Producing reproducible code: A DAA DS framework

A note - 04JULY2022

This document is a shortened version of an attempt toward synthesising many disparate sources on doing reproducible data scientific work. Why this has been necessary is covered in the 'Introduction'. To summarise my thoughts on producing reproducible work: shared principles that have been battle-tested by computer science, software engineering, etc, should form the bedrock of reproducible work in Data Science (version control, shared writing styles, reproducible environments) and additional 'standards' from other fields should also be included (package synthesis, assumption listing, minute keeping).

This document contains the background for the main thrust of the work, producing a framework and procedure for doing reproducible work in the WWL Data Science team. The full document, including the framework, is located in: Data Science -> Projects -> Current -> Reproducible -> Test_files_Scripts -> Github_collection.

Introduction

Data science is a burgeoning and yet still nascent field. Its popularity has created an influx of DS' from an incredible array of disparate backgrounds; economics to biochemistry, and all fields in between! In fact, it's quite likely that DS as a profession contains more generalist skilled individuals than any other scientific career, not least because of the variety of skills required to be proficient, but also due to the fields novelty - I think it reasonable to posit that most, current, established DS were either joining from a second career and/or received no formal pedagogy/taught themselves autodidacticly.

A generalist background clearly has merits, especially in the arena of data scientific work where a wide ranging skill set is absolutely necessary. However, the lack of pedagogical standardisation has caused problems with readability, reproducibility, and replicability. Interestingly, lacking pedagogical standardisation has not obviously inhibited the "output" of DS work from improving. Indeed, the widely recognised tools of Data Scientists' (statistics, AI, coding, problem solving) have enjoyed rising competency and improvements, likely through a convergence in the lessons taught by both online and degree courses, and replicating pedagogy related to these skills from other disciplines.

However, how can we actually know that improvements in Data Scientists' output have been accrued? A definition of improvements would be required, and is beyond this the scope of this brief piece, but I think it's safe to assume that most definitions would include one or more of the following:

- more output (perhaps in a shorter timeframe),
- better performing models,
- more appropriate figures and statistical analyses.

But none of these determine whether the work that went into producing the output (data collection, datasets, the code), were appropriate for the question posed.

This is the crux of the matter: **the "correctness" of the output cannot be objectively assessed without transparency of the code.**

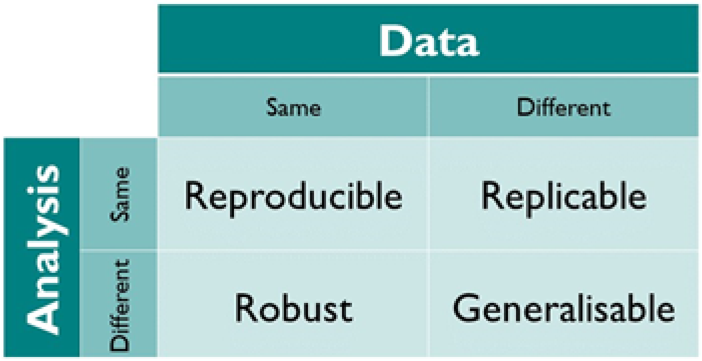
Formal courses on Data Science touch on reproducibility, however serious highlighting of the importance to practical Data Science appears to have been left to the individual to self-teach, or that they learn on-the-job. Here, I argue for reproducibility and describe a framework toward producing robust, reliable and reproducible work in the DAA Data Science team.

Definitions: Reproducible and Replicable

Definitions of key terminology in this space is a real minefield, not least because the field you work in may determine the definition. For a full review on this topic, do read this peer-reviewed paper by Hans Plesser (2018).

The definitions I will be using are taken from the Turing Way:

- *Reproducible*: A result is reproducible when the same analysis steps performed on the same dataset consistently produces the same answer.
- *Replicable*: A result is replicable when the same analysis performed on different datasets produces qualitatively similar answers.
- *Robust*: A result is robust when the same dataset is subjected to different analysis workflows to answer the same research question (for example one pipeline written in R and another written in Python) and a qualitatively similar or identical answer is produced. Robust results show that the work is not dependent on the specificities of the programming language chosen to perform the analysis.
- *Generalisable*: Combining replicable and robust findings allow us to form generalisable results. Note that running an analysis on a different software implementation and with a different dataset does not provide generalised results. There will be many more steps to know how well the work applies to all the different aspects of the research question. Generalisation is an important step towards understanding that the result is not dependent on a particular dataset nor a particular version of the analysis pipeline.



		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Figure 1: *How the Turing Way defines reproducible research*

NHS data policy: highlighting reproducibility

Within the NHS, having dedicated Data Scientists' (or even data literate Data Teams') is a relatively nascent phenomenon. There is a tremendous amount of work to be done to bring NHS data standards up to that of private organisations. Encouragingly though, there's definitely evidence that those with influence in the UK are advocating for the potential that Data Teams' across the NHS can unlock.

The UK government has recently published NHS-centred documents that emphasise the importance, and give detailed guidance, on data best practices. Below are links pointing to these documents, which I'll be keeping up-to-date when other relevant documents become available. It's definitely worth reading these; they're relevant to the work you do, but more importantly they will align you with the wider NHS data community.

- Better, broader, safer: using health data for research and analysis - GOV.UK, April 2022
- Data saves lives: reshaping health and social care with data - GOV.UK, June 2022

The DAA

Our department have published their three year plan, Caring For Our Data, 2022 - 2025, which includes a section on Data Science. It is an encouraging document that promotes a bright vision for the Trust - you should read it. Unfortunately, reproducible working is not explicitly mentioned, so a revision would include greater emphasis on its necessity. However, there are certainly good examples of how working reproducibly will aid with some of the plans objectives:

Working ethically

"...A key consideration within AI is ethics. Any Healthcare AI practice should put all patients and people first. This means that it is our responsibility to, for example, minimise representation bias within any of our algorithms." - Page 15.

- There are high-profile examples of published research that has been retracted due to code writing that did not adequately account for representation bias in their sample. Results can easily be misinterpreted when a sample set does not accurately represent your population.
- Writing code reproducibly allows others to review your work, gives visibility to any assumptions you've made, and keeps you accountable.
- Deep dives:
  - Towards a Code of Ethics for AI - Paula Boddington, 2017

Transparent usage of data

"Improving the transparency of how data is (sic) used and shared" - Page 16

- This dovetails nicely with working ethically. Transparency with our code keeps us accountable, makes code review possible, and therefore makes working unethically less likely to occur.
- Further than this, transparency of code forms the foundation of reproducible work. Full review of the methodology allows constructive critique and ensures appropriate methods and models were used.

- Transparency also fosters learning. Personally, I have primarily learned to code through forums, then reading scripts from others I work with, and more recently using GitHub to pull others repositories and explore their code. Transparent code is simply incredible for self-improvement.

Data Quality

"Our people are confident their decisions are being made on robust data" - Page 11

- If you've worked with data in the NHS then it shouldn't shock you to hear me say how terrible the quality is: hand-typed fields, columns with different names but the same data, columns with impossible to understand values, impossible tails to distributions.. it goes on. A lot of our time is taken up with getting the data cleaned to an appropriate state for initial exploration.
- With so much pre-processing to do, it's unlikely you'll do a perfect job first time round. The issue then becomes whether you continue through to the "end" of your analysis without including/cleaning/excluding/factoring/etc - your output may well be correct for the data you used, but incorrect because the data you used was not appropriate. Working reproducibly will certainly help with this.
- In order to understand what you've done, writing code in a readable manner becomes a requirement. I'll come on to writing styles later on, but suffice to say that a facet of reproducible work is readable work.