2015 ASCR EOD Workshop: Key Challenges and Opportunities

E. Wes Bethel, LBNL

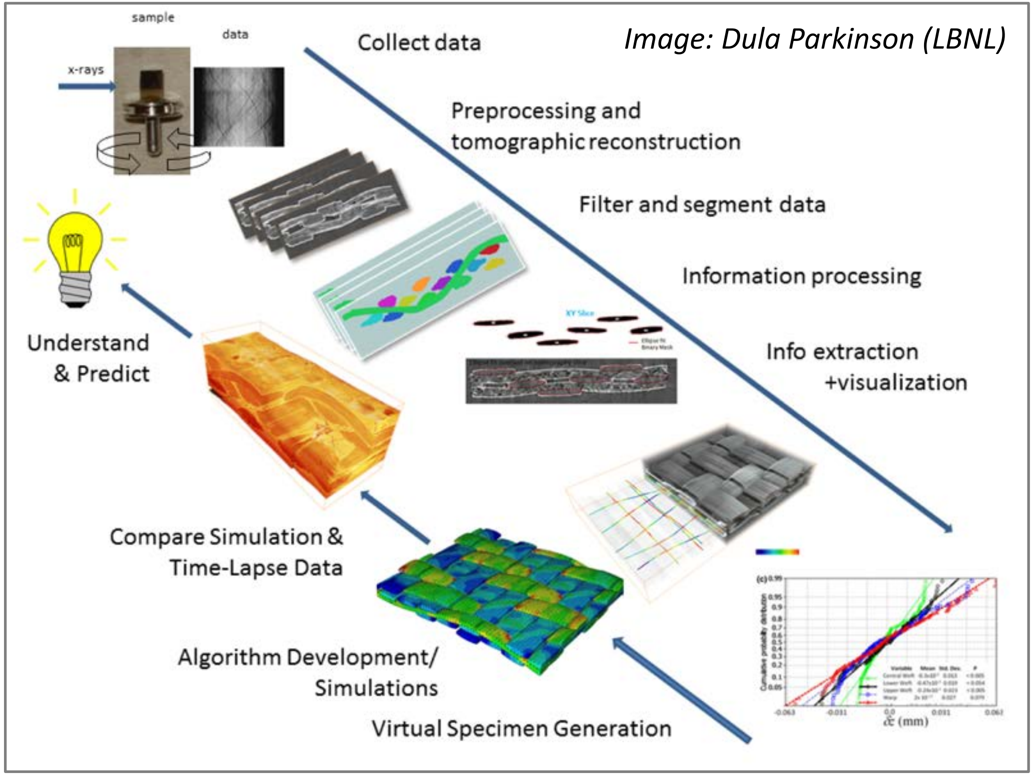
19 Sep 2018, SSIO Workshop, Gaithersburg MD

Definitions, Context

- **Reports/sources**
  - 2015 ASCR Workshop on Management, Analysis, and Visualization of Experimental and Observational Data
  - Requirements Review series (2015-2016): determine requirements for an exascale ecosystem that includes computation and data.
  - ASCAC (2013): intertwining of data and compute
    - Science examplars with computing and data challenges
    - Uses ”data lifecycle” as a framing mechanism
  - NSF (2016): Realizing the Potential of Data Science
    - Uses “data lifecycle” for framing a discussion about R&D of methods/infrastructure, for team/center formation, as an vehicle for workforce development.
    - https://www.nsf.gov/cise/ac-data-science-report

- **Definitions**
  - EOD == Experimental and Observational Data
  - EOS == Experimental and Observational Science
- **Today: summary of key findings, thoughts on implications for SSIO community**

Main Message: EOS Impeded by Data Lifecycle Challenges



- **Where will it be stored?**
  - Volume: multiple EB/yr
- **Can it be absorbed and processed quickly enough?**
  - Rate: time-critical needs
- **Is data usable?**
  - No metadata = unusable data
- **How is data used?**
  - Product vs source, shared vs. private
- **How long will it live?**
  - Minutes, years, decades?
  - Don’t forget about the software
- **How will we do it?**
  - The critical role of software
  - Collaboration and sharing
- **Who is going to do it?**
  - Workforce development, retention

ALS-U Enables Next-Generation Science

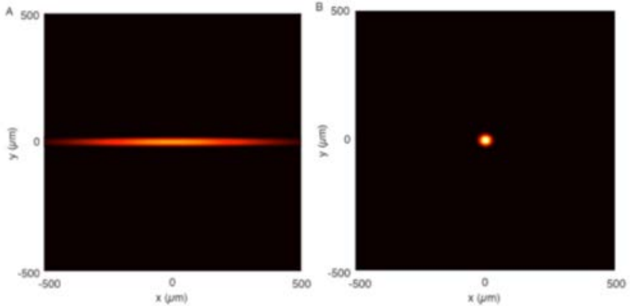
ALS-U produces a much brighter and focused x-ray source

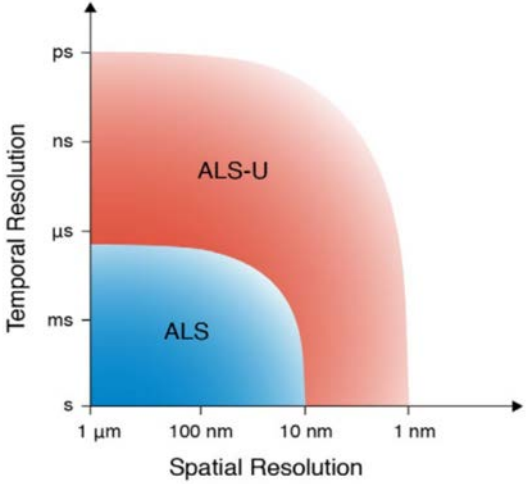
- **Basis for experiments for years to come**
- **Maintain US leadership for "the foreseeable future"**

3D nanoscale imaging with high spectral sensitivity over <u>broad space and time scales</u>.

- **ALS: homogeneous and simply organized systems**
- **ALS-U: heterogenous and hierarchical systems, evolving over time**

The ALS-U provides an x-ray beam (right) that is much more highly focused and brighter than the ALS beam (left).

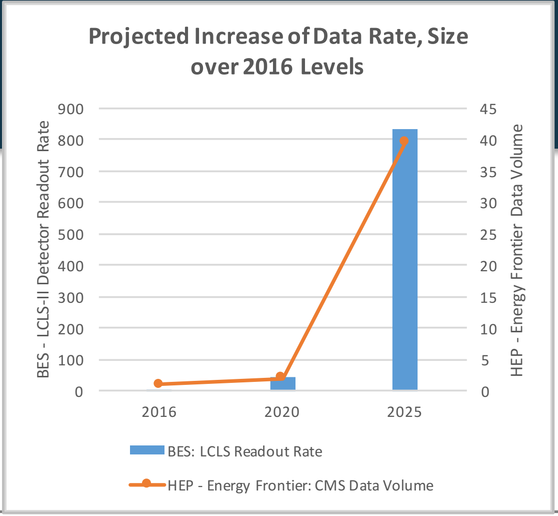




Source: ALS-U: Solving Scientific Challenges with Coherent Soft X-Rays, Jan 2017.

Multiple Exabytes of Data per Year

- **Detectors and other sensors increasing in resolution and speed faster than processors and memory.**
- **Science User Facilities (SUFs) are estimating O(10s) PB/yr**
- **Across SC SUFs in the coming years, *<u>the aggregate forecast is multiple EB/yr.</u>***
- **Are we ready?**



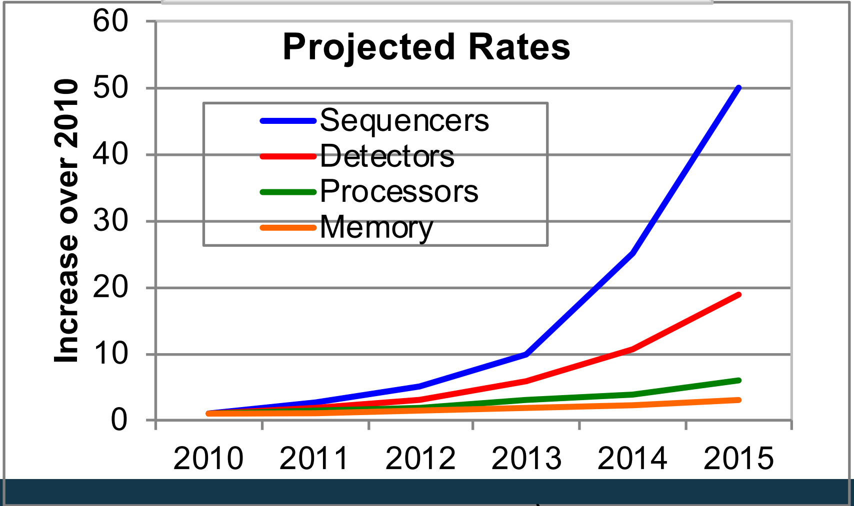
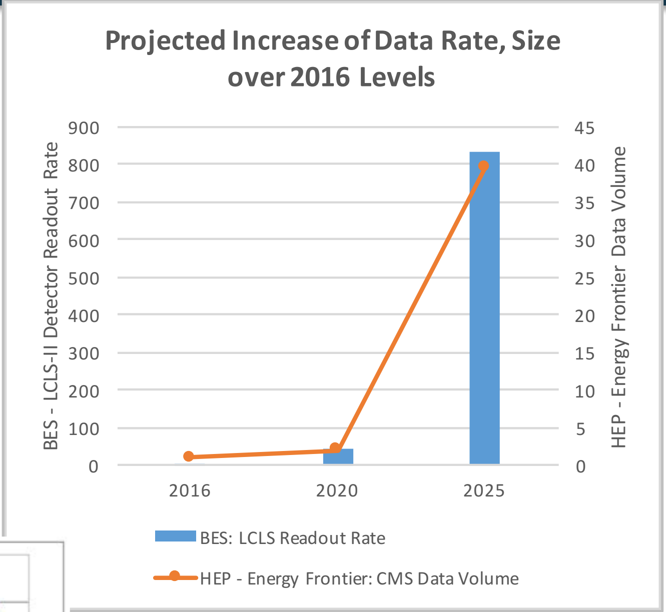


Image: Kathy Yelick (LBNL)

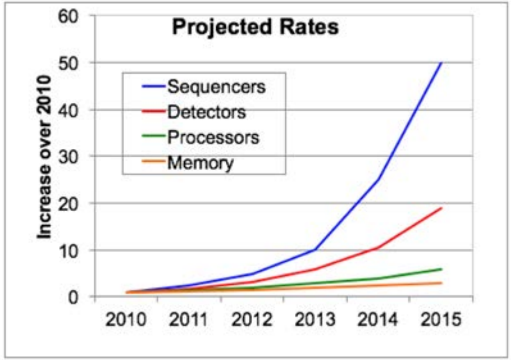
EOD Size Increasing Due to Better Instrumentation, Detectors, Readback

- **Increasing detector resolution and readback rate drives exponential data growth rate in light sources, other EOS**
- **Affects many sciences areas. Here:**
  - LCLS-II readback rates: 120 Hz (2016), 5 KHz (2020) to 1 MHz (2025)
  - CMS data volume: 5 PB (2016), 10 PB (2020) to 197* PB (2025)
- **<u>Each facility: 10s PB/month within a few years</u>**



*197 PB is an average of a low and high scenario called out in [1].

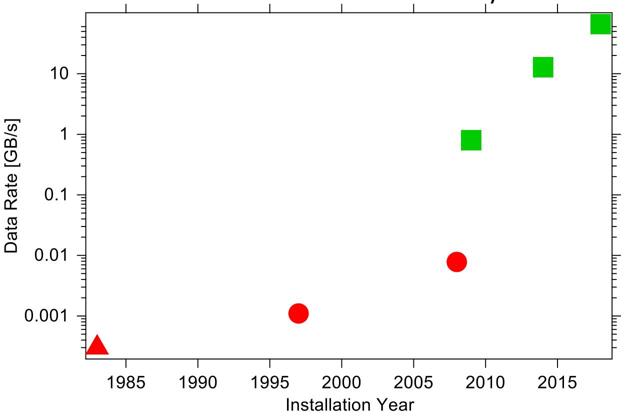
Data Sources:
1. 2015 HEP Exascale Requirements Review Workshop Report
2. 2015 BES Exascale Requirements Review Workshop Report
3. 2015 Workshop on Management, Analysis, and Visualization of Experimental and Observational data – The Convergence of Data and Computing



Previous projections comparing rate of growth based on formula (left) compared to projections based on detector characteristics (above).

History of Raw* Detector Data Rates at NCEM/Foundry

*may not be sustained to disk



4D STEM
200 TB/hr

K2-IS
46 TB/hr

TEAM detector
3 TB/hr

Direct Detection

CCD/phosphor
28 GB/hr
TEAM/Titan

CCD/phosphor
4 GB/hr
CM300

Film
1 GB/hr
AEM

Indirect Detection

Slide Source:
P. Denes
(LBNL)

Global View of Problem Space: Data Lifecycle

ALS Use Case: Experiment Planning and Optimization

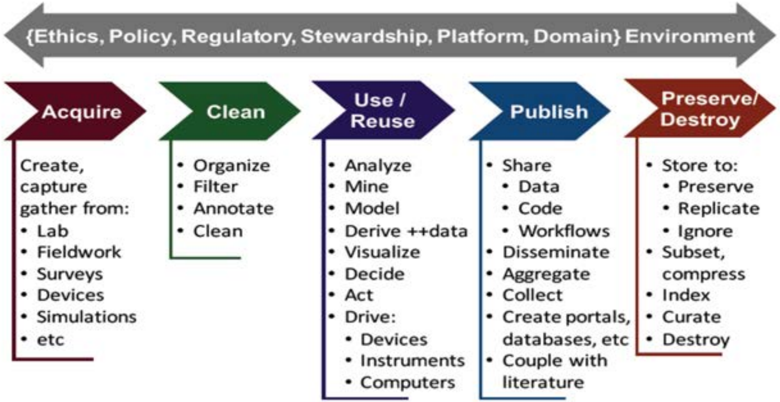
What are the best settings for acquisition of a given material sample on a given beamline?

How to adjust, optimize an experiment in progress?

Recommender Systems

Require access to and use of curated collections of experimental data: *training data* for ML methods

Collection of curated and trained models, V&V



Source: NSF 2016 report: Realizing the Potential of Data Science

Curated Data Collections

Require metadata "standards", metadata and data models/formats

Methodology for collecting, managing metadata and scientific data

Long-term storage, dissemination

Math, CS, Data Science Innovations

Individual methods: eg., search, see, analyze, store, share

System view: combinations of methods

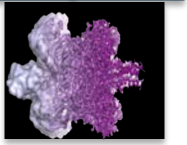
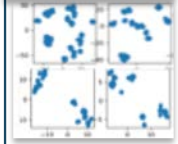
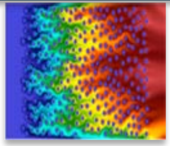
Platform portability and performance

Sustainability

Community adoption and use

Software engineering practices

Systems, Networks, Methods, and Services for Complex Workflows across the Data Lifecycle

Acquire	Transfer	Clean	Use/Reuse	Publish	Preserve
					
Collect from sensors, experiments, simulations	Move from instrument to center	Organize, annotate, filter, encrypt, compress	Analyze, mine, model, learn, infer, derive, predict	Disseminate, aggregate, using portals, databases	Index, curate, age, track provenance, purge

Edge	Network	Processing	Analytics	Distribution	Management
• Co-design detectors and analysis • In-situ analysis for simulation and experiment • Robotics	• Networking "beyond Moore" • Autonomous networks • Named data networking	• Feature selection • Domain-specific compression • Metadata learning • QA/QC	• Statistical learning methods • Type-specific analytics • Simulate / invert • Scalable software	• Scientific databases • Directories and search • User interfaces • Availability • Data integrity	• Fast indexing and data management • Review and prioritize • Social models

Slide Source: K. Yelick (LBNL)

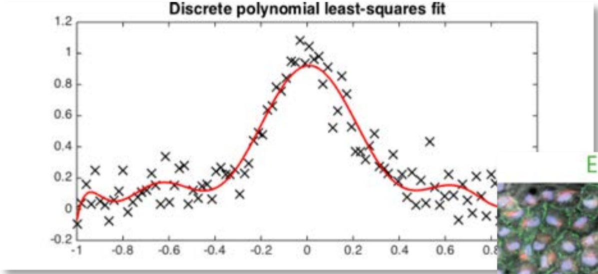
Data: Product or Source?

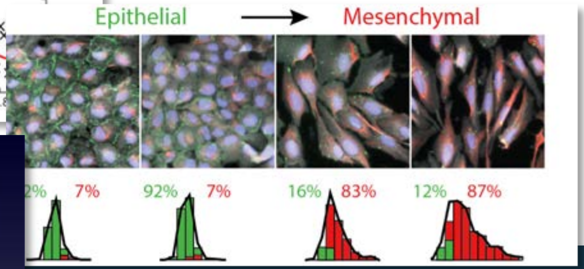
Modeling/simulation:
Solution to equations
produces data.

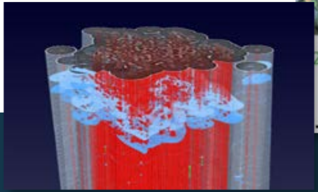
Navier-Stokes momentum equation *(convective form)*

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = -\frac{1}{\rho}\nabla \bar{p} + \nu \nabla^2 \mathbf{u} + \tfrac{1}{3}\nu \nabla(\nabla \cdot \mathbf{u}) + \mathbf{g}.$$

Data Analytics,
Learning:
From data, derive a
model, model parms,
quantitative
information

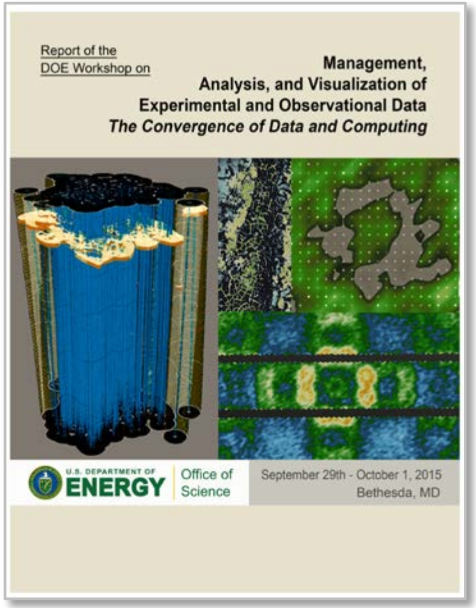






Diversity of Data and Use Cases, Diversity of Challenges

- **Workflows: data+processing pipelines**
  - For accommodating production science
  - Time-sensitive data movement, computations
  - “Fractal” in nature, unbiquitous
- **Curated data collections**
  - How to capture provenance
  - Next-generation search methods
- **New types of methods, approaches**
  - Supervised learning: requires lots of high quality training data
  - Computations, analysis that rely on multi-modal data
- **New types of operational use**
  - Experiment planning
  - Experiment optimization (time sensitive)



2015 EOD Workshop Report – Key Findings

- **Challenges:**
  - All EOS projects struggle with a flood of data
  - EOS projects have unmet, time-critical data needs
  - There is a risk of EOS data being unusable
  - Collaboration and sharing are central to EOS projects
  - EOS data lifecycle needs not being met
  - Software plays a central role in all EOS projects
  - Workforce development, retention concerns

- **Opportunities:**
  - Data reuse: new science after initial experiment
  - Faster science
  - Better science
  - Cost savings: economy of scale