhD Candidate
Political Science, U-M Ann Arbor

Co-Authors: Quintin Beazer, Florida State University; Holger L. Kern, Florida State University

Title: Diverting Attention: Media Coverage of Economic Conditions on Russian State-Controlled TV, 2003-2016

Abstract: How do authoritarian governments use media to shape perceptions about economic performance? Given the economy's importance to regime legitimacy, we argue that authoritarian governments have both the incentive and the ability to reduce news coverage of economic performance during protracted periods of economic decline or crisis. In this paper, we analyze 14 years of nightly news broadcasts from a state-controlled television station in Russia, Channel One, to examine how the composition of pro-governmental news coverage changes in response to souring economic conditions. We identify three primary categories of news segments that act as replacements for news about a weak domestic economy: coverage of positive political events (pro-regime promotion), non-political human interest stories (filler), and negative foreign and international economic news (benchmarking).

Category: Social Science **Poster ID**: 42

Presenter: Colleen McClain, PhD Candidate
Survey Methodology, Institute for Social Research, U-M Ann Arbor

Co-Authors: Zeina Mneimneh, Survey Research Center, Institute for Social Research, U-M Ann Arbor; Trivellore Raghunathan, Survey Research Center, Institute for Social Research, U-M Ann Arbor; Lisa Singh, Computer Science, Georgetown University

Title: Ground Truthing in Social Media Research: Assessing Assumptions Required for Demographic Inference from Twitter

Abstract: The increasing number of studies that use social media to predict and quantify social trends have raised questions on how and whether "big" textual data sources can supplement or replace traditional survey data collection. Yet, extracting social media data for social science research raises questions related to measurement, linkage, and inference. Inferring measures related to human behaviors, attitudes, or beliefs to a specific population requires adjusting for known mismatches between social media samples and the target population. Owing to this challenge, a number of researchers have sought to build predictive models to draw inferences about demographic characteristics of social media users; and many such studies focus on Twitter, since data is often public but lacks user-provided demographic labels. Many of these existing studies, however, suffer from one of two limitations: 1) relying on "ground truth" or validation assumptions that are resource-heavy and prone to human error (such as manual annotation), or 2) restricting the validity of inferences to a subset of users who disclose optional personal information, or who could be linked to other specific data sources. All of these methods and their variations have advantages and disadvantages. Since any use of the data for population inference requires the need for accurate and efficiently generated demographic information, however, we argue that careful consideration of how the "ground truth" is obtained is important for understanding the quality of derived estimates. In this light, we review existing literature on predicting demographics from social media content, discuss potential errors arising from use of various ground truth methods, and identify gaps in existing techniques. We discuss development of preliminary efforts to reduce the number of assumptions necessary for inference from Twitter. Finally, we conclude by discussing the challenges and limitations of our approach and pose directions for further work to improve quality of inference.

Category: Social Science **Poster ID**: 74

Presenter: Mei Fu, Graduate Student
School of Information, U-M Ann Arbor

Co-Authors: Michael Traugott, Communication Studies, Political Science, U-M Ann Arbor

Title: Comparison between the First Republican Primary Debate and Its News Coverage

Abstract: Presidential debates can play an important role in developing a candidate's positive image, especially during the early part of the campaign cycle. Meanwhile, news media are the major source of information from which most of the citizens learn about and form their impressions of the candidates. The investigation of the first debate and its news coverage can help us better understand how much and well the news media covered the candidates, including which topics and how much speaking time they received in the debate, which will eventually be related to the candidates' name recognition and their image development.

In this study, we: 1) tested the hypothesis that the proportion of the news coverage for each candidate is related to the proportion of their actual speaking time in the debate (H1); 2) studied how news topic coverage corresponds to actual debate content (RQ1); and 3) studied whether the candidates receive more coverage in their local newspaper than in the national news (RQ2). We found that Trump had a strong coverage advantage of 7.8 percentage points on average in the four national media compared with a 3.6 percentage point advantage in the debate speaking time. Some topics in the debate show significant growth within 3 days after the debate. Also, the local newspapers tend to provide more coverage for the senator or governor from that state. Since this was the first in a series of 12 Republican primary debates, it is important to see whether these observations persist in the following ones, as well as whether they appear in the general election debates between Donald Trump and Hillary Clinton. Compared with former studies, we utilized computer programming and text mining techniques such as topic extraction to conduct the study more efficiently and in a quantifiable way. Relative merits of computer analysis compared to the traditional method will also be assessed.

Category: Transportation Research **Poster ID**: 7

Presenter: Benjamin; Joshi, Kunal King, Data Scientist, Engineering Analyst
Ubiquiti, Inc.

Co-Authors:

Title: Vehicle Diagnostics via Machine Learning on Service Data

Abstract: Data collected in complex objective-focused environments can often encode the collective knowledge of human experts. One such dataset is that of vehicle diagnostics and repair, where the actions described in the data reflect the collective experience of many different technicians at different repair locations and times. By applying machine learning to this task, one can use the data for for predictive and prescriptive purposes. We explain and show how the available latent knowledge in vehicle repairs data is used to identify the likely cause and the recommended repairs for a malfunctioning vehicle based on the symptoms it exhibits. Our approach combines a knowledge-based approach with modern machine learning to create an application capable of reporting normalized results to the user for new inputs. The user interface itself is simple, and modeled on the familiar Web search engines. Our techniques have also been shown to effectively learn from new situations and adapt based on new data that is entered into the backend repositories.

Category: Transportation Research **Poster ID**: 11

Presenter: Shan Bao, Associate Research Scientist
University of Michigan Transportation Research Institute, U-M Ann Arbor

Co-Authors:

Title: A Data Driven Method to Investigate Drivers' Adaptation Behavior and Decision Making when Interacting with Automated and Connected Vehicle Technologies

Abstract: The rapid evolution of automated and connected vehicle technologies will reform drivers' behavior patterns, therefore brings new challenges and concerns to the transportation society. One of the main barriers to the technology development in the automated and connected area is related to human factors issues. There are still some critical questions that have not been addressed, such as how drivers interact with those technologies, what are the significant factors associated with drivers' decision making, and the quantification of positive and negative safety consequences. The purpose of this study is to examine and quantify drivers' adaptation behavior when interacting with automated and connected vehicle technologies while driving on the real roads, and further understanding the positive and negative safety consequences of those advanced vehicle technologies. To achieve the purposes of this study, naturalistic driving data from two Field Operational Tests (FOT) that both were conducted by the University of Michigan Transportation Research Institute were used in this study. The two FOT studies are the Advanced Collision Avoidance System (ACAS) FOT and the Safety Pilot Model Deployment (SPMD) study. The ACAS system included both a forward crash warning (FCW) system and an adaptive cruise control (ACC) system while the Safety Pilot data represents the best information available on how active connected vehicle technology impacts on driver behavior in a vehicle-to-vehicle connected environment. Results from the analysis showed that automated vehicle technologies (e.g., ACC) may be able to change the behavior of drivers, especially aggressive drivers, when compared to crash warning systems.

Category: Transportation Research **Poster ID**: 30

Presenter: Ding Zhao, Assistant Research Scientist
Mechanical Engineering, U-M Ann Arbor

Co-Authors:

Title: Statistically Certified Testing Approach for Self-Driving Car based on Naturalistic Driving Data

Abstract: It is essential to develop a fair, complete, and practical test approach for self-driving cars. We developed a statistically certified approach to defining the safety of self-driving based on a large scale of real-world driving data. We then develop the traffic primitives that comprise the fundamental traffic scene elements, and spatially map the segmented and clustered primitives into operable tests. Accelerated evaluation is then formed to synthesize primitives with probabilities, statistically, guarantee the equivalence between the on-track tests and on-road tests while significantly reducing the test duration. By using both approaches, we could reduce the testing duration by 100 to 100,000 times compared to direct on-road tests, thus potentially propose to evaluate the safety of a self-driving designs.

Category: Transportation Research **Poster ID**: 31

Presenter: Tak Choi Yu, Graduate student
College of Engineering & Computer Science, U-M Dearborn

Co-Authors: Yi Lu Murphey, Electrical & Computer Engineering, U-M Dearborn; Chang Liu, Electrical & Computer Engineering, U-M Dearborn; Muhammad Tayyab, Electrical & Computer Engineering, U-M Dearborn; Divyendu Narayan, Electrical & Computer Engineering, U-M Dearborn

Title: Accurate Pedestrian Path Prediction using Neural Networks

Abstract: This paper presents a study on predicting pedestrian path in a short time horizon, e.g. less than 5 seconds. Our study is conducted within the context of pre-collision detection and avoidance between vehicle and pedestrian using only the positioning data transmitted through V2P communications. An important component in a pre-collision detection system is to accurately predict pedestrian positions in a very short future time period. Three methods are presented, a dead reckoning prediction method, a pattern recognition neural network and a time series neural network, both are designed to predict the future position of a pedestrian based on recent movements. An innovative feature extraction method has been developed for generating feature vectors that are invariant to trip location and prediction time, which are important for training a pattern recognition neural network. All three methods are evaluated on trip data recorded from two pedestrians using a Smartphone application software.

Category: Transportation Research **Poster ID**: 32

Presenter: Yongquan Xie, Postdoctoral Researcher
College of Engineering & Computer Science, U-M Dearborn

Co-Authors: Chengqi Bian, Electrical & Computer Engineering, U-M Dearborn; Yi Lu Murphey, Electrical & Computer Engineering, U-M Dearborn; Dev S. Kochhar, Ford Motor Company

Title: An SVM Parameter Learning Algorithm Scalable on Large Data Size for Driver Fatigue Detection

Abstract: Support vector machine (SVM) is a classification model that separates instances by maximizing their distance to a classifying hyper-plane. SVM has been applied successfully to solve a wide range of application problems. However the effectiveness of SMV largely depends on the parameters used by its kernel functions. For a SVM with the radial basis function (RBF) being the kernel function, two parameters control the SVM training. Traditionally, grid-search technique is applied to selecting the proper values of the two parameters. The grid-search method is computationally expensive when the size of training samples is large. In this paper, we present a parameter learning algorithm, Distributed Learning and Searching (DL&S) . It is composed of two stages: distributed searching for significant parameters and finding optimal parameters fit for all training data. We applied the DL&S algorithm to solve an important automotive safety problem, driver fatigue detection. We present a driver fatigue detection system using a SVM trained on driver performance data, lane position, lane heading, and lateral distance. We apply the DL&S algorithm to select optimal parameters and use them to train a SVM for driver fatigue detection. Our experimental results show that the SVM generated by the optimal parameters selected by the DL&S algorithm can perform nearly as well as the SVM generated by the parameter pair found by the gridsearch, and, more importantly, the DL&S algorithm consumed only 7.5% of the computational cost needed by grid-search.

Category: Transportation Research **Poster ID**: 53

Presenter: Gary Marple, Assistant Professor
Mathematics, U-M Ann Arbor

Co-Authors:

Title: Data-Driven Construction of High-Fidelity Mobility Maps

Abstract: Physics-based simulation is increasingly playing an important role in constructing mobility maps and in the design of unmanned ground vehicles. Numerical simulations have received significant attention in the recent past as they mitigate the need for extensive experimentation, which can quickly become expensive, or offer predictive capabilities when experiments cannot be performed. However, physics-based high-fidelity simulations are extremely challenging owing to the multi-scale and multi-physics nature of these multi-body dynamics problems. The terrain typically is modeled as a granular media i.e., as a collection of particles of arbitrary shape; each responding to body forces such as gravity, inertia or drag, as well as repulsive or dissipative forces caused by contact. Often, performing even one grain-to-vehicle level simulation for a given set of parameters takes several days of CPU time. The natural question that arises is, “Can we generate mobility maps of arbitrary terrains without performing complex physics-based simulations?” This is a critical question that we are working to address because military vehicles should be deployable worldwide and must be operationally mobile in all environments including unknown soft soil terrains and combat zones. The basic premise is that we can generate a large amount of offline data (or sample library) using existing physics-based simulations, train a machine learning model on this sample set, use the trained model for future predictions, thereby, significantly reducing the computational expense. Such a trained model could possibly be used for online or real-time dynamics simulations as well.

Category: Transportation Research **Poster ID**: 56

Presenter: Vikas Rajendra, Graduate Student
Computer & Information Science, U-M Dearborn

Co-Authors: Omid Dehzangi, Computer & Information Science, U-M Dearborn

Title: Wearable Galvanic Skin Response for Identification of Distraction During Naturalistic Driving Using Continuous Decomposition Analysis

Abstract: Main reason for fatalities due to road accidents is distracted driving. While driving, certain levels of distraction can cause the driver to lose his/her attention. Thus, the number of accidents can be reduced by early detection of distraction. Several studies have been investigated to automatically detect driver distraction. However, these methods might detect the distraction rather late. Although neurophysiological signals using Electroencephalography (EEG) have shown to be another reliable indicator of distraction, EEG signals are very complex and the technology is intrusive to the drivers. In this study, we investigate a continuous measure of phasic Galvanic Skin Responses (GSR) using a wearable wristband to identify distraction during a naturalistic driving. We first decompose the raw GSR signal into its phasic and tonic components using continuous decomposition. We generated a spectrogram transformation for both non-distracted and distracted states to visualize the behavior of the phasic GSR signal. This helped to extract relevant spectral and temporal features to capture the changes/patterns at the physiological level. We then performed feature selection using support vector machine recursive feature elimination (SVM-RFE) a binary class technique in order to: 1) rank the discriminative features for all subjects, and 2) create a reduced feature subset to explore the highest distraction detection accuracy. We explored two scenarios Normal vs. Phone and Normal vs. Text to evaluate the rank of features to be selected. We employed support vector machine (SVM) to generate the 10-CV identification accuracy. The SVM-RFE selected the set of reduced features that successfully characterized and identified the distracted from non-distracted states with a marginal decrease in accuracy while reducing the computational complexity and the redundancy in the input space toward early notification of distraction state to the driver. Our experimental results demonstrated an accuracy of 94.81% using all features and 93.01% using the rank of SVM-RFE.

Category: Transportation Research **Poster ID**: 63

Presenter: Yongquan Xie, Research Fellow
Electrical & Computer Engineering, U-MDearborn

Co-Authors: Chengqi Bian, Electrical & Computer Engineering, U-M Dearborn; Yi Lu Murphey, Electrical & Computer Engineering, U-M Dearborn; Dev S Kochhar, Ford Motor Company

Title: An SVM Parameter Learning Algorithm Scalable on Large Data Size for Driver Fatigue Detection

Abstract: Support vector machine (SVM) is a classification model that separates instances by maximizing their distance to a classifying hyper-plane. SVM has been applied successfully to solve a wide range of application problems. However the effectiveness of SVM largely depends on the parameters used by its kernel functions. For a SVM with the radial basis function (RBF) being the kernel function, two parameters control the SVM training, c and γ. Traditionally, grid-search technique is applied to selecting the proper values of the two parameters. The grid-search method is computationally expensive when the size of training samples is large. In this paper, we present a parameter learning algorithm, Distributed Learning and Searching (DL&S). It is composed of two stages: distributed searching for significant parameters and finding optimal parameters fit for all training data. We applied the DL&S algorithm to solve an important automotive safety problem, driver fatigue detection. We present a driver fatigue detection system using a SVM trained on driver performance data, lane position, lane heading, and lateral distance. We apply the DL&S algorithm to select optimal parameters and use them to train a SVM for driver fatigue detection. Our experimental results show that the SVM generated by the optimal parameters selected by the DL&S algorithm can perform nearly as well as the SVM generated by the parameter pair found by the grid-search, and, more importantly, the DL&S algorithm consumed only 7.5% of the computational cost needed by grid-search.

Category: Transportation Research **Poster ID**: 68

Presenter: Qi Luo, Graduate Student
Industrial & Operations Engineering, U-M Ann Arbor

Co-Authors: Shukai Li, Peking University; Robert Hampshire, Transportation Research Institute, U-M Ann Arbor; Sharon Di, Civil Engineering, Columbia University

Title: Design of Multimodal Network for Transportation-as-a-Service

Abstract: Transportation-as-a-Service (TaaS) is a mobility solution that integrates multiple modes of transport into seamless trip chains. In a trip chain, the smooth transfer from one mode to another is critical. In particular, we study a multimodal system in which passengers use free-floating bikes for first/last-mile, and transit to one-demand transport (shuttles or taxis) at certain hubs. We formulate the fleet management as a bi-level optimization on a closed queuing network. The objective is to minimize the capital expenditures as well as the operating expenses. On the pricing level, we use a corrected approximation method to calculate the stationary distributions of vehicles in a fixed network, and find the minimum expense and its corresponding prices. On the hub selection level, a greedy algorithm finds the near-optimal transfer hub locations. Under mild assumptions, we are able to find the near-optimal prices and network structures in a case study of New York City.

Category: Transportation Research **Poster ID**: 70

Presenter: Heejin Jeong, PhD Candidate
Industrial & Operations Engineering, U-M Ann Arbor

Co-Authors: Youngchan Jang, Industrial & Operations Engineering, U-M Ann Arbor; Neda Masoud, Civil & Environmental Engineering, U-M Ann Arbor

Title: Car Crash Injury Level Classification

Abstract: This study aims to classify the injury severity in motor-vehicle crashes with three contributing factors, including driver, and environmental and temporal factors. The dataset used for this study is the Michigan Traffic Crash Facts. This dataset contains data on fatal injuries caused by traffic crashes. To deal with the imbalanced classes, several techniques have been used, including under-sampling, over-sampling, and hierarchical clustering. Using eleven classification learning models (i.e., Generalized linear model, Decision tree, Random forest, Support Vector Machine, k-nearest neighbors, Conditional Inference Tree, Neural network, Gaussian process, XGBoost, Gradient boosting model, and Naïve Bayes classifier), we classify the level of injury severity (i.e., fatal/serious vs. non-fatal) and attempt to improve the classification performance by two training-testing methods including Bootstrap aggregation (or bagging) and majority voting among selected models. Furthermore, due to the imbalance present in the dataset, we use the geometric mean (G-mean) of specificity and sensitivity to evaluate the classification performance. We show that the classification performance is the highest (i.e., G-mean = 72.4) when bagging is used with Gradient boosting, without any additional treatment for imbalanced data. The effect of treatments for the imbalanced data is maximized when hierarchical clustering is combined with bagging (i.e., 56.6 % increase in G-mean). The effect of training-testing methods is maximized when majority voting among selected (high-performance) models is used, with no additional treatment for imbalanced data (i.e., 4 % increase of G-mean).

Category: Transportation Research **Poster ID**: 77

Presenter: Matthew Aguirre, Graduate student
Industrial & Operations Engineering, U-M Ann Arbor

Co-Authors: Wenbo Sun, Industrial & Operations Engineering, U-M Ann Arbor; Ryan Hancock, Industrial & Operations Engineering, U-M Ann Arbor; Judy Jin, Industrial & Operations Engineering, U-M Ann Arbor; Shan Bao, Human Factors, University of Michigan Transportation Research

Title: Analysis of Distractive Driving Behaviors Based on Kinematic Driving Data

Abstract: Distracted driving has become an increasing concern in recent years for improving driving safety. This has attracted more research interests in analyzing driving behaviors by using available kinematic driving data such as vehicle speed, yaw rate, and lane offset with the consideration of traffic and road conditions. These signals provide rich information to represent driving behaviors, which helps distinguish normal driving from irregular driving under distractions. Our goal is to develop effective data analytics methods for modeling and detecting the distracted driving behaviors (such as texting while driving) based on kinematic driving data.

A training dataset of 20 texting events from drivers driving in their normal everyday settings has been extracted from the IVBSS database given by the University of Michigan Transportation Research Institute. Three nonparametric approaches are taken as the data is a high dimensional and nonstationary time series. The first is an offline method known as energy distance based change point detection with probabilistically pruned objective developed by James Matteson in 2015, achieving detection rates of about 82% on our training dataset. The second is an online feature-based detection method involving total variance and relative variance at one second time windows. Optimizing thresholds for each of these features based on our training dataset with the constraint of 82% detection rate attained a minimum false alarm rate of around 6%. Finally, an online Kalman filter-based method is explored utilizing the fusion of three different models including a constant velocity model, a constant acceleration model, and a constant turn rate and acceleration model. Various features from this Kalman filter approach also show potential to detect texting while driving.