

South Dakota State University

# Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange

---

SDSU Data Science Symposium Programs

---

2020

## 2020 SDSU Data Science Symposium Program

South Dakota State University

Follow this and additional works at: [https://openprairie.sdstate.edu/ds\\_symposium\\_programs](https://openprairie.sdstate.edu/ds_symposium_programs)

---

# SDSU Data Science Symposium 2020



**SOUTH DAKOTA  
STATE UNIVERSITY**

**February 10-11, 2020  
South Dakota State University  
Brookings, South Dakota**

# Sponsors

---

## GOLD



CAPITAL SERVICES

---

## SILVER



TOTAL CARD, INC.



## BRONZE



**First PREMIER Bank**  
Member FDIC  
**PREMIER Bankcard**



GREAT WEST CASUALTY COMPANY

---

*The Difference is Service®*

# Welcome

---

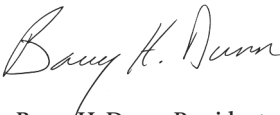


Dear Attendees,

As President of South Dakota State University, I am happy to welcome you to the 2020 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,

A handwritten signature in black ink that reads "Barry H. Dunn".

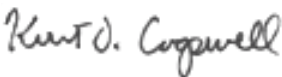
**Barry H. Dunn, President**  
South Dakota State University



Dear Attendees,

On behalf of the SDSU Department of Mathematics and Statistics, it is my pleasure to welcome you all to the 2020 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,

A handwritten signature in black ink that reads "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

1 Welcome from President	16 Topic Contributed Abstracts
1 Welcome from Department Head	21 Contributed Abstracts
3 General Information	25 Poster Abstracts
6 Symposium Schedule	32 Workshop Abstracts
9 Biographies	33 Symposium Committee
16 Keynote	

# General Information

## Proof of Attendance

If needed, please send a request to [sdsu.seminars@sdsu.edu](mailto:sdsu.seminars@sdsu.edu)

## Nametags

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

## Registration/Check-in Hours

Located at the University Student Union in the Volstorff Lounge.

Monday, February 10, 12:30-5:00 p.m.  
Tuesday, February 11, 7:30 a.m.-noon

## Luggage Check Hours

Located at the University Student Union in the Volstorff Lounge.

Monday, February 10, 12:30-5:00 p.m.  
Tuesday, February 11, 7:30 a.m.-noon

## 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 **7628#** to enter the pay lot.

## Transportation

### BATA Bus

Visit the Brookings Area Transportation Authority's website for more information about their services:

[www.brookingsareatransit.com](http://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@sdsu.edu](mailto:sdsu.seminars@sdsu.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.

## 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

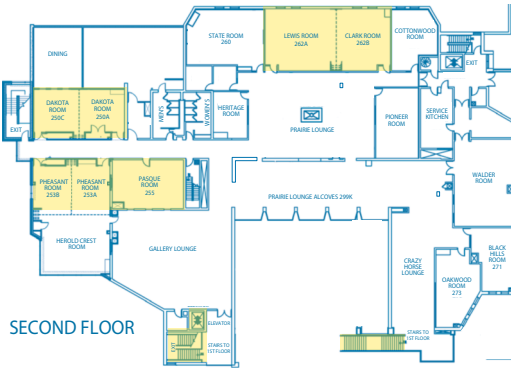
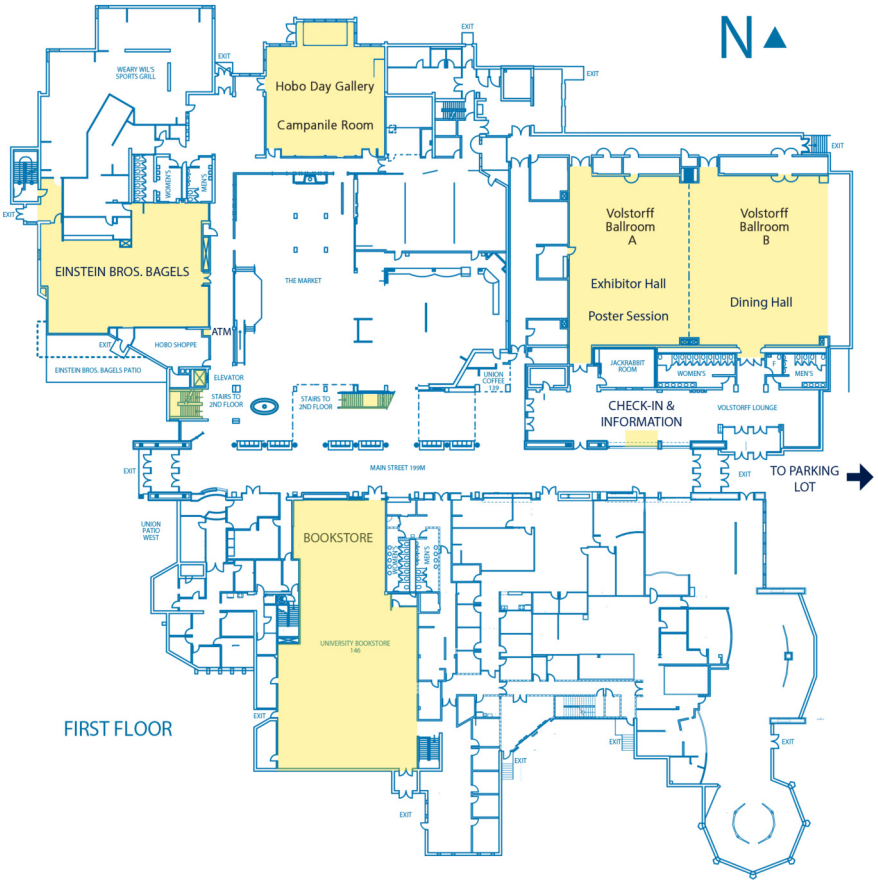
- Union Pay Lot
- Parking

## ATTRACTIONS

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

I-29 ►

# Union Map



Monday, February 10, 2020			
Time	Pasque Room (255)	Dakota Room A/C (250)	Pheasant Room
12:30-5:00 p.m.		Check-in/Registration, Volstorff Lounge	
1:00-3:00 p.m.	<b>WORKSHOP I</b> <i>Data Visualization</i> Dr. Xijin Ge	<b>WORKSHOP II</b> <i>Web Crawling</i> Peter Clausen	<b>WORKSHOP III</b> <i>R Package Creation Fundamentals</i> Riaz Khan
1:00-5:00 p.m.	Job Fair/Recruiting   Exhibitors, Campanile Room & Hobo Day Gallery A&B		
6:00-6:30 p.m.	Banquet, Performing Arts Center		
6:30-8:00 p.m.	Social Time (cash bar)		
7:15-7:30 p.m.	Dinner		
7:30-8:30 p.m.	Welcome, Dr. Bruce Berdanier, Dean of the Jerome J. Lohr College of Engineering  <b>Panel Discussion: Perspectives on Data Science in Organizations</b> , Steve Cross, Great West Casualty, moderator; Charles Henderson, Capital Services; David Groeneveld, Advanced Remote Sensing; Valarie Bares, Sanford Research; Dhiraj Sharan, Query AI		

9

Tuesday, February 11, 2020			
Time	Pasque Room (255)	Dakota Room A/B (253)	Campanile & Hobo Day Gallery (A&B)
7:30 a.m.-noon		Check-in/Luggage Check, Volstorff Lounge	
7:45-8:20 a.m.		Breakfast, Volstorff B	
8:20-8:30 a.m.	Opening Session: <i>Welcome and Introduction</i> , Dr. Kurt Cogswell, Department Head, Volstorff B		
8:30-9:15 a.m.	<b>SESSION 1   Tools</b> Chair: Tobey Eich  <b>SAS Viya Tour</b> Melodie Rush, SAS Institute	<b>SESSION 2   Opportunities for Improving Health through Big Data &amp; Data Science</b> Chair: Dr. Arielle Selya, Sanford  <b>Data Science Infrastructure in the Cloud for Sanford Imagenetics</b> Murat Sincan, Sanford Research  <b>Data-Driven Healthcare: The Sanford Data Collaborative</b> Arielle Selya, Sanford Research	<b>SESSION 3   Precision Agriculture</b> Chair: Paul May, SDSU  <b>Multi-Resolution Approximations for Precision Agriculture</b> Paul May, SDSU  <b>Towards Deep Learning for Weed Detection: Deep Convolutional Neural Networks Architecture for Plant Seeding Classification</b> Martinson Ofori, Dakota State University
9:30-10:30 a.m.			<b>SESSION 4   Advances in Probabilistic Modeling for Machine Learning</b> Chair: Igor Melnykov, UM-Duluth  <b>Semiparametric Imputation Using Conditional Gaussian Mixture Models</b> Danhyang Lee, University of Alabama  <b>On Finite Mixture Modeling of Change-Point Processes</b> Yana Melnykov, University of Alabama



9:30 -10:30 a.m.				<p>An Approach to Initializing the EM Algorithm in Gaussian Mixtures with an Unknown Number of Components Igor Melnykov, Minnesota-Duluth</p> <p>Gaussian Mixture Modeling and Model-Based Clustering Under Measurement Inconsistency Rong Zheng, Western Illinois University</p>
10:30-11:00 a.m.	<b>Networking Break   Exhibitors, Volstorff A</b>			<p><b>SESSION 6   Nonparametric Methods for Data Science</b> Chair: Danica Ommen, Iowa State</p> <p><b>Some of These Things is Not Like the Others: Visual Statistics and Testing in Statistical Groups</b> Susan Vanderplas, Nebraska-Lincoln</p> <p><b>Asymptotically Distribution-Free Change-Point Detection for Non-Euclidean &amp; Multivariate Data</b> Iyanna Chu, Iowa State</p> <p><b>The Development of Pairwise Sample Comparison Methods of Particle Micromorphometry of AL Powders for Nearest Neighbor Classifiers</b> Cami Fuglsby, SDSU</p>
11:00 a.m.-noon	<p><b>SESSION 1   Tools</b> Chair: Tobey Eich</p> <p><b>SAS Viya tour</b> Melodie Rush, SAS Institute</p>	<p><b>SESSION 2   Opportunities for Improving Health through Big Data &amp; Data Science</b> Chair: Arielle Selya, Sanford Research</p> <p>Propensity score methods for reducing bias in observational studies: An example from PRAMS study Sooyong Kim, Sanford Research</p> <p><b>Floor Discussion</b> Arielle Selya, Sanford Research</p>	<p><b>SESSION 5   Natural Language Processing</b> Chair: Sayeed Sajal, Milnot State University</p> <p><b>Large-Window Techniques for Geospatial Raster Data</b> Anne Denton, North Dakota State</p> <p><b>Survey Sentiment Analysis Using Open Source Technologies</b> Jessica Meyer, Optum</p> <p><b>Simple Introduction to Natural Language Processing and Its Clinical Applications in the Era of Artificial Intelligence</b> Yanshan Wang, Mayo Clinic</p>	
noon-1:00 p.m.	<b>Lunch, Volstorff B   Poster Session, Volstorff A</b>			
2:00-3:00 p.m.	<b>Poster Session   Student Poster Competition, Volstorff B</b>			
2:00-3:25 p.m.	<b>Job Fair/Recruiting   Exhibitors, Volstorff B</b>			
<i>Schedule continues on next page.</i>				

	<p><b>SESSION 7   Application</b> Chair: Speed Session by Grad Students</p>	<p><b>SESSION 8   Tools</b> Chair: David Zeng</p> <p><b>Jupyter Notebook &amp; Python for Data Scientists</b> David Zeng, Dakota State University</p> <p><b>Introduction to Git and Github</b> Seema Bhandari, Dakota State University</p>	<p><b>SESSION 9   Forecasting</b> Chair: Ally Pelletier</p> <p><b>Data Science in the Life Insurance Industry</b> Gary Hatfield, Securian Financial, University of Minnesota</p> <p><b>Using Prophet Methodology for Forecasting</b> Ally Pelletier, Polaris</p> <p><b>Multi-Step Forecast of Implied Volatility Surface Using Deep Learning</b> Zhiquang Wang, SDSU</p>	<p><b>SESSION 11   Methods</b> Chair: Karissa Palmer</p> <p><b>Bootstrap Control Chart for Pareto Percentiles</b> Ruth Burkhalter, University of South Dakota</p> <p><b>Stress-Strength Inference for the Multicomponent system Based on Progressively Type-II Censored Samples From Pareto Distributions</b> Lauren Sauer, University of South Dakota</p> <p><b>Imputing Data Without Replication</b> Jixiang Wu, SDSU</p>
<p><b>3:25-3:30 p.m.</b></p>	<p><b>Break</b></p>			
<p><b>SESSION 7   Application</b> Chair: Speed Session by Grad Students</p>	<p><b>SESSION 12   Finance</b> Chair: Thomas Brandenburger</p> <p><b>Bias In, Bias Out? How to Fight Biases in AI and Create Trust</b> Gerald Fahner, FICO</p>	<p><b>SESSION 13   Data Science &amp; Startups</b> Chair: Samantha Nystrom</p> <p><b>Survey Sentiment Analysis Using Open Source Technologies</b> Jessica Meyer, Optum</p> <p><b>Get Answers and Insights From Your Data</b> Dhiraj Sharan, Query AI</p>	<p><b>SESSION 14   Methods</b> Chair: Tye Kinnert</p> <p><b>A Patient Centered Approach to IT Enabled Diabetes Self-Management: The Case of Saudi Arabia</b> Hassan Alyami, Dakota State University</p> <p><b>A Dense-Inception Network for Medical Image Classifications</b> James Boit, Dakota State University</p> <p><b>Exploring Task Decomposition with Ensemble Approach Using Reinforcement Learning</b> Raissa Nusrat, University of South Dakota</p>	
<p><b>3:30-4:30 p.m.</b></p>	<p><b>Closing Session: Thomas Brandenburger, Volstorff B   Poster Winners Announced</b></p>			

# Biographies

## Keynote Speaker



### **Alicia Carriquiry, ISU, CSAFE**

Alicia Carriquiry is a distinguished professor and President's Chair in Statistics and director of Center for Statistics and Applications in Forensic Evidence at Iowa State University.

## Executive Panel Moderator



### **Steve Cross, Great Western Casualty**

With over 22 years of analytical experience in consulting, process improvement, solution design, and data development, Steve Cross has utilized his analytical rigor in the identification, refinement, and application development of new or existing products and concerns to improve product offerings in the most ethical manner. Cross came to Great West Casualty with experience in a variety of vertical markets, specializing in financial, insurance, credit services, health-care, government, marketing, and automotive applications. He has been a speaker at the DMA, NCDM, as well as dozens of corporate conferences. At Great West, Cross has worked on refining the pricing models, creating solutions for internal and external stakeholders, and improving the training focus for analytics and big data. His combination of technical and consultative expertise allows him to wear many hats, giving him a unique perspective. Cross has an undergraduate degree in mathematics/economics from Bradley University and graduate degrees in statistics and genetics from University of Nebraska, Lincoln.

## Invited Speakers



### **Valerie Bares, Sanford Research**

Dr. Valerie Bares earned her Ph.D. in computational science and statistics from SDSU in 2017. Soon after graduation, she joined Sanford Research in the Behavioral Sciences group as a biostatistician. At Sanford, Bares has worked with several areas of research such as behavioral science, cardiology, oncology, orthopedics, and sports science; including both big and small data. She is currently the program director and senior biostatistician of the Research

Design and Biostatistics Core.



### **Lyanna Chu, Iowa State University**

Lyanna Chu earned a Ph.D. in biostatistics at UC Davis in 2019 and before that, she studied at UCLA. Chu is currently working on problems related to modern data types, specifically in the context of change-point detection.



### **Anne Denton, North Dakota State University**

Anne Denton is professor in the Computer Science Department at North Dakota State University. She earned a Ph.D. in physics from the Johannes Gutenberg University, Mainz, Germany, in 1996 and a M.S. in computer science from NDSU in 2003. Her research interests are in data mining of diverse scientific data sets that are too complex to be analyzed using classical statistics techniques. Currently, she is working with collaborators in soil science, agricultural engineering, hydrology, and atmospheric science on projects that involve the climate impacts on agriculture. Denton has published more than 60 peer-reviewed journal and conference publications and has led projects funded at a total of more than one million dollars.



### **Tobey Eich, Premier Bankcard**

Tobey Eich is a managing consulting analyst IV at Premier Bankcard.



### **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. Fahner 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.



### **Cami Fuglsby, South Dakota State University**

Cami Fuglsby is a third-year Ph. D. student under Dr. Christopher Saunders in computational science and statistics at SDSU. She is focusing her research on the study of forensic identification of source problems using machine learning and pattern recognition techniques. Fuglsby serves on the forensic document examination sub-committee through the Organization of Scientific Area Committees for Forensic Science under the National Institute of Standards and Technology as their statistician, where her main focus is on the conclusion language used in examinations. She has received a number of awards for her work, including the first American Statistical Association poster prize at the International Workshop on Simulation and Statistics. Fuglsby's work on statistical methods is supported by the National Institute of Justice.



### **David Groeneveld, Advanced Remote Sensing, Inc.**

David Groeneveld recently moved to the Sioux Falls area from Santa Fe, N.M., coming here specifically to start a business in 2018. His previous startup, HydroBio, built on a long career consulting in issues concerning water, environment and remote sensing. HydroBio developed his patented irrigation prescription method that employs weather and satellite image data to accurately estimate crop water use found to be as accurate for irrigation prescription as soil probes in every pixel of the field. That IP is now irrigating around the world for Bayer. For Advanced Remote Sensing, Inc. (ARSI), David's vision is to solve the most vexing problem

in satellite remote sensing for agriculture; noise-inducing effects from the atmosphere that prevent many potential low-cost revolutionary applications. ARSI was funded by NSF and USDA SBIR Phase I programs in 2019. In 2020, with two anticipated SBIR Phase II awards, ARSI will complete an atmospheric correction algorithm and begin application of game changing solutions for agriculture.



**Gary Hatfield, Minnesota Center for Financial and Actuarial Mathematics**

Dr. Gary Hatfield is a senior director and actuary for Securian Financial. He leads a team of actuaries, data scientists, and quants known internally as the Actuarial Research and Analytics Center of Excellence. He also serves on the company's most senior risk committee. He previously played a key oversight role in asset liability management and hedging. He earned his Ph.D. in mathematics from the University of Minnesota where he continues to be involved at the School of Mathematics as an assistant professor and an advisory board member for the Minnesota Center for Financial and Actuarial Mathematics. Hatfield also serves the curriculum committee for the quantitative finance and investments track of the FSA exams. He is a Fellow of the Society of Actuaries, a member of the American Academy of Actuaries and is also a CFA charter holder. Previous to entering the insurance field in 1998, he taught mathematics at Gustavus Adolphus College in St. Peter, Minn.



**Charles Hendrickson, Capital Services**

Charles Hendrickson has been with Capital Services for over 19 years, joining the company in 2000 as chief financial officer. He moved into the executive vice president role prior to becoming president and CEO. Through his tenure in bank controller and CFO positions, he has gained extensive experience in asset liability management, bank financial planning, and acquisitions. It was his initiative to understand and better report credit card profitability that resulted in Capital's industry-leading ViPRSM system. He also championed the in-house development of Capital's unique scorecard and statistical analysis competencies. Hendrickson began his career in public accounting with McGladrey and Pullen and later held a bankcard finance role at Citibank before his tenure in commercial banking. He has a degree in business economics from St. John's University in Minnesota and is also a certified public accountant. He serves on the board of directors of Capital.



**Sooyong Kim, Sanford Research**

Sooyong Kim is a senior research specialist at Selya lab in Sanford Research. She received her degree as a medical doctor at Korea University in 2015 and earned a Master of Public Health at the University of North Dakota in 2019. Her research interests lie in perinatal epidemiology, especially systemic and maternal behavioral factors that affect maternal and neonatal outcomes.





### **Danhyang Lee, The University of Alabama**

Dr. Danhyang Lee is currently an assistant professor in the Department of Information Systems, Statistics and Management Science at the University of Alabama's Culverhouse College of Business. She received her Ph.D. degree in statistics from Iowa State University in 2019. She has a broad research interest in applied statistics, including survey sampling and missing data analysis. Specifically, her current research deals with issues in multi-level modeling, missing data analysis, survey data integration, small area estimation, and Bayesian inference.



### **Igor Melnykov, University of Minnesota, Duluth**

Dr. Igor Melnykov, originally from Kharkiv, Ukraine, moved to the U.S. in 1999 and earned a Ph.D. in statistics from Bowling Green State University in 2005. He works in the mathematics and statistics department at the University of Minnesota, Duluth, and has held positions at Colorado State University, Pueblo, and Nazarbayev University in Kazakhstan. His research interests include cluster analysis, classification, multiple hypothesis testing, and asymptotic theory.



### **Yana Melnykov, The University of Alabama**

Dr. Yana Melnykov is a senior statistician in the Institute of Business Analytics and an assistant professor of statistics in the ISM department at the University of Alabama where she worked on a project supported by Lockheed Martin. The corresponding paper, "Studying contributions of variables to classification," was published in *Statistics and Probability Letters*. Melnykov's research interests lie in the areas of change point inference and finite mixture modeling.



### **Jessica Meyer, Optum**

Jessica Meyer is a senior data scientist with 5+ years' experience in the field of data analysis, big data and machine learning. She has worked in a variety of industries including retail and healthcare. Meyer holds a B.S. in business administration from Metropolitan State University and a M.S. in data science from the University of St. Thomas. In her spare time, she works as an adjunct associate at the University of Columbia SPS Analytics program, volunteers on the Data Science Committee reviewing speaker proposals for the annual Grace Hopper Tech Conference and co-hosts the WiT Twin Cities podcast.



### **Danica Ommen, Iowa State University**

Dr. Danica Ommen is an assistant professor of statistics at Iowa State University.



### **Ally Pelletier, Polaris**

Ally Pelletier is currently working for Polaris in Minneapolis, Minn. Previously, she was involved in reporting and modeling for the digital department at the *Star Tribune*. Her work included recommendation models, customer retention, and map development. Previous to working at the *Star Tribune*, Pelletier worked as a data science consultant with RProfet. In this position, 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. Pelletier earned a bachelor's in mathematics education from Concordia College in Moorhead, Minn., and an master's in statistics from SDSU. During her time at SDSU she held a research assistantship and an internship in digital media, where 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.



### **Melodie Rush, SAS Institute**

Melodie Rush is the statistician for the Customer Loyalty Team at SAS Institute. She received both her B.S. in statistics and her master's in science of management with a technical option in statistics from North Carolina State University. Before joining SAS in 1996, Rush worked for the Research Triangle Institute as a statistician. Her responsibilities included implementing national and local surveys of various topics, such as health care, employee benefits, and drug abuse.

As part of her research, she has published work for both the American Statistical Association and the American Public Health Association. After joining SAS, Rush has developed presentations and methodology for doing many types of analysis, including data mining, forecasting, data exploration and visualization, quality control and marketing. She has spent the last 16 years helping companies identify and solve problems in each of these analytical areas.



### **Sayeed Sajal, Minot State University**

Dr. Sayeed Sajal is an assistant professor in the Department of Computer Science at Minot State University. He earned his master's and Ph.D. in electrical and computer engineering from the North Dakota State University. He also has an MBA degree in finance and marketing. In addition, Sajal has five years of industry experience in the wireless telecommunications industry. His research focuses are on secured wireless communications, cyber-security, data science, and sensor design. Sajal passionately loves to teach as he believes in the "ripple effect of teaching" even after his death. Currently, he is collaborating in material science, industrial engineering, statistics, hydrology, biology, and computer science, and his two multi-million multi-year NSF grants are pending for approval. Sajal has published more than 30 peer-reviewed journal and conference publications and has led projects funded. He is also looking for collaboration in data science and cybersecurity projects.



### **Arielle Selya, Sanford Research**

Dr. Arielle Selya currently works at the Behavioral Sciences Group at Sanford Research. Selya studies substance use and addiction, particularly cigarette and e-cigarette use among adolescents. She is very interested in advanced methodology, and regularly uses structural equation modeling, propensity score methods for causal inference, machine learning, and system dynamics simulation modeling approaches in her methods. Selya also does research on health services, implementation science, nutrition, and education.



### **Dhiraj Sharan, CEO, Query.AI**

Dhiraj Sharan is the founder and CEO of Query.AI, a startup that unlocks the power of data through the simplification of access and analysis. His career spans 20 years of innovation and leadership in four startups and large companies like HPE, Aruba and Novell. He holds 10 patents and has been a Cisco EIR Entrepreneur. Sharan earned a bachelor's in computer science from IIT (BHU, India) and certificate in management from Harvard University.



### **Murat Sincan, Sanford Research–Sanford Imagenetics**

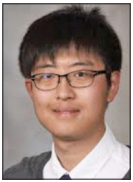
Murat Sincan, M.D., is a director of computational medical informatics for Sanford Imagenetics. Dr. Sincan focuses on both genetic and non-genetic components that influence a patient's health. His current project involves the development and use of natural language processing, which pulls clinical concepts from information living in an EMR. Dr. Sincan also conducts research by working with external partners, such as LA BioMed and North Dakota State

University. They look at theories and different aspects of precision medicine with which Sanford Health is engaged.



### **Susan Vanderplas, University of Nebraska, Lincoln**

Susan Vanderplas earned a Ph.D. in statistics from Iowa State University in 2015, then spent time in the private sector before returning to academia in 2018. She is an assistant professor at University of Nebraska, Lincoln, with research interests in statistical graphics and statistical forensics.



### **Yanshan Wang, Mayo Clinic**

Dr. Yanshan Wang is an assistant professor in the Division of Digital Health Sciences at Mayo Clinic. His research is centered on developing novel natural language processing (NLP), artificial intelligence (AI), and machine learning methodologies for clinical applications. His research goal is to develop EHR data-driven AI solutions to accelerate clinical research and to facilitate clinical practice. Since joining Mayo in 2015, he has served as investigator for multiple

extramural NIH-funded projects and intramural operational projects. He has published over 50 peer-reviewed articles in high-impact medical informatics journals (e.g., JBI, JAMIA), and conferences (e.g., AMIA Annual Symposium, AMIA summit, IEEE BIBM). Wang is also active in organizing conference workshops and shared tasks in the medical informatics community, including the international health NLP workshops and the renowned national NLP clinical challenge (n2c2).



### **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. Zeng earned a Ph.D. in information systems from University of California, Irvine, and an M.S. in computer science from California State University, Long Beach.



### **Rong Zheng, Western Illinois University**

Dr. Rong Zheng earned a Ph.D. in applied statistics at the University of Alabama in 2017 with a concentration on data classification and model-based clustering, statistical quality control and nonparametric statistics. Zheng also earned a bachelor's degree in mathematics and applied mathematics from Henan University, China, in 2012. Her tenure-track faculty position with Western Illinois University started in 2017. She teaches R programming, SAS programming, cat-

egorical data analysis, business statistics, applied time-series forecasting, and applied regression analysis.

## Workshop Instructors

### **Peter Claussen, Gylling Data Management, Inc., SDSU**

Peter Claussen, M.A., is a biometrician who works with statistical analysis with R, python and SAS and experimental design.



### **Xijin Ge, South Dakota State University**

Dr. Xijin Ge earned a BSc. and M.S. in physics from the Beijing University of Science and Technology, China, and a Ph.D. from the University of Tokyo, Japan. After postdoc training, he has been teaching in the SDSU Department of Mathematics and Statistics since 2007. His research interests include bioinformatics, genomics and data science. [www.sdstate.edu/directory/xijin-ge](http://www.sdstate.edu/directory/xijin-ge)



### **Md Riaz Ahmed Khan, South Dakota State University**

Being a high school student who enjoyed physical science and mathematics the most, Riaz went to Bangladesh University of Engineering and Technology and got Bachelor's degree in Electrical Engineering in 2011. Over the next three years, he worked as an electrical engineer before starting graduate school. Riaz got master's in electrical engineering (2016) and master's in statistics (2018) from SDSU. With an engineering and statistics background, Riaz enjoys computer programming and analyzing data, and does on a daily basis. He is the major author of R package ROCit, and gave contribution to the R package Rprofet, both available on CRAN.

# Keynote

**Alicia Carriquiry, Iowa State University, CSAFE**

## ***Machine Learning and the Evaluation of Forensic Evidence***

The emergence of DNA analysis as an effective forensic tool in the 1990s was a revelation, in that for the first time it was possible to quantify the degree of association between a crime scene sample and a suspect. It also had the effect of shining a light on other forensic practices, most of which lack the rigorous and widely accepted scientific foundations of DNA profiling and for which error rates are largely unknown. In the U.S. criminal justice system, jurors choose between two competing hypothesis: the suspect is the source of the evidence found at the crime scene or is not. We discuss how statistical learning algorithms can help address the question of the source of the evidence, and use firearms examination as an illustration. Can we tell whether the defendant's gun fired the bullet recovered from the scene of the crime?

# Topic Contributed Abstracts

**Seema Bhandari, Dakota State University**

## ***Introduction to Git and Github***

GIT is a free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency. The most widely used modern version control system provides a team of collaborators a platform to share and track the changes in source codes. This presentation introduces GIT, GitHub, and how they would help collaborators manage version control of their documents, starting with version control is and why you should use it, and introduces the most common steps in a common Git workflow. Github is the largest web-based git repository hosting service and it enables code collaboration with anyone online. Github also adds extra functionality on top of git. The presentation demonstrates how data scientists would benefit from GIT and Github.

**Lynna Chu, Iowa State University**

## ***Asymptotically Distribution-Free Change-Point Detection for Non-Euclidean and multivariate data***

We present a new framework for the testing and estimation of change-points, locations where the distribution abruptly changes, in a sequence of multivariate or non-Euclidean observations. While the change-point problem has been extensively studied for low-dimensional data, advances in data collection technology had produced data sequences of increasing volume and complexity. Motivated by the challenges of modern data, we study a non-parametric framework that utilizes similarity information among observations and can be applied to various data types as long as an informative similarity measure on the sample space can be defined. Analytical p-value approximations formulas are also provided, making the methods easy-off-the-shelf tools for real applications. The effectiveness of the new approaches are illustrated in an analysis of New York taxi data.



**Anne Denton, North Dakota State University**  
***Large-Window Techniques for Geospatial Raster Data***

In conventional geographic information systems, topographical variables and confocal statistics are typically computed over small sliding windows, often of size 3x3. Such window sizes are appropriate for conventional remotely sensed images with pixel sizes of around 30m. In recent years, high resolution raster data from drones and low-orbit satellites has become ubiquitous. Without resampling high-resolution images, window-sizes orders of magnitude larger in size have to be considered to gain physically relevant information. Conventional computational techniques don't scale to the number of pixels involved. I will show that many complex derived quantities, such as topographic attributes and even fractal dimension can be evaluated efficiently for much larger windows, provided the relevant aggregates are computed using iteratively doubling window sizes. The effectiveness is demonstrated for problems in agriculture and hydrology.

**Gerald Fahner, FICO**  
***Bias In, Bias Out? How to Fight Biases in AI and Create Trust***

Alarming headlines appear at an increasing rate about “biased AI” leading to nontransparent decisions, discrimination and lack of trust.

Newcomers to AI application development are confronted with (and often struggle with) multiple traps that can inadvertently lead to biased and unreasonable automated decisions. By comparison, credit score developers have decades of experience in the “art and science” of mitigating biases through transparent and credible models and decisions that provide benefits to all stakeholders while complying with fair lending regulations.

In this presentation you will learn about traps to be avoided and get our unique perspective on how to achieve comprehension and trust while beating back biases through a special synthesis of data, expertise and technology. The battle to minimize biases is ongoing and we will also share our latest thinking how we can up the ante in this endeavor.

**Cami Fuglsby, South Dakota State University**  
***The Development of Pairwise Sample Comparison Methods of Particle Micromorphometry of AL Powders for Nearest Neighbor Classifiers***

Due to the online sharing of instructional manuals and published books on how to construct an improvised explosive device (IED), bomb makers are now informed on the easily accessible household materials that can be used to make explosive chemical mixtures. This presentation is focused on the development of efficient scoring rules for measuring the similarity (and dissimilarity) of two sets of Al powder based on two distributions of particles' size and shape. In this study, aluminum (Al) powder samples were obtained from legitimate industrial manufacturers, various “in-house” production methods, and seized IEDs. The amateur methods were replicated to produce Al powder from easily available sources, including Al foil, cans of metallic spray paints, Al ingots melted from Al cans and filed or lathed, pyrotechnics, and catalyst packets from binary exploding targets. The resulting datasets are too large and complex to analyze with standard statistical methods in an efficient and scalable manner. Our current work has focused on developing methods that will compare two sets of particles and summarize the difference in the distribution of particles; with each particle characterized by 17 morphometric measurements. The approach we have relied on is focused on using a set of low dimensional projections where we can measure the discrepancy between the two distributions of the projections of particles. We then assess the performance of an omnibus score for a given class of projections and a corresponding distributional comparison method by comparing the within-

source and between-source score distributions with ROC curves. This approach gives us a gross measure of the performance of the score when used for common-but-unknown source-identification problems that commonly arise in forensic science applications such as with IEDs.

**Sonyoon Kim, The University of North Dakota**  
***Electronic Cigarettes May Not Be a “Safer Alternative” of Conventional Cigarettes During Pregnancy: An Analysis of PRAMS Data***

Background: Conventional cigarette (CC) smoking is one of the most preventable causes of adverse birth outcomes. Although electronic cigarettes (ECs) are considered to be safer than CCs during pregnancy, the evidence is yet to be presented. This study examines the effects of EC use on birth outcomes compared to those of CC smokers using multiple analytic methods to adjust for risk factors of tobacco use. Methods: Data were extracted from 3,727 exclusive CC smokers and exclusive EC users who participated in the Phase 8 survey of the Pregnancy Risk Assessment Monitoring System (PRAMS). Adverse outcomes included small-for-gestational-age (SGA), low birthweight (LBW), and prematurity. Analyses were conducted using three different methodologies to account for nine covariates: 1) weighted, unadjusted logistic regression, 2) weighted, adjusted logistic regression, 3) inverse propensity weighted (IPW) logistic regression. Results: While unadjusted logistic regression analyses showed variable favorable effects of EC use on SGA (OR 0.64; 95% CI 0.39–1.04) and LBW (OR 0.58; 95% CI 0.40–0.82) compared to CC smoking, accounting for nine covariates using IPW rendered the results statistically insignificant. Using ECs does not significantly decrease the odds of neonates being SGA (OR 0.64, CI 0.28–1.45), LBW (OR 0.65, CI 0.35–1.20) or born prematurely (OR 0.80, CI 0.37–1.75) compared to smoking CCs. Conclusion: The study fails to support the common perception that ECs are safer than CCs during pregnancy. Until the safety of prenatal EC use is demonstrated, all pregnant women should be encouraged to abstain from all tobacco products including ECs. The use of advanced methodologies is recommended for future research to more robustly account for shared risk factors.

**Danhyang Lee, The University of Alabama**  
***Semiparametric Imputation Using Conditional Gaussian Mixture Models***

Imputation is a popular technique for handling item nonresponse often found in data application. Parametric imputation is based on a parametric model for imputation and is less robust against the failure of the imputation model. Nonparametric imputation is fully robust but is not applicable when the dimension of covariates is large due to the curse of dimensionality. Semiparametric imputation is another robust imputation based on a flexible model where the number of model parameters can increase with the sample size. We propose another semiparametric imputation based on a more flexible model assumption than the Gaussian mixture model. In the proposed model, we assume a conditional Gaussian model for the study variable of interest given the auxiliary variables, but the marginal distribution of the auxiliary variables is not necessarily Gaussian. This method is applicable to high dimensional covariate problem by including a penalty function in the conditional log-likelihood function. The method is applied to 2017 Korean Household Income and Expenditure Survey conducted by Statistics Korea.

**Yana Melnykov, The University of Alabama**  
***On Finite Mixture Modeling of Change-Point Processes***

Change point estimation in standard process observed over time is an important problem in literature with applications in various fields. We study this problem in a heterogeneous population. A model-based clustering procedure relying on skewed matrix-variate mixture is proposed. It is capable of capturing the heterogeneity pattern and estimating change points from all data groups simultaneously. The appeal of such approach also lies in its flexibility to

model the skewness and dependence in data with good interpretability. Two novel algorithms called matrix power mixture with abrupt change model and matrix power mixture with gradual change model are developed. The approaches are illustrated by simulation studies across a variety of settings. The models are then tested on the U.S. crime data with promising results.

**Igor Melnykov, University of Minnesota, Duluth**  
***An Approach to Initializing the EM Algorithm in Gaussian Mixtures with an Unknown Number of Components***

The EM algorithm is a common tool for finding the maximum likelihood estimates of parameters in finite mixture models. As the algorithm is often sensitive to the choice of the initial parameter vector, efficient initialization is an important preliminary process for the future convergence of the algorithm to the best local maximum. Currently, no initialization method has superiority over others in all practical settings. Considering Gaussian mixture models, we propose a procedure for initializing mean vectors and covariance matrices. The suggested approach can be used in a stepwise manner when the number of components is unknown.

**Jessica Meyer, Optum**  
***Survey Sentiment Analysis Using Open Source Technologies***

Organizations need to continuously improve their products, services, and how they do business in order to stay competitive. Not only do they need to know where to make those improvements, they also need to be able to measure the success of those changes. Getting feedback is essential to this process and the best tool companies have is the survey. The information obtained in the survey's can be used to gauge performance, find areas of improvement and used in a recommender system to suggest other products and services to customers. In this talk, we'll discuss the latest developments in NLP and how I've leveraged open source technologies to build sentiment analysis applications.

**Ally Pelletier, Polaris Industries**  
***Using Prophet Methodology for Forecasting***

Forecasting can be a difficult endeavor but is helpful for understanding trends and goal setting. "Prophet" is a forecasting procedure developed by Facebook and is available in either R or Python. It can be easily adjusted to account for seasonality, holidays, and outliers by either data scientists or data analysts. We'll walk through an example forecast and talk about how to present and use forecasts in a digital setting.

**Arielle Selya, Sanford Research**  
***Data-Driven Healthcare: The Sanford Data Collaborative***

The Sanford Data Collaborative was established by Sanford Research and Sanford Enterprise Data Analytics to help pave the way in data-sharing and collaborative access to real-life, timely health care data. This program, aimed at collaborating with regional and national institutional researchers, opens the door to exploring new and innovative methods to analyzing data, developing mutually-beneficial collaborations, and making an impact on population health and health services delivery. By partnering with external researchers who have unique skillsets and expertise, we enhance the opportunity to view health services delivery through a different lens, heightening Sanford's awareness of available analytic trends, and develop valuable collaborations. Realizing the power of multidisciplinary teams, we can drive innovation behind improving health services delivery and patient outcomes. Previous years' projects include innovations around: predictive algorithms and risk scores, enhancing patient engagement measurement, natural language processing of patient surveys, nursing turnover analysis, emergent care utilization, etc.

## **Murat Sincan, Sanford Imagenetics**

### **Data science infrastructure in the cloud for Sanford Imagenetics, a population scale genetic medicine initiative**

The sequencing of the human genome has facilitated a significant increase in our understanding of disease. By using individual genetic information to prevent, diagnose, and treat disease with better precision, genomics-enabled medicine promises health care that is personalized, predictive, proactive, and preventive rather than reactive (Roundtable on Translating Genomic-Bas . . .). Realizing this goal requires integration and joint analyses of large genomic and clinical datasets which demand computational storage and processing capacities which are not found in a typical healthcare provider organization. To facilitate this type of analyses that are at the care of precision healthcare, Sanford Imagenetics implemented a cloud based platform utilizing commercial Infrastructure as a Service (IaaS) with Amazon Web Services and Platform as a Service (PaaS) with Databricks. Databricks Unified Analytics platform (Databricks Delta: Unified Data Manage . . .) allows Sanford Imagenetics to store and analyze population scale genomic and clinical data in a HIPAA compliant platform that can dynamically scale according to our needs, and with a cost structure proportional to active usage.

## **Dhiraj Sharan, Query AI**

### ***Get Answers and Insights from your Data***

Organizations today have multiple data repositories for myriad use-cases like IT operations, cybersecurity, business and compliance. The challenge though is access, correlation and analysis, especially since skilled analyst time is the most precious resource. In his talk Dhiraj will show an “AI Analyst” technology that lets you talk to your data and get answers and insights you need.

## **Susan Vanderplas, University of Nebraska-Lincoln**

### ***One of these things is not like the others:***

### ***Visual Statistics and Testing in Statistical Graphics***

A statistical chart is a function of the data, and can be treated like any other statistic. Visual inference techniques utilize this conceptualization of charts to examine the impact of different plot design decisions, including chart type and plot aesthetics. This talk will include a description of the use of visual inference in different settings: exploratory analysis, education, and experimentation. We will also consider some of the challenges and recent developments in the use of visual inference to explore the design and use of charts and graphs.

## **Yanshan Wang, Mayo Clinic**

### ***A Simple Introduction to Natural Language Processing and***

### ***Its Clinical Applications in the Era of Artificial Intelligence***

With the rapid adoption of electronic health records (EHRs), it is desirable to harvest information and knowledge from EHRs to support automated systems at the point of care and to enable secondary use of EHRs for clinical and translational research. Following the Health Information Technology for Economic and Clinical Health Act (HITECH Act) legislation in 2009, many health care institutions adopted EHRs, and the number of studies using EHRs has increased dramatically. However, much of the EHR data is in a free-text form. Compared to structured data, free text is a more natural and expressive method to document clinical events and facilitate communication among the care team in the health care environment. Natural language processing (NLP), a subfield of artificial intelligence (AI), has become a critical technique in automatically extracting and encoding clinical information from free text EHRs for clinical decision support, quality improvement, and clinical research. This talk will provide a simple introduction to NLP, along with some real-world clinical applications in the nation's top-ranked hospital, Mayo Clinic.

**David Zeng, Dakota State University**  
***Jupyter Notebook & Python for Data Scientists***

Jupyter is a free, open-source, interactive web browser-based tool known as a computational notebook, which data scientists can use to combine source codes, computational output, explanatory text and multimedia resources in a single document. For data scientists, Jupyter Notebook, combined with Python, has emerged as the de facto standard. As the most popular form of interactive computing, Jupyter notebooks provide an environment in which users execute code, see what happens, modify and repeat in a kind of iterative conversation between the data scientist and data. This introductory presentation demonstrates examples of how Jupyter (with Python) would help data scientists and learners of data analytics. New tools such as JupyterLab and JupyterHub will be introduced as well.

**Rong Zheng, Western Illinois University**  
***Gaussian Mixture Modeling and Model-Based Clustering Under Measurement Inconsistency***

Finite mixtures present a powerful tool for modeling complex heterogeneous data. One of their most important applications is model-based clustering. It assumes that each data group can be reasonably described by one of mixture model components. This establishes a one-to-one relationship between mixture components and clusters. In some cases, however, this relationship can be broken due to the presence of observations from the same class recorded in different ways. This effect can occur because of recording inconsistencies due to the use of different scales, operator errors, or simply various recording styles. The idea presented in this paper aims to alleviate this issue through modifications incorporated into mixture models. While the proposed methodology is applicable to a broad class of mixture models, in this paper it is illustrated on Gaussian mixtures. Several simulation studies and an application to a real-life data set are considered, yielding promising results.

## Contributed Abstracts

**Hassan Alyami, Dakota State University**  
***A Patient-Centered Approach to IT Enabled Diabetes Self-Management: The Case of Saudi Arabia***

Diabetes harms millions of people and causes them long-term health complications that can lead to disabilities and/or death. Diabetes self-management technology solutions concerning behavioral changes have been presented as potential methods to enhance diabetes self-management and self-efficacy and ensure efficiency. Previous research on diabetes self-management has lacked emphasis on patients' abilities to use technology, the design and development of an artifact, and cultural dependency. We will use multi-method, multi-theoretical approaches as our study objectives along two dimensions: 1) Understand the impediments to diabetes self-management with a focus on the challenges patients face in raising the efficiency of using technology, and 2) Design, develop, and evaluate an artifact. The goal of the artifact is to induce behavioral change to improve patients' technology self-efficacy and expectations regarding self-management. In this study, we are basing our artifact on Information system design theory (ISDT) and information-motivation-behavioral skills (IMB), which will allow us to create the appropriate artifact to meet patients' expectations and encourage behavioral changes.



**James Boit, Dakota State University**

**David Zeng**

***A Dense-Inception Network for Medical Image Classification***

Recently, the advent of deep convolutional neural network (DCNN) have led to impressive accomplishments on image recognition tasks which was preceded by the breakthrough performance of Inception network for solving image and recognition tasks. In healthcare economies, deep learning approaches have shown great promise in medical diagnosis and treatment. In this work, we proposed a new model architecture that integrates Dense Blocks into the Inception Module that we refer to as DINET. The novel model captures multi-scale learned features while increasing parameter efficiency throughout the network. Our aim is to investigate the impact of maintaining a balanced tradeoff between model efficiency and effectiveness of preventing the problem of vanishing gradient. For our medical images classification problem, specifically using Chest x-ray images, experiment results using our DINET model shows modest performance improvements. In our future research directions, other versions of DINET modules will be experimented to shed more light on the impact of strategically positioning the DINET modules in the network to achieve the trade-offs of model efficiency and performance while increasing interaction of multi-scale learned features.

**Ruth Burkhalter, University of South Dakota**

***Bootstrap Control Chart for Pareto Percentiles***

Lifetime percentile is an important indicator of product reliability. However, the sampling distribution of a percentile estimator for any lifetime distribution is not a bell shaped one. As a result, the well-known Shewhart-type control chart cannot be applied to monitor the product lifetime percentiles. In this presentation, Bootstrap control charts based on maximum likelihood estimator (MLE) are proposed for monitoring Pareto percentiles. An intensive simulation study is conducted to compare the performance among the proposed MLE Bootstrap control chart and Shewhart-type control chart.

**Gary Hatfield, University of Minnesota**

***Data Science in the Life Insurance Industry***

Data Scientist has emerged as one of the hottest and most talked about jobs in the world today. In my talk, I will provide an overview of how data science has emerged in the insurance industry. I will give some examples of how data science is being applied in life insurance and describe how the Actuarial profession is adapting.

**Paul May, South Dakota State University**

**Hossein Moradi**

***Multi-Resolution Approximations for Precision Agriculture***

Precision agriculture is the leveraging of data for better farming practices. An important aspect of precision agriculture is analyzing the effect of covariates on crop yield. Agricultural data sets can be very large, making likelihood-based inference on traditional spatial models computationally burdensome. The Multi-Resolution Approximation allows for fast inference on Gaussian processes by using a particular covariance structure. We show through a simulation study and the analysis of a real agricultural data set that the Multi-Resolution Approximation can be used to estimate covariate effects with near-identical accuracy as traditional likelihood estimation, and with great computational advantage.

**Raisa Nusrat, University of South Dakota**

**Daniel Elliott and KC Santosh**

***Exploring Task Decomposition with Ensemble Approach using Reinforcement Learning***

In the context of machine learning, Reinforcement Learning (RL) hardly requires an introduction and the ensemble method is an effective approach that can improve the performance of a model. In a supervised learning environment, it has been shown that an ensemble of individual learners is capable of decomposing the input space of a classification task. Our research is intended to ascertain whether a model consisting of an ensemble of Q-learners can decompose a specific control task into distinct components and train itself to designate individual learners for a specific component without any additional information or supervision from the human trainers.

A number of artificial intelligence concepts are based on human behavior and how we learn. Task decomposition also has a relevance to human behavior. For a group of people working at a problem, it often happens that the problem is broken down into discrete smaller jobs which are assigned to the individuals in the group. We intend to investigate if this concept can be translated into RL domain by using an ensemble of Q-learners.

While the relevant reviewed papers indicated enhancement of RL performance with ensemble approach, it does not offer any prospect that might be used for task decomposition, with-out any additional task parameters. With our research we want to utilize this unexplored potential and look into the possibility of an ensemble of learners decomposing the input space of a control task.

**Martinson Ofori, Dakota State University**

**Omar El-Gayar**

***Towards Deep Learning for Weed Detection: Deep Convolutional Neural Networks Architectures for Plant Seedling Classification***

Traditional means of on-farm weed control has been known to use manual labor. This process is time consuming, costly and contributes to major yield losses. There are also environmental hazards with conventional or uniform application method of controlling weed infestation. To solve this using computer vision, researchers often use remote sensing weed maps, but this is ineffective due to problems such as solar and cloud cover in satellite imagery.

In this study we leverage the automatic feature extraction capabilities of deep convolutional neural networks (DCNN) to classify plant seedlings. Theoretically, we demonstrate that DCNNs can successfully segment crops and weeds in various phenological growth stages, and identify limitations with these techniques that can further guide future research. In practice, this paper will be relevant to both researchers and producers of computer vision equipment, especially low-cost solution for ground-based site-specific weed control.

**Ayorinde Ogunyiola, South Dakota State University**

**Maaz Gardzi, Semhar Michael, SDSU, Asif Ishtiaque, University of Michigan;**

**Candace May, SDSU; Ryan Stock, Northern Michigan University; and Sumit Vij, Wageningen University**

Climate smart agriculture (CSA) is a promising approach to securing food security for the growing world population under global climate change. CSA includes technological and managerial innovations that can help poor countries simultaneously achieve three development goals: (1) enhance the ability of people and communities to adapt to climate change (adaptation), (2) reduce greenhouse gas emissions (GHG) from food production and distribution processes (mitigation), and (3) improve food production (food security). The CSA community acknowledges the existence of trade-offs between the three objectives. For instance, increasing the use of agrochemicals (e.g. chemical fertilizers) can help maintain crop yields (increase food security) in the face of climate change, but may also result in greater overall GHG emissions (make mitigation ineffective).

This paper examines how large international organizations—who are actively backing and implementing the CSA agenda—are addressing such trade-offs? We answer this question by using topic modelling to classify and label hundreds of official documents on CSA published by six large international organizations. By understanding how these organizations are operationalizing CSA in over 60 developing countries, this paper generates discussion to facilitate effective achievement of the “triple-wins” or all three objectives of CSA and inform areas where agenda improvements and inter-organizational collaborations are required for moving society toward effective and resilient agricultural practices and technologies.

**Lauren Sauer, University of South Dakota**  
***Stress-strength Inference for the Multicomponent System Based on Progressively type-II censored samples from Pareto Distributions***

A system of  $k$  components, where the strengths of all  $k$  components are independent and have identical distribution and each component is subject to a common random stress, is investigated. This system is alive only if at least  $s$  ( $\leq k$ ) component strengths exceed the stress. This is also called a multicomponent stress-strength problem. In this talk, the maximum likelihood estimate of the multicomponent system reliability and the related confident intervals of the system reliability are presented based on progressively type-II censored samples from Pareto distributions. An intensive Monte Carlo simulation study is conducted to compare the impact from difference progressive censoring schemes.

**Zhiguang Wang – South Dakota State University**  
**Multi-step Forecast of the Implied Volatility Surface using Deep Learning**

Modeling implied volatility surface (IVS) is of paramount importance to price and hedge an option. The literature is abundant in predicting realized volatility and the VIX using time series models, but lack in predicting the whole IVS. Machine learning architectures have shown strengths in learning option pricing formulas and estimating implied volatility cross-sectionally. We attempt to bridge the gap between machine learning-based implied volatility modeling and multivariate multi-step implied volatility forecasting. We contribute to the literature by modeling the entire IVS using recurrent neural network architectures, namely Convolutional Long Short Term Memory Neural Network (ConvLSTM) to produce multivariate and multi-step forecasts of the S&P 500 implied volatility surface. The ConvLSTM model is capable of understanding the spatiotemporal relationships between strikes and maturities (term structure), and of modeling volatility surface dynamics non-parametrically.

**Jixiang Wu, South Dakota State University**  
**Imputing Data without Replication**

It is sometimes important to revisit the historical crop field trial data. However, many historical data are available in a format of entry means under different environments rather than repeated field plot data. In this presentation, I will present a recently proposed methodology, which can be used to impute replicated trial data sets to reveal the original information harbored in the original data. As a demonstration, we used a data set, which includes 28 potato genotypes and six environments with three replications to numerically evaluate the properties of this new method. We compared the phenotypic means and predicted random effects from the imputed data with the results from the original data.

# Poster Abstracts

**Loknath Ambati, Dakota State University**

**Omar El-Gayar**

## ***A Comparative Study of Machine Learning Approaches for Human Activity Recognition***

The goal of this project is to study the performance of Machine Learning (ML) techniques used in Human Activity Recognition (HAR). Specifically, we aim to 1) evaluate and benchmark the performance of various ML techniques used for HAR against established ML performance metrics using multiple datasets, and 2) map the characteristics of various HAR datasets to appropriate ML techniques. From a theoretical perspective, the research will shed light into the strengths and weaknesses of various ML techniques that can provide insights into future research aimed at improving these techniques for HAR. From a practical perspective, the research provides guidance into the applicability of various ML techniques to HAR datasets. Studies into improving HAR performance could lead to a significant improvement in the selfcare and self-management interventions. These improvements would open doors for creative innovations in healthcare and other commercial applications that require the detection of human activity.

**Ganga Prasad Basyal, Dakota State University**

**David Zeng**

## ***Impact of Data Quality and Quantity on Its Effectiveness on Multi-stage Transfer Learning Using MRI Medical Images***

Multi-Stage Transfer Learning (MSTL) has been becoming a very promising area of research in the field of medical imaging. Model architecture based on Multi-Stage TL provided promising results surpassing the previous standards. In our study, we provide an overview of Multi-Stage TL and its implementation in medical imaging followed by reviewing the research work in the field of transfer learning in medical imaging. Our objective is to investigate and understand the different effects of data quality and quantity on Multi-Stage Transfer learning using the MRI images. We propose an MSTL model comprises of four different stages, in the first stage the model adapts the features and weights from a pre-trained network, second stage will include the domain adaptation having similar domain data with previous weights being fine-tuned, third stage is split into three separate layers each investigating the impact of Data Quality, Quantity and image features. In the final stage, we will apply the weights learned from the previous stages into the completely new dataset (Target/Problem area) and analyze its effects. Our study discusses the utilization of Multi-Stage transfer in medical imaging using the CNN architectures such as Inception V-3, AlexNet, and ResNet and investigate the current challenges in medical imaging domain such as computational complexity, domain adaptation and effectiveness of data quality and quantity using TL and Multi-Stage TL and proposed the future research areas.

**Benjamin Derenge, South Dakota State University**

**Maverick Maynard**

## ***Analysis of 2018 Central Park Squirrel Census***

This report is an analysis of data gathered in the 2018 Central Park Squirrel Census. First we will discuss some background information on the data and how the data itself is organized, along with a brief description of the organization that retained this information. Then, we will provide our intention through this project and what we hope to discover through the process. Next, will be an overview of how the data was cleaned and in some cases restructured in order to create the best and most accurate results, in addition, to ease the computational process in

generating plots and statistical inferences from the data, using R and Excel. Next comes the result of our computations, and then a conclusion and summary of the report.

This report was created to inform people across the world of the behavior of squirrels in New York City's Central Park, and to provide additional information about the tendencies squirrels may have in distinct situations. We believe that there exist underlying statistics to the happenings of squirrels in Central Park. This was generated in which people of varying education levels could understand.

**Joshua Eason, Cleveland State University**

**Sathish Kumar**

***Evaluation of Text Mining Techniques Using Twitter Data for Hurricane Disaster Resilience***

Data obtained from social media microblogging websites such as Twitter provide the unique ability to collect and analyze conversations of the public in order to gain perspective on the thoughts and feelings of the general public. Sentiment and volume analysis techniques were applied to the dataset in order to gain an understanding of the amount and level of sentiment associated with certain disaster-related tweets, including a topical analysis of specific terms. This study showed that disaster-type events such as a hurricane can cause some strong negative sentiment in the period of time directly preceding the event, but ultimately returns quickly to normal levels. An analysis of the volume of tweets during the same time revealed that the public responds in near real-time to events with conversation on Twitter. This information can be an effective tool in which to arm emergency management personnel with vital human intelligence information to inform decision-making processes ahead of future storm, or disaster-related events. In addition, this study performed empirical performance evaluation experiments on Latent Dirichlet Allocation (LDA) topic models which were generated from Twitter data collected from Hurricane Florence. The performance evaluation experiments showed that LDA topic models struggle to accurately reflect the true latent conversation topics present within a medium-term, event-based dataset. Although the study successfully modeled LDA topic models, it could not produce models that were interpretable by human beings as distinct groups of topic words that were tightly coupled to one another.

**Rajesh Godasu, Dakota State University**

**David Zeng and Kruttika Sutrave**

***Multi-Stage Transfer Learning System with Lightweight Architectures in Medical Image Classification***

Transfer Learning is currently popular in Medical Image classification. Transfer Learning methods are extensively applied with CNN's such as Res-net, Densenet, VGG16, Inception, etc. for various medical diagnoses. However, these models are computationally expensive and over parameterized. Another challenge we identified is limited labeled datasets are available in the medical image domain preventing the major advancements in Transfer Learning for Medical image classification. We propose a Multi-Stage Transfer Learning System using Lightweight Architecture to tackle limited target dataset problem with quicker training time. Preliminary results suggest that our model performed well on CT Head images by improving the accuracy over traditional single-stage transfer learning.

**Confidence Idim, Minot State University**

**Sayeed Sajal**

***Collaborative Filtering in E-commerce Business***

Since the making of the internet easily available to the public, the amount of data that has been produced by users of the internet has been mind-blowing. Industries like the e-commerce

industry that benefited from the publicizing of the internet have also inherited the problem of the insane amount of data produced. Most companies when created make it their primary aim to ensure customer satisfaction in order to make a profit. To ensure the customer's satisfaction, most of the e-commerce companies had to solve the large data issue which occurs due to the insane amount of data available on the internet. This might sound like a good thing to the general public, but for e-commerce companies it made retrieval of useful information about customer-specific need difficult. In order to solve this problem, e-commerce companies had to invest and research algorithms that could fix the issue. Collaborative filtering was the algorithm to solve the large data issue and it is still being used presently. In this paper, we have discussed the improvement of collaborative filtering. We have also discussed the types of collaborative filtering and how e-commerce has affected present trends in collaborative filtering. We will conclude the paper with the challenges we faced during our research.

**Thomas Jernejcic, Dakota State University**

***An Alternative to the One-Size-Fits-All Approach to ISA Training:  
A Design Science Approach to ISA Regarding the Adaption to  
Student Vulnerability Based on Knowledge and Behavior***

Any connection to the university's network is a conduit that has the potential of being exploited by an attacker, resulting in the possibility of substantial harm to the infrastructure, to the university, and to the student body of whom the university serves. While organizations rightfully "batten down the hatches" by building firewalls, creating proxies, and applying important updates, the most significant vulnerability, that of the student, continues to be an issue due to lack of knowledge, insufficient motivation, and inadequate or misguided training. Utilizing the Design Science Research (DSR) methodology, this research effort seeks to address the latter concern of training by seeking to design a methodology that will sufficiently support the automatic adaptation of security training, which will be based on the assessment of student vulnerability determined by the student's overall Information Security Awareness (ISA) knowledge and computer security behavior.

**Sangam KC, Dakota State University**

***Federated Transfer Learning: Current Issues and New Perspectives***

Federated learning is an advancement of machine learning where the model training is accomplished with a collection of decentralized local data maintaining the data privacy. Transfer Learning is the process of reusing the features learned with a pre-trained model to a different problem domain. We survey the literature on the cutting edge, federated transfer learning which combines the characteristics of both federated and transfer learning. We identify key technical issues in its application areas of healthcare, agriculture, city management and resource Allocation and discuss the challenges which includes system heterogeneity, expensive communication, model poisoning attacks and model aggregation. At last, we discuss the new opportunities including new tools for quantifying heterogeneity, new methods on convergence, bandwidth efficiency and collaborative mobile clustering learning that are promising for future research.

**Ziad Kadry, Minot State University**

**Sayed Sajal**

***Simplifying User Interfaces for Data Science and Machine Learning Applications***

Machine learning is in a growing state as more businesses and individuals realize the power it carries in bringing a deeper understanding to large sums of existing data and make predictions based on discovered correlations that weren't apparent before building a model. To accelerate the growth of this field, simplifying the process of machine learning will potentially lead to increasing the efficiency of the process of machine learning, in addition to lowering the ceiling of

previous knowledge needed to start building models which will bring in more newcomers to the field of data science helping it grow as a community and a science. My research focuses on finding the implications of developing an open-source GUI (Graphical User Interface) set on top of a popular machine learning framework like TensorFlow.

Having a modular open-source GUI based machine learning system built to translate function calls to simple drag and drop operations that could be fit to use any of the popular machine learning specific and general data processing python modules could potentially accelerate the process of building models and reduce the number of human errors involved in manually writing python code. This approach to machine learning will also eliminate the need to learn to code for most of its applications, which could bring in many new students who strayed away from the field due to having to knowcoding concepts. Further research is needed to evaluate the time and cost needed to develop such a framework.

**Shaurya Khurana, University of South Dakota  
KC Santosh**

### ***Can Edge Map Be Sufficient for Deep Learning Models to Understand Chest Radiographs?***

In computer vision, edge maps can help understand line-rich objects, such as buildings and cars when image data are considered. The edge map does not include color/texture properties in it. Organizing their relative positioning of these lines/curves is not trivial, and therefore the use of deep learning could possibly avoid missing information. In this work, using deep learning models, we present the use of an edge map for understanding abnormalities (such as tuberculosis and pneumonia) in chest x-rays. Our results will be followed by the discussion, where we state the primary motivation behind the use of the edge maps, not the textures.

**Zhuoning Li, Minnesota State University Mankato**

### ***U.S. Soybean Market Forecasting Using Statistics & Machine Learning Techniques***

The agricultural product stock market is very stochastic and difficult to predict. The market is especially affected due to different political and economic policies. This year, the soybean trading market has been affected the most due to the trade war between the U.S. and China. According to USDA, 17% of the U.S. agriculture produce exports to China and 62% of those products were soybeans. Thus, the soybean market has a remarkable change from previous years. In this study, Long-Short Term Memory (LSTM), Time Series Regression model and GARCH model are explored to analyze the soybean market. Google trend and other factors are evaluated as important indicators to the market.

**Amul Neupane, University of South Dakota  
KC Santosh**

### ***Foreign Object Detection and Localization in Chest X-rays using Deep Learning***

Pulmonary abnormalities, such as tuberculosis (TB), asthma and/or chronic obstructive are global threats. Nearly 1.6 million died from TB alone according to the 2019 World Health Organization report. Computer scientists together with medical experts have designed and reported automated screening systems for chest X-ray (CXR) images. However, most of the research did not consider detecting foreign objects, such as buttons, coins, ring, pins, bone pieces and other medical devices (e.g. pacemaker) all together that can hinder the performance of automatic screening system. The circle-like foreign objects, such as coins are often confused with nodules, which is one of the primary indicators of tuberculosis. Thus, in an automated screening process foreign objects need to be separated. Unlike the previous works, we will employ deep learning models, such as Faster R-CNN (Faster Region Proposal Convolutional Neural Network) to detect almost all kinds of foreign objects in CXR images. This research is



focused on the detection of foreign objects that are of almost all shapes, sizes and texture in CXR using convolutional neural network. Instead of relying on handcrafted features, we now let machine to find distinguished features to achieve an error as low as possible (technically,  $10^{-4}$ ). We also localize their spatial position in CXR, so that the further process of screening can be advanced and at the same time misdiagnosis and confusion can be eliminated.

**Ayorinde Ogunyiola, South Dakota State University**  
**Maaz Gardzi, Semhar Michael, SDSU, Asif Ishtiaque, University of Michigan;**  
**Candace May, SDSU; Ryan Stock, Northern Michigan University;**  
**and Sumit Vij, Wageningen University**

Climate smart agriculture (CSA) is a promising approach to securing food security for the growing world population under global climate change. CSA includes technological and managerial innovations that can help poor countries simultaneously achieve three development goals: (1) enhance the ability of people and communities to adapt to climate change (adaptation), (2) reduce greenhouse gas emissions (GHG) from food production and distribution processes (mitigation), and (3) improve food production (food security). The CSA community acknowledges the existence of trade-offs between the three objectives. For instance, increasing the use of agrochemicals (e.g. chemical fertilizers) can help maintain crop yields (increase food security) in the face of climate change, but may also result in greater overall GHG emissions (make mitigation ineffective). This paper examines how large international organizations—who are actively backing and implementing the CSA agenda—are addressing such trade-offs? We answer this question by using topic modelling to classify and label hundreds of official documents on CSA published by six large international organizations. By understanding how these organizations are operationalizing CSA in over 60 developing countries, this paper generates discussion to facilitate effective achievement of the “triple-wins” or all three objectives of CSA and inform areas where agenda improvements and inter-organizational collaborations are required for moving society toward effective and resilient agricultural practices and technologies.

**Lawrence Sethor Segbeho, South Dakota State University**  
**Frank Schaarschmidt and Gemechis Djira**  
***Asymptotic Simultaneous Estimations for Contrasts of Quantiles***

Although the expected value is popular, many researches in the health and social sciences involve skewed distributions and inferences concerning quantiles. Most standard multiple comparison procedures require the normality assumption. For example, few methods exist for comparing the medians of independent samples or quantiles of several distributions in general. To our knowledge, there is no general-purpose method for constructing simultaneous confidence intervals for multiple contrasts of quantiles. In this paper, we develop an asymptotic method for constructing such intervals and extend the idea to that of time-to-event data in survival analysis. Small-sample performance of the proposed method is assessed in terms of coverage probability and average width of the simultaneous confidence intervals. Good coverage probabilities are observed for most of the distributions considered in the simulations. The proposed method is applied to biomedical data and time-to-event data in survival analysis.

**Omar Sharif, USD**  
**KC Santosh**  
***Understanding Chest X-rays Using Key Points***

Understanding images can definitely help screen chest x-rays, where abnormality comes into play. Automating chest x-ray screening is crucial, where we do not have resources, such as hospitals and radiologists. In this study, we consider the use of key points that are detected

based on the texture changes in the chest x-rays (primarily due to abnormalities e.g. tuberculosis) and check whether these key points can be considered as a tool for screening them. Our results will be discussed with previously reported works (on benchmark datasets).

**Brent Van Aartsen, Dakota State University**  
**Omar El-Gayar**

### ***Systematic Review of Web Usage Mining Techniques and Future Research Options***

Web usage mining (WUM) is an application of data mining techniques on web log data in order to understand the who, what, why, and how of those using a website. Through this systematic review, we look at the research of WUM techniques from 2014 - 2019 in order to understand the current state of WUM research as well as answer our research questions. Our research questions are (RQ1) what data sources are used in web usage mining, (RQ2) what data analysis methods are used to extract the knowledge, (RQ3) what are the applications of Web usage mining, and (RQ4) what future research can be done in the web usage mining area? Using a PRISMA approach to narrow the initial 778 search results, we completed a full analysis of 68 unique articles from four databases: Web of Science, ProQuest, ScienceDirect, and IEEE Xplore. Our article searches focused on the keywords (i) “web usage mining,” (ii) “WUM,” and (iii) “web usage AND mining.” The completion of the article analysis revealed research into WUM is on the decline. The analysis also revealed Personalization and Recommender Systems are the two most heavily researched applications of WUM.

**Alexis VanderWilt, Dakota State University**  
**Cherie Noteboom**

### ***Do Demographics and the Type of Data Visualization Influence the Interpretation of Data?***

The literature review indicates that as students get older, they improve their critical thinking skills. Data can be hard to explain with words or numbers, so there needs to be a way to present data that allows it to be translated into information and knowledge that can be used to guide people to make well-informed decisions. The data should be presented with a medium that aids each person in making these decisions. We used survey and data visualization research to create a survey to investigate our research question with different data visualizations. We considered types of graphs, colors, fonts, and styles when creating our visualizations. The responses to our survey were analyzed to determine the difference between undergraduate and graduate students' responses to see if the level of life experience and age influences the correct interpretation of the data visualizations. Future data analysis will allow us to analyze if there is a significant difference in the interpretation of the data visualizations based on the subgroups of gender and age.

**Simon Weller, Concordia College**

### ***An Empirical Analysis of Best Practices for Sugar Beet Growth in the Southern Red River Valley for the Minn-Dak Farmers Cooperative***

This project involves analyzing data for the Minn-Dak Farmers Cooperative (MDFC) in Wahpeton, N.D. MDFC is owned by approximately 500 shareholders/growers who collectively grow and harvest 115,000 acres of sugar beets every year. These sugar beets are then processed at the MDFC's plant in Wahpeton, and the resulting sugar is distributed and sold throughout the region. MDFC's goal then is to maximize profits by optimizing sugar beet production through the efficient use of land, which is a limited resource. The goal of this project is to analyze historical agricultural and geographic data to aid MDFC in helping their growers make decisions that will lead to increased sugar beet production. The project primarily involves spatial analysis and multiple linear regression to determine which agricultural inputs are most effec-

tive for the increase of production. The agriculture industry is a highly volatile industry in that it relies so heavily on consistent weather. There will always be a level of risk and uncertainty when it comes to agriculture. It is important to note that these findings can only make general predictions, and that there is going to be uncertainty. Actual results can vary due to unforeseen circumstances such as unexpected droughts, flooding, or new diseases.

**Shuk Ping Wong, Minnesota State University Mankato**  
***Bank Loan Default Predictive Models with***  
***Logistics Regression & Support Vector Machine***

Risk management is one of the most crucial areas for banks. Banks are constantly working on effective models to estimate the likelihood of whether a customer could default to maintain a sustainable and profitable business. Although credit scoring is a common indicator for bankers, some financial datasets simply do not come with this variable. This study builds a logistic regression model and a support vector machine (SVM) model to predict whether the loan borrower will default based on different categorical variables. The performance of the models is compared based on accuracy and efficiency. The importance of variables is ranked as a discussion with the result.

**Zhuoyu Yang, Minot State University**  
**Sayed Sajal**

***Predicting severity and frequency of automobile accidents,***  
***and identification of accident hotspots in the U.S.***

Americans are now driving more than ever (U.S. Department of Transportation, 2019). In 2010, close to 33,000 lives were lost and another estimated 3.9 million people were injured in automobile accidents; all things considered, these accidents accounted for \$836 billion in damages. Since then, the rate of automobile-related deaths per 100 million miles traveled has not shown signs of improvement (National Center for Statistics and Analysis, 2019). This research aims to conduct an exploratory data analysis on a dataset containing 2.25 million automobile accident records collected over a span of three years from February 2016 to March 2019, to help predict the severity and frequency of traffic accidents, as well as to identify potential accident “hotspots” across the U.S.

**Rong Zhou, South Dakota State University**  
**Zhaohui Xu and Dazhi Meng**

***Cascading Failure Mechanism in Biological Systems***

We establish the genes correlation networks of three species under different physiological states, normal soybean and those with root rot disease, anaerobic and aerobic growth of yeast and prostate cancer of human, based on the gene expression profiles. After investigating these networks using a cascading failure model, we find that the dynamic stability of the networks under different states differ significantly. Furthermore, the structural key genes which contribute greatly to these differences are identified. Finally, the biological functions of the key genes which may result in the root rot disease of soybean are annotated. This helps to reveal the relevant functional mechanisms that might be responsible for the changes and thus provides a useful tool toward understanding the mechanisms of various life processes.

# Workshop Abstracts

## ***Workshop One: Data Visualization***

**Dr. Xijin Ge, South Dakota State University**

This hands-on workshop assumes that the attendees have basic knowledge of R. The goal is to introduce beginners and intermediate level R users to some packages that can easily generate high-quality, interactive, or web-based graphics. Please make sure you have a recent version of R and Rstudio installed, and also install these packages below (plus devtools and tidyverse) ahead of time. Main topics include:

1. Some basic principles of effective data visualization
2. ggplot2, a quick introduction.
3. Simplifying ggplot2 with ggfortify
4. Interactive plots with plot\_ly and ggplotly
5. Streamline EDA with DataExplorer and GGally
6. Heat maps with complexHeatmap visualizes data matrices  
(Install via [www.bioconductor.org](http://www.bioconductor.org))
7. VennDiagrams and UpSetR visualizes overlapping sets
8. Introducing animations to your plots with gganimate \*
9. Shiny and Dashboards\*
10. Interactive network visualization with visNetwork\*
11. Other tricks and cools plots I learned from my students\*

\* *Time permitting*

## ***Workshop Two: Open Source Tools for Web Scraping***

**Peter Claussen, Gylling Data Management, Inc., and SDSU**

Web scraping or web mining involves interacting with distributed files and information systems through abstract interfaces, where the analyst has little direct control over the computer hardware or services. Programming practices that support web scraping include:

- Language-independent file transfer protocols (i.e. HTTP)
- Self-documenting document structuring languages (HTML, XML, JSON)
- Abstract programming interfaces (API) through which data providers allow systematic queries to data repositories
- Text mining via pattern matching (regular expressions)

This workshop will cover open-source tools available to assist with these practices, with an emphasis on libraries that can be interface via either Python or R

## **Workshop Three: R package creation fundamentals** **Md Riaz Khan**

The goal of the workshop is to learn about the fundamentals of creating an R package using R Markdown. Main topics include:

- R package- what and how
- Preparing the system for package building
- Package creating and package metadata
- Writing functions
- Package documentation using roxygen2
- User defined class and S3 methods
- Adding example dataset to the package
- Vignette using R Markdown
- Package testing and release

### *Prerequisite*

- Basic knowledge of R programming
- Most recent versions of R and R Markdown installed

# Symposium Committee

A big thank you to everyone involved with organizing the 2020 Data Science Symposium. Your commitment and dedication made this event possible!

### *Department Head*

**Kurt Cogswell**

### *Conference Chair*

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

### *Committee Members*

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

**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**  
**Michael Biondo**  
**Josilyn Ulvestad**

### *Volunteers*

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

**Tye Klinnert**  
**Sierra Lutz**  
**Rachel MacDowell**  
**Mohan Manamel**  
**Paul May**  
**Samantha Nystrom**  
**Karissa Palmer**  
**Saraswati Rimal**  
**Samjhana Shakya**  
**Slavik, Timothy**  
**Negassi Tesfay**  
**Roberto Villegas-Diaz**





**SOUTH DAKOTA  
STATE UNIVERSITY**

---

**Department of  
Mathematics and Statistics**