2019

2019 SDSU Data Science Symposium Program

South Dakota State University

SDSU Data Science Symposium 2019

South Dakota
State University

February 4-5, 2019
South Dakota State University
Brookings, South Dakota

Sponors

PLATINUM

MetaBank®

CAPITAL SERVICES

GOLD

tci TOTAL CARD, INC.

unify

SANFORD® HEALTH

SILVER

Fishback Financial Corporation

RStudio®

MerchantBoost®

BRONZE

GREAT WEST CASUALTY COMPANY
The Difference is Service®

bluestem brands, inc.

Avera

MAYO CLINIC

Welcome

Dear Attendees,

As President of South Dakota State University, I am happy to welcome you to the 2019 Data Science Symposium. SDSU and its Mathematics and Statistics Department are pleased to offer this opportunity for students, scholars, and practitioners of data science. This is an excellent opportunity to come together to discuss the latest developments in data science and how they might work to the benefit of all in this state, our region, and beyond.

As demonstrated by the range of presenters at this symposium, the breadth of impact of data science is truly remarkable. Business, finance, government, health care, precision ag, biology, economics, journalism, sociology–the list of areas impacted daily by data science is incredibly diverse. I wish you all a great symposium.

Sincerely,

Barry H. Dunn

Barry H. Dunn, President
South Dakota State University

Dear Attendees,

On behalf of the SDSU Department of Mathematics and Statistics, it is my pleasure to welcome you all to the 2019 SDSU Data Science Symposium. With a remarkable array of high-impact speakers, and a broad range of attendees including students, industry representatives, and faculty members, there will be many opportunities for productive discussion and learning that I hope will be of value to us all. My thanks go out to all the speakers and poster presenters, and of course to the organizing committee as well for their great work in making this symposium possible.

Sincerely,

Kurt D. Cogswell

Kurt Cogswell, Department Head
Department of Mathematics and Statistics

Mathematics and Statistics at SDSU

The SDSU Department of Mathematics and Statistics offers students the opportunity to work with outstanding faculty on high-impact research while enrolled in one of our full range of academic programs, including:

- Computational Science and Statistics Ph.D.
- M.S., B.S., and A.S. in Data Science
- M.S. in Statistics
- M.S. in Mathematics (Statistics Specialization available)
- B.S. in Mathematics (Data Science Specialization available)

Faculty conduct theoretical and applied research in diverse areas with a wide range of collaborators, including:

- Applied Mathematics
- Applied Statistics
- Bayesian Statistics
- Bioinformatics
- Computational Statistics
- Data Science
- Education Analytics
- Finance
- Forensic Statistics
- Health Care
- Mathematics Education
- Mathematical Modeling
- Numerical Analysis
- Pattern Recognition
- Precision Agriculture
- Pure Mathematics

All of this occurs in the beautiful new AME Building!

Table of Contents

General Information

Proof of Attendance
If needed, please send a request to **sdsu.seminars@sdstate.edu**

Nametags
Please wear your nametags at all times while accessing conference services or sessions.

Registration/Check-in Hours
Located at the Univesity Student Union in the Volstorff Lounge.
Tuesday, February 5, 7:30 a.m.-5:30 p.m.

Luggage Check Hours
Located at the Univesity Student Union in the Volstorff Lounge.
Tuesday, February 5, 7:30 a.m.-5:30 p.m.

Internet
Complimentary Wi-Fi available for all attendees. Please select "SDSU Guest" as the network. No password required.

Symposium Parking
Attendees are encouraged to park in the Union Pay Lot, which is located to the east of the University Student Union. Parking is complimentary. Attendees must enter the code **5111#** to enter the pay lot.

Transportation

BATA Bus
Visit the Brookings Area Transporation Authority's website for more information about their services:
www.brookingsareatransit.com

Taxi
AAA Cabs LLC, (605) 690-5456
On Demand Taxi Service, (605) 592-6343

Dining
All meals will be at the Union, in Volstorff Ballroom B. Other on campus options are available in the Union. Symposium banquet will be held in the Performing Arts Center.

Health
For emergencies, call 911
University Police Department,
605-688-5117

Avera Medical Group Brookings
400 22nd Avenue South
Monday-Friday: 8:00 a.m.-7:30 p.m.
Sunday: Closed
(605) 697-9500

Sanford Health Brookings Clinic
922 22nd Avenue South
Monday-Friday: 8:00 a.m.-7:30 p.m.
Sunday: Closed
(605) 697-1900

Brookings Health System
300 22nd Avenue South
Open 24 hours
(605) 696-9000

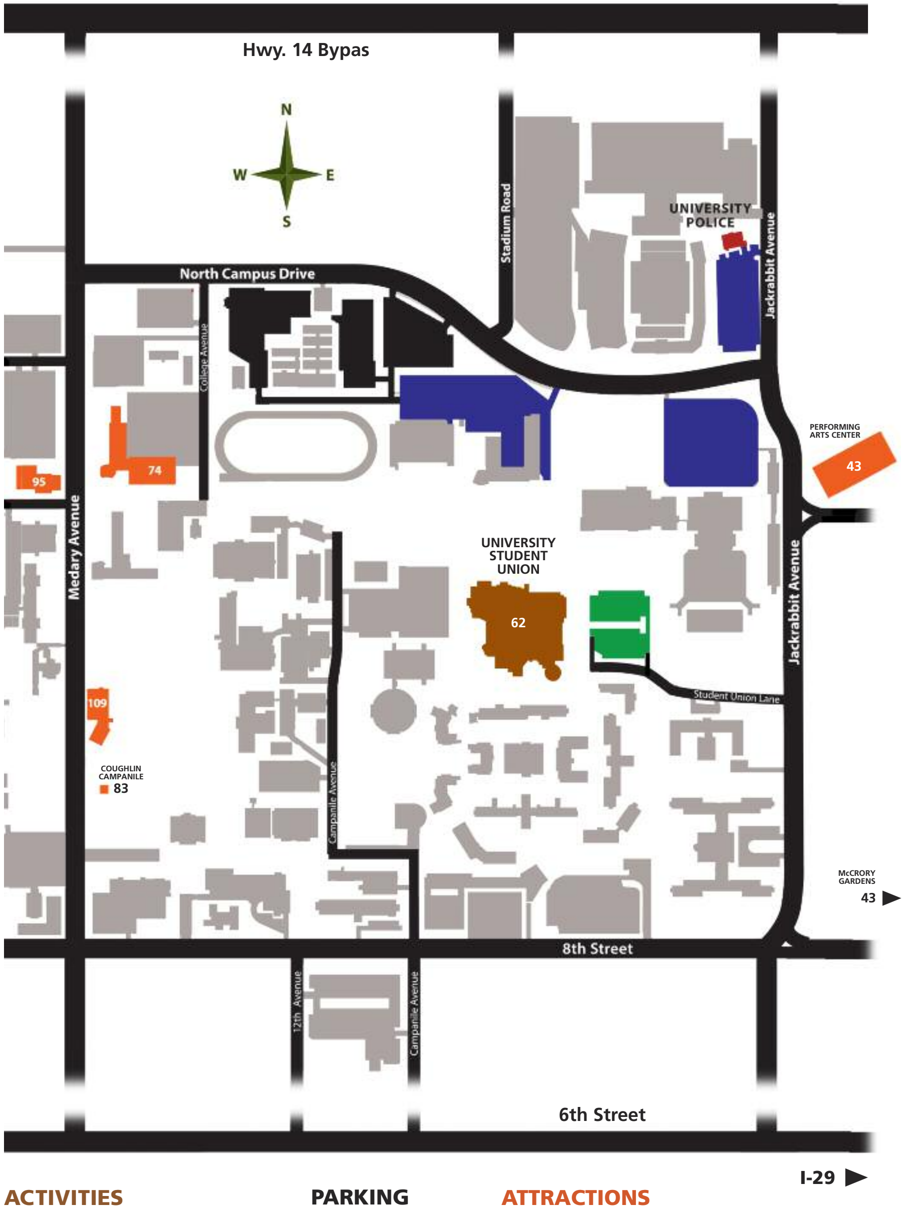
Questions
For any questions or concerns, please email **sdsu.seminars@sdstate.edu** or visit the Registration/Check-in table in the Volstorff Lounge during our hours.

SDSU Bookstore
Data Science attendees receive a 20% discount on their purchases. Please show your Data Science ID badge.
Monday-Friday: 8:00 a.m.-6:00 p.m.
jackrabbitcentral.com

Printing Services
BluePrint Design & Print
Union Lower Level
Monday-Friday: 8:00 a.m.-5:00 p.m.
(605) 688-5496

ATM
Union

SDSU Campus Map



ACTIVITIES

22 Performing Arts Center
62 University Student Union

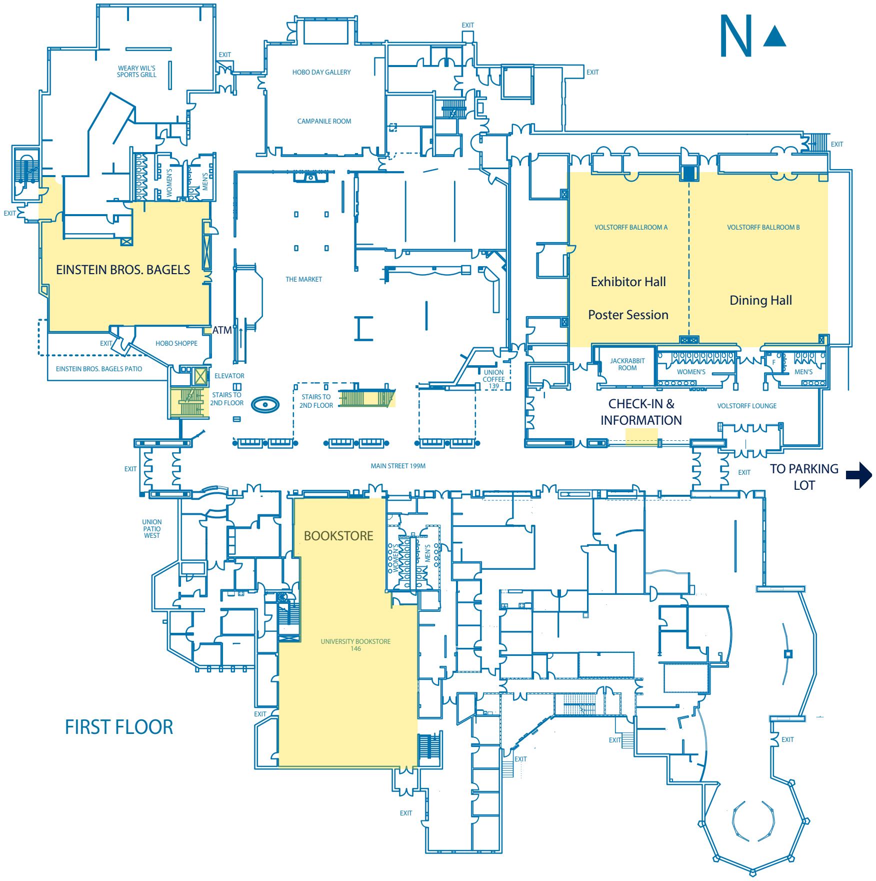
PARKING

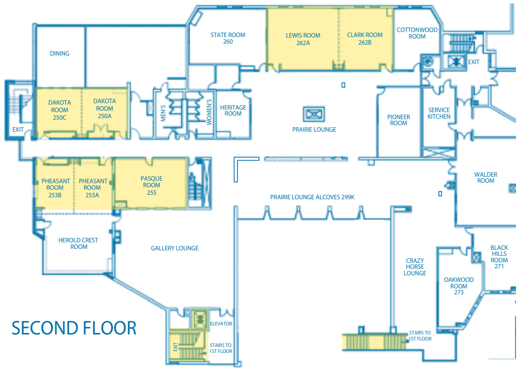
🟩 Union Pay Lot
🟦 Parking

ATTRACTIONS

43 Performing Arts Center
74 Dairy Bar
83 Coughlin Campanile
95 Agricultural Heritage Museum
109 South Dakota Art Museum

Union Map





Monday, February 4, 2019		
Time	**Pasque Room (255)**	**Dakota Room A/C (250)**
1:00-5:00 p.m.	**Check-in/Registartion, Volstorff Lounge**	
1:00-2:45 p.m.	**WORKSHOP | Tidyverse** ***Tidyverse: R packages for data science*** Dr. Adam Sullivan	**WORKSHOP | Deep Learning** ***Deep Learning with Python*** Dr. David Zeng
3:00-5:00 p.m.		
1:00-5:00 p.m.	**Job Fair/Recruiting | Exhibitors, Campanile Room & Hobo Day Gallery A&B**	
	Banquet, Performing Arts Center	
6:00-6:30 p.m.	**Social Time (cash bar)**	
6:30-8:00 p.m.	**Dinner**	
7:15-7:30 p.m.	**Welcome**	
7:30-8:30 p.m.	**Keynote: *Date and Analytics — What you don't know can HELP you!*, Kate Bischoff**	

Tuesday, February 5, 2019			
Time	**Pasque Room (255)**	**Pheasant Room A/B (253)**	**Dakota Room A/C (250)**
7:30 a.m.-noon	**Check-in/Luggage Check, Volstorff Lounge**		
7:45-8:30 a.m.	**Breakfast, Volstorff B**		
8:30-8:40 a.m.	**Opening Session: *Welcome and Introduction*, Vice President Daniel Scholl, Volstorff B**		
8:45-9:45 a.m.	**Keynote: *The Next Great Era: The Age of Artifical Intelligence*, Clay Campbell, Unity Consulting, Volstorff B**		
!0:00-10:50 a.m.	**SESSION 1 | Tools** ***An Exploration of Mulitple Tools for Creating Reproducuble Research*** Adam Sullivan, Brown University ***minque: An R Package for Analyzing Various Linear Mixed Models*** Jixiang Wu, SDSU	**SESSION 2 | Healthcare** ***Applicaion of Transfer Learning Techniques for Medical Image Classification*** James Bolt, DSU ***Identification of Potential Markets– The Sanford Opportunity Score*** Clark Casarella, Sanford Health	**SESSION 3 | Finance** ***FICO$^{®}$Scores Through the Economic Cycle and Understanding Consumer Sensitivities to Economic Fluctuations*** Gerald Fahner, FICO
10:50-11:00 a.m.	**Networking Brreak | Exhibitors, Volstorff A**		
11:00 a.m.-noon	**SESSION 4 | Tools** ***Approximate Bayesian decision-making with complex data: analysis of forensic fingerprint data*** Cedric Neumann, SDSU	**SESSION 5 | Healthcare** ***Leveraging data for the who, what, why, and how of population health*** Emily Griese & Doug Nowak, Sanford Health ***Predicting unplanned medical visits among patients with diabetes using machine learning*** Arielle Selya, Sanford Health	**SESSION 6 | Other** ***Recommendation systems and Personalization Methods and Use Cases*** Ally Pelletier, Star Tribune ***Application of Data Science in Churn Analysis*** Deepak Sanjel, MNSU
noon-1:00 p.m.	**Lunch, Volstorff B | Poster Session, Volstorff A**		
1:00-1:50 p.m.	**SESSION 7 | Methods** ***Multi-Linear Algebraic Eigen decompositions and Their Applications in Data Science*** Randy Hoover, SDSMT ***Dimension Reducation for Big Data*** Hossein Moradi, SDSU	**SESSION 8 | Healthcare** ***Challenges in modeling cognition and demention*** Terry Therneau, Mayo Clinic	**SESSION 9 | Finance** ***Effective Model Validation*** Lance Cundy, Fed Minneapolis ***Data Science Ops: Working Excellently*** Thomas Cleberg, Mutual of Omaha
2:00-3:00 p.m.	**Poster Session | Student Poster Competition, Volstorff B**		
2:00-3:25 p.m.	**Job Fair/Recruiting | Exhibitors, Volstorff B**		
3:30-4:30 p.m.	**SESSION 10 | Methods** ***A survey of parametrical modeling under Progressive type-I Interval Censoring*** Yuhlong Lio, USD ***Neural Shrubs: Using neural networks to improve decision trees*** Kyle Caudie, SDSMT	**SESSION 11 | Other** ***Survival Analysis Methods to Predict Loss Rates in Credit Card Portfolios*** Landon Thompson, United Health ***The Fit of Data Visualization Design for Knowledge Activation*** Kari Sandouka, DSU	**SESSION 12 | Finance** ***Dos and Don'ts of Data Science*** Ryan Burton,Capital Services ***Data Use in the Measurement of Systemic Risk in Financial Systems*** Katherine Kime, UNK
4:30-5:00 p.m.	**Closing Session: Thomas Brandenberger, Volstorff B | Poster Winners Announced**		

Biographies

Keynote Speakers

Kate Bischoff, tHRive Law & Consulting LLC

Kate Bischoff is an overly enthusiastic, sarcastic, and opinionated management-side employment attorney and SHRM-SCP-certified HR pro. She works closely with management, HR folk, and technology companies to improve organizations and make it easier to recruit and retain talent through having easy-to-understand policies, easy-to-use technology, and easy-to-explain compliance initiatives. Kate has been recognized by The New York Times, CNN.com, Wall Street Journal, USA Today, National Public Radio, and other journalistic sources as a leading authority on harassment, technology in the workplace, and employment law.

Clay Campbell, Unify Consulting

Clay Campbell is the director of Innovation and Applied A.I. for Unify Consulting in Seattle, Wash. A Madison, S.D., native, Clay has spent over a decade working with some of the largest companies on the planet solving some of their hardest problems. Most notably he has established himself as a thought leader in the field of Applied A.I. He spends his free time building self-driving cars and treasure hunting.

Invited Speakers

Ryan Burton, Capital Services Inc.

Ryan is the portfolio analytics and risk director at Capital Services where he is responsible for managing analytical applications throughout the company including reporting, modeling, and strategy development. He received his master's in statistics from South Dakota State University researching trended data's impact on credit scoring. He is motivated by making smarter decisions using analytics to improve outcomes for all parties impacted.

Thomas Cleberg, Mutual of Omaha

Thomas Cleborg is a principal data scientist at Mutual of Omaha and a graduate of SDSU and Dakota State University. In addition to developments in reinforcement learning, neural architectures and applied machine learning, Thomas has a strong interest in the responsible application of algorithms, model lifecycle management systems and practices, and interpretability of machine learning models.

Gerald Fahner, FICO

Dr. Gerald Fahner is senior principal scientist in FICO's scores division. He specializes on innovative algorithms that turn data and domain knowledge into superior insights, predictions, and decisions. Gerald is also responsible for the core algorithms underlying FICO's Scorecard development platform. His work on causal modelling won the best paper award at the Credit Scoring and Credit Control XI conference. Prior to joining FICO in 1996, he served as a researcher in artificial intelligence, neural networks and robotics at the International Computer Science Institute in Berkeley, and earned a computer science doctorate from University of Bonn.

Emily Griese, Sanford Health

Emily Griese, Ph.D., is the director of collaborative research at Sanford Health. In this role, she works across Sanford's research, enterprise data and analytics, and quality arms to support coordinated population health strategies throughout the enterprise. She worked to establish and currently directs the Sanford Data Collaborative, a first of its kind data sharing initiative with academic partners to innovate and improve the way healthcare is provided to the patients and communities Sanford serves.

Dr. Griese is a NIH-funded scientist in the Population Health Research Group at Sanford Research and an assistant professor of pediatrics at the University of South Dakota Sanford School of Medicine. She received her Ph.D. in psychological research from the University of Nebraska-Lincoln and completed her postdoctoral fellowship in population health at Sanford Health. Her research focuses on social determinants of health and their impact on health trajectories over time, working specifically with rural communities.

Hossein Moradi, South Dakota State University

Hossein Moradi is an assistant professor of statistics at SDSU. His research interest is the development of statistical models for spatial and spatiotemporal data, functional data, and multivariate data for making inferences from massive datasets, all motivated by the desire to tackle scientific problems. Hossein was a member of Research Network for Statistical Methods for Atmospheric and Oceanic Sciences as postdoctoral fellow. He earned a Ph.D. in system modeling and analysis and a master's in operations research at Virginia Commonwealth University, a master's in mathematical statistics from Tarbiat Modares University and a bachelor's in statistics from University of Mazandaran.

Cedric Neumann, South Dakota State University

Cedric Neumann is an associate professor of statistics at the SDSU. Cedric's main area of research focuses on the statistical interpretation of forensic evidence, more specifically fingerprint, shoeprint and traces. Prior to joining SDSU, Cedric worked for the Forensic Science Service (FSS) in the United Kingdom. As head of the R&D Statistics and Interpretation Research Group, he contributed to the development of the first validated fingerprint statistical model. This model was used to support the admissibility of fingerprint evidence in U.S. courts. Cedric has a Ph.D. in forensic science from the University of Lausanne (Switzerland). He received several awards for his implementation of his thesis work at the United States Secret Service.

Ally Pelletier, Star Tribune

Ally is a data scientist for the Star Tribune in Minneapolis where she is involved in reporting and modeling for the digital department. Her current work includes recommendation models, customer retention, and map development. Previously, Ally worked as a data science consultant with RProfet where she was deeply involved in all aspects of the modeling process. She is a subject matter expert in credit modeling as well as the development of the regular reporting processes and documentation necessary to create data driven decisions. Ally earned a B.A. in mathematics education from Concordia College in Moorhead, Minn., and an M.S. in statistics from SDSU. She held a research assistantship and an internship in digital media at SDSU. In her research she developed new statistical power calculations for measuring mixtures of non-normal distributions to measure the profitability in A/B testing in credit card customer behavior.

Honghao Shan, Experian
Terry Therneau, Mayo Clinic

As a statistician engaged in clinical research programs, my interests reflect both medical and statistical areas. The former has been focused for the last several years on liver disease, liver transplant, hematology with particular emphasis on plasma cell malignancy, and physical medicine. Statistically, the major impetus of my work has been in survival analysis. Currently research includes: 1) correlated random effects models, applied to large family based genetics studies; 2) the analysis and understanding of microarray and proteomics data ; 3) expected survival computation and competing risks; and, 4) the application of tree based methods to survival data. https://www.mayo.edu/research/faculty/ therneau-terry-m-ph-d/bio-00025991

Workshops

Adam Sullivan, Brown University

Adam Sullivan is an assistant professor of biostatistics at Brown University. His interests include flipped/blended learning, online learning, open education, R programming and Statistics Education. He is currently director of the master’s program in biostatistics at Brown University and the faculty statistician at the Hassenfeld Child Health Innovation Institute. His interests include statistics education, online/blended learning, and pediatric research. During his time at Brown he has created numerous courses and has won three teaching awards for this work.

Prior to Brown, Sullivan was a key contributor to the creation of Harvard’s T.H. Chan School of Public Health’s first blended course which also integrated biostatistics and epidemiology. He received his Ph.D. in biostatistics from Harvard University in 2015, his master/s in mathematics from South Dakota State University in 2010 and a bachelor/s in mathematics and secondary education from Houghton College in 2003. He has worked as an educator in all levels from high-school mathematics to graduate-level biostatistics.

David Zeng, Dakota State University

Dr. David Zeng is an assistant professor of information systems at Dakota State University. He teaches graduate courses in the Master of Science in Data Science Program jointly offered by DSU and SDSU. The courses he teaches include predictive analytics for decision making, programming for data analytics (Python), deep learning, and BI & visualization. His research focuses on the economics of IT-enabled services, application of Deep Learning (transfer learning) in healthcare, and generative neural networks in contests and games. His papers have been published in top peer-reviewed journals and awarded as best papers in conferences. David received his Ph.D. in information systems from University of California, Irvine. He earned a M.S. in computer science from California State University, Long Beach.

Keynote Speakers

Clay Campbell, Unify Consulting
The Next Great Era: The Age of Artificial Intelligence

Everybody wants to live in the future, what if I told you it was already here. Artificial Intelligence will change the standards of how we live, interact with each other and survive the planet. The advances we have seen in the past couple of years- self driving cars, robot assistance, speech recognition and computers that can win Jeopardy!, are not the crowning achievements of the AI age. They are just the warm-up act. As we enter a new machine age we will see wonder beyond our belief. Complex machines that can process natural language, computers that can automatically refine their methods and improve over time, computer vision and real time prediction are already here and integrated into our daily lives. We’re going to see AI do more and more faster and faster, for some of us our lives will get better, for others they will go through substantial and altering life and cultural changes.

Kate Bischoff, tHRive Law & Consulting LLC
Slowing Down the Analytics Train

Data scientists have come very far in making business and life in general. Organizations have improved processes, done better market targeting, and increased profitability. Netflix makes recommendations, Amazon knows when we run out of laundry detergent, and Google translates. Yet, analytics can do some pretty scary stuff too, like classifying African Americans as animals or discriminate against women. So, how should we get the benefits of analytics while reducing the risk?

Invited Speakers

Ryan Burton, Capital Services

Dos and Don'ts of Data Science

In an ideal world, we avoid mistakes in our work. Some are preventable and others are unavoidable. Common mistakes in data science that can be minimized include assuming correlation implies causation, modeling with an unrepresentative sample, and focusing on the mean without understanding the distribution. Ryan Burton will give an overview of simple yet common mistakes in data science and guidance on how to avoid them.

Thomas Cleberg, Mutual of Omaha

Data Science Ops: Working Excellently

Teams producing analytical models have many and varied responsibilities in the organizations in which they operate. Creating an environment in which their work is sharable, collaborative, monitored, transparent, responsible, useful and correct is a challenge but possible. In this talk, we will discuss a general framework for creating a model management framework that addresses these challenges, open source projects that seek to enable data science teams to work excellently, and an example of how they can be combined modularly to create an environment that meets the specific needs of a team and organization.

Gerald Fahner, FICO

FICO® Scores Through the Economic Cycle and Understanding Consumer Sensitivities to Economic Fluctuations

FICO® Scores assess default likelihood under "normal" conditions. Additional risk comes from economic variabilty and manifests itself in changes to the score distribution as well as changes to the Odds-to-Score relationaship. We examine credit bureau data from the Great Recession in the U.S. as compared to a stable period. We'll share a framework that provides new and actionable insights into consumer sensitivity to such economic fluctuations such as:

- Application of counter factual analysis, machine learning and scorecards to rank-order consumer's sensitivities
- Consumer segmentation by economic sensitivites
- Sensitivity profiling reveals interesting differences between the most and the least sensitive consumers
- How you can "economic shock absorbers" under your lenidng stragies
- Improving the accuracy of portfolio imulations varying economic conditions.

Emily Griece, Sanford Health
Doug Nowak

Leveraging data for the who, what, why, and how of population health

While healthcare has some of largest and more diverse data, moving from data insights into execution can be difficult in its often fast-paced, shifting environment. This talk will focus on the crucial role data and analytics have in helping to define and execute on the often ill-defined concept of population health.

Hossein Moradi, South Dakota State University

Dimension Reduction for Big Data

In many research areas, such as health science, environmental sciences, agricultural sciences, etc., it is common to observe data with huge volume. These data could be correlated through space and/or time. Studying the relationship for such complex data calls for a fairly advanced modeling techniques. Reducing the dimensionality of the data in both covariates and response space can help researcher to better handle the computational cost.

Cedric Neumann, South Dakota State University

Approximate Bayesian decision-making with complex data: analysis of forensic fingerprint data

Bayesian inference allow us to use information contained in a dataset to update a prior belief about some parameter of interest (e.g., a population mean) and make some inferences about the value of the parameter. The result allows us to quantify the uncertainty about the value of the parameter in a more logical and coherent way than traditional frequentist techniques. Unfortunately, standard Bayesian methods cannot be applied in all scenarios. This is the case for many scenarios that require unreasonably complex models to describe the data and where the corresponding likelihood function cannot be derived.

A class of methods, called Approximate Bayesian Computation, allows for approximate Bayesian inference to be performed in these scenarios. ABC methods are simulation based and allow for coherent decision making. ABC methods can be useful to analyze the results of experiments from a wide range of disciplines (animal science, plant science, healthcare, finance) where the data may be unbalanced, high-dimensional, or encapsulate many different variable types. Examples of the application of ABC forensic evidence will be provided.

Ally Pelletier, Star Tribune

Recommendation Systems and Personalization Methods and Use Cases

Recommender systems work all around us to help customers find products and content they may find valuable. In this session, first we will talk through recommendation systems from the most simple to the more advanced and their use cases. After that we will go over a few of the algorithms and the details of how they are implemented and used in media.

Terry Therneau, Mayo Clinic

Challenges in modeling cognition and dementia

The Mayo Clinic Study of Aging has collected longitudinal data from and age/sex stratified sample of Olmsted County residents over the age of 60, with a goal to better understand the natural course and determinants of cognitive decline, mild cognitive impairment and dementia with age. The study purposefully oversamples those at the older ages, and over the course of 10+ years has accumulated date from 22443 visits of 5425 unique residents. Yet even with this large data set answering many of the questions remains challenging simply because of the long time scale.

We will look the simple question of how biomarkers of amyloid and tau evolve over time, both at the population and the patient level, and some of the statistical challenges that this represents. On autopsy, amyloid plaques and tau fibrils are the hallmarks of Alzheimer's disease. An underlying physical hypothesis (not universally accepted) is that amyloid precedes tau which precedes dementia. What does the data have to say about this?

Accepted Oral Presentations

James Boit, Dakota State University
David Zeng

Application of Transfer Learning Techniques for Medical Image Classification

Recently, the healthcare industry is in a dynamic transformation accelerated by the availability of new artificial intelligence, machine learning, and deep learning (DL) technologies, tools and strategies facilitated by powerful graphical processing unit computing. The deployment of DL models in healthcare organizations for medical image analysis, is creating healthcare use cases, for example, increasing timeliness and accuracy of diagnosis thus improving healthcare outcomes and enhancing better medical decisions. Transfer learning (TL) techniques leverages DL algorithmic architecture to perform medical image analysis and classification which reduces physician's workload, decreases time and costs for interpretation and thus helping physicians to focus more on improving patient care. In this research, we apply TL techniques in medical image classification tasks, namely feature extraction and fine-tuning strategies. We investigate the effectiveness of TL techniques on medical image classification using the Chest X-ray dataset. For our DL model, we used DenseNet-121 architecture, a deep convolutional neural network (DCNN) comprised of 121-layers, as the baseline model to perform a binary classification on the medical images. Applying fine-tuning strategy, freeze-unfreeze method of DCNN top layers and with data augmentation is an effective technique to mitigate overfitting and improve model performance. Using the fine-tuning strategy, we determined improved model performance at Conv5_block16_2_conv as the Optimal Cut-Off Layer. This research will extend its focus on additional fine-tuning approaches, such as hyperparameter optimization, finding optimal data augmentation and generating high resolution medical images using generative adversarial networks to determine the optimal behavior of an effective TL for medical image classification.

Clark Casarella, Sanford Health

Identification of Potential Markets – The Sanford Opportunity Score

The Sanford Opportunity Score (or SOS) is a county-level assessment and scoring of inpatient, outpatient, and clinic encounters originating from the county inside the Sanford Health footprint. The ultimate goal of this score is to aid strategic planning at Sanford Health in the rapid identification of key areas where Sanford has a large potential gain in patient volume and care delivery. This phenomenological model and score takes into account and places varying weights on several quantities, namely, the volume of inpatient encounters, the current and historical market share, the density of physicians, the potential to transform clinic visits to inpatient or outpatient encounters, the payor mix and average reimbursement rate, and the bulk demographics of the market. The SOS has been utilized for two years by the office of Strategic Planning as the cornerstone of the Strategic Briefcase for use across 10+ service lines within Sanford Health.

Kyle Caudle, South Dakota School of Mines
Randy Hoover, Karen Braman

Neural Shrubs: Using neural networks to improve decision trees

Decision trees are a method commonly used in machine learning to either predict a categorical response or a continuous response variable. Once the tree partitions the space, the response is either determined by the majority vote – classification trees, or by averaging the response values – regression trees. This research builds a standard regression tree and then instead of averaging the responses, we train a neural network to determine the response value. We have found that our approach typically increases the predicative capability of the decision tree. We have two demonstrations of this approach that we wish to present as a poster at the SDSU Data Symposium.

Lance D. Cundy, Federal Reserve Bank of Minneapolis

Effective Model Validation

Model validation is a critical step in use of statistical and economic models. It is the set of processes and activities intended to verify that models are performing as expected, in line with their design objectives and business uses. In this session, we will discuss validation as it relates to financial and economic models. An effective validation framework should include three core elements: (i) evaluation of conceptual soundness, including developmental evidence, (ii) ongoing monitoring, including process verification and benchmarking, and (iii) outcomes analysis, including benchmarking. We will explore each of these elements and discuss statistical tools to support validation efforts.

Randy Hoover, South Dakota School of Mines

Multi-Linear Algebraic Eigendecompositions and Their Application in Data Science

Multi-dimensional data analysis has seen increased interest in recent years. With more and more data arriving as 2-dimensional arrays (images) as opposed to 1-dimensioanl arrays (signals), new methods for dimensionality reduction, data analysis, and machine learning have been pursued. Most notably have been the Canonical Decompositions/Parallel Factors (commonly referred to as CP) and Tucker decompositions (commonly regarded as a high order SVD: HOSVD). In the current research we present an alternate method for computing singular value and eigenvalue decompositions on multi-way data through an algebra of circulants and illustrate their application to two well-known machine learning methods: Multi-Linear Principal Component Analysis (MPCA) and Mulit-Linear Discriminant Analysis (MLDA).

Katherine Kime, University of Nebraska Kearney

Data Use in the Measurement of Systemic Risk in Financial Systems

The financial crisis of 2008 led to an increase in research on the subject of systemic risk in various financial systems—broadly speaking, the possibility that losses experienced by one participant in the system would start a cascade of losses in part or all of the rest of the system. While many papers are theoretical in nature and do not use actual data sets, others do. In this talk we will give a partial survey of empirical studies and the data sets used, for example Furfine's use of payment flow data from the Federal Reserve's large-value transfer system, Fedwlre, and the use of the Mexican interbank exposures market and the Mexican Large Value Payments System by Martinez-Jaramillo et al. We will discuss proposed systemic risk measures, by Bisias et al, and by the Federal Reserve.

Y. L. Lio, University of South Dakota

Survey on parametrical modeling under Progressive type-I interval Censoring

Progressive type-I interval censoring scheme has been applied not only to industry reliability but also to medical clinical studies for accommodating the possible subject loss and broken before lifetime expires. In this talk, current methodologies developed will be presented through real data applications.

Kari Sandouka, Dakota State University
Cherie Noteboom

The Fit of Data Visualization Design for Knowledge Activation

Big data provides insight into products, services, business processes and management control activities. The ability to leverage data for actionable knowledge is the key to gaining and maintaining a competitive advantage. More is not always better; as big data is not easily processed manually, there is an increased reliance on computational tools and technologies to extend, partner, supplement and support human cognitive abilities. The combination of human intelligence with technological capabilities is the answer to make smarter decisions in the face of uncertainty and complexity. As the volume of data increases so does the need for new and improved analytical tools facilitating interactions between humans and technology (Bumblauskas, Nold, Bumblauskas, & Igou, 2017; Davenport et al., 2012). The power of visualization comes from coupling soft system attributes (perceptive skills, cognitive reasoning, and domain knowledge) with hard system attributes (computing and data storage). The design of visualizations is often separated from the user, resulting in misinterpretation and leading to error-prone decision-making (Few, 2006). Obtaining a better understanding of what it means for a human to see and think will help in developing more effective knowledge management tools for handling big data.

This research proposes a model for determining the fit of data visualizations grounded in the theories of situational awareness and task-technology fit. The model identifies design attributes that facilitate the task of sensemaking as a mediating factor to knowledge activation. Knowledge activation will only occur if a fit exists between task, technology, and individual.

Deepak Sanjel, Minnesota State University Mankato

Application of Data Science in Churn Analysis

Several subscription-based companies, such as cable TV, internet providers, cell phone etc. will face customer churn despite their best efforts to prevent it, and when it happens, Data Science can be used to dig into what led the customer to leave, and what can be done to prevent a similar customer churning in the future. A case study with customer retention data and commonly used Data Science methods will be discussed.

Arielle Selya, Sanford Health
Eric L. Johnson, University of North Dakota

Predicting unplanned medical visits among patients with diabetes using machine learning

Diabetes poses a variety of medical complications to patients, resulting in a high rate of unplanned medical visits, which are costly to patients and healthcare providers alike. However, unplanned medical visits by their nature are very difficult to predict. The current project draws upon electronic health records (EMR's) of adult patients with diabetes who received care at Sanford Health between 2014 and 2017. Various machine learning methods were used to predict which patients have had an unplanned medical visit based on a variety of EMR variables (age, BMI, blood pressure, # of prescriptions, # of diagnoses on problem list, A1C, HDL, LDL, and a ranked variable for tobacco use severity). A radial-basis support vector machine (SVM) was the most accurate method, achieving a hit rate of 68.5% and a correct rejection rate of 62.9% during cross-validation testing. Follow-up testing of the trained SVM indicated that, of the modifiable prediction variables, high blood pressure and low levels of high-density lipoprotein (HDL) were most strongly predictive of unplanned medical visits. Future directions include refining and validating the predictive model, towards the ultimate goal of developing and implementing clinical recommendations for preventing unplanned medical visits among adult patients with diabetes.

Adam Sullivan, Brown University

An Exploration of Multiple Tools for Creating Reproducible Research

In a great deal of statistical methods we employ the Central Limit Theorem which we assume to be true if we replicated our research over and over again. In the traditional research world, this assumption was made but rarely was there a way to truly replicate work.

The world of data science changes this paradigm. All areas of work from business to health care collect more types of data on a more regular basis. This should lend us the ability to reproduce results and test our assumptions. How do we adequately set up a data workflow to allow for things to be more reproducible? How do we encourage sharing of findings and code so that others in our fields or companies can replicate and reproduce our work?

This presentation will discuss many strategies to creating reproducible research, both from the lense of a researcher preparing their data workflow to allow them to reproduce but also tools for sharing this research to allow others to reproduce it as well.

We will look further into tools like RStudio, RCloud Social, Jupyterhub and Github. All provide platforms for sharing and collaborating on projects while making reproducible research easy to employ.

Landon Thompson, South Dakota State University/United Health
Thomas Brandenburger, Capital Card Services

Survival Analysis Methods to Predict Loss Rates in Credit Card Portfolios

Understanding risky behavior associated with a given cardholder is crucial in managing a successful portfolio in the credit lending industry. Measuring this risk is done at many different time points in the customer lifecycle. This paper explores constructing a model to accurately predict the risk of a cardholder for the lifetime of the account. First, a brief introduction to the customer lifecycle will take place. Next, the structure of the data and the issues associated with it will be discussed. After this, Survival Analysis methodologies will be overviewed and applied to a case study from Capital Card Services. Next, an extension of the single event framework to account for multiple events will then be discussed. Finally, an application of these models to address problems that arise in the credit card industry will be examined.

Jixiang Wu, South Dakota State University

minque: An R Package for Analyzing Various Linear Mixed Models

Linear mixed model (LMM) approaches offer much more flexibility comparing ANOVA (analysis of variance) based methods. There are three commonly used LMM approaches: maximum likelihood, restricted maximum likelihood, and minimum norm quadratic unbiased estimation. These three approaches, however, sometimes could also lead low testing power compared to ANOVA methods. Integration of resampling techniques like jackknife could help improve testing power based on both our simulation studies. In this presentation, I will introduce a R package, minque, which integrates LMM approaches and resampling techniques and demonstrate the use of this packages in various linear mixed model analyses.

Accepted Poster Presentations

Abdelbaset Abdalla, South Dakota State University
Semhar Michael

Finite Mixture Regression Models for Stratified Sampling Design

Despite the popularity and importance, there is limited work on modeling data, which come from complex survey design using finite mixture models. In this work, we explored the use of finite mixture regression models when the samples were drawn using a complex survey design. In particular, we considered modeling data collected based on stratified sampling design. We developed a new design-based inference where we integrated sampling weights in the complete-data log-likelihood function. The expectation-maximization algorithm was developed accordingly. A simulation study was conducted to compare the new methodology with the usual finite mixture of a regression model. The comparison was done using bias-variance components of mean square error. Additionally, a simulation study was conducted to assess the ability of the Bayesian information criterion to select the optimal number of components under the weighted modeling approach. The methodology was implemented on real data with good results.

Girma Ayana, South Dakota State University
Melanie Caffe

Genomic selection for grain yield and quality traits in oat (Avena sativa L)

Genomic selection is the process by which the genetic improvement of plant is accomplished using marker based genomic prediction (GP) of value of an individual as a genetic parent. Traditionally, it has required a series of selection for 10 years or more to release improved seeds. GS has potential of reducing the years of breeding through prediction of progenies performance and saves resources. We used a separate panel of 222 oat (Avena sativa L) lines genotyped with 38,000 SNP markers for three generations. Genomic selection was applied to over seven phenotypic traits in the oat breeding program of South Dakota State University at four locations for three years. GS prediction accuracy, correlation between observed and model prediction, from cross-validation approach were compared using six GS models such as RRBLUP, GAUSS, PLSR, ELNET, RF, and AVEWe found that the AVE method was giving better prediction with average accuracy of 0.25, 050, 0.56, 0.66, 59, and 0.48 for yield, protein content, plump groat, groat oil content, plant height, and groat beta glucan content, respectively. Overall, all six GS models appear to be applicable for predicting quality traits, but we recommend AVG method for quantitative traits like yield.

Audrey Bunge, South Dakota State University
Thomas Brandenburger

Cluster Analysis of Spotify Users

Spotify is considered one of the best music streaming providers in the world. Spotify users can access different information regarding not only tracks, albums, and artists, but also personal listening habits by using a lesser known Spotify feature, the Web Application Programming Interface. The personal listening habits obtained include the top 50 artists of all time, and the top 50 tracks of all time. Market basket analysis is used to condense the genres from the top 50 artists per user. Following the use of each user's top 50 artists and with some data manipulation, we conduct K-means cluster analysis on the 50 top tracks of all time for each user. Once clusters are obtained, we can identify similar listening habits among users.

Kevin R. Callies, Dakota State University
Cherie Noteboom

Employee Acceptance of Employer Control Over BYOD Devices

Organizations face new and growing security challenges as consumer technology continues to be integrated into organizational workflows. Bring your own device (BYOD) is a phenomenon that is here to stay; however, securing employee's personally owned devices may require the organizations to consider exerting some control over the employee's device. In order for organizations to secure access to their sensitive information in this way, they must first garner the employee's consent. This research seeks to model employee acceptance of employer control by constructing a model of employee acceptance based upon the extant acceptance literature. Once a model has been created it is operationalized into several measures and verified using a quantitative survey methodology.

Krishna Ghimire, South Dakota State University
Melanie Caffe-Treml

Evaluation of root system architecture (RSA) of oat (Avena sativa L.) genotypes

Root systems play an essential role in plant by allowing the plant to absorb water and nutrient from the soil. Root system architecture (RSA) describes the spatial arrangement of a root system in the soil, the overall shape and structure of the root system. The objective of this study was to evaluate and compare the root system architecture of a set of 12 oat genotypes. Oat seed were planted in plastic pots filled with 1:1 mixture of garden soil and sand. The experiment was conducted in the greenhouse with a photoperiod of 16 hours of light and a temperature between 21 and 24 °C. The experiment followed a randomized complete block design with 4 replications. Thirty-five days after planting, the roots were removed from the pots, cleaned and photographed. Fifteen root traits were evaluated form each of the image using an online tool, Digital Imaging of Root Traits (DIRT). Significant differences among genotypes were observed for various root traits including projected root area, average root density, root tip count, accumulated width over 10% and 50% depth. Ideal root traits include roots with increased depth of root system, increased deep branching, steep root angle, small root diameter and higher root surface area, as it helps with the nutrient and water absorption from wider and deeper soil profile. Genotypes like Hayden, SD120665, Horsepower and Natty showed desirable root traits and these can be used as genetic resources to develop roots with deeper distribution which has been associated with increased yield potential.

Austin Hanson, South Dakota State University

Stability in Anomaly Detection for Keystroke Dynamics: Exploring the Possibility with Plateau Regression of an Individual's Keystroke Dynamics Changing Over Time

Anomaly detection methods of keystroke dynamics have been proposed for enhanced biometric security for passwords. This study was proposed to confirm the underlying assumptions of these anomaly detectors and see if an individual's keystroke dynamics changing over time, which would skew the training of anomaly detection algorithms and lead to low adequacy. The data used for this study was from a public benchmark data set from Killouhy and Maxion. [5] The data consisted of 51 subjects from the campus of Carnegie Mellon, that completed 8 data-collection sessions completing 50 password repetitions that had to be the password typed correctly. The data was analyzed on the session level for the total time to type the password, as it best describes the entire keystroke dynamics for each individual. Summary statistics had to be derived for each session and median regression by quantile regression was implemented across session using the conditional median across the session or a linear plateau method to derive where subjects became stationary in sessions. The summary statistic was then used in a linear mixed-effect model to test for significance in the slope of the session. It was found that session has a significant decrease in slope and thus there is evidence to show an individual's typing dynamics change over time. Recommendations on how to proceed with the anomaly detector are to explore the idea of implementing adaptive training techniques for the anomaly detector.

Md Riaz Ahmed Khan, South Dakota State University

ROCit- An R package for performance assessment of binary classifier with visualization

Sensitivity (or recall or true positive rate), false positive rate, specificity, precision (or positive predictive value), negative predictive value, misclassification rate, accuracy, F-score- these are popular metrics for assessing performance of binary classifier for specific threshold. Receiver operating characteristic (ROC) curve is a common tool for assessing overall diagnostic ability of the binary classifier. ROCit is an R package that provides flexibility to easily evaluate a binary classifier. ROC curve, along with area under curve (AUC) can be obtained using different methods, such as empirical, binormal and non-parametric. ROCit encompasses a wide variety of methods for constructing confidence interval of ROC curve and AUC. ROCit also features the option of constructing empirical gains table, which is a handy tool for direct marketing. The package offers options for commonly used visualization, such as, ROC curve, KS plot, lift plot. Along with in-built default graphics setting, there are rooms for manual tweak by providing the necessary values as function arguments. ROCit is a powerful tool offering a range of things, yet it is very easy to use.

Kyle Lifferth, South Dakota State University

Predicting charge off using text analysis

This research attempts to predict charge off for loans based on an open text box filled out by the loan applicant to encourage investors to invest in their loans. To predict charge off, latent sematic analysis is used with classification tools, techniques such as random forest, singular value decomposition, and tokenization, and TF-IDF. Preliminary exploratory data analysis shows that text length does not seem to have a strong association with charge off.

Tareq Nasralah, Dakota State University
Omar El-Gayar, Yong Wang

Mining Twitter to Assess the Perceptions of the Opioid Crisis

Opioid addiction has become one of the largest and deadliest epidemics in the United States. Opioids are a group of drugs that include illegal heroin and strong pain relievers by legal prescription, such as morphine and oxycodone. In this research, we plan to study opioids' epidemic by analyzing online social media communities. Understanding the public perceptions and the addicted users' experience of opioid addiction and misuse can help get insights to provide better opioids prevention, treatment, and recovery strategies. In this research, we plan to study opioids' epidemic by analyzing recent tweets data for users who are addicted or non-addicted to opioid prescriptions. The goals of this study can be described along two general objectives that lead to the main contribution of the study. First, this study aims to help in facing the recent deadly opioids' epidemic in the U.S. Second, this study aims to help in addressing the important question of how the daily posts and activities of online social media users can provide better opioids prevention, treatment, and recovery strategies; and strengthen the public health data reporting and collection for opioids epidemic. The main contribution of this study aims is to gain a better understanding of content related to opioids epidemic as represented on social media to inform the strategies and other public health efforts.

Dheeman Saha, South Dakota State University
Laura White, Udaykumar Sekaran, Sandeep Kumar, Senthil Subramanian

Impact of nitrogen rates and landscape position on bacterial taxa and metabolic pathways during switchgrass production

Biofuel crops offer alternative sources of energy and are crucial for sustainability and reduced reliance on foreign oil. Switch grass is an important component of marginal land biofuel systems. Its production requires the application of nitrogen fertilizers which when applied in excess result in greenhouse gas emission, and soil health concerns. Both these processes are impacted by soil microbial communities that play a role in soil biogeochemical processes. The goal in this project is to determine the impact of nitrogen fertilizer rates (high, medium, and low), and landscape positions (crest, toe) on soil microbial communities, in order to enable better nitrogen management in switchgrass production. The composition of bacterial taxa and potential metabolic pathways present in soil samples were evaluated using next-generation sequencing data. Sequencing libraries were constructed from 16S ribosomal RNA (rRNA) gene amplicons, and the resulting sequences were analyzed using Qiime2 to determine bacterial taxonomy composition in the soil samples. The resulting Biological Observation Matrix (BIOM) files were analyzed using PICRUSt, which predicts the metagenome functional content from the 16S rRNA gene sequencing. Together the results revealed specific bacterial taxa and potential metabolic pathways that were influenced by fertilizer rates and landscape position. These results will be presented and discussed.

Kari Sandouka, Dakota State University

Interactive Visualizations: A Literature Review

Visualization has proven to be an effective method at reducing the cognitive processing of data. The primary goal of visualization is to amply cognition where the visual representations facilitate analytical reasoning, support decision-making and allow users to gain insight into complex problems (Card, Mackinlay, & Shneiderman, 1999; Thomas & Cook, 2006; Yi, Kang, Stasko, & Jacko, 2008). Different types of visualizations correspond to different kinds of information. Properly designed graphs reduce bias while simultaneously supporting the goal of visualization. Visualization experts, such as Stephen Few, have produced graph selection frameworks (Borner, 2015). These frameworks assist visualization designers in selecting the correct graph type for a given task, in order to properly design effective visualizations. As the volume of data grows and the complexity of data increase, it is unclear if these frameworks still apply. It is unknown if the selection matrices are as effective with dynamic visualizations as they are with static visualization. The effectiveness of a visualization hinges on two things: its ability to clearly and accurately represent information and the ability to interact with the information to figure out what it means (Few, 2009). The current frameworks for graph selection do not account for interaction techniques. This research discusses extant and emerging literature of interacting with visualizations.

Prakriti Sharma, South Dakota State University
Jesse Wittnebel, Melanie Caffe Treml

High-Throughput Phenotyping of Oat Breeding Nurseries Using Unmanned Aerial Systems

Current strategies for phenotyping thousands of breeding lines under field conditions demand significant investment in both time and labor.Unmanned aerial systems (UAS) can be used to collect Vegetative Index (VI) with high throughput and could provide an efficient way to predict forage yield on thousands of breeding lines. The main objective of the study was to evaluate the use of vegetative indices derived from images for estimating crop biomass. A UAS equipped with a RGB camera was flown over experimental plots in Southshore several times throughout the growing season. A radio-calibration was performed to correct band values on the stitched images before deriving Visual Normalized Difference Vegetation Index (VNDVI). Asignificant positive correlation ($r = 0.6$) between VNDVI and crop biomass was observed for the last flight before forage harvest. This suggests that UAS could provide an efficient way to measure the forage yield potential of oat breeding lines. However, to confirm this, it will be necessary to repeat this experiment in additional environments.

Sebastian Sowada, South Dakota State University

Utilizing Uplift Modeling to Develop a Credit Line Increase Strategy

Traditionally, response models have been used to target individuals who are most likely to respond (yes/no, good/bad) to a direct action (marketing campaign, credit line increase). However, response models are incapable of predicting who is most affected by the direct action. In this paper we will analyze the practice of uplift modeling, the predictive modeling technique used by statisticians to measure the incremental effect of a treatment. We will examine the theoretical background of two types of uplift modeling: difference in scores, and uplift trees. This paper will then describe different validation approaches used specifically to uplift modeling, such as the Qini statistic and Net information value. This paper will conclude in a case study using data on credit card customers. We will use uplift modeling to predict customers who are most likely to have a positive incremental impact from a credit line increase. Finally, using both types of uplift modeling, we will compare the uplift brand of models to a traditional response model and detail why uplift modeling is more appropriate in this situation.

Haidong Wang, University of South Dakota
KC Santosh

Face reidentification in real-time large time series data

Face data happens everywhere and face matching/verification is the must, such as it helps track criminals; unlock your mobile phone; and pay your bill without credit cards (e.g. Apple Pay). Within the scope, the proposed project is to build a face detection and re-identification based on Convolutional Neural Networks (CNN). In our training, separate face data are used (from several different video clips, where frames are used to create image data of size 224,224,3 using OpenCV). For the test, a real-world data (large video) was used, where multiple people are captured. In our implementation, we have used 'keras' (tensor flow backend) to build CNN Resnet 50 architecture that has 23 millions parameters and 174 layers, where VGG pre-trained weight of human faces were applied. We have used support vector machine (SVM) for face re-identification. Our aim is to demonstrate end-to-end simulation at the symposium.

Workshops

Tidyverse: R packages for data science
Dr. Adam Sullivan

Deep Learning with Python
Dr. David Zeng

Currently Deep Learning is required for at least 80% of jobs for data scientists and machine learning engineers. This workshop introduces deep Learning concepts, models, and applications with Keras, the most popular high-level library for Deep learning in Python. Topics include Keras' sequential models, convolutional neural networks for image classification, recurrent neural networks (LSTM) for natural language processing. Some advanced and popular topics such as transfer learning with pre-trained models, reinforcement Q-learning with OpenAI Gym, and training deep learning models with GPU may be covered as well. While some experience in Python or data analytics may be beneficial, no previous knowledge about deep learning is required. All working documents (with tutorials and Python codes) will be provided.

Symposium Committee

A big thank you to everyone involved with organizing the 2019 Data Science Symposium. Your commitment and dedication made this event possible!

Department Head

Kurt Cogswell

Conference Chair

Dr. Tom Brandenburger
thomas.brandenburger@sdstate.edu
605-688-6196

Committee Members

Dr. Semhar Michael
semhar.michae@sdstate.edu
605-688-6316

Dr. Gary Hatfield
gary.hatfield@sdstate.edu
605-688-5846

Dr. Rong Fan
Rong.fan@sdstate.edu
605-688-6196

Events Management

Sheila Ohlsen
Linda Wendt

Volunteers

Thank you to the many volunteers that shared their time and talent to make this conference possible.

Seth Anhir-Donkor
Philip Ato Sam
Eric Bae
Brendan K. Branick
Clarissa Giefer
Mahtab Hajebi
Alex Harms
Mahzabin Khan
Paul May
Nikita Medvedev
Samantha Nystrom
Lawrence S. Segbehoe
Emma Spors
Augustine Tarkom
Kai Zhang

South Dakota
State University

Department of
Mathematics and Statistics