Proctor Foundation Data Science Handbook

Contributors (many from UC Berkeley in addition to Proctor): Ben Arnold, Jade Benjamin-Chung, Kunal Mishra, Stephanie Djajadi, Nolan Pokpongkiat

2019-10-01

Contents

Welcome!

Welcome to the Francis I. Proctor Foundation at the University of California, San Francisco (https://proctor.ucsf.edu)!

This handbook summarizes some best practices for data science, drawing from our experience at the Francis I. Proctor Foundation and from that of our close colleagues in the Division of Epidemiology and Biostatistics at the University of California, Berkeley (where Prof. Ben Arnold worked for many years before joining Proctor).

We do not intend this handbook to be a comprehensive guide to data science. Instead, it focuses more on practical, "how-to" guidance for conducting data science within epidemiologic research studies. Although many of the ideas of environment-independent, the examples draw from the R programming language. For an excellent overview of data science in R, see the book R for Data Science.

Much of the material in this handbook evolved from a version of Dr. Jade Benjamin-Chung's lab manual at the University of California, Berkeley. In addition to the Proctor team, many contributors include current and former students from UC Berkeley.

The last two chapters of the handbook cover our communication strategy and code of conduct for team members who work with Prof. Ben Arnold, who leads Proctor's Data Coordinating Center. They summarize key pieces of a functional data science team. Although the last two chapters might be of interest to a broader circle, *they are mostly relevant for people working directly with Ben.* Just because they are at the end does not make them less important.

It is a living document that we strive to update regularly. If you would like to contribute, please write Ben (ben.arnold@ucsf.edu) and/or submit a pull request.

The GitHub repository for this handbook is: https://github.com/proctor-ucsf/dcc-handbook

Chapter 1

Introduction: Work Flow and Reproducible Analyses

Contributors: Ben Arnold

This handbook collates a number of tips to help organize the workflow of epidemiologic data analyses. There are probably a dozen good ways to organize a workflow for reproducible research. This document includes recommendations that arise from our own team's experience through numerous field trials and observational data analyses. The recommendations will not work for everybody or for all applications. But, they work well for most of us most of the time, else we wouldn't put in the time to share them.

Start with two organizing concepts:

- **Workflow**. Defined here as the process required to draw scientific inference from data collected in the field or lab. I.e., the process by which we take data, and then process it, share it internally, analyze it, and communicate results to the scientific community.

- **Reproducible research**. A fundamental characteristic of the scientific method is that study findings can be reproduced beyond the original investigators. Data analyses that contribute to scientific research should be described and organized in a way that they could be reproduced by an independent person or research group. A data analysis that is not reproducible violates a core principle of the scientific method.

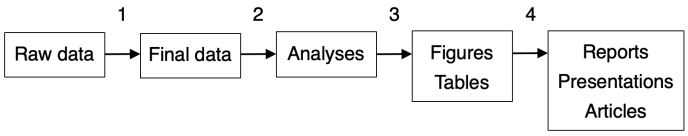

Figure 1.1: Overview of the four main steps in a typical data science workflow

1.1 Workflow

Broadly speaking, a typical scientific data science work flow involves four steps to transform raw data (e.g., from the field) into summaries that communicate results to the scientific community.

When starting a new project, the work flow tends to evolve gradually and by iteration. Data cleaning, data processing, exploratory analyses, back to data cleaning, and so forth. If the work takes place in an unstructured environment with no system to organize files and work flow, it rapidly devolves into into a disorganized mess; analyses become difficult or impossible to replicate and they are anything but scientific. Projects with short deadlines (e.g., proposals, conference abstract submissions, article revisions) are particularly vulnerable to this type of organizational entropy. Putting together a directory and workflow plan from the start helps keep files organized and prevent disorder. Modifications are inevitable – as long as the system is organized, modifications are usually no problem.

Depending on the project, each step involves a different amount of work. Step 1 is by far the most time consuming, and often the most error-prone. We devote an entire chapter to it below (Data cleaning and processing)

1.2 Reproducibility

As a guiding directive, this process should be reproducible. If you are not familiar with the concept of reproducible research, start with this manifesto (Munafo et al. 2017). For a deeper dive, we highly recommend the recent book from Christensen, Freese, and Miguel (2019). Although it is framed around social science, the ideas apply generally.

1.3 Automation

We recommend that the workflow be as automated as possible using a programming language. Automating the workflow in a programming language, and essentially reducing it to text, is advantageous because it makes the process

transparent, well documented, easily modified, and amenable to version control; these characteristics lend themselves to reproducible research.

At Proctor, we mostly use R. With the development of Rstudio, R Markdown and the tidyverse ecosystem (among others), the R language has evolved as much in the past few years as in all previous decades since its inception. This has made the conduct of automated, reproducible research considerably easier than it was 10 years ago.

If you have a step in your analysis workflow that involves point-and-click or copy/paste, then STOP, and ask yourself (and your team): *How can I automate this?*

Chapter 2

Workflows

Contributors: Ben Arnold

A data science work flow typically progresses through 4 steps that rarely evolve in a purely linear fashion, but in the end should flow in this direction:

Table 2.1: Workflow basics

Steps	Example activities	⇒ Inputs	⇒ Outputs
1	**Data cleaning and processing**		
.	make a plan for final datasets, fix data entry errors, create derived variables, plan for public replication files	untouched datasets	final datasets
2-3	**Analyses**		

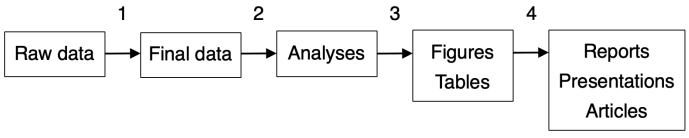

Figure 2.1: Overview of the four main steps in a typical data science workflow

Steps	Example activities	⇒ Inputs	⇒ Outputs
.	exploratory data analysis, study monitoring, summary statistics, statistical analyses, independent replication of analyses, make figures and tables	final datasets	saved results (.rds/.csv), tables (.html,.pdf), figures (.html/.png)
4 .	**Communication** results synthesis	saved results, figures, tables	monitoring reports, presentations, scientific articles

In many modern data science workflows, steps 2-4 can be accomplished in a single R notebook or Jupyter notebook: the statistical analysis, creation of figures and tables, and creation of reports.

However, it is still useful to think of the distinct stages in many cases. For example, a single statistical analysis might contribute to a DSMC report, a scientific conference presentation, and a scientific article. In this example, each piece of scientific communication would take the same input (stored analysis results as .csv/.rds) and then proceed along slightly different downstream workflows.

It would be more error prone to replicate the same statistical analysis in three parallel downstream work flows. This illustrates a key idea that holds more generally:

Key idea for workflows: Whenever possible, avoid repeating the same data processing or statistical a