Data Science and Astronomy : careers and pathways

Authors: Manodeep Sinha, Rachel Webster, Rebecca Lange, Simon Mutch, Darren Croton.

May 2020.

This short white-paper is prepared as input into the Mid-term Review of the Australian Decadal Plan for Astronomy prepared under the auspices of the Australian Academy of Science. Astronomy is a discipline which routinely utilises massive datasets, both for observational and theoretical studies, and hence trains and employs a significant group of ECRs to undertake its research programs. Many of these ECRs eventually take up employment outside Astronomy in areas as diverse as banking and finance, robotics, medical research and development including medical imaging and bioinformatics, climate change and meteorology, etc. A significant number are also employed in the technology giants, such as Google, Microsoft, Atlassian etc. The rise of such essential technical roles (e.g., data scientists) has occurred relatively recently and on a short timescale. Consequently, there is no universally accepted procedure for recognising the contribution of such technical employees. Some industries, in particular the large digital technology firms, have appropriate recognition of the skillsets involved in these activities, many other organisations are struggling to understand where, in their matrix of roles, employees such as data scientists, might fit.

The aim of this paper is to articulate the critical role that data scientists now play in nearly every aspect of the Astronomy discipline's research activities, particularly in government agencies and universities, and to initiate a conversation that encourages appropriate recognition and stability in their employment. Without a stable career for data scientists within astronomy, there is a looming crisis arising from losing highly technically skilled researchers to industry jobs.

Background

Data science will be used throughout this paper to include a wide range of activities related to 'Big Data' and large-scale supercomputing. Job activities that fall under this label include multi-dimensional visualisation, data stewards, data architects, and research software engineers. Algorithms for handling the vast datasets include but are not limited to artificial intelligence algorithms (AI), (deep) machine learning (ML) and other statistical techniques.

Over the past 12 months a wide range of reports highlighting the importance of software in modern research have been prepared by learned academies, such AAST and AAS, specialist government and semi-government organisations, such as Data61 and Australian Standards, and international organisations such as the OECD, and the European Union. An incomplete, but useful list appears at the end of this paper. These reports have addressed a wide range of issues, but the over-riding conclusions can be summarised as:

1. The landscape for the generation, delivery and use of Big Data is evolving very rapidly, with many organisations struggling to incorporate appropriate responses.
2. Big Data and the use of AI will change the economic landscape of many businesses and require them to integrate the opportunities afforded by the analysis of vast datasets related to their activities.
3. There is a chronic shortage of skilled data scientists that will only become more acute over the next few years. It has been estimated that 'Australian industry needs up to 161,000 new specialist Artificial Intelligence (AI) workers by 2030 in machine learning, computer vision, natural language processing and other AI technologies.'[1] , (note: Astronomy is unlikely to provide a large number of these, but can train the AI data scientists working at the highest levels)
4. Public institutions, such as government and universities, are struggling to structure employment for highly skilled data scientists in a sufficiently attractive way to ensure that they can at least compete for the best practitioners. Data science is such a new profession that it is not even classified as a job title in the ANZSCO classification scheme.

Astronomy PhD → Data Scientist

We are undertaking a survey of Australian Astronomy PhDs from 2008-2019 to understand who they are, and where they go after their PhDs. As yet the survey is not complete, but we have preliminary data on 200 of the ~350 graduated PhDs from that period. Of those, 30% are female and 50% are still working as astronomers. We expect that once the data collection is complete, we will find a declining trend of continuing employment within Astronomy – more recent graduates are more likely to be in Astronomy, with the overall percentage remaining in Astronomy getting smaller with increasing time from PhD graduation date. 30% of the cohort are working as data scientists, 10% in various educational roles and the final 10% have not been traced. Thus, Australian Astronomy is graduating about 30 highly sought-after PhDs per annum, with more than a third moving into highly technical roles in data science[2].

Data Scientist Career in Academia or Government

Massive datasets (or data rates) require appropriate people and software to fully utilise the data. But such people do not fit into the traditional academic/professional roles where productivity is measured by the number of papers or number of people supervised. Data scientists are also highly valued by industry, making it even more difficult to retain them within academia/government. The skill set required includes:

- Strong research skills (PhD + experience)
- High Performance Computing (HPC)
- Visualisation
- Big data flows
- Research Data management (including data provenance)

[1] Hajkowicz, Data61 AI Roadmap 2019

[2] Australian Astronomy PhD Destinations: 2008-2019. A survey undertaken by Webster, Millar et al with the support of the ASA and Astro3D.

- Statistical and data analytics
- Problem solving

Within academia, it is difficult for front-line researchers to cover both the science interpretation and the technical analysis. Increasingly the ‘hands-on’ data analysis is being undertaken by junior researchers – PhDs and ECRs. As they complete their degrees, there is a loss of deep technical skills from the research community. Yet it is clear that ‘bleeding edge’ coding and analysis skills are required for many research programs. Indeed, much of the innovation in the area of multi-disciplinary data science are driven in these cross- disciplinary research programs. Multi-skilled teams are now required for most large collaborative projects, but presently there is a poor career structure for the data scientists. Many are on soft money, with short-term positions of 2-3 years, with an indeterminate future. The traditional academic path still has metrics based on refereed papers, which is not a sensible output for a data scientist. Yet these scientists are critical, not only to the future of Astronomical research but the research endeavours in many other disciplines including government agencies. Promotion and hiring structures in academia (and also the government) need to be re-structured to recognise the significant outputs of data scientists and reward them with a proper career structure.

What are the skillsets that need to be fostered and recognised?

The productivity and impact of data scientists can be determined using appropriate academic metrics, including:

(i) KPIs related to projects;

(ii) Core competencies: identify skills required for a research project; innovation; cross-pollination of techniques across research boundaries; how many researchers/projects were enabled;

(iii) Complexity of the projects

(iv) International linkages eg International Virtual Observatory Alliance; Open source software, e.g. Astropy; IEEE (institute of Electrical and Electronics Engineers)

(v) International collaborations, with key software development groups

(vi) Publication: GitHub repositories; papers in dedicated software journals such as Journal of Open Source Software (JOSS)

(vii) Heuristics (e.g., cognitive complexity) to measure code quality and re-usability

Training Astronomers for the Digital Era

The Astronomy community has supported the development of a number of training opportunities, although there is not really coherent framework for a consistent set of basic skills or the possibility of deeper skill sets. The sector is evolving rapidly with significant changes in Information Technology and AI software developments on decade timescales. Thus, more senior academics also need the opportunity to develop their skillset to keep pace with the current practices/tools. Core competencies for PhD students need to more accurately reflect the importance and critical role of software and recognise and appropriately reward the ECRs/MCR who develop these skills.

Astronomers are fully embedded in the digital era, and frequently working with extremely large datasets and/or collaborations. Writing software is now a vital part of a researcher's toolbox. The list of potential skills that astronomers might develop include: version control; shell scripting; python/C/C++; open-source tools; HPC access and/or distributed computing/cloud; basic data management; FAIR data principles; strong statistical skills to understand data and the limitations of the data; ethics in Artificial Intelligence and Data Science; resourcing for computing; cloud skills for business.

Progress has been made towards wider training for Astronomers in digital tools. Ideally a suite of courses would already be available to Physics undergraduates, focusing on some of the basic techniques such as hydrodynamical codes, n-body codes and planet formation simulations. ADACS has a suite of online and in-person courses to provide the elementary skills and could continue to develop a comprehensive database of training opportunities and short courses.

As the skillsets develop, more specific and informal meet-ups and hacks can be arranged to ensure the regular development of ideas and provide forums for understanding the deeper requirements of individual researchers. Specifically, opportunities for more senior researchers to learn new techniques and 'get their hands dirty' are required. With online activity now the norm, plenty of opportunities exist, free of geographical limitations, to tailor the collaborative environment to be fit for purpose.

However, incorporating large-scale training for Astronomers does require a coordinated approach to maximise the benefits to the community.

References: (incomplete!)

'Artifical Intelligence' Hajkowicz + 2019, Data61 AI Roadmap
https://data61.csiro.au/en/Our-Research/Our-Work/AI-Roadmap

'An artificial Intelligence Standard Roadmap: Making Australia's Voice Heard' (2020) Australian Standards
https://www.standards.org.au/getmedia/ede81912-55a2-4d8e-849f-9844993c3b9d/1515-An-Artificial-Intelligence-Standards-Roadmap12-02-2020.pdf.aspx

Artificial Intelligence and Public Standards UK (2020)
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/868284/Web_Version_AI_and_Public_Standards.PDF

OECD Global Science Forum: Report being finalised (Michelle Barker, Chair)
"Buliding digital workforce capacity and skills for data-intensive science"
https://www.innovationpolicyplatform.org/www.innovationpolicyplatform.org/digital-skills-data-intensive-science-oecd-project/index.html

The Effective and Ethical Development of Artifical Intelligence (2019) ACOLA
https://acola.org/hs4-artificial-intelligence-australia/

Recognising the Importance of Software in Research – Research Software Engineers (RSEs), a UK Example
Directorate-General for Research and Innovation (European Commission); Study on Open Science: Monitoring Trends and Drivers
https://op.europa.eu/s/n42b

Career Paths and Prospects in Academic Data Science: Report of the Moore-Sloan Data Science Environments Survey (2018)
A Joint report Berkeley Institute for Data Science at UC-Berkeley, the eScience Institute at UW-Seattle, and the Center for Data Science at New York University
https://osf.io/preprints/socarxiv/xe823/