UP-STAT 2017
ABSTRACTS

FR. HAUS MEMORIAL MATHEMATICS LECTURE — SH 1013 A-B

3:30-4:30 **Diverse Applications of Probabilistic Record Linkage**

Robert Bell
Research and Machine Intelligence, Google
E-mail: rbellnj@gmail.com

The classical usage of probabilistic record linkage is to link common entities between two files using multiple, imperfect linkage variables like name, address, and birth date. The technique handles linking variables that are non-unique and/or error prone. The methodology also applies to many unexpected problems, invariably with interesting twists. I will illustrate with examples from some or all of public health, the decennial census, survey data collection, criminal justice, and software development.

POSTER SESSION

6:00-7:00 **The efficacy of antibiotic combinations in treating MRSA §**

Alexander Webber and Ryan Carpenter
Department of Mathematics, SUNY Geneseo
E-mail: aew11@geneseo.edu

The purpose of this study was to investigate which combinations of antibiotics are most efficacious when treating a patient for a hospital-acquired methicillin-resistant Staphylococcus aureus (MRSA) infection. Data were obtained from the freely available MIMIC III database, from the Beth Israel Deaconess Medical center in Boston Massachusetts. Data sets from the database were merged by patient identifiers and filtered for variables pertinent to MRSA infections, antibiotic usage, and patient demographic information. All data formatting and analyses were completed using the statistical software R Studio. We share the results of how the combinations of antibiotics and the doses per combination correlate with patient recovery time. We also explore how combinations containing one or more penicillin-type antibiotic(s) correlate with efficacy in treating MRSA.

A novel exact method for significance of higher criticism via Steck's determinant §

Jiefei Wang
Department of Biostatistics, SUNY at Buffalo
E-mail: jwang96@buffalo.edu

The higher criticism statistic is used in large-scale inference to assess whether there is a rare and weak signal present in the data set. Higher criticism statistics can be employed in gene expression microarray analysis, image detection, and many other fields where one is interested in testing a joint null hypothesis. The p-value for the higher criticism test is often given by simulation or approximation which can be time consuming or unreliable in small samples. We propose a novel straightforward way using Steck's determinant to
evaluate the p-value for a higher criticism test. The method allows users to quickly and accurately perform an exact test of significance for higher criticism.

On the strong control of Top-*K* error rates §

Ziqiang Chen, Jeffrey Miecznikowski, and Daniel Gaile
Department of Biostatistics, SUNY at Buffalo
David Tritchler

Dalla Lana School of Public Health, University of Toronto
E-mail: ziqiangc@buffalo.edu

We consider the control of novel error rates in the context of a two stage study design in which *M* feature-level assay values are measured for each of *N* samples in the first stage, and the Top-*K* features are identified and carried forward to a second stage validation study. We consider the test of the null hypothesis that the Top-*K* list is comprised completely of "null" features (e.g., features that are not truly differentially expressed across conditions). Existing methods of error control, e.g., family-wise error rate (FWER) and false discovery rate (FDR), do not provide proper error control for testing such a hypothesis. We extend our previous work, in which we provided a seminal exact test for the weak control of the "Top-*K* FWER", to explore what is meant by "strong" control in the context of our hypothesis of interest. We elucidate, via proofs and simulation studies, the conditions under which we can control an error rate that is relevant to the Top-*K* hypothesis test. In doing so, we demonstrate that our proposed method does indeed have utility in the context of Top-*K* two stage study designs, as it provides reasonable control of an error rate that is more relevant for than the error rates that are controlled by existing methods such as FWER and FDR.

testforDEP: An R package for distribution-free tests and visualization tools for independence §

En-shuo Hsu, Department of Biostatistics, SUNY at Buffalo
E-mail: enshuohs@buffalo.edu

This presentation summarizes testforDEP, a portmanteau R package containing several tests and visualization tools to examine independence between two variables. This new package combines classical tests including Pearson's product-moment correlation test, Kendall's rank correlation test, and Spearman's rank correlation test, with modern tests consisting of Vexler's density-based empirical likelihood ratio test, Kallenberg's data-driven rank test, the maximal information coefficient test, Hoeffding's test, and an empirical likelihood-based test. The function testforDEP provides an interface to the test and returns the test statistics, corresponding p-values, and bootstrap confidence intervals. The function AUK provides an interface for Kendall plots and computes the area under the Kendall curve. Here, we provide an introduction to the tests and visualization tools, discuss the optimizations used, and provide an example to demonstrate the package interface.

Predicting Premier League football match outcomes using data analysis §

Joshua Kolodny
Department of Philosophy, SUNY Geneseo
Xiaobai Guo
Department of Physics, SUNY Geneseo
E-mail: jk52@geneseo.edu

For this data analysis project, we investigated the extent to which soccer matches in the English Premier League are predictable, and what methods of prediction yielded the best results. To do this, we tested several methods against each other with regard to how well they could predict 2016-2017 soccer match outcomes. We used the following methods: multinomial logistic regression, tree regression, educated guessing, random chance, a machine learning algorithm and Nate Silver's predictions on his sports blog FiveThirtyEight. To train the models, we used the data that contained Premier League information from 2010-2015 which took into account a variety of factors: budget differences, number of years at top league, home field advantage, FIFA ratings, and difference in the number of wins for the previous month of play.

Using R to analyze e-sports §

Ruotian Zhang
Department of Computer and Information Sciences, SUNY Fredonia
Reneta Barneva
Department of Applied Professional Studies, SUNY Fredonia
E-mail: zhan9511@fredonia.edu

Big data and e-sports are terms that have gained enormous popularity in the last few years. We are combining these hot topics using big data to analyze one of the newest popular video games – Overwatch, released by Blizzard Entertainment in May 2016 and one of the most popular new games in e-sports. There exist championships in which professional teams of six individuals take part. We explore the data from the professional games in order to determine what skill set is important for the players and how the tournaments could be organized in an attractive way. For this purpose, we developed software in the R language. The study may be of interest to both sports professionals and computer scientists.

SESSION 1A — SH 1053

Innovations to Improve Introductory Statistics Courses

9:30-9:50 **Challenges and successes of flipping an introductory statistics course §**

Joseph Ciminelli
Department of Biostatistics and Computational Biology, University of Rochester Medical Center
E-mail: Joseph_Ciminelli@urmc.rochester.edu

Introductory statistics courses have increasingly become a requirement or recommendation for students to complete during their undergraduate education. In many instances, this influx of student enrollment in introductory statistics courses has led to large class sizes where traditional lectures dominate as the most straightforward approach for disseminating information. We investigated whether a flipped classroom—a more active learning style than lectures—is a more beneficial environment for students learning introductory statistics. In our designed flipped classroom, students watched video lectures outside of class and completed collaborative, application-based workshops in class that relied upon the use of real data. Data sets were used over multiple weeks, allowing students to become familiar with the complete process of data exploration, estimation, inference, and reporting. Floating teaching assistants provided real-time feedback and support for students as they completed activities within small groups. Students positively responded to this active learning environment, with survey evidence supporting the continuation of such an application-based flipped classroom for introductory statistics. In this presentation, we will discuss the logistical setup of the flipped classroom, along with presenting student survey results.

9:55-10:15 **The relationship between anxiety and performance in a statistics class**

Susan Mason, Sara Ribble, and Brianna Chupa
Department of Psychology, Niagara University
Elizabeth Reid
Department of Mathematics, Elmira College
E-mail: sem@niagara.edu

For many students, a required course in statistics is seen as a necessary evil. The students fear a heavy workload, challenges beyond their skill levels, public embarrassment, failing grades, even the end of their dreams of graduate school and a career. Moderate levels of anxiety can motivate students to work hard and can, therefore, contribute to successful performance. However, when the fear of failure is so great that it paralyzes a student, it can be a self-fulfilling prophecy. This presentation builds on information presented at the 2016 UP-STAT Conference, where we outlined a course structure and support system designed to reduce student anxiety levels and optimize student performance. We review a series of studies evaluating the system, and offer suggestions for future research and application. Factors associated with lower levels of anxiety include the student's background and skills, high grade expectations, in-class review of the syllabus, a supportive class atmosphere, practice working on problems, and peer mentoring. Areas for future research include student self-assessments as predictors of performance, the relationship between grit and success in statistics, and the value of emphasizing understanding over memorization.

SESSION 1B — SH 1028

Model Selection for Various Purposes via Statistical Learning Methods and their Application

9:30-9:45 **A comparative study of subgroup identification methods for differential treatment effect: performance metrics and recommendations §**

Yang Chen and Marianthi Markatou
Department of Biostatistics, SUNY at Buffalo
Demissie Alemayehu
Pfizer Inc.
E-mail: ychen57@buffalo.edu

Subgroup identification with differential treatment effects serves as an important step towards precision medicine, as it provides evidence regarding how individuals with specific characteristics respond to a given treatment. This knowledge not only supports the tailoring of treatment strategies but also prompts the development of new treatments. This presentation provides a brief overview of the issues associated with the methodologies aimed at identifying subgroups with differential treatment effects, and studies in depth the operational characteristics of five data-driven methods that have appeared recently in the literature. The performance of the methods under study to identify correctly the covariates affecting treatment effects is evaluated via simulation and under various conditions. Two clinical trial data sets are also used to illustrate the application of these methods. Discussion and recommendations pertaining to the use of these methods are provided, with emphasis on the relative performance of the methods under the conditions studied.

9:45-10:00 **Tuning parameter selection in the LASSO with unspecified propensity §**

Yang Yang and Jiwei Zhao
Department of Biostatistics, SUNY at Buffalo
E-mail: yyang39@buffalo.edu

The least absolute shrinkage and selection operator (LASSO), introduced about two decades ago, is one of the most significant model selection methods. Although it has attracted much attention in both theory and computation, many difficulties are still encountered in real applications. For instance, in a real data set, we may have various missing values. To correctly adopt the LASSO, we have to incorporate the missing data mechanism, or the propensity, in the penalized likelihood. Also, how to choose the tuning parameter is still an open problem, especially with a messy data set. Two distinct contributions make our work different from that in the existing literature. First, we allow the data set to have missing values by imposing a very general and flexible propensity. Compared to the missing data methods with a concrete propensity, this assumption is easier to satisfy in reality and it makes our methodology more robust. Many existing ignorable or nonignorable missing data situations, as well as some biased sampling problems, belong to a special case under our consideration. Second, to determine the tuning parameter, we examine four different methods including cross-validation (CV), Bayesian information criterion (BIC), and two others focusing on estimation stability and variable selection stability. To illustrate our methods, we conduct comprehensive simulation studies and apply our methods to real data from a melanoma study.

10:00-10:15 **Boostrapping estimates of clustering stability with applications to model selection and large-scale data §**

Han Yu, Brian Chapman, Arianna DiFlorio, Ellen Eischen, David Gotz, Matthews Jacob, and Rachael Hageman Blair
Department of Biostatistics, SUNY at Buffalo
E-mail: hyu9@buffalo.edu

Clustering is a challenging problem in unsupervised learning. In lieu of a gold standard, stability is a valuable surrogate to performance and robustness. In this work, we propose a non-parametric bootstrapping approach to estimating the stability of a clustering method at multiple levels, including the stability of the individual clusters and observations. This flexible framework can be used in connection with two possible bootstrap approaches for stability. The first approach, scheme 1, is used to assess confidence (stability) around clustering from the original

dataset based on bootstrap replications. A second approach, scheme 2, searches over the bootstrap clusterings for an optimally stable partitioning of the data. The two schemes accommodate different model assumptions, which can be motivated by an investigator's trust (or lack thereof) in the original data. A hierarchical visualization can be extrapolated from the stability profiles that give insights into the separation of groups, and visualizations can be projected for the inspection of individual stability. We also demonstrate the applicability of this framework to large-scale data by using the Bag of Little Bootstraps procedure. Our approaches show good performance in simulation and on real data.

SESSION 1C — SH 1004

Text Mining: Methods and Applications

9:30-9:50 **The application of topic modeling in the Community Views on the Criminal Justice System project §**

Sujeong Seo
College of Science, Rochester Institute of Technology
E-mail: ss1526@rit.edu

The Community Views on Criminal Justice System (CVCJS) initiative was established to collect the Rochester community's perceptions on experiences with the Rochester Police Department (RPD) and the criminal justice system, and share those findings to inform local Gun Involved Violence Elimination (GIVE) strategies. Over one year, 19 focus groups were created across multiple backgrounds founded in police-citizen groups, re-entry groups, and community, neighborhood, and youth organizations. These groups had discussion sessions across six major topics, afterwards answering survey questions. Without preference of note takers or facilitator, and only referencing the quantitative results of the questions, the research team attempted to identify how people answered within those six topics. This presentation will be discussed via an empirical study with major text mining methods such as unsupervised topic modeling (i.e., Latent Dirichlet Allocation [LDA] analysis), one of the most widely used algorithms in the field of computer science, statistics, and machine learning. Data visualization techniques generated by R software will be presented along with text mining results. Further explained is how topic modeling has become a ubiquitous tool for understanding large corpora, providing an easy medium to present the analysis to researchers in social science who are not experienced in statistics and machine learning.

9:55-10:15 **Text mining with a Bayesian hidden topic Markov model §**

K. Tyler Wilcox
College of Science, Rochester Institute of Technology
E-mail: ktw5691@rit.edu

Topic models are a state-of-the-art probabilistic approach to extracting structural information about discrete data and have been used with great success for text mining. A recent development in topic modeling, the Hidden Topic Markov Model (HTMM), incorporates hidden Markov models, resulting in contiguous topic assignments and word sense disambiguation. This is a departure from the standard "bag-of-words" assumption of the seminal Latent Dirichlet Allocation (LDA). While LDA can handle word synonymy, it restricts each word to belong to a single topic. HTMM is capable of assigning a given word to multiple topics depending on location within a document. An expectation-maximization (EM) algorithm for inference was proposed by Gruber, Rosen-Zvi, and Weiss that takes advantage of existing EM techniques for hidden Markov models, but a full derivation was never published. I derive the special state space used by HTMM and approach inference for HTMM in two ways: First, a full derivation of an EM algorithm for frequentist inference and second, a novel Gibbs sampler for Bayesian inference.

SESSION 1D — SH 1013 A

Analysis Data from Electric Power System Networks and Astronomy

9:30-9:50 **Statistical approaches to anomaly detection and intuitive visualization of Phasor Measurement Unit (PMU) data from electric power systems**

Ernest Fokoué
College of Science, Rochester Institute of Technology
Esa Rantanen
Department of Psychology, Rochester Institute of Technology
Jacob Hunt
Department of Biology, Rochester Institute of Technology
E-mail: epfeqa@rit.edu

The monitoring by human operators of electric power system networks is an extremely complex task due to the overwhelmingly amount of information human operators need to be aware of at any given time. The so-called Phasor Measurement Unit (PMU) is one of the leading technologies used to collect data from large electric power system networks. The many advantages of PMU data include much finer-grain and faster analysis of the system than with older technology. However, the extremely high sampling rate at which the data are produced poses new challenges to human capacity to perceive and understand the information conveyed by the data. At any time, the system can be in one of five possible states, and it is critical for the human operator to know the state accurately to take the proper measures to keep the system in good working order. In this talk, I will harness the power of statistical machine learning and data mining methods to derive and develop several approaches to anomaly detection/identification: time series forecasting for early warning, dimensionality reduction/feature extraction, and intuitive visualization, with the overarching aim of presenting the multichannel data to the operator in a form that is both intuitively appealing, meaningful and useful, and providing powerful predictive information for fast and accurate intervention on the whole network. Among other things, we explore and present intuitively annotated stacked matrix plots of raw series along with the corresponding first and second differences, all aimed at revealing to the operator useful and meaningful summaries out of the large amount of data produced by the PMUs. We also present correlation plots to capture network interconnectedness, and we construct and display appropriately colored heat maps revealing regions of high risks.

9:55-10:15 **Inferring the rate and distribution of compact binary mergers observed through gravitational wave detectors using Markov chain Monte Carlo §**

Daniel Wysocki and Richard O'Shaughnessy
College of Science, Rochester Institute of Technology
E-mail: dw2081@rit.edu

With the advent of gravitational wave astronomy, following the first gravitational wave detection in September 2015 by the LIGO Scientific Collaboration and the Virgo Collaboration, we now have a means to measure the properties of gravitational wave emitters, including merging compact objects such as black holes and neutron stars. The expected number of merger events in a given volume of space and observation time can be modeled as an inhomogeneous Poisson process, with an overall rate and a probability distribution over the systems' various physical properties. This rate and distribution will provide information on the population of binary star systems, and can be compared with theoretical population models. In this talk, we describe a general Bayesian framework for inferring this rate and distribution using Markov chain Monte Carlo (MCMC). This method works on noisy data, and relies on a measure of the volume in space to which our detectors are sensitive (which is itself a function of the physical properties). This method relies on a parameterization of the probability distribution of system parameters, which can be restrictive (e.g., a power law) or flexible (e.g., a Gaussian mixture model). With some slight modifications, this method may be used in a variety of terrestrial applications, including inference on climate events and crime patterns.

SESSION 1E — SH 1013 B

Applications of Statistics in Epidemiology

9:30-9:50 **Predicting cholera-positive cases in Haiti §**

Jessica Young
College of Science, Rochester Institute of Technology

E-mail: jay4588@rit.edu

While Western countries typically run census surveys frequently, poorer countries such as Haiti do not have the money to do so; thus research into how Haitians live is severely lacking. Furthermore, studies that do exist tend to be not only old and outdated, but also lacking in depth. Using new census data recently collected from Haiti, I attempt to predict if certain behaviors and living situations can be used as indicators for determining if someone has cholera. Challenges for exploring these data center on getting the surveys into a format suitable for analysis and the severe class imbalance between the number of cholera positive people and cholera negative people. Numerous solutions to these problems are attempted including using different sampling techniques, using ensembles with models like CART and support vector machines, and Bayesian model averaging. Better survey designs and questions to add to future surveys are also discussed.

9:55-10:15 **Using Bayesian multilevel models to quantify variation in suboptimal lymph node examination after colectomy §**

Adan Becerra
Department of Public Health Sciences, University of Rochester
E-mail: adan_becerra@urmc.rochester.edu

The goals of this study were to characterize the variation in suboptimal lymph node examination for patients with colon cancer across individual surgeons, pathologists, and hospitals and to examine if this variation affects 5-year, disease-specific survival. A retrospective cohort study was conducted by merging the New York State Cancer Registry with the Statewide Planning and Research Cooperative System, Medicaid, and Medicare claims to identify resections for stages I-III colon cancer from 2004-2011. Bayesian multilevel logistic regression models characterized variation in suboptimal lymph node examination (< 12 lymph nodes). Multilevel competing-risks Cox models were used for survival analyses. The overall rate of suboptimal lymph node examination was 32% in 12,332 patients treated by 1,503 surgeons and 814 pathologists at 187 hospitals. Patient-level predictors of suboptimal lymph node examination were older age, male sex, nonscheduled admission, lesser stage, and left colectomy procedure. Hospital-level predictors of suboptimal lymph node examination were a nonacademic status, a rural setting, and a low annual number of resections for colon cancer. The percentages of the total clustering variance attributed to surgeons, pathologists, and hospitals were 8%, 23%, and 70%, respectively. The tertiles of surgeon-specific rates of suboptimal lymph node examination were 18%–27%, 27%–35%, and 35%–41%. The tertiles of pathologist-specific rates of suboptimal lymph node examination were 9%–39%, 39%–59%, and 59%–74%. The tertiles of hospital-specific rates of suboptimal lymph node examination were 2%–32%, 32%–55%, and 55%–96%. Increasing pathologist and hospital-specific rates of suboptimal lymph node examination were associated with worse 5-year, disease-specific survival. There was a large variation in suboptimal lymph node examination between surgeons, pathologists, and hospitals. Collaborative efforts that promote optimal examination of lymph nodes may improve prognosis for colon cancer patients. Given that 93% of the variation was attributable to pathologists and hospitals, endeavors in quality improvement should focus on these 2 settings.

SESSION 1F — SH 1008

Statistics in Sports

9:30-9:50 **Applying multi-resolution stochastic modeling to individual tennis points §**

Calvin Floyd
College of Science, Rochester Institute of Technology
E-mail: cmf1089@rit.edu

Individual tennis points evolve over time and space, as each of the two opposing players are constantly reacting and positioning themselves in response to strikes of the ball. However, these reactions are diminished into simple tally statistics such as the amount of winners a player has, or the percentage of first serves a player manages to keep in-bounds. In this talk, I propose a new way to evaluate how an individual tennis point is evolving, by measuring how much a player can expect each shot to contribute to a won point, given whether the player is striking or returning, and where both players are located. This measurement, named “Expected Shot Win Rate” (ESWR), derives from stochastically modeling each shot of individual tennis points. This will be modeled with multiple resolutions,

differentiating between the continuous player movement and discrete events such as strikes occurring and duration of shots ending. This multi-resolution stochastic modeling approach allows for the incorporation of information-rich spatio-temporal player-tracking data, while allowing for computational tractability on large amounts of data. In addition to calculating ESWR, this methodology will be able to highlight the strengths and weaknesses of specific players.

SESSION 2A — SH 1013 B

Model-Based Clustering with Longitudinal Data Applications

10:25-10:45 Clustering deviations in trajectories to understand factors impacting young females' physical activity levels §

Amy LaLonde, Tanzy Love, and Tongtong Wu
Department of Biostatistics and Computational Biology, University of Rochester Medical Center
E-mail: Amy_LaLonde@urmc.rochester.edu

Physical activity levels are declining for people of all ages, including adolescent girls. Previous analyses of longitudinal data from the Trial of Activity for Adolescent Girls from the University of Maryland field site examined the effects of various individual-, social-, and environmental-level factors on the change in physical activity levels over the course of nine years. Not surprisingly, over 90% of participants failed to meet the recommended 60 minutes of average daily moderate-to-vigorous physical activity. Using Bayesian model-based clustering within a linear mixed effects model fit to these trajectories, we find that the patterns of physical activity are starkly different for about 5% of the study sample. These young girls are exercising on average 30 more minutes per day after controlling for the variables deemed in previous analyses to predict physical activity levels. Such findings encourage further exploration of the underlying characteristics or factors impacting the physical activity levels among females.

10:50-11:10 Nonlinear mixed-effects mixture regression models for clustering longitudinal data §

Chongshu Chen
Department of Biostatistics and Computational Biology, University of Rochester Medical Center
E-mail: Chongshu_Chen@urmc.rochester.edu

Finite mixture models are increasingly used to model the distribution of mixtures of subpopulations in multivariate regression settings. Since observations collected from the same subject over time may be dependent, mixture model-based methods that assume independence of observations may not be optimal for modeling and clustering of longitudinal data, such as growth curve trajectories. We introduce a general class of finite mixture regression models with random effects to study the repeated measurements over time and capture variations between subjects and dependencies within units by using random effects. The models further allow the mixing proportions to depend on covariates or features. We propose a version of the stochastic approximation of EM (NR-SAEM) algorithm for maximum likelihood estimation and determine each individual's group membership probability. We illustrate the proposed method by using simulated and real data on repeated hormone levels over a period of the menstrual cycle. In the setting of a mixture regression model with linear random effects, results of estimates for the standard EM algorithm and the NR-SAEM algorithm are presented. Some limitations are that a parametric representation and specification of the variance-covariance structure are needed for the proposed method. It also requires the analyst to specify the number of components in the mixture regression model. The proposed NR-SAEM algorithm consistently estimates the true parameters and this procedure is less sensitive to starting values than the mixture stochastic approximation of EM (MSAEM) algorithm since the estimation procedure avoids using simulated random effects that potentially adds extra variation during the estimation process.

SESSION 2B — SH 1008

Winning with Statistics in Sports

10:25-10:40 A statistical analysis of the NFL Draft: valuing draft picks and predicting future player

success §

Nick Citrone and Samuel Ventura
Department of Statistics, Carnegie Mellon University
E-mail: rcitrone@andrew.cmu.edu

Each year, NFL teams spend an extraordinary amount of time and money scouting, interviewing, and evaluating players who have declared for the NFL Draft, and yet top selections still bust at a high rate. The objective of our research is to identify trends and make predictions that can be used by NFL teams to improve draft success. We approach two sub-problems at the NFL Draft statistically: assessing the value of individual draft picks conditional on the position of the player chosen, and predicting the future performance of individual players at the NFL level. For our analysis on maximizing draft value by position, we use local non-parametric regression to model the approximate value of each pick conditional on position and identify points in the draft when taking particular positions yields significantly higher relative projected approximate value than others. The data come from all 3,820 NFL draft picks made between 1999 and 2013, and were obtained from Pro Football Reference. We find statistically significant differences in the expected approximate value per season when drafting different positions in each round. Our research to predict individual player success uses the same NFL Draft data set in addition to college football player statistics from College Football Reference and NFL Combine scores from nflsavant.com. The model uses Pro Football Reference's approximate value per season as the response variable. Interestingly, we find that linear regression outperforms more complicated statistical models. The results obtained in our study can help NFL franchises identify potentially undervalued players in the NFL Draft by using their predicted approximate value per season. Additionally, teams can optimize which positions they select with certain picks by using our results on the relative value of selecting particular positions at different stages in the draft.

10:40-10:55 nflscrapR: An R package for easy access to NFL data and new model for expected points and win probability §

Ronald Yurko, Max Horowitz, and Samuel Ventura
Department of Statistics, Carnegie Mellon University
E-mail: ryurko21@gmail.com

The lack of publicly available National Football League (NFL) data sources has been a major obstacle in the creation of modern, reproducible research in football analytics. While clean play-by-play and season-level data are available via open-source software packages in other sports, the equivalent datasets are not freely available for researchers interested in the statistical analysis of the NFL. We create a publicly available, open-source R package called nflscrapR that allows easy access to NFL data from 2009-2016. Using a JSON API maintained by the NFL, this package downloads, cleans, parses, and outputs data sets at the individual play, player, game, and season levels. Our package allows for the advancement of NFL research in the public domain by allowing analysts to develop from a common source, enhancing reproducibility of NFL research. We demonstrate the use of our package in several ways. First, we introduce a new model for "expected points" (quantifying the value of individual plays in the context of an NFL game) that uses multinomial logistic regression with a novel observation weighting scheme that takes into account the number of drives until a future score. Second, we use generalized additive models to model each team's win probability given the situational data about each play. Next, we use these models to obtain the expected points added (EPA) and win probability added (WPA) for each individual play. Finally, we are building an accompanying web application that will include advanced splits and statistics, including EPA and WPA, along with in-game probability charts to visualize game trends. This website will make our work more accessible to the public domain, allowing the football analytics community to utilize the work we have done. We note that this work has broader applications (e.g., in player contract valuation) that we intend to explore in future work.

10:55-11:10 The data collection process and play selection in Division I college football §

Taylor Pellerin and Michael Schuckers
Department of Mathematics, St. Lawrence University
E-mail: tjpell13@stlawu.edu

Over the course of a summer fellowship and fall semester senior research, I downloaded a slightly dated but nonetheless massive data set, scraped more recent data and then ran multiple different regression models. The original data set I downloaded from cfbstats.com contained play-by-play statistics for every NCAA Division I

College Football game spanning the 2005 to 2013 seasons. I then built a set of linear and ridge regression models which looked at how well the run-pass decision and a few other factors did in predicting the change in expected points caused by each play of the games, where expected points is taken using a nearest neighbor approach. Nearest neighbor is taken to be the average points gained at the end of a drive for each scenario of down, distance and spot on the field. All scoring and turnover possibilities were handled, with a hefty negative weight being given to turnovers and defensive points. With this set of models, the next step turned to gathering more data. The website that had provided the first 9 years of statistics became a paid service, so scraping the rest of the data became necessary. To do this, I built an R package that, given a season schedule containing the date and teams involved in each game, produces a table of all of the play by play statistics, formatted in the same way as the data provided by cfbstats.com. The actual information is pulled from the same JSON that are used by ESPN.com to fill out their play-by-play statistics pages. This was done primarily using the jsonlite and deplyr R-packages. With these extra data, I then reran all of the original models, as well as a handful of others with new predictive factors, in order to make the analysis more robust.

SESSION 2C — SH 1028

Explorations of Echo State Networks

10:25-10:45 Design exploration on echo state networks §

Seyed Langroudi
Department of Computer Engineering, Rochester Institute of Technology
E-mail: sf3052@rit.edu

Deep learning algorithms are becoming the state of the art for designing different machine learning tasks such as classification and prediction. For instance, a deep learning algorithm performed image classification with 97% accuracy in the ImageNet data set, which outperformed human capability to accomplish the same task in this data set. However, accuracy is not the only criterion relevant to researchers; other criteria such as training time, memory requirement, and power dissipation are also essential. Unfortunately, deep learning architectures consume considerable power, memory and training time. An echo state network is one candidate to address these problems. The input layer, reservoir layer and output layer are three layers of an echo state network. The output of the input layer is fully connected to the reservoir layer. The reservoir layer includes neurons (nodes) that are connected forwardly and recurrently. The output of the reservoir layer is fully connected to output layer. The important feature of the echo state network is that only the output layer needs to be trained. Therefore, training algorithms in echo state network models are simpler than those for other recurrent neural networks. Echo state networks have been used in different application domains such as speech recognition, computer vision and natural language processing. Unfortunately, the echo state network cannot compete with other recurrent neural networks in terms of accuracy. To improve performance, we have to answer these questions: Which applications are suitable for echo state networks? What are the advantages and disadvantages of echo state networks? What network topologies should be used in the reservoir layer? Is it possible to statistically analyze echo state network models?

10:50-11:10 Statistics behind echo state networks §

Qiuyi Wu
College of Science, Rochester Institute of Technology
E-mail: qw9477@rit.edu

Generating Shakespeare masterpiece text, generating C programming language code, speech recognition, and image reconstruction are some of the applications of learning becoming more popular using deep convolutional neural networks and recurrent neural networks. Unfortunately, implementation of these algorithms is suffering from longer training time and an increased memory requirement. To solve these problems, an echo state network can be used as a partially-trained recurrent neural network. The echo state network model includes three layers: an input layer, a reservoir layer, and an output layer. The crucial layer in the echo state network is the reservoir layer with contains recurrent and forward connections between nodes. These connections are randomly selected without being learned. The output layer linearly combines the output signals with reservoir layer signals. Only the output layer weights need to be trained with simple linear regression algorithms or other learning algorithms. Thus, echo state network models have simplified training algorithms compared to other recurrent neural networks and are more efficient than kernel-based methods as they can incorporate temporal stimuli and the output relies on the current input and all

previous input states. Achieving good accuracy using the echo state network is a challenging task. In order to successfully implement artificial intelligence-related tasks with echo state networks, we have to find answers to these questions: What kind of randomness we can try? Does the shape of the distribution of the weights matter? What's the range of random numbers? Is there any reason why previous studies used the -0.5 to 0.5 range for random numbers? Shall we use a fixed leaking rate or a dynamic one?

SESSION 2D — SH 1013 A

Statistics and Technology

10:25-10:45 Determining unreported crime rate in Rochester from emergency call data §

Justin Comparetta
College of Science, Rochester Institute of Technology
E-mail: jc3339@rit.edu

Rochester, NY is one of six dozen cities nationwide to employ automatic gunshot recognition technology as part of its emergency response to gun violence. With strategically placed microphones throughout the city, this technology (named ShotSpotter) uses machine learning to differentiate gunshots from other loud noises (such as cars backfiring, fireworks, etc.). Upon detection of a suspected gunshot, officers are alerted to the location of the gunfire by triangulation of the signal. Using ShotSpotter data collected from 2015-2016 and comparing it to the number of gunshots that are reported via 911, we are able to ascertain the rate at which gunfire that goes unreported, by location, in the city. By combining this non-reported gunshot rate with the complete set of reported crime data that are publicly available from the Rochester Police Department, we are able to geographically correlate these areas with overall crime levels in the city. Finally the above data sets are combined with census data to train a model to predict gun violence.

10:50-11:10 Differential aspects and performance analysis of time difference of arrival (TDOA) technique of sound localization and modern audio recognition techniques of an assistive wearable device for the deaf or hard of hearing population §

Hrishikesh Karale and Gary Behm
College of Science, Rochester Institute of Technology
Subrina Farah
Department of Family Medicine, University of Rochester
E-mail: hhk9433@rit.edu

Cherry is an assistive wearable device aimed to aid people who are deaf or hard of hearing. Only 20% of the total number of people with hearing disabilities use aids in their daily lives. Some of the factors that keep people from using hearing aids include stigma towards people wearing these devices, value in terms of functionality provided by the aid, maintenance, efficiency etc. Our aim is to enhance the hearing experience for people who do not get a lot of value from their current hearing aid. The device consists of 4 microphones that are used to collect real time audio from the user's surrounding environment. The data are then processed using the time difference of arrival (TDOA) technique of sound localization. The task of locating the sound source is more difficult due to a number of complications such as finite source dimensions, reflections from the bottom of the surface, and nearby objects. Perfect solutions might not be possible, since the accuracy depends on the following factors: geometry of the microphone and source, accuracy of the microphone setup, uncertainties in the location of the microphones, lack of synchronization of the microphones, inexact propagation delays and many other issues related to the noise and bandwidth of the emitted pulses. Considering the limitation of the device, it was necessary to have audio data-based machine learning techniques to quantify the performance of the device. Therefore, multiple statistical techniques, including principal component analysis, discriminant analysis, and integrated phoneme subspace method (compound method) for audio feature extractions have been applied to observe differential aspects of the device and the captured audio data on detecting sound localization. A pattern recognition approach based on the hidden Markov model technique was conducted to achieve performance analysis.

SESSION 2E — SH 1004

Detecting Hidden Structure in Data: Anomaly Detection and Cadre Learning

10:25-10:45 Anomaly detection in chip manufacturing §

Andrés Vargas
Department of Mathematical Sciences, Rensselaer Polytechnic Institute
Ridwan Al Iqbal
Department of Computer Science, Rensselaer Polytechnic Institute
E-mail: vargaa5@rpi.edu

Detecting faults in a manufacturing process and finding the root causes quickly is essential for affordable operation of any manufacturing process. We can detect faults in a process by looking for hidden "anomalies" in a time-series data stream emitted by the process. The time points where anomalies occur can then serve as educated guesses for the time points where faults occur. We propose two different approaches for detecting faults in a chip-manufacturing process. One approach is to generate a test statistic from a moving window principal component analysis (PCA) at different time points, and then calculate the probability that each test statistic came from a distribution different from that of test statistics generated at previous points in time. Another approach is to use PCA to de-correlate each attribute of the multivariate time series, and then run independent ARMA (auto-regressive moving average) models on each of the de-correlated time series. In our talk, we discuss the theoretical details behind each of these approaches, and then present our results using data from a real-life chip manufacturing process.

10:50-11:10 Supervised learning of predictive cadres §

Alexander New
Department of Mathematical Sciences, Rensselaer Polytechnic Institute
E-mail: newa@rpi.edu

We consider supervised regression problems in which the population under study may be softly partitioned into a set of cadres. The cadres create clusters of observations based on only a few features. Within these cadres, the behavior of the target variable is more simply modeled than it is on the population as a whole. We introduce a discriminative model for a population that, when trained on a set of observations, simultaneously learns cadre assignment and target prediction rules. Our formulation allows sparse priors to be put on the model parameters. These priors allow for independent feature selection processes to be performed during both the cadre assignment and target prediction processes, which results in simple and interpretable ensemble models. A block coordinate descent algorithm for parameter learning is developed. We present simulated results showing that, under certain conditions, our method significantly exceeds the performance of simpler methods that learn clustering and target prediction rules separately. Further experimental results show that our method is competitive with powerful nonlinear models such as regression forests. Applied to cheminformatics, our model accurately predicts polymer glass transition temperatures. It identifies chemically meaningful cadres, each with interpretable models. Future work includes learning analytically distinct patient cohorts in electronic healthcare records analysis and expanding the model to classification tasks.

SESSION 2F — SH 1053

Issues in and Tools for Statistics Education

10:25-10:45 MET, SET, and so next?

Yusuf Bilgic
Department of Mathematics, SUNY Geneseo
E-mail: bilgic@geneseo.edu

Statistics education is vital for quantitative and stochastic literacy that mostly deals with data-based processes, reasoning and decisions. Common Core State Standards – Mathematics (CCSSM) places a heavy emphasis on the statistics content at all grade levels. The Conference Board of the Mathematical Sciences (CBMS), an umbrella

organization consisting of sixteen professional societies in the mathematical sciences including ASA and NCTM, released a report, The Mathematical Education of Teachers II (MET II Report), in 2012 to promote research, improve education, and expand the uses of mathematics. More recently, the Statistical Education of Teachers (SET Report), commissioned by the American Statistical Association (ASA), clarified the MET II recommendations to distinguish the practices of statistics and mathematics as well as to highlight the extras in the statistical preparation of teachers. In light of these latest reports, in my talk I will address some issues: In what ways are the statistical problem-solving processes different? Are the features of teachers' statistical preparation distinct from their mathematical preparation? What can we, as educators, mathematicians and statisticians, do to reach out to all the students who are struggling in a statistics classroom? I will share an instructional scaffold designed to address the struggling student in statistics instruction as well. Participants are encouraged to take a look at the two reports: (1) Met 2 Report (2012) found at http://cbmsweb.org/MET2/met2.pdf; and (2) SET Report (2015) found at www.amstat.org/asa/files/pdfs/EDU-SET.pdf.

10:50-11:10 **BiomarkerChallenge: An interactive R package and Shiny app to teach common statistical techniques used in evaluating genetic biomarkers §**

Luther Vucic and Daniel Gaile
Department of Biostatistics, SUNY at Buffalo
E-mail: vucicl@gmail.com

We developed a web based platform to facilitate an authentic group learning activity to teach statistical design and analysis concepts relevant to a common class of genomics based biomarker experiments. A freely available R package, BiomarkerChallenge, was authored to provide the desired functionality. A Shiny-based web applet was developed to facilitate a guided workflow and graphics to further the students' understanding. The BiomarkerChallenge package contains the functions necessary to create and analyze simulated array and validation datasets. Functions include purchasing array and validation data, principal component analysis, and t-test analysis with graphics that display which biomarkers are truly relevant. The applet contains guided tutorials for the exercise and all relevant statistical concepts. It also includes a dashboard that sets up the exercise and displays relevant information and graphics to further classroom discussion. Statistical concepts included are power, error control, multiple testing, dimension reduction, distance estimation, Fisher's exact test and clustering techniques.

SESSION 3A — SH 1004

Time-Course Models with Complex Variance Structure in High-Dimensional Data Analysis

11:20-11:35 **On the equivalence of regularized high-dimensional regression in time-course data analysis §**

Yun Zhang and Xing Qiu
Department of Biostatistics and Computational Biology, University of Rochester Medical Center
Juilee Thakar, Department of Microbiology and Immunology, University of Rochester Medical Center
E-mail: Yun_Zhang@urmc.rochester.edu

Time-course gene expression data are often used to study dynamic biological processes. Statistical analyses for time-course data are typically designed for group comparisons or regression analyses. However, these statistical analyses do not fully utilize the temporal information among sampling time points. Recent studies show that using functional data analysis techniques highly increases modelling efficiency in analyzing time-course data. We propose to use a functional concurrent regression model to study the temporal association between gene expression curves and their functional covariates. Furthermore, with genomics data, we face the "large p, small n" problem. A practical way to eliminate this problem is to use regularization methods with either L_1 penalty (ridge) or L_2 penalty (LASSO), or both (elastic-net). To our knowledge, a computationally efficient algorithm for regularized functional regression models is not currently available. In this report, we establish an equivalence between the functional model and the standard multivariate model in high-dimensional space, so that well-established methods for regularized multivariate regression can be directly applied. Termed FUNNEL (FUNctioNal ELastic-net), the equivalence transformation from functional regression to multivariate regression is implemented in

equiv.regression() in the FUNNEL R package. The methods are illustrated with an application to the influenza H3N2 virus infection. Optimal penalty parameter selection through cross-validation is also discussed.

11:35-11:50 **An efficient Monte Carlo sampling method for spherical polygons with applications to wireless communication studies §**

Jiatong Sui and Xing Qiu
Department of Biostatistics and Computational Biology, University of Rochester Medical Center
Hongjun Li
College of Science, Beijing Forestry University
E-mail: Jiatong_Sui@urmc.rochester.edu

Virtually all statistical research can benefit from an efficient, robust Monte Carlo sampling method for the hypothesized distribution of interest. While many such sampling methods exist for distributions defined on R^1, few are available for distributions defined on a manifold, such as the uniform distribution on a spherical polygon. Random sampling for an arbitrary spherical polygon is important not only because all spherical polygons can be decomposed into disjoint triangles, but also because a large family of smooth shapes on S^2 can be approximated by finite spherical triangular mesh; so that the uniform distribution of an arbitrary smooth closed manifold can be approximated efficiently. The commonly used generic approach is to generate spatial samples that are uniformly distributed on the unit sphere, then check if the sample is in the specific polygon by means of cross-product. This approach is inefficient because computing the cross-product for every sampling point is costly; and it is exacerbated if the area of the spherical triangle is small relative to the unit sphere, so it takes many samplings on the unit sphere to produce one useful sample on the spherical triangle. In this talk, we present an algorithm based on the spherical coordinate system and spherical calculus. We have derived the joint distribution of the polar angle and azimuthal angle so that we can generate uniformly distributed samples in a spherical triangle directly. The advantage over other techniques is that the implementation is straightforward without sampling on the whole sphere or any checking stage, hence it is simple and efficient. Finally, we illustrate the usefulness of the proposed sampling method by an algorithm that computes the moments of distance from uniformly distributed customers within the shortest distance (Voronoi cells) to wireless base stations.

11:50-12:05 **A novel network reconstruction method based on a high-dimensional linear state space model with applications to time-series microbiome data §**

Yu Gu and Xing Qiu
Department of Biostatistics and Computational Biology, University of Rochester Medical Center
Yogeshwar Kelkar
Department of Biology, University of Rochester
Hulin Wu
Department of Biostatistics, University of Texas Health Science Center at Houston
E-mail: Yu_Gu@urmc.rochester.edu

Motivated by increasing interest in understanding how microbes interact with each other, recent studies of time-course human microbiome data suggest that the composition of microbiome changes over short time periods due to synergistic and antagonistic interactions among microbiome and also between microbiome and the environment. Such studies along with appropriate mathematical models provide the opportunity to uncover dynamic interaction networks within the microbiome. However, the high-dimensional nature of these data poses significant challenges to the development of such mathematical models. We propose a high-dimensional linear State Space Model (SSM) with a new Expectation-Regularization-Maximization (ERM) algorithm to construct a dynamic Microbial Interaction Network (MIN). System noise and measurement noise can be separately specified through SSMs. In order to deal with the problem of the high-dimensional parameter space in the SSMs, the proposed new ERM algorithm employs the idea of the adaptive LASSO-based variable selection method so that the sparsity property of MINs can be preserved. Simulation studies show that the proposed ERM algorithm performs well for both variable selection and parameter estimation. We applied the proposed method to identify the dynamic MIN from a time-course vaginal microbiome study of women. Our results recapitulate some known microbiome interactions such as the synergistic relationship between *Finegoldia sp.* with anaerobic bacteria such as *Sneathia sp.* and *Anarococcus sp.*, as well as some novel findings such as the positive role of *L.crispatus* in promoting the growth of facultatively anaerobic *Lachnospiraceae* species. This method is amenable to future developments, which may include interactions between microbes and the environment.

SESSION 3B SH 1053

Improving Educational Effectiveness through Data

11:20-11:40 An analysis of certification rate in edX massive open online courses (MOOC) §

Xiaoyu Wan, Warner School of Education, University of Rochester
E-mail: xwan3@u.rochester.edu

In online MOOC courses, the certified rate is not a single criterion to reflect a student's learning outcome. Nonetheless, there are plenty of indications that certification indicates learning, especially engagement and completion with course activities. The aim of this study is to examine the characteristics of the certified learners and students' engagement with the course that could contribute to the certification rate of edX massive open online learning courses. The research used chi-square tests to investigate the significance of associations between certified rate and gender, education level, year of birth, and nationality. A multiple regression analysis was employed to predict the certified rate by using factors of students' engagement with the course activities. The findings revealed that most certified learners were male but female learners had a comparatively higher completion rate. The majority of certified learners were young and middle age learners from the United States, with high educational attainment. There were significant associations between certified rate and gender, education level, year of birth, and nationality. The numbers of days of interaction with the course and interaction with chapters were two strong predictors for the certification rate.

11:45-12:05 Finding more silent motivational resources for depression symptoms and academic outcomes among adolescents §

Jungming Lee
Warner School of Education, University of Rochester
E-mail: mhlee77@gmail.com

Depression is a commonly experienced mood among adolescents even though they may not show clinically significant depressive symptoms. These depression symptoms among adolescents can adversely affect adolescents' school life, general health, and more seriously predicts their adulthood depression. In addition, occasionally they could lead to suicide. The Korean Government Report of 2012 noted that 30.5% of adolescents had experienced depression, and 18.3% of the adolescents had had a thought about committing suicide. Therefore, it is necessary to explore adolescent individuals' resources to cope with depressive tendencies and examine the mediating effect of the resources between depressive tendencies and adolescents' school related performance. The purpose of this study is twofold: one is to investigate individuals' motivational resources that have an effect on class enjoyment among Korean adolescents who reported relatively high depression symptoms; the other is to examine the mediating effects of the most predictable resources between depressive symptoms and class enjoyment among 1,055 Korean adolescents, testing for moderation by academic self-regulation motivation. As motivational resources, the present study explores academic behavioral regulation or various types of motivation identified in self-determination theory in terms of the quality of motivation, relationship with significant others (the current teacher, parents, and friend), and various types of self-schema (academic self-regulatory efficacy), self-esteem, and risk taking tendencies. Hierarchical regression analysis showed that the most effective motivational resource on class enjoyment is *identified self-regulation*, the second is *amotivation*, and the third is *academic self-regulatory efficacy*; additionally, the relationship with their teacher and the relationship with their parents emerged as important among relatively high depressed adolescents. In contrast, among relatively low depressed adolescents, *amotivation* was the most important resource, followed by *academic self-regulatory efficacy* and *intrinsic motivation*, and the relationship with a friend. In summary, the results showed that different personal resources influenced adolescents' class enjoyment in terms of adolescents' level of depression, specifically, the main effect of *the types of academic self-regulation* and *academic self-regulatory efficacy* on depression tendency and class enjoyment among adolescents. Structural equation modeling analysis revealed a partial meditation effect of self-regulatory efficacy in both low- and high- depression groups on academic self-regulation, although we failed to show measurement and structure invariance between low and high academic self-regulative motivation in multi-group analysis. The study suggests that recognition of motivational resources to manipulate adolescents' depression can inform us as to how to protect adolescents from being psychologically harmed by depression and provide ways to eliminate or modify their personal and environmental conditions.

SESSION 3C — SH 1028

Applications of Multiple Imputation

11:20-11:40 Statistical methods to handle non-detects in qPCR §

Valeriia Sherina
Department of Biostatistics and Computational Biology, University of Rochester Medical Center
E-mail: Valeriia_Sherina@urmc.rochester.edu

Quantitative real-time PCR (qPCR) is one of the most widely used methods to measure gene expression. Despite extensive research in qPCR laboratory protocols, normalization and statistical analysis, little attention has been given to qPCR non-detects – those reactions failing to produce a minimum amount of signal. Most current software replaces these non-detects with a value representing the limit of detection. Recent work suggests that this introduces uncertainties in estimation of both absolute and differential expression. Single imputation procedures have recently been developed to handle qPCR non-detects. While better than previously used methods, they underestimate the residual variance, often leading to anti-conservative inference. We propose to treat non-detects as non-random missing data, model the missing data mechanism, and use this model to impute Ct values or obtain direct estimates of relevant model parameters. To account for the uncertainty inherent in the imputation, we propose a multiple imputation procedure, which provides a set of plausible Ct values for each non-detect. In the proposed modeling framework, there are three sources of uncertainty: ambiguity of the model, the missing data mechanism, and measurement error. All three sources of variability were incorporated in the multiple imputation and direct estimation algorithms. The developed methods are implemented in the R/Bioconductor package nondetects. An extensive simulation study was performed to show the benefits of these approaches when estimating gene expression and assessing model misspecification.

11:45-12:05 Multiple imputation vs. single imputation for missing data in scoring health measures

Subrina Farah, Kevin Fiscella, and Mechelle Sanders
Department of Family Medicine, University of Rochester Medical Center
E-mail: Subrina_Farah@urmc.rochester.edu

Missing data is a common issue and validated methods are required to deal with missing data that can properly account for statistical uncertainty due to missingness. Statistical inference of intervention effects or measures of association should account for statistical uncertainty attributable to missing data. This means that methods used for imputing missing data should yield valid Type I error probabilities for hypothesis testing and confidence intervals should have nominal coverage properties. Multiple Bayesian methods and methods such as multiple imputation satisfy this condition, along with various likelihood-based and other validated methods. Single imputation methods like last observation carried forward and baseline observation carried forward are discouraged as the primary approach for handling missing data in the analysis since these methods make strong assumptions about the missingness mechanism and ignore uncertainty about the imputed values. Data have been considered from Get Ready and Empowered about Treatment (GREAT) study on persons living with HIV. For the control and intervention groups, 359 participants' longitudinal data have been considered at three time points, before, during and after intervention, and 9 different health measure scores have been collected. A handful of missing data have been observed for the entire study. For missing value analysis, a single imputation method, last observation carried forward, and multiple imputation methods including those based on the multivariate normal distribution (MVN), Multiple Bayesian and Markov Chain Monte Carlo (MCMC) procedures, have been applied. In calculating most health measures and performing hypothesis tests (paired t-tests and 2-sample t-tests), simulation studies have shown that multiple imputation assuming a MVN distribution leads to more reliable inferences than single imputations.

SESSION 3D — SH 1013 A

The Bradley-Terry Model for Paired Comparisons

11:20-11:40 Prior distributions for the Bradley-Terry model of paired comparisons

John Whelan
College of Science, Rochester Institute of Technology
E-mail: john.whelan@astro.rit.edu

The Bradley-Terry-Zermelo model has been widely used to evaluate paired comparison experiments, with applications ranging from taste tests to rating chess players. It is, among other things, the basis of the KRACH rating system used by several college hockey websites. In addition to the traditional maximum-likelihood implementation, several authors have applied a Bayesian approach where posterior probability distributions are deduced for the strengths of the objects being compared. Such an application requires a choice of prior distribution for these strengths. While most authors have considered families of prior distributions that can reflect the experimenters' additional knowledge about the objects, we are interested in the application as a rating system for sports teams, who should be judged on equal terms. Therefore we evaluate choices of prior distribution according to four simple desiderata: (1) the distribution is invariant under interchange of teams, (2) the distribution is invariant under interchange of winning and losing, (3) the distribution is normalizable, and (4) the prescription is unchanged by adding or removing teams. We find that most of the previously considered priors fail one or more of these desiderata, and consider two families which satisfy them.

11:45-12:05 Major League Baseball and the Bradley-Terry model: a Bayesian perspective §

Gabriel Phelan and John Whelan
College of Science, Rochester Institute of Technology
E-mail: gxp3900@rit.edu

A popular model to analyze objects competing in paired comparisons is the Bradley-Terry-Zermelo model, first described by Zermelo in 1929 and rediscovered in 1952 by Bradley and Terry. With its ample data and lack of ties, Major League Baseball is an ideal setting to apply such a model. A Bayesian approach is advantageous in that it captures all uncertainty over model parameters in the form of a posterior distribution. However, it also requires the specification of a prior distribution and the consideration of computational complexities that are usually not of concern in a frequentist analysis. Considering weakly-informative prior families that treat the teams equally, we use computational techniques like approximate inference and efficient Monte Carlo methods to perform inference on the strengths of Major League teams throughout history, as well as the hyperparameters of the prior families.

SESSION 3E — SH 1013 B

Design of Experiments

11:20-11:35 V_{50} experimental testing of personal protective equipment §

Darsh Thakkar
College of Science, Rochester Institute of Technology
E-mail: dt1412@rit.edu

Binary response experiments are common in epidemiology, biostatistics as well as in military applications. The Up-and-Down method, Langlie's Method, Neyer's method, K in a Row method and 3 Phase Optimal Design are methods used for sequential experimental design when there is a single continuous variable and a binary response. In this talk, we will discuss a new sequential experimental design approach called the Break Separation Method (BSM). BSM provides an algorithm for determining sequential experimental trials that will be used to find a median quantile and fit a logistic regression model using maximum likelihood estimation. BSM results in a small sample size and is designed to efficiently compute the median quantile.

11:35-11:50 Determining an economic and effective experimental design based on average variance of prediction value §

Rupansh Goantiya
College of Science, Rochester Institute of Technology

E-mail: rxg7520@rit.edu

The relationship between average prediction variance and number of runs for I-optimal designs involving 2-5 factors is investigated in this talk. It was found that the average variance of prediction value decreased with the increase in the number of runs. The inverse relationship between the number of runs and average variance of prediction clearly ranked experimental designs with a higher number of runs above the designs with a lesser number of runs in terms of the quality of the information obtained. However, designs with a higher number of runs incur cost and it may not be economical to have an experimental design with zero or approximately zero value for average variance of prediction. Exponential cost curves were fitted to determine the cost associated with the increments in number of runs. Dual vertical axes plots were then produced with the number of runs on the horizontal axis and cost and average variance of prediction on the two parallel vertical axes to identify an economic and effective experimental design.

11:50-12:05 **Evaluating different experimental designs when using robust regression §**

Pranay Kumar
College of Science, Rochester Institute of Technology
E-mail: pk6528@rit.edu

In design of experiments, after the data are collected from an experiment, the analyst generally uses analysis of variance (ANOVA) or linear regression to study the relationship between the response and factors manipulated. ANOVA and linear regression are based on the assumption that the residuals are normally distributed, but there are many situations when this normality assumption is violated. Usually the presence of outliers in the data or the distribution based on the nature of the response can cause the residuals to be non-normal. Robust regression is one of the most useful and effective ways to deal with the problem of non-normality that arises when using linear regression to fit models based on data collected. In this talk, we will focus on the use of robust regression in experimental design when the errors are not normally distributed, but instead, have a lognormal distribution. After extensive literature review we conclude that there is a lack of information on the selection of an experimental design that is the most appropriate when the fitting method is robust regression.

SESSION 3F — SH 1008

Environment and Health

11:20-11:40 **Patterns in public drinking water contamination in several US states since 2000 §**

Xupin Zhang
Warner School of Education, University of Rochester
E-mail: xzhang72@u.rochester.edu

Contamination of drinking water is an important environmental and public health concern. This problem has risen in importance in the United States in recent years due, in part, to concerns that public drinking water system infrastructure has deteriorated over the last several decades. We use drinking water sampling results obtained from several states for the years after 2000 to evaluate the extent of drinking water contamination and its trend. The dozens of routinely, albeit irregularly, sampled drinking water contaminants motivate the use of modern principal component analysis techniques that deal explicitly with missing data. We use these techniques to reduce the dimensionality of the data and identify underlying latent factors of drinking water contamination. We identify several meaningful latent factors. We find that since 2000, several of these factors have risen on average. Moreover, rural public water systems have lesser contamination than urban areas on average, and rural areas have experienced a greater rise than urban areas.

11:45-12:05 **Using machine learning for predicting obesity §**

Zhen Tan
Department of Biochemistry and Biophysics, University of Rochester Medical Center
E-mail: Zhen_Tan@urmc.rochester.edu

Having a healthy weight is an extremely significant part of overall health. Being obese contributes to numerous health conditions that limit the quality and length of life, including hypertension, stroke etc. In this study, the relationship between body mass index and meal preparation patterns is examined using the ATUS Eating & Health (EH) data. Cluster analysis from unsupervised learning is used to cluster the patterns and to make some predictions. Moreover, differences in grocery shopping patterns by income are examined by different machine learning algorithms: KNN, neural network, and decision trees. Suggestions regarding planning and sticking to a healthy diet are made.

SESSION 4 — SH COM

Keynote Lecture

1:25-2:35 **Lessons from the $1,000,000 Netflix Prize**

Robert Bell
Research and Machine Intelligence, Google
E-mail: rbellnj@gmail.com

In October 2006, the DVD rental company Netflix released more than 100 million user ratings of movies for a competition to predict new ratings based on prior ratings. The size of the data (over 17,000 movies and 480,000 users) and the nature of human-movie interactions produced many modeling challenges. One allure to data analysts around the world was a $1,000,000 prize for a team achieving a ten percent reduction in root mean squared prediction error relative to Netflix's existing algorithm. Besides producing a photo finish worthy of a movie, the 33-month competition spurred numerous advances in the science of recommender systems and machine learning, more generally. After describing some of the techniques used by the leaders, I will offer lessons and raise some questions about building massive prediction models.

SESSION 5A — SH 1004

Applications of Statistics in Environmental Studies

2:50-3:10 **Ice wedge thermal variation in east Antarctica: a time series approach**

Maria Caterina Bramati
Department of Statistical Sciences, Cornell University, and Department of Methods and Models for Economics, Territory, and Finance, Sapienza University of Rome
Rossana Raffi and Alessio Baldassarre
Department of Earth Sciences, Sapienza University of Rome
E-mail: mariacaterina.bramati@uniroma1.it

This research aims at studying the thermal variation of ice wedges at various depths. In particular, the analysis of the air, ground surface, ice-wedge top and bottom temperatures are undertaken. The active layer depth is calculated through seasons and years using hourly data at three sites in northern Victoria Land: Baker Rocks, Boomerang Glacier and Mount Jackman. The recording period is from 2004 to 2013 at Baker Rocks and Boomerang Glacier, and from 2006 to 2013 at Mount Jackman. Daily mean ground surface temperatures (DMGST) and daily mean air temperatures (DMAT) are highly correlated at Baker Rocks ($r^2 = 0.96$), at Boomerang Glacier ($r^2 = 0.95$), and at Mount Jackman ($r^2 = 0.92$). This shows that the ground surface temperature at each measurement site responds strongly to air temperature. Moreover, hourly ground surface temperature and daily mean ground surface temperature are generally lower than the air temperature in the winter season, which shows the absence of a significant snow cover. Standard deviations of the hourly temperature show that high temperature variability can exist over one month, with higher variability in winter than in summer. Frequent and large temperature fluctuations are common throughout winter with either a sharp drop or a rapid increase both in air and ground surface temperature. Variations of 25°C to 30°C were recorded over periods of one to four days. The overall variability of temperatures is decomposed using spectral analysis in order to isolate seasonal effects from cycles and long term

trends. The time series approach in the frequency domain is quite new in this field and it represents, therefore, the main contribution to the existing literature.

3:15-3:35 **A Markov chain method to examine the pattern and distribution of rainfall §**

Maruf Raheem
Department of Engineering and Mathematics, Sheffield Hallam University
E-mail: rahemarsac@yahoo.com

A three-state Markov chain was employed to examine the pattern and distribution of daily rainfall in the Uyo and Eket communities of Nigeria using 15 years (1995-2009) of rainfall data obtained from the University of Uyo meteorological centre. Chi-square and WS test statistics were used to test the goodness of fit of the Markov chain to the data. Each year was divided into three different periods: pre-monsoon (January 1-March 31), monsoon (April 1-September 30), and post-monsoon (October 1-December 31). A day was regarded as a dry day if the rainfall was not more than 2.50 mm, a wet day if the rainfall was between 2.51 mm and 5.00 mm, and a rainy day if the rainfall was above 5.00 mm. Based on the three conditions of rainfall (dry, wet, and rainy) and the statistical techniques applied, it was observed that the expected length (duration) of dry, wet and rainy days are: 10 days, 1 day, and 1 day, respectively with a weather cycle of 12 days for the pre-monsoon; 2 days, 1 day, and 2 days, respectively with a weather cycle of 5 days for the monsoon, and 6 days, I day, and 1 day, respectively with a weather cycle of 8 days for the post-monsoon in the Uyo metropolis. For Eket, the data were grouped monthly and a two-state model was applied; the probability of states of dry and rainy days were obtained at long run monthly. The findings showed that August and July had the highest level of rainfall.

SESSION 5B — SH 1028

Curriculum Development for Data Science Education

2:50-3:10 **Teaching introductory statistics in an era of big data**

Bernadette Lanciaux
College of Science, Rochester Institute of Technology
E-mail: bllsma@rit.edu

The curriculum of introductory statistics classes was developed in the 1980s but the computing technology and volume and variety of data have changed substantially since then. Our current introductory statistics class does not prepare students to do they type of data analytics necessary to gain insights from the rich data they are likely to encounter within their field. Current courses on how to analyze so-called *big data* are only available to students with significant programming experience. The technology exists to make data science accessible to many more students. Introductory statistics needs to incorporate some of the foundational concepts of data science if it is going to remain relevant. At RIT, we are reimagining our introductory statistics curriculum and developing Data Science 101 that can introduce beginning students to the field and perhaps whet their appetite enough that they may be inspired to take more courses and perhaps major in data science.

3:15-3:35 **What would a data science general education course contain?**

Kirk Anne
Computing and Information Technology, SUNY Geneseo
E-mail: kma@geneseo.edu

In this presentation, I will present possible content of a course that would be for first year students that introduces different facets of data science and how data science relates to all disciplines. After presenting possible options, I will lead the audience in a discussion of how to organize such a course and what content should be required and what could be optional.

SESSION 5C — SH 1013 A

New Methods in Multivariate and Nonlinear Regression

2:50-3:10 **H-canonical regression**

Joseph Voelkel and Wei Qian
College of Science, Rochester Institute of Technology
E-mail: joseph.voelkel@rit.edu

Multivariate texts use a certain eigenanalysis in MANOVA and multivariate-regression testing. Canonical correlation, not surprisingly, is treated as a separate topic. However, in reduced-rank regression, the most natural formulation of so-called canonical regression leads to the same solution as canonical correlation, and in fact, all three approaches yield the same eigenanalysis. So, this standard approach would to appear to be the most natural one to use in a multivariate regression problem in which we want to approximate a solution in lower dimensions. However, we illustrate through an example—a 2×3^3 experiment ($n = 54$ runs) on an engineering control system, in which the response consists of $p = 311$ temperature readings over time and is modeled as a second-order function of the factors ($q = 14$ predictors)—evidence that clearly indicates that this standard approach does not lead to a reasonable solution. We propose a better solution, called H-canonical regression, which is designed for the situation in which all p responses are measured on, and are to be compared on, the same scale. We show the value of this approach; connect it to a special case of reduced-rank regression; show its relationship to principal components analysis in a limiting case; and compare it to other competing methods. We recommend that this method be used more generally for the same-scale response problem. We also show how this method can be naturally extended to multi-scale response problems, one which includes as a special case the above-mentioned standard approach.

3:15-3:35 **Methodological challenges in nonlinear regression**

Leonid Khinkis, Adina Oprisan, and Milburn Crotzer
Department of Mathematics and Statistics, Canisius College
E-mail: khinkis@canisius.edu

Nonlinear statistical models are used in many fields including environmental science. However, they pose specific methodological challenges uncommon in linear models. These include potential computational difficulties in obtaining the least-squares estimates of the model parameters along with the validity of confidence regions, confidence intervals, and other inferences incorporating these estimates. There exist a number of methods, both local (Bates and Watts; Hamilton et al.) and global (Pazman and Pronzato) attempting to address the mentioned challenges. We offer our approach to some of these problems and will use some simple models from the statistical literature to illustrate it. The approach is based on a concept of global curvature and utilizes global properties of nonlinear regression models.

SESSION 5D — SH 1013 B

New Methods in Bioinformatics

2:50-3:10 **Two complementary methods for relative quantification of ligand binding site burial depth in proteins: the 'Cutting Plane' and 'Tangent Sphere' methods**

Vicente Reyes
Ronin Institute
E-mail: Vicente.Reyes@ronininstitute.org

We describe two complementary methods to quantify the degree of burial of ligand and/or ligand binding site (LBS) in a protein-ligand complex, namely, the 'cutting plane' (CP) and the 'tangent sphere' (TS) methods. To construct the CP and TS, two centroids are required: the protein molecular centroid (global centroid, GC) and the LBS centroid (local centroid, LC). The CP is defined as the plane passing through the LBS centroid (LC) and normal to the line passing through the LC and the protein molecular centroid (GC). The "exterior side" of the CP is the side opposite GC. The TS is defined as the sphere with center at GC and tangent to the CP at LC. The percentage of

protein atoms (a) inside the TS, and (b) on the exterior side of the CP are two complementary measures of ligand or LBS burial depth since the latter is directly proportional to (b) and inversely proportional to (a). We tested the CP and TS methods using a test set of 67 well characterized protein-ligand structures, as well as the theoretical case of an artificial protein in the form of a cubic lattice grid of points in the overall shape of a sphere and in which LBS of any depth can be specified. Results from both the CP and TS methods agree very well with data reported by Laskowski et al., and results from the theoretical case further confirm that both methods are suitable measures of ligand or LBS burial. Prior to this study, there were no such numerical measures of LBS burial available, and hence no way to directly and objectively compare LBS depths in different proteins. LBS burial depth is an important parameter as it is usually directly related to the amount of conformational change a protein undergoes upon ligand binding, and ability to quantify it could allow meaningful comparison of protein dynamics and flexibility.

3:15-3:35 **A visual-computational method for atomistic functional and evolutionary comparison of biological molecular dynamics**

Gregory Babbitt
College of Science, Rochester Institute of Technology
Jamie Mortensen, Erin Coppola, and Justin Liao
Department of Biomedical Engineering, Rochester Institute of Technology
E-mail: gabsbi@rit.edu

Traditional informatics in genomics and molecular evolution work only with static representations of sequence and structure, thereby ignoring molecular motion over time. Yet, molecular dynamics underpins most aspects that define molecular behavior in the cell. Here, we leverage advances in video graphics processors to develop comparative methods of analysis and visualization applied to molecular dynamic simulations. Our statistical method/software, DROIDS (Detecting Relative Outlier Impacts in Dynamic Simulations) is designed in conjunction with AMBER molecular dynamic simulation software (Assisted Model Building with Energy Refinement) and Chimera/VMD molecular visualization packages. DROIDS will implement massively-parallel Kolmogorov-Smirnov statistics on atomic fluctuations (i.e., B factors) to identify significant functional differences as well as evolutionary changes affecting the rapid harmonic dynamics of DNA, RNA, and protein, and their interaction. DROIDS provides a fundamental framework for addressing functional evolutionary questions surrounding biophysically-encoded information in the cell. It will represent the first molecular evolutionary inference test applied to molecular dynamics. With case examples, we demonstrate how DROIDS can quantify the functional evolution of binding forces across heterogeneous macromolecular complexes (e.g., nucleosome or transcription factor binding sites) as well as significant changes in protein conformational rigidity due to the impacts of mutation, the binding of pharmaceuticals or the binding of toxins.

SESSION 5E — SH 1053

Analysis of Internet Measurements / Theory of Influence

2:50-3:10 **Quasi-parametric dependence between new internet measurement and globalization**

Bruce Sun
Department of Mathematics, SUNY Buffalo State
E-mail: sunbq@buffalostate.edu

Under the new measurement for the internet supplied from the Cooperative Association for Internet Data Analysis (CAIDA), we relax the orthogonal and normality conditions with methodology based on empirical codependence structures and marginal distribution functions. The nonparametric dependence between diffusion of the internet and different globalization indices is examined using a panel of 10 countries for ten years, as long as error terms do not follow multivariate normal distributions. The model selections are ranked and we find the best dependence model. Also, the shock from the annual cycles can be forecasted, and we verify that the prompt reflection can be traced by the traffic fluctuations.

3:15-3:35 **Introduction to the Theory of Influence**

Mihail Barbosu

College of Science, Rochester Institute of Technology
E-mail: mxbsma@rit.edu

The starting point in the Theory of Influence is the idea of influence of a over b, $i(a,b)$, at a given moment in time, with $0 \leq i(a, b) \leq 1$; if $i(a, b) = 0$, a has no influence over b while if $i(a, b) = 1$, we say that a dominates b. In this talk we introduce the notion of influence in a finite-dimensional and an infinite-dimensional dynamical system, we present some properties of $i(a, b)$, and we discuss how this concept can be used in clustering algorithms.

§ Indicates presentations that are eligible for the student presentation awards