Workshop 1: Integrating R and Python - An Introduction to Getting the Best of Both Worlds
Monday 14th September 1000 – 1300 hrs BRITANNIA Conference Room, Tower Hotel, London

Presenter: Chris Musselle

Overview:
R and Python are both mature and strong open source tools for conducting numerical analysis and scientific computation. Each shines in its own way: with R’s extensive collection of packages for statistical analysis, machine learning and visualisation; and Python’s widespread utility as a general programming language, making it more applicable when you need to source, filter and parse custom data formats, or interact with other applications.

Though there are on-going efforts from both communities to extend the functionality of the libraries available, there are often times when a programmer may wish to use both languages in an analysis pipeline. This could be down to the strength of certain libraries for specific types of analysis, syntactic preferences, or simply the prior knowledge already invested in doing tasks in a certain language.

This workshop aims to introduce some of the ways currently available to integrate the two languages into a single analysis pipeline. The workshop will look at writing and calling command line functions in R and Python as well as some more sophisticated ways of calling out to R from within Python itself. The workflow engine ‘Snakemake’ will also be introduced to provide a way of faithfully reproducing these multi-step data analysis pipelines, and keep track of input and output dependencies.

Outline:
- Introduction
  - Overview of a Data Analysis pipeline
  - Where Python can help
  - What tools are available
  - Integration strategies
- Command Line Options for running Python and R.
  - Writing command line functions
  - Using system calls
  - Using Snakemake as a workflow engine
- Options for running R from inside Python.
  - The rpy2 module
  - PypeR

Prerequisites:
As this is an interactive workshop with short exercises, participants will require a laptop to get the most out of it. Prior to the workshop, participants should make sure they have the following installed:
1. The Anaconda Python distribution: This is the simplest way to get the Python language and many of the extended libraries setup and configured in a one click installer. It is also available cross platform. The workshop will use Python 3. http://continuum.io/downloads#py34
2. Snakemake, a Python library for data analysis workflow management. Installation instructions can be found at https://bitbucket.org/johanneskoester/snakemake/wiki/Home
3. R version 3.1.2

The workshop is an intermediate level workshop aimed at people with a basic knowledge of the R and Python languages. Prior knowledge of Python syntax and data structures will be beneficial but is not essential, as some preliminary material will be sent out ahead of time to make sure people are up to speed in this area.

Note: This workshop is interactive and will include exercises. Participants will require a laptop

Chris Musselle works as a Consultant for Mango Solutions and holds a PhD in Computer Science from the University of Bristol. Whilst having expertise in a number of programming languages, Chris’ specific interest and experience is in Python.