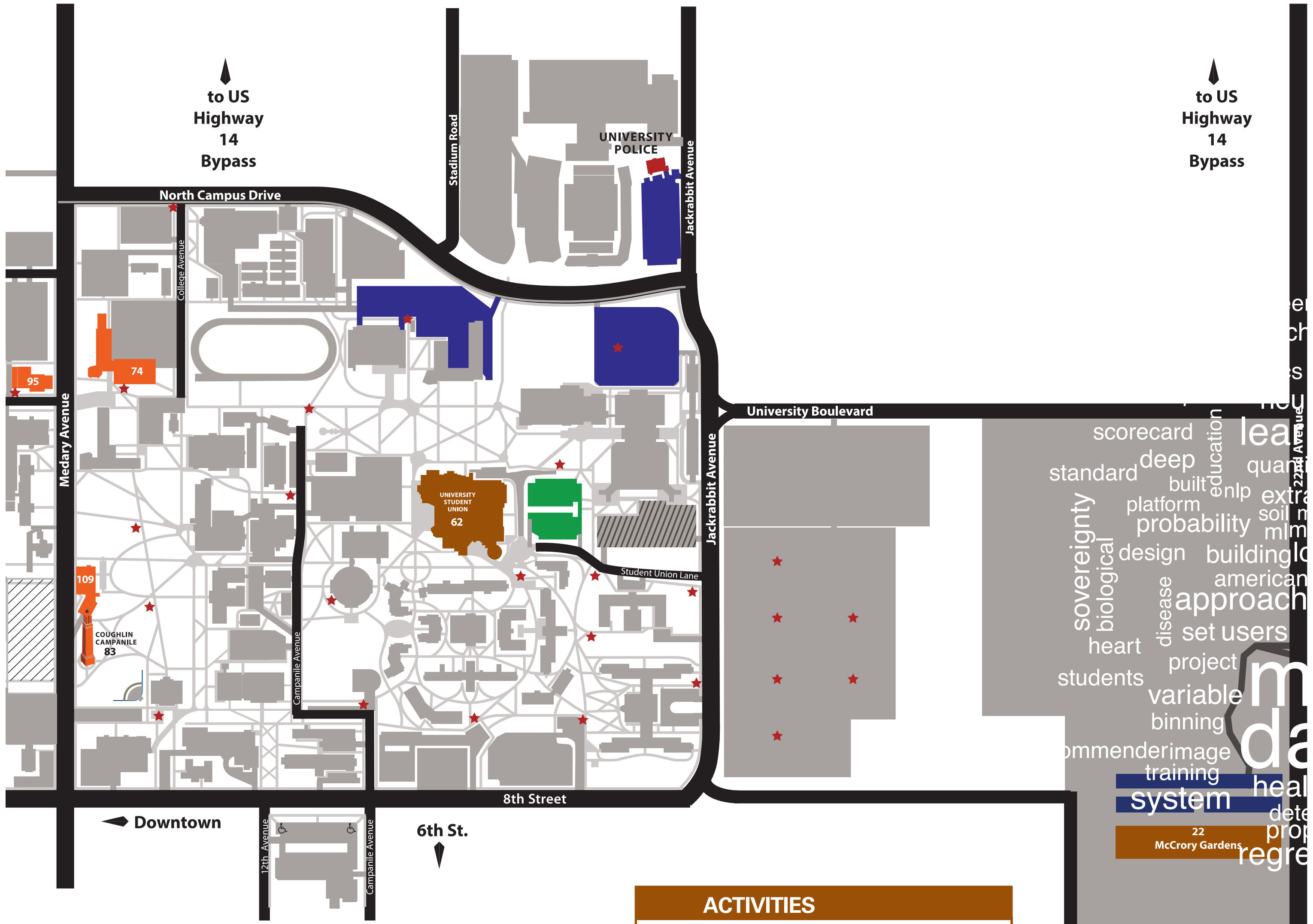
South Dakota State University

2018 SDSU Data Science Symposium

February 11 - 12, 2018

SDSU Campus Map



N
W E
S

Legend

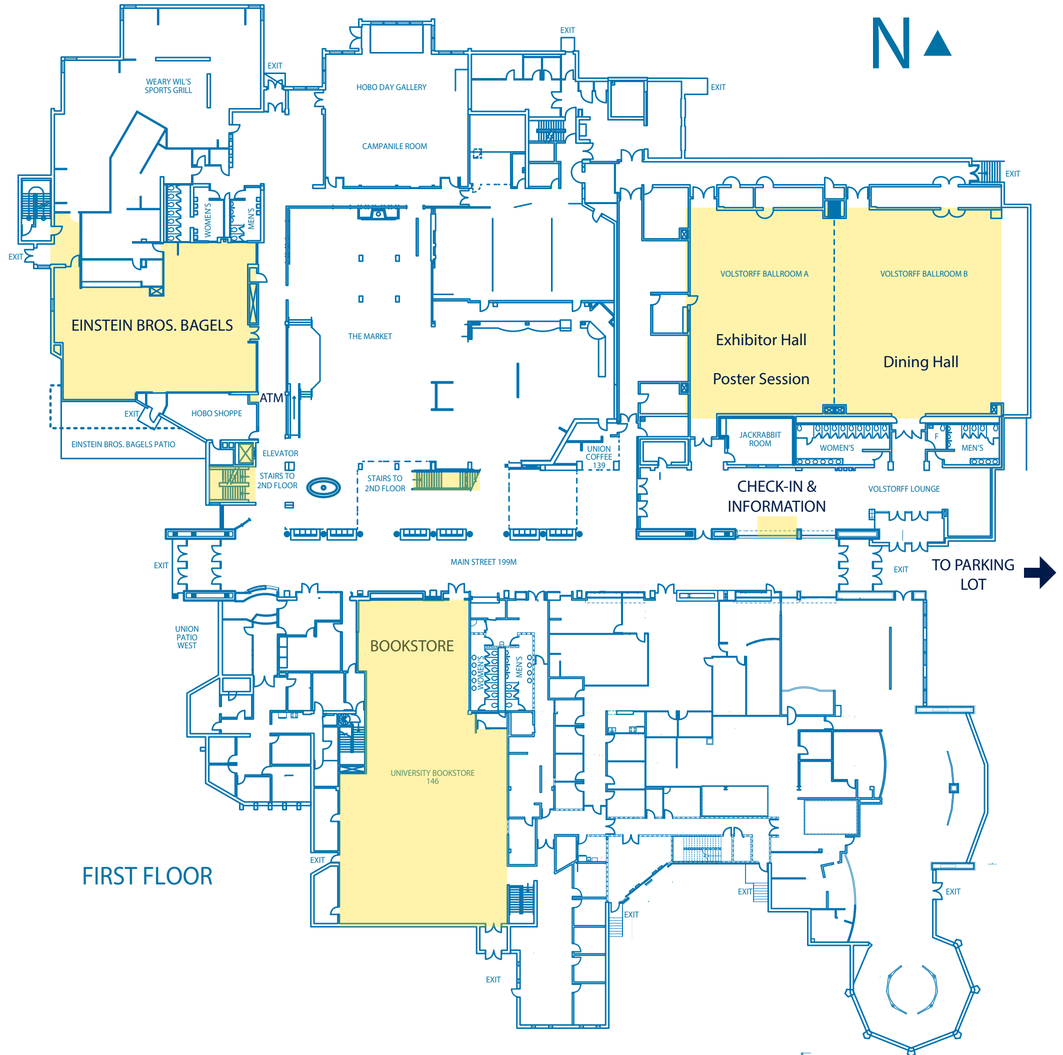
- 🟫 Activities
- 🟥 University Police
- 🟧 Attractions
- ★ Emergency Call Box

ACTIVITIES	
62	University Student Union
22	McCrory Gardens 631 22nd Ave, 57006

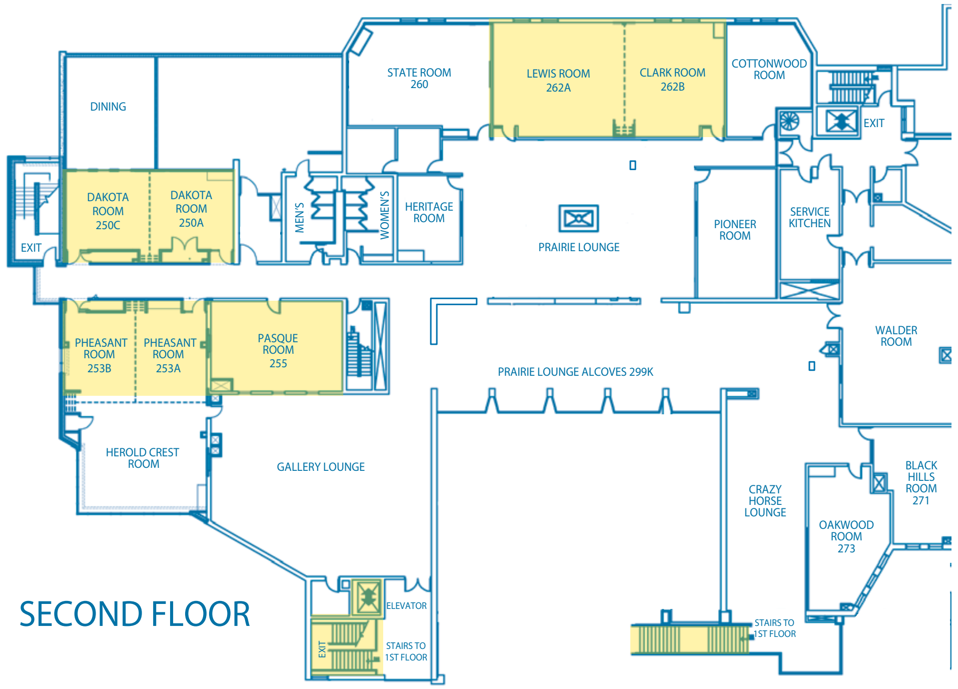
PARKING	
🟩	Union Lot Student Union Lane, 57007
🟦	Parking

ATTRACTIONS	
74	Dairy Bar
83	Coughlin Campanile
95	Agricultural Heritage Museum
109	South Dakota Art Museum

Union Map



FIRST FLOOR



SECOND FLOOR

SOUTH DAKOTA STATE UNIVERSITY
Be Great. Start Here.™

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. 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

SOUTH DAKOTA STATE UNIVERSITY

Dear Attendees,

As President of South Dakota State University, I am happy to welcome you to the 2018 Data Science Symposium. SDSU and its Mathematics and Statistics Department are pleased to offer this opportunity for students, scholars, and practitioners of data science. This is an excellent opportunity to come together to discuss the latest developments in data science and how they might work to the benefit of all in this state, our region, and beyond.

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

Sincerely,

Barry H. Dunn

Barry H. Dunn, President

SOUTH DAKOTA STATE UNIVERSITY

2018 SDSU Data Science Symposium

Dear Attendees,

On behalf of the SDSU Department of Mathematics and Statistics, it is my pleasure to welcome you all to the 2018 SDSU Data Science Symposium. With a remarkable array of high-impact speakers, there will be many opportunities for productive discussion and learning that I hope will be of value to us all. My thanks go out to all the speakers and poster presenters, and of course to the organizing committee as well for their great work in making this symposium possible.

Sincerely,

Kurt D. Cogswell

Kurt D. Cogswell, Department Head

Sponsors

Platinum

CAPITAL SERVICES

Gold

MetaBank

SANFORD HEALTH

Silver

tci TOTAL CARD, INC.

R Studio

jmp Statistical Discovery.™ From SAS.

Bronze

GREAT WEST CASUALTY COMPANY
The Difference is Service®

Fishback Financial Corporation

BULLDOG MEDIA GROUP
A DIGITAL MARKETING COMPANY

POLARIS Data Science

General Information

If you need proof of attendance, please send a request to sdsu.seminars@sdstate.edu

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

Registration/Check-in Hours
Located at the Union in the Volstorff Lounge.
Monday, February 12 7:30 am – 5:30 pm

Luggage Check Hours
Located at the Union in the Volstorff Lounge.
Monday, February 12 7:30 am – 5:30 pm

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

Parking
Symposium Parking
Attendees are encouraged to park in the Union Pay Lot, which is located to the east of the University Student Union, Student Union Lane, Brookings, SD 57007. Parking is complimentary. Attendees must ender the code 1896# to enter the pay lot.

Transportation
BATA Bus
Visit BATA's website for more information about their services.
www.brookingsareatransit.com

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

Dining
All meals will be at the Union, in Volstorff B. Other on campus options are available in the Union.

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

Avera Medical Group Brookings
400 22nd Ave. S, Brookings, SD 57006
Monday – Friday: 8 am – 7:30 pm
Sunday: Closed
(605) 697-9500

Sanford Health Brookings Clinic
922 22nd Ave S, Brookings, SD 57006
Monday – Friday 8 am – 8:30 pm
Sunday Closed
(605) 697-1900

Brookings Health System
300 22nd Ave, Brookings, SD 57006
Open 24 hours
(605) 696-9000

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

SDSU Bookstore
Data Science attendees receive a 20% discount on their purchase. Please show your Data Science ID badge.
Monday - Friday 8 am - 6 pm
jackrabbitcentral.com

Printing Services
BluePrint Design & Print
Union
Monday – Friday: 8 am – 5 pm
(605) 688-5496

ATM
Union

Keynote Speakers

Steve Cross

With over 22 years of analytical experience in Consulting, Process Improvement, Solution Design, and Data Development, Steve 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.

Steve came to Great West Casualty with experience in a variety of vertical markets, specializing in financial, insurance, credit services, healthcare, government, marketing, and automotive applications. He has been a speaker at the DMA, NCDM, as well as dozens of corporate conferences.

At Great West, Steve 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. Steve has an undergraduate degree in Mathematics/Economics from Bradley University and graduate degrees in Statistics and Genetics from University of Nebraska Lincoln.

Benson Hsu

Benson Hsu, MD, MBA, is the Chief Medical Analytics Officer at Sanford Health. In this role, he serves as a senior advisor to corporate leadership on the strategic application of healthcare analytics. Previously, he served as the Vice President of Enterprise Data and Analytics, leading the centralized data and analytics team of over 70 members with responsibility over the application of data and analytics across all lines of business including clinical delivery, quality, population health, health plan, and business strategy at Sanford Health – a $4.2 billion integrated health system in the Upper Midwest with 40+ hospitals, 275+ clinics, 1,400+ physicians, 28,000+ employees, and a health plan with 180,000+ members serving over 2.5 million people in 300 communities across 250,000 square miles.

Dr. Hsu continues to maintain an active practice as a board certified pediatric critical care physician and serves as the Medical Director for the Pediatric Transport program. He has a background in health services research and is an Associate Professor of Pediatrics at the University of South Dakota with over 50 peer reviewed publications and abstracts. Additionally, he has presented in over 25 regional, national, and international conferences. Prior to his medical career, Dr. Hsu was a management consultant in the healthcare, Internet, and financial arenas.

As a clinical leader, Dr. Hsu has held numerous leadership positions in both state and national medical societies. He currently serves on the Executive Committee for the Section on Critical Care at the American Academy of Pediatrics and develops policy statements for the Academy as a member of the Committee on Hospital Care.

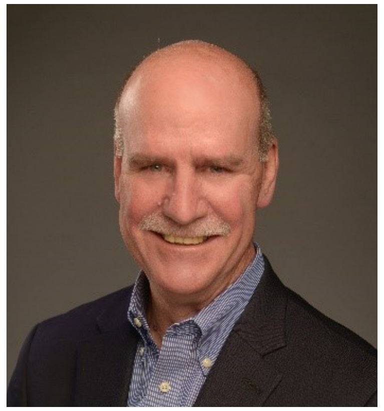
Dr. Hsu received his AB in Economics from Princeton University and his MBA from Duke University Fuqua School of Business, graduating with highest honors as a Fuqua Scholar. He earned his MD from the University of Missouri School of Medicine and completed post-graduate training in Pediatrics and Pediatric Critical Care Medicine at the University of Wisconsin-Madison.

Keynote Speakers

Dr. Gerald Fahner, Senior Principal Scientist, FICO

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

Invited Speakers



Mark Gorman

Mark Gorman is CEO and Founder of The Gorman Group Consultancy. He has nearly 20 years of experience in applying big data, data science and data analytics to both strategic and operational business issues. Mark’s focus is on assisting business organizations and the technology companies supporting them in increasing their adoption of emerging data science and business intelligence technologies.

As a former business executive, Mark has been responsible for setting strategic direction in identifying, developing and delivering successful new products and services to the market. While doing so, Mark gained deep experience and expertise in assuring his clients could rely on organizationally specific business requirements, operationally applicable analytic algorithms and technologically appropriate support capabilities. He is committed to helping firms take advantage of emerging data and technology solutions that will drive profitability, revenue growth, operational efficiency, and customer satisfaction.

Before forming The Gorman Group, Mark authored multiple research white papers, and provided industry research and advisory services for TowerGroup, a financial research and advisory services firm out of Needham, MA. Prior to that, he worked for Fair Isaac, a vendor of solutions based on predictive analytic algorithms and decision support software. At Fair Isaac he was responsible for setting strategic direction in the development and deployment of automated decision management solutions both in the U.S. and select international insurance markets.

Prior to working for Fair Isaac, Mark worked for two leading Life and Annuity insurance companies, a Property and Casualty subsidiary, and their Asset Management and Brokerage subsidiaries in a variety of marketing, product management, and product research and development roles.

Matt Nissen

Matt Nissen is the Business Planning Manager at CAPITAL Services. In this role, Matt develops collection's strategies on delinquent accounts, CAPACITY forecasts for both Customer Service and Collection's, and Queuing Theory models for Workforce Management Planning. Matt was a Data Scientist for CAPITAL Services before Business Planning Manager. As a Data Scientist, Matt built Marketing Scorecards and Financial ProForma Models. Matt holds a Master's in Statistics from SDSU and received an undergraduate degree in Mathematics with a minor in Statistics and Economics from SDSU.

Kevin Potcner

Kevin Potcner is an Academic Ambassador for JMP Statistical Discovery Software. His primary experience has been supporting engineers and scientists on how to extract the most value from their data through teaching and consulting. Kevin has held academic positions at The Rochester Institute of Technology and the University of Florida, and is currently an adjunct professor at the University of San Francisco and an instructor for California State University Data Science program. He is on the Editorial Boards of the International Journal of Data Sciences, Internal Journal of Business Analytics & Intelligence, and Applied Marketing Analytics Journal. He holds a BS in Printing Sciences and an MS in Applied Statistics both from The Rochester Institute of Technology.

Emily Griese - Sanford Health

Emily Griese, PhD is the Director of Collaborative Research at Sanford Health. In this role, she works across Sanford's Research, Enterprise Data and Analytics, and Quality arms to support coordinated population health strategies throughout the enterprise. She has worked to establish and currently directs the Sanford Data Collaborative, a first of its kind data sharing initiative with academic partners to innovate and improve the way healthcare is provided to the patients and communities Sanford serves.

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

Ryan Swanstrom

Since creating the first data science specific blog on the internet in 2012, Ryan Swanstrom has been named as a thought-leader in big data and listed as one of the most influential people on the internet for data science. Ryan currently works as the Director of Data Science for Unify Consulting, where he helps companies use data to solve problems. He lives in South Dakota with his bride and five children. You can follow Ryan's Data Science 101 blog at http://101.datascience.community or follow him on Instagram @ryan.swanstrom, Twitter @ryanswanstrom or LinkedIn.

David Zeng

Dr. David Zeng received his PhD in Information Systems at University of California, Irvine, in 2011. Currently he is an assistant professor/researcher in the Masters of Science in Analytics Program in College of Business and Information Systems, Dakota State University, South Dakota. He teaches Predictive Analytics, Programming for Data Analytics (Python) and Deep Learning for future Data Scientists. Dr. Zeng's research interests include Applications of Machine Learning (Deep Neural Networks) in Healthcare, Deep Neural Networks in Contests/Games, Game Theory and Information Economics of IT-enabled Services. His current research projects include "Deep Learning for Pre-test Probability of Coronary Artery Disease", funded by Faculty Research Initiative Grant, 2017~2018, at DSU and "Learning High-dimensional Representation of Healthcare Concepts.

Nirav Bhagat

Nirav Bhagat is CEO and Founder of Metafunding, Inc, a marketing and IT services company for financial services clients. Metafunding combines marketing strategy, quantitative analysis and software engineering talent to help its clients (profitably) grow. Nirav has held leadership and analytics positions at RetailMeNot and PwC. MBA, University of Chicago. BSEE, Rice University.

André de Waal - SAS

André de Waal received his Ph.D. in theoretical computer science from the University of Bristol during 1994. He spent the next year in Germany and Belgium continuing his research in Logic Programming and Automated Theorem Proving. During 1996 he returned to South Africa to take up his position as lecturer at the School of Computer Science and Information Systems at the then Potchefstroom University for Christian Higher Education (which later became the North-West University), where he was later promoted to Associated Professor. During 1999 he became one of the founder members of the Centre for Business Mathematics and Informatics at the same university. He became responsible for the Data Mining Program in the Centre and shifted his research focus to include Neural Networks and Predictive Modeling. He joined SAS Institute in Cary, NC during December 2010 to take up the position of Analytical Consultant in the Global Academic Program.

Ally Pelletier - Star Tribune

Ally is currently working as a Data Scientist for the Star Tribune in Minneapolis, MN. She is involved in reporting and modeling for the digital department at the Star Tribune. Her current work includes recommendation models, customer retention, and map development. Previous to working at the Star Tribune, Ally 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. Ally earned a BA in Mathematics Education from Concordia College in Moorhead, MN and an MS in Statistics from South Dakota State University. During her time at SDSU she held a research assistantship and an internship in digital media. In her research she developed new statistical power calculations for measuring mixtures of non-normal distributions to measure the profitability in A/B testing in credit card customer behavior.

Invited Speakers

Jason Rogowski – Polaris

Jason Rogowski completed his mathematics degree at Gustavus Adolphus College and masters in Biostatistics at the University of Minnesota, starting his career in the medical device industry. After several years designing and measuring clinical trials, he changed career paths and moved into customer and marketing analytics at Deluxe Corporation, then Olson, a marketing analytics agency. He's been at Polaris for 3 years, first as the manager of the digital analytics team, now the Director of Data Science, guiding a team whose goal is to drive business efficiency with predictive analytics for everything from manufacturing processes to targeted marketing.

Sunday | February 11

Time	Great Hall	Meeting Rooms West/East
12:00 - 5:00 pm	**Check-in/Registration** *(Location: Foyer)*	
1:00 - 2:45 pm	Workshop 1 *Data Science in the Cloud with Microsoft Azure ML* Ryan Swanstrom	Workshop 2 *Python for Data Science* David Zeng
3:00 - 5:00 pm		
	Banquet *(Great Hall)*	
6:00 - 6:30 pm	**Social Time** *(Cash Bar)*	
6:30 - 8:00 pm	**Dinner**	
7:15 - 7:30 pm	**Welcome**	
7:30 - 8:30 pm	**Keynote: Steve Cross** *Data and Analytics - What you don't know can HELP you!*	

Workshop Information

Workshop Information
February 11, 2018
McCrory Gardens

The goal is to introduce participants to new topics in a concentrated way that allows the participant to get "hands on" in that technology.

Workshop 1: Data Science in the Cloud with Microsoft Azure ML
Presented by Ryan Swanstrom, Director of Data Science for Unify Consulting

Azure ML is a web-based tool for machine learning. It has a simple drag and drop interface, yet it is still powerful enough to integrate with R and/or Python. This course will teach you how to create experiments with Azure ML. This course is laboratory based and will include time to build machine learning experiments. You will leave the course with everything you created.

Workshop 2: Python for Data Science
Presented by Dr. David Zeng, Assistant Professor/Researcher in the Masters of Science in Analytics

Outline of the Workshop
1. Setup and Python/Jupyter Notebook Basics
2. 10 minutes to Pandas/DataFrame (actually about 30 mins)

Monday | February 12

Time	Clark Room 262 B	Pheasant Room 253 A/B	Dakota Room 250 A/C	Pasque Room 255
7:30 - noon	**Check-in / Luggage Check** *(Location: Volstorff Lounge)*			
8:00 - 9:00 am	**Breakfast** *(Location: Volstorff B)*			
9:00 - 9:10 am	**Opening Session** *Welcome and Introduction : President Barry Dunn (Location: Volstorff B)*			
9:10 - 10:00 am	**Keynote:** *Combining Machine Learning with Domain Expertise for FICO® Score Development* *Gerald Fahner - FICO (Location: Volstorff B)*			
10:00 - 10:50 am	**Panel Discussion** Moderator: Mark B. Gorman *Making Analytics Great Again (Location: Volstorff B)*			
10:50 - 11:00 am	**Networking break / Exhibitors** *(Location: Volstorff A)*			
11:00 am - 12:00 pm	**Session 1: Tools** Chair: Dheeman Saha *Use case of Amazon Web Services SageMaker in Digital Analytics* Ally Pelletier - Star Tribune *Setting up Python and R in classroom* Anton Bezuglov - BVU	**Session 2: Healthcare** Chair: Quinn Fargen *Collaboration and Innovation: Pushing forward Data-Driven Population Health* Emily Griese - Sanford Health*	**Session 3: Methods** Chair: Jyotshana Paudyal *Table 1: SAS and R tools to build summary statistics tables* Paul Thomson - Sanford Health *Using Atypicality to Identify Outliers* Austin O'Brien - DSU	**Session 4: Financial/ Methods** Chair: Juan Xie *Enhancing Collection Effectiveness through Net Lift Analysis* Matt Nissen - Capital Services *Estimating the Quantile Elasticity of Intertemporal Substitution with Instrumental Variables Quantile Regression* Lance Cundy - UoI
12:00 - 1:00 pm	**Lunch** *(Volstorff B)* **and Poster session** *(Location: Volstorff A)*			

*** 50 minute talks**

Monday | February 12

Time	Clark Room 262 B	Pheasant Room 253 A/B	Dakota Room 250 A/C	Pasque Room 255
1:00 - 2:00 pm	**Keynote:** *Navigating Data Integration in Healthcare: What, Why, How, and Who?* *Benson Hsu - Sanford Health (Location: Volstorff B)*			
2:00 - 3:00 pm	**Session 5: Tools** Chair: Runan Yao *JMP and the Predictive Modeling Workflow* Kevin Potcner - JMP*	**Session 6: Healthcare** Chair: Eric Bae *Analyzing ligand-binding poteins using their structural information* Galkande Premarathna - MNSU *Mining Users Feedback: Discovering the Gaps in Mobile Patient Portals* Cherie Noteboom - DSU	**Session 7: Methods** Chair: Damon Bayer *Why Data Science is Difficult* Ryan Swanstrom - Unify Consulting *Modeling Vegetation Growth on Termite Mounds* Matthew Biesecker - SDSU *Shiny app as a solution for streamlining complex bioinformatic analyses* Xijin Ge - SDSU	**Session 8: Financial** Chair: Riaz Khan *Using Attribution Modeling to Find Profitable Lending Customers* Nirav Bhagat - Metafunding*
3:00 - 3:30 pm	**Networking break / Exhibitors** *(Location: Volstorff B)*			
3:30 - 5:00 pm	**Session 9: Tools** Chair: Riaz Khan *Sentiment Analysis of Donald Trump's Tweets* André de Waal - SAS* *Building R Package* Yuhlong Lio - USD	**Session 10: Healthcare** Chair: Girma Ayana *No Show Predictive Model: A Bayesian Approach* Robert Menzie - Sanford Health *AI in Healthcare: Automated Chest X-ray Screening* KC. Santosh - USD *Deep Neural Networkds to Predict Self-perception of Cardiovascular Disease ~ A Technical Demo* David Zeng - DSU	**Session 11: Methods Towards A Distant** Chair: Abdelbaset Abdalla *Supervision Paradigm for Clinical Information Extraction: Creating Large Training Datasets for Machine Learning* Yanshan Wang - Mayo *Making the analytics journey without getting lost in the cloud* Jason Rogowski - Polaris *Selecting Categorical and Quantitative Variables in Linear Regression Analysis* Jixiang Wu - SDSU	**Session 12: Applications** Chair: Damon Bayer *Prediction on Oscar Winners Based on Twitter Sentiment Analysis Using R* Sayeed Sajal - Minot state *Predicting the Origins of Artwork Found in Rural Churches* Nathan Axvig - Concordia *An In-Class Geospatial Data Analytics Project Inspired by a Comedian* Russ Goodman - Central College
5:00 - 5:30 pm	**Closing Session** *Thomas Brandenburger (Location: Volstorff B)* Winners of poster announcement			

*** 50 minute talks**

Gerald Fahner (FICO)
Combining Machine Learning with Domain Expertise for FICO® Score Development
Machine learning models—often regarded as a black boxes—are attractive due to their high degree of automation and predictive power, whereas highly scrutinized credit scoring operations must employ transparent models. We investigate the potential of modern ML techniques, and of blending ML and Scorecard technology, to ascertain maximum predictive power of the FICO® Score subject to regulatory and explainability requirements. Our first case study benchmarks the US FICO® Score against modern ML approaches and discusses explainability challenges with "unfettered" ML models such as Gradient Boosted Decision Trees (GBDT). Our second case study concerns a recent development of a new FICO® Score outside the US where we combined the raw predictive power of ML with the advantages of Scorecard technology (ability to impute domain knowledge, ease of explanation) in a "best of both worlds" approach. This session shares experiences and methodologies that can be interesting for anyone who seeks to effectively leverage ML and domain knowledge to develop highly predictive yet still explainable models.

Mark Gorman (The Gorman Group Consultancy)
Making Analytics Great Again
Panel Discussion

Matt Nissen (Capital Services)
Enhancing Collection Effectiveness through Net Lift Analysis
Do you want to collect more dollars while spending less? Net lift analysis can help target customers with the highest propensity to be positively influenced by a collections treatment while balancing the costs of collecting. A collections effectiveness case study was conducted that outlines the basics of net lift analysis. Strategies using net lift analysis resulted in an increase in dollars collected and a decrease in collections costs. Key takeaways are the importance testing, using the correct metrics, and evolving as you learn.

Benson Hsu
Navigating Data Integration in Healthcare: What, Why, How, and Who?
As healthcare transitions from a fee-for-service model towards a value-based paradigm, healthcare organizations from payers to provider are exploring new data sets to augment existing traditional health (or more specifically, illness) data sets. In this discussion, we will explore exactly what types of data can serve to improve understanding of our communities, why these data sets are essential in this transition of our approach to healthcare payment and delivery, how these data sets can be integrated as well as the barriers to integration, and who ultimately "owns" this data as various (and at times, conflicting) constituents strive to serve the population. Is there genuine data democracy?

Kevin Potcner (JMP)
JMP and the Predictive Modeling Workflow
A typical real-world predictive modeling workflow includes (sometimes iteratively) data cleaning and exploration, model fitting, model validation, model comparison, final model selection and deployment of the final predictive model.
In this session we illustrate the predictive modeling workflow by analyzing a real dataset. After data preparation and initial exploration, we will create a number of predictive models such as multiple linear regression, regression tree, partition based methods, Neural Net, among others. We will "publish" models to the Formula Depot, and explore and select the best model(s) using the Prediction Profiler and JMP's Model Comparison tool.

Ryan Swanstrom (Unify Consulting)
Why Data Science is Difficult
It should come as no surprise that many data science and analytics projects fail. There are a whole number of reasons, and this talk will cover some of them. We will walk through the journey of planning a pizza party, studying for a test, and a few other fun stories. All of it will relate back to challenges with data science in the real world.

David Zeng (Dakota State University)
Deep Neural Networks to Predict Self-perception of Cardiovascular Disease ~ A Technical Demo
I train a multiple-hidden-layer deep neural network that predicts self-perception of heart health (including Cardiovascular Disease) with a large data set of 1729 features and about 30,000 samples. The dataset is based on the CDC Demographics, Dietary, Examination, Laboratory, and Questionnaire datasets collected from 1999 to 2016. Substantial data cleaning and pre-processing are done with Python pandas library.
The objective of this research is three-fold:
* Better understanding of how well DNN would improve the accuracy of prediction on perception of cardiovascular disease;
* Framework of developing more sophisticated DNN models to predict medical outcomes;
* Foundation for learning multi-dimensional/distributed representation of healthcare concepts that are both interpretable and scalable.
The presentation focuses on the technical (training a deep neural network with latest developments in the field of deep learning) aspects of the research.

Narav Bhagat (Metafunding)
Using Attribution Modeling to Find Profitable Lending Customers
A career in marketing data science will entail answering questions about how best to allocate spend within a marketing portfolio. As marketing budgets continue to shift towards online channels and away from print, the availability and quantity of campaign data make the data scientist's job more challenging than ever. Within financial services (and credit card issuers), a strategy that optimizes the mix of paid direct mail, paid online, organic search and on-site optimization has historically led to more efficient marketing costs. That strategy can been derived through attribution modeling, a close cousin to marketing mix models. In this presentation, Nirav will present examples of the conversion path of customers over paid, earned and owned marketing channels. He'll introduce the data engineering and analytics tools used to give credit to each channel for conversions. And, he'll introduce a few areas where science and tools are nascent.

Andre de Waal (SAS)
Sentiment Analysis of Donald Trump's Tweets
Social media generates huge amounts of data every day and most of the data is unstructured. This is a untapped resource that may provide significant benefits to companies able to exploit this data. SAS Visual Analytics is a big data tool that facilitates the visualization of large data sets. In this talk we demonstrate how insight can be derived from the analysis of Donald Trump's tweets. First, a word cloud is built and then a sentiment analysis is done on all of his tweets. Tweets are grouped into topics and the sentiment surrounding each topic is analyzed. This leads to the discovery of novel and interesting insights.

Abstracts | Invited Speakers

Ally Pelletier (Star Tribune)
Patrick Johnston
Use case of Amazon Web Services SageMaker in Digital Analytics
Amazon Web Services (AWS) provides cloud computing resources to companies around the world. These services include data storage, computing, and analytics. In November 2017, AWS released SageMaker. SageMaker is an end-to-end service which allows developers and data scientists to explore data, train models, and deploy models in production all within an easy to use platform. SageMaker comes with built-in common algorithms but also allows data scientists to bring their own algorithms in many languages including R and Python.
During this presentation we will walk through a use case of SageMaker in the digital analytics industry. The use case will include all steps of the modeling process including data preparation, training, and deployment.

Emily Griese (Sanford Health)
Collaboration and Innovation: Pushing forward Data-Driven Population Health
As healthcare continues to transition to value-based care, understanding how to effectively leverage data for population health is essential. Sanford Health recognizes this vital step and is leveraging models of innovation and collaboration to ensure success in a shifting healthcare climate. This talk will address current efforts including the Sanford Data Collaborative as well as introduce advanced analytics teamwork occurring across our footprint.

Jason Rogowski (Polaris)
Making the analytics journey without getting lost in the cloud
The Polaris data science team was founded with a simple objective: be predictive. Driving business value with initial quick wins was relatively easy, but scaling our ability to drive change was limited by the small size of the team. Furthermore, many of the business units needed fundamental reporting automation more than a neural network. We will discuss how we pivoted our strategy and are taking a more holistic self-service enablement approach alongside predictive algorithm development.

Abstracts | Oral Presentations

Nathan Axvig (Concordia College)
Predicting the Origins of Artwork Found in Rural Churches
A few years ago, I was contacted by Rodney Oppegard, a church historian. He had spent many years collecting information on ecclesiastical furnishings and artwork found in the Lutheran churches of rural North Dakota, and while his data set was extensive it was by no means complete. Some artwork was unsigned or the signature obscured, other pieces had been transferred to different churches, and in some cases the church itself had been destroyed by fire years before, leaving only incomplete records and fading memories as clues to the original church's configuration. Mr. Oppegard wanted to know whether there was a mathematical way to use existing data to "fill in the holes" of his data set. In this talk, I will outline how geospatial and rudimentary archival data were used to construct and evaluate models for determining which of several popular artists was responsible for a particular church's alter painting.

Anton Bezuglov (Buena Vista University)
Nathan Backman
Setting up Python and R in Classroom
Python and R are perhaps the two most popular free analytics platforms for Data Science. Unfortunately, for novice users such as students, installation of proper components and tool configuration is usually a problem. Even more difficult it is to ensure that everyone in the classroom works in the same environment, uses the same package versions, etc. Here, we present a classroom setup, where the students access their Python and R notebooks through HTTP via JupyterHub server. In this setup, the students have full permissions to their home directories as well as read/execute access to shared notebooks and data directories. The instructors can use the latter two directories to share in-class work, lectures, and data with all the students. The JupyterHub server runs on a virtual Linux machine to facilitate resource management and backups. This approach has been tested on smaller classes of up to 15 simultaneous users (both students and faculty). For larger classes, JupyterHub can be deployed on multiple nodes using Docker Swarm. Overall, this setup is an excellent platform where the users can focus on their work: learning or research without wasting time on package configuration, backups, and resource management.

Matthew Biesecker (South Dakota State University)
Modeling Vegetation Growth on Termite Mounds
Coupled systems of nonlinear reaction-diffusion PDE's have been model pattern formation since the 1950's. In particular, Turing proved that minor perturbations to initial conditions can result in exotic pattern formations. More recently, systems of PDE have been used to model plant/groundwater interactions. In this talk, we will discuss a recent work on mathematical models used to model the growth of vegetation on or around termite mounds in the African Savannah.

Lance Cundy (University of Iowa)
Estimating the Quantile Elasticity of Intertemporal Substitution with Instrumental Variables Quantile Regression
I estimate the quantile elasticity of intertemporal substitution (QEIS) of consumption using instrumental variables quantile regression. The elasticity of intertemporal substitution represents the willingness of a consumer to substitute future consumption for present consumption. In this paper, agents have a quantile utility preference instead of standard expected utility. This allows for the capture of heterogeneity along the conditional distribution of agents. The QEIS considers structural breaks in the data and is estimated for each regime using linearized Epstein-Zin preferences and by the use of fixed effects, instrumental variables, and quantile regression. The estimator is a feasible estimator based on smoothed sample moments. In order to estimate the model, the Nielsen

Consumer Panel dataset is used. This dataset is built from transactional data that follows households in the United States and their grocery purchases from 2004 through 2014. Because of the transactional nature of the dataset, there is a low source of measurement error in consumption, and aggregation bias can be minimized. To estimate the model, consumption is aggregated weekly, and consumption growth is measured over a four-week time period in order to match four-week Treasury bills. Results give evidence of heterogeneity of the QEIS along the quantiles of the conditional distribution.

Xijin Ge (South Dakota State University)
Shiny app as a solution for streamlining complex bioinformatics analyses
Rapid innovation in biotechnology, especially DNA sequencing, holds great promise for revolutionizing medicine. The main bottleneck is how to analyze and interpret the massive amount of data effectively. Many stand-alone software packages exists, mostly as R packages. We developed iDEP(Integrated Differential Expression and Pathway analysis), a large Shiny application that integrates hundreds of R packages, and large gene annotation database. Available at (http://ge-lab.org/idep/), iDEP streamlines complex bioinformatics pipelines as a friendly web interface, that can turn data into biological insights within minutes, instead of months.

Russ Goodman (Central College)
An In-Class Geospatial Data Analytics Project - Inspired by a Comedian
This talk will share the details of an intriguing and appealing in-class project for students in an introductory Data Analytics class for advanced mathematics majors. The project originated with a joke from a popular comedian about whether "La Quinta" is Spanish for "Next to Denny's" and developed into an investigation of that quip. In this project, students learn to acquire the appropriate geospatial data, learn some new skills in Excel, RStudio or other data analysis software, experience quite a bit of problem-solving, and work hard to communicate their results.

Yuhlong Lio (University of South Dakota)
Building R Package
In this talk, a set of procedures for building R package will be discussed and some recently built R packages will be introduced.

Robert Menzie (Sanford Health)
Clark Casarella
No Show Predictive Model: A Bayesian Approach
Patients not showing up to their appointments is a detriment to both the patient and the health care system. As health care systems transition from fee-for-service programs to value-based program, clinical visits (especially primary care) will become the gateway to improving overall patient health outcomes. In order to ensure patients are receiving the appropriate treatment and maintaining a healthy lifestyle they must be completing their scheduled visits. The main goal of the model is to predict patient no-show probabilities with the intent of taking the model one step further by linking it to actionable data points and decisions. The model employs the use of a logistic regression and Bayesian update approach. The regression is devised of patient demographical, behavioral and diagnosis characteristics, as well as visit logistics. The logistic regression creates a priori probability based on requisite factors. Then due to the highly behavioral impetus of missing appointments, a Bayesian update is applied to the priori probability to obtain a final, posterior probability. The Bayesian application to this model significantly contributes to the patient's probability and details the importance behind patient-level interventions. The output of the model has a high level of accuracy that allows clinics not only to see which patients have a high risk of not showing up, but also the factors that physicians may be able to remedy down the road. The model was built using a standard 10-fold cross-validation. The test set was then ran through the model and used to determine the weighting for the Bayesian update. Lastly the data was validated using the remaining 10%, which resulted in an AUC of .927. Combining the accuracy of this model with the prescriptive ability of the factors, can allow for a significant reduction of no-shows, not only by enhancing appointment logistics (calls, overbooking, etc.) but also by improving patients' lifestyle.

Cherie Noteboom (Dakota State University)
Mining Users Feedback: Discovering the Gaps in Mobile Patient Portals
Patient portals are positioned as a central component of patient engagement through the potential to change the physician-patient relationship and enable chronic disease self-management. In this article, we extend the existing literature by discovering design gaps for patient portals from a systematic analysis of negative users' feedback from the actual use of patient portals. Specifically, we adopt topic modeling approach, LDA algorithm, to discover design gaps from online low rating user reviews of a common mobile patient portal, EPIC's mychart. To validate the extracted gaps, we compared the results of LDA analysis with that of human analysis. Overall, the results revealed opportunities to improve collaboration and to enhance the design of portals intended for patient-centered care.

Austin O'Brien (Dakota State University)
Using Atypicality to Identify Outliers
This presentation will outline the development and use of a probabilistic measure for outlier detection, referred to as atypicality. Given a set of objects, we can create a corresponding set of similarity scores between them. Assuming the set of scores has a normal distribution, we can estimate the score distribution's parameters. We compute atypicality by comparing the likelihood of an object given these estimated parameters to the likelihood of bootstrapped samples. The atypicality measure is then used as a p-value in a hypothesis test, where the null hypothesis states that the object in question is similar to the remaining objects; the alternative hypothesis is that the object is an outlier. This can be used in a variety of applications, especially where we have multiple objects in multi-dimensional space.

Galkande Premarathna (Minnesota State University, Mankato)
Leif Ellingson
Analyzing ligand-binding proteins using their structural information
It is known that a protein's biological function is in some way related to its physical structure. Many researchers have studied this relationship both for the entire backbone structures of proteins as well as their binding sites, which are where binding activity occurs. However, despite this research, it remains an open challenge to predict a protein's function from its structure. There are many useful applications from protein function predictions, such as effective drug discovery with fewer side effects, development of structure-based drug designs, disease diagnosis, and many more. This presentation will discuss how this ligand-binding protein prediction problem is approached by taking a higher level object-oriented approach, which is named as Covariances of Distances to Principal Axis (CDPA) that summarizes the description of the binding site so that it reduces the amount of information lost compared to most of the other approaches. Thereby, a model-based method is considered, where the nonparametric model is implemented by using the features of the binding sites for a given ligand group for understanding and classification purposes. Then the results obtained using the model-based approach are compared to the alignment-based method used by Ellingson and Zhang (2012) and Hoffmann et al. (2010).

Sayeed Sajal (Minot State University)
Israt Jahan

Prediction on Oscar Winners Based on Twitter Sentiment Analysis Using R

In the new era of development, social media is not only getting popularity but also useful to reveal hidden information. Most of the people use social media to connect friends and family, to express their emotions, to give feedback, and to raise concerns as quickly as possible. We can reveal important information from people responses in social media. OSCAR nominations were announced for the year 2017 and all the nominees are very active on Twitter. All the tweets regarding their nominations are publicly available. Here, in this paper, we analyzed the public tweets from Twitter and predict who will be the OSCAR winner in 2017. More specifically, we analyzed all the tweets of the OSCAR nominees in the category of "Actor in leading role" since the day of OSCAR nomination announcement. After analyzing the data, we predicted who will be the winner based on the twitter sentiment analysis using R programming.

KC Santosh (University of South Dakota)

AI in Healthcare: Automated Chest X-ray Screening

Unstructured data, i.e. image is worth a thousand words. Image analysis has several different applications; healthcare, for instance. Fundamental image processing mechanics let us focus on how we can actually represent visual images to be processed in machine learning algorithms. More specifically, the talk aimed to provide how data scientist works with an emphasis on image processing and pattern recognition. In this context, we will present an automatic chest X-rays screening system to detect pulmonary abnormalities using chest X-rays (CXR) in non-hospital settings. In particular, the primary motivator of the project is the need for screening HIV+ populations in resource-constrained regions for the evidence of Tuberculosis (TB). The system analyzes thoracic edge map, shapes as well as symmetry that exists between the lung sections of the posteroanterior CXRs. For classification, we have used several different classifiers, such as support vector machine, Bayesian network, multilayer perceptron neural networks, random forest and convolutional neural network. Using CXR benchmark collections made available by the National Institutes of Health (NIH) and National Institute of Tuberculosis and Respiratory Diseases, India, the proposed method outperforms the previously reported state-of-the-art methods by more than 5% in terms of accuracy and 3% in terms of area under the ROC curve (AUC). On the whole, the talk will consider state-of-the-art works in image analysis, pattern recognition and machine learning under the framework of healthcare and/or medical imaging. Having all these topics, we will provide/summarize how AI and machine learning have helped healthcare advance than it used to be.

Paul Thompson (Sanford Research)
Solomon Adu

Table 1: SAS and R tools to build summary statistics tables

In presenting clinical data, Table 1 is used to present demographic, clinical, and pre-treatment data for trials. This includes continuous variables (age, BP, BMI - presented as means&stddev or median-IQR) and discrete variables (gender, race, cancer stage - presented as proportions-counts). While the actual statistical/data analytic techniques involved are trivial, constructing such tables takes a lot of time as results must be assembled into standard tables. Automating this is a huge time saver. SAS and R methods for building this table in a consistent manner are presented.

Yanshan Wang (Mayo Clinic)
Elizabeth J. Atkinson, Shreyasee Amin, Hongfang Liu

Towards a Distant Supervision Paradigm for Clinical Information Extraction: Creating Large Training Datasets for Machine Learning

Background In the era of big data, a large number of clinical narratives exist in electronic health records. Automatic extraction of key variables from clinical narratives has facilitated many aspects of healthcare and biomedical research. Conventional approaches are based on rule-based natural language processing (NLP) techniques that rely on expert knowledge and exhaustive human efforts of designing rules. Recently machine learning has seen a big performance gain compared to conventional NLP approaches. Despite the impressive improvements achieved by machine learning models, large manual labeled training data are the crucial building blocks of conventional machine learning methods and key enablers of recent deep learning methods. However, large training data are not always readily available and usually expensive to obtain from human annotators. This problem becomes more significant for use cases in clinical domain due to the Health Insurance Portability and Accountability Act (HIPAA) where methods, such as crowdsourcing, are not applicable, and requirements of annotators being medical experts. Method In this paper, we propose a distant supervision paradigm for clinical information extraction. In this paradigm, rule-based NLP algorithms are used to generate large training data with labels automatically. Machine learning models are subsequently trained on these distant labels with word embedding features. Results We study the effectiveness of the proposed framework on two clinical information extraction tasks i2b2 smoking status extraction shared task and a fracture extraction task at our institution. We tested three prevalent machine learning models, namely, Convolutional Neural Networks, Support Vector Machine, and Random Forrest. Conclusion The experimental results show that the proposed distant supervision paradigm is effective for the machine learning models to learn rules towards gold standard from distant labels. Moreover, the machine learning models trained on the distant labels generated by a rule-based NLP algorithm could perform better than the NLP algorithm given sufficient data. Additionally, we showed that CNN was more sensitive to the data size than the conventional machine learning models and that all the tested machine learning methods were viable options for the distant supervision paradigm.

Jixiang Wu (South Dakota State University)

Selecting Categorical and Quantitative Variables in Linear Regression Analysis

Variable selection is an important means to construct a model that predicts a target/responsible variable with a set of predictable variables. The predictable variables could include quantitative, binary, and/or categorical variables; however, commonly used variable selection methods such as forward selection, backward selection, and stepwise selection are more focused on quantitative variables. It will be a helpful addition to multiple linear regression if categorical variables can be integrated with the commonly used variable selection methods. We proposed a generalized variable selection method that can be used to select both categorical and quantitative variables simultaneously. The detailed results will be presented at the symposium.

1. Nehal Adhikari (University of South Dakota)
Quality Control for the Strength of Carbon Fibers
Carbon fibers are ingredients for the rigid composite material used in aerospace and other applications. It is very important to ensure the strength of carbon fibers to meet the required standard. In this presentation, Burr type-X distribution is used to demonstrate good-of-fit test for the strength of carbon fibers by least square method. Bootstrap quality control charts are developed via least square method estimates. Stimulation study is carried out through R program to show the bootstrap quality control charts are capable to detect out of control case efficiently. Practical example will be given to address the applications.

2. Abhilasha Bajracharya (South Dakota State University)
Riaz Khan, Semhar Michael, Reinaldo Tonkoski
Data Center Load Forecast using Hidden Markov Models
The energy cost of data centers tantamount to their overall operational cost. A possible solution to this immense cost could be proper scheduling of the power resources. This can be achieved by forecasting the data center loads. However, highly variable nature of the data center loads makes it challenging to use the traditional methods of load forecasting. In this paper, a stochastic method based on Hidden Markov process is developed to model the data center load and is used for a day-ahead forecasting. This method is out-standing because of its flexibility in addressing the variable nature of the data center load. The utility of the model is illustrated using a dataset from National Renewable Energy Laboratory - Research Support Facility (NREL - RSF). Two models created based on the proposed method yielded Mean Absolute Percentage Errors (MAPE) of 1.49% and 3.89%.

3. Damon Bayer (South Dakota State University)
Semhar Michael
Variable Selection Techniques for Clustering on the Unit Hypersphere using von Mises-Fisher Distributions
Mixtures of von Mises-Fisher distributions have been shown to be an effective model for clustering data on a unit hypersphere, but variable selection for these models remains an important and challenging problem. In this paper, we derive two variants of the Expectation-Maximization (EM) framework, which are each used to identify a specific type of irrelevant clustering variable in these models. The first type are noise variables, which are not useful for separating any pairs of clusters. The second type are redundant variables, which may be useful for separating pairs of clusters, but do not enable any additional separation beyond the separability provided by some other variable. Removing these irrelevant variables is shown to improve cluster quality in simulated as well as benchmark text datasets.

4. James Boit (Dakota State University)
Malaria Surveillance System Using Social Media
Social media, for example Twitter has increasingly provided opportunities for massive data collection of topical issues affecting the modern society. Opinions and data in public health issues are very prevalent on twitter and provide an invaluable source of interesting information that can be mined for decision making in public health organizations. This paper aims to analyze the extent to which malaria data is reported on twitter and then develop a malaria surveillance system to detect and monitor malaria incidences. Therefore, this paper will discuss a proposed malaria surveillance system (MSS) that leverages twitter data in sub-Saharan Africa. The MSS system will comprise of a data collection module, analysis engine and a metrics module. Finally, the Malaria surveillance system will be evaluated using data and reports from existing traditional surveillance systems.

5. Merritt Burch (South Dakota State University)
Vivek Shrestha, Boris Shmagin, Donald Auger
Systemic Analysis of Biological Data from an Isogenic Maize Line
Vast amounts of data are generated in modern biological research, which creates a challenge to their analysis. Here we use a combination of factor analysis and principal component analysis against morphological measurements taken from a collection of maize plants that are descended from a single doubled-haploid plant. We are looking to identify relationships and structures within the data using uncorrelated subsets that explain much of the variability present. With these multivariate statistical techniques, we expect to summarize systematic patterns and complex relationships in our data and show that these analyses are useful in other biological research fields.

6. Peter Dolan (University of Minnesota Morris)
Grid Feature Extraction Techniques
R, with the proper packages installed, provides a powerful platform for image manipulation and feature extraction. Strengths and weaknesses of three techniques for extracting characteristics from images containing a rectangular grid (with and without defects) are contrasted. Techniques discussed are the 2 dimensional fast Fourier transform (FFT), the Hough transformation for line detection, and topological data analysis (TDA) utilizing Vietoris-Rips persistence barcodes.

7. Peter Gilbertson (Dakota State University)
Jodi Cisewski, Cora MacPherson, Christina Park, Anita Johnson, Jack Moye, Jr.
Exploratory Tools for National Children's Study Data
The National Children's Study Archive is an information, data, and sample repository for the National Children's Study, designed to make data and samples freely available for scientific research, with an approved research request. The study, which was active from 2009 to 2014, collected data and samples from over 12,000 mothers, fathers, and children across the US at 40 study locations. These participants are currently represented by nearly 14,000 variables, over 200,000 biological samples, and over 4,000 environmental samples in the Archive.

In order to facilitate meaningful evaluation and consumption of this wealth of data, the Archive has developed new exploratory tools. The Protocol Browser allows users to flow through the visit progression and visit instrumentation and identify available study datasets. The Participant Explorer allows users to investigate study participation by participant type (woman, child, father), demographics (e.g., education level, marital status), and data collection point. The Sample Explorer allows researchers to use demographics and study visit information to explore the available biological (blood, hair, nails, saliva, urine, vaginal swab, breast milk, cord blood, meconium, and placenta) and environmental (air, dust, water) primary and derivative samples that were collected from a subset of NCS families. Both the Participant and Sample Explorers provide researchers with the sample sizes available that fit the types of participants or samples of research of interest. The Variable Locator allows users to search the available NCS datasets for questions and variables of interest, returning specific data elements available. With these research tools, researchers targeting unique populations and topic areas can quickly establish if a particular topic area is represented in NCS data and samples. Once identified, researchers can use the Archive's proposal submission and review process to begin using NCS resources to pursue their scientific objectives.

8. Eric Guthrie (South Dakota State University)
Interpolating Demographic Estimates for Alternate Subpopulations and Geographies
Applied demography employs population studies in the effort to answer real world questions and the problems that business and civic leaders face on an ongoing basis. To answer these questions the applied demographer sometimes performs primary research, but more often they attempt to leverage and extend the use of publicly available data to answer the questions presented in an efficient and time constrained manner. The work described here looks at

a problem presented by the Michigan Department of Education and the solution presented by the Michigan State Demographer. The problem required estimates for a single-year age group at a non-standard poverty level. These data are not published by the U.S. Census Bureau, but a novel solution that deploys geographic interpolation methods was developed to serve an intermediate need until a custom tabulation of Census data could be delivered. With the delivery of a custom tabulation of Census data and the interpolated dataset that was produced by the State Demographer, there was a unique opportunity to test the results of the interpolation against what would be a gold standard dataset. The results reveal that the process of interpolating estimates devised as a solution could produce estimates that could be useful for a variety of purposes.

9. Tingting He (South Dakota State University)
Krishna Ghimire, Jyotirmoy Halder and Jixiang Wu
Identification of efficient experimental design(s) by comparing different commonly used designs for large data sets
Following basic principles of experimental design and using appropriate filed design usually play a significant role in successful plant breeding program. Breeders usually seek suitable field designs and use them for their experiments to minimize the experimental error. Finding out the most appropriate design for the experiment may greatly improve the data analysis and help breeders to take right decision. For the current study, 64 corn hybrids (genotypes) were evaluated in six counties in North Carolina. The main objective of our study is to analyze and compare results using different models or experimental designs (CR, RCBD, Sub-block, and Rectangular) and finally determine which design(s) is the most suitable for each county. Linear mixed model (LMM) with jackknife resampling technique was used for data analysis. Among all the tested designs, we found that Rectangular design is more suitable for handling large data sets and CR design was least effective. Rectangular design was well over 2 times and close to 2 times better than CR and RCB design respectively in terms of Relative Efficiency (RE) for most of the counties.

10. Calla Holzhauser (South Dakota State University)
Semhar Michael
Forecasting for the All Women Count! Program
All Women Count! program is a no-cost breast and cervical cancer screening program for qualifying women. We are interested in estimating the number of women who will use the program for the next 5 years. Forecasting was done using several commonly used models for each county. In addition, a Gaussian mixture of regression time series model is used to perform clustering and forecasting. Four models were tested and the model with the lowest test root mean square error was chosen to carry out the forecasting by county. The model chosen most often was the ordinary least squares regression closely followed by ridge regression and linear regression with autoregressive integrated moving average errors. Model selection for the mixture model was done using the Bayesian information criterion and found 5 clusters were optimal. The five clusters identified the counties with increasing and decreasing participation. The results will help the South Dakota Department of Health with future planning and implementation of the program.

11. Md Riaz Ahmed Khan (South Dakota State University)
Collaborative Filtering approach of recommender system with application in Amazon's jewelry products
Personalized recommendation is one of the key drivers of today's e-commerce. Among numerous available items, recommender system makes recommendation to the users based on their past behaviors. With increasing popularity, recommender system found its application in movies, music, videos, jokes, driving routes, restaurants and all kinds of general products.
One common approach to build a recommender system is Collaborative Filtering (CF). In this work, we walk through the different steps of making recommender system based on different CB techniques (user based, item based, hybrid). We used Amazon's review data of jewelry products to build different recommender systems and evaluate their performances.

12. Bigyan Khanal (Dakota State University)
David Zeng
Application of Deep Neural Network for Calculation of Pretest Probability of the Heart Diseases
Application of deep learning techniques for calculation of pre-test probability(PTP) scoring of heart diseases based on the structured and unstructured data reduces unnecessary cardiac imaging tests, medical costs, and other potential risks. For any patient, the calculation of PTP will increase the performance of a given diagnostic test for heart diseases and be performing no test for a patient with fewer than a certain threshold of the likelihood function. In the model for cardiac imaging decision-making system, the input layer is the features of the patients' medical history and the different physical and biological attributes. The output layer is defined as the decision based on the probability for every heart diseases tests and the several combinations of them. The Deep Neural Network (DNN) model whose network uses the efficient Adam gradient descent optimization algorithm with a logarithmic loss function, was implemented using Keras. The number of neurons, number of hidden layers, batch size, convergence criterion, activation function, learning rate and regularization parameters is defined as the model requirement. The pretest probability of the heart diseases is compared relatively against the score from traditional linear measures for the model validation.

13. Runan Yao (South Dakota State University)
Xijin Ge
Discover Characteristics and Behaviors Influencing Weight Loss Using Sanford Profile Data
Profile by Sanford is a membership-based weight loss program. By one-on-one interactions with a weight loss coach, it helps each member make their own weight loss plan includes nutrition, activity and lifestyle. This research utilizes massive data collected via Profile by Sanford to analyze member behavior. Data includes activity, hip and waist circumference, body weight measurements, food items, meal plans, medications, demographics, coach meetings and other records. Discovery of characteristics and behaviors influencing weight loss will benefit current and future members of Profile.

14. Joseph Robertson (South Dakota State University)
The Impact of Data Sovereignty on American Indian Self-Determination: A Proof of Concept using Data Science
The basis for my work is further examining the ideas of American Indian self-determination and tribal sovereignty beyond the legal and political definitions to provide a more modern approach to regaining eroded sovereignty through data sovereignty. Deloria (1976) theorized that maintaining the term sovereignty strictly in a legal/political context becomes a limiting concept designed at preventing solutions. Appropriately, this concept began as my initial framework design to adjudicate Federal Indian Law, policy, and theories by quantification using computational statistics. McKinley et al. (2012) asserts that in order to construct more enduring governments and viable economic institutions, education in Indigenous communities should uphold the values, interests, and cultures of the communities and nations. Thus, new understandings of Indigenous nationhood can be not only conceptualized, but can serve to protect and preserve community, sovereignty, and cultural traditions. Data Sovereignty is a realization of this new theory and praxis.
The underlying foundation that is critical in this concept is higher education as it relates to nation building. A nation building approach coupled with higher education is deemed to be the most effective way to begin quantifying previously only qualitative assumptions about how American Indians serve their community.

Poster Presentation

My 2016 story map, The Impact of Data Sovereignty on American Indian Self Determination was quite a lengthy discussion on American Indian history, Federal Indian Law and Policy, and the primers of how data sovereignty could further the development of unifying a data platform for data collection, management, and storage in Indian Country. Data sovereignty reflects how these concepts relate to cultural capital, educational praxis, and finally GIS/spatial analysis in developing this platform through statistical design theory.

I will explore two case studies in realizing this theory and practice, but also discuss further open source initiatives I would like to pursue above and beyond this. I feel by developing this praxis through my doctoral dissertation, it will become more clear how powerful GIS and statistical design theory is for an American Indian scholar developing smart solutions for tribal communities. Thus, indigenous scholarship through educational nation building models the power of data sovereignty in real time.

15. Eric Stratman (South Dakota State University)
Thomas Brandenburger
A New WOE and Scorecard Building R Package
In almost every data science project that involves modeling, cleaning and transforming the data is the most time-consuming part. One method of data transformation is called weights of evidence (WOE) binning. WOE binning is commonly used with credit scoring because most of the data is highly skewed and contains numerous missing values. Despite WOE binning being a common practice in the credit scoring industry, there are relatively few R packages that perform WOE binning and also build a scorecard. Due to the lack of packages in R capable of these methods, we believed it would be beneficial to the R community to build an R package that performed WOE binning and scorecard creation. The R package created has built-in functions to allow the user to complete all parts of the credit scoring process. These functions include: custom binning, WOE transformation, variable clustering, scorecard creation, and WOE visualization. By creating this R package, users will have the ability to create a scorecard using WOE transformation for their data science projects.

16. Yohannes Tecleab (South Dakota State University)
Gary Hatfield
Spatial regression for corn and soybean yield in South Dakota
Farm terrain attributes influence soil properties and hence are important determinants of crop yield. Here a multi-year data on corn and soybean yield from one farm in western South Dakota is analyzed to understand the potential predictors. Soil samples from 467 stations within the farm are sampled and the chemical composition of the soil analyzed. Data in yield and altitude was collected from the yield monitor. Aspect and slope of each yield sampling point are derived from altitude. Soil sampling stations had sparser sampling resolution compared to the yield data. To balance the sample sizes, Voronoi polygons are created around each soil sampling station and the mean values of yield, aspect and slope are taken at each polygon. Spatially weighted regression model is fitted on the resulting balanced dataset and the best predictors of yield are identified.

17. Runan Yao (South Dakota State University)
Xijin Ge
Discover Characteristics and Behaviors Influencing Weight Loss Using Sanford Profile Data
Profile by Sanford is a membership-based weight loss program. By one-on-one interactions with a weight loss coach, it helps each member make their own weight loss plan includes nutrition, activity and lifestyle. This research utilizes massive data collected via Profile by Sanford to analyze member behavior. Data includes activity, hip and waist circumference, body weight measurements, food items, meal plans, medications, demographics, coach meetings and other records. Discovery of characteristics and behaviors influencing weight loss will benefit current and future members of Profile.

Committee and Volunteers

Committee
A big thank you to everyone involved with organizing the 2018 Data Science Symposium. Your commitment and dedication made this event possible!

Department Head
Kurt Cogswell

Conference Co-Chair
Dr. Gary Hatfield
gary.hatfield@sdstate.edu | 605-688-5846
Dr. Tom Brandenburger
thomas.brandenburger@sdstate.edu | 605-688-6196

Committee Members
Dr. Cedric Neumann
cedric.neumann@sdstate.edu | 605-688-5833
Dr. Semhar Michael
semhar.michael@sdstate.edu | 605-688-6316
Dr. Xijin Ge
xijin.ge@sdstate.edu | 605-688-5854

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

Abdelbaset Abdalla
Girma Ayana
Damon Bayer
Eric Bae
Quinn Fargen
Riaz Khan
Jyotshana Paudyal
Dheeman Saha
Juan Xie
Runan Yao

South Dakota
State University

2018 SDSU
Data Science
Symposium