# Virtual Student Research Symposium in Statistics and Data Science

**November 7, 2020**

*Hosted by the Boston Chapter of the American Statistical Association, in collaboration with the student chapters of Bentley, Boston University, and Tufts*

## SCHEDULE

10:00 AM – Introduction and Keynote

**Francesca Dominici**, Harvard University
*A Particulate Solution: Data Science in the Fight to Stop Air Pollution and Climate Change*

11:00 AM – Session 1: Business Analytics and Data Science

**Emily Siff**, Session Chair, Program in Biological and Biomedical Sciences, Yale University
11:00 AM **Brianna Johnson**, Program in Data Science, UMass Dartmouth
11:15 AM **Sai Srikanth**, Program in Data Science, Northeastern University
11:30 AM **Harshita Srinivas**, Program in Data Science, UMass Dartmouth
11:45 AM **Alexander Chau**, Program in Data Analytics, Bentley University
12:00 PM **Anuja Das**, Program in Data Analytics, Bentley University
12:15 PM **Dwyer Deighan**, Program in Data Science, UMass Dartmouth

12:30 PM – Career Panel

**Kelly Kung,** Host, Department of Mathematics and Statistics, Boston University
**Matthew Austin**, Partner, AJO Partners
**Kristin Baltrusaitis**, Research Associate, Center for Biostatistics in AIDS Research, Harvard T.H. Chan School of Public Health
**Ina Jazic**, Senior Biostatistician, Vertex
**Ryan Rosario**, Data Scientist, Google
**Jeffrey Swartzel**, Scientist, Data & Modeling Sciences, Procter & Gamble

1:15 PM – Lunch with drop-in chats

Room 1: Applying to graduate programs in (bio)statistics and data science
**Scott Field**, Assistant Professor, Program in Data Science, UMass Dartmouth
**Judith Lok**, Associate Professor, Department of Mathematics and Statistics, Boston University
**Mihaela Predescu**, Associate Professor, Program in Business Analytics, Bentley University

Room 2: Careers in pharma and biotech
**Ina Jazic**, Senior Biostatistician, Vertex

Room 3: Careers in data science
**Ryan Rosario**, Data Scientist, Google
**Jeffrey Swartzel**, Scientist, Data & Modeling Sciences, Procter & Gamble

Room 4: Careers in finance and business analytics
**Matthew Austin**, Partner, AJO Partners

Room 5: Careers in clinical research
**Kristin Baltrusaitis**, Research Associate, Center for Biostatistics in AIDS Research, Harvard T.H. Chan School of Public Health

1:45 PM – Session 2: Biostatistics

**Wenying Deng**, Session Chair, Department of Biostatistics, Harvard University
1:45 PM **Qingyan Xiang**, Department of Biostatistics, Boston University
2:00 PM **Emily Siff**, Program in Biological and Biomedical Sciences, Yale University
2:15 PM **Sarah Conner**, Department of Biostatistics, Boston University
2:30 PM **Gauri Kamat,** Department of Biostatistics, Brown University
2:45 PM **Tyler Erickson**, Department of Biostatistics, Duke University

3:00 PM – Break

3:15 PM – Session 3: Statistics

**Qingyan Xiang**, Department of Biostatistics, Boston University
3:15 PM **Wenying Deng**, Department of Biostatistics, Harvard University
3:30 PM **Ante Bing**, Department of Mathematics and Statistics, Boston University
3:45 PM **Justin Zhu**, Department of Statistics, Harvard University
4:00 PM **Wenrui Li**, Department of Mathematics and Statistics, Boston University
4:15 PM **Kelly Kung,** Department of Mathematics and Statistics, Boston University

4:30 PM –Happy hour with drop-in chats

Room 1: Social room for (bio)statistics

Room 2: Social room for data science

## KEYNOTE

**Francesca Dominici**, Clarence James Gamble Professor of Biostatistics, Population and Data Science, Co-Director of Data Science Initiative, Harvard University

*A Particulate Solution: Data Science in the Fight to Stop Air Pollution and Climate Change*

What if I told you I had evidence of a serious threat to American national security – a terrorist attack in which a jumbo jet will be hijacked and crashed every 12 days. Thousands will continue to die unless we act now. This is the question before us today – but the threat doesn't come from terrorists. The threat comes from climate change and air pollution.

We have developed an artificial neural network model that uses on-the-ground air-monitoring data and satellite-based measurements to estimate daily pollution levels across the continental U.S., breaking the country up into 1-square-kilometer zones. We have paired that information with health data contained in Medicare claims records from the last 12 years, and for 97% of the population ages 65 or older. We have developed statistical methods and computational efficient algorithms for the analysis over 460 million health records.

Our research shows that short and long term exposure to air pollution is killing thousands of senior citizens each year. This data science platform is telling us that federal limits on the nation's most widespread air pollutants are not stringent enough.

This type of data is the sign of a new era for the role of data science in public health, and also for the associated methodological challenges. For example, with enormous amounts of data, the threat of unmeasured confounding bias is amplified, and causality is even harder to assess with observational studies. These and other challenges will be discussed.

## ABSTRACTS

**Ante Bing**, Department of Mathematics and Statistics, Boston University

*Learn-As-you-GO (LAGO): an innovative adaptive design for multi-component trials*

Different types of adaptive designs have been implemented in clinical trials in recent years. These contemporary designs allow for prospectively planned modifications and for example provide researchers the possibility to end the study intermittently, while preserving the integrity and validity of the study. Some examples of adaptive designs include adaptive randomization design, group sequential design, and dropping treatment arms. Multi-stage studies with a multi-component intervention package are usually separated into a pilot study, which aims to find the optimal intervention package, followed by a randomized trial to evaluate the efficacy of the previously found intervention package. Our new "Learn-As-you-GO" (LAGO) design is an innovative adaptive design that allows for the intervention package in later stages to depend on the results collected during previous stages, while all outcomes are used to estimate efficacy after the trial's conclusion. More specifically, LAGO studies consist of multiple stages. After each stage, the data collected so-far are analyzed, and a revised intervention package is calculated based on these data. This new intervention package is then rolled out in the next stage of the study.

LAGO for binary outcomes [1] has been developed and was illustrated with the BetterBirth study, which aimed to improve maternal and child health around birth in India. This work did not include center effects, although center covariates such as center size were included. In my presentation, I will introduce LAGO, briefly examine LAGO with binary outcomes, and introduce LAGO designs with continuous outcomes and LAGO designs with center effects. I will present detailed simulation results. The simulation results indicate that the estimators for treatment effects resulting from a LAGO study, with continuous outcomes and in the presence of center effects, are likely consistent and asymptotically normal. I will also discuss some recent challenges and future work.

Reference:
Nevo, D. Lok, J. Spiegelman, D. (2019, Jan 23). Analysis of "Learn-As-You-Go" (LAGO) Studies. Accepted by the Annals of Statistics. (https://arxiv.org/abs/1808.06310)

**Alexander Chau**, Program in Data Analytics, Bentley University (alexchau21@gmail.com)

*Efficiency Changes – An Analytical Look at the NBA's Mentality Shift*

The National Basketball Association (NBA) has always been about gaining a competitive edge over opponents. With the increased awareness and emphasis in analytics, the NBA's shift toward data-driven efficiency has changed everything from team composition to shot selection. In this presentation, NBA data from the past 10 seasons are evaluated to identify where and how efficiency changes have taken place to see analytics in action.

My research has identified significant changes in how players decide where to shoot the ball based on the differences in the expected points per shot. Determined by average shooting percentages, the most efficient shots to take in the modern NBA are close 2-point shots (0-3 feet) and 3-point shots (22-23.75 feet+). Given that the NBA has been pushing toward analytics with the rise of "stats geeks" like Daryl Morey of the Houston Rockets implementing their own data-driven systems, it is no surprise the league has evolved to accept these changes over the past 10 years.

From the data analyzed, there is proof the mid-range 2-point shot has been heavily phased out of the arsenal of players in the NBA as the demand for 3-point sharpshooters has increased over the past few years. Even

for players that are used in the "Center" position (which has notoriously represented the worst shooters in the league), the modern day Center in the NBA is increasingly becoming a player that can also down 3-point shots. As the 3-point shot increases in popularity, the rise in efficiency follows. This data signifies the NBA is constantly changing and adapting to new tactics as competition only become more cutthroat.

**Sarah Conner**, Department of Biostatistics, Boston University

*Estimation and modeling of the restricted mean time lost in the presence of competing risks*

Survival data with competing or semi-competing risks are common in observational studies. As an alternative to cause-specific and subdistribution hazard ratios, we describe the between-group difference in cause-specific restricted mean times lost (RMTL). This measure of association gives the mean difference in life expectancy lost to a specific cause of death or in disease-free time lost, in the case of a non-fatal outcome, over a pre-specified period. To adjust for covariates, we introduce an inverse probability weighted estimator and its variance for the marginal difference in RMTL. We also introduce an inverse probability of censoring weighted regression model for the RMTL. In simulation studies, we examined the finite sample performance of the proposed methods under proportional and non-proportional subdistribution hazards scenarios. We illustrated both methods with competing and semi-competing risks data from the Framingham Heart Study. We estimated sex differences in atrial fibrillation-free times lost over 40 years. We also estimated sex differences in mean lifetime lost to cardiovascular disease and non-cardiovascular disease death over 10 years among individuals with atrial fibrillation.

**Anuja Das**, Program in Data Analytics, Bentley University (das_anuj@bentley.edu)

*Change detection in self-exciting point processes through scrambled trends and its applications in terrorism modelling*

Terrorism has been a consistent global security threat for centuries. The increasingly complicated relationship between the multiplying terror factions worldwide almost always leads to attacks and consequent retaliations that seem to be turning more nuanced and becoming steadily more violent. While it is a phenomenon that seemingly lies dormant for a period of time and strikes when you least expect it, the threat never completely subsides. But are these ebbs and flows of activity that have the ability to bring entire countries to their knees truly unpredictable?

Our research develops a new class of statistics within the field of change point detection. By changing the flow of time, we identify changes in underlying intensities for point processes. More specifically, we target self-exciting events like terrorism that exhibit many layers of intensity variations owing to the strong likelihood of retaliatory behavior. While there is extensive research that delves into change point analysis of point processes, self-exciting patterns remain notoriously challenging to tame. Our statistics demonstrate improvements over their competitors and are able to identify potential change points within both the immigrant and offspring regimes without paying a hefty price in terms of model complexity. Additionally, similarities among terror categories are explored through estimated change point proximities, calculated through the Hausdorff metric. Bootstrapped interval estimates are also offered.

**Dwyer Deighan**, Program in Data Science, University of Massachusetts Dartmouth

*Genetic-algorithm-optimized neural networks for gravitational wave classification*

Gravitational-wave detection strategies are based on a signal analysis technique known as matched filtering. Matched filtering is known to be optimal under certain conditions, yet in practice, these conditions are only approximately satisfied while the algorithm is computationally expensive. Despite the success of matched

filtering for signal detection, due to these limitations, there has been recent interest in developing deep convolutional neural networks (CNNs) for signal detection. Designing these networks remains a challenge as most procedures adopt a trial and error strategy to set the hyperparameter values. We propose and develop a new method for hyperparameter optimization based on genetic algorithms (GAs). We compare six different GA variants and explore different choices for the GA-optimized fitness score. We show that the GA can discover high-quality architectures when the initial hyperparameter seed values are far from a good solution as well as refining already good networks. For example, when starting from the architecture proposed by George and Huerta, the network optimized over the 20-dimensional hyperparameter space has 78% fewer trainable parameters while obtaining an 11% increase in accuracy for our test problem. Using genetic algorithm optimization to refine an existing network should be especially useful if the problem context (e.g. statistical properties of the noise, signal model, etc) changes and one needs to rebuild a network. In all of our experiments, we find the GA discovers significantly less complicated networks as compared to the seed network, suggesting it can be used to prune wasteful network structures. While we have restricted our attention to CNN classifiers, GA hyperparameter optimization can be applied within other machine learning settings, including alternative architectures for signal classification, parameter inference, or other tasks.

**Wenying Deng**, Department of Biostatistics, Harvard University

*Scalable Variable Selection with Theoretical Guarantee using Bayesian Neural Networks*

This work shows that variational Bayesian neural network (BNN) is an effective tool for high-dimensional variable selection with rigorous uncertainty quantification. Given the nice known theoretical properties of Gaussian process, we first derive the asymptotic distribution of proposed variance importance and use random Fourier features to approximate Gaussian kernel. We show that a properly configured variational BNN (1) learns the variable importance effectively in high dimensions. (2) conducts variable selection in a computationally efficient way. Extensive simulation confirms that the proposed algorithm outperforms existing classic and neural-network-based variable selection methods, particularly in high dimensions.

**Tyler Erickson**, Department of Biostatistics, Duke University

*Analysis of the Discrepancy between Bronchoscopists and Pathologists on the Adequacy of Bronchoscopy Biopsy Samples*

The purpose of this study was to help bronchoscopists understand the variability in the number of alveoli tissue samples collected from lung transplant patients during a bronchoscopy to ultimately improve a patient's likelihood of receiving a diagnosis from a pathologist. There can be a discrepancy between the number of samples collected by the bronchoscopist and the number of usable samples as determined by the pathologist. A minimum of 4 usable samples (as called by the pathologist) is necessary or the patient's bronchoscopy will be marked as ungradable and no diagnostic conclusions drawn about the status and presence of rejection of the transplanted lung(s). In pursuing this research goal, we needed to describe the prevalence and magnitude of the discrepancy of sample-calls between bronchoscopists and pathologists. We continued the research by then identifying patient and procedure characteristics associated with the occurrence of an inadequate number of alveoli samples. The analysis cohort consists of 142 bronchoscopy procedures performed between May 2017 and May 2018 at Duke University Medical Center in 128 lung transplant recipients. Univariable and multivariable logistic regression was conducted. Forward selection with an alpha-to-enter criterion of 0.05 was used to identify a subset of the characteristics to be included in the multivariable model. We found that there could be a potentially large difference in the number of samples called by bronchoscopists and pathologists, indicating the prevalence and need for further study. Upon final analysis, it was shown that a patient's sex, frequency of previous biopsies, use of general

anesthesia, number of attempts to collect a sample and presence of clear lung fluid were all correlated with an inadequate number of alveoli samples for pathologists. Bronchoscopists could use this information to potentially change sampling procedures, by safely increasing number of samples collected, for patients who have similar characteristics.

**Brianna Johnson**, Program in Data Science, University of Massachusetts Dartmouth

*Variation of Medicare Costs for Intracranial Hemorrhages and Cerebral Infarctions Across the United States*

Currently in the United States, one stroke occurs every 40 seconds, with most victims being over the age of 65. These individuals are primarily on Medicare, a government funded healthcare system that spends over half of a trillion dollars per year. Although treatment for stroke patients has become standard, there is still a $44,000 disparity between treatment costs, making it essential to determine possible factors and put an end to hospital monopolies. This paper uses mathematical modeling and spatial data techniques to highlight the largest geographic disparities nationwide, as well as look into past allegations of the reasons for variance. Research conducted in this paper will provide reasoning for unnecessary hospital bills through data visualization, as well as outline possible solutions to end the drastic increase for out of pocket Medicare costs nationwide.

**Gauri Kamat,** Department of Biostatistics, Brown University

*Leveraging Random Assignment to Impute Missing Covariates in Causal Studies*

Baseline covariates in randomized experiments are often used in the estimation of treatment effects, for example, when estimating treatment effects within covariate-defined subgroups. In practice, however, covariate values may be missing for some data subjects. To handle missing values, analysts can use imputation methods to create completed datasets, from which they can estimate treatment effects. Common imputation methods include mean imputation, single imputation via regression, and multiple imputation. For each of these methods, we investigate the benefits of leveraging randomized treatment assignment in the imputation routines, that is, making use of the fact that the true covariate distributions are the same across treatment arms. We do so using simulation studies that compare the quality of inferences when we respect or disregard the randomization. We consider this question for imputation routines implemented using covariates only, and imputation routines implemented using the outcome variable. In either case, accounting for randomization offers only small gains in accuracy for our simulation scenarios. Our results also shed light on the performances of these different procedures for imputing missing covariates in randomized experiments when one seeks to estimate heterogeneous treatment effects.

**Kelly Kung**, Department of Mathematics and Statistics, Boston University

*The Causal Effect of Drug-Induced Homicide Prosecutions Reported in Media on Drug Overdose Deaths*

With the on-going overdose crisis and increased risk of drug overdoses from the COVID-19 pandemic in the United States, governments have passed many policies in hopes of decreasing drug overdose death rates. Examples of such policies include drug-induced homicide (DIH) laws and prosecutions, whereby those distributing drugs to overdose victims are charged with the deaths. Such prosecutions are presented as an overdose prevention measure, but their impact has not been empirically assessed. We estimated the impact of having a DIH prosecution reported by the media between 2000 - 2017. Using drug overdose death data from the CDC, we estimated how the risk of unintentional drug overdose deaths in 50 states depends on the absence/presence of prosecutions reported by the media. To control for other relevant policy interventions, we used a difference-indifference-like model, but on a risk ratio scale (since we expected effects to be larger

in states with a greater drug overdose death problem). Time effects were smoothed and could vary by U.S. region. We estimated that having any DIH prosecutions reported by the media was associated with a 7.8% increase (risk ratio of 1.078, 95% CI: (1.066, 1.091)) in unintentional drug overdose deaths. Further analysis suggested that in the states analyzed, there was a total of approximately 32,674 (95% CI: (27,843, 37,449)) deaths attributable to DIH prosecutions reported in the media in the 50 states from 2000 - 2017. The analysis suggests that DIH prosecutions may actually aggravate the crisis they are intended to solve.

**Wenrui Li**, Department of Mathematics and Statistics, Boston University

*Estimation of the Epidemic Branching Factor in Noisy Contact Networks*

Many fundamental concepts in network-based epidemic modeling depend on the branching factor, which captures a sense of dispersion in the network connectivity and quantifies the rate of spreading across the network. Moreover, contact network information generally is available only up to some level of error. We study the propagation of such errors to the estimation of the branching factor. Specifically, we characterize the impact of network noise on the bias and variance of the observed branching factor for arbitrary true networks, with examples in sparse, dense, homogeneous and inhomogeneous networks. In addition, we propose a method-of-moments estimator for the true branching factor. We illustrate the practical performance of our estimator through simulation studies and with contact networks observed in British secondary schools and a French hospital. This is joint work with Eric D. Kolaczyk and Daniel L. Sussman.

**Emily Siff**, Program in Biological and Biomedical Sciences, Yale University

*COVID-19 Survival: What are the Driving Forces? An Unsupervised Machine Learning Approach*

As of November 2020, the COVID-19 pandemic remains a public health emergency. Although many clinical and biological aspects of COVID-19 have been investigated, researchers have not yet statistically pinpointed which factors are the most crucial determinants of mortality. Isolating these factors would immensely help in prioritizing resources and guiding treatment options. Utilizing data from Rhode Island Lifespan hospitals, our investigation of 300 adults with PCR-confirmed COVID-19 focuses on two critical issues: (1) which factors most strongly affect COVID-19 mortality and (2) which factors conjunctively determine COVID-19 outcomes. To estimate the driving factors, we will use cosine similarity multidimensional scaling (Principle Coordinates Analysis) and then group participants into categories using shared factors (quintiles), such as COVID-19 severity (mild, moderate, severe, fatal). To estimate conjunctive factors, we will carry out lasso regression, a regularization technique, and then apply Latent Dirichlet Allocation (LDA). Altogether, this study will advance COVID-19 research by using innovative, rigorous statistics to discern the driving forces behind COVID-19 outcomes.

**Sai Srikanth**, Program in Data Science, Northeastern University

*Co-registration of 3D Mass Spectrometry Images with 2D Histological Images*

Mass Spectrometry Imaging (MSI) is widely used in proteomics, metabolic, drug discovery, and detection of tumors using statistical and Machine Learning (ML) methods. Optical (microscopic) images of the tissues hold valuable information regarding the morphology of the tissue. Integrating these image modalities can argument the ML methods and enable more reliable and reproducible methods. Several preparation steps that are involved in the acquisition of the MS image induce local elastic deformations that prevent direct overlap of the MS and the optical images. In order to integrate the images, the MS images need to be registered to the optical images prior to performing any downstream analyses. While methods for registering multi-modal images exist, they are not typically available to experimentalists outside the lab that developed them. We propose a novel workflow that integrates the process of registration and

downstream analyses (e.g. classification). The registration is carried out using deep learning based convolutional neural networks that capture the transformation within the layer weights.

**Harshita Srinivas**, Program in Data Science, University of Massachusetts Dartmouth

*Gan-based image super-resolution for high fidelity natural images*

Recent progress in the field of image synthesis using the generative adversarial approach is very promising. However, extending this towards the reconstruction of high-fidelity images from their lower resolution images pose new challenges while upscaling them with higher upscale factors, simultaneously retaining the fine texture as of original high-resolution images. In this line, the below approach of combining the two most successful models into a new model for super-resolution image synthesis would be thriving in achieving a high quality of diverse images. Our new model, an amalgamation of SRGAN with BigGAN by retaining their respective perspectives towards the generation of super-resolution images has made this possible. Combining BigGAN's Generator and Discriminator architecture after slight modifications by restructuring class-conditional generative models and conditional batch normalization into unconditional generator and batch normalization along with the SRGAN's pipeline retaining the content loss, perceptual loss, and adversarial loss in the training process has shown good image synthesis. Here, DIV2K, Diverse 2k 1000 images, collection of diverse images which includes flora, fauna, objects, humans, architectures, sceneries have been used. This dataset has been down sampled (x4) using bicubic downscaling and has been used as our Low-Resolution input images and high definition of the same dataset as High-Resolution images for Super-Resolution image synthesis. The objective behind the usage of this dataset is mainly their diversity. We attempt to compare the SSIM, PSNR, Inception Score, FID of this new model with the state-of-the-art models.

**Qingyan Xiang**, Department of Biostatistics, Boston University

*From truncation by death to the survival-adjusted median: a summary measure in the presence of death*

Many clinical studies focus on a clinical outcome such as a neurocognitive score, also in situations where patients may die. Unmeasured clinical outcomes due to death are often called "truncation by death". We will argue that treating outcomes as "missing" or "censored" due to death could be misleading for the treatment effect evaluation. It is possible that the median (neurocognitive) score in the survivors and median (neurocognitive) score in the always-survivors present a trade-off between good neurocognitive outcomes and the probability of survival: the treatment group with good median scores has a worse probability of survival, even in situations where both the probability of survival and the probability of any good (neurocognitive) score are better in the treatment arm. In this talk, we advocate to not always see death as a mechanism through which health outcomes are missing, but rather as part of the outcome measure. To account for the survival status when addressing the causal effects of treatments on the (neurocognitive) score, we propose the survival-adjusted median as an alternative summary measure of health outcomes in the presence of death. The survival-adjusted median is the outcome threshold such that 50% of the population is alive with an outcome above that threshold. We will show that the survival-adjusted median provides a convincing summary measure for treatment effect evaluation.

**Justin Zhu**, Department of Statistics, Harvard University

*Modeling Time-Varying Treatment Effects with Zero-Inflated Data*

Gaussian Process (GP) models have gained popularity for its flexibility to handle correlation among data sampled from distributions in the exponential family. The correlation frequently characterizes time-dependent data, such as step count data across different time horizons. In this project, we plan to analyze

the effectiveness of Gaussian Process on zero-inflated Poisson (ZIP) step count data. To do so, we evaluate the performance of Gaussian Process in fitting a series of generative models approaching the ZIP distribution. Our approach is to create a switched likelihood function in our Gaussian Process with the switched parameter a proxy of our pi variable denoting the probability of zero-inflation. A flexible GP should be able to successfully model any mixture of generative distributions from the exponential family and zero-inflated data. By being able to model the underlying generative distribution, we can better identify the time-varying treatment effect of mobile health interventions for step count data.