CODATA-RDA 'Research Data Science' summer school report

Research data in the 21st century

Antonella Zane

University of Padova

In the period 1-12 August 2016 I attended a summer school on methods, tools and competences in data science indispensable for the research in the 21st century.
The course, which was held in Trieste at the **Abdus Salaam International Centre for Theoretical Physics (ICTP)**[1], was free and ideally dedicated to young researchers from across the world. It was organised with the collaboration of **CODATA** (The Committee on Data for Science and Technology)[2], **RDA** (Research Data Alliance)[3] and **TWAS** (The World Academy of Science)[4]. Sponsors such as **Godan** (Global Open Data and Agriculture and Nutrition)[5] and **GEO** (Group on Earth Observation)[6] issued attendees from developing countries with allowances for travelling, board and accommodation. Partners in this project were the non-profit organisations **Data Carpentry**[7] and **Software Carpentry**[8] which provide researchers with competences and skills in research computing.
Out of 315 applicants, 80 were selected based on their resumes and admission questionnaires. For 11 hard-working days they had theory lessons, hands-on computer workshops, teamwork and optional evening seminars on *Author Carpentry*[9] held by the course resident librarian Gail Clement (CalTech University) - for 93 total hours.
Fifteen lecturers[10] from different countries and a dozen tutors helped manage workshops for the great number of attendees. All activities were in English, and the friendly and engaging atmosphere favoured opinion exchanges between people.
The aim of the organising institutions was **to promote the quality, availability policies and necessary competences to improve the use of research data** that today, thanks to new technologies, may be collected and reproduced more efficiently than ever before:

1 https://www.ictp.it/.
2 http://www.codata.org/.
3 https://rd-alliance.org/.
4 http://twas.org/.
5 http://www.godan.info/.
6 http://www.earthobservations.org/index.php.
7 http://www.datacarpentry.org/.
8 http://software-carpentry.org/.
9 http://libguides.caltech.edu/authorcarpentry.
10 http://indico.ictp.it/event/7658/speakers.

Contemporary research – particularly when addressing the most significant, transdisciplinary research challenges – cannot be done effectively without a range of skills relating to data. This includes the principles and practice of Open Science and research data management and curation, the use of a range of data platforms and infrastructures, large scale analysis, statistics, visualisation and modelling techniques, software development and annotation and more. We define 'Research Data Science' as the ensemble of these skills.[11]

The challenges

One of the aspects our lecturers often emphasised is the great challenge posed by Big Data, i.e., large amounts of perhaps unstructured and heterogeneous data that different systems increasingly produce, as well as their inter-relationships, which are progressively more common in everyday life (Internet of Things, GPS, social networks, etc.).
In the research field, data management requires new competences and new professionals, like data scientists and data engineers. The former create models and update them continually, the latter create systems.
I attended this course for two reasons: firstly because, as a former researcher, I was keen on finding out how research data are managed today and learning about the tools currently available to carry out this activity. Secondly, as a librarian, I wanted to know both what the researchers' true needs are and what services the University Library System has to support them.
The aim of this report is to spread the knowledge about this intriguing initiative that will probably be repeated in future years and which may arouse the interest of anyone seeking advanced training in a stimulating international context.

The value of data

Data and metadata are essential to the knowledge generated by research activities, and may elicit new research themselves. Their preservation and maintenance over time are indispensable to ensure results are re-used and reproduced.
Good research needs good data. One whole day of the course, with lectures and a lot of teamwork, was dedicated to Research Data Management (RDM), the data management and assessment activity carried out throughout the period in which data are of scientific interest, with the aim of getting the most out of them, in order to share and reuse them. RDM includes the Data Management Plan, a document containing detailed information about data produced within projects, and which is increasingly called for by institutions funding research projects.

11 http://indico.ictp.it/event/7658.

This new paradigm requires researchers, who traditionally are absorbed in carrying out research activities they are best suited for, to plan ahead their use, production and management of data, not only in view of their specific targets, but also taking into account data *as such.* The data they create then potentially flow into the scientific community's shared wealth, which is at the basis of *data-driven research*. In this new paradigm, therefore, also librarians play a role. As Anelda van der Walt[12], a member of the Summer school organising team said,

> "*Library Carpentry aims to teach librarians skills they will need to support their researchers in the 21st century – just as researchers are learning about new tools and methodologies, librarians also need new skills to work with bigger data, harness the power of the internet, etc."*

Working the black seam

To efficiently and effectively manage big data, researchers may now access a broad ecosystem of time-saving computer programmes and systems ensuring greater efficiency (statistics, data processing even with parallel calculations and their visualisation) than traditional spreadsheets.
Below is an overview of the programmes, systems and concepts we used, for each of which I report some aspects of my experience.

Unix Shell

Shell (command line) is an interactive programme that enables you to start other programmes while natively incorporating its own set of instructions with which new, even very complex scripts can be created.
In the course we worked in *GNU/Linux (Linux Mint)* and *Bash*, one of the most common shells. We used *Nano* as script editor.
Why use shells and do more work? Because with one command I can do many things! The command line enables different, more powerful programmes to interact with each other (*cat*, *curl*, *cut*, *find*, *grep*, *head*, *sort*, *uniq*, …), thus chaining the entry or exit data flow from each of them (pipeline), for instance, to list files and arrange them by size, verify their consistency, check, arrange and compare the contents of 2 files, chain more commands in one line, and get results in a matter of seconds.

Git

It is a version control system – basically, this programme helps you keep the development of your software and your documents under control. Git is also command line software (although graphic

12 https://www.linkedin.com/in/aneldavanderwalt.

interfaces are available).
In the course, we created a repository, familiarised with the main commands and simulated different situations that may arise when you use Git both by yourself and in a team with other people.

R/RStudio (ggplot2, tmap, Shiny)

R is both a language and a programme that offers a wide range of statistical and visualisation techniques thanks to packages that can be easily installed, like *ggplot2* and *tmap*, that enable users to customise graphs and plots. R does not only produce still graphs, but also interactive visualizations with its web application *Shiny. R* is a free alternative to the analogous proprietary programme *S*.
For workshop training purposes we downloaded the public dataset *gapminder* from a CRAN[13] mirror and even carried out some advanced table and graph visualisation exercises with these data.

Visual Design

One day was devoted to visual data analysis, both from the theoretical viewpoint and from that of software tools suited for the graphic processing of big data (e.g. *Tableau*, *VizQL*, *Protovis*, etc.)
We trained with the *5-design sheet* methodology and manually drew different graphic solutions to represent data according to models *What, Why* and *How.*

SQL and SQLite

SQL is the language to communicate with a relational database. SQLite is a lightweight, easy-to-install DMBS available for the most popular operating systems, and is frequently used as back-end for other programmes. Although it cannot handle big data, it is a perfect solution for small- and medium-sized applications and for personal study.
In the course, to access and handle data, we used a database created with SQLite, both with its specific Firefox plugin *SQLite Manager* and with direct SQL commands.

Machine Learning and Recommender System

Machine Learning (ML) is a sector of artificial intelligence that makes machines 'intelligent', i.e., capable of teaching themselves using algorithms that learn from the data provided.
The difference between traditional programming and ML lies in the fact that in the former, when designing software, programmers define all the logical conditions and predefined reactions of the system, while in the latter computers partly teach themselves what to do. The techniques used by ML are recommendations (e.g., suggestions provided by Netflix and Amazon according to the customers' preferences), clustering (e.g., Google News) and classifying (e.g. spam filters).

13 FTP and WEB Server networks that store identical copies of coded versions.

At the course, we worked with datasets provided by *MovieLens*[14], a website that recommends its users which films to watch based on their preferences (ratings and tags).

Internet of Things (IoT)

It has been calculated that by 2020 there will be more than 25 billion devices connected via dedicated platforms (e.g. smartphones, smart TVs), leading to amazing developments in the realms of home automation and intelligent cities, for instance.

Data Science Applications and use cases

What can I do with Big Data? Given their quantity, I can gather them and carry out statistical analyses (e.g. data warehouse) as well as *indexing, querying and searching, knowledge discovery (data mining, stats modeling) and data-driven (predictive, deep learning) research*. Data science applies to all data – for instance, data science is collecting information on crimes occurred in a particular quarter at a particular time to gauge the police force required to tackle them.
The USA will soon need competences in this field, and they are already looking for about 190,000 predictive analysts and 1.5 million managers/analysts who can take decisions by analysing big data.

Neural networks

In the second to last day we analysed data with the support of neural networks, and mentioned some of the most typical approaches and techniques.
We used *R* with *kohonen* and *neuralnet* packages.

Landscape of Research Computing

What can I do if I have a conference in one week and I must still process a huge amount of data? I can increase my calculating capacity using *nodes* (a node is a processor, which is often a virtual machine itself), perhaps borrowed from cloud providers.
In a practical session, we were given credentials to access high throughput computing resources of the Open Science Grid[15], on which we practised with the following workload management systems - *Condor* for calculus jobs, and *DAGMan* for dependency management. Lastly, we created our own Virtual Machine in the *OpenStack*[16] environment of the *Jetstream*[17] centre.

14 https://movielens.org/.
15 https://www.opensciencegrid.org/.
16 https://www.openstack.org/.
17 http://jetstream-cloud.org/.

Further study

In a relatively short time, thanks to the enthusiasm, competence and experience of lecturers and organisers, we also learnt about the main features of a dozen families of software tools/systems we can combine to manage, process and visualise our research data.
Clearly, I cannot draw up an exhaustive report of what was seen and examined in the very intense 11 days of the course – it is not the purpose of this document anyway – so please get in touch with me for any further information.
The syllabus of the course and all teaching materials presented can be found here http://indico.ictp.it/event/7658/.
If you wish to organise a Data and Software Carpentry workshop at your premises here is what you may need:

- information about the people involved and planning a workshop: http://software-carpentry.org/workshops/operations;
- informing colleagues about the workshop: http://software-carpentry.org/workshops/pitch;
- application form: http://software-carpentry.org/workshops/request/.

The lecturers' recommended further reading

- Bouiton, G., Babini, D., Hodson, S., Li, J., Marwala, T., Musoke, M. G. N., Uhlir, P. F. & Wyatt, S., *Open data in a big data world*: http://www.science-international.org/sites/default/files/reports/open-data-in-big-data-world_short_en.pdf (short version) http://www.science-international.org/sites/default/files/reports/open-data-in-big-data-world_long_en.pdf (extended version)
- Gray, D., Brown, S. & Macanufo, J. (2010), *Gamestorming: A Playbook for Innovators, Rule-breakers, and Changemakers* , O'Reilly Media, 290 p.
- Munzer, T. (2014), *Visualization Analysis and Design,* CRC Press, 428 p.

Tools/Technologies/Projects mentioned

- colorbrewer2.org – an online tool to help you choose good colour combinations for maps and other graphic elements;
- hadoop.apache.org – framework for the distributed processing of big data, such as unstructured data gathered by social networks and Internet of Things;
- www.kaggle.com (technological start-up that provides data scientists with working environments and a set of services for data management).
- Lupi, G. (2015), *Sketching with Data Opens the Mind's Eye* http://news.nationalgeographic.com/2015/07/2015704-datapoints-sketching-data/.
- Lupi, G.& Posavec, S., *"Dear Data" Project:* http://www.dear-data.com/about/.

Acknowledgements

I would like to thank my coursemates Elena Bertossi and Tanja Wissig for encouraging me and for reading this document.