Building your own Computational Pipeline for Social Scientists

Introduction

Computational social science (CSS) brings computational approaches to social science questions.

This student-led training explores a practical introduction to building computational workflows with Nextflow.

Nextflow is a 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.

The guided practical component of the training is implemented in the code spaces environment 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 any of these materials according to the terms of the license.

The training is based off carpentries incubator and The Turing Way. The course material page was designed using the workbench.

🗺 Objectives

- ☑ Understand the role of auxiliary tools in supporting best practice within the Open Research framework.
- ☑ Design/adapt a research project directory to own research.
- ☑ Understand the fundamental components of a workflow implemented in Nextflow, including channels, processes and operators.
- ☑ Launch a computational workflow using remote workspace, GitHub Codespaces.
- ☑ Curate own configurations and settings on a practice workflow to ensure portability, scalability, automatic resource management and re-entrancy.
- ☑ Discuss ways to implement own research directory.
- ☑ Explore links to the wider network for computational researchers across Scotland.

🗺 Intended learning outcomes

- ☑ Connecting with researchers in your field
- ☑ Ability to design and implement research projects from data to report
- ☑ Navigate reproducible, efficient, and communicable code via GitHub

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.

Course materials

The guided practical material for the half day course can be found in the sgsss-workflow scripts repository. The teaching and learning material lives in the course training page. Material was prepared by (PGR) Eleni Omiridou, University of Glasgow. The course was first run in July 2025.

Materials

This workshop makes use of an online teaching and learning space. This online environment involves using the GitHub Codespaces training environment:

[ⓞ Open in GitHub Codespaces]

The material and supporting resources can be found in the following folders:

- set-up - Instructions to set up the material for the workshop.
- code - Files for the guided practical material on building a workflow.
- template - Template folder to adapt to your own workflow.
- workflow - Full demo - ready-set-workflow !

Guided Practical Resources (Asynchronous)

Link to Pre-recorded Material:

Building your own Computational Pipeline for Social Scientists

Follow along the pre-recorded sessions and companion practical material:

Workflow 00: Setup Workflow 01: Intro Workflow 02: Hello Nextflow Workflow 03: Parameters Workflow 04: Channels Workflow 05: Modules + Optional Topics

Schedule

Section	Time	Topics Covered

Section	Time	Topics Covered
1. Introduction	00h 25m	What are the FAIR research principles? How do FAIR principles apply to software? How does folder organisation help me?
2. Hello Nextflow	00h 50m	What is Nextflow? Why should I use a workflow management system? What are the features of Nextflow? What are the main components of a Nextflow script? How do I run a Nextflow script?
Break	10m	
3. Parameters	01h 00m	How can I change the data a workflow uses? How can I parameterise a workflow? How can I add my parameters to a file?
4. Channels	01h 40m	How do I move data around in Nextflow? How do I handle different types of input, e.g. files and parameters? How can I use pattern matching to select input files?
Break	10m	
5. Modules	02h 00m	How do I run tasks/modules in Nextflow? How do I get data, files and values, into a module?
Finish Introductory Material	02h 20m	
6. Modules Part 2	optional	How do I get data, files, and values, out of processes? How do I handle grouped input and output? How can I control when a process is implemented? How do I control resources, such as number of CPUs and memory, available to processes? How do I save output/results from a process?
7. Workflow	optional	How do I connect channels and processes to create a workflow? How do I invoke a process inside a workflow?
8. Operators	optional	How do I perform operations, such as filtering, on channels? What are the different kinds of operations I can perform on channels? How do I combine operations? How can I use a CSV file to process data into a Channel?
9. Reporting	optional	How do I get information about my pipeline run? How can I see what commands I ran? How can I create a report from my run?

Section	Time	Topics Covered
10. Nextflow configuration	optional	How do I configure a Nextflow workflow? How do I assign different resources to different processes? How do I separate and provide configuration for different computational systems?
11. Auxiliary Tools	optional	When should I use a pre-built container? How can I customise a container? What is a remote codespace?
12. Resuming a Workflow	optional	How can I restart a Nextflow workflow after an error? How can I add new data to a workflow without starting from the beginning? Where can I find intermediate data and results?
13. Portability of Workflow	optional	How can I move my analysis to a computer cluster?

Acknowledgements

I am grateful to the Scottish Graduate School of Social Sciences (SGSSS) for funding and organising this course.

This lesson is a template lesson that uses The Carpentries Workbench. Built with sandpaper (0.16.12), pegboard (0.7.9), and varnish (1.0.5).

Further information

Please do not hesitate to get in contact if you have queries, suggestions or corrections regarding the material: (PGR) Eleni Omiridou

Any errors, omissions or inaccuracies are my own.

> All examples and descriptions are licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

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, Zsofia 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. DOI 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](#)