Final report - JISC Data Champions 2017-2018

Sergio Martínez Cuesta
University of Cambridge Data Champions

Summary

The main aim of my JISC Data Champions project was to develop and deliver training and engagement activities in research data management and analysis for students and early-career researchers in colleges and departments at the University of Cambridge. Early in the project I came to realise that stand-alone training sessions focused exclusively on data management concepts were not successful as researchers found them too abstract, uninteresting or detached from their day-to-day research activities. To find new ways to improve take-up and facilitate adoption of good practices, I went on to explore surveys of computational training needs collected throughout the university and became involved in international training projects such as the Data and Software Carpentries. After discussing with other University of Cambridge Data Champions[1], we realised that by means of learning computational and programming concepts, researchers would automatically learn basic data management concepts too. As a result, I decided to embed training in data management within short courses focused on data analysis, visualization and version control, where data management itself played only a small part. Researchers had also the opportunity to bring their own datasets to the newly developed sessions and put the new concepts learnt into practice directly. Overall, training became more targeted, personalized and engaging for their work.

Activities

Training

1. *Introduction to GitHub for chemists[2], 24th October 2017, Department of Chemistry, University of Cambridge*
Git is a free and open source distributed version control system, which allows researchers to create repositories to host projects and track changes automatically. GitHub is the web platform that allows the upload, management and sharing of git repositories online. This 1h course was developed in collaboration with Clair Castle and had theory and practise sessions (30 min each) in version control. We had 28 chemistry researchers and students attending, some of which brought their own devices to follow the practical part. We also had researchers with broad expertise in version control, who contributed to the event by sharing their knowledge. A follow-up session was later delivered on the 14th December 2017 to members of the Balasubramanian group where I do my day-to-day research. All training materials (including feedback) are openly available as an online course.

2. *Make your academic life easy with ORCID: an introduction[3], 23rd February 2018, Department of Chemistry, University of Cambridge*
Many funders and publishers now require researchers to get an Open Researcher and Contributor ID (ORCID): an identifier that is unique and comes with researchers wherever they go. It eliminates *ambiguity* on the author's name so researchers can get full credit for their work. This 1h session was also developed in collaboration with Clair Castle and had ~20 members of the Department of Chemistry attending. It was not only of interest to chemistry researchers but also to administration personal

e.g. assistants, who manage the online presence of research group leaders and heads of units in the department. Training materials are openly available.

3. *Introduction to R[4], 15th March 2018, Wolfson College, University of Cambridge*
R is one of the most widely used programming languages for data analysis, statistics and visualisation in academia and industry. It is open-source, available in all computing platforms and supported by a broad community of software developers and researchers who contribute R packages and libraries to many fields of research. This 1.5h course provided a short beginners introduction to manage data and analyses in R. I demonstrated basic examples on how to input, explore, plot and output data using a dataset containing anonymised information from 100 lung cancer patients aged 42-44 from different states in the US. We had ~10 researchers of different backgrounds attending the workshop. They brought R and RStudio installed on their laptops to be able to follow along and some shared their own research datasets. Many found useful the ability to integrate all their analyses in individual scripts, which readily facilitates reproducibility in research.

4. *Data visualisation with R and ggplot2[5], 31st May 2018, Wolfson College, University of Cambridge*
This 1h course was a continuation of the previous R course and provided a short beginners introduction to creating and managing graphics using the popular library ggplot2. I demonstrated basic examples on how to import data, perform different types of plots and export graphics using R standard functions and the library ggplot2. We also had ~10 researchers of different backgrounds attending this workshop.

Team building

I was involved in creating several data-related groups at the University of Cambridge during 2017-2018:

- *Chemistry Data Champions[6]*. A group of six researchers engaging members of the Department of Chemistry in good management of chemical data and exchange of open data formats.
- *CRUK-CI Data Champions[7].* Together with members of the Bioinformatics core we run a course to teach students and researchers how to avoid data disasters[8].
- *Wolfson College Data Science Group[9].* A community of students and researchers interested in presenting and discussing research involving big data across different disciplines.

Dissemination

1. All materials developed during this award were made openly available in GitHub and uploaded to ELIXIR's (Europe's distributed infrastructure for life-science data) training platform TeSS[10].

2. *Cambridge Data Champions Forums[11], 14th September 2017 and 11th January 2018, University of Cambridge*
These forums involved discussions with Cambridge Data Champions where ideas for activities were put forward.

3. *Genome Informatics 2017*[12], *1st-4th November 2017, Cold Spring Harbour Laboratory, New York*
This is one of the best world conferences on computational methods in genomics and data management and analysis in life sciences. Curious to know more about how computational researchers in the US tackle data management in life sciences, I presented the computational methods developed in my research, and had the opportunity to discuss data management activities and strategies with attendees.

4. *Engaging Researchers in Good Data Management*[13], *15th November 2017, St Catharine's College, University of Cambridge*
I was part of the organisation team for this conference together with colleagues from the Office of Scholarly Communication in Cambridge, Jisc and SPARC Europe. Marta Busse-Wicher and I presented a talk "A World of Activities by Cambridge Data Champions" giving an overview about the Cambridge programme as well as presenting some of the activities developed to engage students and researchers.

5. *Applied Bioinformatics in Life Sciences, 8th-9th March 2018, KU Leuven, Belgium*
Similarly to *Genome Informatics 2017*, this was a fantastic opportunity to interact with other computational biologists and bioinformatics researchers from Europe. There were keynote talks from speakers involved in developing ELIXIR - Europe's life sciences data management infrastructure. I presented my computational research and introduced my data engagement and training activities too.

Spending information

Expenditure	Activity	Date	Cost (£)
Speaker and refreshments costs	Introduction to GitHub	Oct 2017	141.81
Conference registration fee	Genome Informatics	Nov 2017	933.71
Flights to NY JFK	Genome Informatics	Nov 2017	464.80
Gatwick airport hotel accommodation before NY JFK flight	Genome Informatics	Nov 2017	72.50
UK return train tickets to Gatwick airport	Genome Informatics	Nov 2017	48.40
NY local trains, underground and food costs	Genome Informatics	Nov 2017	44.55
ESTA fee to enter the US	Genome Informatics	Nov 2017	10.70
Poster printing costs	Genome Informatics	Nov 2017	18.00
Conference registration fee	Engaging Researchers	Nov 2017	95.00
Conference registration fee	Applied Bioinformatics	Mar 2018	428.89
Trains and hotel costs	Applied Bioinformatics	Mar 2018	241.52
Trains and hotel costs 2	Applied Bioinformatics	Mar 2018	0.69
Total:			**2500.57**

Acknowledgements

Special thanks go to the community of Cambridge and Jisc Data Champions for ideas and input for activities. Thanks to Liam Bond from the Grants office at the CRUK-CI for managing the grant, my colleagues in the Balasubramanian research group and our audience in the training activities for useful feedback.

References

[1]. https://www.data.cam.ac.uk/intro-data-champions
[2]. https://github.com/semacu/20171024_GitHub_Chemistry_Cambridge
[3]. https://github.com/semacu/20180223_ORCID_Chemistry_Cambridge
[4]. https://github.com/semacu/20180315_IntroductionToR_Wolfson_Cambridge
[5]. https://github.com/semacu/20180531_DataVisualisationRggplot2_Wolfson_Cambridge
[6]. https://www-library.ch.cam.ac.uk/data-champions
[7]. https://github.com/crukci-bioinformatics/data-champions
[8]. https://datachampcam.github.io/avoid-data-disaster/
[9]. https://www.facebook.com/groups/395500144179243/
[10]. https://tess.elixir-europe.org/
[11]. https://github.com/semacu/20170914_CambridgeResearchDataChampionsForum
[12]. https://github.com/semacu/20171101-04_GenomeInformatics
[13]. https://github.com/semacu/20171115_EngagingResearchersinGoodDataManagement