Clarkson University

David A. Walsh ‘67 Arts & Sciences
Mini-Conference

**Friday August 31, 2018**

# DATA SCIENCE MEETS HEALTHCARE, BIOLOGY AND ENVIRONMENTAL SCIENCES

| | | | |
|---|---|---|---|
| | | ***Breakfast and Registration*** | 08.00 - 09.00 |
| Welcome | Chuck Thorpe | Introductory remarks | 09.00- 09.15 |
| Workshop | Sumona Mondal | Data mining and statistical decision-making | 09.15 - 10.45 |
| | | ***Break*** | 10.45 - 11.00 |
| Snell 241 | Runa Bhaumik | Extracting meaningful information from text | 11.00 - 12.30 |
| | | ***Lunch*** | 12.30 - 01.30 |
| Keynote Address | Bimal Sinha | Big data - statistical issues | 01.30 - 02.00 |
| Short Talks:<br>Big Data Theory | Joseph Skufca | Machine learning, medical diagnosis, and biomedical engineering research: a challenging move from silica to clinic | 02.00 - 02.25 |
| | Marko Budišić | Topology and operator theory for non-mathematicians | 02.25 - 02.50 |
| Snell 213 | Supraja Gurajala | Can we build accurate spatio-temporal event models with social media data? | 02.50 - 03.15 |
| | | ***Coffee Break*** | 03.15 - 03.45 |
| Short Talks:<br>Big Data<br>Applications | Suresh Dhaniyala | Prediction of air quality from photographs | 03.45 - 04.10 |
| | Susan Bailey | Using patterns of molecular evolution to identify sites under selection across genes and genomes | 04.10 - 04.35 |
| Snell 213 | Tom Langen | Citizen science reveals negative effects of roads and road traffic on amphibians across spatial scales and regions in the eastern United States | 04.35 - 05.00 |

Clarkson
UNIVERSITY
defy convention™

For more information about the workshop please contact Sumona Mondal (smondal@clarkson.edu), Shantanu Sur (ssur@clarkson.edu), or Devin Kapper (dkapper@clarkson.edu)
Breakfast (8.00am-9.00am) and lunch (12.30pm-1.30pm) will be provided to workshop/conference attendees.

This workshop is funded by the David A Walsh ‘67 Arts & Sciences Mini-Conference Grant.

# Clarkson University

# David A. Walsh ‘67 Arts & Sciences Mini-Conference

**Friday August 31, 2018**

## WORKSHOP TALKS

ASSOCIATE PROFESSOR

DEPARTMENT OF MATHEMATICS

CLARKSON UNIVERSITY

### DATA MINING AND STATISTICAL DECISION MAKINGS 9.15 AM - 10.45 AM

**SUMONA MONDAL, Ph.D.**

Data mining techniques are at the forefront and of high importance in the research community due to the difficulty of analyzing and interpreting high volumes of big data. In this workshop, we present a discussion and hands-on experience on how to use techniques which are popular as data mining tools. We will also demonstrate how to conduct statistical decision makings to make the most meaningful inferences of the results.

Sumona Mondal is an Associate Professor of Statistics in the Mathematics Department at Clarkson University. Professor Mondal’s research interests are Multivariate Theory, Statistical Inference, Design of Experiments, and Applications of Data Mining Techniques on biological and engineering data. Mondal is an active member in the scientific community, publishing over 25 peer-reviewed articles regularly in statistics, mathematics, engineering, bio-medical journals and giving more than 40 talks at conferences across the globe.

RESEARCH ASSISTANT PROFESSOR

DEPARTMENT OF PSYCHIATRY

UNIVERSITY of ILLINOIS at CHICAGO

### EXTRACTING MEANINGFUL INFORMATION FROM TEXT 11.00 AM - 12.30 PM

**RUNA BHAUMIK, Ph.D.**

Text mining has gained popularity due to the enormous amount of text data, which are produced in a variety of forms such as patient records, healthcare insurance data, news outlets, social networks, etc. International Data Corporation (IDC) predicts that the data volume will grow to 40 zettabytes by 2020, leading to a 50-time growth from the beginning of 2010. Text data, being an unstructured form, cannot be simply processed and perceived by computers. Therefore, efficient and effective techniques and algorithms are required to discover useful patterns.

Dr. Bhaumik is a research assistant professor at the Biostatistical Research Center in the Department of Psychiatry at the University of Illinois at Chicago. Her research focuses on Longitudinal Data Analysis, Multivariate Statistical analysis, Graph Theory, and applications of Machine Learning Algorithms to Neuroscience and other fields.

Clarkson University

David A. Walsh ‘67 Arts & Sciences Mini-Conference

**Friday August 31, 2018**

# Big data - statistical issues

**1.30 PM - 2.00 PM**

With the advent of an enormous amount of data on many aspects of our daily lives, prompted by modern web-based technology (texting, tweeting, messaging, Facebook, etc.), it is indeed a great challenge for the data analysts to efficiently extract and analyze meaningful and relevant information. In this talk, I will briefly discuss some pertinent statistical issues.

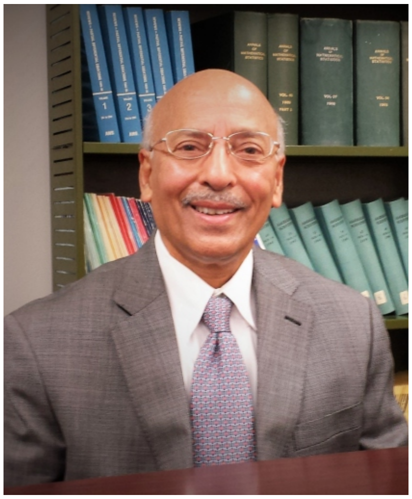
About the Speaker:
University of Maryland Presidential Research Professor
University System of Maryland Board of Regents Professor
Fellow, Institute of Mathematical Statistics
Fellow, American Statistical Association
Research Mathematical Statistician, US Census Bureau

Professor Sinha is the Founder of the Statistics Graduate Program at UMBC. A 1973 Ph.D. in statistics from the University of Calcutta, Professor Sinha is an ex-faculty of the Indian Statistical Institute and the University of Pittsburgh. A Professor of Statistics at UMBC since 1985, Professor Sinha’s research activities span topics in theoretical and applied statistics, including multivariate analysis, linear models, ranked set sampling, environmental statistics, statistical meta-analysis, and data analysis under confidentiality protection. He has coedited several volumes, and coauthored four books (John Wiley, Springer, Academic). He is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics, and an elected member of the International Statistical Institute. His research has been funded by the US Environmental Protection Agency for about twenty years. Professor Sinha’s research contribution in the area of environmental statistics has been recognized through a Distinguished Achievement Award from the Environmental Statistics Section of the American Statistical Association. In acknowledgment of his research productivity, Professor Sinha was named a Presidential Research Professor in 2008. Furthermore, he received the University System of Maryland Board of Regents Excellence in Research award in 2012. Professor Sinha has served on the editorial board of several national and international statistics journals, and mentored over 30 Ph.D. students.

**BIMAL SINHA, Ph.D.**
**PRESIDENTIAL RESEARCH PROFESSOR**

**DEPARTMENT OF MATHEMATICS AND STATISTICS**

**UNIVERSITY OF MARYLAND BALTIMORE COUNTY**



The Arts & Sciences Seminar Series is a weekly colloquium series that has been supported by the School of Arts & Sciences Advisory Council at Clarkson University especially through generous gifts from David A. Walsh ‘67.

Please contact ansseminar@clarkson.edu

SA&S 300: Arts and Sciences Seminar is a one credit course intended to foster an interdisciplinary outlook in undergraduates majoring in the School of Arts and Sciences.

Clarkson UNIVERSITY defy convention™

Clarkson University

David A. Walsh ‘67 Arts & Sciences Mini-Conference

Friday August 31, 2018

# BIG DATA THEORY

PROFESSOR/CHAIR

DEPARTMENT OF MATHEMATICS

CLARKSON UNIVERSITY

## Machine learning, medical diagnosis, and biomedical engineering research: a challenging move from silica to clinic

2.00 PM - 2.25 PM

### JOSEPH SKUFCA, Ph.D.

A large number of papers are appearing in the biomedical engineering literature that describe the use of machine learning techniques to develop classifiers for detection or diagnosis of disease. However, the usefulness of this approach in developing clinically validated diagnostic techniques so far has been limited and the methods are prone to overfitting and other problems which may not be immediately apparent to the investigators. This commentary is intended to help sensitize investigators as well as readers and reviewers of papers to some potential pitfalls in the development of classifiers, and suggests steps that researchers can take to help avoid these problems. Building classifiers should be viewed not simply as an add-on statistical analysis, but as part and parcel of the experimental process. Validation of classifiers for diagnostic applications should be considered as part of a much larger process of establishing the clinical validity of the diagnostic technique.

Joe Skufca graduated from the US Naval Academy and served 20 years as a submarine officer. He retired from the Navy in 2005, completed his PhD in Applied Mathematics, and began as a faculty member in the Math Department at Clarkson University. His work stretches broadly across applied mathematics and dynamical systems, with focus on applied modeling. He now serves as Chair of the Department of Mathematics and is one of the founding co-directors of the Clarkson's Interdisciplinary Master's Program in Data Analytics.

Clarkson University

David A. Walsh '67 Arts & Sciences Mini-Conference

**Friday August 31, 2018**

# BIG DATA THEORY

**ASSISTANT PROFESSOR**

**DEPARTMENT OF MATHEMATICS**

**CLARKSON UNIVERSITY**

## Topology and operator theory for Non-Mathematicians

**2.25 PM - 2.50 PM**

**MARKO BUDISIC, Ph.D.**

I will present two modern mathematical algorithms, persistent homology, and Koopman mode decomposition, that have made a significant splash in their respective fields. Although the original papers describing these techniques can seem abstruse to non-mathematicians, both algorithms extend familiar tools in, resp., network structure analysis, and frequency analysis. The talk will give an easy entry-point into these techniques and illustrate the potential that these tools bring to the table in biological and health-related settings.

Marko Budisic is an Assistant Professor of Mathematics at Clarkson, with background in applied dynamical systems and control theory. His research interests are in applying techniques from nonlinear dynamics and applied topology to data arising in fluid mechanics, engineering, biology, and other fields. Starting in Spring'19, he will be teaching a PICMath class at Clarkson where students will act as mathematical consultants to solve real problems for industrial partners.

Clarkson University

David A. Walsh ‘67 Arts & Sciences Mini-Conference

**Friday August 31, 2018**

# BIG DATA THEORY

## Can we build accurate spatio-temporal event models with social media data?

**2.50 PM - 3.15 PM**

**SUPRAJA GURAJALA, ABD**

**ASSISTANT PROFESSOR**

**DEPARTMENT OF COMPUTER SCIENCE**

**SUNY POTSDAM**

Big Data from online social networks (OSNs) provides a rich possibility for quantified analysis in such diverse topics as social behavior and social sensing of physical events. Models based on OSN data have been used for monitoring natural events like earthquakes and predicting human-driven events such as elections. As such models begin to be used for increasingly sensitive issues like public health, crime prediction, etc, it is important to understand the spatio-temporal accuracy of these models. Here, using global air quality events as sample data, we build robust OSN models and evaluate their ability to predict events over space and time. We collected a ~ 25 million tweet data set for this analysis and built optimized bag-of-words (BoW) based regression models for accurate spatio-temporal event prediction. We then evaluated the performance of these models for use outside of their training data set. The models were tested to determine their accuracy as a function of relative lengths and distances between the times and geographical locations of the training and test data sets.

Supraja Gurajala is a faculty member of the Computer Science Department in SUNY Potsdam.  She joined the department in Jan 2017. She is currently also a PhD student in Clarkson University's computer science department, working under the guidance of Prof. Jeanna Matthews. Her research interests are in the fields of social media data analytics, computer networks, and security.

# Clarkson University

# David A. Walsh ‘67 Arts & Sciences Mini-Conference

**Friday August 31, 2018**

# BIG DATA APPLICATIONS

BAYARD D. CLARKSON DISTINGUISHED PROFESSOR

DEPARTMENT OF MECHANICAL & AERONAUTICAL ENGINEERING

CLARKSON UNIVERSITY

## Prediction of air quality from photographs

**3.45 PM - 4.10 PM**

**SURESH DHANIYALA, Ph.D.**

Understanding the effects of aerosol particles on human health has been a challenge for several decades due to limited availability of techniques to measure a population's particulate exposure. Existing air quality monitoring networks are limited in number and require extensive manpower and equipment to operate. This research takes a physics-based approach in estimating an important air quality parameter called $PM_{2.5}$ from analysis of visual change of a scenery with time. Combining theoretical aerosol scattering and extinction equations with the characteristics of a camera and the background scenery, we develop a governing equation that relates camera signal to the properties of aerosol, the incident light and the image being captured. From inversion of this integral equation, we establish an expression for turbidity. Using 3-years' worth of images captured from a camera at a fixed location (downtown Chicago), we are able to test our model. We use a classification learner to first select images from clear weather conditions and then analyzed the photographs to calculate turbidity values based on changes in the images of buildings. We show that our calculated turbidity values are well correlated with on-ground measurements of $PM_{2.5}$ data from nearby EPA monitoring sites. I will discuss our model development approach, data collection techniques, and the machine learning tools used to generate our model. I will also provide preliminary results from our work and the implications of our research in expanding air monitoring globally.

Suresh Dhaniyala is the Bayard D. Clarkson Distinguished Professor in the Mechanical and Aeronautical Engineering Department at Clarkson University. Prof. Dhaniyala's interests are in the fields of air quality monitoring, sensors, and health effects of airborne particles. He received a NSF CAREER award and Clarkson's John W. Graham Jr. Award for research accomplishments and the Pi Tau Sigma award for teaching. He has more than 50 peer-reviewed publications, authored 3 book chapters, and 3 patents on aerosol sensing techniques.

Clarkson University

David A. Walsh ‘67 Arts & Sciences Mini-Conference

Friday August 31, 2018

# BIG DATA APPLICATIONS

## Using patterns of molecular evolution to identify sites under selection across genes and genomes

4.10 PM - 4.35 PM

### SUSAN BAILEY, Ph.D.

ASSISTANT PROFESSOR
DEPARTMENT of BIOLOGY
CLARKSON UNIVERSITY

Evolution is the outcome of many of stochastic events, beginning with the random process of mutation, and moving on to the complexities of ecological and environmental interactions that drive natural selection. For this reason, the general patterns and processes of evolution, particularly at the molecular level, can sometimes be difficult to characterize without a large amount of data. Fortunately, with the rise of cheap and fast genome sequencing technology, we now have access to thousands of terabytes of new publicly available DNA sequence data each year. I will discuss traditional bioinformatics approaches used to scan genes and genomes for signs of selection and present a new type of approach that uses patterns of parallel evolution in a large collection of genomes to identify nucleotide sites under selection that have often been overlooked in the past. I show that preliminary results from this new approach correlate well with experimental measures of selection at the same nucleotide sites and discuss the challenges and next steps in refining this method.

Susan Bailey is an Assistant Professor of Biology at Clarkson University. Prior to this position, she was a post-doctoral researcher in the Bioinformatics Research Center at Aarhus University, Denmark and obtained her PhD in Evolutionary Biology from the University of Ottawa, Canada. Her research focuses on understanding the processes driving evolution using mathematical and statistical models in combination with experimental evolution of microbes.

Clarkson University

David A. Walsh ‘67 Arts & Sciences Mini-Conference

**Friday August 31, 2018**

# BIG DATA APPLICATIONS

PROFESSOR/CHAIR

DEPARTMENT of BIOLOGY

CLARKSON UNIVERSITY

## Citizen science reveals negative effects of roads and road traffic on amphibians across spatial scales and regions in the eastern United States

**4.35 PM - 5.00 PM**

**TOM LANGEN Ph.D.**

The North American Amphibian Monitoring Program (NAAMP) is a citizen science program for which trained volunteers do yearly anuran (frog and toad) call surveys along specified routes multiple times a year. We examined the effects of habitat composition, roads and road traffic, and habitat configuration on the distribution and species richness of anurans among thirteen eastern and central United States. Undergraduates from nine biology and environmental science courses collated occupancy data and characterized landscape structure at five spatial scales at 1617 sampling locations from NAAMP, surveyed within 1999 to 2013. We found that anuran species richness and individual species distributions were negatively impacted by high road densities and traffic volumes. The negative effect of road density on anuran richness was strongest at the smallest spatial scales (300–1000 m), and this pattern was consistent across regions. Our results indicate that the underlying effects of roads on amphibians likely involve direct mortality, behavioral barriers to movement, and reduction in the quality of roadside habitats. Our results also demonstrate the scientific value of well-designed, large-scale citizen science programs, and the educational value of multi-university collaborative problem-based learning.

Dr. Tom A. Langen is the Department Chair and Professor of Biology at Clarkson University. Dr. Langen conducts research on the environmental impact of roads, and on the effectiveness of public-private partnerships for wetland restoration. He also is involved in several projects focused on developing teaching modules for undergraduates in ecology and environmental science that use 'big data' to teach concepts related to large spatial and temporal scales.