Efficient Use of Electronic Health Records for Translational Research
Tianxi Cai
Harvard University

While clinical trials remain a critical source for studying disease risk, progression and treatment response, they have limitations including the generalizability of the study findings to the real world and the limited ability to test broader hypotheses. In recent years, due to the increasing adoption of electronic health records (EHR) and the linkage of EHR with specimen bio-repositories, large integrated EHR datasets now exist as a new source for translational research. These datasets open new opportunities for deriving real-word, data-driven prediction models of disease risk and progression as well as unbiased investigation of shared genetic etiology of multiple phenotypes. Yet, they also bring methodological challenges. For example, obtaining validated phenotype information, such as presence of a disease condition and treatment response, is a major bottleneck in EHR research, as it requires laborious medical record review. A valuable type of EHR data is narrative free-text data. Extracting accurate yet concise information from the narrative data via natural language processing is also challenging. In this talk, I'll discuss various statistical approaches to analyzing EHR data that illustrate both opportunities and challenges. These methods will be illustrated using EHR data from Partner's Healthcare.

Predictive Models in Forensic Science
Alicia Carriquiry
Iowa State University

Learning About Learning
Ronald Christensen
University of New Mexico

This talk will use Support Vector Machines for binary data to motivate a discussion of the wider role of reproducing kernels in Statistical Learning. If time permits, comments on bagging and boosting will be included.

TBD
David Dunson
Duke University

TBD
Edward George
University of Pennsylvania

Prediction of Complex Traits in Animal Breeding: From Henderson to Bayesian Machine Learning
Daniel Gianola
University of Wisconsin

Model-Based Prediction in General and as Applied to the Outcomes of College Football Games
David Harville
Iowa State University

Success in model-based prediction depends critically on knowledge of the underlying phenomena and on being able to distinguish what is important from what is not. Typically, the data are observational in nature, giving rise to potential pitfalls of a kind that can be difficult to recognize and avoid. Performance in “repeated application” is likely to be important; at a minimum, this requires that the predictions be what Dawid referred to as well-calibrated.

For purposes of illustration, the prediction of the outcomes of college football games will be discussed. Some numerical results will be presented for the 2017 season. Among the issues encountered in this application is the “lumpiness” of the distribution of scores noted by Mosteller; this issue has been exacerbated by the introduction of overtime.

Maximizing the Usefulness of Statistical Classifiers for Two Populations

Daniel Jeske
University of California, Riverside

The usefulness of two-class statistical classifiers is limited when one or both of the conditional misclassification rates is unacceptably high. Incorporating a neutral zone region into the classifier provides a mechanism to refer ambiguous cases to follow-up where additional information might be obtained to clarify the classification decision. Through the use of the neutral zone region, the conditional misclassification rates can be controlled and the classifier becomes useful. An application to prostate cancer will be used to illustrate how neutral zone regions can extract utility from a potentially disappointing classifier that might otherwise be abandoned.

Predicting the Emotions of Tennis Players from Single-Camera Video

Stephanie Kovalchik
Victoria University and Tennis Australia

Mentality has been one of the most elusive concepts in quantitative sport performance analysis. In this talk, I present a framework for capturing emotional information for tennis players from match broadcasts. The framework yields predictions for seven emotional states relevant to sport: ‘anxiety’, ‘anger’, ‘annoyance’, ‘dejection’, ‘elation’, ‘focus’, and ‘fired up’. Two feature sets are included in the model training for each emotion in sport: the Facial Action Coding System and 17 facial action units. Multiple prediction approaches were trained and tested using these features using a labeled dataset of 1,700 facial images of professional male and female tennis players extracted from 505 match videos. We applied the prediction models to establish emotional profiles for the ‘Big 4’ (Roger Federer, Rafael Nadal, Andy Murray, and Novak Djokovic) at the 2017 Australian Open. Rafael Nadal exhibited the most ‘anxiety’ of the four players (32%, 95% CI 29 to 35%), while Roger Federer was the only player whose predominant state was ‘neutral’ (24%, 95% CI 21 to 27%). All players except for Roger Federer showed significant emotional reactions to the outcomes of points. Further, several emotional states of Rafael Nadal and Novak Djokovic were significantly predictive of their chances of winning the next point. Our predictive framework for extracting emotional data from single-camera video in professional tennis shows the feasibility of bringing the quantitative study of the inner game into sports performance analysis.

Accounting for uncertainties in predictive inference

Jing Lei
Carnegie Mellon University

When prediction is made from complex data using more advanced methods, adequately accounting for the model and sample uncertainty is necessary for reliable inference. I will introduce a new predictive inference framework based on a generic tool that converts any point estimator to an interval predictor. The resulting prediction bands have valid average coverage under essentially no assumptions, while retaining the optimality of the initial point estimator under standard assumptions. Secondly, I will discuss how to make better use of cross-validation for improved bias-variance trade-off, by adopting a hypothesis testing framework. Both methods are applicable to a very wide range of prediction algorithms.

TBD
Carl Morris
Harvard University

Getting Beyond the Mean in Predictive Inference
Hal Stern
University of California-Irvine

Predictive modeling in many domains attempts to draw inferences about the parameters characterizing performance of the set of units of interest. Examples include predictions of team strength in sports, animal breeding potential, and risk of disease incidence. Often units are ranked based on the posterior mean of the parameters or a related estimator. This can be misleading though in settings where there is considerable variation in the information available for different units. Alternative estimands can be useful in this case. Examples are provided in the settings described above.

Charles Roy Henderson: Farm Boy, Athlete, Scientist
Dale Van Vleck
University of Nebraska-Lincoln

Charles Henderson has had more impact on animal agriculture than any other person through universal use of his Mixed Model Equations to obtain predictors of genetic merit. Those equations also made possible obtaining BLUE and BLUP for many other applications. They provide the foundation for many statistical analyses packages. The presentation will include a brief biography and comments on steps along the unusual path that eventually led to his influence on statistical estimation and prediction. His early life and academic and athletic achievements in Iowa and the interim of 10 years after college to his PhD program in animal breeding and statistics at Iowa State College beginning in 1946 will be summarized. His scientific career began at age of 37 in 1948 as an associate professor of animal husbandry at Cornell University. His contributions to statistical analyses and especially to genetic prediction began immediately and include: Henderson's Methods 1, 2 and 3 for variance component estimation, his Mixed Model Equations to obtain BLUE and BLUP, and his remarkable algorithm to obtain easily the inverse of the genetic relationship matrix which is needed to be able to use information from all relatives for genetic prediction. His contributions should not be forgotten.

TBD
Martin Wainwright
University of California-Berkeley

Interaction Selection: Its Past, Present, and Future

Hao Helen Zhang
University of Arizona

In regression and classification problems, interaction selection plays an important role to improve model prediction and interpretability. The topic of interaction selection has recently drawn much attention in the literature, partially due to its wide applications in identifying important gene-gene and gene-environmental effects in biological and medical research. In this talk, we will first start with a comprehensive review on basic principles for interaction selection, classical methods, and their properties. Then we will discuss unique and enormous challenges associated with interaction selection for high dimensional data, from both computational and theoretical perspectives. A variety of penalty-based interaction selection methods for data of moderate dimensions will be reviewed. Finally, we present the recently works on scalable methods of interaction selection and screening for ultra-high dimensional settings. In particular, a new hierarchy-preserving regularization solution path algorithm will be introduced, along with some new theoretical and empirical results.

Spatial Prediction: A Graphical Overview
Dale Zimmerman
University of Iowa

One important application of predictive inference is to spatial prediction, i.e., to a setting where the observations and the predictands are indexed by their locations in a common physical space (often, but not always, Euclidean). Such a setting lends itself nicely to the graphical display (using maps over that space) of several important concepts of predictive inference. In this talk, we give an overview of spatial best linear unbiased prediction (BLUP) --- also known as kriging --- and we display several interesting concepts associated with it (the "screening effect," "perfect interpolation," and others). We highlight the impact that the spatial correlation function has on BLUP and the impact that estimating such a function has on empirical BLUP, and we consider the consequences of these for good spatial sampling design for prediction.