Why R? 2019 Abstract Book

Why R? Foundation

Why R? 2019 Abstract Book

Why R? Foundation

Contents

Introduction **3**
Participants . . . 3

List of presentations **4**
API . . . 4
Automating Google Slides creation . . . 4
Bringing interactivity into engineering courses with BERT-based Excel-R applications . . . 4
Google PageSpeed with R . . . 4
BIO 1 . . . 5
AmyloGram: the R package and a Shiny server for amyloid prediction . . . 5
Machine Learning usage for prediction of state change in bipolar disorder . . . 5
BIO 2 . . . 6
Multicenter study, 33 TB of data and the goal: predicting epilepsy . . . 6
On Role and Methods of Data Analysis in Drug Abuse Prevention . . . 6
R in Ministry of Health . . . 6
Business . . . 8
Integrating R and Python for reproducible business analytics . . . 8
R for Entrepreneurs : supply chain automation case . . . 8
EDA . . . 9
How to deal with nested lists in R? Using the purrr, furrr and future packages in practice. . . . 9
Geo . . . 11
Features of districts of Warsaw visible from space . . . 11
Geospatial data analysis and visualization in R . . . 11
Spatial econometrics with self-made weighting matrixes - uncovering similarity of sample with machine learning results and categorical variables. . . . 11
Keynotes . . . 13
The landscape of spatial data analysis in R . . . 13
Lightnings 1 . . . 16
Crazy Sequential Representations - The 10958 Problem . . . 16
RUcausal: An R package for Representing Uncertainty in causal discovery . . . 17
RME: interpretable explainations for sequence models . . . 17

Introduction

The purpose of Why R? is to provide the space for the professional networking and knowledge exchange, between practitioners and students, from the area of statistical machine learning, programming, optimization and data science fields.

Participants

The Why R? events are aimed at experienced data science practitioners although each conference gathers a significant percentage of students (~30%). Our participants recruit from very diverse backgrounds with a clear domination of mathematics (especially statistics) and computer science. All of them have occupations related to data science, including professional R developers (programmers), data engineers, machine learning practitioners and business analysts. One of the key advantages of Why R? is its ability to attract participants both from academia and industry.

List of presentations

API

Automating Google Slides creation

Author: Piotrek Ciurus (Azimo)

Description

I would like to talk about automating Google Slide creation using R. First, complete data workflow will be presented. Second, two possible ways will be reviewed: exporting data file with automation using Google Apps Script and direct slide generation from R script. Finally, I will present practical example of business application.

Bringing interactivity into engineering courses with BERT-based Excel-R applications

Author: Florent Bourgeois (University of Toulouse, Laboratoire de Genie Chimique)

Description

With Excel being the computing tool most used by the engineering community, developing Excel applications that call R functions is highly desirable for engineers as it merges Excel's interactivity with a high level numerical environment. This paper was written with engineering trainers in mind. It should provide them with an applied and illustrative guide for easy development of applications that merge Excel and R using BERT as the interoperability solution. Simple examples are provided that exemplify the ease with which such applications can be created. Such applications, which are interactive by design since they use Excel as their front-end, can help engineering educators increase the attractiveness and dynamics of their engineering courses.

Google PageSpeed with R

Author: Leszek Sieminski (Ringier Axel Springer Polska)

Description

One of more both important and tedious tasks in digital marketing is to optimize website loading. Firstly, introduce the concept and show what tools are required to perform an analysis (with an example). Secondly, I will describe how to enhance the tools' capabilities by using web API and show good practices on a real-world package. Finally, I will also shortly describe the architecture of the code and how to use the results to improve website loading time.

Agenda:

1. What is web performance optimization and why is it important?
2. Tools: Google PageSpeed Insights
3. Report example
4. Problems with the tool & how the API solves them
5. "pagespeedParseR? package
6. Working with API: good practices
7. Package architecture
8. Results & how to use them to improve the website

BIO 1

AmyloGram: the R package and a Shiny server for amyloid prediction

Author: Jaroslaw Chilimoniuk (University of Wroclaw)

Co-authors: Michal Burdukiewicz, Piotr Sobczyk Stefan R?diger, Malgorzata Kotulska and Pawel Mackiewicz

Description

Amyloids are proteins associated with important clinical disorders (e.g., Alzheimer's or Creutzfeldt-Jakob"s diseases). Despite their great diversity, all amyloid proteins can undergo their aggregation initiated by 6- to 15-residue segments. The structure and the function of proteins are encoded in the linear sequence of amino acids. But the aggregation propensity seems to not depend on the exact amino acid residues, but rather on their physicochemical properties. Therefore, we created a model of amyloidogenicity incorporating this knowledge.

We have created 524,284 reduced amino acid alphabets based on diversified combinations of the physicochemical properties of amino acid residues. Using a very fast implementation of the random forest classifier from the ranger package we cross-validated all reduced alphabets and identified one that provided the best discrimination between amyloids and non-amyloids. Our feature selection method found 65 motifs that are the most relevant to the discrimination of amyloid and non-amyloid sequences.

Our toolkit, the biogram R package, provides a set of useful tools for encoding protein sequences into features understandable by machine learning algorithms. Our software, inspired by natural language processing, extracts n-grams of amino acids from proteins and selects only the most informative ones using developed by us Quick Permutation Test (QuiPT).

The reduction of amino acid alphabet turned out very efficient. Most of the predictors based on them outperformed those trained on the full amino acid alphabet. Among 65 most informative amino acid motifs identified during the analysis, 15 were independently confirmed in experimental studies. AmyloGram effectively recognizes patterns responsible for the aggregation (AUC = 0.90, Matthews correlation coefficient: 0.63) outperforming existing amyloid-predicting software.

AmyloGram was further used to predict the properties of amyloid sequences from the AmyLoad database. We analyzed 23 peptides and validated experimentally using ThT assay. 15 out of 23 peptides had the same amyloid properties as predicted by AmyloGram. Moreover, predictions of AmyloGram were verified experimentally as our tool led to the discovery of a novel amyloid protein, MspA, produced by Methanospirillum hungatei JF-1.

Our analysis not only confirmed that amyloidogenicity depends on the general physicochemical properties of proteins but also revealed which features are the most relevant to the initiation of amyloid aggregation.

Despite large computational requirements, the whole study was conducted in R, mostly using functions from packages biogram and ranger. AmyloGram itself is distributed as the R package and a shiny web-server at: www.smorfland.uni.wroc.pl/shiny/AmyloGram/.

Machine Learning usage for prediction of state change in bipolar disorder

Author: Olga Kaminska (Britenet)

Description

State change of patients in bipolar disorder may cause irreversible changes. The aim of the project was to predict the state change towards depression / mania in the examined patients. Each person goes through this disease differently, that's why personalization of the algorithm is so important. The study used methods of both supervised and unsupervised learning through data from real patients. The prepared solution is the first step to create a complete prognostic system that would make life easier for doctors and patients.

BIO 2

Multicenter study, 33 TB of data and the goal: predicting epilepsy

Author: Glowacka Jagoda (Transition Technologies S.A.)

Co-authors: Kamil Sijko, Konrad Wojdan

Description

In EPISTOP project, 101 patients with TSC mutation causing uncontrolled growth of benign tumors were followed up since birth till 2 year of life to observe the epileptogenesis process. The blood were sampled from those patients in 3 or 4 defined moment of seizures development. After that, all of the samples were sequenced. Moreover, EEG, MRI and neuropsychological studies were performed to asses patients? condition and the clinical data were collected. Together 33 TB of data were gathered and more than 60 thousands of features were tested to select potential signs of epilepsy, given the patient condition and the outcome at 2 year of life. Multiple steps were performed in order to extract pattern from the high-dimensional data. Eventually, the goal of the analysis were to correctly point out those patients with increased risk of developing epilepsy in first 2 years of life.

As a part of EPISTOP team, we were responsible for statistical analysis of clinical data and for joined analysis of multiple sources data. All analyses were performed in the R envinroment. At the WhyR? conference we would like to share our experience while working in the multicenter project involving 16 international partners from hospitals and laboratories from 10 countries. We would like to address medical-data related issues such as constantly changing data and patients? statuses or the problem of number of features much greater than number of samples. We will revise our methodology, sharing the good pratices that helped us to keep our research reproducible, together with the unique evidence-based 'NOT-to-do'.

On Role and Methods of Data Analysis in Drug Abuse Prevention

Author: Jakub Weiner (Revolution Train)

Description

The talk will revolve around the data analytical approach to the phenomena of drug exposiure of Central European (CZ, GE, PL, SK) youth as described by the data collected within Revolution Train project and the external sources. Furthermore, a proposed always-on cloud & blockchain infrastructure will be presented with a view on building a solid research approach for the innovations in primary prevention.

R in Ministry of Health

Author: Piotr Nowosielski (Ministry of Health Republic of Poland)

Co-authors: Michal Walczynski, Mariusz Zieba, Klaudiusz Witczak, Filip Wojciechowski

Description

The R language was introduced in the Analyses and Strategies Department of the Ministry of Health in 2015. Since then MS Excel is no longer used as an analytic tool.

1. Why R?
    a) Short transition description from Excel to R
        i. Big data handling
        ii. Dierent data sets
    b) Short transition description from classical statistics to machine learning methods
        i. Basic analyses already shown on the online platform .
        ii. More complex problems waiting for being solved
2. Implementation examples:

a) Patient's paths analyses

Deriving care pathways for a specic disease from administrative data might be troublesome due to the abundance of non-informative entries in patients' long-term history. Therefore, such data necessitates an aggregation method that provides a comprehensive overview of individuals' progression through the public healthcare system accompanied by a compelling visualization. The data for consecutive years were aggregated using stochastic processes with the underlying assumptions: * the process does not acknowledge individuals' history except the last visit in the healthcare system, i.e. it is memoryless (also known as the Markov property); * care pathways are modelled in form of state transition probabilites, whereas exclusive events in history such as a visit in primary care, ambulatory care, an undergone medical procedure etc. constitute the states; * state transition probabilites are non-zero which accounts for the fact that the patients can move between each and every state, inclusive of recurrent visits to the same state; * transition matrices are transformed into directed complete graphs wherein the vertices correspond to particular states and the edges to state transition probabilites.

These simplifying assumptions can be treated as neglibile in view of an extensive sample size.

b) Stroke departments prediction model

In order to receive thrombolysis (life saving service), the patient who underwent the stroke should arrive within 4 hours of the incident on a stroke unit. The Ministry of Health data shows the patient's place of residence at the commune (Polish gmina) level. Due to a low population density in some areas of Poland the model structure must take more aspects into consideration. This model contains approximately 6.5 thousand linear restrictions by 6.5 millions of decision variables. In addition, the objective function has been dened to minimize the distance between a patient (in a straight line) to a unit. Hence, naturally, linear programming methods are applied.

Business

Integrating R and Python for reproducible business analytics

Author: Richard Louden (The Oakland Group)

Description

Analytics without reproducibility, the ability to reproduce an output from its component parts, results in inherent risk. This is especially true in a business environment where staff can and will move to new jobs, leaving projects and work that may be vital for the business. In addition, analytics without collaboration can lead to wholly unsuitable results. This collaboration may come in the form of utilising different programming languages or additional input regarding business context. Both aspects require strong ways of working and suitable toolsets in order to be effective, which will be the main subject of this talk. Utilising previous experience of large businesses, from a management consulting and retail background, I aim to show how reproducibility can be improved with some simple methodologies, and collaboration aided with recent tools. Examples of such includes establishing a strong base to work from via utilising projects and working with paths correctly and improving collaboration with colleagues using tools such as reticulate and the draft redoc package. The overall aim of this talk is to inspire those who do not currently utilise such practices to improve both their own workflows and those utilised within their company, for improved reproducibility and collaboration.

R for Entrepreneurs : supply chain automation case

Author: Francois Jacquet (artinlean sp. zoo)

Description

I would like to share the journey I took to build from scratch a production grade machine learning workflow for automated stock and sales management, with emphasis on :

- Hurdles in accessing/cleaning data at a commercial company and how I could overcome them
- How suprisingly fast it was for me to go from idea to working model in R
- How painful and key it is to make it production grade (and one can in R too)
- Exchange with you on some principles I applied in my journey that helped a lot (Lean start-up, Agile, continuous training, keeping higher purpose. . .)

EDA

How to deal with nested lists in R? Using the purrr, furrr and future packages in practice.

Author: Lidia Kolakowska (Sotrender)

Description

Cleaning and preparing data for analysis is one of the most time-consuming stages in the analyst's work. This process is further prolonged and can be frightening when data is available as nested lists. An example can be data received for analysis in JSON format, downloaded directly from API or non-relational database such as MongoDB.

Based on data on political advertisements, provided by Facebook, in my presentation I will show you how to optimize the processing of data nested in lists using publicly available packages. In addition, I will discuss how to easily write anonymous functions, which are to be iterated after each element of the list, as well as two lists simultaneously. An anonymous function (also known as a lambda expression) is a definition of a function that is not associated with an identifier. This means that it is a function that is created and used, but never assigned to a variable. Using the purrr package, it will show you how to define this type of function in a very short way.

By jumping to a higher level with time-consuming code processing, I will show examples of using furrr and future packages, which with minimal editing will allow you to process the same operations in parallel. And all this in a pipe!

Tag words: nested lists, json data format, iterating over two list at same time, dealing with NULLs in nested lists, filtering lists elements, parallel processing, processing in pipelines, joining data from lists without primary keys

main packages: dplyr, purrr, future, furrr, fs, jsonlite ### MasteR of Tables

Author: Tomasz Żółtak (Educational Research Institute, Warsaw, Poland)

Description

There is a widespread opinion that preparing good looking tables in R is hard. That's not true! Simply some great tools to work with tables in R are not so widely known. On this talk you'll have an opportunity to learn what are these tools and how to use them. Talk will consist of 4 parts:

- 1. tabels package - preparing tables has never been so easy! Learn a very flexible way to describe table content (either summary statistics or counts, or percentages) using formula notation that will allow you to prepare even complicated tables in the blink of an eye.
- 2. Make table look prettier and exactly as you want (either in LaTeX/PDF or in HTML) with kableExtra package (with some remarks also on xtable package).
- 3. Make conditional formatting with formattable package (HTMLonly).
- 4. Pretty model summaries tables at hand with stargazer or texreg packages. ### R Tools for Automated Exploratory Data Analysis

Author: Mateusz Staniak (Warsaw University of Technology)

Co-authors: Przemyslaw Biecek

Description

Before a predictive model is built, the data set must needs to be well understood. This process is usually referred to as the Exploratory Data Analysis (EDA). In the era of countless easily available, but noisy and large data sets, automation of EDA is a task that could greatly speed up data analysis and aid non-experts who need to deal with data. In this talk, I will describe many R packages for fast, automated EDA (autoEDA)

with their strenghts and weaknesses. The talk is based on a paper "The Landscape of R Packages for Automated Exploratory Data Analysis" which was accepted to the R Journal.

Geo

Features of districts of Warsaw visible from space

Author: Krystian Andruszek (Faculty of Economic Sciences, University of Warsaw, Data Science Lab)

Co-authors: Piotr Wójcik, Ewa Sobolewska

Description

Daytime satellite images in high-resolution are commonly used to derive features of regions or smaller areas using convolutional neural networks (CNN). One can identify meaningful features like the number and density of buildings, the prevalence of shadow area as a proxy for building height, the number of cars, density and length of roads, type of farmland, roof material, etc. CNNs are a special kind of multilayer neural networks applied in image recognition. CNNs are identifying boundaries (edges), which separate areas of different colors. Based on low level concepts (a curve, a straight line) one can build more high level concepts (a square, circle, etc.) and even more abstract concepts. Training a neural network is a very demanding and time consuming process that requires powerful computational resources. One of solutions is transfer learning, in which the model is not trained from scratch, but uses some pre-trained model, that was trained before on a large benchmark dataset to solve a similar problem.

The aim of this presentation is to explain the idea of meaningful features extraction from images with the use of convolutional neural networks and transfer learning and show a practical example based on the districts of Warsaw.

Geospatial data analysis and visualization in R

Author: Çizmeli Servet Ahmet (PranaGEO LTD)

Description

Analysis of geospatial data requires specialized software tools. The R ecosystem provides a rich set of powerful open source packages that make it possible to work with geospatial data. In this workshop we provide hands-on exercises with real datasets. We will learn how to import geospatial data in R, make interactive maps, convert between different formats and map projections. We will experiment on spatial queries and perform basic statistical analyses. This is a hands-on workshop so attendees are expected to bring a laptop.

Dr. Servet Ahmet Çizmeli is environmental scientist and tech entrepreneur. He has 20 years of research and education experience in the area of environmental sciences, remote sensing and GIS. He has extensive experience in field data collection and geospatial data analysis and modeling tools. He is the co-founder of melda.io, a cloud-base data science and reproducible research platform.

Spatial econometrics with self-made weighting matrixes - uncovering similarity of sample with machine learning results and categorical variables.

Author: Maria Mikos (University of Warsaw, Department of Economic Sciences)

Description

Crucial part of spatial econometrics are weighting matrixes. However, spatial dependency is not the only relation, that can be adapted in this form. R package spdep:: provides a method to build own matrixes and convert them to listw class. Therefore, this function opens a possibility to utilize user-build objects in modeling. Filtering not only for geographical dependence, but also for heterogeneity of sample, they can significantly reduce the overbias of standard models. They can be used as an alternative for dummy-variable in OLS and exchange adjacency matrix in Spatial Durbin Model. Using Iris dataset and NUTS4 panel data two case-studies were presented. Categorical variable and machine learning results were used to uncover similarity of data. OLS modeling was augmented with self-made weighting matrixes and, as a result, lowest

values of Information Criteria were obtained. The author stressed that weighting matrixes build on categorical data and clustering results can significantly improve econometrical estimation.

Keynotes

The landscape of spatial data analysis in R

Author: Jakub Nowosad (Adam Mickiewicz University, Poznan)

Description

Spatial information surrounds us, from location data stored on your phone to national park borders and satellite images of the Earth. The ubiquity of geographic data and its importance for research explains R's long history of supporting spatial data analysis and the growing 'rspatial' community. Since R inception in the 1990s, hundreds of packages related directly or indirectly to spatial data analysis have been created, made by people from a range of disciplines including spatial statistics, cartography, remote sensing, soil science, geomorphometry, ecology, transportation, archeology, epidemiology, and econometrics. They often have different goals and sometimes use different words to explain the same concepts. Spatial data analysis packages also evolved together with the rest of the R packages with the use of pipes, tidyverse, or Rcpp. As a result, currently, there are many rspatial packages with different APIs, some more prominent than others. Important steps in the history of rspatial were the creation of sp and raster packages, in 2005 and 2010, respectively. The sp package became the basis for hundreds of other packages, allowing spatial (vector) data to be included in a wide range of applied software. The raster package enabled working with large raster data sets. Recently, several new packages have appeared such as sf, the successor to sp. Much development is ongoing, including for terra (described as a successor to raster) and stars (the system for working with irregular spatial data sets). The aforementioned packages are often used as a cornerstone for many applied packages, including those used to retrieve spatial data (e.g., rnaturalearth, rgbif, tidycensus), spatial visualizations (e.g., mapview, ggplot2, cartography, tmap), transportation (e.g., osmr, stplanr), or ecology (e.g. ade4, landscapemetrics). This talk will show several examples of how rspatial can be used to solve different problems. It will also give some hints on how not to get lost in the rspatial landscape and discuss some future directions or the rspatial developments.

Bio

Jakub is an assistant professor in the Department of Geoinformation at the Adam Mickiewicz University in Poznan, Poland. His main research is focused on developing and applying spatial methods in order to broaden our understanding of processes and patterns in the environment. He has extensive teaching experience in the fields of spatial analysis, geostatistics, statistics, and machine learning. Jakub is also an active member of the ####rspatial community and a co-author of the Geocomputation with R book ### Random forests: The first-choice method for every data analysis?

Author: Marvin N. Wright (Leibniz Institute for Prevention Research and Epidemiology, Bremen, Germany)

Description

Random forests rarely win prediction challenges. However, if implementation time and/or computing time are taken into account they are hard to beat. Consequently, in tis talk, I argue that every data analysis should start with a random forest. The talk starts with an overview of settings where random forests excel and where they fail and continues with an in-depth analysis of the reasons for this behavior. Further, I will give an overview of random forest implementations in R and update our previously published runtime comparison (Wright & Ziegler 2017).

Bio

Marvin is a Postdoc at the Leibniz Institute for Prevention Research and Epidemiology in Bremen, Germany. He is the author of several R packages, including the random forest implementation ranger. He holds a Ph.D. in Biostatistics from the University of Lübeck, supervised by Andreas Ziegler. Previously, Marvin worked at the University of Lübeck, was a visiting researcher at the University of Copenhagen and also spent some time in the automotive industry and at health insurance. His main research interests are interpretable

machine learning, genetic epidemiology and survival analysis. ### Modelling and Analysing Interval Data in R

Author: Paula Brito (Faculty of Economics, University of Porto)

Co-author: A. Pedro Duarte Silva (Católica Porto Business School & CEGE, Univ. Católica Portuguesa, Portugal)

Description

Symbolic Data Analysis is concerned with analysing data with intrinsic variability, which is to be taken into account. In Data Mining, Multivariate Data Analysis and classical Statistics, the elements under analysis are generally individual entities for which a single value is recorded for each variable - e.g., individuals, described by age, salary, education level, etc. However, when the elements of interest are classes or groups of some kind - the citizens living in given towns; car models, rather than specific vehicles - then there is variability inherent to the data. Symbolic data goes beyond the usual data representation model, considering variables whose observed values for each element are no longer necessarily single real values or categories, but may assume the form of sets, intervals, or, more generally, distributions.

In this talk we focus on the analysis of interval data, i.e., when the variables' values are intervals of IR. Parametric probabilistic models for interval-valued variables have been proposed and studied by Brito & Duarte Silva (2012). These models are based on the representation of each observed interval by its MidPoint and Log-Range, and Multivariate Normal and Skew-Normal distributions are assumed for the whole set of 2p MidPoints and Log-Ranges of the original p interval-valued variables. The intrinsic nature of the interval-valued variables leads to different structures of the variance-covariance matrix, represented by different possible configurations.

The R package MAINT.Data implements the proposed models and methodologies. This package is built around S4 classes and methods, introducing a specific data class for representing interval data. It includes functions and methods for modelling and analysing interval data, in particular maximum likelihood estimation and statistical tests for the different configurations. Methods for (M)ANOVA and Discriminant Analysis of this data class are also provided. For the Gaussian model, Model-Based Clustering, robust estimation, and outlier detection are implemented; corresponding methods for Robust Discriminant Analysis are also available. MAINT.Data is available at CRAN as part of the R software project.

Bio

Paula Brito is Associate Professor at the Faculty of Economics of the University of Porto, and member of the Artificial Intelligence and Decision Support Research Group (LIAAD) of INESC TEC, Portugal. She holds a doctorate degree in Applied Mathematics from the University Paris Dauphine, and an Habilitation in Applied Mathematics from the University of Porto. Her current research focuses on the analysis of multidimensional complex data, known as symbolic data, for which she develops statistical approaches and multivariate analysis methodologies. In this context, she has been involved in two European research projects and coordinates the Portuguese participation in the H2020 FinTech project. Paula Brito was president of the International Association for Statistical Computing (IASC) in 2013-2015. She has been invited speaker at several international conferences, is regularly member of international program committees, has been chair of COMPSTAT 2008 and will be chair of the upcoming conference IFCS 2021. Web-page: www.fep.up.pt/docentes/mpbrito . ### *tfprobably* correct - adding uncertainty to deep learning with TensorFlow Probability

Author: Sigrid Keydana (RStudio)

Description

In deep learning, it's the numbers that count - mean squared error, accuracy, F2 score, IOU and what not (possibly, or hopefully, in relation to training time and hardware resources). These metrics are commonly obtained for point estimates, like a numeric or a class prediction. In real life though, as they say, "nothing is

certain besides death and taxes". Luckily, we can obtain uncertainty estimates from deep neural networks with tfprobability, the R interface to TensorFlow Probability: tfprobability provides distribution layers that are used seemlessly in a Keras network. Beyond deep learning, it lets you build complex hierarchical models, to be fitted with MCMC or variational inference. As expected, TensorFlow Probability being built on top of TensorFlow, we profit from GPU and distributed training. This talk gives an overview of tfprobability, its features and applications, and shows how you can add back uncertainty where it belongs.

Bio

Sigrid is an Applied Researcher at RStudio. She has experience as a psychologist, software developer and data scientist. She is passionate about exploring the frontiers of deep learning and especially helping users employ the power of deep learning from R. ### Is data science experimenting on people?

Author: Steph Locke (Locke Data)

Description

As more and more businesses and public sector organisations are using data science to change the way they do things or to automate processes, we're changing interfaces and interactions.

- So are we experimenting on people?
- What might our ethical and legal obligations actually be, either way?
- How do we ensure our we don't perpetuate a culture of unethical uses of data science?
- How to we check out models for the perpetuation of discrimination and bias?

Steph will elaborate on some of these questions to help us think more deeply about our role and will propose actions for addressing some of the questions. Attend this keynote to think!

Bio

Steph is the founder of a consultancy in the UK. Her talks, blog posts, conferences, and business all have one thing in common – they help people get started with data science. Steph holds the Microsoft MVP award for her community contributions. In her spare time, Steph plays board games with her husband and takes copious pictures of her doggos. ### Always Be Deploying. How to make R great for machine learning in (not only) Enterprise.

Author: Wit Jakuczun (WLOG Solutions)

Description

For many years software engineers have put enormous effort to develop best practices to deliver stable and maintainable software. How R users can benefit from this experience? I will try to answer this question going through several concepts and tools that are natural for software engineers but are often undervalued by R users.

I will start with a description of the deployment process because this is the ultimate step that exposes all weaknesses. You will learn about structuring R project, using abstractions to manage model's features, automating models building process, optimizing the performance of the solution and the challenges of the deployment process itself.

Lightnings 1

Crazy Sequential Representations - The 10958 Problem

Author: Anne Bras (Erasmus University, the Netherlands)

Co-authors: Vincent van der Velden

Description

Inder Taneja (an Indian mathematician) attempted to write the integers from 1 up to 11111 in terms of 1 to 9 (in increasing and decreasing order) by using addition, subtraction, multiplication, division, exponentiation, parenthesis and/or digit concatenation. For example:

```
9617 = 1+2^3*(45+(6+7)*89) = 9*876+5+(4*3)^(2+1)
9618 = 1*(2+3+4+5)*(678+9) = (9+8+7*(6+54+3))*21
9619 = 1+(2+3+4+5)*(678+9) = 9*87+(6*5+4^3)^2/1
```

These representations are generally referred to as crazy sequential representations (CSR). Interestingly, within the 1 up to 11111 range, only one CSR remains to be identified, namely the increasing CSR for 10958.

Pure brute force approaches to "the 10958 problem" are unfeasible, as billions of lexicographical unique expressions can be generated. However, various techniques can be used to reduce the number of candidate expressions, as was done in this project.

Efficient algorithms (generalizing the concept of CSR to arbitrary bases) were designed and implemented from scratch. Millions of CSR were identified (for the integers from -2147483647 up to 2147483647, in base 3 up to 62). Given the nature of CSR, one might consider CSR to be proof-of-work, as identification is complex, while verification is trivial. ### D3 + DALEX = deep interactive model explanations in R

Author: Hubert Baniecki (MI2 MiNI PW)

Description

Static and dull plots might not be enough anymore. We are inevitably growing into fast paced society. The demand for responsive virtual reality has vastly increased these days.

The answer to that trend is the generic plotD3 function. It uses r2d3 package as an interface to D3 visualizations in R. D3.js is a JavaScript library for manipulating documents based on data. It really helps you bring your data to life. Using r2d3 allows us to produce interactive and animated plots that provide more value and great scalability.

We can greatly benefit from implementing and using plotD3 functions in R packages. Authors can present their ideas in various shapes or forms and easily indicate valuable information. Users get to experience visualisation on another level. It all comes at a cost of learning D3 and maintaining more code.

Explainable AI is a hot topic nowadays. Tools like DALEX, iBreakDown, auditor and many others came into existence to answer one simple question 'Why?'. Improving these explainers with interactive visualization makes them simply powerful. plotD3 allows for interactions between multiple plots that are not possible with more traditional tools out there like ggplot2. It is also free from client-server relationship. This is where it shines and as an example I can present the **dime** package.

Links: https://modeloriented.github.io/dime/ ### Don't walk, run! runner package for rolling window functions.

Author: Dawid Kaledkowski ()

Description

`runner` package contains standard running functions (aka. windowed, rolling, cumulative) with additional options. `runner` provides extended functionality like date windows, handling missings and varying window

size. `runner` brings also rolling streak and rollin which, what extends beyond range of functions already implemented in R packages. Presentation will reveal how easy it is to use `runner`, what package can do and others can't and how good is the computing performance.

RUcausal: An R package for Representing Uncertainty in causal discovery

Author: Ioan Gabriel Bucur (Radboud University Nijmegen)

Co-authors: Tom Claassen, Tom Heskes

Description

Causal discovery is a fundamental problem in scientific research. Understanding the causal links between a set of observed variables is crucial for predicting the effects of interventions and policies. RUcausal is an R package intended to provide robust methods for deriving causal relations from observational data. It contains an efficient implementation of the state-of-the-art 'Bayesian Constraint-Based Causal Discovery' (BCCD) algorithm, which takes as input the data correlation matrix and outputs a single partial ancestral graph (PAG) representing the class of all possible causal graphs over the measured variables, along with estimates for the reliability of each inferred causal relation. RUcausal includes an interface for specifying relevant background knowledge regarding the structure of the graph (e.g., forbid an edge between two variables) or regarding the causal relations between variables (e.g., a variable like gender cannot be causally influenced by other variables in the system). Furthermore, the package provides a routine for generating multivariate Gaussian data from specified causal models and a routine for visualizing the output PAG, which uses the plotting library Rgraphviz.

RME: interpretable explainations for sequence models

Author: Mateusz Kobylka (MI2 DataLab)

Description

Explainable Artificial Intelligence is evolving rapidly and many useful tools have been developed to investigate and expound decisions of complex machine learning models in the past few years, yet there are many fields where XAI is still in its infancy. In the lighting talk I would like to preview a novel method of explaining AI