DRAFT

Health Data Research UK Summer School 2019

Contents

1. Introduction

The inaugural HDR UK Summer School will be hosted in St Andrews on 19-23 August 2019. It aims to train participants in core data science skills and on how to apply these to different data sets, from genomic to population level.

The program consists of a mixture of plenary lectures, by leading health researchers, which will focus on different aspects of health data science. Hands-on sessions will consist of computer exercises that will enable the participants to apply statistical methods to the analysis of health data under the guidance of the lecturers and teaching assistants. Networking opportunities are integral to the program and group discussions will be encouraged throughout the event.

Based on feedback from the HDRUK community, we have put together a programme that will cater to individuals with different levels of expertise.

The “Core Data Science Skills” track will address the needs of participants with little expertise in core data science skills while the “Applied Data Science Skills” track will cater to the more experienced participant, already having working experience of Data Science tools and resources.

Schools Engagement Activities (SEA) developed by the participants during the Summer School will be delivered on the final day (Friday 23rd) to students aged 16-18 studying STEM subjects. These developed activities will form the basis of an open source repository for HDRUK.

A hackathon-style activity will support the development of training materials for a Health Data Science focused curriculum of short courses; such curriculum will be created and maintained by HDR UK Fellows, in partnership with the Software Sustainability Institute and clinical experts.

By the end of the programme participants will have gained:

1. An appreciation of how data science is being applied in health
2. An understanding of the challenges associated when working with health data, including ethical considerations
3. Hands-on experience of using key data science software and tools
4. A working knowledge of how these can be applied to a variety of data sets

2. Agenda

For further detail on each session please see section 4.

Monday		
12.00-13.00	Arrival and registration	
13.00-14.00	Lunch	
14.00-14.10	Welcome Professor Peter Diggle, Director of Training, Health Data Research UK Professor Colin McCowan, Professor of Health Data Science, University of St Andrew's	
	Core Data Science Skills (40 max)	**Applied Data Science Skills (including discussion groups)**
14.10-17.30	HealthyR Dr Riinu Ots, Senior Data Manager, Surgical Informatics Research, University of Edinburgh Dr Tom Drake Dr Kenneth McLean Mr Stephen Knight	Analysing eHR data Dr Andreas Karwath, Birmingham Dr Chiara Batini, Leicester
17.30-19.30	Welcome reception & posters	

Tuesday		
08:30 - 9:30	Plenary: Title (TBC) Professor Cathie Sudlow	
09.30-12.30	HealthyR Dr Riinu Ots, Dr Tom Drake, Dr Kenneth McLean, Mr Stephen Knight	Analysing genomic data Dr Chiara Batini, Leicester Dr Andreas Karwath, Birmingham
12.30-13.30	Lunch	
13.30-15.30	HealthyR Dr Riinu Ots, Dr Tom Drake, Dr Kenneth McLean, Mr Stephen Knight	Analysing genomic data Dr Chiara Batini, Leicester Dr Andreas Karwath, Birmingham
15.30-16.00	Tea/coffee	
16.00-17:30	Parallel sessions: SEA preparation/Discussion meetings on topics proposed by Fellows (DM)/Hackathon for training materials development	
17:30 onwards	Social: Beach Art Competition	

Wednesday	
08:30 - 9:30	Plenary: Climate Information for Public Health Action Professor Madeleine Thomson, Senior Science Lead, Climate Change and Health, Wellcome Trust

09.30-12.30	HealthyR Dr Riinu Ots, Dr Tom Drake, Dr Kenneth McLean, Mr Stephen Knight	Introduction to image analysis: deformable image registration Dr Bartlomiej Papiez, Oxford
12.30-13.30	Lunch	
13.30-16:00	HealthyR Dr Riinu Ots, Dr Tom Drake, Dr Kenneth McLean, Mr Stephen Knight	Geospatial data-handling in R Professor Sarah Rodgers, Liverpool Dr Richard Fry, Swansea
16.00-16.30	Tea/coffee	
16.30-18.00	Parallel sessions: SEA preparation/DM/Hackathon for training materials development	
18:00 onwards	Social: Putting at the Himalayas	

Thursday		
8:30-9:30	Achieving 21st century precision cancer control - it all about the data! Professor Mark Lawler, Queen's University Belfast	
09.30-12.30	Introduction to Machine Learning Dr Adriano Barbosa Dr Pilar Cacheiro, QMUL	Applied Machine Learning Matthew Willetts Alexander Camuto, Alan Turing Institute
12.30-13.30	Lunch	
13.30-15.30	Introduction to Machine Learning Dr Adriano Barbosa Dr Pilar Cacheiro, QMUL	Applied Machine Learning Matthew Willetts Alexander Camuto, Alan Turing Institute
15.30-16.00	Refreshment break	
16.00-18:00	Parallel sessions: SEA preparation/DM/Hackathon for training materials development	
19.00	Formal dinner After dinner speaker: Stephen Senn	

Friday	Schools engagement event
09.30-10.00	Welcome and Introduction
10.00-10.30	Living is a Risky Business Professor Jen Rogers, Associate Professor of Statistics, University of Oxford
10.30-12.00	Rotate around three of five stations
12.00-12.45	Lunch
12.45-13.45	Rotate around remaining two stations
13.45-14.00	Prize winners and close of meeting

3. Speaker biographies

Madeleine Thomson is the Senior Science Lead for Climate Change and Health, Our Planet Our Health Programme, Wellcome Trust, UK. Madeleine is a leading expert on climate change and health, having recently held senior research positions at the International Research Institute for Climate and Society (IRI) and the Mailman School of Public Health at Columbia University. Until July 2019 she served as Director of the IRI/PAHO-WHO Collaborating Centre (430) for Early Warning Systems for Malaria and Other Climate Sensitive Diseases. She is a Visiting Professor at Lancaster University, UK.

Further speaker biographies to be added soon.

4. Session descriptions

Note to participants: For sessions not held in a computer lab you will require your own laptop and will need to download the relevant software listed here in advance.

HealthyR*

- Leads: Dr Riinu Ots, Dr Tom Drake, Dr Kenneth McLean, Mr Stephen Knight http://healthyr.surgicalinformatics.org/)
- Within these sessions you will be introduced to a quick start course in R. More details can be found on their website.
- Software requirements: This session will use a computer lab on site, and therefore does not require participants to download software.

Analysing eHR data

- Leads: Dr Andreas Karwath, Rutherford Research Fellow, University of Birmingham & Dr Chiara Batini, Rutherford Research Fellow, University of Leicester
- Within this session we will get a first glimpse of eHR data and possible analysis techniques (simple regression, visualization, clustering, etc.) and more importantly, what relevant questions can be answered.
- Software requirements: preferably: git, Editor (Emacs, nano, or vi - if Windows: Notepad++ or so), Docker with the ability to install arbitrary docker images (including data)
    - Otherwise: Python 3.6+, libraries such as sklean, scipy, pandas, etc. (More detail upon request)

Analysing genomic data

- Leads: Dr Chiara Batini, UKRI Innovation Research Fellow, University of Leicester & Dr Andreas Karwath, UKRI Innovation Research Fellow, University of Birmingham
- Within this session we will have (i) an introduction to genomics that will cover how this discipline evolved and basic concepts to understand genomic data at present; (ii) a group discussion on which questions can be answered with genomic data in the context of health sciences; (iii) an introduction to different technologies, their pros and cons and what kind of data they generate; and (iv) a

practical session that will cover either a sequencing experiment and the initial steps of a NGS pipeline or an example of association testing between genetic and phenotypic variation. Finally, we will briefly discuss follow-up steps to understand the impact of an identified variant.

- Software requirements: Docker
    - We will use a docker container that will include all software needed in the practicals, and web browser if and when needed.

Introduction to image analysis: deformable image registration

- Lead: Dr Bartlomiej Papiez, Rutherford Research Fellow, Oxford Big Data Institute
- The objective of this session will be to introduce deformable image registration as one of the basic image analysis techniques in medical imaging. The first part will be overview of typical image registration frameworks (theory and some popular methods), the second part will introduce possible applications in cancer imaging; third part will give a glimpse of practical use of image registration for exemplar images using one of the presented image analysis toolboxes.
- Software requirements: Jupyter notebook, python (standard packages)

Geospatial data-handling in R

- Leads: Professor Sarah Rodgers, Professor of Health Informatics, University of Liverpool & Dr Richard Fry, Senior Lecturer, Swansea University
- Our objective for this session is to bring spatial thinking into data science. This workshop is suitable to give all non-spatial R coders an introduction to what you need to think about to incorporate open source map data into analyses, and to visualise your data.

    - 13:30-14:15 General introduction to Geography and Health - why place matters
        - An introduction to the principles and theory of geography
        - How spatial data can help us in our research (data linkage UPRNs/RALFs, etc.)
        - Ecological fallacy and the Modifiable Area Unit Problem (what happens if you don't do this – theory)
        - R as a GIS
    - 14:15-15:45 - **Practical**
        - Using R as a GIS
            - The R Spatial Libraries
                - tidy data and simple features
            - Different spatial data types in R
            - Loading spatial data
            - Displaying spatial data
            - Joining spatial and non-spatial data
            - Creating maps of basic rates using small area geographies
            - Creating maps of refined rates using spatial data and population statistics
            - Interactive mapping in R
    - 15.45 - 16:00 Summary and questions

- Software requirements: R version 3.4.4 or above; R Studio 1.1.4 or above; operating system: Windows 7+, Ubuntu 14+, Mac OS X 10.12+

Introduction to Machine Learning

- Leads: Dr Adriano Barbosa, UKRI Rutherford Fund Innovation Fellow/Lecturer, QMUL & Dr Pilar Cacheiro, Research Fellow, QMUL

    - 09:30 - 10:00 General Introduction to Machine Learning;
    - 10:00 - 10:30 Exploring the R-studio environment and the Caret package
    - 10:30 - 11:00 K-Means Clustering;
    - 11:00 - 11:30 K-Nearest Neighbour Algorithm
    - 11:30 - 14:30 Study case: Diabetes Prediction using:
        - Linear Regression
        - Decision Trees
        - Random Forest Classifiers
        - Support Vector Machines
    - 14:30 - 15:00 Study case: Breast Cancer Prediction
    - 15:00 - 15:30 Study case: Heart Disease Prediction

- Software requirements: This session will use a computer lab on site, and therefore does not require participants to download software.

Applied Machine Learning

- Leads: Matthew Willetts & Alexander Camuto, Alan Turing Institute
- During this session, we aim to cover contemporary methods for imaging data, both supervised and unsupervised - so CNNs and GANs/VAEs, as well as current state of the art methods for sequence data.
- Software requirements at https://github.com/MatthewWilletts/HDRUK_AppliedML