ASCR WORKSHOP ON IN SITU DATA MANAGEMENT

Enabling Scientific Discovery
from Diverse Data Sources

January 28–29, 2019

Tom Peterka
tpeterka@mcs.anl.gov
Argonne National Laboratory
Mathematics and Computer Science Division

DOI: 10.2172/1493245

U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

Workshop Charge

Seek community input on the development of in situ capabilities for managing the execution and data flow among a wide variety of coordinated tasks for scientific computing.

Definition of In Situ Data Management (ISDM)

The practices, capabilities, and procedures to control the organization of data and enable the coordination and communication among heterogeneous tasks, executing simultaneously in an HPC system, cooperating toward a common objective.

Definition of In Situ Data Management (ISDM)

The practices, capabilities, and procedures to control the organization of data and enable the coordination and communication among heterogeneous tasks, executing simultaneously in an HPC system, cooperating toward a common objective.

Why In Situ?

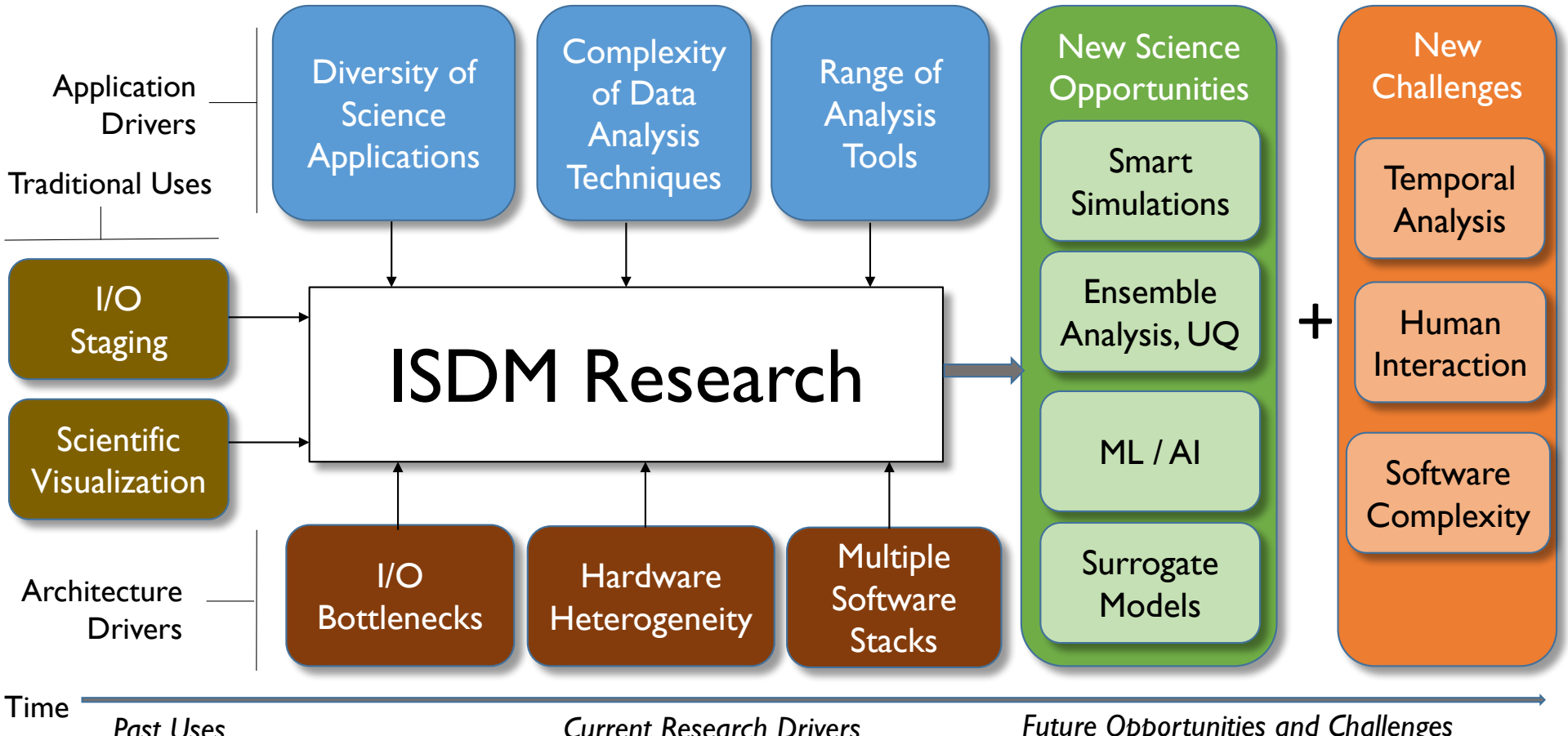
- ISDM can make critical contributions to managing and reducing large data volumes from computations and experiments.

  Successful ISDM can minimize data movement, save storage space, and boost resource efficiency—often while simultaneously increasing scientific precision.

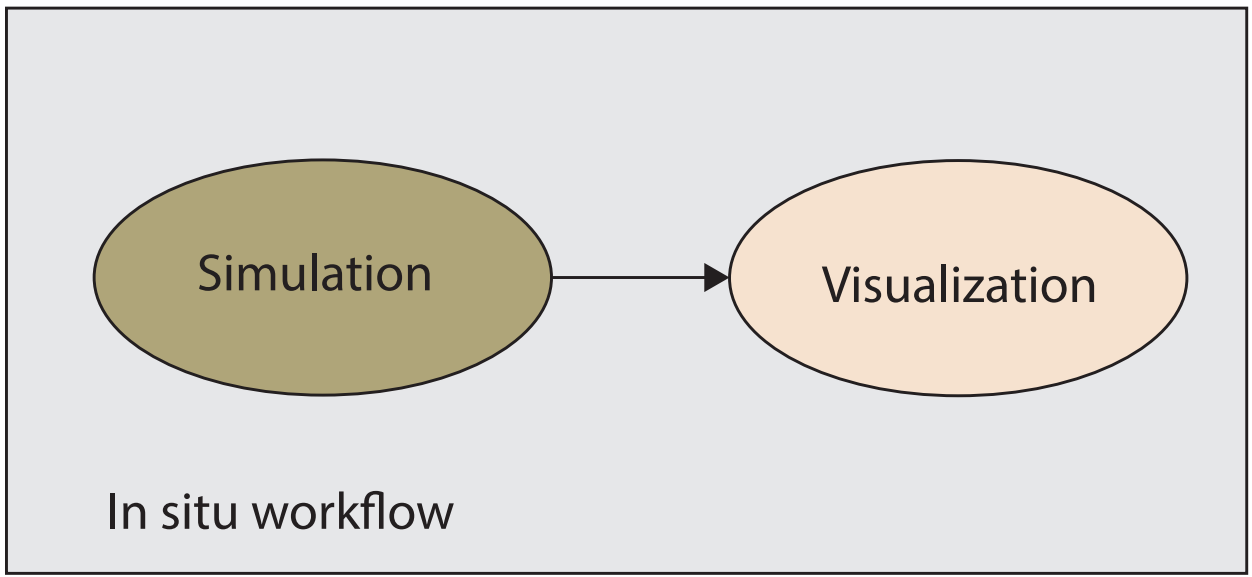
- The in situ methodology enables scientific discovery from a broad range of data sources, over a wide scale of computing platforms.

  Successful ISDM will benefit real-time decision making, design optimization, and data-driven scientific discovery.

Overview

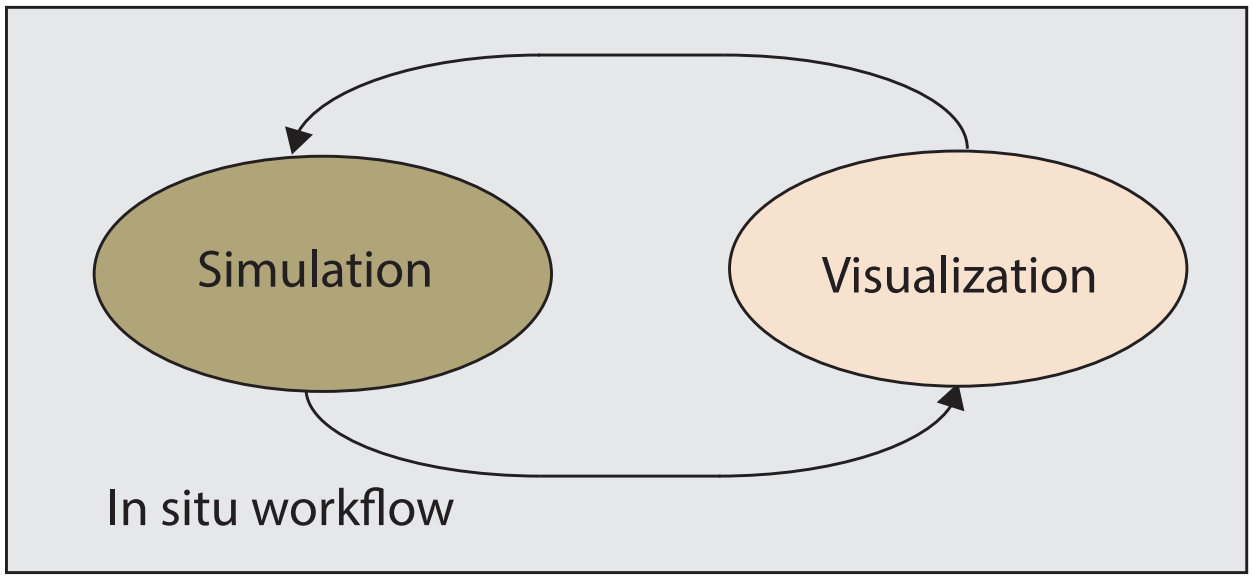

Application Drivers
Traditional Uses
Diversity of Science Applications
Complexity of Data Analysis Techniques
Range of Analysis Tools
I/O Staging
Scientific Visualization
ISDM Research
Architecture Drivers
I/O Bottlenecks
Hardware Heterogeneity
Multiple Software Stacks
New Science Opportunities
Smart Simulations
Ensemble Analysis, UQ
ML / AI
Surrogate Models
+
New Challenges
Temporal Analysis
Human Interaction
Software Complexity
Time
Past Uses
Current Research Drivers
Future Opportunities and Challenges

In Situ Yesterday

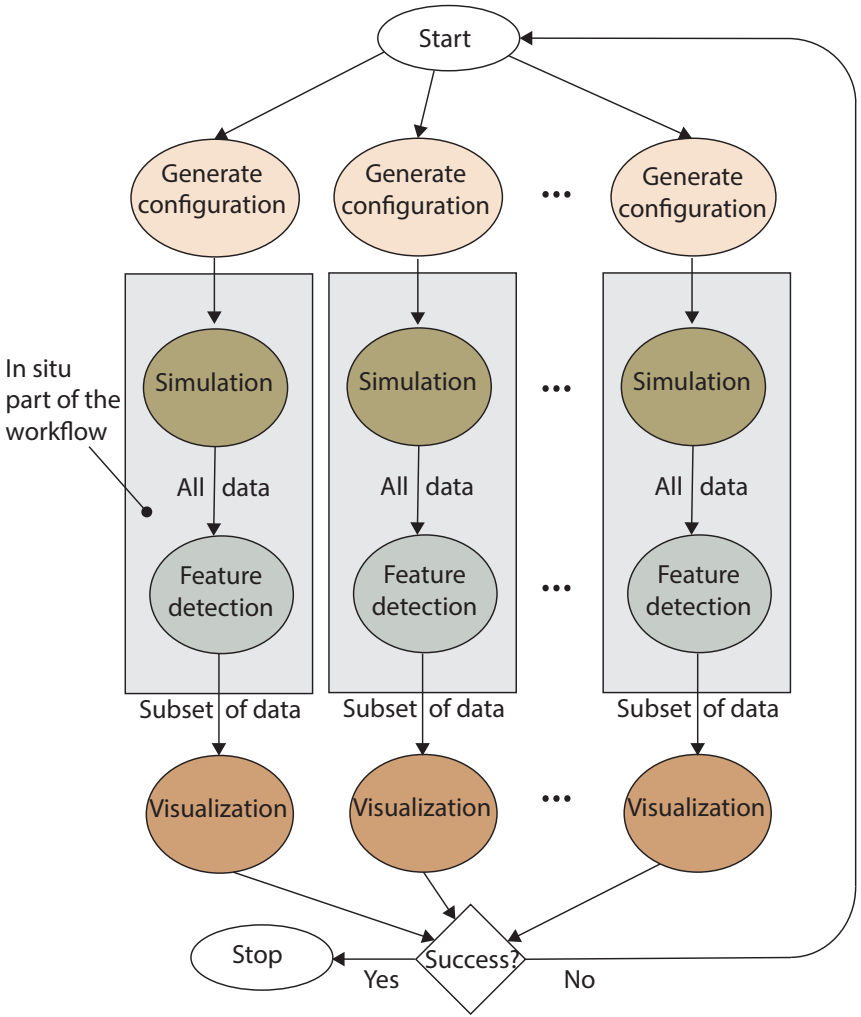


[Zajac, 1964]

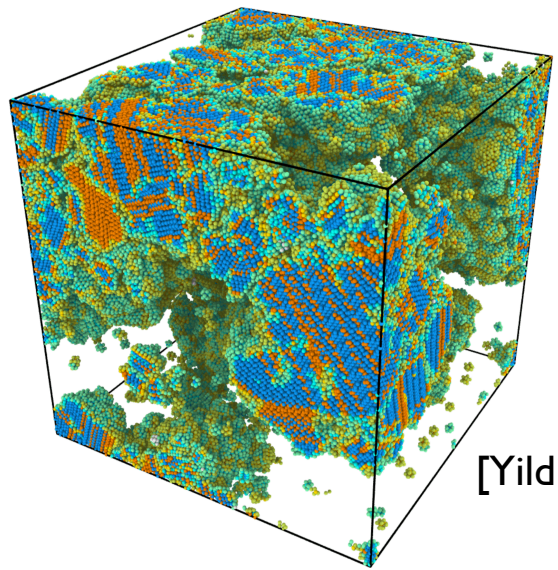
In Situ Yesterday



[Parker et al., 1995]



In Situ Today



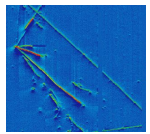
[Yildiz et al., 2019]

BES workflow of dynamic ensemble of simulations and in situ detection of stochastic events

One Vision of Future ISDM

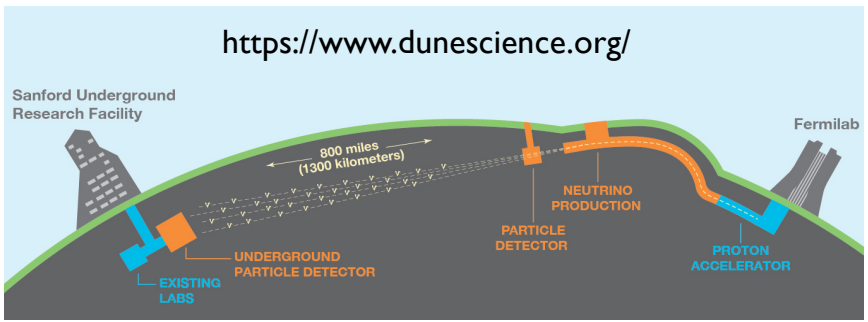
Sanford

In situ detection / processing

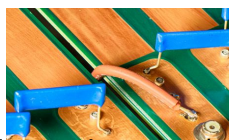


Filtered /compressed data

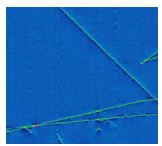
https://www.dunescience.org/



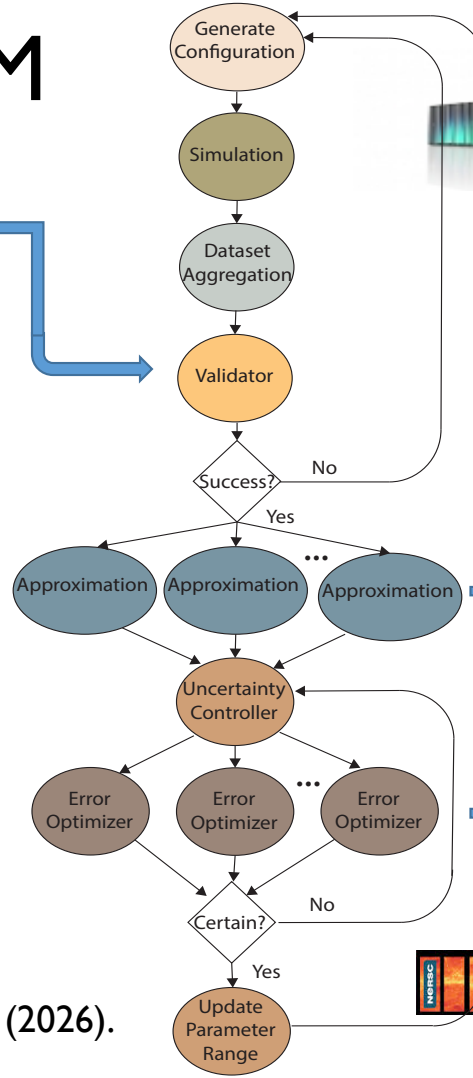
FermiLab



In situ detection / processing

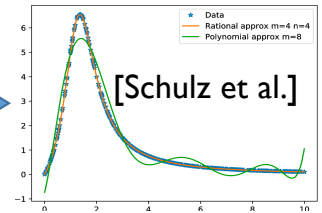


Filtered /compressed data

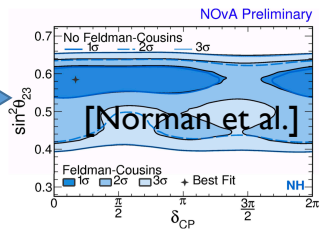


https://www.alcf.anl.gov/

ALCF



Surrogate model for simulated events



Feldman-Cousins correction of simulated and measured events

https://www.nersc.gov/

NERSC

Neutrino event generation and parameter optimization for DUNE (2026).

ASCR In Situ Data Management (ISDM) Workshop

Organizing Committee and Logistics

Name	Affiliation	Role
Tom Peterka	ANL	Chair
Debbie Bard	NERSC	Organizer
Janine Bennett	SNL	Organizer
Wes Bethel	LBNL	Organizer
Ron Oldfield	SNL	Organizer
Line Pouchard	BNL	Organizer
Christine Sweeney	LANL	Organizer
Matthew Wolf	ORNL	Organizer
Laura Biven	DOE-ASCR	Program Manager

Date	Location
Jan. 28-29, 2019	Bethesda North Marriott, Rockville, MD

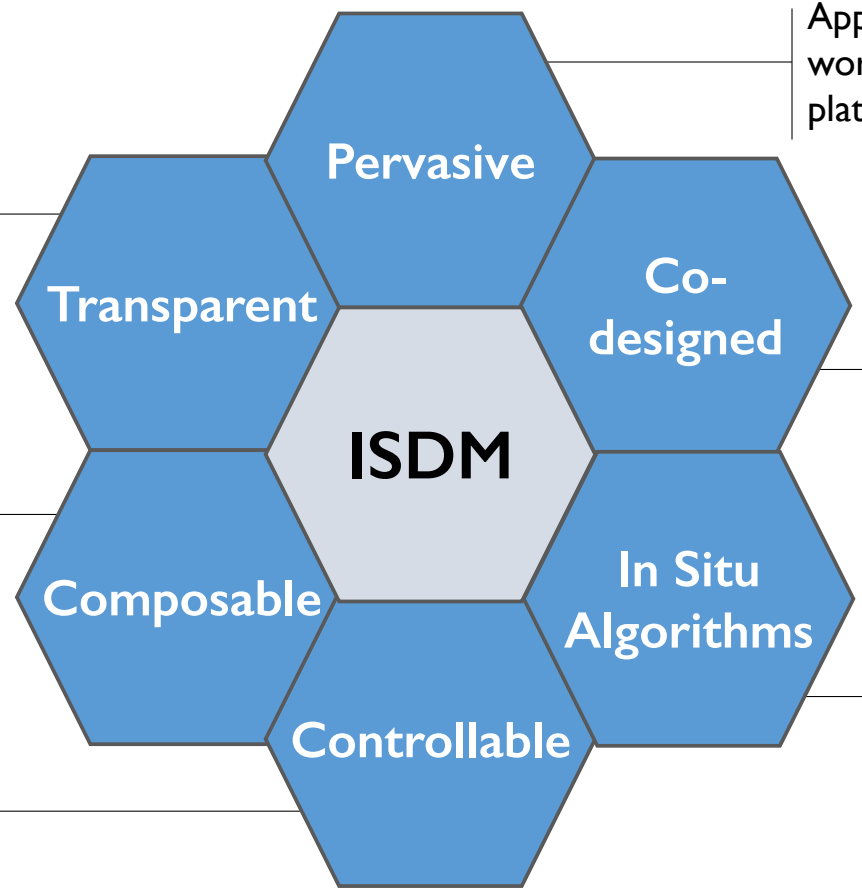
Abstract

This workshop seeks community input on the development of in situ capabilities for managing the execution and data flow among a wide variety of coordinated tasks for scientific computing. The workshop considers ISDM in addition to the traditional roles of accelerating simulation I/O and visualizing simulation results, to more broadly support future scientific computing needs. In particular, the convergence of simulation, data analysis, and artificial intelligence will require machine learning, data manipulation, creation of data products, assimilation of experimental and observational data, analysis across ensemble members, and, eventually the incorporation of tasks on non-von Neumann architecture.

PRDs

Priority Research Directions

Pervasive, controllable, composable, and transparent ISDM, co-designed with the software stack and with fundamentally new algorithms.



Increase confidence in reproducible science, repeatable performance, and feature discovery through provenance.

Apply ISDM and in situ workflows at a variety of platforms and scales.

Develop interoperable ISDM components for agile and sustainable programming.

Coordinate ISDM development with underlying system software.

Understand the design space of autonomous decision-making and control of in situ workflows.

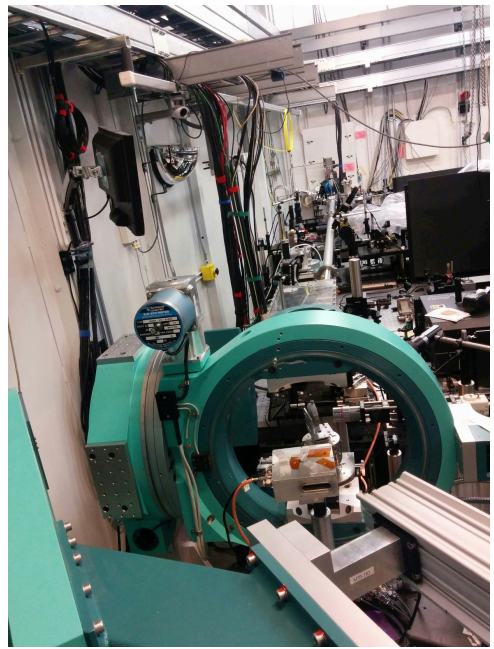
Redesign analysis algorithms for the in situ paradigm.

How can ISDM methodologies help meet the needs for real-time, high-velocity data applications at the edge and other non-high-performance computing platforms? How can ISDM enable science at experimental and observational facilities?

Pervasive ISDM

Apply ISDM methodologies and in situ workflows at a variety of platforms and scales.

A changing landscape of use cases is driving new applications of ISDM. The ability to execute the same ISDM tasks and workflows across a spectrum of computational platforms, spanning high-performance supercomputers to experimental detectors and even embedded devices, will reduce human effort and improve portability by applying consistent computing methods.



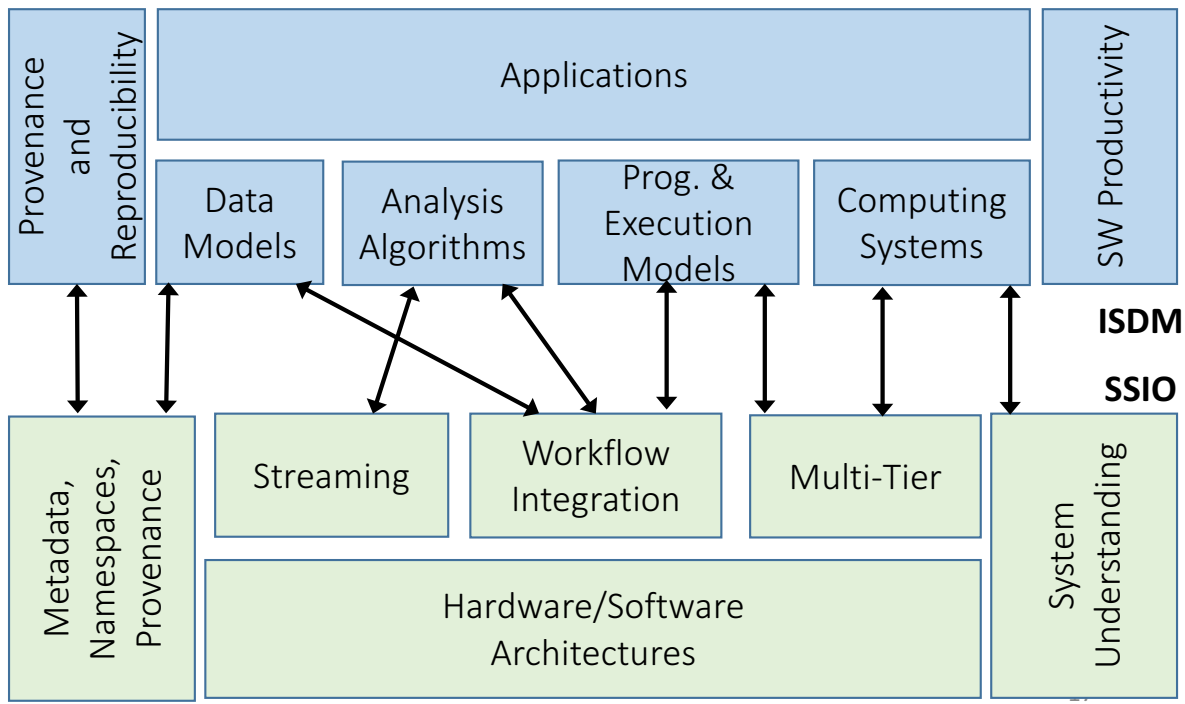
Experimental apparatus at Argonne Advanced Photon Source Sector 7. Diffraction images can be reconstructed while experiments are ongoing.

What abstractions, assumptions, and dependencies on system services are needed by ISDM? What information must be exchanged between the ISDM tools and the rest of the computing software stack to maximize performance and efficiency?

Co-designed ISDM

Coordinate the development of ISDM with the underlying system software so that it is part of the software stack.

Understanding the interlayer dependencies so that ISDM becomes part of the software stack can facilitate connections between software layers, communicate semantic meaning, and realize efficient performance in high-performance computing and other software stacks.

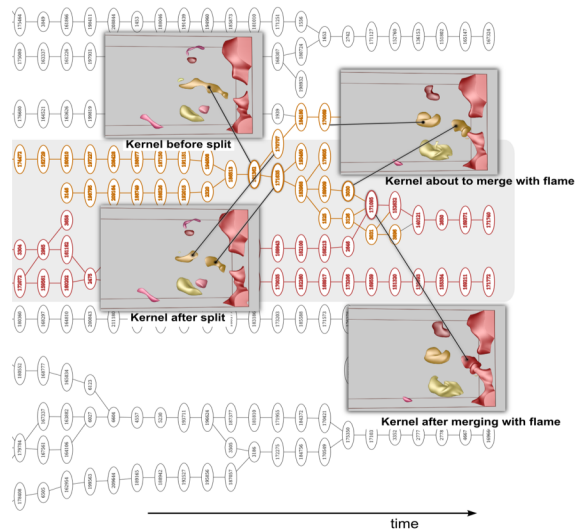


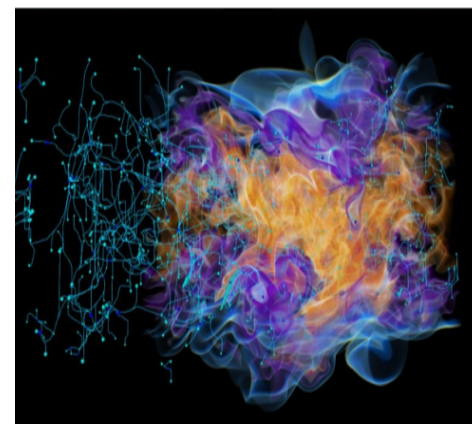
How should in situ algorithms be designed to make the most of the available resources? What new classes of data transformations can profit from in situ data access in the presence of constraints imposed by other tasks?

In Situ Algorithms

Redesign data analysis algorithms for the in situ paradigm.

The in situ environment for data processing and analysis differs substantially from the post hoc environment, requiring fundamentally new algorithms and approaches. Progress will benefit from multidisciplinary approaches that holistically consider the opportunities, constraints, and user needs of in situ analysis.





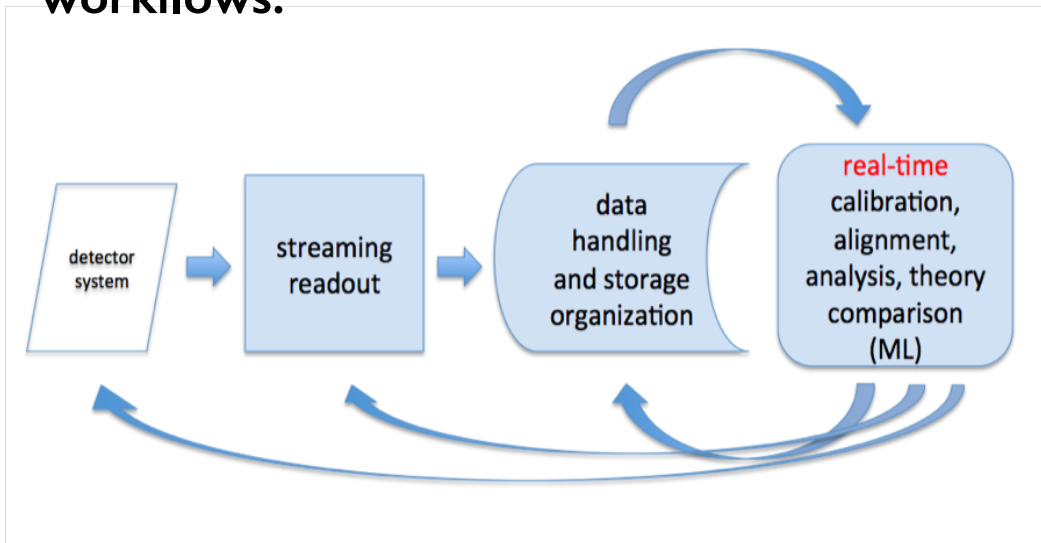
In situ topological feature detection in turbulent combustion simulations used to segment and track localized intermittent ignition and extinction features. (images courtesy of J.H. Chen).

What metrics best describe the ISDM design space? How can that space be defined, codified, and evaluated to support design decision-making and control?

Controllable ISDM

Understand the design space of autonomous decision-making and control of in situ workflows.

Understanding the space of ISDM parameters is crucial to making intelligent design decisions, both by humans and autonomously. The capability to optimize a constrained ISDM design space will enable predictable performance and scientific validity. Design metrics will promote knowledge sharing across communities.



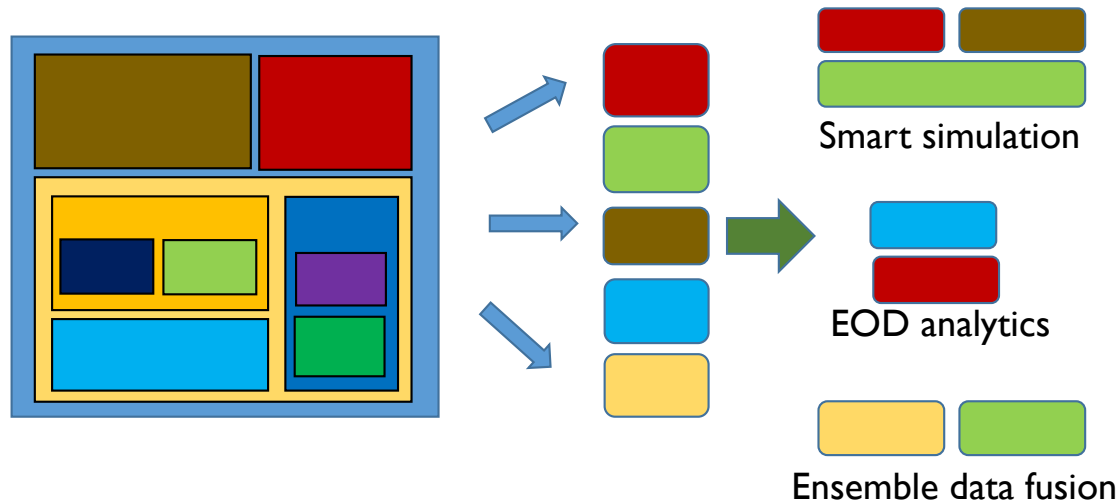
Model of how information flows for experimental computing, illustrating how real-time data analysis is required to guide the detector system, readout system, and data handling (image courtesy of Amber Boehnlein).

Can the composition of ISDM software components maximize programmer productivity and usability? What design decisions of ISDM software components promote their interoperability in order to ensure the long-term utility of ISDM software for the science community?

Composable ISDM

Develop interoperable ISDM components and capabilities for an agile and sustainable programming paradigm.

The flexible composition of interoperable ISDM software components will enable developers and end users to choose from an array of widely available tools, thereby increasing productivity, portability, and usability, and will ultimately result in agile and reusable software.



How can provenance and metadata support data discoverability, reuse, and reproducibility of results? How can these artifacts be captured automatically and analyzed in situ, at the scale of DOE science?

Transparent ISDM

Increase confidence in reproducible science, deliver repeatable performance, and discover new data features through the provenance of ISDM.

In situ provenance and metadata are crucial to understanding scientific results, assessing correctness, and connecting underlying models and algorithms with workflow execution. The ability to capture and query provenance and metadata at scale and in situ will enable many diverse science needs.

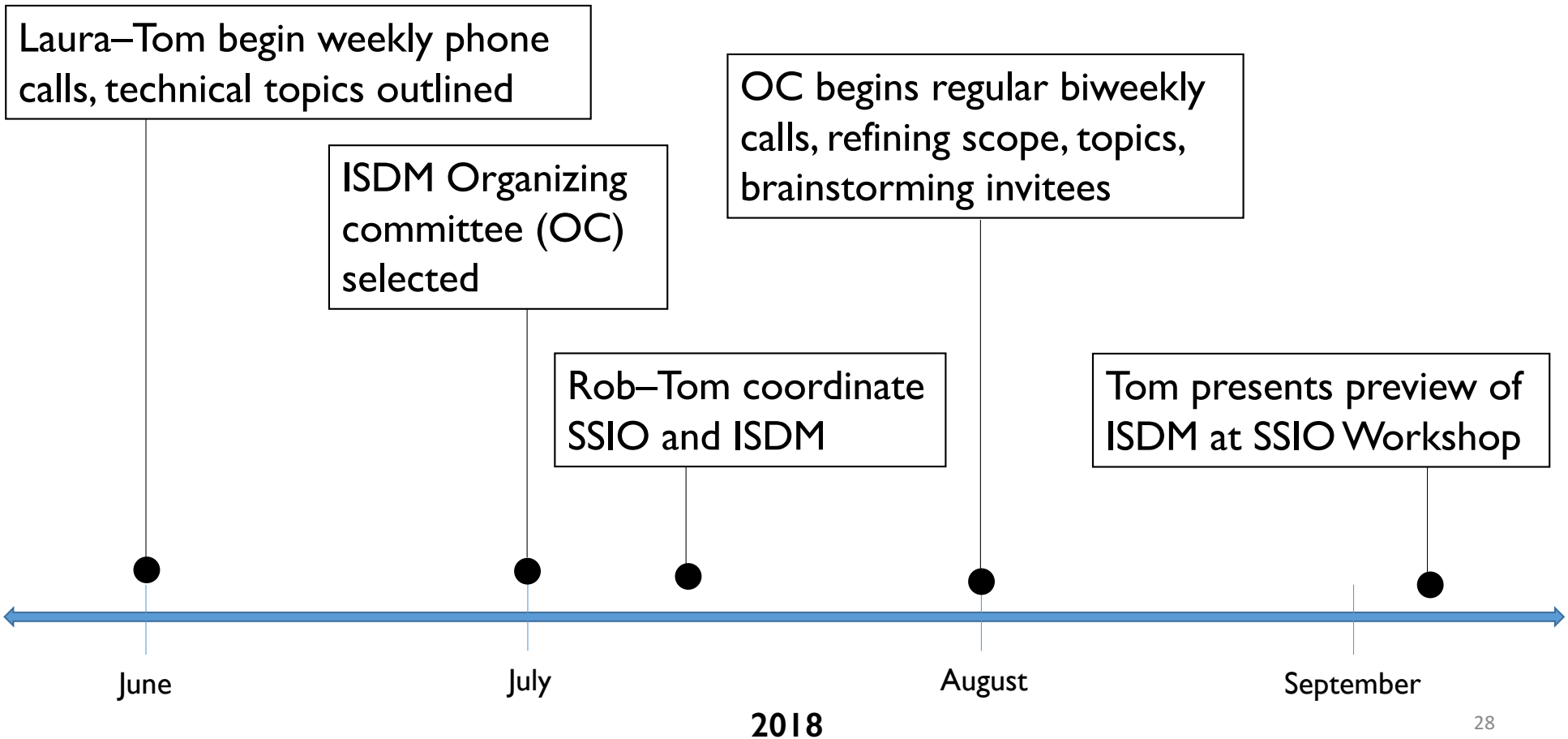
Performance provenance for NWChem computational chemistry simulation (image courtesy of Huub Van Dam, Wei Xu, Cong Xie, and Wonyong Jeong).

Process

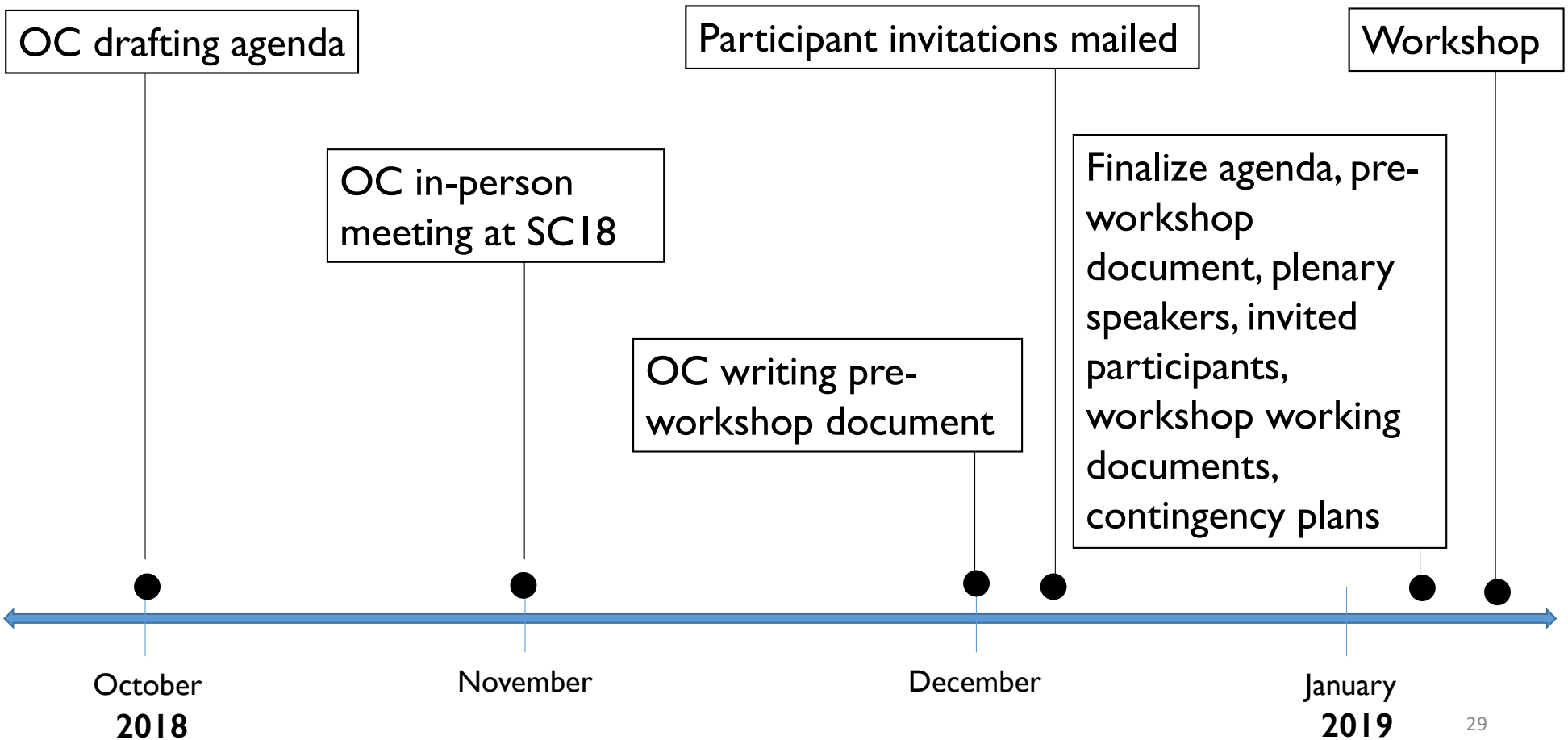
Breakout Session Research Areas	PRDs					
	Pervasive ISDM	Co-designed ISDM	In Situ Algorithms	Controllable ISDM	Composable ISDM	Transparent ISDM
DATA MODELS						
Multimodal data science and ML	X		X	X		X
Intent expression and reuse	X	X	X	X	X	
COMPUTATIONAL PLATFORMS						
System software support		X				X
Heterogeneous hardware		X				
ANALYSIS ALGORITHMS						
Reduced representations	X		X			
Run-time control		X	X	X		
New platforms and outputs	X	X	X			X
PROVENANCE AND REPRODUCIBILITY						
Scalable, portable provenance capture	X	X				X
In situ provenance processing			X			X
Provenance for ML and reproducibility					X	X
PROGRAMMING AND EXECUTION MODELS						
Elastic, dynamic resources		X		X		
Scheduling and optimizing execution		X		X		X
Composable workflows	X	X	X		X	
Streaming	X		X			X
SOFTWARE ARCHITECTURE						
Use cases driving design	X				X	
Usability and sustainability					X	
Software tool interoperability					X	
User confidence					X	
Science facilities partnerships	X				X	

Getting from Workshop Topics to PRDs

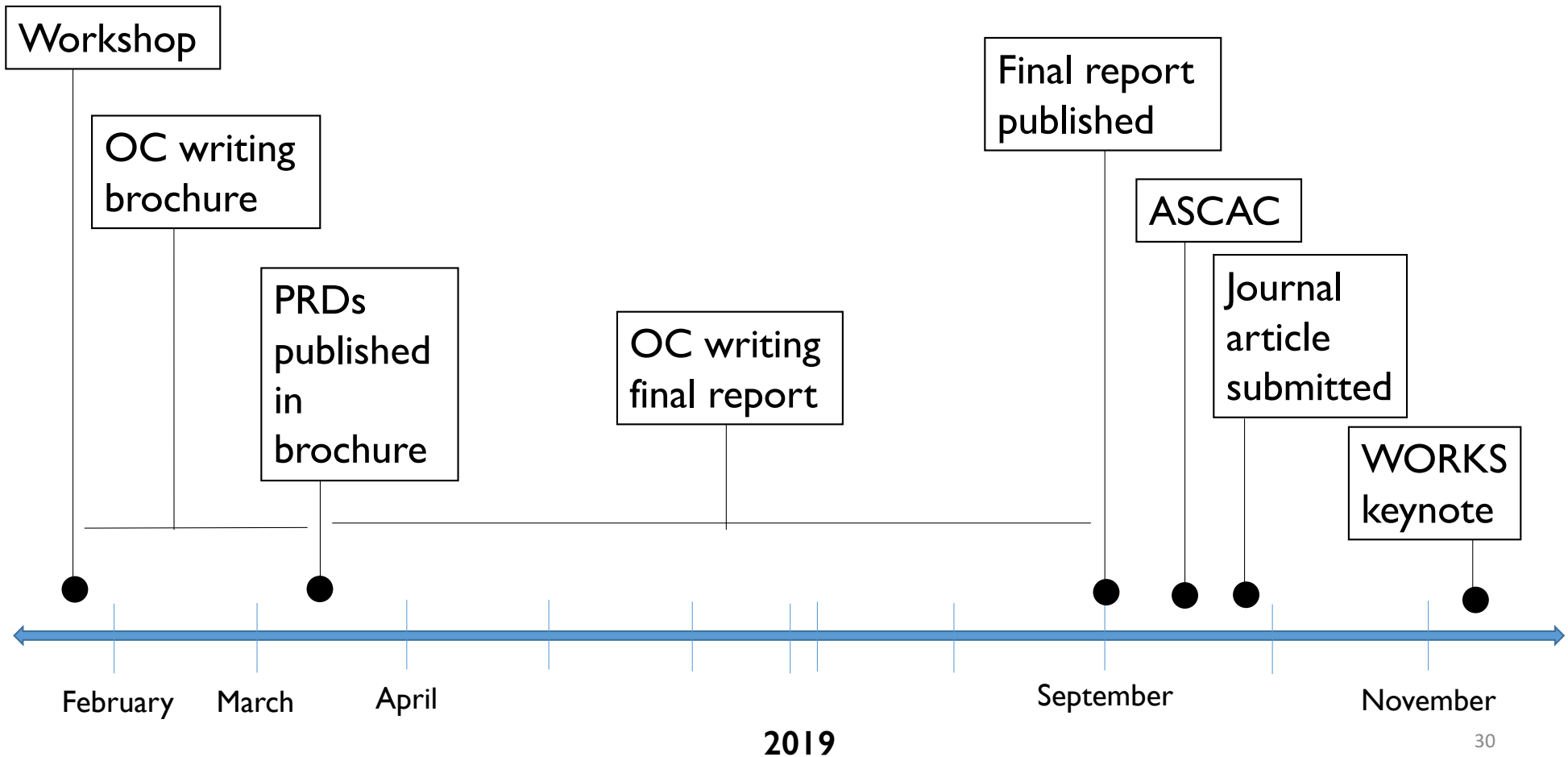
How it all Started…



Leading up to the Workshop



After the Workshop



At the Workshop

Science Applications

Interface to applications and science workflows

Data Models: Connection and Communication

Structure, semantics, and movement of in situ data

Computational Platforms and Environments

Interface to hardware and system software stacks and future platforms

Analysis Algorithms

Portable, high-performance algorithms that can be used in situ and elsewhere

Provenance and Reproducibility

Information for diagnostics, performance studies, and scientific reproducibility

Programming and Execution Models

Programming and executing disparate constituent tasks in an ISDM framework

Software Architecture for Usability and Sustainability

Software that can be built, deployed, sustained, and used to support DOE science

Breakout Session: 4 Parts

Part 1: Preview	Part 2: Divergent / Ideation	Part 3: Synthesis / Prioritization	Part 4: Convergent / Problem Solving
Review topic page from pre-workshop document	Breadth-first generation of challenges / opportunities and why they are important (impact)	Synthesize candidate PRD titles and statements from challenges / opportunities	Develop PRD details
10 minutes	45 minutes	15 minutes	15 minutes

1.5 hours

Data Capture

Key Challenges and Opportunities

Please describe the underlying science challenges and opportunities that motivate this PRD

State of the art

Please answer the following questions:

- Who else is doing this?
- What are the technology and research gaps?

New Research Direction

Please answer the following questions:

- What will you do to address the challenge?
- What research questions will you ask / answer?
- What are the potential risks?
- What would success look like?
- What assumptions about users, hardware, or other parts of the software stack motivate this as a priority / are required for success?

Potential Scientific Impact

Please answer the following questions:

- What new scientific capabilities will follow?
- What new methods and techniques will be developed?

Data Capture

Key Challenges and Opportunities

Please describe the underlying science challenges and opportunities that motivate this PRD

State of the art

Please answer the following questions:

- Who else is doing this?
- What are the technology and research gaps?

New Research Direction

Please answer the following questions:

- What will you do to address the challenge?
- What research questions will you ask / answer?
- What are the potential risks?
- What would success look like?
- What assumptions about users, hardware, or other parts of the software stack motivate this as a priority / are required for success?

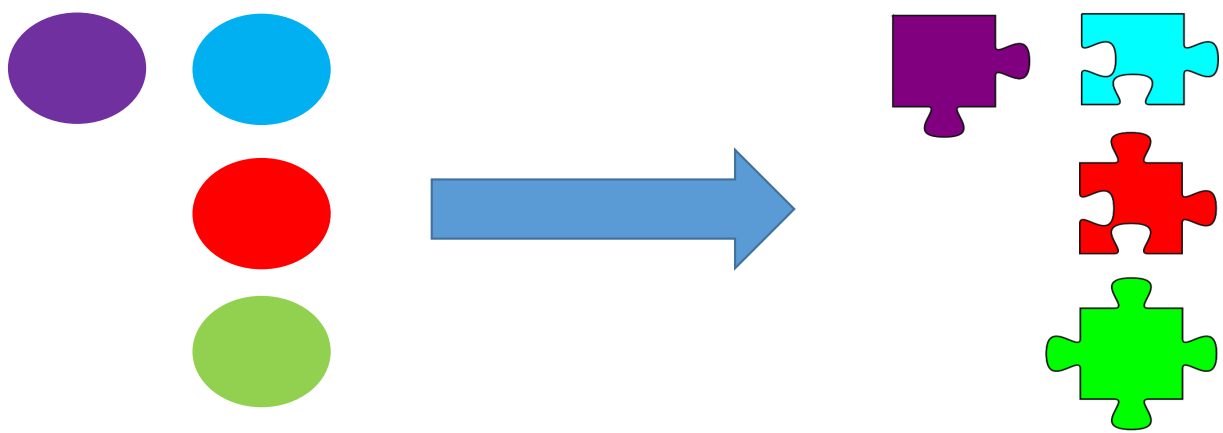
Potential Scientific Impact

Please answer the following questions:

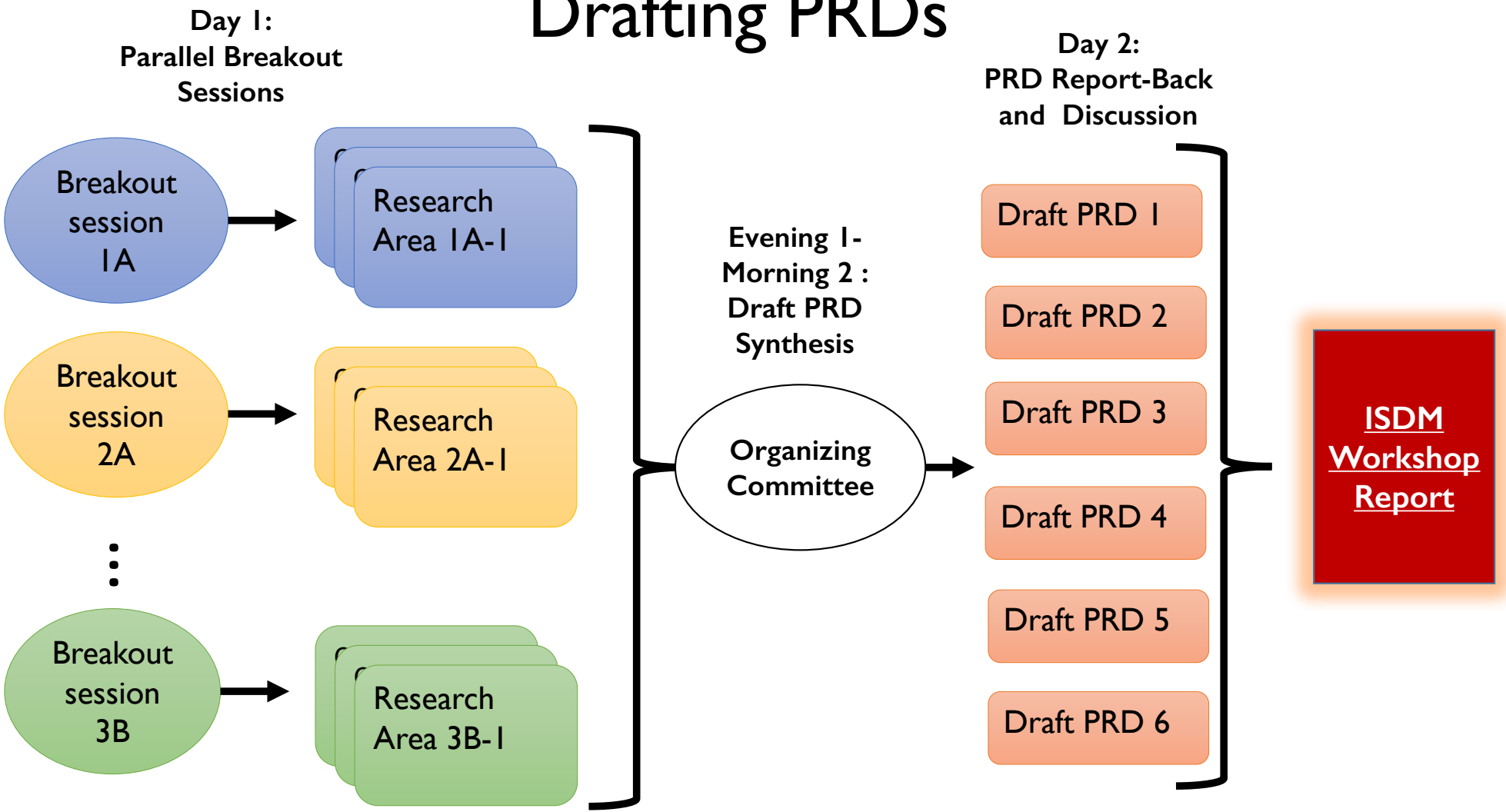
- What new scientific capabilities will follow?
- What new methods and techniques will be developed?

Assumptions and Dependencies Matter

- Track the consequence of assumptions back to priorities if/when assumptions change.
- Follow relationships between parts of the research portfolio.
- Promotes a software stack view of the portfolio.
- Components of the portfolio can work together to achieve capabilities.



Drafting PRDs



Resources

Brochure (4 pages)

Full report (100 pages)

https://science.osti.gov/ascr/Community-Resources/Program-Documents

DOI: 10.2172/1493245

Thank You

Organizing Committee

Debbie Bard, Janine Bennett, Wes Bethel, Ron Oldfield, Line Pouchard, Christine Sweeney, and Matthew Wolf

Program Manager

Laura Biven

Workshop Participants

Jim Ahrens, Ann Almgren, Katie Antypas, Edmun Begoli, Amber Boehnlein, Ron Brightwell, Jacqueline Chen, Hank Childs, Warren Davis, Marc Day, Jai Dayal, Ewa Deelman, Lei Ding, Nicola Ferrier, Fernanda Foertter, Berk Geveci, Katrin Heitmann, Bruce Hendrickson, Jay Hnilo, Adolphy Hoisie, Dan Jacobson, Shantenu Jha, Ming Jiang, Kirk Jordan, Scott Klasky, Kerstin Kleese van Dam, Eric Lancon, Earl Lawrence, Steve Legensky, Jay Lofstead, Andrew Lumsdaine, Kwan-Liu Ma, Pat McCormick, Ken Moreland, Dmitriy Morozov, H. Sarp Oral, Manish Parashar, Valerio Pascucci, Amedeo Perazzo, Beth Plale, Thomas Proffen, Mike Ringenburg, Silvio Rizzi, Rob Ross, Tim Scheibe, David Schissel, Malachi Schram, Katie Schuman, Nicholas Schwarz, Han-Wei Shen, Eric Stephan, Victoria Stodden, Michela Taufer, Justin Wozniak, John Wu, and Hongfeng Yu.