Presenter_Name	Title	Abstract
April Galyardt	Mixed membership models for continuous data	Mixed membership models, including grade-of-membership models, have been used previously to successfully analyze discrete data (Cooil and Varki, 2003; Erosheva, 2005; Erosheva et al, 2007; Airoldi et al, 2008). This class of models works by describing heterogeneity at the individual-level, rather than at the population-level as a latent class model would. Each individual has a latent membership vector describing the degree to which they belong to each class, and a probability distribution over the manifest variables which is a mixture of the distributions characterizing each class. However, the multinomial distribution appears to be unique, in that a convex combination of the extreme profile densities is equivalent to a convex combination of the extreme profile parameters. This observations leads to two possible generalizations of the grade-of-membership model, one based on convex combinations of densities, and another based on convex combinations of parameters. Properties of the unconditional distribution of data for each generalization are explored.
Arthur J. Kendall	A modern look at Historical Papers	The statistical tools available today allow us to readily accomplish tasks that are far beyond what was practical when Mosteller and Wallace (1964) first looked at the authorship of The Federalist papers. The approach they started operates on the assumption that individuals implicitly write in ways that their works can be distinguished. This presentation discusses exploring the Federalist Papers using additional variables, visualizations, and algorithms that were not feasible when Mosteller and Wallace did their work. Different sets of variables describing the texts are used to look at this corpus from a variety of perspectives using statistical methods that are widely available today. Time will be set aside for the audience to suggest additional ways to explore textual data.
Benjamin Shih	Unsupervised Discovery of Student Learning Tactics	Unsupervised learning algorithms can discover models of student behavior without any initial work by domain experts, but they also tend to produce complicated, uninterpretable models that may not predict student learning. We propose a variant of simple E-M clustering for hidden Markov models that can discover student learning tactics while incorporating student-level outcome data, constraining the results to interpretable models that also predict student learning. This approach is robust, domain-independent, and does not require domain experts.
Bill Shannon	Functional Data Analysis of Actigraphy/Actical Data	The goal of this research is to objectively associate patient activity-level patterns with clinical treatment response and physiological/clinical covariate characteristics of subjects. 37 patients (out of 750 to be recruited) wore actical watches continuously for one week (Monday – Friday) and their total activity level recorded every minute. These subjects were classified by PHQ-9 survey scored by a psychiatrist into not depressed (N = 25); mildly depressed (N = 7); severely depressed (N = 5). Analyzing activity data as functions and not numbers or sumamry statistics we uncover more understanding of patient subgroup behaviors (i.e., aberrant patient activity patterns in severely depressed patients). The behavior we can uncover can not be found without functional data analysis methods.
Bill Shannon	STATISTICAL Tools to Process and Analyze Human Microbiome Data	Billions of short sequences of DNA from bacteria living in and on humans will be generated by Human Microbiome Project (HMP) investigators in the next few years using next generation sequencing technology. The goal is to use theses sequences to improve the health of people by better understanding how these bacterial populations interact with their human host to impact disease processes. Data generation (i.e., the collection of these short DNA sequences) no longer limits advances in this field; rather, the bottleneck now and in the foreseeable future is in data analysis. Without statistically validated methods to speed up the analysis of this data, accurate conclusions from microbiome experiments will be hard to make. This talk will introduce the HMP and some of the existing data analysis challenges.

Presenter_Name	Title	Abstract
Dale Josephs	Effects of Feature Selection on the Classification of Spoken and Written English	Research is ongoing in multiple disciplines for an effective, efficient method to automatically distinguish spoken English from written text without relying on its associated metadata. This paper demonstrates the effect of feature selection and textual pre-processing on the results of Support Vector Machine (SVM)-based classification of text. Using a publically accessible subset of the American National Corpus that contained transcripts of spoken English as well as originally written documents, Information Gain and TF-IDF term rankings were compared for predictive confidence and accuracy across a variety of pre-processing conditions.
Dan Knights	Supervised Classification of Human Microbiota	Recent advances in DNA sequencing technology have allowed the collection of high-dimensional data from human-associated microbial communities on an unprecedented scale. A major goal of these studies is the identification of important groups of microorganisms that vary according to physiological or disease states in the host, but the incidence of rare taxa and the large numbers of taxa observed make that goal difficult to obtain using traditional approaches. Fortunately, similar problems have been addressed by the machine learning community in other fields of study like microarray analysis and text classification. In this talk we demonstrate that existing supervised classifiers can be applied to microbiota classification, both for selecting subsets of taxa that are highly discriminative of the type of community, and for building models that can accurately classify unlabeled data. To encourage the development of new approaches to supervised classification of microbiota, we discuss several structures inherent in microbial community data that may be available for exploitation in future research.
Daniel Aloise	An improved column generation algorithm for minimum sum-of-squares clustering	Given a set of entities associated with points in Euclidean space, minimum sum-of-squares clustering (MSSC) consists in partitioning this set into clusters such that the sum of squared distances from each point to the centroid of its cluster is minimized. A column generation algorithm for MSSC was given by du Merle, Hansen, Jaumard and Mladenovic in SIAM J. Sci. Comput. 21, 1485-1505, 2000. The bottleneck of that algorithm is the solution of the auxiliary problem of finding a column with negative reduced cost. We propose a new way to solve this auxiliary problem based on geometric arguments. This greatly improves the efficiency of the whole algorithm and leads to exact solution of instances 10 times larger than previously done.
David Friedenberg	Detecting Cluster Structure using Diffusion Maps	Diffusion Maps are a powerful technique for representing the connectivity of a dataset in a reduced dimensional space. They have been successfully used in applications such as data parameterization, regression, and density estimation. In this talk, we will review the Diffusion Map framework and show that Diffusion Maps can also be used to capture clustered structure in complex datasets. We introduce a method for automatically and simultaneously choosing the tuning parameters in the context of Clustering. Additionally, we introduce a Self-tuning Diffusion Map which replaces the standard global tuning parameter with a series of local tuning parameters sensitive to structure at different scales.
Di Liu	Anomaly Detection in a Large Phone Call Dataset	We present work on detecting anomalies in a large cell phone call data set with 4 million users. The scale of the data allows us to examine natural but complicated patterns in social interactions and groups. We were able to detect different types of anomalies based on summary statistics from the data set. In particular, we find many users with calling patterns similar to telemarketers, and we detect with high accuracy users who have changed phone numbers. We also discuss the computational challenges of making inference from such a large data set. Our evaluation shows that our methods are effective, and give results which are informative to the phone company.

Presenter_Name	Title	Abstract
Douglas Steinley	Finding Multiple Cluster Structures: All Variables Are Not Created Equal	Often times, when conducting general exploratory cluster analysis, all variables are included in the analysis. The authors show why an all inclusive approach may not lead to the best results in a cluster analytic setting. A variable selection procedure is presented that is able to find multiple cluster structures in the presence of variables with no cluster information. The validity of the technique is explored through extensive simulation studies. Internet usage patterns are analyzed and serve as a concrete example of the procedures in a data analytic context.
Elena Deych	Using Mantel Correlation for Comparing Sequencing Regions in Human Microbiome Project Datasets	Next generation sequencing offers a new but expensive tool for understaning how microbial populations in and on the human can influence health. The goal of the analyses presented in this talk was to answer two questions: Does sequencing of different 16S variable regions produce equivalent results? Are any of the body sites redundant in terms of bacterial populations? If variable regions and body sites produce the same estimate of the microbiome within each subject, resources can be preserved by reducing the number of variable regions and body sites that need to be sequenced.
Elizabeth Ayers	Incorporating Covariates to Classify Student Skill Knowledge Parameters	Cognitive diagnosis models have become a popular means of estimating student skill knowledge. These models use only student responses and information about which skills are required by each item, student covariate information is ignored. However, the affect of student covariates, such as gender or SES, on skill knowledge mastery is an important question in education. We propose extending the DINA model (Junker and Sijtsma, 2001) by modeling the latent skill knowledge indicators using a logistic regression to estimate the effects of covariates on skill knowledge. Covariate information improves the prediction of the latent skill knowledge indicators. When applying our methods to a subset of data from an online tutor, we have reasonable and interpretable parameter estimates.
Jia Wang	Functional Data Analysis of Actigraphy Data for Identifying Fatigue Patterns in Depressed and Non-Depressed Patients	Our objectives are to develop, validate, apply and distribute statistical algorithms and software based on functional data analysis for analyzing actigraphy data to 1) Objectively associate activity patterns with subject characteristics (e.g., depression score) 2) Produce informative visualization tools, and 3) Help researchers/clinicians incorporate actigraphy into their work. This poster presents preliminary work we have done using functional linear modeling and functional principal components analysis to analyze differences in circadian rhythms between non-depressed and seveerely depressed patients.
Justin H. Gross	Current Issues in Automated Content Analysis for Political Science	Content analysis (text analysis, in particular) represents an essential and evolving set of research methods for political scientists, bridging a gap between qualitative and quantitative studies. From political tracts and manifestos to speeches, press releases, debates, treaties, laws, and judicial decisions, the spoken and written word have long offered us insights into political actors in their natural habitat, so to speak. Computer-assisted analysis and (more recently) fully automated clustering and classification have come to supplement traditional manual coding techniques. Yet the enormous opportunities emerging through the expanse of the Internet demand additional innovation. I describe some of these new challenges faced by political scientists, note some recent contributions of political methodologists attempting to learn from large bodies of text, and identify a few developments in other applied fields and machine learning that hold promise for add ressing the needs of political scientists going forward.
Markus Breitenbach	Validating taxonomic structure using Bayes-Optimal Misclassification Probability Bounds	Validating taxonomic structure using Bayes-Optimal Misclassification Probability Bounds Markus Breitenbach and Tim Brennan This paper examines several approaches we used to evaluate the stability of cluster solutions found in large samples of prison inmates assessed on an array of criminological and criminal behavior patterns. In a prior analysis we discovered 8 clusters in a sample of about 1200 inmates. These clusters re-appeared in a more recent, independent sample of more than ten-thousand inmates. We will present and discuss a method to measuring cluster separation by a bound on the

Presenter_Name	Title	Abstract
		probability of misclassification determined by the "minimum error minimax probability machine" (Huang et.al. 2004). These findings will also be related to the results of using Bagged cluster procedures and the McIntyre-Blashfield approach to establishing cluster stability.
Mel Janowitz	A Modifications of single-linkage clustering	A common criticism of single-linkage clustering is the fact that objects are clustered using rather weak evidence. One solution to this problem is to modify the way that similatity of a pair of clusters is calculated. Another method might involve measuring the strength of any link between a pair of clusters, and just removing any sufficiently weak links. This will be concretely illustrated using a graph theoretic model and removing edges that are bridges or suitable generalizations of bridges. The material is still in a preliminary form, and appears in my book Ordinal and Relational Clustering.
Nema Dean	Empty K-Means: A Flexible Skill Set Profile Clustering Method	In cognitive diagnosis modeling, the goal is to estimate students' current skill masteries based on student responses and question design. These models do not scale well to medium or high numbers of skills. As an alternative, skill set profile clustering groups students based on skill mastery estimates using, for example, k-means. For K skills, there are 2^K possible true skill set profiles (for complete/zero mastery). However, in practice, not all possible profiles will be present. Moreover, some skill mastery estimation methods are hindered in the presence of a high proportion of multiple skill questions. We modify the k means algorithm to allow for the possibility of empty clusters; in addition, appropriate starting centers that account for the question design are derived. This empty k means method is flexible and removes the restriction on the number of skills and the question design. We show results for an online intelligent tutor.
Stanley L. Sclove	Logistic Regression: A Review and Indication of Application in Other Models	Logistic regression is reviewed in the context of generalized linear models and link functions. The logit link function is derived from discriminant analysis. Bias correction of the inverse link (the logistic function) is obtained. Other link functions for binary variables are discussed. It is indicated how logistic regression can be used in finite mixture models and hidden Markov models.
Stephen France (2nd author)	Selecting Attributes for Sentiment Classification Using Feature Relation Networks	A major concern when incorporating large sets of diverse n-gram features for sentiment classification is the presence of noisy, irrelevant, and redundant attributes. These concerns can often make it difficult to harness the augmented discriminatory potential of extended feature sets. We propose a rule-based multivariate text feature selection method called Feature Relation Network (FRN) that considers semantic information and also leverages the syntactic relationships between n-gram features. FRN is intended to efficiently enable the inclusion of extended sets of heterogeneous n-gram features for enhanced sentiment classification. Experiments were conducted on three online review test beds in comparison with methods used in prior sentiment classification research. FRN outperformed the comparison univariate, multivariate, and hybrid feature selection methods; it was able to select attributes resulting in significantly better classification accuracy irrespec tive of the feature subset sizes. Furthermore, by incorporating syntactic information about n-gram relations, FRN is able to select features in a more computationally efficient manner than many multivariate and hybrid techniques.

Presenter_Name	Title	Abstract
Susan Huse	Effect of clustering method on the estimated richness of microbial communities	With the advent of next-generation sequencing technology, microbial ecologists are uncovering new organisms faster than taxonomists can assign names. Clustering DNA sequence tags from the SSU rRNA gene amplicons is a method for creating operational taxonomic units (OTUs) that does not rely on names and is now standard in microbial ecology. The common method of generating OTUs by multiple sequence alignment and complete-linkage clustering significantly increases the number of predicted OTUs, inflating richness and diversity estimates of microbial communities. An average-linkage clustering based on pairwise alignments more accurately predicts expected OTU richness in preparations of known composition for amplicons longer than ~200nt. For shorter amplicons, such as V6 tags, a 2% single-linkage preclustering methodology smoothes the noise and minor variations in the sequencing data and further reduces the unintended inflation. These alternative clustering m ethods reduce the OTU richness in environmental samples by as much as 30–60%, but they do not reduce the fraction of OTUs in the long-tailed rank abundance curves characteristic of many microbial communities.
T. Siva Tian	Dimensionality Reduction for Classification with High-Dimensional Data	High-dimensional data refers to data with a large number of variables, often larger than the number of observations. High-dimensional data are encountered in a wide range of areas such as engineering, biometrics, psychometrics, and neuroimaging. Classifying these data is a difficult problem because the enormous number of variables poses challenges to conventional classification methods and renders many classical techniques impractical. A natural solution is to add a dimensionality reduction step before a classification technique is applied. In order to deal with multivariate data, two approaches are proposed. One is a simulated annealing (SA) based method and the other is a multivariate adaptive stochastic search (MASS) method. They both utilize stochastic search algorithms to select a handful of optimal transformation directions from a large number of random directions in each iteration. One advantage of the proposed methods is that they can accurately project the data onto very low-dimensional non-linear, as well as linear, spaces. These methods are designed to mimic variable selection type methods, such as the Lasso, or variable combination methods, such as PCA, or a method that combines the two approaches. Particularly, MASS can adaptively adjust the model complexity level, and hence performs well in situations where variable selection or variable combination methods fail. We demonstrate the strengths of SA and MASS on an extensive range of simulation and real studies by comparing them to many classical and modern classification methods. Classification problems associated with functional data are also addressed. We propose a functional adaptive classification (FAC) approach which takes the functional response into consideration and produces highly accurate and interpretable results. FAC is also based on a stochastic search procedure guided by the evaluation of model complexity. This often results in a simple relationship between functional covariates and the reduced data and makes the model interpretable. Simulation studies and an fMRI time course study are also provided to show the effectiveness of the proposed method.
Tim Brennan	Classifying Criminals: A multi-axial taxonomy of a prison inmate population	Classifying Criminals: A multi-axial taxonomy of a prison inmate population Tim Brennan and Markus Breitenbach We describe a project to identify taxonomic structure in a sample of prison inmates (N = 1072). We adopted a multi-axial classification strategy in a complex feature space of demographic, social and psychological factors and full criminal offense histories. We report results for two axes: 1) a criminal behavior axis and 2) An explanatory social psychological axis. Cluster identification used as sequence of bagged K-Means and a semi-supervised internal connectivity clustering as alternate methods. Cluster stability was examined using the McIntyre-Blashfield cross sample approach; as well as stability of results across clustering methods. We describe the criminal taxonomies emerging in these two axes and challenges related to on-going validation and cluster visualization.