Professionalising Data Science Roles Workshop

Pre-workshop material

1. Overview of the Project

The National Audit Office report on 'Challenges in using data across government' highlights the current gap in data skills at several levels. The skills gap is compounded by a lack of clarity in definitions for data science roles and their skill requirements.

The National Data Strategy highlights that there is no widely agreed definition of data skills and the role descriptors are used inconsistently across different institutions. To address this urgent issue, we are carrying out the Turing Skills Policy project '**Professionalising traditional and infrastructure research roles in data science**'. Our goal is to standardise and strengthen specialist research infrastructure roles alongside the traditional roles in data science. Outputs from this project will contribute to addressing the gap in the policy landscape by informing policymakers, institutions and the workforce of the skills needs and career pathways in data science. As a Turing Way project, learnings and resources from this project will be shared in *The Turing Way* handbook.

Overview of the workshop

On 2 October 2023, we are hosting the **Professionalising Data Science Roles Workshop** at The Alan Turing Institute as part of this project. This workshop will facilitate discussion among experts from across different sectors to identify core competencies and skills for specialist roles established as well as needed in the future in diverse areas of the data science ecosystem.

We will use "skill areas" as defined by the Alliance for Data Science Professionals (AfDSP) standards as the starting point to enable discussions. Through two breakout discussions, we will map skills areas for different roles and further expand on the specific evidence or professional requirements. Outputs from this workshop will be shared openly with permission and attributions.

2. Skills Areas and Frameworks for Data Science Roles

The AfDSP standards guide the professional accreditation of data scientist roles by defining different 'skill areas' and pairing them with evidential requirements as professional responsibilities. These standards define data scientists as an 'overarching profession'. Data science as a profession is continually evolving to address emerging challenges in the data landscape (details in 2.1). Throughout the lifecycle of a project, the data science team requires more specialised skills and roles that the designated data scientists can not address effectively alone. We provide a summary of how different teams at The Alan Turing Institute make up our data science and research teams (details in 2.2).

Using **AfDSP standards' skill areas as a starting point**, this project will expand the professional requirements, which will be presented as skills frameworks (details in 2.3) for different specialist roles. Below we provide an example skills framework we have developed for the Research Community Management which will used as a reference for developing frameworks for different specialist roles (details on Zenodo: Sharan & Karoune, doi: 10.5281/zenodo.8337627).

2.1 Challenges Leading to Skills Gap in Data Science

In the ever-evolving landscape of data science, research is becoming more complex and challenging. To drive forward high-impact data science research, appropriate support and attention must be given to some major interrelated issues such as research quality, research sustainability, interdisciplinarity, and Equality, Diversity, Inclusion and Accessibility (EDIA).

- **Research quality** - The reproducibility crisis raised awareness in the research community of the need for more transparency and accountability in research to help improve the overall quality of research. Open research practices are considered the way forward in this regard but to implement these practices in data science there is a need for specialist trainers and other open science professionals to upskill and contribute to research teams.
- **Research sustainability** - The need for more sustainable research also stems from the open science movement and includes the problems of cost-effectiveness and the environmental cost of research. To address this issue, researchers must consider and implement sustainable archiving practices such as the FAIR data principles and also ensure high-quality research. These practices again require specialised skills and are often fulfilled by roles such as data stewards or librarians.
- **Interdisciplinarity** - Data science is by nature interdisciplinary and requires deliberate knowledge exchange and collaboration with different stakeholders. Alongside researchers, it is also important to involve educators, engineers, policymakers, government, professionals and the public at different stages of research to make it as impactful and responsible as possible. This aspect of data science surfaces the complexities of working across domains that can be addressed by specialist roles to ensure the involvement of the right stakeholders throughout the research.
- **Equality, diversity, inclusion and accessibility (EDIA)** - EDIA needs to be embedded in all the research that we undertake. This is more complex than we think as inequity is in every part of data science - project design through to research dissemination and their impact on society. These interrelated issues require data science teams to embed stronger considerations and skills to ensure inclusive and ethical research and broader adoption of open science and reproducible practices.

Important issues that go hand-in-hand are the institutional recognition and incentivisation of different kinds of research roles and contributions. With these main considerations, this workshop aims to understand and curate what skills are needed to undertake the expanded and more specialist data science roles. These curated resources will be communicated broadly to inform national policies.

2.2 Data Science Teams: An Example from The Alan Turing Institute

To embed and maintain considerations for research quality, research sustainability, interdisciplinarity and EDIA, as well as to take on thorny research questions, we apply a (big) 'team science' approach in our research at The Alan Turing Institute.

⚠️ ***Important notes:*** *There are many aspects that go into ensuring these considerations. The team science approach is what we take to integrate diverse experts in addressing these problems collaboratively! However, different institutes may take different approaches.* ⚠️

Most research projects at the Turing involve multiple partners, which speaks to one of our institute's principles - "collaborate and convene". Our collaborative ways of working reflect the complexity of the projects and the stakeholders that we are convening, which include partnerships across sectors, such as academic organisations, research institutions, government, third sector and industry partners. The scale, complexity and interdisciplinarity of these projects require specialists who bring leadership and skills in data science, engineering, collaboration, community building and management, alongside specific domain skills. Our data science teams vary depending on the type of research questions and objectives, as well as intended real-world impacts.

Some of our largest and most ambitious projects involve data science teams that involve the following experts.

1. **Domain experts**
    - Principal Investigators - often from Turing programme teams and partner organisations
    - Postdoctoral researchers or research fellows
    - Research associates or early career researchers
2. **Cross-cutting research infrastructure specialists**
    - Research Data Scientists/Software Engineers (Research Engineering Group/REG)
    - Research Community Managers (RCMs)
    - Research Application Managers (RAMs)
    - Data Wranglers
    - AI/Applied Ethicists
    - Participatory researchers (citizen science and public engagement)
    - Project/Programme Managers
    - Partnership development manager

Each of these specialists may not be involved for the whole length of the project or solely working on one project, but they input into goals and objectives, collaborate in the right area of work and deliver on specific pieces of work, such as through specific work packages.

2.3 Example of skills framework for research community managers

The Research Community Managers (RCMs) at the Turing work with different AI/data science projects and their communities, ensuring that collaborations amongst diverse stakeholders are effective and impactful. They involve all community members in defining challenges, developing milestones, integrating research best practices and co-creating solutions that they can all benefit from. As a team, they connect data, software and expertise from projects across the institute, while engaging researchers meaningfully with each other's work. As data science professionals, RCMs themselves bring specific data science skills that are defined in the AfDSP standards.

Through our work at the Turing, and further research carried out under this project, we have identified two **core skill sets**, communication and engagement, where Research Community Management (RCM) roles provide specific expertise and leadership on, and three **peripheral skill sets** (strategic development, technical expertise and management skills) that are required in their role but other research roles take leadership on. Specifically, RCMs contribute to matching the overall *skills make-up of the team* by either 1) directly bringing expertise, or 2) identifying and involving other collaborators in fulfilling other skill areas.

The two core skill sets could be mapped to AfDSP standards' **skill area C**: problem definition and communication with stakeholders, and **skill area E**: evaluation and reflection. Based on the AfDSP evidential requirements related to the specific skill areas, we have further expanded these skill sets to create the RCM Skills Framework that provide a greater definition for professional requirements of the RCM roles in data science. The detailed RCM Skills Framework has been shared online on Zenodo under the CC-BY 4.0 License, doi: 10.5281/zenodo.8337627 (Sharan & Karoune, 2023).

Appendix Table 1: Example of RCM skill sets mapped to the AfDSP standards and further expanded to create an RCM Skills Framework, part of which is shown for a core skill set "engagement"

Table 1.1 Research Community Management roles have five overarching skill sets. Communication and Engagement are two core skill areas. The remaining three skill areas (strategic development, technical expertise and project/programme management) partly intersect with other roles within the team such as data stewards, research engineers and project managers, allowing RCMs to facilitate knowledge transfer across those specialist groups. **1.2** AfDSP Standards Table with Skill areas **A:** data privacy & stewardship; **B:** definition, acquisition, engineering, architecture, storage & curation; and **C:** Problem definition & communications with stakeholders; **D:** problem-solving, analysis, statistical modelling, visualisation; **E:** evaluation & reflections. Each skill area is mapped to evidential requirements in a numbered list. The RCM core skill sets together map to the AfDSP Skill Areas C and E. **1.3** We specifically highlight 'Engagement' as one of the core skill sets for RCMs. Based on the evidential requirements provided by AfDSP, we produce a comprehensive and expanded definitions for RCM skills such as community onboarding, reproducibility practices and open collaboration needed to carry out the professional responsibilities.

Appendix 1: Example of RCM skill sets mapped to the AfDSP standards and further expanded to create an RCM Skills Framework, part of which is shown for a core skill set "engagement"

Table 1.1: RCM Skill Sets

Core (blue cells) and peripheral skill sets
Communications
Engagement* (community and stakeholder)
Strategic development
Technical Expertise
Management Skills

Table 1.2: Alliance for Data Science Professionals Standards

Skill Areas	Evidential Requirements
Skill Area A. Data Privacy and Stewardship	1. Ensuring the protection of personal and sensitive data. 2. Managing sensitive data. 3. Data stewardship and standards.
Skill Area B. Definition, acquisition, engineering, architecture, storage and curation.	1. Data collection and management. 2. Data engineering. 3. Deployment.
Skill Area C. Problem definition and communication with stakeholders	1. Problem definition. 2. Relationship management.*
Skill Area D. Problem solving, analysis, statistical modelling, visualisation.	1. Identifying and applying technical solutions and project management approaches. 2. Data preparation and feature modelling. 3. Data Analysis and model building.
Skill Area E. Evaluation and Reflection	1. Project evaluation. 2. Ethical behaviour.* 3. Sustainability and best practices.* 4. Reflective practice and ongoing development.*

Table 1.3: 'Engagement' skill area expanded in the RCM Skills Framework

Engagement*
Creating and sharing engagement process and technical documents (openly via The Turing Way)
Scientific engagement with researchers and (volunteer) onboarding
Roadmap for project and community engagement
Workshops for knowledge share and training/skill building
Short term and long-term **engagement strategy** (stakeholder mapping and planning)
Mentoring opportunities (upskilling and onboarding community members on visible roles as well as connecting them with other projects at the Turing)
Community Policies and their alignment with organisational policies
Recognition and visibility of community members (rewards, incentive alignments and support)
Onboarding diverse community members and leaders (EDIA embedding in research)
Community survey and reporting (community health, engagement, infrastructure)
Creating collaborative opportunities such as events and platforms for feedback and discussions
Open collaboration and ethical participatory approaches
Reproducibility skills (The Turing Way practices)

This is an image of the original editable version of the table available on our shared drive.