Management, Analysis, and Visualization of Experimental and Observational Data
The Convergence of Data and Computing

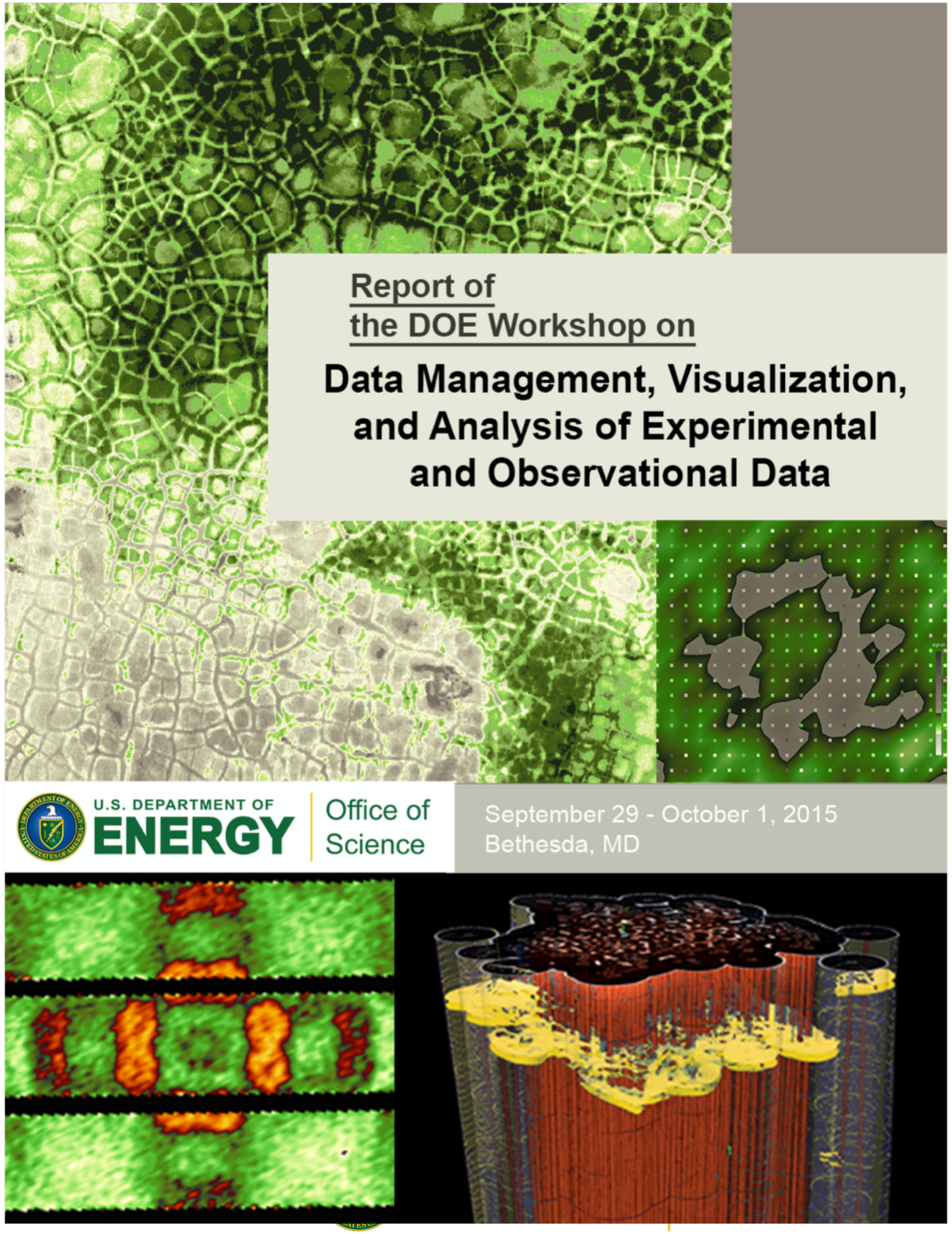
E. Wes Bethel

Lawrence Berkeley National Laboratory

02 August 2016

Workshop Objectives

- Better understand data-centric issues facing Office of Science Science User Facilities
- Identify for meeting those needs and science objectives
- Foster dialogue between EOD projects and ASCR



Report of the DOE Workshop on

Data Management, Visualization, and Analysis of Experimental and Observational Data

U.S. DEPARTMENT OF ENERGY | Office of Science

September 29 - October 1, 2015
Bethesda, MD

Executive Summary

- Gaining scientific knowledge from experimental data is increasingly difficult
  - DOE/SC operates dozens of Science User Facilities (SUFs), each generates vast amounts of data, which is quickly growing in size, speed, complexity
  - O(EB)/yr projected data size across SC SUFs within a short amount of time
  - ASCR-sponsored workshop in Sep 2015 to focus on data-centric issues of SC SUFs
- Convergence of data and computing: data- and computing-centric needs increasingly intertwined, symbiotic
  - Advances in computing help us to obtain better data from experiment, and do more with it, thereby increasing the value of each experiment
  - Similar to vision spelled out in the National Strategic Computing Initiative (Mar 2016)
- Acute, urgent data-centric needs in SUFs and science programs
  - Our ability to collect data far exceeds our ability to analyze and store data
  - Each EOS facility pursuing its own path towards meeting data-centric challenges: duplication of effort, increased costs program-wide, can benefit from more cross-program interaction

What’s coming

- Methodology for assessing needs
- What happened at the workshop
- Science use cases
- Findings
- Recommendations
- Next steps

Methodology for Collecting SUF Data Needs

- Provided a detailed use case template to science representatives
  - A “questionnaire” of sorts
  - More detail on next slides
- Science User Facilities (SUFs) polled:
  - SUFs: EMSL (PNNL), ARM (PNNL), ALS (LBNL), LCLS (SLAC), SNS & HFIR (ORNL), SPEM/STEM (ORNL), APS (ANL), DUNE (FNAL), HEP/Cosmic Frontier
  - Computationally focused projects: cosmology/astro (Johns Hopkins), climate (PNNL)
- Their responses (LaTeX + images) form 10 chapters of the workshop report

Describe Doing Science with Data (1/3)

- Present/near term, future term views of science project
  - Science objective & motivation
  - What SC computing facilities used, if any?
  - How do you “do science” with data?
  - Diagram showing data lifecycle, processing stages
  - Collaboration/interaction points in that process
  - How much data (what %) are you able to use now? How much would you like to be able to use in the future?

Data Lifecycle and The Five V’s (2/3)

- Data lifecycle
  - For the data lifecycle illustration, please provide discussion that describes key points in the process
  - Now and in the future
- Data-centric Requirements: The Five V’s, One L, One M, …
  - Velocity: data rates (e.g., DAQ), transfer rate requirements
  - Volume: data size
  - Variety: the different types (modes) of data one might want to collect and use in scientific analysis
  - Veracity: measure of noisiness, trustworthiness, problems/issues with data
  - Value: not all data are equal, some are ephemeral, some very long-lived
  - Lifespan (or longevity): what is the shelf life
  - Market for data: who are the consumers

Impediments, Gaps, Challenges (3/3)

- 3-5 impediments or barriers facing project now, and going into the future
- An ”open ended” invitation to speak up about problems and concerns
- Sample list of thoughts
  - Real-time processing/throughput requirements to support experiment tuning
  - Community centric data repos
  - Plan for distributing data
  - Resource shortages
  - Resilience in workflows
  - Metadata/provenance collection
  - Other barriers to making use of data

What’s coming

- Methodology for assessing needs
- What happened at the workshop
- Science use cases
- Findings
- Recommendations
- Next steps

Methodology for Collecting SUF Data Needs

- Provided a detailed use case template to science representatives
  - A “questionnaire” of sorts
  - More detail on upcoming slides
- Science User Facilities (SUFs) polled:
  - SUFs: EMSL (PNNL), ARM (PNNL), ALS (LBNL), LCLS (SLAC), SNS & HFIR (ORNL), SPEM/STEM (ORNL), APS (ANL), DUNE (FNAL), HEP/Cosmic Frontier
  - Computationally focused projects: cosmology/astro (Johns Hopkins), climate (PNNL)
- Their responses (LaTeX + images) form 10 chapters of the workshop report

Who Was At the Workshop?

Science User Facility	Lab	Office
Environmental Molecular Sciences Lab	PNNL	BER
Climate Modeling (Computing)	PNNL	BER
Atmospheric Radiation Measurement Climate Research Facility	PNNL	BER
Advanced Light Source	LBNL	BES
Linac Coherent Light Source	SLAC	BES
Spallation Neutron Source High Flux Isotope Reactor	ORNL	BES
Scanning Tunneling Electron Microscopy	ORNL	BES
Advanced Photon Source	ANL	BES

Science User Facility	Lab	Office
Deep Underground Neutrino Experiment	FNAL	HEP
Cosmology/astrophysics (Computing)	JHU	HEP
Cosmic Frontier	ANL	HEP

Math/CS/Facilities - ASCR
Computing/Network Facilities: NERSC, OLCF, ALCF, ESnet
CS/math: data management, data analysis (ML, graph analytics, statistical analysis, etc.)/visualization, data mining, operating systems/runtime, workflow, optimization, UI/human factors, applied mathematics

Guests
NSF Computational Facility
UK Science Grid

Process for Identifying Needs, Gaps in Technology

- What happened at the workshop?
  - Science presentations
  - Breakouts: group discussions
  - Lightning round responses
    - What are primary needs
    - What are primary research topics (gaps) to meet those needs
  - Science feedback to math/CS/facility responses
- After the workshop
  - Review and input from the broader community

What's coming

- Methodology for assessing needs
- What happened at the workshop
- Science use cases
- Findings
- Recommendations
- Next steps

EOS Projects and Data Lifecycle

- 11 Science Use Cases
  - Significant diversity, but also significant common themes (later slides)
- The following slides show several distinct illustrations of science use cases and data lifecycle

Cosmic Frontier (HEP)

"…the main new trend for science user facilities is the evolution and integration of HPC systems within a data-centric usage model."

Mission focus areas:

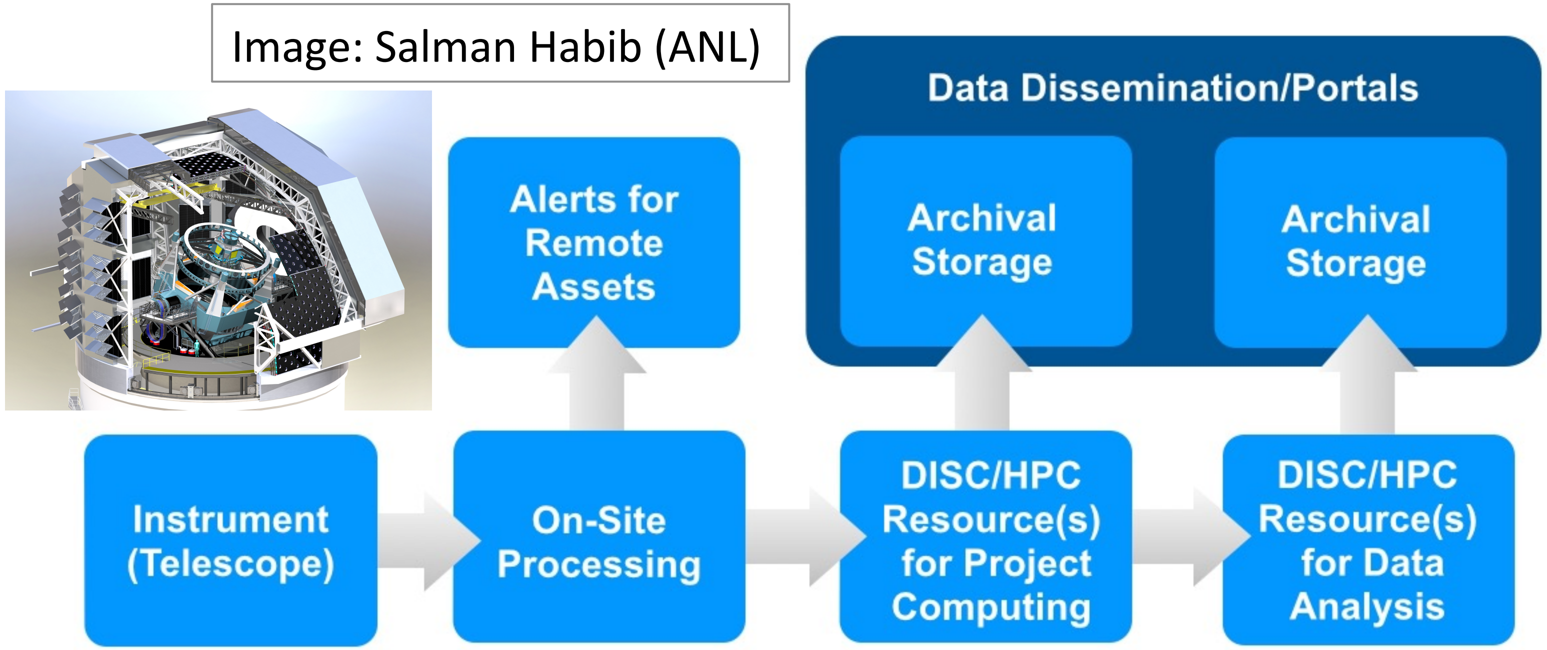
- Detection and mapping of galactic/ extra-galactic sources of radiation
- Understand nature of cosmic acceleration, dark matter

- Telescopes funded/operated by other agencies, e.g., NSF
- Large communities engaged in research (consume data)
- Broad and prolonged use of data products
- Raw data sizes O(1-100) PB, processed to produce catalogues
- LSST: 15TB/day DAQ rate

Impediments/Gaps

- Lack of trained manpower, career paths
- Integration of data movement, storage, archival within workflows
- Strategies for data archival, curation
- Lack of standard data formats
- Software stack for next-gen arch's
- Machine learning/stats methods for analysis

Image: Salman Habib (ANL)



Data Dissemination/Portals
Alerts for Remote Assets
Archival Storage
Archival Storage
Instrument (Telescope)
On-Site Processing
DISC/HPC Resource(s) for Project Computing
DISC/HPC Resource(s) for Data Analysis

Advanced Light Source (ALS)

> "...it is getting to the point where users cannot just download their data: their hard drive isn't big enough, and if it was, they wouldn't have adequate computing power to do anything with it."

Mission focus:
Synchrotron light source for imaging, scattering, and spectroscopy experiments in chemical, geological, life, material and physical sciences.
Users come, do experiment, want to leave with data in hand.

Impediments/Gaps

- Diversity of science: one accelerator, O(40) beamlines, diverse experiments at each: no "one size fits all"
- Usability and accessibility of computing, data
- Desire to steer data collection
- Data volume, rate exceed capacity and capability
- New data vis/analysis methods

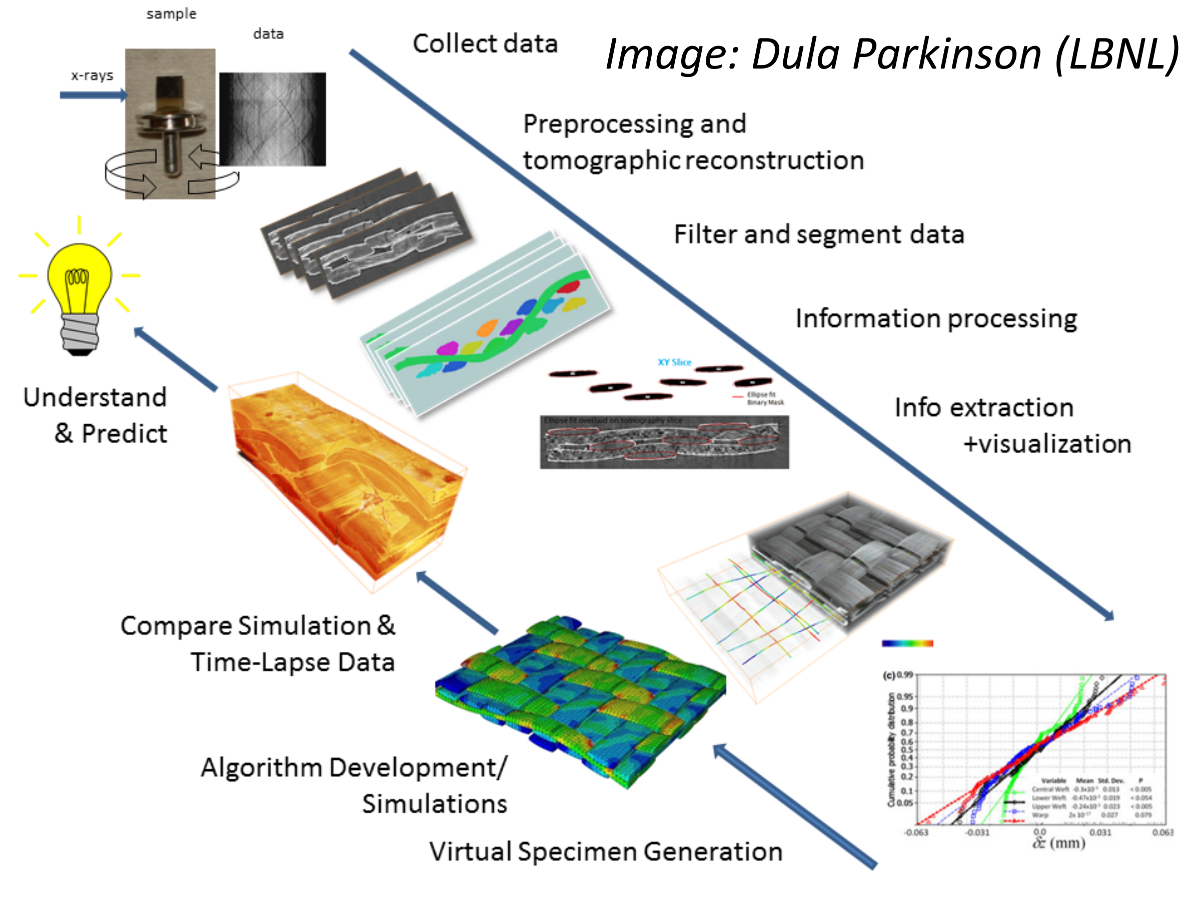
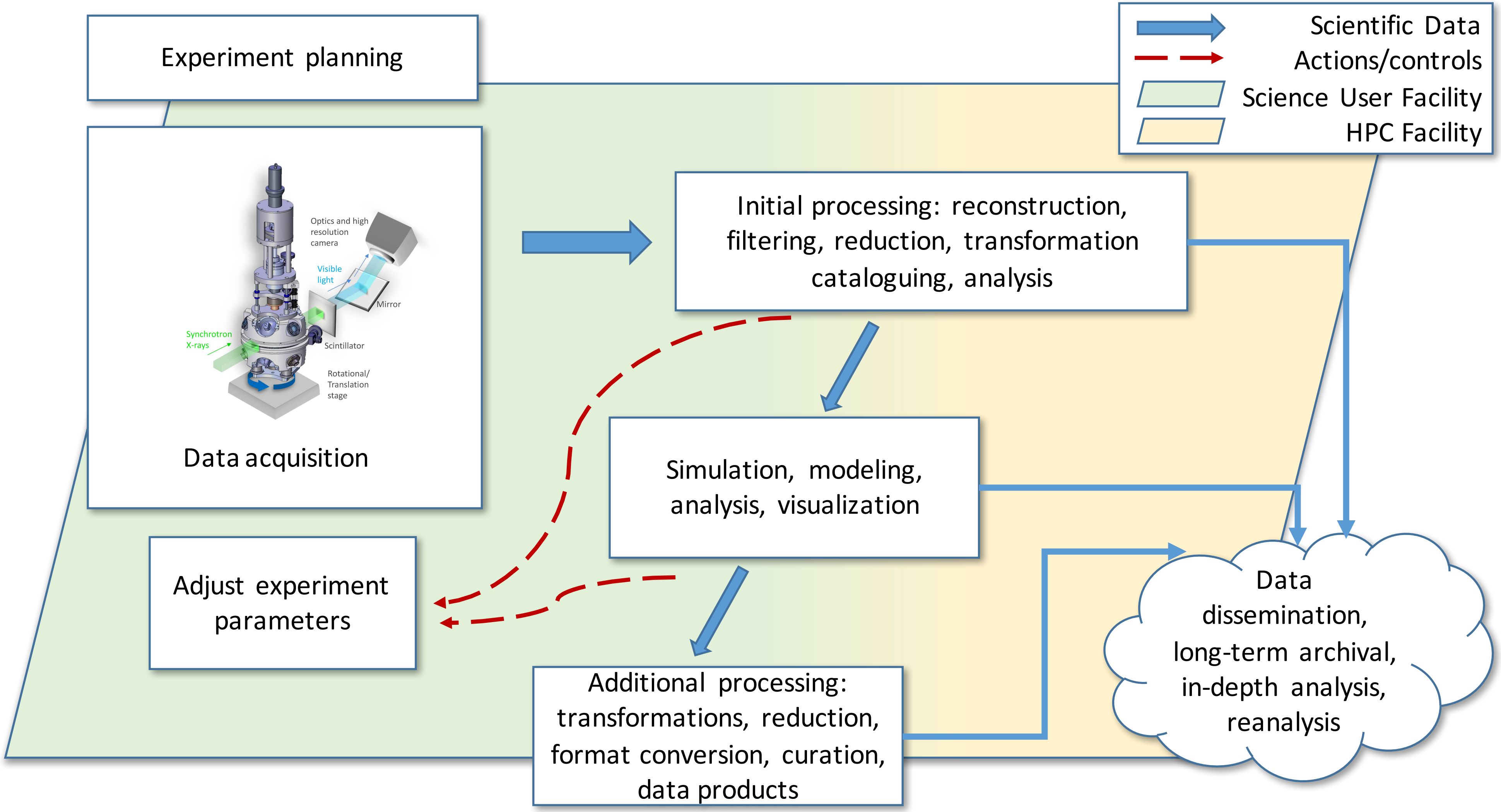

sample
data
x-rays
Collect data
Preprocessing and tomographic reconstruction
Filter and segment data
Information processing
XY Slice
Info extraction +visualization
Understand & Predict
Compare Simulation & Time-Lapse Data
Algorithm Development/ Simulations
Virtual Specimen Generation

Image: Dula Parkinson (LBNL)

Environmental Molecular Science Laboratory (BER)

The Data Lifecycle of a Canonical EOS Project


Optics and high resolution camera
Visible light
Mirror
Synchrotron X-rays
Scintillator
Rotational/ Translation stage

Scientific Data

Actions/controls

Science User Facility

HPC Facility

Experiment planning

Data acquisition

Initial processing: reconstruction, filtering, reduction, transformation cataloguing, analysis

Simulation, modeling, analysis, visualization

Additional processing: transformations, reduction, format conversion, curation, data products

Adjust experiment parameters

Data dissemination, long-term archival, in-depth analysis, reanalysis

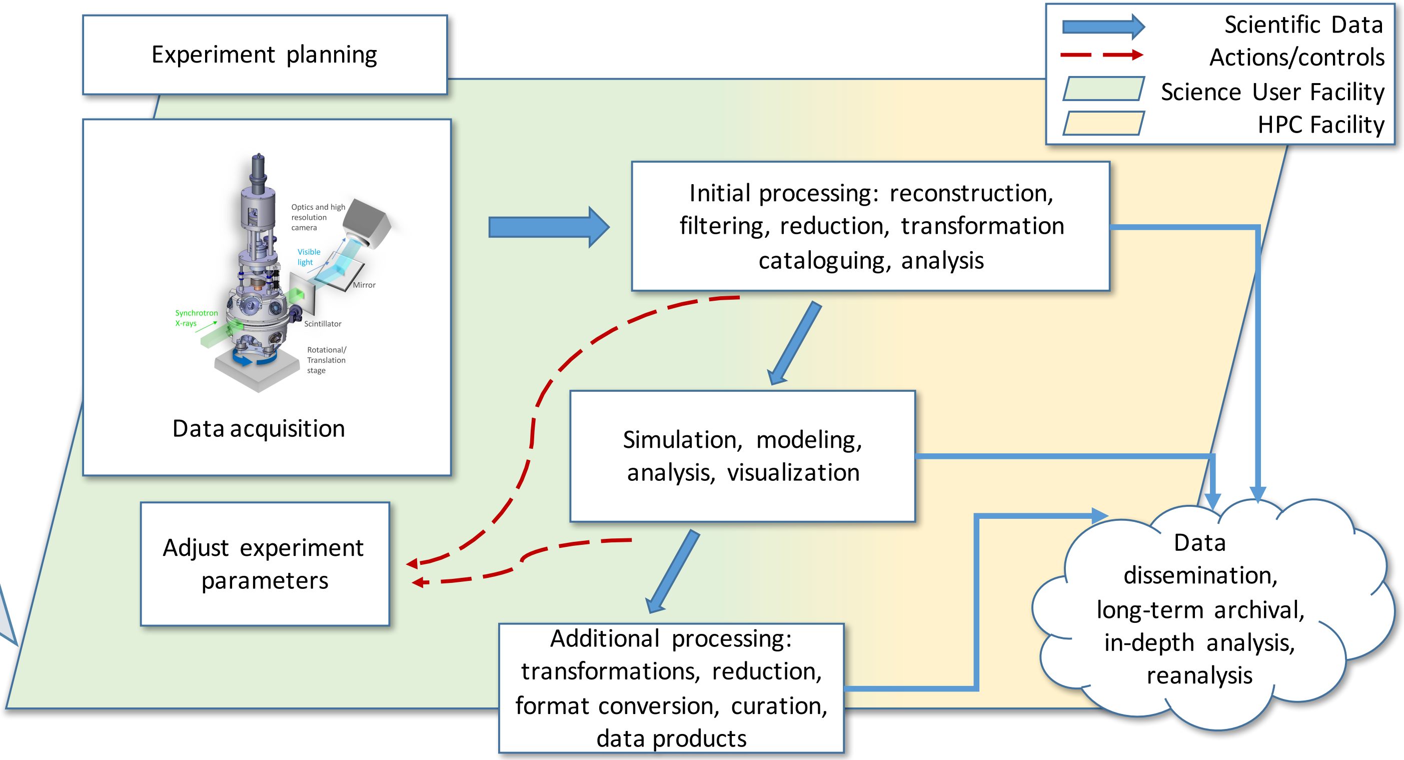
The Data Lifecycle of a Canonical EOS Project

Computational and experimental ecosystem

“All EOS requires computing”

Some is located close to the instrument (green), some may be at HPC facilities located elsewhere (yellow).

This illustration intends to show that the different processing stages might happen at either local or remote locations, but to allow for either configuration.



Optics and high resolution camera
Visible light
Mirror
Synchrotron X-rays
Scintillator
Rotational/ Translation stage

Scientific Data

Actions/controls

Science User Facility

HPC Facility

Experiment planning

Data acquisition

Initial processing: reconstruction, filtering, reduction, transformation cataloguing, analysis

Simulation, modeling, analysis, visualization

Adjust experiment parameters

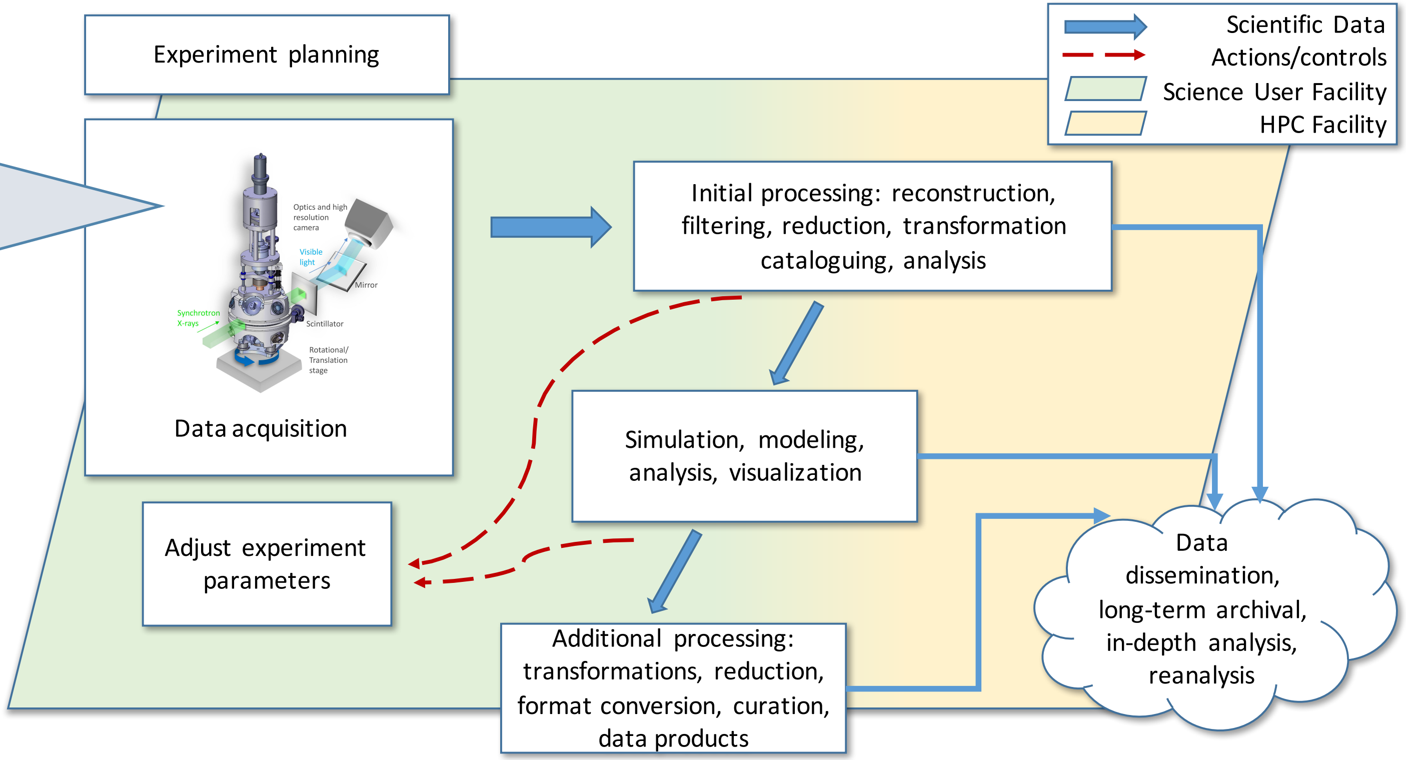
Additional processing: transformations, reduction, format conversion, curation, data products

Data dissemination, long-term archival, in-depth analysis, reanalysis

The Data Lifecycle of a Canonical EOS Project

Obtaining data from the instrument/device

Straightforward (lab analysis of sample) to complex (sky survey)



Optics and high resolution camera
Visible light
Mirror
Synchrotron X-rays
Scintillator
Rotational/ Translation stage

Experiment planning

Data acquisition

Initial processing: reconstruction, filtering, reduction, transformation cataloguing, analysis

Simulation, modeling, analysis, visualization

Additional processing: transformations, reduction, format conversion, curation, data products

Adjust experiment parameters

Data dissemination, long-term archival, in-depth analysis, reanalysis

Scientific Data

Actions/controls

Science User Facility

HPC Facility

The Data Lifecycle of a Canonical EOS Project

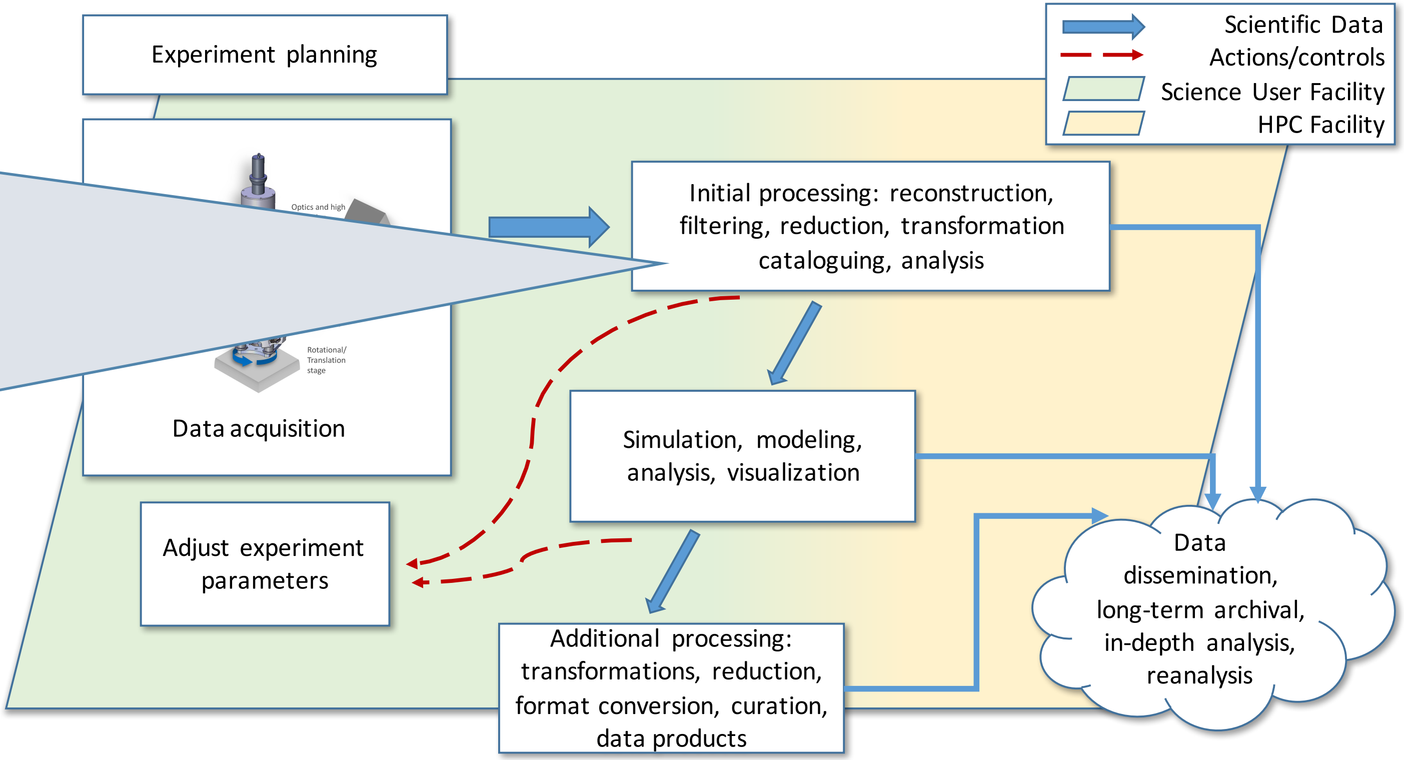
Primary analysis and initial processing (low latency)

Verify the experiment is working as expected

Early-stage processing/analysis to reduce data size, to extract most meaningful information from data, data products, alerts

May result in changes to experimental parameters

Typically occurs “close to” the instrument at the SUF



Scientific Data

Actions/controls

Science User Facility

HPC Facility

Experiment planning

Data acquisition

Initial processing: reconstruction, filtering, reduction, transformation cataloguing, analysis

Simulation, modeling, analysis, visualization

Adjust experiment parameters

Additional processing: transformations, reduction, format conversion, curation, data products

Data dissemination, long-term archival, in-depth analysis, reanalysis

U.S. DEPARTMENT OF ENERGY | Office of Science

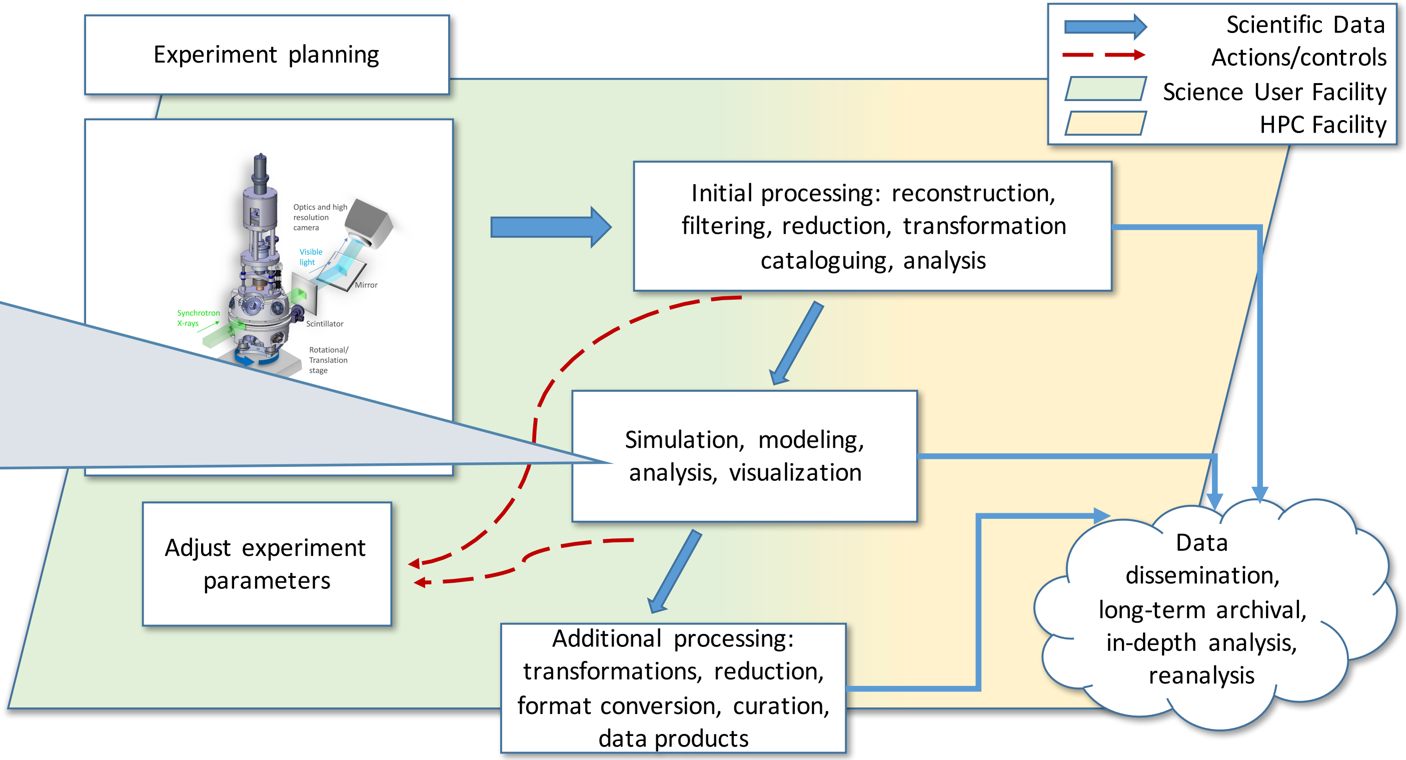
The Data Lifecycle of a Canonical EOS Project

EOD validation, deeper analysis

Use of computational models, compare experiment with simulation, deeper hypothesis testing, data products

Adjustment of experiment parameters to obtain better data

Likely requires significant low-latency HPC resources



Experiment planning

Initial processing: reconstruction, filtering, reduction, transformation cataloguing, analysis

Simulation, modeling, analysis, visualization

Adjust experiment parameters

Additional processing: transformations, reduction, format conversion, curation, data products

Data dissemination, long-term archival, in-depth analysis, reanalysis

Scientific Data

Actions/controls

Science User Facility

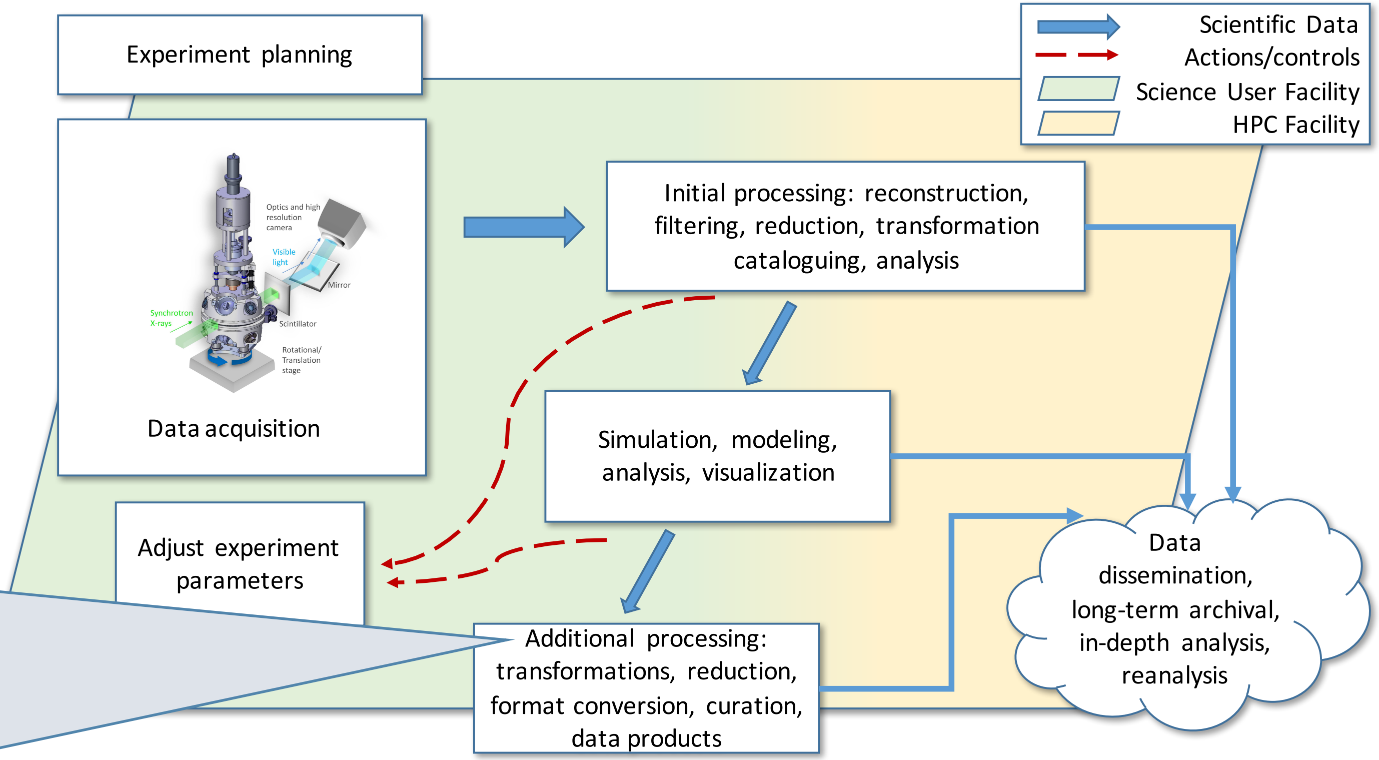
HPC Facility

Optics and high resolution camera
Visible light
Mirror
Synchrotron X-rays
Scintillator
Rotational/ Translation stage

The Data Lifecycle of a Canonical EOS Project

Preparation of curated data products

Could require significant HPC resources (e.g., Sky Surveys and catalogues)



Experiment planning

Data acquisition

Initial processing: reconstruction, filtering, reduction, transformation cataloguing, analysis

Simulation, modeling, analysis, visualization

Adjust experiment parameters

Additional processing: transformations, reduction, format conversion, curation, data products

Data dissemination, long-term archival, in-depth analysis, reanalysis

Scientific Data

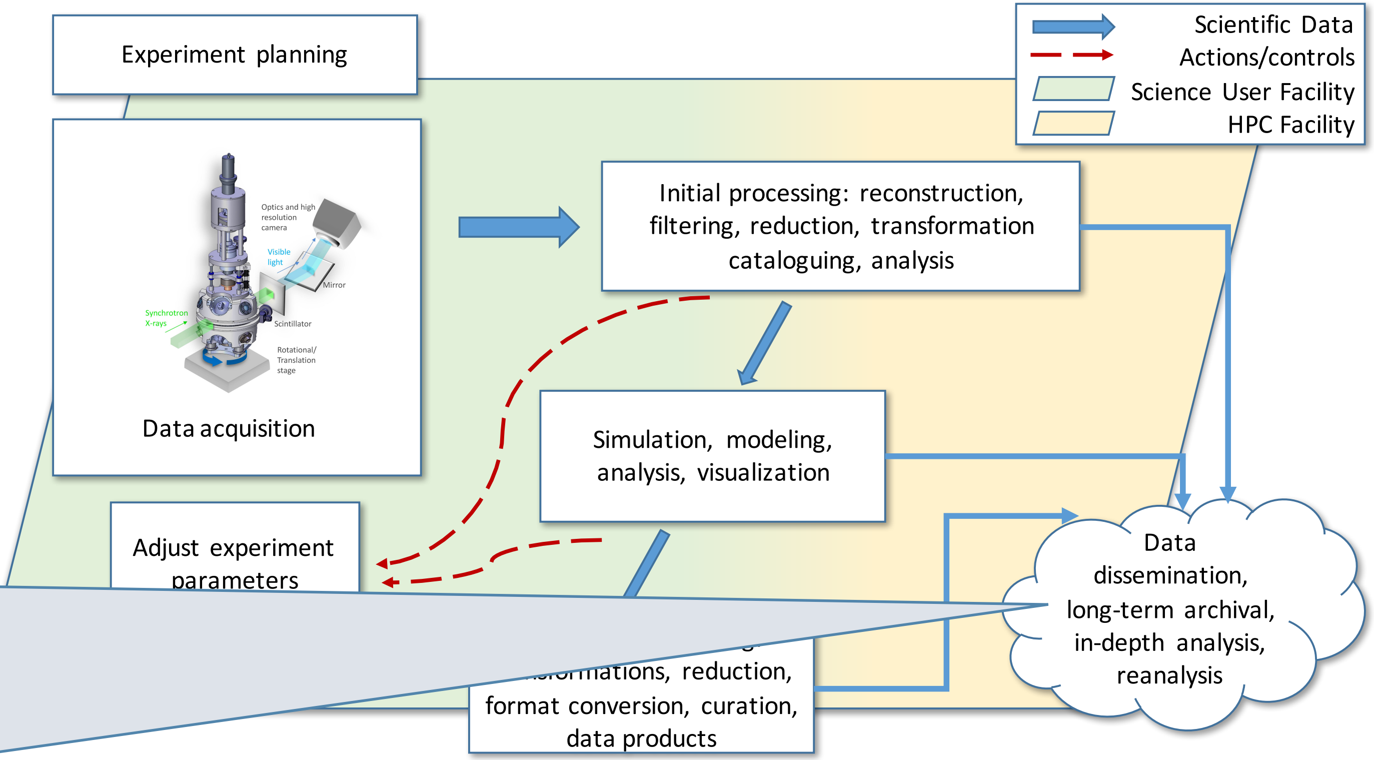
Actions/controls

Science User Facility

HPC Facility

Optics and high resolution camera
Visible light
Mirror
Synchrotron X-rays
Scintillator
Rotational/ Translation stage

The Data Lifecycle of a Canonical EOS Project


Experiment planning
Optics and high resolution camera
Visible light
Mirror
Synchrotron X-rays
Scintillator
Rotational/ Translation stage
Data acquisition
Adjust experiment parameters
Initial processing: reconstruction, filtering, reduction, transformation cataloguing, analysis
Simulation, modeling, analysis, visualization
…nsformations, reduction, format conversion, curation, data products
Data dissemination, long-term archival, in-depth analysis, reanalysis
Scientific Data
Actions/controls
Science User Facility
HPC Facility

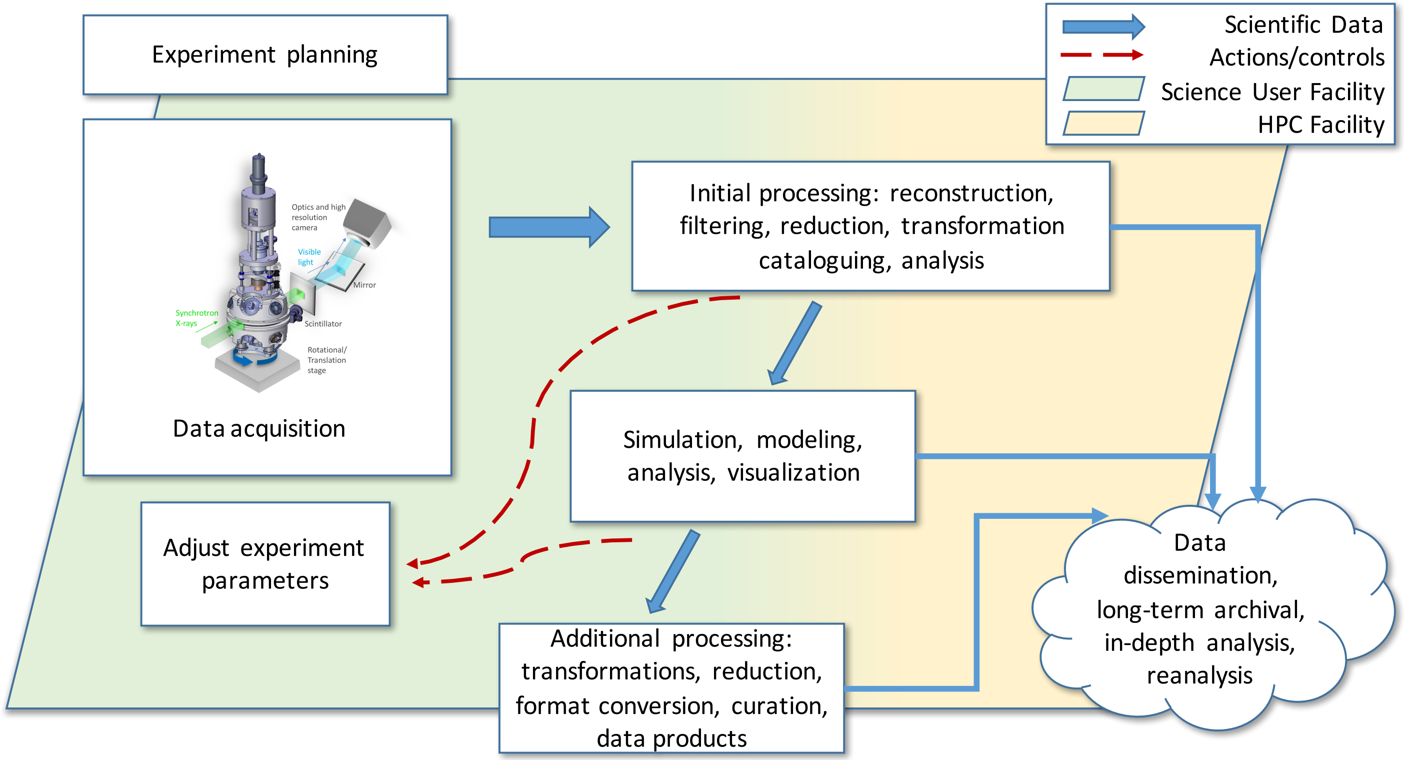
EOD can have a very long shelf life as it is distributed to “the world” and subject to reanalysis for future studies outside the scope of the initial experiment

Where is this data hosted?
Is it well documented?
Is there code to work with it?

The Data Lifecycle of a Canonical EOS Project

Data products:

- EOS projects support scientific research where data is collected from experiments.
- Consumer may be single PI or an entire community
- Potentially long lifespan, potentially quite large
- Want to know things like who created data product, under what conditions, etc. (Metadata/ provenance)
- Reference datasets
- Sharing with colleagues
- Including with publications



Optics and high resolution camera
Visible light
Mirror
Synchrotron X-rays
Scintillator
Rotational/ Translation stage

Scientific Data

Actions/controls

Science User Facility

HPC Facility

Experiment planning

Data acquisition

Initial processing: reconstruction, filtering, reduction, transformation cataloguing, analysis

Simulation, modeling, analysis, visualization

Adjust experiment parameters

Additional processing: transformations, reduction, format conversion, curation, data products

Data dissemination, long-term archival, in-depth analysis, reanalysis

What's coming

- Methodology for assessing needs
- What happened at the workshop
- Science use cases
- Findings
- Recommendations
- Next steps

EOD Growth Outpacing Computing Growth

- Detectors and other sensors increasing in resolution and speed faster than processors and memory.
- Within a few years from now
  - Each SUF estimating O(10s) PB/yr
  - ***Aggregate across SC: multiple EB/yr***

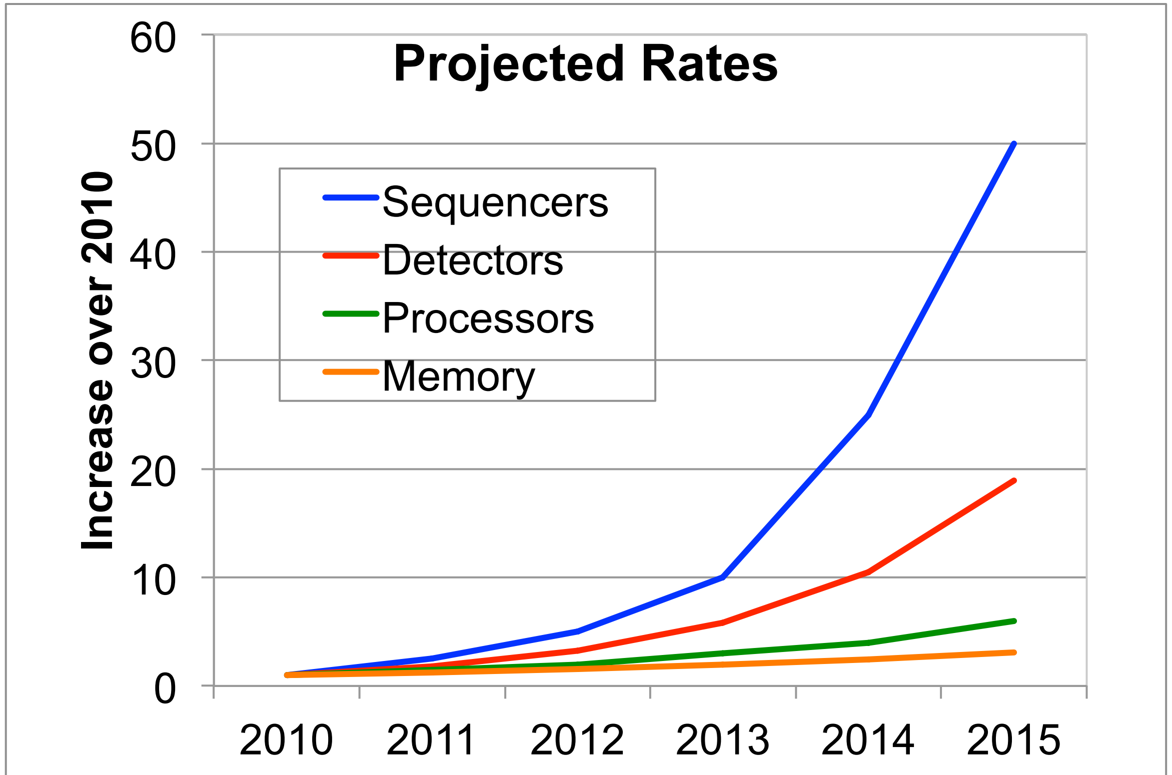


Image: Kathy Yelick (LBNL)

The Struggle to Keep Up with the Demands and Opportunities Resulting from a Data Flood

- Data acquisition rates at each facility approaching 10s of PB/yr
  - Aggregate across all SC: multiple exabytes/yr
  - Data also in increasing in complexity and diversity
- Driver: increase in resolution of sensors, increases in flux and brightness, multiple simultaneous instruments
  - Example: Sky survey projects use multiple instruments that look at different wavelengths of light
- Challenges: analyzing and visualizing data, moving data, harvesting metadata, storing and sharing data
- Convergence thought: want data that is free from errors and that focuses on a specific science objective
  - Bring to bear computational methods to ensure the best possible data are collected during an experiment
- Opportunity cost: potential loss of science due to inability to capture, store, analyze, visualize, and share data

> A key limitation today is our [in]ability to analyze and visualize the acquired data due to its volume, velocity, and variety.
>
> *B. Toby (ANL)*
> *Advanced Photon Source*
> *X-ray imaging/microscopy*

EOS Projects' Use of Large-Scale, High Performance Computing Facilities

- Meeting data challenges requires more resources: compute, networking, storage, suitable software, and operational policies
- EOS workloads/needs are different than traditional HPC workloads:
  - Computational: characterized as low-latency, high-throughput, data- and compute-intensive (more on next slide).
  - Data products and distributions: experimental data has a long shelf life; many projects require community access to massive datasets for a long period of time.
- Challenge: many HPC facilities have operational policies optimized for large-concurrency batch jobs, rather than fast-turnaround, high-throughput distributed workloads

> Procedures for moving data from place to place, including tools for automating resilient workflow for orchestrating distributed data-related operations are a bottleneck.
>
> *P. Rasch (PNNL), Climate modeling and analysis*

Time-Critical Data Needs

- EOS projects require low-latency, high-throughput response from infrastructure for data movement, analysis, processing and storage.
- Drivers:
  - (Convergence topic) Experiment optimization/tuning: near real-time analysis of experiment results (using HPC platforms) to adjust experiment parameters to obtain better data.
  - Increasing throughput to keep pace with data acquisition rates.
- Challenges:
  - Many HPC facilities have operational policies optimized for large-concurrency batch jobs, rather than fast-turnaround, high-throughput distributed workloads
  - Algorithms and software infrastructure designed for workloads of the last decade (or last century) likely inadequate to accommodate these data loads and response times.

> Predicting optimal [experiment] parameters could optimize data collection schemes and ultimately provide better quality data.
>
> *B. Toby (ANL)*
> *Advanced Photon Source*
> *X-ray imaging/microscopy*

The Risk of Unusable Data

- Without adequate metadata, scientific data has limited usefulness:
  - Unknown origin, unknown processing applied to data
  - Limits ability to validate or reproduce results
  - Virtually impossible to ”search” data that has no metadata (this data is not a text document or web page)
- In many projects, data-centric operations (management, analysis, movement, distribution) are the responsibility of an individual science user
  - As a result, only a fraction of collected data is ever analyzed, and only a fraction of that is ever published or made available
- Drivers
  - Compliance with Executive Order for data dissemination
  - Opportunity for new science unforeseen by original experimentalists when data is shared, reused
- Challenges:
  - Collecting metadata is difficult
  - There are no standard methods, tools
  - Not an intrinsic part of the culture (yet)

> One very real problem is that data is almost never usable by anyone other than the person who produced it. This problem must be solved if making data publicly available is to have any useful purpose.
>
> *H. Steven Wiley (PNNL)*
> *Environmental Molecular Sciences Laboratory*

Collaboration and Sharing are Central to EOS

- Sharing of data, tools, and methodologies are central to modern EOS, their mission is to collect data and share it
  - Yet existing infrastructure is insufficient, inadequate
- Drivers:
  - Need for sharing within a single project group, and sharing with the broader scientific community
  - Common methods/tools for working with data require common data models/formats, as well as simple ways to capture metadata, provenance, methodologies
- Challenges:
  - Approaches for sharing are *ad hoc*: each site (or individual) “rolls their own” data formats and tools, approaches for sharing
  - ”Home brew” methods/tools not likely to be adopted/used by others
  - High barrier to entry results in many simply not bothering to share
- Opportunities
  - Potential for reducing costs of duplicated effort: in software (individual vs. community), and data repositories (a few centralized ones vs. many small individual ones)
  - Potential for new methods/algorithms/software to emerge around well-defined data models and validated against curated data

> Current technologies are inadequate for sharing data between group memebers. The community needs a more fluid means for sharing data and working together.
>
> *H. Steven Wiley (PNNL) Environmental Molecular Sciences Laboratory*

EOS Data Lifecycle Needs Not Being Met

- EOS projects have significant, complex data lifecycle needs that go beyond what is provided by HPC facilities
  - Data lifecycle refers to all stages of data collection, movement, processing, analysis, management, curation, and sharing.
- Drivers:
  - EOS mission is to collect data and share it
  - Reference datasets: widely used by many authors; new discoveries possible from reanalysis; validation of new methods and instruments
- Challenges:
  - *Ad hoc* nature of meeting these needs: the responsibility often falls on the shoulders of the individual scientist
  - No program-wide means for long-term data archival, sharing
- Opportunities
  - Potential to increase science discoveries per experiment

> ..our only archival process right now is that provided by the published journal.
>
> Providing more access to data, in a manner that can be used by more scientists, will improve efficiency, increase scientific impact, and result in more discoveries per experiment.
>
> *G. Granroth and T. Proffen (ORNL)*
> *Spallation Neutron Source*

The Central Role of Software in EOS Projects

- Software is a critical element for all EOS projects in all aspects of working with data
- Drivers:
  - EOS projects vulnerable to inefficiencies and increased costs that can result from software-related issues (e.g., manual vs. automated execution)
- Challenges:
  - How to create the scientific software needed to run the science facility
  - The current *ad hoc* nature of software development, where there is little, if any, software reuse across facilities, or, in many cases, within the same facility.
  - Increasing complexity: distributed, complex scientific workflows
- Opportunities
  - Software methods, like analysis algorithms, play a key role in improving the quality of data collected during an experiment
  - Improving scientific productivity: by encapsulating complexity, by automating tasks, by resiliently responding to faults

> Each beamline operates with unique capabilities and an independent scientific mission..Computational needs and strategies may differ, but computation is required for nearly every aspect of the facility.
>
> *B. Toby (ANL)*
> *Advanced Photon Source*
> *X-ray imaging/microscopy*

Workforce Development and Retention

- The single most precious resource we have in the sciences is our personnel
- Challenges:
  - Specialized knowledge/training across multiple disciplines is required to be effective
  - Data scientists are in high demand in industry
  - Inadequate or insufficient career paths
- Opportunities
  - EOS is a problem rich environment
  - EOS offers a unique training and workforce development environment

Many data-centric problems require close interaction between the domain scientist and data scientist. Generally, such dual background is the exception and some amount of professional training is required to fill this gap.

S. Kalinin (ORNL)
Scanning probe and scanning transmission electron microscopy

What’s coming

- Methodology for assessing needs
- What happened at the workshop
- Science use cases
- Findings
- Recommendations
- Next steps

Recommendations

- Lots of slides, one per finding/recommendation area, not enough time to present on 02 Aug 2016 at JLab
- In brief:
  - ASCR HPC facilities designed and operated for large-concurrency computational workloads; demands of EOS are different, they need to evolve
  - Collaboration/sharing in EOS is the norm: the computational infrastructure/ ecosystem needs to evolve in several key ways to facilitate
  - Metadata/provenance are a huge unsolved problem
  - Program-wide visibility of and action on data issues can have a huge positive impact
  - EOS projects need automated, reproducible ”data pipelines”
  - Practices in software development, deployment can be improved
  - Develop and nurture and EOS-focused data science workforce

Summary of Math, CS Research Topics

- Lots of slides, one per focus area

What’s coming

- Methodology for assessing needs
- What happened at the workshop
- Science use cases
- Findings
- Recommendations
- Next steps

Call to Action

- We need your help in planning next steps
  - Don't really want to have another workshop…
- The questions Amber posed on my behalf:
  - How do you deal with data now, what are priorities?
  - Where do you want to be in 5-10 years?
  - What are the pain points you feel, what works and doesn't work?
- Other questions:
  - What type of "business model" would work best?
  - Examples:
    - Single PI ASCR grants to focus attention on issues and with science stakeholders
    - SciDAC-like projects
    - CAMERA-like projects
    - CoDesign-like projects
    - How to increase surface area between Math/CS/Data people and your program(s)?