Reproducible research at the Turing

The Alan Turing Institute is the UK's national institute for data science, headquartered at the British Library in the heart of London's Knowledge Quarter. Our mission is to make great leaps in data science research in order to change the world for the better. Five founding universities – Cambridge, Edinburgh, Oxford, UCL and Warwick – along with the UK Engineering and Physical Sciences Research Council created The Alan Turing Institute in 2015. Our researchers collaborate across disciplines to generate impact, both through theoretical development and application to real-world problems. We are fuelled by the desire to innovate and add value.

Cross-disciplinary collaborations are hard. Implicit knowledge and scientific output standards vary dramatically between computer science, mathematics, engineering, economics, psychology and social science. Projects funded by fellowships such as the rOpenSci program can fill these gaps. Community management is currently undervalued and difficult to fund. The *reproducible research liaison officer* funded by this fellowship (see budget for details) will allow me to continue to build momentum towards an academic reward system that values reproducible research.

My colleagues in the research software engineering team and I **seek to position The Alan Turing Institute as a world leader for reproducible research** by embedding a culture of "reproducible by default" at the institute and providing training and tools to make reproducible research "too easy not to do".

Reproducibility ensures that the wider academic, industry and governmental communities we operate in can easily validate, replicate and build on the work we produce. This will help us extend the Turing's impact beyond the immediate outputs of the institute itself, providing a "force multiplier" effect on the progress of Data Science in the UK and beyond. The policy officer of The Alan Turing Institute, Helena Quinn, will make sure that our outputs are jargon free and fit for use by civil servants and national funding agencies.

We want the research outputs of the Turing to be seen as coming with a "quality seal of approval". We want to lead by example to demonstrate how institutions can ensure that their work is reliable and trustworthy. We will build the Turing Reproducibility Champions program to raise the profile of excellent reproducible and reusable research.

The benefit to the global community will be highly curated, openly licensed (for remixing and reuse) documentation to support universities and research funders adopt, implement *and enforce* reproducible research policies. The Alan Turing Institute is ideally placed to spread the impact of our reproducible research programme to the wider community, with Knowledge Quarter partners including the British Library, The Wellcome Trust, UCL, The Francis Crick Institute and Figshare. The *reproducible research liaison officer* will organise open events for the wider community to share our learning and promote the resources developed by this program.

Key focus: Accessible Data

We will ensure that Turing researchers who have spent a long time collecting and curating datasets are able to make them available to others as easily as possible. We will build and provide clear guidance on ethical constraints and best practises regarding persistent data archiving. We will make it easy for data curators to receive citations for their hard work, for example by assigning digital online identifiers, or by publishing data description papers. We

will work with the Turing communications team to ensure that the public datasets are promoted widely to members of the global data science community.

We will harness the expertise of members of the Turing working on encrypted data analyses to translate their work into guidance for data scientists publishing on sensitive or identifiable data. Not all data can be shared openly and it is imperative that our documentation provides options for all cases while also being clear, standardised and easy to follow. We will encourage the adoption of data management plans for all projects funded by the Turing. If they are created and followed from the beginning of the project, data management plans not only ensure that the database is an excellent resource for future research, they also ensure that the research team are more efficient in their analysis process.

Key focus: Open Source Code

To reproduce published findings, it is necessary to have the exact steps the researchers undertook in generating their results. We will support all members of the Alan Turing Institute in sharing the code they write with readers of their published results. As a minimum, we would like to ensure that publications from the Turing can be verified by any independent researcher.

Beyond reproducible research, sharing analysis code allows members of the Turing and the wider community to "stand on the shoulders of giants" and benefit from the expertise of others. We will ensure that Turing members are working efficiently through two key avenues. Firstly, as with accessible data, we will help to provide documentation and dissemination channels to clearly communicate what effort has been undertaken to date.

We will also support members of the Turing by harnessing the power and expertise of the research software engineering team. They will work with researchers to develop code to a higher standard than is currently required of academic researchers. Members of the Turing and the data science community at large will benefit from reliable, documented, tested, version controlled analysis code. Funding for this aspect of the project has already been acquired through the Alan Turing Institute seed funding programme (see budget below for details).

Key focus: Community building and skills training

By design, the Turing comprises researchers of diverse backgrounds and expertise. Two often cited reasons for not ensuring that research is reproducible are 1) the lack of training provided in the software or protocols required for others to verify the work and 2) having “no time” to acquire the skills and knowledge.

We will provide getting-started resources and trainings for version control with git and GitHub, code testing, intro to coding in Python, R and Unix, along with "demystify" sessions around software engineering. The rOpenSci funded reproducible liaison officer will organise weekly office hours to provide in-person support along with regular newsletters to disseminate external resources. She will be responsible for facilitating communication between the Turing researchers, graduate students, business team and research software engineers, and to inspire community members to take the next step in ensuring that their work is reproducible (through the Turing Reproducibility Champions program, described below). We will continue to share our experiences with the Mozilla Open Science community, and we look forward to contributing to the rOpenSci community.

Project plan

Requirement Gathering: The most important step in our plan is to see what already exists, what needs to be curated and if there are already tools that would do the job. We will focus on collecting information around:

- What tools already exist?
- What are researchers in the Turing using?
- What resources would best serve Turing researchers?
- Can we contribute to existing projects to make them even better?
- Do we need to create any new software?

Turing Reproducibility Champions: The Turing reproducibility champions program will highlight 4-5 examples of research currently being conducted at the Turing. Specifically, we will ask researchers who are already working reproducibly to work with us to develop case studies. For example, the work that Tomas Petricek and May Yong recently showcased on Accounting for Democracy. We will work with project leaders to ensure others (both within the Turing and externally) are able to access the research code and data, and work with the Turing communications team to promote these examples widely. The reproducibility champions will inspire their colleagues to adopt more reproducible practises and will allow us to refine a protocol for best supporting researchers as they publish their work.

Budget

Requested fellowship funds:

- Reproducible research liaison officer: £20,000 ($25,000 USD) salary, £8,000 ($10,000 USD) onboarding costs. **Total: £28,000 ($36,000 USD).**

We will recruit and hire for either 6 months full time, or 12 months part time, a reproducible research liaison officer. She is likely to have a doctoral degree in a computational field to ensure that she is able to easily communicate with Turing researchers about their needs and current working practises. It is necessary that she has very strong communication skills and will be responsible for supporting the documentation of data and software, along with writing lay summaries to promote reproducible research from Turing members. The reproducible research liaison officer will work entirely openly from the start and seek to build connections between the Turing and others working towards a similar goal. She will undertake the requirement gathering exercises and manage the Turing Reproducibility Champions program, and will report to rOpenSci fellowship lead investigator Dr Kirstie Whitaker.

Funds already acquired:

- Research software engineering time: 2 days per week for 12 months, £36,400 ($47,000).
- Lead investigator (Dr Kirstie Whitaker) salary through 3 year research fellowship to Turing Institute from EPSRC grant (EP/N510129).

Timeline

- **October – December**: Recruit reproducible research liaison officer. Survey Turing researchers to assess their needs & what resources are already available. Identify and build relationships with external communities working towards reproducible research. Recruit 4-5 reproducibility champions.

- **January – April**: Deliver trainings and support new members of the Turing research community as they begin their projects. Develop common protocols and documentation based on feedback of reproducibility champions.

- **May – September**: Write white paper on our experiences incentivising reproducible research at an institutional level. Working title: *How to do research in the 21st Century: The Turing Way*. Continue to support reproducibility champions and students. Organise external events to share our learning with the wider community.