Statistics, the Impact of Big Data

60th Anniversary of the

Florida State University Department of Statistics

1959 - 2019

April 12-13, 2019

at the Augustus B. Turnbull III Florida State Conference Center

Friday, April 12: Room 103 (simulcast in room 101)

12:30pm - 1:00pm	**Registration**
1:00pm - 1:05pm	**Opening Remark: Associate Dean, Tim Logan**
1:05pm - 1:10pm	**Welcome Speech: Chair, Xufeng Niu**
1:10pm - 2:30pm	**Session I (Session Chair: Jayaram Sethuraman)** **Gregory Campbell, Michael Schell, Hulin Wu, Yichuan Zhao**
2:30pm - 3:00pm	**Break**
3:00pm - 4:20pm	**Session II (Session Chair: Robert Clickner)** **Larry Crow, Douglas Jones, James Lynch, Barbara Stevens**
4:20pm - 4:40pm	**Break**
4:40pm - 6:00pm	**Session III (Session Chair: Doug Zahn)** **Edsel Pena, Benedikt Johannesson, John Robinson, Rao Chaganty**
6:00pm - 7:00pm	**Break, Group Photos**
7:00pm - 9:30pm	**Banquet (Session Chair: Wei Wu)** **Greetings from former chairs and retired faculty** **Prize conferring for logo contest winners** **Banquet Talk by Blanton Godfrey**

Saturday, April 13: Parallel Sessions in rooms 103 & 101

8:30am - 9:00am	**Refreshments**
9:00am - 10:20am	**Session IV-A (Room 103, Session Chair: Fred Huffer)** **Balasubramanian Narasimhan, Stephen Lee, Shanti Gomatam, Ivo Dinov** **Session IV-B (Room 101, Session Chair: Xufeng Niu)** **Leif Ellingson, Naomi Brownstein, Derek Tucker, Mingfei Qiu**
10:20am - 11:00am	**Break, Group Photos**
11:00am - 12:00pm	**Session V-A (Room 103, Session Chair: Daniel McGee)** **Michiko Wolcott, Billy Franks, Ryan Scolnik** **Session V-B (Room 101, Session Chair: Lifeng Lin)** **Guanyu Hu, Kumaresh Dhara, Xin Li**
12:00pm - 1:00pm	**Pizza Lunch (Room 103)**
1:00pm - 2:00pm	**Panel Discussion (Room 103)** **(Coordinator: Kristine Rosenberger)** **Panelists: Larry Crow, Benedikt Johannesson, Hulin Wu, Yichuan Zhao**

Friday, April 12: 103 Turnbull Center (Simulcast in room 101)

12:30pm - 1:00pm **Registration**

1:00pm - 1:05pm **Opening Remark (Associate Dean: Tim Logan)**

1:05pm - 1:10pm **Welcome Speech (Chair: Xufeng Niu)**

1:10 pm - 2:30pm **Session I (Session Chair: Jayaram Sethuraman)**

(1:10pm - 1:30pm)

Gregory Campbell, GCStat Consulting LLC
Statistics For "Big Data" In Biomedicine

Society is awash in data in this new data-rich world. In this era of "Big Data" there are many sources of possible data sources that could be leveraged in biomedicine. The effort to convert such data to evidence and then utilize such evidence in an efficient manner can be a challenge but the payoff can be enormous in terms of savings of time and resources. Statisticians and statistical methodologies are well-positioned to play a crucial role in the conversion of big data into evidence. There are however inherent biases (especially selection biases) in the analysis of "big data". Also at issue is the generalizability of the results. The important role that statistics plays in large datasets is illustrated historically in a number of applications from the National Institutes of Health and the Food and Drug Administration. This includes statistical methodology examples from neuroimaging data such as functional MRI (fMRI) and other imaging systems, the Framingham Heart Study, a large microarray (genetic) project, genetic tests and the use of observational ("real-world data") to augment information from randomized clinical trials. Currently there is a wealth of data from companies who organize electronic medical records as well as all information generated from wearable technology. Statistics can and should play a key role in data science, neural networks, and deep learning. An important consideration is data quality and for clinical trials for premarket regulatory purposes this can be addressed using Clinical Data Interchange Standards Consortium (CDISC) standards. Statisticians are uniquely poised for a bright future that will generate even much more useful data!

(1:30pm - 1:50pm)

Michael J. Schell, Moffitt Cancer Center
Identifying The Most Impactful Statistical Literature

This talk identifies 3, 655 "high-impact" statistical papers from 137 journals. These papers have either a minimum of 100 citations in the ISI Web of Science, including at least 20 in 2014-15 or 2015-2016, or ≥40 citations in 2015-2016. Among the 11 statistical journals that contribute ≥ 50 papers each (58% of the total), these high-impact papers represent 4.7% of citable items. To render this large list of papers more useful, we have developed a multi-level hierarchical statistical classification system, with 6 Kingdoms, 51 Phlya, and 282 Families at

the highest three levels. The 6 Kingdoms are: Estimation (N=756), Experimental Studies (N=521), Regression (N=920), Testing (N=245), Specialized (N=1016), and Broad Topics (N=364). The fraction of citation counts that represent "applied" use is estimated, and count-based metrics are given for impact, trajectory, diffusion, and applied area of primary use of the paper. These adjunctive metrics facilitate a focused search for papers of interest. Applied statisticians should find it helpful in staying current and upgrading one's statistical practice. Methodological statisticians can also profit, as the list provides the currently most impactful statistical literature, and provides an overview of the major contributions in statistics.

(1:50pm - 2:10pm)

Hulin Wu, University of Texas Health Science Center at Houston
Big Data Challenges And Opportunities For Statisticians

The newly emerging concepts of Big Data and Data Science have made a very big splash among academic world, industries and governments in the past several years. In particular, it has a big impact to the statistical community, since statisticians are traditionally considered as a unique profession to deal with, especially analyze data. There is a potential threat that the emerging professionals of Data Scientists may replace statisticians' job. In this talk, I will discuss the differences between the traditional statistics and the emerging field of data science via analyzing and understanding the new concept and definition of data science and Big Data. I will also propose and discuss our strategies in order to identify our opportunities in the era of Big Data.

(2:10pm - 2:30pm)

Yichuan Zhao, Georgia State University
Empirical Likelihood For The Bivariate Survival Function Under Univariate Censoring

The bivariate survival function plays an important role in multivariate survival analysis. Using the idea of influence functions, we develop empirical likelihood confidence intervals for the bivariate survival function in the presence of univariate censoring. It is shown that the empirical log-likelihood ratio has an asymptotic standard chi-squared distribution with one degree of freedom. A comprehensive simulation study shows that the proposed method outperforms both the traditional normal approximation method and the adjusted empirical likelihood method in most cases. The Diabetic Retinopathy Data are analyzed for illustration of the proposed procedure.

2:30pm-3:00pm

Break

Friday, April 12: 103 Turnbull Center (Simulcast in room 101)

3:00pm - 4:20pm

Session II (Session Chair: Robert Clickner)

(3:00pm - 3:20pm)

Larry H. Crow, Crow Reliability Resources, Inc.
Estimating System Reliability After Corrective Actions With Limited Data

The initial reliability of new, complex, state of the art systems is generally very low relative to the requirement and low relative to the capability of the design. These systems are typically subjected to reliability growth development testing where problem failure modes are found, and corrective actions taken. It is important for management and engineering to track and manage the system reliability growth so that appropriate resources can be allocated to attain the reliability requirement. If corrective actions are incorporated into the system during test phases, then the most widely used statistical reliability growth tracking model worldwide is based on the nonhomogeneous Poisson process (NHPP). However, some systems are tested over a single test phase and all corrective actions are incorporated at one time, with no available data after the corrective actions. An estimate of the system reliability after these corrective actions are incorporated is called a projection and involves applying a fix effectiveness factor to the failure rate estimate of each observed failure mode. The problem, and key statistical issue addressed in this presentation is that not all of the failure times for an observed failure mode are representive of the failure rate for that mode. This is particularly significant if an observed failure mode only has one failure occurrence during the test. This presentation discusses a methodology for estimating the failure rates for the observed problem failure modes with limited data that utilizes both the Crow NHPP model and the multinomial distribution. This methodology is then used to estimate the improved system reliability after corrective actions.

(3:20pm - 3:40pm)

Douglas H Jones, Rutgers University
Bayes Updating In Educational Norm-Reference Testing

Concerning educational testing, test scores provide the basis for assessment of learning. The purpose of the assessment may be criterion-reference or norm-reference. In criterion-reference testing, test scores indicate the level of mastery of subject matter for a particular age/grade level in the K-12 setting. While in norm reference testing, test scores indicate the level of test performance in a particular subject matter compared to performances of all test-takers in age/grade levels. Other examples of a norm-reference test are the IQ test, SAT, ACT. Examples of a criterion-reference test is the PHD qualifying test (written and oral) at the FSU Stat Department.

For either types of testing, criterion-reference or norm-reference, the raw test score is transformed to a scale. An obvious scale is the percentage correct: number questions correct (raw score) divided by total number of questions. Howev-

er, this scaled score cannot be used to compare scores across different forms of the test or different age/grade levels in the same subject matter. Therefore, even though there is the distinction between criterion-reference and norm-reference testing, scores need to be scaled using complete data from a population. The usual scaled score is the quantile of the raw score (or percentage correct) as determined by the performances of all test-takers.

Educators have found that quantiles are very handy: The quantile of one test form can be transformed to the quantile of another test form. This permits comparison of test scores from year to year and between age/grade levels. So no matter the type of test, criterion vs norm, using scaled scores, officials could ideally measure scholastic progress and thus manage educational reform. Parents could measure Johnny's potential for getting into Harvard, maybe (joke). College accrediting agencies could better determine whether a college unit is assuring learning, and possibly determine which colleges are doing an exceptional job, although everyone finds this latter comparison distasteful or political.

Now statisticians also find that quantiles are very handy for data analysis. They know that the quantile function is the inverse of the cumulative distribution function. Educators believe they need only the normal CDF for scaling all their tests.

We will apply newer methods for estimating the CDF of test scores. In particular we explore Bayes Nonparametic estimators and apply this to real data from the Terra Nova Assessment series.

K-12 Test Publishers must update test norming tables every 4 years or less (US State Laws and "No Child Left Behind"). Before the law, test data was updated every 10 years using over 250K children. Now schools charge $10-$15 per child. We explore the possibility of using Bayes estimation and optimal design to reduce cost. This is accomplished by borrowing information from prior norming tables with Bayes estimation of

CDF - A norming table is a CDF.

(3:40pm - 4:00pm)

James Lynch, University of South Carolina
Some Stochastic Models and Analysis Methods For Fiber Bundle Models: Rosen's Experiments on Fibrous Composites

Rosen's experiments (1963, 1964) on fibrous composites in the 1960's gave fundamental insights into their failure. These experiments and some statistical methods and models for these experiments seem to have been overlooked by some in the physical science community that study Fiber Bundle Models (FBMs). Here we summarize the results given in Li and Lynch <https://arxiv.org/abs/1903.02546> who have sketched an overview of these methods and models to

correct this oversight and indicate their use in understanding Rosen's Specimen A experiments.

(4:00pm - 4:20pm)

Barbara Stevens, Central Intelligence Agency
A Career in Statistics/Data Science at CIA

This talk will highlight what a 30-year career at CIA can look like for graduates coming into the workforce with degrees in statistics or computer science. I began work in a small data science unit and have progresses through many different jobs to my current position which is the Chief of Data Science for the CIA. I'd like to give some career highlights and give a sense for what working at CIA is a good choice for students entering today's workforce with data science skills.

4:20pm - 4:40pm

Break

4:40pm - 6:00pm

Session III (Session Chair: Douglas Zahn)

(4:40pm - 5:00pm)

Edsel A. Pena, University of South Carolina
Nonparametric Confidence Regions

In this talk I will revisit the classical problem of constructing confidence regions for a finite-dimensional parameter in nonparametric models. In particular, focus will be on the issue of optimality of the resulting confidence regions using invariance arguments. This work will have relevance in the construction of multiple confidence regions which will be relevant in high-dimensional problems akin to the problem of multiple hypothesis testing. This is joint work with my student Taeho Kim.

(5:00pm - 5:20pm)

Benedikt Jóhannesson, Former Minister Of Finance Of Iceland
Does It Help To Have A Ph.d. In Statistics If You Become Minister Of Finance?

On October 6, 2008, a day that lives in infamy, the world caved in and Iceland collapsed. A week later I gave a speech in which I declared: "Many have worried about the unfair advantages the rich have in Iceland. Now we don't have that problem anymore. – We are all poor." It was a joke, and everyone laughed, but for the next three or four years economic and mental depression almost suffocated Iceland.

Gradually things improved economically. However, no active politician seemed to share my conviction that for a solid future in a global economy we had to Simplify and Unify!

In 2016 I formed a new political party in Iceland, the Liberal Reform Party. I was the first Chairman and in parliamentary elections that fall we won over 10% of the public vote and I became Minister of Finance in a coalition government.

Did it help me to have a background in mathematics and statistics? I was not creating any mathematical models. But:
1. I understood the importance of such models, and their inherent weaknesses.
2. Unlike many mathematicians and even some statisticians, I have always been interested in numbers.
3. When you have a Ph.D. in statistics or mathematics, people give you more respect than someone with a degree in political science.

So the answer is: Yes!

Lastly, in our mathematical world we often search for optimal solutions. In the real world such a solution may not exist, and even if it does, we can never be sure we found it.

(5:20pm - 5:40pm)

John W. Robinson, Minnesota Department of Commerce
An Insurance Regulator' Concerns With Data Analytics

Data Analytics has become a must-have in the world of business decision-making. Actuaries have been performing statistical analysis since the advent of the profession, but we have only recently incorporated the new tools and techniques into our rigorous examination system. We believe that actuaries are well-suited to the field, because we combine the technical knowledge of statistics with our understanding of the business to which it is being applied, in our case, insurance.

However, it is not all about the Statistics. We have to be concerned that the Statistics is not being used in socially unacceptable ways. This brief presentation will highlight a few concerns, from the perspective of someone charged with rule-making for the insurance industry.

(5:40pm - 6:00pm)

N. Rao Chaganty, Old Dominion University
Models For Selecting Differentially Expressed Genes In Microarray Experiments

There have been many advances in microarray technology, enabling researchers to quantitatively analyze expression levels of thousands of genes simultaneously. Two types of microarray chips are currently in practice - the spotted cDNA chip developed by microbiologists at Stanford University in the mid-1990's and the oligonucleotide array first commercially released by Affymetrix Corporation in 1996. Our focus is on the spotted cDNA chip, which is more popular than the later microarray. In a cDNA microarray, or "two-channel array," the experimental sam-

ple is tagged with red dye and hybridized along with a reference sample tagged with green dye on a chip which consists of thousands of spots. Each spot contains preset oligonucleotides. The red and green intensities are measured at each spot by using a fluorescent scanner. In this talk, we aim to discuss bivariate statistical models for the red and green intensities, which enable us to select differentially expressed genes.

6:00pm - 7:00pm

Break
Group Photos

7:00pm - 9:30pm

Banquet (Session Chair: Wei Wu)
Greetings from former chairs and retired faculty
Prize conferring for logo contest winners

(8:30pm - 9:00pm)

Banquet Talk
A. Blanton Godfrey, North Carolina State University
Statistical Challenges In Reducing Maternal And Child Mortality

Every year, an estimated 303,000 mothers die from preventable causes. Over 5.6 million children die each year before the age of five. Sustainable Development Goal 3 for 2030 includes reducing maternal and child mortality by 50%. This goal has been endorsed by all 193 United Nations members. For the past five years, we have been working with colleagues in the Gillings School of Global Public Health at the University of North Carolina at Chapel Hill to create a new Institute for Implementing Global Health Solutions to support this goal. The World Health Organization, UNICEF, the World Bank, the Global Financing Facility, the UN Family Planning Agency, and Lancet are partners for this work. Part of this new initiative is a Dynamic Interactive Data Visualization and Utilization Lab – that is the part we are leading and will be the focus of this presentation.

One of our challenges is exploring, understanding, and presenting the incredible amount of data on maternal and child mortality and the causes of death in 193 countries over the past twenty-five years. Many countries have impressive results in reducing maternal and child mortality. What is most important is how they have done this, what are the best practices, and can these best practices be replicated in other locations to accelerate reductions of maternal and child mortality. Fortunately, we have a good start. Through a generous gift from the Bill and Melinda Gates Foundation, the Institute for Health Metrics and Evaluation at the University of Washington has created an excellent database on maternal and child mortality. Truly useful dynamic interactive data visualization software is becoming more and more available every year, and it is becoming easier to see where the progress is. The hard part is still discovering what was actually done to create these results and how to determine if other countries with similar results, cultures and envi-

ronments could use similar approaches. In this presentation we will demonstrate many of these new visualization tools while describing many of the statistical challenges we are still facing.

8:30am - 9:00am

Refreshments

9:00am - 10:20am

Session IV-A: (Room 103, Session Chair: Fred Huffer)

Room 103
(9:00am - 9:20am)

Balasubramanian Narasimhan, Stanford University
Tools For Convex Optimization

Domain specific languages, which are specialized languages implemented in a general purpose programming language, have become useful tools for expressing, manipulating, and solving problems in specific contexts. In convex optimization, a modeling methodology called disciplined convex programming imposes a set of conventions one must follow for constructing convex programs and is implemented in several languages: `CVX` in `Matlab`, `CVXPY` in `Python`, and `Convex.jl` in `Julia`. `CVXR`, our R language implementation, allows the user to formulate convex optimization problems in a natural mathematical syntax rather than the restrictive standard form required by most solvers. One specifies an objective and set of constraints by combining constants, variables, and parameters using a library of functions with known mathematical properties. `CVXR` then applies signed disciplined convex programming to verify the problem's convexity. Once verified, the problem is converted into standard conic form using graph implementations and passed to a cone solver. We demonstrate the `CVXR` modeling framework with applications in statistics and machine learning. This is joint work with Anqi Fu and Stephen Boyd.

Room 103
(9:20am - 9:40am)

Stephen Lee, University of Idaho
Text Mining

Text Mining methods are discussed and employed to find interesting words, ngrams, phrases, patterns, and documents in sequence data including English Texts and Genomic DNA data. Results are presented using Directed Graphs, Word Clouds, and Multidimensional Scaling. Application areas include the discovery of functional elements or genomic words in living organisms like human, plant, bacteria and mold; and the revealing of characteristics streaming tokens, words, and phrases in text corpora like the Bible and the Koran.

Room 103
(9:40am - 10:00am)

Shanti Gomatam, Food & Drug Administration
Safety Assessment Via Large Cardiovascular Outcomes Trials

In July 2008 an FDA advisory committee was convened to discuss the role of cardiovascular assessment in the pre- and post-market settings. Subsequent to this meeting the FDA issued a draft Guidance document "Diabetes Mellitus --- Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes" in December 2008 to provide guidance on addressing concerns about cardiovascular risk of treatments for Type 2 Diabetes. The findings and impact of some cardiovascular outcomes trials conducted in the intervening 10 years to satisfy the guidance requirements will be discussed.

Room 103
(10:00am - 10:20am)

Ivo Dinov, University of Michigan
DataSifter: Sharing of Sensitive Data via Statistical Obfuscation

There are no practical, reliable, and effective mechanisms to share sensitive information to inspire novel methodological developments without compromising intellectual property, confidentiality, personal data. In many fields, like health, financial, intelligence, socioeconomics, high-dimensional data is prevalent and there is a profound need to develop advanced data interrogation techniques to extracting useful and actionable information the balancing the utility of the data with the risk of exposing private, personal, or secure organizational information. Excessive scrambling or encoding of the information makes it less useful for modelling, or analytical processing. Insufficient preprocessing may uncover sensitive information and introduce a substantial risk for re-identification of individuals or trade secrets by various stratification techniques. To address this problem, we developed a novel statistical method (DataSifter) that provides on-the-fly de-identification of sensitive structured and unstructured high dimensional data, such as clinical data from electronic health records (EHR). DataSifter technology enables administrative control over the balance between risk of data re-identification and preservation of the data information content. Under careful set up of user-defined privacy levels, our simulation experiments and real biomedical case-studies suggest that the DataSifter protects privacy while maintaining data utility for different types of outcomes of interest. The application of DataSifter on ABIDE data provides a realistic demonstration of how to employ the proposed algorithm on EHR with more than 500 features (www.DataSifter.org).

9:00am-10:20am

Session IV-B: (Room 101, Session Chair: Xufeng Niu)

Room 101
(9:00am - 9:20am)

Leif Ellingson, Texas Tech University
Classification Of Protein Binding Sites Using Their Structural Information Via Object Data Analysis

It is known that a protein's biological function is in some way related to its physical structure. However, it remains an open challenge to predict a protein's function from its structure. Many approaches to this problem involve computing univariate descriptors for a protein's structure, resulting in much information being lost. Another option is to utilize higher level object representations of this information. In both cases, a problem that must be overcome is that the molecular structures of proteins are provided with respect to arbitrary coordinate systems, which commonly results in the need to perform computationally expensive pairwise alignments of the protein structures. The pairwise comparisons obtained from such procedures are often also difficult to use for model-based classification.

The methodology considered here represents protein binding site structures as covariance matrices in a manner that is invariant to rotation and translation, thus eliminating the need for pairwise alignment. It also simultaneouslyani reduces information loss compared to univariate representations and allows for model-based classification using Mahalanobis distance. This approach is illustrated for two benchmark data sets. If time permits, the impact of the choice of distance on the space of symmetric positive definite matrices will be discussed.

Room 101
(9:20am - 9:40am)

Naomi Brownstein, Moffitt Cancer Center
Multimodality Tests And Their Implementation In Sas Software

Multimodality tests investigate if an unknown density contains more than one mode and are popular in applications such as cluster analysis. Two popular multimodality tests are the Dip Test of Unimodality and Silverman's Critical Bandwidth Test. These tests are easy to conduct in R but are not currently included in base SAS software. This talk introduces implementations of the Dip and Silverman Tests as macros in SAS software, capitalizing on the capability of SAS to execute R code internally. An overview of the installation steps and example macro calls on real and simulated data are provided.

Room 101
(9:40am - 10:00am)

James Derek Tucker, Sandia National Laboratories
Elastic Data Depth

Statistical inference on functional data is becoming more prevalent as collection methods become more advanced. The problem of statistical analysis and

modeling of functional data in process control is important in determining when a production has moved beyond a baseline. Similarly, in many biomedical applications, doctors use long, approximately periodic signals (such as the electrocardiogram) to diagnose and monitor diseases. In this context, it is desirable to identify abnormalities in these signals. In particular, we develop a method known as Elastic Data Depth. Data Depth is the notion of ranking observations by the "centralness". We define two measures of depth that are able to measure both types of variability (amplitude and phase) that are found in functional data. We will present results in detecting outliers on a simple simulated data example and a real world electrocardiogram data set.

Room 101
(10:00am - 10:20am)

Mingfei Qiu, Federal Home Loan Bank of Atlanta
How Statistical Courses Are Applied to Our Work

Statistical tools are widely used in the banking industry, especially in the risk management area. This talk will briefly introduce how the topics in the courses, such as logistic regression, hypothesis testing and distribution theory, are implemented into the credit risk, operational risk and market risk management.

10:20am - 11:00am

Break
Group Photos

11:00am - 12:00pm

Session V-A: (Room 103, Session Chair: Daniel McGee)

Room 103
(11:00am - 11:20am)

Michiko I. Wolcott, Msight Analytics, LLC
A Culture of Empirically Informed Decision Making: The First Step

In today's world, the need for data is largely a matter of "how" rather than "if." While the strategic importance of data is widely acknowledged, many continue to struggle with achieving organizational maturity with respect to data; statisticians and analytics professionals, challenged with moving the cultural needle, often become disappointed and frustrated. In this talk, we will start to scratch the surface on our roles and responsibilities in a culture that embraces data and analytics, and some critical aspects in a change toward such culture.

Saturday, April 13: Parrallel Sessions (103 & 101)

Room 103
(11:20am - 11:40am)

Billy Franks, Astellas Pharma B.V.
Neural Networks – Do They Take Human Thinking Out Of Artificial Intelligence?

Much ado has been made about neural networks and deep learning as methods to derive insights from data in a highly automated way. This talk will highlight the structure of neural networks, contrasted to logistic regression, and then walk the audience through a basic simulated data set to highlight the unexpected behavior of neural networks on discovering very basic linear structures. No solutions will be provided, but hopefully interesting questions will be raised.

Room 103
(11:40am - 12:00pm)

Ryan Scolnik, FortressIQ
Annual FSU Statistics Data Science Competition

For the past two years, the FSU Statistics department has run a data science competition in an effort to encourage students to apply their learnings from the classroom to data in the real world. The competitions are open to all active graduate students in the department and the winners are awarded cash prizes. This presentation will outline the competition process, highlight student work from previous years and provide information on how to become involved as an alumnus.

11:00am - 12:00pm

Session V-B: (Room 101, Session Chair: Lifeng Lin)

Room 101
(11:00am - 11:20am)

Guanyu Hu, University of Connecticut
Subsampled Bayesian Approach for Big Data

Markov Chain Monte Carlo (MCMC) requires to evaluate the full data likelihood at different parameter values iteratively and are often computationally infeasible for large data sets. In this paper, we propose to approximate the log-likelihood with subsamples taken according to nonuniform subsampling probabilities, and derive the most likely optimal (MLO) subsampling probabilities for better approximation. Compared with existing subsampled MCMC algorithm with equal subsampling probability, our MLO subsampled MCMC has a higher estimation efficiency with the same subsampling ratio. We also derive a formula using the asymptotic distribution of the subsampled log-likelihood to determine the required subsample size in each MCMC iteration for a given level of precision. This formula is used to develop an adaptive version of the MLO subsampled MCMC algorithm. Numerical experiments demonstrate that the proposed method outperforms the uniform subsampled MCMC.

Room 101
(11:20am - 11:40am)

Kumarsh Dhara, University Of Florida
Variable Selection In Enriched Dirichlet Process With Applications To Causal Inference

Dirichlet process mixtures are often used to model the joint distribution of a response and predictors. However, the clusters formed when fitting the model often depends heavily on the covariates. Enriched Dirichlet process priors (EDP) overcomes these issues by modeling the joint distribution of response and predictors using a nested structure. EDP has been recently used in causal inference. It is common that a large number of covariates are available for modeling the response but only a few of them are important. In this paper, we propose a variable selection approach while using an enriched Dirichlet process. Removing irrelevant covariates helps in efficient and simpler modeling of the joint structure of the response and covariates.

Room 101
(11:40am - 12:00pm)

Xin Li, Oak Ridge National Lab
Graph-Bootstrapping: Applications In Functional Imaging Of Materials

With the advent of hardware, the functional imaging of materials community is rapidly tending towards simultaneous capture of multiple imaging channels and complex modes of operation involving high-dimensional and information-rich datasets, bringing forward the challenges of visualization and analysis, particularly for cases where the underlying dynamic and physical model is poorly understood. To address this issue, as initial efforts, recently developed methods based on nearest neighbor graph construction and manifold clustering have shown potential for exploratory data analysis, facilitating extraction of physics between material and imaging systems. Due to the inherent absence of labels within the field of unsupervised machine learning, the ultimate validation of our work is the discovery of interesting, useful and externally validated results. External validation, such as on data with physical meaning, is an example of how we hope to develop mathematically sound and broadly applicable techniques, while simultaneously confirming the physical interpretations.

The relationships within measurements can be preserved by constructing the nearest neighbor graph. A recent method LargeVis efficiently calculates the approximate nearest neighbor graph and layouts the graph in a low-dimensional manifold space via solving a principled probability model. The graph-construction-manifold-layout frame is also interpreted via algebraic topology theory by the UMAP method. We argue that the intrinsic low dimensionality of the physics suggests the presence of the low-dimensional manifold can be derived from

the high-dimensional measurements. In practice, we found the initial manifold layout from the nearest neighbor graph tend to be few bulks, made the following clustering task challenging. Trying to separate manifold clusters and present them in a clearer way, we empirically proposed Graph-Bootstrapping procedure that iteratively reconstructs the nearest neighbor graph based on previous manifold positions and then recalculates manifold coordinates based on the reconstructed graph. We have tested this procedure on LargeVis and UMAP algorithms. With HDBSCAN clustering, this procedure proves effective on different domain applications.

We hope these new applications and encouraging results will possibly provide a playground for establishing comprehensive statistical inference framework for manifold learning and clustering of graphs in a practically approachable way, and stimulate synergistic collaborations between microscopists, physicists, material

12:00pm - 1:00pm

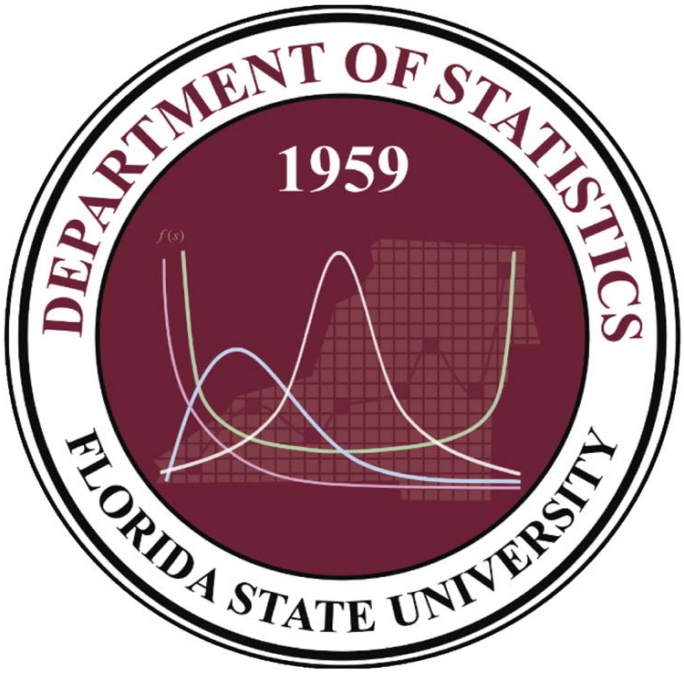
Pizza Lunch (Room 103)

1:00pm - 2:00pm

Panel Discussion (Room 103)
(Coordinator: Kristine Rosenberger)
Panelists: Larry Crow, Benedikt Johannesson, Hulin Wu, Yichuan Zhao

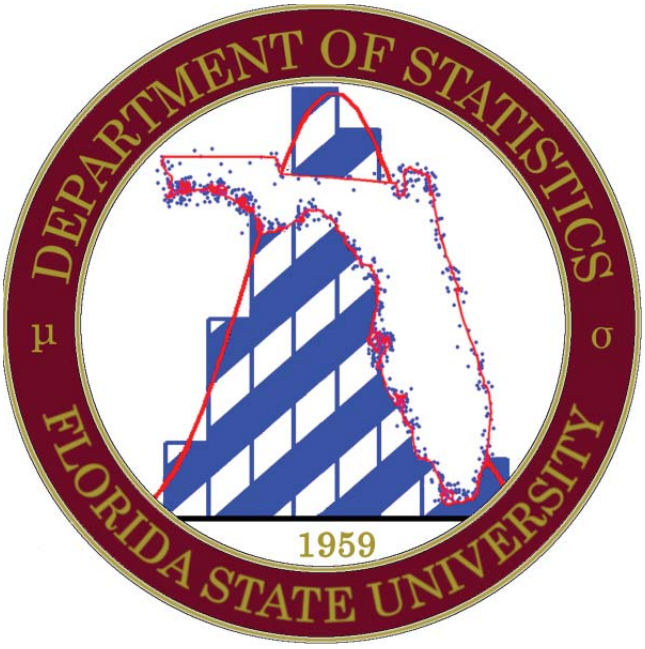
Departmental Logo Design Contest Winners

1st Place Winner



2nd Place Winner

3rd Place Winner



Welcome Back to Florida State University!

The Florida State University Department of Statistics is delighted to have our alumni and friends join us for our 60th anniversary celebration! We hope you enjoy your time in Tallahassee, Florida! For those who are not joining us for conference meals, we have provided information on some local restaurants and attractions nearby:

MADISON SOCIAL
– TALLAHASSEE –

Madison Social on the edge of FSU campus delivers classic pub fare with a twist. On weekends, the menu highlights gourmet brunch items, Blood Marys, mimosas and sangria. Patrons from Seminole fans to happy-hour professionals rave about the hand-crafted cocktails and beer cocktails. 705 S. Woodward Avenue.

COOSH'S

Locally owned and operated, Coosh's brings a taste of Cajun country to Tallahassee. Family-friendly, with daily specials. Staples include gumbo, jambalaya and crawfish etoufee. Their address is 705 South Woodward B102.

Harry's
SEAFOOD • BAR GRILLE
"A New Orleans Style Restaurant"

Harry's Seafood Bar and Grille

Harry's of Tallahassee is the place for Cajun, Creole, and Southern-style cooking. They are located at 301 South Bronough Street, and definitely worth the drive. Their wide menu selection means that there's truly something for everyone!

ITALIAN
Centrale
PARLOUR

Centrale Italian Parlour

Casual Natured. Italian Hearted. Meet Centrale, a new Italian Parlour serving up old-school Italian fare with a fun, fresh edge. Their address is 815 W. Madison St..

NEFETARI'S
Fine Cuisine & Spirits

Nefetari's Fine Cuisine and Spirits

Inspired by the Egyptian Queen Nefetari, this restaurant endeavors to treat each of their customers as royalty. They offer international dishes, as well as an assortment of vegetarian/vegan and gluten-free options. Their address is 812 South Macomb Street.

LITTLE
MASA

Little Masa is a Fast-Casual counter service style restaurant serving all of your Asian favorites including: wok-classics, noodles of Asia, and sushi. The address is 619 South Woodward Ave. The original and popular Masa restaurant is at 1650 N. Monroe St.