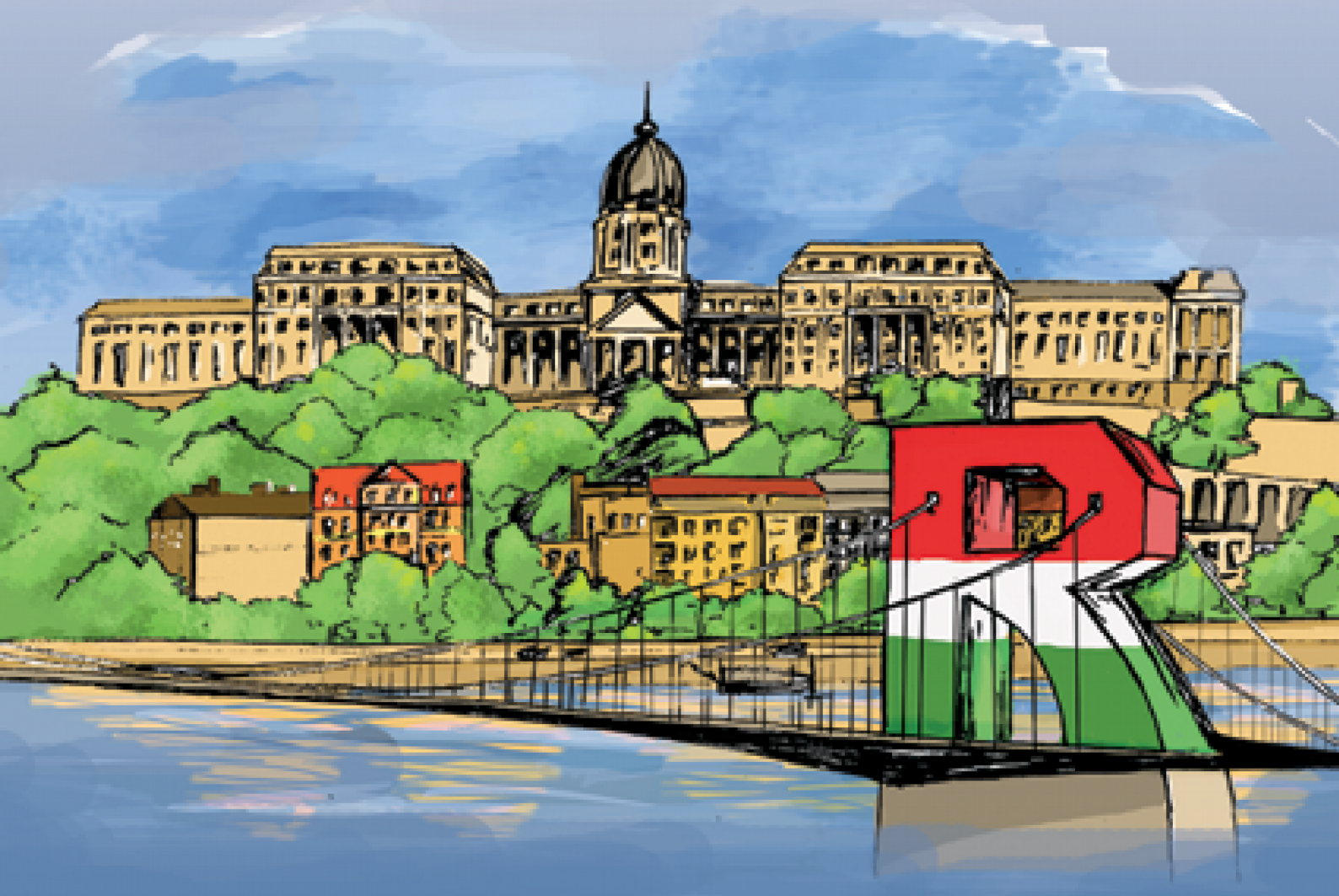
European R Users Meeting

eRum 2018

Budapest, May 14-16



https://2018.erum.io

Platinum Sponsor

RStudio®

Gold Sponsors

MANGO SOLUTIONS

QUANTIDE
knowledge from data

Silver Sponsors

H2O.ai

emarsys

Open Analytics

BUDAPEST BI FORUM

Bronze Sponsors

WLOG SOLUTIONS

Jumping Rivers

R consortium

R-bloggers

Shift

R

Table of Contents

Welcome to eRum 2018!

Dear participant,

I'm very delighted to welcome you to the European R Users Meeting 2018 conference – taking place at multiple venues in Budapest, Hungary between May 14-16, 2018; where we expect around 500 R enthusiast attendees from all around Europe and some from overseas as well!

The eRum 2018 conference is building on the traditions of eRum 2016 (Poznan, Poland) and the satRday 2016 event (Budapest, Hungary) with the promise of an affordable, yet high-quality, nonprofit conference focusing on R and providing a networking event for the European R community – in years when the useR! conference is hosted outside of Europe.

I'm extremely proud that this year Budapest and the Hungarian R community can host this event, and it was a great pleasure to see all the supportive feedback on the early works of this conference, the positive replies from the potential keynote and invited speakers, then the huge interest in our call for papers, sponsorship opportunities and registration as well. Your kind support and contributions to the success of the conference are highly appreciated!

Although the next three days will be extremely busy – offering a broad variety of programs (over one hundred workshops / half-day tutorials, keynotes, invited and contributed talks, Shiny demo sessions and posters) starting early in the morning and lasting until late afternoon mostly focusing on scientific topics, I'd like to urge you to also take part in the social and networking events at nights – the most important aspects of conferences in my opinion, which is really difficult to do remotely, and probably that's why we all still prefer to attend conferences in person instead of watching talks remotely (but for those who cannot make it, we provide live-streaming).

So keeping in mind the busy schedule, I keep it brief with the housekeeping tasks – please

- **Don't forget to bring your badge** to all conference/networking events, wear those where it can be easily seen – otherwise you will not be able to enter the venues. No exceptions.
- Comply with the eRum 2018 **code of conduct** and related (eg harassment) policies.
- Follow the **most recent news** at *https://2018.erum.io* and *@erum2018* on Twitter.
- Please use the **#eRum2018** hashtag when tweeting about the conference.
- Feel free to use the public **WiFi** connections (CEU Guest & AkvariumKlub), no passwords.
- In case of any questions, problems or concerns, ask for help at the registration desk.

I hope to meet you all in person, and I wish you have a very educational and fun time here!

Best,
Gergely Daróczi – on behalf of the Organizing Committee

Organizing and Program Committee

The Organizing Committee of this nonprofit conference is chaired by Gergely Daróczi, who volunteered to kick-off eRum 2018 with the help of the local R community and R enthusiast sponsors, just like at the first satRday event in Budapest 1.5 years ago. The administrative, financial and legal background is provided by Upshift R Kft, a Hungarian limited company.

Besides Gergely, Mariachiara Fortuna (Quantide) and Henriett Daróczi (Upshift R) took care most of the organizational tasks, and Klaudia Korniluk worked on all the design materials. Balázs Horváth (Glowing Bulbs) contributed the video and photography tech background for the event, and the VIDra team (Combit) delivered the live-streaming and video recording infrastructure. The Central European University helped with providing the workshop rooms at a discounted price.

We are also very grateful to the eRum 2016 Organizing Committee (especially Adolfo Álvarez, Maciej Beręsewicz, Marcin Kosiński and Przemyslaw Biecek) not only for coming up with the idea of this conference series, but also for their active help regarding contacts, ideas, feedback etc.

The list of invited speakers was decided by and the more than 150 abstract submissions for the Call for Papers were reviewed by the Program Committee members, who also helped finalizing the detailed conference program:

- Adolfo Alvarez (Poland)
- Ágnes Salánki (Hungary)
- Andrew Lowe (Hungary)
- Bence Arató (Hungary)
- Branko Kovač (Serbia)
- Eszter Windhager-Pokol (Hungary)
- Gergely Daróczi (Hungary)
- Heather Turner (United Kingdom)
- Kevin O'Brien (Ireland)
- Imre Kocsis (Hungary)
- László Gönczy (Hungary)
- Maciej Beresewicz (Poland)
- Mariachiara Fortuna (Italy)
- Przemyslaw Biecek (Poland)
- Szilárd Pafka (United States of America)

Contact information:

- organizers@erum.io
- program_committee@erum.io
- speakers@erum.io
- invoicing@erum.io

Venues

The workshop day takes place at the Central European University (Budapest, Nádor utca 15), shown as the graduation cap / mortarboard icon in the below map, and the Welcome Reception, then the two conference days will happen in the Akvárium Klub (Budapest, Erzsébet tér 12), shown as the microphone icon below:

The location of the Tuesday and Wednesday social events are also highlighted on the map: the boat icon stands for the starting point of the Conference Dinner (right on the riverside, at Budapest, Petőfi square; Port 9 – Zsófiahajó), and the pizza icon shows the Budapest R-Ladies Meetup location (Budapest, Kossuth Lajos street 7-9).

All these venues are very centrally located in Budapest and are definitely within walking distance to the main conference venue.

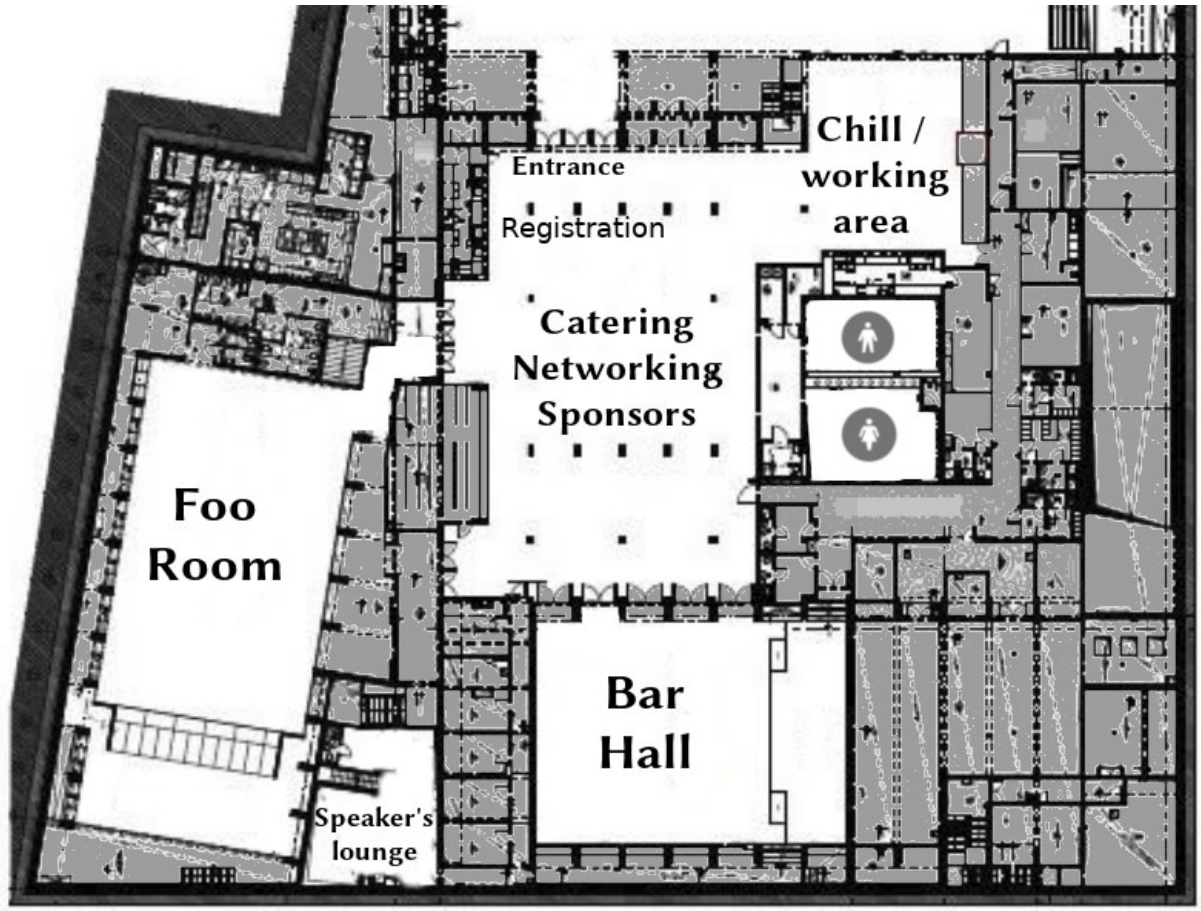
Central European University

Please note that the Central European University (CEU) has multiple buildings in the area, so make sure to enter the main building at Nádor street 15. You will find the registration and further information on the actual location of the workshops a couple meters after the entrance.

The two main auditoriums (Aud A and Aud B) can be found behind the registration desk. Coffee and lunch will be served on the same floor, but in the N13 building that you can access right from the registration desk via the turnstile system – that will be disabled for the coffee and lunch breaks. Rooms 101, 103 and 106 can be found on the first floor; and rooms 202 and 203 on the second floor – that you can access via the stairs or elevator right behind the registration desk.

Akvárium Klub

The Welcome Reception and the two conference days will take place at the Akvárium Klub – a cultural center with a great atmosphere, featuring a networking area under an artificial pool (including the "Registration" desk right at the "Entrance" as shown on the below map), two big auditoriums ("Foo Room" and "Bar Hall"), several "Speaker's lounges" and a separate zone for relaxing and working. The outside entrance can be found downstairs at Erzsébet Square.



Other Locations

Please see the venues of the Conference Dinner and R-Ladies Meetup on page 12 and 15.

Detailed Schedule

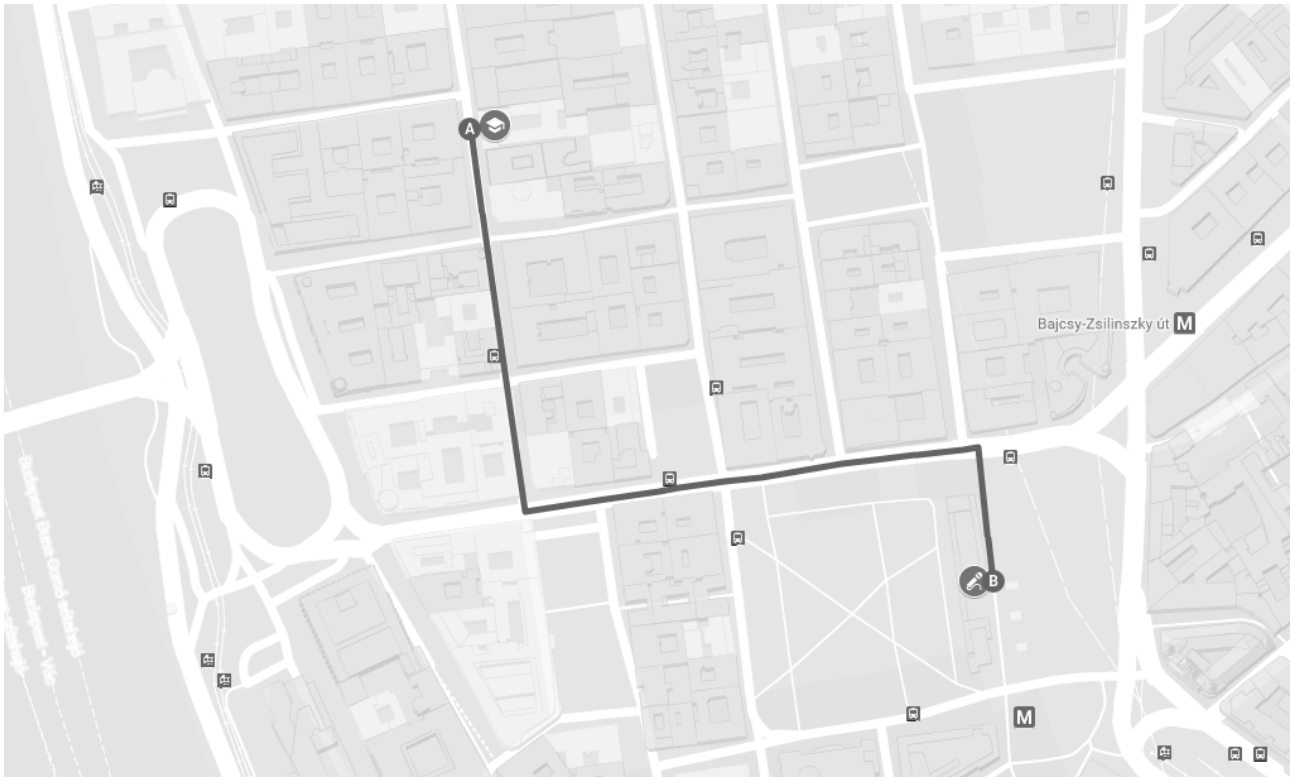
Workshop Day (May 14, 2018 – Monday) 8:00 – 17:00

All the workshops are hosted at the Central European University's new building at Budapest, Nádor street 15. The abstracts can be found on the page number as shown between the square brackets after the talk title.

	N15 Aud A 150 seats	**N15 Aud B** 150 seats	**N15 103** 75 seats	**N15 101** 50 seats	**N15 106** 40 seats	**N15 203** 25 seats	**N15 202** 20 seats
8:00	**Registration (N15 entrance), Coffee (N13)**						
8:30							
9:00	Efficient R programming [16]	DALEX: Descriptive mAchine Learning Explanations [17]	Clean R code - how to write it and what will the benefits be [18]	Building an Interpretable NLP model to classify tweets [19]	Geocomputation with R [19]	Graphs: A datastructure to query [20]	Forwards Package Development Workshop for Women [21]
9:30							
10:00							
10:30	**Coffee break (N13)**						
11:00	Efficient R programming [16]	DALEX: Descriptive mAchine Learning Explanations [17]	Clean R code - how to write it and what will the benefits be [18]	Building an Interpretable NLP model to classify tweets [19]	Geocomputation with R [19]	Graphs: A datastructure to query [20]	Forwards Package Development Workshop for Women [21]
11:30							
12:00							
12:30	**Lunch break (N13)**						
13:00							
13:30	Deep Learning with Keras for R [22]	Automatic and Interpretable Machine Learning in R with H2O and LIME [22]	The beauty of data manipulation with data.table [23]	Building a package that lasts [24]	Building a pipeline for reproducible data screening and quality control [24]	Plotting spatial data in R [25]	Forwards Package Development Workshop for Women [21]
14:00							
14:30							
15:00	**Coffee break (N13)**						
15:30	Deep Learning with Keras for R [22]	Automatic and Interpretable Machine Learning in R with H2O and LIME [22]	The beauty of data manipulation with data.table [23]	Building a package that lasts [24]	Building a pipeline for reproducible data screening and quality control [24]	Plotting spatial data in R [25]	Forwards Package Development Workshop for Women [21]
16:00							
16:30							
18:00	**Welcome Reception** (see next page)						

Welcome Reception (May 14, 2018 – Monday) 18:00 – 22:00

The Welcome Reception takes place in the Akvárium Klub after the workshops finished at the Central European University – so you have to walk ~500 meters (~5 minutes) between the two locations:



Relax, you will have plenty of time, as the workshops end at 17:00 and the registration for the Welcome Reception starts at 18:00 and you are not too late until 19:00 (although you might miss a free drink or two).

This event is free and open to all conference attendees without any formal dress code. Light dinner and drinks (soft drinks, beer, wine) will be provided for free, a' la carte options available on demand.

Schedule:

- 18:00 – 19:00 Registration with a welcome drink and sandwiches
- 19:00 – 19:45 Shiny demos presented in the main auditorium
- 20:00 – 22:00 Poster session and networking in the smaller auditorium and café

For the full list of Shiny demos and posters to be presented at the event, please see the Abstracts section starting on page 66 and 71.

Conference Day 1 (May 15, 2018 – Tuesday) 8:00 – 17:00

		Foo Room	Bar Hall	
8:00		**Registration**		
8:10				
8:20				
8:30				
8:40				
8:50		Conference opening		
9:00	Gergely Daróczi	Martin Mächler [26]		
9:10				
9:20				
9:30				
9:40				
9:50		Jeroen Ooms: Using Rust code in R packages [28]	Edwin Thoen: A recipe for recipes [33]	Branko Kovač
10:00				
10:10		Lionel Henry: Harness the R condition system [34]	Ildiko Czeller: The essentials to work with object-oriented systems in R [34]	
10:20				
10:30				
10:40		**Coffee break**		
10:50				
11:00	Przemyslaw Biecek	Stefano M. Iacus: Sentiment Analysis on Social Media and Big Data [26]		
11:10				
11:20				
11:30				
11:40				
11:50		Olga Mierzwa-Sulima: Taking inspirations from proven frontend frameworks to add to Shiny with 4 new packages [28]	Marcin Kosiński: Multi-state churn analysis with a subscription product [34]	Henrik Bengtsson
12:00				
12:10		Mikołaj Olszewski: Not all that Shiny by default [35]	• Bence Arató: The Big Connection - using R with big data • Florian Privé: … statistical tools with big matrices … on disk • Matthias Kaeding: RcppGreedySetCover … • Emil Lykke Jensen: Make R elastic [52– 53]	
12:20				

		Foo Room	Bar Hall	
12:30				
12:40				
12:50				
13:00		**Lunch break**		
13:10				
13:20				
13:30	Bence Arató	Nathalie Villa-Vialaneix: Learning from (dis)similarity data [26]		Kevin O'Brien
13:40				
13:50				
14:00				
14:10				
14:20		Erin LeDell: Scalable Automatic Machine Learning in R [29]	Sander Devriendt: Sparsity with multi-type Lasso regularized GLMs [36]	
14:30				
14:40		Szilard Pafka: Better than Deep Learning - Gradient Boosting Machines (GBM) in R [29]	Francois Mercier: Nonlinear mixed-effects models in R [36]	
14:50				
15:00		Andrie de Vries: Tools for using TensorFlow with R [37]	Stanislaus Stadlmann: bamlss.vis - an R package for interactively visualising distributional regression models [38]	
15:10				
15:20				
15:30		**Coffee break**		
15:40				
15:50	David Smith	Matthias Templ: Compositional analysis of our favourite drinks [30]	Tom Reynkens: Estimating the maximum possible earthquake magnitude using extreme value methodology: the Groningen case [39]	Kevin O'Brien
16:00				
16:10		Przemyslaw Biecek: Show me your model 2.0 [30]	Andrew Collier: Taking the Bayesian Leap [39]	
16:20				
16:30		Heather Turner: Modelling Item Worth Based on Rankings [40]	• Timothy Wong: Generalised Additive Model for Field Operation Demand Modelling [54–56] • Krzysztof Jędrzejewski: IRT and beyond ... • Lubomír Štěpánek: Classification ...of facial emotions ... • Johannes Gussenbauer: The R-Package 'surveysd'	
16:40				
16:50		Federico Marini: Interactivity meets Reproducibility: the ideal way of doing RNA-seq analysis [41]	• Samuel Borms: ... R for textual sentiment time series ... • Peter Laurinec: Time Series Representations ... • Ekaterina Fedotova: ... efficient processing of spatial data ... • Jakub Houdek: How to tell if a hockey player performs well • Chris von Csefalvay: Soylent Green is populations [57–59]	
17:00				
19:00		**Conference Dinner** (see next page)		

Conference Dinner (May 15, 2018 – Tuesday) 19:00 – 22:00

The conference dinner requires a separate ticket purchase – that is shown on the badge with a boat icon. As the available seats are extremely limited, please don't attempt to attend this event if your badge does not have a boat icon. Make sure to bring your badge with you for the dinner!

As you may have guessed, the dinner takes place on a sightseeing boat that will do a 2-hours cruise on the Danube, so you have to walk from the conference venue around 1,000 meters (~10 minutes) to the riverside as shown on the below map to Budapest, Petőfi tér (Port 9 – Zsófiahajó):

Conference Dinner schedule:

- 19:00 – Onboarding, welcome drink on the terrace of the boat
- 19:15 – The Capatain's very brief welcome speech
- 19:20 – Departure (heading North to Margit Bridge, then South until Petőfi Bridge)
- 19:30 – Dinner served
- 21:30 – Return to the port

The soup, main dish, salad and dessert options will be displayed in a buffet format; and the complimentary soft drinks, beer, wine and a welcome champagne will be served to the tables. There's no official dress code, and we don't expect extreme weather conditions.

Conference Day 2 (May 16, 2018 – Wednesday) 8:00 – 17:00

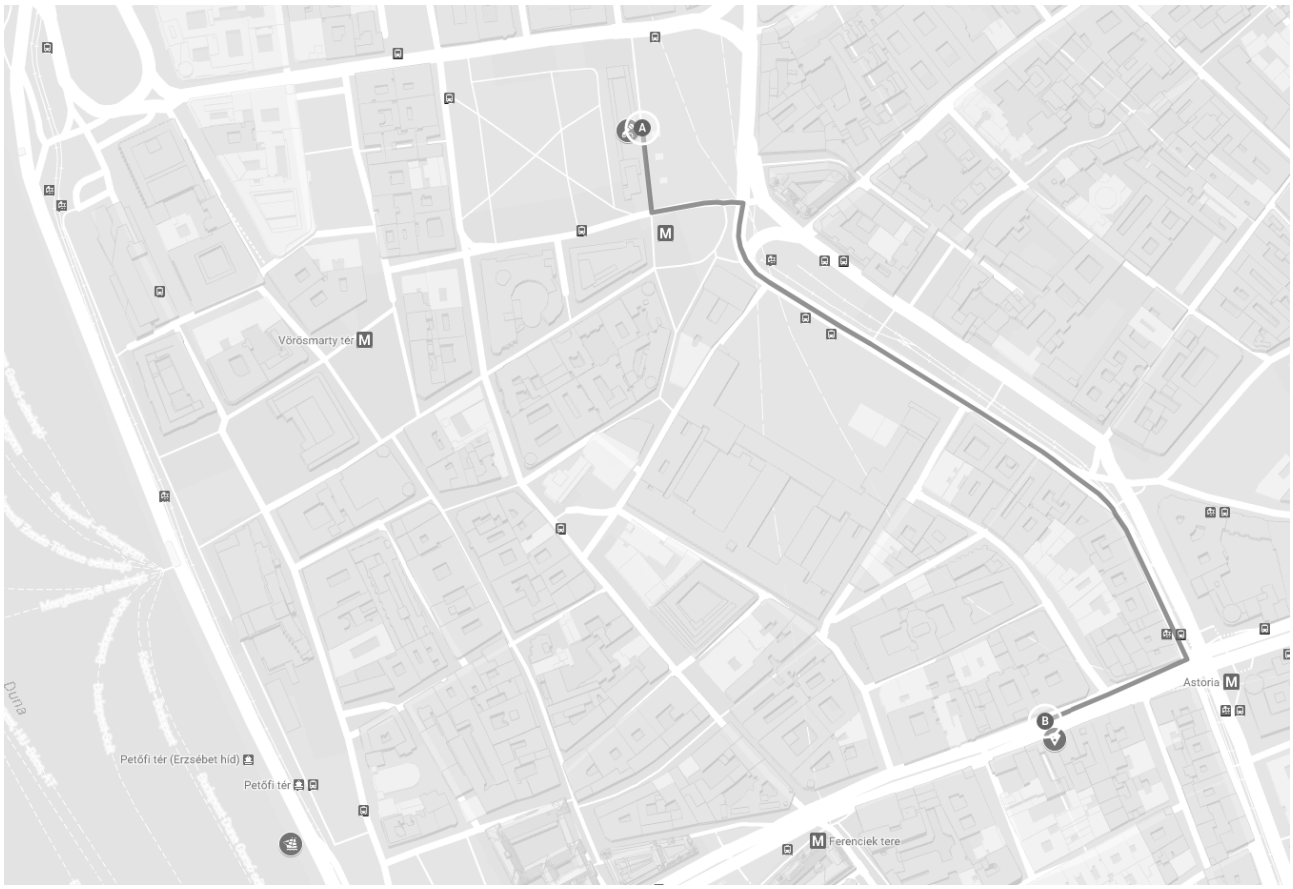
		Foo Room	Bar Hall	
8:30				
8:40		**Registration**		
8:50				
9:00	Heather Turner	Roger Bivand: A practical history of R (where things came from) [27]		
9:10				
9:20				
9:30				
9:40				
9:50		Henrik Bengtsson: A Future for R: Parallel and Distributed Processing in R for Everyone [31]	Noa Tamir: Data Culture in Practice [42]	Andrew Collier
10:00				
10:10		Dénes Tóth: radii.defer - Deferred execution of nested functions [42]	Aimee Gott: Using R to Build a Data Science Team [43]	
10:20				
10:30				
10:40		**Coffee break**		
10:50				
11:00	Eszter Windhager-Pokol	Barbara Borges: Drilldown data discovery with Shiny [31]	Leopoldo Catania: Predicting Cryptocurrencies Time–Series with the eDMA package [43]	Andrew Collier
11:10				
11:20		Colin Gillespie: Getting the most out of GitHub and friends [32]	eDavid Ardia: Markov-Switching GARCH Models in R: The MSGARCH Package [44]	
11:30				
11:40		David Smith: Speeding up R with Parallel Programming in the Cloud [44]	Andreas Scharmüller: Time series modeling of plant protection products in aquatic systems in R [45]	
11:50				
12:00		Simon Field: Exploiting Spark for high-performance scalable data engineering and data-science on Microsoft Azure [45]	Claus Thorn Ekstrøm: Predicting the winner of the 2018 FIFA World Cup predictions [46]	
12:10				
12:20		Goran Milovanović: Wikidata Concepts Monitor: R in action across Big Wikidata [46]	• Hannah Frick: ... the Wealth of R Packages [60–61] • Mikkel Freltoft Krogsholm: Write Rmazing Code! • Tamas Szilagyi: Robust Data Pipelines ... • Alicja Fraś: Nested apply as an alternative ...	
12:30				

		Foo Room	Bar Hall	
12:40				
12:50				
13:00		**Lunch break**		
13:10				
13:20				
13:30				
13:40	Roger Bivand	Achim Zeileis: R/exams -- A One-for-All Exams Generator [31]		
13:50				
14:00				
14:10				
14:20				
14:30		Mark van der Loo: Tracking changes in data with the lumberjack package [32]	• Mikolaj Olszewski: What teaching R taught me ... • Tatjana Kecojevic: ... R workshop in the cloud • Titus Laska: Quality Assurance in Healthcare ... • Mira Céline Klein: Writing R packages for clients .. • Luke Johnston: ... open scientific workflow [61–65] • Tamás Nagy: Manage your meta-analysis workflow • Andrea Schnell: Establishing analytical pipelines ..	Ágnes Salánki
14:40				
14:50		Edwin de Jonge: validatetools - resolve and simplify contradictive or redundant data validation rules [47]		
15:00				
15:10				
15:20		**Coffee break**		
15:30				
15:40	Roger Bivand	Arthur Charpentier: Demographics with Genealogical Data [33]	Andrea Melloncelli: What software engineers can teach to data scientists: code safety with automatic tests [48]	Ágnes Salánki
15:50				
16:00		Robin Lovelace: Geocomputation for Active transport planning: a case study of cycle network design [49]	Wit Jakuczun: Know your R usage workflow to handle reproducibility challenges [49]	
16:10				
16:20		Mira Kattwinkel: openSTARS - prepare GIS data for regression analysis on stream networks [50]	Omayma Said: Fitting Humans Stories in List Columns: Cases From an Online Recruitment Platform [50]	
16:30				
16:40		Tomislav Hengl: Machine Learning (ranger package) as a framework for spatial and spatiotemporal prediction [51]	Zuzana Hubnerova: Asymptotic Powers of Selected ANOVA Tests in Generalized Linear Models [52]	
16:50				
17:00		**Closing remarks**		
17:45		**R Ladies Meetup** (see next page)		
18:30				
20:30				

R Ladies Meetup (May 16, 2018 – Wednesday) 17:45 – 21:00

We're delighted to invite everyone (independently whether being identified as a woman), who accepts the code of conduct, to attend the R-Ladies Budapest Meetup, where all R-Ladies and allies will have a chance to meet the fantastic members of our local R community. The meetup is free but requires registration at https://www.meetup.com/R-Ladies-Budapest/events/250031700

The event is sponsored by Emarsys, whose office is within walking distance of the conference venue: Budapest, Kossuth Lajos u. 7-9, 1053



Schedule:

- 5:45 PM – 6:30 PM: Registration (pizza and drinks will be served)
- 6:30 PM – 8:30 PM: Presentations
- 8:30 PM: Networking

Speakers:

- Ellen Talbot (R-Ladies Manchester & R-Ladies Liverpool)
- Erin LeDell (R-Ladies San Francisco, R-Ladies Global)
- Heather Turner (R Forwards)
- Isabella Gollinni (R Forwards)
- Olga Mierzwa-Sulima (R-Ladies Warsaw)

Abstracts and Speaker Biographies

The list of abstracts below follows the order of the talks as listed in the detailed schedule.

Workshops

A hands-on tutorial for 20-100 persons on a beginner or advanced R topic for 180 mins with a 30 mins coffee break between the two 90 mins long sessions in a classroom environment, where attendees work on their own laptop.

Efficient R programming [Mon 9:00]

Speaker: Colin Gillespie (Data Scientist/Senior Lecturer @ Jumping Rivers/Newcastle University)

Colin Gillespie is Senior lecturer (Associate professor) at Newcastle University, UK. He has been running R courses for over eight years at a variety of levels, ranging from beginners to advanced programming. He is co-author of the recent book Efficient R programming, O'Reilly.

Section: HPC, Big Data

This tutorial will cover a variety of techniques that will increase the productivity of anyone using R. Topics include optimizing your set-up, tips for increasing code performance and ways to avoid memory issues. Ranging from guidance on the use of RStudio to ensure an efficient workflow to leveraging C++, this tutorial provides practical advice suitable for people from a wide range backgrounds.

An overview of the topics covered are:

- Efficient set-up: the .Rprofile and .Renviron files, the importance of a good IDE, and switching BLAS libraries.
- Efficient hardware: assessing your computer hardware with the benchmarkme package.
- Efficient collaboration: coding guidelines and the importance of version control.
- Efficient programming: common R data types, good programming techniques, parallel computing and the byte compiler.
- Efficient learning: practical suggestions for improving your general R knowledge.
- Efficient C++ programming: A brief introduction to Rcpp.

Pre-requisites

Participants should be familiar with for loops, if statements and writing simple functions.

Justification

R is now used in many disparate settings. However, the majority of R programmers have no formal computer training, and instead have "learned on the job". This tutorial aims to fill in some of these gaps.

Potential attendees

This tutorial will be of interest to most conference attendees and so I would expect it to be reasonably popular. I gave a similar tutorial at useR!2017 and it was the most popular tutorial by around 100 participants.

DALEX: Descriptive mAchine Learning EXplanations. Tools for exploration, validation and explanation of complex machine learning models [Mon 9:00]

Speaker: Przemyslaw Biecek (Associate Professor @ Warsaw University of Technology, Poland) and Mateusz Staniak

Data Scientist with background in both mathematical statistics and software engineering. Research activities are mainly focused on high-throughput genetic profiling in oncology. Also interested in evidence based education, evidence based medicine, general machine learning modeling and statistical software engineering. An R enthusiast: three books, dozen packages, lots of talks, classes and workshops.

Section: Machine Learning, Graphics

Complex machine learning models are frequently used in predictive modelling. There are a lot of examples for random forest like or boosting like models in medicine, finance, agriculture etc. In this workshop we will show why and how one would analyse the structure of the black-box model.

This will be a hands-on workshop with four parts. In each part there will be a short lecture (around 20 minutes) and then time for practice and discussion (around 20 min).

Introduction

Here we will show what problems may arise from blind application of black-box models. Also we will show situations in which the understanding of a model structure leads to model improvements, model stability and larger trust in the model.

During the hands-on part we will fit few complex models (like xgboost, randomForest) with the mlr package and discuss basic diagnostic tools for these models.

Conditional Explainers

In this part we will introduce techniques for understanding of marginal/conditional response of a model given a one- two- variables. We will cover PDP (Partial Dependence Plots) and ICE (Individual Conditional Expectations) packages for continuous variables and MPP (Merging Path Plot from factorMerger package) for categorical variables.

Local Explainers

In this part we will introduce techniques that explain key factors that drive single model predictions. This covers Break Down plots for linear models (lm / glm) and tree-based models (randomForestExplainer, xgboostExplainer) along with model agnostic approaches implemented in the live package (an extension of the LIME method).

Global Explainers

In this part we will introduce tools for global analysis of the black-box model, like variable importance plots, interaction importance plots and tools for model diagnostic.

R packages:

- mlr (Bernd Bischl and others)
- live (Staniak Mateusz, and Przemysław Biecek)
- FactorMerger(Sitko Agnieszka, and Przemyslaw Biecek)
- pdp (Greenwell, Brandon)
- ALEPlot (Apley, Dan)

Clean R code - how to write it and what will the benefits be [Mon 9:00]

Speaker: Ildiko Czeller (Data Scientist @ Emarsys Technologies Kft, Hungary) and Jenő Pál

Ildi Czeller is a mathematician who has worked as a data scientist at Emarsys in Budapest for almost 3 years now. She writes code mainly in R using the ggplot2, shiny, data.table, purrr and rmarkdown packages. She has a major role in developing an in-house R package ecosystem of 5+ packages.

Section: Reproducible Research, Use-cases, Teaching

By the end of the tutorial participants should be able to:

- recognize code improvement possibilities,
- refactor analysis code by extracting some parts into functions,
- reason whether a piece of code could be regarded as clean,
- understand the benefits of applying clean code principles.

During the tutorial the participants will perform a guided data analysis task and make several refactoring steps as we progress. They will immediately experience the benefit of applying the shown techniques. The tutorial will cover launguage-agnostic and R-specific topics as well.

Some of the language-agnostic topics covered:

- extract code into functions
- organize code into files and folders
- organize functions within an R file
- how to choose variable and function names
- single responsibility principle
- what is a pure function

Some R-specific topics covered:

- how to extract some ggplot2 layers into functions
- how to create functions operating on different columns of a data frame
- parametrized rmarkdown documents
- useful RStudio keyboard shortcuts

Pre-requisites

Participants should be able to create simple R functions and use R for data analysis. We believe the tutorial is useful for beginners as well as for more experienced R programmers.

Justification

Most R users do not have a formal background in software engineering, however, coding is a significant part of their job. We believe every R user can benefit from writing cleaner and simpler code which also makes it more reusable and less error-prone. Writing clean code also helps with reproducibility. Refactoring early and often makes the life of your future self and your collaborators as well as others wishing to understand your code easier.

Building an Interpretable NLP model to classify tweets [Mon 9:00]

Speaker: Grace Meyer (Techincal Adviser- Data Science @ Oxera, United Kingdom) and Kasia Kulma

Grace Meyer is a commercial analytics expert with proven success in advising business strategy based on data driven insights. At Oxera, she applies machine learning to strategic projects and leads the data science and programming team. Dr. Kasia Kulma is a Data Scientist at Aviva with experience in building recommender systems, customer segmentations, predictive models and is now leading an NLP project. She is the author of the blog R-tastic. Grace & Kasia are both mentors in R-Ladies London.

Section: Machine Learning, Text mining

Unstructured text data is rapidly increasing in volume and variety and with advances in Machine Learning it's becoming available to be tapped for insights and patterns. One of the use-cases of predictive modelling in text analytics would be to classify an author based on text alone. However, even the most accurate model may be difficult to interpret and therefore understand how reliable it is or whether it produces insights that can be generalized. One of the solutions here is to apply the Local Interpretable Model-agnostic Explanations (LIME) framework to the classifiers to generate interpretable explanations.

In this workshop, we will take you step-by-step through tidytext principles and text-analytics pipeline to create a predictive model classifying tweets by Clinton or Trump. We will go over the data collection, exploration, feature engineering and model building. Finally, we will apply the LIME framework to better understand and interpret what drives model predictions.

R packages: readr, dplyr, tm, tidytext, text2vec , caret, xgboost, lime

Data used: Clinton-Trump-tweets @ https://www.kaggle.com/benhamner/clinton-trump-tweets

---

Geocomputation with R [Mon 9:00]

Speaker: Jannes Muenchow (Postdoc @ Friedrich Schiller University Jena, Germany) and Robin Lovelace, Jakub Nowosad

The speaker: - has a special interest in and passion for predictive mapping of landslide susceptibility and biodiversity (using statistical and machine learning models). - worked as a geo-data scientist for a location analyst consulting company. - is the creator and maintainer of the R package RQGIS and a co-author of the forthcoming book "Geocomputation with R".

Section: Spatial, Statistics

Geographic data is special and has become ubiquitous. Hence, we need computational power, software and related tools to handle and extract the most interesting patterns of this ever-increasing amount of (geo-)data. This workshop gives an introduction how to do so using R. It will introduce the audience how the two most important spatial data models - vector and raster - are implemented in R. The workshop will also give an introduction to spatial data visualization. Maps are a compelling way to display complex data in a beautiful way while allowing first

inferences about spatial relationships and patterns. Additionally, we will bridge R with Geographic Information Systems (GIS), i.e., we show how to combine the best of two worlds: the geoprocessing power of a GIS and the (geo-)statistical data science power of R. We will do so with a use case presenting spatial and predictive modeling.

By the end of this workshop, the participants should:

- know how to handle the two spatial data models (vector and raster) in R.
- import/export different geographic data formats.
- know the importance of coordinate reference systems.
- be able to visualize geographic data in a compelling fashion.
- know about geospatial software interfaces and how they are integrated with R (GEOS, GDAL, QGIS, GRASS, SAGA).
- know about the specific challenges when modeling geographic data.

Tutorial content

- The R spatial ecosystem
- Vector data model: simple features (**sf**)
- Raster data model (**raster**)
- Geographic data visualization (**ggplot2**, **mapview**, **tmap**)
- Bridges to GIS (**RQGIS**, **RSAGA**, **rgrass7**)
- Spatial modeling case study

Graphs: A datastructure to query [Mon 9:00]

Speaker: Benjamin Ortiz Ulloa (Data Visualization Engineer @ VT-ARC, United States)

A data visualization engineer who is very passionate about graph data structures.

Section: Databases, Big Data, Community, Text mining

When people think of graphs, they often think about mapping out social media connections. While graphs are indeed useful for mapping out social networks, they offer so much more. Graphs provide a datastructure that scales very well and can be queried in intuitive ways. Data structures in the real world resemble vertices and edges more than they resemble rows and columns. **Gremlin** and **Cypher** are query languages that take advanatage of the natural structure of graph databases. In this tutorial I will show how we can use **igraph** in a similar manner as these graph query languages to get new insites into our data.

The workshop will be divided in 4 parts:

1. A survey of graphs and how they are used (45 min)
    - Social Network Analyses
    - Natural Language Processing
    - NoSQL
    - Epidemiology
2. An introduction of **igraph** (70 min)
    - Graph structures
    - **igraph** syntax
    - Graph IO
    - Summary Statistics
    - Base Plotting
    - ggraph

3. Case Study: NLP (45 min)
    - Filter logic
    - Traversal logic
    - tidytext
    - magritr
4. Graph Ecosystem (30 min)
    - Gephi
    - D3
    - TinkerPop

R packages: igraph, magrittr, tidyr, ggplot2, ggraph

References:

- Practical Gremlin: https://github.com/krlawrence/graph
- R igraph manual pages: http://igraph.org/r/doc/
- Text Mining in R: https://www.tidytextmining.com/
- ggraph: https://github.com/thomasp85/ggraph
- Exploring Graphs: https://beemyfriend.github.io/graphs.html

---

Forwards Package Development Workshop for Women [Mon 9:00]

Speaker: Isabella Gollini (Assistant Professor in Statistics @ University College Dublin, Ireland) and Heather Turner

Dr Isabella Gollini is an Assistant Professor in Statistics at University College Dublin, Ireland. She is the author and contributor of three R packages: tailloss, GWmodel, lvm4net. Dr Heather Turner is a freelance consultant, with experience of developing packages for research and business. Her gnm package won the John M. Chambers Statistical Software Award in 2007. Isabella and Heather are core team members of the R Foundation Forwards taskforce for women and under-represented groups.

Section: Community, Infrastructure

An analysis of CRAN maintainers in 2016 estimated that 11.4% were women. This proportion is much lower than the proportion of women that attended the R conference useR! in the same year (28%). In addition a survey of the participants at that useR! conference found that women were less likely than men to have experience of contributing to or writing packages. Also women were less likely to use R recreationally, so perhaps have less opportunity to develop package development skills.

This workshop is designed to address this skills gap. It is for women who have done some R coding and are ready to take the next step in providing it to others to use.

During the tutorial participants will learn how to - make code into an R package, - do collaborative coding with GitHub, - write a vignette or an article, - build a package web page, - submit a package to CRAN.

Participants can bring their own code that they wish to make into a package, or work with our example.

R packages: knitr, devtools, pkgdown, rmarkdown, roxygen2, testthat

Deep Learning with Keras for R [Mon 13:30]

Speaker: Aimee Gott (Senior Data Science Consultant @ Mango Solutions, United Kingdom) and Douglas Ashton, Mark Sellors

Aimee, Douglas and Mark work in the data science team at Mango Solutions. Aimee is the lead trainer at Mango and has taught courses across all aspects of data science with a particular focus on R. Douglas is a principal consultant and specialises in machine learning/deep learning, working with customers to embed these techniques in their analytic workflows. Mark is head of data engineering and works with Mango customers to set up their infrastructure ready for advanced analytics techniques.

Section: Big Data, Machine Learning

If you don't work at one of the big tech giants, then deep learning may seem out of reach. Fortunately, in recent years the barrier to entry has dropped dramatically. Libraries, such as TensorFlow, have made it much easier to implement the low level linear algebra. While Keras builds on this to provide a high level Python API specifically for building neural networks. A single data scientist can now quickly build a deep network, layer by layer, without losing time on implementation details. You don't need terabytes of data and GPU clusters to get started; even relatively small problems can now benefit from deep learning.

A key aim of Keras is to reduce the time from idea to implementation. Many data scientists choose to use the R language for its first class statistics functionality, powerful data manipulation, and vibrant community. While Python is also a fantastic choice for data science, learning it is a significant investment when what you really want to be doing is trying out your idea. For this reason RStudio created the R Interface to Keras. This allows an R user to quickly experiment with neural networks to see if they are right for their problem.

In this workshop we will get you up and running with Keras for R. We will cover some theoretical background but the focus is on implementation. We will demonstrate how to setup different types of neural network to solve simple problems with time series, and give you the opportunity to build your own with guided exercises.

A cloud based RStudio Server environment will be provided so attendees only require a laptop with internet access and a modern browser. Basic R knowledge is required, and it will help if attendees are familiar with packages such as dplyr for data manipulation.

Automatic and Interpretable Machine Learning in R with H2O and LIME [M13:30]

Speaker: Jo-fai Chow (Data Scientist @ H2O.ai, United Kingdom)

Jo-fai (or Joe) is a data scientist at H2O.ai. Before joining H2O, he was in the business intelligence team at Virgin Media in UK where he developed data products to enable quick and smart business decisions. He also worked remotely for Domino Data Lab in the US as a data science evangelist promoting products via blogging and giving talks at meetups.

Section: Machine Learning

General Data Protection Regulation (GDPR) is just around the corner. The regulation will become enforceable a week after the eRum conference (from 25 May 2018). Are you and your organization ready to explain your models?

This is a hands-on tutorial for R beginners. I will demonstrate the use of two R packages, h2o & LIME, for automatic and interpretable machine learning. Participants will be able to follow and build regression and classification models quickly with H2O's AutoML. They will then be able to explain the model outcomes with a framework called Local Interpretable Model-Agnostic Explanations (LIME).

References:

- Hall et al (2017): Ideas on interpreting machine learning.
- Ribeiro et al (2016): Introduction to Local Interpretable Model-Agnostic Explanations.
- Ribeiro et al(2016): Why Should I Trust You? Explaining the Predictions of Any Classifier.
- Wikipedia (2018): General Data Protection Regulation.

The beauty of data manipulation with data.table [Mon 13:30]

Speaker: János Divényi (Lead Data Scientist @ Emarsys, Hungary)

János Divényi is a PhD candidate in economics at the Central European University (CEU) who works as lead data scientist at Emarsys in Budapest. He writes code in R (and Python), likes to think carefully about causality, and seeks intuitive understanding of complicated stuff. He is an occasional speaker of the local R meetup, and has more than 5 years' experience of teaching from various institutions (CEU, BME, MCC).

Section: Data Munging, Reproducible Research, Use-cases

The data.table package is a powerful tool for manipulating data, especially if the underlying data set gets large (~ above 1 GB). In spite of its clear advantages the package is underused. Many R users are afraid of it because of its "ugly" syntax. This workshop aims to dismantle this belief by showing the beauty in the package logic, and illustrating its strengths in performance.

By the end of the tutorial participants should be able to:

- use data.table for common data manipulation tasks (data input/output, aggregation, reshaping, joins)
- build on their understanding of data.table syntax to solve more complicated tasks
- compare the performance of the package to other widely used alternatives

The workshop would start with a short introduction into the logic and syntax of the data.table package. The participants would discover the power and beauty of the package through guided data manipulation tasks borrowed mainly from Emarsys use cases.

Pre-requisites

Participants should be able to solve data manipulation tasks in R (using base R or the hadleyverse/tidyverse). No knowledge of the data.table package is required.

Building a package that lasts [Mon 13:30]

Speaker: Colin FAY (Data Analyst, R trainer, Social Media Manager @ ThinkR, France)

Colin Fay is Data Analyst, R trainer and Social Media Expert at ThinkR, a French agency focused on everything R-related. Colin is a prolific open source developer, author of more than 12 R packages actively maintained on GitHub (6 of them being on CRAN): attempt, proustr, tidystringdist... He also contributes to several other packages. He is a member of the RWeekly team, a collaborative news bulletin about R, and the cofounder of the Breizh Data Club, an association of French data professionals.

Section: Infrastructure, Reproducible Research

You've got the killer idea for an R package, but you're not sure where to start? Then you've come to the right place!

During this workshop, we'll go through the whole process of building a package that lasts. In other words, we'll review the best practices to create a package that works (obviously), but more importantly a package that is extensively tested to prevent bugs, that will be easier to maintain on the long run, and that will be fully documented.

At the end of this workshop, the attendees will have a road map for writing robust packages, designed for industrial use and/or for CRAN.

Plan of the workshop

- Package basics : understand and organise your package structure
- Best practices for writing functions in packages (optimised for speed and for maintenance)
- Using tests : why you should write tests for your package, and how to do it.
- Documentation : Using roxygen to document your functions, create a vignette to explain what your package does, and enhance your documentation with pkgdown.
- Continuous integration and code coverage with GitHub, Travis and Codecov

R packages: devtools, usethis, attempt, testthat

Required skills of participants

- Basic knowledge of R
- Functions
- Basic Markdown

Required work to do before workshop

- Install a recent version of R and RStudio on your laptop
- Participants can come with a series of functions they want to put in a package, or use the examples that will be provided by the speaker during the workshop.

Building a pipeline for reproducible data screening and quality control [M 13:30]

Speaker: Claus Ekstrøm (Professor @ Biostatistics, University of Copenhagen, Denmark) and Anne Helby Petersen

Anne Helby Petersen holds a MS in statistics and is the primary author of the dataMaid R package. She is experienced in communicating statistical topics to a wide audience as a teaching assistant at the University of Copenhagen. Claus Thorn Ekstrøm is the creator/contributor to several R packages

(dataMaid, MESS, MethComp, SuperRanker) and is the author of "The R Primer" book. He has previously given tutorials on Dynamic and interactive graphics and the role of interactive graphics in teaching.

Section: Data Munging, Reproducible Research

One of the biggest challenges for a data analyst is to ensure the reliability of the data since the validity of the conclusions from the analysis hinges on the quality of the input data. This tutorial will cover the workflow of data screening and -validation that transforms raw data into data that can be used for statistical analysis. In particular, we will discuss organizing research projects, tidy data formats, internal and external validity of data, requirements for reproducible research, the dataMaid R package for customized data screening, the assertr and assertive R packages for data validation and data validation rule sets, and how to produce code books that summarize the final result of the data screening process and provide a starting point for the subsequent statistical analyses.

Plotting spatial data in R [Mon 13:30]

Speaker: Martijn Tennekes (Data Scientist @ Statistics Netherlands, Netherlands)

Martijn Tennekes has a PhD in game theory, and has been working at Statistics Netherlands for eight years on data visualization, big data, and R. He authored three data visualization packages (treemap, tabplot, and tmap).

Section: Spatial, Graphics

In this workshop you will learn how to plot spatial data in R by using the **tmap** package. This package is an implementation of the grammar of graphics for thematic maps, and resembles the syntax of **ggplot2**. This package is useful for both *exploration* and *publication* of spatial data, and offers both *static* and *interactive* plotting.

For those of you who are unfamiliar with spatial data in R, we will briefly introduce the fundamental packages for spatial data, which are **sf**, **sp**, and **raster**. With demonstrations and exercises, you will learn how to process spatial objects from various types (polygons, points, lines, rasters, and simple features), and how to plot them. Feel free to bring your own spatial data.

Besides plotting spatial data, we will also discuss the possibilities of *publication*. Maps created with **tmap** can be exported as static images, html files, but they can also be embedded in **rmarkdown** documents and **shiny** apps.

R packages: tmap, sf, sp, raster, rmarkdown, shiny

Tennekes, M. (2018) tmap: Thematic Maps in R. Forthcoming in the Journal of Statistical Software (JSS).

Keynotes

45 minutes invited, plenary talks followed by 5 minutes of Q&A in the main auditorium.

Speaker: Martin Mächler (Senior Scientist in Statistics @ ETH Zurich, Switzerland) [Tue 9:00]

Martin is a Mathematician (Ph.D. ETH Z) and Statistician, Lecturer and Senior Scientist at Seminar für Statistik, ETH Zurich, R Core member, Secretary General of the R Foundation. Authored more than 20 R packages (such as Matrix, cluster, robustbase, cobs, VLMC, bitops or copula). Emacs ESS Core Developer since 1997 and Project Leader since 2004, author of several books and over 50 scientific journal articles.

Sentiment Analysis on Social Media and Big Data [Tue11:00]

Speaker: Stefano M. Iacus (Professor of Statistics @ University of Milan, Italy)

Stefano is a full professor in Statistics, former R Core Team member (1999-2014) and maintainer of several R packages e.g: (sde, cem, rrp and opefimor). Founder and president of Voices from the Blogs running sentiment analysis and text mining projects. Author of several scientific books, book chapters and journal articles.

Extracting semantic content from texts has a long history in statistics but it has become quite a popular theme very recently with the advent of social media, especially those like Twitter, which mainly dispatch text messages. Applications range from marketing to election forecasting, well being or happiness analysis, market mood, etc. Despite a huge development of automatic methods, NLP or ontology based algorithms, social media data are not as easy to analyze as one might think compared to the analysis of romance or other forms of digital texts. The reason being the creativity and continuous change of online language. Supervised methods seem to be the only option. We will present some recent theoretical developments and a series of real life applications of these techniques to Twitter and other big data.

Learning from (dis)similarity data [Tue13:30]

Speaker: Nathalie Villa-Vialaneix (Researcher @ INRA, France)

Nathalie is a researcher at the French National Institute for Agronomical Research (INRA) in the Unit of Applied Mathematics and Computer Sciences in Toulouse. She is the maintainer of the SOMbrero, SISIR and RNAseqNet R packages and author of a number of others. She received her PhD in Mathematics from the University Toulouse 2 (Le Mirail), in 2005. She is a board member of the biostatistics platform in Toulouse and a former board member of the French Statistical Association (SFdS).

In some applications and in order to better address real-world situations, data can be known through pairwise measures of resemblance or difference between the objects of interest (similarities, dissimilarities, kernels, networks...). This talk will describe a general framework to deal with such data, especially focusing on the unsupervised setting and exploratory analyses. Also, solutions for combining multiple relational data - each providing a different view on a

specific aspect of the data - will be described. The talk will provide an overview of applications of this framework to self-organizing maps (R package SOMbrero), constrained hierarchical clustering (R package adjclust) and PCA (R package mixKernel), with illustrations on case studies in the fields of biology and social sciences.

A practical history of R (where things came from) [Wed 9:00]

Speaker: Roger Bivand (Professor @ Norwegian School of Economics, Norway)

Roger received his PhD in geography from the London School of Economics and post-doctoral degree from Adam Mickiewicz University. His current research interests are in developing open source software for analysing spatial data. He has been active in the R community since 1997. He is an auditor of the R Foundation, editor of the Journal of Statistical Software, Journal of Geographical Systems, Geographical Analysis and Norsk Geografisk Tidsskrift; and Editor-in-Chief of the R Journal.

Not infrequently, we wonder why choices such as stringsAsFactors=TRUE or drop=TRUE were made. Understanding the original uses of S and R (in the 1900s), and seeing how these uses affected the development of R lets us appreciate the robustness of R's ecosystem. This keynote uses readings of the R sources and other information to explore R's history. The topics to be touched on include the "colour" books (brown, blue, white, green), interlinkages to SICP (Scheme) and LispStat, the lives of R-core and CRAN, Ancients and Moderns (see Exploring the CRAN social network).

R/exams: A One-for-All Exams Generator [Wed 13:40]

Speaker: Achim Zeileis (Professor of Statistics @ Universität Innsbruck, Austria)

Being an R user since version 0.64.0, Achim is co-author of a variety of CRAN packages such as zoo, colorspace, party(kit), sandwich, or exams. He is a Professor of Statistics at the Faculty of Economics and Statistics at Universität Innsbruck. In the R community he is active as an ordinary member of the R Foundation, co-creator of the useR! conference series, and co-editor-in-chief of the open-access Journal of Statistical Software.

A common challenge in large-scale courses is that many variations of similar exercises are needed for written exams, online tests conducted in learning management systems (such as Moodle, Blackboard, etc.), or live quizzes with voting via smartphones or tablets. The open-source package exams for R provides a one-for-all approach to automatic exams generation, tying together various open-source packages (in R and beyond). It is based on individual exercises for multiple-choice or single-choice questions, numeric or text answers, or combinations of these. The format can be either in R/LaTeX or R/Markdown containing questions/solutions with some random numbers, text snippets, plots/diagrams, or even individualized datasets. The exercises can be combined to exams and easily rendered into a number of output formats including PDFs for classical written exams (with automatic evaluation), import formats for various learning management systems, live voting (via ARSnova), and the possibility to create custom output (in PDF, HTML, Docx, ...).

It is illustrated how the Department of Statistics at Universität Innsbruck manages its large introductory mathematics course using PDF exams that can be automatically scanned and evaluated, online tests in the OpenOLAT learning management system, and live quizzes in the ARSnova audience response system. Furthermore, it is demonstrated how psychometric item response theory (IRT) can be utilized to gain further insights into the difficulty of the questions, ability of the students, and the "fairness" of the exam across participants.

Invited Talks

18-minutes invited talks followed by 2 minutes of Q&A in the main auditorium.

Using Rust code in R packages [Tue 9:50]

Speaker: Jeroen Ooms (Postdoctoral researcher @ rOpenSci, USA)

Jeroen graduated in 2014 at the UCLA department of statistics and is now a post doctoral researcher at UC Berkeley with the rOpenSci group. His official job description involves development of algorithms and software to enable processing, security and archiving of research data to facilitate data-driven open science. In practice he writes R packages that do cool and important stuff. Some popular ones are opencpu, jsonlite, curl, V8, openssl, mongolite, commonmark, pdftools and hunspell. Recently he developed an interest in cryptography and the decentralized web.

Taking inspirations from proven frontend frameworks to add to Shiny with 4 new packages [Tue 11:50]

Speaker: Olga Mierzwa-Sulima (Senior Data Scientist @ appsilon, Poland)

Olga is a senior data scientist at Appsilon Data Science and a co-founder of datahero.tech. She leads a team of data scientists and build data science predictive/explanatory solutions and deploy them in production, usually wrapped in a Shiny App UI. She develops Appsilon's open-source R packages. Olga holds a MSc degree in Econometrics from the University of Rotterdam. She co-organizes the largest meetup of R users in Poland and is a co-founder of R-Ladies Warsaw chapter.

Section: Businesses, Dashboards, Machine Learning, Time-series

There is no need to praise Shiny for its influence on results presentation. It's no longer only a tool internally used by data science teams. Currently it's becoming an alternative for business and is replacing both the BI solutions and custom made web applications. In order to face the competition, it needs constant development and new features. As with many other technology stacks, Shiny could benefit from community contributions for further development of the packages themselves and growth of independent libraries. In this presentation we will talk about four novel packages that add interesting capabilities to Shiny such as beautiful UI shiny.semantic, semantic.dashboard, routing shiny.router, and internationalization shiny.i18n. All four packages have been developed to meet the needs that had to be addressed in business projects. This indisputably shows the gaps in the current Shiny capabilities that these packages fill. During the development process of the packages we took inspiration from proven frontend frameworks such

as Meteor or Django. We will demo these packages and show how their development adds to the open source community, thereby helping companies adopt R/Shiny.

Scalable Automatic Machine Learning in R [Tue 14:20]

Speaker: Erin LeDell (Chief Machine Learning Scientist @ H2O.ai, USA)

Erin LeDell is the Chief Machine Learning Scientist at H2O.ai, an artificial intelligence company in Mountain View, California, USA, where she works on developing H2O, an open source library for scalable machine learning. Before joining H2O.ai, she was the Principal Data Scientist at Wise.io and Marvin Mobile Security, and the founder of DataScientific, Inc. Erin received her Ph.D. in Biostatistics from University of California, Berkeley and has a B.S. and M.A. in Mathematics.

Section: Machine Learning

In this presentation, we provide an overview of the field of "Automatic Machine Learning" (AutoML) and introduce the AutoML functionality in the scalable and distributed machine learning library, H2O. We will present our unique methodology for automating the machine learning workflow, which includes feature pre-processing and automatic training/tuning of many models, with the goal of maximizing model performance.

H2O AutoML provides an easy-to-use interface which automates the process of training a large selection of candidate models without any user configuration or knowledge about specific machine learning algorithms. The interface is designed to have as few parameters as possible so that all the user needs to do is point to their dataset, identify the response column and optionally specify a time-constraint. The user can also specify which model performance metric that they'd like to optimize and use a metric-based stopping criterion for the AutoML process rather than a specific time constraint. By default, several Stacked Ensembles will be automatically trained on the collection individual models to produce a highly predictive ensemble model which, in most cases, will top the AutoML Leaderboard.

H2O AutoML is available in all H2O interfaces including the h2o R package, Python module and the Flow web GUI. We will also provide simple R code examples to get you started using H2O AutoML.

Gradient Boosting Machines (GBM) in R [Tue 14:40]

Speaker: Szilard Pafka (Chief Scientist @ Epoch, United States)

Szilard has a PhD in Physics for using statistical methods to analyze the risk of financial portfolios. For the last decade he's been the Chief Scientist of a tech company in California doing everything data (analysis, modeling, data visualization, machine learning etc). He is the founder of the LA R meetup, the author of a machine learning benchmark on github (1000+ stars), a frequent speaker at conferences, and he has taught graduate machine learning courses at two universities (UCLA, CEU).

Section: Big Data, Businesses, HPC, Machine Learning

With all the hype about deep learning and "AI", it is not well publicized that for structured/tabular data widely encountered in business applications it is actually another machine learning

algorithm, the gradient boosting machine (GBM) that most often achieves the highest accuracy in supervised learning tasks. In this talk we'll review some of the main GBM implementations available as R packages such as gbm, xgboost, h2o, lightgbm etc, we'll discuss some of their main features and characteristics, and we'll see how tuning GBMs and creating ensembles of the best models can often achieve unparalleled prediction accuracy.

Compositional analysis of our favourite drinks [Tue 15:50]

Speaker: Matthias Templ (Senior lecturer @ ZHAW, Switzerland)

Matthias Templ is lecturer at the Zurich University of Applied Sciences, Switzerland. His research interest includes imputation, statistical disclosure control, compositional data analysis and computational statistics. He published two books and more than 45 papers. Additionally, he is the author of several R packages. In addition, Matthias Templ is the editor-in-chief of the Austrian Journal of Statistics. With two of his colleagues he owns and founded the company data-analysis OG.

Section: Statistics, Use-cases

Compositional data are nowadays widely accepted as multivariate observations carrying relative information. Compositional data follow the principle of scale invariance, typically being represented in proportions and percentages. In other words, for compositional data the relevant information is contained in the (log-)ratios between the components (parts). Compositional data are present in almost any field of research. Examples for compositional data are, for example, concentration of chemical elements in soil samples, time budget data, expenditures, tax or wage components or percentages and ratios reported in various tables.

Through data from our favourite drinks, we will show the usefulness of the representation of data in isometric coordinates and the analysis of these coordinates instead of analysing the raw data on the simplex. As a side note of the talk, we want to answer such important questions of life: will the quality of beer mainly depend on age and how it should be stored? Should you drink blended coffee, or is Scottish Whisky really different to Irish or American Whiskey? We use the package *robCompositions* for all practical examples.

Show me your model 2.0 [Tue 16:10]

Speaker: Przemyslaw Biecek (Associate Professor @ Warsaw University of Technology, Poland)

Applied statistician working with high-dimensional models for personalised oncology. Interested in Machine Learning, obsessed with DataVis. Huge believer of data-literacy education for not-only-kids (http://betabit.wiki). Head of MI2 DataLab in Warsaw (https://mi2datalab.github.io/MI2DataLab_webpage). Currently working on descriptive explainers for black-box machine learning models.

Section: Machine Learning

According to many Kaggle competitions, winning solutions are often obtained with elastic tools like random forest, xgboost or neural networks. These algorithms have many strengths but also share a major weakness, which is the lack of interpretability of a model structure. Still we may extract some knowledge about the model structure. During this talk I will overview core

techniques for exploration of machine learning models, like: Partial Dependency Plots, Individual Conditional Expectations, Merging Path Plots, Local Interpretable Visual Explanations and Break Down plots.

A Future for R: Parallel and Distributed Processing in R for Everyone [Wed 9:50]

Speaker: Henrik Bengtsson (Associate Professor @ University of California, San Francisco (UCSF), United States)

Henrik Bengtsson has a background in Computer Science (MSc) and Mathematical Statistics (PhD) and is an Associate Professor at the UCSF Department of Epidemiology and Biostatistics. He has extensive experience in applied statistics, computational genomics, and large-scale processing. He has worked with R since 2000 and since contributed 30+ packages to CRAN and Bioconductor.

Section: Big Data, HPC, Infrastructure, Reproducible Research

In programming, a *future* is an abstraction for a *value* that may be available at some point in the future. The non-blocking nature of futures makes them ideal for *asynchronous* evaluation of expressions in R, e.g. in parallel on the local machine, on a set of remote machines, or in the cloud.

I am presenting the simple, unified, cross-platform future ecosystem for parallel and distributed processing in R. The most fundamental construct is f <- future({ expression }) for evaluating the expression asynchronously and later in the program retrieve the value using v <- value(f). How and when futures are resolved depends is specified by plan() without further modification of the R code. Higher level map-reduce constructs exists, e.g. future_lapply(), and doFuture backends for foreach(). The future framework, available on CRAN, has been used in production for several years.

R packages: future, future.apply, doFuture, future.batchtools

---

Drilldown data discovery with Shiny [Wed 11:00]

Speaker: Barbara Borges Ribeiro (Software Engineer @ RStudio, Spain)

Barbara is a software engineer at RStudio working primarily in the Shiny package. She holds a double major in Statistics and Computer Science from Macalester College. After four freezing Minnesota winters, she is back in her warm homeland of Portugal (but to the disappointment of many, she's not a soccer fan).

Data science is often thought as carefully building up from data. However there are many cases where going the other away around, and drilling down into the data, can also be extremely useful. Have you ever seen a plot where something seems off? Maybe it's a few egregious outliers or a quirk in the expected trend. Instead of going back to the drawing board immediately, you can leverage the power of Shiny to allow you to interactively start from an aggregate visualization (or summary) and then drill down into the lower-level, finer-grained data. Whether it is by interactively creating new tabs, modal windows or other methods, drilling down allows you to discover data that's been right under your nose, without having to leave your Shiny app. In addition to drilling down, you can also drill through (by looking at snapshots of the data at

different periods), and even drill up (by creating aggregate values from the underlying data). These capabilities allow for more satisfying data presentation or data reporting Shiny apps, since its consumers can investigate the data to their heart's content. In this talk, I will demo a mock personal finance Shiny app that takes advantage of functions like insertUI/removeUI, appendTab/removeTab, and showModal/removeModal.

Getting the most out of GitHub and friends [Wed 11:20]

Speaker: Colin Gillespie (Data Scientist/Senior Lecturer @ Jumping Rivers/Newcastle University, United Kingdom)

Colin Gillespie is Senior lecturer (Associate professor) at Newcastle University, UK. He has been running R courses for over eight years at a variety of levels, ranging from beginners to advanced programming. He is co-author of the recent book Efficient R programming, O'Reilly.

Section: Infrastructure

Over the last few years, the popularity of git and GitHub has been on the increase. In this talk, we'll discuss a number of Github's friends and how to incorporate them into your workflow. During the course of the talk, we'll cover (amongst other things) how we can perform automatic checks on commits via travis & codecov, hosting books via the bookdown package, allowing twitter to notify your followers whenever you push and maintaining your own R package repostory.

Tracking changes in data with the lumberjack package [Wed 14:30]

Speaker: Mark van der Loo (Methodologist @ Statistics Netherlands, Netherlands)

Mark van der Loo works as a consultant and researcher at the department of statistical methods of Statistics Netherlands. He has (co)authored and published several R packages related to data cleaning, including 'validate', 'dcmodify', 'errorlocate', 'extremevalues', and 'stringdist'. Mark is coauthor of the book 'Statistical Data Cleaning with Applications in R' published by Wiley, Inc (2018).

Section: Reproducible Research, Data Munging, Infrastructure

A data analyses may contain many steps where data is modified or corrected. For reasons of quality control and efficiency it is imperative to be able to understand the effect of each step on the final result. One way to do this is to somehow measure (log) the effect that each data processing step has on the data.

The lumberjack package offers an elegant solution by endowing the function composition (pipe) operator with the capability of logging changes in data flowing through it. This means that the effect of any data processing function that adheres to the 'tidy' data-in-data-out style can be monitored almost effortlessly. The package offers several basic loggers as well as a framework for users and package authors to refine logging by defining their own loggers.

In this talk I will go into the design principles behind the package, demonstrate the workflow, and show several examples of loggers that usefully summarize changes in data.

Demographics with Genealogical Data [Wed 15:40]

Speaker: Arthur Charpentier (Professor @ Universite de Rennes, France) and Ewen Gallic

Professor at the faculty of Economics at Universite de Rennes, in France. Editor of 'Computational Actuarial Science with R' (CRC Press, 2014) and of the blog https://freakonometrics.hypotheses.org/.

Section: Big Data, Spatial, Use-cases, Social Sciences, Databases, Finance

In our study, we try to understand French migration (within France) in the XIXth century. Through a partnership with a genealogical site, we have obtained almost a billion 'records'. In the first part of the talk, we will discuss those data, and how to study them, with R. The most difficult tasks is that most tree are ascendant (from children to grand parents), but when studying migration, we must have a descendant approach (from grand parents to children). Furthermore, since we use collaborative data, there are a lot of doublons, and most of them are difficult to track (typos). In the second part, we will discuss semiological aspects : how to visualize complex information, with R.

References: Charpentier and Gallic (2018): Studying French Migration in the XIXth century with collaborative genealogical data.

Regular Talks

18-minutes contributed talks followed by 2 minutes of Q&A in one of the auditoriums.

A recipe for recipes [Tue 9:50]

Speaker: Edwin Thoen (Data Scientist @ funda, Netherlands)

I am a data scientist at funda, which is the Netherlands' principal website for selling and renting houses. I use R on a daily basis and I am the author and maintainer of the padr package. I contributed also to GGally and recipes.

Section: Data Munging, Machine Learning, Reproducible Research, Statistics

The recipes package (Kuhn and Wickham, 2018) is a set of “Preprocessing Tools to Create Design Matrices”. It delivers a framework in which preprocessing steps on one or more variables are captured in individual objects, called steps and checks. The former do transformations on variables, the latter assert that expectations about the variables are met. A recipe object is created on a train set, to which the steps and checks are added one by one. Once the recipe is done, the prep method is used to estimate all the relevant statistics from the variables. Finally, the actual transformations are applied to data sets using the bake function on the recipe. New data, such as test sets or future observations to score, also run through the recipe via bake. This ensures that the exact same preparation is used on all data sets.

To leverage recipes fully, one should add their own steps and checks to the ones that are shipped with the package. However, whereas the use of the package is intuitive and quick to pick up, writing custom steps and checks requires some understanding of the package inner workings. In this talk I will give a quick introduction to the package and I will elaborate on how to create your

own steps and checks. Providing a framework, or if you like a recipe, for them. After attending this talk you should be able to create your own steps and checks.

Harness the R condition system [Tue 10:10]

Speaker: Lionel Henry (Programmer @ RStudio, Belgium)

Lionel developed a passion for R programming while studying political science and statistics. He is now a programmer at RStudio in the tidyverse and r-lib team.

Among the many unusual features of the R language, the condition system is probably one of the most obscure. While it is mainly used for exception handling, it is much more powerful and general. This presentation will get advanced R programmers up to speed on the difference between messages and printed output, the difference between exiting and inplace handlers, how does condition muffling work (e.g. suppressWarnings() and suppressMessages()), how to implement error recovery, and how to use error objects to make unit testing more robust and to pass useful metadata to error handlers.

The essentials to work with object-oriented systems in R [Tue 10:10]

Speaker: Ildiko Czeller (Data Scientist @ Emarsys Technologies Kft, Hungary)

Ildi Czeller is a mathematician who has worked as a data scientist at Emarsys in Budapest for almost 3 years now. She writes code mainly in R using the ggplot2, shiny, data.table, purrr and rmarkdown packages. She has a major role in developing an in-house R package ecosystem of 5+ packages.

Section: Teaching, Use-cases

All R users have used S3, the oldest and most prominent object-oriented system in R even if they were unaware of it, for example by using the summary function both for data frames and for linear models. The two main building blocks of an object-oriented system are objects with specific type (class) and functions (methods) which behave differently depending on the class of their parameters. Most R users probably also had an experience where they got unexpected results which would have been easier to understand with a foundation in object-oriented systems in R. This talk aims to fill some of the gaps so that you can work confidently with existing code utilizing S3 or S4.

The three widely used object-oriented systems are S3, S4 and R6. This talk will focus on S3 which is the most widely used and assume no prior knowledge of object-oriented systems. I will start with a visual explanation of the most important concepts and then I will show you how understanding the basics can help you in your day-to-day work. I will guide you with examples and show hands-on tricks to understand, debug and get the documentation of existing code utilizing S3 or S4.

Multi-state churn analysis with a subscription product [Tue 11:50]

Speaker: Marcin Kosiński (Statistician @ Gradient Metrics, Poland) and Thomas Vladeck, Kyle Block

Challenges seeker and devoted R language enthusiast. In the past, keen on the field of large-scale online learning and various approaches to personalized news article recommendation. Co-organizer of the +1600 members R Enthusiasts meetups in Warsaw and Polish R Users Conferences 2017/2018 (whyr.pl). Interested in R packages development and survival analysis models. Currently explores and improves methods for quantitative marketing analyses and global surveys at Gradient Metrics.

Section: Businesses, Machine Learning, Statistics

Subscriptions are no longer just for newspapers. The consumer product landscape, particularly among e-commerce firms, includes a bevy of subscription-based business models. Internet and mobile phone subscriptions are now commonplace and joining the ranks are dietary supplements, meals, clothing, cosmetics and personal grooming products.

Standard metrics to diagnose a healthy consumer-brand relationship typically include customer purchase frequency and ultimately, retention of the customer demonstrated by regular purchases. If a brand notices that a customer isn't purchasing, it may consider targeting the customer with discount offers or deploying a tailored messaging campaign in the hope that the customer will return and not "churn".

The churn diagnosis, however, becomes more complicated for subscription-based products, many of which offer multiple delivery frequencies and the ability to pause a subscription. Brands with subscription-based products need to have some reliable measure of churn propensity so they can further isolate the factors that lead to churn and preemptively identify at-risk customers.

During the presentation I'll show how to analyze churn propensity for products with multiple states, such as different subscription cadences or a paused subscription. If the time allows I'll also present useful plots that provide deep insights during such modeling, that we have developed at Gradient Metrics - a quantitative marketing agency (http://gradientmetrics.com/).

R packages: survminer, survival

Not all that Shiny by default [Tue 12:10]

Speaker: Mikołaj Olszewski (Data Scientist @ Pearson, Poland) and Mateusz Otmianowski

Mikołaj Olszewski is an experienced Data Scientist working at Pearson and the co-founder of a startup providing high quality Data Science trainings. He's passionate about R & Shiny, and crazy about learning. Loves to learn new things and teach others.

Section: Web Apps, Dashboards

Shiny is a popular web application framework that allows to quickly build analytical dashboards using only R. It has many great built-in features that address the needs of most of users. But, what if you need to go beyond that and build custom solutions packed with features that are not available straight away? How far can you push Shiny without breaking it?

It turns out that quite far. In the Exploratory Data Science team at Pearson, we use Shiny to create self-serve analytical tools for internal clients. While we create these tools based on initial

requirements, we often need to be able to add additional features as the tool matures. This forces us to think outside the box about how we can build upon the framework offered by Shiny.

In this talk, we'll present our most complex Shiny app yet. We'll talk about the process we went through while building it, from prototyping through development, user testing, and deployment. We will cover advanced features of Shiny (e.g. a modularised application structure), communication between Shiny server and JavaScript, as well as sophisticated UI solutions (e.g. click drilldowns, interactive tutorials and custom welcome screens). Most importantly, we will cover what we have learned through rounds of usability testing that entirely changed our approach to data application design.

Sparsity with multi-type Lasso regularized GLMs [Tue 14:20]

Speaker: Sander Devriendt (PhD student @ KU Leuven, Belgium) and Katrien Antonio, Edward Frees, Tom Reynkens

Sander Devriendt is a PhD student at the actuarial research group of KU Leuven under the supervision of prof. Katrien Antonio. His research focusses on actuarial predictive modeling for mortality, insurance pricing, fraud, reserving and loss modeling.

Section: Machine Learning, Statistics

Current datasets often contain many variables. To cope with this, sparse regression methods have been developed to obtain more interpretable and better performing models than their classical counterparts. Standard regularization methods such as the Lasso or Elastic Net are already implemented in *R* to obtain sparse GLM results. However, variables often have different structural properties demanding different regularization techniques to obtain logical results. We propose a multi-type Lasso approach to solve this problem, where coefficients of different variables can be regularized by different penalties. Our proposed estimation procedure uses the theory of proximal to split the objective function into smaller subproblems. Each of these subproblems will only contain one penalty such that existing algorithms can be applied. We show how this setup is implemented in *R* and how different tools and packages work together to get the most out of our approach. The setup is being implemented in an *R* package where bottleneck calculations happen in *C++* through the *Rcpp* API.

R packages: mgcv, Rcpp, RcppArmadillo, speedglm, glmnet, parallel

References:

- Gertheiss, J. and Tutz, G. (2010). *Sparse modeling of categorical explanatory variables*. The Annals of Applied Statistics, 4(4), 2150-2180.
- Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical learning with sparsity: the Lasso and generalizations*. Chapman and Hall/CRC Press.
- Parikh, N. and Boyd, S. (2013). *Proximal algorithms*. Foundations and Trends in Optimization, 1(3), 123-231.

Nonlinear mixed-effects models in R [Tue 14:40]

Speaker: Francois Mercier (Senior Principal Scientist @ F. Hoffman-La Roche, Switzerland)

Francois Mercier has a background in statistics and is currently leading a group of Pharmacometricians at Roche/Genentech. Over his 20 years of experience in the Pharma industry, he has accumulated a large experience in model-informed drug development to support decision making in disease areas like multiple sclerosis or solid tumor cancer.

Section: Medical / Pharma, Statistics

Mixed-effect models are commonly used when data consist of clusters of observations. If the model is nonlinear in the parameters, nonlinear mixed-effects (NLME) models are considered. Growth or shrinkage data where the change trajectory is nonlinear is a common type of data modeled with NLME models, with the subject treated as a cluster and the repeated measurements as the individual observations nested within the cluster. Ordinary differential equations (ODEs) provide a natural framework to describe dynamic systems but sometimes they accept closed-form solutions.

In the recent years, a number of R packages have been created to fit NLME models using various approaches. These are the *nlme* (Pinheiro et al. 2017), *saemix* (Comets et al. 2017), *nlmixr* (Wang et al. 2017), and *brms* (Buerkner et al. 2016) packages; .each of them has its own specificities. In this talk, we compare the features (algorithms, grammar, documentation, examples), advantages (predictions, embedded model diagnostic tools), and limitations (scope, computational time, dependencies and other constraints) of these 4 solutions.

To illustrate this analysis, we consider two examples taking from the pharmaceutical industry. The first one is a real case of pharmacokinetic data measured in a small number of subjects (N<40), followed up for a long period of time with a rich sampling scheme (more than 8 assessment time points). The second one is a real case of longitudinal tumor size data measured in a large number of solid cancer patients (N>100), followed up for a short period of time (with a median of 4 assessment time points).

R packages: *nlme*, *saemix*, *nlmixr*, *brms*

Tools for using TensorFlow with R [Tue 15:00]

Speaker: Andrie de Vries (Solutions Engineer @ RStudio, United Kingdom)

Andrie started using R in 2009 for market research statistics and joined Revolution analytics in 2013, where he helped enterprise with their challenges in adopting R for machine learning. After the acquisition of Revolution analytics by Microsoft in 2015, he worked with customers on their implementation of neural network and machine learning project. In 2017 he joined RStudio as a Solutions Engineer. Andrie is co-author of "R for dummies".

Section: Machine Learning

In this session I demonstrate how to use R to train and deploy deep learning models with TensorFlow and Keras. You can easily construct scalable data pipelines, train models, evaluate training runs, and deploy your models to a variety of platforms. Until recently these tools were Python centric, but using new R packages (provided by RStudio and available from CRAN), any R programmer can now easily use TensorFlow.

Key takeaways:

1. Use R to create rich machine learning models with TensorFlow. Several new R packages enable you to train models, including tensorflow, keras and tfestimators. You can also use the packages tfdatasets (create scalable input pipelines), tfdeploy (deploy models), and tfruns (track, visualize and manage training runs). These packages provide you with a complete tool chain for end-to-end data preparation, model training and deployment.
2. RStudio desktop and server has great tooling to make development easy - you can easily track, visualize and manage TensorFlow training runs.
3. Deploy and publish your TensorFlow models to a variety of platforms, including Google CloudML, Paperspace and RStudio Connect.

R packages: tensorflow, keras, tfestimators, tfdatasets, tfruns, tfdeploy, cloudml

References:

- R Interface to TensorFlow (website). https://tensorflow.rstudio.com
- Chollet and Allaire, Deep Learning with R (2018): https://www.amazon.com/Deep-Learning-R-Francois-Chollet/dp/161729554X

---

bamlss.vis - an R package for interactively visualising distributional regression models [Tue 15:00]

Speaker: Stanislaus Stadlmann (PhD Student @ Georg-August University of Göttingen, Germany)

Stanislaus Stadlmann is a PhD student in Statistics from the University of Göttingen who has written R packages mostly surrounding the Shiny universe, e.g. Gotta Read 'Em All, bamlss.vis and ghp.

Section: Statistics, Web Apps, Graphics

A newly emerging field in statistics is distributional regression, where not only the mean but each parameter of a response distribution can be modeled using a set of predictors (Klein et al., 2015). Notable frameworks called Generalized Additive Models for Location, Scale and Shape and Bayesian Additive Models for Location, Scale and Shape were introduced by Rigby and Stasinopoulos (2001) in the form of a frequentist perspective and Umlauf, Klein, and Zeileis (2017) using a Bayesian approach, respectively. In distributional regression models, the interpretation of covariate effects on response moments and the expected conditional response distribution is more difficult than with traditional methods such as Ordinary Least Squares (OLS) or Generalized Linear Models (GLM) since the moments of a distribution often do not directly equate the modeled parameters but are rather a combination of them with a varying degree of complexity. This talk will introduce a framework for the visualization of distributional regression models fitted using the bamlss R package (Umlauf et al., 2017) as well as feature an implementation as an R extension titled bamlss.vis. The main goals of this framework are the abilities to: 1. See and compare the expected distribution for chosen sets of covariates 2. View the direct relationship between moments of the response distribution and a chosen explanatory variable, given a set of covariates. Additionally, the user can obtain the code which created the visualizations described above to potentially reproduce them later.

Estimating the maximum possible earthquake magnitude using extreme value methodology: the Groningen case [Tue 15:50]

Speaker: Tom Reynkens (Postdoctoral researcher @ KU Leuven, Belgium) and Jan Beirlant, Andrzej Kijko, John H.J. Einmahl

Tom Reynkens is currently a postdoctoral researcher in actuarial science at KU Leuven, Belgium, and holds a PhD in Mathematics on the subject of extreme value theory from the same university. He is the main author of the ReIns package which complements the book "Reinsurance: actuarial and statistical aspects". Together with the authors of this book, Tom gave workshops for actuaries in Switzerland, Poland and South Africa where he was responsible for the practical sessions with R.

Section: Statistics

The area-characteristic, maximum possible earthquake magnitude is required by the earthquake engineering community, disaster management agencies and the insurance industry. The Gutenberg-Richter law predicts that earthquake magnitudes follow a truncated exponential distribution. In the geophysical literature several parametric and non-parametric estimation procedures were proposed. Estimation of the maximum possible earthquake magnitude is of course an extreme value problem to which classical methods for endpoint estimation could be applied. However, recent extreme value theory (EVT) methods for truncated tails at high levels constitute a more appropriate setting for this estimation problem. In this talk, we use the methods from the extreme value and geophysical literature to estimate the maximum possible magnitude of earthquakes induced by gas extraction in the Groningen province of the Netherlands. Moreover, a **Shiny** application has been developed (https://treynkens.shinyapps.io/Groningen_app/) to let users perform the analysis where they can e.g. change the considered time period.

All considered EVT-based endpoint estimators are implemented in the **ReIns** package which complements the book "Reinsurance: actuarial and statistical aspects". In addition to EVT tools to deal with truncated data, **ReIns** provides implementations of classical EVT plots and estimators, and EVT methods that handle censored data.

R packages: ReIns, shiny

References:

- Albrecher H., Beirlant J., and Teugels J. (2017). *Reinsurance: Actuarial and Statistical Aspects*. Wiley, Chichester.
- Beirlant J., Kijko A., Reynkens T., and Einmahl J.H.J. (2018). “Estimating the Maximum Possible Earthquake Magnitude Using Extreme Value Methodology: The Groningen Case.” *Natural Hazards*, accepted. https://doi.org/10.1007/s11069-017-3162-2.
- Reynkens T. and Verbelen R. (2017). *ReIns: Functions from "Reinsurance: Actuarial and Statistical Aspects"*. https://CRAN.R-project.org/package=ReIns

Taking the Bayesian Leap [Tue 16:10]

Speaker: Andrew Collier (Data Scientist @ Exegetic Analytics, South Africa)

Consulting Data Scientist working on various interesting and challenging projects in industry and academia. Spends a lot of time thinking about R, Python, SQL and Cloud Computing. Always trying to find ways to explain these things in simple terms.

Section: Machine Learning, Statistics

What are the minimum things you need to know to start applying Bayesian techniques in R? This talk will provide an entry level discussion covering the following topics:

- What can Bayes do for me? (A brief introduction to Bayesian methods)
- What is Stan? (Writing models in Stan)
- Using Stan in R. (Using the rstan package)

The talk will be peppered with useful tips for dealing with the initial challenges of using Stan.

R packages: rstan

Modelling Item Worth Based on Rankings [Tue 16:30]

Speaker: Heather Turner (Consultant @ Freelance, United Kingdom) and Ioannis Kosmidis, David Firth, Jacob van Etten

Heather is a freelance consultant providing support to clients across sectors in statistics and R programming. She is an Associate Fellow of the University of Warwick and a member-at-large on the board of the R Foundation, as part of which she chairs the Forwards task force for women and under-represented groups.

Section: Statistics

Given a set of rankings, for example the finishing orders in a set of races or consumer preferences from market research, we are often interested in estimating the underlying worth of the items being ranked. A well-established approach is to fit a Plackett-Luce model, jointly attributed to Plackett (1975) and Luce (1959).

This talk introduces a new package, **PlackettLuce**, for fitting the Plackett-Luce model to rankings data. In contrast to some of the other packages or approaches to fitting the model in R, **PlackettLuce** can efficiently handle all of the following:

- ties in the rankings;
- rankings of only a subset items;
- clustered rankings (rankings for distinct sets of items), and
- one or more items that are always ranked first or last in their rankings.

PlackettLuce also enables inference on the worth estimates by providing model-based standard errors and a method for obtaining quasi-standard errors, which don't depend on the identifiability constraints.

Another key feature of the package is a method for fitting Plackett-Luce trees, which will be illustrated in a novel application to identify growing conditions for which trial varieties of bean plants have significantly different worth values.

References:

- Luce, R. Duncan (1959): Individual Choice Behavior: A Theoretical Analysis. https://doi.org/10.2307/2282347.
- Plackett, Robert L. (1975): The Analysis of Permutations. Appl. Statist 24 (2):193–202. https://doi.org/10.2307/2346567.
- Heather Turner, Ioannis Kosmidis and David Firth (2018): PlackettLuce: Plackett-Luce Models for Rankings. R package version 0.2-2. https://CRAN.R-project.org/package=PlackettLuce

---

Interactivity meets Reproducibility: the ideal way of doing RNA-seq analysis [Tue 16:50]

Speaker: Federico Marini (Postdoctoral Fellow @ Center for Thrombosis and Hemostasis Mainz (CTH); Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), Germany) and Harald Binder

After obtaining my PhD in Biostatistics/Bioinformatics last July, I intend to continue the development of methods and software as a postdoctoral fellow. Being a Virchow Fellow at the Center for Thrombosis and Hemostasis, I can sit at the interface with many other disciplines and focus on translational bioinformatics. With platelets transcriptomics as a use case, I plan to develop packages and web applications for interactive and reproducible research in RNA-seq analysis and other -omics data.

Section: Bioinformatics, Reproducible Research, Web Apps

Next generation sequencing technologies, such as RNA-Seq, generate tens of millions of reads to quantify the expression levels of the features of interest. A wide number and variety of software packages have been developed for accommodating the needs of the researchers, mostly in the R/Bioconductor framework. Many of these focus on the identification of differentially expressed (DE) genes (DESeq2, edgeR,) to discover quantitative changes between experimental groups, while other address alternative splicing, discovery of novel transcripts, or RNA editing. Moreover, Exploratory Data Analysis is a common step to all these workflows, and despite its importance for generating highly reliable results, it is often neglected, as many of the steps involved might require a considerable proficiency of the user in the programming languages. Principal Components Analysis (PCA) is used often to obtain a dimension-reduced overview of the data.

Our proposal addresses the two steps of Exploratory Data Analysis and Differential Expression analysis with two different Bioconductor packages, pcaExplorer and ideal. We developed web applications in the Shiny framework also including support for reproducible analyses, thanks to an embedded text editor and a template report, to seamlessly generate HTML documents as a result of the user's exploration.

This solution, which we also outlined in (Marini and Binder 2016), serves as a practical proof of principle of integrating the essential features of interactivity and reproducibility in the same tool, fitting both the needs of life scientists and experienced analyists, thus making our packages good candidates to become companion tools for each RNA-Seq analysis, and also serving as a base for the development of other applications (e.g. for single cell RNA-seq).

References: Marini and Binder 2016 - Development of Applications for Interactive and Reproducible Research: a Case Study. Genomics and Computational Biology

---

Data Culture in Practice [Wed 9:50]

Speaker: Noa Tamir (Data Science Team Lead @ Babbel, Germany)

Noa manages a team of data scientists at Babbel and co-orginses R-Ladies Berlin. Her role at Babbel is to break down technical and conceptual complexities to improve the user experience of the world's first language learning app. Previously she worked as a data science technical lead at King on games like Candy Crush. Noa is passionate about making data science & analytics approachable, effective and inclusive.

Section: Businesses, Community, Teaching

Every company wants to be data driven, but not everyone in the organization is going to be trained in data science and analytics, just as not everyone codes, designs or writes contracts. So how do we get to a point where we have a data driven organizational culture without teaching everybody statistics and R? In this talk Noa will present some strategies and examples to get you started on build you own data culture.

---

radii.defer: Deferred execution of nested functions [Wed 10:10]

Speaker: Dénes Tóth (owner @ Kogentum Ltd., Hungary) and Balázs Urbán

Denes Toth is a freelanceR with strong academic background formerly working on the field of test development and cognitive neuroscience. At present he participates in automatized data analytic, visualization and optimization projects as lead developer.

Section: Use-cases, Web Apps

Suppose you develop a graphical user interface which allows the user to create interactive, multi-facetted and/or multi-page charts which visualize the results of arbitrary statistical analyses. To create the charts, you query the data from a database, perform statistical computation on it, repeat it for several data chunks, and plot the results. Suppose the main performance bottleneck is the download of data. Also suppose that the user can modify the parameters of a chart several times before actually requesting the final plot. Welcome to the world of *radii.defer*, an **R** package by which you can create a nested chain of arbitrary (unevaluated) **R** functions. A *Defer* object contains a function and its arguments, and a caching scheme. The arguments of a *Defer* function may be not only standard **R** objects but other (usually unevaluated) *Defer* objects as well, and the arguments can be updated whenever you wish. When a *Defer* object is executed, it does not execute any of its *Defer* arguments unless it is absolutely necessary. Furthemore, *radii.defer* plays well with the *future* package allowing asynchronous execution which does not block the main **R** thread. In the talk we will present the main features of the package, and as an example, we will illustrate its use in a complex, cloud-based *Shiny* application which motivated the development of the package.

Using R to Build a Data Science Team [Wed 10:10]

Speaker: Aimee Gott (Senior Data Science Consultant @ Mango Solutions, United Kingdom)

Aimee is a senior consultant in the team at Mango Solutions where she is heavily involved in the development and delivery of Data Science training. She has worked with a number of customers across industry to make R and data science a part of their day to day work.

Section: Teaching, Use-cases

The rise of data science has brought with it a high demand for data scientists. More companies than ever before are looking for the broad skillset of THE data scientist –or the ever out of reach unicorn-- but in reality, most data scientists are moving to the field from diverse backgrounds such as statistics, physics, chemistry, engineering and computer science. So, how does a company ensure their data science team has the right mix of skills to achieve their business goals? Ensuring the statistician has the software engineering skills they need and the computer scientist can understand advanced analytics are just the tip of the iceberg.

At Mango, we have been working with a number of companies, across different industries, to help them build their in-house data science capability - from completely new teams to teams already in place, looking to understand how to do data science in practice.

So how do we build on the skills teams with diverse backgrounds already have? How can we use the capability of R to grow strong data science teams? In this talk we will discuss some of the approaches that we have taken at Mango to educate teams in everything from software development best practices to the intricacies of machine learning – all without leaving R!

Predicting Cryptocurrencies Time–Series with the eDMA package [Wed 11:00]

Speaker: Leopoldo Catania (Assistant Professor @ Aarhus BBS, Denmark)

I am Assistant Professor in Econometrics at Aarhus University and CREATES. My works concern the development of univariate and multivariate econometrics models applied to quantitative risk management, time-varying dependence and volatility. I believe in open source software. I am the writer and maintainer of three R packages: MCS, GAS and eDMA. I'm also an author of the MSGARCH package for R. GAS and eDMA are detailed in two articles forthcoming in the Journal of Statistical Software.

Section: Finance, Reproducible Research, Statistics, Time-series

Cryptocurrencies have recently gained a lot of interest from investors, central banks and governments worldwide. The lack of any form of political regulation and their market far from being “efficient”, requires new forms of regulation in the near future. In this paper we study the predictability of cryptocurrencies time–series. We perform a Dynamic Model Averaging analysis using many predictors leading us to more than 31 millions of Dynamic Linear Models estimated at each point in time during the forecast horizon. The whole analysis and the associated huge computational complexity relies on the eDMA package of Catania and Nonejad [2017, Journal Statistical Software, (in press)] available in the R environment. Results are reported for the four most representative coins and indicate different levels of predictability depending on: i) the time

period, ii) forecast horizon, iii) and type of coin. A Portfolio optimization application shows that trading strategies which incorporate cryptocurrencies are profitable even after the inclusion of transaction costs.

R packages: eDMA

References: Catania and Nonejad (2018): Dynamic Model Averaging for Practitioners in Economics and Finance: The eDMA Package

---

Markov-Switching GARCH Models in R: The MSGARCH Package [Wed 11:20]

Speaker: David Ardia (Assistant Professor of Finance @ University of Neuchâtel, Switzerland) and Bluteau, K., Boudt, K., Catania, L., Trottier, D.-A.

David Ardia is Professor of Quantitative Finance and Head of the Master of Science in Finance at the University of Neuchâtel, Switzerland. He is the author of several scientific articles and open-source packages such as 'DEoptim', 'MSGARCH', 'PeerPerformance', and 'sentometrics'.

Section: Finance, Time-series

We describe the package MSGARCH, which implements Markov-switching GARCH models in R with efficient C object-oriented programming. Markov-switching GARCH models have become popular methods to account for regime changes in the conditional variance dynamics of time series. The package MSGARCH allows the user to perform simulations as well as Maximum Likelihood and MCMC/Bayesian estimations of a very large class of Markov-switching GARCH-type models. The package also provides methods to make single-step and multi-step ahead forecasts of the complete conditional density of the variable of interest. Risk management tools to estimate conditional volatility, Value-at-Risk and Expected Shortfall are also available. We illustrate the broad functionality of the MSGARCH package using exchange rate and stock market return data.

R packages: MSGARCH

References: Ardia et al. (2017) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2845809

Speeding up R with Parallel Programming in the Cloud [Wed 11:40]

Speaker: David Smith (Cloud Developer Advocate @ Microsoft, United States)

David Smith is a developer advocate at Microsoft, with a focus on data science and the R community. With a background in Statistics, he writes regularly about applications of R at the Revolutions blog (blog.revolutionanalytics.com), and is a co-author of "Introduction to R", the R manual. Follow David on Twitter as @revodavid.

Section: HPC, Machine Learning

There are many common workloads in R that are "embarrassingly parallel": group-by analyses, simulations, and cross-validation of models are just a few examples. In this talk I'll describe the doAzureParallel package, a backend to the "foreach" package that automates the process of spawning a cluster of virtual machines in the Azure cloud to process iterations in parallel. This

will include an example of optimizing hyperparameters for a predictive model using the "caret" package.

R packages: doAzureparallel, foreach, caret

Time series modeling of plant protection products in aquatic systems in R [W11:40]

Speaker: Andreas Scharmüller (PhD student @ University Koblenz-Landau, Germany) and Mira Kattwinkel, Ralf B. Schäfer

Master in Freshwater Ecology from the university of Natural Resources in Vienna, Austria. Currently enrolled in a PhD program at the University Koblenz-Landau, Germany. PhD focus: Big data and statistics in the field of water chemistry. Avid R and open source software user.

Section: Big Data, Statistics, Time-series, Spatial, Databases, Use-cases

Plant protection products (PPP) including fungicides, herbicides and insecticides are amongst other factors an important stressor in aquatic ecosystems. We analyzed their seasonal occurrence patterns and possible related adverse effects on aquatic organisms in small running waters. Therefore more than 450.000 water chemistry samples had been collected by German federal agencies between 2005 and 2015. The data were processed in R, mostly by using the data.table-package, before they were uploaded into a relational PostgreSQL data base. For subsequent analysis of seasonal occurrence patterns of PPP in streams we used Generalized Additive Models (GAM) from the mgcv package.

We hypothesized (i) to find increased in-stream occurrences of PPP during months of increased application. Likewise we hypothesized (ii) adverse effects on aquatic organisms to peak during, or shortly after periods of application. Furthermore we hypothesized (iii) to find a long-term decrease in occurrences of compounds, whose authorization has expired during or shortly before the monitoring period.

We defined the occurrence of PPP as the number of detections per compound per month and sampling site. In order to quantify adverse effects of single PPP on different aquatic organism groups Toxic Unit (TU) metrics were used. They are calculated by dividing in-stream compound concentrations by endpoints of experimental toxicity tests. Such test data were obtained through large publicly accessible data bases. Additional data on chemicals were collected using the webchem package and data on organisms were accessed through the taxize package.

We were able to trace seasonal patterns of PPP in small streams and draw conclusions on the effect on aquatic communities therein. With the help of the mentioned R packages and data base systems we were able to establish fast and reproducible procedures for the analysis of large environmental data sets.

R packages: data.table, taxize, webchem, mgcv, ggplot2

Exploiting Spark for high-performance scalable data engineering and data-science on Microsoft Azure [Wed 12:00]

Speaker: Simon Field (Azure Data Specialist @ Microsoft, United Kingdom)

Simon has 20+ years experience working with organisations across EMEA in the field of Data Warehousing, Business Intelligence, Big Data and Advanced Analytics, assisting them in resolving the many data integration, consolidation, migration, data-model and governance issues. Simon has worked extensively with large-scale data processing technologies from Teradata, IBM, HP, Microsoft and joined Microsoft through the Revolution Analytics acquisition in 2015.

Section: Big Data, HPC, Data Munging, Dashboards, Databases, Machine Learning

Spark has become an incredibly popular data processing engine to support scalable data science. There are several R packages that abstract Spark processing from R. Microsoft Azure provides several different options for working R on Spark depending optimised for different needs. In this session we will outline the options and best-practices regarding which one to use when. We will provide some examples of using different packages with Spark in Azure.

R packages: SparkR, SparklyR, dplyr, rSparkling, RevoScaleR

Predicting the winner of the 2018 FIFA World Cup predictions [Wed 12:00]

Speaker: Claus Thorn Ekstrøm (Professor @ Biostatistics, University of Copenhagen, Denmark)

Claus Thorn Ekstrøm is professor at the section of biostatistics, University of Copenhagen. He is the creator and contributor to a number of R packages (eg, dataMaid, MESS, MethComp, SuperRanker) and is the author of "The R Primer" book.

Section: Statistics, Use-cases

The 2018 FIFA World Cup will be played in Russia this summer. The World Cup is the source of almost endless predictions about the winner and the results of the individual matches.

Different statistical models form the basis for predicting the result of individual matches. We present an R framework for comparing different prediction models and for comparing predictions about the World Cup results. Everyone is encouraged to contribute their own function to make predictions for the result of the 2018 World Cup.

Each contributer will be shown how to provide two functions: a function that predicts the final score for a match between two teams with skill levels a_1 and a_2, and a function that updates the skill levels based on the results of a given match. By supplying these two functions to the R framework the prediction results can be compared and the winner of the best football predictor can be found when the 2018 World Cup finishes.

Wikidata Concepts Monitor: R in action across Big Wikidata [Wed 12:20]

Speaker: Goran Milovanović (Data Scientist @ Wikimedia Foundation Deutschland, Serbia)

Goran S. Milovanović is a Data Scientist with Wikimedia Deutschland. His current work is focused on the distributive semantics of Wikidata usage across Wikimedia projects (Wikipedia, Wikimedia Commons, etc). He is the principal developer of the Wikidata Concepts Monitor, a system that relies on R, SPARQL, HiveQL, SQL, and Shiny to track and analyze the data on Wikidata item usage.

Goran holds a PhD in Psychology and has entered Data Science following an academic career in Cognitive Science.

Section: Dashboards, Big Data, Community, Machine Learning, Social Sciences, Statistics, Use-cases

Wikidata, one of the most prominent Wikimedia projects nowadays, presents a formal ontology that adheres to the RDF standard. All Wikimedia projects (Wikipedia, Wiktionary, Wikivoyage, Wikimedia commons, etc) can reach out to and make use of Wikidata by relying on the features of MediaWiki and Wikibase. Currently, more than 800 Wikimedia websites - some of which are found among the most dynamic places online at all - maintain a client-side Wikidata usage tracking. The resulting data sets present a challenging Big Data problem that calls for SQL transfers to Hadoop via Apache Sqoop, massive HiveQL ETL operations and data wrangling in R just in order to enable for the tracking and analytics of individual Wikidata item usage statistics across the Wikimedia projects.

The Wikidata Concepts Monitor (WDCM) system was developed in R to orchestrate various technologies (SPARQL to access Wikidata, SQL, Apache Sqoop, Hive, and Spark) in order to provide for Wikidata usage tracking and analytics. The WDCM then takes a step beyond and performs machine learning (LDA, t-SNE) across the matrices that encompass Wikidata items and Wikimedia projects in order to gain insight into the distributive semantics (i.e. topic models, dimensionality reduction for similarity maps) of Wikidata usage. The system encompasses a set of Shiny dashboards where the results of the WDCM statistical machinery are visualized and tools for analytical work provided in order to help the understanding of the utterly complex social and behavioral phenomenon of Wikidata usage.

References:

- WDCM Portal: http://wdcm.wmflabs.org/
- Wikitech: Wikidata Concepts Monitor (WDCM). The WDCM technical documentation. URL: https://wikitech.wikimedia.org/wiki/Wikidata_Concepts_Monitor
- Wikidata Concepts Monitor (WDCM). The WDCM Wikidata wiki project page. URL: https://www.wikidata.org/wiki/Wikidata:Wikidata_Concepts_Monitor

---

validatetools: resolve and simplify contradictive or redundant data validation rules. [Wed 14:50]

Speaker: Edwin de Jonge (Statistical consultant / Methodologist @ Statistics Netherlands / CBS, Netherlands)

Edwin de Jonge is a statistical consultant at Statistics Netherlands. He is (co)author of several R-packages including ffbase, whisker, daff, tableplot and docopt. His joint work with Mark van der Loo includes building R packages for structured data-cleaning: automatically detecting errors (errorlocate), correcting errors (deductive, deducorrect) and validating data (validate and validatetools). This work has resulted into the book "Statistical data cleaning with Applications in R" (2018).

Section: Reproducible Research, Businesses, Data Munging, Statistics

Using many rules to check the validity of your data often results in a large collection of rules that may generate unwanted interactions. This may seem obvious, but often happens (unknowningly). validatetools helps detecting and resolving accumulated redundant or (partially) contradictory rules. We will demonstrate the included methods as well as describe the inner workings of detecting and resolving the issues with Mixed Integer Programming.

R packages: validate, validatetools

References: Statistical Data Cleaning with Applications in R, Wiley, 2018, M. van der Loo, E. de Jonge

---

What software engineers can teach to data scientists: code safety and ... with automatic tests [Wed 15:40]

Speaker: Andrea Melloncelli (IT manager @ Quantide SRL, Italy)

I'm a Physics by tuition. I specialize in R, in particular I take care of the code and the system part. My main focus is importing modern development methodologies in the Data Science word using R and unix. I work with Quantide as consultant and R trainer. In my job, my first concern is providing and take care of infrastructure that uses R. I have often worked providing a big data infrastructure of Hadoop with Spark. My second duty is to produce well coded dashboards with RMarkdown and Shiny.

The development through the tests is a well-established methodology in the field of software engineering for several reasons:

First of all, it validates the real functionality of what we are developing, and avoid regressions in the code functionality: in fact, the automatic tests are run frequently and every time that a new portion of code is added, they check that everything is still working as expected (and eventually point out where it's not going so) Secondly, tested code enables refactoring: it means that using tests you can modify the working code to improve its readability or abstraction without adding further functionality. This provides the re-usability of that code. Furthermore tests describe the code: clearly written tests in fact provide a working documentation of the low-level functionalities of the code being written Finally, Test Driven Development is a programming methodology that starts with the tests to describe the design of the code before it validates its functionality.

However, this methodology has not yet established itself in the world of data science, even if RStudio and several others have already created tools for supporting the package development with automatic tests. Since validating code and making refactoring are very important qualities in data science, I would like to show within this talk how to use tests in the development of a data science project and what are the great advantages that it can lead, in term of clarity of the code, robustness and safeness of code changings.

Outline:

1. Why do tests
2. The Testthat package (https://github.com/r-lib/testthat)
3. Test Driven Development (https://en.wikipedia.org/wiki/Test-driven_development)

4. The importance of refactoring
5. Tests as working documentation
6. Shiny tests

Geocomputation for Active transport planning: a case study of cycle network design [Wed 16:00]

Speaker: Robin Lovelace (Research Fellow in Transport and Big Data @ University of Leeds, United Kingdom)

Robin is a University Academic Fellow at the Leeds Institute for Transport Studies (ITS). Robin has published books on Microsimulation with R (2015) Efficient R Programming (2016) and Geocomputation with R (forthcoming). His research also has real-world impact. Robin is Lead Developer of the Propensity to Cycle Tool, which is being used to develop cycling networks across England (see www.pct.bike) and the stplanr package. Twitter: @robinlovelace.

Section: Graphics, Infrastructure, Reproducible Research, Social Sciences, Spatial, Use-cases

Although it has academic origins, R is now widely used for many real-world applications including finance, epidemiology, store location analysis and conservation. Indications suggest that the language continues to gain adoption in other areas due to its advantages of cost, reproducibility and flexibility over proprietary products. This talk will explore the potential uptake of R for transport planning, a domain in which powerful yet accessible scriptable languages have much to offer. Transport planning relies on data analysis, a range of spatial and temporal data forms, and visualisation, areas that R excels in, especially with add-on packages such as **sf**. To illustrate the point the talk will describe work commissioned by the UK's Department for Transport to develop a Propensity to Cycle Tool (PCT). The PCT is now being used around England to help design strategic cycle networks and improve the effectiveness of precious public investment in transport infrastructure (Lovelace et al. 2017). Based on the experience developing the PCT we will discuss the limitations of R as a tool for transport planning, and the potential of recent developments in packages such as **stplanr** and **dodgr** packages to address them (Lovelace and Ellison, under review). The talk will conclude by outlining reasons for transport planning authorities to demand the use of open source software and reproducibility to ensure democratic accountability and the cost-effective use of public funds.

R packages: sf, stplanr, dodgr

References:

- Lovelace, R., Goodman, A., Aldred, R., Berkoff, N., Abbas, A., Woodcock, J., 2017. The Propensity to Cycle Tool: An open source online system for sustainable transport planning. Journal of Transport and Land Use 10. https://doi.org/10.5198/jtlu.2016.862
- Lovelace, R., Ellison, R., under review stplanr: A Package for Transport Planning. The R Journal.

Know your R usage workflow to handle reproducibility challenges [Wed 16:00]

Speaker: Wit Jakuczun (Onwer / Senior Data Scientist @ WLOG Solutions, Poland)

Founder and owner of a consulting company WLOG Solutions that is a strategic partner in R based large scale analytics. In the company he is responsible for translating customer needs to mathematics and vice versa. Have 12+ years of experience in delivering predictive, simulation and optimization models.

Section: Reproducible Research, Businesses, Use-cases, Infrastructure

R is used in a vast ways. From pure ad-hoc by hobbysts to an organized and structured way in an enterprise. Each way of R usage brings different reproducibility challenges. Going through range of typical workflows we will show that understanding reproducibility must start with understanding your workflow. Presenting workflows we will show how we deal reproducibiilty challenges with open-source R Suite (http://rsuite.io) solution developed by us to support our large scale R development.

R packages: rsuite, packrat, miniCRAN, checkpoint

---

openSTARS: prepare GIS data for regression analysis on stream networks [W16:20]

Speaker: Mira Kattwinkel (Scientist @ University of Koblenz - Landau, Germany) and Ralf Schäfer

Topics: Assessing anthropogenic stressors and understanding and predicting their ecological effects on populations, communities and biodiversity in general Tools: Geostatistics on stream networks, GIS, dynamic population and community modelling, Bayesian parameter inference

Section: Spatial, Statistics

Statistical data analysis on stream networks needs to take into account the spatial correlation of measurements sampled at connected stream sites. The *SSN* package for Spatial Statistical Modeling on Stream Networks provides tools to fit glms with spatial autocorrelation. However, so far the tool 'STARS' provided for GIS data preparation is based on the commercial software ArcGIS. *openSTARS* offers an alternative in **R** based on open source GRASS GIS and R packages for spatial data. Thereby it merges the whole work flow of data preparation and statistical analysis on stream networks in one software. *openSTARS* provides functions to derive stream networks from a digital elevation model, detect and correct non-dichotomous junctions, which cannot be handled by *SSN*, derive stream catchments and intersect them with land use or other GIS data as potential explanatory variables. Additionally, locations for model predictions can be generated along the stream network. We exemplify the application of *openSTARS* for water quality monitoring data.

R packages: openSTARS, SSN, rgrass7

Fitting Humans Stories in List Columns: Cases From an Online Recruitment Platform [Wed 16:20]

Speaker: Omayma Said (Data Scientist @ WUZZUF, Egypt)

I work as a Data scientist at Wuzzuf where we develop online career services including recruitment, job boards, assessments and more. Prior to that I obtained my Master's degree in "System Engineering

and Engineering Management" with focus on electronics.I is most interested in digging deep into data, finding new insights and communicating results effectively.

Section: Businesses, Use-cases, Statistics, Dashboards

When designing products and taking data-informed decisions, the stories of individuals are often obscured by the tendency to focus solely on aggregated data and models. This talk will focus on the paradigm I embrace, and the tools I build on top of the tidyverse packages, to combat this tendency.

At WUZZUF, we provide online career services including job boards, applicant tracking systems, and assessments. I'll talk about a situation where we observed a low supply of talent for senior software developers jobs, outlining my approach to testing hypotheses about the root causes of the supply/demand imbalance. Rather than reacting to the problem by looking at high-level metrics and jumping to conclusions about the root causes *(seen as an acquisition issue by marketers or a recommendation issue by engineers)*, I started a deeper analysis, oriented by job hunting and hiring as human experiences.

The tidyverse tools enabled me to see the stories of individuals while maintaining context through the patterns of groups. I'll highlight how combining tidyr, purrr, list columns, and 'shiny' let me go beyond the preliminary assumptions and reveal deeper issues. At the end, this led to actionable insights regarding our search/recommendation, talent acquisition and job postings.

Machine Learning (ranger package) as a framework for spatial and spatiotemporal prediction [Wed 16:40]

Speaker: Tomislav Hengl (Senior researcher @ ISRIC - World Soil Information, Wageningen University, Netherlands) and Tom Hengl, Marvin Wright, Madlene Nussbaum

Tom has over 20 years of experience in doing research for the purpose of mapping and modelling environmental data at regional and global scales. Tom leads development of mobile phone and web apps for serving and sharing soil data from professional users to non-specialists. Current areas of interest: Machine Learning and BigGeodata, automated mapping, spatiotemporal prediction based on machine learning.

Section: Machine Learning, Statistics, Spatial, Big Data

We describe a framework to use Random Forest to generate spatial and spatiotemporal predictions i.e. as an alternative to model-based geostatistics. Spatial auto-correlation, especially if still existent in the cross-validation residuals, indicates that the predictions are maybe biased, and this is sub-optimal. To account for this, we use Random Forest (as implemented in the ranger package) in combination with geographical distances to sampling locations to fit models and predict values. We describe eight typical situations of interest to spatial prediction applications: (1) prediction of 2D continuous variable without any covariates, (2) prediction of 2D variable with covariates, (3) prediction of binomial variable, (4) prediction of categorical variable, (5) prediction of variables with extreme values, (6) weighted regression, (7) predictions of multivariate problems, and (8) prediction of spatiotemporal variable. Our results indicate that RFsp can produce comparable results to model-based geostatistics. The advantage of RFsp over model-based

geostatistics is that RFsp requires much less statistical assumptions and is easier to automate (and scale up through parallelization). On the other hand, computational intensity of RFsp can blow up as the number of training points and covariates increases. RFsp is still an experimental method and application with large data sets (>>200 spatial locations) is probably not recommended. To download all data sets and more detail code examples please refer to https://github.com/thengl/GeoMLA/

R packages: ranger, GSIF, caret, rgdal, geoR, gstat

References:

- Hengl, T., Nussbaum, M., Wright, M. and Heuvelink, G.B.M., 2018? Random Forest as a Generic Framework for Predictive Modeling of Spatial and Spatio-temporal Variables, PeerJ (submitted).
- Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1), 1–17.

Asymptotic Powers of Selected ANOVA Tests in Generalized Linear Models [Wed 16:40]

Speaker: Zuzana Hubnerova (associate professor on leave @ Brno University of Technology, Faculty of Mechanical Engineering, Czech Republic)

Zuzana was teaching probability and statistics at the Brno University of Technology, CZ but is now living in Budapest. She focused her research on properties of generalized linear models. Moreover, Zuzana enjoys applying statistical methods in various fields such as environmetrics, econometrics, psychometrics.

Section: Statistics

This talk aims at tests of equality of expected values in balanced one-way ANOVA type generalized linear models based on deviance or score statistic. As the power of these tests cannot be derived analyticaly, their asymptotic approximation is derived. Presented R function power.anova.glm.test allows a user to compute the approximation of power of tests or determine parameters to obtain target power.

Lightning Talks

A variation of the pecha kucha and ignite formats with fixed number of 15 slides, each shown for 20 seconds, in an auditorium, followed by 1 quick question if time allows.

The Big Connection - using R with big data [Tue 12:10]

Speaker: Bence Arató (director @ BI Consulting Kft., Hungary)

I've been in the data industry for 20+ year, mostly working as an analyst, advisor and architect. I often design data analytics systems for my clients and also teach numerous data-related workshops and classes.

Section: Big Data

This talk will show a few ways R can be used to access and process data stored in big data-type backends such as *Apache Impala* and *Apache Spark*.

R packages: dplyr, implyr, sparklyr, SparkR

An R package for statistical tools with big matrices stored on disk [Tue 12:15]

Speaker: Florian Privé (PhD student @ Univ. Grenoble Alpes, France)

I am a PhD student in predictive human genetics, fond of Data Science and an R(cpp) enthusiast. I am also the founder and co-organizer of the Grenoble R user group.

Section: Big Data, Bioinformatics, Statistics, Machine Learning

The R package bigstatsr provides functions for fast statistical analysis of large-scale data encoded as matrices. It provides for instance matrix operations, Principal Component Analysis (PCA), sparse multivariate linear supervised models, utility functions and much more. Package bigstatsr can handle matrices that are too large to fit in memory thanks to memory-mapping to binary files on disk. This is very similar to the format big.matrix provided by the R package bigmemory.

RcppGreedySetCover: Scalable Set Cover [Tue 12:20]

Speaker: Matthias Kaeding (researcher @ RWI - Leibniz Institute for Economic Research, Germany)

Matthias is a phd student at the research data centre at the RWI - Leibniz Institute for Economic Research, Essen. His thesis deals with the analysis of large data sets.

Section: Big Data, Data Munging, Use-cases

The set cover problem is of fundamental importance in the field of approximation algorithms: Given a collection of sets, find the smallest sub-collection, so that all elements from a universe are covered. A diverse range of real world problems can be represented as set cover instance such as location-allocation, shift planning or virus detection. An optimal solution to the problem via linear programming is available. Due to the computational costs involved, this is not a feasible solution for large problems. A quick approximation is given by the greedy algorithm.

This talk introduces RcppGreeySetCover, an implementation of the greedy set cover algorithm using Rcpp. The implementation is fast due to the reliance on efficient data structures made available by the Boost headers and the C++ 11 standard library. Preprocessing of input data is done efficiently via the data.table package. Input and output data is a tidy two column data.frame. As a case study, we apply the package on a large scale (>= 100 million rows) hospital placement problem, which initially motivated the creation of the package.

Make R elastic [Tue 12:25]

Speaker: Emil Lykke Jensen (CEO @ MediaLytic, Denmark)

Developed an online media monitoring and analysis platform running on Shiny and Elastic Search serving high end governmental and industry organizational users in Denmark.

Section: Databases

This talk will focus on how to use Elastic Search with R and will consists of three parts. First it will give you a short introduction to Elastic Search: what is it, why/why not use it and how do you use it with R (introducing the elastic package). Secondly it will show you just how fast Elastic Search is in comparison to SQL (specifically MySQL) and it will outline the syntax for querying with Elastic Search. Thirdly it will give you examples on how to use Elastic Searchs buildt-in aggregations using the Danish Central Business Register database.

Modelling Field Operation Capacity using Generalised Additive Model and Random Forest [Tue 16:30]

Speaker: Timothy Wong (Senior Data Scientist @ Centrica plc (British Gas), United Kingdom) and Terry Phipps; Matthew Pearce

Timothy currently serves as Senior Data Scientist at Centrica plc, the largest energy service provider in the United Kingdom which is commonly-known as British Gas. He specialises in machine learning and statistics. He is also an R package author (CRAN contributor).

Section: Big Data, Businesses, Machine Learning, Statistics, Time-series, Use-cases

In any large-scale customer-facing business, accurately predicting demand ahead of time is of paramount importance*. Workforce capacity can be flexibly scheduled at local area accordingly. In this way, we can ensure having sufficient workforce to meet volatile demand.

In this case study, we focus on the gas boiler repairing field operation in the UK. We have developed a prototype capacity forecasting procedure which uses a mixture of machine learning techniques to achieve its goal. Firstly, it uses Generalised Additive Model (GAM) approach to estimate the number of incoming work requests. It takes into account the non-linear effects of multiple predictor variables. The next stage uses a large random forest (RF) to estimate the expected number of appointments for each work request by feeding in various ordinal and categorical inputs. At this stage, the size of the training set is considerable large (approx. 250 million rows) and does not fully-fit in memory. In light of this, the random forest model was trained in chunks / parallel to enhance computational performance.

Once all previous steps have been completed, probabilistic input such as the ECMWF Ensemble weather forecast to give a view of all predicted scenarios. Planning operators can then use the model output to fine-tune workforce assignment at the local level to meet changing demand. British Gas operates a service and repair business with more than 6,000 qualified engineers ready to serve customers who are urgently in need across the country.

IRT and beyond - what to do when you want to modify a model, but the package you use do not let you? [Tue 16:35]

Speaker: Krzysztof Jędrzejewski (Data Scientist @ Pearson, Poland)

Data scientist working with R for over 3 years. Also a lecturer at postgraduate studies in Big Data.

Section: Machine Learning, Statistics, Use-cases

Item Response Theory (IRT) is a fundamental concept used in analysing educational data. IRT modelling is used for estimating the difficulty of quiz questions, and the skill levels of learners. However, it can also be applied to the variety of other areas, e.g. clinical trials [1] or influence of genes on some medical conditions [2].

The simplest way to implement IRT modelling in your R workflow is to use one of the dedicated R packages. However, this is also the least elastic approach because these packages either do not allow to introduce any additional variables, or allow it only to a limited degree.

A more flexible way of customising an IRT-based model is to treat it as logistic regression with random effects. When more complex models need to be fit, one may need to use Bayesian modelling software such as Stan, or estimate model parameters with gradient descent using libraries such as TensorFlow.

In this talk, I'll show how to estimate IRT model parameters in R using each of these approaches. I'll also highlight the advantages and disadvantages of each method. Most of the approaches that I'll show may also be applied to other areas, not only IRT.

R packages: TAM, lme4, rstan, tensorflow

References:

- [1] CHAKRAVARTY, Eliza F.; BJORNER, Jakob B.; FRIES, James F. Improving patient reported outcomes using item response theory and computerized adaptive testing. The Journal of rheumatology, 2007, 34.6: 1426-1431.
- [2] TAVARES, Héliton Ribeiro; ANDRADE, Dalton Francisco de; PEREIRA, Carlos Alberto de Bragança. Detection of determinant genes and diagnostic via item response theory. Genetics and Molecular Biology, 2004, 27.4: 679-685.

Classification and attractiveness evaluation of facial emotions for purposes of plastic surgery using machine-learning methods and R [Tue 16:40]

Speaker: Lubomír Štěpánek (biostatistician, software developer, junior lecturer, PhD candidate @ First Faculty of Medicine, Charles University & Faculty of Biomedical Engineering, Czech Technical University in Prague, Czech Republic) and Pavel Kasal, Jan Měšťák

A Master's student in Statistics. A PhD candidate in Biomedical Informatics focused on medical decision-making systems and facial attractiveness evaluation for purposes of plastic surgery. A medical doctor, former oncologist, and a junior lecturer in Introductory Informatics and R courses.

Section: Machine Learning, Medical / Pharma, Statistics, Graphics

Current studies conclude that human facial attractiveness perception is data-based and irrespective of the perceiver. However, the analyses of associations between facial geometry and its visual impact exceed the power of classical statistical methods. What is more, current plastic surgery deals with aesthetic indications such as an improvement of the attractiveness of a smile or other facial emotions.

In this work, we have applied machine-learning methods and R language to explore how accurate classification of photographed faces into sets of facial emotions (based on Ekman-Friesen FACS scale) is, and, furthermore, which facial emotions are associated with higher level of facial

attractiveness, measured using Likert scale by a board of observers. Facial image data were collected for each of a patient (exposed to an emotion incentive), then processed, landmarked and analysed using R. Bayesian naive classifiers using e1071 package, regression trees via tree and rpart packages and finally neural networks by neuralnet package were learned to allow assigning a new face image data into one of the facial emotions.

Neural networks manifested the highest predictive accuracy of a new face categorization into facial emotions. The geometrical shape of a mouth, then eyebrows and finally eyes affect in descending order an intensity of a classified emotion, as was identified using decision trees.

We performed machine-learning analyses using R to compare which classification method conducts the best prediction accuracy when classifying face images into facial emotions, and to point out which facial emotions and their geometry affect facial attractiveness the most, and therefore should preferentially be addressed within plastic surgeries.

References: Kasal P., Fiala P., Štepánek L. et al. Application of Image Analysis for Clinical Evaluation of Facial Structures. In: Medsoft 2015. (2015), pp. 64–70

The R-Package 'surveysd' [Tue 16:45]

Speaker: Johannes Gussenbauer (Methodologist @ Statistics Austria, Austria) and Alexander Kowarik, Matthias Till

I am a Mathematician at Statistics Austria specialising in estimation and imputation techniques for sampling surveys as well as machine learning. I have been an R-user since my undergraduate studies at university and am using it as the main tool for my analysis.

Section: Statistics, Social Sciences

There is an urgent need for regional indicators, especially on poverty and social exclusion, by policy makers in the EU. Surveys which were designed to estimate these indicators on national level, such as EU-SILC, usually do not provide the required precision for reliable estimates on regional levels like NUTS2 and below.

With the R-Package **surveysd**, we present a package for estimating standard errors for yearly surveys with and without rotating panel design. Using surveys with complex survey designs e.g. multistage sampling design is supported and can be freely defined. The implemented method for standard error estimation uses bootstrapping techniques in combination with multiple consecutive waves of the survey. This leads to a reduction in the standard error, especially for estimates done on a subgroup of the survey data. The package enables the user to estimate point estimates as well as their standard error on arbitrary subgroups of his/her data. Also the applied point estimate can freely be chosen, although some predefined point estimates are already implemented. Finally the results can be visualized in two different ways to gain quick insight in the quality of the estimated results.

An integrated framework in R for textual sentiment time series aggregation and prediction [Tue 16:50]

Speaker: Samuel Borms (Ph.D. student @ Université de Neuchâtel, Switzerland) and David Ardia, Keven Bluteau, Kris Boudt

Samuel Borms spent one year working as a consultant in corporate finance and financial risk management after his studies in Business Economics at the Solvay Brussels School. Currently, he is a Ph.D. student at the Université de Neuchâtel and the Vrije Universiteit Brussel, where he investigates the dynamics and information value of textual sentiment, with a focus on forecasting applications in finance and economics. He is the creator of the sentometrics R package.

Section: Big Data, Text mining, Time-series

This lightning talk's aim is to provide a hands-on introduction to optimized textual sentiment indexation in R using the 'sentometrics' package. Driven by data availability, sentiment analysis is increasingly used to capture the information value within textual data of all sorts. We set forward a methodology that accounts for the metadata of texts and the various degrees of freedom in aggregating sentiment to construct plenty textual sentiment time series. Such time series, or indices, can then be related to other variables given appropriate econometric techniques. We propose to use the elastic net regression as a way to deal with the many (correlated) sentiment indices. Above workflow is fully implemented in the R package 'sentometrics'. The package allows to compute lexicon-based sentiment scores of numerous texts at once, to aggregate the textual sentiment scores into multiple time series, and use these time series to predict any other variable. The talk will consist in a fast-paced but clear example of how to use the R package, from which the methodology becomes evident at the same time. For the example, we take a corpus of texts from two major U.S. journals to obtain daily forecasts of a monthly index of economic policy uncertainty.

Reference: Ardia, D., Bluteau, K., Borms, S., and Boudt, K. (2017). "The R Package sentometrics to Compute, Aggregate and Predict with Textual Sentiment".

Time Series Representations for Better Data Mining [Tue 16:55]

Speaker: Peter Laurinec (PhD. Student @ FIIT STU, Slovakia (Slovak Republic))

PhD. student at the Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava. My theme of the PhD. thesis is about improving forecasting accuracy of electricity load by cluster analysis of consumers. I am blogging about time series forecasting and data mining (https://petolau.github.io/).

Section: Time-series, Machine Learning

A large amount of time series data are created from different sources nowadays. These sources can be smart meters, meteorological stations, web traffic etc. Large datasets of time series have to be processed in a fast way and mined for important and accurate decision making. Time series representation methods help more effective time series data mining thanks to their useful features as significant dimensionality reduction, the emphasis of fundamental (essential) shape

characteristics and implicit noise handling. In the talk, various time series representation methods will be presented. Their implementations and usage by TSrepr package will be shown. The use case with smart meter data clustering will be shown in order to demonstrate the usefulness of various representation methods for a consumers profile extraction and improvement of consumption forecasting accuracy. The simple extensibility of the TSrepr functionality will be also shown.

R packages: TSrepr

Pragmatic approach for efficient processing of spatial data: application to climatology [Tue 17:00]

Speaker: Ekaterina Fedotova (Senior Researcher @ Moscow Power Engineering Institut, Russian Federation)

I'm a senior researcher in the laboratory of the Moscow Power Engineering Institute. Our work is focused on the climate-energy nexus. At the moment R is one of my key instruments for calculations and reporting

Section: Spatial, Time-series

A number of powerful tools is available nowadays in R for work with NetCDF files. However, all of these methods imply use of data with high temporal resolution. Attempts of performance optimization are very likely to fail due to not quite obvious pitfalls of existing approaches. Some of these pitfalls were discovered during solving of a practical climate problem.

The considered problem was to estimate multi-annual averages of climatic parameters (air temperature, precipitation etc.) using an ensemble of the climate models. The solution was intended to employ the existing approaches as much as possible. A simple approach was developed to counteract discovered pitfalls and to utilize the data of the coarse temporal resolution while remaining on the safe side.

R packages: netcdf4, raster, RCMIP5

The work was supported by the grant MK-1494.2017.8 for young scientists

---

How to tell if a hockey player performs well (enough) [Tue 17:05]

Speaker: Jakub Houdek (Student @ University of Economics, Prague, Czech Republic)

*I am a student at the University of Economics in Prague. My major is operations research and econometrics while the minor is intelligent systems.

I have worked at the Czech Academy of Sciences as a research assistent or at Home Credit as a risk analyst. Recently i started working for a small company as a data analyst.

So far i have managed to be a part of a few publications, had a poster at useR!2017 and had a talk at my university about hockey analytics.*

Section: Statistics

Main topic of this talk is simple - it is desired to tell whether a selected player performs well or at least well enough. The process of the analysis is quite complex and demands several more or less complicated algorithms. First we introduce an algorithm for data pre-processing.

Data is obtained using **nhlscrapr** package which extracts data from NHL's *Real Time Scoring System. RTSS* contains informations about every single event that has occurred on ice in a selected season. This data is then processed using Macdoland's linear regression model. The result is a special *+/-* statistic which is adjusted to a single player's performace, thus we call it *adjusted plus-minus (APM).*

However, there are several hundreds of players in NHL and it is not really possible to examine every single player. For this reason a logistic regression model, which aggregates player performances, was created and tells us, which team is supposed to be well performing. Finally, our estimates are compared with real results.

This approach is quite unique as the aggregation serves as method of backtesting.

This leads us to a question - *can a player's performance be analysed?*

R packages: nhlscrapr, SparseM, MatrixModels

References:

- Macdonald, B. (2011). A regression-based adjusted plus-minus statistic for NHL players. Journal of Quantitative Analysis in Sports, 7(3).
- Thomas, A., & Ventura, S. L. (2017). Nhlscrapr: Compiling the nhl real time scoring system database for easy use in r. Retrieved from https://CRAN.R-project.org/package=nhlscrapr

---

Soylent Green is populations! Using synthetic populations in research and analytics [Tue 17:10]

Speaker: Chris von Csefalvay (Clinical Computational Epidemiologist @ Transcend Insights, United States)

Born in Budapest and educated in Oxford and Leiden, Chris von Csefalvay has been analysing, munging, transforming and interpreting healthcare related data for much of the last decade. His areas of interest comprise dispersion analysis of viral hemorrhagic fevers, biological defence and zero knowledge proofs in the decommissioning process of WMDs. With his wife Katie, an artist and art historian, and their kitten River, they split their time between Budapest and Southern California.

Section: Medical / Pharma, Bioinformatics, Social Sciences

In many fields, especially in healthcare, data sets often contain personally identifiable information (PII) that may on one hand be relevant to the analysis but, on the other, legislation such as HIPAA may limit its use. The use of synthetic populations - populations generated from an original sample that share certain statistical properties and relationships of the original sample but are reconstituted in a way that makes retrieval of the original PII impossible - is a possible solution to this problem. This presentation discusses the underlying methodologies of generating accurate synthetic populations and caveats to using synthetic populations, and briefly demonstrates the generation of a synthetic population using the **synthpop** package. Researchers in fields ranging

from healthcare through financial services to public sector information research can no doubt benefit from a thorough understanding of methods to generate synthetic populations.

Navigating the Wealth of R Packages [Wed 12:20]

Speaker: Hannah Frick (Data Scientist @ Mango Solutions, Great Britain)

Hannah Frick is a statistician turned data scientist, co-author and maintainer of the trackeR and psychomix R packages and a member of the R-Ladies leadership team.

Section: Community, Infrastructure

One of the main strengths of R are the many different add-on packages, covering various areas in statistics and machine learning as well as data wrangling and visualisation, interacting with APIs, etc. Over 12'000 packages are available through the Comprehensive R Archive Network (CRAN) and thousands more through BioConductor and GitHub. While it is great to have so many options, such a wealth of packages is not always easy to navigate, especially for someone new to the R ecosystem. How do you find a package for your task at hand? If there are several candidates, how do you choose which one to use? A typical workflow is to first gather a set of packages with relevant functionality before, via high-level comparisons, narrowing that set down to a handful of packages which are then explored in more detail. This talks highlights different approaches to discovering, comparing and choosing packages and the tools available for this: curated lists such as the CRAN task views, search-focussed approaches such as RSeek, community-driven approaches like Crantastic!, R-bloggers and R-weekly, metrics for high-level comparisons and suggested guidelines for in-depth review.

Write Rmazing Code! [Wed 12:25]

Speaker: Mikkel Freltoft Krogsholm (Senior Data Scientist @ Think Big Analytics, Denmark)

Mikkel writes R code all the time - on the job and in his spare time. Mikkel has worked for some of the biggest companies in the world so he knows how to write R code that is ready for production. He also follows the research community and does a lot of citizen science himself - and he follows the discussions on data sharing and reproducible research.

Section: Reproducible Research, Businesses

This talk is about how to unf*ck your code. And by unf*ucking I mean making sure that it works every time, under every condition and is written in a way that makes sense to you and to others. Because if it doesn't, then your code is f*cked. If you are a researcher then it means doing reproducible research. If you work in business it means writing production ready code. And if you are just writing code alone in the dark it means writing code your future self will understand. This talk is about coding styles, comments, documentation, packaging, tests and docker. This talk aims a making good programmers out of good data scientists.

Robust Data Pipelines with Drake and Docker [Wed 12:30]

Speaker: Tamas Szilagyi (Sr. Data Analyst @ ING, Netherlands)

Tamas is a passionate R user, working as a Senior Data Analyst at ING Investments in Amsterdam. Born in Hungary, he completed his studies in Belgium and Scotland, and worked in Canada before moving to the Netherlands. He has over 5 years of experience in various analytical roles and taught himself to program "on the job". He is a strong advocate of open-source tools and writes regularly about his hobby data projects on his blog: http://tamaszilagyi.com/

Section: Big Data, Infrastructure

In its early days a programming language for statistical computing and not much else, the R ecosystem of today is diverse enough to allow for end-to-end production systems written solely in R. In this session I will talk about how to set up a data pipeline going from data ingestion to serving up a prediction API. The main ingredient will be the workflow management package drake which - while still in its early days - takes the best from other similar frameworks written in Python such as Luigi or Airflow. For ease of deployment, I will also show how to encapsulate such an application in a Docker container and touch uplon a few best practices for doing so.

R packages: drake, plumbr, caret

Nested apply as an alternative to double for loops [Wed 12:35]

Speaker: Alicja Fraś (PhD Student @ Poznan University of Economics and Business, Poland)

I used to work as a senior business analyst at McKinsey in Warsaw. While pregnant and on maternity leave I conducted few courses in R and started PhD developing my programming and statistics skills. The PhD thesis concerns the topic of mutual fund fees and its determinants.

Section: Finance, Databases, Use-cases

Apply family functions are one of the R language advantages. They not only make the code look cleaner and more understandable, but oftentimes decrease the time of the code execution. Conducting calculations for my PhD I struggled with the need of operating on both columns and rows of the dataset. Double for loops with embedded custom functions to execute in each iteration were unclear, slow and caused errors. Efficient alternative may be utilizing nested apply family functions. I will show some of examples, that turned out to be very useful in my work. Combining apply and sapply together helped me decrease execution time thrice when compared to double for loop. One smart line of tapply nested in apply substituted four lines of aggregate function looped with for - with the same output. In most cases I first used apply to access the single column, and then - depending on the task - sapply or tapply, to perform given operation on the rows of the given column. In some cases I also implemented indexing rows in sapply, to operate only on selected rows of the columns. Additional advantage of this approach is no need to create a storage object (e.g. a data frame) before, as apply function accommodates and does not demand predefined type and dimension of the data structure.

What teaching R taught me about R [Wed 14:30]

Speaker: Mikolaj Olszewski (co-founder @ iDash s.c., Poland)

Mikołaj Olszewski is an experienced Data Scientist and the co-founder of a startup providing high quality Data Science training. He's passionate about R & Shiny, and crazy about learning. He loves to learn new things and teach others.

Section: Teaching

As an experienced instructor and a person who co-own a company that provides Data Science training, I find teaching a highly rewarding activity. Realising that learners understand the material you teach is priceless and highly motivating. Especially when running training or workshops on complex material, like R.

None of the programming languages is easy to teach but R seems to be outstandingly hard especially for unexperienced instructors. There are many things that can disturb or even ruin a well planned training. This could be a different behavior of the same code on computers that were brought to the session or a wide distribution of participants prior skills.

I believe that the success and spread of given technology depends highly on number of instructors that can teach it effectively. That's why in this short lightning talk I want to present some practices and guidelines that worked best for me during my career as a training instructor and motivate others to start teaching R.

Setting up your R workshop in the cloud [Wed 14:35]

Speaker: Tatjana Kecojevic (Data Scientist and instructor @ DataTeka, United Kingdom)

Dr. Tatjana Kecojević is a longtime R user with a doctorate from Statistics from the University of Manchester. She has spent many years working in U.K. higher education as a Senior Lecturer and has a comprehensive research record in area of quantile regression. She is a cofounder of DataTeka a company dedicated to helping people better understand and make sense of their data through thorough and insightful training strategies. She founded RLadies-Manchester and coorganises RLadies Belgrade and M

Section: Teaching, Big Data, Infrastructure

Setting up the computer environment for all participants of an R workshop can be a daunting and exasperating task. Installing and setting up the softwares and obtaining necessary files could give a confusing tone right at the beginning of the workshop before you have even started with the programme delivery. Not messing up with setup and being able to start using R in the same fashion as other people in the group is desirable not just to the workshop participants, but to the instructor.

There are a few options available for setting up RStudio in a browser. Using RStudio Server enables you to setup working environment tailored to your specific audience. It helps by removing any frustrations for the participants caused by downloads of the required files and installations from other workshop activities that you really want them to focus on. Setting up RStudio Server on Amazon Web Service (ASW), called an EC2 (Elastic Compute Cloud) instance, is not difficult. You can customise the working environment to meet your needs by installing globally required packages, adding the data and the script files you want participants to have

available when running RStudio in their browsers. Another alternative that has become recently available is RStudio Cloud. Although RStudio Cloud is an alpha version and it is still under development you could already do pretty much everything you are able by setting up your own RStudio Server.

This talk will illustrate the possibilities and benefits of using RStudio in a browser that could provide a great environment for teaching and learning data science with R.

Quality Assurance in Healthcare with R [Wed 14:40]

Speaker: Titus Laska (Data Scientist @ Federal Institute for Quality Assurance and Transparency in Healthcare, Germany) and Dirk Schumacher, Michael Höhle

The speaker is part of the Biometrics Unit of the German Federal Institute for Quality Assurance and Transparency in Healthcare (IQTIG).

Section: Medical / Pharma, Reproducible Research, Statistics, Use-cases, Web Apps

The Federal Institute for Quality Assurance and Transparency in Healthcare (IQTIG) founded in 2015 is the independent scientific institution responsible for the mandatory quality assurance in the German healthcare system. It is commissioned by the Joint Federal Committee, the leading institution in the self-government of healthcare, and the Ministry of Health.

The national quality measurement and improvement system in Germany is based on about 350 indicators, evaluated yearly or quarterly in order to improve selected areas of healthcare quality. The results of the analyses and routine reports have -- in some instances -- direct regulatory consequences and thus need to be of the highest quality.

R is used as the primary tool in the Biometrics Unit for routine statistics as well as for the research activities of the unit. One of the largest *R* components currently in routine use is an internal package that computes the results of the indicators for around 1,500 hospitals throughout Germany in more than 20 medical areas, including obstetrics, transplantation medicine, hip and knee replacement. The package was developed to validate the results published by the institute and is subsequently used for primary computation.

Internal packages are developed using an automated build infrastructure, automated tests and code reviews. Interactive statistical applications are published internally with **shiny**. All our analyses and statistical developments are carried out reproducibly with packages like **knitr**.

In this talk we will describe where and how *R* is being used within the IQTIG organization and how open source software supports our mission to provide transparency and continuous improvement support for the German healthcare system.

Writing R packages for clients: Guidelines at INWT Statistics [Wed 14:45]

Speaker: Mira Céline Klein (Data Scientist @ INWT Statistics GmbH, Germany)

Mira Céline Klein works as a Data Scientist at INWT Statistics, a fast-growing data and analytics company based in Berlin. In her daily work, she applies various statistical methods to customers' data using R. Mira holds masters' degrees in psychology and statistics.

Section: Businesses, Infrastructure, Reproducible Research

In our data and analytics company, we have been writing R packages for clients (and ourselves) for several years. Virtually all of our projects include an R package. In some projects, package development is the main goal, whereas in others the package only supports a statistical analysis and may never be used again.

Even in the latter case, we would always write a package because it reduces the number of errors and eases the communication in a team. Writing packages is not difficult, but there is always a way to do it even better. In this talk I want to share some of our lessons learned which have evolved into a company-wide set of best practices.

The most important learning is that tests and checks are as important for a data analysis as they are for software development. Helpful tools and strategies are also to rely on R CMD check, writing tests using the testthat package, and testing a package under different operating systems. To facilitate the work in a team, a version control system is essential as well as a common style guide.

R packages: devtools, testthat, roxygen2, lintr

An R toolkit to simplify and automate an open scientific workflow [Wed 14:50]

Speaker: Luke Johnston (Postdoctoral researcher @ Aarhus University, Denmark)

Luke is a postdoctoral researcher studying diabetes epidemiology at Aarhus University in Denmark.

Section: Reproducible Research

Many fields of science are slow to adopt open scientific practices (e.g. sharing code), especially in the biomedical fields. Given the increasing calls and demands for science to be more open (open data, open source) and reproducible, these practices will increasingly become requirements to publish papers and obtain funding. However, one of the biggest challenges for many researchers is that it is difficult to adhere to open and reproducible science principles. While there are a few packages and workflows currently available, e.g. ProjectTemplate or makeProject, these packages tend to focus on template creation or rely on extensive documentation rather than automation. The prodigenr package aims to simplify and automate many of these open science tasks. Presently, the prodigenr package automates the creation of an open reproducible research project, with templates for manuscripts, posters, and slides currently available. The structure and workflow of the newly created project emphasizes adhering to reproducible practices (e.g. documenting dependencies, use of R Markdown for manuscripts with R code chunks, etc). The ultimate goal is to create a package (or packages) similar in ideology to the devtools package, but aimed at mainly biomedical researchers and scientists using (or wanting to use) open scientific practices. This package is also being developed through the Mozilla Open Project Leader Training. The aim is to create an ecosystem of packages that encourage and automate open science principles. The aim of presenting this package and ideas at eRum is to get feedback and suggestions from the community (particularly surrounding practices in reproducibility and open science), how to best target biomedical researchers, and to find potential collaborators.

R packages: prodigenr

Manage your meta-analysis workflow like a boss: Introducing the {metamanager} package [Wed 14:55]

Speaker: Tamás Nagy (Adjunct Professor @ Eövös Loránd University, Hungary)

Tamás has a PhD in psychology, and currently works in the Eotvos Lorand University as an adjunct professor. His research interests include stress and emotion psychophysiology and scientific computing. Previously, Tamás was the lead researcher of Synetiq, a Hungarian research startup that uses biometric sensors to measure emotional reactions to media content. Tamás has been using R since 2014, and teaches R programming and data analysis for graduate students.

Section: Data Munging, Reproducible Research, Social Sciences

According to the hierarchy of scientific evidence, a meta-analysis - when executed properly - can provide the most reliable answer for a research question. However, conducting a meta-analysis is a time consuming enterprise, that requires not just domain specific knowledge and analytical experience, but considerable data management skills as well. To aid reproducible research, it should be possible to handle all data management tasks directly from R. Although there are excellent R packages to for conducting the statistical part of a meta-analysis - like the famous {metafor} package -, there is a lack of packages dedicated to meta-analysis data management. We present the {metamanager} package, that was developed to efficiently manage meta-analysis projects using R. Using the package helps to implement the requirements of the PRISMA checklist, and provides convenient functions for conducting reproducible meta-analysis. Key functionality involves refining search terms, conducting literature search in multiple scientific databases, downloading and tidying article meta-data, creating files for screening and coding articles, calculating coder agreement, etc. The functions can be easily arranged into human readable pipelines using the %>% operator. The package uses google drive to store data, which ensures continuous access and provides a straightforward interface for manual data entry, while handles version control and collaborative editing. It is also convenient for research teams where not all collaborators are proficient in R. During the presentation I am going to show the functionality of the package through a real-life example.

R packages: tidyverse, googlesheets, googledrive

Reference: https://github.com/nthun/metamanager

Establishing analytical pipelines tools and culture [Wed 15:00]

Speaker: Andrea Schnell (Economist, Data Analyst @ Statistical Office of Zurich, Switzerland)

Economist/Data Analyst at the Statistical Office of Zurich, exploring business statistics and regional economics. Former positions include economic research in the banking sector and in research institutes. Co-Organizer of the Zurich R User Group.

Section: Reproducible Research, Community, Use-cases, Data Munging, Infrastructure

This talk covers the change of statistics production from an ad hoc spreadsheet based process to an analytical pipelines framework. Ingredients: internal R packages that streamline R output according to corporate design requirements, raise awareness for reproducibility issues and integration of git processes. In addition to the technical procedures and skills we engage in community building and knowledge sharing, both fostering collaboration within our organization and beyond.

Shiny Demos [Mon 19:00]

5 minutes of free-form presentation in an auditorium on a Shiny application – without any dedicated time for Q&A, but with a follow-up session in a dedicated area (with tables) of the Poster Session where the presenters can continue demoing their application to the interested attendees and answer questions etc.

SmaRP: Smart Retirement Planning A Shiny App to evaluate retirement benefits and optimization thereof, in the context of the Swiss system

Speaker: Francesca Vitalini (Solutions Consultant @ Mirai Solutions GmbH, Switzerland)

*Francesca Vitalini is a computational biophysicist by training, field in which she holds a PhD from Freie Universität Berlin. In the past three years she has been working as a Solutions Consultant for Mirai Solutions GmbH, where she develops analytical solutions for business clients in **R**. Francesca supports smart data analytics for educated decision-making.*

Section: Businesses, Finance, Web Apps, Use-cases

The Swiss social security system, one of the most robust, is based on a three-pillar regime. The first Pillar, common to most developed countries, is a state-run pay-as-you-earn system with minimum benefits. The voluntary contribution (Pillar III) is a privately-run, tax-deductible insurance fund. At the heart of the Swiss system is the so-called Pillar II, a compulsory, tax-deductible company occupational pension insurance fund. Voluntary Pillar II buy-ins are regulated but allow for benefits improvement while reducing the tax burden during the working career. The complexity is further increased by a municipality-dependent taxation. Altogether this calls for an early-stage conscious approach towards retirement planning. However, it is not straight-forward to assess effects of early retirement, moving to a different canton or applying a different voluntary pension schema.

SmaRP, Smart Retirement Planning, supports the users in an educated decision-making. Using R and Shiny, we developed a parameterisable pension calculator web application, which provides real-time computation of the total retirement funds over time, explicitly accounting for the various contributing blocks. It features a flexible yet intuitive user interface with several options for detailed personalisation. An interactive visualisation is implemented using googleVis and the underlying data is made available for download. The user is additionally presented with a custom report generated with rmarkdown/knitr. Unlike other pension calculators, the details of the model approach and underlying assumptions are disclosed to make results transparent and reproducible. The modular structure of the Shiny application allows for seamless extension and supports modular execution, making the application fully adaptive.

In this talk we will give a deep-dive into the framework by showcasing illustrative ex- amples and scenarios everyone can relate to.

R packages: Shiny, rmarkdown, knitr, dplyr, googleVis

Going async with Shiny

Speaker: Dávid Gyurkó (Data Scientist @ Hungary)

Data scientist - FreelanceR - Interested in web development - Frequent on the Stackoverflow tag: [shiny].

This talk introduces useR-s to asynchronous (async) programming in Shiny using the future and promises packages. The talk will include the answers to the questions: what is async? why bother with it? and how can I go async with Shiny?

R packages: future, promises

Informed clinical decisions based on population pharmacometric models with the aid of Shiny

Speaker: Agnieszka Borsuk-De Moor (PhD Student @ Medical University of Gdańsk, Poland)

PhD Student in PK/PD modeling focused on population models of drugs used in intensive care and anesthesiology and individualized drug dosing

Section: Medical / Pharma

Population pharmacokinetic/pharmacodynamic (PK/PD) models capture the variability of the target population of patients and enable the search for potential covariance influence on drug's PK/PD parameters in individuals. This knowledge could help to adjust drug dosing in clinical practice. However, the complexity of the models, their underlying hiercharchy and the variety of softwares used for analysis make population models incomprehensible for clinicians, while they could be valuable for individualizing therapy and increasing its safety. Doctors need an easy tool which will be able to answer typical questions emerging in clinical setting. Shiny package offers a readily-available solution to this issue that does not require any programming skills nor pharmacometric expertise. In this talk the use of Shiny application representing pharmacometric model will be discussed in practice.

R packages: shiny, ggplot2

Shiny Dashboard on streaming data

Speaker: András Tajti (developer @ OTP Bank Plc., Hungary)

Started using R along studying survey-statistics - especially he igraph package for network analysis. After university, started working as a need-to-be developer at Andego, using mostly R. In the process of implementing daily fraud-detection processes, company-network detection solutions, treasury application, started collecting utility functions, from which the data cleaning functions are worth to share.

Section: Big Data, Web Apps, Dashboards

Following reactions to certain events never been as easy as today: just go toTwitter, and collect everything with the tag of your interest. But pure tweets can be cumbersome to understand for a single viewer, especially in real time. For this problem, one can create a dashboard wit the number of tweets, most popular tags, reaction count, anything which can be computed fast enough to get on the board 'in real time'. Although by default, Shiny does not have a digestStream reactive function, there are multiple ways to get around this problem. and I will show some of them.

Towards Native Declarative Data Collection with Question and Survey Instant Feedback in R&Shiny

Speaker: Kamil Wais (Assistant Professor @ University of Information Technology and Management in Rzeszow, Poland, Poland)

Data Scientist with R & Shiny programming skills, specializing in new research techniques in Social Science based on Internet technologies and Open Data. Author of the genderizeR R package. Previously, On-line Research Product Manager in one of the largest global institutes of market and opinion research, and short-term visiting Assistant Professor at the Center for Social Research & Center for Research Computing at the University of Notre Dame (Indiana, USA).

Section: Dashboards, Social Sciences, Web Apps

R is great for data analysis and Shiny is great for interactive data visualisation, but could we use R&Shiny for efficient declarative data collection? Moreover, how can we develop web data products in R&Shiny, that are based on real-time declarative data collection with after-question and after-survey instant feedback?

Users of such web data products should be able to immediately access the feedback relevant to their answers. To increase the value of the feedback, it should be dynamically customised to each respondent. This can be achieved by pre-programmed templates of feedback scenarios, which can be adaptively customised by the respondent’s answers to this or previous questions. Employing large analytical and data visualisation capabilities in R, we could try to adapt any type of instant feedback to each user. Using R, we could also combine different feedback sources: a respondent’s answers to a given question and to other questions, other users’ answers, external open data (imported into our app or available via APIs), and aggregated or summarised outcomes from reference studies.

What are the possibilities and obstacles for developing such data products natively in R&Shiny? How the idea of QAF (Question, Answer, and Feedback) objects can be implemented in R&Shiny? What is the roadmap for developing ODGAR framework for On-line Data Gathering, Analysing, and Reporting? Is it possible to build mobile app in R&Shiny? I will try to answer these questions using experience gained from developing early stage prototypes.

R packages: shiny, ODGAR, shinyjs, shinydashboard, shinywidgets, shinysense, shinymaterial, shinygui, shinyBS, shinyFeedback, limeRick.

The Zurich RealEstateApp - An R-Shiny application for the Zurich real estate market

Speaker: Max Grütter (Data Scientist @ Statistical Office of the Canton of Zurich, Switzerland)

I received my doctorate in labour market economics from the University of Zurich and worked in public administration and a private consultancy firm in Switzerland. For two years now, I have been a data analyst at the Statistical Office of the Canton of Zurich, focusing on the real estate market and an enthusiastic R-user.

Section: Statistics, Web Apps

The Statistical Office of the Canton of Zurich is developing an internal-**R Shiny app** for improving the speed and quality of real estate market queries.

Since 1974, the Statistical Office has privileged access to detailed information on the Zurich real estate market in form of administrative data. Despite a wide range of statistics, simple inquiry tools and in-depth analyses, the Statistical Office is receiving more and more complex queries on the real estate market. Many of these requests are similar and can be met using standard processes. Nevertheless, they cause a (too) high personnel expenditure in production, quality assurance and communication.

To solve this problem, the office is currently developing an R-Shiny application, which allows to create complex analyses from different areas of the real estate market without any specific programming knowledge. Spatial, temporal and content criteria can be individually selected using various selection tools. In a real-time calculation, the **Zurich RealEstateApp** produces the results and displays them in tabular and graphical form. A download function for the tables and graphics is also provided (Excel, csv, pdf). The application ensures that the results are methodically and content wise clean and harmless in terms of data protection law, making it easy and reliable for customers to access valuable facts for their decision-making process.

The described project will be presented at the erum2018 conference in the form of a presentation. In addition to the most important steps during the development and the central selection tools of the app, application examples and initial findings from the test phase with beta users will be presented. It is planned that interested parties will be able to test the app themselves during the conference.

R packages: tidyverse, shiny, leaflet

What is the best place to be? Location optimization with R and Google Maps

Speaker: Bartosz Czernecki (Assistant Professor @ Adam Mickiewicz University, Poland) and Jakub Nowosad

Bartosz has been using R and Fortran for 9+ years for atmospheric research and GIS. His experience was gained in the R&D dept. of the Polish met service (HPC, visualisation, modeling). Received his PhD in geosciences (climatology). Currently working as assistant professor in the Department of Climatology at Adam Mickiewicz University in Poznan. Author of numerous impact-factored

scientific papers. Co-founder of iqdata.pl where he couples scientific background with business and industry needs.

Section: Businesses, Use-cases, Spatial, Social Sciences, Web Apps

Have you ever tried to find a hotel that is close to the conference venue but then you realized that close in terms of a distance doesn't always mean the same in terms of time? Or have you searched the house market for a place that is the closest to both yours and your partner's workplaces? Simple distance measures could not be the best in these cases. For example, you are unable to cross a river in every point - you need to find the closest bridge. Additionally, transportation availability varies for different places and largely depends on a preferred mode of transportation (e.g. foot, cycle, motor vehicle, public transport). Private motor vehicle could be the most efficient in the suburbs, but not the most suitable in a city center. Distance Matrix API of Google Maps accessed through the googleway and spatial packages can be useful to answer the above questions. It allows for computation of distances and times between points, but also can be modified to create isochrone maps. They present areas of the similar "distance" minutes spend traveling (or kilometers along roads) from the point of interested. In this talk, we will show a case study of finding the most optimal place to stay for the eRum conference by the means of interactive visualization features of spatial R for location optimization.

R packages: googleway, leaflet, sp, sf, raster, rgdal

Visualizing vehicle usage with the leaflet package

Speaker: Peter Szabolcs (Data scientist @ Vodafone Shared Services Budapest, Hungary) and Tamas Molnar

Peter has a background in finance from Corvinus University, but he developed strong feelings towards data science as well while teaching statistics courses at the university. After graduation Peter became a business analyst at Tresorit and played a big role in building up it's BI system from scratch and also performed a few data science related projects. In 2017, he joined Vodafone SSC Budapest's Advanced analytics team, where Peter is now working as a senior data scientist.

Section: Businesses, Spatial, Use-cases

One good use case for visualisation is to check the movements of some fleet cars on a map.Visualizing spatial data is always fun and exciting for the developer, but sometimes the end user finds it hard to make a good use of it. With this in mind, the VSSB Advanced Analytics team wanted to create a solution that is easy to use, dynamic and is very much centered around needs of the business. The concept has been brought into life on a dataset containing real information about how people were using the cars of the company's fleet. The mapping is conducted with the help of the leaflet package and the solution is embedded into a Shiny environment. A related work in progress is to apply social network analysis on the driver population. In case of success this attempt would be presented as well.

Posters [Mon 20:00]

Presenters bring their printed posters to be mounted on the wall or provided poster holder for 4 hours on the Monday night social event. Attendees can freely walk around in the dedicated Poster Session area to check the posters and chat with the presenters. The presenters are suggested to stand by or not too far from their posters to answer any question others might have and to grab all opportunities to speak about their presentation.

Crazy Sequential Representations of Natural Numbers

Speaker: Anne Bras (PhD Student @ Erasmus Medical Center, Netherlands) and Vincent van der Velden

Background in computer science and medicine. Currently a PhD candidate working on high-throughput high-dimensional data analysis. Frequent user of the R environment and attendee of various R conferences.

Others have shown that almost all natural numbers from 0 to 11111 can be written in terms of increasing and decreasing orders of 1 to 9 by only using plus, minus, product, potentiation, division and/or brackets.

Random examples for increasing order:

- 9617 = 1+2^3*(45+(6+7)*89)
- 9618 = 1*(2+3+4+5)*(678+9)
- 9619 = 1+(2+3+4+5)*(678+9)

Random examples for decreasing order:

- 9617 = 9*876+5+(4*3)^(2+1)
- 9618 = (9+8+7*(6+54+3))*21
- 9619 = 9*87+(6*5+4^3)^2*1

Generally this collection is revered to as the crazy sequential representations of natural numbers. Surprisingly, only one crazy sequential representation remains to be identified (10958 in terms of increasing orders of 1 to 9).

Goals:

- Identify missing crazy sequential representations for 10958
- Extend the existing collection from 11111 up to 2147483647

Considering 9 digits and 6 operations, billions of lexicographical unique equations can be generated, which makes typical brute-force approaches unfeasible for someone with limited computational resources. During this presentation I will elaborate on various techniques used to reduce the number of candidate equations (to a manageable size) and elaborate on the technical difficulties encountered (including symbolic algebra, big number arithmetic, cluster computing, etc).

Finding groups in time-to-event data by means of the clustcurv package

Speaker: Nora M. Villanueva (PhD student @ Dep. Statistics and O. R. University of Vigo, Spain) and Marta Sestelo, Luís Meira-Machado

I received her MSc degree in Statistical Techniques from the University of Vigo in 2012. I have combined my work at Gradiant with research in the University of Vigo. Particularly, my research lines are nonparametric statistical inference and software development, in which I continues to work in order to obtain my Ph.D. in the Department of Statistics and Operation Research.

Section: Statistics, Medical / Pharma

One important goal in survival analysis is the comparison of survival curves between groups. In many longitudinal survival studies one is often interested in assessing whether there are differences in survival among different groups of participants. For example, in a longitudinal medical study with a survival outcome, we might be interested in comparing survival between participants receiving different treatments, different age groups, racial/ethnic groups, geographic localization, etc. Several nonparametric methods have been proposed in the literature to test for the equality of survival curves for censored data. However, none of these methods can be used to determine groups among a series of survival curves. Naive approaches, such as pairwise comparisons, lead to a large number of comparisons making difficult the interpretations. Based on this, a new method is proposed which allows determining groups of survival curves with an automatic selection of their number. This method is implemented in the R *clustcurv* package and is illustrated using data from a colon cancer study.

References:

- Kaplan E.L., Meier P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53:457 -- 481.
- Macqueen J.B. (1967). Some methods of classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281--297.

---

survidm: Inference and Prediction in an Illness-Death multi-state Model

Speaker: Marta Sestelo (Postdoctoral Research and Lecturer @ Dep. Statistics and O. R., CINBIO and SiDOR group, University of Vigo, Spain) and Luís Meira-Machado

I am a Senior Researcher at Gradiant and Lecturer at University of Vigo. I am focused on the development of new methodologies and algorithms linked with statistics. I am interested in estimation and inference methods of flexible models, in developing practical tools for data analysis and in gaining better understanding of real life issues through statistical knowledge. At the moment, my lines of research are closely related to computational statistics, survival analysis, nonparametric regression

Section: Medical / Pharma, Statistics

Multi-state models are a useful way of describing a process in which an individual moves through a number of finite states in continuous time. The illness-death model plays a central role in the theory and practice of these models, describing the dynamics of healthy subjects who may move to an intermediate “diseased” state before entering into a terminal absorbing state. In these models one important goal is the modeling of transition rates which is usually done by studying the relationship between covariates and disease evolution. However, biomedical researchers are also interested in reporting other interpretable results in a simple and summarized manner. These

include estimates of predictive probabilities, such as the transition probabilities, occupation probabilities, cumulative incidence functions, prevalence and the sojourn time distributions. An R package was built providing answers to all these topics.

References:

- Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., and Andersen, P.K. (2009). Multi-state models for the analysis of time-to-event data. Statistical Methods in Medical Research, 18, 195-222.
- Putter, H. and Spitoni, C. (2016). Non-parametric estimation of transition probabilities in non-Markov multi-state models: The landmark Aalen-Johansen estimator. Statistical Methods in Medical Research, 1-12.

The comparison of robust and non-robust geostatistical model of meteorological data from Estonia

Speaker: Michael Kala (Student @ Czech technical university, Faculty of Civil Engineering, Czech Republic) and Petra Pasovská

We are students of a master study programme Geomatics on the FCE, CTU. Our specialization includes the problematique of geostatistics. For solving geostatistics problems, we use R-project.

Section: Statistics

Geostatistics is a branch of statistics dealing with estimates and predictions of stochastic phenomena on Earth. It applies general statistical principles for modeling and drawing conclusions about geostatistical problems. The aim of this project is to estimate parameters of the geostatistical model of meteorological data using both methods of robust and non-robust statistics and to compare these two approaches. The origin of the used data is the dataset ECA&D (Europian Climate Assessment & Dataset), that is distributed freely. For the analysis, data from Estonian meteorological station were used (temperature, the depth of snow etc.). Analyses are made in the environment of R-studio emphasizing the use of geoR and georob packages.

R packages: geoR, georob

Reference: Europian Climate Assessment & Dataset, http://www.ecad.eu/

Hydrological and data-driven simulation of river flows in R

Speaker: Milan Čistý (professor @ Slovak University of Technology, Slovakia (Slovak Republic)) and Veronika Soldánová

Milan Čistý is professor in Hydrology and Water management on Slovak University of Technology

Section: Machine Learning, Time-series, Use-cases

Our proposal of the poster is the case study of using R in hydrology. It is based on the following ideas. We will present the creation of a river flow generator, which uses as an inputs temperature and precipitation data. Our work points out the difference between two types of modelling: flow simulation and flow forecasting. In the case of the flow simulation (e.g., their generation in the future), flows from the days before the prediction cannot be used as the input data. They are simply not available when we want to generate flows in a river in future time horizons. Most

studies on machine learning applications in rainfall-runoff modelling were flow-prediction studies, so this study will also re-evaluate the performance ratio of machine learning models versus hydrological modelling from point of view of real rainfall-runoff modelling. We emphasize word "real" in the previous sentence because prediction on the base of previous flows included in input data is basically just an extrapolation. To increase the accuracy of the flow simulation (or generation), the authors have designed and tested various variants of machine learning models, which will be evaluated and described on the poster together with R codes (keras, XGBoost, RF...). The proposed models are compared with the hydrological model (hydroTUW R package is used). The comparison is made on stream flows in Slovakia. hydroGOF package and other R packages are used for this purpose. The proposed approach which will be described on the poster allows for a significant increase in the accuracy of the flow simulation. As will be demonstrated in the case study, improvement in the precision of the modelling is significantly affected not only by the selection of model but also by hydrology-inspired feature engineering, hydrology-inspired feature selection, and the appropriate architecture of the entire modelling.

Predicting performance of concrete structures by machine learning

Speaker: Zsuzsanna Szabó (Researcher @ Mining and Geological Survey of Hungary, MTA Premium Postdoctorate Research Program, Hungary)

Environmental physicist-geochemist, PhD, Data Science and R Enthusiast with 10 years of international academic research experience in Hungary, Norway and France. Presently focuses on cement deterioration subsurface within a prestigious MTA research grant.

Section: Machine Learning, Reproducible Research, Statistics, Use-cases

Being able to predict the degradation of concrete is important in economic, environmental and human safety point of view. An academic research project titled “Geochemical interactions of concrete in core samples, experiments and models” has begun in 2017 October which primarily aims to understand and/or simulate geochemical interactions (mineral dissolution and precipitation processes) in concrete–rock–water systems. One of the several research tools available for the prediction of these reactions is numerical geochemical modeling (based on chemical equations, equilibrium and rate constants). To support these experimentally validated theoretical models a data based solution is also explored by machine learning algorithms (empirical models), first, applied on a publicly available dataset. By the use of a neural network algorithm and the usually available data, the compressive strength of concrete is proven to be possible to estimate. Most of these works, however, primarily focus on the preparation of concrete (the mix) and its age (as predictors). By reproducing one of these study cases for the dataset ‘concrete’ in the R package ‘AppliedPredictiveModeling’, lessons to learn are collected for the improvement of the geochemical perspective project. Based on the experiences of this first analysis, a local dataset is expected to access which is suitable to predict concrete integrity even when affected by intense geochemical reactions like sulfate or acid attack.

Modelling of job offers

Speaker: Krzysztof Marcinkowski (Student @ Poznań University of Economics and Business, Poland) and Katarzyna Zadroga

We are econometrics students interested in data science and statistic. We are members of student scholarly associations: Estymator. We finished internships as a junior data sientists.

Section: Text mining, Statistics

The main goal of our poster is to classify jobs offers, obtained from one of Poland's leading classified ads service, to topics based on text description. We will apply text mining methods, in particular focusing on using the Latent Dirichlet Allocation algorithm. This approach will allow us to predict unknown categories of offers basing on extracted keywords. We will use tidytext, topicmodels and tm packages. Last but not least, we will use dimension reduction techniques implemented in homals package to prepare informative visualization. This poster is a part of our master's thesis devoted to usage of Internet data sources to provide job vacancy statistics.

R packages: tidytext, tm, homals, topicmodels.

References:

- Silge Julia and David Robinson. 2016. "Text Mining with R A Tidy Approach."
- J. Ćwik, J. Mielniczuk, 2009, 'Statystyczne systemy uczące się. Ćwiczenia w oparciu o pakiet R.', Wydawnictwo: OWPW

Hierarchical Cluster Analysis

Speaker: Piotr Opiela (student @ Uniwersytet Ekonomiczny w Poznaniu, Poland) and Greta Białkowska, Magdalena Maślak

We are a students of Informatics and Econometrics at Poznań University of Economics and Business also we are a members of "Estimator" student club. We finished internships as a junior data sientists.

Section: Statistics

The main aim of this poster is to present the best vacation resorts based on multivariate analysis, including cluster analysis and synthetic measures. We collected data about destinations which included: average temperature of water, humidity, distance from Poland, average price of the expidition, ratio of prices in the destination to prices in Poland, GDP per capita.

We processed data with base R and tidyverse packages. Further, we conducted cluster analysis based on Ward method and k-means, followed by calculations of synthetic measures based on the Generalized Distance Measure which allowed us to rank the destinations. In this step we used clusterSim, cluster and factoextra packages.

R packages: tidyverse, cluster, factoextra, clusterSim

Reference: http://uc-r.github.io/hc_clustering#replication

Generating Time Series' Latent Factors with Variational Autoencoders

Speaker: Sami Diaf (Student @ Hildesheim Universität, Germany)

Economist, Statistician and Data Scientist ! mixed background with several years of interdisciplinary research.

Section: Time-series, Finance, Statistics

The poster will detail how to generate latent factors using a Bayesian autoencoder architecture (Kingama & Welling, 2014) which can be used for forecasting purposes. Applied to the Euro/Dollar daily exchange rate series, this methodology yielded two normally distributed latent factors exhibiting distinct behaviors.

R packages: keras, dplyr, ggplot2

References:

- D. P. Kingma, M.Welling :Autoencoding Variational Bayes (2014) https://arxiv.org/pdf/1312.6114.pdf
- C. Doersch: Tutorial on Variational Autoencoder (2016) https://arxiv.org/pdf/1606.05908.pdf
- S. Keydana: Variational Autoencoders for Anomaly Detection (2017) https://rpubs.com/zkajdan/308801
- M. Shiffman: Under the Hood of the Variational Autoencoder (in Prose and Code) (2016) http://blog.fastforwardlabs.com/2016/08/22/under-the-hood-of-the-variational-autoencoder-in.html
- F. Chollet: Building Autoencoders in Keras https://blog.keras.io/building-autoencoders-in-keras.html

---

Finding relevant atmospheric predictors of surface precipitation type

Speaker: Bartosz Czernecki (Assistant Professor @ Adam Mickiewicz University, Poland) and Andrzej Wyszogrodzki, Jan Szturc, Leszek Kolendowicz, Marek Półrolniczak

Bartosz has been using R and Fortran for 9+ years mostly for atmospheric research and GIS. His experience was gained in the R&D dept. of the Polish met service (HPC, visualization and modelling tasks). Received his PhD in geosciences (climatology) in 2014. Currently working as assistant professor at Adam Mickiewicz University, Poznan (PL). Author of numerous impact-factored scientific papers. Co-founder of IQData where he couples his scientfic background with a needs of business and industry.

Section: Big Data, Databases, Machine Learning, Spatial

The possibility of differentiating between types of atmospheric precipitation is of a key importance in contemporary climatological studies. Both liquid and solid precipitation, being one of the elements of the water circulation, may be viewed as extreme weather phenomena that are possibly dangerous, and therefore they play an important role for the economy. Present knowledge of physical dependencies conditioning the occurrence of surface precipitation type is burdened with a considerable error. Therefore numerous studies try to parametrize atmospheric factors that impacts this phenomena.

In this study authors aimed to use different data sources such as: dual-polarimetric doppler radars, satellite-derived products, lightning detection network, mesoscale numerical weather prediction model and in-situ observations. Due to different: (1) data reliability, (2) notable differences in

temporal and (3) spatial resolution, and also (4) different data formats (HDF-5, NetCDF, GRIB, GeoTIFF, ASCII), a common database was created. Some limitation of R for raster dataset was found for rotated geographical projections. The application of machine learning algorithms, especially based on the 'caret' (to avoid overfitting) and 'Boruta' packages, let to find all relevant features that might be applicable for modelling of surface precipitation type.

R packages: dplyr, tidyr, rgdal, sp, sf, raster, rasterVis, parallel, data.table, ncdf4, hdf5, gtools

---

Application of 'R' to assess the impact of biometeorological conditions for the landscape perception and objectification of its evaluation – experiment with eyetracking glasses

Speaker: Leszek Kolendowicz (professor @ Adam Mickiewicz University in Poznań, Poland) and Marek Półrolniczak, Ilona Potocka, Mateusz Rogowski, Szymon Kupiński

Leszek Kolendowicz and Marek Półrolniczak are climatologists and work at the Climatology Department of Adam Mickiewicz University (UAM) in Poznań. Ilona Potocka and Mateusz Rogowski are geographers and work at the UAM Tourism Department, while Szymon Kupiński works at the Supercomputing and Networking Center in Poznań.

Section: Social Sciences, Spatial, Graphics, Statistics

The aim of the study was to determine the influence of biometeorological conditions on the perception of the landscape according to personal characteristics and actual general physical and mental state of the observer as well as the features of the landscape itself. The study area is situated in the northern part of Poznań in Warta river valley. View point is located on the roof of the building of the Faculty of Geographical and Geological Sciences of Adam Mickiewicz University. During the study, 28 persons (14 women and 14 men) was qualified for the project. Using the mobile eyetracker which recorder eyeball movement for assessing the AOI (area of interest) they perceived the same section of a panorama in various seasons and under different types of weather. 9 research cycles throughout the entire research period (January 2017 – February 2018) were conducted to take into consideration the seasonal changes in the landscape. As a result, we expect to determine variables concerning the magnitude and nature of changes in perception taking into consideration the degree of influence of the current weather state. As the research has demonstrated responders with negative feelings spent four times more average with eye gaze on development area, trees in flower and almost three times more on buildings. These ones who average spend two times more dwell time on sand, felled-tree and flowers said that landscape was non-stimulating. In the study, calculations and visualization were made using R program packages.

Effect of patient compliance on long-term survival

Speaker: Mariann Borsos (Senior biostatistician @ Adware Research Kft, Hungary)

I have a Master's degree in Applied Mathematics from Eotvos Lorand University with specialization in Mathematical Statistics. I am a biostatistician with 10 years' experience in later phase (Phase III, IV and post-marketing) human clinical trials. I have been an R user for 4 years.

Section: Medical / Pharma, Statistics

In most medical conditions it is not sufficient to find the correct diagnosis and the appropriate medical treatment as patient compliance still remains a relevant barrier to treatment effectiveness. Patient noncompliance can take many different forms and can be measured in a number of ways. In our presentation we chose the disease-free survival as measurement of treatment effectiveness.

We focus on the problem that patient compliance should be calculated from data collected during the whole investigated interval till the event time. For censored patients it is clear that degree of compliance can be calculated from the whole investigated period but for those who suffered an event during this period calculation of compliance arises several methodological questions and the problem of biased estimates.

We present two possible solutions. First, we investigate the method generally known for time-dependent covariates where final results are determined as a weighted average of several survival models applied on shorter sub-periods of the original time period. Second, we present our model where the investigated period is divided into two parts: first we determine the compliance from data of the first period and then apply the survival model only for the second period. We perform a series of the above survival models changing the cut-point and investigate the effect of the choice of cut-point on the results.

R packages: survival, km.ci, survminer

References:

- Friedo at al.: Survival analysis: time-dependent effects and time-varying risk factors, Kidney International, 74 (2008)
- Mark Egan Tomas J. Philipson: Health Care adherence and personalized medicine, Working Paper 20330 (2014)

Facial attractiveness evaluation for purposes of plastic surgery using web-based shiny application

Speaker: Lubomír Štěpánek (biostatistician, software developer, junior lecturer, PhD candidate @ First Faculty of Medicine, Charles University & Faculty of Biomedical Engineering, Czech Technical University in Prague, Czech Republic) and Pavel Kasal, Jan Měšťák

A Master's student in Statistics. A PhD candidate in Biomedical Informatics focused on medical decision-making systems and facial attractiveness evaluation for purposes of plastic surgery. A medical doctor, former oncologist, and a junior lecturer in Introductory Informatics and R courses.

Section: Graphics, Medical / Pharma, Statistics, Web Apps

According to current studies, facial attractiveness perception seems to be data-driven and irrespective of the perceiver. However, the ways how to evaluate facial attractiveness complexly and how to make comparisons between facial images of patients before and after plastic surgery procedure are still unclear and require ongoing research.

In this study, we have developed a web-based shiny application providing facial image processing, i. e. manual face landmarking, facial geometry computations and regression model fitting

allowing to identify geometric facial features associated with an increase of facial attractiveness after undergoing rhinoplasty, one of the most common plastic surgeries.

Profile facial image data were collected for each of a patient, processed, landmarked and analysed using the web-based application. Multivariate linear regression was performed to select predictors increasing facial attractiveness after undergoing rhinoplasty. Facial attractiveness was measured using Likert scale by a board of independent observers.

The shiny web framework enables to develop a complex web interface and, because of shinyjs package, a clickable interaction useful for the landmarking as well. Given the collected data, enlargement both of a nasolabial and nasofrontal angle within rhinoplasty were determined as significant predictors increasing facial attractiveness.

We built a web-based shiny application enabling basic facial image processing and evaluating facial attractiveness. Furthermore, we performed a regression analysis using the application to point out which facial geometric features affect facial attractiveness the most, and therefore should preferentially be treated within plastic surgeries.

R packages: shiny, shinyjs

Reference: Kasal P., Fiala P., Štepánek L. et al. Application of Image Analysis for Clinical Evaluation of Facial Structures. In: Medsoft 2015. (2015), pp. 64–70.

A user-friendly interface for spatial data analysis with R

Speaker: Mónica Balzarini (Full Professor, Senior Research @ Universidad Nacional de Córdoba. CONICET, Argentina) and Mariano Córdoba, Pablo Paccioretti, Franca Giannini Kurina, Cecilia Bruno

Professor, National University of Córdoba since 2003. Senior Research at the National Council of Science and Technology of Argentina, CONICET Research and Teaching Experience: Agricultural Statistics. Statistical Modeling, Multivariate Analysis, and Spatial Statistics applied to agricultural sustainable developments. Professor of Ph.D. courses on Applied Statistics. Main Advisor of Ph.D. Thesis on Biometry for experimental and observational studies in Agriculture.

Section: Statistics, Spatial

Information technologies that generate different types of data associated with spatial localization have been promoted in the last decades. The optimal use of georeferenced data depends on the capacities for efficiently and simply analyzing spatial variability. To facilitate the implementation of univariate and multivariate algorithms for spatial data, a "Spatial Statistics" module was developed in the statistical software InfoStat using R. The development merges the best of the ease-to-use menu-driven in InfoStat with the power of R and includes procedures for spatial data pre-processing and analytics. Autocorrelation indexes, correlation coefficients, semivariograms, interpolation, principal components and clustering for spatial data can be performed throughout a dashboard. In addition, we develop the FastMapping app to automate interpolation and mapping of spatial variability. FastMapping automates the selection of the best geostatistical model for kriging prediction by cross-validation. The application was developed using the Shiny R package, and it is compatible with different browsers. The use of both developments is illustrated by the

implementation of a logical sequence of algorithms to obtain a multivariate spatial variability map from a database of five soil variables intensively recorded in an agricultural field under precision agriculture.

Mathematical modelling of cocoa bean fermentation using a Bayesian framework

Speaker: Mauricio Moreno-Zambrano (PhD student @ Department of Life Sciences & Chemistry, Computational Systems Biology Lab, Jacobs University Bremen, Germany) and Sergio Grimbs, Matthias S. Ullrich, Marc-Thorsten Hütt

I graduated in 2010 from Escuela Politecnica del Ejercito (Ecuador) with a B.Sc. in Biotechnology. Afterwards, I pursued a M.Sc. in Statistics at the University of Leuven (Belgium), where I got graduated in 2015. The current focus of my research in my Ph.D. in Computational Systems Biology, is the construction of biochemical reaction models of the process of cocoa bean fermentation.

The process of cocoa bean fermentation represents a key step in cocoa processing in terms of development of chocolate's flavor and aroma. Opposed to other fermentation processes in the food industry, it embraces a great complexity since it involves the sequential activation of mixtures of several microbial populations under non-controlled conditions. However, cocoa bean fermentation is a prototypical situation for the application of modelling by means of coupled non-linear ordinary differential equations. Here, a quantitative model of cocoa bean fermentation is constructed based on available microbiological and biochemical knowledge. The model is formulated as a system of eight coupled ordinary differential equations where two types of state variables are distinguished: 1) Metabolite concentrations of glucose, fructose, ethanol, lactic acid and acetic acid, and 2) Population sizes of yeast, lactic acid bacteria and acetic acid bacteria. By the use of a Bayesian framework for the estimation of the parameters, we demonstrate that the model is capable of quantitatively describe existing fermentation time series. Thus, the proposed model is a valuable tool towards a mechanistic understanding of this complex biochemical process.

R packages: rstan, dplyr

Incremental Dynamic Time Warping in R

Speaker: Leodolter Maximilian (PhD candidate @ Austrian Institute of Technology, Austria) and Brändle Norbert, Plant Claudia

I did my masters in mathematics at the Technical University of Vienna. Now I am doing my PhD in cooperation with the Austrian Institute of technology and the University of Vienna. In 2017 I submitted my first package to CRAN.

Section: Time-series, Text mining, Statistics

Dynamic Time Warping (**DTW**, [Sakoe et al. 1978]) is a distance measure for two time series of different lengths that allows non-linear alignments of time series and has been applied in many fields (text mining, activity recognition, transport research, etc.) for detection of motifs, classification and clustering. In contrast to the Euclidean distance, the complexity of DTW computation is quadratic in the number of time points. There are many approaches on speeding

up the calculation of DTW (early abandoning, approximation, lower bound). Many applications generate streaming data, e.g. transport mode detection. For streaming data, an efficient incremental computation of DTW is desirable, which our R package **IncDTW** (Incremental Dynamic Time Warping) facilitates by recycling already computed results. We used the Rcpp package to implement the heart of the DTW algorithm in C++. We collected smartphone accelerometer data while travelling with different transport modes and demonstrate that DTW outperforms the traditional Euclidean distance in detecting transport mode specific patterns, and how the computation time benefits from the incremental calculations. In terms of computation time we compare our implementation with existing ones in R and demonstrate our method to be up to 20 times faster.

R packages: IncDTW, Rcpp, dtw

Reference: H. Sakoe and S. Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 26, 1 (Feb 1978), 43–49. https://doi.org/10.1109/TASSP.1978.1163055

R use in Hydrology: an example with R and the hydrological model GEOtop

Speaker: Emanuele Cordano (Freelancer Researcher @ Rendena100 di Cordano Emanuele (self-employed), Italy) and Giacomo Bertoldi (Eurac Research, Institute for Alpine Environment, Bolzano, Italy, www.eurac.edu), Samuel Senoner

I'm Emanuele Cordano, environmental engineer and hydrological modeller. Since 2011, I have been working with R to analyze hydro-climatic data time series and I became an R enthusiast. i'm author of some R packages, especially in hydrology and climatology. I work as a freelancer for local (Trentino-South Tyrol, Italy) and European (JRC) research using R in projects on water cycle modeling and water resource management.

Section: Reproducible Research, Time-series, Spatial

Eco-hydrological models are increasingly used in the contexts of hydrology, ecology, precision agriculture for a detailed description of the water cycle at various scales: local scale, an hillslope or a watershed. However, with increasing computing power and observations available, bigger and bigger amount of raw data are produced. Therefore the need to develop flexible and user-oriented interfaces to visualize multiple outputs, perform sensitivity analyzes and compare against observations emerges. This work presents two R open-source packages: **geotopbricks** and **geotopOptim2**. They offer an I/0 interface and R visualization tools the GEOtop hydrological distributed model (http://geotopmodel.github.io/geotop/), which solves water mass and energy budget equations to describe water cycle in the Earth´s critical zone. The package *geotopbricks* (https://github.com/ecor/geotopbricks and https://CRAN.R-project.org/package=geotopbricks) is able to to read the GEOtop I/O data of the model. The package *geotopOptim2* (https://github.com/EURAC-Ecohydro/geotopOptim2) calling the **hydroPSO** (https://CRAN.R-project.org/package=hydroPSO) package can be used for model calibration against observations. Further details and complete R package dependencies are listed in *geotopOtim2* description file. As a demonstration example, an analysis of modeled and observed soil moisture and

evapotranspiration time series in some alpine agricultural sites (https://github.com/EURAC-Ecohydro/MonaLisa) are presented.

R packages: geotopbricks, stringr, geotopOptim2, hydroPSO, hydroGOF, shiny and leaflet

Reference: Endrizzi, S., Gruber, S., Dall'Amico, M., and Rigon, R. (2014): GEOtop 2.0: simulating the combined energy and water balance at and below the land surface accounting for soil freezing, snow cover and terrain effects, Geosci. Model Dev., 7, 2831-2857.

EstWeibull: an R package for estimation of parameters for two-parameters Weibull distribution

Speaker: Tereza Konecna (Ph.D. student @ Brno University of Technology, Faculty of Mechanical Engineering, Czech Republic)

My name is Tereza Konecna and I am Ph.D. student at Brno University of technology. I study a statistics, optimization and computer methods of image processing. The topic of my thesis is Generalized spatial models.

Section: Statistics

The Weibull distribution is frequently applied in various fields, ranging from economy, business, biology, hydrology, or engineering. This work deals with the development of an R-package, that aim at the two-parametric Weibull distribution, which is unique by using the method of quantiles and the one-way ANOVA (for example, the packages ExtDist a EWGoF, use the method of moments or estimation by maximum likelihood).

R-package EstWeibull utilizes the Weibull model, exactly the two-parametric Weibull distribution. The package is developed for the estimation of random sample parameters using one of the following methods: method of quantiles, the maximum likelihood estimations, or Weibull probability plot. The package includes the goodness of fit test statistics for fully specified distribution and for composite hypothesis based on EDF statistic - the Kolmogorov-Smirnov and the Anderson-Darling.

The package also contains the parameter estimates by maximum likelihood method in one-way ANOVA type models - estimations for the model with constant scale parameter, constant shape parameter and the model with both parameters constant. The tests with nuisance parameters are included too - namely the score test, the Wald test, and the likelihood ratio test, with null hypothesis that the parameter scale or shape is constant. We use stats and rootSolve packages.

References:

- KONECNA, T., Model with Weibull responses. Brno: Brno University of Technology, Faculty of Mechanical Engineering, 2017. 64 p. Supervisor doc. Mgr. Zuzana Hubnerova, Ph.D.
- MURTHY, D. N. P.; M. XIE; R. JIANG., Weibull models. 4. Hoboken, N.J.: J. Wiley, c2004.

Distinct operative molecular processes across breast cancer subtypes.

Speaker: Daniel Tiezzi (PI @ 1. University of São Paulo; 2. FATEC, Brazil) and Francisco Jose Candido dos Reis; Jurandyr Moreira de Andrade; Lucas Baggio Figueira

PhD in Medical Sciences (2007) PI - University of São Paulo - Breast Disease Division (since 2008) TCGA PanCan and Cervical Cancer AWGs Graduate Student - Software Development - FATEC (since 2017)

Section: Bioinformatics

Virtually all malignant tumours carry somatic mutations. Distinct mutational processes tumour have been exposed imprint specific structural alteration in the DNA. Recent studies have reported there are at least 30 signatures associated to underlying mutational processes in cancer. Breast cancer is a heterogeneous disease based on molecular profile (transcriptome). However, it is not clear whether mutational processes each tumour have been exposed can differentiate then in terms of biological behaviour. We used the *SomaticSignatures* and *deconstructSigs* packages in **R** over WES data to infer mutational signatures associated to breast carcinomas from TCGA database and used an unsupervised hierarchical clustering to identify samples exposed to similar mutational processes. We identified 8 operative mutational processes (S1 - S8) across 1,068 breast cancer samples. S1, S2, S3 and S8 are highly correlated to COSMIC signatures 2/13 (APOBEC activity), 1 , 3 and 5, respectively. We applied hierarchical clustering to the mutational spectrum matrix by samples. Most samples (356) clusters at C4. The C4 signature is enriched by S3 (COSMIC 1) signature, associate to age related endogenous mutational process initiated by spontaneous deamination of 5-methylcytosine. Luminal A and B tumours preferentially cluster at C4. On the other hand, Basal-like and HER2 tumours cluster preferentially at C7 and C2, respectively. The C7 is enriched for S7 signature (strongly associated with germline and somatic BRCA1 and BRCA2 mutations). The C2 is enriched for S1 signature (activity of the AID/APOBEC family of cytidine deaminases). This is the first report demonstrating specific breast cancer subtypes are exposed to distinct molecular processes.

R packages: SomaticSignatures, deconstructSigs, BSgenome.Hsapiens.UCSC.hg19, GenomicRanges, iRanges

References: Nik-Zainal *et al.* Mutational processes molding the genomes of 21 breast cancers. Cell. 2012 May 25;149(5):979-93.

Reproducibility of data analysis and reporting with R, R Markdown, knitr and LaTeX using the example of the standardized Austrian school leaving exam

Speaker: Michael Themessl-Huber (Psychometrician @ Federal Ministry of Education, Science and Research, Austria) and Philipp Gewessler, Jan Steinfeld, Larissa Bartok, Martina Frebort

Studies of psychology (MSc.) and statistics (BSc.) at the University of Vienna. Work experience: Junior Statistician at Medical University of Vienna for the nutritionDay project; Psychometrician at the Federal Ministry of Education, Science and Research.

Section: Social Sciences, Reproducible Research

This contribution outlines a process that makes all analysis and reporting steps reproducible and comprehensible. In addition, this procedure also allows flexible adaptation of analyses to individual requirements. The data analysis process is presented on the basis of the quality assurance process for task development in the context of the standardized school leaving exam in

Austria. One aspect of quality assurance is the collection of data from pilot studies in which tasks are tested on a representative student population (BIFIE, 2013). After the respective exam, a post-test-analysis (pta) is performed, in which the exam results and given tasks are analysed. The interim and final data analyses require a flexible and quick reporting. In order to make data analyses and reporting independent of the data analyst, R packages were developed for both the piloting and the pta. They contain functionality for reading, cleaning and checking the (raw) data, as well as generating reports automatically. The reporting features are designed to be flexible enough to add or remove sections by specifying the appropriate function arguments to suit the needs of different recipients. Git was used as versioning tool. The documentation was written with bookdown, whereas the automatic report generation was implemented using LaTeX and knitr. In this poster presentation, the cycle of the report generation will be shown, technical details will be explained, the analyses steps will be clarified and exemplary tables/figures from pilots and pta reports will be presented.

R packages: bookdown, devtools, ggplot2, knitr, rmarkdown, xtable

---

Differential impact of anthropogenic pressures on Caspian Sea ecoregions

Speaker: Matteo Lattuada (PhD candidate @ Systematics Justus Liebig University, Germany)

I am an outgoing young researcher with a great interest in biodiversity conservation and sustainable development. I love to do research in these fields using R as my most trustful companion.

Section: Spatial

In the Caspian Sea, a renowned endemic rich ecosystem, species likely diversified in space along various abiotic parameters such as temperature, salinity and depth. In a previous study, these parameters were used to infer 10 ecoregions, which are nested into three physico-geographically defined sea areas: i) North Caspian ii) Middle Caspian, and iii) South Caspian. Recently, a biodiversity decline probably caused by a combination of natural and anthropogenic pressures, was reported from all the three sea areas. In this study, we analyzed the anthropogenic impact of 9 human-derived pressures in the 10 ecoregions by the use of Cumulative Effect Assessment (CEA) methods. Aggregating the results in the sea areas, we found that the North Caspian shows higher average CEA scores compared to the Middle Caspian and the South Caspian. Furthermore, we detected differences in the anthropogenic pressure contribution to the CEA score among the sea areas: the North Caspian is mostly affected by poaching (46% of the CEA score), whereas the Middle and South Caspian by pollution (45% and 51%). This was further supported by the average ecoregion CEA scores, where the highest impact appears in the transitional ecoregion between the North and the Middle Caspian. This can pose a risk for the survival of the endemic species with the distribution limited only to the North Caspian. We propose that the potential explanation for the spatial pattern of the CEA score may be the pollution removal by water currents and the concentration of anthropogenic activities near river mouths. Finally, our study can be used as baseline for a reproducible spatially explicit environmental monitoring with the aim of implementing ecosystem management plans focused on the sustainable use of the Caspian Sea resources.

R packages: Raster, rgdal, gstat, sp, ggplot2

Notes

Notes

Platinum Sponsor

RStudio®

Gold Sponsors

MANGO SOLUTIONS

QUANTIDE
knowledge from data

Silver Sponsors

H2O.ai

emarsys

Open Analytics

BUDAPEST BI FORUM

Bronze Sponsors

WLOG SOLUTIONS

Jumping Rivers

R consortium

R-bloggers

Shift

R


Budapest
eRum 2018