
*June 27–30, 2016*
*Stanford University*
*California, USA*

**Program Committee**

Jenny Bryan, University of British Columbia
Dianne Cook, Monash University
Peter Dalgaard, Copenhagen Business School
Dirk Eddelbuettel, Ketchum Trading
Susan Holmes, Stanford University
Torsten Hothorn, Universität Zürich
Julie Josse (Chair), Agrocampus Ouest Rennes
Patrick Mair, Harvard University
Jeroen Ooms, University of California, Los Angeles
Hilary Parker, Stitch Fix
Hana Ševčíková, University of Washington, Seattle
Torben Tvedebrink, Aalborg University
Heather Turner, University of Warwick

# BOOK OF ABSTRACTS

Platinum Sponsors







Gold Sponsors



# Silver Sponsors

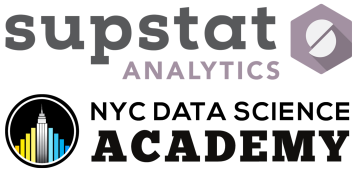



Bronze Sponsors

---







Publishers


SAGE
Publishing


CAMBRIDGE
UNIVERSITY PRESS


WILEY


Springer

# Other Sponsors



# useR! 2016 Program at a glance

| | Monday, June 27, 2016 | Tuesday, June 28, 2016 | Wednesday, June 29, 2016 | Thursday, June 30, 2016 |
|---|---|---|---|---|
| 8:00-8:30 | Registration (all day) | Registration (all day) | Registration (morning) | |
| 8:30-9:00 | | Opening Session | | |
| 9:00-10:00 | Morning Tutorials (including coffee break) | Richard Becker | Hadley Wickham | Deborah Nolan |
| 10:00-10:30 | | Coffee Break including Poster Exhibits (display at Sponsor Pavilion) | | |
| 10:30-11:00 | | • Bayesian (Barnes)<br>• R in Business (Econ 140)<br>• Bioinformatics (Lane)<br>• Kaleidoscope (McCaw)<br>• Performance (Siepr 130) | • Statistics Methods (Barnes)<br>• Performance (Econ 140)<br>• Bioinformatics (Lane)<br>• Kaleidoscope (McCaw)<br>• Lightning Talks (Siepr 120)<br>• Database (Siepr 130) | • Teaching (Barnes)<br>• Lightning Talks (Econ 140)<br>• Statistics & Big Data (Lane)<br>• Kaleidoscope (McCaw)<br>• Graphics (Siepr 130) |
| 11:00-12:00 | | | | |
| 12:00:1:00 | Lunch | Lunch including Poster Exhibits (displayed at Sponsor Pavilion) | | |
| 1:00-2:00 | | • Regression (Barnes)<br>• R & Other Languages (Lane)<br>• Kaleidoscope (McCaw)<br>• Case Study (Siepr 120)<br>• Teaching (Siepr 130) | • Reproducible Research (Barnes)<br>• Generalized Mixed Models (Econ 140)<br>• Sponsor Session (Lane)<br>• Kaleidoscope (McCaw)<br>• Spatial (Siepr 120)<br>• Packages & Development (Siepr 130) | • Sponsor (Lane)<br>• Lightning Talks (McCaw)<br>• Lightning Talks (Siepr 130) |
| 2:00-2:15 | Afternoon Tutorials (including coffee break) | | | |
| 2:15-2:30 | | | | Simon Urbanek |
| 2:30-3:00 | | Coffee Break including Poster Session (displayed at Sponsor Pavilion) | | |
| 3:00-3:30 | | | | Closing Remarks |
| 3:30-3:45 | | Donald Knuth | Daniela Witten | Conference ends |
| 4:00-4:15 | Short Break | | | |
| 4:15-4:30 | R Initiatives | | | |
| 4:30-4:45 | | Short Break | Bus to Cruise (board by 5:00 latest) | |
| 4:45-5:30 | | • Analytics (Barnes)<br>• Miscellaneous (Econ 140)<br>• Interactive (Lane)<br>• Kaleidoscope (McCaw)<br>• Miscellaneous (Siepr 120)<br>• Machine Learning (Siepr 130) | | |
| 6:00-6:15 | | | Cruise and Conference Dinner (Sponsored by Microsoft) Return to campus at 10:30 | |
| 6:30-8:30 | | Welcome Reception (Sponsored by RStudio) | | |

# Contents

## Part I: Poster

Contents

Contents

Contents

## Part II: Lightning Talk

Contents

Contents

## Part III: Oral Presentation

Contents

Contents

Contents

Contents

Contents

Contents

Contents

Contents

Contents

# Part I

# Poster

Presentation type: Poster

# High-performance R with FastR


## Adam Welc

*Oracle Labs*



**Abstract:** R is a highly dynamic language that employs a unique combination of data type immutability, lazy evaluation, argument matching, large amount of built-in functionality, and interaction with C and Fortran code. While these are straightforward to implement in an interpreter, it is hard to compile R functions to efficient bytecode or machine code. Consequently, applications that spend a lot of time in R code often have performance problems. Common solutions are to try to apply primitives to large amounts of data at once and to convert R code to a native language like C. FastR is a novel approach to solving R's performance problem. It makes extensive use of the dynamic optimization features provided by the Truffle framework to remove the abstractions that the R language introduces, and can use the Graal compiler to create optimized machine code on the fly. This talk introduces FastR and the basic concepts behind Truffle's optimization features. It provides examples of the language constructs that are particularly hard to implement using traditional compiler techniques, and shows how to use FastR to improve performance without compromising on language features.



*Keywords:* performance, compilation, fastr

Presentation type: Poster

# DiLeMMa - Distributed Learning with Markov Chain Monte Carlo Algorithms with the ROAR Package


## Ali Mehdi Zaidi

*Microsoft*



**Abstract:** Markov Chain Monte Carlo algorithms are a general technique for learning probability distributions. However, they tend to mix slowly in complex, high-dimensional models, and scale poorly to large datasets. This package arose from the need for conducting high dimensional inference in large models using R. It provides a distributed version of stochastic based gradient variations of common continuous-based Metropolis algorithms, and utilizes the theory of optimal acceptance rates of Metropolis algorithms to automatically tune the proposal distribution to its optimal value. We describe how to use the package to learn complex distributions, and compare to other packages such as RStan.



*Keywords:* MCMC, simulation, parallel computing, distributed computing, Metropolis

Presentation type: Poster

# Analyzing and visualizing spatially and temporally variable floodplain inundation


## Alison A. Whipple & Joshua H. Viers

*University of California, Davis, University of California, Merced*



**Abstract:** With the continuing degradation of riverine ecosystems, advancing our understanding of the spatially and temporally variable floodplain conditions produced by a river's flood regime is essential to better manage these systems for greater ecological integrity. This requires development of analysis and visualization techniques for multi-dimensional spatio-temporal data. Research presented here applies 2D hydrodynamic modeling output of a floodplain restoration site along the lower Cosumnes River, California in R to analyze this spatio-temporal raster data and develop informative and engaging visualizations. Modeling output is quantified and compared within and across modeled flood events in space and time using metrics such as depth, velocity, and duration. To aid comparison and interpretation, rasters of model time steps are also summarized by integrating across space as well as across time. Data manipulation and summary is performed primarily within the raster package. This research presents new methods for quantifying and visualizing hydrodynamic modeling outcomes, improving understanding of the complex and variable floodplain inundation patterns that drive ecosystem function and process.



*Keywords:* spatio-temporal analysis, raster analysis, visualization, floodplain, hydrodynamic modeling

Presentation type: Poster

# Visualization of health and population indicators within urban African populations using R

**Amos Mbugua Thairu, Martin Mutua, Marylene Wamukoya, Patricia Elungata, Thaddeus Egondi, Zacharie Dimbuene & Donatien Beguy**

*African Population and Health Research Center (APHRC)*

**Abstract:** The Demographic and Health Surveys (DHS) Program has collected and disseminated open data on population and health through more than 300 surveys from various countries. One of our research interests is to investigate the linkage between urban poverty and health in African countries. Using the DHS raw data we have computed indicators focusing on exploring how the indicators differ between different groups in the urban areas. These groups are based on wealth tertiles and consist of the urban poor, urban middle and the urban rich.

Following the analysis we have developed the Urban Population and Health Data Visualization Platform which is an interactive web application using Shiny. Online deployment of the platform through the APHRC website is underway and we believe it will assist policymakers and researchers to perform data explorations and gather actionable insights. By sharing the code through github we hope that it will contribute towards promoting the adoption of R particularly by universities and researchers in Africa as an alternative to costly proprietary statistical software.

The platform showcases the power of R and is developed using R and various R packages including shiny, ggplot, googleVis, RCharts, DT for graphics and dplyr for data manipulation.

*Keywords:* Data Visualization, Open Data, Population and Health Surveillance, Interactive Web Application

Presentation type: Poster

# Educational Disparities, Biomedical Efficacy and Science Knowledge Gaps: can the Internet help us reduce these inequalities?

**Andreea Loredana Moldovan & Nick Allum**

*Department of Sociology, University of Essex*

**Abstract:** The economic, health, and knowledge disparities between the world's "haves" and "have-nots" are some of the key issues we face in this day and age. (World Economic Forum, 2011) Unfortunately, very little communication research has been applied to understanding what we can do to help reduce these inequalities. Even more worryingly, some studies have found that feeding more information to the public through traditional media has the adverse effect of widening gaps based on educational disparities. ( Tichenor, Donohue, Olien, 1970) We study the impact that Internet use has on the disparity between lowly and highly educated citizens in terms of their science (biomedical) knowledge, as well as their sense of efficacy regarding medical research. For this, we employ Wave II of the Wellcome Trust Monitor Survey (2012), which is fielded to a nationally representative sample of the UK population. We conduct a series of moderated regression models with mean centring using the 'lmres' function in the 'pequod' package. (Mirisola, A. & Seta, L., 2016) We also use the 'simpleSlope' and 'PlotSlope' functions in order to do a simple slope analysis, as well as to create two and three-way interaction plots. These functions are comprehensive of what the statistical literature recommends for such tests, and they save time and effort by reducing the number of analytical steps. R helped us find that increased Internet use in the lower education group can help significantly narrow both knowledge and efficacy gaps that emerge from educational disparities. Implications for science communication are discussed.

*Keywords:* education, inequality, efficacy, knowledge, pequod, moderated regression

Presentation type: Poster

# Statistics and R for Analysis of Elimination Tournaments


## Ariel Shin & Norm Matloff

*University of California, Davis*



**Abstract:** There is keen interest in statistical methodology in sports. Such methods are valuable not only to sports sociologists but also those in sports themselves, as exemplified in the book and movie "Moneyball." These statistics enhance comparisons among players and possibly even enable prediction of games. However, elimination tournaments present special statistical challenges. This paper explores data from the national high school debate circuit, in which the first author was an active national participant. All debaters participate in the 6 pre-elimination rounds, but subsequently the field successively narrows in the elimination rounds. This atypical format makes it difficult to use classical statistical methods, and also requires more sophisticated data wrangling. This paper will use R to explore questions such as: Does gender affect the outcome of rounds? Does geography play a role in wins/losses? What constitutes an upset? Is there a so-called "shadow effect," in which the weaker the expected competitor in the next round, the greater the probability that the stronger player will win in the current stage? Among the purposes of this project is to use it as an R-based teaching tool, and help the debate community understand the inequalities that exist in relation to gender, region, and school. Typical graphs that can be generated may be viewed at `https://github.com/ariel-shin/tourn`. Our R software will be available in a package "tourn."



*Keywords:* elimination tournaments, data wrangling, gender differences, shadow effect

Presentation type: Poster

# Community detection in multiplex networks : An application to the C. elegans neural network

**Brenda Betancourt & Rebecca Steorts**

*Duke University*

**Abstract:** We explore data from the neuronal network of the nematode C. elegans, a tiny hermaphroditic roundworm. The data consist of 279 neurons and 5863 directed connections between them, represented by three connectomes of electrical and chemical synapses. Our approach uses a fully Bayesian two-stage clustering method, based on the Dirichlet processes, that borrows information across the connectomes to identify communities of neurons via stochastic block modeling. This structure allows us to understand the communication patterns between the motor neurons, interneurons, and sensory neurons of the C. elegans nervous system.

*Keywords:* Multiplex neural networks, community detection, Dirichlet process

Presentation type: Poster

# A Large Scale Regression Model Incorporating Networks using Aster and R


## Yun Wang & Brian Kreeger

*Wells Fargo, Teradata Aster*



**Abstract:** Leveraging the Aster platform and the TeradataAsterR package, end users can overcome the challenges of memory/scalability limitations of R and the costs of transferring large amounts of data between platforms. We explore integration of R with Aster, a MPP database from Teradata, focusing on a predictive analytical case study from Wells Fargo. It's always crucial for Wells Fargo to understand customer behaviors and why they do it. In this analysis, we utilized Aster's graph analysis functionalities to explore customer relationship, and check how network effect changes customers' behaviors. A logistic regression model was built, and a R shiny application was also used to visually represent impact of important attributes from the model.



*Keywords:* Big Data, MPP, applications, TeradataAsterR, customer relationship, Graph Analysis

Presentation type: Poster

# Profile Analysis of Multivariate Data Using the profileR Package


## Christopher David Desjardins & Okan Bulut

*Center for Applied Research and Educational Improvement, University of Minnesota, Centre for Research in Applied Measurement and Evaluation, University of Alberta*



**Abstract:** Profile analysis is a multivariate data analysis technique employed in the social sciences that is the statistical equivalent of a repeated measures extension of the MANOVA model. Profile analysis is mainly concerned with test scores; more specifically with profiles of test scores obtained from an assessment. A test score profile shows differences in subscores on tests that are commonly administered in medical, psychological, and educational studies to rank participants of a study on some latent construct. Practitioners in these fields are typically interested in quantifying both an individual's overall performance on a test (i.e., their level) and variation between scores on subtests within the test (i.e., their pattern). A suite of profile analytic procedures for decomposing observed scores into both level and pattern effects exists for the R programming language in the profileR package (Bulut and Desjardins 2015). This package includes routines to perform criterion-related profile analysis, profile analysis via multidimensional scaling, moderated profile analysis, profile analysis by group, and a within-person factor model to derive score profiles. This presentation will showcase several of these methods, illustrating their application with various data sets included within the package, as well as describing the future direction for the profileR package.



*Keywords:* profile analysis, psychometrics, educational assessment

Presentation type: Poster

# Urban Mobility Modeling using R and Big Data from Mobile Phones


## Daniel Emaasit

*University of Nevada Las Vegas*



**Abstract:** There has been rapid urbanization as more and more people migrate into cities. The World Health Organization (WHO) estimates that by 2017, a majority of people will be living in urban areas. By 2030, 5 billion people—60 percent of the world's population—will live in cities, compared with 3.6 billion in 2013. Developing nations must cope with this rapid urbanization while developed ones wrestle with aging infrastructures and stretched budgets. Transportation and urban planners must estimate travel demand for transportation facilities and use this to plan transportation infrastructure. Presently, the technique used for transportation planning includes the conventional four-step transportation planning model, which makes use of data inputs from local and national household travel surveys. However, local and national household surveys are expensive to conduct, cover smaller areas of cities and the time between surveys range from 5 to 10 years in even some of the most developed cities. This calls for new and innovative ways for Transportation Planning using new data sources.

In recent years, we have witnessed the proliferation of ubiquitous mobile computing devices (inbuilt with sensors, GPS, Bluetooth) that capture the movement of vehicles and people in near real time and generate massive amounts of new data. This study utilizes Call Detail Records (CDR) data from mobile phones and the R programming language to infer travel/mobility patterns. These CDR data contain the locations, time, and dates of billions of phone calls or Short Message Services (SMS) sent or received by millions of anonymized users in Cape Town, South Africa. By analyzing relational dependencies of activity time, duration, and land use, we demonstrate that these new "big" data sources are cheaper alternatives for activity-based modeling and travel behavior studies.



*Keywords:* Urban Mobility, Big Data, Cellular Data, Travel Demand Modeling, Activity-Based Models

Presentation type: Poster

# Web-based automated personalized homework with WebWork and R


## Davor Cubranic & Bruce Dunham

*University of British Columbia*



**Abstract:** WeBWorK is an open-source online homework system for math and sciencesncourses. It is used by over 1000 universities around the world. While WeBWorK includes a problem library of over 20,000 homework problems, few of those cover content from undergraduate statistics. Five years ago, we started a project to adopt WeBWorK in a range of first, second, and third year statistics classes, developing our own homework problems where necessary. Homework problems in WeBWorK are written in PG, a Perl-based DSL that combines problem definition, a Latex-like syntax for user-facing content, and sophisticated answer checkers. It quickly became apparent that implementing for it the necessary library of statistical functions and graphical support would be a huge undertaking, requiring writing the equivalent of R's "base", "stats", and "graphics" packages. Instead, we decided to provide a way to use R from PG. This included a set of PG "macros" for calling into R, retrieving results and converting them into PG data types, as well as displaying R's graphical output by WeBWorK's problem renderer. For performance and security, R is run on a separate host using Rserve, so we also wrote a Perl library implementing the Rserve client and reading serialized RDS and RData content. With this foundation, we have been able to develop WeBWorK homework content for ten statistics courses over three departments in the university, taken by over a thousand students over the past four years, and with very positive pedagogical outcomes so far.



*Keywords:* Statistical education, automated homework, cross-language integration

Presentation type: Poster

# Integrating R & Tableau


## Douglas Friedman, Jody Schechter & Bryan Baker

*Booz Allen Hamilton, Booz Allen Hamilton, US Army Corps of Engineers*



**Abstract:** Tableau is regularly used by our clients for the purposes of visualization and dashboarding, but they also often require the analytics and statistical functionality of R to analyze their data. While Tableau supports the integration of R, it is not always a straightforward process to blend the functionality of the two together. We plan to discuss our lessons learned from building Tableau applications that integrate with R, including best practices for performance optimization, sessionizing interaction on Tableau production servers, and reducing network latency issues. We will also discuss the limitations of Tableau's R integration capability.

Our goal is help others working to avoid common frustrations and roadblocks when integrating R and Tableau.



*Keywords:* Tableau, analytical tool development

Presentation type: Poster

# hurdlr: An R package for zero-inflated and over-dispersed count data


## Earvin Balderama & Taylor Trippe

*Loyola University Chicago*



**Abstract:** When considering count data, it is often the case that many more zero counts than would be expected of some given distribution are observed. It is well-established that data such as this can be reliably modeled using zero-inflated or hurdle distributions. However, it is also not uncommon that count data, especially ecological or environmental data, contain some number of extremely large observations which typically would be considered outliers and excluded from analyses due to difficulties in model fitting. In lieu of throwing out data or risk mis-specifying the distributional form, observations above a given threshold can be modeled by, e.g., an extreme value distribution, or any distribution with a long right tail. To utilize this modeling technique, we develop an R package "hurdlr" to utilize the double-hurdle model of Balderama, Gardner, and Reich (2014), which accounts for both the zero-inflation and extreme over-dispersion present in many count data sets. The hurdlr package functions are flexible and versatile: it can be applied with various count distributions and are able to allow for one or multiple hurdles. A Bayesian hierarchical framework is used for estimation, and covariate information can be included to inform top-level parameters.



*Keywords:* Bayesian estimation, Discrete count distributions, Ecology, Hurdle models, Over-dispersion, Zero-inflation

Presentation type: Poster

# All-inclusive but Practical Multivariate Stochastic Forecasting for Electric Utility Portfolio


## Eina Ooka

*The Energy Authority*



**Abstract:** Electric utility portfolio risk simulation requires stochastically forecasting various time series data: power and gas prices, peak and off-peak loads, thermal, solar and wind generation, and other covariates, in different time granularities. All these together presents modeling issues of autocorrelation, linear and non-linear covariate relationships, non-normal distribution, outliers, seasonal and weekly shapes, heteroskedasticity, temporal disaggregation and dispatch optimization. As a practitioner, I'll discuss how to organize and put together such a portfolio model from data scraping, simulation modeling, all the way to deployment through Shiny UI, while pointing out what worked what didn't.



*Keywords:* multivariate stochastic simulation, time series, seasonal and trend decomposition, outliers, heteroskedasticity, temporal disaggregation, regularization, rshiny, portfolio simulation, ML predictive model, time series regression, copula

Presentation type: Poster

# Statistical assessment of the similarity of amino-acid sequences


## Elena Rantou

*FDA*



**Abstract:** One of the classes of data considered in order to support equivalence of a generic to a reference listed drug is the comparison of amino-acid chain distributions. Sequences of amino-acids with certain molar ratio characteristics are used to explore novel comparison approaches, for these distributions. Different similarity measures, such as Tanimoto distances can produce a similarity matrix comparing the sequences. These measures will be compared based on their performance. Furthermore, we should search for important characteristics (features) that produce a meaningful separation of the sequences into clusters. This can be accomplished using weighted sampling, K-means and self-organizing maps (SOM). Additionally, clustering can be explored through building probability profiles for sequences of fixed lengths. In all these cases, a population of thousands of peptide chains from a single simulation resulted in hundreds of thousands of residue sequences. Data cleaning/organizing and pattern identification through these sequences of equal length, is computationally intensive and is carried using string detection functions such as 'str_detect' from the R-package 'string'.

When the circumstances necessitate cleavage of the amino-acid sequences at a certain residue, it is important to develop efficient coding, in order to investigate the properties of the distributions of the cleaved sequences and their molecular weights. The cleavage and sequencing of such immense size - data sets, is efficiently handled by the 'rstring' and 'Biostrings' R-packages and storage container functions such as 'AAStringSet'. This group of functions also facilitates the task of building empirical probability distributions of all unique amino acid sequences of a specified length.

The performance of different metrics will be assessed and all approaches will be discussed in the context of using similarity of the amino-acid sequences, in order to demonstrate bioequivalence between a complex-molecule drug and its generic version. Furthermore, the issue of seeking computationally efficient pathways for dealing with such data sets will be addressed.



*Keywords:* Amino-acid sequences, Similarity, Bioequivalence, Generic drugs, Molecular weight

Presentation type: Poster

# Monitoring nonlinear profiles with R: an application to Quality Control

**Emilio L. Cano, Javier M. Moguerza & Mariano Prieto Corcoba**

*Rey Juan Carlos University and the University of Castilla-La Mancha, Rey Juan Carlos University, ENUSA Industrias Avanzadas*

**Abstract:** In many situations, processes are often represented by a function that involves a response variable and a number of predictive variables. In this work, we show how to treat data whose relation between the predictive and response variables is nonlinear and, therefore, cannot be adequately represented by a linear model. This kind of data are known as nonlinear profiles. Our aim is to show how to build nonlinear control limits and a baseline prototype using a set of observed in-control profiles (Phase I analysis). Using R, we show how to afford situations in which nonlinear profiles arise and how to plot easy-to-use nonlinear control charts. This new class of control charts can be incorporated to a Statistical Process Control (SPC) strategy in order to deal with complex systems using tools that are familiar to process owners, such as control charts. The tool is also suitable for the Control phase of a DMAIC Six Sigma cycle. The SixSigma R package makes use of regularizacion theory in order to smooth the profiles. In particular, a Support Vector Machine (SVM) approach is followed, with an unattended parameters setting option. The package also allows to represent smoothed and non-smoothed profiles, and to compute the so-called prototype and confidence bands, which are actually the counterparts of center line and control limits in classical control charts, to monitor new profiles (Phase II analysis). The methods have been described in the book "Quality Control with R", within Springer's Use R! Series.

*Keywords:* quality control, control charts, nonlinear profiles, SVM, Six Sigma

Presentation type: Poster

# Applied Biclustering Using the BiclustGUI R Package


## Ewoud De Troyer & Ziv Shkedy

*Hasselt University (Center of Statistics)*



**Abstract:** Big and high dimensional data with complex structures are emerging steadily and rapidly over the last few years. A relative new data analysis method that aims to discover meaningful patterns in a big data matrix is *biclustering*. This method applies clustering simultaneously on 2 dimensions of a data matrix and aims to find a subset of rows for which the response profile is similar across a subset of columns in the data matrix. This results in a submatrix called a bicluster. The package `RcmdrPlugin.BiclustGUI` is a GUI plug-in for R Commander for biclustering. It combines different biclustering packages to provide many algorithms for data analysis, visualisations and diagnostics tools in one unified framework. By choosing R Commander, the BiclustGUI produces the original R code in the background while using the interface; this is useful for more experienced R users who would like to transition from the interface to actual R code after using the algorithms. Further, the BiclustGUI package contains template scripts that allow future developers to create their own biclustering windows and include them in the package. The BiclustGUI is available on CRAN and on R-Forge. The GUI also has a Shiny implementation including all the main functionalities. Lastly the template scripts have been generalized in the `REST` package, a new helping tool for creating R Commander plug-ins.



*Keywords:* biclustering, GUI, envelope package, R Commander, high dimensional data

Presentation type: Poster

# RCAP Designer: An RCloud Package to create Analytical Dashboards


## Ganesh K Subramaniam

*AT&T Labs*



**Abstract:** RCloud is an open source social coding environment for Big Data analytics and visualization developed by AT&T labs. We discuss RCAP Designer, an RCloud package that provides a way for Data Scientists to build R web applications similar to Shiny in the RStudio environment.

RCAP designer creates a workflow where the source R code is created within the RCloud environment in an R notebook. The package allows the data scientist to transform this notebook into an R dashboard application. This does not require developing web code (JavaScript, CSS, etc.). A number of widgets have been developed for creating the page design, several kinds of contents (R plots, interactive plots, an iframe, etc) and the different event controls for the page. For example, to include an R plot, one would drag and drop the RPlot widget onto the canvas. After the appropriate sizing of the plot window, the widget is configured to select the R plot function from the current workspace, and automatically link it to the control parameters. Once the design elements are saved, RCAP uses RCloud to render the page on the fly.

High level RCAP design considerations: On the server (R) side, RCAP produces the appropriate wrapping for the user's R code with the necessary templates to push the results back to the client side. This includes all of the RCloud commands and various error catching mechanisms. These wrapped functions are exposed to the JavaScript via OCAP. The user can just do normal plotting code and RCAP makes sure it appears on the page. The JavaScript supplied by the widgets is in charge of the layout. It lays out the grid, loads the text, iframes and any other static content. The event controller widgets in RCAP use the reactive programming paradigm.RCAP is a statistician's convenient web publishing tool for R analytics and visualizations developed within the RCloud environment.


References:


Subramaniam. G, Larchuk. T, Urbanek. S and Archibad. R (2014). iwplot: An R Package for Creating web Based Interactive. In useR! 2014, The R User Conference, (UCLA, USA), Jul. 2014

Woodhull. G, RCloud – Integrating Exploratory Visualization, Analysis and Deployment. In useR! 2014, The R User Conference, (UCLA, USA), Jul. 2014

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

RStudio, Inc, shiny: Easy web applications in R, 2014, URL: http://shiny.rstudio.com

Presentation type: Poster

*Keywords:* R Cloud, R, Dynamic Dashboards, Interactive Visualization

Presentation type: Poster

# The markovchain R package


## Giorgio Alfredo Spedicato

*UnipolSai*



**Abstract:** The markovchain package provides functions for handling Discrete Time Markov Chains (DTMCs) using S4 methods. A general overview of its capabilities will be offered: managind DTMCs classes, performing probabilistic structural analysis, fitting and general inference. Applied examples will be presented in addition.



*Keywords:* Markov Chains; R package; markovchain

Presentation type: Poster

# Logistic modelling of increased antibacterial resistance with sales

**Hannes Gislason, Marita Debess Magnussen, Shahin Gaini & Karl G. Kristinsson**

*University of the Faroe Islands, Department of Science and Technology, Medical Faculty, University of Iceland, Department of Internal Medicine and Research Unit of the Medical Department, National Hospital of the Faroe Islands; Department of Infectious Diseases, Odense University Hospital and University of Southern Denmark, Department of Clinical Microbiology, Landspitali University Hospital*

**Abstract:** Resistant and multi-resistant bacteria are seriously affecting modern health care. Therefore, it is critically important to understand and control the increased resistance. Binomial data with resistant or non-resistant values, where the resistance proportion is bounded between 0 and 1, can be modelled by logistic regression. In its simplest form, the log odds of resistance equals a linear equation with one variable representing sales and two parameters for the intercept and slope, respectively. We used the glm-function with family = quasibinomial to account for larger variances (overdispersion), than expected from the binomial distribution. The R-packages visreg and ggplot2 were used, respectively, to calculate the regression curves with confidence limits, and to visualize the results. Resistance against 7 antibiotics for E. Coli isolates (n = 210) in the Faroe Islands, 2009-2012, was compared with the corresponding antibacterial mean-sales, 2008-2011. A prop.trend.test for trend is clearly significant (p-value < 2.2e-16), while the logistic regression is highly overdispersed ($\approx 32$) indicating low model fit (slope p-value = 0.05). Considering different biological resistance mechanisms, we exclude resistance outliers and extreme sales to minimize overdispersion ($\approx 1$), and find for 5 of 7 resistances the parameter 3.24 for slope (Std. Error = 0.226, p-value = 0.0007), OR [95% CI] = 25.5 [16.6, 40.4].Similarly, including data from Iceland and Denmark, we show how about 7 of 18 antibiotic resistances in the 3 countries closely follow a logistic prediction for increased resistance with sales, while we also detect different mechanisms for the remaining resistances.

*Keywords:* Logistic regression, overdispersion, quasibinomial, antibacterial resistance and sales, E. coli.

Presentation type: Poster

# Writing a dplyr backend to support out-of-memory data for Microsoft R Server


## Hong Ooi

*Microsoft*



**Abstract:** Over the last two years, the dplyr package has become very popular in the R community for the way it streamlines and simplifies many common data manipulation tasks. A feature of dplyr is that it's extensible; by defining new methods, one can make it work with data sources other than those it supports natively. The dplyrXdf package is a backend that extends dplyr functionality to Microsoft R Server's xdf files, which are a way of overcoming R's in-memory limitations. dplyrXdf supports all the major dplyr verbs, pipeline notation, and provides some additional features to make working with xdfs easier. In this talk, I'll share my experiences writing a new back-end for dplyr, and demonstrate how to use dplyr and dplyrXdf to carry out data wrangling tasks on large datasets that exceed the available memory.



*Keywords:* big-data, dplyr, data-munging

Presentation type: Poster

# shinyGEO: a web application for analyzing Gene Expression Omnibus (GEO) datasets using shiny

**Jasmine Dumas, Michael Gargano & Garrett M. Dancik**

*DePaul University, Eastern Connecticut State University, Eastern Connecticut State University*

**Abstract:** Identifying associations between patient gene expression profiles and clinical data provides insight into the biological processes associated with health and disease. The Gene Expression Omnibus (GEO) is a public repository of gene expression and sequence-based datasets, and currently includes > 42,000 datasets with gene expression profiles obtained by microarray. Although GEO has its own analysis tool (GEO2R) for identifying differentially expressed genes, the tool is not designed for advanced data analysis and does not generate publication-ready graphics. In this work, we describe a web-based, easy-to-use tool for biomarker analysis in GEO datasets, called shinyGEO.

shinyGEO is a web-based tool that provides a graphical user interface for users without R programming experience to quickly analyze GEO datasets. The tool is developed using 'shiny', a web application framework for R. Specifically, shinyGEO allows a user to download the expression and clinical data from a GEO dataset, to modify the dataset correcting for spelling and misaligned data frame columns, to select a gene of interest, and to perform a survival or differential expression analysis using the available data. The tool uses the Bioconductor package 'GEOquery' to retrieve the GEO dataset, while survival and differential expression analyses are carried out using the 'survival' and 'stats' packages, respectively. For both analyses, shinyGEO produces publication-ready graphics using 'ggplot2' and generates the corresponding R code to ensure that all analyses are reproducible. We demonstrate the capabilities of the tool by using shinyGEO to identify diagnostic and prognostic biomarkers in cancer.

*Keywords:* Bioinformatics, Gene Expression, Survival Analysis, Shiny

Presentation type: Poster

# Presidential Rankings: Visualization and Comparisons


## Jefferson Davis

*Indiana University*



**Abstract:** Ranking presidents has been a pastime among American historians ever since Arthur Schlesinger's 1948 survey set the prototype. Although these rankings reliably draw news coverage, there are issues to keep in mind. Whose opinions should be surveyed? Liberals and conservatives, political scientists and economists, Americans and non-Americans might all have very different ideas about what makes a good and effective president. Also, how do we combine multiple rankings? Economist Kenneth Arrow showed in 1951 that there is no valid way to amalgamate multiple rankings into a single one. What to do? Visualization methods provide a way to tackle these problems. We begin by computing distances between individual rankers (the Kendall distance.) Then, with multidimensional scaling, we use these distances to plot the rankers and see what patterns develop. This gives a way to see the answers to such questions as

- How much difference actually is there between different rankers?
- Do political scientists and economists differ much in their rankings?
- Do rankings change over time?

These methods, of course, are used to visualize many different sorts of ranked data—from restaurant reviews to Hollywood movies. With this in mind, we use an R dataframe of movies with American presidents as characters to answer one last question:

- If we left presidential rankings up to Hollywood, rather than political scientists, would anything change?



*Keywords:* Visualization, rank orders, American history

Presentation type: Poster

# Data Quality Profiling - The First Step with New Data


## Jim Porzak

*DS4CI.org*



**Abstract:** The first step, when getting a new data set, is to take a look at the data for completeness, accuracy, and reasonableness. This talk will describe a method based on Jack Olson's Data Quality - The Accuracy Dimension. The input data set can be either a raw text or spreadsheet file or from a source with columnar meta-data like a SQL table or an R data frame. The only setup is to connect to the data source. Using RMarkdown, dplyr, grid, and ggplot2 we produce a report where each column is profiled by data types, summary statistics (if numeric or date), distribution plot, counts, and the head and tail values. This facilitates a quick visual scan of each column for data quality issues. The simple visual format also aids communication with the data provider to dig into quality issues and, hopefully, clean up the data set before wasting time and effort on an analysis flawed by bad data. We provide examples both good and suspect columns.



*Keywords:* data quality, data profiling

Presentation type: Poster

# Teaching statistics to medical students with R and OpenCPU

**Jörn Pons-Kühnemann & Anita Windhorst**

*Institute for Medical Informatics, Justus Liebig University, Giessen, Germany*

**Abstract:** In general medical students do not have or aim at a deeper understanding of statistics. Nevertheless some knowledge of basic statistical reasoning and methodology is indispensable to apprehend the meaning of results of scientific studies published in medical journals. Also, some familiarity with the correct interpretation of probability statements concerning medical tests is crucial for physicians.

In order to supplement our regular statistics classes at the medical faculty we started to develop an online system providing a pool of assignments. Each student gets an individual assignment with a modified data set, asking therefore for a slightly different solution. This enables the system to verify the student's personal achievement and a data base may keep record of his/her performance.

Our system utilizes OpenCPU installed on a Linux server. The front-end is developed with HTML and JavaScript, while the back-end involves R and MySQL.

The state of the development, the problems, and the students response will be presented.

*Keywords:* Statistics education, Online training, R, MySQL, OpenCPU

Presentation type: Poster

# Developing R Tools for Energy Data Analysis


## Kara Downey, Seth Wayland & Kai Zhou

*Opinion Dynamics Corporation*



**Abstract:** Energy efficiency program evaluators use data from a variety of sources, which range from utility billing databases to surveys to logs from smart thermostats. We estimate reductions in energy usage attributable to various energy savings programs. Our evaluations are used to certify that utilities are meeting state or federally mandated efficiency standards; to reimburse utilities for funds spent on what amounts to reducing demand for power; or to assess whether programs are worth continuing or expanding. It is therefore critical that our estimates be reliable and reproducible. The raw data we receive is usually very messy, and the types of problems we find in the data are sufficiently niche that off-the-shelf data cleaning software is of limited help. Our poster demonstrates a suite of custom R packages and Shiny apps that we've developed to streamline the process of prepping energy use data for analysis. Our current tools include:

- A package, noaaisd, which pulls appropriate hourly weather data from the NOAA website and appends it to geocoded customer usage data
- A Shiny app that allows users to quickly explore individual smart thermostat logs and save information about the quality of each log's data

Features under development include:

- A package to automate the cleaning of utility billing data (this requires, among other things, the ability to detect and appropriately correct gaps or overlaps in billing periods for individual customers, and the ability to flag abnormal billing periods or energy consumption)
- A Shiny app to help build baseline energy usage models



*Keywords:* energy, energy efficiency, data preparation, Shiny, package development

Presentation type: Poster

# ROSETTAHUB, the next generation data science platform


## Karim Chine

*ROSETTAHUB LTD*



**Abstract:** Fragmentation in the data science space reduces the productivity of data scientists and compromises their ability to share, collaborate and make their results reproducible. Frictions are partly due to the large diversity of tools and environments they use, to their complex dependencies and to the difficulty of interconnecting them. RosettaHUB is an innovative platform that significantly reduces those frictions and offers data scientists a streamlined experience in their day-to-day interaction with tools, infrastructures, data and peers. R, Python, Julia, SQL, Scala, Spark, ParaView etc. can be used simultaneously within RosettaHUB which acts as a one stop shop exposing web based and user friendly clouds management consoles, workbenches and notebooks. It offers a Google-docs like experience to data scientists and lets them use the different tools from anywhere and collaborate in real-time. It makes it possible to create and share resources on any infrastructure including EC2, GCE and private clouds. It keeps track of all interactions with the environment and allows the reproduction of all the created artifacts. Within RosettaHUB, R, Python, Julia, Spark etc. share the same workspace and operate in the same memory, they can call each other seamlessly, share all their variables and access advanced collaborative visualization functionalities and web spreadsheets, they can be used to create and share cross-language interactive and collaborative web applications and services. RosettaHUB is fully programmable and all its capabilities are accessible from R (RosettaHUB package), from Python (RosettaHUB module), from Excel and Word (RosettaHUB add-in), from Java, C# and Node.js (SDKs).



*Keywords:* Cloud computing, Workbench, Notebook, Python, Spark, Spreadsheet, EXCEL, Word, EC2, GCE, Docker, Interactive Web Applications,Collaboration, reproducibility, cross-language, SDKs, Web Services

Presentation type: Poster

# Making Shiny Seaworthy: A weighted smoothing model for validating oceanographic data at sea.

**Kevin W. Byron & Mathew L. Nelson**

*City of San Diego*

**Abstract:** The City of San Diego conducts one of the largest ocean monitoring programs in the world, covering 340 square miles of coastal waters and sampling at sea 150 days each year. Water quality monitoring is a cornerstone of the program and requires the use of sophisticated instrumentation to measure a suite of oceanographic parameters (e.g., temperature, depth, salinity, dissolved oxygen, pH). The various sensors or probes can be episodically temperamental, and oceanographic data can be inherently non-linear, especially within stratifications (i.e., where the water properties change rapidly with small changes in depth). This makes it difficult to distinguish between extreme observations due to natural events (anomalous data) and those due to instrumentation error (erroneous data), thus, requiring manual data validation at sea.

This Shiny app improves the manual validation process by providing a smoothing model to flag erroneous data points while including anomalous data. Standard smoothing models were unable to model stratification without including erroneous data, so we elected to use a custom weighted average model where observations with a greater deviation from the local mean have less weight.

We coupled this model with an interactive Shiny session using ggplot2 and R Portable to create an offline web application for use at sea. This Shiny app takes in a raw data file, presents a series of interactive graphs for removing/restoring potentially erroneous data, and exports a new data file. Additional customization of the Shiny interface using the shinyBS package, Javascript, and HTML improve the user experience.

*Keywords:* Shiny, Modeling, Smoothing Functions, Oceanography, Marine Biology

Presentation type: Poster

# mvarVis: An R package for Visualization of Multivariate Analysis Results


## Kris Sankaran & Lan Huong Nguyen

*Department of Statistics, Stanford University, Institute for Computational and Mathematical Engineering, Stanford University*



**Abstract:** mvarVis is an R package for visualization of diverse multivariate analysis methods. We implement two new tools to facilitate analysis that are cumbersome with existing software. The first uses htmlwidgets and d3 to create interactive ordination plots; the second makes it easy to bootstrap multivariate methods and align the resulting scores. The interactive visualizations offer an alternative to printing multiple plots with different supplementary information overlaid, and bootstrapping enables a qualitative assessment of the uncertainty underlying the application of exploratory multivariate methods on particular data sets.

Our approach is to leverage existing packages – FactoMineR, ade4, and vegan – to perform the actual dimension reduction, and build a new layer for visualizing and bootstrapping their results. This allows our tools to wrap a variety of existing methods, including one table, multitable, and distance-based approaches – principal components, multiple factor analysis, and multidimensional scaling, for example. Since our package uses htmlwidgets, it is possible to embed our interactive plots in Rmarkdown pages and Shiny apps. All code and many examples are available on our github.



*Keywords:* multivariate analysis, visualization, bootstrap, htmlwidgets

Presentation type: Poster

# Time Flies - Use R to Analyze the Changing Airline Industry


## Longyi Bi

*Carlson Wagonlit Travel*



**Abstract:** A project to use historical flight data or flight schedule data to uncover airline industry evolution and competitive landscape changes at three levels: airport, carrier and route. Different packages and techniques are used to manage and analyze the data and to visualize the trends and to highlight facts; a shiny dashboard app is developed to allow interactive slice and dice queries. This is an example to use R to swiftly process large datasets, to draw unusual insights, and to build product prototypes for the travel industry.



*Keywords:* travel, airline industry, visualization, dashboard, business application

Presentation type: Poster

# MethylMix 2.0: a bivariate Gaussian mixture model for identifying methylation driven genes

**Marcos Prunello, Olivier Gevaert**

*Biomedical Informatics Research, Stanford University*

**Abstract:** DNA methylation is a key mechanism that regulates gene transcription and its importance in carcinogenesis has been widely explored. Both hypo and hyper-methylated genes can deregulate gene expression and variations in methylation are associated with several diseases. We previously proposed a method called MethylMix, available as an R package at Bioconductor, which aims to derive key methylation-driven genes in cancer which have an effect on gene expression. MethylMix consists of a three-step algorithm: it first identifies subgroups of samples with common methylation patterns, then compares each subgroup to normal tissue samples to define hypo- or hyper-methylation states, and finally reports a gene as being methylation-driven if it is transcriptionally predictive defined as a significant negative linear association between methylation level and gene expression. Now we present MethylMix 2.0, introducing some extensions to the original method. MethylMix 2.0 uses a bivariate Gaussian mixture model to jointly model DNA methylation and gene expression in order to identify the different subgroups of cancer samples. Also, while MethylMix focuses only on genes with a strong negative linear relationship between DNA methylation and gene expression, MethylMix 2.0 extends the search for driver genes to include genes which show other types of association. We applied MethylMix 2.0 on six large cancer cohorts and show that MethylMix 2.0 identifies known and new hypo- and hyper-methylated genes and can also be used to identify samples subtypes. MethylMix 2.0 is intended to be submitted soon as an R package to Bioconductor

*Keywords:* DNA methylation, gene expression, differential genes, mixture models

Presentation type: Poster

# Data Analysis Pipeline for the Molecular Diagnosis of Brain Tumors


## Martin Sill, Volker Hovestadt, Daniel Schrimpf, David Jones, David Capper, Stefan Pfister & Andreas von Deimling

*Division of Biostatistics, German Cancer Research Center (DKFZ), Division of Molecular Genetics, German Cancer Research Center (DKFZ), Department of Neuropathology, University Hospital Heidelberg; CCU Neuropathology, German Cancer Research Center (DKFZ), Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), Department of Neuropathology, University Hospital Heidelberg; CCU Neuropathology, German Cancer Research Center (DKFZ), Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), Department of Neuropathology, University Hospital Heidelberg; CCU Neuropathology, German Cancer Research Center (DKFZ)*



**Abstract:** More than 100 brain tumor entities are listed in the World Health Organization (WHO) classification. Most of these are defined by morphological and immunohistochemical criteria that may be ambiguous if the tissue material is of poor quality. This can make a histological diagnosis challenging, even for skilled neuropathologists. Molecular high-throughput technologies that can complement standard histological diagnostics have the potential to greatly enhance diagnostic accuracy. Genome-wide DNA methylation, acting as a 'fingerprint' of cellular origin, is one such promising technology for tumor classification. nWe have collected brain tumor DNA methylation profiles of almost 3,000 cases using the Illumina HumanMethylation450 (450k) array, covering over 90 brain tumor entities. Using this data, we trained a Random Forest classifier which predicts brain tumor entities of diagnostic cases with high accuracy. 450k methylation data can also be used to generate genome-wide copy number profiles and predict target gene methylation. Based on several R and Bioconductor packages, we have established a data analysis pipeline which takes 450k methylation data as input and automatically generates diagnostic reports containing quality control metrics, tumor class probabilities, copy number profiles and target gene methylation status. Besides sharing R packages which contain parts of the pipeline with cooperating institutes, we also offer a web interface that allows researchers from other institutes to apply the pipeline to their own data.nPractical experience from different cooperating institutes show that application of our pipeline to 450k methylation array data represents a cost efficient method to greatly improve diagnostic accuracy and clinical decision making.



*Keywords:* machine learning, bioinformatics, methylation data, brain tumor diagnosis

Presentation type: Poster

# Giving a boost to renewable energy development: predicting reef fish community distributions in the Main Hawaiian Islands using boosted regression trees

**Matthew Poti, Kostantinos Stamoulis, Jade Delevaux, Mary Donovan, Alan Friedlander, Matthew Kendall, Bryan Costa & Arliss Winship**

*CSS-Dynamac; NOAA National Centers for Coastal Ocean Science, University of Hawaii at Manoa Fisheries Ecology Research Lab, University of Hawaii at Manoa Department of Natural Resources and Environmental Management, University of Hawaii at Manoa Fisheries Ecology Research Lab, University of Hawaii at Manoa Fisheries Ecology Research Lab, NOAA National Centers for Coastal Ocean Science, NOAA National Centers for Coastal Ocean Science, CSS-Dynamac; NOAA National Centers for Coastal Ocean Science*

**Abstract:** Spatially explicit information describing reef fish communities is critical for effective management of coral reef ecosystems. However, geographic coverage of in situ data is often limited. To overcome this information gap, statistical modeling can be used to make predictions across space by relating sample data to predictor variables describing the associated environment. As part of a marine biogeographic assessment to inform the Bureau of Ocean Energy Management's renewable energy policy decisions in Hawaii, spatial predictions of several reef fish community metrics were generated from visual survey data compiled by University of Hawaii's Fisheries Ecology Research Lab and environmental predictors representing seafloor topography, benthic habitats, geography, and oceanography. Boosted regression trees, an ensemble approach combining machine learning with tree-based statistical models, were fit to the data in R using the 'dismo' package. Model parameters were tuned by fitting models for a range of learning rate, tree complexity, and bag fraction values, and identifying for each combination of values the number of boosting trees that minimized predictive deviance. Predictors that contributed least to model performance were eliminated using a recursive model simplification procedure. Non-parametric bootstrapping, in which the survey data were randomly resampled and a model was fit on each sample, was then used to create an ensemble of spatial predictions across the study area. The coefficient of variation was calculated to visualize the spatial precision of model predictions. Model performance was evaluated by calculating the percent deviance explained by the model when evaluated on data withheld from model fitting.

*Keywords:* coral reef fish, predictive modeling, boosted regression trees, machine learning

Presentation type: Poster

# Approaches to R education in Canadian universities

**Michael A. Carson & Nathan Basiliko**

*Laurentian University*

**Abstract:** The R language is a powerful tool used in a wide array of research disciplines and owes a large amount of its success to its open source and adaptable nature. This has caused rapid growth of formal and informal online and text resources that is beginning to present challenges to novices learning R. Students are often first exposed to R in upper division undergraduate classes or during their graduate studies. The way R is presented has consequences for the fundamental understanding of the program and language itself. That is to say there is a dramatic difference in user comprehension of R if learning it as a tool to do an analysis opposed to learning another subject (e.g. statistics) using R. While some universities do offer courses specific to R it is more commonly incorporated into a pre-existing course or a student is left to learn the program on his or her own. To better establish how students are exposed to R, an understanding of the approaches to R education is critical. In this survey we evaluated the current use of R in Canadian university courses to determine what methods are most common for presenting R. While data are still being collected we anticipate that courses using R to teach another concept will be much more common than courses dedicated to R itself. This information will influence how experienced educators as well as programmers approach R, specifically when developing educational and supplemental content in online, text, and package specific formats.

*Keywords:* Education, University Use, Adaptive Learning, Supplemental Content

Presentation type: Poster

# Energy prediction and load shaping for buildings


## Michael Anthony Wise

*Microsoft Corporation*



**Abstract:** Energy costs for Microsoft's 120 building main-campus are very high, particularly because of the almost exclusive usage of electric heating there. About 10% of these are demand charges (a peak-usage surcharge), and become very pronounced in the winter. To reduce these, we have modeled building energy consumption to predict demand peaks using random forest and boosted trees regression as implemented in the randomForest and gbm packages (sometimes together with caret) and then piloted in our operations center.

Now in a second phase more advanced models were developed to allow this peak-flattening without manual intervention. Transitioning a predictive-model to a command-and-control model like this was complex, and capturing the physical reality required the use of multiple cascaded models, also using tree-based regression techniques. Optimization (to find the best control parameters) and simulation (to gauges the overall impact of intervention) were used and the problems typical for dynamical systems (stabilization, non-convergence, etc.) had to be overcome; these will be addressed in the talks.

All of the development and modelling work was done in R and Shiny using RStudio and later RTVS, afterwards the R-code and ggplot2 plots were deployed to various platforms including Azure ML, PowerBI and R Services for SQL Server.



*Keywords:* Machine Learning, Control Theory, Smart Buildings

Presentation type: Poster

# Encounters of the Chinook kind: visualizing fish movement with R


## Myfanwy E. Johnston

*UC Davis*



**Abstract:** Biotelemetry (the electronic tagging and tracking of organisms) is notorious for producing large data sets. Finding ways to visualize these data sets is an ongoing effort in ecology; ideally we want to convey not only summary statistics (movement rates, turning angles, etc), but also insight into animal behavior itself. The movements and migrations of anadromous fish, which are often constrained within the linear boundaries of a river system or delta and only monitored intermittently with acoustics, present a special challenge to effective visualization. Excellent individual examples of visualization of fish migration can be found piecemeal in the literature and 'in the wild' (read: on the internet), but no central repository or reference volume currently exists on the topic. Thus, fisheries ecologists and behaviorists often find themselves having to reinvent the wheel when it comes to data visualization of their acoustic datasets, or worse, never visualize them at all. In hopes of providing a common starting point for these researchers and all other interested parties, this poster collects and presents examples of visualization of fish movement and migration using tools in R, including applications with ggplot2, googleVis, leaflet, and Shiny, among others. Code and reproducible examples of all visualization workflows and datasets are available in the github repository: http://github.com/Myfanwy/SeeFishMove.



*Keywords:* visualization, fish, migration, movement, ecology, behavior

Presentation type: Poster

# Bridging the Data Visualization to Digital Humanities gap: Introducing the Interactive Text Mining Suite

**Olga Scrivner & Jefferson Davis**

*Indiana University*

**Abstract:** In recent years, there has been growing interest in data visualization for text analysis. While text mining and visualization tools have been successfully integrated into research methods in many fields, their use still remains infrequent in mainstream Digital Humanities. Many tools require extensive programming skills, which can be a roadblock for some literary scholars. Furthermore, while some visualization tools provide graphical user interfaces, many humanities researchers desire more interactive and user-friendly control of their data. In this talk we introduce the Interactive Text Mining Suite (ITMS), an application designed to facilitate visual exploration of digital collections. ITMS provides a dynamic interface for performing topic modeling, cluster detection, and frequency analysis. With this application, users gain control over model selection, text segmentation as well as graphical representation. Given the considerable variation in literary genres, we have also designed our graphical user interface to reflect choice of studies: scholarly articles, literary genre, and sociolinguistic studies. For documents with metadata we include tools to extract the metadata for further analysis. Development with the Shiny web framework provides a set of clean user interfaces, hopefully freeing researchers from the limitations of memory or platform dependency.

*Keywords:* data visualization, shiny application, digital humanities

Institute of Information Theory and Automation

Presentation type: Poster

# Multiple-Output Quantile Regression in R


## Pavel Boček & Miroslav Šiman

*Institute of Information Theory and Automation*



**Abstract:** The presentation introduces two recent multiple-output quantile regression methods, generalizing quantile regression to the case of multivariate responses, and shows how they could be performed in R. The directional multiple-output quantile regression can be employed thanks to the presented new R package modQR. The elliptical multiple-output quantile regression can be transformed to a convex optimization problem and then solved with the aid of the solvers for semidefinite programming already available in R.



*Keywords:* multivariate quantile, quantile regression, multiple-output regression

Presentation type: Poster

# ALZCan: Predicting Future Onset of Alzheimer's Using Gender, Genetics, Cognitive Tests, CSF Biomarkers, and Resting State fMRI Brain Imaging.


## Pravin Ravishanker

*Bellarmine College Preparatory, San Jose, CA*



**Abstract:** Due to a lack of preventive methods and precise diagnostic tests, only 45% of Alzheimer's patients are told about their diagnosis. I hypothesized that one can create an accurate diagnostic/prognostic software tool for early detection of Alzheimer's using functional connectivity in resting-state fMRI brain imaging, genetic SNP data, cerebrospinal fluid (CSF) concentrations, demographic information, and psychometric tests.

Using R programming language and data from ADNI, an ongoing, longitudinal, global effort tracking clinical/imaging AD biomarkers, I examined 678 4D fMRI scans and 56847 observations of 1722 individuals across three diagnostic groups. ICA on fMRI scans yielded graph structures of connectivity between brain networks. For diagnosis, 4 support vector machines and 6 gradient boosting machines were trained 10 times each for fMRI, genetic, CSF biomarker, and cognitive data. For prognosis, 3 linear regression models predicted cognitive scores 6 to 60 months into the future. Forecasted cognitive scores and demographic information were used for prognosis.

ALZCan had 81.82% diagnostic accuracy. Prognostic accuracy for 6, 12, 18 months in future was 75.4%, 68.3%, 68.6%. AD patients showed significantly lower transitivity and average path length between functional brain networks. I examined relative influence/predictive power of multiple biomarkers, confirming previous findings that gender has higher influence than genetic factors on AD diagnosis. Overall, this study engineered a novel neuroimaging feature selection method by using machine learning and graph-theoretic functional network connectivity properties for diagnosis/prognosis of disease states. This analytical tool is capable of predicting future onset of Alzheimer's and Mild Cognitive Impairment with significant accuracy.



*Keywords:* Alzheimer's, Neuroinformatics, Machine Learning, Multi Voxel Pattern Analyses, Functional Network Connectivity

Presentation type: Poster

# Social Vulnerability to Climate Change: Automation and Validation of Indices

**Ram Barankin, Robert E. Bowen & Paul Kirshen**

*UMass Boston*

**Abstract:** Coastal areas all over the world are experiencing the effects of climate change. In particular, sea level rise, high storm surges, and intensive precipitation, are causing floods in these areas, resulting in excessive damage. Various academic studies and governmental projects have been studying the vulnerability of different systems, such as social systems, infrastructure systems, and ecosystems, to climate change effects. In particular, social vulnerability studies often use different social attributes (e.g., income and age), usually obtained from the US Census Bureau. Those can indicate whether a certain community is vulnerable to climate change effects. Hence, these attributes are considered to be indicators of social vulnerability and aggregating them into one value produces a social vulnerability index (SVI) that varies across geographical units. New packages in R (such as ACS), dramatically improve the efficiency of acquiring large capacity of data. Consequently, the analysis can be done for large scales and in fine resolutions. This feature, along with the use of dimensionality reduction methods (such as, Principal Component Analysis), and when combined with the use of table manipulation functions (such as within Data Table package), allow the automation of SVI's construction in R. Consequently, the researcher can define the geography of interest, and quickly produce the relevant SVI. The current study uses the described algorithm to learn about social vulnerability to extreme flooding. In addition it uses statistical predictive models (such as regression and Structural Equation Modeling) to learn about the validity of SVIs and the weights (importance) of the various indicators.

*Keywords:* Climate change; Social vulnerability; Census data; ACS; Dimensionality reduction methods; Predictive statistical model; Social science

Presentation type: Poster

# The Use of Ensemble Learning Methods in Open Source Data Challenges


## Rebecca Z. Krouse & Agustin Calatroni

*Rho Inc., Federal Systems Division, Chapel Hill, NC*



**Abstract:** As data collection grows in size and complexity across a variety of industries, open source data challenges are becoming more widespread. We present our experience developing prediction models within the context of data challenges. With the goal of maximizing predictive performance, we explored ensemble learning methods to train our models. We demonstrate the use of these methods using R packages such as h2o and h2oEnsemble and cloud computing platforms. In order to obtain an approximation of our predictive ability prior to challenge submission, we developed wrapper code to perform cross validation on the H2O ensembles. We display our process for determining the expected level of performance of the trained model on external data sources.



References:

Spencer Aiello, Tom Kraljevic, Petr Maj and with contributions from the H2O.ai team (2015). h2o: R Interface for H2O. R package version 3.8.1.3. https://CRAN.R-project.org/package=h2o

Erin LeDell (2016). h2oEnsemble: H2O Ensemble Learning. R package version 0.1.6. https://github.com/h2oai/h2o-3/tree/master/h2o-r/ensemble/h2oEnsemble-package



*Keywords:* ensemble learning, machine learning, data challenges, cross validation, h2o, h2oEnsemble, cloud computing

Presentation type: Poster

# R Microplots in Tables with the latex() Function


## Richard M Heiberger

*Temple University, Department of Statistics, Fox School of Business*



**Abstract:** Microplots are often used within cells of a tabular array. We describe several simple R functions that simplify the use of microplots within LaTeX documents constructed within R. These functions are coordinated with the latex() function in the Hmisc package or the xtable function in the xtable package. We show examples using base graphics, and three graphics systems based on grid: lattice graphics, gg2plot graphcs, and vcd graphics. These functions work smoothly with standalone LaTeX documents and with Sweave, with knitr, with org mode and with Rmarkdown.



*Keywords:* microplots, tables, R, latex

Presentation type: Poster

# Helping Non-programmers Use R


## Robert Anthony Muenchen

*University of Tennessee*



**Abstract:** Unfortunately, many researchers who could benefit from the use of R will never be good programmers. Graphical user interfaces such as R Commander, Rattle, and Deducer allow non-programmers to use R with minimal training. However, those tools store what they do in the form of R programs that their users might not understand. This makes re-use and repeatable research challenging for them. The free and open source KNIME software addresses this problem through the use of a workflow (flowchart) style of user interface. Users can simply choose from a list of common models to use. If a model is needed that does not come with the software, an R programming icon can be used in which any R code can be typed. Such custom nodes can be easily shared by other users. This approach allows for analyses that can range from quite simple to extremely complex, even blending in other programming languages such as Python or MATLAB. For handling big data, work can easily be passed to a Hadoop Map Reduce or Apache Spark system. This presentation will demonstrate how R integrates into KNIME.



*Keywords:* Graphical User Interface, Workflow, Flowchart, KNIME

Presentation type: Poster

# Bayesian inference for Internet ratings data using R


## Ruby Chiu-Hsing Weng

*Department of Statistics, National Chengchi Univ.*



**Abstract:** Internet ratings data are usually ordinal measurements nfrom 1 to 5 (or 10) rated by Internet users on the quality nof all kinds of items. The traditional graphical displays of the ratings data does not account for the inter-rater difference. Some model-based methods with MCMC approach have beennsuggested to address this problem. In the present work we propose a real-time Bayesian inferencenalgorithm for parameter estimation. Two real data sets and the R implementation of the abovenmentioned algorithm will be presented



*Keywords:* Bayesian inference; Internet ratings data; online algorithm

Presentation type: Poster

# Using R in the evaluation of psychological tests

**Rudolf Debelak, Johanna Egle, Lena Köstering & Christoph P. Kaller**

*University of Zurich, Schuhfried GmbH, Universitätsklinikum Freiburg, Universitätsklinikum Freiburg*

**Abstract:** Psychological tests are used in many fields, including medicine and education, to assess the cognitive abilities of test takers. According to international standards for psychological testing, psychological tests are required to be reliable, fair, and valid. This presentation illustrates how R can be used to assess the reliability, fairness, and validity of psychological tests using the Tower of London task as an example. In clinical neuropsychology, the Tower of London task is widely used to assess a person's planning ability. Our data consist of 798 respondents who worked on the 24 test items of the Tower of London – Freiburg Version. By employing the framework of factor analysis and item response theory, it is demonstrated that the number of correctly solved problems in this test can be considered as a reliable and sound indicator for the planning ability of the test takers. It is further demonstrated that the individual problem difficulties remain stable across different levels of age, sex and education, which provides evidence for the test's fairness. All computations were carried out with the R packages psych, lavaan and eRm, all of which are freely available on CRAN.

*Keywords:* Item Response Theory, Factor Analysis, Psychological Assessment

Presentation type: Poster

# 'IMGTStatClonotype': An R package with integrated web tool for pairwise evaluation and visualization of IMGT clonotype diversity and expression from IMGT/HighV-QUEST output

**Safa Aouinti, Dhafer Malouche, Véronique Giudicelli, Patrice Duroux, Sofia Kossida & Marie-Paule Lefranc**

*IMGT®, the international ImMunoGeneTics information system®, Montpellier University; Unité Modélisation et Analyse Statistique et Economique, IMGT®, IMGT®, IMGT®, IMGT®*

**Abstract:** The adaptive immune response is our ability to produce up to $2.10^{12}$ different immunoglobulins (IG) or antibodies and T cell receptors (TR) per individual to fight pathogens. IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org), was created in 1989 by Marie-Paule Lefranc (Montpellier University and CNRS) to manage the huge and complex diversity of these antigen receptors and is at the origin of immunoinformatics, a science at the interface between immunogenetics and bioinformatics. Next generation sequencing (NGS) generates millions of IG and TR nucleotide sequences, and there is a need for standardized analysis and statistical procedures in order to compare immune repertoires. IMGT/HighV-QUEST is the unique web portal for the analysis of IG and TR high throughput sequences. Its standardized statistical outputs include the characterization and comparison of the clonotype diversity in up to one million sequences. IMGT® has recently defined a procedure for evaluating statistical significance of pairwise comparisons between differences in proportions of IMGT clonotype diversity and expression, per gene of a given IG or TR V, D or J group. The procedure is generic and suitable for detecting significant changes in IG and TR immunoprofiles in protective (vaccination, cancers and infections) or pathogenic (autoimmunity and lymphoproliferative disorders) immune responses. In this talk, I will present the new R package ('IMGTStatClonotype') which incorporates the IMGT/StatClonotype tool developed by IMGT® to perform pairwise comparisons of sets from IMGT/HighV-QUEST output through a user-friendly web interface in users' own browser.

*Keywords:* Statistical significance, differences in proportions, big data, multiple hypothesis testing, IMGT, IMGT/HighV-QUEST, immunoglobulin, antibody, T cell receptor, next generation sequencing (NGS)

Presentation type: Poster

# Visualizations and Machine Learning in R with Tessera and Shiny

**Sarah M. Reehl, Allison M. Thompson & Lisa M. Bramer**

*Pacific Northwest National Lab*

**Abstract:** In a divide and recombine (D&R) paradigm, the Tessera tool suite of packages (https://tessera.io), developed at Pacific Northwest National Laboratory, presents a method for dynamic and flexible exploratory data analysis and visualization. At the front end of Tessera, analysts program in the R programming language, while the back end utilizes a distributed parallel computational environment. Using these tools, we have created an interactive display where users can explore visualizations and statistics on a large dataset from the National Football League (NFL). These visualizations allow any user to interact with the data in meaningful ways, leading to an in depth analysis of the data through general summary statistics as well as insights on fine grain information. In addition, we have incorporated an unsupervised machine learning scheme utilizing an interactive R Shiny application that predicts positional rankings for NFL players. We have showcased these tools using a variety of available data from the NFL in order to make the displays easily interpretable to a wide audience. Our results, fused into an interactive display, illustrate Tessera's efficient exploratory data analysis capabilities and provide examples of the straightforward programming interface.

*Keywords:* Visualizations, Tessera, Shiny, Big Data, Machine Learning

Presentation type: Poster

# request: A DSL for http requests


## Scott Alan Chamberlain

*rOpenSci/UC Berkeley*



**Abstract:** Making http requests within R has gotten much easier with httr, jsonlite, and other tools. However, especially for consumers, the experience could be simpler. With inspiration from Python's httpie, the request R package makes a variety of assumptions that will fit most use cases: data will be in json format, you want to make a GET request, and you probably want back data.frame's. request allows non-standard evaluation to easily build up requests, but doesn't make you use it. request handles paging automatically, includes a retry helper, and more.



*Keywords:* http, api, web, dsl

Presentation type: Poster

# Curde: Analytical curves detection


## Simon Gajzler, Simon Gajzler & Lukas Streda

*CTU*



**Abstract:** The main aim of our work is to develop the new R package curde. The package is used to detect line or conic curves in a digital image. The package contains the Hough transformation for a line detection using the accumulator. The Hough transform is a feature extraction technique and its purpose is to find imperfect instances of objects within a certain class of shapes. This technique is not suitable for curves with more than three parameters. For conic fitting, robust regression is used. For noisy data, solution based on Least Median of Squares (LMedS) is highly recommended. In this package, algorithms for non-user image evaluation is implemented. The whole process of the non-user image evaluation includes the image preparation. The preparation consists of various methods such as image grayscaling, thresholding or histogram estimation. The conversion from the grayscaled image to binary is realised by the calculation of the Sobel operator convolution and by the application of the threshold technique. After that the convolution technique is applied. The new R package curde will be the integration of all previous techniques to the one complex package.



*Keywords:* Computer vision, Sobel operator, Convolution, Grayscale, Hough line detection, Robust regression

Presentation type: Poster

# Multi-stage Decision Method To Generate Rules For Student Retention


## Soma Datta & Susan Mengel

*University Of Houston Clear Lake, Texas Tech University*



**Abstract:** The retention of college students is an important problem that may be analyzed by computing techniques, such as data mining, to identify students who may be at risk of dropping out. The importance of the problem has grown due to institutions' requirement of meeting legislative retention mandates, face budget shortfalls due to decreased tuition or state-based revenue, and fall short of producing enough graduates in fields of need, such as computing. While data mining techniques were applied with some success, this article aims to show how R can be used to develop a hybrid methodology to enable rules to be created for the minority class with coverage and accuracy range which were not available as per existing literature. A multiple stage decision methodology (MSDM) used data mining techniques for extracting rules from an institution's student data set to enable administrators to identify at risk students. The data mining techniques included partial decisions trees, K-means clustering, and Apriori association mining to be implemented in R. MSDM was able to identify students with up to 89% accuracy on student datasets, where the number of at risk students was fewer than the retained students that made the at risk model difficult to build. The motivation for using R was twofold. First, to generate rules for minority class, and second, use R to make it reproducible.



*Keywords:* student retention, rule generation, data mining, computing, R, recursive partition, Apriori, k-mean clustering, reproducible research

Presentation type: Poster

# Rapid development of shiny apps for in-house data mining in biological data


## Stefan Reuscher

*Nagoya University*



**Abstract:** The advances in DNA-sequencing technology led to a vast amount of datasets from diverse biological sources. Making those complex datasets accessible for data mining to non-bioinformaticians in a timely manner is still challenging. In a typical scenario unprocessed data is generated, pre-processed and distributed by a third party (e.g. a DNA-sequencing center) before biological questions are addressed. While R offers excellent capabilities to do that, its steep learning curve makes it rather unattractive for pure biologists, who "fear the command line" and prefer graphical user interfaces.

Using shiny apps we present examples of how to make biological data (gene expression datasets) and their associated analyses methods available to biologists. Data mining in gene expression datasets can involve techniques like dimensional reduction, clustering, testing for significant differences or term enrichment analyses. In addition to conducting those analyses, data visualization is necessary to quickly evaluate results. Those tasks normally require at least a working knowledge of R and some commonly used packages.

The combination of shiny apps with common analyses packages allows the rapid deployment of applications tailored to a specific dataset that allows effective data mining with a graphical user interface. By including raw data and pre-calculated results objects shiny apps can be used as an efficient, self-contained way of data distribution and analysis.



*Keywords:* bioinformatics, data mining, data visualization, transcriptomics

Presentation type: Poster

# Sequence Analysis with Package TraMineR


## Teck Kiang Tan

*Institute for Adult Learning, Singapore*



**Abstract:** Sequence analysis started in biological science to examine pattern of protein DNA and subsequently applied in social sciences to study the pattern of sequences from individual's life course. Many social science studies concerned with time series are recorded in sequences. Past studies using sequence analysis include footsteps of dances, class careers, employment biographies, family histories, school-to-work transitions, occupational career pattern, and other life-course trajectories.

The TraMineR is a package specially designed for carrying out sequence analysis for the social sciences (Gabadinho, Studer, Muller, Buergin, & Ritschard, 2015). It is a data mining tool that is most appropriate to mine and group social sequence data. It contains toolbox for the manipulation, description and rendering of sequences and functions to produce graphical output to describe state sequences, categorical sequences, sequence visualization, and sequence complexity. It also offers functions for computing distances between sequences with different metrics, which includes optimal matching, longest common prefix and longest common subsequence. In combination with cluster analysis and multidimensional scaling, typology can be formed to understand the life-course trajectories by grouping the sequences into groups.

I will briefly outline the key functionalities of TraMineR and demonstrate the procedure for carrying out social sequence analysis with real life examples to highlight the usefulness of the TraMineR package. Other R packages related to sequence analysis will also be covered during the session.



*Keywords:* Sequence analysis, Package TraMineR, Life Course Studies

Presentation type: Poster

# R Shiny Application for the Evaluation of Surrogacy in Clinical Trials

**Theophile Bigirumurame, Ziv Shkedy, Geert Molenberghs, Marc Buyse, Tomasz Burzykowski & Wim Van del Elst**

*Hasselt University, Hasselt University, Hasselt University; International Drug Development Institute (IDDI), Hasselt University, Hasselt University*

**Abstract:** In clinical trials, the determination of the true endpoint or the effect of a new therapy on the true endpoint may be difficult, requiring an expensive, invasive or uncomfortable procedure. Furthermore, in some trials the primary endpoint of interest (the "true endpoint"), for example death, is rare and/or takes a long period of time to reach. In such trials, there would be benefit in finding a more proximate endpoint (the "surrogate endpoint") to determine more quickly the effect of an intervention.

We present a new R Shiny application for the evaluation of surrogate endpoints in randomized clinical trials using patients data. The Shiny application for surrogacy consists of a set of friendly user function which allow the evaluation of different types of endpoints (i.e., continuous, categorical, binary, survival endpoints) and produce a unified and interoperable output. With this new Shiny App, the user does not need to have the R software installed on his computer. It is a web based application. It can also be run from any device with internet connection.

We demonstrate the usage and capacities of this Shiny App for surrogacy using several examples clinical trials in which validation of a surrogate to the primary endpoint in the trials was of interest.

*Keywords:* Endpoint, surrogate, clinical trial

Presentation type: Poster

# Prediction of key parameters in the production of biopharmaceuticals using R

**Theresa Scharl, Michael Melcher, Gerald Striedner & Friedrich Leisch**

*BOKU Vienna, BOKU Vienna, University of Natural Resources and Life Sciences, University of Natural Resources and Life Sciences*

**Abstract:** In this contribution we present our workflow for model prediction in E. coli fed-batch production processes using R. The major challenges in this context are the fragmentary understanding of bioprocesses and the severely limited real-time access to process variables related to product quality and quantity. Data driven modeling of process variables in combination with model predictive process control concepts represent a potential solution to these problems. In R the statistical techniques best qualified for bioprocess data analysis and modeling are readily available.

In a benchmark study the performance of a number of machine learning methods is evaluated, i.e., random forest, neural networks, partial least squares and structured additive regression models. For that purpose a series of recombinant E. coli fed-batch production processes with varying cultivation conditions employing a comprehensive on- and offline process monitoring platform was conducted. The prediction of cell dry mass and recombinant protein based on online available process parameters and two-dimensional multi-wavelength fluorescence spectroscopy is investigated. Parameter optimization and model validation are performed in the framework of a leave-one-fermentation-out cross validation. Computations are performed using among others the R packages robfilter, boost, nnet, randomForest, pls and caret. The results clearly argue for a combined approach: neural networks as modeling technique and random forest as variable selection tool.

*Keywords:* machine learning, random forest, neural networks, structured additive regression models, bioengineering, chemometrics

Presentation type: Poster

# Video Tutorials in Introductory Statistics Instruction


## Thomas Edward Burk

*University of Minnesota*



**Abstract:** I use the Rcmdr package in the introductory statistics course I teach for non-majors. For the past several years I've used video tutorials, in addition to written documents covering the same material, for the lab portion of the course where students use Rcmdr and R to analyze data. All course materials are made available via a content management system that allows me to analyze to what degree students are utilizing various delivery mechanisms. This poster will present how I've assembled the video tutorials as well as usage patterns over the last three course offerings. The associations between tutorial usage type/frequency and student performance in the course are also explored.



*Keywords:* Instruction, videos, student performance

Presentation type: Poster

# Using R with Taiwan Government Open Data to create a tool for monitor the city's age-friendliness

**Ting-Wei Lin, Wen Tsai Hsu, Zheng Wan Lin, Yu Wen Kao, Po Shang Yang & Chi Tse Teng**

*National Taiwan University, Providence University, Providence University, Providence University, National Chung-Hsing University, Providence University*

**Abstract:** Due to rapidly growing aging population,to create a aging-friendly city is a important goal of modern government. Some indexes to reflect the city's age-friendliness may help the local government to monitor and improve the policy practice and these information also should be open and interactive to the citizen who caring about this issue. And the R language provides a great flexibility in dealing with the diversity of the file formats from government. Besides, the data visualization and web application supported by R can make the analysis result more understandable and interactive.

According to WHO 2015 age-friendly city guidelines, there are eight aspects for a comfort of elder living (outdoor spaces, transportation, housing, social participation, social respect, civic participation, communication, health and community support). And we use the Taiwan OpenGovernment data to integrate indexes with normalization and to visualize the indexes geographically. In the end, we create a shiny application with interactive Plotly to let the result easily be approached. The result may show how R can easily to utilize the government data and provide a great application turning WHO guideline into a monitor tool helping the government practice in age-friendly policy.

*Keywords:* OpenGovernment, Shiny, Age-friendliness, geographic mapping

Presentation type: Poster

# Reproducible research works_with_R


## Toby Dylan Hocking

*McGill University*



**Abstract:** Doing truly reproducible research using R is difficult! If your code uses CRAN packages, and those packages have been updated since you last ran your code, then you may get different results. Rather than declaring dependencies using only package names, for example library(glmnet), I have found that it helps to also declare the versions, for example works_with_R("3.2.3", glmnet="1.9.5"). In this lightning talk I will briefly explain how these declarations make it easier to (1) install packages, (2) load packages, and (3) perform reproducible research.



*Keywords:* install.packages, CRAN, versions, github, library, reproducible research

Presentation type: Poster

# Imputing Gene Expression to Maximise Platform Compatibility

**Weizhuang Zhou, Lichy Han & Russ B. Altman**

*Stanford University*

**Abstract:** Microarray measurements of gene expression constitute a large fraction of publicly shared biological data, and are available in the Gene Expression Omnibus (GEO). Many studies use GEO data to shape hypotheses and improve statistical power. Within GEO, the Affymetrix HG-U133A and HG-U133 Plus 2.0 are the two most commonly used microarray platforms for human samples; the HG-U133 Plus 2.0 platform contains 54,220 probes and the HG-U133A array contains a proper subset (21,722 probes). When different platforms are involved, the subset of common genes is most easily compared. This approach results in the exclusion of substantial measured data and can limit downstream analysis. To predict the expression values for the genes unique to the HG-U133 Plus 2.0 platform, we constructed a series of gene expression inference models based on genes common to both platforms. Our model predicts gene expression values that are within the variability observed in controlled replicate studies and are highly correlated with measured data. Using six previously published studies, we also demonstrate the improved performance of the enlarged feature space generated by our model in downstream analysis.

*Keywords:* LASSO, gene inference, microarray

Presentation type: Poster

# Partition-Assisted Clustering: Application to High-Dimensional Multi-Sample Single-Cell Data Analysis

**Ye Li, Dangna Li, Nikolay Samusik, Xiaowei Wang, Garry P. Nolan & Wing H. Wong**

*Stanford University, Stanford University, Stanford University, Peking University, Stanford University, Stanford University*

**Abstract:** Cytometry advances the study of cellular phenomena at the single-cell level by providing the necessary information to find subpopulation complexity of a heterogeneous sample. The analysis of cytometry datasets is becoming more challenging as cytometry advances increase the data size and dimension. Flow cytometry typically monitors at most 12 genes per cell; however, in recent years, mass cytometry (CyTOF) has taken the routine upper limit to 45 genes per cell. Biologists traditionally analyze the data by manually drawing polygon enclosures around subpopulations on a series of 2D projections of the data; this procedure becomes exponentially harder and time-consuming with increasing dimensions. To aid both experts and non-experts in finding the subpopulation complexity of cytometry samples, we introduce and apply partition-assisted clustering (PAC), which is implemented as an R package called PAC, to enable consistently accurate and efficient analysis of low and high-dimensional single-cell datasets. PAC utilizes the data density implicitly and enables the discovery of cell subpopulations in the dataset without computational bias due to systematic elimination of data points in the analysis.

*Keywords:* cytometry, clustering, data density

# Part II

# Lightning Talk

Presentation type: Lightning Talk

# Automated risk calculation in clinical practice and research - the riskscorer package


## Alexander Meyer, Stefan Vogel, Simon Sündermann, Jörg Kempfert & Volkmar Falk

*German Heart Institute Berlin*



**Abstract:** Clinical risk scores are important tools in therapeutic decision making as well as for analysis and adjustments in clinical research. Often risk scores are published without an easily accessible interface for calculation. And if tools exist, mostly these are web based user interfaces and therefor not suitable for either batch processing in research or integration into the hospital's clinical information system infrastructure.

We developed the _riskscorer_ package for easy and automatic clinical risk score calculation with the following features in mind:

- simple programming interface
- extensibility
- flexible handling of differing data codings
- individual patient risk calculation as well as the possibility of batch processing
- an HTTP web-service interface based on the plumber (https://github.com/trestletech/plumber) package for easy integration into an existing clinical information system infrastructure

Currently three surgical risk scores are implemented: STS score (http://riskcalc.sts.org/), EuroScore I and EuroScore II (http://www.euroscore.org/). It is already used in our research and integration into our clinical information system is planned. The riskscorer package is under continues development and we have released the source code under the MIT license on the GitHub platform (https://github.com/meyera/riskscorer).

The integration of automated risk score calculation into the clinical workflow and into reproducible and efficient data analysis pipelines in research has the potential to improve patient outcomes.



*Keywords:* risk sccores, reproducible research, web service, plumber

Presentation type: Lightning Talk

# rempreq: An R package for Estimating the Employment Impact of U.S. Domestic Industry Production and Imports


## Allan Miller

*UC Berkeley Extension*



**Abstract:** The impact of imports and technological change on domestic employment is a long-term and ongoing topic for academic and government research, and discussion in the popular media[1][2].

The U.S. Bureau of Labor Statistics (BLS) publishes a current and historical Employment Requirements Matrix (ERM), which details the employment generated directly and indirectly across all industries by a million dollars production of a given industry's primary product[3]. The BLS data can give an indication of the relative impact of different industries' primary production, and is broken down by years (1997-2014), over 200 sectors, and by domestic-only versus total production including imports. The ERM is often used in research as a component of general Leontief Input-Ouput models, and external sources of economic data for final demand.[4] R, with it's support for input-output modelling and general matrix operations, is well-suited to research in this area[5][6].

The package rempreq includes both the current and historic tables, for total production and domestic production only. It also includes functions for accessing any particular year (or years), selecting industries, and for including domestic versus total output. These can be used to conveniently generate estimated time series for the employment impact for various types of production, the impact of imports on employment, and investigate changes in the technological structure of industries related to employment in those industries over time.

This presentation will include an introduction to rempreq, sample demonstrations of its use, and future plans for the extension of the package.



*Keywords:* employment impact, input-output analysis, economics, international trade

Presentation type: Lightning Talk

# Peirce–theory-of-signs in R


## Alon Friedman

*University of South Florida*



**Abstract:** According to the community site for R packages, crantastic, the most popular packages often provide us advanced statistical models, toolboxes for data manipulation, and different visualization capabilities. We see little packages that applies to the study of language. The philosophy of language, or semiotics, discusses the nature of the meaning in the language, its cognition and relationship to reality. Few R packages address these issues in today's data rich environment. We raise the question: can we explore a semiotics theory in R and what is the outcome?

Semiotics, the study of signs, is derived from the philosophical speculations on signification and languages (Chandler, 2004). In the nineteenth century, deliberation of the meaning of the term continued via two schools of thought that provided different interpretations. The first was promulgated by the American philosopher Charles Sanders Peirce, who proposed the study of the sign as "semiosis." Ferdinand de Saussure, on the other hand, studied the sign as a dyadic relationship that is connected by a linguistics structure. We focus on the work of Peirce with a triadic foundation of the term sign, where anything can be sign as long as someone interprets it as signifying something, referring to standing for something other than itself. Peirce's semiotic theory is based on deductive logic, where the process of reasoning stands for one or more statements (premises) to reach a logically certain conclusion. His approach has merit both in its defined scope and its appeal to those who are interested in the development of modern logic.

In order to convert Peirce's sign theory to R, we developed a new object classification. This classification holds the same attributes of the classes and methods as discussed by Becker, et.al (1988) but it also provides additional description to better organize the object properties and its triangulation relationships. We found that the advantage of using Peirce's sign theory in R is the ability to convert data to a triangulation structure based on object classification. In our presentation, we demonstrate Peirce's monadic logic and its iconic and visual representation in R.



*Keywords:* Semiotics, C.S. Peirce, triangulation, meta analysis

Presentation type: Lightning Talk

# Two Cultures: From Stata to R


## Annie J Wang

*Analyst Institute*



**Abstract:** It's one of the most popular programming languages for statistics; it's widely taught and used in industry and academia alike; and it has a vibrant user community. It's Stata! For many researchers, Stata is their native tongue. This talk discusses what happens when you join an analytics organization that runs on Stata and then transition that team to R. It's about more than just swapping out egen for dplyr::mutate: using R requires a cultural shift in a team's attitudes around programming, analysis, quality assurance, and collaboration. We'll go through the most important and most challenging of these shifts, including places where the culture of Stata can make us better R users.



*Keywords:* R, stata, industry applications, social science, researchers, pedagogy

Presentation type: Lightning Talk

# Scaling R for Business Analytics


## Arlene Mari Zaima

*Teradata*



**Abstract:** There's no question that R is the fastest growing analytic language amongst data miners and data scientists. Organizations are also embracing R for business analytics to attract the new generation of analytic talent entering into the industry. However data and processing limitations associated with R become a real challenge as analyst wrestle with billions of records and analyze complex relationships, while working with new data sources to enhance business analytic solutions. Vendors are addressing this challenge with parallel R technology and claim to "lift all limitations of R", but no data platform will "auto-magically" scale R. This session drills into the different ways to scale R and its benefits and challenges. The takeaway from this session is a set of questions that can be used to evaluate scalable R technologies to align with your business requirements.



*Keywords:* Scaling R, parallel, big data, business analytics

Presentation type: Lightning Talk

# NetworkRiskMeasures: risk measures for (financial) networks, such as DebtRank, Impact Susceptibility, Impact Diffusion and Impact Fluidity.

**Carlos Leonardo Kulnig Cinelli & Thiago Christiano Silva**

*Central Bank of Brazil*

**Abstract:** The recent financial crisis has made clear to academics and regulators that it is not enough to assess systemic risks looking only at the health of individual institutions — due to interconnectedness, exposures that may seem harmless at the individual level may turn out to be systemically dangerous when taking into account the system as whole. Complex network theory and computer simulations can help one assess how risks could propagate in financial networks. Although an important subject, to the best of our knowledge the R community still lacks a package that implements systemic risk analysis tools for networks. The NetworkRiskMeasures package addresses this issue by providing a unified framework for analyzing risk in financial networks. It compiles several measures and algorithms used to estimate risk, both at the micro and macro levels, such as Default Cascades, DebtRank, Impact Susceptibility, Impact Diffusion and Impact Fluidity. In this presentation, we will first formally introduce some notions of financial risk measures and network theory. Then, using networks estimated by maximum entropy and minimum density methods, we illustrate how one can perform network risk assessment in practice using the NetworkRiskMeasures package.

*Keywords:* Netoworks, Systemic Risk, Contagion

Presentation type: Lightning Talk

# The Best Time to Post on Reddit


## Daniel David Leybzon

*UCLA*



**Abstract:** I used R to visualize the best time to post on Reddit using data collected with Google's BigQuery. I queried a publicly accessible dataset containing almost 200 million Reddit posts (all of the posts between January 2006 and August 2015), extracting the score (upvotes minus downvotes) and the timestamp for each post. I proceeded to separate the posts into the day of the week and the time of day it was posted, summed the scores for that period of time, and divided the result by the total number of posts made in that period of time. I visualized this normalized data as a heatmap using ggplot2.



*Keywords:* visualization, r, reddit, ggplot2, bigquery

Presentation type: Lightning Talk

# shinyjs: Easily improve UX in your Shiny apps without having to learn JavaScript


## Dean Attali

*University of British Columbia*



**Abstract:** Whether you're a seasoned Shiny app developer or you're just excited that you were able to complete the tutorial, shinyjs can probably improve your apps within seconds. There are many nseemingly mundane functions that many Shiny users often want to perform: hiding/showing an element, disabling/enabling an input, resetting an input back to its original value, and many nothers. The shinyjs package makes all of these extremely easy by calling a single function. All of this is done using JavaScript (JS) under the hood, but you don't have to know that! If you do know JS, shinyjs also lets you interface with JS in a simple way and provides an easy way to call your own JS functions as if they were R code. This talk will introduce you to the wonderful world of shinyjs, and hopefully leave you excited to build more Shiny apps!



*Keywords:* shiny

Presentation type: Lightning Talk

# Empowering Business Users with Shiny


## Derek Damron

*Allstate Insurance*



**Abstract:** Relationships between data scientists and business users can often be very transactional in nature (i.e. give us some data and we'll give you a solution). This approach to analytics can produce meaningful results but removes business users from the analytical process, which often hinders adoption and prevents user insight from enhancing the analysis.

Shiny is a powerful tool that can be used to create compelling output from analytic work but it can also be used to cultivate interactive feedback loops between data scientists and business users. These feedback loops help ensure that data scientists are answering the right questions and that business users are given the opportunity to invest themselves in the analysis, which often expedites the execution and adoption of the data science work. The iterative development of these Shiny applications also works well within the agile framework that is becoming common for data science projects.

In this talk we will discuss some examples of how Shiny has been used at Allstate to empower business users and create an organizational appetite for data science.



*Keywords:* shiny, visualization, building relationships, driving change

Presentation type: Lightning Talk

# Estimating causal dose response functions using the causaldrf R package


## Douglas Galagate & Joseph L. Schafer

*U.S. Census Bureau*



**Abstract:** Causal inference aims at the fundamental question of how changing the level of a cause or treatment can affect a subsequent outcome. Whether data analysts want to admit it or not, many analyses in behavioral, social, biomedical, and other fields of science are aimed at understanding causal relationships, even when the data or methods are not well suited to the task.

This presentation gives an overview of the causaldrf R package which addresses the relatively under-explored problem of estimating causal effects when the treatment is real-valued and continuous.



*Keywords:* observational data, causal inference, potential outcomes, propensity scores, treatment effects, dose-response function

Presentation type: Lightning Talk

# Chunked, dplyr for large text files


## Edwin de Jonge

*Statistics Netherlands (CBS)*



**Abstract:** During a data analysis project it may happen that a new version of the raw data comes available or that data changes are made outside of your control. `daff` is a R package that helps to keep track of such changes. It can find differences in values between data.frames, store these differences, render them and apply them as a patch to a new data.frame. It can also merge two versions of a data.frame having a common parent version. It wraps the daff.js library of Paul Fitzpatrick (`http://github.com/paulfitz/daff`) using the V8 package.



*Keywords:* large data file processing, dplyr

Presentation type: Lightning Talk

# gtfsr: A package to make transit system analysis easy


## Elaine Allen McVey

*TransLoc*



**Abstract:** The General Transit Feed Specification (GTFS) is a standard to represent transit systems (routes, schedules, stops, etc.). The gtfsr package makes it easy to analyze transit systems by: (1) providing an interface to APIs that provide GTFS feeds, (2) defining a standard data object for GTFS data, (3) validating multiple aspects of data/feed quality, (4) providing convenience functions for common data joins, (5) enabling useful visualizations, comparisons, and analyses of single or multiple transit systems. nA particularly interesting and challenging aspect of this project is determining how to assess data quality from various angles in a situation where it varies widely. Although the data formats in GTFS feeds are fairly straightforward, standardizing the way GTFS data is structured and validated in R has the potential to enable common R tools for sophisticated transit system analysis.



*Keywords:* transit, package, api, gtfs

Presentation type: Lightning Talk

# Convenient educational & psychological test reporting with the QME package & a Shiny UI

**Ethan Christopher Brown, Kory Vue & Andrew Zieffler**

*University of Minnesota*

**Abstract:** The QME package and Shiny UI for QME were created to provide a relatively easy-to-use tool for performing psychometric analyses in R under the Classical Test Theory framework. Although there are already several packages in R that can be used to carry out different parts of a psychometric analysis, there is not currently a package that provides practitioners with the comprehensive functionality to import and score test data, compute item- and test-level statistics, and perform distractor analysis. Faced with these challenges, faculty and graduate students at the University of Minnesota have been collaborating on finding better software solutions to drive broader adoption of psychometric practices. The QME package and Shiny UI fill this gap by allowing users to go from their raw un-scored item responses to a practical report with relative ease, providing a compelling open source alternative to commercial packages such as ITEMAN and SPSS.

*Keywords:* psychometrics, shiny, reporting, classical test theory

Presentation type: Lightning Talk

# Interact with Python from within R


## Florian Schwendinger

*WU Vienna University of Economics and Business*



**Abstract:** `PythonInR` makes it easy to send R objects to Python, retrieve Python objects to R, call Python functions/methods and evaluate Python code. Furthermore, it allows to generate virtual Python objects in R and to import entire Python packages to R, by automatically generating the appropriate interfaces. A shortcoming of automatically generated interfaces is that they depend on the type transformations automatically done by `PythonInR`. Since it is normally not possible to get a 1:1 mapping between the data structures of two programming languages, one has to decide how to deal with this ambiguity. The new version of `PythonInR` deals with this challenge in two ways. Firstly, it allows the user to specify additional type information, so called typehints. From a technical point of view, typehints are just attributes, which allow to provide additional information in a structured way (e.g. `th.string(x)`) without changing the class information and print behavior. Secondly, since Python, unlike R, has no built-in data structures for matrices and vectors, `PythonInR` reimplements those data structures with the appropriate methods, so they can easily be transformed into their Python equivalents, if the necessary packages are available. Equipped with an 1:1 mapping of the most common types it is possible to programmatically generate bindings in a meaningful way. In the talk, I would like to present how to use `PythonInR` and show how to import entire packages into R.

In the talk, I would like to present how to use PythonInR and show how to import entire packages into R.



*Keywords:* R, Python, PythonInR

Presentation type: Lightning Talk

# Event Detection with Social Media Data

**Frederick J. Boehm, Robert W. Turner & Bret M. Hanlon**

*University of Wisconsin-Madison*

**Abstract:** Social media represents a new mechanism by which individuals form opinions about social and political events. Most social media users now have the ability to communicate with other users, even if they are separated in geography and ideology. As an initial step in studying communications on social media, we present a workflow in R to detect events. Specifically, we use latent Dirichlet allocation (LDA) of tweets from Twitter at well-defined time intervals to detect important political and social events. A critical step in our workflow is the data wrangling of raw tweet downloads from Twitter, which we accomplish with the new parseTweetFiles R package. We illustrate our methods on tweets near and during the time of the National Football League's 2015 Super Bowl game. With this collection of tweets, we detect short-lived topics related to 1) important Super Bowl plays and players and 2) Super Bowl half-time show performers. We then discuss implications of our methods for event detection and place our findings in the context of scholarly discussions of social media discourse.

*Keywords:* Twitter, Social Media, Event Detection, Latent Dirichlet Allocation, Unsupervised learning

Presentation type: Lightning Talk

# Visualization of Uncertainty for Longitudinal Data

**Bénédicte Fontez, Nadine Hilgert, Susan Holmes & Gabrielle Jeanne Weinrott**

*INRA Montpellier, Montpellier SupAgro, Stanford University, INRA Montpellier*

**Abstract:** Data in agronomy and other life sciences are often sparse and longitudinal and contain inherent uncertainty that needs to be taken into account. For practical reasons, the results of exploratory analysis of data of this kind should be presented in a way that is interpretable and accessible to scientists in the field. Latent Factor Models can be quite useful to expose the underlying structure of a data set (see West, 2003). A Bayesian framework opens the possibility of incorporating expert knowledge and information about the level of uncertainty in the form of prior distributions (see Rowe, 2000; Minka, 2000; or Ghosh & Dunson, 2008), and the ability to recover posterior density quantiles.

Our R package performs Bayesian Latent Factor Analysis on longitudinal data and includes novel graphics to visualize data uncertainty. The Bayesian inference is done using a No-U-Turn-Sampler (see Hoffman & Gelman, 2011) with the *rstan* package, the R wrapper for the STAN programming language (http://mc-stan.org/). The package provides graphics that resemble the outputs of classical Principal Component Analysis (a close relative of Latent Factor Models, see Tipping & Bishop, 1999), but with integrated projection uncertainty regions, enabling facilitated interpretation of the effect of uncertainty on the analysis results.

*Keywords:* Visualization, Uncertainty, Longitudinal Data, Exploratory Data Analysis, Bayesian Latent Factor Analysis, STAN

Presentation type: Lightning Talk

# Optimizing Food Inspections with Analytics


## Gene Leynes & Tom Schenk

*City of Chicago*



**Abstract:** In 2013 the City of Chicago was the recipient of a Bloomberg Philanthropies grant to develop a smart data platform. The aim of the platform is to develop tools to help city government increase efficiency through data driven decision making. Based on this work the city released a machine learning application in 2014 that helps inspectors prioritize their workload by predicting which food establishments are most likely to have violations.

The food inspection project has been released as an open source project on GitHub, and all of the data and code that was used to develop and evaluate the model has been made public. The model brings together several disparate data sources such as garbage cart requests, various 311 complaints, weather information, as well as past inspection results and business information.

The project is intended to be reproducible and the hope is that it will be replicated in other cities. We used many open source tools such as Knitr and GitHub to make it easy for others to replicate the work. Indeed, many cities are showing interest, and two organizations (one for profit and one not for profit) have already begun the replication work.

Another interesting aspect of this work is the public private collaboration that made this research possible. Aside from the initial grant, much of the initial research was done on a volunteer basis by members of the Allstate Insurance data science team. They used entirely open data which is freely available on Chicago's open data portal https://data.cityofchicago.org to develop the initial model. Also the local community has been very engaged in this and other similar projects.

The predictive model runs nightly in R, and the results are exported to a Shiny application which is used by the director of food inspections. The model relies heavily on the data.table package for efficient data processing and management.

This work has been featured on PBS NewsHour, in Atlantic Monthly's CityLab website, our local Chicago Sun Times, and more.

The most exciting aspect of this work is that it has such broad application. Chicago, like many other municipalities, has struggled with years of budget cutbacks and reduced staff levels, but the workload has remained unchanged. Many departments struggle to keep up with inspections that are important for public safety, such as elevators, building permits, and home lead inspections. We are currently working to improve the process of these inspections by illuminating the most likely sources of problems to help focus on issues that have the greatest impact for the public.


References:


https://chicago.github.io/food-inspections-evaluation/

http://mayorschallenge.bloomberg.org/ideas/the-chicago-smartdata-platform/

Presentation type: Lightning Talk

http://www.pbs.org/newshour/bb/chicago-revamps-restaurant-inspections-by-tapping-into-social-media/
http://www.citylab.com/cityfixer/2016/01/chicago-is-predicting-food-safety-violations-why-arent-other-cities/422511/
http://chicago.suntimes.com/news/7/71/838316/restaurant-inspections-predictive-analytics

*Keywords:* reproducible research, open data, government, data.table, reproducible operations

Presentation type: Lightning Talk

# Let's meet on satRday!


## Gergely Daroczi

*Budapest Users of R Network*



**Abstract:** The idea of organizing cheap regional conferences, as a link between local R User Groups and international conferences on R, was brought up at the EARL 2015 conference in Boston, which was quickly followed-up by a short survey in the R community – with quite a lot and positive responses. So we decided to organize the first few SQLSaturday-like R events on three continents in 2016: having one-day long, cheap or totally free conferences on the weekends with 100-300 attendees. This lightning talk will present the overall goals of this initiative, the survey results collected before submitting the project to the R Consortium and also the news on the forthcoming events in this year.



*Keywords:* R community, R User Groups, conferences

Presentation type: Lightning Talk

# Text Mining and Sentiment Extraction in Central Bank Documents


## Giuseppe Bruno

*Bank of Italy*



**Abstract:** The deep transformation induced by the World Wide Web (WWW) revolution has thoroughly impactedna relevant part of the social interactions in our present global society. The huge amount ofnunstructured information available on blogs, forum and public institution web sites puts forward differentnchallenges and opportunities. Starting from these considerations, in this paper we pursue a two-foldngoal. Firstly we review some of the main methodologies employed in text mining and for the extractionnof sentiment and emotions from textual sources. Secondly we provide an empirical application by consideringnthe latest 20 issues of the Bank of Italy Governor's concluding remarks from 1996 to 2015. Byntaking advantage of the open source software package R, we show the following:

- 1. checking the word frequency distribution features of the documents;
- 2. extracting the evolution of the sentiment and the polarity orientation in the texts;
- 3. evaluating the evolution of an index for the readability and theformality level of the texts;
- 4. attempting to measure the popularity gained from the documents in the web.

The results of the empirical analysis show the feasibility in extracting the main topics from the consideredncorpus. Moreover it is shown how to check for positive and negative terms in order to gauge thenpolarity of statements and whole documents. The R employed packages have proved suitablenand comprehensive for the required tasks. Improvements in the documentation and the package arrangement are suggested for increasing the usability.



*Keywords:* Text Mining, Wordcloud, Polarity, Sentiment Analysis, Zipf's Law

Presentation type: Lightning Talk

# Introduce R package: Tree Branches Evaluated Statistically for Tightness (TBEST)


## Guoli Sun & Alexander Krasnitz

*Stony Brook University, Cold Spring Harbor Laboratory*



**Abstract:** We formulate a method termed Tree Branches Evaluated Statistically for Tightness (TBEST) for identifying significantly distinct tree branches in hierarchical clusters. For each branch of the tree a measure of distinctness, or tightness, is defined as a rational function of heights, both of the branch and of its parent. A statistical procedure is then developed to determine the significance of the observed values of tightness. Based on our benchmark analysis, TBEST is a tool of choice for detection of significantly distinct branches in hierarchical trees grown from biological data.



*Keywords:* R package introduction, hierarchical clustering, bioinformatics

Presentation type: Lightning Talk

# Automated clinical research tracking and assessment using R-Shiny

**Hao Zhu, Timothy Tsai, Ilean I. Isaza & Thomas G. Travison**

*Hebrew Seniorlife, Institute for Aging Research, Hebrew Seniorlife, Institute for Aging Research, Hebrew Seniorlife, Institute for Aging Research, Hebrew Seniorlife, Institute for Aging Research; Boston Claude D. Pepper Older Americans Independence Center; Harvard Medical School*

**Abstract:** Many database tools enjoying widespread use in academic medicine, such as REDCap, provide only limited facilities for study monitoring built-in. They often do, however, provide the ability for analysts to interact with data and metadata via Application Program Interfaces (APIs). This offers the possibility of automated monitoring and web-based reporting via the use of external tools and real-time interaction with the API.

In this presentation we provide a framework for automated data quality and participant enrollment monitoring in clinical research using R Shiny Server. To illustrate, we demonstrate linkage of R Shiny to a REDCap database via its API, streamlining the process of data collection, fetching, manipulation and display in one process. Automated display of to-the-minute enrollment, protocol adherence, and descriptions of enrolled samples, are provided in a format consistent with NIH reporting requirements. Additionally, we demonstrate the flexibility of this system in providing interactivity to address investigator queries in real-time; for instance, investigators may select sample subgroups to display, or inquire as to the risk profile of an individual with an extreme baseline measurement on an important screening variable. This approach offers the potential to greatly increase efficiency and maintenance of data quality in research studies.

*Keywords:* shiny, REDCap, clinical data management, bioinformatics

Presentation type: Lightning Talk

# GeoFIS: an R-based open source software for analyzing and zoning spatial data


## Hazaël Jones, Bruno Tisseyre, Serge Guillaume, Jean-Luc Lablée & Brigitte Charnomordic

*Montpellier SupAgro, Montpellier SupAgro, IRSTEA, IRSTEA, INRA*



**Abstract:** There is an emerging need to integrate spatial data into easy to use decision support tools. R provides many packages for analysing and modelling spatial data that are going to be useful for decision support in various fields, such as Geography, Environment and Digital Agriculture.

The GeoFIS software (https://mulcyber.toulouse.inra.fr/projects/geofis/) provides a simple scalable framework to view and analyze spatial data. The user-friendly interface is designed to be supplemented easily by the addition of R-functions. It imports and filters georeferenced data or GIS layers, to analyze their spatial structure and to represent them with zoning algorithms (based on Euclidean or fuzzy distances allowing to include expert knowledge, see Pedroso et al, 2010). A simplified representation using homogeneous zones helps users in their decision processes.

The GeoFIS interface is written in Java and uses the open source GeoTools library to display data layers. Geostatistical analyses are implemented through calls to R packages (sp, gstat. rgeos). The calling protocol is based on R-serve (http://www.rforge.net/Rserve/) and the encapsulation of S4 R objects into Java classes. Zoning algorithms include some C++ calls to accomodate large data sets. GeoFIS is used by students, engineers and researchers with little or no knowledge of R necessary. An R commander like interface allowing GeoFIS calls from inside R is under study.



*Keywords:* Georeferenced data, Visualization, Decision support, Expert knowledge, Uncertainty, Fuzzy Distance, Segmentation, Free software, R-serve, Simple interface

Presentation type: Lightning Talk

# Weather Alerts Data with R


## Ian Cook

*TIBCO Software Inc.*



**Abstract:** Weather data is of interest to many R users. Existing R packages, including weatherData, provide access to sources of current and historical weather conditions data. But there has been no R package to retrieve current weather alerts data, such as the advisories, watches, and warnings issued by government weather agencies. This talk introduces the new R package weatherAlerts, which retrieves active weather alerts from the United States National Weather Service (NWS). I will discuss potential applications for visualization and analysis, demonstrate uses in the Internet of Things, and seek international collaborators to expand the geographical scope of the package.

Furthermore, weather alerts are associated with affected geographical areas, which are often specified as references to predefined polygons in external sources of spatial data. Common applications of weather alerts data require merging the data with this spatial data, so that each weather alert includes a polygon or polygons specifying its affected geographical area. This talk introduces the companion package weatherAlertAreas, which compactly represents all predefined polygon areas used in NWS alerts, and eliminates the substantial effort that was previously required to merge NWS alerts data with corresponding spatial data.



*Keywords:* weather, open data, spatial data

Presentation type: Lightning Talk

# Tie-ins between R and Openstreetmap data


## Jan-Philipp Kolb

*Gesis Leibniz Institute for the social sciences*



**Abstract:** An abundance of information emerged through collaborative mapping as a consequence of the development of Openstreetmap (OSM) in 2004. Currently, all kinds of R-packages are available to deal with different types of spatial data. But getting the OSM data into the R-environment can still be challenging, especially for users who are new to R. One way to access information is the usage of application programming interfaces (APIs) like the Overpass API.

In this presentation, I will focus on the possibilities to access, assess and process OSM-data with R. Therefore, I will provide the tie-ins of the R-language and OSM. Since XML-protocols are often used to describe spatial information, we will use the package geosmdata, which is a wrapper to transfer such information to R-dataframes, using the XML-package. Furthermore, I will showcase the importance of such connections via a brief case-study related to social sciences.



*Keywords:* geosmdata, Nominatim, OpenStreetMap, osmar, Overpass API, raster, rjson, sp, XML

Presentation type: Lightning Talk

# R's Role in Healthcare Data: Exploration, Visualization and Presentation


## Jeff Mettel

*Loopback Analytics*



**Abstract:** Over the past six or seven years, the healthcare industry has been transformed through a broad shift away from paper-based workflows to electronic ones. This new electronic infrastructure has enabled the collection of vast amounts of clinical and financial data. Problematically, the ability to analyze such data and convert it into meaningful insights capable of driving continuous improvements has not kept pace with the speed of data acquisition.

With healthcare costs continuing to grow, national policy changes have recently placed a new mandate for healthcare providers to leverage data to reign in expenses. R, which has been honed and developed across a number of industries prior to healthcare, has proven to be an invaluable tool in addressing this challenge. The integrated data manipulation capabilities, focus on exploratory data analysis and deep integration with broader data analysis pipelines, including final end-user analytics and visualizations, have proven invaluable.

This presentation will highlight the growing role of R in healthcare analytics. Specifically, it will focus on:

- R's role in rapid data exploration, facilitated through R-Markdown, particularly around data sets where the exact questions are still being formulated
- R's role in end-user-ready visualizations, with a focus on the ggplot package
- R's role in crafting an explanatory narrative around investigatory analysis



*Keywords:* healthcare, analytics, visualization, ggplot

Presentation type: Lightning Talk

# Hash Tables in R are Slow


## Jeffrey Horner

*Vanderbilt University Department of Biostatistics*



**Abstract:** An array hash is a cache-conscious data structure that takes advantage of hardware prefetchers for improved performance on large hash tables, those large enough to fit in main memory and larger than fast fixed size cpu caches.

However, their implementation is a radical departure from standard chained hash tables. Rather than using chains of hash buckets for collision resolution, array hashes use segments of contiguous memory called dynamic arrays to store keys and values. Adding and deleting items from the hash involve copying the entire segment to new areas in memory. While this may seem wasteful and slow, it's surprisingly efficient in both time and space[2].

In R, hashed environments are implemented using lists with each list element (a CONS cell) acting as the hash bucket. The CONS cell is the binding agent for a symbol and value. Hashed environments are searched using the pointer address of the symbol rather than the symbol's printed name.

R-Array-Hash takes advantage of this by implementing an integer array hash[1] to store addresses of symbols and their associated values. Care is also taken to account for whether or not a binding is locked, active, etc.

Similarly, R-Array-Hash reimplements R's string cache using a string array hash. This introduces the most radical change to R's API: CHAR() no longer returns an address that points to the area at the end of the SEXP. Rather it returns an address located in one of the contiguous dynamic arrays of the string hash table. Therefore, care must be taken in C code to use the address immediately since additions and deletions to the string hash could render the result of CHAR() useless. There are many areas of the code that sidestep this by calling translateChar(), which has been changed to always copy the string pointed by CHAR().



*Keywords:* high performance computing, natural language processing, text mining

Presentation type: Lightning Talk

# Clustering of Hierarchically-Linked Multivariate Datasets

**Terrance D. Savitsky & Jeffrey M. Gonzalez**

*US Bureau of Labor Statistics*

**Abstract:** We present the growclusters package for R that implements a maximum posterior estimation of partitions (clusters) using a penalized optimization function derived from the limit of a Bayesian probability model under a multivariate Gaussian mixture on the mean, either under a Dirichlet process (DP) mixing measure or a hierarchical DP (HDP) mixing measure in the limit of a function of the global variance (to zero). We illustrate this package using data collected from a federal survey of business establishments. A special feature of this data is that it is collected under an informative sampling design. Under an informative sampling design the probability of inclusion depends on the surveyed response. We demonstrate a feature of the growclusters package that incorporates the sampling weights to "undo" the effects of the informative design to yield asymptotically unbiased estimation of the clusters.

*Keywords:* surveys, establishment surveys, informative sampling design

Presentation type: Lightning Talk

# A Shiny App is Worth 1000**3 Words: A Case Study in Displaying Three-Dimensional Dose Combination Response Data


## Jocelyn Sendecki

*Nonclinical Biostatistics, Janssen Research and Development, LLC*



**Abstract:** We all know the adage "A picture is worth a thousand words," but when it comes to multi-dimensional data, there are challenges to maintaining this visual brevity; this case study provides an example. A dose combination experiment was conducted wherein cytokine expression in T cells was measured in response to two compounds administered in combination across a set range of doses. Response surface regression was used to find the dose combination with the highest level of response. Biological response to this "best" dose combination was then compared against response to single-dose comparator compound. Tables of summary statistics, conditional boxplots, snapshots of 3D scatterplots, and paragraphs describing the first derivative are all perfectly adequate to present results from this experiment in a manuscript. However for fast communication of complicated results, a Shiny app succeeds admirably. This particular Shiny app, which utilizes the shiny, rgl, and shinyRGL packages, provides a dynamic and flexible solution to the difficulty of displaying literally all sides of this multi-dimensional analysis to live audiences. Rather than having to sit through an explanation of local extrema versus saddle points, scientist can quickly generate the shape of biological response over dose combinations and examine how the best dose combination stacks up against the comparator. Moreover, users can also take advantage of the portability that deployment on a Shiny server or Shinyapps.io provides to those without R access, allowing them to present results without requiring additional software or a spare statistician



*Keywords:* Shiny, rgl, three-dimensional graphics

Presentation type: Lightning Talk

# Using R for Game Development Analysis


## Kenneth Buker

*University of South Florida*



**Abstract:** Heatmaps for the game genre FPS (first person shooters) are a core graphical component of analysis when it comes to E-sport development. These are generated from game coordinate data collected by the game engine to visualize problem areas for both map design and game flow. It is critical to be able to visualize and quantify problems efficiently and objectively. Developers with small budgets don't have the resources, expertise, and often access to create from scratch a tailored heatmap generation tool in engine. Open source R provides a free and powerful platform in which to analyze and generate these heatmaps. While there are limited packages built on game theory probabilities on CRAN, there is little too no development of applied analysis and visualization of video game data.

In this presentation, I will present the two stages of this package, the import and conversion, as well as the output graphic it produces. This was used specifically in "Insurgency" to focus development resources into fixing problem areas that arose during its competitive development and was presented at the Game Developer Conference in March of 2016. This new application of R shows its robust nature as well as its success in exploring unknown systems and industries.



*Keywords:* Video Game Analysis, Bayesian Inference, Multi-variable System Analysis, Heat-map, Esports Analysis,

Presentation type: Lightning Talk

# Forecasting Revenue for S&P 500 Companies Using the baselineforecast Package


## Konstantin Golyaev & Gagan Bansal

*Microsoft*



**Abstract:** Most businesses require accurate revenue forecasts for efficient operations. Reliable forecasts facilitate the operational planning process and enable long-term investments. At Microsoft, we forecast revenue for multiple divisions across lines of business and geographies. To this end, we developed an R package baselineforecast that builds on top the excellent forecast package that deals with univariate time series.

Our package has four major advantages. First, it computes concurrent forecasts at different horizons without having to refit the models. Second, it admits arbitrary external features external, e.g. macroeconomic data such as oil prices or unemployment rates. Third, it can forecast arbitrary number of series simultaneously. Finally, it employs an ensemble approach in which rolling forecasts from univariate time series methods act as features for a sophisticated machine learning algorithm such as elastic net regression or gradient boosted regression trees.

We applied our approach to publicly available data concerning quarterly revenue for the S&P 500 companies. Our primary data source is companies' Income Statements, which we augmented with macroeconomic time series from the Federal Reserve Economic Data (FRED), such as Real GDP, unemployment levels, U.S. leading indicator, WTI oil price, and a few others. We trained a set of univariate time series models, as well the boosted regression trees using forecasts from the above as features together with macroeconomic data from FRED. In an out-of-sample test the boosted trees regression outperformed the best time series model at longer forecast horizons, while maintaining virtually identical performance for short-term forecasts.



*Keywords:* time series forecasting, machine learning, revenue forecasting, elastic net, boosted regression trees

Presentation type: Lightning Talk

# FirebrowseR an 'API' Client for Broads 'Firehose' Pipeline


## Mario Deng & Sven Perner

*Pathology of the University Hospital of Luebeck and Leibniz Research Center Borstel, Lübeck and Borstel, Germany*



**Abstract:** The Cancer Genome Atlas is one of the most valuable resources for modern cancer research. One of the major projects for processing and analysing its data is the Firehose Pipeline, provided the Broad Institute. The pre-processed and analysed data of this pipeline is made available through the Firebrowse website (http://firebrowse.org/) and a RESTful API, to download such data sets. FirebrowseR is an R client to connect and interact with this API, to directly download and import the request data sets into R. Using FirebrowseR, only requested data sets are download from the API, reducing the overhead for downloading; compared to the classic download options, such as CSF, MAF or compressed files. Further FirebrowseR capsules the provided data sets into a standardised format (a data.frame or JSON object), making the steps of data wrangling and importing needless.



*Keywords:* firebrowse, tcga, bioinformatics, cancer genetics

Presentation type: Lightning Talk

# Bespoke eStyle Statistical Training for Africa: challenges and opportunities of developing an online course

**Miranda Yolanda Mortlock & Vincent Mellor**

*University of Queensland*

**Abstract:** The development of 'BeST' an online course for African scientists and early career researchers aimed to provide support for experimental design principles and the use of R software. It is available at yieldingresults.org and is supported by the Australian Centre for International Research (ACIAR). A team of developers produced materials with an emphasis on visual and practical materials. The site is continuously available and is in modular format and aims to assist in developing designs and following through with analysis and reporting. The early evaluation by clients and students will be presented. Options and challenges for future support and collaboration of the site will be discussed.

*Keywords:* Experimental design, analysis, online training, online course, Africa, science, agriculture, early career

Presentation type: Lightning Talk

# Interactive dashboards for visual quality control of air quality data


## Nathan Pavlovic

*Sonoma Technology, Inc.*



**Abstract:** While automated quality control is a key component of modern data processing workflows, visual review by a trained eye can further ensure data quality. In particular, graphical representations allow analysts to quickly review data in context. Such review has relied on specialized, often costly software or inefficient processing with spreadsheet software. The interactive graphing capabilities provided by the R Shiny package presents an opportunity to explore and interact with data in practical and user-friendly ways that are relatively simple to implement and flexible to the needs of particular datasets.

Using Shiny, we developed data quality control dashboards to facilitate analyst review of ambient air quality and meteorological datasets such as criteria pollutant concentrations, wind profiles, and weather forecasts. These dashboards allow analysts to systematically visualize and validate data using a straightforward user interface, and interactively mark data points that are suspect or invalid. The dashboards log all validation activities, display multiple plots to allow comparison among parameters or data sources, and support collaboration among multiple analysts when deployed to Shiny Server. In this presentation, we will provide an overview of these dashboards, highlighting useful features and the functions used we used to create them.



*Keywords:* environmental data, air quality, data visualization, interactive graphics

Presentation type: Lightning Talk

# Building a High Availability REST API Engine for R


## Nick Elprin

*Domino Data Lab*



**Abstract:** Modern businesses require APIs that have rock solid uptime, where deploying a new version never drops a request, where you can promote and roll back versions, and that perform with low latency and high throughput. Domino has built our R API endpoint functionality leveraging open source tools such as nginx and tresletech's plumber package, to support modern data science teams desire to reduce time from modeling to productionalization. In this talk, we discuss lessons we have learned building this functionality using the R ecosystem. We describe some of the technical challenges building such a platform, and some best practices for researchers who want to make their R models easily deployable as APIs. Domino's technology has served millions of requests for clients ranging from online media to energy companies. We will tell you how we did it.



*Keywords:* api, endpoints, REST, production, plumber

Presentation type: Lightning Talk

# Outlier Detection Methods


## Rajiv Shah

*University of Illinois at Chicago*



**Abstract:** This talk reviews some of the most relevant statistical approaches for outlier and anomaly detection in R. It covers statistical approaches, clustering based approaches, nearest neighbor approaches, random forest, and autoencoders. This talk will pull from a variety of existing R packages including DMwR, fclust, dbscan, isolation forest, and autoencoder.

The talk is relevant because outlier detection is a necessary step to clean the data and in other instances, the outliers may be of interest themselves. For example, identification of credit card fraud involves identifying outliers. While the need for outlier detection has increased, there has also been a rise in the number of techniques for identifying outliers.

This talk provides a theoretical background for each approach. This provides an understanding of the assumptions and limitations of each approach. This will then be demonstrated with examples of different datasets to show how performance varies for differing approaches. The specific methods include: Extreme Value, Expectation Maximization, Kmeans, Fuzzy Clustering, DBSCAN, Isolation Forests, an Autoencoders.

This talk includes sharing/demonstrating two interactive shiny apps that illustrate these algorithms. The first is for low dimensional data and is available at: http://projects.rajivshah.com/shiny/outlier/. The second is a shiny app that must be run locally on a computer (due to computational requirements), for higher dimension methods. By including a shiny application, this allows people to try these different methods out on datasets. In my experience, these sort of talks resonate much better, because the audience can be directly involved.



*Keywords:* Outlier and Anomaly detection, Shiny app

Presentation type: Lightning Talk

# Scalable semi-parametric regression with mgcv package and bam procedure

**Matteo Fasiolo, Yannig Goude, Raphaël Nedellec & Simon Wood**

*University of Bristol, EDF R&D, EDF R&D, University of Bristol*

**Abstract:** The mgcv package proposes a flexible framework for fitting Generalized additive regression models.

However, classical fitting procedure can be computationally intensive. The bam procedure brings about substantial computational savings, by adapting standard fitting algorithms to provide scalability to "big" data sets [1].

In particular, parallel approaches have been implemented to exploit multi-core architectures and to reduce memory footprint. We will present the results of joint work between the University of Bristol and one R&D team of EDF (the major French electrical utility). The new bam procedure has been used to model electrical load time series freely availablenfrom the NYC ISO. The new optimization algorithm (FREML) of bam allows the user to fit scalable additive models on data up to millions of observations and thousands of estimated parameters.

*Keywords:* Load forecasting, parallel, mgcv, semi-parametric regression

Presentation type: Lightning Talk

# Using R at a rapidly scaling healthcare technology startup


## Sandy Griffith & Josh Kraut

*Flatiron Health*



**Abstract:** Building technology at a rapidly scaling startup requires many trade-offs. This talk will discuss the learnings, challenges and opportunities of building a culture of R at a startup where R is not the primary language, and the doubling time of company size and scope is measured in months, not years. We will illustrate these concepts with examples from Flatiron Health, an oncology-focused technology company. Topics include using R in production, trade-offs between flexibility and scalability, short- vs. long-term solutions, and language choice on cross-functional teams. We will contrast examples where R was used for prototyping, but the final technical solution diverged away from R; cases where an R solution persisted over time; and scenarios where parallel development in R and an alternate medium were desired for audit controls. Through case studies, we outline a framework for evaluating these technical infrastructure decisions in a forward-thinking yet pragmatic manner. We'll also touch on the establishment of team- and company-wide best practices for R usage and strategies for building a sharable framework suitable for both new and advanced users.



*Keywords:* healthcare, technology, industry, oncology, culture, medicine

Presentation type: Lightning Talk

# Thinking about Energy Markets with interactivity


## Soumya Kalra

*Academic - Rutgers*



**Abstract:** This talk will focus on using the functionality within in R to show key risk metrics graphically in the energy industry using the shiny interface. This talk will serve as an introduction to both interactivity in R as well as quick overview of financial risk currently in energy trading markets specifically focused on spillover effects from oil to the financial markets overall. The link below is similar work done by me in the past using R and presented in Chicago.



*Keywords:* shiny, dashboard, energy markets, quantitative risk, energy data, oil markets, energy risk

Presentation type: Lightning Talk

# Understanding human behavior for applications in finance and social sciences: Insights from content analysis with novel Bayesian learning in R

**Stefan Feuerriegel, Nicolas Pröllochs & Dirk Neumann**

*Carnegie Mellon University, University of Freiburg, University of Freiburg*

**Abstract:** Research in finance and social sciences nowadays utilizes content analysis to understand human decisions in the face of textual materials. While content analysis has received great traction lately, the available tools are not yet living up to the needs of researchers. As our contribution, we propose, implement and demonstrate a novel approach to study tone, sentiment and reception of textual materials in R. Our approach utilizes Bayesian learning to extract words from documents that statistically feature a positive and negative polarity. This immediately reveals manifold implications for practitioners, finance research and social sciences: researchers can use R to extract text components that are relevant for readers and test their hypothesis based on these. On the other hand, practitioners can measure which wording actually matters to their readership and enhance their writing accordingly. We demonstrate the added benefits in two case studies from finance and social sciences. We also incorporate our algorithm together with common baselines for sentiment analysis in a new R package. It overcomes possible shortcomings in the existing choice of packages by providing a comprehensive toolset for sentiment analysis —supporting both a broad range of dictionary-based approaches and machine learning. Our R package effortlessly performs sentiment analysis of written materials and offers built-in functionality tailored for content analysis.

*Keywords:* Content Analysis, Information Extraction, Information Processing, Natural Language Processing, Sentiment Analysis, Text Mining, Bayesian Learning

Presentation type: Lightning Talk

# Performance Above Random Expectation: A more intuitive and versatile metric for evaluating probabilistic classifiers


## Stephen R Piccolo

*Brigham Young University*



**Abstract:** Many classification algorithms generate probabilistic estimates of whether a given sample belongs to a given class. Various scoring metrics have been developed to assess the quality of such probabilistic estimates. In many domains, the area under the receiver-operating-characteristic curve (AUC) is predominantly used. When applied to two-class problems, the AUC can be interpreted as the frequency at which two randomly selected samples are ranked correctly, according to their assigned probabilities. As its name implies, the AUC is derived from receiver-operating-characteristic (ROC) curves, which illustrate the relationship between the true positive rate and false positive rate. However, ROC curves—which have their roots in signal processing—are difficult for many people to interpret. For example, in medical settings, ROC curves can identify the probability threshold that achieves an optimal balance between over- and under-diagnosis for a particular disease; yet it is unintuitive to evaluate such thresholds visually. I have developed a scoring approach, Performance Above Random Expectation (PARE), which assesses classification accuracy at various probability thresholds and compares it against the accuracy obtained with random class labels. Across all thresholds, this information can be summarized as a metric that evaluates probabilistic classifiers in a way that is qualitatively equivalent to the AUC metric. However, because the PARE method uses classification accuracy as its core metric, it is more intuitively interpretable. It can also be used to visually identify a probability threshold that maximizes accuracy—thus effectively balancing true positives with false positives. This method generalizes to various other applications.



*Keywords:* machine learning, medical diagnosis, classification, bioinformatics

Presentation type: Lightning Talk

# madness: multivariate automatic differentiation in R


## Steven Elliot Pav

*Gilgamath Consulting*



**Abstract:** The madness package provides a class for automatic differentiation of 'multivariate' operations via forward accumulation. 'Multivariate' means the class computes the derivative of a vector or matrix or multidimensional array (or scalar) with respect to a scalar, vector, matrix, or multidimensional array. The primary intended use of this class is to support the multivariate delta method for performing inference on multidimensional quantities.



*Keywords:* automatic differentiation, delta method, statistical inference, CRAN package

Presentation type: Lightning Talk

# Getting R into your bathroom


## Torben Tvedebrink, Poul Svante Eriksen & Søren Buhl

*Department of Mathematical Sciences, Aalborg University*



**Abstract:** Have you ever considered how to use R when decorating your home? In this presentation I will show how R can be used to generate beautiful mathematical patterns based on complex numbers. In my case, we will use the patterns for the mosaics of the bathroom floor in our house.

The patterns are based on the work of the Danish statistician and actuarian Thorvald N. Thiele. Thiele was a pioneer in statistics and his contributions are described in full by Professor SteffennLauritzen in "Thiele: Pioneer in Statistics. Oxford University Press, 2002.". Thiele was among the founders of the first Danish insurance company, whose floor in the main entrance is designed by Thiele's own patterns.

The structures in the patterns comes from properties of the Gaussian integers, $Z[i]$, where $Z[i] = a + bi$ with $a$ and $b$ being integers. In the ggthiele (in progress) project we utilise ggplot2 to produce the patterns, which are based on quadratic residue classes. Some examples of the patterns can be seen at http://people.math.aau.dk/~tvede/useR2016/thiele.pdf.



*Keywords:* Complex numbers, Patterns, ggplot2

Presentation type: Lightning Talk

# MAVIS: Meta Analysis via Shiny


## William Kyle Hamilton & Burak Aydin

*University of California, Merced, Recep Tayyip Erdoğan University*



**Abstract:** We present a Shiny (RStudio & Inc., 2014) web application and R (R Core Team, 2015) package to simplify the process of running a meta-analysis using a variety of packages from the R community, including the popular metafor package (Viechtbauer, 2010). MAVIS (Hamilton, Aydin, and Mizumoto, 2014) was created to be used as a teaching tool for students and for scientists looking to run their own meta-analysis. Currently MAVIS supports both fixed and random effects models, methods for detecting publication bias, effect size calculators, single case design support, and generation of publication grade graphics. With this application we've created an open source browser based graphical user interface (GUI) which has lowered the barrier of entry for novice and occasional users.



*Keywords:* shiny, meta-analysis, education, metafor

Presentation type: Lightning Talk

# Maximum Monte Carlo likelihood estimation of conditional auto-regression models


## Zhe Sha

*University of Oxford*



**Abstract:** Likelihood of conditional auto-regression (CAR) models is expensive to compute even for a moderate data size around 1000 and it is usually not in closed form with latent variables. In this work we approximate the likelihood by Monte Carlo methods and propose two algorithms for optimising the Monte Carlo likelihood. The algorithms search for the maximum of the Monte Carlo likelihood and by taking the Monte Carlo error into account, the algorithms appear to be stable regardless the initial parameter value. Both algorithms are implemented in R and the iterative procedures are fully automatic with user-specified parameters to control the Monte Carlo simulation and convergence criteria.

We first demonstrate the use of the algorithms by simulated CAR data on a $20 \times 20$ torus. Then methods were applied to a data from forest restoration experiment with around 7000 trees arranged in transects in study plots. The growth rate of trees was modelled by a linear mixed effect model with CAR spatial error and CAR random effects. A approximation to the MLE was found by our proposed algorithms in a reasonable computational time.



*Keywords:* CAR models, Monte Carlo likelihood, response surface design, importance sampling, spatial statistics

# Part III

# Oral Presentation

Presentation type: Oral Presentation

# R markdown: Lifesaver or death trap?


## A. Jonathan R. Godfrey & Timothy P. Bilton

*Institute of Fundamental Sciences, Massey University*



**Abstract:** The popularity of R markdown is unquestionable, but will it prove as useful to the blind community as it is for our sighted peers? The short answer is "yes" but the more realistic answer is that it depends on so many other aspects some of which will remain outside the skill sets of many authors. Source R markdown files are plain text files, and are therefore totally accessible for a blind user. The documents generated from these source files for end-users differ in their accessibility; HTML is great and a pdf generated using LaTeX is very limited. International standards exist for ensuring most document formats are accessible, but the TeX ncommunity has not yet developed a tool for generating an accessible pdf document from any form of LaTeX source. There is little hope for any pdf containing mathematical expressions or graphical content. In contrast, the HTML documents created from R markdown can contain many aspects of accessibility with little or no additional work required from a document's author. A substantial problem facing any blind author wishing to create an HTML document from their R markdown files is that there is no simple editor available that is accessible; RStudio is not an option that can be used by blind people; until such time as an alternative tool becomes available, blind people will either have to use cumbersome work-arounds or rely on a small application we have built specifically for editing and processing R markdown documents.



*Keywords:* reproducible research, text editor, blind

Presentation type: Oral Presentation

# New Paradigms In Shiny App Development: Designer + Data Scientist Pairing


## Aaron Seth Horowitz

*McKinsey & Company*



**Abstract:** With the help of Shiny, advanced analytics practitioners have been liberated from professional application development constraints: long-turn development cycles, difficult interactions with IT groups unfamiliar with statistical modelling, challenges in making their content more accessible to broad audiences, and steep resource/time costs. However, this has pushed the burden of UX design, graphical presentation and scaling decisions onto the shoulders of the data scientist, who may or may not have a good background in these fields.

Now that supporting capabilities exist, such as packages that make interfacing with JavaScript visualization libraries easier (htmlwidgets) and the recent release of new shiny features, the work effort can be split, and much more compelling products can be produced. We plan to discuss a real-life example of creating a shiny application with HTML Templates, modules, etc. with support from a web-design expert. We'll describe the process of how we worked together to build basic prototypes, the benefits of shared work, and the challenges involved with such diverse skill sets. Finally, we'll show an example application built for a pricing and promotions model, and describe the impact this toolset had for us and our clients.



*Keywords:* shiny, application development, design, ux, visualization, analytics consulting, tool development

Presentation type: Oral Presentation

# Capturing and understanding patterns in plant genetic resource data to develop climate change adaptive crops using the R platform

**A. Bari, Y.P. Chaubey, M.J. Sillanpää, F.L. Stoddard, H. Khazaei, S. Dayanandan, A.B. Damania, S.B. Alaoui, H. Ouabbou, A. Jilal, M. Maatougui, M. Nachit, R. Chaabane & M. Mackay**

*Data and Image Analytics - DIM / CGIAR, Concordia University, University of Oulu, University of Helsinki, University of Saskatchewan, Concordia University, University of California, Davis, Institut Agronomique et Vétérinaire Hassan II, Institut National de la Recherche Agronomique (INRA), Institut National de Recherche Agronomique de Tunis (INRAT), University of Queensland*

**Abstract:** Genetic resources consist of genes and genotypes with patterns reflecting their dynamic adaption to changing environmental conditions. Detailed understanding of these patterns will significantly enhance the potential of developing crops with adaptive traits to climate change. Genetic resources have contributed in the past to about 50 percent increase in crop yields through genetic improvements, further improvement and development of climate change resilient crops will largely depend on these natural resources. However, the datasets associated with these resources are very large and consist mostly of records of single observations or/and continuous functions with limited information on key variables. Analysis of such complex and large datasets requires new mathematical conceptual frameworks, and a flexible evolving platform for a timely and continuous utilization of these resources to accelerate the identification of genetic material or genes that could be used for improving the resilience of food crops to climate change. In this global collaborative research and during the development of the theoretical framework, numerous modelling routines have been tested, including linear and nonlinear approaches on the R platform. The results were validated and used for the identification of sources of important traits such as drought, salinity and heat tolerance. This paper presents the conceptual framework with applications in R used in the identification of crop germplasm with climate change adaptive traits. The paper addresses the dynamics as well as the specificity of genetic resources data, which consists not only of records of mostly single observations but also functional data.

*Keywords:* Applied mathematics, R language platform, Crop genetic resources, Adaptation to climate change, Patterns in large datasets

Presentation type: Oral Presentation

# R: The last line of defense against bad debt.


## Alberto Martin Zamora

*McKinsey & Company*



**Abstract:** During the last decade, data has changed the behaviors of individuals and corporations alike. On the latter, Advanced Analytics has gained considerable momentum – not only as a source of competitive advantage in the short-term, but also the risk of becoming obsolete in the medium-term. In this context, data scientists cannot offer solutions on data-reach problems without leveraging the opportunities of statistical learning with a tool like R, which allows to rapidly transform prototypes into useful solutions, thanks to the functional nature of R. To be specific we will focus on a particular and pressing issue across industries, geographies and organizations: collections & bad debt. We will show how machine learning algorithms leveraging R helped shape better solutions on a "millennial" problem (i.e. how am I getting paid back?) During this talk, we will show how, with the help of R as our main power horse, we approach a collection problem from its inception to the actual business implementation. First, we will describe how we can preprocess the data that may be useful for the purpose of predicting which customers are going to fail to pay their bills. Then, we will explore the relationship between the past payment behavior of a customer and his ability to satisfy future obligations. Finally, we will conclude sharing briefly how the output of a prediction model can be translated into effective business strategies using a project we have been involved on recently as an example.



*Keywords:* Business, Machine Learning, Collections, Bad Debt

Presentation type: Oral Presentation

# Meta-Analysis of Epidemiological Dose-Response Studies with the dosresmeta R package


## Alessio Crippa & Nicola Orsini

*Karolinska Institutet*



**Abstract:** Quantitative exposures (e.g. smoking, alcohol consumption) in predicting binary health outcomes (e.g. mortality, incidence of a disease) are frequently categorized and modeled with indicator variables. Results are expressed as relative risks for the levels of exposure using one category as referent. Dose-response meta-analysis is an increasing popular statistical technique that aims to estimate and characterize an overall functional relation from such aggregated data. A common approach is to contrast the outcome risk in the highest exposure category relative to the lowest. A dose-response approach is more robust since it takes into account the quantitative values associated with the exposure categories. It provides a detailed description of how the risk varies throughout the observed range of exposure. Additionally, since all the exposure categories contribute to determine the overall relation, estimation is more efficient. Our aim is to give a short introduction to the methodological framework (structure of aggregated data, covariance of correlated outcomes, estimation and pooling of individual curves). We describe how to test hypothesis and how to quantify statistical heterogeneity. Alternative tools to flexibly model the quantitative exposure will be presented (splines and polynomials). We will illustrate modelling techniques and presentation of (graphical and tabular) results using the dosresmeta R package.



*Keywords:* dose-response, meta-analysis, heterogeneity, flexible

Presentation type: Oral Presentation

# Simulation of Synthetic Complex Data: The R-Package simPop


## Alexander Kowarik, Matthias Templ, Bernhard Meindl & Olivier Dupriez

*Statistics Austria, TU Vienna; Statistics Austria, Statistics Austria, World Bank*



**Abstract:** The production of synthetic datasets has been proposed as a statistical disclosure control solution to generate public use files out of protected data, and as a tool to create "augmented datasets" to serve as input for micro-simulation models. The performance and acceptability of such a tool relies heavily on the quality of the synthetic populations, i.e., on the statistical similarity between the synthetic and the true population of interest. Multiple approaches and tools have been developed to generate synthetic data. These approaches can be categorized into three main groups: synthetic reconstruction, combinatorial optimization, and model-based generation. We introduce simPop, an open source data synthesizer. SimPop is a user-friendly R-package based on a modular object-oriented concept. It provides a highly optimized S4 class implementation of various methods, including calibration by iterative proportional fitting and simulated annealing, and modeling or data fusion by logistic regression and other methods.



*Keywords:* microdata, simulation, synthetic data, population data, statistical disclosure control

Presentation type: Oral Presentation

# A spatial policy tool for cycling potential in England

**Ali Abbas, Nikolai Berkoff, Alvaro Ullrich, James Woodcock & Robin Lovelace**

*MRC Epidemiology Unit, University of Cambridge, Independent web developer, MRC Epidemiology Unit, University of Cambridge, MRC Epidemiology Unit, University of Cambridge, Institute for Transport Studies (ITS)/CDRC, University of Leeds*

**Abstract:** Utility cycling is an increasingly common objective worldwide. The Propensity to Cycle Tool (PCT) http://www.pct.bike is a planning support system created using open source software; including R (Shiny) for data processing and (Leaflet) interactive visualisation. The project is funded by the UK Department for Transport.

We have developed the sustainable transport planning package (stplanr). Given two points: origin and destination (OD), it displays a straight line connecting them. To get a route, it relies on two APIs GraphHopper and CycleStreets. The GraphHopper API is global, whereas CycleStreets is UK specific. Cyclestreets API incorporates hilliness, giving faster and quieter routes. We have used MapShapper library to simplify the boundaries of the geographical data (shape files).

A geographical based multi-layered application has been developed using Shiny and Leaflet packages. The PCT represents current cycling and cycling potential based on OD data from the England 2011 Census. Cycling potential and the corresponding health and environmental benefits are modelled as a function of route distance, hilliness and other factors at OD and area level. One of the main hurdles was to incorporate complex spatial big data sets, and allow multiple web-users to concurrently use the tool. In order to load, manipulate and interrogate the data, we use on-demand innovative mechanisms to visualize it.

This talk explains the design, build and deployment of the PCT with an emphasis on reproducibility (e.g. creation of the stplanr package for data pre-processing), scalability (solved with the new JavaScript interface package MapShapper) and lessons learned

*Keywords:* Big Data, Spatial Analysis, Reproducible Research, Open Source

Presentation type: Oral Presentation

# Transforming a Museum to be data-driven using R


## Alice Daish

*British Museum*



**Abstract:** With the exponential growth of data, more and more businesses are demanding to become data-driven. Seeking value from their data, big data and data science initiatives; jobs and skill sets have risen up the business agenda. R, being a data scientists' best friend, plays an important role in this transformation. But how do you transform a traditionally un-data-orientated business into being data-driven armed with R, data science processes and plenty of enthusiasm?

The first data scientist at a museum shares her experience on the journey to transform the 250-year-old British Museum to be data-driven by 2018. How is one of the most popular museums in the world, with 6.8 million annual visitors, using R to achieve a data-driven transition?

- Data wrangling
- Exploring data to make informed decisions
- Winning stakeholders' support with data visualisations and dashboard
- Predictive modelling
- Future uses including internet of things, machine learning etc.

Using R and data science, any organisation can become data driven. With data and analytical skills demand higher than supply, more businesses need to know that R is part of the solution and that R is a great language to learn for individuals wanting to get into data science.



*Keywords:* data-driven, R, data science, museum, data wrangling, modelling, data analysis, data visualisations,

Presentation type: Oral Presentation

# Integrated R labs for high school students

**Amelia McNamara, James Molyneux & Terri Johnson**

*Smith College*, *University of California, Los Angeles*, *University of California, Los Angeles*

**Abstract:** The Mobilize project developed a year-long high school level Introduction to Data Science course, which has been piloted in 27 public schools in the Los Angeles Unified School District. The curriculum is innovative in many ways, including the use of R and the associated curricular support materials. Broadly, there are three main approaches to teaching R. One has users learning to code in their browser (Code School and DataCamp), another has them working directly in the R console (swirl), and a final approach is to have students follow along with an external document (OpenIntro). The integrated R labs developed by Mobilize bridge between working at the console and following an instructional document. Through the mobilizr package, students can load labs written to accompany the course directly into the Viewer pane in RStudio, allowing them to work through material without ever leaving RStudio. By providing the labs as part of the curricular materials we reduce the burden on teachers and allow students to work at their own pace. We will discuss the functionality of the labs as they stand, as well as developments in the .Rpres format that could allow for even more interactive learning.

*Keywords:* Statistics education, statistical computing

Presentation type: Oral Presentation

# Introducing Statistics with intRo


## Andee Kaplan & Eric Hare

*Iowa State University*



**Abstract:** intRo is a modern web-based application for performing basic data analysis and statistical routines as well as an accompanying R package. Leveraging the power of R and Shiny, intRo implements common statistical functions in a powerful and extensible modular structure, while remaining simple enough for the novice statistician. This simplicity lends itself to a natural presentation in an introductory statistics course as a substitute for other commonly used statistical software packages, such as Excel and JMP. intRo is currently deployed at the URL http://www.intro-stats.com. In this talk, we describe the underlying design and functionality of intRo, including its extensible modular structure, illustrate its use with a live demo, and discuss future improvements that will enable a wider adoption of intRo in introductory statistics courses.



*Keywords:* R, Shiny, teaching, interactive

Presentation type: Oral Presentation

# Rethinking R Documentation: an extension of the lint package


## Andrew M Redd

*University of Utah*



**Abstract:** In this presentation I will present an extension to the lint package to assist with documentation of R objects. R is the de facto standard for literate programming thanks to packages such as. However, R still falls behind competing languages in the area of documentation. In Steve McConnell's classic Code Complete (2004) his first principle of commenting routines is "keep comments close to the code they describe." The native documentation system for R requires separate files. Packages have been developed that improve the situation, however unlike Doxygen, on which they are based, they do not allow full mixing code with documentation. I propose a paradigm shift for R documentation, which I have implemented in the R package lint. This strategy allows for several subtle changes in documentation, while seeking to preserve as much previous capability as is reasonable. First is to store documentation as an R object itself, allowing for documentation to be dynamically generated and manipulated in code. Documentation can also be kept as an attribute of the function or object that it documents and can exist independent from a package. The second change is that the documentation engine makes full use of the R parser. This integrates code with documentation comments and allows tailoring meaning to location. These extensions give more capability to programmers and users of R easing the burden of creating documentation. I welcome comments and discussion on the strategy of documentation and the direction of implementation.



*Keywords:* documentation, programming style

Presentation type: Oral Presentation

# FiveThirtyEight's Data Journalism Workflow With R


## Andrew Williams Flowers

*FiveThirtyEight*



**Abstract:** FiveThirtyEight is a data journalism site that uses R extensively for charts, stories, and interactives. We've used R for stories covering: p-hacking in nutrition science; how Uber is affecting New York City taxis; workers in minimum-wage jobs; the frequency of terrorism in Europe; the pitfalls in political polling; and many, many more.

R is used in every step of the data journalism process: for cleaning and processing data, for exploratory graphing and statistical analysis, for models deploying in real time as and to create publishable data visualizations. We write R code to underpin several of our popular interactives, as well, like the Facebook Primary and our historical Elo ratings of NBA and NFL teams. Heck, we've even styled a custom ggplot2 theme. We even use R code on long-term investigative projects.

In this presentation, I'll walk through how cutting-edge, data-oriented newsrooms like FiveThirtyEight use R by profiling a series of already-published stories and projects. I'll explain our use of R for chart-making in sports and politics stories; for the data analysis behind economics and science feature pieces; and for production-worthy interactives.



*Keywords:*

Presentation type: Oral Presentation

# CVXR: An R Package for Modeling Convex Optimization Problems

**Anqi Fu, Steven Diamond, Stephen Boyd & Balasubramanian Narasimhan**

*Stanford University*

**Abstract:** CVXR is an R package that provides an object-oriented modeling language for convex optimization. It allows the user to formulate convex optimization problems in a natural mathematical syntax rather than the restrictive standard form required by most solvers. The user specifies an objective and set of constraints by combining constants, variables, and parameters using a library of functions with known curvature and monotonicity properties. CVXR then applies signed disciplined convex programming (DCP) to verify the problem's convexity and, once verified, converts the problem into a standard conic form using graph implementations and passes it to an open-source cone solver such as ECOS or SCS. We demonstrate CVXR's modeling framework with several applications.

*Keywords:* Convex Optimization, Convex Analysis, Disciplined Convex Programming, Machine Learning, Numerical Methods

Presentation type: Oral Presentation

# R in machine learning competitions


## Anthony Goldbloom

*Kaggle*



**Abstract:** Kaggle is a community of almost 450K data scientists who have built almost 2MM machine learning models to participate in our competitions. Data scientists come to Kaggle to learn, collaborate and develop the state of the art in machine learning. This talk will cover some of the lessons from winning techniques, with a particular emphasis on best practice R use.



*Keywords:* R, machine learning, competitions, Kaggle

Presentation type: Oral Presentation

# Real-time analysis of the intraday financial volatility: Big data, simulations and stochastic volatility using R

**Antonio Alberto Santos**

*University of Coimbra, Portugal*

**Abstract:** Financial volatility is a key element for economic agents that make decisions in financial markets. To define the measures of volatility through financial models, data need to be collected; models need to be estimated; and the relevant results need to be presented in an integrated way. Using the capabilities of R, these tasks can be performed in an integrated form, allowing a more efficient use of the data, models and measures to characterize the volatility evolution in the financial markets. A package in R that integrates the three tasks of collecting and treating big financial datasets; estimating the models and defining relevant measures of volatility; and presenting the results in an intuitive and iterative form, is certainly useful. The capabilities of R to retrieve publicly available data from different sites and to organize the data conveniently are used to deal with big data sets. Estimation of the parameters within the stochastic volatility model, and forecasting of the volatility is usually done through Bayesian statistics and Markov chain Monte Carlo methods. A mix of code in R and C is used to accomplish these tasks. The presentation of the measures of volatility forecasts can take advantage of the resources available in R. This is done by an R Shiny web application. A package in R was developed to perform the three aforementioned tasks, and some of the main functions will be described in the presentation.

*Keywords:* Big data, Bayesian estimation, Forecasting, Particle filter, Markov chain Monte Carlo, Simulation, Stochastic volatility

Presentation type: Oral Presentation

# Efficient in-memory non-equi joins using data.table


## Arun Srinivasan

*Open Analytics*



**Abstract:** A join operation combines two (or more) tables on some shared columns based on a condition. An equi-join is a case where this combination condition is defined by the binary operator ==. It is a special type of $\theta$-join which consists of the entire set of binary operators: {=, ==}. This talk presents the recent developments in the data.table package to extend its equi-join functionality to any/all of these binary operators very efficiently. For example, X[Y, on = .(X.a >= Y.a, X.b Y.a, X.b < Y.a)] performs a range join. Many databases are fully capable of performing both equi and non-equi joins. R/Bioconductor packages IRanges and GenomicRanges contain efficient implementations for dealing with interval ranges alone. However, so far, there are no direct in-memory R implementations of non-equi joins that we are aware of. We believe this is an extremely useful feature that a lot of R users can benefit from.



*Keywords:* Non-equi joins, Interval joins, Range joins, In-memory, Large data, Performance

Presentation type: Oral Presentation

# ggduo: Pairs plot for two group data


## Barret Schloerke, Di Cook & Ryan Hafen

*Purdue University, Monash University, ryanhafen.com*



**Abstract:** The R package 'GGally' provides several amalgam plots that build on the basic 'ggplot2' plotting system. Functions produce multivariate plots like generalized scatterplot matrices, and parallel coordinate plots are provided, network plots, survival models, and glyph maps for spatiotemporal data. This new work introduces a function, ggduo, to produce generalized plots for two groups of variables (e.g. a matrix of X variables and a matrix of Y variables), as might be modeled by multivariate regression, canonical correlation analysis or even multivariate time series. It builds on the structure of the ggmatrix functions used to produce the generalized pairs plot. The new function will help analysts to look at their data to support better modeling.



*Keywords:* plot matrix pairs duo two group plotmatrix visualization ggplot2 GGally ggpairs ggduo ggmatrix viz

Presentation type: Oral Presentation

# DataSHIELD: Taking the analysis to the data

**Becca Wilson, Paul Burton, Demetris Avraam, Andrew Turner, Neil Parley, Oliver Butters, Tom Bishop, Amadou Gaye, Vincent Ferretti, Yannick Marcon, Jonathan Tedds & Simon Price**

*University of Bristol, University of Bristol, University of Bristol, University of Bristol, University of Bristol, University of Bristol, MRC Epidemiology Unit, University of Cambridge, National Institutes of Health, McGill University, McGill University, University of Leicester, University of Bristol, University of Bristol*

**Abstract:** Irrespective of discipline, data access and analysis barriers result from a range of scenarios:

- ethical-legal restrictions surrounding confidentiality and the sharing of, or access to, disclosive data;
- intellectual property or licensing issues surrounding research access to raw data;
- the physical size of the data is a limiting factor.

DataSHIELD (http://www.datashield.ac.uk) was born of the requirement in the biomedical and social sciences to co-analyse individual patient data from different sources, without disclosing sensitive information. DataSHIELD comprises a series of R packages enabling the researcher to perform distributed analysis on the individual level data, whilst satisfying the strict ethical-legal-governance restrictions related to sharing this data type. Furthermore, under the DataSHIELD infrastructure — set up as a client-server model — raw data never leaves the data provider (the server) and no individual level data can be seen by the researcher (the client). Base functionality in the DataSHIELD R packages includes descriptive stats (e.g. mean), exploratory stats (e.g. histogram), contingency tables (1-dimensional and 2-dimensional frequency tables) and modelling (survival analysis using piecewise exponential regression, glm). The modular nature of DataSHIELD has allowed the scoping of additional data types to expand DataSHIELD functionality with respect to genomic, text and geospatial data. Different infrastructure models are also possible — tailored for pooled co-analysis, single site analysis and linked data analysis. DataSHIELD has been successfully piloted in two European biomedical studies, sharing data across 14 different biobanks to investigate healthy obesity and the effect of environmental determinants on health. It is of proven value in the biomedical and social science domains, but has potential utility wider than this.

*Keywords:* distributed R, privacy-protected data analysis, disclosure, sensitive data

Presentation type: Oral Presentation

# RServer: Operationalizing R at Electronic Arts


## Ben Weber

*Electronic Arts*



**Abstract:** The motivation for the RServer project is the ability for data scientists at Electronic Arts to offload R computations from their personal machines to the cloud and to enable modeling at scale. The outcome of the project is a web-based tool that our scientists can use to automate running R scripts on virtual machines and perform a variety of reporting and analysis tasks. We are using RServer to operationalize data science at EA.

The core of RServer is a java application that leverages WAMP to provide a web front-end for managing deployed R scripts. At Electronic Arts, we host RServer instances on our on-demand infrastructure and can spin up new machines as necessary to support new products and analyses. In order to deploy scripts to the server, team members check in their R scripts and supporting files to Perforce and modify the server's schedule file.

In addition to the web tool, we've developed an internal R package that provides functionality for connecting to our data sources, password management, and additional features for RServer, such as making R Markdown reports accessible via URLs. Some uses of RServer include running ETLs, creating and emailing R Markdown reports, and hosting dashboards and Shiny applications. This infrastructure enables us to power analytics in a way that is usually reserved for tools like Tableau, while utilizing the full power of R. This presentation will include a live demo of RServer. We are releasing an open-source version of RServer at useR! that supports Git.



*Keywords:* R Tools, Operationalizing R, Cloud Computing, Version Control, R Markdown, Git

Presentation type: Oral Presentation

# ETL for Medium Data


## Ben S Baumer

*Smith College*



**Abstract:** Packages provide users with software that extends the core functionality of R, as well as data that illustrates the use of that functionality. However, by design the type of data that can be contained in an R package on CRAN is limited. First, packages are designed to be small, so that the amount of data stored in a package is supposed to be less than 5 megabytes. Furthermore, these data are static, in that CRAN allows only monthly releases. Alternative package repositories – such as GitHub – are also limited in their ability to store and deliver data that could be changing in real-time to R users. The etl package provides a CRAN-friendly framework that allows R users to work with medium data in a responsible and responsive manner. It leverages the dplyr package to facilitate Extract-Load-Transfer (ETL) operations that bring real-time data into local or remote databases controllable by R users who may have little or no SQL experience. The suite of etl-dependent packages brings the world of medium data – too big to store in memory, but not so big that it won't fit on a hard drive – to a much wider audience.



*Keywords:* medium data, etl, dplyr, sql

Presentation type: Oral Presentation

# How to do one's taxes with R


## Benno Süselbeck

*University of Muenster, Center for Information Processing*



**Abstract:** In this talk it is shown how to generate a return of tax (German VAT) with R and send it over the internet to the tax administration. As this is certainly not a standard application for R (special software exists for this purpose) it may be worthwhile to have a closer look at the techniques used to realize such kind of transaction and to reveal any analogies to distributed data analysis. If confidential data cannot be analysed in the environment where it is created or stored, it has to be transferred over the internet to some kind of nexecution service, e.g. a cluster system. Encryption is necessary to protect the data as well as appending a digital signature to guarantee ownership nand prevent modification. Additionally some kind of packaging has to be applied to the data together with metadata giving directions for the receiver to handle the delivery. When returning the result the same techniques are used. So again privacy and authorship are ensured. For the tax example all these procedures have to observe well established cryptographic standards for encryption, hashing and digital signatures which change from time to time according to new results in cryptographic research. I demonstrate an implemenation in R for this kind of transaction in a data science context, trying to use the same rigorous standards mentioned above whenever possible. This leads to an overview of existing R packages and external software useful and necessary to realize a corresponding program. Finally some proposals for a possible standardization of a secure distributed data analysis scenario are presented.



*Keywords:* Secure Data Analysis, Distributed Data Analysis, Cryptography, Cryptographic Standards, Encryption, Digital Signature, XML Schema, XML Signature

Presentation type: Oral Presentation

# Group and Sparse Group Partial Least Square Approaches Applied in Genomics Context

**Benoit Liquet, Pierre Lafaye de Micheaux, Boris Hejblum & Rodolphe Thiebaut**

*University Pau et Pays de L'Adour, ACEMS: Centre of Excellence for Mathematical and Statistical Frontiers, QUT, Australia, CREST, ENSAI, Campus de Ker-Lann, Inria, SISTM, Talence; Inserm, U897, Bordeaux;Vaccine Research Institute, Inria, SISTM, Talence; Inserm, U897, Bordeaux; Vaccine Research Institute*

**Abstract:** In this talk, I will concentrate on a class of multivariate statistical methods called Partial Least Squares (PLS). They are used for analysing the association between two blocks of 'omics' data, which bring challenging issues in computational biology due to their size and complexity. In this framework, we will exploit the knowledge on the grouping structure existing in the data, which is key to more accurate prediction and improved interpretability. For example, genes within the same pathway have similar functions and act together in regulating a biological system. In this context, we developed a group Partial Least Squares (gPLS) method and a sparse gPLS (sgPLS) method. Our methods available through our sgPLS R package are compared through an HIV therapeutic vaccine trial. Our approaches provide parsimonious models to reveal the relationship between gene abundance and the immunological response to the vaccine.

*Keywords:* Partial Least Square, Group Sparsity, Variable Selection, SVD

Presentation type: Oral Presentation

# Connecting R to the OpenML project for Open Machine Learning

**Bernd Bischl, Jakob Bossek, Giuseppe Casalicchio, Benjamin Hofner, Pascal Kerschke, Dominik Kirchhoff, Michel Lang, Heidi Sebold & Joaquin Vanschoren**

*LMU Munich, University of Münster, Friedrich-Alexander University Erlangen, University of Münster, TU Dortmund, TU Dortmund, University of Zurich, Eindhoven University of Technology*

**Abstract:** OpenML is an online machine learning platform where researchers can automatically log and share data, code, and experiments, and organize them online to work and collaborate more effectively. We present an R package to interface the OpenML platform and illustrate its usage both as a stand-alone package and in combination with the mlr machine learning package. We show how the OpenML package allows R users to easily search, download and upload machine learning datasets. Users can easily log their auto ML experiment results online, have them evaluated on the server, share them with others and download results from other researchers to build on them. Beyond ensuring reproducibility of results, it automates much of the drudge work, speeds up research, facilitates collaboration and increases user's visibility online. Currently, OpenML has 1,000+ registered users, 2,000+ unique monthly visitors, 2,000+ datasets, and 500,000+ experiments. The OpenML server currently supports client interfaces for Java, Python, .NET and R as well as specific interfaces for the WEKA, MOA, RapidMiner, scikit-learn and mlr toolboxes for machine learning.

*Keywords:* machine learning, data repositories, reproducible reserach, meta learning, experimental databases

Presentation type: Oral Presentation

# The options and challenges of spatial analysis in R


## Bhaskar Vishnu Karambelkar

*Northwestern University / ThreatConnect Inc.*



**Abstract:** A bird's eye overview of spatial analysis on R with a focus on cartography. Spatial analysis is playing a big role in all forms of industries and domains, from civic planning, to weather predictions to climate change monitoring. It is equally important in tactical operations and strategic ones and one of the most interesting domains of the 21st century.

What are the classical and modern spatial analysis options available in R? Where does cartography and mapping fit into the picture? How does R compare to dedicated software like ArcGIS, QGis? What options does R provide for interactive, web mapping and analysis? Is R suitable for processing lots of data? These are some of the questions someone looking to perform spatial analytics in R will have.

This talk is aimed at novice GIS professionals looking to use R for spatial analysis as well as seasoned pros of other GIS software looking to use R for their spatial analytics needs. The talk will review the options R provides for spatial analysis and look at the strengths and weakness of R as a spatial analysis tool. I will cover cartography/mapping options as well as spatial analysis model in brief. I will also examine how spatial analysis can be augmented by pulling data from various open APIs.



*Keywords:* Spatial Analysis, Cartography, mapping

Presentation type: Oral Presentation

# Multivoxel Pattern Analysis of fMRI Data


## Bradley Russell Buchsbaum

*Rotman Research Institute, Baycrest*



**Abstract:** Analysis of functional magnetic resonance imaging (fMRI) data has traditionally been carried out by analyzing each voxel's time-series independently with a linear model. While this approach has been effective for creating statistical maps of brain activity, recent work has show that greater sensitivity to distributed neural signals can be achieved with multivariate approaches that analyze patterns of activity rather than methods that work only on voxel at a time. This has led to an explosion of interest in so-called "multivoxel pattern analysis" (MVPA) which is essentially the application of machine learning algorithms to neuroimaging data. The R programming environment is well-suited for MVPA analyses due to its large and varied support for statistical learning methods available on CRAN. Many of these methods are can be conveniently accessed using a standard interface provided by the 'caret' library. Here we present a new library (rMVPA) that makes MVPA analyses of fMRI data available to R users by leveraging the 'caret' and 'neuroim' packages. The rMVPA analyses implements multiple methods for multivariate analysis of fMRI data including the spherical searchlight method, region of interest analyses, and a new hierarchical ensemble approach to MVPA.



*Keywords:* fMRI, MVPA, machine learning, medical images, neuroimaging, neuroscience

Presentation type: Oral Presentation

# Simulation and power analysis of generalized linear mixed models


## Brandon LeBeau

*University of Iowa*



**Abstract:** As computers have improved, so has the prevalence of simulation studies to explore implications for assumption violations and explore statistical power. The simglm package allows for flexible simulation of general(ized) linear mixed models (multilevel models) under cross-sectional or longitudinal frameworks. In addition, the package allows for different distributional assumptions to be made such as non-normal residuals and random effects, missing data, and serial correlation. A power analysis by simulation can also be conducted by specifying a model to be simulated and the number of replications. This package can be useful for instructors or students for courses involving the general(ized) linear mixed model, as well as researchers looking to conduct simulations exploring the impact of assumption violations. The focus of the presentation will be on showing how to use the package, including live demos of the varying inputs and outputs, with working code. In addition to the syntax, a Shiny application will be made to show how the features can be made accessible to students in the classroom that are unfamiliar with R. The Shiny application will also provide a nice use case for the package, a live vignette of sorts.



*Keywords:* simulation, linear mixed model, power analysis

Presentation type: Oral Presentation

# Practical tools for exploratory web graphics


## Carson Sievert

*Iowa State University*



**Abstract:** Interactive statistical graphics toolkits play an important role in the exploratory phase of a data analysis cycle. Web graphics are rarely used during this phase, and are commonly reserved solely for the presentation of findings, mainly due to a lack of tools for quick iteration. The R package ggplot2 is a popular tool for data visualization with an elegant framework allowing useRs to quickly iterate through many plots with a minimal amount of friction. The R package plotly converts ggplot2 graphics to a web-based version, adding automatic support for interactivity such as pan, zoom, and identification (i.e., tooltips). It also has support for more advanced interactive techniques, such as linked brushing, thanks to infrastructure provided by the R packages shiny and htmlwidgets. In this talk, I'll present numerous examples that demonstrate these techniques, and how to they can used to derive insights from data.



*Keywords:* plotly, ggplot2, shiny, EDA

Presentation type: Oral Presentation

# How to keep your R code simple while tackling big datasets


## Charles Arthur Piercey

*TidalScale, Inc.*



**Abstract:** Like many statistical analytic tools, R can be incredibly memory intensive. A simple GAM (generalized additive model) or K-nearest neighbor routine can devour many multiples of memory size compared to the starting dataset. And, R doesn't always behave nicely when it runs out of memory.

There are techniques to get around memory limitations, like using partitioning tools or sampling down. But these require extra work. It would be really nice to run elegantly simple R analytics without that hassle.

Using a really big, public dataset, from CMS.gov, Chuck will show GAM, GLM, Decision Trees, Random Forest and K Nearest Neighbor routines that were prototyped and run on a laptop then run unchanged on a single simple Linux instance with over a Terabyte of RAM against the entire dataset. This big computer is actually a collection of smaller off-the-shelf servers using TidalScale to create a single, virtual server with several terabytes of RAM.



*Keywords:* R, analytics, big data

Presentation type: Oral Presentation

# Compiling parts of R using the NIMBLE system for programming algorithms

**Christopher J. Paciorek, Perry de Valpine & Daniel Turek**

*UC Berkeley*

**Abstract:** The NIMBLE R package provides a flexible system for programming statistical algorithms for hierarchical models specified using the BUGS language. As part of the system, we compile R code for algorithms and seamlessly link the compiled objects back into R, with our focus being on mathematical operations. Our compiler first generates C++, including Eigen code for linear algebra, before the usual compilation process. The NIMBLE compiler was written with extensibility in mind, such that adding new operations for compilation requires only a few well-defined additions to the code base. We'll describe how one can easily write functions in R and automatically compile them, as well as how the compiler operates behind the scenes. Functions can be stand-alone functions or can be functions that interact with hierarchical models written in BUGS code, which NIMBLE converts to a set of functions and data structures that are also compiled via C++. Finally, we'll show how the system has been used to build a full suite of MCMC and sequential Monte Carlo algorithms that can be used on any hierarchical model.

*Keywords:* compilation, domain specific language, algorithms

Presentation type: Oral Presentation

# Crowd sourced benchmarks


## Colin Stevenson Gillespie

*Newcastle University*



**Abstract:** One of the simplest ways to speed up your code is to buy a faster computer. While this advice is certainly trite, it is something that should still be considered. However it is often unclear to determine the benefit of upgrading your system. The 'benchmarkme' package aims to tackle this question by allowing users to benchmark their system and compare their results with other users. This talk will discuss the results of this benchmarking exercise. Additionally we'll provide practical advice about how to move your system up the benchmark rankings through byte compiling and using alternate BLAS libraries.



*Keywords:* Efficiency, speed, cpu

Presentation type: Oral Presentation

# Continuous Integration and Teaching Statistical Computing with R


## Colin Witter Rundel

*Dept of Statistical Science, Duke University*



**Abstract:** In this talk we will discuss two statistical computing courses taught as part of the undergraduate and masters curriculum in the Department of Statistical Science at Duke University. The primary goal of these courses is to teach advanced R along with modern software development practices. In this talk we will focus in particular on our adoption of continuous integration tools (github and wercker) as a way to automate and improve the feedback cycle for students as they work on their assignments. Overall, we have found that these tools, when used appropriately, help reduce learner frustration, improves code quality, reduces instructor workload, and introduces powerful tools that are relevant long after the completion of the course. We will discuss several of the classes' open-ended assignments and explore instances where continuous integration made sense and well as cases where it did not.



*Keywords:* education, continuous integration, docker, github, wercker

Presentation type: Oral Presentation

# FlashR: Enable Parallel, Scalable Data Analysis in R

**Da Zheng, Joshua Vogelstein, Carey E. Priebe & Randal Burns**

*Johns Hopkins University*

**Abstract:** In the era of big data, R is rapidly becoming one of the most popular tools forndata analysis. But the R framework is relatively slow and unablento scale to large datasets. The general approach of speeding up an implementation in R is to implement the algorithms in C or FORTRAN and provide an R wrapper. There are many works that parallelize R andnscale it to large datasets. For example, Revolution R Open parallelizes a limited set of matrix operations individually, which limits its performance. Others such as Rmpi and R-Hadoop exposes low-level programmingn-interface to R users and require more explicit parallelization. It is challenging to provide a framework that has a high-level programming interface while achieving efficiency. FlashR is a matrix-oriented R programming framework that supports automatic parallelization and out-of-core execution for large datasets. FlashR reimplements matrix operations in the R base package and provides some generalized matrix operations to improve expressiveness. FlashR automatically fuses matrix operations to reduce data movement between CPU and disks. We implement machine learning algorithms such as Kmeans and GMM in FlashR to benchmark its performance. On a large parallelnmachine, both in-memory and out-of-core execution of these R implementations in FlashR significantly outperforms the ones in Spark Mllib. We believe FlashR significantly lowers the expertise for writing parallel and scalable implementations of machine learning algorithms and provides new opportunities for large-scale machine learning in R. FlashR is implemented as an R package and is released as open source (http://flashx.io/).

*Keywords:* Auto-parallelization, out-of-core execution, large-scale machine learning,

Presentation type: Oral Presentation

# The challenge of combining 176 x #otherpeoplesdata to create the Biomass And Allometry Database (BAAD)

**Daniel Stein Falster, Richard G FitzJohn, Remko A Duursma & Diego Darneche**

*Biological Sciences, Macquarie University, Australia, Imperial College, London, Hawkesbury Institute for the Environment, Western Sydney University, Australia, Monash University, Australia*

**Abstract:** Despite the hype around “big data", a more immediate problem facing many scientific analyses is that large-scale databases must be assembled from a collection of small independent and heterogeneous fragments – the outputs of many and isolated scientific studies conducted around the globe. Together with 92 other co-authors, we recently published the Biomass And Allometry Database (BAAD) as a data paper in the journal Ecology, combining data from 176 different scientific studies into a single unified database. BAAD is unique in that the workflow – from raw fragments to homogenised database – is entirely open and reproducible. In this talk I introduce BAAD and illustrate solutions (using R) for some of the challenges of working with and distributing lots and lots of #otherpeople’s data.

*Keywords:* Data aggregation, R, reproducible

Presentation type: Oral Presentation

# An embedded domain-specific language for ODE-based drug-disease modeling and simulation


## David A James, Wenping Wang & Melissa Hallow

*Novartis, Novartis, University of Georgia*



**Abstract:** We present a domain-specific mini-language embedded in R for expressing pharmacometrics and drug-disease models. Key concepts in our RxODE implementation include

- A simple syntax for specifying models in terms of ordinary differential equations (ODE).
- A compilation manager to translate, compile, and load machine code into R for fast execution.
- An 'eventTable' closure object to express inputs/perturbations into the underlying dynamic system being modeled.
- Model reflectance to describe a model's structure, parameters, and derived quantities (useful for meta-programming, e.g., automatic generation of shiny applications).

We present examples in the design of complex drug dosing regimens for first-in-human studies via simulations, the modeling of unabated Alzheimer disease progression, and time-permitting, the modeling of visual acuity among age-related macular degeneration patients in the presence of disease-mitigating therapies. We also compare our approach in RxODE to similar work, namely, deSolve, mrgsolve, mlxR, nlmeODE, and PKPDsim. In closing, we use this 40th anniversary of the S language to reflect on the remarkably solid modeling framework laid out almost 25 years ago in "Statistical Models in S" (Chambers and Hastie (1992)), and to identify new challenges for specifying and fitting increasingly more complex statistical models, such as models of dynamic systems (as above), models for multi-state event history analysis, Bayesian data analysis, etc.



*Keywords:* Programming languages, parsing, differential equations, PK/PD

Presentation type: Oral Presentation

# broom: Converting statistical models to tidy data frames


## David Garrett Robinson

*Stack Overflow*



**Abstract:** The concept of “tidy data" offers a powerful and intuitive framework for structuring data to ease manipulation, modeling and visualization, and has guided the development of R tools such as ggplot2, dplyr, and tidyr. However, most functions for statistical modeling, both built-in and in third-party packages, produce output that is not tidy, and that is therefore difficult to reshape, recombine, and otherwise manipulate. I introduce the package “broom,” which turns the output of model objects into tidy data frames that are suited to further analysis and visualization with input-tidy tools. The package defines the tidy, augment, and glance methods, which arrange a model into three levels of tidy output respectively: the component level, the observation level, and the model level. These three levels can be used to describe many kinds of statistical models, and offer a framework for combining and reshaping analyses using standardized methods. Along with the implementations in the broom package, this offers a grammar for describing the output of statistical models that can be applied across many statistical programming environments, including databases and distributed applications.



*Keywords:* computing, data wrangling, modeling, tidy data

Presentation type: Oral Presentation

# R at Microsoft


## David Mark Smith

*Microsoft*



**Abstract:** Since the acquisition of Revolution Analytics in April 2015, Microsoft has embarked upon a project to build R technology into many Microsoft products, so that developers and data scientists can use the R language and R packages to analyze data in their data centers and in cloud environments.

In this talk I will give an overview (and a demo or two) of how R has been integrated into various Microsoft products. Microsoft data scientists are also big users of R, and I'll describe a couple of examples of R being used to analyze operational data at Microsoft. I'll also share some of my experiences in working with open source projects at Microsoft, and my thoughts on how Microsoft works with open source communities including the R Project.



*Keywords:* Microsoft R, business, applications

Presentation type: Oral Presentation

# SpatialProbit – for fast and accurate spatial probit estimations.


## Davide Martinetti & Ghislain Geniaux

*INRA, Centre de recherche Provence-Alpes-Côte d'Azur, UR ECODEVELOPPEMENT*



**Abstract:** This package meets the emerging needs of powerful and reliable models for the analysis of spatial discrete choice data. Since the explosion of available and voluminous geospatial and location data, existing estimation techniques cannot withstand the course of dimensionality and are restricted to samples counting having less than a few thousand observations.

The functions contained in SpatialProbit allow fast and accurate estimations of Spatial Autoregressive and Spatial Error Models under Probit specification. They are based on the full maximization of likelihood of an approximate multivariate normal distribution function, a task that was considered as prodigious just seven years ago (Wang et al. 2009). Extensive simulation and empirical studies proved that these functions can readily handle sample sizes with as many as several millions of observations, provided the spatial weight matrix is in convenient sparse form, as is typically the case for large data sets, where each observation neighbours only a few other observations.

SpatialProbit relies amongst others on Rcpp, RcppEigen and Matrix packages to produce fast computations for large sparse matrixes.

Possible applications of spatial binary choice models include spread of diseases and pathogens, plants distribution, technology and innovation adoption, deforestation, land use change, amongst many others.

We will present the results of the SpatialProbit package for a large database on land use change at the plot level.



*Keywords:* Spatial statistics, discrete choice model, probit, land use

Presentation type: Oral Presentation

# Adding R, Jupyter and Spark to the toolset for understanding the complex computing systems at CERN's Large Hadron Collider

**Dirk Duellmann**

*CERN*

**Abstract:** High Energy Physics (HEP) has a decades long tradition of statistical data analysis and of using large computing infrastructures. CERN's current flagship project LHC has collected over 100 PB of data, which is analysed in a wold-wide distributed computing grid by millions of jobs daily. Being a community with several thousand scientists, HEP also has a tradition of developing its own analysis toolset. In this contribution we will briefly outline the core physics analysis tasks and then focus on applying data analysis methods also to understand and optimise the large and distributed computing systems in the CERN computer centre and the world-wide LHC computing grid. We will describe the approach and tools picked for the analysis of metrics about job performance, disk and network I/O and the geographical distribution and access to physics data. We will present the technical and non-technical challenges in optimising a distributed infrastructure for large scale science projects and will summarise the first results obtained.

*Keywords:* statistical analysis, large computing infrastructures, optimisation, distributed systems

Presentation type: Oral Presentation

# Extending CRAN packages with binaries: x13binary


## Dirk Eddelbuettel & Christoph Sax

*Debian and R Projects, Christoph Sax Data Analytics*



**Abstract:** The x13binary package provides pre-built binaries of X-13ARIMA-SEATS, the seasonal adjustment software by the U.S. Census Bureau. X-13 is anwell-established tool for de-seasonalization of timeseries, and used bynstatistical offices around the world. Other packages such as seasonal can now rely on x13binary without requiring any intervention by the user. Together, these packages bring a very featureful and expressive interface for working with seasonal data to the R environment. Thanks to x13binary, installing seasonal is now as easy as any other CRAN package as it no longer requires a manual download and setup of the corresponding binary. Like the Rblpapi package, x13binary provides interesting new ways inndeploying binary software to aid CRAN: A GitHub repository provides the underlying binary in a per-operating system form (see the x13prebuilt repository). The actual CRAN package then uses this repo to download andninstall the binaries once per installation or upgrade. This talk will detail our approach, summarize our experience in providing binaries via CRAN and GitHub, and discuss possible future directions.



*Keywords:* X-13ARIMA-SEATS, time series, deseasonalization, package, CRAN, binary, distribution

Presentation type: Oral Presentation

# Providing Digital Provenance: from Modeling through Production


## Nick Elprin & Eduardo Ariño de la Rubia

*Domino Data Lab*



**Abstract:** Reproducibility is important throughout the entire data science process. As recent studies have shown, subconscious biases in the exploratory analysis phase of a project can have vast repercussions over final conclusions. The problems with managing the deployment and life-cycle of models in production are vast and varied, and often reproducibility stops at the level of the individual analyst. Though R has best in class support for reproducible research, with tools like KnitR to packrat, they are limited in their scope. In this talk we present a solution we have developed at Domino, which allows for every model in production to have full reproducibility from EDA to the training run and exact datasets which were used to generate. We discuss how we leverage Docker as our reproducibility engine,



*Keywords:* reproducibility, docker, training, modeling

Presentation type: Oral Presentation

# Spatial data in R: simple features and future perspectives


## Edzer Pebesma & Roger Bivand

*University of Muenster, Norwegian School of Economics*



**Abstract:** Simple feature access is an open standard for handling feature data (mostly points, lines and polygons) that has seen wide adoption in databases, javascript, and linked data. Currently, R does not have a complete solution for reading, handling and writing simple feature data. With funding from the R consortium, we will implement support for simple features in R. This talk discusses the challenges and potential benefits when doing so. It also points out challenges when analysing attribute data associated with simple features. In particular, the question whether a property refers to a property at every location of a feature (such as the land cover of a polygonndelineating a forest) or merely to a summary statistic computed over the whole feature (such as population count over a county) is to be solved. Its consequences for further analysis and data integration will be illustrated, and solutions will be discussed.



*Keywords:* spatial statistics, spatial databases, spatial data

Presentation type: Oral Presentation

# "AF" a new package for estimating the attributable fraction


## Elisabeth Eva britt Dahlqwist

*Karolinska Institute*



**Abstract:** The attributable fraction (or attributable risk) is a widely used measure that quantifies the public health impact of an exposure on an outcome. Even though the theory for AF estimation is well developed, there has been a lack of up-to-date software implementations. The aim of this article is to present a new R package for AF estimation with binary exposures. The package AF allows for confounder-adjusted estimation of the AF for the three major study designs: cross-sectional, (possibly matched) case-control and cohort. The article is divided into theoretical sections and applied sections. In the theoretical sections we describe how the confounder-adjusted AF is estimated for each specific study design. These sections serve as a brief but self-consistent tutorial in AF estimation. In the applied sections we use real data examples to illustrate how the AF package is used. All datasets in these examples are publicly available and included in the AF package, so readers can easily replicate all analyses.



*Keywords:* R package, attributable fraction, causal inference, epidemiology

Presentation type: Oral Presentation

# Analysis of big biological sequence datasets using the DECIPHER package


## Erik Scott Wright

*University of Wisconsin - Madison*



**Abstract:** Recent advances in DNA sequencing have led to the generation of massive amounts of biological sequence data. As a result, there is an urgent need for packages that assist in organizing and evaluating large collections of sequences. The DECIPHER package enables the construction of databases for curating sequence sets in a space-efficient manner. Sequence databases offer improved organization and greatly reduce memory requirements by allowing subsets of sequences to be accessed independently. Using DECIPHER, sequences can be imported into a database, explored, viewed, and exported under non-destructive workflows that simplify complex analyses. For example, DECIPHER workflows could be used to quickly search for thousands of short sequences (oligonucleotides) within millions of longer sequences that are contained in a database. DECIPHER also includes state-of-the-art functions for sequence alignment, primer/probe design, sequence manipulation, phylogenetics, and other common bioinformatics tasks. Collectively, these features empower DECIPHER users to handle big biological sequence data using only a regular laptop computer.



*Keywords:* DNA, bioinformatics, databases

Presentation type: Oral Presentation

# Revolutionize how you teach and blog: add interactivity


## Filip Schouwenaars

*DataCamp*



**Abstract:** R vignettes, blog posts and teaching materials are typically standard web pages generated with R Markdown. DataCamp has developed a framework to make this static content interactive: R code chunks are converted into an R-session backed editor so readers can experiment. This talk will explain the inner workings of the technology, as well as a the tutorial R package that makes the transition to interactive web pages seamless. Some hands-on examples will showcase the remarkable ease with which you can convert R Markdown documents, vignettes and Jekyll-powered blogs into interactive R playgrounds



*Keywords:* reporting, blogging, teaching, interactivity, markdown, html

Presentation type: Oral Presentation

# Automating our work away: one consulting firm's experience with KnitR.


## Finbarr Timbers

*Darkhorse Analytics*



**Abstract:** As consultants, many of the projects that we work on are similar, with many steps repeated verbatim across projects. Previously, our workflow was based largely in Microsoft Office, with our analysis done manually in Excel, our reports written in Word, and our presentations in Powerpoint. In 2015, we began using R for much of our analysis, including making slide decks and reports in RMarkdown.

Our presentation discusses why we made the change, how we managed it, and advice for other consulting firms looking to do the same.



*Keywords:* RMarkdown, KnitR, Industry, Productivity, Automation

Presentation type: Oral Presentation

# Tools for Robust R Packages


## Gábor Csárdi

*Mango Solutions, UK*



**Abstract:** Building an R package is a great way of encapsulating code, documentation and data, in a single testable and easily distributable unit. At Mango we are building R packages regularly, and have been developing tools that ease this process and also ensure a high quality, maintainable software product. I will talk about some of them in this presentation. Our goodPractice package gives advice on good package building practices. It finds unsafe functions like sapply and sample; it calculates code complexity measures and draws function call graphs. It also incorporates existing packages for test coverage (covr) and source code linting (lintr). It can be used interactively, or in a continuous integration environment. The argufy package allows writing declarative argument checks and coercions for function arguments. The checking code is generated and included automatically. The progress package allows adding progress bars to loops and loop-like constructs (lapply, etc.) with minimal extra code and minimal runtime overhead. The pkgconfig package provides a configuration mechanism in which configuration settings from one package does not interfere with settings from another package.



*Keywords:* package development, package linter

Presentation type: Oral Presentation

# viztrackr: Tracking and discovering plots via automatic semantic annotations


## Gabriel Becker, Sara Moore & Michael Lawrence

*Genentech Research, University of California, Berkeley, Genentech Research; R-core development team*



**Abstract:** Data analyses often produce many different data visualizations. Keeping track of these plots is crucial for both correctness and reproducibility of analytic results. Analysts typically resort to direct use of filenames and paths to organize and label their plots. Unfortunately, such ad hoc approaches do not scale well to longer and more complex analyses. Furthermore, locating specific plots months or years after the fact, when the chosen naming scheme has likely been forgotten, can be time consuming and painful. We propose a system which automatically tracks visualizations and annotates them with meaningful, searchable metadata. Beyond the benefits to individual analysts, the ability to search through plots created by others to discover analyses relevant to a particular dataset or research question is a powerful tool for facilitating collaboration and advancing science within multi-analyst, multi-project research departments and the wider scientific community. We present the viztrackr framework, a tool for tracking, automatically annotating, discovering, and reproducing statistical plots created in the R statistical programming language.



*Keywords:* computing, graphics, reproducibility, discoverability, provenance

Presentation type: Oral Presentation

# swirl-tbp: a package for interactively learning R programming and data science through the addition of "template-based practice" problems in swirl

**Kyle Marrotte & Garrett M. Dancik**

*Eastern Connecticut State University*

**Abstract:** The R package 'swirl' allows users to learn R programming by completing interactive lessons within the R console. Lessons (written in the YAML mark-up language) can include educational content such as text, graphics, and links, and multiple choice or open-ended questions. If a user does not answer a question correctly, hints may be provided until the correct answer is given. Although 'swirl' is a valuable package for learning important concepts in R and data science, users are limited in their ability to practice these concepts as 'swirl' lessons are static, so that a user repeating a lesson will see the same questions each time. This motivates an extension to 'swirl' that includes 'template-based problems' that would allow a user to practice on an endless supply of problems for a given topic.

We describe and implement a new package, 'swirl-tbp', that introduces 'template-based practice' problems to the 'swirl' framework. Specifically, 'swirl-tbp' extends 'swirl' by allowing instructors to include template-based problems in 'swirl' lessons. Template-based problems are problems that include numbers, variable names, or other features that are randomly generated at run-time. As a result, a user can be provided with an endless supply of practice problems that differ, e.g., with respect to the numbers used. This allows users to repeatedly practice problems in order to reinforce concepts and practice their problem-solving skills. We demonstrate the utility of 'swirl-tbp' by showing template-based problems for practicing basic R programming concepts such as vector creation and statistical concepts such as the calculation of probabilities involving normally distributed random variables.

*Keywords:* Education, R Programming, template-based problems, swirl

Presentation type: Oral Presentation

# Shiny Gadgets: Interactive tools for Programming and Data Analysis


## Garrett Grolemund

*RStudio*



**Abstract:** A Shiny Gadget is an interactive tool that enhances your R programming experience. You make Shiny Gadgets with the same package that you use to make Shiny Apps, but you use Gadgets in a very different way. Where Shiny Apps are designed to communicate results to an end user, Gadgets are designed to generate results for an R user. Each Shiny Gadget returns a value that you can immediately use in your code. You use Shiny Gadgets during the course of your analysis to quickly hone iterative tasks in an interactive fashion. For example, you might use a Shiny Gadget to preview the matches that are generated by a regular expression–as you write the expression. Or you might use a Shiny Gadget to identify high leverage points in your model–as you fit the model. Unlike Shiny Apps, Shiny Gadgets do not need to be deployed on a server. Shiny Gadgets are defined right inside of a regular R function. This is important, because it means that Gadgets can directly access the function's arguments, and the return value of the Gadget can be the return value for the function. Despite this difference, almost everything you know about Shiny Apps will transfer over to writing Shiny Gadgets.

Ready to see what Gadgets are all about? Attend this talk for some inspiring examples. The talk will also introduce the miniUI package, a collection of layout elements that are well-suited to Shiny Gadgets.



*Keywords:* Shiny, Interactive tools, Programming, Data Analysis, Workflow

Presentation type: Oral Presentation

# Network Diffusion of Innovations in R: Introducing netdiffuseR

**George Gerald Vega Yon, Stephanie Dyal, Timothy Hayes & Thomas Valente**

*University of Southern California*

**Abstract:** The Diffusion of Innovations theory, while one of the oldest social science theories, has embedded and flowed in its popularity over its 100 year or so history. In contrast to contagion models, diffusion of innovations can be more complex since adopting an innovation usually requires more than simple exposure to other users. At the same time, although computational tools for data collection, analysis, and network research have advanced considerably with little parallel develop of diffusion network models. To address this gap, we have created the netdiffuseR R package.

The netdiffuseR package implements both classical and novel diffusion of innovations models, visualization methods, and data-management tools for the statistical analysis of network diffusion data. The netdiffuseR package goes further by allowing researchers to analyze relatively large datasets in a fast and reliable way, extending current network analysis methods for studying diffusion, thus serving as a great complement to other popular network analysis tools such as igraph, statnet or RSiena. netdiffuseR can be used with new empirical data, with simulated data, or with existing empirical diffusion network datasets.

*Keywords:* diffusion of innovations, social network analysis, survival analysis, graph theory

Presentation type: Oral Presentation

# edeaR: extracting knowledge from process data

**Gert Janssenswillen, Marijke Swennen, Benoît Depaire, Mieke Jans & Koen Vanhoof**

*Hasselt University*

**Abstract:** During the last decades, the logging of events in a business context has increased massively. Information concerning activities within a broad range of business processes is recorded in so-called event logs. Connecting the domains of business process management and data mining, process mining aims at extracting process-related knowledge from these event logs, in order to gain competitive advantages. Over the last years, many tools for process mining analyses have been developed, having both commercial and academic origins. Nevertheless, most of them leave little room for extensions or interactive use. Moreover, they are not able to use existing data manipulation and visualization tools.

In order to meet these shortcomings, the R-package edeaR was developed to enable the creation and analysis of event logs in R. It provides functionality to read and write logs from .XES-files, the eXtensible Event Stream format, which is the generally-acknowledged format for the interchange of event log data. By using the extensive spectrum of data manipulation methods in R, edeaR provides a very convenient way to build .XES-files from raw data, which is a cumbersome task in most existing process mining tools. Furthermore, the package contains a wide set of functions to describe and select event data, thereby facilitating exploratory and descriptive analysis. Being able to handle event data in R both empowers process miners to exploit the vast area of data analysis methods in R, and invites R-users to contribute to this rapidly emerging and promising field of process mining.

*Keywords:* Process mining, process data, event data

Presentation type: Oral Presentation

# Interactive Naïve Bayes using Shiny: Text Retrieval, Classification, Quantification


## Giorgio Maria Di Nunzio

*Department of Information Engineering, University of Padua*



**Abstract:** Interactive Machine Learning (IML) is a relatively new area of ML where focused interactions between algorithms and humans allow for faster and more accurate model updates with respect to classical ML algorithms. By involving users directly in the process of optimizing the parameters of the ML model, it is possible to quickly improve the effectiveness of the model and also to understand why some values of the parameters of the model work better than others through low-cost trial and error and experimentation with inputs and outputs.

In this talk, we show three interactive applications developed with the Shiny package on the problems of text retrieval, text classification and text quantification. These applications implement a probabilistic model that use the Naïve Bayes (NB) assumption which has been widely recognised as a good trade-off between efficiency and efficacy, but it achieves satisfactory results only when optimized properly. All these three applications provide a two-dimensional representation of probabilities that has been inspired by the approach named Likelihood Spaces. This representation provides an adequate data visualization to understand how parameters and costs optimization affects the performance of the retrieval/classification/quantification application in a real machine learning setting on standard text collections.

We will show that this particular geometrical interpretation of the probabilistic model together with the interaction significantly improves not only the performance but also the understanding of the models and opens new perspectives for new research studies.



*Keywords:* Interactive Machine Learning, Naïve Bayes, Shiny, Information Retrieval, Text Classification, Text Quantification

Presentation type: Oral Presentation

# Fry: A Fast Interactive Biological Pathway Miner


## Goknur Giner & Gordon K. Smyth

*The Walter and Eliza Hall Institute of Medical Research*



**Abstract:** Gene set tests are often used in differential expression analyses to explore the behavior of a group of related genes. This is useful for identifying large-scale co-regulation of genes belonging to the same biological process or molecular pathway.

One of the most flexible and powerful gene set tests is the ROAST method in the limma R package. ROAST uses residual space rotation as a sort of continuous version of sample permutation. Like permutation tests, it protects against false positives caused by correlations between genes in the set. Unlike permutation tests, it can be used with complex experimental design and with small numbers of replicates. It is the only gene set test method that is able to analyse complex "gene expression signatures" that incorporate information about both up and down regulated genes simultaneously.

ROAST works well for individual expression signatures, but has limitations when applied to large collections of gene sets, such as the Broad Institute's Molecular Signature Database with over 8000 gene sets. In particular, the p-value resolution is limited by the number of rotations that are done for each set. This makes it impossible to obtain very small p-values and hence to distinguish the top ranking pathways from a large collection. As with permutation tests, the p-values for each set may vary from run to run.

This talk presents Fry, a very fast approximation to the complete ROAST method. Fry approximates the limiting p-value that would be obtained from performing a very large number of rotations with ROAST. Fry preserves most of the advantages of ROAST, but also provides high resolution exact p-values very quickly. In particular, it is able to distinguish the most significant sets in large collections and to yield statistically significant results after adjustment for multiple testing. This makes it an ideal tool for large-scale pathway analysis.

Another important consideration in gene set tests is the possible unbiased or incorrect estimation of P-values due to the correlation among genes in the same set or dependence structure between different sets.



*Keywords:* Pathway analysis, Singular value decomposition, Beta approximation

Presentation type: Oral Presentation

# Exploring the R / SQL boundary


## Gopi Kumar & Hang Zhang

*Microsoft Corporation*



**Abstract:** Databases have a long history of delivering highly scalable solutions for storing, manipulating, and analyzing data, transaction processing and data warehousing, while R is the most widely used language for data analytics and machine learning due to its rich ecosystem of machine learning algorithms and data manipulation capabilities. But, when using these tools together, how do you decide how much processing to do in SQL before switching to R? In this talk, we will explore setting the R / SQL boundary under three scenarios: RODBC connections, dplyr data extractions, and in-database R processing, and examine the consequences of each of these approaches with respect to data exploration, feature engineering, modeling and predictions. We identify common performance killers such as excessive data movements and serial processing, and illustrate the techniques, with examples from both an open source database (Postgres) and a commercial database (Microsoft SQL Server).



*Keywords:* data processing, big data, in-database analytics

Presentation type: Oral Presentation

# rbokeh: A Simple, Flexible, Declarative Framework for Interactive Graphics


## Paul Hafen Ryan

*Purdue University*



**Abstract:** The rbokeh package is an R interface to the Bokeh visualization library. The interface is designed to be simple but flexible, allowing the expressiveness required for rapid generation of ad hoc statistical plots, while providing simple yet useful interactive capabilities such as tooltips and zoom/pan with nearly no additional effort, and additional customized interactivity with a little more effort. In this talk I will introduce rbokeh and show examples of using it to create a wide variety of interactive displays, including Shiny applications and Trelliscope



*Keywords:* visualization, interactive, Bokeh

Presentation type: Oral Presentation

# United Nations World Population Projections with R


## Hana Ševčíková, Patrick Gerland & Adrian Raftery

*University of Washington, United Nations Population Division, University of Washington*



**Abstract:** Recently, the United Nations adopted a probabilistic approach to projecting fertility, mortality and population for all countries. In this approach, the total fertility and female and male life expectancy at birth are projected using Bayesian hierarchical models estimated via Markov Chain Monte Carlo. They are then combined yielding probabilistic projections for any population quantity of interest. The methodology is implemented in a suite of R packages which has been used by the UN to produce the most recent revision of the World Population Prospects. I will summarize the main ideas behind each of the packages, namely bayesTFR, bayesLife, bayesPop, bayesDem, and the shiny-based wppExplorer. I will also touch on our experience of the collaboration between academics and the UN.



*Keywords:* population projections, UN World Population Prospects, predictive distribution

Presentation type: Oral Presentation

# trackeR: Infrastructure for Running and Cycling Data from GPS-Enabled Tracking Devices in R


## Hannah Frick & Ioannis Kosmidis

*University College London*



**Abstract:** The use of GPS-enabled tracking devices and heart rate monitors is becoming increasingly common in sports and fitness activities. The trackeR package aims to fill the gap between the routine collection of data from such devices and their analyses in a modern statistical environment like R. The package provides methods to read tracking data and store them in session-based, unit-aware, and operation-aware objects of class trackeRdata. The package also implements core infrastructure for relevant summaries and visualisations, as well as support for handling units of measurement. There are also methods for relevant analytic tools such as time spent in zones, work capacity above critical power (known as $W'$), and distribution and concentration profiles. A case study illustrates how the latter can be used to summarise the information from training sessions and use it in more advanced statistical analyses.



*Keywords:* sports, tracking, running, cycling, work capacity, distribution profiles

Presentation type: Oral Presentation

# Efficient tabular data ingestion and manipulation with MonetDBLite


## Hannes Mühleisen

*Centrum Wiskunde & Informatica*



**Abstract:** We present “MonetDBLite", a new R package containing an embedded version of MonetDB. MonetDB is a free and open source relational database focused on analytical applications. MonetDBLite provides fast complex query answers and unprecedented speeds for data availability and data transfer to and from R.

MonetDBLite greatly simplifies database installation, setup and maintenance. It is installed like any R package, and the database fully runs inside the R process. This has the crucial advantage of data transfers between the database and R being very fast. Another advantage is MonetDBLite’s fast startup with existing data sets. MonetDBLite will store tables as files on disk, and can reload from these regardless of their size. This enables R scripts to very quickly start processing data instead of loading from, e.g., a CSV file every time. MonetDBLite leverages our previous work on mapping database operations into R (now achieved through dplyr in the MonetDB.R package) as well as previous work on ad-hoc user defined functions for MonetDB with R.

The talk will introduce the package, demonstrate its installation, and showcase a real-world statistical data analysis on the Home Mortgage Disclosure Act (HMDA) dataset. We show how MonetDBLite compares with its (partial) namesake SQLite and other relational databases. We will demonstrate that for statistical analysis workloads, MonetDBLite easily outperforms these previous systems, effectively allowing analysis of larger datasets on desktop hardware.

MonetDBLite has been submitted to CRAN and will hopefully be accepted by useR! 2016.



*Keywords:* relational databases, column store, dplyr, aggregation, join, embedded

Presentation type: Oral Presentation

# Predicting individual treatment effects


## Heidi Seibold, Achim Zeileis & Torsten Hothorn

*University of Zurich, University of Innsbruck, University of Zurich*



**Abstract:** Treatments for complicated diseases often help some patients but not all and predicting the treatment effect of new patients is important in order to make sure every patient gets the best possible treatment. We propose model-based random forests as a method to detect similarities between patients with respect to their treatment effect and on this basis compute personalized models for new patients to obtain their individual treatment effect. The whole procedure focuses on a base model which usually contains the treatment indicator as a single covariate and takes the survival time or a health or treatment success measurement as primary outcome. This base model is used to grow the model-based trees within the forest as well as to compute the personalized models, where the similarity measurements enter as weights.

We show how personalized models can be set up using the cforest() and predict.cforest() functions from the "partykit" package in combination with regression models such as glm() ("stats") or survreg() ("survival"). We apply the methods to patients suffering from Amyotrophic Lateral Sclerosis (ALS). The data are publicly available from https://nctu.partners.org/ProACT and data preprocessing can be done with the R package "TH.data". The treatment of interest is the drug Riluzole which is the only approved drug against ALS but merely shows minor benefit for patients. The personalized models suggest that some patients benefit more from the drug than others.



*Keywords:* Personalised medicine, Treatment effect, Model-based recursive partitioning, Random Forest, partykit

Presentation type: Oral Presentation

# Approximate inference in R: a case study with GLMMs and glmmsr


## Helen Elizabeth Ogden

*University of Warwick*



**Abstract:** The use of realistic statistical models for complex data is often hindered by the high cost of conducting inference about the model parameters. Because of this, it is sometimes necessary to use approximate inference methods, even though the impact of these approximations on the fitted model might not be well understood. I will discuss some practical examples of this, demonstrating how to fit various Generalized Linear Mixed Models with the R package glmmsr, using a variety of approximation methods, with a focus on what difference the choice of approximation makes to the resulting inference. I will talk about some more general issues along the way, such as how we might detect situations in which a given approximation might give unreliable inference, and the extent to which the choice of approximation method can and should be automated. I will finish by briefly reviewing some ideas about how best to share and discuss challenging models and datasets which could motivate the development of new approximation methods.



*Keywords:* Latent variable model, Intractable likelihood, Numerical approximation, Pairwise comparison model, Data sharing

Presentation type: Oral Presentation

# Resource-Aware Scheduling Strategies for Parallel Machine Learning R Programs though RAMBO

**Helena Kotthaus, Jakob Richter, Andreas Lang, Michel Lang & Peter Marwedel**

*Department of Computer Science 12, TU Dortmund University, Dortmund, Germany*

**Abstract:** We present resource-aware scheduling strategies for parallel R programs leading to efficient utilization of parallel computer architectures by estimating resource demands. We concentrate on applications that consist of independent tasks. The R programming language is increasingly used to process large data sets in parallel, which requires a high amount of resources. One important application is parameter tuning of machine learning algorithms where evaluations need to be executed in parallel to reduce runtime. Here, resource demands of tasks heavily vary depending on the algorithm configuration. Running such an application in a naive parallel way leads to inefficient resource utilization and thus to long runtimes. Therefore, the R package "parallel" offers a scheduling strategy, called "load balancing". It dynamically allocates tasks to worker processes. This option is recommended when tasks have widely different computation times or if computer architectures are heterogeneous. We analyzed memory and CPU utilization of parallel applications with our TraceR profiling tool and found that the load balancing mechanism is not sufficient for parallel tasks with high variance in resource demands. A scheduling strategy needs to know resource demands of a task before execution to efficiently map applications to available resources. Therefore, we build a regression model to estimate resource demands based on previous evaluated tasks. Resource estimates like runtime are then used to guide our scheduling strategies. Those strategies are integrated in our RAMBO (Resource-Aware Model-Based Optimization) Framework. Compared to standard mechanisms of the parallel package our approach yields improved resource utilization.

*Keywords:* Resource-Aware Scheduling, Parallelization, Distributed Computing, Performance Analysis, Machine Learning, Hyperparameter Tuning, Model-Based Optimization

Presentation type: Oral Presentation

# A Future for R


## Henrik Bengtsson

*University of California, San Francisco*



**Abstract:** A future is an abstraction for a value that may be available at some point in the future and which state is either unresolved or resolved. When a future is resolved the value is readily available. How and when futures are resolved is given by their evaluation strategies, e.g. synchronously in the current R session or asynchronously on a compute cluster or in background processes. Multiple asynchronous futures can be created without blocking the main process providing a simple yet powerful construct for parallel processing. It is only when the value of an unresolved future is needed it blocks.

We present the future package which defines a unified API for using futures in R, either via explicit constructs $f <- future(expr)$ and $v <- value(f)$ or via implicit assignments (promises) $v \% <- \% expr$. From these it is straightforward to construct classical *apply() mechanism. The package implements synchronous eager and lazy futures as well as multiprocess (single-machine multicore and multisession) and cluster (multi-machine) futures. Additional future types can be implemented by extending the future package, e.g. BatchJobs and BiocParallel futures.

We show that, because of the unified API and because global variables are automatically identified and exported, an R script that runs sequentially on the local machine can with a single change of settings run in, for instance, a distributed fashion on a remote cluster with values still being collected on the local machine.

The future package is cross-platform and available on CRAN with source code on GitHub (https://github.com/HenrikBengtsson/future).



*Keywords:* asynchronous, parallel, distributed, lazy evaluation, future, promise, compute cluster, multicore, multisession, remote access, global variables, R package, cross platform

Presentation type: Oral Presentation

# Using Shiny modules to build more-complex and more-manageable apps


## Ian John Lyttle

*Schneider Electric*



**Abstract:** The release of Shiny 0.13 includes support for modules, allowing you to build Shiny apps more quickly and more reliably. Furthermore, using Shiny modules makes it easier for you to build more-complex apps, because the interior complexity of each module is hidden from the level of the app. This allows you, as a developer, to focus on the complexity of the app at the system-level, rather than at the module-level.

For example, there are open-source shiny modules that: read a time-indexed csv file then parse it into a dataframe, visualize a time-indexed dataframe using dygraphs, and write a dataframe to a csv file to be downloaded. Modules are simply collections of functions that can be organized into, and called from, packages.

The primary focus of this presentation will be on how modules from the “shinypod" package can be assembled to into “simple" shiny apps. As well, there will be demonstrations of more-complex apps built using modules. In this case, Shiny apps are built as interfaces to web-services, allowing you to evaluate the usefulness of suites of web-services without having to be immediately concerned with the API clients.

Time permitting, there could be some discussion of how Shiny modules are put together.



*Keywords:* shiny, modules, interactive

Presentation type: Oral Presentation

# Distributed Computing using parallel, Distributed R, and SparkR

**Edward Ma, Indrajit Roy & Michael Lawrence**

*Hewlett Packard Labs*, *Hewlett Packard Labs*, *Genentech*

**Abstract:** Data volume is ever increasing, while single node performance is stagnate. To scale, analysts need to distribute computations. R has built-in support for parallel computing, and third-party contributions, such as Distributed R and SparkR, enable distributed analysis. However, analyzing large data in R remains a challenge, because interfaces to distributed computing environments, like Spark, are low-level and non-idiomatic. The user is effectively coding for the underlying system, instead of writing natural and familiar R code that produces the same result across computing environments.

This talk focuses on how to scale R-based analyses across multiple cores and to leverage distributed machine learning frameworks through the ddR (Distributed Data structures in R) package, a convenient, familiar, and idiomatic abstraction that helps to ensure portability and reproducibility of analyses. The ddR package defines a framework for implementing interfaces to distributed environments behind the canonical base R API. We will discuss key programming concepts and demonstrate writing simple machine learning applications. Participants will learn about creating parallel applications from scratch as well as invoking existing parallel implementations of popular algorithms, like random forest and kmeans clustering.

*Keywords:* Distributed computing, big data, machine learning, Spark

Presentation type: Oral Presentation

# brglm: Reduced-bias inference in generalized linear models


## Ioannis Kosmidis

*Department of Statistical Science, University College London*



**Abstract:** This presentation focuses on the brglm R package, which provides methods for reduced-bias inference in univariate generalised linear models and multinomial regression models with either ordinal or nominal responses (Kosmidis, 2014, JRSSB and Kosmidis and Firth, 2011, Biometrika, respectively).

The core fitting method is based on the iterative correction of the bias of the maximum likelihood estimator, and results in the solution of appropriate bias-reducing adjusted score equations. For multinomial logistic regression, we present alternative algorithms that can scale up well with the number of multinomial responses and illustrate the finiteness and shrinkage properties that make bias reduction attractive for such models. For families with dispersion parameters (e.g. gamma regression), brglm uses automatic differentiation to compute the reduced-bias estimator of arbitrary invertible transformations of the dispersion parameter (e.g. user-supplied). We also present the implementation of appropriate methods for inference when bias-reduced estimation is being used.



*Keywords:* adjusted score functions, finiteness, iterative bias correction, shrinkage, penalized likelihood

Presentation type: Oral Presentation

# Notebooks with R Markdown


## J.J. Allaire

*RStudio*



**Abstract:** Notebook interfaces for data analysis have compelling advantages including the close association of code and output and the ability to intersperse narrative with computation. Notebooks are also an excellent tool for teaching and a convenient way to share analyses.

As an authoring format, R Markdown bears many similarities to traditional notebooks like Jupyter and Beaker, but it has some important differences. R Markdown documents use a plain-text representation (markdown with embedded R code chunks) which creates a clean separation between source code and output, is editable with the same tools as for R scripts (.Rmd modes are available for Emacs, Vim, Sublime, Eclipse, and RStudio), and works well with version control. R Markdown also features a system of extensible output formats that enable reproducible creation of production-quality output in many formats including HTML, PDF, Word, ODT, HTML5 slides, Beamer, LaTeX-based journal articles, websites, dashboards, and even full length books.

In this talk we'll describe a new notebook interface for R that works seamlessly with existing R Markdown documents and displays output inline within the standard RStudio .Rmd editing mode. Notebooks can be published using the traditional Knit to HTML or PDF workflow, and can also be shared with a compound file that includes both code and output, enabling readers to easily modify and re-execute the code.

Building a notebook system on top of R Markdown carries forward its benefits (plain text, reproducible workflow, and production quality output) while enabling a richer, more literate workflow for data analysis.



*Keywords:* Notebooks, R Markdown, Reproducible Research

Presentation type: Oral Presentation

# The simulator: An Engine for Streamlining Simulations


## Jacob Bien

*Cornell University*



**Abstract:** Methodological statisticians spend an appreciable amount of their time writing code for simulation studies. Every paper introducing a new method has a simulation section in which the new method is compared across several metrics to preexisting methods under various scenarios. Given the formulaic nature of the simulation studies in most statistics papers, there is a lot of code that can be reused. We have developed an R package, called the “simulator", that streamlines the process of performing simulations by creating a common infrastructure that can be easily used and reused across projects. The simulator allows the statistician to focus exclusively on those aspects of the simulation that are specific to the particular paper being written. Code for simulations written with the simulator is succinct, highly readable, and easily shared with others. The modular nature of simulations written with the simulator promotes code reusability, which saves time and facilitates reproducibility. Other benefits of using the simulator include the ability to “step in" to a simulation and change one aspect without having to rerun the entire simulation from scratch, the straightforward integration of parallel computing into simulations, and the ability to rapidly generate plots and tables with minimal effort.



*Keywords:* simulation studies, reproducible, methodological statistics

Presentation type: Oral Presentation

# mlrMBO: A Toolbox for Model-Based Optimization of Expensive Black-Box Functions


## Jakob Richter

*TU Dortmund University*



**Abstract:** Many practical optimization tasks, such as finding best parameters for simulators in engineering or hyperparameter optimization in machine learning, are of a black-box nature, i.e., neither formulas of the objective nor derivative information is available. Instead, we can only query the box for its objective value at a given point. If such a query is very time-consuming, the optimization task becomes extremely challenging, as we have to operate under a severely constrained budget of function evaluations. A modern approach is sequential model based-optimization, aka Bayesian optimization. Here, a surrogate regression model learns the relationship between decision variables and objective outcome. Sequential point evaluations are planned to simultaneously exploit the so far learnt functional landscape and to ensure exploration of the search space. A popular instance of this general principle is the EGO algorithm, which uses Gaussian processes coupled with the expected improvement criterion for point proposal. The mlrMBO package offers a rich interface to many variants of model-based optimization. As it builds upon the mlr package for machine learning in R, arbitrary surrogate regression models can be applied. It offers a wide variety of options to tackle different black-box scenarios:

- Optimization of pure continuous as well as mixed continuous-categorical search spaces.
- Single criteria optimization or approximated Pareto fronts for multi-criteria problems.
- Single point proposal or parallel batch point planning during optimization.

The package is designed as a convenient, easy-to-use toolbox of popular state-of-the-art algorithms, but can also be used as as a research framework for algorithm designers.



*Keywords:* machine learning, hyperparameter optimization, tuning, black box optimization, bayesian optimization

Presentation type: Oral Presentation

# Covr: Bringing Code Coverage to R


## James F Hester

*RStudio*



**Abstract:** Code coverage records whether or not each line of code in a package is executed by the package's tests. While it does not check whether a given program or test executes properly it does reveal areas of the code which are untested. Coverage has a long history in the computer science community (Miller and Maloney in Communications of the ACM, 1963), unfortunately the R language has lacked a comprehensive and easy to use code coverage tool.

The covr package was written to make it simple to measure and report test coverage for R, C, C++ and Fortran code in R packages. It has measurably improved testing for numerous packages and also serves as an informative indicator of package reliability. Covr is now used routinely by over 1000 packages on CRAN, Bioconductor and GitHub.

I will discuss how covr works, how it is best used and how it has demonstrably improved test coverage in R packages since its release.



*Keywords:* R packages, computer science, development, testing

Presentation type: Oral Presentation

# What can R learn from Julia


## Jan Vitek

*Northeastern University*



**Abstract:** Julia, like R, is a dynamic language for scientific computing but, unlike R, it was explicitly designed to deliver performance competitive to traditional batch-compiled languages. To achieve this Julia's designers made a number of unusual choices, including the presence of a set of type annotations that are used for dispatching methods and speed up code, but not for type-checking. The result is that many Julia programs are competitive with equivalent programs written in C. This talk gives a brief overview of the key points of Julia's design and considers whether similar ideas could be adopted in R.



*Keywords:* Dynamic languages, compilers, Julia

Presentation type: Oral Presentation

# A Case Study in Reproducible Model Building: Simulating Groundwater Flow in the Wood River Valley Aquifer System, Idaho

**Jason C. Fisher**

*U.S. Geological Survey*

**Abstract:** The goal of reproducible model building is to tie processing instructions to data analysis so that the model can be recreated, better understood, and easily modified to incorporate new field measurements and (or) explore alternative system and boundary conceptualizations. Reproducibility requires archiving and documenting all raw data and source code used to pre- and post-process the model; an undertaking made easier by the advances in open source software, open file formats, and cloud computing. Using a software development methodology, a highly reproducible model of groundwater flow in the Wood River Valley (WRV) aquifer system was built. The collection of raw data, source code, and processing instructions used to build and analyze the model was placed in an R package. An R package allows for easy, transparent, and cross-platform distribution of its content by enforcing a set of formal format standards. R largely facilitates reproducible research with the package vignette, a document that combines content and data analysis source code. The code is run when the vignette is built, and all data analysis output (such as figures and tables) is created extemporaneously and inserted into the final document. The R package created for the WRV groundwater-flow model includes multiple vignettes that explain and run all processing steps; the exception to this being the parameter estimation process, which was not made programmatically reproducible. MODFLOW-USG, the numerical groundwater model used in this case study, is executed from a vignette, and model output is returned for exploratory analyses.

*Keywords:* groundwater, model, reproducible, package, vignette

Presentation type: Oral Presentation

# Using Shiny for Formative Assessments


## Jason M Bryer

*Excelsior College*



**Abstract:** Shiny has become a popular approach for R developers to create interactive dashboards. Given the rich set of features available in Shiny, it has the capability for data entry and collection. This talk introduces a framework for using Shiny to conduct formative assessments whereby students both complete an assessment online as well as receive immediate feedback and scores on their performance. Examples using multiple choice assessments and Likert type self-report assessments will be provided along with feedback templates using R markdown for rapid development. The implications of this approach in course development will also be discussed.



*Keywords:* shiny, assessment, teaching

Presentation type: Oral Presentation

# Importing modern data into R


## Javier Luraschi

*RStudio*



**Abstract:** This talk explores modern trends in data storage formats and the tools, packages and best practices to import this data into R.

We will start with a quick recap of the existing tools and packages for importing data into R: readr, readxl, haven, jsonlite, xml2, odbc and jdbc. Afterwards, we will discuss modern data formats and the emerging tools we can use today. We will explore sparkr, mongolite and the role of specialized packages like fitbitScraper and getSymbols. This talk will wrap up by assessing gaps and exploring future trends in this space.



*Keywords:* data,json,xml,sparkr,parquet,mongolite,oauth,readr,readxl,haven,

Presentation type: Oral Presentation

# Using Spark with Shiny and R Markdown


## Jeff David Allen

*RStudio*



**Abstract:** R is well-suited to handle data that can fit in memory but additional tools are needed when the amount of data you want to analyze in R grows beyond the limits of your machine's RAM. There have been a variety of solutions to this problem over the years that aim to solve this problem in R; one of the latest options is Apache Spark™.

Spark is a cluster computing tool that enables analysis of massive, distributed data across dozens or hundreds of servers. Spark now includes an integration with R via the SparkR package. Due to Spark's ability to interact with distributed data little latency, it is becoming an attractive tool for interfacing with large datasets in an interactive environment.

In addition to handling the storage of data, Spark also incorporates a variety of other tools including stream processing, computing on graphs, and a distributed machine learning framework. Some of these tools are available to R programmers via the SparkR package.

In this talk, we'll discuss how to leverage Spark's capabilities in a modern R environment. In particular, we'll discuss how to use Spark within an R Markdown document or even in an interactive Shiny application. We'll also briefly discuss alternative approaches to working with large data in R and the pros and cons of using Spark



*Keywords:* spark, cluster, distributed, shiny, rmarkdown, big data

Presentation type: Oral Presentation

# jailbreakr: Get out of Excel, free


## Jenny Bryan & Rich FitzJohn

*University of British Columbia; rOpenSci, Imperial College London; rOpenSci*



**Abstract:** One out of every ten people on the planet uses a spreadsheet and about half of those use formulas: “Let’s not kid ourselves: the most widely used piece of software for statistics is Excel." (Ripley, 2002) Those of us who script analyses are in the distinct minority!

There are several effective packages for importing spreadsheet data into R. But, broadly speaking, they prioritize access to [a] data and [b] data that lives in a neat rectangle. In our collaborative analytical work, we battle spreadsheets created by people who did not get this memo. We see messy sheets, with multiple data regions sprinkled around, mixed with computed results and figures. Data regions can be a blend of actual data and, e.g., derived columns that are computed from other columns.

We will present our work on extracting tricky data and formula logic out of spreadsheets. To what extent can data tables be automatically identified and extracted? Can we identify columns that are derived from others in a wholesale fashion and translate that into something useful on the R side? The goal is to create a more porous border between R and spreadsheets. Target audiences include novices transitioning from spreadsheets to R and experienced useRs who are dealing with challenging sheets.



*Keywords:* spreadsheets, data import, reproducible research

Presentation type: Oral Presentation

# Linking htmlwidgets with crosstalk and mobservable


## Joe Cheng

*RStudio, Inc.*



**Abstract:** The htmlwidgets package makes it easy to create interactive JavaScript widgets from R, and display them from the R console or insert them into R Markdown documents and Shiny apps. These widgets exhibit interactivity "in the small": they can interact with mouse clicks and other user gestures within their widget boundaries.

This talk will focus on interactivity "in the large", where interacting with one widget results in coordinated changes in other widgets (for example, select some points in one widget and the corresponding observations are instantly highlighted across the other widgets). This kind of inter-widget interactivity can be achieved by writing a Shiny app to coordinate multiple widgets (and indeed, this is a common way to use htmlwidgets). But some situations call for a more lightweight solution.

crosstalk and robservable are two distinct but complementary approaches to the problem of widget coordination, authored by myself and Ramnath Vaidyanathan, respectively. Each augments htmlwidgets with pure-JavaScript coordination logic; neither requires Shiny (or indeed any runtime server support at all). The resulting documents can be hosted on GitHub, RPubs, Amazon S3, or any static web host.

In this talk, I'll demonstrate these new tools, and discuss their advantages and limitations compared to existing approaches.



*Keywords:* interactive graphics, javascript, htmlwidgets, linked brushing

Presentation type: Oral Presentation

# Dynamic Data in the Statistics Classroom


## Johanna Hardin

*Pomona College*



**Abstract:** The call for using *real* data in the classroom has long meant using datasets which are culled, cleaned, and wrangled prior to any student working with the observations. However, an important part of teaching statistics should include actually retrieving data. Nowadays, there are many different sources of data that are continually updated by the organization hosting the data website. The R tools to download such dynamic data have improved in such a way to make accessing the data possible even in an introductory statistics class. We provide four full analyses on dynamic data as well as an additional six sources of dynamic data that can be brought into the classroom.



*Keywords:* data scraping, data science, data analysis pipeline, authentic data

Presentation type: Oral Presentation

# Big Data Algorithms for Rank-based Estimation


## John Kapenga, John Kloke & Joesph W. McKean

*Western Michigan University, University of Wisconsin–Madison, Western Michigan University*



**Abstract:** Rank-based (R) estimation for statistical models is a robust nonparametric alternative to classical estimation procedures such as least squares. R methods have been developed for models ranging from linear models, to linear mixed models, to time series, to nonlinear models. Advantages of these R methods over traditional methods such as maximum-likelihood or least squares are that they require fewer assumptions, are robust to gross outliers, and are highly efficient at a wide range of distributions. The R package, Rfit, was developed to widely disseminate these methods as the software uses standard linear model syntax and includes commonly used functions for inference and diagnostic procedures.

Large datasets are becoming common in practice, and the ability to obtain results in real time is desirable. We have developed algorithms for R estimation which improve the speed at the expense of a slight decrease in accuracy in big data settings. In this talk we describe the traditional as well as the big data algorithms for R estimation. We present examples and results from simulation studies which illustrate the algorithms.



*Keywords:* nonparametric; linear models; robust

Presentation type: Oral Presentation

# A Lap Around R Tools for Visual Studio


## John Lam

*Microsoft*



**Abstract:** R Tools for Visual Studio is a new, Open Source and free tool for R Users built on top of the powerful Visual Studio IDE. In this talk, we will take you on a tour of its features and show you how they can help you be a more productive R user. We will look at:

- Integrated debugging support
- Variable/data frame visualization
- Plotting and help integration
- Using the Editor and REPL in concert with each other
- RMarkdown and Shiny integration
- Using Excel and SQL Server
- Extensions and source control



*Keywords:* tools, database, SQL, Excel

Presentation type: Oral Presentation

# Visualizing Simultaneous Linear Equations, Geometric Vectors, and Least-Squares Regression with the matlib Package for R

**John David Fox**

*McMaster University*

**Abstract:** The aim of the **matlib** package is pedagogical — to help teach concepts in linear algebra, matrix algebra, and vector geometry that are useful in statistics. To this end, the package includes various functions for numerical linear algebra, most of which duplicate capabilities available elsewhere in R, but which are programmed transparently and purely in R code, including functions for solving possibly over- or under-determined linear simultaneous equations, for computing ordinary and generalized matrix inverses, and for producing various matrix decompositions. Many of these methods are implemented via Gaussian elimination.

This paper focuses on the visualization facilities in the **matlib** package, including for graphing the solution of linear simultaneous equations in 2 and 3 dimensions; for demonstrating vector geometry in 2 and 3 dimensions; and for displaying the vector geometry of least-squares regression. We illustrate how these visualizations help to communicate fundamental ideas in linear algebra, vector geometry, and statistics. The 3D visualizations are implemented using the rgl package.

*Keywords:* linear algebra, matrix algebra, vector geometry, least squares, dynamic 3D graphics, visualization

Presentation type: Oral Presentation

# Grid Computing in R with Easy Scalability


## Jonathan Adams & David Bronke

*ARMtech Insurance Services*



**Abstract:** Parallel computing is useful for speeding up computing tasks and many R packages exist to aid in using parallel computing. Unfortunately it is not always trivial to parallelize jobs and can take a significant amount of time to accomplish, time that may be unavailable. My presentation will demonstrate an alternative method that allows for processing of multiple jobs simultaneously across any number of servers using Redis message queues. This method has proven very useful since I began implementing it at my company over two years ago. In this method, a main Redis server handles communication with any number of R processes on any number of servers. These processes, known as workers, inform the server that they are available for processing and then wait indefinitely until the server passes them a task.

In this presentation, it will be demonstrated how trivial it is to scale up or down by adding or removing workers. This will be demonstrated with sample jobs run on workers in the Amazon cloud. Additionally, this presentation will show you how to implement such a system yourself with the rminions package I have been developing. This package is based on what I have learned over the past couple of years and contains functionality to easily start workers, queue jobs, and even perform R-level maintenance (such as installing packages) on all connected servers simultaneously!



*Keywords:* grid computing, scalability, redis, message queues

Presentation type: Oral Presentation

# flexdashboard: Easy interactive dashboards for R


## Jonathan McPherson

*RStudio, Inc.*



**Abstract:** Recently, dashboards have become a common means of communicating the results of data analysis, especially of real-time data, and with good reason: dashboards present information attractively, use space efficently, and offer eye-catching visualizations that make it easy to consume information at a glance.

Traditionally, however, dashboards have been difficult to construct using tools readily available to R users, and so are built by a separate engineering team if they're built at all.

In this talk, we present a new package, `flexdashboard`, which empowers R users to build fully-functioning dashboards. To make this possible, `flexdashboard` leverages two existing packages: `R Markdown` and `Shiny`. `R Markdown` provides a means to describe the dashboard's content and layout using simple text constructs; and, optionally, `Shiny` enables interactivity among components and allows the full analytic power of R to be used at runtime.

The talk will focus on the practical steps involved in setting up a dashboard using `flexdashboard`, including:

- Building space-filling layouts using declarative R Markdown directives;
- Using dashboard components, such as tables, value boxes, charts, and more;
- Constructing multi-page dashboards for the presentation of larger or more detailed results; and
- Adding interactivity using Shiny.



*Keywords:* interactivity, dashboards, visualization, shiny, rmarkdown

Presentation type: Oral Presentation

# Classifying Murderers in Imbalanced Data Using randomForest


## Jorge Alberto Miranda

*County of Los Angeles*



**Abstract:** In order to allocate resources more effectively with the goal of providing safer communities, R's randomForest algorithm was used to identify candidates who may commit or attempt murder. And while crime data within the general population may be highly imbalanced, one may expect the rate of murderers within a high-risk probationer population to be much less imbalanced. However, the County of Los Angeles had nearly 130 probationers commit or attempt murder out of nearly 17,000, a ratio close to 1:130). Classic methods were used to overcome class imbalance, including under/over stratified sampling and variable sampling per tree. The results were encouraging. Model validation tests demonstrate an 87% overall accuracy rate at relatively low costs. The agency currently uses a risk assessment tool that was outperformed by randomForest up to 52% (both in overall accuracy and a reduction in false positives). This work is based on research conducted by Berk, R. et al. (2009) originally published by Journal of the Royal Statistical Society.



*Keywords:* randomForest, data imbalance, public safety, crime prevention

Presentation type: Oral Presentation

# Experiences on the Use of R in the Water Sector


## David Ibarra & Josep Arnal

*Universidad de Alicante*



**Abstract:** In this study we present some real cases where R has been a key element on building decision support systems related to the water industry. We have used R in the context of automatic water demand forecast, its application to optimal pumping scheduling and building a framework to offer these algorithms as a service (using RInside, Rcpp, MPI, RProtobuf among others) to easily integrate our work on heterogeneous environments. We have used an HPC cluster with R to solve big problems faster. About water demand forecast we used several tools like lineal models, neural networks or tree based method. On short term we included also weather forecast variables. The selection of the method is carried out dynamically (or online) using out-of-sample recent data. The optimal pumping schedule model is loaded with LPSolveAPI package and solved with CBC. We produce nice HTML5 reports of the solutions using googleVis package.



*Keywords:* Parallel processing, Optimization, Stochastic models, Scheduling, Water supply systems

Presentation type: Oral Presentation

# What's up with the R Consortium?


## Joseph B Rickert

*Microsoft*



**Abstract:** The R Consortium is a business association organized under the Linux Foundation with a mission to support the R Community. Founded just before useR! 2015, it has already become a focus for R Community activities. During its first year, the R Consortium has begun to evaluate and fund projects while dealing with all of the internal start-up issues of developing internal structures, policies and operating procedures. In this talk, I will attempt to provide some insight into the workings of the R Consortium, describe the process behind the recent call for proposals, discuss the projects selected for funding so far, and provide some guidance on writing a proposal for the next round of funding which will close on July 10th.



*Keywords:* R Consortium

Presentation type: Oral Presentation

# Phylogenetically informed analysis of microbiome data using adaptive gPCA in R

**Julia Anne Fukuyama**

*Stanford*

**Abstract:** When analyzing microbiome data, biologists often use exploratory methods that take into account the relatedness of the bacterial species present in the data. This helps in the interpretability and stability of the analysis because phylogenetically related bacteria often have similar functions. However, we believe (and will demonstrate), that the methods currently in use put too much emphasis on the phylogeny when making the ordinations. To address this, we have developed a framework we call adaptive gPCA, which allows the user to specify the amount of weight given to the tree and which will automatically select an amount of weight to give to the tree.

We have implemented this method in R and have made it easy to use with phyloseq, a popular R package for microbiome data storage and manipulation. Additionally, we have developed a shiny app that allows for interactive data visualization and comparison of the ordinations resulting from different weightings of the tree.

*Keywords:* pca, microbiome, ecology, phylogenetic tree

Presentation type: Oral Presentation

# Rho: High Performance R


## Karl Millar

*Google*



**Abstract:** The Rho project (formerly known as CXXR) is working on transforming the current R interpreter into a high performance virtual machine for R. Using modern software engineering techniques and the research done on VMs for dynamic and array languages over the last twenty years, we are targeting a factor of ten speed improvement or better for most types of R code, while retaining full compatibility.

This talk will discuss the current compatibility and performance of the VM, the types of tasks it currently does well and outline the project's roadmap for the next year.



*Keywords:*

Presentation type: Oral Presentation

# R/qtl: Just Barely Sustainable


## Karl W Broman

*University of Wisconsin-Madison*



**Abstract:** R/qtl is an R package for mapping quantitative trait loci (genetic loci that contribute to variation in quantitative traits, such as blood pressure) in experimental crosses (such as in mice). I began its development in 2000; there have been 46 software releases since 2001. The latest version contains 39k lines of R code, 24k lines of C code, and 16k lines of code for the documentation. The continued development and maintenance of the software has been challenging. I'll describe my experiences in developing and maintaining the package and in providing support to users. I'm currently working on a re-implementation of the package to better handle high-dimensional data and more complex experimental crosses. I'll describe my efforts to avoid repeating the mistakes I made the first time around.



*Keywords:* package development, genetics, quantitative trait loci

Presentation type: Oral Presentation

# How can I get everyone else in my organisation to love R as much as I do?


## Kate Ross-Smith

*Mango Solutions*



**Abstract:** Learning R is dangerous. It entices us in by presenting an incredibly powerful tool to solve our particular problem; for free! And as we learn how to do that, we uncover more things that make our solution even better. But then we start to look around our organisation or institution and see how it could make everyone's lives better too. And that's the dangerous part; R's got us hooked and we can't give up the belief that everyone else should be using this, right now. Even though R is free, open source software, there are often barriers to introducing it organisation-wide. This could be because of such things as IT or quality policies, the need for management buy-in or because of perceptions in learning the language. This presentation will first discuss the aspects required to understand these barriers to entry, and the different types of resolution for these. It will then use three projects to show how, by understanding the requirements of the organisation, and developing situation-specific roll-out strategies, these barriers to entry can be overcome. The first example is a large organisation who wanted to quickly (within 6 weeks) show management how Shiny could improve information dissemination. As server policies made a proof of concept difficult to run internally, this project used a cloud hosted environment for R, Shiny and a source database. The second example is around two SME's who required access to a validated version of R, which was provided via the Amazon and Azure marketplaces. The key aspect of these projects is the value to IT departments of being able to distribute a pre-configured machine around the organisation.



*Keywords:* Deployment, roll-out, architecture, IT, management

Presentation type: Oral Presentation

# permuter: An R Package for Randomization Inference

**Kellie Nicole Ottoboni, Jarrod Millman & Philip B. Stark**

*UC Berkeley*

**Abstract:** Software packages for randomization inference are few and far between. This forces researchers either to rely on specialized stand-alone programs or to use classical statistical tests that may require implausible assumptions about their data-generating process. The absence of a flexible and comprehensive package for randomization inference is an obstacle for researchers from a wide range of disciplines who turn to R as a language for carrying out their data analysis. We present permuter, a package for randomization inference. We illustrate the program's capabilities with several examples:

- a randomized experiment comparing the student evaluations of teaching for male and female instructors (MacNell et. al, 2014)
- a study of the association between salt consumption and mortality at the level of nations
- an assessment of inter-rater reliability for a series of labels assigned by multiple raters to video footage of children on the autism spectrum

We discuss future plans for permuter and the role of software development in statistics.

*Keywords:* Permutation tests; Randomization; Inference; Nonparametrics

Presentation type: Oral Presentation

# RcppParallel: A Toolkit for Portable, High-Performance Algorithms

**J.J. Allaire, Kevin Ushey, Kevin Ushey, Dirk Eddelbuettel, Romain Francois & Gregory Vandenbrouck**

*RStudio*, *RStudio*, *Debian*, *Freelance*, *Microsoft*

**Abstract:** Modern computers and processors provide many advanced facilities for the concurrent, or parallel, execution of code. While R is a fundamentally a single-threaded program, it can call into multi-threading code, provided that such code interacts with R in a thread-safe manner. However, writing concurrent programs that run both safely and correctly is a very difficult task, and requires substantial expertise when working with the primitives provided by most programming languages or libraries.

RcppParallel provides a complete toolkit for creating safe, portable, high-performance parallel algorithms, built on top of the Intel “Threading Building Blocks" (TBB) and “TinyThread" libraries. In particular, RcppParallel provides two high-level operations – ‘parallelFor’, and ‘parallelReduce’, which provide a framework for the safe, performant implementation of many kinds of parallel algorithms.

We’ll showcase how RcppParallel might be used to implement a parallel algorithm, and how the generated routine could be used in an R package.

*Keywords:* rcpp, parallel

Presentation type: Oral Presentation

# The phangorn package: estimating and comparing phylogenetic trees


## Klaus Peter Schliep & Liam Revell

*Department of Biology, University of Massachusetts Boston*



**Abstract:** Methods of phylogenetic reconstruction are nowadays frequently used outside computational biology like in linguistics and in form of hierarchical clustering in many other disciplines. The R package phangorn allows to reconstruct phylogenies using Maximum Likelihood, Maximum Parsimony or distanced based methods. The package offers many functions to compare trees through visualization (splits networks, lento plot, densiTree) and to choose and compare statistical models (e.g. modelTest, SH-test, parametric bootstrap). phangorn is closely connected with other phylogenetic R packages ape or phytools in the field of phylogenetic (comparative) methods.



*Keywords:* phylogenetics, trees, computational biology

Presentation type: Oral Presentation

# RosettaHUB-Sheets, a programmable, collaborative web-based spreadsheet for R, Python and Spark


## Latifa Bouabdillah

*ROSETTAHUB LTD*



**Abstract:** RosettaHUB-Sheets combine the flexibility of the bi-dimensional data representation model of classic spreadsheets with the power of R, Python, Spark and SQL. RosettaHUB-Sheets are web based, they can be created programmatically on any cloud. They enable Google-docs like real-time collaboration while preserving the user's data privacy. They have no limitation of size and can leverage the cloud for performance and scalability.

RosettaHUB-Sheets act as highly flexible bi-dimensional notebook as they make it possible to create powerful mash-ups of multi-language scripts and results. RosettaHUB-Sheets are combined with an interactive widgets framework with the ability to overlay advanced interactive widgets and visualizations including 3D Paraview ones. RosettaHUB-Sheets are fully programmable in R, Python and JavaScript, macros similar to Excel VBA's can be triggered by various cells and variables states changes events. RosettaHUB-sheets can be shared to allow real-time collaboration, interactive teaching, etc. RosettaHUB-Sheets' are represented by an SQL database and can be queried and updated in pure SQL.

The RosettaHUB Excel add-in makes it possible to synchronize a local Excel sheet with a RosettaHUB-Sheet on the cloud: Excel becomes capable of accessing any R or Python function as a formula and can interact seamlessly with powerful cloud-based capabilities, likewise, any Excel VBA function or data can be seamlessly exposed and shared to the web. RosettaHUB-sheet are the first bi-dimensional data science notebooks they give access to the most popular data-science tools and aim to contribute to the democratization and pervasiveness of data science.



*Keywords:* Python, web, Spreadsheet, EXCEL, collaboration, VBA, SPARK

Presentation type: Oral Presentation

# Visual Pruner: A Shiny app for cohort selection in observational studies

**Lauren R. Samuels & Robert A. Greevy, Jr.**

*Vanderbilt University School of Medicine*

**Abstract:** Observational studies are a widely used and challenging class of studies. A key challenge is selecting a study cohort from the available data, or “pruning" the data, in a way that produces both sufficient balance in pre-treatment covariates and an easily described cohort from which results can be generalized. Although many techniques for pruning exist, it can be difficult for analysts using these methods to see how the cohort is being selected. Consequently, these methods are underutilized in research. Visual Pruner is a free, easy-to-use Shiny web application that can improve both the credibility and the transparency of observational studies by letting analysts use updatable linked visual displays of estimated propensity scores and important baseline covariates to refine inclusion criteria. By helping researchers see how the pre-treatment covariate distributions in their data relate to the estimated probabilities of treatment assignment (propensity scores), the app lets researchers make pruning decisions based on covariate patterns that are otherwise hard to discover. The app yields a set of inclusion criteria that can be used in conjunction with further statistical analysis in R or any other statistical software. While the app is interactive and allows iterative decision-making, it can also easily be incorporated into a reproducible research workflow. Visual Pruner is currently hosted by the Vanderbilt Department of Biostatistics and can also be run locally within R or RStudio. For links and additional resources, see http://biostat.mc.vanderbilt.edu/VisualPruner.

*Keywords:* observational studies, propensity scores, interactive graphics, visualization, Shiny

Presentation type: Oral Presentation

# Applying R in Streaming and Business Intelligence Applications


## Lou Bajuk-Yorgan

*TIBCO Software*



**Abstract:** R provides tremendous value to statisticians and data scientists. However, they are often challenged to integrate their work and extend that value to the rest of their organization. This presentation will demonstrate how the R language can be used in Business Intelligence applications (such as Financial Planning and Budgeting, Marketing Analysis, and Sales Forecasting) to put advanced analytics into the hands of a wider pool of decisions makers. We will also show how R can be used in streaming applications (such as TIBCO Streambase) to rapidly build, deploy and iterate predictive models for real-time decisions. TIBCO's enterprise platform for the R language, TIBCO Enterprise Runtime for R (TERR) will be discussed, and examples will include fraud detection, marketing upsell and predictive maintenance.



*Keywords:* Real time, streaming, BI, applications, enterprise

Presentation type: Oral Presentation

# Zero-overhead integration of R, JS, Ruby and C/C++


## Lukas Stadler

*Oracle Labs*



**Abstract:** R is very powerful and flexible, but certain tasks are best solved by using R in combination with other programming languages. GNU R includes APIs to talk to some languages, e.g., Fortran and C/C++, and there are interfaces to other languages provided by various packages, e.g., Java and JS. All these interfaces incur significant overhead in terms of performance, usability, maintainability and overall system complexity. This is caused, to a large degree, by the different execution strategies employed by different languages, e.g., compiled vs. interpreted, and by incompatible internal data representations.

The Truffle framework addresses these issues at a very fundamental level, and builds the necessary polyglot primitives directly into the runtime. Consequently, the FastR project, whose goal is to deliver an alternative but fully-compatible R runtime, leverages this infrastructure to allow multiple languages to interact transparently and seamlessly. All parts of a polyglot application can be compiled by the same optimizing compiler, called Graal, and can be executed and debugged simultaneously, with little to no overhead at the language boundary.

This talk introduces FastR and the basic concepts driving Truffle's transparent interoperability, along with a demo of the polyglot capabilities of the FastR runtime.



*Keywords:* polyglot, fastr, performance, ruby, c, c++, JavaScript, interfaces

Presentation type: Oral Presentation

# Taking R to new heights for scalability and performance


## Mark Hornick

*Oracle Corporation*



**Abstract:** Big Data is all the rage, but how can enterprises extract value from such large accumulations of data as found in the growing corporate "data lakes" or "data reservoirs." The ability to extract value from big data demands high performance and scalable tools – both in hardware and software. Increasingly, enterprises take on massive predictive modeling projects, where the goal is to build models on multi-billion row tables or build thousands or millions of models. Data scientists need to address use cases that range from modeling individual customer behavior to understand aggregate behavior or tailoring predictions at the individual customer level, to monitoring sensors from the Internet of Things for anomalous behavior. While R is cited as the most used statistical language, limitations of scalability and performance often restrict its use for big data. In this talk, we present scenarios both on Hadoop and database platforms using R. We illustrate how Oracle Advanced Analytics' R Enterprise interface and Oracle R Advanced Analytics for Hadoop enable taking R to new heights for scalability and performance.



*Keywords:* Big Data, R, Hadoop, Database, Scalability, Internet of Things

Presentation type: Oral Presentation

# Data validation infrastructure: the validate package


## Mark van der Loo & Edwin de Jonge

*Statistics Netherlands*



**Abstract:** Data validation consists of checking whether data agrees with prior knowledge or assumptions about the process that generated the data, including collecting it. Such knowledge can often be expressed as a set of short statements, or rules, which the data must satisfy in order to be acceptable for further analyses.

Such rules may be of technical nature or express domain knowledge. For example, domain knowledge rules include 'Someone who is unemployed can not have an employer (labour force survey)', 'the total profit and cost of an organization must add up to the total revenue (business survey)' and the price of a product in this period must lie within 20% of last year's price (in consumer price index data).

Data validation is an often recurring step in a multi-step data cleaning process where the progress of data quality is monitored throughout. For this reason, the validate package allows one to define data validation rules externally, confront them with data and gather and visualize results.

With the validate package, data validation rules become objects of computation that can be maintained, manipulated and investigated as separate entities. For example, it becomes possible to automatically detect contradictions in certain classes of rule sets. Maintenance is supported by import and export from and to free text or yaml files, allowing rules to be endowed with metadata.



*Keywords:* data cleaning, data quality, data validation rules, Domain Specific Language

Presentation type: Oral Presentation

# Bayesian analysis of generalized linear mixed models with JAGS


## Martyn Plummer

*International Agency for Research on Cancer*



**Abstract:** BUGS is a language for describing hierarchical Bayesian models which syntactically resembles R. BUGS allows large complex models to be built from smaller components. JAGS is a BUGS interpreter written in C++ which enables Bayesian inference using Markov Chain Monte Carlo (MCMC). Several R packages provide interfaces to JAGS (e.g. jags, runjags, R2jags, bayesmix, iBUGS, jagsUI, HydeNet). The efficiency of MCMC depends heavily on the sampling methods used. Therefore a key function of the JAGS interpreter is to identify design motifs in a large complex Bayesian model that have well-characterized MCMC solutions and apply the appropriate sampling methods. Generalized linear models (GLMs) form a recurring design motif in many hierarchical Bayesian models. Several data augmentation schemes have been proposed that reduce a GLM to a linear model and allow efficient sampling of the coefficients. These schemes are implemented in the glm module of JAGS. The glm module also includes an interface to the sparse matrix algebra library CHOLMOD, allowing the analysis of GLMs with sparse design matrices. The use of sparse matrices in a Bayesian GLM renders the distinction between "fixed" and "random" effects irrelevant and allows all coefficients of a generalized linear mixed model to be sampled in the same comprehensive framework.



*Keywords:* MCMC, sparse matrices, GLM, C++

Presentation type: Oral Presentation

# ranger: A fast implementation of random forests for high dimensional data

**Marvin N. Wright & Andreas Ziegler**

*Universität zu Lübeck, Germany*

**Abstract:** Random forests are widely used in applications, such as gene expression analysis, credit scoring, image processing or genome-wide association studies (GWAS). With currently available software, the analysis of high dimensional data is time-consuming or even impossible for very large datasets. We therefore introduce ranger, a fast implementation of random forests, which is particularly suited for high dimensional data. We describe the implementation, illustrate the usage with examples and compare runtime and memory usage with other implementations. ranger is available as standalone C++ application and R package. It is platform independent and designed in a modular fashion. Due to efficient memory management, datasets on genome-wide scale can be handled on a standard personal computer. We illustrate this by application to a real GWAS dataset. We show that ranger is a fast and memory efficient implementation of random forests to analyze high dimensional data. Compared with other implementations, the runtime of ranger proves to scale best with the number of features, samples, trees, and features tried for splitting.

*Keywords:* random forests, machine learning, recursive partitioning, classification, R, C++, Rcpp

Presentation type: Oral Presentation

# Fast additive quantile regression in R

**Matteo Fasiolo, Simon N. Wood, Yannig Goude & Raphael Nedellec**

*University of Bristol, University of Bristol, EDF R&D, EDF R&D*

**Abstract:** Quantile regression represents a flexible approach for modelling the impact of several covariates on the conditional distribution of the dependent variable, which does not require making any parametric assumption on the observations density. However, fitting quantile regression models using the traditional pinball loss is computationally expensive, due to the non-differentiability of this function. In addition, if this loss is used, extending quantile regression to the context of non-parametric additive models become difficult. In this talk we will describe how the computational burden can be reduced, by approximating the pinball loss with a differentiable function. This allows us to exploit the computationally efficient approach described by [1], and implemented by the mgcv R package, to fit smooth additive quantile models. Beside this, we will show how the smoothing parameters can be selected in a robust fashion, and how reliable uncertainty estimated can be obtained, even for extreme quantiles. We will demonstrate this approach, which is implemented by an upcoming extension of mgcv, in the context of probabilistic forecasting of electricity demand.

[1] Wood, S. N., N. Pya, and B. Safken (2015). Smoothing parameter and model selection for general smooth models. http://arxiv.org/abs/1511.03864

*Keywords:* Quantile regression, Additive models, Smoothing

Presentation type: Oral Presentation

# Wrap your model in an R package!


## Michael Rustler & Hauke Sonnenberg

*Kompetenzzentrum Wasser Berlin gGmbH, Kompetenzzentrum Wasser Berlin gGmbH*



**Abstract:** The groundwater drawdown model WTAQ-2, provided by the United States Geological Survey for free, has been "wrapped" into an R package, which contains functions for writing input files, executing the model engine and reading output files. By calling the functions from the R package a sensitivity analysis, calibration or validation requiring multiple model runs can be performed in an automated way. Automation by means of programming improves and simplifies the modelling process by ensuring that the WTAQ-2 wrapper generates consistent model input files, runs the model engine and reads the output files without requiring the user to cope with the technical details of the communication with the model engine. In addition the WTAQ-2 wrapper automatically adapts cross-dependent input parameters correctly in case one is changed by the user. This assures the formal correctness of the input file and minimises the effort for the user, who normally has to consider all cross-dependencies for each input file modification manually by consulting the model documentation. Consequently the focus can be shifted on retrieving and preparing the data needed by the model. Modelling is described in the form of version controlled R scripts so that its methodology becomes transparent and modifications (e.g. error fixing) trackable. The code can be run repeatedly and will always produce the same results given the same inputs. The implementation in the form of program code further yields the advantage of inherently documenting the methodology. This leads to reproducible results which should be the basis for smart decision making.



*Keywords:* groundwater modelling, reproducibility, automation

Presentation type: Oral Presentation

# Teaching R to 200 people in a week


## Michael Andrew Levy

*University of California, Davis*



**Abstract:** Across disciplines, scholars are waking up to the potential benefits of computational competence. This has created a surge in demand for computational education which has gone widely underserved. Software Carpentry and similar efforts have worked to fill this gap with short, intensive introductions to computational tools, including R. Such an approach has numerous advantages; however, it is labor intensive, with student:instructor ratios typically below ten, and it is diffuse, introducing three major tools in two days. I recently adapted Software Carpentry strategies and tactics to provide a deeper introduction to R over the course of a week with a student:instructor ratio above 50. Here, I reflect on what worked and what I would change, with the goal of providing other educators with ideas for improving computational education. Aspects of the course that worked well include live coding during lectures, which builds in flexibility, demonstrates the debugging process, and forces a slower pace; multiple channels of feedback combined with flexibility to adapt to student needs and desires; and iterative, progressively more-open-ended exercises to solidify syntactical understanding and relate functions, idioms, and techniques to larger goals. Aspects of the course that I would change and caution other educators about include increasing the frequency and shortening the duration of student exercises, delaying the introduction of non-standard evaluation, and avoiding any prerequisite statistical understanding. These and other suggestions will benefit a variety of R instructors, whether for intensive introductions, traditional computing courses, or as a component of statistics courses.



*Keywords:* education, instruction, pedagogy

Presentation type: Oral Presentation

# Checkmate: Fast and Versatile Argument Checks


## Michel Lang & Bernd Bischl

*TU Dortmund University, LMU Munich*



**Abstract:** Dynamically typed programming languages like R allow programmers to interact with the language using an interactive Read-eval-print-loop (REPL) and to write generic, flexible and concise code. On the downside, as the R interpreter has no information about the expected data type, dynamically typed programming languages usually lack formal argument checks during runtime. Even worse, many R functions automatically convert the input to avoid throwing an exception. This results in exceptions which are hard to debug. In the worst case, the lack of argument checks leads to undetected errors and thus wrong results. To mitigate this issue, runtime assertions can be manually inserted into the code to ensure correct data types and content constraints, and useful debugging information is generated if the former are violated.

The package checkmate offers an extensive set of functions to check the type and relevant characteristics of the most frequently used data types in R. For example, the function 'assertInteger' also allows to check for missing values, lower and upper bounds, min/exact/max length, duplicated values or names. The package is mostly written in C to avoid any unnecessary performance overhead. Thus, the programmer can write assertions which not only outperform custom R code for such purposes, but are also much shorter and more readable. Furthermore, checkmate can simplify the writing of unit tests by extending the testthat package with many new expectation functions. Third-party packages can link against checkmate's C code to conveniently check arbitrary SEXPs in compiled code.



*Keywords:* package development, software engineering, defensive programming, assertions

Presentation type: Oral Presentation

# Statistics and R in Forensic Genetics


## Mikkel Meyer Andersen, Poul Svante Eriksen & Niels Morling

*Aalborg University, Denmark, Aalborg University, Denmark, University of Copenhagen, Denmark*



**Abstract:** Genetic evidence is often used as evidence in disputes. Mostly, the genetic evidence is DNA profiles and the disputes are often familial or crime cases. In this talk, we go through the statistical framework of evaluating genetic evidence by calculating an evidential weight. The focus will be the statistical aspects of how DNA material from the male Y chromosome can help resolve sexual assult cases. In particular, how an evidential weight of Y chromosomal DNA can be calculated using various statistical methods and how the methods use statistics and R. One of the methods is the discrete Laplace method which is a statistical model consisting of a mixture of discrete Laplace distributions (an exponential family). We demonstrate how inference for that method was initially done using R's built-in glm function with a new family function for the discrete Laplace distribution. We also explain how inference was speeded up by recognising the model as a weighted two-way layout with implicit model matrix and how this was implemented as a special case of iteratively reweighted least squares.



*Keywords:* Forensic genetics, DNA, evidential weight, exponential family, glm

Presentation type: Oral Presentation

# A first-year undergraduate data science course


## Mine Cetinkaya-Rundel

*Duke University*



**Abstract:** In this talk we will discuss an R based first-year undergraduate data science course taught at Duke University for an audience of students with little to no computing or statistical background. The course focuses on data wrangling and munging, exploratory data analysis, data visualization, and effective communication. The course is designed to be a first course in statistics for students interested in pursuing a quantitative major. Unlike most traditional introductory statistics courses, this course approaches statistics from a model-based, instead of an inference-based, perspective, and introduces simulation-based inference and Bayesian inference later in the course. A heavy emphasis is placed on reproducibility (with R Markdown) and version control and collaboration (with git/GitHub). We will discuss in detail course structure, logistics, and pedagogical considerations as well as give examples from the case studies used in the course. We will also share student feedback and assessment of the success of the course in recruiting students to the statistical science major.



*Keywords:* teaching, education, data science, pedagogy

Presentation type: Oral Presentation

# Two-sample testing in high dimensions


## Nicolas Städler, Sach Mukherjee & Frank Dondelinger

*Netherlands Cancer Institute, German Centre for Neurodegenerative Diseases, Lancaster Medical School, Lancaster University*



**Abstract:** Estimation for high-dimensional models has been widely studied. However, uncertainty quantification remains challenging. We put forward novel methodology for two-sample testing in high dimensions (Städler and Mukherjee, JRSSB, 2016). The key idea is to exploit sparse structure in the construction of the test statistics and in p-value calculation. This renders the test effective but leads to challenging technical issues that we solve via novel theory that extends the likelihood ratio test to the high-dimensional setting. For computation we use randomized data-splitting: sparsity structure is estimated using the first half of the data, and p-value calculation is carried out using the second half. P-values from multiple splits are aggregated to give a final result. Our test is very general and applicable to any model class where sparse estimation is possible. We call the application to graphical models Differential Network. Our method is implemented in the recently released Bioconductor package nethet. Besides code for high-dimensional testing the package provides other tools for exploring heterogeneity from high-dimensional data. For example, we make a novel network-based clustering algorithm available and provide several visualization functionalities. Molecular networks play a central role in biology. An emerging notion is that networks themselves are thought to differ between biological contexts, such as cell type, tissue type, or disease state. As an example we consider protein data from The Cancer Genome Atlas. Differential Network applied to this data set provides evidence over thousands of patient samples in support of the notion that cancers differ at the protein network level.



*Keywords:* Two-sample testing in high dimensions, Data splitting, Sparsity, Non-nested hypotheses, Gaussian graphical models, Differential network, Bioconductor package, TCGA protein data

Presentation type: Oral Presentation

# Calculation and economic evaluation of acceptance sampling plans


## Nikola Kasprikova & Jindrich Klufa

*University of Economics in Prague, Czech Republic*



**Abstract:** Sampling inspection is one of the quality control tools used in industry to help keep the quality of the products at satisfactory level while at the same time having the cost in control. When using acceptance sampling inspection, a decision on whether the lot of items is to be accepted or rejected is based on results of inspecting a sample of items from the lot. Acceptance sampling plans which minimize the mean inspection cost per lot of the process average quality when the remainder of rejected lots is inspected were originally designed by Dodge and Romig for the inspection by attributes. Sampling plans for the inspection by variables were then proposed and it has been shown that such plans may be more economical than the corresponding attributes sampling plans. We recall the calculation and economic performance evaluation of the variables sampling plans, show how further improvements in inspection cost could be achieved using EWMA-based statistic and we comment on some of the possibilities available for calculation and evaluation of the plans in R extension package LTPDvar.



*Keywords:* Inspection by variables, inspection cost optimization, LTPD and AOQL acceptance sampling plans

Presentation type: Oral Presentation

# Estimation of causal effects in network-dependent data


## Oleg Sofrygin & Mark J. van der Laan

*University of California, Berkeley*



**Abstract:** We describe two R packages which facilitate causal inference research in network-dependent data: `simcausal` package for conducting network-based simulation studies; and `tmlenet` package for the estimation of various causal effects in `simcausal`-simulated, or real-world network datasets. In addition to the estimation of various causal effects, the `tmlenet` package implements several approaches to estimation of standard errors for dependent (non-IID) data with known network structure. Both packages implement a new syntax that repurposes the list indexing operator '`[[...]]`' for specifying complex network-based data summaries of the observed covariates. For example, `sum(A[[1:Kmax]])` will specify a network-driven summary, evaluated for each unit $i$ as a sum of the variable `A` values for all "friends" of $i$. This new syntax is fully generalizable towards any type of user-defined functions and any type of networks. The practical applicability of both packages is then illustrated with a large-scale simulation study of a hypothetical highly-connected community with an intervention that aimed to increase the level of physical activity by (i) educating a simulated study population of connected subjects, and/or (ii) by intervening on the network structure itself. We will describe how our work can be extended to complex network processes that evolve over time, and discuss possible avenues for future research on estimation of causal effects in longitudinal network settings.



*Keywords:* networks, causal inference, dependent data, simulation, semi-parametric estimation

Presentation type: Oral Presentation

# Implementing R in old economy companies: From proof-of-concept to production


## Oliver Bracht

*eoda GmbH*



**Abstract:** In old economy companies, the introduction of R is typically a button-up process that follows a pattern of three major stages of maturity: At the first stage, guerrilla projects use R parallel to the “official" IT environment. The usage of R is often initiated by interns, student assistants or newly recruited graduates. At the second stage, when the results of the guerrilla projects attract the attention of business departments, R is used as analytic language in proof-of-concept projects. When the proof-of-concept has been successful, the outcome shall be transferred to the production system. At this stage R is being introduced “officially” to the IT environment. While the first and second level of maturity usually do not cause any major problems, the step to the third level is most crucial for the long term success of the implementation of R. This talk will focus on how to master the switch from proof-of-concept to production. It will show based on real world experiences typical road blocks as well as the most important success factors.



*Keywords:* R in business, old economy, R in production, maturity stages

Presentation type: Oral Presentation

# bamdit: An R Package for Bayesian Meta-Analysis of Diagnostic Test Data


## Pablo Emilio Verde

*University of Düsseldorf*



**Abstract:** In this work we present the R package bamdit, its name stands for "Bayesian meta-analysis of diagnostic test-data". bamdit was developed with the aim of simplifying the use of models in meta-analysis, that up to now have demanded great statistical expertise in Bayesian meta-analysis. The package implements a series of innovative statistical techniques including: the Bayesian Summary Receiver Operating Characteristic curve, the use of prior distributions that avoid boundary estimation problems of component of variance and correlation parameters, analysis of conflict of evidence and robust estimation of model parameters. In addition, the package comes with several published examples of meta-analysis that can be used for illustration or further research in this area.



*Keywords:* meta-analysis, diagnostic test data, hierarchical models, conflict of evidence, bias modeling, MCMC, JAGS

Presentation type: Oral Presentation

# Detection of Differential Item Functioning with difNLR function

**Patricia Martinkova, Adela Drabinova & Ondrej Leder**

*Czech Academy of Sciences, Charles University in Prague, Charles University in Prague*

**Abstract:** In this work we present a new method for detection of Differential Item Functioning (DIF) based on Non-Linear Regression. Detection of DIF has been considered one of the most important topics in measurement and is implemented within packages `difR`, `lordif` and others. Procedures based on Logistic Regression are one of the most popular in the study field, however, they do not take into account possibility of guessing or probability of carelessness, which are expectable in multiple-choice tests or in patient reported outcome measures. Methods based on Item Response Theory (IRT) models can count for guessing or for carelessness/inattention, but these latent models may be harder to explain to general audience. We present an extension of Logistic Regression procedure by including probability of guessing and probability of carelessness. This general method based on Non-Linear Regression (NLR) model is used for estimation of Item Response Function and for detection of uniform and non-uniform DIF in dichotomous items. Simulation study suggests that NLR method outperforms or is comparable with the LR-based or IRT-based methods. The new `difNLR` function provides a nice graphical output and is presented as part of Shiny application `ShinyItemAnalysis`, which is available online.

*Keywords:* differential item functioning, item response theory, psychometrics

Presentation type: Oral Presentation

# Using R in a regulatory environment: FDA experiences.


## Paul H Schuette

*FDA*



**Abstract:** The Food and Drug Administration (FDA) regulates products which account for approximately one fourth of consumer spending in the United States of America, and has global impact, particularly for medical products. This talk will discuss the Statistical Software Clarifying Statement (http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm445917.htm), which corrects the misconception that FDA requires the use of proprietary software for FDA submissions. Next, we will describe several use cases for R at FDA, including review work, research, and collaborations with industry, academe and other government agencies. We describe advantages, challenges and opportunities of using R in a regulatory setting. Finally, we close with a brief demonstration of a Shiny openFDA application for the FDA Adverse Event Reporting System (FAERS) available at https://openfda.shinyapps.io/LRTest/.



*Keywords:* Regulatory, Biopharmaceutical, Government

Presentation type: Oral Presentation

# Visualizing multifactorial and multi-attribute effect sizes in linear mixed models with a view towards sensometrics.

**Per Bruun Brockhoff, Isabel de Sousa Amorim, Alexandra Kuznetsova, Søren Bech & Renato R. de Lima**

*DTU Compute, Danish Technical University, Universidade Federal de Lavras, Campus da UFLA, DTU Compute, Danish Technical University, Aalborg University, Universidade Federal de Lavras, Campus da UFLA*

**Abstract:** In Brockhoff et al (2016), the close link between Cohen's d, the effect size in an ANOVA framework, and the so-called Thurstonian (Signal detection) d-prime was used to suggest better visualizations and interpretations of standard sensory and consumer data mixed model ANOVA results. The basic and straightforward idea is to interpret effects relative to the residual error and to choose the proper effect size measure. For multi-attribute bar plots of F-statistics this amounts, in balanced settings, to a simple transformation of the bar heights to get them transformed into depicting what can be seen as approximately the average pairwise d-primes between products. For extensions of such multi-attribute bar plots into more complex models, similar transformations are suggested and become more important as the transformation depends on the number of observations within factor levels, and hence makes bar heights better comparable for factors with differences in number of levels. For mixed models, where in general the relevant error terms for the fixed effects are not the pure residual error, it is suggested to base the d-prime-like interpretation on the residual error. The methods are illustrated on a multifactorial sensory profile data set and compared to actual d-prime calculations based on ordinal regression modelling through the ordinal package. A generic "plug-in" implementation of the method is given in the SensMixed package, which again depends on the lmerTest package. We discuss and clarify the bias mechanisms inherently challenging effect size measure estimates in ANOVA settings.

*Keywords:* Effect Size; Analysis of Variance; F test; d-prime; Sensometrics

Presentation type: Oral Presentation

# bigKRLS: Optimizing non-parametric regression in R

**Pete Mohanty & Robert B. Shaffer**

*Stanford University, University of Texas at Austin*

**Abstract:** Data scientists are increasingly interested in modeling techniques involving relatively few parametric assumptions, particularly when analyzing large or complex datasets. Though many approaches have been proposed for this situation, Hainmueller and Hazlett's (2014) Kernel-regularized Least Squares (KRLS) offers statistical and interpretive properties that are attractive for theory development and testing. KRLS allows researchers to estimate the average marginal effect (the slope) of an explanatory variable but (unlike parametric regression techniques whether classical or Bayesian) without the requirement that researchers know the functional form of the data generating process in advance. In conjunction with Tichonov regularization (which prevents overfitting), KRLS offers researchers the ability to investigate heterogeneous causal effects in a reasonably robust fashion. Further, KRLS estimates offers researchers several avenues to investigate how those effects depend on other observable, explanatory variables.

We introduce bigKRLS, which markedly improves memory management over the existing R package, which is key since RAM usage is proportional to the number of observations squared. In addition, we allow users parallelize key routines (with the snow library) and shift matrix algebra operations to a distributed platform if desired (with bigmemory and bigalgebra).

As an example, we estimate a model from a voter turnout experiment. The results show how the effects of a randomized treatment (here, a get-out-the-vote message) depend on other variables. Finally, we briefly discuss which post-estimation quantities of interest will help users determine whether they have sufficiently large sample size for the asymptotics on which KRLS relies.

*Keywords:* non-parametric regression, big data, R, KRLS

Presentation type: Oral Presentation

# GNU make for reproducible data analysis using R and other statistical software


## Peter John Baker

*Dr*



**Abstract:** As a statistical consultant, I often find myself repeating similar steps for data analysis projects. These steps follow a pattern of reading, cleaning, summarising, plotting and analysing data then producing a report. This is always an iterative process because many of these steps need to be repeated, especially when quality issues are present or overall goals change. Reproducibility becomes more difficult with increasing complexity.

For very small projects or toy examples, we may be able to do all analysis steps and reporting in a single markdown document. However, to increase efficiency for larger data analysis projects, a modular programming approach can be adopted. Each step in the process is then carried out using separate R syntax or markdown files. GNU Make automates the mundane task of regenerating output given dependencies between syntax, markdown and data files in a project. For instance, if we store results from time consuming analyses and radically change a report, we only need to rerun the R markdown file for reporting. On the other hand, if initial data are changed, we rerun everything. In both cases, we can set up our favourite IDE to use Make and simply press the 'build' button.

To extend Make for R, Rmarkdown, SAS and STATA, I have written pattern rules which are available on github. These are used by adding a single line to the project Makefile. An overall strategy and constructing a simple Makefile for a data analysis project will be briefly outlined and demonstrated.



*Keywords:* data analysis, workflow, make, don't repeat yourself, reproducible research

Presentation type: Oral Presentation

# Rectools: An Advanced Recommender System


## Pooja Rajkumar & Norman Matloff

*University of California, Davis*



**Abstract:** Recommendation engines have a number of different applications. From books to movies, they enable the analysis and prediction of consumer preferences. The prevalence of recommender systems in both the business and computational world has led to clear advances in prediction models over the past years.

Current R packages include recosystem and recommenderlab. However, our new package, rectools, currently under development, extends its capabilities in several directions. One of the most important differences is that rectools allows users to incorporate covariates, such as age and gender, to improve predictive ability and better understand consumer behavior.

Our software incorporates a number of different methods, such as non-negative matrix factorization, random effects models, and nearest neighbor methods. In addition to our incorporation of covariate capabilities, rectools also integrates several kinds of parallel computation.

Examples of real data will be presented, and results of computational speedup experiments will be reported; results so far have been very encouraging. Code is being made available on GitHub, at `https://github.com/Pooja-Rajkumar/rectools`.



*Keywords:* Recommender systems, covariates, non-negative matrix factorization, random effects, nearest neighbor, parallel computation

Presentation type: Oral Presentation

# How to use the archivist package to boost reproducibility of your research


## Przemyslaw Biecek & Marcin Kosinski

*University of Warsaw, Warsaw University of Technology*



**Abstract:** The R package archivist allows you to share and reproduce R objects - artifacts with other researchers, either through a knitr script, embedded hooks in figure/table captions, shared folder or github/bitbiucket repositories.

Key functionalities of this package include: (i) management of local and remote repositories which contain R objects and objects meta-data (properties of objects and relations between them); (ii) archiving R objects to repositories; (iii) sharing and retrieving objects by their unique hooks; (iv) searching for objects with specific properties / relations to other objects; (v) verification of object's identity and object's context of creation.

The package archivist extends, in combination with packages such as knitr and Sweave, the reproducible research paradigm by creating new ways to retrieve and validate previously calculated objects. These functionalities also result in a variety of opportunities such as:

- sharing R objects within reports or articles by adding hooks to R objects in table or figure captions;
- interactive exploration of object repositories;
- caching function calls;
- retrieving object's pedigree along with information about session info.



*Keywords:* reproducible research, automated reports, repository with R objects

Presentation type: Oral Presentation

# Deep Learning for R with MXNet

**Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, Zheng Zhang, Qiang Kou & Tong He**

*University of Washington, Carnegie Mellon University, Stanford University, National University of Singapore, TuSimple, New York University, Microsoft, University of Alberta, Massachusetts Institute of Technology, NYU Shanghai, Indiana University, Simon Fraser University*

**Abstract:** MXNet is a multi-language machine learning library to ease the development of ML algorithms, especially for deep neural networks. Embedded in the host language, it blends declarative symbolic expression with imperative tensor computation. It offers auto differentiation to derive gradients. MXNet is computation and memory efficient and runs on various heterogeneous systems. The MXNet R package brings flexible and efficient GPU computing and state-of-art deep learning to R. It enables users to write seamless tensor/matrix computation with multiple GPUs in R. It also enables users to construct and customize the state-of-art deep learning models in R, and apply them to tasks such as image classification and data science challenges. Due to the portable design, the MXNet R package can be installed and used on all operating systems supporting R, including Linux, Mac and Windows. In this talk I will provide an overview of the MXNet platform. With demos of state-of-art deep learning models, users can build and modify deep neural networks according to their own need easily. At the same time, the GPU backend will ensure the efficiency of all computing work.

*Keywords:* deep learning, machine learning, high-performance computing

Presentation type: Oral Presentation

# Modeling Food Policy Decision Analysis with an Interactive Bayesian Network in Shiny


## Rachel Lynne Wilkerson

*Baylor University*



**Abstract:** The efficacy of policy interventions for socioeconomic challenges, like food insecurity, is difficult to measure due to a limited understanding of the complex web of causes and consequences. As an additional complication, limited data is available for accurate modeling. Thorough risk based decision making requires appropriate statistical inference and a combination of data sources. The federal summer meals program is a part of the safety net for food insecure families in the US, though the operations of the program itself are subject to risk. These uncertainties stem from variables both about internal operations as well as external food environment. Local partners often incur risk in operating the program; thus we use decision analysis to minimize the risks. After integrating public, private, and government data sources to create an innovative repository focused on the operations of the child nutrition programs, we construct a Bayesian network of variables that determine a successful program and compute the expected utility. Through an expected utility analysis, we can identify the key factors in minimizing the risk of program operations. This allows us to optimize the possible policy interventions, offering community advocates a data driven approach to prioritizing possible programmatic changes. This work represents substantial progress towards innovative use of government data as well as a novel application of Bayesian networks to public policy. The mathematical modeling is also supplemented by a community-facing application developed in Shiny that aims to educate local partners about evidence based decision making for the program operations.



*Keywords:* food policy, decision analysis, Bayesian networks, shiny

Presentation type: Oral Presentation

# Htmlwidgets: Power of Javascript in R


## Ramnath Vaidyanathan, Yihui Xie, J.J. Allaire, Joe Cheng & Kenton Russell

*Alteryx, RStudio, RStudio, RStudio*



**Abstract:** htmlwidgets is an R package that provides a comprehensive framework to create interactive javascript based widgets, for use from R. Once created, these widgets can be used at the R console, embedded in an R Markdown report, or even used inside a Shiny web application. In this talk, I will introduce the concept of a “htmlwidget", and discuss how to create, develop and publish a new widget from scratch. I will discuss how multiple widgets can be composed to create dashboards and interactive reports. Finally, I will touch upon more advanced functionality like auto-resizing and post-render callbacks, and briefly discuss some of the exciting developments in this area. There has been significant interest in the R community to bring more interactivity into visualizations, reports and applications. The htmlwidgets package is an attempt to simplify the process of developing interactive widgets, and publishing them for more widespread usage in the R community.



*Keywords:* visualizations, javascript, interactive graphics, shiny

Presentation type: Oral Presentation

# bayesboot: An R package for easy Bayesian bootstrapping


## Rasmus Bååth

*Lund University*



**Abstract:** Introduced by Rubin in 1981, the Bayesian bootstrap is the Bayesian analogue to the classical non-parametric bootstrap and it shares the classical bootstrap's advantages: It is a non-parametric method that makes weak distributional assumptions and that can be used to calculate uncertainty intervals for any summary statistic. Therefore, it can be used as an inferential tool even when the data is not well described by standard distributions, for example, in A/B testing or in regression modeling. The Bayesian bootstrap can be seen as a smoother version of the classical bootstrap. But it is also possible to view the classical bootstrap as an approximation to the Bayesian bootstrap.

In this talk I will explain the model behind the Bayesian bootstrap, how it connects to the classical bootstrap and in what situations the Bayesian bootstrap is useful. I will also show how one can easily perform Bayesian bootstrap analyses in R using my package bayesboot (https://cran.r-project.org/package=bayesboot).



*Keywords:* Bayesian, bootstrap, R package, Statistical inference, non-parametric statistics

Presentation type: Oral Presentation

# Superheat: Supervised heatmaps for visualizing complex data


## Rebecca Louise Barter & Bin Yu

*University of California, Berkeley*



**Abstract:** Technological advancements of the modern era have enabled the collection of huge amounts of data in science and beyond. Accordingly, computationally intensive statistical and machine learning algorithms are being used to seek answers to increasingly complex questions. Although visualization has the potential to be a powerful aid to the modern information extraction process, visualizing high-dimensional data is an ongoing challenge. Here, we introduce the supervised heatmap, called superheat, which is a new graph that builds upon the existing clustered heatmaps that are widely used in fields such as bioinformatics. Supervised heatmaps have two primary aims: to provide a means of visual extraction of the information contained within high-dimensional datasets, and to provide a visual assessment of the performance of model fits to these datasets. We will use two case studies to demonstrate the practicality and usefulness of supervised heatmaps in achieving these goals. The first will examine crime in US communities for which we will use the supervised heatmaps to gain an in-depth understanding of the information contained within the data, the clarity of which is unparalleled by existing visualization methods. The second case study will explore neural activity in the visual cortex where we will use supervised heatmaps to guide an exploration of the suitability of a Lasso-based linear model in predicting brain activity. Supervised heatmaps are implemented via the superheat package written in the R programming software and is currently available via github.



*Keywords:* data visualization, heatmaps, exploratory data analysis, model assessment, multivariate data

Presentation type: Oral Presentation

# Reusable R for automation, small area estimation and legacy systems


## Rhydwyn McGuire, Helen Moore & Michael Nelson

*New South Wales Ministry of Health*



**Abstract:** Running a complex model once is easy, just pull up your statistical program of choice, plug in the data, the model and off you go. The problem comes when you then find yourself trying to scale to running that model with different data hundreds or thousands of times. In order to scale and save analysts from spending all their time running models over and over again you need automation. You need a well-designed and tested environment. You need well-engineered R. You also need to sell it to analysts. We wanted to use the tools of software engineering and reusable research to allow statisticians and epidemiologists to be more efficient, but statisticians and epidemiologists are not computer scientists and a lot of this world is new to them. So we had to develop not only for good software practice but to ensure that others could use our tools, even when it comes with a very different focus to what they might be used to.

Using the example of batch small area estimation using generalized additive models, we will talk about the project, the tools we used and how to integrate R into a legacy SAS environment with a minimum of pain, allowing for uptake of the strengths of R without exposing new users to its complexity.



*Keywords:* Reproducible research, spatial statistics, generalized additive models

Presentation type: Oral Presentation

# Run-time Testing Using assertive


## Richard James Cotton

*Weill Cornell Medicine - Qatar*



**Abstract:** assertive is a group of R packages that lets you check that your code is running as you want it to. assert_* functions test a condition and throw an error if it fails, letting you write robust code more easily. Hundreds of checks are available for types and properties of variables, file and directory properties, numbers, strings, a variety of data types, the state of R, your OS and IDE, and many other conditions. The packages are optimised for easy to read code and easy to understand error messages.



*Keywords:* assertions,testing,programming,package development

Presentation type: Oral Presentation

# Introducing the permutations package


## Robin Hankin

*Auckland University of Technology*



**Abstract:** A 'permutation' is a bijection from a finite set to itself. Permutations are important and interesting objects in a range of mathematical contexts including group theory, recreational mathematics, and the study of symmetry. This short talk will introduce the 'permutations' R package for manipulation and display of permutations. The package has been used for teaching pure mathematics, and contains a number of illustrative examples. The package is fully vectorized and is intended to provide R-centric functionality in the context of elementary group theory. The package includes functionality for working with the "megaminx", a dodecahedral puzzle with similar construction to the Rubik cube; the megaminx puzzle is a pleasing application of group theory and the package was written specifically to analyze the megaminx. From a group-theoretic perspective, the center of the megaminx group comprises a single non-trivial element, the 'superflip'. The superflip has a distinctive and attractive appearance and one computational challenge is to find the shortest sequence that accomplishes the superflip. Previously, the best known result was a superflip of 83 turns, due to Clarke. The presentation will conclude by showing one result of the permutations package: an 82-turn superflip.



*Keywords:* computational software, group theory, megaminx, permutations, superflip

Presentation type: Oral Presentation

# Revisiting the Boston data set (Harrison and Rubinfeld, 1978)


## Roger Bivand

*Norwegian School of Economics*



**Abstract:** In the extended topical sphere of Regional Science, more scholars are addressing empirical questions using spatial and spatio-temporal data. An emerging challenge is to alert "new arrivals" to existing bodies of knowledge that can inform the ways in which they structure their work. It is a particular matter of opportunity and concern that most of the data used is secondary. This contribution is a brief review of questions of system articulation and support, illuminated retrospectively by a deconstruction of the Harrison and Rubinfeld (1978) Boston data set and hedonic house value analysis used to elicit willingness to pay for clean air.



*Keywords:* Spatial support, spatial econometrics, hedonic models

Presentation type: Oral Presentation

# R AnalyticFlow 3: Interactive Data Analysis GUI for R


## Ryota Suzuki & Tatsuhiro Nagai

*Ef-prime, Inc.*



**Abstract:** R AnalyticFlow 3 is an open-source GUI for data analysis on top of R. It is designed to simplify the process of data analysis for both R experts and beginners. It is written in Java and runs on Windows, OS X and Linux. Interactive GUI modules are available to perform data analysis without writing code, or you can write R scripts if you prefer. Then you can connect these modules (or scripts) to build an "analysis flow", which is a workflow representing the processes of data analysis. An analysis flow can be executed by simple mouse operation, which facilitates collaborative works among people with different fields of expertise. R AnalyticFlow 3 is extensible: you can easily build custom GUI modules to add functions that you need. Custom module builder is available for this purpose, which itself is a simple, user-friendly GUI to design custom modules. Any R function including your original R script can be converted to a GUI module. It also provides typical tools such as code editor, object/file browser, graphics device, help browser and R console. There are also many useful features including code completion, debugger, object-caching, auto-backup and project manager. R AnalyticFlow 3 is freely available from our website (`http://www.ef-prime.com`). The source code is licensed under LGPL, and works with other open-source libraries including JRI, JUNG and Substance.



*Keywords:* GUI, Java, JRI

Presentation type: Oral Presentation

# mumm: An R-package for fitting multiplicative mixed models using the Template Model Builder (TMB)

**Sofie Pødenphant Jensen, Kasper Kristensen & Per Bruun Brockhoff**

*DTU Compute - Technical University of Denmark, DTU Aqua - Technical University of Denmark, DTU Compute - Technical University of Denmark*

**Abstract:** Non-linear mixed models of various kinds are fundamental extensions of the linear mixed models commonly used in a wide range of applications. An important example of a non-linear mixed model is the so-called multiplicative mixed model, which we will consider as a model with a linear mixed model part and one or more multiplicative terms. A multiplicative term is here a product of a random effect and a fixed effect, i.e. a term that models a part of the interaction as a random coefficient model based on linear regression on the fixed effect. The multiplicative mixed model can be applied in many different fields for improved statistical inference, e.g. sensory and consumer data, genotype-by-environment data, and data from method comparison studies in medicine. However, the maximum likelihood estimation of the model parameters can be time consuming without proper estimation methods. Using automatic differentiation techniques, the Template Model Builder (TMB) R-package [Kristensen, 2014] fits mixed models through user-specified C++ templates in a very fast manner, making it possible to fit complex models with up to $10^6$ random effects within reasonable time. The mumm R-package uses the TMB package to fit multiplicative mixed models, such that the user avoids the coding of C++ templates. The package provides a function, where the user only has to give a model formula and a data set as input to get the multiplicative model fit together with standard model summaries such as parameter estimates and standard errors as output.

*Keywords:* Non-linear mixed models, multiplicative mixed model, TMB, sensory data

Presentation type: Oral Presentation

# Helping R Stay in the Lead by Deploying Models with PFA


## Stuart Bailey & Robert Grossman

*Open Data Group*, *University of Chicago*



**Abstract:** We introduce a new language for deploying analytic models into products, services and operational systems called the Portable Format for Analytics (PFA). PFA is an example of what is sometimes called a model interchange format, a standard and domain specific language for describing analytic models that is independent of specific tools, applications or systems. Model interchange formats allow one application (the model producer) to export models and another application (the model consumer or scoring engine) to import models. The core idea behind PFA is to support the safe execution of statistical functions, mathematical functions, and machine learning algorithms and their compositions within a safe execution environment. With this approach, the common analytic models used in data science can be implemented, as well as the data transformations and data aggregations required for pre- and post-processing data. We will discuss the deployment of models developed in R using PFA, why PFA is strategically important for the R community, and the current state of R libraries for PFA exporting and manipulation of models developed in R.



*Keywords:* R, predictive analytics, Portable Format for Analytics, domain specific language, model deployment

Presentation type: Oral Presentation

# Size of Datasets for Analytics and Implications for R


## Szilard Pafka

*Epoch*



**Abstract:** With so much hype about “big data" and the industry pushing for distributed computing vs traditional single-machine tools, one wonders about the future of R. In this talk I will argue that most data analysts/data scientists don’t actually work with big data the majority of the time, therefore using immature “big data" tools is in fact counter productive. I will show that contrary to widely-spread believes, the increase of dataset sizes used for analytics has been actually outpaced in the last 10 years by the increase in memory (RAM), making the use of single-machine tools ever more attractive. Furthermore, base R and several widely used R packages have undergone significant performance improvements (I will present benchmarks to quantify this), making R the ideal tool for data analysis on even relatively large datasets. In particular, R has access (via CRAN packages) to excellent high-performance machine learning libraries (benchmarks will be presented), while high-performance and parallel computing facilities have been part of the R ecosystem for many years. Nevertheless, the R community shall of course continue pushing the boundaries and extend R with new and ever more performant features.



*Keywords:* high-performance computing, machine learning, (no) big data

Presentation type: Oral Presentation

# Heatmaps in R – Overview and Best Practices


## Tal Galili & Yoav Benjamini

*Tal Aviv University*



**Abstract:** A heatmap is a popular graphical method for visualizing high-dimensional data, in which a table of numbers are encoded as a grid of colored cells. The rows and columns of the matrix are ordered to highlight patterns and are often accompanied by dendrograms. Heatmaps are used in many fields for visualizing observations, correlations, missing values patterns, and more.

This talk will provide an overview of R functions and packages for creating useful and beautiful heatmaps. Attention will be given to data pre-processing, choosing colors for the data-matrix via viridis, producing thoughtful dendrograms using dendextend and colorspace, while ordering the rows and columns with DendSer (and seriation). The talk will cover both static as well as the newly available interactive plotting engines using packages such as gplots, d3heatmap, ggplot2 and plotly.

The speaker is the author of the dendextend R package, a co-author of the d3heatmap package, and blogs at `www.r-statistics.com`.



*Keywords:* heatmap, dendrogram, visualization, interactive visualization, hierarchical clustering, plotly, dendextend

Presentation type: Oral Presentation

# Using Jupyter notebooks with R in the classroom


## Tanya Tickel Schlusser

*unaffiliated*



**Abstract:** When teaching statistics to non-programmers, the challenges of programming in R often exceed the challenge presented by new statistics concepts. This presentation will discuss a recent paper comparing methods for teaching programming (Jacobs, Gorman, Rees, and Craig, 2016), including the use of Jupyter notebooks. Jupyter notebooks are run in a server-client Notebook Application that allows editing and running Jupyter notebooks in a web browser. The audience will be able to execute a live Jupyter notebook running R code, demonstrating the most successful approach in their paper.



*Keywords:* teaching,jupyter

Presentation type: Oral Presentation

# Data Landscapes: a pragmatic and philosophical visualisation of the sustainable urban landscape

**Tatjana Kecojevic & Alan Derbyshire**

*University of Central Lancashire, UK, Manchester School of Architecture, UK*

**Abstract:** The Vernacular Ecology Index (VEI) is a newly proposed assessment method for sustainable urban development. It is composed of five elements (energy, culture, systems, placeness and vernacular) that are indicative of the spirit of the real and illusory within the context of the urban ecosystem. Within the components there are integrated indicators that aim to reflect and measure the viability of the individual element. When synthesized with their counterparts the index indicates strengths and areas in need of improvement within the designated study subject. Most importantly the index acts as a visual illustration of ecological progress as it is a critical intention to involve communities in the process of ecological appraisal, or put simply 'mutual interaction'. One of the primary purposes of the VEI tool is to establish networks of benchmark practice in order to stimulate feedback loops to complimentary regions, ultimately benefitting the broader bioregion. Applying the index to a number of projects in a stipulated locality effectively offers an aerial image of the urban ecosystem's health that could potentially pinpoint ecological strengths and weaknesses of the identified region.

This talk will illustrate how VEI's graphical information model is developed using R's grammar of graphics, which allows clear representation of the five categories with the ability to establish a rating for each of the components. Trough the use of shiny, R enables interactive communications with users for imputing VEI's assessment data that can be presented on Google Maps for building spatial aerial image of a regional assessment.

*Keywords:* sustainability, vernacular, urban development, visualisation

Presentation type: Oral Presentation

# Wrapping Your R tools to Analyze National-Scale Cancer Genomics in the Cloud


## Tengfei Yin & Nan Xiao

*Seven Bridges Genomics*



**Abstract:** The Cancer Genomics Cloud (CGC), built by Seven Bridges and funded by the National Cancer Institute hosts The Cancer Genome Atlas (TCGA), that is one of the world's largest cancer genomics data collections. Computational resources and optimized, portable bioinformatics tools are provided to analyze the cancer data at any scale immediately, collaboratively, and reproducibly. Seven Bridges platform is not only available on AWS but also available on google cloud as well. With Docker and Common Workflow Language open standard, wrapping a tool in any programming language into the cloud and compute on petabyte of data has never been so easy. Open source R/Bioconductor package 'sevenbridges' is developed to provide full API support to Seven Bridges Platforms including CGC, supporting flexible operations on project, task, file, billing, apps etc, users could easily develop fully automatic workflow within R to do an end-to-end data analysis in the cloud, from raw data to report. What's most important, 'sevenbridges' packages also provides interface to describe your tools in R and make it portable to CWL format in JSON and YAML, that you can share easily with collaborators, execute it in different environment locally or in the cloud, everything is fully reproducible. Combined with the R API client functionality, users will be able to create a CWL tool in R and execute it in the cancer genomics cloud to analyze the huge amount of cancer data at scale.



*Keywords:* NCI Cancer Genomic Cloud pilot, Seven Bridges Genomics, CWL, docker, R package

Presentation type: Oral Presentation

# On the emergence of R as a platform for emergency outbreak response


## Thibaut Jombart, The Hackout 1,2,3 Teams & Neil Ferguson

*Imperial College London*, *Various Places*, *Imperial College London*



**Abstract:** The recent Ebola virus disease outbreak in West Africa has been a terrible reminder of the necessities of rapid evaluation and response to emerging infectious disease threats. For such response to be fully informed, complex epidemiological data including dates of symptom onsets, locations of the cases, hospitalisation, contact tracing information and pathogen genome sequences have to be analysed in near real time. Integrating all these data to inform Public Health response is a challenging task, which typically involves a variety of visualisation tools and statistical approaches. However, a unified platform for outbreak analysis has been lacking so far. Some recent collaborative efforts, including several international hackathons, have been made to address this issue. This talk will provide an overview of the current state of R as a platform for the analysis of disease outbreaks, with an emphasis on lessons learnt from a direct involvement with the recent Ebola outbreak response.



*Keywords:* outbreak, epidemic, hackathon, emergency, modeling, data visualisation

Presentation type: Oral Presentation

# Differential equation-based models in R: An approach to simplicity and performance


## David Kneis & Thomas Petzoldt

*TU Dresden, TU Dresden*



**Abstract:** The world is a complex dynamical system, a system evolving in time and space in which numerous interactions and feedback loops produce phenomena that defy simple explanations. Differential-equation models are powerful tools to improve understanding of dynamic systems and to support forecasting and management in applied fields of mathematics, natural sciences, economics and business. While lots of effort has been put into the fundamental scientific tools, applying these to specific systems requires significant programming and re-implementation. The resulting code is often quite technical, hindering communication and maintenance. We present an approach to: (1) make programming more generic, (2) generate code with high performance (3) improve sustainability, and (4) support communication between modelers, programmers and users by:

- automatic generation of Fortran code (package rodeo) from spreadsheet tables containing state variables, parameters, processes, interactions and documentation,
- numerical solution with general-purpose solvers (package deSolve),
- web-based interfaces (package shiny), that can be designed manually or auto-generated from the model tables (package rodeoApp),
- creation of docs in LaTeX or HTML.

Package rodeo uses a stoichiometry-matrix notation (Petersen matrix) of reactive transport models and can generate R or Fortran code for ordinary and 1D partial differential equation models, e.g. with longitudinal or vertical structure. The suitability of the approach will be shown with two ecological models of different complexity: (1) antibiotic resistance gene transfer in the lab, (2) algae bloom control in a lake.



*Keywords:* differential equation models, stoichiometry matrix, numerical solver, code generation, Fortran, R, ecological modelling

Presentation type: Oral Presentation

# xgboost: An R package for Fast and Accurate Gradient Boosting


## Tong He

*Simon Fraser University*



**Abstract:** XGBoost is a multi-language library designed and optimized for boosting trees algorithms. The underlying algorithm of xgboost is an extension of the classic gradient boosting machine algorithm. By employing multi-threads and imposing regularization, xgboost is able to utilize more computational power and get more accurate prediction compared to the traditional version. Moreover, a friendly user interface and comprehensive documentation are provided for user convenience. The package has been downloaded for more than 4,000 times on average from CRAN per-month, and the number is growing rapidly. It has now been widely applied in both industrial business and academic researches. The R package has won the 2016 John M. Chambers Statistical Software Award. From the very beginning of the work, our goal is to make a package which brings convenience and joy to the users. In this talk, I will briefly introduce the usage of xgboost, as well as several highlights that we think users would love to know.



*Keywords:* Machine Learning, xgboost, gradient boosting machine, R package, John M. Chambers Statistical Software Award

Presentation type: Oral Presentation

# Most Likely Transformations


## Torsten Hothorn

*University of Zurich*



**Abstract:** The “mlt" package implements maximum likelihood estimation in the class of conditional transformation models. Based on a suitable explicit parameterisation of the unconditional or conditional transformation function using infrastructure from package “basefun", we show how one can define, estimate and compare a cascade of increasingly complex transformation models in the maximum likelihood framework. Models for the unconditional or conditional distribution function of any univariate response variable are set-up and estimated in the same computational framework simply by choosing an appropriate transformation function and parameterisation thereof. As it is computationally cheap to evaluate the distribution function, models can be estimated by maximisation of the exact likelihood, especially in the presence of random censoring or truncation. The relatively dense high-level implementation in the “R" system for statistical computing allows generalisation of many established implementations of linear transformation models, such as the Cox model or other parametric models for the analysis of survival or ordered categorical data, to the more complex situations illustrated in this paper.



*Keywords:* transformation model, transformation analysis, distribution regression, conditional distribution function, conditional quantile function, censoring, truncation

Presentation type: Oral Presentation

# Fitting Complex Bayesian Models with R-INLA and MCMC


## Virgilio Gómez-Rubio & Carlos Gil-Bellosta

*Universidad de Castilla-La Mancha, Datanalytics*



**Abstract:** The Integrated Nested Laplace Approximation (INLA) provides a computationally efficient approach to obtaining an approximation to the posterior marginals for a large number of Bayesian models. In particular, INLA focuses on those models that can be expressed as a Latent Gaussian Markov Random field. Its associated R package, R-INLA, implements a number of functions to easily fit many of these models. However, it is not easy to implement new latent models or priors. Bivand et al. (2014) proposed a way of using R-INLA to fit models that are not implemented, by fixing some parameters in the model and then combining the fitted models using Bayesian Model Averaging (BMA). This is implemented in the INLABMA R package. An interesting feature of this approach is that it allows Bayesian models to be fitted in parallel. Recently, Gomez-Rubio et al. (2016) have proposed the use of MCMC and INLA together to fit more complex models. This approach allows INLA to fit models with unimplemented (or multivariate) priors, missing data in the covariates and many more latent models. Finally, we will explore how these ideas can be applied to fit models to Big Data. This involves fitting models to separate chunks of data with R-INLA and then combining the output to obtain an approximation to the model with all the data.



References:

Bivand et al. (2014). Approximate Bayesian Inference for Spatial Econometrics Models. Spatial Statistics 9, 146-165.

Gomez-Rubio et al. (2016). Extending INLA with MCMC. Work in progress.



*Keywords:* INLA, MCMC, Bayesian Inference

Presentation type: Oral Presentation

# Interactive Terabytes with pbdR


## Wei-Chen Chen, Drew Schmidt & George Ostrouchov

*pbdR Core Team, Silver Spring, MD, USA, pbdR Core Team, Knoxville, TN, USA, pbdR Core Team, Oak Ridge, TN, USA; Oak Ridge National Laboratory*



**Abstract:** Historically, large scale computing and interactivity have been at odds. A new series of packages have recently been developed to attempt to rectify this problem. We do so by combining two programming models: client/server (CS) and single program multiple data (SPMD). The client/server allows the R programmer to control from one to thousands of batch servers running as cooperating remote instances of R. This can easily be done from a local R or RStudio session. The communication is handled by the well-known ZeroMQ library, with a new set of package bindings available to R by way of the pbdZMQ package. The client and server are implemented in the new remoter and pbdCS packages. To handle computations, we use the established pbdR packages for large scale distributed computing. These packaegs utilize HPC standards like MPI and ScaLAPACK to handle complex coupled computations on truly large data. These tools use the batch SPMD programming model, and constitute the server portion of the client/server hierarchy. So once the client issues a command, it is transmitted to the SPMD servers and executed in a massively parallel fashion.This talk will discuss the package components and provide timing results for some Terabyte size computations running on hundreds of cores of a cluster.



*Keywords:* Command line interfaces, Distributed programming language

Presentation type: Oral Presentation

# Advancing In-memory Analytics and Interoperability for R (and Python)


## Wes McKinney

*Cloudera, Inc.*



**Abstract:** In this talk, I'll discuss how projects like Apache Arrow are ushering in a new architectural roadmap for interoperability between traditional in-memory / single-node programming languages like R and Python and the rest of the modern large-scale data management and database ecosystem. It will show the current state of the art and opportunities for R to become more useful as a component in more complex systems involving many different data processing tools.



*Keywords:* data structures, file formats, big data, R, Python

Presentation type: Oral Presentation

# Colour schemes in data visualisation: Bias and Precision

**William K. Cornwell, Kendra Luong, Rita Sousa & Rich FitzJohn**

*UNSW, Australia, UNSW, Australia, University of Technology, Sydney, Australia, University College London, UK*

**Abstract:** The technique of mapping continuous values to a sequence of colours, is often used to visualise quantitative data. The ability of different colour schemes to facilitate data interpretation has not been thoroughly tested. Using a survey framework built with Shiny and loggr, we compared six commonly used colour schemes in two experiments: a measure of perceptually linearity and a map reading task for: (1) bias and precision in data interpretation, (2) response time and (3) colour preferences. The single-hue schemes were unbiased — perceived values did not consistently deviate from the true value, but very imprecise — large data variance between the perceived values. Schemes with hue transitions improved precision, however they were highly biased when not close to perceptually linearity (especially for the multi-hue 'rainbow' schemes). Response time was shorter for the single-hue schemes and longer for more complex colour schemes. There was no aesthetic preference for any of the colourful schemes. These results show that in choosing a colour scheme to communicate quantitative information, there are two potential pitfalls: bias and precision. Every use of colour to represent data should be aware of the bias–precision trade-off and select the scheme that balances these two potential communication errors.

*Keywords:* graphics, color, colour, shiny

Presentation type: Oral Presentation

# Profvis: Profiling tools for faster R code


## Winston Chang

*RStudio*



**Abstract:** As programming languages go, R has a bit of a reputation for being slow. This reputation is mostly undeserved, and it hinges on the fact that R's copy-on-modify semantics make its performance characteristics different from other many other languages. That said, even the most expert R programmers often write code that could be faster. The first step to making code faster is to find which parts are slow. This isn't an easy task. Sometimes we have no idea what parts of code are expensive, and even when we do have intuitions about it, those intuitions can be wrong. After the slow parts of code have been identified, one can move on to the next step: speeding up that code. In this talk I'll show how to profile and optimize code using profvis, a new package for exploring profiling data. Profvis provides a graphical interface that makes it easy to spot which pieces of code are expensive. I will also discuss why some common operations in R may be surprisingly slow, and how they can be sped up.



*Keywords:* profiling, performance

Presentation type: Oral Presentation

# When will this machine fail?


## Xinwei Xue & James Ren

*Microsoft*



**Abstract:** In this talk, we demonstrate how to develop and deploy end-to-end machine learning solutions for predictive maintenance in manufacturing industry with R. For predictive maintenance, the following questions regarding when a machine fails are typically asked:

What's the Remaining Useful Life (RUL) of an asset? Will an asset fail within a given time frame? Which time window will an asset likely fail?

We formulate the above questions to regression, binary classification and multiclass classification problems respectively, and use a public aircraft engine data to demonstrate the complete modeling steps in R: data labeling, processing, feature engineering, model training and evaluation. R users are often challenged with productizing the models they built. After model development, we will show two ways of productization: 1) deploy with SQL server as stored procedures using the new R services; 2) deploy it by publishing as a web service restful API; Either approach would enable user to call the deployed scoring engine from any applications.

The presentation will be followed by a live demo during the talk.



*Keywords:* Failure prediction; predictive maintenance; model deployment

Presentation type: Oral Presentation

# OPERA: Online Prediction by ExpeRts Aggregation


## Pierre Gaillard & Yannig Goude

*Department of Mathematical Sciences of Copenhagen University, EDF R&D/ University Paris-Sud*



**Abstract:** We present an R package for prediction of time series based on online robust aggregation of a finite set of forecasts (machine learning method, statistical model, physical model, human expertise, ...). More formally, we consider a sequence of observations $y(1),\dots,y(t)$, to be predicted element by element. At each time instance t, a finite set of experts provide prediction $x(k,t)$ of the next observation $y(t)$. Several methods are implemented to combine these expert forecasts according to their past performance (several loss functions are implemented to measure it). These combining methods satisfy robust finite time theoretical performance guarantees. We demonstrate on different examples from energy markets (electricity demand, electricity prices, solar and wind power time series) the interest of this approach both in terms of forecasting performance and time series analysis.



*Keywords:* Online learning; expert aggregation; Energy forecasting

Presentation type: Oral Presentation

# Inside the Rent Zestimates


## Yeng Bun

*Zillow Group*



**Abstract:** Zillow, the leading real estate and rental marketplace in USA, uses R to estimate home values (Zestimates) and rental prices (Rent Zestimates). Every day, we refresh Zestimates and Rent Zestimates on Zillow.com for more than 100 million homes with output from scoring daily re-trained models. The model training and scoring infrastructure rests on top of R, allowing rapid prototyping and deployment to production servers. We make extensive use of R, including the development of an in-house package called ZPL that functions similar to MapReduce on Hadoop, but runs on relational databases. In this presentation, we will go under the hood to see parts of the engine that power the Rent Zestimates.



*Keywords:* R, Rent, Zestimate, big data, parralel computing, production

Presentation type: Oral Presentation

# Authoring Books with R Markdown


## Yihui Xie

*RStudio, Inc.*



**Abstract:** Markdown is a simple and popular language for writing. R Markdown (http://rmarkdown.rstudio.com) has made it really easy to author documents that contain R code, and convert these documents to a variety of output formats, including PDF, HTML, Word, and presentations. There are still some missing pieces in the toolchain, especially when writing long-form articles and books, such as cross-references, automatic numbering of figures/tables, multiple-page HTML output and a navigation system, and so on. The R package bookdown has solved all these problems for several types of output formats, such as HTML, PDF, EPUB and MOBI e-books. The visual style of the book is customizable. When the output format is HTML, the book may contain interactive components, such as HTML widgets and Shiny apps, so readers may interact with certain examples in the book in real time (screenshots of these examples will be automatically taken and used when the output format is non-HTML). In this talk, we will give a quick tour through the bookdown package, and show how to quickly get started with writing a book. We will also talk about various options for editing, hosting, and publishing a book. Our goal is that authors can focus as much as possible on the content of the book, instead of spending too much time on any complicated non-portable syntax of authoring languages, or tools for converting books to different output formats. In other words, "one ring to rule them all."



*Keywords:* R Markdown, Book, Publication, PDF, HTML, E-books

Presentation type: Oral Presentation

# Multiple Hurdle Tobit models in R: the mhurdle package


## Yves Croissant & Fabrizio Carlevaro

*University of La Réunion, University of Geneva*



**Abstract:** mhurdle is a package for R enabling the estimation of a wide set of regression models where the dependent variable is left censored at zero, which is typically the case in household expenditure surveys. These models are of particular interest to explain the presence of a large proportion of zero observations for the dependent variable by means of up to three censoring mechanisms, called hurdles.

For the analysis of censored household expenditure data, these hurdles express a good selection mechanism, a desired consumption mechanism and a purchasing mechanism, respectively. However, the practical scope of these paradigmatic hurdles is not restricted to empirical demand analysis, as they have been fruitfully used in other fields of economics, including labor economics and contingent valuation.

For each these censoring mechanisms, a continuous latent variable is defined, indicating that censoring is in effect when the latent variable is negative. Latent variables are modeled as the sum of a linear function of explanatory variables and of a normal random disturbance with a possible correlation between the disturbances of different latent variables. To model possible departures of the observed dependent variable to normality, we use flexible transformations allowing rescaling skewed or leptokurtic random variables to heteroscedastic normality.

mhurdle models are estimated using the maximum likelihood method for random samples. Model evaluation and selection are tackled by means of goodness of fit measures and Vuong tests.

Real-world illustrations of the estimation of multiple hurdles models are provided using data from consumer expenditure surveys.



*Keywords:* household's expenditure survey analysis, censored regression models, hurdle models, tobit models