Keynotes

Yulia R. Gel, Professor, University of Texas Dallas
Dissecting Blockchain Analytics: What We Can Learn from Topology and Geometry of Blockchain Transaction Graphs

Blockchain technology and, in particular, blockchain-based cryptocurrencies offer us information that has never been seen before in the financial world. In contrast to fiat currencies, all transactions of crypto-currencies and crypto-tokens are permanently recorded on distributed ledgers and are publicly available. As a result, this allows us to construct a transaction graph and to assess not only its organization but to glean relationships between transaction graph properties, crypto price dynamics as well as illegal and illicit activities such as emerging ransomware.
In this talk we discuss horizons and limitations of what new can be learned from topology and geometry of cryptocurrency transaction graphs whose even global network properties remain scarcely explored. By introducing novel tools based on topological data analysis, functional data depth, network motifs, and geometric deep learning, we show that even some subtler patterns in blockchain transaction graphs can provide critical insights for money laundering tracking, price analytics, and market sentiment assessment.

Justin Smith, Sr. Director of Advanced Analytics, St. Luke's Health System
Data Science before, during, (and after!) a global pandemic

This presentation will provide an overview of the rapidly changing field of Data Science from a healthcare perspective. **Background history**: How Data Science is traditionally leveraged to enhance business decision-making and drive patient care. **Current use**: The emergence of COVID19 as a global pandemic upended norms yet Data Science plays an important role in the successful management resources through machine learning and A.I. **Prospects**: What does the future (post-pandemic) hold for Data Science and where is the field going in the near-, mid-, and long-term future.

Invited abstracts

Anu Amallraja – Lead Clinical Bioinformatics Analyst, Avera Cancer Institute
Understanding the multi-scale heterogeneity in cancer genomes
Cancer is an incredibly complex and heterogeneous disease that arises from a person's own cells. It is responsive to environmental changes and to treatment in particular, which makes it hard to treat effectively. Not only are there over 800 sub-types of cancer, individual cells of a patient's tumor can also vary between each other.
Advances in sequencing and computing technologies have allowed us to generate vast amounts of data from tumor and normal tissues that can help us understand not only the development, growth and spread of cancer in general, but also study the specifics of a particular patient's cancer, down to the granularity of a single cell. Bioinformatics tools help us translate sequencing data into biologically meaningful information.
Identifying and understanding the complexity of a patient's cancer using these methods is a very useful tool in the clinic to make therapeutic choices appropriate for that patient. This talk will introduce the various

levels of heterogeneity in cancer genomes, demonstrate examples from a study focused on gastric adenocarcinomas, and discuss the effects this can have for precision medicine approaches.

Travis Burge Scientist/Engineer – Raven Industries
Kelly Vanderwerff- Raven Industries
A vision guidance system on agricultural sprayers reduces operator stress
Manually operating an agricultural sprayer is a stressful activity. Operators work an average of 15 hrs/day in peak season [1], navigating 38-cm (15-in) tires in 76-cm (30-in) rows. Guidance systems can relieve operator fatigue [2]. A commercially-available vision guidance system (VSN®, Raven Industries) is available for agricultural vehicles.
Human subject protocols were approved by the IRB. Three experienced male operators drove manually and with VSN in the same field. Each wore an Empatica E4 wristband to measure their electrodermal activity (EDA). Sprayer steering status was recorded from the sprayer guidance system (RS1TM, Raven Industries). EDA data was filtered with a Hampel filter to remove artifacts [3] (1 s window before and after) and then with an infinite impulse response (MATLAB filtfilt function) [4] with a window of 1 s. Filtered EDA peaks and valleys were calculated (MATLAB findpeaks function) [5] and stressful events were defined as those which exceeded a magnitude threshold of 0.01 μS [6]. The frequency of stressful events, stressful event characteristics (e.g., magnitude, duration and area under the curve) and frequency of steering adjustments were calculated while driving in straight rows (length > 150 m). An ANOVA was performed on each calculated metric with steering type and participant as predictor variables and $p < 0.05$ considered significant. Thirty-four passes in four fields were analyzed (16 manual, 18 VSN). Operators steering with VSN had 49% fewer stressful events per time compared to manual driving (3.6 versus 7.1 events/min, $p < 0.001$, Figure 1). These results suggest that steering with VSN considerably reduces the stress on agricultural operators compared to steering manually.

Jiyul Chang – Senior Lecturer, South Dakota State University
Datasets for Precision Agriculture Practices
Precision agriculture practices as a data-based agriculture require many types of data. In data side of view, precision agriculture practices can be explained in three steps which are 1) data acquisition, 2) data processing, and 3) implementation of final decisions. Producers and agronomists collect data from fields and apply the final decisions to farming. Data scientists need to analyze the big data to make a best decision for maximum benefits. However many different types and formats of data cause difficulties of data analysis. This presentation will talk about types and formats of data that are actually using in precision agriculture practices.

Yuzhou Chen – Postdoctoral Scholar – Princeton University
Topological data analysis of dynamic Ethereum token networks
Forecasting price in the dynamic Ethereum token networks data is indispensable for understanding the blockchain dynamics and measuring the risk connectedness among the cross-cryptocurrency trades. In the last few years, Geometric Deep Learning (GDL), e.g., Graph Convolutional Networks (GCNs), have emerged as a powerful alternative to more conventional time series predictive models. Despite their proven success, GCNs tend to be limited in their ability to simultaneously infer latent temporal relations among entities (i.e., traders and sellers in the underlying blockchain transaction network). In our work, we make the first step on a path of bridging the two emerging directions, namely, time-aware GDL with time-conditioned topological representations of complex dynamic Ethereum token networks. To summarize such time-conditioned topological properties, we develop novel topological representations. We then propose topology-based GDL models which allows us to simultaneously learn co-evolving intra- and inter-

dependencies (i.e., spatial and temporal correlations among nodes) in the dynamic Ethereum token networks data.

In this talk, I will present our topology-based GDL models "TAMP2-S2GCNETs" and "Z-GCNETs" --- the first effort to integrate the extracted time-conditioned topological descriptors (i.e., multipersistence and zigzag persistence image) into GDL that can enhance GDL architectures with the most salient time-conditioned topological information from data. The topology-based GDL model with time-aware topological layer bridges time-conditioned GDL with time-aware persistent homology representations of the data in learning complex multivariate spatio-temporal processes. I will discuss our work to predict Ethereum blockchain price and also show substantial computational gains and high utility of the proposed time-conditioned topological descriptors for encoding the time-conditioned knowledge.

Lance Cundy – Economist, Federal Reserve Bank
Forecasting During Uncertain Times
COVID has generated uncertainty with economic and financial models and their data. With institutions and individuals changing behavior, there is risk that underlying data generating processes have changed. As a result, (i) models built prior to COVID may no longer be appropriate to make reasonable forecasts, and (ii) data from the COVID period may not be suitable to forecast future behavior. In this session, we will discuss items to consider when making forecasts during these uncertain times. Through various examples, we will explore how to assess model performance and examine solutions to overcome the challenges created by uncertainty.

Gerald Fahner - Senior Principal Scientist, FICO
Advances in Explanation-friendly Scorecard Technology
Traditional scorecards are generalized additive models with step functions over predictors as component functions. While these models are flexible to capture nonlinear effects and easy to explain (explanation can be enhanced by imposing shape constraints on the component functions), the discontinuous nature of step functions can be unnatural for some continuous numeric predictors and hence harder to explain. This applies to use cases in business analytics such as credit scoring and perhaps even more so to certain health or behavioral sciences applications.
We introduce smoother scorecards which generalize the step functions by higher-order B-splines to get around discontinuities. The models can be trained to maximize the Divergence objective (a measure of separation between the score distributions of Positives and Negatives) subject to score engineering constraints such as global or piecewise monotonicity. We also discuss how predictive performance and explanation can be improved by adding a roughness penalty term to the objective in order to smooth wiggly splines. The presentation will explain concepts and present case study results from credit risk scoring and/or healthcare to back up our claims and to illustrate the power and elegance of an innovative generalization to traditional scorecards.

Craig Jorgensen – Customer Success Manager, Query AI
Decentralized Data Access and Analysis for Cybersecurity Usecases
Siloed data prevents organizations from gaining timely views into security risks. The Query.AI platform provides real-time, centralized insights from decentralized data so you can accelerate cybersecurity investigations and efficiently respond to threats.

Edward Krueger - Channel Partners
Landon Thompson and Josh Moore, Channel Partners

Equipment Finance Credit Risk Modeling - A Case Study in Creative Model Development & Nimble Data Engineering

This presentation will focus first on providing an overview of Channel and the Risk Analytics team that performed this case study. Given that context, we'll then dive into our approach for building the modeling development data set, techniques and tools used to develop and implement the model into a production environment, and some of the challenges faced upon launch. Then, the presentation will pivot to the data engineering pipeline. During this portion, we will explore the application process and what happens to the data we collect. This will include how we extract & store the data along with how it is integrated into our other systems for decision purposes. We will also talk about how the data is transformed from a raw, sometimes unstructured state, to something more usable by a data science team – and demonstrate how this data was harnessed to help guide model enhancements as key opportunity areas have been identified.

Michael Lim – Vice president of Data Science and Analytics, TransUnion

The Role of Data Science in Fraud Detection

The impact of fraud losses on overall profitability has been growing for many financial institutions in recent years. As technology continues to evolve and digital transactions increase, this new environment creates opportunities for fraudsters to deploy new tactics to commit fraud. The global pandemic certainly accelerated movement to digital transactions, including an increase in consumers interacting with financial institutions in a contactless manner. In this session, we'll review fraud trends during the pandemic and where they are headed as the economy recovers and consumers begin engaging in the consumer credit market again. These will be valuable insights to understand how to best defend against emerging fraud tactics which will translate to better portfolio growth and profitability.

Himel Mallick – Associate Principal Scientist, Biostatistics, Merck Research Laboratories

Tweedie mixed models for spatial transcriptomics and digital pathology

A key analytic task in spatial transcriptomics studies is to identify genes that display spatial expression (SE) patterns, commonly referred to as SE genes, similar in spirit to popular differential expression (DE) analysis. Here we propose spatial Tweedie mixed models based on a self-adaptive Tweedie distribution that flexibly captures a large dynamic range of observed spatial transcriptomics expression profiles. To properly account for the unique characteristics of spatially resolved transcriptomics data, we propose a two-part joint model: (i) a Gaussian process (GP) regression for modeling spatial changes in gene expression, and (ii) an over-dispersed Tweedie model for modeling expression values that captures over-dispersion through a nugget (white noise) term in the underlying GP covariance function. Empirical evidence of the attractiveness of the method is demonstrated via extensive simulation studies and real data analysis. We also apply our method to structurally similar digital pathology data thus providing a unified modeling framework for spatial molecular profiles.

Volodymyr Melnykov – Professor, The University of Alabama

Shuchismita Sarkar - Bowling Green State University and Yana Melnykov - The University of Alabama

Model-based clustering of directed weighted networks

An approach relying on the notion of mixture models is proposed for modeling and clustering directed weighted networks. The proposed methodology can be used in many settings including modeling multilayer networks. Computational issues associated with the developed procedure are addressed by the use of a MCMC procedure. The utility of the methodology is illustrated on synthetic data as well as real-life data containing trade operations among European Union members.

Yana Melnykov – Assistant Professor, The University of Alabama
Volodymyr Melnykov, and Xuwen Zhu – The University of Alabama
Studying contributions of variables to classification
A problem of finding variables responsible for classifying a particular observation as well as detecting those that interfered with the assignment made is considered. We address this problem by providing a formal argument supported by intuitive geometric interpretation.

Danica Ommen – Assistant professor, Iowa State University
Madeline Q. Johnson, Boston Scientific
Forensic Handwriting Identification using Random Forests and Score-based Likelihood Ratios
Handwriting analysis is conducted by forensic document examiners who can visually recognize characteristics of writing to assess the writership propositions. Recently, there have been incentives to investigate how to quantify the similarity between two written documents to support the conclusions drawn by experts. To this end, we use an automatic algorithm within the open-source 'handwriter' package in R to decompose a handwritten sample into small graphical units of writing. These graphs are sorted into exemplar groups or clusters. We assume that the frequency with which a writer produces graphs to each cluster is characteristic of their handwriting. Then, given handwritten document pairs, we can use the difference in their vectors of cluster frequencies as the input for a random forest. The output from the random forest is used as the similarity score. We estimate the densities of the similarity scores computed from multiple pairs of documents where the source attribution is known and use them to obtain score-based likelihood ratios (SLRs). We find that several different types of SLRs can successfully indicate the strength of evidence for writership determinations.

Kayli Rageth – Principal Clinical Intelligence Analyst – Avera
Development and Evaluation of a Rural SARS-Cov-2 Hospital Admissions Predictive Model
Ongoing evolution of the SARS-Cov-2 virus has propelled the world into a state of the unknown. The impact upon healthcare has been immense. Case numbers and hospitalizations have grown at rapid rates as new variants exhibit higher transmissibility. Efforts to gain foresight into the evolving conditions have been underway for organizational and planning purposes alike. Recognizing that there are repercussions from both, underestimates and overestimates of predicted hospitalizations, we have developed a model that leverages sophisticated data science techniques which employ foundational epidemiologic methods. This has allowed us to account for this highly volatile and dynamic landscape and forecast hospital admissions with reasonable accuracy throughout the course of this pandemic.

Ali Rahnavard, Assistant Professor of Biostatics and Bioinformatics, George Washington University
Decoding infectious disease omics data: COVID-19 case study
Infectious diseases are a great challenge to the world due to the high rate of transmissibility and dynamics of viral populations over time and space. The enduring success of treatment strategies (e.g., vaccine, drug) requires characterizing viral evolution during the initial and subsequent treatments and discrete host/viral genomic variants associated with different clinical outcomes and viral phenotypes (e.g., drug resistance and vaccine escape mutants). We developed and applied novel computational approaches to integrate diverse data types (viral genomics, host genomics, clinical data, treatment data, health outcome data), identify patterns of associations among these data types, and thereby identify targets of biomarkers associated with treatments and disease outcomes and how these change over time and with interactions of different viral and host populations. In an investigation of COVID-19 as an exemplar system, we found SARS-COV-2 mutations occur at a very interesting rate; for example, nonstructural protein 3 (nsp3) variation across our data population co-occurred with Spike protein, a target for most COVID-19 vaccines. We also investigated

human body responses to the infection using metabolomics and proteomics profiling. We discovered important pathways that explained organ dysfunctions, such as lung inflammation. We identified biomarkers such as citrulline and Hyaluronan-binding protein 2 indicative of multi-system tissue dysfunction and can be used for COVID-19 diagnosis. The methodology developed is generally applicable across infectious disease outbreaks and systems in humans, agriculture, and nature as more and more omics data become available in such systems.

Bonny Specker – Professor and Chair Emerita, SDSU
Data and Epidemiology: The Need to Educate the Public
Unlike previous pandemics, the COVID pandemic came at a time of political polarization and extensive use of social media. Early in the outbreak, it soon became clear that the general lack of understanding of simple epidemiologic principles led to confusion and mistrust. Using data posted to the South Dakota Department of Health dashboard, summaries of daily data were provided via email and blog to individuals from university emergency management staff to city council members, retirees, and the general public. Weekly videos were produced by the city and posted to social media accounts. The goal of these data updates and videos was to provide graphic summaries of the current situation in Brookings County and the state of South Dakota and educate the public on the general principles of epidemiology.

Kruttika Sutrave and Rajesh Godasu
Towards Long Term Impact of DL Models in Medical Imaging
The current literature suggests expansion in the research area that combines Generative Adversarial Networks (GANs) and Transfer learning (TL). Generalizability and Scalability are two important attributes to evaluate DL models to assess their long-term impact. In this research we analyze if linear combination (Data augmented TL) of these two techniques is more generalizable and scalable, or TL enabled GANs approach has better long-term impact. First, we implement the Data augment TL approach by employing DCGAN to generate synthetic chest Xray images and to pre-train a VGG-16 model. Next, we implement the TL enabled GAN method by initially training WGAN using chest and abdomen images. We then retrain the WGAN using colon Xray images to classify between normal (benign) and cancer (malignant) polyps

Priya Swaminathan - Clinical Bioinformatics Analyst, Avera Cancer Institute
Patient-specific analysis and visualization of cancer pathways
Introduction Analysis of transcriptomics data enables us to quantify gene expression and determine activity of cancer pathways. Evaluation of gene sets and pathways, and visualizing pathway activity for each patient is important to make inferences about treatment outcomes, overall survival, and potential therapeutic targets. However, pathway analysis of cancer patient samples without their respective normal remains challenging. **Methods and Results** Our method displays patient-specific pathway activation relative to reference populations from The Cancer Genome Atlas (TCGA) and The Genotype-Tissue Expression (GTEX). Patient-specific pathway activity scores are shown within the respective percentiles and averages of the TCGA and GTEX cohorts to visualize and compare a patient's individual pathway activation profile within the TCGA tumor and GTEX normal reference population.
Patient-specific pathway scores are obtained from gene expression data of each patient using PROGENy (Pathway RespOnsive GENes for activity inference) and transformed into activity scores for each of 14 cancer pathways. The pathway activity scores of each new patient tumor sample are scaled using the parameters of scaled pathway activity scores obtained from TCGA and GTEX gene expression data of all cancer tissue types. Lastly we visualized the pathway activity scores of each patient sample. **Conclusions**

Our visualization method is intended to look at the different activation patterns of the 14 cancer pathways in our patient samples. Pathway genes of interest can then be evaluated in the patient samples to identify causal mutations and their associations with the pathway activity scores hence aid in treatment selection and development of precision medicine therapeutics.

Susan Vanderplas – Assistant professor, University of Nebraska – Lincoln
How Do You Define a Circle? Perception and Computer Vision Diagnostics
Neural Networks are very complicated and very useful models for image recognition, but they are generally used to recognize very complex and multifaceted stimuli, like cars or people. When neural networks are used to recognize simpler objects with overlapping feature sets, things can go a bit haywire. In this talk, we'll discuss a model built for applications in statistical forensics which uncovered some very interesting problems between model-based perception and human perception. I will show visual diagnostics which provide insight into the model, and talk about ways we might address the discrepancy between human perception and model perception to produce more accurate and useful model predictions.

Yang Wang – Assistant Professor, College of Charleston
Conditional mixture modeling and model-based clustering
Due to a potentially high number of parameters, finite mixture models are often at the risk of overparameterization even for a moderate number of components. This can lead to overfitting individual components and result in mixture order underestimation. One of the most popular approaches to address this issue is to reduce the number of parameters by considering parsimonious models. The vast majority of techniques in this direction focus on the reparameterization of covariance matrices associated with mixture components. We propose an alternative approach that emphasizes modeling cluster locations. The developed procedure enjoys remarkable modeling flexibility, especially noticeable in the presence of non-compact clusters. Due to an attractive closed form formulation, speedy parameter estimation is available by means of the EM algorithm. The utility of the proposed method is illustrated on synthetic and well-known classification data sets.

Sommer West, Data Scientist, Capital Services
Modeling Consumer Risk: A Comparison of Logistic Regression, Scorecard, and Machine Learning Models
Predicting the probability a consumer will not repay their loan is a complicated, but important challenge. In the credit card industry, it is especially important because there is no collateral. Various models can be used to predict this risk of a consumer not paying, but the decision as to which model to use can be a big challenge itself. In general, the types of models range from simple models, like logistic regression, to complex models, like machine learning. There are pros and cons in using either a simple or complex type of model. Typically, complex models perform better in terms of accuracy, but complex models can be hard to explain and even harder to implement. Vigorous research should be done before deciding which type of model to implement, keeping in mind the benefits of using a simpler model with fewer complexities.

Bing Xu- Biostatistician, PhD Avera Cancer Institute
16 Year Life History and Genomic Evolution of an ER+ HER2- Breast Cancer
Metastatic breast cancer is one of the leading causes of cancer related death in women. Limited studies have been done on the genomic evolution between primary and metastatic breast cancer. We reconstructed the genomic evolution through the sixteen year history of an ER+ HER2- breast cancer patient to investigate molecular mechanisms of disease relapse and treatment resistance after long term exposure to hormonal therapy. Genomic and transcriptome profiling was performed on primary breast tumor (2002), initial

recurrence (2012) and liver metastasis (2015) samples. Cell free DNA analysis was performed at eleven timepoints (2015-2017). Mutational analysis revealed a low mutational burden in the primary tumor which doubled at the time of progression, with driver mutations in PI3K-Akt and RAS-RAF signaling pathways. Phylogenetic analysis showed an early branching off between primary tumor and metastasis. Liquid biopsies, while initially negative, started to detect an ESR1 E380Q mutation in 2016 with increasing allele frequency till end of 2017. Transcriptome analysis revealed 721 (193 up, 528 down) genes to be differentially expressed between primary tumor and first relapse. Most significantly down-regulated genes were TFF1 and PGR, indicating resistance to AI therapy. Most up-regulated genes included PTHLH, S100P, and SOX2 promoting tumor growth and metastasis. This phylogenetic reconstruction of the life history of a single patient's cancer as well as monitoring tumor progression through liquid biopsies allowed for uncovering the molecular mechanisms leading to initial relapse, metastatic spread and treatment resistance.

Yingying Zhang –Assistant Professor, University of South Alabama

Semi-supervised clustering of time-dependent categorical sequences with application to discovering education-based life patterns

A new approach to the analysis of heterogeneous categorical sequences is proposed. The first-order Markov model is employed in a finite mixture setting with initial state and transition probabilities being expressed as functions of time. The expectation-maximization algorithm approach to parameter estimation is implemented in the presence of positive equivalence constraints that determine which observations must be placed in the same class in the solution. The proposed model is applied to a dataset from the British Household Panel Survey to evaluate the association between the education background and life outcomes of study participants. The analysis of the survey data reveals many interesting relationships between the level of education and major life events

Accepted oral presentations

Madeline Q. Johnson, Boston Scientific
Danica M. Ommen, Iowa State University/CSAFE

Forensic Handwriting Identification using Random Forests and Score-based Likelihood Ratios

Handwriting analysis is conducted by forensic document examiners who can visually recognize characteristics of writing to assess the writership propositions. Recently, there have been incentives to investigate how to quantify the similarity between two written documents to support the conclusions drawn by experts. To this end, we use an automatic algorithm within the open-source 'handwriter' package in R to decompose a handwritten sample into small graphical units of writing. These graphs are sorted into exemplar groups or clusters. We assume that the frequency with which a writer produces graphs to each cluster is characteristic of their handwriting. Then, given handwritten document pairs, we can use the difference in their vectors of cluster frequencies as the input for a random forest. The output from the random forest is used as the similarity score. We estimate the densities of the similarity scores computed from multiple pairs of documents where the source attribution is known and use them to obtain score-based likelihood ratios (SLRs). We find that several different types of SLRs can successfully indicate the strength of evidence for writership determinations.

Qingqing Li – University of South Dakota
Yuhlong Lio

On statistical estimates of the inverted Kumaraswamy Distribution under adaptive type-I progressive hybrid censoring

The probability distribution modeling is investigated via maximum likelihood estimation method based on adaptive type-I progressively hybrid censored samples from the inverted Kumaraswamy distribution. The point estimates of model parameters, reliability, hazard rate and quantile are obtained and confidence intervals are also developed by using asymptotic distribution as well as bootstrap method. Monte Carlo simulation has been performed to evaluate the accuracy of estimations. Finally, a real data set is given for the application illustration.

Kari Sandouka, Dordt College

Linguistic Cues in Disaster Relief Crowdfunding

Disaster crowdfunding is an emerging phenomenon that has not received attention in academic research. It provides a unique context that allows us to explore the use of prosocial motivators in crowdfunding behavior. The stories for disaster relief crowdfunding are examined through the lens of a written donation request with the Linguistic and Inquiry Word Count (LIWC) software. Within the framework, factors from theories of prosocial motivation are used to understand how linguistic cues persuade potential donors. The results indicate the importance of clear and concise language through the use of concrete language, a positive mood, and an indication of the recipient's wellbeing. The findings indicate the importance of the crowdfunding story to the fundraising effort, providing both theoretical and practical implications.

Jennifer Schon - Northwestern College - Orange City;
Hailey Louw and Misael Bruzzone- Northwestern College, Iowa

Impacts on Student-Athletes' GPA at Northwestern College

This project studied the impact of distance from home and other variables on GPA among students at a small, private college in the Midwest. Using data from the 2013 - 2016 cohorts, we show that distance is the largest deterrent to freshman GPA among student athletes even when compared to financial need and gender. Among student athletes, those who are from a long distance are likely to have a GPA 0.17 points

lower than their counterparts who come from a short distance. We show that athletic participation alone is not a significant determinant of freshman GPA when previous academic performance and financial need are taken into consideration and held constant. Based on our findings, we suggest institutions composed of majority student athletes can expect those from a long distance to underperform compared to their short-distance counterparts in freshman GPA and therefore should be a focus when it comes to academic support more so than athletes in general.

Jennifer Schon - Northwestern College - Orange City;
Theo Jongerius, Northwestern College, Iowa

Factors of Significance for Graduating at a Private College

This research project examined factors that influence odds of graduating at a small, rigorous, private college in the Midwest. De-identified data from the college's database for the 2013 to 2016 cohorts was collected by the Institutional Research office and provided to the researcher. Several statistical and machine learning techniques were utilized to develop predictive models, including logistic regression and neural network analyses. Fields included include number of major changes, time spent relaxing, satisfaction, and exam preparation techniques. Control variables included distance from home, high school gpa, and religious affiliation. The methods of analysis will be compared and contrasted and results will be discussed.

Accepted poster abstracts

A comparative analysis of topic modeling Techniques for short text data
Loknath Sai Ambati – Dakota State University
Omar El-Gayar
Massive amount of short texts such as tweets, reviews, and social media posts are available on the internet nowadays. It is important for a wide variety of applications to be able to analyze short texts for content analysis and for insights from the textual data. However, limited number of words in such short text can be challenging for meaningful content analysis. This study aims to investigate topic modeling techniques for short text data by performing a comparative analysis of various topic modeling techniques for efficient topic extraction. From a theoretical perspective, the research will shed light into the strengths and weaknesses of various topic mining techniques that can provide insights into future research aimed at improving these techniques for short text in various application domains. From a practical perspective, the research provides guidance into the applicability of topic modeling to short text data.

Generative Adversarial Networks in Tumor-Reglated Research: A Review and Agenda for Moving Forward
Andrew James Behrens, Dakota State University
Cherie Noteboom, Dakota State University
Recent advances in Generative Adversarial Networks (GANs) have led to many new variants and uses of GANs. The latest advancements have allowed researchers and practitioners to apply this technique to tumor-related problems with limited data. One of the trends in this problem domain is to develop different variants of GANs suited explicitly to particular problems. The variants of GANs are numerous but share a common characteristic of expanding the dataset by creating synthetic data from the original dataset. This paper aims to develop a research agenda through a systematic literature review that investigates practitioners' and researchers' emerging issues and current works on the topic. Emerging implementation trends and limitations of GANs in tumor-related problems are explored.

An alpha-based prescreening methodology for a common but unknown source likelihood ratio with different subpopulation structures
Dylan Borchert, South Dakota State University
Semhar Michael, Christopher Saunders, and Andrew Simpson
Prescreening is a commonly used methodology in which the forensic examiner includes sources from the background population that meet a certain degree of similarity to the given piece of evidence. The goal of prescreening is to find the sources closest to the given piece of evidence in an alternative source population for further analysis. This paper discusses the behavior of an $\alpha-$based prescreening methodology in the form of a Hotelling T^2 test on the background population for a common but unknown source likelihood ratio. An extensive simulation study with synthetic and real data were conducted. We find that prescreening helps give an accurate estimate of the likelihood ratio when there is a subpopulation structure in the alternative source population.

Development of strategies for estimating a response surface to characterize a black-box algorithm in terms of a white-box algorithm
Cami Fuglsby, South Dakota State University
Christopher Saunders, South Dakota State University; Danica Ommen, Iowa State University; JoAnn Buscaglia, Federal Bureau of Investigation Laboratory; and Michael Caligiuri, University of California, San Diego

In forensic identification of source problems, there is an increasing lack of explainability of the complex black-box algorithms for the assignment of evidential value. Generally speaking, black-box algorithms are designed with prediction in mind. Although the information fed into the algorithm and the features used to make the prediction are often known to the user, the complexity of the algorithm limits the ability of the end user to understand how the input features are used. On the other hand, more transparent algorithms (sometimes referred to as "white-box") are typically less accurate even if they provide direct information on how the input object is directly used for predicting a class or outcome. In this work, we begin the development on a response surface that characterizes the output of a black-box algorithm with the output of a white-box algorithm. Using a set of handwriting samples, we use a complex black-box algorithm across multiple features to produce a set of pairwise scores and a simple, transparent algorithm that uses individual features to produce another set of pairwise scores. A generalized least squares method is used to test the null hypothesis that there is no relationship between the two types of scores. The outcome of the significance tests helps to determine which of the individual feature scores have an influence on the black-box scores.

Deploying Live Dashboard Data using USDA Data APIs to inform farmers/producers
Indira Fuyal, Dakota State University
Paurakh Paudel and David Zeng, Dakota State University
The main objective of this project is to provide the farmers/producers and small business owners with a data intensive information hub to better understand the market trends and patterns in an interactive way to help make informed decisions. This technology is part of a larger project that includes AI powered data services to provide users with decision making support for real time data solutions. Together with Natural Language Generation, there is a huge potential to deploy this technology. In this small demo of this hugely beneficial technology, we make use of the publicly available Open Data Services – ESR Data APIs to extract data in json format. The data in question is from USDA's Export Sales Reporting program which provides a constant stream of up-to-date market information for 40 U.S. agricultural commodities sold abroad. The data is used to analyze the overall level of export demand, determine where markets exist, and assess the relative position of U.S. commodities in foreign markets. Although we are only focused on this data source for this poster presentation, it is only a representation of the potential of such a technology and will be scaled and combined with NLG to be deployed in the service of the public as a part of the aforementioned larger project. To retrieve the data, a Python script is written to make the API call to link the data to Tableau. The data is then used to develop interactive and dynamic visualizations in Tableau. While heaps of data piled up in row after row in multiple tables can be cumbersome and tedious to go through and find meaningful conclusions from, the same data in the form of representative visualization helps paint a vivid picture of the trends and patterns forming within the data. Even an untrained eye can spot a peak or a dip in a line chart or distinguish the market share percentage by simply looking at the size of a slice of a pie in a pie chart. Thus, the aim of this project is to help our target beneficiaries, i.e., producers, farmers, and small businesses to ultimately make data driven decisions delivered to them in an uncomplicated manner through simple visualizations while the technology runs the hard yards behind the scenes. After these simplifications of raw data to relevant visualizations, they are assembled to create dashboards ready for deployment. This can be done by sharing the dashboards to Tableau Public and further publishing them on a hosted web server for the users to access it easily.

Detection of prostate cancer using machine learning techniques: An exploratory study
Laxmi Manasa Gorugantu, Dakota State University
Omar El-Gayar and Nevine Nawar, Dakota State University
Prostate cancer (PCa) is one the most frequent and fatal cancers in men. It can be slow-growing and indolent or fast-growing and aggressive. Testing for PCa remains problematic. Evidence is mounting that

overdiagnosis and over-treatment can result in adverse side-effects yet have little impact in preventing death from PCa. Consequently, the importance of predictive tools that help physicians in the diagnosis of the condition cannot be understated. There are several prediction models using PSA and other risk factors for detecting clinically significant PCa. However, these models tend to predominantly rely on multivariate logistic regression and tend to be limited in the number of risk factors accounted for in the model.
Accordingly, the objective of the research is to investigate the potential of various machine learning techniques using an expanded set of risk factors to improve the sensitivity and specificity of detecting clinically significant PCa. Compared to logistic regression, the machine learning techniques considered could account for the complexity in predicting PCa. Examples of machine such techniques include support vector machines (SVM), decision trees, Bayesian classifiers, and random forest. Risk factors considered include prostate-specific antigen (PSA), digital rectal examination (DRE), as well as age, race/ethnicity, and family history. The proposed model will be evaluated for specificity and sensitivity against state-of-the-art models. By capturing more complex relations between risk factors and incidence of PCa, the resultant predictive model may have the potential for reducing the downstream harms of PSA testing.

Atypicality Based Measures for the Identification of Counterfeit Aspirin
Janean Hanka, South Dakota State University
Megan Guetzloff, Christopher Saunders, Cami Fuglsby, and Brian Logue, South Dakota State University
In this work we are focused on building a pattern recognition system for chemometric data arising from the LC-MS.MS spectrometry analysis of brand name aspirin. Originally, the goal of the work was to build a set of discriminate functions that can separate between known brands of manufactured aspirin, similar to LDA, which finds a set of projections that optimize linear separation relative to within class variation [1]. However, these functions must assign each observation to a known class, even if the likelihood of an observation having arisen from any given class is very small. As an alternative, we investigate the use of atypicality measures as a way around this issue. The atypicality of an observation with respect to a given population (or class) is the chance of drawing a new sample from a given class that has a greater likelihood of being observed than the actual observation that we are considering assigning to said class. A pattern recognition system based off atypicality measures would assign an observation to the class with the smallest atypicality given that it is not above some threshold. This threshold can be considered a method for determining if an observation is likely to not belong to any known classes at all. We will perform a simulation study comparing the effectiveness of atypicality based methods to LDA and QDA methods when the assumptions of the discriminate functions are satisfied, and then apply the three methods to a chemometric data set related to the analysis of aspirin pills.

Multivariate Statistics Applied to Additive Manufacturing of Inconel
Jason Hasse, South Dakota State University
Semhar Michael, Anamika Prasad, South Dakota State University
Additive manufacturing (AM) continues to be a developing area of research which allows for the creation of complex components with minimal material waste. Multivariate statistical methods can be employed to analyze components created with varying input parameters. Further, posthoc analysis can assist researchers with getting higher performance of future components.

Predicting US wildfire possibility and severity
Hoang Long Nguyen, Minot State University
Louise M. DuPont, Minot State University

Over the last few centuries forest landscapes across America have suffered from dramatic climate change. As result, many changes in landscape conditions have resulted in an increase in wildfires across the United States. Unquestionably many of the fires were caused by humans. Using data provided by the U.S. National Interagency Fire Center, we investigate the possibility of predicting future wildfire events based on fires from within the past 20 years. Specifically, Suppression cost which is included DOI Agencies and Forest Services varies from $239,943,000 to $2,274,000,000 from 1985 to 2020 respectively [1]. Data include the size of fire, topography, terrain, and cause would be used to help predict an effective evacuation and mitigation strategy. This could become an essential tool when planning evacuation strategies and wildfire prevention. This research will focus on data visualization of all mattered factors that cause U.S wildfire. Additionally, logistic regression machine learning algorithms will also be applied using related data in order to predict the possibility and severity of wildfire. References: [1] U.S National Interagency Fire Center. (2020). Suppression Cost. Federal Firefighting Costs (Suppression only). Retrieved from https://www.nifc.gov/fire-information/statistics/suppression-costs

Wald Type Tests with the Wrong Dispersion Matrix
Kosman Rajapaksha
A Wald type test with the wrong dispersion matrix is used when the dispersion matrix is not a consistent estimator of the asymptotic covariance matrix of the test statistic. One class of such tests occurs when there are p groups and it is assumed that the population covariance matrices from the p groups are equal, but the common covariance matrix assumption does not hold. The pooled t test, one-way ANOVA F test, and one-way MANOVA F test are examples of this class. Another class of such tests is used for weighted least squares. Two bootstrap confidence regions are modified to obtain large sample Wald-type tests with the wrong dispersion matrix.

Antecedents of Success for Proficiency in Core Skills
Kari Sandouka, Dort University
Miranda Vander Berg
The test-optional movement in the United States generates the need for creative solutions in advising general education courses (e.g., writing, reading, mathematics, etc.). The authors used pattern analysis and statistical modeling to identify antecedents of success for the Educational Testing Service Proficiency Profile assessment. The core skills areas for the ETS Proficiency Profile are reading, writing, mathematics, and critical thinking. The focus of the research includes pattern analysis of general education courses taken by students that increased their scores between the first year and third year assessments. Other factors such as, but not limited to, gender, distance from home, and discipline of study are included. The practical implication of the results indicates a sequence of courses that provide the most growth in students' proficiency of core skills.

Generalized Estimating Equations (GEE) Approach for Clustered Binary Data with Application to COVID-19 Treatment
Sadixa Sanjel
Clustered binary data frequently occur in epidemiology and other applied fields such as clinical trial studies, where observations within the respective samples are correlated. In such situations, the standard logistic regression method is not valid as logistic regression requires the observations to be independent of one other. This situation arises when treating COVD-19 patients. Patients from certain clusters, such as geographic areas or the same family, are highly correlated, and we need to fit the model using the GEE approach. In this paper, Standard Logistic Regression (LS), Generalized Linear Models (GENMOD), and

GEE procedures have been utilized for comparison purposes. Simulated data have been used for case study analysis. The GEE model's accuracy rate is superior when data are binary clustered.
For more information about the GEE and GENMOD methods, see Fitzmaurice, Laird, and Ware (2011); Hardin and Hilbe (2003); Diggle et al. (2002); Lipsitz et al. (1994).

Identifying Subpopulations of a Hierarchical Structured Data using a Semi-Supervised Mixture Modeling Approach

Andrew Simpson, South Dakota State University
Semhar Michael, Christopher Saunders, and Dylan Borchert, South Dakota State University

The field of forensic statistics offers a unique hierarchical data structure in which a population is composed of several subpopulations of sources and a sample is collected from each source. This subpopulation structure creates a hierarchical layer. We propose using a semi-supervised mixture modeling approach to model the subpopulation structure which leverages the fact that we know the collection of samples came from the same, yet unknown, source. A simulation study based on a famous glass data was conducted and shows this method performs better than other unsupervised approaches which have been previously used in practice.

Standardized incidence ratio of the COVID-19 pandemic in a midwestern state

Emma Spors, South Dakota State University
Semhar Michael, South Dakota State University

The Coronavirus disease 2019 (COVID-19) has made a drastic impact around the world, with some communities facing harsher outcomes than others. We looked at what factors contributed to negative outcomes from the pandemic in South Dakota (SD). In addition, we sought to understand how counties in SD fared compared to expected using the SD rate as a reference population. To do this, a penalized generalized linear regression model was used to identify factors that were associated with COVID-19 hospitalization and death rates. The Standardized Incidence Ratio (SIR) values of all counties were computed three times for hospitalization rates and three times death rates to account for population, population and age, and population, age, and socio-demographic factors. With the linear model, we identified that race and education as significant factors associated with the outcomes which were confirmed by current literature. The SIR values highlighted counties that had more or less severe outcomes than expected. The counties with high non-white populations, which mostly included counties with American Indian reservations, typically had worse outcomes than other counties. Counties with high educational attainment typically had better outcomes than other counties. We believe that these results may provide useful information to improve the implementation of mitigation strategies to curb the damage of this or future pandemics by providing a way for data-driven resource allocation.

Predicting automobile accident severity and hotspots using multinomial logistic regression

Zhuoyu Yang, Minot State University
Yongmin Kim, Minot State University

Americans are now driving more than ever [1]. In 2010, close to 33,000 lives were lost and another estimated 3.9 million people were injured in automobile accidents; all things considered, these accidents accounted for $836 billion in damages [2]. Since then, the rate of automobile-related deaths per 100 million miles traveled has not shown signs of improvement [3]. This research expands upon a previous year's poster presented at the South Dakota State University Data Science Symposium 2019 [4]. While the previous research focuses on a data visualization of automobile accident hotspots on a map based on the severity and frequency of accidents, this research aims to train a multinomial logistic regression machine learning model using data related to weather conditions, speed limit, and GPS coordinates to predict the severity of

automobile accidents. The development of such a machine learning model can help inform emergency services better manage resources in anticipation of potential automobile accidents based on prevailing weather conditions, speed limit along a stretch of road, and location data. An updated version of the previous dataset will be used. This dataset contains approximately 1.5 million automobile accident data points, collected over a span of over four years, from February 2016 to December 2020.

References

[1] US Department of Transportation. Federal Highway Administration. (May, 2019). Strong economy has Americans driving more than ever before. Retrieved from https://cms8.fhwa.dot.gov/newsroom/strong-economy-has-americans-driving-more-ever
https://cms8.fhwa.dot.gov/newsroom/strong-economy-has-americans-driving-more-ever

[2] Blincoe, L. J., Miller, T. R., Zaloshnja, E., & Lawrence, B. A. (2015, May). The economic and societal impact of motor vehicle crashes, 2010. (Revised)(Report No. DOT HS 812 013). Washington, DC: National Highway Traffic Safety Administration

[3] National Center for Statistics and Analysis (2019, December). Early estimate of motor vehicle traffic fatalities for the first 9 months (Jan – Sep) of 2019. (Crash Stats Brief Statistical Summary. Report No. DOT HS 812 874). Washington, DC: Highway Traffic Safety Administration.

[4] Identification of Automobile Accident Hotspots using Countrywide Traffic Accident Dataset, B. Z. Yang & S. Z. Sajal, Ph.D., Presented at 2020 South Dakota State University Data Science Symposium. Public Affairs, Public Policy and Public Administration

Fine-tuning Transformer-based Natural Language Generation Algorithms for USDA Grains Reports for Farmers, Producers, and Small Businesses

Winston Zeng

The Transformer architecture for Natural Language Generation (NLG) was nothing short of a revolution for natural language processing. Self- and multi-head attention models have proven their efficacy in a variety of textual tasks, including classification, translation, summarization, and generation. Fine-tuning Transformer-based models on summarization tasks has proven successful with small textual datasets. In this research, we focus on fine-tuning pre-trained transformer-based NLG algorithms for USDA reports on grains to produce high-quality summaries, highlights, narratives, and Q&As, to enable farmers, producers, and small businesses in making informed decisions on production, investments, expansions, and risk management. Based on high-performing transformer-based algorithms pre-trained on large textual datasets such as collections of news articles, we develop, train, and fine-tune NLG algorithms with implementation of the Transformer model architecture on publicly available USDA grain publications and reports with characteristics that require specific considerations in model building and training, such as topic-specific (e.g, corn or soybeans), numerical data-intensive (e.g, supply and demand), and temporally sequential (e.g, weekly progress updates). A multi-stage end-to-end NLG system will be developed in which learning evolves from unsupervised, semi-supervised, to supervised, and the final stage of supervised learning would utilize human-expert-produced summaries and highlights. While the research contributes to the field of applying transformer-based models into specific domains, the practical goal is to automate the processes to efficiently inform intended audiences with high-quality content.