Oak Ridge National Laboratory

Design of a Scientific Data Analysis Support Platform

Nathan Martindale, Jason Hite, Scott Stewart, Mark Adams

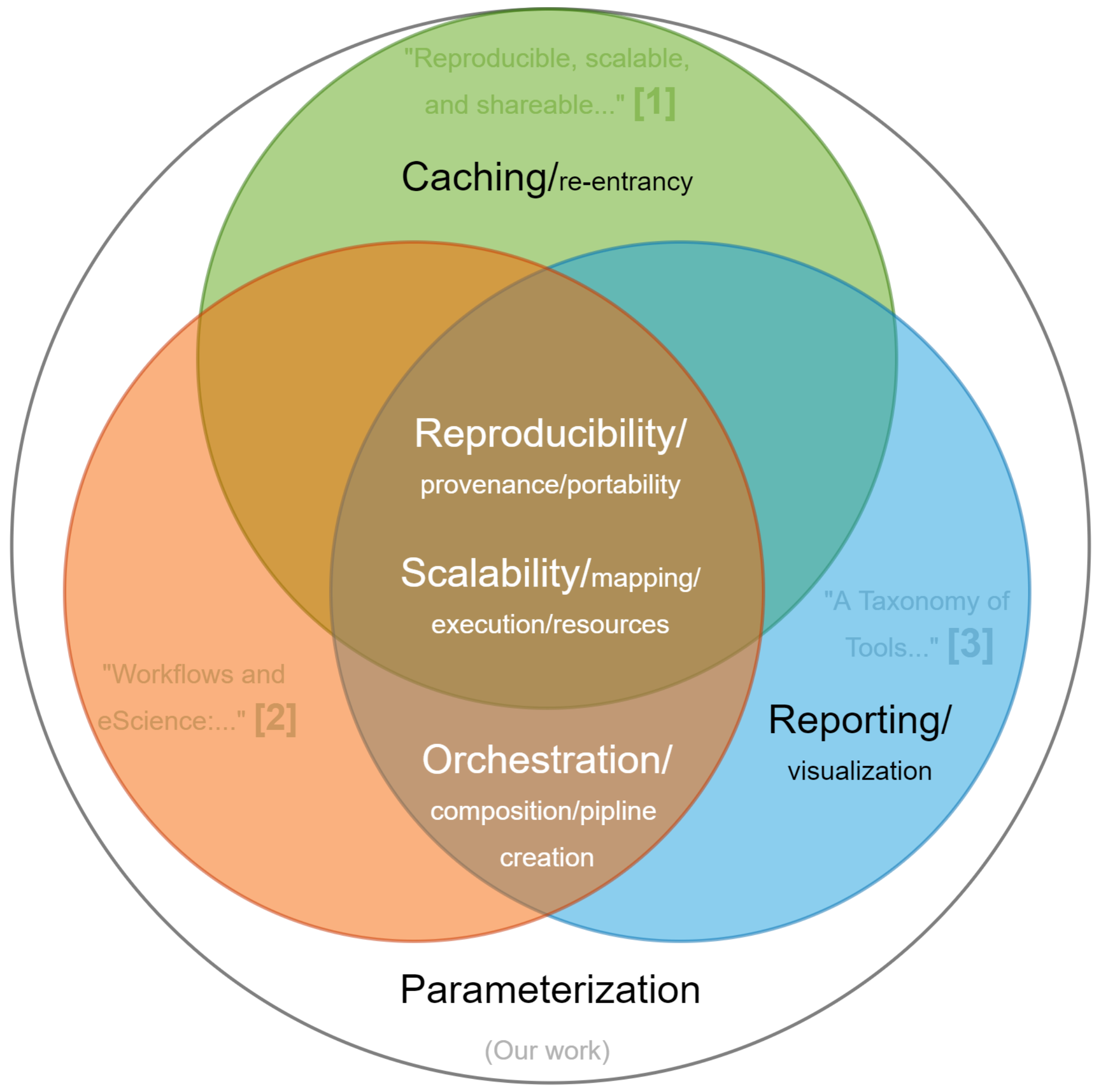
U.S. Department of Energy

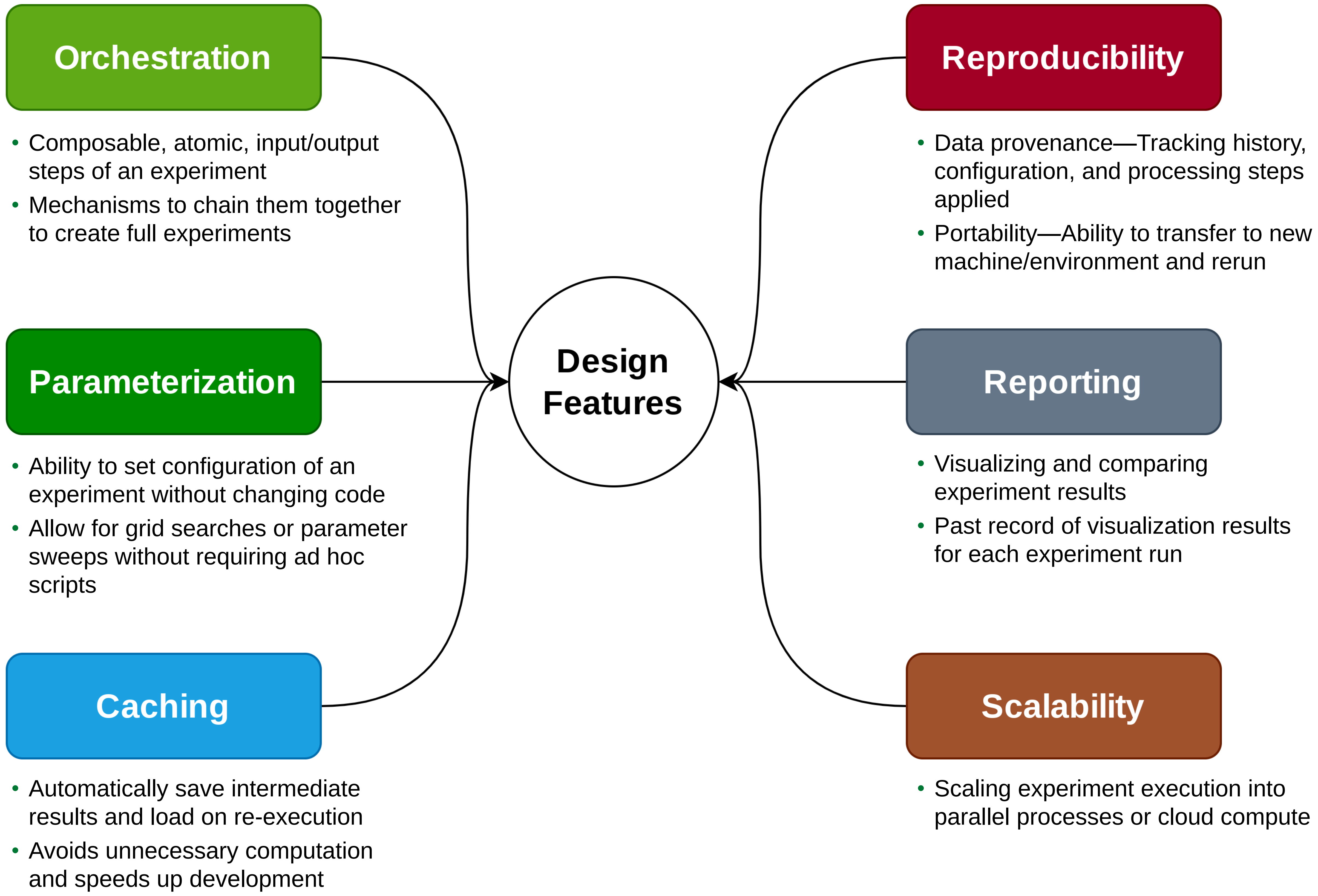
Reproducibility Crisis

- Complexity in software, environments, and lack of good software engineering principles in scientific domains has led to a reproducibility crisis in many computational research–based fields
- Resolving this requires infrastructure that supports good scientific and software practices
- Our work
  - surveys the literature for necessary design features and proposes a combination
  - analyzes design features of existing tools and proposes a new open-source tool, Curifactory

Concepts from Literature

- FAIRness principles
- Software engineering principles (testing, version control, agile)
- Compiled suggested design features from three other works on workflow/experiment management systems [1,2,3]


"Reproducible, scalable, and shareable..." [1]
Caching/re-entrancy
Reproducibility/ provenance/portability
Scalability/mapping/ execution/resources
"A Taxonomy of Tools..." [3]
"Workflows and eScience:..." [2]
Orchestration/ composition/pipline creation
Reporting/ visualization
Parameterization
(Our work)



Design Features

Orchestration

- Composable, atomic, input/output steps of an experiment
- Mechanisms to chain them together to create full experiments

Parameterization

- Ability to set configuration of an experiment without changing code
- Allow for grid searches or parameter sweeps without requiring ad hoc scripts

Caching

- Automatically save intermediate results and load on re-execution
- Avoids unnecessary computation and speeds up development

Reproducibility

- Data provenance—Tracking history, configuration, and processing steps applied
- Portability—Ability to transfer to new machine/environment and rerun

Reporting

- Visualizing and comparing experiment results
- Past record of visualization results for each experiment run

Scalability

- Scaling experiment execution into parallel processes or cloud compute

Existing Tooling

- DVC— Git-like interface for versioning datasets. Every compute input/output is a file, and caching/provenance is free
- MLFlow—STRONG MLOps tool, supporting entire data science life cycle, includes a powerful reporting dashboard and distributed computing
- Sacred—Allows parameterization directly in Python functions, user specification of observers for tracking metadata and artifacts
- Kedro—Ability to deploy to clusters, excellent web dashboard reporting and experiment visualization, ability to export entire project to docker container

	Orchestration	Parameterization	Caching	Provenance	Portability	Reporting	Scalability
DVC	▨	▨	🟩	▨	▨	▨	▨
MLFlow		▨	🟨	🟩	🟩	🟩	🟩
Sacred	▨	🟩	🟨	🟩	▨	▨	
Kedro	▨	▨	🟨	▨	🟩	🟩	🟩
Curifactory	▨	🟩	🟩	🟩	🟩	▨	▨

https://github.com/ORNL/curifactory

Curifactory

- **Orchestration**—Atomic level abstraction "stage," a function with defined inputs and outputs. Stages chained/composed into experiment scripts
- **Parameterization**—Parameters defined and instantiated in Python scripts, allowing inheritance, composition, looping
- **Caching**—Stages provide easy mechanism to store and reload every output to disk. Re-running same experiment will reload rather than recompute
- **Reproducibility**—Metadata tracked every run. Ability to create full store of an experiment containing all information, output report, and every intermediate artifact. Ability to build docker container for specific run
- **Reporting**—Every run outputs HTML report with metadata and user-definable graphics
- **Scalability**—Runs with multiple parameter sets can be run in multiple processes
- Published with BSD-3 clause license, available on PyPI and GitHub

References

[1] Laura Wratten, Andreas Wilm, and Jonathan Göke. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nature Methods 18(10):1161–1168, October 2021. doi:10.1038/s41592-021-01254-9.

[2] Ewa Deelman, Dennis Gannon, Matthew Shields, and Ian Taylor. Workflows and e-Science: An overview of workflow system features and capabilities. Future Generation Computer Systems, 25:524–540, May 2009. doi:10.1016/j.future.2008.06.012.

[3] Luigi Quaranta, Fabio Calefato, and Filippo Lanubile. A Taxonomy of Tools for Reproducible Machine Learning Experiments.

Acknowledgments

The authors would like to acknowledge the US Department of Energy, National Nuclear Security Administration's Office of Defense Nuclear Nonproliferation Research and Development (NA-22) for supporting this work