Reproducible Analytical Pipeline

Christophe Bontemps - SIAP[1]

29 Jun 2022

[1] *This document uses teaching materials developped by Joseph Crispell & Nathan Begbie (ONS-UK)*

What is a Analytical Pipeline? The ideal scenario

What is a Analytical Pipeline? The reality...

What are the issues?

- Lots of manual steps
- Hard to reproduce
- Mistakes are easily made and hard to track
- The steps aren't recorded
- Using multiple independent tools
- How do we keep track of which file versions people have?

What is a Reproducible Analytical Pipeline (RAP)?

- It is easily repeatable
- It is easily extendable
- It is automated
- It minimises mistakes
- It is fast
- It builds trust

Image taken from: The Turing Way book

What does a RAP look like?

What does a RAP look like?

All the functionality in one place

What are the benefits?

- Easy for others to use
- Others can change and adapt
- All steps are recorded
  - Including whilst it is built
- Automated and fast
- Open and promotes trust

Image taken from: The Turing Way book

Is it Data Science?

No, it is a building block that's essential for data science!

What do we need?

- Open-source tools
- Version control with `git`
- To consider reproducibility
- Time to learn

Why open-source instead of proprietary?

Open source tools are:

- Used by millions - huge supportive online community
- Flexible to all data sources
- Free for anyone to use - it is easier to share
- Flexible to all output types

Image taken from: The Turing Way book

What is version control?

Tracking the three **W**s: **W**ho made **W**hich change and **W**hy?

For example, take a look at this presentation

Why use version control?

- One place to store your code
- You and collaborators are free to write and develop locally
- Complete documented history of all changes made
- Easy to share
- Your future self will thank you!

Which version control should I use?

Tool	Cost	Where is master?	Advantage
Github	Free	Online	Huge user base
Gitlab	Free	Local or Online	Continuous integration
Bitbucket	$$$	Cloud	Cloud security
Azure	$$$	Cloud	Cloud security

What do we mean by reproducible?

We want to look back and be able to repeat our work easily and quickly.

What are the benefits?

- Helps build trust
- Not reliant on single individual
- Can be adapted and re-used

Image taken from: The Turing Way book

Making a RAP is difficult

Before we start, here are a few things to consider:

- IT infrastructure available
- Data privacy - where and how am I storing my data?
- Expertise - what training do I need?
- Legacy systems - what are the barriers to transitioning?

But it is worth it!

Image taken from: The Turing Way book

And we don't have to do it all at once

The building blocks of a RAP:

- Version control
- Using open-source tools
- Create reproducible code

... are useful in their own right, each will improve the auditability, speed and quality of your work.

RAP in practice

RAP has been successfully rolled out in 10s of projects across the UK government.

It is now part of the best practice documentation.

Work continues across the government to roll out RAP to more projects.

Packages to help us with RAP

`govdown` - R package to recreate our UK government website template.

`drake` - R package to streamline reproducible pipelines in R.

`snapcraft` - python package to solve those dependency issues!

Mentorship at the Data Science Campus

We aim to raise the aid community's capacity for using Data Science to meet the Sustainable Development Goals.

We'll link up Data Scientist mentors from the campus with mentees. Together we'll set plan out a project, describing:

- What the project is, its impact and scope
- The resources, expertise and time will we'll need
- And, how we'll work together during the project

Mentorship at the Data Science Campus

As we plan out our project, we'll be considering those sometimes difficult aspects of RAP:

- IT infrastructure available
- Data privacy - where and how am I storing my data?
- Expertise - what training do I need?
- Legacy systems - what are the barriers to transitioning?

Useful resources

- The UK government RAP website.
- A free RAP course to teach you all you need to know.
- How the Data Science Campus sets its coding standards.
- A new open-source book from the Alan Turing institute setting out how to do reproducible data science.
- The github page for this presentation and other materials is here.
- Get involved in the UK government RAP champion network here.

Citing *The Turing Way*

Many of the beautiful images used in this presentation were taken from *The Turing Way* book.

Full citation:

The Turing Way Community, Becky Arnold, Louise Bowler, Sarah Gibson, Patricia Herterich, Rosie Higman, . . . Kirstie Whitaker. (2019, March 25). The Turing Way: A Handbook for Reproducible Data Science (Version v0.0.4). Zenodo. http://doi.org/10.5281/zenodo.3233986