UCL R Project Showcase 2021

Book of Abstracts

The UCL R Project Showcase 2021, hosted virtually on 17 June 2021, will bring R users from across UCL together for a free online event. It will showcase the work of undergraduate, Masters and PhD students and early-career researchers through short talks on projects that use R.

The event is organised by the UCL R User Group and as part of the UCL Festival of Code coordinated and enabled by funding provided by the UCL eResearch Domain.

Keynote: *From frogs to dogs: using R to tell a story*	
Speaker	Kirsten McMillan
	Dogs Trust
Title	From frogs to dogs: using R to tell a story
Abstract	The rapidly increasing volume and diversity of ecological data has been driven by many factors, such as automated collection, technological advances, and data repositories. Facing these big data, many researchers have developed and/or refined R programming skills for quality assurance and reproducibility, along with the ability to access open-source tools and advanced techniques regarding data manipulating, modelling and visualisation. In this talk, I will focus on two main topics: amphibian disease dynamics and canine ecology. More specifically, I will demonstrate how R has allowed me to: generate quantitative estimates of how infection parameters of an amphibian infectious disease covary with environmental heterogeneity and identify host-specific determinants of resistance; establish a population baseline for the UK pet dog population, including spatial density and distribution; and provide evidence to the charity sector, of the advantages of applying computational tools, such as text mining and sentiment analysis, to scraped online data.

Session 1: Code Development	
Speaker	David Simons
Department - Faculty	Genetics, Evolution and Environment - Life Sciences
Title	Using R to investigate, report and automate a living review on the association of smoking with SARS-CoV-2 infection, disease severity and mortality
Abstract	A living review was implemented to synthesise the rapidly emerging evidence on the association of smoking and SARS-CoV-2 infection and disease severity. This review is currently on version 11 with an earlier version (v7) published in the journal Addiction. This review has been used to inform national and global public health organisations. The code to reproduce this analysis was regularly released on GitHub and archived on Zenodo to support the principles of Open Science. As the project continued and the analysis was converted from classical meta-analysis to Bayesian meta-analysis an R markdown document was used to combine different scripts, generate outputs and to aid interpretation of the included studies. The most recent version (v11) was written entirely within R, automating the management of references and new data. This project is an example of iterative improvement of an analysis and report writing pipeline conducted within R.

Speaker	Alice Burleigh
Department - Faculty	GOS Institute of Child Health – Population Health Sciences
Title	Automating genetic variant prioritisation in rare disease gene hunting
Abstract	Human genome sequencing is essential for researching, diagnosing, and discovering rare genetic diseases. Once sequencing has been performed on a patient, tens of thousands of genetic variants must be narrowed down in silico to hunt for the disease-causing culprit. This is the data analysis equivalent to looking for a needle in a haystack. I am developing 'FiltEX'- a workflow in R which automates this laborious process. FiltEX filters, sorts, and annotates genetic variants according to various criteria to generate a shortlist of potentially disease-causing variants. FiltEX also combines existing genetic analysis tools Exomiser and OptiType into a novel workflow, enabling the researcher to make more informed decisions about which variants to investigate. The output of FiltEX is an organised spreadsheet which can be easily interpreted and analysed. FiltEX aims to simplify, streamline, and enhance the process of 'needle-finding' in genetic sequencing analysis, consequently accelerating genetic diagnosis times for the patient.
Speaker	Jonathan Bourne
Department - Faculty	UCL Energy Institute – Bartlett (Built Environment)
Title	Everything you ever wanted to know about SETSe but were too afraid to ask
Abstract	The Strain Elevation Tension Spring embedding algorithm (SETSe) is a deterministic method for embedding feature-rich networks. The algorithm uses simple Newtonian equations of motion to embed the network onto a locally euclidean manifold. To create the embedding, SETSe converts node attributes into forces and the edge attributes into springs. SETSe finds an equilibrium position when the forces on the springs balance the forces of the nodes. Some applications of SETSe are; analysing social networks; understanding the robustness of power grids; geographical analysis; predicting node features; understanding power dynamic between individuals and organisations; analysis of molecular structures. This presentation will provide both a brief technical discussion of the algorithm and its implementation, as well as several use cases. There are very few options for graph embeddings using R, and this is something that rsetse seeks to address; the algorithm has been implemented in the package `rsetse` and is available on CRAN.
Speaker	Dean Markwick
Department - Faculty	Statistical Science – Mathematical & Physical Sciences
Title	dirichletprocess: Non parametric Bayesian statistics
Abstract	The dirichletprocess package provides software for creating flexible Dirichlet processes objects in R. Users can perform nonparametric Bayesian analysis using Dirichlet processes without the need to program their own inference algorithms. Instead, the user can utilise our pre-built models or specify their own models whilst allowing the dirichletprocess package to handle the Markov chain Monte Carlo sampling. Our Dirichlet process objects can act as building blocks for a variety of statistical models including and not limited to: density estimation, clustering and prior distributions in hierarchical models.

Session 2: Visualisations	
Speaker	Nicole Watson
Department - Faculty	UCL Energy Institute – Bartlett (Built Environment)
Title	Visualising characteristics of future energy models
Abstract	In the energy field, emerging models known as peer-to-peer energy trading, transactive energy, and collective self-consumption have the

	capacity to revolutionise our future markets. There is general consensus that these models all involve individuals trading or sharing energy directly with one another, however, this fast-moving and cross-disciplinary field lacks a universally accepted definition of each of these models. This talk will present visualisations from a qualitative project aiming to identify the defining characteristics of each of these three models through a systematic literature review. Throughout the project, R was used to create visualisations, including “model blueprints” representing characteristics associated with each model. This involved manipulating qualitative data into a form that could be meaningfully visualised, without losing the richness of that data. The talk will focus on using ggplot to create highly customised visualisations, as well as the challenges and value of using data visualisation to complement qualitative work.
Speaker	Yaning Wu
Department - Faculty	Institute of Epidemiology and Health Care – Population Health Sciences
Title	Can you hear the sirens? Presenting ambulance response patterns during a coronavirus surge using multimedia
Abstract	During your country's lockdown, did you hear more ambulance sirens than usual? Or were they merely more obvious without the sounds of bustling pedestrians? I sought to answer this question by visualising ambulance incident dispatch data from New York City with R, using animated line plots to show how incident numbers and response times changed when comparing February - April of 2020 and 2019. Then, because ambulances are primarily known by their sounds, I sonified this data to bring more accessibility and humanity to the information I presented. If I've done it right, you should hear sirens rising and falling in pitch.
Speaker	Indigo Brownhall
Department - Faculty	CEGE – Engineering Sciences
Title	Creating and Visualising a Sanitation Risk Index in Antanarivo, Madagascar
Abstract	Madagascar ranks 172nd for sanitation provision out of 180 countries. Together with a high population density and frequent flooding, Antananarivo suffers heavily, effecting individual's health, happiness, and dignity. It has been estimated that Madagascar's economy loses nearly $600 million each year from lost productivity and increased health costs due to sanitation related diseases. In this project, I propose three different geospatial risk indexes, which use open-source data to calculate and visualise the inequalities and spatial variance of sanitation related health risks. Furthermore, using R-Shiny. creating a platform where local decision makers can understand the problem, and identify areas where investment is needed. The work is on behalf of Gatherhub, a geospatial charity, and the project was funded by the World Bank and the Sir Halley Stewart Trust.
Speaker	Thomas Savy
Department - Faculty	UCL Cancer Institute – Life Sciences
Title	Using R to elucidate signalling pathways that define glioblastoma cell fate
Abstract	Glioblastoma (GBM) is a lethal cancer, with a median survival of 14-16 months, and currently incurable – partly due to a subpopulation of glioma stem cells (GSCs). GSCs exhibit resistance to mainstream cancer therapies. Recent evidence has shown that these cells can be driven towards an oligodendrocyte-like fate when in a niche environment – the damaged white matter. Oligodendrocyte-like differentiation is associated with a decrease in tumour malignancy, thus providing a window of therapeutic exploitation. To identify pathways involved in differentiation, I took RNA-seq analysis results from a patient-derived xenograft model and used R to visualise data and perform enrichment analysis. In addition, I have utilised R to

	cross-reference receptors present in the RNA-seq analysis and specific receptors registered in online databases. I hope to elucidate pathways that could be targeted through therapy in a niche-independent setting.
Speaker	Max Beesley
Department - Faculty	GOS Institute of Child Health – Population Health Sciences
Title	Using R to investigate single-cell transcriptomic data
Abstract	I use computationally-intensive transcriptomic techniques to characterise the autologous multipotent stem cells (AFSCs) present in the human amniotic fluid. AFSCs can be expanded and differentiated during gestation, making them an ideal candidate for fetal and neonatal autologous regenerative medicine. I use single-cell RNA-sequencing to; determine the AFSCs origin, identify cell markers for accurate isolation and to establish the extent of their multi-potency. This relies heavily on the use of R for graphical representation, statistics and for manipulating huge datasets in the search for candidate genes. One cumulation of this work is in the overlaying of single-cell and bulk RNA-sequencing data, this allows us to identify what cell types have been formed by our AFSC-organoids to then prove the multipotency of the AFSCs. I will explain my computational pipeline and discuss the mixture of R-based techniques I use to fully harness this data and facilitate my group's work.
Speaker	Inga Usher
Department - Faculty	UCL Cancer Institute – Medical Sciences
Title	Exploring the multi-omic landscape of chordoma
Abstract	The focus of my PhD is studying chordoma, a rare cancer that affects the skull and spine. Chordoma is treated primarily by surgery with few other options. Both the tumour and its treatment have negative effects on the lives of patients and survival has not improved significantly in the past few decades. Chordoma is hypothesised to arise from remnants of a structure that is normally only present during embryonic and fetal development but the trajectory between development and the established cancer is incompletely understood. Cancer is a genomic disease and chordoma is no different. By studying data from whole genome sequencing, methylation arrays and single cell RNA sequencing I hope to elucidate the biology of chordoma to identify clinically relevant biomarkers and potential therapeutic vulnerabilities. These multi-omic data converge in R where I perform pre-processing, downstream analysis and visualisation.
Speaker	Susanna Pagni
Department - Faculty	Department of Clinical and Experimental Epilepsy – Brain Sciences
Title	The secrets of DNA: a journey with R
Abstract	DNA is the complex molecule found in every cell of our body that determines who we are: from the colour of our eyes to the shape of our face, the geographical origins of our ancestors, which diseases we are predisposed to, and much more. In medicine, it is essential to analyse DNA, observe it, read it and understand the secrets it holds. Today, this is possible with the support of Informatics. In my presentation, I will show you how, with the use of R, it is possible to visualise DNA, observe the dimensions of each chromosome, see how different genes interact with each other, and finally go into detail to recognise a genetic mutation and predict its consequences.

Session 3: Modelling	
Speaker	Naomi Launders
Department - Faculty	Division of Psychiatry – Brain Sciences
Title	Clustering of physical health conditions in people with severe mental illness
Abstract	People with severe mental illness (SMI), e.g. schizophrenia or bipolar disorder, are at an increased risk of physical health conditions, but little is known of how these diseases cluster.

	Using R, I extracted diagnoses of 24 physical health conditions from electronic medical records of 68,392 patients with SMI and 273,568 without. I perform multiple correspondence analysis (MCA) using the FactoMineR package, and K-means cluster analysis of the MCA dimensions. I determined number of clusters by examining Silhouettes and Calinski-Harabaz results, from the ClusterCrit package. I identified six disease profiles, of which five were common to both cohorts. People with SMI were more likely to belong to a "respiratory and neurological disease" cluster and less likely to belong to the "heart disease" or "hypertension" cluster than comparators. While diseases cluster similarly in people with and without SMI, there is a need for interventions aimed at specific risk-groups in people with SMI.
Speaker	Dan Lewer
Department - Faculty	Epidemiology and Public Health – Population Health Sciences
Title	Making a stochastic model of Covid19 in the homeless population of England
Abstract	We used base R to build a model of Covid19 transmission in the homeless population of England. The code is posted here: https://github.com/maxeyre/Homeless-COVID-19. The model uses careful application of primitive functions, subsetting, and integer matrices to allow efficient computation of a complex model of disease transmission. The results were published in Lancet Respiratory Medicine and supported a campaign for better accommodation for people experiencing homelessness in winter 2020.
Speaker	Lucie Gourmet
Department - Faculty	Genetics, Evolution and Environment – Life Sciences
Title	Evolutionary forces shaping phenotypes of cancer progression
Abstract	Mutations promoting cancer development will be positively selected unlike deleterious changes which are removed by negative selection. We are interested in how the mutational landscape is selected in different tumour phenotypes related to cancer progression, including epithelial-mesenchymal transition, cellular dormancy, and immune evasion. We stratified patients for each category to find the specific pressures determining evolutionary trajectories. We used the ratio of non-synonymous mutations to synonymous mutations to measure selection in each individual phenotype. In addition, we investigated the role of the tumour microenvironment to understand its influence on our phenotypes of interest. In addition to a pan-cancer analysis across 31 primary tumour tissues from the Cancer Genome Atlas, we also conducted tissue-specific analyses which enabled us to detect a context-dependent interplay between the immune system and cancer. We show that EMT, dormancy and immune escape affect one another in specific ways.

Organisers: Dr Ellen Webborn (Energy Institute) and Dr Scott Allan Orr (Institute for Sustainable Heritage)

Judges: Dr Kirsten McMillan (Dogs Trust), Dr Owain Kenway (Head of Research Computing), and Prof Philip Luthert (Institute of Ophthalmology and co-chair of the eResearch Domain)