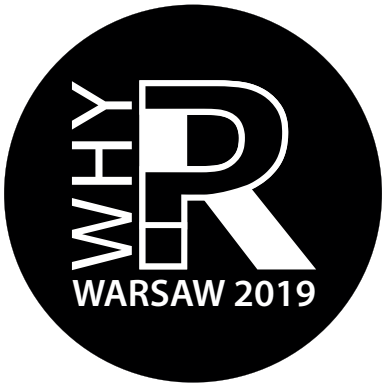

Why R? 2019 Abstract Book

PUBLISHED BY WHY R? FOUNDATION

HTTP://WHYR.PL

September 2019

Contents

1. 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.

Pre-meetings

In 2019, Why R? 2019 was preceded by twelve pre-meetings in seven countries.

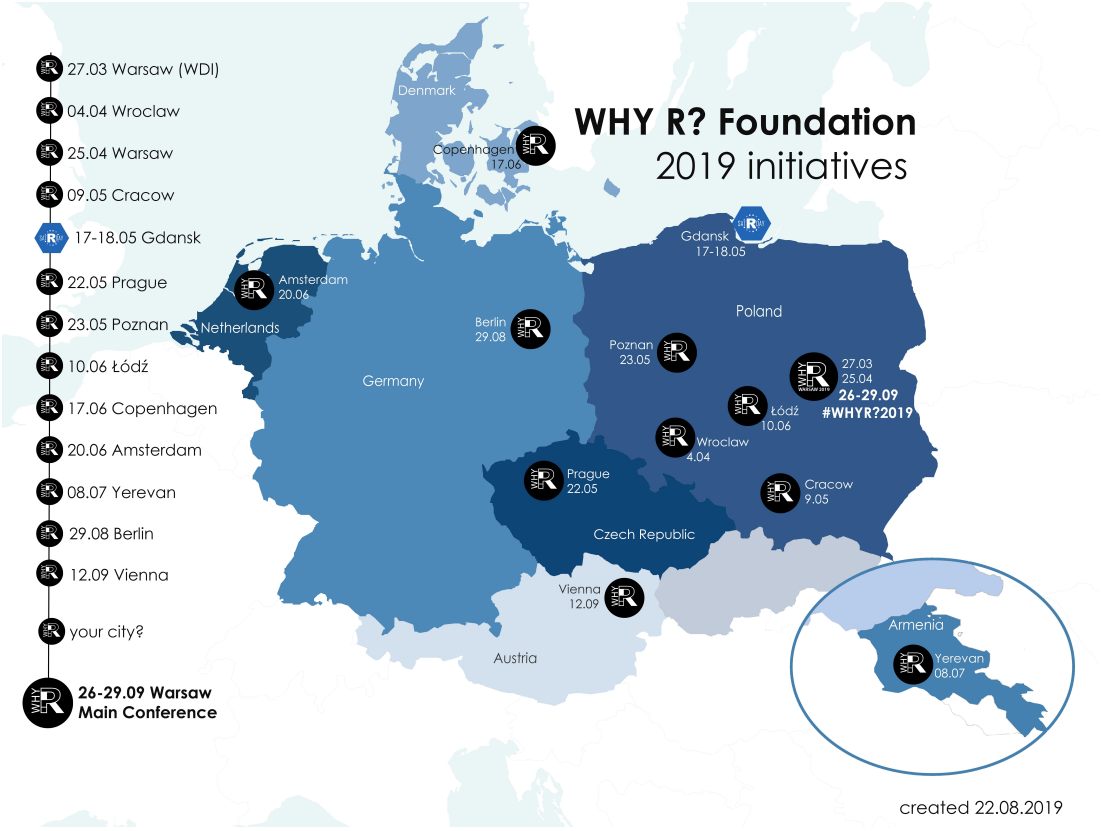

Figure 1.1: Why R? 2019 pre-meetings.

2. 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 Rödiger, 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.

Tidysq for Working with Biological Sequence Data in ML Driven Epitope Prediction in Cancer Immunotherapy

Author: Leon Eyrich Jessen (Technical University of Denmark)

Co-authors: Michał Burdukiewicz, Dominik Rafacz, Weronika Puchała, Jarosław Chilimoniuk, Katarzyna Sidorczuk, Filip Pietluch, Stefan Rödiger

Description

We are amidst a data revolution. Just the past 5 years, the cost of sequencing a human genome has gone down approximately 10-fold. This development moves equally fast within areas such as mass spectrometry, in vitro immuno-peptide screening a.o. This facilitates the search for bio-markers, biologics, therapeutics, etc. but also redefines the requirements for storing, accessing and working with data and the skillset of bio data scientists. In this talk I will present tidysq, an R-package aiming at extending the Tidyverse framework to include (tidy) bio-data-science / bioinformatics. Tidysq will be presented in context with current status in ML driven (neo)epitope prediction within cancer immunotherapy.

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'.

R for experimentalists: HDX-MS example

Author: Weronika Puchala (Institute of Biochemistry and Biophysics Polish Academy of Sciences)

Co-authors: Michal Burdukiewicz, Michal Kistowski, Katarzyna A. Dabrowska, Aleksandra E. Badaczewska-Dawid, Dominik Cysewski, Michal Dadlez

Description

Complex and precise experiments like Mass Spectrometry generate an enormous amount of complicated data. Such datasets require manual pre-processing, which due to their size, is tedious, time-consuming and error-prone. To automatize these steps and also provide a whole analytic workflow, we present HaDeX, an R package for analysis and visualization of Hydrogen/Deuterium Exchange Mass Spectrometry (HDX-MS) data. It facilitates complete data analysis, including quality control, Bayesian framework for differential analysis and ISO-based uncertainty. The sheer volume of data requires highly efficient data processing which is ensured by the data.table package. Our tool also provides a collection of data visualizations that comprehensively summarize HDX-MS results. The package is available on CRAN: https://CRAN.R-project.org/package=HaDeX.

The main audience for HaDeX is HDX-MS practitioners whose area of expertise doesn<92>t include programming and advanced data analysis skills. To help them with their work, HaDeX is also available as a Shiny web server with a wide range of clickable customizable options. For users operating on sensitive data standalone application is available. To ensure publication-quality figures all the plots are fully editable by the user and processed data is easily downloadable in every step. The reproducibility of the analysis perfomed in GUI is ensured with advanced reporting functions. It is important to us that our analytic methodology is transparent and understandable for the users so it is consulted with international experts and discussed in-depth in the package vignette.

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<92>s decisions that is dedicated to models working on sequential data. Recurrent Memory Explainer (RME) is a tool that focuses on the memory of sequential model and tries to explain the decision by pointing important places in the sequence that had an effect on the prediction. RME is based on so-called „memory profiles<94>, which are intuitive and can be easily visualised on simple plot charts. Thanks to the explanations produced by RME, it is possible to determine, for example, whether the more recent or past events had bigger influence on the final decision. The new method of explaining predictions will be presented on the example of medical data and LSTM-based model. The model tries to predict patient's next disease given their disease history, stored in a sequence. RME is able to show which diseases in the history, according to the model, had an impact on the final prediction and how big this impact was. RME was created as a part of master thesis at Warsaw University of Technology, supervised by Przemyslaw Biecek. We would like to thank the LekSeek company for the access to anonymized data.

Selling solutions based on R (which is GPL licensed). Is this possible?

Author: Kamil Sijko (Transition Technologies)

Description

Core R is licensed under GPL-2 / 3 which is really great license from user perspective: you can not only run the program, but also share the app, see and study the source code and even modify it if you wish. But let's look at this from perspective of a person or company that would like to use R skills for living / business. GPL requires you to license your derivative work under GPL. How does this licensing model fit into corporations? Is it possible to integrate R code into closed-sourced solutions? Will this imply changes in licensing? Can we sell R-based solutions to customers? Can they re-sell our work without our consent? I will briefly answer all those questions based on our adventures with integrating R services into our proprietary software and consultations with lawyers and R-community. I will give examples of both good and bad ideas for business models involving R.

Using R6 classes to communicate with a REST API

Author: Patrik Drhlik (Freelance Data Scientist)

Description

I'll show a project where I developed an R package that uses a main R6 class that connects to a remote production or local development PHP server.

The PHP server offers a REST API for getting data into R that can be then analysed by other functions in the package.

The R6 class manages the following: - server authentication using a JWT token - differentiates between a development and a production server - uses a configuration file to store sensitive credentials that shouldn't be in a repository - wraps HTTP requests to always send the JWT token in the HTTP Authorization header - parses all server responses into data frames - exposes an API to get desired data from the server

It would be possible to talk more than 5 minutes about the project. I don't mind any presentation form.

Lightnings 2

autobiograML: towards automated machine learning in protein function prediction

Author: Dominik Rafacz (Warsaw University of Technology)

Co-authors: Katarzyna Sidorczuk, Stefan Rodiger, Przemyslaw Gagat, Michal Burdukiewicz

Description

Background: The advancements in various 'omics' fields have resulted in the discovery of many new protein sequences. Their functional annotations, however, come in at a much slower pace because they require laborious and often expensive experimental procedures. The machine learning models fill in this gap by providing estimates of protein functions. Although they do not replace the experiments, the in silico methods undoubtedly help scientists to understand the ever-growing protein datasets. The challenges of developing appropriate models for protein data exclude from the field scientists with limited machine learning expertise and resources. Therefore, we propose autobiograML, an R package designed to automatically apply our framework for protein function prediction [1, 2]. Methods: autobiograML models the relationships between provided protein sequences (encoded as amino acid motifs) and annotations. The Bayesian framework optimizes the hyperparameters of the model in nested cross-validation. The outer layer of the cross-validation is later used to select the optimal machine learning algorithm. Our software produces not only a model but also a list of important motifs for further studies. Moreover, autobiograML generates Shiny web servers that might be later distributed between less R-savvy users. Results and discussion: Although autobiograML does not cover all the intricacies of machine learning, it offers a reliable way to create a model for the prediction of protein function. Our tool will be a valuable machine learning assistant for many research groups studying areas that are too new to have already well-established computational methods.

[1] Burdukiewicz, M., Sobczyk, P., Rodiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. Scientific Reports 7, 12961.

[2] Burdukiewicz, M., Sobczyk, P., Chilimoniuk, J., Gagat, P., and Mackiewicz, P. (2018). Prediction of Signal Peptides in Proteins from Malaria Parasites. International Journal of Molecular Sciences 19, 3709.

bdl: interface and tools to Local Data Bank API

Author: Krzysztof Kania (Statistics Poland)

Description

Local Data Bank is Poland's largest database of the economy, society and the environment. LDB offers more than 40 thous. statistical features grouped thematically. According to European programme "Open data - access, standards and education" Statistics Poland released REST API to LDB database giving programmers much better access to the data. My package lets download data directly into R environment and provides set of quick tools like summarizing and visualising LDB data on charts and maps.

Links: https://github.com/statisticspoland/R_Package_to_API_BDL

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.

PepBay: Implementation of Bayesian inference in the analysis of peptide arrays

Author: Katarzyna Sidorczuk (University of Wroclaw)

Co-authors: Andreas Weinhäusel, Michal Burdukiewicz

Description

Biomarkers are biological features measurable in biological media, such as human tissues or fluids, that may be indicators of a health state or presence of a disease. To this date, a wide variety of biomarkers in the form of proteins have been identified and routinely used for clinical diagnoses of many diseases, including cancers. The search for new biomarkers may be facilitated by the use of peptide arrays, collections of short protein fragments displayed on a solid surface, as they allow testing thousands of peptides simultaneously.

However, analysis of the data derived from peptide arrays is challenging because of the 'large p small n' problem. It is a common issue in medical studies, where the availability of patients is limited and the number of available covariates is significantly larger than the sample size. Traditional methods fail in such cases, as the correction for multiple testing results in very high p-values. In consequence, it is difficult to distinguish significant effects from noise.

To address these problems, we propose a novel Bayesian approach that utilizes similarities between peptides to better find their associations with the health condition of patients. In addition to its accuracy, our approach also eliminates bias-inducing data pre-processing. To perform Bayesian modeling, we used the R package BEST based on JAGS. In contrast to frequentist methods, it provides a widespread value distribution over, e.g., the effect size that is more interpretable than p-values. Our approach will be implemented in the application PepBay, a tool to facilitate analysis of the results derived from peptide arrays.

R in marketing surveys - how to speed up the analysis of open ended questions

Author: Agnieszka Otreba-Szklarczyk (Herbalife Nutrition)

Description

Open ended questions are very common type of question in marketing surveys. Their analysis consumes a lot of time, especially when the research is conducting in several countries and need to be translate in one language. How to automate this task? The core of this presentation it will be to show how using R (tm, topicmodels, translateR) we can analyze open ended question.

Testing games artificial intelligence algorithms in Shiny

Authors: Łukasz Wawrowski (DOJI S.A., Poznań University of Economics and Business)

Description

R can be the last thing that you think about in the context of games. But in fact it is an very good option if you need to prepare low-cost prototype. In the presentation I will show the set of useful functions those allow to create playable turn-based game similar to well-known Heroes of Might and Magic in Shiny. Moreover a few examples of AI algorithms in games will be presented.

vivo: Is it Victoria In Variable impOrtance detection?

Author: Anna Kozak (MI2DataLab)

Co-authors: Przemyslaw Biecek

Description

Your model predicts something and you would like to know which features were the most important? When a model has many features and plotting all one-dimensional summary statistics is troublesome, vivo indicates which variables are worth paying attention to. The vivo is an R package which calculates instance level feature importance (measure of local sensitivity). The feature importance is based on Ceteris Paribus profiles and can be calculated in a few variants. During the talk I will show how they are different and which to choose.

Links https://cloud.r-project.org/web/packages/vivo/index.html https://github.com/MI2DataLab/vivo

What we don't have but need. Some missing R functions in teaching econometrics

Author: Rafal Wozniak (Faculty of Economic Sciences University of Warsaw)

Description

Despite buoyant number of R packages, some basic areas seem to be undeveloped. Some missing R functions overcomplicate advanced econometrics teaching, and in consequence, can make R less appealing to students. Especially, students not majoring in econometrics might be tempted to use other statistical software. To justify why R is the language of choice in teaching, a new package is proposed. The package tries to bridge the gap of missing R functions. This paper proposes a preliminary version of the new package (fesuw). It consists of a few functions that appear not to be found in available packages. Functions for marginal effects of the binary choice models, ordinal logits, tobit models for a given set of values are proposed alongside with R-squared statistics of static and wide panel models and the linktest for binary dependent models functions.

Modelling 1

Custom loss functions for binary classifications problem with highly imbalanced dataset using Extremely Gradient Boosted Trees

Author: Bartosz Kolasa (DataWalk SA)

Co-authors: Patryk Wielopolski

Description

One of the most common approaches used in binary classification problem is building a model using Extremely Gradient Boosted Trees algorithm that utilizes cross-entropy as a loss function. Unfortunately in context of highly imbalanced dataset this approach is underperforming. While this function is insensitive to the identity of the assigned class in the case of misclassification, in practice it is a very common situation to have skewed sensitivity to error, meaning wrong assignments for one class are much worse than for other. During our talk we would like to present results of experiment where we tested different custom loss function using Extremely Gradient Boosted Trees algorithms for real world application where cost of misclassification error was highly asymmetrical and how to implement such solution in R.

Investment Portfolio Optimization

Author: Michał Podsiadło

Description

The presentation would describe the problem of constructing efficient portfolios and show how to impelement such a solution for the best fitted portfolio so that it would get maximized returns with minimized risks.

Multidimensional Scaling with the smacof package

Author: Barbara Jancewicz (OBM UW)

Description

Multidimensional scaling attempts to recreate respondents' perception map using differences (distances) indicated in their answers. It was popular in psychometrics as well as market research studies. A couple of years ago a new algorithm (smacof) was created so the way in which the algorithm chooses the optimal solution became more intuitive. After few more years the authors of this new algorithm implemented it in R as a package called smacof. I would like to show you how useful multidimensional scaling can be both in academic and commercial setting, as it can help in interpretation and visualisation of complex data. I will also introduce a simple way to verify the quality of the obtained results.

NLP models for the masses with the Quanteda package and a Shiny interface

Author: Ken Benoit, Damian Rodziewicz

Description

In this presentation you'll learn how to build Natural Language Processing(NLP) models with the quanteda package. You will also learn how to deliver NLP technology to non-programming analysts by creating a Shiny interface. Natural Language Processing has become a significant field in applications - customer claim automation, detecting fake news, voice assistants, and social media bots are all NLP work results. The possibility of accessing massive text collections, the rapid development of computational power and emerging tools mean that this field will continue to enjoy great popularity and development. An excellent example of such a tool is Quanteda - an R language package for quantitative analysis of textual data and the winner of the 2018 SAGE Concept Grant.

During the presentation, we will familiarize the audience with the possibilities of Quanteda and present interesting results of a model that was built using it. In the second part, we will tell you what steps we have taken to bring the world of NLP closer to people interested in the field who are not necessarily programmers. More specifically, we will tell you about the Quanteda Graphical User Interface (GUI) project - a web application that opens the door to new communities of users to start their own NLP

projects. In this part, you will have the opportunity to learn about the current state of work on the project, and you'll find out what problems need to be addressed when implementing such solutions.

Modelling 2

Detecting topics in civil service job offers using Latent Dirichlet Allocation model

Author: Adam Bień (SKN Estymator (Poznan University of Economics and Business))

Co-authors: Maciej Beresewicz

Description

The main goal of this presentation is an application of Latent Dirichlet Allocation model (LDA) on data scraped from Chancellery of the Prime Minister's official website (https://nabory.kprm.gov.pl), containing characteristics concerning each job offer published on mentioned page. The LDA model's purpose is to determine what topics the job offers' descriptions consist of and use that information to approximate demand on the labour market.

All information about valid and archived job offers had been retrieved with the "rvest" package. Next the dataset was tidied, the descriptions' words split up, stop words removed, and finally the words transformed to respective stems using packages "tidyverse", "lubridate", "tidytext", "stringi". The modified data was then used in creating the LDA model with "topicmodels" package, which was able to distinguish different topics brought up in descriptions. The topics generated via the model were then adequately labelled and the demand approximated by summing probabilities of affiliation for every job offer to each topic.

How I became an emoji data scientist (using R!)

Author: Hamdan Azhar (PRISMOJI)

Description

Emojis have been called a "new type of language." According to statistics cited by Ad Week, as much as 92% of the online population uses emojis. Twitter reports that since 2014 alone, over 110 billion emojis have been tweeted.

Yet, despite the profusion of emojis in digital life, little research has been done that leverages emojis to understand popular sentiment. We believe that emoji data science, a largely unexplored field, might be a powerful new methodology for both the computational social sciences as well as fast data journalism.

We'll share preliminary research based on an analysis of millions of tweets that explores the relevance of emoji analytics to fields ranging from pop culture (i.e. the Kanye West vs. Taylor Swift dispute), to politics (the US presidential elections as well as Brexit), to gender norms, to the Olympics, and more.

We'll also introduce concepts including emoji valence, hashtag-emoji co-occurrence, and sentiment analysis that combine the fields of computational linguistics and natural language processing to provide building blocks for understanding what emojis mean and what they reveal about our culture.

The impact of Federal Open Market Comitee minutes on financial markets

Author: Ewelina Osowska (University of Warsaw)

Description

Algorithmic trading strategies are nowadays very commonly developed with the use of machine learning algorithms. One can generate the signal to sell or buy particular assets with the advance of time based on number of arbitrages among which still developing is event arbitrage. It is a high-frequency strategy which trade on the market movements surrounding news announcements. As a foundation for analysis minutes provided by the Federal Open Market Commitee are considered. In order to extract the news sentiment, one can use recurrent neural networks (RNNs) which are a special kind of neural networks heavily applied in natural language processing. Having done that, one can examine the relationship between an asset behaviour and news sentiment and assess whether specific news sentiment is a significant explanatory variable in the forecast development.

The aim of this presentation is to explain the idea of event arbitrage strategy, the process of extracting news sentiment which is then used as an explanatory variable in the modelling process and present a practical workshop based on Federal Open Market Commitee minutes with the comparison of different machine learning algorithms.

Using categorical embeddings (deep learning) in boosting models

Author: Tamas Burghard (ingatlan.com)

Description

This talk intends to highlight the applications / pitfalls of using categorical embeddings (from deep learning) in gradient boosting models, which is a kind of transfer learning. Word vectors are commonplace, so this talk focuses only general categorical variables (with high cardinality). This is a mixture of some theory and practice in R, and my experience of generating such embeddings.

Philosophy

Collective intelligence in GitHub teams

Author: Dorota Celinska-Kopczynska (University of Warsaw)

Description

We present the results of the analysis of developers working in GitHub repository hosting service with a focus on similarity and dissimilarity and their impact on team performance. We measure similarity or dissimilarity with hyperbolic self-organizing maps and then provide text mining and sentiment analysis on the available data on discussions among developers. Our results suggest that dissimilarity plays important role in networks related to information flow (e.g., following, issues networks) while similar developers tend to work on code (pullrequests). We also find diverse roles of heterophylic or homophylic connections <97> homogeneity enhances code development while heterogeneity helps in assessing user interactions (e.g., bug reports, feature suggestions, asking for help).

Hacking R as a script kiddie

Author: Colin Gillespie (Jumping Rivers)

Description

Data science is increasing moving away from a users laptop to the cloud. But how secure is this process? When we consider the value of the data that is often being analysed, security should be at the forefront. In this talk, we won't look at complex hacking but instead, focus on the relatively easy hacks that can be performed to access systems. Instead of talking about potential hacks, we'll discuss actual issues that we've discovered (and since fixed). We'll use three R related examples of how it is possible to access a users system. In the first example, we'll investigate domain squatting on the Bioconductor website. This website deals with genomics research, and so the typical users are Government agencies, Universities and large pharmaceutical companies. By registering only thirteen domains (for a total cost of 130), we had the potential to run arbitrary R code on hundreds of unique users. These users included the majority of the top ten Universities in the world, large government organisations and many companies. With a few additional modifications, we could make this almost impossible to detect. In the second example, we'll look at techniques for guessing passwords on RStudio server instances that are hosted on the internet. Lastly, we'll highlight how users can be a little too trusting when running R code from blogs. The talk will conclude with practical advice on how to avoid these issues both as an individual and as an organisation.

R & MicroService

Author: Colin Fay (ThinkR)

Description

To some extent, the microservice philosophy has been there since the beginning of unix: as the first rule of the unix philosophy states, "write programs that do one thing and do it well".

Almost 40 years after the writing of the UNIX philosophy, this mantra is now more modern than ever, as we have entered in the era of micro service ? instead of writing big monolithic applications, we now divide things into very small services that are orchestrated to work together. Which allow more flexibility, easier maintenance, and even more interoperability than ever.

But how can we apply this to our favorite language, R? How can we build microservices with and for R? Why would we even want to do that? That's the very questions Colin will address in this talk about R & Microservices.

Traits of a world-class data scientist

Author: Olga Mierzwa-Sulima (Senior data scientist at Appsilon)

Description

There is a lot of hype around AI on the current market, which drives the demand for data scientists. However, there are a lot of misconceptions about what the data scientist actually does from day to day.

Q: What does a data scientist do all day? A: S/he builds models!

This is (unfortunately) only partly true.

This talk aims to describe the complete process of a typical data science project from formulating a problem to model deployment and measuring impact. We focus on tools and methods in R that are crucial to becoming a successful data scientist. We stress the importance of problem definition and data preparation, which are often underestimated. We cover the pitfalls to avoid in data science projects using real-life examples. In the case of modeling part of the data science cycle, we focus on objectives such as library choice, interpretability, and business relevant metrics. We show that while Kaggle is a great playground to polish your modeling skills, its competitors<92> objectives differ from those of business projects. Lastly, we comment on what happens when models leave the nest.

Scoring

Experiment management using mlflow and R

Author: Michal Rudko (PwC)

Description

There is nothing so painful as to have a perfect script on perfect data version producing perfect metrics only to discover that you don't remember what are the hyperparameters that were passed as arguments, what was the data set version, etc. Even reproducing the code of an ideal experiment is not that straightforward. Let's admit - it has always been somehow problematic to keep track of all of this - very often it required a custom code or even . . . a spreadsheet was used in order not to get lost. In my presentation I will cover how some of these challenges can be overcome by using an experiment management tool called mlflow. I will also share how we used this package to collect experiment metadata, organize it in a meaningful way and make it available to access and collaborate within our team on one of the advanced analytics projects at PwC Poland.

Forecasting rental prices of flats in Krakow

Author: Wolak Jacek (AGH University of Science and Technology)

Co-authors: Jalocha Mateusz

Description

The aim of the study is to investigate the relationship between the location of the apartment and the rental price and to create - on the basis of geospatial variables and distances of apartments from selected objects - the most accurate possible heat map with forecast rental prices of apartments per m^2. Data for the study was extracted from websites with advertisements using the web-scraping method. Due to a small number of offers in some of the 18 districts of Kraków, only four delegations of Sródmiescie, Krowodrza, Podgórze, Nowa Huta were included. In addition, two methods of model construction will be compared, the first of which will be created on the basis of all data, while the second method of prediction will be done using models created for all delegations on the basis of data occurring only in this area. The study was conducted using the programming language R (in the RStudio environment). In order to forecast the prices of rented flats, methods such as multiple regression, bagging, random forest, GBM and XGBoost were used. On the other hand, when selecting the method with the lowest prediction error, the mean square error (RMSE) and absolute error (MAE) were used. As a result, heat maps with results for each method will be presented.

Predict, vote and elect with R

Author: Karol Klimas (Plenti)

Description

In the past decade analytics fuelled by data and advancement in computing technology has become an integral part of any political campaign process. It has influenced everything from methods of polling to stock market reactions after surprises. Data analysts out-predicted many so-called political experts, who relied mostly on gut instinct and experience. However, year 2016 brought not one but two major upsets and such still keep happening. In my presentation I will cover not only predicting but also understanding election results and, more important, the entire election process. As an example I will show how to better gasp the D'Hondt method (also called the Jefferson method), which will allocate seats in the upcoming Polish parliamentary election.

Shiny

A Shiny Real-time Application for Backtesting Investment Strategies on Regulated and Crypto Markets

Author: Pawel Sakowski (QFRG WNE UW)

Co-authors: Przemyslaw Rys

Description

Our study presents results of application of widely known asset allocation model proposed by Markowitz (1952) on regular and crypto markets. Our strategies are constructed with most important world equity indices and largest cryptocurrencies (in terms of market capitalization) on daily data for the last five years.

Results are presented in form of interactive Shiny application. The user can easily get equity lines of analysed strategies and benchmark portfolios, as well as their performance and risk measures for selected strategy parameters. Our solution also reports historical portfolio composition at every point of time together with weights obtained in optimization process and presents dynamics of historical correlation between assets from regular and crypto markets.

Our application also allows to perform a sensitivity analysis with respect to length of historical window, frequency of portfolio rebalancing and degree of financial leverage. This gives a chance for the user to easily manipulate those parameters and to observe how they affect the strategy results.

The results presented in our application illustrate potential of risk diversification offered by new class of investable assets offered by crypto markets. Application is deployed in the cloud and the whole process of updating the data and portfolio rebalancing is performed automatically.

Challenges of Shiny application development at scale

Atuhors: Jakub Małecki, Jakub Stepniak (Analyx)

Description

At Analyx, we build software product called SpendWorx. It is Shiny+R application to support marketing decisions. The software uses multiple mathematical models to predict sales and to optimize marketing spending. We present SpendWorx and we discuss selected technical, organizational and analytical challenges we have encountered during the development. We also present our best practises in development and what we learned along the way.

Improving the communication of environmental data using Shiny

Author: Theo Roe (Jumping Rivers)

Description

The environment agency has a wealth of data relating to the health of rivers in the UK. The problem being that it's difficult to explore the data and thus hard to build any tools using the data. We've built a Shiny app that allows for visual exploration of the river health in the UK.

Shiny application for algorithmic trading

Authors: Tomasz Koc, Piotr Wójcik (University of Warsaw)

Description

The aim of this presentations is to show Shiny application that can be applied to backtest selected algorithmic strategies applied on historical quotations of S&P 500 stocks. Uploading own data with is also possible. The application allows to select assets that will be considered in the strategy, define different entry and exit techniques based on technical analysis indicators (including two moving averages/medians crossover, volatility breakout and double volatility breakout). Trading can be applied on single assets or pairs. Different pair selection methods are available ? based on correlation, regression and cointegration. The strategy performance is assessed with several commonly used measures (incl. aggregated PnL, Information Ratio, maximum drawdown) in gross and net terms (after taking into account transactional

costs). One can also perform strategy parameters optimization based on the selected criterion and verify its performance on the test data.

Vision 1

A Case Study for Image Classification using Transfer Learning

Author: Olgun AYDIN (Senior Data Scientist / PwC)

Description

Deep Learning (DL) is rising star of Machine Learning (ML) and Artificial Intelligence (AI) domains and it has been proven that deep neural networks(DNN) are one of the most crucial inventions for the 21th century. Nowadays, DNNs are being used as a key technology for many different domains: self-driven vehicles, smart cities, security, automated machines. For the purpose of this use case, DNN has been trained by using images of products on the store shelves as input, product categories (cereals, milk, soda, etc.) as label . Transfer Learning(TL) has been used during to train such a deep neural network predicts product category regarding given image of product. For the training of the DNN, open source Freiburg Grocery Dataset has been used. The VGG16 network, developed by Oxford University researchers, has been used to perform TL. Due to the nature of TL, it is necessary to freeze some layers and retrain them with a new structure. For this purpose, only the final layer has been frozen first, then the last five layers have been frozen. These networks have been trained using different combinations of epoch and batch sizes. After comparing the performance of those networks, best performed model has been used for creating user interface. Shiny application has been created using to provide user interface to end users. This Shiny application basically calls the trained model and predicts product class for the image uploaded by the user.

Facial landmarking made (possible and) easy with R!

Author: Lubomir Stepanek (Institute of Biophysics and Informatics, First Faculty of Medicine, Charles University)

Description

R language is definitely a powerful tool well suitable for most analytical tasks arising from an everyday routine of a scientist or data analyst. However, there are fields of data processing which R is not so much native for and other languages or tools such as Python or Octave are preferred to handle them. Some of these fields are image processing and computer vision and particularly facial computer where there are barely original R packages currently available for.

There are of course approaches how to handle missing libraries in R. In general, an API library could bridge the gap between R on the one hand and the special tool (usually built in a different programming language) on the other hand. In this work, we tried to rethink the problem and go even further such that we have developed a web-based shiny application providing facial image processing and especially facial landmarking. The facial landmarking is powered using C++ library dlib dedicated to machine-learning based computer vision algorithms, but only for C++ speaking users.

The connection between R, C++ dlib and R package shiny could open R functionality and computing power to a wider audience using comfortable and clickable way.

Semantic segmentation using U-Net with R

Author: Michal Maj (Data Science Solutions Michal Maj)

Description

In this presentation I will discuss semantic segmentation, which is one of the most interesting computer vision tasks, used e.g in self-driving cars or medical image analysis. I will present U-Net deep neural network model and show how to build it in R using Keras.

Vision 2

DeepSport: A Shiny app for sports video analysis

Author: Pablo Maldonado (Data Start)

Description

In this talk we present a Shiny app that helps coaches prepare their teams better with automatic video tagging and insightful visualizations.

Detection of solar panels based on aerial images using deep learning

Author: Michel Voss (UEP)

Co-authors: Maciej Beresewicz

Description

The main goal of the article is to present the results of a study on the use of deep learning networks to detect solar panels based on aerial images of Poznan. In addition, the main motivation is to obtain more detailed information about the use of solar energy in Poland drawing on big data sources, which until now have not been used for this purpose.

The data was acquired from the Management Board of Geodesy and Municipal Cadastre GEOPOZ in Poznan and included orthophotomaps for 2016 and the layer of buildings and plots of lands. We extracted buildings from the images using R statistical software and the sf package. To detect solar panels we used the Turi Create library written in Python which re-implements the YOLO (You Only Look Once) library.

The object recognition algorithm was trained on a sample of images that included annotations (bounding boxes) about the exact location of solar panels. The results indicate a very high recognition efficiency at the level of 96-99% on the test sample. Based on this procedure we found that around 2% of residential buildings in Poznan in 2016 had solar panels mounted on roofs.

As far as we know, this is the first use of deep learning to detect solar panels in Poland. Currently, similar studies are being carried out by for instance Statistics Netherlands as part of the DeepSolaris project. The study exemplifies a trend involving the use of aerial and satellite images for statistical purposes thanks to advanced machine learning algorithms and open source software.

XAI

Compare predictive models created in different languages with DALEX and friends

Author: Szymon Maksymiuk (MI2 DataLab / MiNI PW)

Description

Friend of yours dared to claim that his model is better only because was built in Python or Julia? Time to help you in your rightful battle against anyone who opposes superiority of your model!

First part of my presentation is dedicated for participants who would like to explain and compare models between R and Python. I will present functionalities I have developed that are now available in DALEXtra package. They let us to build explainers of Python models in a unified way.

During the second part, I am going to familiarize participants with tools and ideas helpful in models comparison, such as residuals distribution comparison. Nowadays, we have plenty of different models that are easy to build and have similar performance, that is why we would like to have more criterions than pure MSE or AUC. Presentation will cover different approaches to that topic. The goal is to show information we can extract from our model's output and the way we can present it. Some of these ideas are implemented as fucntion in DALEXtra package that creates report with results.

modelling ROC curves, simulating profit increase in risk-based pricing regime and combining correlated scorecards

Author: Blazej Kochanski (Politechnika Gdanska)

Description

From time to time a business need appears to have a simple model to understand what is the impact of better machine learning tools in bank lending. As I have been working for some time on this, I would like to share a few effects in this area: modelling ROC curves, simulating profit increase in risk-based pricing regime and combining correlated scorecards. All with R.

survxai: how to explain predictions for survival models?

Author: Aleksandra Grudziaz (MIM UW)

Description

Advanced machine learning models are used in more and more areas. Also in survival modeling. Survival random forest and other complex black-boxes supplant classical Cox Proportional-Hazard model. Although black-boxes are very effective, it is difficult to understand which factors drive their decisions. During this talk, I will introduce the survxai package which contains methods for the explanation and exploration of survival models. This package allows to analyze the models at two levels: global concerning, among others, the behavior of the model when changing the value of the variable and local where we focus on the prediction for a specific observation. Explanations can be visualized by using survival curves. In my talk, I will present the functionality of the package on the example of several models of survival analysis: Cox Proportional-Hazards Model, Random Forest for Survival and Regression for a Parametric Survival Model.

References:

- Aleksandra Grudziaz, Alicja Gosiewska, and Przemyslaw Biecek (2018). survxai: Visualization of the Local and Global Survival Model Explanations. R package version 0.2.0 [https://mi2datalab.github.io/survxai/]
- Aleksandra Grudziaz, Alicja Gosiewska, and Przemyslaw Biecek (2018). survxai: an R package for structure-agnostic explanations of survival models. Journal of Open Source Software, 3(31), 961, https://doi.org/10.21105/joss.00961 Przemyslaw Biecek (2018). DALEX: Descriptive mAchine Learning EXplanations. R package version 0.4 [https://pbiecek.github.io/DALEX/]

3. List of speakers

Ahmet, Çizmeli Servet

Andruszek, Krystian

AYDIN, Olgun

Azhar, Hamdan

Baniecki, Hubert

Bień, Adam

Bourgeois, Florent

Bras, Anne

Brito, Paula

Bucur, Ioan Gabriel

Burghard, Tamas

Celinska-Kopczynska, Dorota

Chilimoniuk, Jaroslaw

Ciurus, Piotrek

Drhlik, Patrik

Fay, Colin

Gillespie, Colin

Grudziaz, Aleksandra

Jacek, Wolak

Jacquet, Francois

Jagoda, Glowacka

Jakuczun, Wit

Jancewicz, Barbara

Jessen, Leon Eyrich

Kaledkowski, Dawid

Kaminska, Olga

Kania, Krzysztof

Keydana, Sigrid

Klimas, Karol

Kobylka, Mateusz

Kochanski, Blazej

Kolakowska, Lidia

Kolasa, Bartosz

Kozak, Anna

Locke, Steph

Louden, Richard

Maj, Michal

Maksymiuk, Szymon

Maldonado, Pablo

Mierzwa-Sulima, Olga

Mikos, Maria

Nowosad, Jakub

Nowosielski, Piotr

Osowska, Ewelina

Otreba-Szklarczyk, Agnieszka

Podsiadło, Michał

Puchala, Weronika

Rafacz, Dominik

Rodziewicz, Ken Benoit, Damian

Roe, Theo

Rudko, Michal

Sakowski, Pawel

Sidorczuk, Katarzyna

Sieminski, Leszek

Sijko, Kamil

Staniak, Mateusz

Stepanek, Lubomir

Voss, Michel

Weiner, Jakub

Wozniak, Rafal

Wright, Marvin N.

Żółtak, Tomasz