



Practical Guide: Building your own Computational Pipeline for Social Scientists



Introduction

Computational social science (CSS) brings computational approaches to social science questions, has been shown to be a powerful tool.

The training explores a practical introduction to building computational workflows with *Nextflow* with the examples focusing on an application within social science. Feel free to follow along some of the practical material, the only prerequisite is that you register for an account with GitHub - no prior software installation is required!

Nextflow is workflow management software which enables the writing of scalable and reproducible scientific workflows. It can integrate various software package and environment management systems such as Docker, Singularity, and Conda. It allows for existing pipelines written in common scripting languages, such as R and Python, to be seamlessly coupled together. It simplifies the implementation and running of workflows on cloud or high-performance computing (HPC) infrastructure.

This training material was based off the training developed and maintained by [Seqera](#) and released under an open-source license ([CC BY-NC-ND](#)) for the benefit of the community. You are welcome to reuse these materials according to the terms of the license. The training includes material from the [carpentries incubator](#) and [The Turing Way](#).

Follow the link to Nextflow documentation below:



The course material for the half day course can be found in the [sgsss-workflow](#) repository, the material lives in a webpage that hosts the [course training](#). Material was prepared by [\(Student\) Eleni Omiridou](#), University of Glasgow. The course was first run in July 2025.

Pre-requisites

- Bring own device to follow practical component
- Set up an account with GitHub, follow link to: [GitHub page](#)
- Optionally, upgrade to a GitHub Education account. For more information follow link: [GitHub Education](#)

Set up - Online Learning Environment

Step 1: Open GitHub Codespaces

Click on the button below to launch a new codespace. If you are not already logged into GitHub in your browser then you may be prompted to do so. Then click the option to create a new codespace. You can opt for all the default options or choose to change ex. machine type.

 [Open in GitHub Codespaces](#)

Note: Steps 2 - 4 require you to copy and paste the code directly into your Terminal!
Press Enter to run the code each time.

Step 2: Organise Template Files

```
mkdir templates  
mv * templates/
```

Step 3: Read in Material Using Git

```
git clone --branch workflow-scripts --single-branch  
https://github.com/omiridoue/sgsss-workflow.git
```

Step 4: Navigate to the new Folder

```
cd sgsss-workflow/scripts
```

Practical Material

The training materials can be found in the following folders:

- [set-up](#) - Instructions to set up the material for the workshop.
- [template](#) - Template folder to adapt to your own workflow.
- [code](#) - All code scripts for the practical exercises.
- [workflow](#) - Full demo - ready-set-workflow !

Asynchronous Material

Link to [pre-recorded material](#)

Follow along with the companion practical material:

- [Workflow 00: Setup](#)
- [Workflow 01: Intro](#)
- [Workflow 02: Hello Nextflow](#)
- [Workflow 03: Parameters](#)
- [Workflow 04: Channels](#)
- [Workflow 05: Modules + Optional Topics](#)

Synthetic Data

The data involves low fidelity synthetic data. This means that the data is generated using functions, and only resembles real-world data in a very basic way. This type of synthetic can be useful for teaching and learning, or helping develop code. Generative AI was used to construct the data generating functions for this synthetic data. Please ensure any further use of this data includes this section. The code used to generate data can be shared upon request, drop a line to [\(PGR\) Eleni Omiridou](#).

The example application is from my research project working with school based secondary data on adolescent health behaviours. It is common for health and education researchers to work with multi-level data involving individual questionnaire data (level 1), nested in schools (level 2) and time (level 3).

AI Disclosure

This document was created with assistance from AI tools. The content has been reviewed and edited by the author. For more information on the extent and nature of AI usage, please contact the author.

Citations

Graeme R. Grimes, Evan Floden, Paolo Di Tommaso, Phil Ewels and Maxime Garcia Introduction to Workflows with Nextflow and nf-core. <https://github.com/carpentries-incubator/workflows-nextflow> 2021.

Ruth M. Ripley, Tom A. B. Snijders, Zsófia Boda, Andras Voros, and Paulina Preciado (2024). Manual for Siena version 4.0. R package version 1.4.13. <https://www.cran.r-project.org/web/packages/RSiena/>.

The Turing Way Community. (2022). The Turing Way: A handbook for reproducible, ethical and collaborative research (1.0.2). Zenodo. <http://doi.org/10.5281/zenodo.3233853>.

The Turing Way Community, & Scriberia. (2020, March 3). Illustrations from the Turing Way book dashes. Zenodo. <http://doi.org/10.5281/zenodo.3332807>.

The lesson material was adapted with permission from seqera labs [nextflow-tutorial](#)