1st Analytics without Borders Conference

	Wilder (AAC Building)	Lindsay 27	Lindsay 30	AAC 141
8:00-8:45	**Registration**			
8:45-9:00	**Welcome remarks**			
9:00-9:45	Keynote: Analytics Without Borders: Corporate and Academic Pathways			
9:45-10:00	**Coffee break**			
10:00-10:50	Career Session: Panel discussion	- Multi-Channel Attribution - A Case Study - Feature Selection Made Simple: The HARVEST Algorithm	- Health analytics practice in VA: two case studies on Osteoporosis and mental diseases - Application of analytics for technological forecasting in drug development	- The Application of Fatal Analysis Reporting System Data on the Road Safety Education of U.S. Minors - Predictive Modeling with Rare Events using Various Approaches
11:00-11:50	Tutorial: AText – Machine Analytics on Text	- Anomaly detection: Techniques and best practices - Learning to Love Bayesian Statistics	- Visualizing the World's Largest Timed Cycling Event and an Interactive Job Searching Tool - Determinants of Mobile Broadband Diffusion: A Focus on Developing Countries	- Penalization with Group-wise Sparsity with Applications to eBay Motors - Data Science for Algae price prediction
12:00-1:15	**Lunch**			
1:15-2:05	Tutorial: Introduction to R	- The Customer Experience Scorecard - Examples of Contributions Analysis		- How analytics have impacted the lives of the Commonwealth of Massachusetts' most vulnerable citizens - Manual Versus Automated Classification of Tweets in the Policing Domain
2:15-3:05	Tutorial: Introduction to Python	- The Future of Analytics: Human Reasoning		- Advanced Analytics in Insurance: An inside look at how Liberty Mutual Used Elastic Net to Improve Their Pricing Model - Life after Grad School: Transitioning from Academic Analytics to Corporate Analytics
3:05-3:15	Coffee break			
3:15-4:00	Keynote: Uplift Modeling using SAS Enterprise Miner			

Presentation Abstracts

Learning to Love Bayesian Statistics
Author: Allen Downey

Abstract:
Bayesian methods are well-suited for business applications because they provide concrete guidance for making decisions under uncertainty. But myths about the Bayesian approach continue to slow its adoption. In this talk I unpack these myths and explain the pros and cons of Bayesian methods compared to classical statistics.

Keywords: Bayesian statistics

Life after Grad School: Transitioning from Academic Analytics to Corporate Analytics
Author: Joseph Dery

Abstract:
Ever wonder what your career will be like after University? Or, how to best position yourself in the corporate world after being a student? Join Joe Dery, Lead Data Scientist for EMC's Business Insights & Analytics group, for a discussion of how to best make the transition from being an analytics student to an analytics professional. Joe will take you through key learnings, potential challenges, and best practices for being as successful as possible in this transition.

Keywords: Analytics, Student, Corporate, Transition

The Application of Fatal Analysis Reporting System Data on the Road Safety Education of U.S. Minors
Authors: Molly Funk , Max Karsok, Michelle Williams

Abstract:
A team of three Bryant University students is advancing to the final round of the SAS Student Symposium at the 2016 SAS Global Forum. The team utilizes various SAS tools and procedures to examine Fatal Analysis Reporting System data. The data contains information regarding fatal motor vehicle crashes in the United States and Puerto Rico from 2011 to 2013. The team profiles ten unique situations of minor involvement in motor vehicle accidents to characterize factors influencing these scenarios. From these profiles, recommendations are made to improve the education of different stakeholders in minor road safety situations.

The Customer Experience Scorecard
Author: Vic Hoffman

Abstract:
There is a tendency for retailers to focus on purchase behavior changes from a marketing perspective, directing attention towards competition and lifestyle changes as the reasons why customers alter behavior. Retailers typically try to identify at risk or defected customers, and then send offers or rewards in an effort to retain them. However, there are aspects of a customer's experience, which may be causing that customer to become dissatisfied, reduce purchases, and eventually leave. Dissatisfaction is also a key reason why a competitor has the opportunity to take a customer's business away. The "Customer Experience Scorecard" addresses retailer controllable activities, in areas such as supply chain, merchandising, pricing, and operations that contribute to a customer's shopping decision. These experiences need to be understood through analytics and addressed as part of the overall strategy to retain customers, drive increased purchasing behavior and cultivate loyalty.

Manual versus Automated Categorization of Tweets in the Public Safety Domain
Authors: Kevin Mentzer, Jennifer Xu, Vignesh Ram

Abstract:
This work examines how local police department use social media to communicate with their community. Using twitter data from 5 Massachusetts police departments, we discuss the challenges of categorizing large text-dominated datasets. We'll show that different police departments use twitter for different purposes. Finally, we'll cover our next steps in this work in-progress.

Anomaly detection: Techniques and best practices
Author: Sri Krishnamurthy

Abstract:
Anomaly detection (or Outlier analysis) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. It is used is applications such as intrusion detection, fraud detection, fault detection and monitoring processes in various domains including energy, healthcare and finance. In this talk, we will introduce anomaly detection and discuss the various analytical and machine learning techniques used in in this field. Through a case study, we will discuss how anomaly detection techniques could be applied to energy data sets. We will also demonstrate, using R, an application to help reinforce concepts in anomaly detection and best practices in analyzing and reviewing results.

Keywords: Anomaly detection, Machine learning, R, Energy

Multi-channel attribution - a case study
Author: John (Zhuang) Li

Abstract:
Given the proliferation of digital media, today's companies promote their products and services through multiple channels. As a result, customers often were exposed to multiple ads from various channels prior to making their purchase decisions. The number one question for the marketers is: What is my ROI for the spend across different channels? This study seeks to quantify the partial value of each marketing contact that contributed to conversion and the resulting revenues by incorporating the touch history and channel-specific time-decay algorithms. The channels include the traditional marketing vehicles such as print and email, as well as typical digital channels including display, paid search and natural search. ROI then is derived by allocated revenues on the one hand, and the associated cost on the other hand. Based on ROI, marketers can adjust the budget planning and allocation to maximize ROI going forward.

Keywords: multichannel attribution, digital marketing, budget optimization

Health analytics practice in VA: two case studies on Osteoporosis and mental diseases
Author: Mingfei Li

Abstract:
As the largest health care systems in the U.S., Veteran Health Administration provides comprehensive cares to more than 8.3 million veterans each year. In this presentation, the VHA care system and database will be introduced briefly. Two collaborative research projects will be introduced as an example of health analytics practice in VA: osteoporosis for males, and medication adoption in mental disease. Both the opportunities and challenges in data analytics will be discussed.

Keywords: health analytics, government data analytics

Data Science for Algae price prediction
Authors: Nilam Shete, Akshay Prakash, Michael J Walsh, Bentley University

Abstract:
The Center for Integration of Science & Industry at Bentley under the leadership of Dr. Michael J Walsh have come up with a pricing model based on nutritional data gathered from various sources. This model is then used to predict the price for a synthetically created algae feed. Technologies used: R, Python. Techniques: Linear Mixed Effects Regression, Machine Learning, Web Scraping, Statistical Modeling.

Keywords: Data Mining, Statistical Modeling, Machine Learning, Web Scraping

Predictive Modeling with Rare Events using Various Approaches
Authors: Alan Olinsky and Phyllis Schumacher, Bryant University

Abstract:
There are many techniques for predictive modeling of a bivariate target variable that can be utilized with large data sets. However, when the target variable represents a rare event, with an occurrence in the data set of approximately 10% or less, then traditional modeling techniques might not be appropriate. There will be a discussion of some different methods that have been developed to improve the prediction outcomes of such rare events.

Keywords: predictive modeling, rare events, big data

Examples of Contributions Analysis
Author: Robert B. Smith, Social Structural Research

Abstract:
This paper develops a paradigm for the analysis of contingency tables that is rooted in Bayes's theorem. The analysis quantifies the effects of the variables' categories as differences in percentage points and, after multiplication by the categories' frequency probabilities, their contributions. The examples focus on Barack Obama's margin of victory over John McCain in the 2008 election. These contributions clarify the overall impacts of the different categories on the vote better than do the differences in percentage points. After close inspection of the survey data, to test conjectures it also applies regression models that are weighted by the survey sample; boxes and a depiction elucidate aspects of these procedures.

Keywords: public opinion research, contributions to margins of victory

Visualizing the worlds largest timed cycling event and an interactive job searching tool
Author: Daniel Stasin

Abstract:
The Cape Cycle Tour is an annual cycling event in South Africa, drawing almost 30 000 athletes on race day. Visualizing the finish times of cyclists allows for a context driven representation of how athletes performed. The second visualization is intended as an interactive job searching tool for international students studying in the United States.

Keywords: Tableau, Data, Visualization

Penalization with Group-wise Sparsity with Applications to eBay Motors
Authors: Qing (Wendy) Wang, Dan Zhao

Abstract:
Penalization methods are commonly used in statistical data analyses to accomplish variable selection and parameter regularization. When the predictor variables can be grouped together, group-wise sparsity is desired. The Group Lasso (Yuan and Lin, 2007) and Sparse Group Lasso (Simon et al., 2013) are proposed to account for group-wise regularization. Satisfactory results of these methods have been shown in bioinformatics and genetics studies when the size of the feature space p is much larger than the sample size n. We wonder whether these methods have similar success in econometrics when p<n, but p and n are both relatively large. We consider several penalization methods, including the Lasso, the Ridge, the Group Lasso, and the Sparse Group Lasso, with application to eBay Motors auction data. We present numerical comparisons of these methods in both ordinary linear regression and generalized linear regression. We show that when the covariates can be categorized into groups, the Sparse Group Lasso method outperforms the others by achieving a much more parsimonious model.

Keywords: eBay Motors, group-wise sparsity, Group Lasso, penalization, Sparse Group Lasso

Feature Selection Made Simple: The HARVEST Algorithm

Author: Herbert I.Weisberg, Victor P. Pontes

Abstract:
Feature selection with high-dimensional data and a very small proportion of relevant features poses a severe challenge to standard statistical methods. We have developed a new patent-pending approach (HARVEST) that is relatively straightforward to apply, and in many circumstances can virtually guarantee useful results. The basic idea is to evaluate each feature in the context of many random subsets of other features. HARVEST exploits the fact that a relevant feature can add real predictive value to some of these subsets. In contrast, an irrelevant feature will add no real value and can only appear to be associated with the outcome by chance. Based on this insight, we have derived a method for testing the relevance of features that includes a rigorous statistical test of significance. Impressive results produced so far by our HARVEST algorithm indicate that it can be effective in predictive analytics, both in science (e.g., genomics) and business (e.g., marketing).

Keywords: feature selection, random subsets, HARVEST

Determinants of Mobile Broadband Diffusion: A Focus on Developing Countries
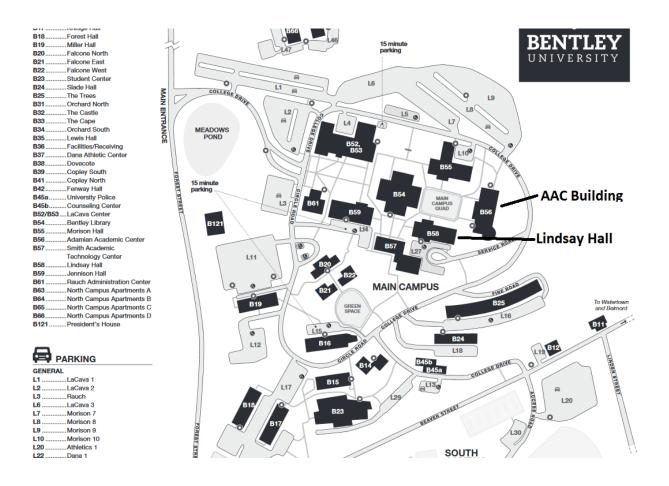
Authors: David J. Yates, Girish J. "Jeff" Gulati, Marco Marabelli

Abstract:
Past research on the broadband digital divide indicates a widening divide in which developing countries are falling further behind countries in the developed world. In response to this problem, the World Bank has advocated a "mobile first" strategy for developing countries. Unfortunately, there is little understanding of what determines mobile broadband adoption or diffusion in developing countries. In this paper, we begin to address this problem by exploring to what extent policy, regulation, government, and governance affect mobile broadband diffusion in the

developing world. Our results show that when controlling for distribution and level of income, there is greater mobile diffusion in developing (i.e., non-OECD) countries that encourage competition in their telecommunication industries and practice sound governance in their public sector. Although governance is an important determinant of mobile broadband diffusion, we find no evidence that political structure (i.e., the level of democracy) matters. We also find that regulation of telecommunications licensing is associated with decreased access to mobile broadband. Further examination of our data suggest that national governments have either modernized and streamlined this regulatory measure or are performing important functions related to mobile services (e.g., spectrum allocation) without the need to regulate licenses for telecommunication service providers. We discuss these important results in light of prior literature and suggest new avenues of research that stem from our findings.

- Note: Wilder Pavilion is in Adamian (AAC Building)