# SatRdays London 2023

Schedule and Abstracts


22 April 2023

Invalid Date

# Schedule

| Time | Title | Room |
|---|---|---|
| 09:00 - 09:30 | Registration | |
| 09:30 - 09:45 | Welcome | |
| 09:45 - 10:30 | Julia Silge | |
| 10:30 - 10:55 | Michael Stevens & Botan Ağın | |
| 10:55 - 11:25 | Break | |
| 11:25 - 11:50 | Russ Hyde | |
| 11:50 - 12:15 | Ella Kaye and Heather Turner | |
| 12:15 – 13:20 | Lunch | |
| 13:20 – 13:30 | Welcome from CUSP London | |
| 13:30 - 13:55 | Andrew Collier | |
| 13:55 - 14:20 | Jack Davison | |
| 14:20 - 14:50 | Break | |
| 14:50 - 15:15 | Vyara Apostolova and Laura Cole | |
| 15:15 - 15:50 | Oli Hawkins | |
| 15:50 - 16:00 | Close | |

# Abstracts

## Keynote Speakers

### Julia Silge

*Data Science Leader at Posit*

**What is "production" anyway? MLOps for the curious**

Many data scientists understand what goes into training a machine learning or statistical model, but creating a strategy to deploy and maintain that model can be daunting. You may have even heard that R is not appropriate for production use. In this talk, learn what the practice of machine learning operations (MLOps) is, what principles can be used to create a practical MLOps strategy, what people mean when they say "production", and what kinds of tasks and components are involved. See how to get started with vetiver, a framework for MLOps tasks in R (and Python) that provides fluent tooling to version, deploy, and monitor your models.

## Oliver Hawkins

*Editorial Data Scientist at the Financial Times*

**Why R is good for journalism**

Data journalism has been an established discipline within newsrooms for more than a decade. But as the number of potential sources of data keeps growing, data analysis is becoming an increasingly important part of journalism more generally. Whether you're covering politics, economics, finance, health, education, crime or the environment, understanding and explaining the relevant data has become a vital part of telling the story.

The Financial Times' visual and data journalism team works with reporters to make data analysis a central feature of the newspaper's coverage. And R has become essential to that task. It's the tool we use to produce and manage datasets, to develop and deploy data pipelines, and to find answers to the questions we address in our visual storytelling.

This talk will explore how we use R for editorial work at the Financial Times, and how we tackle some of the key challenges of computational journalism. Why is R a good choice for data journalism? How does it integrate with other technologies? How do you follow good software development practices while meeting news deadlines? And how do you collaborate and develop skills in a team with different levels of programming experience?

## Botan Ağın

*Data Analyst at SamKnows*

## Michael Stevens

*Data Analyst at SamKnows*

**AutRmatic reporting: billions of internet measurements, hundreds of reports and one repository to rule them all**

SamKnows has been pioneering internet performance measurements for over 14 years. The reason we exist is to provide a source of truth for how the internet is really performing. The data we collect can be used as a common language between government regulators, internet service providers, academics, and content providers to optimise and improve internet performance for everyone.

Day to day SamKnows uses R to handle a huge range of automated and self-serve workloads. Keeping track of each report's recipients, delivery schedule, dependencies and deployment procedure can be tricky, especially in the nightmare scenario of suddenly needing to migrate all of your jobs to a new server or cloud environment.

In this presentation, we will talk about how we structure our regularly-scheduled reports as standardised entities within a monorepo. We will explain how this approach reduces the latency in setting up a report, makes it easier for new team members to contribute, and lets us uphold standards while retaining the flexibility to deliver work in diverse formats with a range of complexity levels and opportunities for manual intervention. We will go into detail on specific workflows that take the terabytes of data collected by SamKnows from cloud and on-premises data sources, process them into an R Markdown document, formatted spreadsheet, and raw CSV output, and distribute them through cloud file storage, FTP servers, email, Slack and more.

## Vyara Apostolova

*Senior Analyst and Data Analytics co-lead at UK National Audit Office*

## Laura Cole

*Head of Modelling at UK National Audit Office*

**ScRutinising government spending**

The National Audit Office supports Parliament in holding government to account both via its Financial Audit and Value for Money work. The Analysis Hub is a central team that utilises a range of analytical techniques to support both strands of work. The proposed presentation will showcase two examples of how we in the Analysis Hub use R to support our mission to hold government to account.

We use R to reproduce complex models that departments employ to produce accounting estimates for their financial accounts. Our R reproductions allow us to assess if departments have implemented their selected methodology correctly and to highlight any model integrity issues. We also implement additional sensitivity testing, including via Monte Carlo simulations to capture the uncertainty around model outputs. The presentation will cover an overview of our approach and a demo of a reproduction of a dummy model.

We have also built a R-shiny app, Covid-19 Cost tracker, that brings together data from across the UK government on the costs of measures in response to the Covid-19 pandemic. It is one of the very few sources of comprehensive information on Covid-19 related spending and the only one as an interactive tool. With it the public can examine spending by department and category of spend as well as interact with bubble graphs to explore the costs of individual policies. The presentation will include an overview of how the data analytics team and audit team collaborated to produce the output and a demo of the app.

## Andrew Collier (presenting) & Bianca Peterson

*Data Scientist at Fathom Data*

**Sidekicks of the Tidyverse**

In the realm of the Tidyverse, there are functions which are always in the spotlight. These are the titans: well known and loved, frequently invoked and virtually indispensable. There are other, lesser-known functions which stand quietly in the shadows. Unacknowledged, somewhat obscure and almost forgotten. Waiting for their moment to shine.

I'll talk about a few of these lesser known (but equally useful) functions, lauding their virtues and showing how they can help you succeed on your next Data Science quest.

## Jack Davison

*Air Quality Measurements Data Analyst at Ricardo Energy & Environment*

**"Put it on a map!" – Developments in Air Quality Data Analysis**

An understanding of air quality is crucial as it can have significant public health, environmental and economic effects. However, air quality data is complex, constantly changing in space and time, and influenced by a myriad of factors such as meteorology and human activity. This makes air quality analysis challenging, and communicating the results of this analysis more challenging still!

Just over a decade ago, the {openair} package was authored to provide an open-source toolkit to help air quality practitioners get the most out of their data, and is still used widely in academia, consultancy and industry today. While {openair} itself has not changed hugely in recent years, much thought has been put into extending it through leveraging more recent tools and packages.

In this talk I will discuss how we have recently married {leaflet} and {openair} to create effective, interactive air quality maps. In particular, I'll discuss the development of the {openairmaps} package – a toolset which makes it easy to create interactive "directional analysis" maps to help explore the geospatial context of pollution monitoring data.

## Russ Hyde

*Data Scientist at Jumping Rivers*

**Does code quality even matter in data science?**

It depends!

If you need to quickly summarise some data for an ad-hoc request, then knock out the code in whatever manner gets the job done.

But what happens when you start getting a lot of similar requests, or you are working on a more substantial project, or you are collaborating within a larger team? Now, productivity should be viewed 'across the team' and 'across all projects'. What can you do to help yourself and your colleagues, and what tools exist to help?

Code quality concerns those aspects of software that make it easier to work with, easier to explain to others and easier to maintain or extend.

In this talk, I'll take you through the source code for an evolving analysis project. We'll discuss how to (and how not to) modularise code. Along the way, we'll talk about actions and calculations, body-tweaking, duplicate stomping and a few tools that help automate the boring low-level stuff that teams sometimes disagree about.

## Ella Kaye

*PhD Candidate (Statistics) at University of Warwick*

## Heather Turner

*Research Software Engineering Fellow*

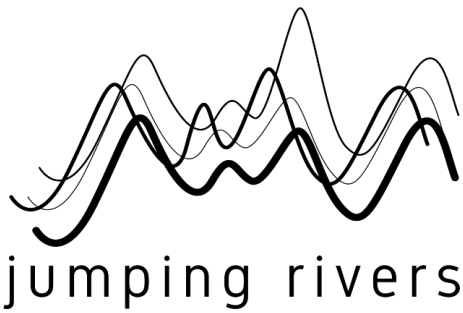
**Sustainability and EDI (Equality, Diversity and Inclusion) in the R Project**

The R Project is over 20 years old, but its future is not secure - many of the R Core Team are nearing retirement and there are not enough new contributors to sustain the work. We present a number of initiatives, organised under Heather Turner's 'Sustainability and EDI (Equality, Diversity and Inclusion) in the R Project' fellowship, to encourage and train a new, more diverse, generation of contributors. These include R contributor office hours, collaboration campfires, bug BBQs, translatathons and an updated R development guide. This presentation is also a call to action to encourage others to get involved in supporting this language, a fundamental piece of software in many disciplines, used by an estimated 2 million people.

# Sponsors

CUSP London is the Centre for Urban Science and Progress based at King's College London, UK. Their mission is to support interdisciplinary research and innovation using Data Science in and for London.
Find them on Twitter @CuspLondon

---




Jumping Rivers is an analytics company specialising in creating bespoke solutions for modern business problems. Their team of data science and engineering experts come from many different backgrounds, and their wealth of knowledge and experience allows them to think outside the box and solve problems in new and innovative ways.
Find them on Twitter @jumping_uk

---


Posit's goal is to make data science more open, intuitive, accessible, and collaborative. They provide tools that make it easy for individuals, teams, and enterprises to leverage powerful analytics and gain insights they need to make a lasting impact.
Find them on Twitter @posit_pbc

---


The central mission of the R Consortium is to work with and provide support to the R Foundation and to the key organizations developing, maintaining, distributing and using R software through the identification, development and implementation of infrastructure projects.
Find them on Twitter @RConsortium