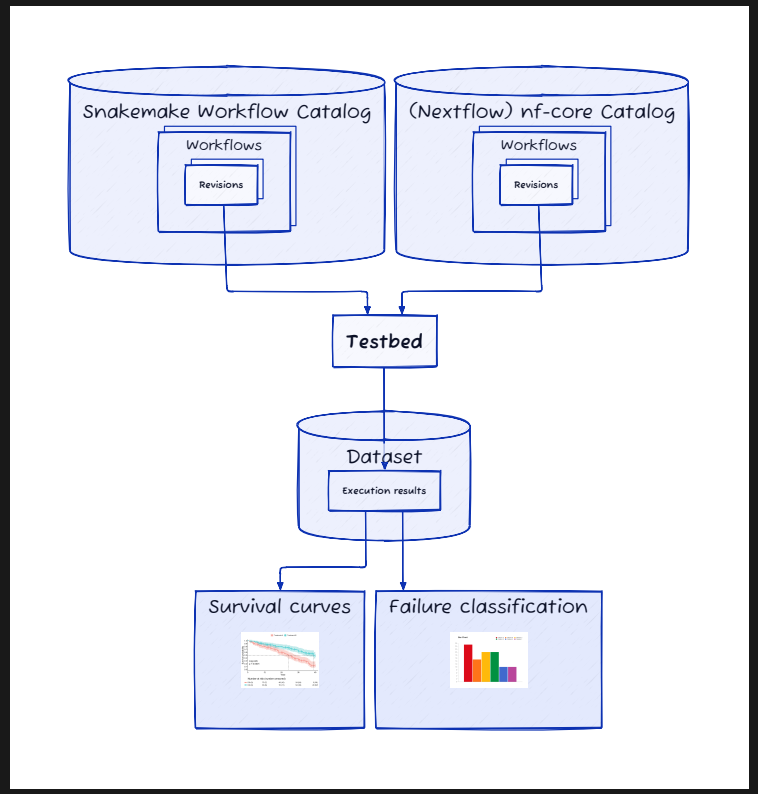
PACKAGING A TESTBED FOR WORKFLOWS

Samuel Grayson, Reed Milewicz, Dan Katz, Darko Marinov

Reproudcible HPC

PROJECT



Automatic Reproduction of Workflows in the Snakemake Workflow Catalog and nf-core Registries. **Grayson**, Marinov, Katz, Milewicz. ACM REP 2023

HOW TO MAKE REPRODUCIBLE?

- Docker
- Spack
- Parsl
- Non-root assumption

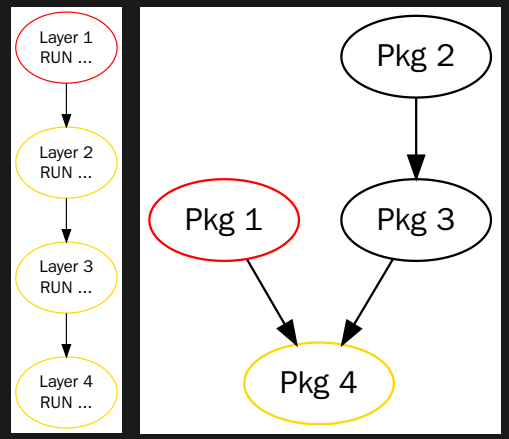
SAME VS SIMILAR ENVIRONMENT

- Both are valuable
- Same is more likely to work
- Same is useful if similar gives different answer

PKG MGR VS DOCKER

- Workflows may use Docker
  - Either Docker-in-Docker or sibling containers
- Where to store images?
  - DockerHub: 6 months of inactivity → deletion (2021)
  - Every service will face cost constraints eventually
- How to compose libraries in separate Docker containers?

- `docker build`
  - Second line is `apt-get update`
  - When pkgs non-pinned, 50% stale in 10 days (Shaffer et al. 2021)
  - Have to use pkg mgr anyway
  - Docker cache is linear



RESULTS

- Testbed defined by Spack environment
- Parsl works on many settings
- → Experiment usable on **Laptop, NCSA Delta, and Azure**.

Quantity	All	SWC	nf-core
# workflows	101	53	48
% of workflows with ≥ 1 non-crashing release	53%	23%	88%
# releases	584	333	251
% of releases with no crash	28%	11%	51%

Kind of crash	All	SWC	nf-core
Missing data/config input	32.2%	43.8%	16.7%
Conda environment unsolvable	10.8%	18.9%	0.0%
Unclassified reason	7.9%	12.0%	2.4%
Timeout reached	7.0%	5.7%	8.8%
Singularity error	6.0%	6.6%	5.2%
Other (workflow script)	5.7%	1.5%	11.2%
Other (workflow task)	1.2%	0.0%	2.8%
Network resource changed	0.7%	0.0%	1.6%
Missing software dependency	0.5%	0.9%	0.0%
No crash	28.1%	10.5%	51.4%
Total	100%	100%	100%