LOCATION AWARE SCIENTIFIC WORKFLOWS

Prior Knowledge Required
- Nextflow
- Wits Cluster
- AWS
- Linux

Initialization and Setup
- Configure Linux Virtual Machine (VM)
- Download Required Software
  - Bash
  - Java
  - Nextflow
  - Optional Extras
    - Docker
    - Singularity
    - AWS, etc
- Setup Agile Version Control (Git)
  - Story Mapping
  - SCRUM Boards
  - Project Conventions
  - Architectural Decision Records
- Import Basic Nextflow Model
  - Main File Setup
  - Configuration File Setup
    - Set different Profiles that sets the executor

Techniques for Execution Configuration (a)
- Input Data Analysis
  - Analysis size of input data
  - Analysis location(s) of input data
- Identify Network Constraints
- Identify Need for Multiprocessing/Clustering
- Identify Optimal Execution Location
  - Method to retrieve location from metadata
- Identify Optimal Executor
  - Based on location, select specific 'Profile' within Configuration File to match that of the input data

Nextflow Library Implementation (b)
- Code Nextflow Processes
  - Implement execution configuration steps from (a) as processes/methods that are callable
- Set Pipeline Parameters
  - Set input of pipeline to the input data set and the output to the specified Profile executor in the configuration file
  - Called within initial steps before any scripts are run, to adjust a config file using the processes coded above. This library should be called at the start of any workflow to optimize performance

Library Testing Using Workflows (c)
- Test Designed Library (b) on Simple Datasets
  - Use Nextflow Library on small data sets with little to no actual processes to ensure it works
- Develop/Modify New Workflow using Library on Complex Datasets
  - Add library to existing complex workflow where the library gets used as the predecessor to any script execution

Experimental Evaluation (d)
- Test Computational Efficiency
  - Execution time comparisons when using the library or not for both simple and complex workflows
- Create execution report
  - Either a basic command line result or preferably a PDF output with preferred execution location, selections, times, etc

Conclusion
- Project report
- Open Day poster
- Submit/Deliver a final presentation