3-23-2018

3rd Analytics Without Borders Conference Program

Bentley University

3rd Analytics without Borders Conference

3rd Analytics without Borders Conference Program

Location: Wilder Pavilion, AAC 165
Bentley University, 175 Forest Street, Waltham, MA

Date: March 23rd, 2018

8:00: Registration check-in at Wilder Pavilion

Coffee and breakfast available

8:45: Welcoming remarks at Wilder Pavilion by Heikki Topi, Dominique Haughton, Bentley University

9:00-9:50: **Keynote 1 at Wilder Pavilion**

Data Science, A.I., Business Analytics, Machine Learning, Operations Research, and More – What Are They? How Are They Used in Practice?

by Victor S.Y. Lo, Head of Data Science and Artificial Intelligence, Workplace Solutions, Fidelity Investments

Introducer: Dominique Haughton, Bentley University

Parallel sessions Track 1 at Wilder Pavilion (Chair: Greg Vaughan)

10:00-10:50: Track 1 Parallel session 1

Analytic Storytelling: From Numbers to Narratives by P.H. Benjamin Lohnes, *Epsilon*

11:00-11:50: Track 1 Parallel session 2

Climate Value of Offshore Wind Energy by Zana Cranmer, Erin Baker, *Bentley University*

Using data analytics to quantify NIH contribution to new drug approvals by Ekaterina Galkina Cleary, *Bentley University*

Parallel sessions Track 2 at Lindsay 26A (Chair: Maria Skaletsky)

10:00-10:50: Parallel session 1

AI – Should we be worried? by Richard Glass, *Bryant University*

11:00- 11:50: Parallel session 2

Applying Behavioral Economics to the Business of Higher Education by Roland B. Stark, *Integrative Statistics, Maynard, MA, USA*

Parameter Optimization for Constant Proportion Portfolio Insurance (CPPI) Investment Strategy by Olga Biedova, *Bentley University*

Parallel sessions Track 3 at AAC141 (Chair: Alan Olinsky)

11:00- 11:50: Parallel session 1

Pass if He Can't Pass: The Analytics of NBA Player Success by Joel Wald, *Bentley University*

New Approaches to the Study of Shirking in Major League Baseball by Richard J. Paulsen, *Bentley University*

Tutorial technical workshop at JEN 309

10:00-11:50: Tutorials session: R and Python, By David Oury, *Bentley University*

12:00-1:00: Simple Lunch

1:00-1:50: **Keynote 2 at Wilder Pavilion**

Issues in Classification by Eric Breck, Senior Software Engineer at Google

Introducer: Nathan Carter, Bentley University

Parallel sessions Track 1 at Wilder Pavilion (Chair: Kevin Mentzer)

2:00-2:50: Track 1 parallel session 3

It Might be Legal, but is it Ethical? Data Mining Ethics by Rachael Juskuv, *Bryant University*

3:00-3:50: Track 1 parallel session 4

Under the Covers of the Ethereum Blockchain by Kevin Mentzer, *Bryant University*

General Introduction to Blockchain Technology by Ryan Harris, *Bryant University*

Parallel sessions Track 2 at Lindsay 26A (Chair: Richard Glass)

2:00-2:50: Track 2 Parallel session 3

Financial Statement Auditors use of Data Analytics by Jared Koreff, *University of Central Florida*

Efficient Big Data Model Selection with Applications to Fraud Detection by Greg Vaughan, *Bentley University*

3:00-3:50: Track 2 parallel session 4

Sentiment Analysis on Apple using Twitter Data by Rai & Jujjavarapu, *UMass-Dartmouth*

Parallel sessions Track 3 at AAC141 (Chair: David Oury)

2:00-2:50: Track 3 Parallel session 2

Employee or Contractor: Courts' Misclassification Decisions from Analytics Lens by Feizollahi, *Georgia State University*

Evaluating consistency of Boston government residential house assessment using machine learning Jinting Zhang, *Bentley University*

3:00-3:50: Track 3 parallel session 3

One-day ahead stock price prediction using machine learning models by Rai & Jujjavarapu, *UMass-Dartmouth*

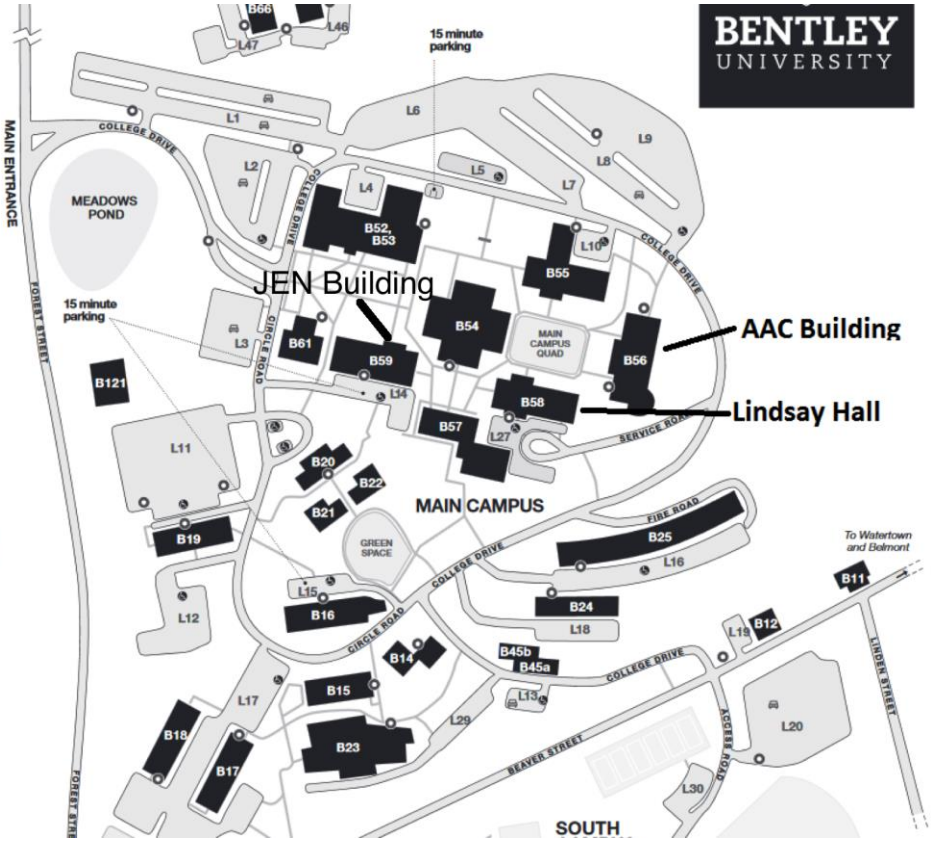
Clustering with Genetic Algorithms in Division III Collegiate Wrestling by Nathan Carter, *Bentley University*

B18 ..... Forest Hall
B19 ..... Miller Hall
B20 ..... Falcone North
B21 ..... Falcone East
B22 ..... Falcone West
B23 ..... Student Center
B24 ..... Slade Hall
B25 ..... The Trees
B31 ..... Orchard North
B32 ..... The Castle
B33 ..... The Cape
B34 ..... Orchard South
B35 ..... Lewis Hall
B36 ..... Facilities/Receiving
B37 ..... Dana Athletic Center
B38 ..... Dovecote
B39 ..... Copley South
B41 ..... Copley North
B42 ..... Fenway Hall
B45a ..... University Police
B45b ..... Counseling Center
B52/B53 .... LaCava Center
B54 ..... Bentley Library
B55 ..... Morison Hall
B56 ..... Adamian Academic Center
B57 ..... Smith Academic Technology Center
B58 ..... Lindsay Hall
B59 ..... Jennison Hall
B61 ..... Rauch Administration Center
B63 ..... North Campus Apartments A
B64 ..... North Campus Apartments B
B65 ..... North Campus Apartments C
B66 ..... North Campus Apartments D
B121 ..... President's House

PARKING

GENERAL

L1 ..... LaCava 1
L2 ..... LaCava 2
L3 ..... Rauch
L6 ..... LaCava 3
L7 ..... Morison 7
L8 ..... Morison 8
L9 ..... Morison 9
L10 ..... Morison 10
L20 ..... Athletics 1
L22 ..... Dana 1



Morning Tracks	Track 1 (Wilder Pavilion) Chair: Greg Vaughan	Track 2 (LIN 26A) Chair: Maria Skaletsky	Track 3 (AAC141) Chair: Alan Olinsky	Workshop (JEN309)
Session 1 10-10:50	Analytic Storytelling: From Numbers to Narratives (Lohnes)	AI – Should we be worried? (Glass)		R and Python workshop
Session 2 11-11:50	Climate Value of Offshore Wind Energy (Cranmer , Baker)	Applying Behavioral Economics to the Business of Higher Education (Stark)	Pass if He Can't Pass: The Analytics of NBA Player Success (Wald, Li)	
	Using data analytics to quantify NIH contribution to new drug approvals (Cleary et al)	Parameter Optimization for Constant Proportion Portfolio Insurance (CPPI) Investment Strategy (Biedova)	New Approaches to the Study of Shirking in Major League Baseball (Paulsen)	
Lunch break				
Afternoon Tracks	*Track 1 (Wilder Pavilion) Chair: Kevin Menzer*	*Track 2 (LIN 26A) Chair: Richard Glass*	*Track 3 (AAC141) Chair: David Oury*	
Session 3 2-2:50	It Might be Legal, but is it Ethical? Data Mining Ethics (Juskuv)	Financial Statement Auditors use of Data Analytics (Koreff)	Employee or Contractor: Courts' Misclassification Decisions from Analytics Lens (Feizollahi et al)	
		Efficient Big Data Model Selection with Applications to Fraud Detection (Vaughan)	Evaluating consistency of Boston government residential house assessment using machine learning (Zhang)	
Session 4 3-3:50	Under the Covers of the Ethereum Blockchain (Mentzer, Gaugh)	Sentiment Analysis on Apple using Twitter Data (Rai , Jujjavarapu)	One-day ahead stock price prediction using machine learning models (Rai & Jujjavarapu)	
	General Introduction to Blockchain Technology (Harris, Castillo)		Clustering with Genetic Algorithms in Division III Collegiate Wrestling (N. Carter, A. Iyengar, M. Lanham, S. Netsler, D. Schrader, A. Zadeh)	

Presentation Abstracts

Morning Keynote Presentation

Data Science, A.I., Business Analytics, Machine Learning, Operations Research, and More – What Are They? How Are They Used in Practice?

Speaker: Victor S.Y. Lo, Head of Data Science and Artificial Intelligence, Workplace Solutions, Fidelity Investments

Abstract:

This keynote will discuss the definitions and relationships of many seemingly related (or seemingly different) fields: Data Science, A.I., Business Analytics, Machine Learning, Operations Research, and more. Additionally, we will introduce applications of these techniques in business practice as well as career prospects of these fields.

Speaker Biography:

Victor S.Y. Lo is a seasoned Big Data, Marketing, Risk, and Finance leader and innovator with over two decades of extensive consulting and corporate experience employing data-driven solutions in a wide variety of business areas, including Customer Relationship Management, Market Research, Advertising Strategy, Risk Management, Financial Econometrics, Insurance, Product Development, Transportation, and Human Resources. He is actively engaged with Big Data Analytics, causal inference, and is a pioneer of Uplift/True-lift modeling, a key subfield of data science.

Victor has managed teams of quantitative analysts in multiple organizations. He is currently Head of Data Science and Artificial Intelligence in Workplace Solutions at Fidelity Investments. Previously he managed advanced analytics teams in Personal Investing, Corporate Treasury, Managerial Finance, and Healthcare and Total Well-being at Fidelity Investments. Prior to Fidelity, he was VP and Manager of Modeling and Analysis at FleetBoston Financial (now Bank of America), and Senior Associate at Mercer Management Consulting (now Oliver Wyman).

For academic services, Victor has been a visiting research fellow and corporate executive-in-residence at Bentley University. He has also been serving on the steering committee of the Boston Chapter of the Institute for Operations Research and the Management Sciences (INFORMS) and on the editorial board for two academic journals. Victor earned a master's degree in Operational Research and a PhD in Statistics, and was a Postdoctoral Fellow in Management Science. He has co-authored a graduate level econometrics book and published numerous articles in Data Mining, Marketing, Statistics, and Management Science literature.

Afternoon Keynote Presentation:

Issues in Classification

Speaker: Eric Breck, Senior Software Engineer at Google

Abstract:

As machine learning methods become increasingly critical in today's world, it is becoming ever more important to ensure that models and systems that rely on machine learning do not perpetuate unwanted biases that may be latent in various data sources. In this talk, we'll examine potential sources of unwanted bias, dive into some real world examples, and discuss possible mitigation strategies. Above all, we'll argue for the importance of awareness of the possible issues that can arise and the value of questioning assumptions during model development.

One-day ahead stock price prediction using machine learning models
Author(s) Bharatendra Rai and Ratna Jujjavarapu

Abstract
Stock price prediction is an interesting and challenging field of research. We use thirty input features derived from historical data to study performance of four popular predictive models viz., extreme gradient boosting (XGB), random forest (RF), linear discriminant analysis (LDA), support vector machine (SVM). This study involves stock price data from three well-known companies viz., Apple, Bank of America and Ford. We assesses prediction model performance using three different threshold levels for classifying one-day ahead stock price as up, down or neutral. Prediction model performances are compared using three metrics viz., area under curve, sensitivity and accuracy.

Keywords Machine learning, Prediction models, Model performance

Sentiment Analysis on Apple using Twitter Data
Author(s) Bharatendra Rai, Ratna Sirisha Jujjavarapu

Abstract
Nowadays, social media has become a platform for people to express their feelings or opinions on various topics. And most of these post of are mostly informal, unstructured, brief. Based on this situation, this research attempts to use tweets on Apple to review users opinion on Apple by using sentiment Analysis.

Efficient Big Data Model Selection with Applications to Fraud Detection
Author(s) Gregory Vaughan

Abstract
As the volume and complexity of data continues to grow, more attention is being focused on solving so-called big data problems. One field where this focus is pertinent is credit card fraud detection. Model selection approaches can identify key predictors for preventing fraud. Stagewise Selection is a classic model selection technique that has experienced a revitalized interest due to its computational simplicity and flexibility. Over a sequence of simple learning steps, stagewise techniques build a sequence of candidate models that is less greedy than the stepwise approach.

We introduce a new stochastic stagewise technique that integrates a sub-sampling approach into the stagewise framework, yielding a simple tool for model selection when working with big data. Simulation studies demonstrate the proposed technique offers a reasonable trade-off between computational cost and predictive performance. We apply the proposed approach to synthetic credit card fraud data to demonstrate the technique's application.
Keywords: big data; stagewise estimation; sub-sampling; fraud detection; clustered data

Using data analytics to quantify NIH contribution to new drug approvals
Author(s) Ekaterina Galkina Cleary, Jennifer M. Beierlein, Navleen Surjit Khanuja, Laura M. McNamee, Fred D. Ledley

Abstract
This study quantifies the extent of government investment into research on new drug discovery and development. Using relational databases, we examined National Institutes of Health (NIH) funding related to the 210 novel drugs approved by the FDA since 2010. The analysis identified >2 million research publications directly involving these drugs or describing basic research related the drugs' targets. We linked these publications to >200,000 fiscal years of NIH funding supporting this research and a total investment of >$100 billion. We estimate that each "first-in-class" drug coming to market is enabled by as much as $800 million in NIH-funded research, highlighting the scale of the public sector's contribution to therapeutic discovery. This work has implications for evidence-based policy making, including the risk that reduced research funding would delay the emergence of new drugs. Analytic challenges surrounding data quality and infrastructure will be discussed.

Keywords: basic science, NIH funding, drug development, translational science, data analytics

Financial Statement Auditors use of Data Analytics
Author(s) Jared Koreff

Abstract
Several accounting bodies, including the American Institute of Certified Public Accountants (AICPA), CPA Canada, International Auditing and Assurance Standards Board (IAASB) and Public Company Accounting Oversight Board (PCAOB) have all expressed interest in incorporating more data analytics into the financial statement audit process. This presentation will provide an overview of the use of data analytics by financial statement auditors. While data analytics offer great promise in identifying audit-relevant information, identifying such information does not guarantee that it will be effectively incorporated into an auditor's decision making process. This session will also present the results of an experiment examining the impact of data analytical models (anomaly vs. predictive) analyzing different types of data (financial vs. non-financial) on auditors' judgments.

Keywords: Financial statement auditors, CPAs, predictive modeling, nonfinancial data

Evaluating consistency of Boston government residential house assessment using machine learning
Author(s) Jinting Zhang

Abstract
Governments use the market approach to assess residential house price, which is comparing the property to similar properties that have recently sold. But how well the government did with the market approach and whether such approach can provide a consistent assessment on similar residential houses? We use 2017 Boston government's residential house assessment database and apply machine learning including clustering and classification algorithm to test whether the similar residential house has similar assessed price and explore what kinds of house are more likely to be overestimated or underestimated by the government.

Pass if He Can't Pass: The Analytics of NBA Player Success
Author(s) Joel Wald and Mingfei Li

Abstract
Each year, NBA teams must make large investments in highly unknown players during the NBA draft. Our task was to make their decisions more effective and simpler via data analysis and analytical models. After defining NBA Player success, we predicting it using a combination of college statistics and NBA draft combine measurements as inputs for logistic regression. While many factors play a role in player success, one common theme across many different models was the statistical significance of college assists per game. As a result, we stand by our presentation title: Pass if he can't pass!

Keywords Sports analytics, NBA draft, player career success

Under the Covers of the Ethereum Blockchain
Author(s) Kevin Mentzer, Michael Gough

Abstract
In this talk we will discuss the structure of the Ethereum blockchain and demonstrate how to query the blockchain to better understand the distributed ledger concept. We will use the CryptoKitties application to show attendees how to extract data related to a single application.

Keywords Blockchain, Ethereum, Cryptocurrencies

Employee or Contractor: Courts' Misclassification Decisions from Analytics Lens
Author(s) Mohammad Javad Feizollahi, Charlotte Alexander

Abstract
We investigate federal district court decisions in worker misclassification cases from 2008 to 2016. Independent contractors lack the benefits and protections available to employees. They may file suit, seeking "employee" classification, but the law poorly defines the two categories. Moreover, judges' classification opinions have never been studied systematically. We use text mining and analytical tools to identify the related misclassification lawsuits, extract key features, and understand how courts distinguish between employees and independent contractors and the factors that influence their decisions.

Clustering with Genetic Algorithms in Division III Collegiate Wrestling

Authors: N. Carter, A. Iyengar, M. Lanham, S. Netsler, D. Schrader, A. Zadeh.

Abstract: Coaches of NCAA Division III Collegiate Wrestling agreed that the system for dividing schools into regions unfairly disadvantaged some wrestlers in national competition. A team of faculty from several universities worked together on the problem and used an algorithm to cluster the schools into regions to seek minimal travel times and the most equal distribution of region sizes and power. The work is based on genetic algorithms and produces results demonstrably better than those the NCAA has historically used.

Keywords: sports analytics, cluster analysis, genetic algorithms

Parameter Optimization for Constant Proportion Portfolio Insurance (CPPI) Investment Strategy
Author(s) Olga Biedova

Abstract
n the current capital market environment, investors constantly face the challenge of finding the successful and stable investment mechanisms. Highly volatile equity markets and extremely low bond returns bring about the demand for sophisticated yet reliable risk management strategies. Investors are looking for risk management solutions to efficiently protect their investments. One of the investment strategies that is widely used by practitioners is the Constant Proportion Portfolio Insurance (CPPI) strategy. The CPPI strategy has been introduced and has been extensively studied in the literature. In most of the studies, the restrictive assumptions of continuous portfolio rebalancing, as well as the Black-Scholes type dynamics of the underlying risky asset, are made. We consider the CPPI strategy under the more relevant assumptions of discrete portfolio rebalancing. We also rely on more realistic market scenarios using real market data for our numerical simulations. We present an extended comparative quantitative analysis of the performance of the CPPI investment strategy in various market scenarios and within a range of input parameter values. We present an approach to numerical optimization of the parameters of the CPPI strategy based on bootstrap simulations and with respect to investor relevant optimization criteria.

Analytic Storytelling: From Numbers to Narratives
Author(s) P.H. Benjamin Lohnes

Abstract
The power of analytics may be unquestioned, but the ability to turn raw data into real insight requires more than an understanding of modeling and statistical concepts. An analyst who transforms data into critical information transitions from being a mere "number cruncher" to a true data storyteller. In this talk we will cover the elements of storytelling, some of the possible story types used to tell analytic stories, and how to develop a framework in crafting your own data stories. This talk is aimed at current students, new grads, or the experienced data professional looking to make more of an impact.
Keywords: storytelling, analytics, narrative, communication

It Might be Legal, but is it Ethical? Data Mining Ethics
Author(s) Rachael Juskuv

Abstract
Information is readily available and easy to obtain in data mining. Within our academic and professional circles, we find that full accessibility can be useful, but also lead to privacy and ethical dilemmas that we may have not prepared for. This presentation will discuss ways to ensure ethical consideration is part of data mining accountability.

Keywords: data mining, ethics, values, data decisions

New Approaches to the Study of Shirking in Major League Baseball
Author(s) Richard J. Paulsen

Abstract
The impact of contract length on player performance has received a good deal of attention in the sports economics literature. Players are said to shirk following the signing of a guaranteed contract if performance falls below some level of expected performance. While much of the literature uses OLS to identify shirking behavior by players under multi-year contracts, this study employs fixed effects regressions and two-stage least squares regressions to analyze player performance from 2009 to 2017. Using a fixed effects regression, marginal evidence of shirking is found with Wins Above Replacement as the outcome variable. Employing three different measures of marginal revenue product, OLS regressions are used and find evidence of shirking, where shirking is measured as the difference between expected and realized marginal revenue product. Two-stage least squares is then employed to address potential endogeneity in the marginal revenue product specifications. Using 2SLS, economically larger shirking effects are found. Last, the impact of contract length on shirking behavior is found to vary based on whether or not a player has joined a new team, and by player age.

Keywords Contract Length, Shirking, Major League Baseball

Applying Behavioral Economics to the Business of Higher Education
Author(s) Roland B. Stark

Abstract
Colleges and universities are keenly interested in the factors that motivate students or their parents with respect to key decisions: whether to apply, to enroll if accepted, to stay in school, or to contribute as alumni. Abundant evidence shows that traditional means of identifying these factors tend to produce misleading results. "Derived importance" approaches that draw on principles of behavioral economics fare much better. We demonstrate a variety of such methods and recount a series of applied situations in which they have played an important role in illuminating the reasons for key choices. Techniques discussed include group differences on objective vs. subjective measures; correlation; regression; vignette research including conjoint analysis; and market basket research.

Keywords behavioral economics, survey research, market research, enrollment management, causality

General Introduction to Blockchain Technology
Author(s) Ryan Harris, Lydia Castillo

Abstract
Introduce the general idea of Blockchain Technoloy (also known as Distributed Ledger Technology) as a concept. Discuss what is needed to know to understand how the technology works and provide examples of its usage.

Keywords Blockchain, Basics, General

The Climate Value of Offshore Wind Energy

Author(s) Zana Cranmer, Erin Baker

Abstract

What is the value of permitting offshore wind farms? Can it help us mitigate climate change? We present a method for estimating the prospective value of permitting offshore wind farms, by comparing climate damages and abatement costs in cases with and without offshore wind energy. The value depends on the cost of offshore wind technology, the climate policy under consideration, the severity of damages from climate change, and the discount rate. This value can be balanced against the local environmental impacts of offshore wind in order to inform a sustainable regulatory process for siting offshore wind farms. In the absence of a binding climate policy, the present value of permitting offshore wind farms ranges from $283 billion to $2.9 trillion under central assumptions about damages and discount rate. The implied value of technological change in offshore wind energy ranges from $347 billion, when damages are low and discount rates high, to $33 trillion, when damages are high and discount rates low.