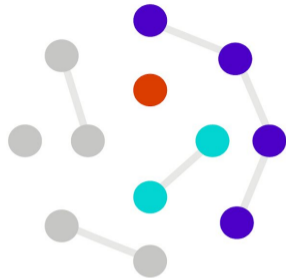


Leveraging Continuous Integration for enhanced eResearch on High Performance Computing Clusters



Ignatius Menzies - Software Engineer; Data Science Platform

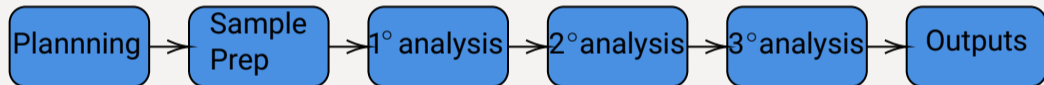
You can't spell Reproducibility without CI



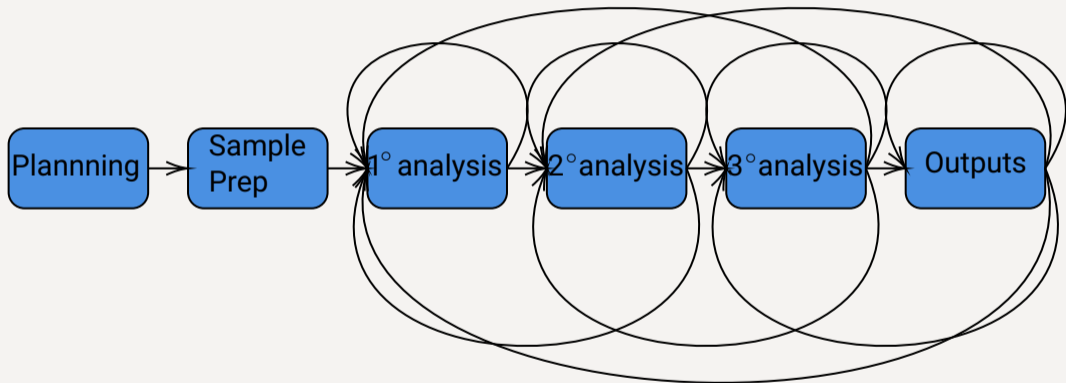
Research is messy iterative and non-linear



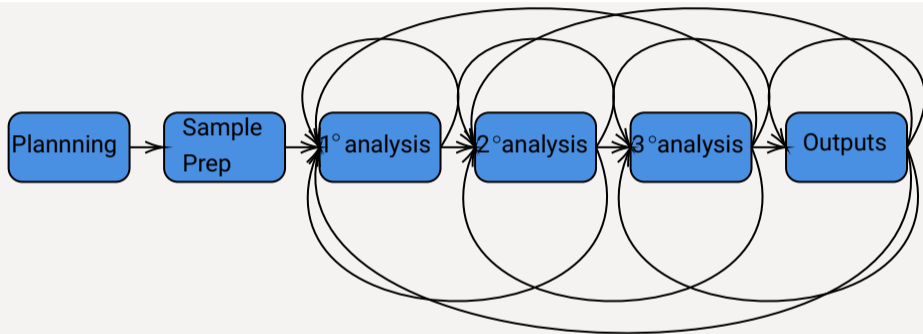
Research is messy iterative and non-linear



Research is messy iterative and non-linear

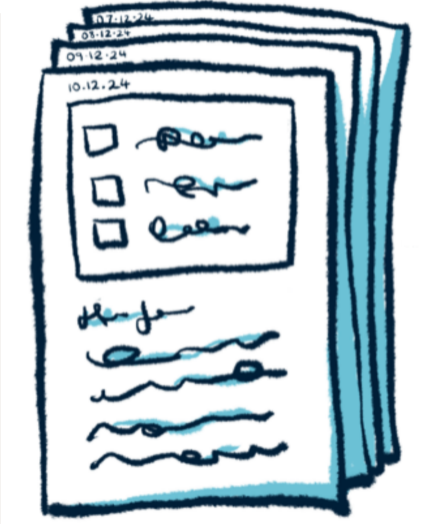


Research is messy iterative and non-linear

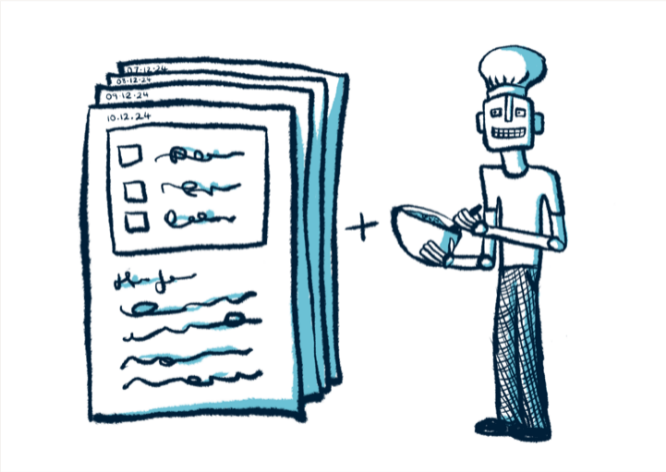


- High risk of errors
- Inefficient use of researcher time
- Not fun - can lead to shortcuts

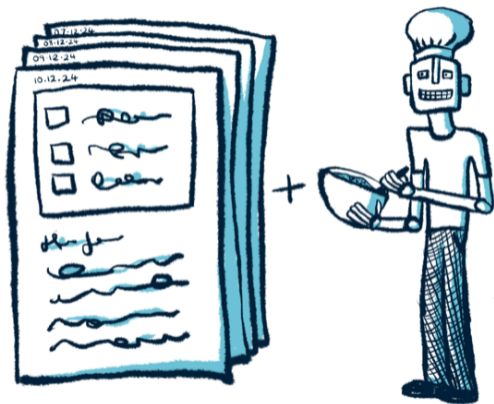
Reproducible research: a cooking analogy



Reproducible research: a cooking analogy



Reproducible research: a cooking analogy



Continuous Integration (CI) is like a robot sous-chef for your research.

CI in software development

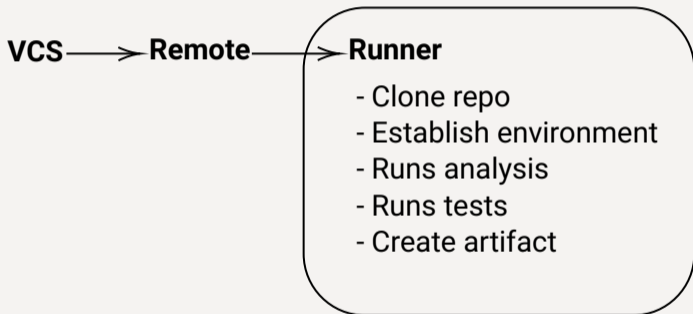


VCS

→ Remote

→ Runner

- Clone repo
- Establish environment
- Compiles software
- Runs tests
- Create artifact



Pass or fail



The screenshot shows the GitHub Actions interface for the repository 'Garvan-Data-Science-Platform / jointcalling-ref38'. The 'Actions' tab is selected, displaying a list of workflow runs. The workflow is named 'Fetch checkout and run on NCI'. Three runs are visible, each for the step 'Add happy to workflow steps' on the 'feature/adds-happy' branch, triggered by 'eriurn'.

Run ID	Status	Branch	Actor	Time
#33	In progress	feature/adds-happy	eriurn	5 days ago
#32	Failed	feature/adds-happy	eriurn	last week, 15m 17s
#31	Completed	feature/adds-happy	eriurn	last week, 15m 55s

Continuous Integration (CI) provides automated testing and rapid feedback.



But what about HPC?

Integrating CI and HPC is challenging:

- Data is typically too large to be moved to runner
- CI runner has insufficient compute capacity
- Interface to HPC is typically interactive
- Scheduler adds indirection

Existing approaches for integrating CI and HPC

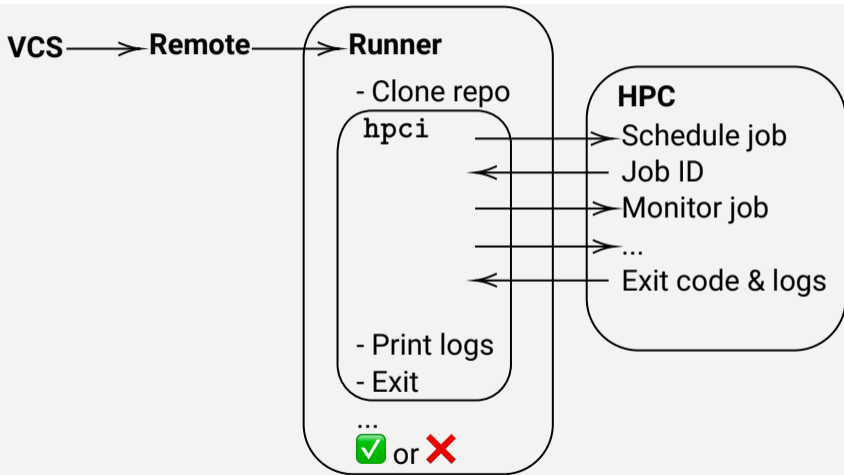


- Jacamar CI (<https://gitlab.com/ecp-ci/jacamar-ci>)
- Jenkins-CI with HPC
(<https://journals.sagepub.com/doi/full/10.1177/1087057116679993>)
- HPC-Rocket (<https://github.com/SvenMarcus/hpc-rocketg>)



HPC + CI = hpci

HPC + CI = hpci



How we're using `hpci` at Garvan



`hpci` helps provide automated validation for bioinformatics workflows.

How we're using hpci at Garvan



```
41m 12s
Schedule and monitor jobs
1070 resources: mem_mb=4000, mem_mb=3015, disk_mb=100, disk_mb=50, tmpdir=/jobs/125892412.gadi-pbs, project=gadi,
storage=gdata/a56, runtime=600, threads=1
1071
1072 [Tue Oct 1 11:55:20 2024]
1073 Finished job 0.
1074 19 of 19 steps (100%) done
1075 Complete log: .snakemake/log/2024-10-01T112628.396498.snakemake.log
1076 Observed results match expected results
1077
1078 =====
1079 Resource Usage on 2024-10-01 11:55:46:
1080 Job Id: 125892412.gadi-pbs
1081 Project:
1082 Exit Status: 0
1083 Service Units: 3.93
1084 NCPUs Requested: 4 NCPUs Used: 4
1085 CPU Time Used: 00:00:45
1086 Memory Requested: 8.0GB Memory Used: 623.89MB
1087 Walltime requested: 10:00:00 Walltime Used: 00:29:27
1088 JobFS requested: 100.0MB JobFS used: 46.0B
1089 =====
1090 Job Exit Status: 0
```

How we're using hpci at Garvan



```

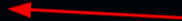
Schedule and monitor jobs 34m 23s
1109 ERROR: argument -command: invalid choice: test/results (choose from add, artifacts, cache, check-ignore,
'checkout', 'commit', 'completion', 'config', 'daemon', 'dag', 'data', 'pull', 'push', 'fetch', 'status', 'dataset',
'ds', 'destroy', 'diff', 'du', 'experiments', 'exp', 'freeze', 'unfreeze', 'gc', 'get', 'get-url', 'git-hook', 'import',
'import-db', 'test-db', 'import-url', 'init', 'install', 'list', 'ls', 'list-url', 'ls-url', 'metrics', 'move', 'mv',
'params', 'plots', 'queue', 'remote', 'remove', 'rm', 'repro', 'root', 'stage', 'studio', 'unprotect', 'update',
'version', 'doctor')
1110 Something has changed. Check diff for details
1111
1112 =====
1113                Resource Usage on 2024-09-19 18:21:24:
1114 Job Id:          125025341.gadi-pbs
1115 Project:         ██████████
1116 Exit Status:     1
1117 Service Units:  3.73
1118 NCPUs Requested: 4                NCPUs Used: 4
1119                                     CPU Time Used: 00:00:35
1120 Memory Requested: 8.0GB           Memory Used: 562.81MB
1121 Walltime requested: 10:00:00       Walltime Used: 00:28:00
1122 JobFS requested:  100.0MB          JobFS used: 103.0B
1123 =====
1124 Job Exit Status: 1
1125 Error: Process completed with exit code 1.

```

How we're using hpci at Garvan



```
Run happy 3m 17s
1.043349
54  SNP PASS 735 735 0 905 1 168 0 1.000000 0.998643
0.185635 0.999321 2.693467 2.620000 1.588028
1.661765
55 [PASSED] SNP PASS Recall ABOVE THRESHOLD (1.0 > 0.99)
56 [PASSED] SNP PASS Precision ABOVE THRESHOLD (0.998643 > 0.99)
57 [PASSED] INDEL PASS Recall ABOVE THRESHOLD (0.987013 > 0.95)
58 [PASSED] INDEL PASS Precision ABOVE THRESHOLD (1.0 > 0.95)
59
60 =====
61 Resource Usage on 2024-10-25 11:13:23:
62 Job Id: 127646506.gadi-pbs
63 Project:
64 Exit Status: 0
65 Service Units: 0.40
66 NCPUs Requested: 4 NCPUs Used: 4
67 CPU Time Used: 00:02:18
68 Memory Requested: 32.0GB Memory Used: 4.23GB
69 Walltime requested: 10:00:00 Walltime Used: 00:01:29
70 JobFS requested: 100.0MB JobFS used: 30.91KB
71 =====
72 Job Exit Status: 0
```



How we're using hpci at Garvan



```
Run happy 2m 43s
52  SNP  ALL      735    735     0     910     1    173     0     1.000000     0.998643
    0.190110     0.999321     2.693467     2.625498     1.588028
    1.645349
53  SNP  PASS     735    735     0     905     1    168     0     1.000000     0.998643
    0.185635     0.999321     2.693467     2.620000     1.588028
    1.661765
54
55  =====
56  Resource Usage on 2024-10-21 16:55:59:
57  Job Id:      127175534.gadi-pbs
58  Project:     ██████████
59  Exit Status: 2
60  Service Units: 0.40
61  NCPUs Requested: 4                NCPUs Used: 4
62                                     CPU Time Used: 00:02:06
63  Memory Requested: 32.0GB          Memory Used: 4.2GB
64  Walltime requested: 10:00:00      Walltime Used: 00:01:29
65  JobFS requested: 100.0MB          JobFS used: 3.16KB
66  =====
67  Job Exit Status: 2
68  Error: Process completed with exit code 2.
```

HPC + CI = hpci



Automated testing reduces the likelihood of errors in production workflow runs.

The tests run on the same hardware as the production workflows, ensuring consistency.

Frequent, small validation jobs are more efficient than errors during much larger production jobs.



Huge thanks to my Garvan collaborators:

- John Reeves and Victor Liu from Cancer Ecosystems Program; and,
- Joe Coptly and Eric Urng from DSP Production Bioinformatics.

Get in touch if you are interested in using `hpci`, or if you have any other ideas where `hpci` could be useful (*i.menzies@garvan.org.au*).

<https://github.com/Garvan-Data-Science-Platform/hpci/>