Keynotes

Zip codes and address changes - Using innovative data analytics to predict and impact health outcomes

Emily Griese, PhD Chief Operating Officer, Sanford Health Plan, Sioux Falls, SD

Healthcare continues to face headwinds including increasingly sicker populations and rising costs to deliver care. Because of this there is a need innovate in the approach to caring for patients, recognizing traditional healthcare and a one size fits all approach won't work. This talk will focus on innovative, data-driven approaches to uncovering the unique needs of patients. Specifically, leveraging advanced analytics and new sources of data reflective of non-medical factors influencing health and health outcomes.

Unraveling Complex Problems: Applying Systems Thinking in Data Science

Ryan Nichols, Data Science and Analytics Advisor, TransUnion

It would be an understatement to say that our world and its' problems are as complex as ever. When developing models it is essential that we create some simplified view of a system in order to focus on relevant information, while remaining aware of interdependencies with external systems. While this is somethings we all do to explain events, correlations, or estimate the impact of a decision; how often do you practice this approach of thinking intentionally? As data scientist, we find ourselves increasingly looked to as a primary source for answers and solutions. This task comes with the responsibility to consider perspectives outside our specialization in hopes of gaining understanding around our most urgent problems. In this session, we will explore the importance of understanding complexity in a concise and deliberate way to maximize the net benefit.

Invited abstracts

An 'Arg' to Comprehend Genetic Constraint

Suvobrata Chakravarty, Associate Professor, SDSU

As germline variants are filtered through natural selection, macro-genetic data on millions of germline variants obtained from next-generation sequencing (NGS) from a large human population provide an important avenue to address fundamental questions on selection pressure and genetic constraints such as probing the depletion of germline variants to gain functional insight into genetic constraints. Our study focuses on translating genetic constraints at the level of amino acid to probe protein function.

Talk to your data with ChatGPT and RTutor.ai

Steven Ge- Associate Professor, SDSU

The ability to gain insights from data is an essential skill for diverse careers across industries and academia. Analyzing such datasets often requires writing specialized computer programs, a time-consuming process. Capitalizing on recent advances in artificial intelligence (AI), we developed a platform that enables users to interact with data in natural languages through an interpreter – a large language model fluent in the languages of both humans and computers. Under a reactive programming environment (Shiny), we use OpenAI's Davinci model (ChatGPT's sibling) to translate user requests in English and a few dozen

languages into R or Python scripts, which are then executed alongside user-uploaded data to produce results or error messages instantaneously. RTutor.ai enables people without coding experience to conduct preliminary analysis and visualization while boosting productivity for those with experience. This proof-of-concept platform has the potential to democratize data science and facilitate learning.

Analysis of State and Parameter Estimation Techniques using Dynamic Perturbation Signals
Timothy M Hansen, South Dakota State University
The trend in electric power systems is the displacement of traditional synchronous generation (e.g., coal, natural gas) with renewable energy resources (e.g., wind, solar photovoltaic) and battery energy storage. These energy resources require power electronic converters (PECs) to interconnect to the grid and have different response characteristics and dynamic stability issues compared to conventional synchronous generators. As a result, there is a need for validated models to study and mitigate PEC-based stability issues, especially for converter dominated power systems (e.g., island power systems, remote microgrids). This presentation will introduce methods related to dynamic state and parameter estimation via the design of active perturbation signals for converter dominated power systems.

Equipment Finance Securitization – Driving business value through advanced analytics
Ed Krueger, Landon Thomson, and Sebastian Sowada - Channel
This presentation will focus first on providing an overview of Channel and the Data & Analytics team that performed this case study. We will introduce the topic of securitization. Our use case will demonstrate the benefits financially as well as reputationally. We will discuss the challenges of securitizing a portfolio with a lack of historical performance and share our creative solutions that include predictive modeling, data engineering, financial analysis, and data science. We will also share the outcome of our work, and the details of our inaugural securitization.

Improving Customer Experience Through Natural Language Processing
Valerie Reed and Paige Pennock - First Premier Bank
Unstructured data such as spoken words can be transcribed into text, transformed into structured data, then analyzed using Natural Language Processing (NLP) methods. Putting meaning behind this complex data is beneficial to a business by identifying newly arising issues and efficiently monitoring call topics. This can be done by two common NLP methods, sentiment analysis and topic modeling. Sentiment analysis is a form of text mining used by data scientists to determine emotion behind text. SAS Viya has developed a tool using a combination of NLP and machine learning methods to help users perform sentiment analyses. Topic modeling identifies groups of text that pertain to a similar subject. Identifying the point in a call with the highest friction and understanding the potential cause of that friction can help identify potential process improvements. Through word scoring and topic modeling we can identify this specific point in the call and determine what topics were discussed prior to the friction. In this presentation, we will cover the process, methods, and benefits of performing sentiment analysis and go through a case study utilizing topic modeling.

Statistical Discrimination Methods for Forensic Source Interpretation of Aluminum Powders in Explosives
Danica M. Ommen, Iowa State University
Christopher P. Saunders, South Dakota State University, JoAnn Buscaglia, Federal Bureau of Investigation

Aluminum (Al) powder is often used as a fuel in explosive devices; therefore, individuals attempting to make illegal improvised explosive devices often obtain it from legitimate commercial products or make it themselves using readily available Al starting materials. The characterization and differentiation between sources of Al powder for additional investigative and intelligence value has become increasingly important. Previous research modeled the distributions of micromorphometric features of Al powder particles within a subsample to support Al source discrimination. Since then, additional powder samples from a variety of different source types have been obtained and analyzed, providing a more comprehensive dataset for applying the two statistical methods for interpretation and discrimination of source. Here, we compare two different statistical techniques: one using linear discriminant analysis (LDA), and the other using a modification to the method used in ASTM E2927-16e1 and E2330-19. The LDA method results in an Al source classification for each questioned sample. Alternatively, our modification to the ASTM method uses an interval-based match criterion to associate or exclude each of the known sources as the actual source of a trace. Although the outcomes of these two statistical methods are fundamentally different, their performance with respect to the closed-set identification of source problem is compared. Additionally, the modified ASTM method will be adapted to provide a vector of scores in lieu of the binary decision as the first step towards a score-based likelihood ratio for interpreting Al powder micromorphometric measurement data.

Advancing Machine Learning Through Multilinear Subspace Methods
Cagri Ozdimir, SDSMT
Randy C. Hoover, Kyle Caudle, Karen Braman, and Jackson Cates, SDSMT
The last few decades have seen significant advances in various sensing instruments generating massive volumes of n-dimensional data. To best explore, analyze, and provide insights from such data, new mathematical tools are needed in an effort to bridge the gap between traditional machine/deep learning models and their multilinear counterparts. This talk presents a new approach to dealing with such data using a multilinear (tensor-tensor) perspective and provides insight into how such a bridge can be built. In particular, we formulate the mathematical theory of tensor decompositions via an algebra of circulants detailing novel extensions of traditional linear algebraic tools. We then provide insights into several different application are within the machine/deep learning community and illustrate how multilinear extensions can be achieved. We conclude with an informal discussion surrounding such developments and possible application spaces not yet investigated.

An Overview of Deep Learning and Universality
Michael Puthawala – Assistant Professor - SDSU
In recent years machine learning, and in particular deep learning has emerged as a powerful and robust tool for solving problems in fields ranging from robotics, to medicine, materials science, cosmology and beyond. As work on applications has advanced, so too has theory advanced to guide, explain, interpret deep learning. In this talk I will provide an overview on universality of neural networks. I will explain what it means for a neural network to be a universal approximator, qualitatively and quantitatively, and give examples of these results applied to existing networks. Finally, I will conclude by discussing some recent work on manifold universality in the nascent field of geometric machine learning.

Parallel and Distributed Query Engine for Federated Searching
Kyle Putnam – Principal Software Engineer – Query.AAI
Network and IT data volumes used to be manageable and were housed in data lakes inside corporate networks. Fast forward to today, with the explosion of Cloud and SaaS, data volumes are enormous, getting larger daily, and highly distributed. For any organization, its data now resides outside its perimeter, across

multiple cloud providers and 3rd party SaaS vendors. This makes querying and correlating user, device and application data a big challenge. We built an efficient and cost–effective distributed query engine that lets organizations seamlessly query their data and get the relevant answers and context they desire. Our distributed query engine is built to auto-scale, run parallel queries, perform contextual lookups, and find optimal query execution plans. Queries are executed over vendors' APIs. End-user applications like federated searching are built leveraging a GraphQL API interface to our query engine. Operational use-cases include analyzing phishing, suspicious logins, and other cybersecurity threats to the organization.

Active Learning to Minimize the Possible Risk from Future Epidemics
KC Santosh, University of South Dakota

In medical imaging informatics, for any future epidemics (e.g., Covid-19), deep learning (DL) models are of no use as they require a large dataset as they take months and even years to collect enough data (with annotations). In such a context, active learning (or human/expert-in-the-loop) is the must, where a machine can learn from the first day with minimum possible labeled data. In unsupervised learning, we propose to build pre-trained DL models that iteratively learn independently over time, where human/expert intervenes only when it makes mistakes and for only a limited data. In our work, deep features are used to classify data into two clusters (0/1: Covid-19/non-Covid-19) on two different image datasets: chest x-ray (CXR) and Computed Tomography (CT) scan of sizes 4,714 and 10,000 CTs, respectively. Using pre-trained DL models and unsupervised learning, in our active learning framework, we received the highest AUC of 0.99 and 0.94 on CXR and CT scan datasets, respectively. Not to be confused, our primary objective is to provide a strong assertion on how active learning could potentially be used to predict disease from any upcoming epidemics.

Utilizing Cloud Resources to Develop and Deploy Machine Learning Solutions
Eric Stratman, ValidiFI

Many organizations across a variety of industries want to incorporate data science into their decision making. By using data science methodologies, organizations can transform into the next evolution of their business. Instead of constantly being reactive, they can become more proactive in their business decisions by implementing data science solutions. One issue that is commonly encountered within many of these organizations, is that after they have developed a machine learning solution they are unsure how to implement such solution. These machine learning solutions need to not only predict and score automatically, but they need to be scalable to meet an organization's demands. The use of cloud infrastructure technology can solve these problems. Cloud resources such as Azure and AWS have the tools and resources available to help data scientists analyze data, develop machine learning models, and automate data science solutions. My presentation provides an introduction on how to utilize cloud resources in the Azure platform to implement machine learning solutions to help organizations evolve and make better decisions.

CancerTrialMatch: a web application for the curation and matching of clinical trials at a precision oncology center
Padmapriya Swaminathan
Shivani Kapadia, Anu Amallraja, Casey Williams, and Tobias Meissner, Avera Cancer Institute Center for Precision Oncology, Sioux Falls, SD, USA

The adoption of next-generation sequencing for cancer patients has made molecular profiling possible and identify biomarkers. At the Avera Cancer Institute, biomarker-based clinical trials are often presented as treatment options to oncologists at the molecular tumor board. This necessitated a method to capture

structured trial data and match them to patients based on their disease and sequencing profile in a systematic manner.
We developed an open-source web application, CancerTrialMatch, that enables to (i) add trials through a semi-automated curation interface, (ii) browse and search trials, (iii) and match patients to biomarker-based trials.
This application uses R Shiny to create simple interfaces, a mongo database to store trial data, various R libraries to query data and perform computations, and Docker to manage software installation and application instantiation. The curation interface is semi-automated because querying the clinicaltrials.gov API will return discrete data for many fields, except biomarkers and disease subtypes. The user has to manually input disease type based on the OncoTree classification, as well as biomarker information for mutations, copy numbers, fusions, TMB, MSI/PD-L1 status, RNA expression, and disease-specific markers such as ER/PR/HER2 status.
We believe that this computational resource will reduce the person-hours required for trial management for a patient, aid in increasing clinical trial enrollment by systematically providing treatment options, and hence an important tool in the clinical application of precision oncology.

A New Goodness-of-Fit Statistic for Binormal ROC Curve
Larry Tang – University of Central Florida
In the context of assessing binary classification model performance for a single random variable, Receiver Operating Characteristic curves (abbreviated as ROC curves) graphs the false positive rate on the x-axis against the true positive rate on the y-axis. We study a new goodness-of-fit statistic which uniquely measures a lack of fit between a parametric binormal and an empirical ROC curve in the perpendicular direction. In the simulation studies, we measure perpendicular distances from the theoretical ROC curve to the empirical ROC curve. Based on the simulation results we estimate the associated one-sided tests' critical values and confidence intervals based on sample size and order statistics. Given the critical values, we perform new simulation experiments with alternative negative distributions to assess the power of these perpendicular distance goodness of fit tests focusing on the binormal model. Further research will shed light on the cases where the power of the perpendicular distance-based goodness of fit tests justifies the additional complexity and computational time compared to vertical distance-based goodness of fit tests.

Don't solve the wrong problem! Cautionary tales of data science in the industry
Jeremy Werne, Allstate
Being a data scientist is awesome, but the job isn't what most expect. In this talk we will explore how professional data science actually works in the industry, bust or confirm common myths about data science jobs, and study how to frame data science problems correctly.

Accepted oral presentations

Can machine learning predict particle deposition at specific intranasal regions based on computational fluid dynamics inputs/outputs and nasal geometry measurements?
Mohammad Mehedi Hasan Akash, South Dakota State University

Zachary Silfen, Boston University, Diane Joseph-McCarthy, Boston University, Arijit Chakravarty, Fractal Therapeutics, Saikat Basu, South Dakota State University
Along with machine learning modeling, numerical simulations of respiratory airflow and particle transport can be used to improve targeted deposition at the upper respiratory infection site of numerous airborne diseases. Given the need for more patient data from varied demographics, we propose a machine learning-enabled protocol for determining optimal formulation design parameters that may match nasal spray device settings for successful drug delivery. We measured 11 anatomical parameters (including nasopharyngeal volume, nostril heights, and mid-nasal cavity volume) for 10 CT-based nasal geometries representative of the population for this aim. We also ran 160 computational fluid dynamics simulations of drug delivery on the same geometries for various breathing situations, using varied pressure gradients to drive inhaled air transport to evaluate drug deposition at the various upper airway areas for nasal inhalers. Using this test data, we constructed 18 machine-learning models to estimate the targeted deposition at the different regions of the upper airway. This study contributes to developing a customized, efficient intranasal delivery system for prophylactics, treatments, and immunizations; the findings will apply to a broad spectrum of respiratory disorders.

AI-driven wheat yield and protein content forecasting using UAV remote sensing
Mohammad Maruf Billah, South Dakota State University
Maitiniyazi Maimaitijiang, South Dakota State University, Shahid Nawaz Khan, South Dakota State University, Swas Kaushal, South Dakota State University, Jonathan Kleinjan, South Dakota State University
Preharvest forecasting of wheat grain yield, test weight and protein content is critical in terms of in season decision making and field management practices, as well as field-based high-throughput phenotyping toward enhanced yield and grain quality. In recent years, the rapid advancement of Unmanned Aerial Vehicle (UAV) and sensor technologies enabled high-resolution spatial, spectral, and temporal data collection with a lower cost. Coupled with cutting-edge artificial intelligence and deep learning (AI/DL) algorithms, UAV remote sensing has become an important tool in a variety of agricultural applications. This study aims to investigate the potential of UAV-based multitemporal multispectral data for preharvest wheat yield, test weight and protein content estimation under the framework of AI/DL. UAV-based multispectral images were collected throughout the 2022 winter wheat growing season over seven experimental winter wheat fields across South Dakota, USA. Plot-level canopy spectral and texture features were derived from UAV multispectral imagery. Traditional machine learning approaches such as Partial Least Squares Regression, Support Vector Regression, and Random Forest Regression were employed to develop prediction models using plot-level averaged spectral and texture features. Additionally, deep learning methods such as Convolutional Neural Networks (CNN) and hybrid CNN and Long-Short Term Memory (CNN-LSTM) were also implemented using plot-level reflectance imagery as input for prediction model development. This research highlights the potential of coupling high-resolution UAV remote sensing with cutting-edge AI/DL in predicting wheat yield, test wight and protein content. The results from this work deliver valuable insights for high-throughput phenotyping and crop field management with high spatial precision.

The effect of boom leveling on spray dispersion
Travis Burgers, South Dakota State University
Miguel Bustamante, Universidad Adolfo Ibáñez, Juan F Vivanco, Universidad Adolfo Ibáñez
Self-propelled sprayers are commonly used in agriculture to disperse agrichemicals. These sprayers commonly have two boom wings with dozens of nozzles that disperse the chemicals. Automatic boom height systems reduce the variability of agricultural sprayer boom height, which is important to reduce

uneven spray dispersion if the boom is not at the target height. A computational model was created to simulate the spray dispersion under the following conditions: a) one stationary nozzle based on the measured spray pattern from one nozzle, b) one stationary model due to an angled boom, c) superposition of multiple stationary nozzles due an angled boom, and d) superposition of multiple nozzles given the inputs of measured boom heights and the position of the sprayer over a field in time. The effect of boom leveling on spray dispersion was compared for three boom leveling systems on two sprayers (Systems A and B on a John Deere R4045, Systems B and C on an AGCO RoGator 1100C). For each boom leveling system, the measured boom height and sprayer position in time for one run was used (medium terrain course at 26 kph (16 mph)) [Burgers et al. Appl. Eng. Agric. 2021]. For each run, a coverage map was calculated with the measured boom heights and a reference level boom; the spray application error was calculated as the difference between them. The area for which the application error was less than 10% was 34.4 and 56.6% for Systems A and B on the R4045, respectively, and 45.0 and 59.3% for Systems C and B on the RoGator, respectively. This model can be used to quantify and compare coverage maps from boom leveling systems.

Temporal Tensor Factorization for Multidimensional Forecasting
Jackson Cates,SDSMT
Karissa Scipke, SDSMT, Randy Hoover, SDSMT, Kyle Caudle, SDSMT
In the era of big data, there is a need for forecasting high-dimensional time series that might be incomplete, sparse, and/or nonstationary. The current research aims to solve this problem for two-dimensional data through a combination of temporal matrix factorization (TMF) and low-rank tensor factorization. From this method, we propose an expansion of TMF to two-dimensional data: temporal tensor factorization (TTF). The current research aims to interpolate missing values via low-rank tensor factorization, which produces a latent space of the original multilinear time series. We then can perform forecasting in the latent space. We present experimental results of the proposed method with other state of the art methods on the Jericho-E-Usage energy dataset.

Skip-GCN : A Framework for Hierarchical Graph Representation Learning
Jackson Cates,SDSMT
Justin Lewis, SDSMT, Randy Hoover, SDSMT, Kyle Caudle, SDSMT
Recently there has been high demand for the representation learning of graphs. Graphs are a complex data structure that contains both topology and features. There are first several domains for graphs, such as infectious disease contact tracing and social media network communications interactions. The literature describes several methods developed that work to represent nodes in an embedding space, allowing for classical techniques to perform node classification and prediction. One such method is the graph convolutional neural network that aggregates the node neighbor's features to create the embedding. Another method, Walklets, takes advantage of the topological information stored in a graph to create the embedding space. We propose a method that takes advantage of both the feature embeddings and topological by an intersection of the two methods. We first represent information across the entire hierarchy of the network by allowing the graph convolutional network to skip neighbors in its convolutions. Then using multilinear algebra, we can capture correlations across the hierarchies to create our node embeddings by representing our convolutions as a tensor. We can follow up the captured node embeddings by a dense layer to perform node classification or link prediction.

Would AI Stocks Estimate Be as Surprised to USDA Stock Reports as Private Market Analysts?
Asif Mahmud Chowdhury, South Dakota State University
Matthew Elliott, South Dakota State University

The USDA survey-based Quarterly Grain Stocks reports are the primary source of information regarding the relative supply of U.S. corn, soybeans, and wheat for the last fifty years. Previous research has examined the accuracy of the USDA stock reports and their relevancy to the market, given alternative sources of estimates (e.g., Isengildina-Massa et al., 2021). For example, private industry analysts also estimate expected quarterly grain stock reports before USDA releases their reports. Market information firms such as Bloomberg and Reuters publish a subset of these estimates a few days before the USDA reports. Previous research has found that when industry analysts have significant differences in stock expectations compared to what the USDA releases for grain stocks, market prices tend to adjust rapidly to what the USDA found in their survey. Many media outlets and previous research attribute the differences in expectations and changes in market prices to a ""market surprise.""Karali et al. (2020) found compelling evidence that the discrepancy in USDA reports from private analysts' expectations plays a vital role in explaining grain futures price movements on report days. Market analysts, USDA officials, and researchers have given four reasons for market surprises in the grain stocks reports. First, USDA surveys may fail to account for grain in transit when surveying stocks. Second, many private analysts use standard conversion rates (e.g., average test weight per bushel of reported corn) for products derived from grain inputs to estimate their expected grain stocks after a quarter. However, these conversion rates may vary because of the quality of the grain and be less (more) than what private analysts estimate. Third, errors in estimating what portion of existing stocks is from old or new crop production may cause surprises in the final annual report before a change in the marketing year. For example, USDA asks in their survey how much old crop corn is on hand on September 1st, although some crops taken in by grain wholesalers can be new crops by this date. Fourth, USDA survey-based stock reports contain survey noise. It is still being determined whether market analysts can correctly consider survey noise when reconciling their estimates versus the USDA and smooth future estimates, assuming some portion of the previous report was due to noise and survey error. What is the primary reason market analysts are frequently surprised by USDA QAS reports? Given the recent surge in grain movement data, available grain quality data, and data on the output of significant demand sources of grain, particularly at a state level, it is possible to use advances in analyzing high dimensional data (e.g., random forest, gradient boosting) to develop an objective artificial intelligent (AI) market analyst. This paper aims to explore additional public data sources related to commodity demand and supply in the corn, wheat, and soybean markets and apply AI techniques to determine whether data analytics improves the prediction of QAS reports released by USDA for corn, soybeans, and wheat. Our primary research objective is to determine if AI can more accurately predict QAS estimates from USDA than the survey of Market analysts that Bloomberg and Reuters have historically provided. Our secondary objective is to attempt to decompose the surprise into by source of surprise. We will use Random Forest ML model to predict the stock estimate of the three major commodities (Corn, Soybean, and Wheat) with all the publicly available data before the national announcement of the Quarterly Stock Report. We will compare the stock estimate provided by our AI techniques to private market analysts, which have been a critical component of information before the announcement days. Our research findings will also decompose the variables most important for explaining market surprises. Specifically, does the amount of grain in transit, changes in demand due to grain quality, or the mixing of new crops and old crops in stock estimates mainly explain the surprise? Further, our findings may determine if private analysts have problems reconciling noise in previous USDA surveys when making future estimates for future reports.

Finding Needles in Haystacks: Rare Category Detection using Semi-supervised Active Learning
Rohan Loveland, SDSMT

Rare category detection addresses the problem of exploring data sets that are too large for unaided analysis. The Farpoint algorithm utilizes machine learning techniques to discover sparsely represented classes,

through interactive queries and semi-supervised clustering. Application of the algorithm to skewed MNIST datasets is used to empirically characterize performance.

Toward Online Data-Driven Control: The Theory, Methodology, and System Stability
Huitian Lu, South Dakota State University
Moses Rabai, South Dakota State University, Sean Edeki, South Dakota State University
Because of the rapid development in data computation science and computer technology, online (real-time) date-driven control is becoming possible. The control strategy in the data-driven system comes from abstracting or learning from data. This so-called data-driven control with AI technology possesses distinctive advantages for nonlinear dynamical system control without a system-identified model. The data-driven method has advantages in dealing with nonlinear dynamics characteristics. High-state dimension, system nonlinearity, time-variant parameters, signal noise, uncertainties, and stochasticity are often studied in real-world application presentations. This work briefly presents the discussions on methods from model-based control (MBC) toward data-driven control (DDC) and the strategies adopted in data-driven methodology. The data-driven model-predictive control (MPC), data-driven machine-learning control (MLC), and some commonly adopted mathematical algorithms for optimization. Other discussions on specific concerns in the practice of DDC. The stability of a system with a data-driven dynamic approach is also studied as a necessary condition for an automatic (closed-loop) control system and the control function.

On Reliability Estimation of Lomax Distribution under Adaptive Type-I Progressive Hybrid Censoring
HassanOkasha Al-Azhar University – Egypt
Y. L. Lio. University of South Dakota,Mohammed Albassam, King Abdulaziz University
Bayesian estimates involve the selection of hyper-parameters in the prior distribution. To deal with this issue, the empirical Bayesian and E-Bayesian estimates may be used to overcome this problem. The first one uses the maximum likelihood estimate (MLE) procedure to decide the hyper-parameters; while the second one uses the expectation of the Bayesian estimate taken over the joint prior distribution of the hyper-parameters. This study focuses on establishing the E-Bayesian estimates for the Lomax distribution shape parameter functions by utilizing the Gamma prior of the unknown shape parameter along with three distinctive joint priors of Gamma hyper-parameters based two asymmetric loss functions include a general entropy and LINEX loss functions. To investigate the effect of the hyper-parameters' selections, mathematical propositions have been derived for the E-Bayesian estimates of the three shape functions that comprise the identity, reliability and hazard rate functions. Monte Carlo simulation has been performed to compare all procedures. Two real data sets from industry life test and medical study are applied for the illustrative purpose.

Robustness Analysis of Convolutional Layers in Image Classification Neural Networks
Gabriel Picioroaga, University of South Dakota
Reliable neural networks are designed so that their outputs satisfy a Lipschitz condition with respect to small deformations of the inputs. The so-called Lipschitz inequality is controlled by a constant which naturally implies that too large a change in the output requires some change in the input. Large Lipschitz constants are hypothesized as the culprit in classification errors, either intentional (e.g. adversarial attacks) or not. At each layer in a network one can find a (local) Lipschitz constant depending on the activation function and the transformation by which features are extracted (convolution, matrix multiplication, compression). The total constant is at most the product of all local ones. In our work we showed that at a convolutional layer one can derive a non-Lipschitz like inequality. We consider particular deformations which are obtained by convolving with a fixed kernel, e.g. a blurring filter. By taking advantage of

mathematical properties of convolution (commutativity and L² -norm estimates) we prove that the convolutional layer's outputs of a signal and its deformation satisfy a non-Lipschitz inequality via a constant that depends only on the layer's parameters and the L¹-norm of the blurring kernel. Although the bounding constant is less than 1, the error in the output seems to propagate down the network and thus heuristically explain classification errors of blurry images. In our experiments in Matlab with various image classification neural nets, pre-trained and ad-hoc trained (imported with all parameters except the fully connected layer) we have noticed consistent errors in classifying blurred images. With ad-hoc trained networks we have noticed improvement in classification of heavily blurred images after the data sets were augmented with slightly blurred copies. This is joint work with Nate Harding.

Ensemble of Score Likelihood Ratios for the common source problem
Federico Veneri, Iowa State University/CSAFE
Danica M. Ommen, Iowa State University/CSAFE
Machine learning-based Score Likelihood Ratios have been proposed as an alternative to traditional Likelihood Ratios and Bayes Factor to quantify the value of evidence when contrasting two opposing propositions. Under the common source problem, the opposing proposition relates to the inferential problem of assessing whether two items come from the same source. Machine learning techniques can be used to construct a (dis)similarity score for complex data when developing a traditional model is infeasible, and density estimation is used to estimate the likelihood of the scores under both propositions. In practice, the metric and its distribution are developed using pairwise comparisons constructed from a sample of the background population. Generating these comparisons results in a complex dependence structure violating assumptions fundamental to most methods. To remedy this lack of independence, we introduce a sampling approach to construct training and estimation sets where assumptions are met. Using these newly created datasets, we construct multiple base SLR systems and aggregate their information into a final score to quantify the value of evidence. Our experimental results show that this ensembled SLR can outperform traditional SLR in terms of the rate of misleading evidence, discriminatory power and show they are more reliable.

Accepted poster abstracts

2D respiratory sound analysis to detect lung abnormalities
Rafia Sharmin Alice, University of South Dakota
KC Santosh, University of South Dakota
In this paper, we analyze deep visual features from 2D data representation(s) of the respiratory sound to detect evidence of lung abnormalities. The primary motivation behind this is that visual cues are more important in decision-making than raw data (lung sound). Early detection and prompt treatments are essential for any future possible respiratory disorders, and respiratory sound is proven to be one of the biomarkers. In contrast to state-of-the-art approaches, we aim at understanding/analyzing visual features using our Convolutional Neural Networks (CNN) tailored Deep Learning Models, where we consider all possible 2D data such as Spectrogram, Mel-frequency Cepstral Coefficients (MFCC), spectral centroid, and spectral roll-off. In our experiments, using the publicly available respiratory sound database named ICBHI 2017 (5.5 hours of recordings containing 6898 respiratory cycles from 126 subjects), we received the highest performance with the area under the curve of 0.79 from Spectrogram as opposed to 0.48 AUC from the raw data from a pre-trained deep learning model: VGG16. Our study proved that 2D data representation

could help better understand/analyze lung abnormalities as compared to 1D data. In addition, our results can be compared with previous works.

Covariance based clustering for classification
Theophilus Anim Bediak, South Daktoa State University
Semhar Michael, South Dakota State University
Classification involves assigning observations to known classes based on some common features. Commonly known methods for solving classification problems include linear and quadratic discriminant analysis. Both methods assume each class follows a multivariate normal distribution with common covariance matrix for LDA and with class specific covariances for QDA. The equal covariance assumption in LDA is often described as almost unrealistic and simplistic in real-life applications, resulting in less flexibility and high bias. Estimating class-specific covariance matrices becomes a problem when there are large number of classes. This work proposes a method which is a compromise between LDA and QDA. Here a covariance-based clustering method using mixtures of Wishart are used to identify classes that have common covariance matrix. The method is applied to glass fragments classification problem as a means to forensic source identification.

A Characterization of Bias Introduced into Forensic Source Identification when there is a Subpopulation Structure in the Relevant Source Population.
Dylan Borchert, South Dakota State University
Semhar Michael and Christopher Saunders, South Dakota State University
In forensic source identification the forensic expert is responsible for providing a summary of the evidence that allows for a decision maker to make a logical and coherent decision concerning the source of some trace evidence of interest. The academic consensus is usually that this summary should take the form of a likelihood ratio (LR) that summarizes the likelihood of the trace evidence arising under two competing propositions. These competing propositions are usually referred to as the prosecution's proposition, that the specified source is the actual source of the trace evidence, and the defense's proposition, that another source in a relevant background population is the actual source of the trace evidence. When a relevant background population has a subpopulation structure, the rates of misleading evidence of the LR will tend to vary within the subpopulations, sometimes to an alarming degree. Our preliminary work concerning synthetic and real data indicates that the rates of misleading evidence are different among subpopulations of different sizes, which can lead to a systematic bias when using a LR to present evidence. In this presentation we will summarize our preliminary results for characterizing this bias.

Models for Predicting Maximum Potential Intensity of Tropical Cyclones
Iftekhar Chowdhury, South Dakota State University
Gemechis Djira, South Dakota State University
Tropical cyclones (TCs) are considered as extreme weather events, which has a low-pressure center, namely an eye, strong winds, and a spiral arrangement of thunderstorms that produces heavy rain, storm surges, and can cause severe destruction in coastal areas worldwide. Therefore, reliable forecasts of the maximum potential intensity (MPI) of TCs are critical to estimate the damages to properties, lives, and risk assessment. In this study, we explore and propose various regression models, to predict the potential intensity of TCs in the North Atlantic at 12, 24, 36, 48, 60, and 72- hour forecasting lead time. In addition, a popular machine learning method, the Gradient Boosted Regression Tree (GBRT) algorithms is also used to further predict the TCs intensity changes in every 12-h over its entire lifespan (up to 72-hour) and the model is optimized by the Bayesian Optimization algorithm. The MPI is determined by using maximum wind speed (VMAX) measured in knots associated with a TC. The model utilizes the data of 160 TCs (n = 8,011) over North

Atlantic Basin obtained from the extended best track database from 2000 to 2021. We used data from 2000-2016 for model calibration and data from 2017-2021 for model validation. The results indicate that radius of maximum wind (RMAX), minimum central pressure (MCP), and latitude are the significant predictors that captures MPI of observed TCs. Some other implications of the results are also discussed.

Application of Gaussian Mixture Models to Simulated Additive Manufacturing
Jason Hasse, South Dakota State University
Semhar Michael, South Dakota State University, Anamika Prasad, Florida International University
Additive manufacturing (AM) is the process of building components through an iterative process of adding material in specific designs. AM has a wide range of process parameters that influence the quality of the component. This work applies Gaussian mixture models to detect clusters of similar stress values within and across components manufactured with varying process parameters. Further, a mixture of regression models is considered to simultaneously find groups and also fit regression within each group. The results are compared with a previous naive approach.

Deep Generative Modeling for Communication Systems Testing and Data Sharing
Kyle Caudle, South Dakota School of Mines and Technology;
Randy Hoover, Larry Pyeatt, and Trevor Krason, South Dakota School of Mines and Technology;
A common problem that limits distribution of real-world RF data is data sensitivity. Data often contain proprietary information or, in the case of military applications, the data may actually be classified. The goal of the proposed research is to investigate generative models and maps that will remove or obfuscate sensitive information from a population while otherwise being faithful to the target distribution. In addition to seeking to solve the population obfuscation problem, several potential approaches will be presented in this poster as well as initial experiments for investigating and testing these approaches.

Spatial Data Analysis for Traffic Safety Network Screening
Akosua Okyere-Addo, South Dakota State University
S. M. Rahat Rashedi, South Dakota State University
The roadway system represents a major investment, both public and private, and a valuable resource that enables mobility and accessibility to users. Due to degradation of aging infrastructure and increasing traffic, transportation agencies are seeking to effectively update or improve the system. With rising costs, tight budgets, and limited land resources, agencies are seeking effective techniques for identifying critical mobility and safety concerns. Historically, assignment of crashes to portions of the network, whether segments or intersections, has been the primary manner to link crash and road elements. The primary goal is to explore a potentially more efficient and effective means of developing roadway connected crash cluster identification results as an input to network screening and diagnostics. Beyond this, using the data linkages, we intend to explore crash typology (e.g., severity, collision type) distributions and clusters with respect to network (traffic and roadway) characteristics. The research uses GIS and spatiotemporal analysis techniques relying on crash locations as a basis rather than elements of the road network (e.g., intersections, non-intersections). The intent of the analysis is to develop crash clusters that can be flagged for further analysis and potential mitigation. Crash, roadway geometrics, and traffic data of 5 years (2015-2019) were collected for Story County of Iowa. These readily available data are being analyzed using Geospatial Information System (GIS) and some statistical models such as the Markov Switching Models (MSM). Existing methodologies will also be used to develop results that will be compared with the new methodology. There have been many network screening methods since the 1970s and the more widespread inception of computing resources. Though these methods have some advantages, they do have several

shortcomings as well as detailed in the Highway Safety Manual (HSM). Due to these shortcomings, the HSM and other sources promote methods that address these shortcomings such as the Empirical Bayesian (EB) and Hierarchical Bayesian (HB) methods. Historically, assignment of crashes to segments or intersections, has been the primary manner to connect crashes and road elements. The use of GIS and spatial and temporal analysis, alternative methods for connecting crashes to the roadway network have been developed to replace the older methods like screening with additional differentiating criteria.
Impact – The primary outcome of this research is the development of a new method of generating distributions and clusters of crashes along a roadway network for use in traffic safety screening. Practitioners should be able to utilize the process to develop network screening related to their jurisdictions, given sufficient and appropriate input data.

Spatial Data Analysis for the Development of Expected Adverse Weather Charts for Transportation Construction Projects

S M Rahat Rashedi, South Dakota State University
Akosua Ofosua, Okyere-Addo, South Dakota State University

Seasonal and daily weather events impact construction projects across the various climate regions of South Dakota in differing fashions. Additionally, the impacts for similar weather events can impact grading, surfacing, and structural construction activities in various ways. Adverse weather conditions can cause major delays which may lead to time extensions and increase project cost. To address these issues, South Dakota Department of Transportation (SDDOT) developed Working Day Weather Charts in 1998. However, advances in construction practices and weather prediction as well as climatic changes have occurred over the interim 25 years. This study is focused on developing updated zones, tables, charts, and recommendations for roads and bridges construction in South Dakota. The tables and charts are planned to be developed on both weekly and monthly basis to determine the impact of adverse weather events on construction projects and for use in future contracts. Weather, soil, and hydrographic data for South Dakota state are being considered for this study. The primary importance is on the weather data which is collected for 30 years (1991-2020) period from National Oceanic and Atmospheric Administration (NOAA). The important weather data parameters are temperature, snow, rainfall, and wind. The soil data have been collected from the broad-based inventory of soils and non-soil areas of the United States namely State Soil Geographic (STATSGO2). The key focus is to analyze the soil parameters in combination with adverse weather events that affect the construction of roads and bridges. The hydrographic data is focused on the peak flow at major water bodies in South Dakota that may cause flooding or ponding which affects road and bridge construction. Additionally, interviews with SDDOT personnel and construction contractors were conducted to determine factors important to the industry. Starting with data exploration of all the available data, key parameters will be analyzed to develop updated expected adverse weather day chart and updated zones. A considerable amount of work has been done on effects of weather on construction type categories and various Department of Transportation agencies evaluate the use of adverse weather in contract time calculations. The Virginia Department of Transportation place contract determination guidelines online. The VDOT document provides steps in determining contract time but contained little information on the impact of adverse weather on contract time calculations. Another document from the National Research Council of Canada on construction work protocols during winter in 1971. Beyond that, a recent (2022) publication from the National Cooperative Highway Research Program (NCHRP) covers a systematic approach for determining construction contract time. However, in most papers, little information is documented on the impact of adverse weather and how to implement that in tables and charts for construction type activities across South Dakota. The results can directly help SDDOT engineers and contractors to estimate the appropriate contract time and warranted time extension due to unexpected

adverse weather for variety of transportation construction projects across the diverse geographical terrains and climates of South Dakota.

What are your Strengths?: An Analysis of the Correlation of Strengths and Majors
Sherryl Mae Rowe, Dordt University
A crucial aspect of working with others in the workforce is knowing one's strengths and weaknesses. The analytical company Gallup has allowed students and professionals to discover these traits through the CliftonStrengths test. For each of the one hundred seventy-seven questions, the test gives two statements on opposite sides of a scale. From there, users can rank whether one side of the scale describes them better or is neutral. Analytical software then ranks a person's interaction with thirty-four identified strengths. The focus of this research includes looking at the correlation between students' strengths and their majors. Other factors, such as extracurriculars, gender, predicted success rate, and whether they are a first-generation student, were also included. The practical implication of these results presents the commonly seen strengths a student within a particular major would possess, which could allow professors to help grow the weaknesses while educating their students on how to thrive within a group work setting.

Two-Stage Approach for Forensic Handwriting Analysis
Ashlan J Simpson, Iowa State University
Danica M Ommen, Iowa State University
Trained experts currently perform the handwriting analysis required in the criminal justice field, but this can create biases, delays, and expenses, leaving room for improvement. Prior research has sought to address this by analyzing handwriting through feature-based and score-based likelihood ratios for assessing evidence within a probabilistic framework. However, error rates are not well defined within this framework, making it difficult to evaluate the method and can lead to making a greater-than-expected number of errors when applying the approach. This research explores a method for assessing handwriting within the Two-Stage framework, which allows for quantifying error rates as recommended by a federal report by PCAST (Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature Comparison Methods). The coincidence probabilities produced here can be used in later research to asses error rates using a ROC curve.

Finite Mixture Modeling for Hierarchically Structured Data with Application to Keystroke Dynamics
Andrew Simpson, South Dakota State University
Semhar Michael, South Dakota State University
Keystroke dynamics has been used to both authenticate users of computer systems and detect unauthorized users who attempt to access the system. Monitoring keystroke dynamics adds another level to computer security as passwords are often compromised. Keystrokes can also be continuously monitored long after a password has been entered and the user is accessing the system for added security. Many of the current methods that have been proposed are supervised methods in that they assume that the true user of each keystroke is known apriori. This is not always true for example with businesses and government agencies which have internal systems that multiple people have access to. This implies that unsupervised methods must be employed for these situations. One may propose using finite mixture models to model the keystroke dynamics but we show that there is often not a one-to-one relationship between the number of mixture components and the number of users. Also, users usually type numerous times during the session or block of time while using the system which means the keystroke dynamics from the session can be assumed to have arisen from the same user. We propose a novel method that accounts for the lack of a one-to-one relationship between the number of users and the number of components as well as accounts for known

information based on when keystrokes were typed. Based on simulation studies and the motivating real-data example the proposed model shows good performance.

A novel approach to detect COVID-19 fake news by mining biomedical information from news articles

Jordan Smith, St. Cloud State University
Faizi Fifita, Mengshi Zhou, St. Cloud State University

Containing the spread of fake news is crucial to the management of the COVID-19 pandemic response. Machine learning models have been used to automatically detect COVID-19 fake news from news content. Current machine learning features have limitations in preserving the biomedical information reported in the news articles. To address this limitation, we propose a novel approach to predict COVID-19 fake news by combining biomedical information extraction (BioIE) with machine learning models. We extracted 158 novel features using advanced BioIE algorithms. Fifteen machine learning classifiers were trained to predict the COVID-19 fake news using those BioIE-based features. Among the fifteen classifiers, random forest achieved the best performance with an Area Under the ROC curve (AUC) of 0.892. We demonstrated that BioIE-based features have higher prediction power as compared to the existing machine learning features. We next developed a multi-modality model by combining BioIE-based features with existing features. Our new model outperformed a state-of-art multi-modality model (AUC 0.916 vs. 0.875). In summary, our study indicates that mining biomedical information from news articles has great potential in identifying COVID-19 fake news.

Comparing Crime Rates Before and After the Covid-19 Pandemic in the United States

Anna C Stevens, St. Cloud State University
Shiju Zhang, St. Cloud State University

We will compare the rates of several types of crime in each state before and after the covid-19 pandemic. The results will be presented on the United States map for clear visualization. In addition, an interactive app will be created to allow for smooth workflow.

The Relevance of Shame Across Time and Location

Miranda Vander Berg, Dordt College
Kari Sandouka, Dordt College

Twitter is used among various entities professionals, politicians, and the general public as an online social network. Many tweets are informational, but others are reactive based on judgment that leads to public shaming. In response to the book “The Shame Machine” (by Cathy O’Neil), we look at Tweets to determine a linguistical and content analysis of shame. The research focuses on content analysis to define if a tweet contains language that is deduced as public shaming. Other factors relating to the tweet are the time, date, location of the author, and if it’s the initial post or a response to the post, are included. The practical implications of the results indicate how social media, particularly Twitter, opens the door for shaming discourse.