Abigail Lebrecht

Mumsnet

Talk title:

Stopwords personalisation and Text cleaning at Mumsnet

About:

Abigail has been a Data Scientist for nearly a decade and using R for even longer. She started using R while studying for a PhD in Queueing Theory from Imperial College London. She is currently Principal Data Scientist at Mumsnet and has experience working with a wide variety of types of data across numerous industries, particularly e-commerce and marketing.

Abstract:

Textual analytics and natural language processing techniques are powerful tools that give businesses a deeper understanding of users' views and help improve user experience. These techniques can derive sophisticated meaning and insight from text, but require clean textual data. Often, only a generic pre-defined list of stopwords is advised as a sufficient method to remove irrelevant words and phrases from text to optimise analysis. However, this is not sufficient for conversational textual data on social networks or message boards. Data scientists spend at least 60% of their time cleaning data, yet most tutorials focus on analysis techniques, rather than developing efficient textual cleaning methodologies.

This is a particularly complex issue at Mumsnet, the UK's largest website for parents and 10th biggest social network with 14 million unique monthly users. Mumsnet users are known for developing their own acronyms and unique language usage when posting. This makes standardised stopwords lists insufficient for effective textual analysis. In this talk we will share how we have developed solutions to this problem using R at Mumsnet, including techniques to personalise stopword lists and optimising our approach to cleaning textual data, to create a fast and standardised approach.

Alexis Iglauer

PartnerRe

Talk title:

Medical underwriting triage: An end-to-end machine learning case study

About:

Alexis leads the Life & Health Analytics team at PartnerRe, which provides data science expertise to internal and external clients across the globe. He enjoys being on the intersection of data science, technology, actuarial science and business. Alexis grew up in Pretoria, and now lives in Zürich.

Abstract:

The output of PartnerRe's data science team is often deployed to production. Examples include automated reports used by decision makers, tools for reinsurance pricing (e.g. simulation engines) and trained machine learning models that support work flows in our company. In our context, production' means that something is used by non-technical users in a business critical context. Consequently, production content needs to be highly available and users expect complete reproducibility of all results.

To make R fit for this purpose, we have developed a set of best practices around how we structure and version control our code and the environments our code runs in. In this talk, we focus on the open source infrastructure supporting this process: version control, continuous integration, containerisation and container deployment and orchestration.

Amanda Beedham

RSA Insurance

Talk title:

Harnessing AI to Create Insight from Text

About:

In her 20 year career, Amanda Beedham has provided predictive analytics consultancy and training across multiple industries and has embraced the development and adoption of Data Science.

Amanda redefined a customer loyalty programme for a North American airline using segmentation, she delivered purchaser and redeemer models for an international asset management company and used hierarchical time series to evaluate sales for an automobile paint manufacturer.

Amanda now works as a Data Scientist at RSA where she has developed and deployed leading edge GBMs to insurance problems. Amanda's current focus is applying the next generation of text analytics techniques at RSA.

Abstract:

Word Embedding is outperforming older methods of Text Mining by removing the laborious process of building dictionaries of related words.

Facebook's Starspace has become recently available in R alongside Text2vec's Glove Embeddings. These libraries allow users to build and apply neural language models natively in R.

At RSA, we have been investigating Word Embedding techniques on a variety of use cases and have been amazed by the results. The aim of this talk is to demonstrate that Word Embedding is relatively easy to perform in R and to showcase why these techniques are so powerful.

The following use cases will be discussed:

• Invoice data - predicting spend category from invoice line text using Starspace Tag Embedding.

• Invoice data - using invoice line text to predict cost using Starspace Embeddings and GBMs.

• Claims data - identifying claim type using GloVe Word Embedding and clustering.

The talk will discuss how to create supervised and unsupervised embeddings in R and how these can be input into GBMs (for regression modelling) and clustering (for grouping documents). The discussion will also show how to find associated words and documents and how to visualise embeddings using T-Sne.

Amit Kohli

ACDI/VOCA

Talk title:

R is for Reconciliation: how machine learning is helping Colombia move forward

About:

Amit is an environmental engineer turned data strategist. He started his career collecting and using environmental data to clean soil and water resources, then disseminated water resources & uses data with UN-FAO's AQUASTAT and now leads Data initiatives @ ACDI/VOCA. Amit's commitment to international development comes from his experience growing up in 'developing countries' and witnessing every extreme in living conditions.

He is passionate about R and is dedicating to promoting its dissemination having led R-User groups in Ghana and the UK. He is the author of the ShinyTester, TileMaker, and BulletChartR R packages and blogs at https://www.amitkohli.com and http://www.acdivoca.org/data-digest.

Abstract:

With a final peace deal between the government and the Revolutionary Armed Forces of Colombia (FARC) signed in November 2016, Colombia is currently turning the page on half a century of armed conflict. Despite the promise of peace, tremendous difficulties persist in conflict-affected regions. Against this backdrop, the USAID-funded Program of Alliances for Reconciliation (PAR), implemented by ACDI/VOCA, accepts the challenge of carving out a space for reconciliation, bringing together actors from all corners of society and transforming mentalities marked by war, intolerance, and fear.

But what does reconciliation mean? To some it means revenge, to some it means justice, to some it means monetary compensation, or honouring the dead, or moving on as a country. In order to better understand and respond to these sometimes oppositional needs and seek to generate changes in perceptions, attitudes and behaviors, the "DecidoSer" (I decide to be) psychometric approach is developed. Thematically, "DecidoSer" develops four drivers: respect, trust, empowerment and dialog.

This talk will document how machine learning created a two-way highway of technique validation; on the one hand, helping identify which techniques have the most impact, and on the other which techniques required reformulating, rethinking or eliminating for some demographics.

Amit Arora

Hughes Network Systems

Talk title:

Exploring and predicting network service availability using R

About:

After working in telecommunication software for 16 years I switched to data science. I pursued M.S. in Data Science for Georgetown University, completed in Dec 2018. Currently I work as lead data scientist at Hughes Network Systems. I have setup a data science practice at Hughes. Along with my team I analyze realtime streaming data from thousands of Terminals, create models for predicting network service availability, predictive analytics and also analyze data from business systems (NLP, object detection).

R is an essential part of our data science toolkit, RStudio Connect is now our standard platform for sharing data science apps.

Abstract:

Hughes Network System uses tools and technologies available in the R ecosystem to create an enterprise grade application for exploring and predicting broadband service availability throughout the United States and Canada.

The availability of historical broadband service availability data has provided a rich dataset that is now mined using R for both exploratory analysis (e.g. geospatial heat-maps) as well as predictive analysis (predict best suited ISP and technology given service requirements).

The application stack is cloud native and consists of the entire data science pipeline from ingestion to machine learning to an end user facing application. Each building block is containerized (containerit, rocker) and runs inside a kubernetes pod on the Google Cloud Platform(GCP). All interface to GCP is through packages available via the "cloudyr" project. We use tidyverse for exploratory analysis, plotly, leaflet, htmlwidgets for visualizations, H2O for machine learning (ML) and plumber for providing an API endpoint for the ML models. A Shiny application hosted inside RStudio Connect provides a front end to the end user. Horizontal scaling, logging and security are provided through native GCP constructs.

This application thus presents a simple yet an extremely effective example of a cloud native production application written completely in R.

Ana Henriques

Talk title:

Using R in Production at PartnerRe

About:

Ana Henriques is the Analytics Tool Lead in PartnerRe's Life & Health Department. Originally from Portugal, with a background in Pharmaceutical Sciences, Ana moved to Switzerland where she initiated her career in Data Science. Ana is now focused on business-side delivery of platforms and tools to support data science and related functions.

Abstract:

The output of PartnerRe's data science team is often deployed to production. Examples include automated reports used by decision makers, tools for reinsurance pricing (e.g. simulation engines) and trained machine learning models that support work flows in our company. In our context, production' means that something is used by non-technical users in a business critical context. Consequently, production content needs to be highly available and users expect complete reproducibility of all results.

To make R fit for this purpose, we have developed a set of best practices around how we structure and version control our code and the environments our code runs in. In this talk, we focus on the open source infrastructure supporting this process: version control, continuous integration, containerisation and container deployment and orchestration.

Andreas Cardeneo

Allianz Lebensversicherungs-AG

Talk title:

The documentation is the code: Reproducible workflows and reliable decision making

About:

I am a data scientist by profession, an operations research guy by conviction, a computer scientist by experience and an industrial engineer by education. I am a father of two, a swimmer and an espresso drinker. Over the last years in data science I learned to appreciate the R language and its community that encourages me to contribute.

Abstract:

As data analysts and data scientists we communicate retsults and recommendations based on data analyses and models to a diverse set of stakeholders. While these might not be interested that much in the algorithmic and mathematical details, they tend to question the underlying data, its preparation and the whole analytical process.

We frequently experience this situation with both internal and external stakeholders from management, business, IT, data protection, regulators, and auditors and have learned that only a well documented and understood process will yield the trust into our conclusions necessary for decision making.

In this presentation, I would like to show and explain how we use systems, tools, conventions, and processes to address reproducible analytics and reliable decision making. Here, I refer to reproducibility as the analyst's ability to recreate analytical results while reliable decision making refers to the addressee's ability to confidently base decisions on the analytical results.

In particular, I will present an R package that supports generating R workflows from declarative data and workflow descriptions in markdown documents. Besides technical aspects, I will highlight the application of the package in analytics projects and show how it facilitates addressing non-technical stakeholders.

Andy Nicholls

GSK

Talk title:

Making Better Decisions

About:

Andy Nicholls works in the specialist Statistic Data Science unit at GSK, helping Biostatistics to become are more Data Science-focused organisation. He is an active member in the R community, co-authoring R in 24 hours and leading the R Validation Hub initiative within the Pharmaceutical industry.

Abstract:

Trial design has been a key focus in the pharmaceutical industry for a long time. But the discussion is often around how design choices affect immediate outcomes. For example, will the trial deliver a significant p-value on the primary endpoint? In practice the decision as to whether to progress a drug is not as simple.

Annika Westphal

ORX Association

Talk title:

Making R's data visualisation capabilities available to R non-users

About:

I'm a Statistical Assistant Manager at ORX Association, the leading financial services operational risk association. As part of my role, I conduct research on the quantification of operational risk, including for capital calculation and stress testing purposes. I'm also leading a research stream on the use of machine learning techniques for operational risk. I use R for data analysis and visualisation, to gain insight into unstructured data through machine learning, and to automate reporting. I also build user applications that make R's data analysis capabilities available to non-users in my company.

Abstract:

One of R's key strengths are its graphic and charting capabilities – creating high quality charts that capture the reader's attention. A current challenge is finding people with the right skillset as knowledge of R becomes increasingly more in demand across multiple industries.

This presentation shows how R's data analysis and visualisation capabilities can be made available to non-users with the help of shiny. As part of our research on operational risk, ORX Association regularly conducts surveys among operational risk professionals. The focus of this presentation will be a case study on the practical application of R, showing how we've built an internal data analysis tool. The tool allows the user to directly import the data from our third-party survey tool; performs the necessary data analysis; and provides the user with a wide number of options to customise graphs.

The presentation will highlight the benefits that this data analysis tool has provided to our business. R has allowed us to become more effective and efficient through speeding up our data analysis process; enabling a wider number of people in the company to use R's data visualisation capabilities; and to deliver our research at pace.

Avision Ho

Department for Education

Talk title:

Why moving from a Nobel Prize algorithm to a simpler, bespoke one isn't a bad thing for business

About:

When not cooking and baking in his spare time for his Instagram, Avision can be found searching for things to do to twiddle his thumbs, including working on data science projects and sharing these on his GithHub page, reading books, watching film and creating memes.

Abstract:

Consider that you have a set of people going to a conference and a set of talks taking place simultaneously. Each person cannot go to more than one talk but they have preferences over which talk they want to attend. How do you efficiently match people to talks based on their preferences?

This problem was based on a real example where 400 people with preferences had to be matched to 5 talks of limited sizes. The presentation will cover how this was tackled first through the Nobel Prize-winning Gale-Shapley algorithm, but due to evolving business requirements, led to this being shelved in favour of a more bespoke algorithm built in R specifically for this problem. It details the business process to obtain buy-in and how an R Shiny web application was developed to ensure this algorithm could be used longer-term by the events team who were not R coders.

This talks covers aspects such as obtaining buy-in from risk-averse stakeholders on the benefits and risks of automating existing business processes; working closely with them within an Agile framework to keep them informed, involved and on-board; and communicating with wider users to build consensus and trust in the matching algorithm.

Ben Travers

Stephens Scown LLP

Talk title:

Successfully Preparing your R-Based Product for Investors or Sale: A Legal Perspective.

About:

An innovative, leading IP & IT Lawyer, Ben Travers qualified in 2006 and heads up Stephens Scown LLP's Intellectual Property and Information Technology team. He founded the team in 2012 and it has grown to become one of the largest specialist teams in the UK. Ben advises on contentious and non-contentious matters. He is active in the technology marketplace, helping innovative and ambitious businesses to achieve organic and sustainable growth.

Abstract:

Businesses are increasingly turning to statistical computing, analysis and machine learning tools generated in R to carry out research and development or to gain incite into their consumers' behaviour, patients' symptoms, market trends, etc. from data they have gathered. This information is even more valuable than the data it stems from. But, with a mix of free and open source software packages in the R libraries, assembled in a free environment, how can you ensure that your product is ready for market, compliant with applicable laws and that your exposure to risk is minimal?

This talk will give you incite into what your clients and investors' professional advisors perceive as key risks when purchasing or investing in your R-based products and the ways in which you can mitigate those risks. We will review questions of IP ownership, licensing and security. We will also discuss the ownership of materials generated using your programs, focusing in particular on the output of machine learning and artificial intelligence.

This talk will not give specific legal advice and you should seek independent legal advice in relation to your business ventures.

Ben Byrne

Roche Products

Talk title:

Can R-Shiny get drugs to patients faster?

About:

Ben is a statistical programmer with a keen interest in advanced analytics. Ben studied a master's in applied statistics, specialising in machine learning after securing a scholarship with Cardiff University, graduating with a distinction in 2017. Ben also graduated with a first-class undergraduate degree in Mathematics, Operational Research and Statistics with a year in industry at Roche.

Following his postgraduate studies, Ben rejoined Roche as a Statistical Programmer, analysing clinical trial data using SAS and R.

Ben is an avid Portsmouth football fan, although he finds supporting Portsmouth as much of a struggle as the most challenging bit of code!

Abstract:

This presentation will cover a case study from an Oncology molecule filing team throughout Roche's earliest FDA RTOR (Real-Time Oncology Review) application, where utilization of R-Shiny interactive data displays helped to influence a faster filing. This new way of working had benefits prior to final analysis, during this time, and also after. Due to the added interactivity provided by R-Shiny, ahead of the final analyses we were able to spot critical issues in the data/programming/derivations which, if missed, could have put the filing at risk! Then after unblinding we presented the IDDs to key scientific stakeholders in addition to our static TLG outputs, and utilized them in the statistical and clinical interpretation meetings, allowing dynamic real-time answers to their exploratory questions at the click of a button. Following on from this we used the IDDs to significantly cut down the number of extra unplanned requests needing to be programmed for the clinical study report and filing documents, saving time and effort. This presentation and live demo of our App will provide our experience of IDDs within a key study readout, as well as recommendations and lessons learned, including speculation as to where the industry could be heading with this technology.

Charlotte Wise

Essence

Talk title:

Beyond the average: a bayesian approach for setting media targets

About:

I manage a small team of analysts at Essence, a global media agency and part of GroupM, WPP. I hold a degree in PPE and an MSc in Economics, both focused in econometric modelling. I was introduced to R by my first manager, an R enthusiast with a contagious passion. Over the years since I have refined my R skills, specialising in data visualisation along with econometric and machine learning modelling solutions. I'm a keen advocate of R in my current team, and run regular internal "tidyverse and beyond" training sessions on the modern use of R within data science.

Abstract:

Marketeers can find it difficult to determine whether a media campaign has reached its full potential, which is incremental to optimising future performance. It can be tempting to benchmark your performance against an average of the past, but with multiple factors contributing to campaign success, and a scarce number of historical campaigns to benchmark against, this can quickly become meaningless.

At Essence, this problem is exacerbated through the global span of our campaigns and large variety of products we advertise. Even if an averaging approach got to the bottom of setting targets for a single product in the UK, where we have a plethora of historical data, what about in LATAM, where the media landscape is rapidly changing? Or if we have a large number of products, and relatively little historic data on each?

Our approach provides a bayesian solution to this problem. We implement a hierarchical bayesian model to estimate 'benchmarks' for media performance across multiple markets, for multiple products. Using tidyverse functionality and brms (an R-wrapper for STAN), we illustrate how we determined key trends then built and cross-validated our model. Our aim is to demonstrate how hierarchical bayesian inference can overcome common business problems of data sparsity.

Chris Billingham

MAG-O

Talk title:

Battle of the Bands - Starring Tidytext and Tensorflow

About:

Chris is Lead Data Scientist with MAG-O, Manchester Airport Groups in-house digital agency with a recent focus on NLP and Online Attribution. He lives in Buxton with his wife, Sarah, and two children, Lyra and Leo. In his spare time he likes run, then obsess over his running data.

Abstract:

Using the combined power of Tidytext and Tensorflow, see how you can turn these titans of Data Science to one of the more intractable problems in the world of music. Following the problem from its very inception to its ultimate end, you'll come on journey of knowledge and understanding, whilst enjoying some gifs along the way.

Chris Mainey

Healthcare Evaluation Data (HED) - University Hospitals Birmingham NHS FT

Talk title:

Driving R adoption in an NHS information service, barriers and solutions

About:

I'm a analyst, staitsician and data scientist for the NHS. With background in SQL-based analysis and BI tools, I took a job with Healthcare Evaluation Data (HED), that required a signficant stats component. Currently completing

Abstract:

Healthcare Evaluation Data (HED) is an online NHS benchmarking using national data. We provide web-based interactive reporting tools for organisations to compare performance, featuring various statistical modelling approaches and live manipulation of large datasets. We have relied on SAS for many of our models, but are transitioning to R for modelling, analysis work and developing support material.

This talk will explain how we are moving from one keen user, to wider adoption, focussing on:

- Setting up R in a 'locked-down' NHS environment
- Best practise for building statistical models on large data sets with limited hardware
- Spreading use of R through package development and training
- Encouraging R use for analysis and ad-hoc reporting
- Open-source principles, information governance and intellectual property

Barriers we have faced include scepticism about security risks, unhelpful hardware and network settings, a perception that R was not fit for production, fear of losing SAS at the 'standard,' and the learning curve for new R users.

We have embarked on a structured training program for our teams and have set up a local R user group. We are aiming to further engage with the R community as user confidence grows.

Christel Swift

BBC

Talk title:

Building a shiny app to show affinity between programmes

About:

Christel Swift has spent her career in music and media measurement. She was an elected committee member of the Media Research Group and was technical advisor for the UK media currencies for TV, Radio and Outdoor for many years.

She has been working at the BBC for the last 6 years, in Marketing Science and in Data Science. In her current role as Senior Data Scientist in the Central Insights team, she works across the various BBC products (TV, Radio, News, Homepage, Children...) with both analytics and survey data, helping stakeholders understand their audience better.

Abstract:

The BBC has a wealth of audio and video assets, ranging from Radio 4's Gardener's Question Time to BBC1's Strictly Come Dancing. Wouldn't it be great to find out what radio programmes viewers of Eastenders are most likely to be interested in? Or to find bundles of Comedy programmes that go well together? This can be done with a heady mix of Market Basket Analysis, Network visualisation and Shiny.

David Smith

Microsoft

Talk title:

A DevOps process for deploying R to production

About:

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.

Abstract:

So you've built an amazing model in R. It generates great predictions on your desktop. Now, how do you get it into production?

Deploying R functions within Docker containers, and exposing them with the 'plumber' package, is a simple and effective way to integrate R into applications via a REST API. We can further provide scale for high-volume workloads by deploying that container to Kubernetes. As the complexity grows, however, managing and updating deployments can be challenging.

In this talk, we describe a CI/CD based process in Azure DevOps to automate the entire build, test and deploy process for R-based models in production. The process emphasizes model reproducibility, by capturing and tracking changes in the code, data, tests and configurations that define the model. You will learn how, with a check-in to GitHub as the trigger, your model can be automatically retrained, optimized, built, validated, and — with human approval if necessary — released. Once deployed to the production environment — in this talk we'll focus on scalable open-source frameworks like Kubernetes — the model will be subject to continuous monitoring and performance tracking until such time that a new model version is warranted, and the DevOps lifecycle begins again.

David Baker

Toynbee Hall

Talk title:

haRnessing The Open Source Community to Help Local Communities

About:

David John Baker is a music researcher and educator passionate about research questions at the intersection of music theory and music science. His research looks to understand how the people learn melodies in order to improve pedagogical practices in aural skills education. His overlapping quantitative skill set with the world of data science has also led him to both music industry projects and work in charity. Over the past year he served as the Research and Evaluation Residential Volunteer Worker at Toynbee Hall in London, England.

Abstract:

Since its inception in 1884, evidence based research has been central to Toynbee Hall's mission as a charity throughout East London communities. Over a century ago we played a key role in Charles Booth's creation of the first data visualizations for public good publishing a series of poverty maps in 1903. Over 100 years later, we still are engaging with our local community to solve problems, but now utilizing R to enable us to accomplish our goals in an efficient, cost effective manner.

In this talk, I tell the story of how Toynbee Hall has adopted R as a tool within the charity sector. Not only has R allowed for rapid analyses of data on a diversity of projects, but I also show how embracing open source software allowed for our team to host a series of data hackathons that allowed us to recruit freelance data scientists to help analyze publicly available datasets to contribute to building materials that we use in our policy advocacy campaigns. I document both the successes and potential pitfalls our team encountered during this time and strongly advocate for using R as a means to grow a community of volunteer analysts for public good.

Detlef Nauck

BT

Talk title:

Model Factories and Test-Driven Machine Learning

About:

Dr Detlef Nauck is a Chief Research Scientist for Data Science with BT's Research and Innovation Division located at Adastral Park. Detlef has 30 years of experience in data analytics, machine learning, and AI. At BT, he is leading a group of international research scientists working on improving the use of Data Science. Detlef focuses on establishing best practices in Data Science for conducting analytics professionally and responsibly leading to new ways of analysing data for making better decisions. Part of his role is leading the initiative on the development and use of responsible and ethical AI in the company.

Abstract:

Data Scientists and machine learning specialists are familiar with testing principles during the model building phase like cross-validation, but they are often unfamiliar with test-driven software engineering principles. While testing a learned model gives an idea how well it might perform on unseen data this is not sufficient for model deployment. Trying to learn from test driven software development practices we look across the machine learning life cycle to understand where we need to test and how this can be done. The testing of data, for example, is essential as it not only drives the machine learning phase itself, but it is paramount for producing reliable predictions after deployment. Testing the decisions made by a deployed machine learning model is equally important to understand if it delivers the expected business value.

To operate test-driven machine learning in production we look at the concept of model factories. They add an orchestration layer that automates as much of the testing and model building as possible and supports governance.

We look at examples of R in production at BT to illustrate the challenges that have to be overcome in complex organisations or use cases and show how a model factory can be assembled.

Doug Ashton

Mango Solutions

Talk title:

Rapid Reproducible R Projects

About:

Doug joined Mango Solutions in 2014 after a research career in statistical physics at Bath and Utrecht. He has gone on to specialise in machine learning methodology. As a data science consultant Doug has applied these techniques for numerous clients, as well as developing training courses on machine learning workflows and tools, such as Keras.

Abstract:

Often times data science has to happen quickly. This could be because your project is a proof-of-concept, or perhaps there is a limited window of opportunity. When the pressure is on, reproducibility tends to suffer. This leads to lost knowledge, is harder to scale, and the cycle starts again the next time you're under pressure. During a recent competition we organised an R project, with multiple team members, to generate a number of sales forecasts in a short space of time. This talk will cover the principles we followed to balance reproducibility with creativity, sharing knowledge, and evaluating all the models we produced. We'll look at the tools we used from the R world. RStudio notebooks for exploratory work, rOpenSci's Drake for data pipelines, R packages for reusable, testable code, and GitLab providing the project hub where everything was brought together.

Duncan Garmonsway

Government Digital Service

Talk title:

Get good data out of bad spreadsheets: tidyxl and unpivotr

About:

I am a data scientist in the UK Government Digital Service and have previously worked for Scottish Power and the New Zealand public sector. I studied English Literature at University College London and Applied Statistics at Victoria University of Wellington.

I wrote Spreadsheet Munging Strategies, a guide to getting data out of complex spreadsheets using my R packages tidyxl and unpivotr. I am an ROpenSci reviewer.

Abstract:

Unlock legacy data from spreadsheet prison with the R packages tidyxl and unpivotr. Multi-layered headers? behead() them one at a time. Meaningful colours? Turn colour into data. Sweat no more over error-prone, laborious and demoralising scripts. Create robust, flexible, tidyverse-friendly pipelines and get payoff from importing your first spreadsheet into a tidy R dataframe.

Eduardo Contreras Cortes

Ernst & Young LLP

Talk title:

Words that will inspire

About:

Eduardo Contreras Cortes is a data science manager at EY. Eduardo has more than 10 years of experience in banking and consulting including companies like Lloyds Banking Group and Citigroup. He holds a Bachelor in Actuarial Science and a Masters Degree in Statistics from the University of Edinburgh

Abstract:

A project that analysed 2,500+ TED talks using text analytics and machine learning with R to find the factors that make some talks more popular than others and provide suggestions on how to improve your public speaking

Edward Watkinson

Royal Free London (NHS Foundation Trust)

Talk title:

R in the Hospital - Starting the journey

About:

I am an analyst, economist and data scientist currently working for the Royal Free London Hospital Group, after previously working in the NHS at a national level, and as a web developer in manufacturing.

I am interested in how data analysis can improve patient care. I use analysis and prediction to inform decisions about healthcare provision, and R has massively helped. I use R to both streamline/automate tasks; as a significant part of the data science pipeline to dig into complex problems; and to model and predict changes in the healthcare we provide to patients.

Abstract:

Many non-profit organisations want to harness the power of data science, but most (especially the NHS) don't have the funds to invest in large-scale roll-outs of tools. In this talk we share the first steps in our journey to adopt R as the core tool on our analytical workbench for helping to run our hospitals, and show how useful it has been in the cash-strapped NHS. We will share how we started small, and persuaded analytical staff and decision-makers of R's power, including brief examples of our first 'showcase' projects. We will also share some of the challenges we had and how we overcome them, and our vision for R in our hospitals in the future.

Elena Furlan

Aviva

Talk title:

Using fastText-based embeddings to categorise online search queries

About:

I am a Data Scientist at Aviva, working with a team of six on building machine learning solutions to deliver personalised marketing content to our customers, particularly through online channels.

I have a background in Statistics and 5 years of analytics consulting experience across different industries.

Abstract:

Product-related text queries to online search engines are strongly indicative of the users' needs and intent, and are thus an important source of information for developing personalised offerings and customer experience, as well as achieving budget optimisation in paid search. However, extracting useful features from such queries is a challenging task due to their amount, semantic similarity between some of the queries, considerable noise caused by misspellings, etc. In this talk, we will demonstrate how word embeddings obtained with the Facebook's fastText algorithm can be used to automatically categorise thousands of distinct search phrases related to insurance products into semantically meaningful and business purpose-relevant groups. We will also explain why we chose fastText for our task by comparing it with other natural language processing techniques, such as word2vec-based embeddings and topic analysis.

Gwilym Morrison

Royal London

Talk title:

From Data to Deployment: overcoming the challenges of embedding R models in Production

About:

Gwilym currently leads a Data Science team in Royal London's Intermediary business. His background is in Molecular Biology, but having made the transition to business from academia has since worked in a number of industries, including Telecoms, Construction, Banking and Insurance.

Abstract:

In many ways, creating models and analysis is the easy part of a successful Data Science project. The tough bit is getting the artefacts produced into production, particularly if the use case in question requires real-time decisioning. These challenges include selling the solution to senior managers and colleagues in IT, architecting environments in which to deploy those models and overcoming technical challenges relating to models in production. In this talk you'll hear about how Royal London has deployed Machine Learning models to revolutionise Life Insurance Underwriting and how it has overcome the challenges of bringing those models through from data to deployment and successfully made a difference to the business and its customers.

Harold Selman

Ordina

Talk title:

Running R in a Spring Boot 2 application using GraalVM at the Dutch National Police

About:

Harold Selman is Lead Data Scientist at Ordina in the Netherlands. With a background in applied mathematics and education, he started his career four years ago at Ordina as Data Scientist. He learned that explaining his work to others is at least as important as the work itself. Currently, he is working on a project to make digital dossiers with smart document analysis to support the Dutch National Police. Before, he worked on several projects within fraud detection and profiling.

Abstract:

A lot of the best Data Science is done in R, but getting R to run as a streaming application in a complex environment using only open source tooling is an absolute nightmare. Making R interface with Kafka, or S3, or anything else, means giving up version control to external packages, while letting R talk to external services through a Docker container in networked VMs requires near superhuman resolve. As Data Scientists we woke up screaming, filled with Docker-induced-terror at the Kafka-esque transformation of their job. Our daily work rapidly morphing into "anything but" our core competency... It was time we found another path. Then a colleague mentioned he had heard something about integrating R into a Java application... Could this be our Salvation? Our Light in the Darkness? Our Holy GraalVM?

In this talk, we will explore how we brought GraalVM into practice at the Dutch National Police by integrating R code into a Spring Boot 2 application written in Java and Scala. We will discuss how GraalVM and FastR helped bring our Data Scientists back to the work they enjoy, while leaving the stream integration to languages and frameworks much better suited to the job.

Hasnain Mahmood

London Clearing House (London Stock Exchange Group)

Talk title:

Risk Management using Behavioural Analysis in R

About:

Hasnain is a Senior Quantitative Associate at LCH, currently residing in the Change and Innovation stream which is part of the In-Business Risk Management team. His day-to-day activities revolve around risk modelling, research and development of risk and business analytics.

He has over 5 years in Quantitative Finance. Specialising in automated trading strategies, modelling and forecasting weather influence on financial products in the Energy, Oil and Gas industry.

Abstract:

LCH (London Clearing House) is a rates and multi-asset clearing service provider, part of the London Stock Exchange Group. We increase the stability and efficiency of the global financial market by guaranteeing the settlement of trades between counterparties by providing risk management capabilities.

Put simply, we make financial markets safer.

Margining is a risk framework that mitigates counterparty risk. We collect collateral from trading counterparties to cover their risk in the market. Margins are calculated as nominal amounts for each counterparty's portfolio against current market conditions.

We conducted a quantitative research study to forecast the margins charged to counterparties. We already had a good understanding of the relationship between counterparty positions and margins, however our goal was to identify unique factors that drive counterparty trading behaviour. Our project had both pioneering and progressive goals as we increasingly seek to drive business decisions with data science.

Our quantitative research team uses an R-focused technology stack. We use BitBucket for version control, and our code reviews are supplemented with Confluence for documentation regarding the methodology and implementation. All of our codes have been productionised to run on internal servers and are scheduled using Jenkins to create an automated system.

Hayfa Mohdzaini

Universities & Colleges Employers Association (UCEA)

Talk title:

Delivering a GDPR-compliant pay benchmarking service without breaking the bank

About:

Hayfa is passionate about communicating complex data in simple terms and enjoys finding ways to make the data analysis and publication process more efficient. A few years ago she convinced her employer Universities & Colleges Employers Association (UCEA) to use R to automate production of their pay benchmarking reports in Word and Excel. Hayfa started her career in IT, testing software at IBM, but has since spent much of her career either working in HR or writing for HR practitioners. She has a bachelor's degree in computer science and masters in HR.

Abstract:

Cloud computing, open source software and tiered pay as you go subscription models have all helped lower the cost of creating apps. It's not surprising there seems to be an app for everything these days.

But how do you justify the cost of developing and maintaining an important, confidential and time critical app that universities only refer to a few times a year? This was the problem that the Universities & Colleges Employers Association (UCEA) faced with its pay benchmarking reports, which are produced annually to inform management and professorial pay decisions at over 140 universities. Enter the Rize package, the R package to dockerise R shiny apps.

In this talk I will take you through the options we considered before choosing Rize, and how it forms part of a GDPR-compliant infrastructure where UCEA member universities can access the benchmarking app through single sign-on. I will share with you our experiences of the new solution to date and the extent to which it is more user friendly and cost effective than other solutions in the long run.

Jack Pamely

BCA

Talk title:

Machine Learning and Dev Ops: API Deployment with and without training wheels

About:

I have been developing Machine Learning algorithms and data science services at BCA as a Data Scientist for the past 3 years. During my career with data I have developed a general passion regarding all things Machine Learning and become a daily user and advocate for the R language.

Abstract:

BCA have been on a journey from training and deploying machine learning models manually, using the infrastructure provided by Azure Machine Learning (ML) to a continuous integration approach, where models are automatically retraining and deployed to Azure's Kubernetes Service (AKS). Jack and Morgan will take you through this journey explaining the benefits/drawbacks of each approach; examining the Dev Ops tech stack used in each and reviewing how each approach is implemented. This will suit anyone thinking about model deployment, whether getting started with exposing models via an API or looking for an automated, continually integrated and robust deployment method.

James Laird-Smith

The Financial Times

Talk title:

Introducing scheduler: making recurring calendar events a little easier

About:

I'm a data scientist working at the Financial Times where I am part of the team responsible for customer analytics. My academic training is in empirical finance in which I hold undergraduate and postgraduate degrees from the University of Cape Town. I grew up in South Africa, but I now live and work in London. I do open source development in R during my spare time. I have a passion for information and making information understandable and useful to people. Outside of that I have interests in statistics, data visualisation and text mining. I also really like waffles.

Abstract:

Recurring calendar events occur frequently in everyday life. Some are simple, like birthdays, but others are complicated, like working out when the last Friday of the month occurs in every second quarter. This complexity makes them difficult both to conceptualise and to program with. This talk will introduce the newly created scheduler package, which aims to make dealing with calendar events easier by providing a simple grammar for expressing them using R code. scheduler has functions to represent all kinds of date elements found in the lubridate package: years, months, weeks, days, weekdays, quarters, semesters etc. It also has functions to represent dates in terms of their relations to one another, such as the "third Friday of July" or the "last Thursday in August". Finally, scheduler has functions to specify ranges of dates, such as "after the first Monday in July" but "before the last weekday of August". Finally scheduler has functions allowing users to compose these various elements together using set operations to form arbitrarily complex patterns. All of this means R users and businesses now a have a toolkit for easily dealing with these recurring calendar events in a way that integrates well with the tidyverse.

James Smythe

Culture of Insight

Talk title:

Powering Up Formula One's Market Research with R

About:

Founder of a great little company that automates work for market research teams, saves them time and builds apps that make their projects more valuable.

Also a dad who likes to chase around after his family, and race kayaks when he has a spare moment.

Previously worked in research roles at Carat, Global Radio and Kantar.

Abstract:

In 2016, F1's new owners introduced a new, consumer-centric strategy. A Research and Insights team was established to launch a comprehensive research programme, in six months, with a team of five servicing 100 data-hungry executives. New Research Director Matt Roberts knew that demand for insights could mean his team becoming overwhelmed.

Culture of Insight was tasked with automating the collation of data from F1's full breadth of sources on a race-by-race basis, and presenting them in a portal that allowed F1, its teams and partners to explore, fulfil demanding expectations, stimulate their curiosity, and keep them coming back.

The power of R, through from connectivity, and data wrangling to Shiny, created a whole process that delivers new data at the speed, and to the technical standards required by F1.

The Portal has been used over 6,800 times in its first year, by F1, partners and teams. The day after the 2018 Monaco GP, 64 users logged in for the latest data, the highest daily figure since the Portal's launch.

Jeremy Horne

Freelance

Talk title:

Building a successful data science team with R at the heaRt!

About:

I'm Jeremy, a specialist in customer analytics, reporting automation, machine learning and data visualisation, all underpinned through the R environment, which I began using in 2005, when I was a fresh-faced grad working in the city. Since 2014, I have been in senior media agency positions, deploying R-based solutions to marketing challenges and empowering my teams to be R advocates and specialists, initially as Evaluation Director at MEC and more recently as Head of Analytics at MC&C.

Abstract:

R has grown phenomenally since I wrote my first line of code in 2005. Back then, it was no more than a useful "tool" for your armoury and barely anyone had even heard of "data science".

For me, it was in 2010 that I started to adopt R regularly. As it slowly took over my life, I began to encourage team members (and clients) that it was central to building their teams of the future and followed my own advice in 2014, joining an organisation had never even installed R on a single machine, yet had a vision to spend less time reporting and more time analysing their clients' data. I adopted R to automate reports, analyse data, build models and create apps for clients, a theme which continued when I moved agencies in 2017, again embedding R into every project and transitioning processes from outside of R.

Team building and changes of this scale can be daunting, but are not hard if you follow a process and have a strategy. In this talk, I'll take you through my vision, how I've set up my teams and the R functionality that can give your data science offering the quickest wins.

Johannes Tang Kristensen

Arla Foods

Talk title:

How much milk do our cows produce? Lessons learned from putting our first R model into production

About:

I am currently working as a Senior Data Scientist at Arla Foods, one of the largest dairy companies in the world. This is my first industry job and since I started, 2.5 years ago, I have been on an exciting journey with many new challenges, insights and experiences, some of which I will share in my talk. Prior to my current job, I held various academic positions at different universities. R has been my main statistical language for the last 12 years.

Abstract:

Working in a dairy company, the single most important raw material we have is the milk our farmers produce. Therefore, knowing how much milk we can expect to receive weeks and months ahead is crucial for our entire supply chain. We were given the task of building a model that could produce such forecasts and replace the current manual forecasting process. Starting as a proof-of-concept, the value of replacing the current process with a model was quickly shown and the project was approved as an actual IT development project.

Today we have a forecasting solution that is used globally in our company and in this talk I will present our model and production setup and explain the reasoning behind our choices, including what we would do differently today. The main challenges we faced were: 1) Tracking down the right data sources and ensuring that we could get (correct) daily updates; 2) Expanding the model to meet new requirements; 3) Building a custom model manager for running the R code in order to retrain the model weekly and expose the forecasts through an API; 4) Deploying Shiny dashboards showing the forecasts as served by the API in our cloud setup.

Jonas Muench

Bayer Business Services GmbH

Talk title:

Leveraging the power of R in a regulated life science environment

About:

I am working as an IT consultant at Bayer Business Services, the global competence center for integrated business solutions of all Bayer Divisions.

My professional background is in the life sciences, where I finished my doctorate in the field of Neurobiology in 2017. Since then I am working at the interface between Bayer Pharma and IT as consultant and project manager. Recently I took over the role Business Analyst in a DevOps Team focussing on the needs of Clinical Pharmacometrics.

Abstract:

At Bayer we see an increasing interest in the application of R, mainly because of its easy adaptability and high versatility to support various business processes in Research and Development, Business Intelligence and Product Supply. This flexibility, however, comes at a price as fragmented R development environments with scattered data silos can become difficult to control in large organizations. To overcome this challenge, our Data Science team therefore implemented an innovative centralized self-service platform where users can build up their own data science environments in an instance by using extensive automation and cloud resources. However, in a highly regulated pharmaceutical R&D environment different concepts need to be applied. Besides the validation of infrastructure, IDE and R packages, a special challenge is to design a GxP-compliant data management. Additionally, we aim to realize this in an agile fashion to enable quick adaptions to the needs of a growing user base. In this talk, I would like to present the current status of our validated R environment for R&D at Bayer and an outlook to the future.

Julia Fumbarev

BWM Group

Talk title:

Process analysis and optimal allocation of parking space with R

About:

Julia Fumbarev has two Master's Degree, Electrical Engineering and Business Management, from the Technical University of Munich, Germany. She now works as a Data Scientist at BMW Group in Munich, where she focuses on machine learning, applied AI and big data analytics. Before joining BMW she worked for two years as a Data Scientist at Volkswagen Data:Lab.

Abstract:

Optimal allocation of parking space is a difficult problem in vehicle distribution. An unnecessary movement of a vehicle can cause additional cost in the range of several hundred Euros. When scaled to high volume production and delivery of cars, the importance of optimized processes becomes clear. At the BMW Group, we are faced with the delivery of millions of cars a year. In order to optimize our parking space allocation we employed process mining methods that evaluate the movements that took place. We present the analysis and how weak points in the process such as bottle necks or loops could be identified. In particular, we evaluated standby times, transport times and process times and secondly, the process flow and the associated resources. This enabled us to uncover inefficiencies and quantify the optimization potential. Following this thorough analysis, we implemented a recommendation algorithm that suggests an optimal allocation of the parking spaces. The metric optimized by our algorithm is the standing time in the storage locations. All analysis is based on the BMW Group's internal distribution. We show that the algorithm can reduce the number of transfers and improves the use of parking space, as measured in terms of standing time

Kasia Kulma

Mango Solutions

Talk title:

Integrating empathy in the Data Science process

About:

Kasia Kulma holds is a Data Scientist at Mango Solutions and holds a PhD in evolutionary biology from Uppsala University. She has experience in building recommender systems, customer segmentations, web applications and running NLP projects. She is the author of the blog R-tastic and is a mentor and organiser in R-Ladies London. She is an R-enthusiast interested in data (science) ethics and evidence-based medicine.

Abstract:

Despite the fast growth of analytics talent and more sophisticated technical skill-sets, the success rate of data science projects remains low. It may be because people-related factors are top challenges in such projects, e.g. lack of clear question, company politics or results not being used by decision makers. It's a reminder that the role of data science is to empower business to make better decisions, and to achieve it we need to ensure that we answer the right questions and have the right information. This is where empathy comes into play: enabling open and collaborative work even in complex business structures. In my talk, I demonstrate that empathy has a clearly defined role at every step of Data Science process: from pitching project ideas and gathering requirements to implementing solutions, informing & influencing the stakeholders and gauging the impact of the product. Next, I show that empathy is not a talent that we are born with, but a skill that everyone can strengthen and improve. Finally, I present the framework that integrates empathy at every stage of the data science process by aligning our problem understanding with stakeholders' agenda and ensures the right communication within and between the teams.

Kelly O'Briant

RStudio

Talk title:

The R in Production Handoff: Building bridges from data science to IT

About:

Kelly O'Briant is a solutions engineer at RStudio interested in configuration and workflow management with a passion for R administration. She has a background in data science and software design. Kelly loves several programming languages, but has found that R always sparks joy.

Abstract:

We know that adopting documentation, testing, and version control mechanisms are important for creating a culture of reproducibility in data science. But once you've embraced some basic development best practices, what comes next? What does it take to feel confident that our data products will make it to production? This talk will cover case studies in how I work with R users at various organizations to bridge the gaps that form between development and production. I'll cover reasons why CI/CD tools can enhance reproducibility for R and data science, showcase practical examples like automated testing and push-based application deployment, and point to simple resources for getting started with these tools in a number of different environments.

Kevin Kuo

RStudio

Talk title:

Towards open collaboration in insurance analytics

About:

Kevin is a software engineer at RStudio and is the founder of Kasa AI, a community organization for open research in insurance analytics. He develops open source packages for big data and machine learning, and is an author of the sparklyr and mlflow packages. As a former actuary, he is actively involved in the actuarial science research community.

Abstract:

Recent advances in AI technology have had, and will continue to have, profound impact on all industries, including insurance. However, insurance analytics professionals, and actuaries in particular, being some of the most quantitatively minded businesspeople, have a unique opportunity to embrace AI and open research and modernize the profession. We introduce Kasa AI, a not-for-profit community initiative for open research and software development for insurance analytics. Inspired by rOpenSci and Bioconductor, we hope to bring together the insurance community to solve the most impactful problems. We provide an overview of the current projects, including using neural networks for individual claims reserving and forecasting, and explaining machine learning models for pricing in a regulatory context.

Kieran Martin

Roche Products Limited

Talk title:

R in Pharma: A tailored approach to converting programmers to R in an industry resistant to change

About:

I have had a varied career switching from different industries, from the ONS to Insurance, and more recently Roche.

Through this time I have worked on variety of analytical projects, but one constant has been my usage of R. I have been a keen advocate of R, and have been part of efforts to promote it's use within Roche.

Abstract:

The pharmaceutical industry has historically, overwhelmingly made use of one software package, SAS. Almost every submission made to regulatory agencies such as the FDA and the EMA have been made using SAS code, and SAS data formats.

For those of us who are enthusiastic about promoting usage of R, this can lead to an uphill struggle, but things are changing. R has become more and more popular, and appetite for R usage is growing.

In this talk I will discuss some of the efforts to promote R within Roche, with a focus on two particular methods:

- Targeted training, focused on problems programmers have to solve every day
- Tidyverse training, and a focus on functional programming
- In house data and as close to real examples as possible, to show direct applications for work in R
- New R packages. Multiple different packages have been developed to support different activities. In this talk I will focus primarily on diffdf, a package that is available on CRAN, which used allows detailed dataset comparison, meeting an unmet need

Megan Stamper

Financial Times

Talk title:

Building a new data science pipeline for the FT with RStudio Connect

About:

As a Data Scientist at the Financial Times I have extensive experience working with subscription-based customer data. Recently the team and I have been working on a wide range of projects, including building lifetime value models, recommendation engines and attribution models.

I came to the field of data science after completing my PhD in Applied Mathematics and Oceanography. I'm passionate about the ethics of data science in business and in using data visualisation to tell compelling stories with data.

Abstract:

We have recently implemented a new Data Science workflow and pipeline, using RStudio Connect and Google Cloud Services. This has vastly decreased our pipeline complexity, allowing us to bring our models and products into scheduled production more quickly. In addition, our workflow, working closely together as a team on all projects on a regular two-week sprint cycle, has increased the range of projects we have been able to take on and complete.

To detail some of the key lessons we've learned (and some of the difficulties!), we'll walk you through one of our recent sprints, where we productionalised the generation of a suite of behavioural and demographic features so that they can be more easily plugged in to a range of models and used across the business by the FT's platform and product teams.

Mehrdad Mamaghani

Swedbank

Talk title:

Deployment of Deep Anomaly Detection in R

About:

Mehrdad Mamaghani holds a PhD in applied mathematical statistics from Stockholm University along with 10+ publications. Previously, Mehrdad has worked within the pharmaceutical and communication industries. At Swedbank, along with rest of the Analytics & AI group, Mehrdad and his colleagues conduct extensive work and research to better leverage the data within the bank as well as creating frameworks for more efficient and customer-oriented banking processes using deep learning techniques and advanced hardware platforms.

Abstract:

Anomaly detection has numerous applications in a wide variety of fields. In banking, with ever growing heterogeneity and complexity, it is difficult to discover deviating cases using traditional investigation techniques and pre-defined scenario searches. In this talk we'll have a walk-through on how Swedbank's deep learning models run on a state-of-the-art platform can help to detect unseen anomalies and deviations utilizing a large spectrum of features. We will specifically focus on those parts related to model building and engineering in R and how such pipelines in R can be taken to production.

Michael Hurst

HEOR Ltd

Talk title:

Machine learning in healthcare: beyond performance

About:

Michael Hurst joined HEOR Ltd in 2013 and now takes up the position of principal data scientist helping to form data strategy in the company as a whole.

With a background in computer science, Michael now heads up a small team within HEOR. This team is tasked with working with all different types of healthcare data (e.g. clinical trials, GP records) in the aim to create meaning from large vast data.

Michael's strenghts lie in data pre-processing (with prior experience in big datasets (upto ~1b records)), data visualisation and machine learning.

Abstract:

Has a GP ever told you to stop smoking or decrease your cholesterol? By making these changes, risk can be significantly reduced. But how can this risk be quantified and how does and how should machine learning fit into this process?

Historically, risk equations have been developed to quantify risk of an event based on patient characteristics. Based on raw patient data through linear regression models, risk is fluid and measured based on linear relationships between patient-level factors. Recently, research has been undertaken to understand how machine learning can be trained to replace traditional risk equations due to their ability to identify complex non-linear relationships in the data. Machine learning methods are often pitted against traditional methods using standard performance metrics and typically they are often only able to achieve marginal gains. Even when gain is achieved, clinicians must take the results on face-value and can't utilise known factors to be pro-active in-patient care.

Within this presentation, the difficulties of using machine learning in healthcare will be explored and why the focus should be on using R to break open the black box for relationship identification as opposed to relying on performance metrics.

Mitchell Stirling

Heathrow Airport Ltd

Talk title:

Understanding Airport Baggage Demand through R modelling

About:

Mitchell is a senior analyst at Heathrow Airport with seven years experience working in Operations, Commercial and Strategic positions. Previously he worked at London 2012 in the workforce planning department and at the Velodrome during the Games themselves.

Away from work he likes to use his statistical background to model the outcomes of sporting events as well as the less mathematical pursuits of music and cinema.

He lives in Reading with his wife Steph and two-year old daughter, Aurora.

Abstract:

Heathrow Airport is entering a new phase of growth. For the past 20 years growth at the airport has been constrained by it's existing runways and permitted air movements. Within 10 years a third runway is planned and the airport plans to move from 80 million passengers to 150 million passengers.

Along the way Heathrow's planners are looking at potential scenarios for occupancy and use of infrastructure to maximise existing assets and reduce the need for expensive capital works early in the programme.

To explore how these scenarios would impact the demand on baggage systems Heathrow have worked with Mango to convert a legacy PERL script in to an R package and make a number of improvements that cut down manual intervention, flag errors earlier, stabilising of the process and allow for greater variation in key inputs.

Mohammed Amin Mohammed

The NHS

Talk title:

Promoting the use of R in the NHS - progress and challanges

About:

I am the lead for the NHS-R Community project which is described here: https://nhsrcommunity.com/

I am an academic with an interest in applied health services research to improve the quality and efficiency of the NHS.

Abstract:

The English National Health Service (NHS) is one of the leading healthcare systems in the world . It was launched in 1948 with the guiding principle of being free at the point of delivery – a kind of crowd funded open-source freeware equivalent of healthcare. The NHS in England deals with about 1 million people every 36 hours and is continually generating vast amounts of data about the health and care of people . This data is one of the most precious, yet under tapped, resources in the NHS. But mining these mountains of data is a colossal task.

This is where R comes in. R was conceived in 1992 as a free open-source statistical programming environment, which is now widely used in industry and academia. But its use in the NHS is almost non-existent. Whilst there are several reasons for this, the absence of R at scale in the NHS, means that the NHS is unable to take advantage of the huge benefits of R. We have set up a NHS-R Community. Our aim is to promote the use of R in the NHS, and help to make the NHS better.

Nassos Stylianou

BBC News

Talk title:

How the BBC News data team uses R for graphics

About:

I tell stories using data for BBC News. I have worked on some of the most successful data and interactive stories at the BBC, including a calculator allowing people to find out how their food choices impact on the environment and a deep dive into the UK's housing market recovery, which received an award from the Royal Statistical Society.

Abstract:

Co-presentation by Nassos Stylianou and Clara Guibourg, Senior Data Journalists at BBC News

Over the past year, the BBC's data journalists have fundamentally changed how they produce graphics for publication on the BBC News website.

After having used R for complex data extraction, wrangling and analysis in major BBC stories for a number of years, as well as to build prototypes, we have recently shifted to using R and its ggplot2 package to create production-ready charts.

In this talk we'll explain why and how that change has come about, document our process and discuss what we learned along the way.

We will also touch upon how the initial success of this transition to R for graphics led us to develop an internal course to spread the use of R and ggplot2 to other members of the Data and Visual Journalism team who had no prior knowledge of R or previous programming experience.

Robert Duff

Transport for London

Talk title:

Let me in! Let me on! Quantifying highly frustrating events on the Underground

About:

Wrangling/visualising my way through life!

I've had the pleasure of working at Transport for London (TfL) for over a decade and in recent years have been an active and leading member of the Data Science community that we've built up from scratch. R has played a huge role in this and I love its quirks and how it can help unlock analytical potential.

Outside of work,I'm a proud Welshman & father of two. I look forward to the weekends immensely to enjoy family time & walks in the countryside.Avid fan of North American Sports! The stats on offer are incredible!

Abstract:

Crossrail 2 will provide additional rail capacity in a south west and north east corridor through London. To help get the project closer to reality and over the line we have turned to R help inform the business case submission.

At TfL we are in a great position where we have some very enviable datasets to play with!

Using two of these: our ticketing transactional data and train movement data plus a whole heap of expert knowledge from our underground stations we aim to capture information regarding two key elements of the passenger journey currently not quantified:

- Being held outside a station

- Left behind on the platform (inability to board)

Our aim is to paint a picture of how the customer experience looks like now and demonstrate that the new Crossrail 2 services will change things for the better!

Richard Marshall

Hiscox

Talk title:

Using data to drive better decisions

About:

Richard works as a data strategy consultant at specialist insurer Hiscox. His work focuses on realising tangible value through the use of data and analytics and helping shape Hiscox's journey on becoming a more data driven company. Prior to this role, Richard worked as a Fine Art underwriter with Hiscox in Lloyd's of London after completing a PhD in Biomedical Imaging from the University of Birmingham.

Sam Hall & Sam Collins

TravelSupermarket

Talk title:

R For A Data Driven SEO Workflow

About:

Sam Hall - Data Scientist specialising in digital marketing optimisation.

Sam Collins - SEO Manager, responsible for all organic traffic to Travelsupermarket.

Abstract:

Organic traffic from Google and other search engines is a crucial revenue driver for any online business, but it also poses some difficult data problems that traditional SEO tools struggle to handle. At TravelSupermarket, we've turned to R to help us deal with the volume of available data and turn that into actionable insight. We want to share how we use R for ingesting SEO data, exploring trends and why R is useful when Google stops you performing traditional randomised experiments.

Sandro Matos

Merkle Aquila

Talk title:

Preventing human trafficking through the power of advanced analytics

About:

I'm a senior data scientist from Merkle Aquila. I've worked across several analytical projects from predictive models to dashboards in many industries from telecommunications to insurance and retail. I'm originally from Portugal and decided to move to the beautiful city of Edinburgh in 2015. I love travelling, trying new food and receiving postcards from around the world.

Abstract:

Since July 2018 Merkle Aquila has been running a pro-bono partnership with Stop the Traffik (STT), a global charity pioneering the cause of intelligence led prevention of human trafficking. Through this pro-bono initiative, a group of us have been donating our expertise in analytics and strategic thinking to STT.

In 2019 our focus is on using visualization platforms and open source tools such as R to help STT achieve their vision. Specifically we hope to leverage:

1. STT's existing HT incident data as well as data from a newly created Trafficking Analysis Hub to identify key patterns and trends across different geographies;

2. Natural Language processing to identify general public sentiment and levels of awareness in different geographies to drive STT's program strategy.

Sebastian Wolf

Freelancer i.a. for Roche

Talk title:

RSelenium or shinytest: How to make shiny apps ready for use in a regulated environment

About:

Sebastian Wolf is a scientific software developer and shiny enthusiast. He his currently contracting on a project for Roche. The project includes a framework to analyze clinical trials with R and shiny. His most recent shiny projects were: bioWARP, "the largest shiny app in the world" and stravachaser "A virtual city cycle race". He runs a blog on shiny and "How to test R" on Medium.

Abstract:

Quality assurance (QA) inside Pharma is a highly regulated environment. There are two kinds of software currently used in Pharma QA. "Validated" Excel sheets and special purpose applications. Validating Excel is hard. Coding special purpose apps is hard, too. R shiny is a lightweight alternative. It includes a great user interface and is easy to set up. In Pharma QA using shiny is challenging though. The QA decides whether a drug or reagent is market ready and harm-free. Returning a correct, valid output is the challenge for each shiny app used in such an environment. Testing shiny is mandatory to make apps ready for productive use. With RSelenium and shinytest there are two ways to do so. Each of them has certain drawbacks, and of course benefits. You will learn and understand when to choose which of the two approaches. The talk will present two shiny apps. The apps each got tested with a different approach. Thanks to solid tests they are now in production in Pharma. You will see how far you can go with shiny in a regulated field. All challenges using shiny in any regulated field can be overcome.

Stephen Gormley

Amgen Ltd.

Talk title:

An Enrolment Modelling R Package

About:

I have worked in the pharmaceutical/biotechnology industry for ~20 years and have an undergraduate degree in Statistics and post-graduate degree in Software Engineering. During this time, I have been a SAS macro developer, a statistical SAS programmer and a clinical trial product lead programmer (primarily in the Oncology therapeutic area). For the past two years I have been working in the Data Science team within the Centre for Design and Analysis at Amgen Ltd. My primary role within this group is to re-design raw complicated mathematical/statistical code into more robust, repeatable, testable, maintainable and well documented R packages (and visualisations).

Abstract:

The Data Science team within the Centre for Design and Analysis at Amgen Ltd. have developed an enrolment modelling R package which, for a multicentre clinical trial, predicts the probability of achieving enrolment target times and calculates an optimal site allocation given a variety of country specific constraints.

This presentation shall set out the business use case for the underlying mathematical models and provide details of the design choices encountered during R package development, testing, deployment and documentation, which shall include:

- S3 as the primary design choice (and why not R6 or S4?).
- A few of the key R packages used, including:
  - Improved efficiency with Rcpp in the optimisation routines.
  - Further enhancements to the optimisation routines using genetic algorithms DEoptim::DEoptim and GA::gaisl.
  - Documentation using roxygen2.
  - Extensive testing using testthat.
  - Appropriate code coverage using covr.
- The business appropriate SDLC followed, including the installation and operational qualification for deploying to a production R server.

Note: The package was built from complex scripts which were based on the enrolment process models developed in Anisimov & Fedorov (2007), Anisimov (2011). This presentation will not detail the underlying mathematical methodology that forms the basis of the package.

Susanna Liberti

eBay

Talk title:

Using R to ”glue” together data from different sources

About:

I am the Head of Business Insights for the Classifieds business of eBay Scandinavia, I have a strong background as Technical leader that I developed in 20 years of experience. I have been working with different kind of data, in different industries: e-commerce, green energy, forensic, Telecommunications and others. In my long experience, I always found fascinating unveiling the stories embedded in data and supporting data driven decisions.

Abstract:

In eBay Classifieds data are the foundation of our strategy and what we use to measure success. Being effective at merging data from different sources is more and more important in our business: most of the time we need to access Finance data from our Finance systems, tracking data (frontend and backend) coming from our Platforms and data coming from external providers.

In this presentation, I will show how we used R to access, merge, clean and visualize data from our multiple data sources; I will talk you through our challenges and how we approached them and explain how this approach helped us building better visualizations in a more flexible way.

Premal Desai

The Gym Group

Talk title:

Flexing analytical muscles - how data science, culture and commercial rigour comes together to drive better return on investment

About:

Premal is currently Head of Data & AI projects at The Gym Group based in London. Prior to this, Premal was Global Head of Strategic Marketing & Analytics at Thomson Reuters where he setup and established a data-centric culture and organisation. He previously worked as a Strategy & Marketing consultant at PA Consulting and as a Business Alchemist for Orange. He is an Alumni of Harvard Business School, London Business School and graduated from the London School of Economics. Most importantly, he is a lifelong and passionate Liverpool FC fan!

Theo Boutaris

Weber Shandwick

Talk title:

Deep Milk: The Quest of identifying Milk-Related Instagram Posts using Keras

About:

Theo Boutaris is the Data Science Associate Director at Weber Shandwick. He is one of the first 100 people globally with a Golden Badge on the R Programming Language on stackoverflow.com and is the author of tableHTML, a popular R package hosted on CRAN. Theo is applying his knowledge on Weber Shandwick's analytical work across all areas of audience intelligence and commercial performance analysis. In his free time he contributes to the Open Source community and plays sports, particularly tennis.

Abstract:

Undoubtedly, Instagram contains a vast amount of information (text, images, videos) that companies harvest in order to understand their customers.

Traditionally, companies would firstly attempt to analyse posts' text to understand people's behaviours. However, text does not always agree with what people show in an image. Authors on social media use sarcasm, irony, metaphors and slang language when publishing a post and on many occasions analysing just the text would extract the wrong insights.

The goal of our analysis focuses on whether people who post about milk-containing food on Instagram are happier than the people who post about other foods. Our initial findings showed that a significant proportion of people who mentioned the word milk in the text, were actually talking about topics completely unrelated to food. Therefore, in order to capture the sentiment of just those authors talking about milk foods, we applied Deep Learning (with CNNs) on Images to make sure they showed milk related food.

This talk will take you through our journey of using Keras to identify milk-containing foods, reaching to the conclusion that people posting about milk foods are actually happier than those posting about food in general.

Thomas Laber

Austrian Post

Talk title:

serveRless - how to deploy R code in a modern cloud infrastructure

About:

I am the Lead Data Scientist at the Austrian Post where we build the data science infrastructure and the data science team. After studying Business Informatics at the Vienna University of Economics and Business and TU Wien, I worked as a consultant for Accenture in Vienna and Detecon in San Francisco. My goal is to nudge this lovely formerly state-owned monopoly with its great number of tenured employees into the direction of data-driven decision making.

Abstract:

R is a great language for rapid prototyping and experimentation, but putting an R model in production is still more complex and time-consuming than it needs to be. With the growing popularity of serverless computing frameworks that offer Functions-as-a-Service (like AWS Lambda, Azure Functions) or Container-as-a-Service (ECS and ACI) we see a a huge chance to allow R developers to more easily deploy their code into production. We want to show you how you can use serverless computing to easily put models in production. We will discuss the pros and cons of various approaches and how we implemented a completely serverless data science platform for R in Microsoft Azure that you can arbitrarily scale up and down. While we come from an Azure background, porting the ideas over to AWS or Google Cloud should be straight forward.

Timothy Wong

Centrica plc

Talk title:

Large-Scale Time Series Forecasting in Apache Spark

About:

Timothy is Senior Data Scientist at Centrica plc, an energy service company based in the United Kingdom. His areas of interest include statistic, machine learning and big data. He was trained as a policy researcher and served at the United Nations prior to joining Centrica. He holds two Master's degrees.

Abstract:

Accurately forecasting power demand is important for securing energy supply. Time series forecasting methods and other machine learning algorithms can be used to create energy forecasts. We have developed a forecasting framework based on multi-model approach at customer account level. The framework uses a wide range of algorithms (e.g. GLM, ElasticNet, Seasonal ARIMA-X, Decision Tree, Random Forest and Gradient Boosting Machine). Models are pre-trained on AWS EMR cluster using Spark/SparklyR. The process is run at massively parallel scale (> 3000 vCores). Once the model training algorithm has completed, the model objects are persisted on AWS S3 so that they can be reused at a later date. To trigger a forecast, the deploy pipeline will load the pre-trained model object from S3 and create a forecast based on the prevailing inputs. The output is stored as partitioned parquet files on S3, which can be converted into table view through AWS Athena.

Yizhar Toren

Shopify

Talk title:

Hard talks: Explaining Bayes to Business (marketing spend use-case)

About:

I've been evangelising R since I first started using it at university: As a TA, as a consultant, as a statistician and and a data scientist working on prototypes and production systems. Passionate about Bayesian statistics, functional programming and and all around math geek.

Abstract:

Marketing spend attribution is a tough question. In many (and indeed in most) cases, linking marketing spend to conversion rates is harder than you think: from partial / biased / inconsistent tracking of exposures and leads, through issues with A/B testing and even the interpretation of the results.

Fortunately, in the past couple of years a new set of powerful tools have been made available in R (packages like `prophet`, `causalimpact`, etc.). These packages offer "out-of-the-box" access to powerful forecasting & inference methods, but they come with a price: the Bayesian frame of mind with the different trade-offs it presents are particularly hard to explain to business users (model structures, choice of priors, uncertainty bounds vs. point estimates,...)

In this talk I review the ongoing conversation we have with our marketing department around the data, models and interpretation of marketing spend KPIs: Where we started, where we are today and what were the challenges were met along the way.

Zhanna Mileeva

N Brown Group

Talk title:

Model Performance Assessment

About:

Zhanna received a PhD in Physics and has over 10 years' experience in Material Science, Research and Development, bringing innovative solutions to multi-disciplinary academic and industrial projects. In pursuit of her passion for Advanced Analytics and Visualization, Zhanna dived into the world of Data Science. Being a part of N Brown Group's Data Division, she is currently focused on increasing business profitability.

Abstract:

Data Science and Machine Learning are rapidly becoming the driving force in making informed decisions and bringing a positive impact to business. Together they can provide better understanding, promote new thinking, disrupt the business, improve customer experience and offer better solutions.

The capabilities are enormous; however, the power of predictive technology should be carefully applied so users are not misled by its outcomes.

The application of appropriate metrics to model accuracy assessment is vital for comparable, measurable and reproducible results. It can also form a basis for further refinement and continuous improvement.

The talk focuses on several model performance metrics comparing them and reviewing their applicability. It also introduces a coefficient of model efficiency and a dynamic approach to model accuracy assessment.