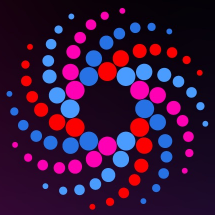
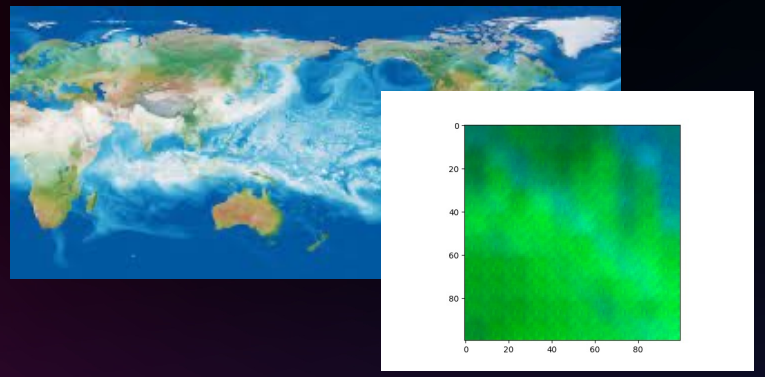


SC22
Dallas, TX | hpc accelerates.

RECUP: A (meta)data framework for reproducing hybrid workflows with FAIR

Line C. Pouchard, Brookhaven National Laboratory
Tanzima Z. Islam, Texas State University
Bogdan Nicolae, Argonne National Laboratory
Robert Ross, Argonne National Laboratory

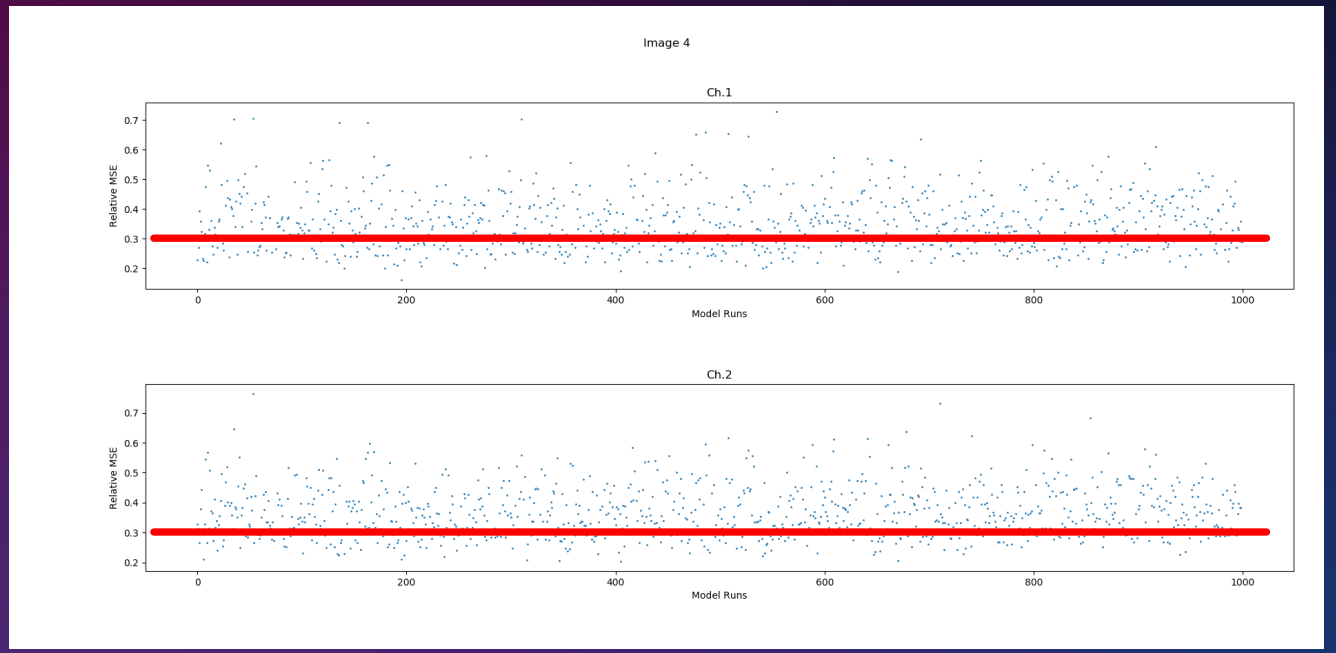
Reproducibility at scale: what kind?



Performance reproducibility: minimal run-to-run variation across multiple runs of the same application using a consistent set of configurations

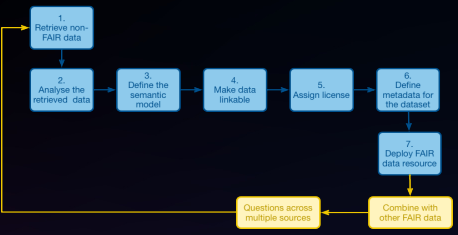
Result reproducibility: the statistical reproducibility of results within certain error bounds

Stengel,et.al: "Adversarial super-resolution of climatological wind and solar data," 2020, doi: 10.1073/pnas.1918964117.



Can the FAIR-ification of digital objects help?

What should be made available for SW, data, and workflows to become FAIR?



Computing environments, submission scripts, libraries and their version number
Metadata: scientific metadata, performance counters, instrumentation choices, instrument metadata

Metadata exist and is captured in various *non-interoperable* data formats, schemas, and services
- data services, containers
- machine learning platforms and their versions: Tensorflow, pytorch, etc.

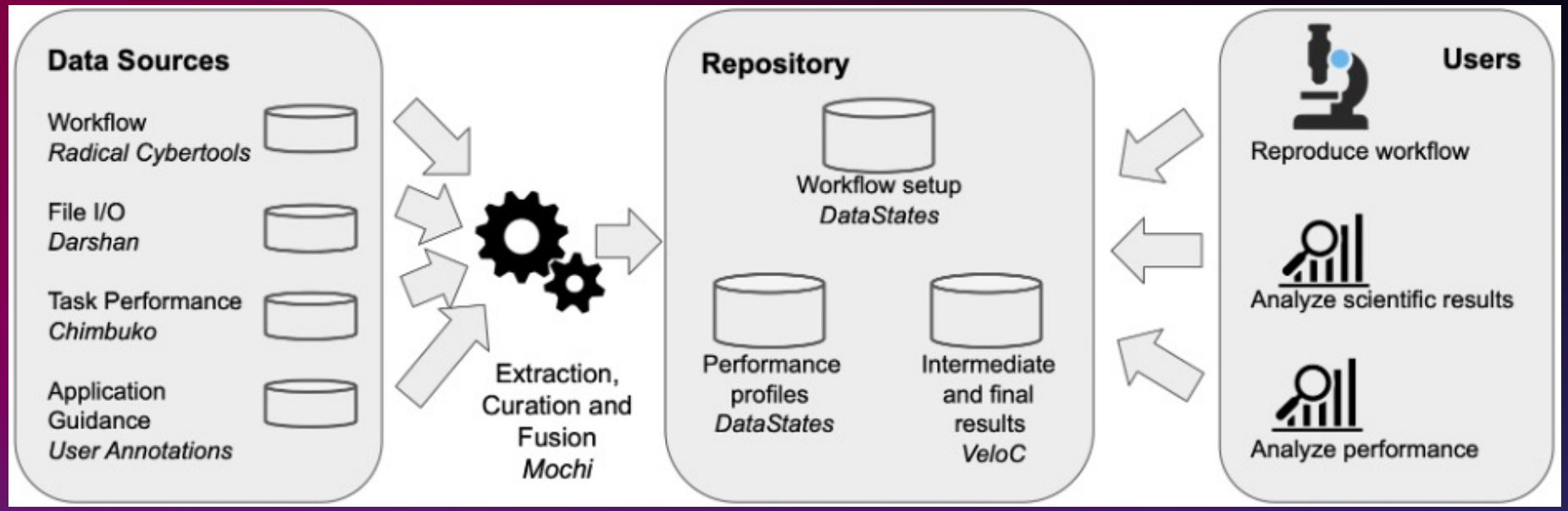
Metadata standards: WFCommons

Automatic capture of provenance:
- metadata and their relationship to data
- SW dependencies
- versioning

Persistent Identifiers: many schemes e.g. ARK, DOI, minIDs, easyIDs, etc.

```
In [5]: print_info()

System_Info:
        OS :  Ubuntu 18.04
        CUDA :  10.0
        numpy :  1.14.5
        GPU :  GeForce GTX 1080Ti

Platform_Info:
        pkatform :  tensorflow-gpu
        version :  1.14.0

Hyperprameters:
        model_type :  MLP
        layers_num :  5
        layer_info :
            layer1_num :  400
            layer1_activation :  tanh
            layer2_num :  400
            layer2_activation :  tanh
            layer3_num :  200
            layer3_activation :  tanh
            layer4_num :  200
            layer4_activation :  tanh
            layer5_num :  100
            layer5_activation :  tanh
        loss :  L1
        optimizer :  Adam
        batch_size :  200
        learning_rate :  0.0001
        epochs :  50

Random seed: 2
```

RECUP infrastructure enabling FAIR and reproducibility



(1) identify and capture the rich information necessary for reproducing hybrid workflows at scale: fuse, organize, store, index
(2) make the captured information FAIR to enable key workflow reproducibility tasks: re-runs, re-use workflows, data
(3) use the (meta)data to isolate where one workflow's execution deviated relative to another
(4) design reproducibility metrics for scientific and performance results

Thank you for your attention

Acknowledgements: the authors gratefully acknowledge the funding support from the U.S. Department of Energy Office of Science. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan.