
*December 26, 2019*
*Indian Institute of Technology, Mumbai*
*Maharashtra, India*

**Program Committee**

Balasubramanian Narasimhan, Stanford University
Meghna Patnaik, Indian Statistical Institute
Deepayan Sarkar, Indian Statistical Institute
Susan Thomas, Indira Gandhi Insitute of Development Research

# BOOK OF ABSTRACTS

Sponsors


IISA
International Indian Statistical Association


R Studio®


H2O.ai


Google Research


R consortium

# Contents

## Part V: Contributed Talks

## Conference Schedule

| |
|---|
| 08:30 Check in |
| 09:00–11:00 **Invited Tutorial**<br>R Packages for Communicating Reproducible Research, *Martin Morgan* |
| 11:00–11:15 Break |
| 11:15–12:15 **Contributed talks**<br>• How Do People Engage With UK Parliamentary Debates, *Pushkal Agarwal*<br>• What is in a name: an analysis of diversity in India, *Sabir Ahamed*<br>• Are Business School Reviews Helpful? A Text Analysis and Machine Learning Approach with Evidence from India, *Soumyajyoti Datta*<br>• Namma App: Exploring mobile monitoring air quality data using the Shiny package, *Adithi Upadhya*<br>• Simulation study using R for comparing Kaplan-Meier method with weighted Kaplan-Meier methods, *Jagathnath Krishna* |
| 12:15–13:30 Break |
| 13:30–14:30 **Invited talk**<br>Hidden variables: latent structure in bacterial communities in the human microbiome, *Susan Holmes* |
| 14:30–14:45 Break |
| 14:45–15:45 **Special contributed talks**<br>• Efficiency in data processing: data.table basics, *Jan Gorecki*<br>• Disciplined Convex Programming in R, *Balasubramanian Narasimhan* |
| 15:45–16:30 **Contributed talks**<br>• Exploring probability distributions for bivariate temporal granularities, *Sayani Gupta*<br>• Variable Clustering using Ortho-oblique Rotation, *Krishna Mohan Roy*<br>• Developing an R Package for New Two-Stage Methods in Forecasting Time Series with Multiple Seasonality, *Anupama Lakshmanan*<br>• Shiny tool for building Interactive Decision Tree, *Snehasish Sarkar* |
| 16:30–17:30 **Posters**<br>• R vs Python for cancer genomics data analysis in oral cancer, *Sachendra Kumar*<br>• Sleuth: Pipeline for Differential analysis of RNA-Seq data, *Jyotsana Mehra*<br>• Analysis of Barbados malnutrition data using Bioconductor, *Moumita Karmakar*<br>• Meta-Heuristic Optimisation methods for model-fitting in R, *Sourav Garg* |

# Part I

# Invited Tutorial

Presentation type: Invited Tutorial

# R Packages for Communicating Reproducible Research


## Martin Morgan

*Research Professor, Biostatistics, SUNY, Buffalo and Director of the Bioconductor project, USA*



**Abstract:** This tutorial is for all who wish to write R packages. R is a fantastic language for you to develop new statistical approaches for the analysis and comprehension of real-world data. R packages provide a way to capture your new approach in a reproducible, documented unit. An R package is surprisingly easy to create, and creating an R package has many benefits. In this tutorial we create an R package. We start with a data set and a simple script transforming the data in a useful way; perhaps you have your own data set and script? We replace the script with a function, and place the function and data into an R package. We then add documentation, so that our users (and our future selves) understand what the function does and how the function applies to new data sets. With an R package in hand, we can tackle more advance challenges: vignettes for rich narrative description of the package; unit tests to make our package more robust; and version control to document how we change the package. The final step in the development of our package is to share it with others, through github, through CRAN, or though domain-specific channels such as Bioconductor.



*Keywords:* R packages, reproducible research

# Part II

# Invited Talk

Presentation type: Invited Talk

# Hidden variables: latent structure in bacterial communities in the human microbiome


## Susan Holmes

*Professor of Statistics, Stanford University, USA*



**Abstract:** The analyses of complex biological systems often results in output that may seem just as complex, with little useful knowledge extracted as a result of the multiple layers of information. Analogies with methods in textual analyses (Natural Language Processing) such as the use of latent variables methods provides useful interpretations as shown by Sankaran and Holmes, 2018. The use of multi-scale strategies is providing useful predictions of preterm birth and a deeper understanding of resilience of the human microbiome after antibiotic perturbations.

Our team has shown that Bayesian and Bootstrap approaches can provide non-parametric answers to the statistical challenges and have supplemented these with effective uncertainty visualization techniques distributed as Bioconductor/R packages (phyloseq, adaptiveGPCA, reelapse, bootLong). This presentation will include joint work with Kris Sankaran, Julia Fukuyama, Ben Callahan, Claire Donnat, Joey McMurdie, Pratheepa Jeganathan, Lan Huong Nguyen and David Relman's group at Stanford.



*Keywords:* microbiome, bioinformatics, bayesian methods, resampling

# Part III

# Special Contributed Talks

Presentation type: Special Contributed Talks

# Efficiency in data processing: data.table basics


## Jan Gorecki

*H2O.ai*



**Abstract:** We will briefly go through the db-benchmark report to see performance-wise state of data processing tools for operations such as aggregation and join. Then we will discuss basic concepts of data.table. How we can use extended version of data.frame '[' operator to achieve flexibility of expressing data processing operations in comprehensive and concise way.



*Keywords:* data.table, data processing, data aggregation

Presentation type: Special Contributed Talks

# Disciplined Convex Programming in R


## Balasubramanian Narasimhan

*Senior Research Scientist and Director, Data Coordinating Center, Department of Biomedical Data Sciences, Stanford University*



**Abstract:** Convex optimization plays an important role in statistics and machine learning. We introduce CVXR, an R package for Disciplined Convex Programming (DCP), a way of formulating convex optimization problems in a natural mathematical syntax. Problem objectives and constraints can be constructed by combining constants, variables and parameters using a library of functions with known mathematical properties. The DCP calculus verifies the convexity, converts it using a graph implementation into a form that solvers, either open source or commercial, can handle. We will illustrate with a number of examples. This is joint work with Anqi Fu and Stephen Boyd.



*Keywords:* convex optimization, estimation, machine learning

# Part IV

# Posters

Presentation type: Posters

# R vs Python for cancer genomics data analysis in oral cancer


## Sachendra Kumar

*Indian Institute of Science, Bengaluru, India*



**Abstract:** Recent advances in cancer genomics research area opens opportunity to develop new methods and tools for analysing cancer genomics data. Bioconductor R packages and Python script are widely used to analyse high-throughput genomic data by bioinformatician and data scientist. Therefore, the comparative data analysis using R and Python could help to explore its application in genomic data analysis. The comparative data analysis was performed on subset of The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma (TCGA-HNSC) using multiple genomic data sets (gene expression and mutation) in oral cancer. The result suggests both R and Python have their pros and cons and could be improved further for the cancer genomic data analysis.



*Keywords:* R, Python, bioinformatics

Presentation type: Posters

# Sleuth: Pipeline for Differential analysis of RNA-Seq data


## Jyotsana Mehra

*Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi, India*



**Abstract:** In the field of Computational Biology, with the advancements of data collection techniques there is an overwhelming data accumulation. For non-native programmers, such as biologist, handling, analyzing and visualizing biological data in itself is a difficult task. To solve this problem researchers have introduced tools and packages for the state of the art programming languages like R, Python etc. In our work we present a pipeline that takes RNA-Seq data and does differential analysis using Sleuth, an R package. In the pipeline we also make use of dependencies like Kallisto, a program which we use for quantification of transcripts in RNA-Seq data and Bioconductor which provides tools for analysis of genomic data. Using the aforementioned pipeline on samples we can identify the significant genes as biomarkers playing an important role in the identification of particular disease.



*Keywords:* differential analysis, RNA-Seq, bioinformatics

Presentation type: Posters

# Analysis of Barbados malnutrition data using Bioconductor


## Moumita Karmakar

*Texas A&M University, Texas, USA*



**Abstract:** Barbados malnutrition study is a type of 450k methylation data. There are a total of 94 samples from generation 1(G1). G1 individuals suffered from severe protein malnutrition during the first year of life and are followed up to 48 years in the study. Actual biological age of individuals with their nutrition exposure status (control (CON) or postnatal malnourished (MAL)) are available. There are 47 female (20 CON and 27 MAL) and 47 (24 CON and 23 MAL) male subjects. We also have information available on Metabolic syndrome variables like glucose intolerance test (gtt), fasting plasma glucose level (fpg) etc. for the same individuals(measured two different time points). Prenatal malnutrition is positively associated with risk of metabolic syndrome. Here our goal is to evaluate the association/direction of association between the postnatal malnutrition/malnutrition limited to first year of life and metabolic syndrome variables in the Barbados Nutrition Study (BNS) cohort. The analyses were performed using R Language 3.03 and Bioconductor 2.13.



*Keywords:* Bioconductor, bioinformatics, methylation data

Presentation type: Posters

# Meta-Heuristic Optimisation methods for model-fitting in R


## Sourav Garg

*Indian Institute of Management, Indore, India*



**Abstract:** Evolutionary and Meta-Heuristic optimisation techniques are commonly used for non-differential, multi-dimensional functions with many local optima where traditional methods fail to reach the global optimum. Such methods provide us with accurate estimates of model parameters. We are employing these optimizing methods in a semi-parametric model fitting exercise where the objective is to model the profits of promising start-ups with respect to their spending patterns. We are comparing different methods based on certain goodness of fit measures to find the optimal one.



*Keywords:* optimization, estimation

# Part V

# Contributed Talks

Presentation type: Contributed Talks

# How Do People Engage With UK Parliamentary Debates?


## Pushkal Agarwal

*Department of Informatics, King's College, London, UK*



**Abstract:** In this paper, we attempt to characterise how people engage with video data of Parliamentary debates by using more than two years of Google Analytics data around these videos. We analyse the patterns of engagement—how do they land on a particular video? How do they hear about this video, i.e., what is the (HTTP) referrer website that led to the user clicking on the video? Once a user lands on a video, how do they engage with it? For how long is the video played? What is the next destination? We employ Non-Negative Matrix Factorization (NMF) and Principal Component Analysis (PCA) libraries of R on the video views (6 Million in last 2 years) matrix to identify different archetypes of users, and identify 3 archetypes (Direct, Social and Search). Interestingly, these different archetypes appear to have different levels of engagement with the Parliamentary videos.



*Keywords:* nonnegative matrix factorization, principal components, video analysis

Presentation type: Contributed Talks

# What is in a name: an analysis of diversity in India


## Sabir Ahamed

*Pratichi Insitute, Pratichi Trust, India*



**Abstract:** Parents took one of the toughest decision in naming their wards. Analysis of baby name has gained popularity in the US. In India, childrens' name is not available in big data set, data on the birth certificate are also not readily available for analysis. In India, naming their Kids also depends on religion, caste, and language etc. The proposed presentation aims to draw a picture of diversity in India. Drawing on a data of 29 Lakh class IX students, the presentation will show the top gender-wise 50 names along with the social groups, ( Schedule Caste, Schedule Tribe, and Muslim etc,). The analysis will make an attempt the prepare a tidy data frame by separating the names, creating sets of words by `unset_token`, visualization the data using word frequencies, wordcloud etc. To demonstrate a generational gap in naming, we will compare the name of their guardian.



*Keywords:* text mining, visualization

Presentation type: Contributed Talks

# Are Business School Reviews Helpful? A Text Analysis and Machine Learning Approach with Evidence from India


## Soumyajyoti Datta

*Indian Institute of Management, Indore, India*



**Abstract:** In order to develop a pool of well trained mangers and leaders equipped with the scientific means of solving complex and multi-faceted business problems , the policy makers and several entrepreneurs have taken initiatives to establish business schools across the nations like India. It is also concurrent with the phenomenal rise of social media. Reviews about the management institutes, which have been submitted at various online portals remain an important and unexplored area for scientific investigation. This data is huge, dynamic, unstructured making the traditional methods of analysis inappropriate. Nevertheless, due to the immense business value for the management training industry, the current study, through a substantially large corpus of Google reviews, has exploited the natural language processing and machine learning techniques to accomplish meaningful industry insights. Sentiment extraction and LDA topic modeling along with OLS regression have been employed using the various packages of R. The study has implications for the academic administrators, potential candidates, academic marketers and online study portal administrators.



*Keywords:* text mining, natural language processing, visualization

Presentation type: Contributed Talks

# Namma App: Exploring mobile monitoring air quality data using the Shiny package


## Adithi Upadhya

*ILK Consultancy, USA*



**Abstract:** R Shiny has become increasingly popular among R users for developing flexible and interactive platforms.The Shiny app we present here is for visualising data collected using mobile measurements (sensors on a moving platform). We used this app for visualising the data collected on air pollutants like $PM_{2.5}$, Black Carbon and Ultrafine particles, with portable instruments (which are capable of providing high temporal resolution output, for instance, 1 hertz data) mounted in a CNG-fuelled car. The mobile air pollution monitoring campaign is carried out in select neighbourhoods in Bengaluru city. The individual pollutant files are merged with the GPS data obtained separately.The resulting files can be easily accessed through the app and used to visualise the pollutant data. With the click of a button, the app generates: summary statistics, time series plots and spatial map of pollutant concentrations. Mapping parameters (pollutants) can be selected from a drop-down menu.

One of the challenges this app addresses is that of managing high frequency data ($\sim$ 1 hertz) generated using a mobile platform. The app allows team members to easily access data from the storage location and visualise the data in near-real time, without requiring knowledge in R. The app reduces the time consumed for analysing each pollutant individually. It helps check the quality of the data at near real time and instantly visualise pollution hot spots. The time series plots of each pollutant help understand the temporal patterns in addition to the health of the instruments. This app will help air pollution researchers and amateurs interested in conducting individual mobile experiments using portable air-quality sensors, which are easily available in the market, and also to obtain high-quality data.



*Keywords:* shiny, time series, air pollution

Presentation type: Contributed Talks

# Simulation study using R for comparing Kaplan-Meier method with weighted Kaplan-Meier methods


## Jagathnath Krishna

*Regional Cancer Centre, Thiruvananthapuram, India*



**Abstract:** Kaplan-Meier (K-M) method is used to estimate survival from time-to-event data and it assumes loss to follow-up (LFU) as random. But when the study outcome influenced by the factors for loss, LFU can be non-random and it leads to over-estimated survival. In order to reduce the over estimation of survival probabilities, weighted K-M methods has been introduced by several researchers (Jan et al. 2004, Huang 2008). Even though these methods reduce over estimation, most of the times these estimates results in under estimated survival. Hence Jagathnath et al. 2019 introduced modified weighted K-M (MWKM) by giving weightage to loss to follow (LFU). The present study aimed to address the issue of over estimation and under estimation existing survival methods and to compare various weighted K-M methods. An R program was developed for MWKM method and other weighted KM methods for estimating the survival probabilities. Simulation study and real data analysis using R program were done. The simulation study was conducted for different proportion of follow-up and censoring and observed that MWKM gives better survival estimate with least bias compared to other methods.



*Keywords:* survival analysis, censored data

Presentation type: Contributed Talks

# Exploring probability distributions for bivariate temporal granularities


## Sayani Gupta

*Monash University, Australia*



**Abstract:** Smart meters measure energy usage at fine temporal scales, and are now installed in many households around the world. We propose some new tools to explore this type of data, which deconstruct time in many different ways. There are several classes of time deconstructions including linear time granularities, circular time granularities and aperiodic calendar categorizations. Linear time granularities respect the linear progression of time such as hours, days, weeks and months. Circular time granularities accommodate periodicities in time such as hour of the day, and day of the week. Aperiodic calendar categorizations are neither linear nor circular, such as day of the month or public holidays.

The hierarchical structure of many granularities creates a natural nested ordering. For example, hours are nested within days, days within weeks, weeks within months, and so on. We refer to granularities which are nested within multiple levels as "multiple-order-up" granularities. For example, hour of the week and second of the hour are both multiple-order-up, while hour of the day and second of the minute are single-order-up.

Visualizing data across various granularities helps us to understand periodicities, pattern and anomalies in the data. Because of the large volume of data available, using displays of probability distributions conditional on one or more granularities is a potentially useful approach. This work provides tools for creating granularities and exploring the associated within the tidy workflow, so that probability distributions can be examined using the range of graphics available in the ggplot2 package. In particular, this work provides the following tools:

- Functions to create multiple-order-up time granularities. This is an extension to the lubridate package, which allows for the creation of some calendar categorizations, usually single-order-up.
- Checks on the feasibility of creating plots or drawing inferences from two granularities together. Pairs of granularities can be categorized as either a harmony or clash, where harmonies are pairs of granularities that aid exploratory data analysis, and clashes are pairs that are incompatible with each other for exploratory analysis.



*Keywords:* distribution theory, visualization

Presentation type: Contributed Talks

# Variable Clustering using Ortho-oblique Rotation


## Krishna Mohan Roy

*Bridgei2i.com*



**Abstract:** As dimensionality increases, it is more difficult to determine irrelevant inputs than to identify redundant inputs. The goal of variable reduction algorithms is to reduce the dimension into lower space. Variable reduction techniques like PCA, transform original dimensions into other dimensions where the goal is to maximize the explained variance using less number of dimensions. Techniques like PCA are challenging in terms of business interpretation and it becomes difficult for an analyst to explain the principal components to the business users. It makes the results of further analysis (say key driver analysis) less actionable.

We have built an R package which performs variable clustering (to be called **varclusbi2i** henceforth). The package allows users to split clusters based on eigenvalue or minimum variance explained by clusters. It also allows user to visualize and save the results. **Varclusbi2i** can solve the problem of explanation by retaining the original variables as it forms clusters of variables that are similar in nature and thus the user can select representative from each cluster based on the importance of variables within the cluster. **Varclusbi2i** is an oblique principal component analysis to get non-overlapping clusters of variables. It can help a statistician to quickly select important variables or reduce the number of variables used for building models. It clusters variables (i.e. forms group of variables) that are highly correlated within the cluster and highly un-correlated with variables in the other clusters. The algorithm used by **varclusbi2i** is binary and divisive—i.e. initially all the variables begin in one cluster and then they split until the second eigenvalue of clusters becomes greater than given threshold. It is a non-hierarchical way of clustering as variables can be re-assigned to other clusters as well.



*Keywords:* variable clustering, dimension reduction

Presentation type: Contributed Talks

# Developing R Package for New Two-Stage Methods in Forecasting Time Series with Multiple Seasonality


## Anupama Lakshmanan

*Indian Institute of Management, Bengaluru, India*



**Abstract:** Complex multiple seasonality is an important emerging challenge in time series forecasting. We propose a framework that segregates the task into two stages. In the first stage, the time series is aggregated at the low frequency level (such as daily or weekly) and suitable methods such as regression, ARIMA or TBATS, are used to fit this lower frequency data. In the second stage, additive or multiplicative seasonality at the higher frequency levels are estimated using classical, or function-based methods. Finally, the estimates from the two stages are combined.

In this work, we build a package for implementing the above two-stage framework for modeling time series with multiple levels of seasonality within R. This would make it convenient to execute and possibly lead to more practitioners and academicians adopting it. The package would allow the user to decide the specific methods to be used in the two stages and the separation between high and low frequency. Errors are calculated for both model and validation period, which may be selected by the user and model selection choices based on different criterion will be facilitated. Forecast combination may also be integrated with the developed routine. The schematics will be presented along with demonstration of the package in several real data sets.



*Keywords:* time series, modeling

Presentation type: Contributed Talks

# Shiny tool for building Interactive Decision Tree


## Snehasish Sarkar

*Bridgei2i.com*



**Abstract:** CHAID and CART are popular algorithm used for building decision tree.There are plethora of packages available in R, all of them directly give a decision tree as an output based on the target and independent variables as an input. The user has no control whatsoever when it comes to selecting the variables and the nodes at which they want to split. Most of the times these splits are not useful for taking business decisions as they are chosen purely based on mathematical significance. This is where our solution comes handy and gives users control over selection of variables and node for split.

**ITreeBi2i** is an R package which gives the power in users hand. The algorithm enables the user to decide the variable that would be used to create the split at any given node. The user through an R Shiny platform can view a score (based on Gini Index or Chi-Square) for each variable at the time of split and can make a decision to split with his desired variable. This helps the user to select the best variable not only through a statistical significance but also by using his business rules/constraints. As node is getting splitted, same can be visualized in a tree structure in R Shiny platform. User has a flexibility to merge, delete any node by a simply click after seeing the split result.

An user friendly shiny application backed up by strong decision tree algorithm is very much useful for business user.



*Keywords:* interactivity, web application, decision trees