EARL Online 2020, Friday 16th October agenda

Time	Name	Talk title	Company
9am-9.10am	Rich Pugh	Welcome to EARL	Mango Solutions
9.10am-9.55am	Annarita Roscino	Keynote: My journey from data practitioner to data & analytics leader	Zurich
9.55am-10am	Break		
10am-10.30am	Dr Lisa Clark	For love or money? How to generate high revenue from a team of R developers	Virgin Media
10.30am-11am	Owen Garrity	Mind the gap – how forecasting can help to see gaps between financial targets and demand in advance	The Very Group
11am-11.30am	Dr Adriana De Palma	Using R to predict global biodiversity loss	Natural History Museum
11.30am-12pm	Jeremy Horne	Building an engaging email marketing campaign using R	RocketMill
12pm-12.30pm	Eryk Walczak	Using R to study financial regulations	Bank of England
12.30pm-1pm	Break		
1pm-1.30pm	Sergey Mastitsky	Still parsing User Agent strings for your models? Use this instead!	Aviva
1.30pm-2pm	Harshitha Ravindra	R-Powered dashboards using shiny for digital health RWE studies	Vitaccess
2pm-2.30pm	Emma Vestesson	Using R for personalised medicine	UCL and the Health Foundation
2.30pm-3pm	Dave Goody	How R shiny apps drive operational decision making	Department for Education
3pm-3.30pm	Steve Marriot and Samir Khimani	Practical experience with insurance pricing in R	NHBC
3.30pm-4pm	Joe Fallon & Gavin Thompson	Supporting the self-employed during COVID-19	HMRC
4pm-4.45pm	tbc	Keynote	
4.45pm	Mango Solutions	Closing remarks	

Detailed agenda

Name, job title, company and talk title	Synopsis	About	Image
Dr Lisa Clark, Data Science Manager, Virgin Media *For love or money? How to generate high revenue from a team of R developers*	With a huge growth in easy-to-use (yet highly sophisticated) algorithms and related technical talent, business leaders are coming to expect high-revenue returns from analytical projects. Equally, as AI and Data Science become household concepts, businesses are adopting smart technologies, insisting on the requirement and use of machine learning models in every industry. For businesses, however, complex models can be lengthy to develop and top skills are an expensive resource. For keen programmers in this field, the result is often a strain between quick return on investment and exciting code. Drawing on direct examples from industrial projects using R, I will discuss the encountered trade off between elegant modelling, which often resulted in higher efficacy and also higher developer morale, and "dirty" coding, which got the job done quickly and still generated high revenue. I will detail lessons learned from marketing models (e.g churn), through network algorithms, to fun with the Knapsack model. The aim is to show how we use R in a working environment to make both high revenue gains and cost saving measures, making the tough decision as to when a model is "good enough".	Lisa Clark leads the high-performing data science team in Virgin Media Business, which delivers data science solutions across the whole company. Lisa is an experienced data scientist, having more than 20 years' experience within private business, public admin and academic settings. She originally trained as a theoretical physicist, specialising in cosmology, later becoming an expert in energy systems and then econometrics. Her real expertise lies in mathematical and computational analysis of large and complex data sets, but also thrives in the end-to-end flow of data, from collection through manipulation to revenue generation.	

Owen Garrity, Data Scientist, The Very Group, *Mind the gap – how forecasting can help to see gaps between financial targets and demand in advance*	The Very Group is the UK's largest integrated retail and financial service provider, and operator of Very.co.uk and Littlewoods.com. With over 100 years of history the company has a passion for change and improvement. In the past we have relied on SAS for many different modelling approaches. However, over the past few years there has been an initiative to move some of these pieces of work to open source, alongside any new work to be conducted in open source. One such case was identifying gaps between financial targets and most likely scenario given historical data. Once knowing the gap and from where it's coming better decision can be made to prevent it. This talk will convey how we approached this business use case by utilising R.	I am Data Scientist at The Very Group, currently working within the Retail Data Science team. Before this I was a consultant aiding charities, B2B and B2C with their data driven needs. My experience with R stems back to my Mathematics degree alongside the use of Matlab and derive. Since then I have been utilising R for EDA, Modelling, and reporting back to stakeholders in the form of Shiny dashboards and Rmarkdown reports.	
Dr Adriana De Palma, Co-Investigator Researcher, Natural History Museum, *Using R to predict global biodiversity loss*	The PREDICTS Project aims to better understand how Earth's biodiversity is impacted by humans. This collaborative effort led by the Natural History Museum has resulted in a database of 3.25 million species measurements from nearly 50,000 species in about 100 countries. R has been an indispensable tool for our data processing, modelling and communication of results. To handle our large data sets efficiently, we use packages like {foreach} and {doParallel} for parallelisation and {data.table} for fast wrangling. We also make use of R's statistical prowess through tools like {lme4} and {gamm4} for mixed-effects modelling, which helps us build robust models for predicting biodiversity change now and into the future. Additionally, we've found that {shiny} and {plotly} have been excellent tools for communicating our results to policy makers and the public, who can explore projections in both space and time. This talk will describe how R has	I am a computational ecologist with a focus on understanding biodiversity responses to human drivers of change, including both land-use change and extreme weather events. I am interested in synthesising extensive carefully-collated databases to identify, and test the efficacy of, potential management practices and policy decisions at different scales to mitigate the impacts of global change on biodiversity and ecosystem services. I have an interest in policy engagement and have developed online tools to better communicate our peer-reviewed scientific results to policy makers and the public.	

	been our major tool of choice for understanding how humans impact the planet and for getting the message into the hands of people who have the power to change things.		
Jeremy Horne, Head of Data Science, RocketMill *Building an engaging email marketing campaign using R*	Digital media is more than just "doing some Google" or "making a Facebook ad". Email marketing is a simple and cost-effective way of reaching your previously engaged (and of course GDPR compliant) audience. It is also arguably easier to ignore – people may be less "in market" or the information may not be relevant at that time. So how do you construct an engaging email campaign? Your historical email data contains a wealth of information that you can use for optimisation. What time of day are people most engaged? What day of week? Does this differ by gender or demographic? Which "click-through" links have received the highest-level engagement? Which links are most likely to lead to a purchase? This talk takes you through an example of how we worked with one of our clients to answer these questions in R, alongside a collection of text analytics techniques to determine the optimal word count, which words and phrases lead to the highest click rate and how important it is to mention your brand in the message. I'll also talk about how we used a machine learning model to determine the type of content each person is most likely to engage with.	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. I moved into media in 2014, deploying R-based solutions to marketing challenges and empowering my teams (and three year old son) to be R advocates and specialists. I joined RocketMill in 2019 to strengthen the agency's data offering by setting up a new data science team and later that year also set up BrightonR – Brighton's first dedicated R user group.	

Eryk Walczak, Senior Research Data Scientist, Bank of England Using R to study banking regulation	The banking reforms that followed the financial crisis of 2007–08 led to an increase in UK banking regulation from almost 400,000 to over 720,000 words, and to concerns about their complexity. We define complexity in terms of the difficulty of processing linguistic units, both in isolation and within a broader context, and use natural language processing and network analysis to calculate complexity measures on a novel dataset that covers the near universe of prudential regulation for banks in the United Kingdom before (2007) and after (2017) the reforms. Linguistic, i.e. textual and network, complexity in banking regulation is concentrated in a relatively small number of provisions, and the post-crisis reforms have accentuated this feature. In particular, the comprehension of provisions within a tightly connected 'core' requires following long chains of cross-references. This talk will describe the process of acquiring and analysing unstructured data coming from various sources (e.g. web scraping and pdf documents).	Eryk Walczak is a senior research data scientist in the Advanced analytics division. He is also a PhD candidate in the Department of Experimental Psychology at the University College London. Prior to joining the Bank, Eryk worked in analytic roles for a fintech and social media company. His research interests involve applying data science and experimental methods to study macroeconomics. Eryk has been using R since 2012. He developed several R packages and contributes to R-bloggers.	

| Dr Sergey Mastitsky, Data Science Lead, Aviva *Still parsing User Agent strings for your models? Use this instead!* | User-Agent strings (UAS) are header fields in HTTP requests used to identify the device and browser making the request. By parsing a UAS, one can extract many data points (type and make of the visitor's device, operating system and its version, etc.) that can be used as valuable inputs for Machine Learning models (e.g., encoding customer affluence and tech savviness). However, UAS are lacking standardised formatting, which makes their parsing a formidable task, requiring high-quality regular expressions. In addition, the variety of values one can encounter in UAS is astronomically large and constantly growing, making one-hot encoding of the respective features impractical. In this talk, I will demonstrate how these problems can be overcome by embedding UAS into a low-dimensional space using Natural Language Processing techniques. I will also provide an overview of the respective R packages and illustrate the use of UAS embeddings in unsupervised and supervised learning applications. | Sergey is a Data Scientist with multiple years of experience in academic and industrial sectors. He currently manages a team of Data Scientists and engineers, who are helping to deliver personalised product offerings and experience to customers of the insurance company Aviva (UK). Sergey has published 4 books on data analysis and visualisation using R, and has been authoring a popular blog about this language since 2011. | |

| **Harshitha Ravindra**, Data Visualisation Lead, Vitaccess Limited R-Powered Dashboards using Shiny for Digital Health RWE Studies | Vitaccess is a multi-award-winning digital healthcare research and strategic consultancy. We offer a world-class real-world evidence (RWE) data portfolio and data science research via our state-of-the-art digital platform, experience-based Market Access, and Health Economics and Outcomes Research (HEOR) solutions plus localization and linguistic validation in life science. The data we collect from patients through apps creates disease-specific patient registries containing multiple variables on disease characteristics, treatments, Patient-Reported Outcomes, and general health-related quality of life. These data captured from patients enables us to visualise the data for longitudinal and cross-sectional studies allowing to facilitate different stakeholders with different analysis to make important decisions. To demonstrate these different analyses, and their importance, few examples of data visualisation using diverse packages of R (like shiny/ shiny dashboard/ htmlwidgets/ ggplot2/ plotly/ leaflet) will be presented (a) Patient demographics data (b) patients reported outcomes (c) performance/ engagements of the study. We will further present how these different visualisations add value to the analysis performed and insight inferred. | I am working at Vitaccess as Data Visualisation lead. I have over eight years of diverse experience as a programmer and researcher in data visualisation, data modelling, python, R and statistics. In most recent years, I have been part of building several data visualisation tools and performed data analysis on public health data. I am pursuing my Ph.D. at the International Institute of Information Technology - Bangalore, India.
Currently, I am working on building a business use case dashboard using real-time patient data based on R shiny. | |

Emma Vestesson, Senior Data Analyst, UCL and the Health Foundation Using R for personalised medicine	Accurate dosing of chemotherapy is vital to treat cancer without severe side-effects. In addition to this, knowing when the drug has left the body allows clinicians to know when it's safe to discharge a patient. What happens to the drug in the body depends on eg age, sex and kidney function and can be estimated using non-linear mixed effects model. These are complex models that can't be calculated by hand by a clinician so in comes Shiny; a way to give clinicians the power of models but without writing code. I've built a Bayesian estimation shiny app that predicts when a drug has left the body to be used in clinical practice. I'll discuss some technical challenges encountered eg clearing old inputs when new patient data is entered or making a dynamic input widget based on user input. I'll cover design choices when building an app that is useful and safe in a hospital setting. It also includes working with pharmacists when designing the user interface and what outputs where the most useful to them. I'll discuss the challenge of deploying an app in a hospital and working with patient data.	Emma is a Senior Data Analyst at the Health Foundation and a part-time PhD student at the UCL Great Ormond Street Institute for Child Health. Her research is focussed on using electronic health records and machine learning to understand and improve prescribing. She is also an R-ladies London organiser and a proud member of #rdogladies.	
Joe Fallon & Gavin Thompson, HMRC**,** Supporting the self-employed during COVID-19	The Self-Employment Income Support Scheme (SEISS) is one element of the UK Government's targeted support to individuals and businesses as a result of the Covid-19 situation. Our team used R to bring together disparate tax return data, creating the definitive database of who was eligible and what their grant should be, in order to support 3.4 million people. Additionally, our insight into the data helped to reshape the policy as it was being developed.		

	We adopted the principles behind reproducible analytical pipelines, enabling us to build a robust pipeline to process over 100 different deliveries of data, that could be rapidly changed to keep up-to-date with policy changes and to take account of data quality issues. It also enabled us to move processing swiftly to the cloud when computational power became an issue. In this talk we will share our experiences and explore which approaches helped us to manage this complex project, delivering to tight deadlines with evolving requirements.		