Enabling Multi-Instrument Pixel-Level Science with a High Throughput Computing, Data Access and Analysis Facility

Margaret W. G. Johnson

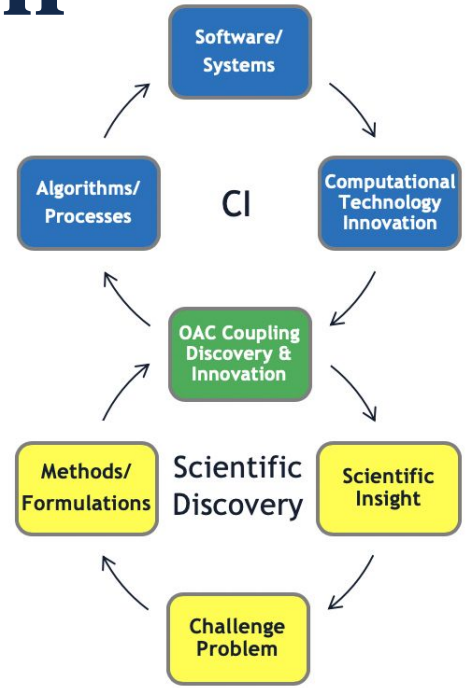
2019 NSF Workshop on Connecting LFs and CI

September 16, 2019

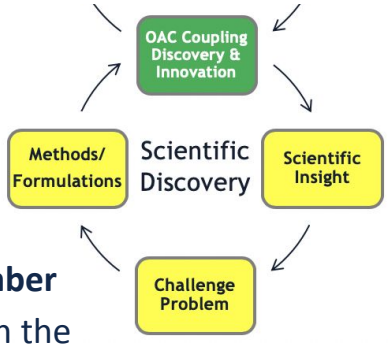
ILLINOIS
NCSA | National Center for Supercomputing Applications

Guidance from the NSF Blueprint for a National CI Ecosystem

- The blue cycle represents the evolution of facility-related capabilities, which are in general oriented towards benefitting multiple scientific domains.
- The yellow cycle represents work on specific challenge problems, which are often specific to a domain.
- The green block represents a materialized capability, often in the form of a physical facility, that provides acumen and resources that can be applied to multiple grand challenge problems.
- We have seen great historical success in the development of MPI machines in support of simulation science.
- This overall pattern can be applied to instrumental science and replicate the success we have seen for simulation science.



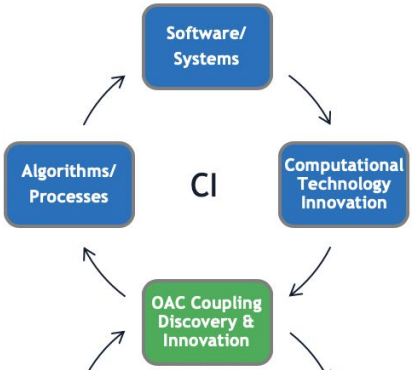
Motivation



- There are significant scientific projects that can be accomplished by allowing **a large number of modestly-funded individual investigators** to focus on their science topics by relying on the capabilities and expertise of a facility providing a level of data and technology support.
  - Success of the MPI/HPC centers and the HEP community
- There exists at NCSA diverse instrumental science domains needing exactly the same combination of technical expertise and technical solutions, including:
  - **Climate** scientist generating 2.4 PB of fused climate data from multiple instruments.
  - Discussions in the **multi-messenger astrophysics** community about joint processing between large facilities, but with no facility "owning" the provisioning and expertise for the processing.
  - Combining datasets from large **astronomical** instruments (e.g., Euclid, LSST, WFIRST); pixel-level processing to produce data not produced by the instrumentation project (e.g., specialized coadds).
  - Multiple cases of **emerging instrumental science** fields needing data science, e.g., bioimaging

Common CI Specialized Skills and Techniques

- Many observational data processing and data management methods are generic to a wider range of scientific disciplines:
  - data engineering
  - resource management
  - artificial intelligence
  - data fusion
  - virtual data (recomputation to trade off persistent storage)
  - and other topics related to large data.
- There is a general need to support this processing. There is very little of it that requires large MPI processing capability.



Characteristics of the Facility

- This would be a **shared CI facility** that would support the instrument facilities and the science exploitation of their data products.
  - Sufficiently large to **provide economies of scale** in resources and expertise
- The facility would
  - provide for application and dissemination of **data science and engineering expertise**.
  - working with the communities, **advance the state-of-the-art** relevant to the facility, jointly and economically benefiting the scientific domains that it serves.
  - **accelerate science** by supporting a large number of projects and grand challenges.