Report from Dagstuhl Seminar 23132

AI-Augmented Facilities: Bridging Experiment and Simulation with ML

Peer-Timo Bremer[*1], **Brian Spears**[*2], **Tom Gibbs**[*3], **and Michael Bussmann**[*4]

1 **Lawrence Livermore National Laboratory, US.** bremer5@llnl.gov
2 **Lawrence Livermore National Laboratory, US.** spears9@llnl.gov
3 **Nvidia – Santa Clara, US.** tgibbs@nvidia.com
4 **Helmholtz-Zentrum Dresden-Rossendorf, DE.** m.bussmann@hzdr.de

Abstract

In the last week of March 2023, Schloss Dagstuhl hosted a Dagstuhl Seminar on “AI-Augmented Facilities: Bridging Experiment and Simulation with ML”. The seminar brought together experimental and computational scientists, experts on edge and HPC computing, and machine learning and computer science researchers to jointly develop a strategic vision on how to move towards AI-augmented facilities in a unified manner. The goal was to suggest a common research agenda with an emphasis on areas where joint efforts are needed for future progress. Starting with some overarching perspectives the seminar was dominated by lively discussions that resulted in a strategic write-up to be published separately.

Seminar March 26–31, 2023 – https://www.dagstuhl.de/23132

2012 ACM Subject Classification Applied computing → Physical sciences and engineering; Computing methodologies → Machine learning; Computer systems organization → Embedded and cyber-physical systems; Mathematics of computing → Probability and statistics; Information systems → Data management systems

Keywords and phrases Self-driving experiments, Smart facilities, AI models, Multimodal data, Accelerated scientific discovery, AI software stack for experiments

Digital Object Identifier 10.4230/DagRep.13.3.106

1 Executive Summary

Peer-Timo Bremer (Lawrence Livermore National Laboratory, US)
Brian Spears (Lawrence Livermore National Laboratory, US)
Tom Gibbs (Nvidia – Santa Clara, US)
Michael Bussmann (Helmholtz-Zentrum Dresden-Rossendorf, DE)

The Dagstuhl Seminar connected three traditionally different communities: experimental and computational scientists, experts in HPC and edge computing, and machine learning researchers, to discuss a new vision for future AI-augmented facilities. This document summarizes the activities during the week of in person discussion including the outline of a position paper that is under development to publish the joined findings. The seminar proceeded in roughly three stages: an introduction with two keynotes and a general discussion

[*] Editor / Organizer

AI-Augmented Facilities: Bridging Experiment and Simulation with ML, *Dagstuhl Reports*, Vol. 13, Issue 3, pp. 106–131
Editors: Peer-Timo Bremer, Brian Spears, Tom Gibbs, and Michael Bussmann
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

on the goals, an expansive phase of collecting ideas and defining the scope of the position paper, and finally working groups on creating explicit outlines and collecting materials for various sections of the paper.

2 Table of Contents

3 Keynotes and Topic Introduction

The week started with a session to introduce all participants with a little bit of their background to facilitate later discussions and provide an overview of the available expertise. This was followed by two introductory keynotes from the organizers briefly describing the current state of affairs in AI-augmented facilities from both the US (Brian Spears) and the EU (Michael Bussmann) perspective which consumed Monday morning.

3.1 Facilities & AI a US Perspective

Brian Spears (Lawrence Livermore National Laboratory, US)

The pace of data generation in modern science has greatly accelerated, but the pace of transformational discovery is still too slow. This is clear at a variety of state-of-the-art facilities: laser experiments are slow or noisy; advanced manufacturing (AM) is open loop; accelerators need time consuming tuning, sometimes by hand. However, AI-enable self-driving systems can accelerate our science and discovery processes. AI sentinels that help collect data, compare to prediction, and choose next steps can provide a step change in experimental and manufacturing operations. They will bring accelerated closed-loop operations, transformational data rates, physics-informed experiment updates on sub-second timescales, and digital twins for facility modeling and optimization. This not only accelerates discovery, but it deepens the quality of knowledge that we can discover. As examples, it will provide stabilized lasers and autonomous optimization of high-energy-density physics, self-correcting AM processes and high-throughput operations, and repeatable, robust accelerator conditions. Beyond individual systems, self-driving ecosystems composed of interconnected sets of these facilities will offer capabilities greater than the sum of their parts offering rapid discoveries that are hard to conceive in today's slower and isolated science regime. To achieve this goal, the science community needs to work together to build the scientific tools to execute self-driving operations. With community input, like that provided by Dagstuhl, we can make self-driving science systems a reality.

3.2 Facilities & AI an EU Perspective

Michael Bussmann (Helmholtz-Zentrum Dresden-Rossendorf, DE)

In the EU, the use of AI at large-scale research infrastructures is on the rise in a broad variety of fields from Particle Physics to Photon Science, Neutron Science, Life Science, Astrophysics to Laser Science and more. The ESFRI Roadmap and the EU digitalization strategy highlight the importance of data and meta data and the potential of AI. Focusing on the example of Germany, We highlight how the Helmholtz Association as the largest research organization in Europe and in particular the Helmholtz Research Field Matter plan to develop autonomous, intelligent facilities using AI at key points in the data lifecycle of research facilities. The topic Data Management & Analysis and the Helmholtz Incubator for Information and Data Science play key roles, looking at such diverse topics as data lifecycle management, the tight integration of simulation, experiments and machines, online and large-scale data analysis, visual analytics, optimization, automation and resilience. Embedded in a national AI strategy with key components such as the National Research Data Infrastructure, ErUM-Data and many more, embedded in EU-wide and international cooperations and initiatives, the landscape of AI-augmented facilities is being shaped into a EU-wide, cross-community effort to enable excellent science at optimum conditions across the whole spectrum of research infrastructures. We argue that this can be a blueprint for international collaboration on AI-augmented facilities.

4 Working Groups Results

Starting Monday afternoon the seminar switched to a mixture of interactive working groups followed by sessions to report the results and plan the next agenda items. The sections below will briefly summarize the individual sessions including (subsets of) the raw notes when possible and list or participants where available.

4.1 Monday Afternoon

In two closely related sessions on Monday afternoon all seminar participants first collected a list of prototypical science drivers that motivate the need for AI-augmented facilities. Subsequently, this discussion branched out to list stakeholders and specific applications that could be used later as examples. Finally, the discussion converged on defining more the goal of the seminar more explicitly: To collect the insights, existing solutions, and strategic ideas into a perspective paper to be jointly published. Consequently, the remainder of Monday afternoon was spend creating a first paper outline that simultaneously served as a guide for the schedule on Tuesday morning. An overview of the topics discussed is provided in Figure 1 in form of a topic graph.

4.2 Tuesday Morning

Following the initial paper outline different groups in parallel started to flesh out individual sections. This started with three parallel breakouts on the overall needs, the approach, and the expected outcomes as the cornerstone of the paper.

4.2.1 Group 1: Needs

Working Group 1 was tasked to explore the need for AI augmented facilities in more detail to ultimately serve as the motivation for the perspective. Figure 2 documents some of the notes including a conceptual diagram of the state-of-the-art created as straw-man for the discussion. The list of high level needs collected during the outbrief included:

- Faster Science
- Better Science / per $ or €
- Data Interpretation / optimum operation
- Optimised sciences
- Addressing grand challenges

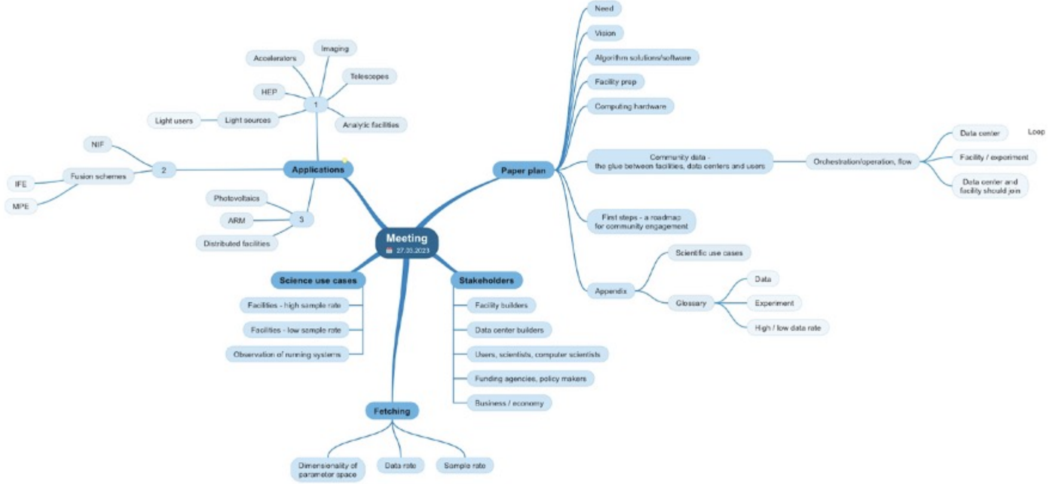


Figure 1 Topic graph of the discussion on Monday afternoon outlining both the science drivers, stakeholders and applications as well as a plan for a perspective paper as the goal for the seminar.

- More accurate, precise and reproducible results
- Finding complex patterns in big data
- Automated workflows for scientific facilities
- Optimise energy efficiency of large facilities
- Finding new and unexpected science in data

4.2.2 Group 2: Approach

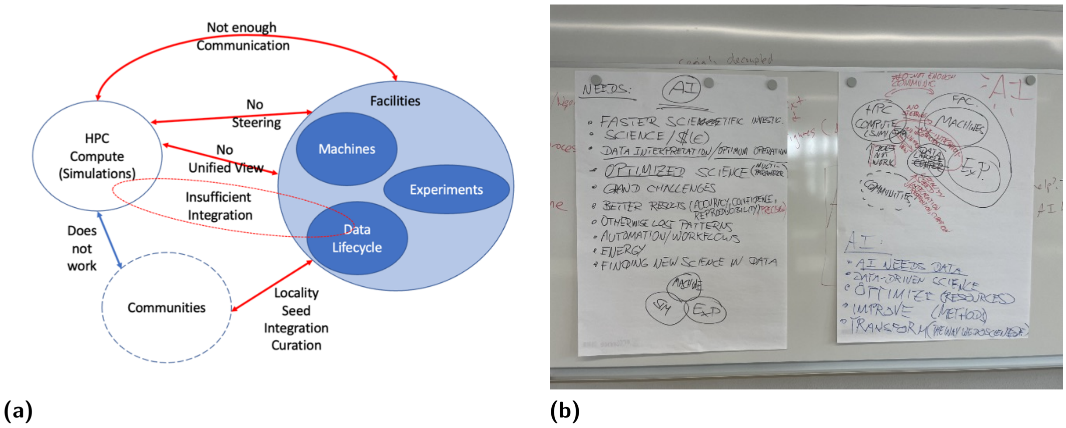
The discussion around which approaches might be fruitful to pursue resulted in a set of high level questions that would need to be answered followed by directions for solutions.

Questions:

1. How will we control complex, serial, and decoupled science experiments/observations ?
2. What high-level (abstract) approaches will we use to be more responsive to (said) experiments/observations?
3. How will we respond to grand challenges?
4. What expertise (existing or to be developed) do we need?
5. How do we enable scientists to shorten the "time to science"?

Answers:

1. Match simulation availability to experimental demands → surrogate models to bring physics into the control loop
2. Use inductive AI methods to merge multimodal and heterogeneous data → Autoencoders, deep neural networks, reinforcement learning, etc.
3. Build efficient connections between computing and experiments → couple AI, optimization, ...
4. Engineer and demonstrate robustness and uncertainty quantification for application on real machines and systems → AI techniques for UQ, robustness, to increase machine uptime, etc.
5. Offload experimental data to computing resources
6. Decide next steps based on both computational and experimental knowledge



NEEDS: AI

- FASTER SCIENTIFIC INVESTIG.
- SCIENCE / $ (€)
- DATA INTERPRETATION / OPTIMUM OPERATION
- OPTIMIZED SCIENCE (MULTI-PARAMETER)
- GRAND CHALLENGES
- BETTER RESULTS (ACCURACY, CONFIDENCE, REPRODUCIBILITY) PRECISION
- OTHERWISE LOST PATTERNS
- AUTOMATION / WORKFLOWS
- ENERGY
- FINDING NEW SCIENCE IN DATA

AI:

- AI NEEDS DATA
- DATA-DRIVEN SCIENCE
- OPTIMIZE (RESOURCES)
- IMPROVE (METHODS)
- TRANSFORM (THE WAY WE DO SCIENCE)

Figure 2 (a) Straw-man diagram of the current state of the art and its challenges; (b) Notes of the Needs working group.

4.2.3 Group 3: Envisioned Outcomes

To better understand what success might look like the group members decided to organize the discussion into three time frames: near-term (1-3 years), mid-term (5 years), and long term (10 years) goals.

1-3 Years

- Define Requirements for
  - Computing needs
  - Experimental needs
- Interactive (ML) compute
  - Capability for 5000 Jupyter notebooks
- Edge computing integrated into control system (EPICS++)
- Data and meta-data curation (per domain)
- Connect control to HPC
- New generation of AI-ready control system
- ML enabled simulations
  - Inner loop
  - Outer loop

5 Years

- Self-driving experiment
  - Self-tuning
  - Self-optimizing
  - Archive for stability/reproduction
- Integrated simulations for ML guidance and inference
- Science metrics other than uptime
- Reliable surrogate models updated on the fly
- AI enabled streaming
  - Analysis
  - Anomaly detection
- Distributed experiments
- Robust/ reliable / explainable ML for science

10 Years

- Data standards with translators
- Autonomous collection of data
- Scientist in the loop decision making
- Coupled experiments

4.3 Tuesday Afternoon

Tuesday afternoon developed the paper draft further by exploring the software and hardware needs as well as the necessary changes in large scale facilities. The charges for the different group after the morning discussions were as follows:

- Software needs
  - AI techniques and software tools (classes of AI tools – huge networks, tiny ones, ...)
  - Orchestration for computing and experiment interoperation
  - More ...
- Computing hardware
  - Needs for compute on the instrument
  - Needs for compute at the edge or facility
  - Needs for data center
  - Needs for distributed computing
  - Kinds of architectures, systems
  - More ...
- Facility preparation
  - Making facilities look fully integrated – unified compute and experiment
  - Making facilities AI ready (AI ready diagnostics and instruments)
  - Making facilities prepared for computing
  - Making facilities networked to resources

4.3.1 Group 1: Software Needs

The software working group first collected a list of general capabilities that will be required before discussing in more detail the different dimensions that differentiate various AI approaches.
Necessary capabilities

- AI tools and methods
  - AI benchmarks
  - Tools for knowledge extraction
- Orchestration
  - Online and coupled automation
- ML-Ops and RSE
  - Language of choice: Julia, Python
  - Frameworks to build AI
  - Workflows
  - Notebooks / Interactive development
  - Model / Data parallel
  - Big parallel models
  - Hyperparameter optimisation
  - Model optimisation
- Target users
  - Facilities
  - Scientists / Users
- Automation
- Tools for model calibration and validation
- Tools for continual learning (especially for model drift)

Dimension of AI models

- Time to train
- Time to inference
- Cost and energy
- Size of the model (number of parameters)

  - Data size
  - Compute for training
  - Compute for inference
- Model science performance
  - Accuracy
  - Precision
  - Stability
  - Convergence
  - Fidelity
  - Composability
  - Modularity
- Data
  - Sparsity
  - Multiple modalities
  - Heterogeneity
  - Locality
  - Privacy
- Model capability
  - Multimodal
  - Explainability
  - Deterministic
  - Uncertainty
  - Federated models
  - Dynamic models
- Privacy preserving federated learning as a Service
- Generalisability / Ability to disentanglement (symbolic knowledge / data-driven / representation learning)
- Physics-Informedness
- Procedural knowledge

4.3.2 Group 2: Computing Hardware

The hardware discussion resulted in a list of technology directions that must be considered for AI augmented facilities.

- Object store-type, DB focused, storage
- SWaP (size, weight, and power) based AI-hardware (sensors, embedded systems etc.)
- Chiplet-based (low latency/high bandwidth) embedded-AI accelerator
  - Composability across multiple vendors
- Latency optimized accelerator – multi mode(a)l
- AI hardware for edge training
- Capacity: Large scale, on-demand, aI-training
- Capability system driven AI computing
- Network protocols?!
  - Ethernet

4.3.3 Group 3: Facility Preparation

While full integration and potential distributed “super-facilities” are the long term goal, the facility preparation was discussed in terms of experimental and compute facilities yet with a focus on ensuring inter-facility communication

Experimental Facilities

- AI ready diagnostics / control
  - Digitized
  - Networked vs. online (real-time availability)
  - Compute enabled (ASIC)
  - Sufficient bandwidth
  - (Semi-)autonomous calibration
  - Monitoring and change detection
- Data acquisition system
  - Acquisition
  - Provenance
  - Meta-data
  - System state
- Local compute
  - High precision and AI compute
  - On-site data reduction
  - Large model inference
  - On-site storage
  - Data storage and exploration before transmission
  - Networking
- Software defined systems
- Flexible adjustment of controls (readiness to update)
- Retrain operators to engage with AI

Computing Facilities

- Data transfer / interface with the world
- Networking
- Robotic operating system
- workflow service layer
- Reconfigurable storage
- Hetrogeneous nodes with flexible connection for strong vs. weak scaling
  - High-precision
  - Low-precision
- Make computer center time-responsive
- Data availability and/or streaming dataflow
- Hardware and software policies, i.e., queue priorities

4.4 Wednesday Morning

Wednesday morning focused on some outstanding topics such as communication, first steps, and potential early demonstration targets.

4.4.1 Group 1: Communications and Data Movement

- ESNet
- One platform to file bug in a transparent process
- How to influence Open source packages
  - Resilience
  - Robustness
  - Life-cycle management of the AI software stack
- Storage across sites in a transparent way
  - Access API
  - Lifecycle of data

- High bandwidth vs. low latency
  - What is the priority (depending on the dataflow)
  - Networks are restricted by reality
  - Critical boundary conditions
- Wireless connections
  - Policy issues
  - Security
- Remote control must be viable
- Dealing with asynchronous information

4.4.2 Group 2: First steps

The second group discussed which communities to engage and on which projects to start moving towards the joint vision of AI augmented facilities.

Early Demonstration Targets

- Magnetic confinement fusion: Surrogate models and experimental controls
- "InterTwin": Multi-Science digital twin engine
- Laser-driven inertial fusion (LLNL, LBNL, Rochester) : AI surrogates
- Rock-IT: Digital twin for photon science with applicationsin catalytics
- BOFAB: Advanced Photon Lightsource + SLAC
- Digital earth
- OPTIMA: Real world evidence + AI for cancer
- ACCLAIM: AI for accelerator research
- ERVM-WAVE: AI for safe operation of Einstein telescope
- Helmholtz AI: Autonomous accelerators
- Synthetic data for particle physics
- MALA: Materials predictions from fundamental physics
- KITTEN: Energy responsible project for accelerators at KIT
- Center of Excellence for Research on AI- and Simulation-Based Engineering at Exascale (RAISE)
- ErUM-Data-Hub: central networking and transfer office for the digital transformation in the exploration of universe and matter.

General topics of interest for first steps

- Determining generalizable patterns of application of AI
- Workshops/Hackathons/Datathons/Studyathons/Ideathons
- Repositories for AI models
- Standardized intermediate representations
- BADGER/SLAC: Control optimizations, BlueSky
- Control standards for robotics
- Openness and interoperability
- High-throughput AI hardware
- Energy + Sustainability + Optimization
- Multi-target optimization
- Multi-X surrogate demonstrations
- Federated learning
- Hosting for training data?

4.4.3 Group 3: 1-3 Year Projects

What are the capabilities and time frames to make these capabilities a regular occurrence, i.e., the experimental/computing facility has an access mode for this

Experimental Facilities

- Proxy hardware (1Y)
- Access to test facilities (1-2Y)
  - connect different science teams
- Triggered analysis (2-3Y)
- Experimental steering (3Y)

Computing Facilities

- Access policies (1Y)
- Make schedules reliable and open (1Y)
- On demand / interactive computing (1-2Y)

Other important initial activities not directly linked to facilities

- Building working groups across multiple professional societies
- Staff education
  - Curriculum development
  - Summer schools
  - Maybe connect to academically minded staff members interested in teaching or even explicitly hiring these
- Workshops to find “killer” applications
- Social aspects
  - Courses aimed to get people comfortable rather than necessarily convey knowledge
  - Schedule coordination
- Joint proposals

5 Perspective Paper Development

Given the discussions outlined above the seminar broke into writing groups largely split by sections that started to develop outline and/or initial text. Below we report on the state of these developments by the end of the seminar. Once finalized the goal is to publish the polished results as a persectives paper in a yet to be determined venue.

5.1 Section 1: Needs

Michael Bussmann (Helmholtz-Zentrum Dresden-Rossendorf, DE)
Jean-Luc Vay (Lawrence Berkeley National Laboratory, US)
Arvind Ramanathan (Argonne National Laboratory – Lemont, US)

Gaps that AI can address

AI has enabled transformational progress in a number of scientific, engineering, and technology domains; however, enabling rapid progress in the adoption of AI strategies will not be entirely realized, unless these critical gaps are addressed.

1. AI is needed for optimizing the life-cycle of scientific data: While commercial applications (such as recommender systems) have ready access to quality datasets, scientific data is often more decentralized, distributed and less deterministic. Hence, one of the key needs of AI is in identifying the critical aspects from data, including achieving the ability to optimally reduce and represent vast scientific datasets, while also identifying potentially anomalous data points in them. Further, AI methods need to automatically enable the transformation of data → (actionable) information → (inferable) knowledge, which can significantly impact today's scientific processes.
2. AI needs for enabling close, near real-time coupling of modeling, simulations and experiments: Multimodal, multi-view, multi-scale data is extremely common in scientific processes. Specifically, experiments and sensor networks within integrated research infrastructure produce data at specific length- and time-scales for investigating complex phenomena and AI techniques can potentially act as 'glue' to stitch together such datasets. However, to truly enable experimental design at scale, there needs to be a strong coupling of how simulations and experiments inform each other about the complex phenomena of interest. This in turn requires AI methods to not only act as effective surrogate models, but also to respect foundational scientific principles that produce the underlying data distributions.
3. AI is needed for optimizing facility operations and control: Current scientific instrumentation within research facilities is largely informed by user access and individual research goals, which makes facilities to be extremely expensive and less optimal for cost-effective management. However, costs of managing research infrastructure can be significantly lowered by analyzing user access patterns and modeling, which can be used to optimize how experiments and simulations can be scheduled and automated. In essence, much like the car assembly line in the past, there is a need to industrialize the scientific process through meaningful engagement of scalable automation (such as robotics).
4. AI is needed for cross-domain, data-driven, automated, accelerated discovery: Current scientific progress is enabled in part by serendipitous discoveries within individual labs or consortiums, and still relies on domain-specific expertise drawn from intense training and experience. However, as is now evident, rapid progress in science and technology is enabled by 'cross-pollination' – wherein knowledge drawn from across disciplines and technology can mutually inform and benefit new foundation discoveries. Enabling such discoveries requires access to enormous datasets and analyses of patterns across such data which is enabled by modern AI techniques. Furthermore, instead of having individuals synthesize this knowledge, and design experiments, automated, data-driven knowledge distillation and hypotheses generation can direct experimental campaigns that can in turn accelerate the pace of scientific discoveries to address some of the aforementioned grand challenges.

5.2 Section 2a: Vision – Approach

Brian Spears (LLNL – Livermore, US)
Sunita Chandrasekaran (University of Delaware – Newark, US)
Matthew Streeter (Queen's University of Belfast, GB)

The pace and quality of scientific discovery can be greatly improved by erasing the boundary between experimentation and computation. Each step in the scientific process can be revisited with newer hypotheses through the application of machine learning to accelerate scientific discovery, which then enables orchestration of the whole activity. The scientist is then the conductor of this orchestra, and is placed back in control of the increasing complexity and huge scale of modern science.

Coupling scientific computing with large-scale experimentation using modern AI methods will introduce a new class of unified, AI-augmented facility. In such a facility, AI surrogates can be used to capture the intricacies of detailed simulations, but in a way that is accessible to experiments in seconds, not hours or months. Such facilities will then have access to simulations as a real-time commodity for informing experiments. Likewise, AI-guided analysis using representation learning can capture and reduce complicated diagnostic information on the fly. This promises to provide distilled and interpretable empirical data back to the computational world, again in moments, not months. AI-driven representations of both computational and experimental science, scientists can decide far more effectively – about the best next simulation to run or the most insightful next experiment to execute. In fact, these decisions can be accelerated and improved by using formal optimization algorithms that exploit compact AI representations to navigate to superior experimental outcomes; the decisions can even be made semi-autonomous or automated, allowing an optimized loop to execute well-informed experiments incredibly rapidly, keeping pace with modern facilities that perform experiments on the timescales of minutes down to milliseconds.

Its tightly AI-coupled components and efficient information processing ecosystem allow for scientific discovery on much shorter timescales than previously imaginable. Likewise, the AI-enabled operating environment simplifies the execution of simulation and experiment. This frees expert users of the previously disconnected facilities to think and innovate rather than labor over job submission, diagnostic data reduction, and similar operational tasks. Furthermore, non-expert users will find access to complicated science machines to have been democratized, allowing these new facilities to serve a far wider range of scientific communities.

We envision the technical roadmap of our AI-enabled facility to be an energy-efficient framework/design/infrastructure driven by novel hardware and software components that are connected together to form a unified system. With a goal to offer an AI-acceleration solution everywhere, newer dedicated AI-hardware will be built to address the need for a variety of commuting fronts on the diagnostic end, near the facility, at a data center and across a distributed compute network. We will also build software for orchestrating a variety of operations in both high performance compute and experiments. Such AI-orchestrated software will provide the necessary infrastructure to drive informed decisions thus enabling newer science. Facilities augmented by AI will also have a modernized strategy for handling volumes of heterogeneous data that carries context from generation until analysis.

5.3 Section 2b: Vision – Outcomes

Roger H. French (Case Western Reserve University – Cleveland, US)
Niko Bier (DLR, Institute of Aerodynamics and Flow Technology, DE)
Shantenu Jha (Brookhaven National Lab – Upton, US& Rutgers University – Piscataway, US)

AI will be applied at different levels in scientific facilities.

1. On a level close to and integrated into e.g. optical sensors AI will be used to directly provide scientific information rather than raw data to the facility. This will not only lead to significantly decreased requirements in bandwidth and data storage but will also drastically speed up the scientific experiment as such.
2. On a facility level AI will be used to generate digital twins of the facility and thus to control scientific experiments. This will significantly increase the efficiency of a facility by better coordinating the tools and sensors involved during the experiment.Moreover the status of the facility and its components will be accessible in real-time and predictive maintenance will become more efficient and will increase the availability and safety of the facility. The digital twin of a facility will also allow for the use of virtual sensors making it possible to collect previously not accessible data.
3. On an overall level distributed scientific facilities and HPC resources will be merged together into an AI-driven facility. In these facilities AI will be used to orchestrate the experiments. It will not only significantly speed up scientific experiments but will also be used for planning and conducting of experiments. AI will not only perform experiments better, but will conduct better experiments.

In such a new "generative science" completely new scientific questions will emerge and appropriate experiments will be conducted. Scientists will be able to focus on understanding and answering fundamental questions on the basis of an unprecedented quality of scientific experiments and their results.

5.4 Section 3: Software

Ravi Madduri (Argonne National Laboratory – Lemont, US)
Jeyan Thiyagalingam (Rutherford Appleton Lab. – Didcot, GB)

The software requirements for the AI Augmented Facilities are driven by users and operators of the facilities. In Software Development Life Cycle (SDLC), these processes are typically referred to as stakeholder analysis and requirement analysis. For the purpose of this document, we identified facility operators and scientists using the facilities as stakeholders. This section offers requirement analysis and identities of current methodologies and places where AI can augment and lead to better results.

We assume that AI-Augmented facilities have the following stages, regardless of the scientific domain (whether they are, e.g., photonic or laser or environmental or datacenter facilities) or whether they are centralized or distributed facilities.

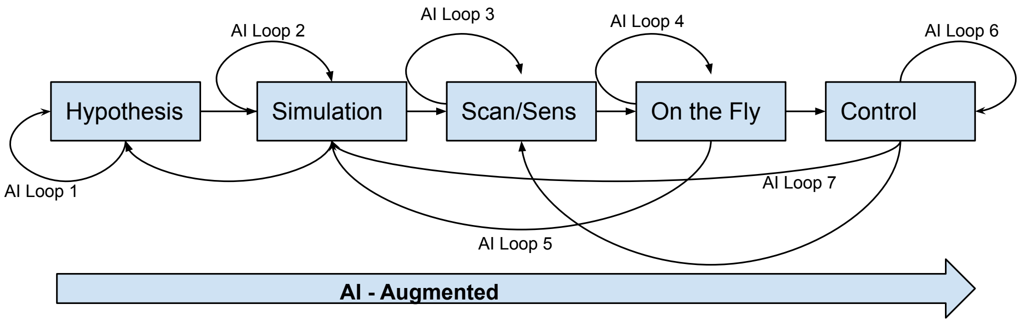


Figure 3 Examples of AI Loops at different stages: Synthetic GAN to simulate a facility that help validate or come up hypothesis (Omniverse) – AI Loop 1, Surrogate Models that help come up with better results from Simulation – AI Loop2, etc.

1. Basic hypothesis or expectation of the scientific outcomes from experiments or data acquisition
2. Simulation studies to understand the hypothesis better, and to develop insights into expected results off the potential experiments or data acquisition
3. Empirical data/fielding of experiments
4. Consolidation and interpretation of the data for developing better understanding or to underpin scientific discoveries
5. Tuning or controlling of the data acquisition step for improving the quality of the data

In the contemporary setting, some of these are human-controlled, which may eventually be replaced by an automated system component in the context of an AI-augmented facility. There are a number of challenges here that different stakeholders face here,especially in the absence of any AI-specific capabilities. These include the following:

- **Expensive Simulations:** The simulations can be a serious bottleneck to the overall progress, especially when they are of high-fidelity in nature, or when high-quality outputs are desirable,
- **Divergence between Simulated and Experimental Data:** The actual results from the experiments may as well be different to the hypothesis or expected results, owing to a number of reasons
  - Scope-insensitive calibration of the data acquisition instruments or sensors,
  - Incorrect a priori of hypothesis for simulations, and
  - Physical limitations of the data acquisition, such as potential dosage level or acquisition resolution of the detectors.
- **Data Analysis and Enhancements:** Understanding, interpreting and analyzing the experimental or sensed data, and correlating that with the a priori formed as part of the hypothesis, and
- **Feedback-driven Optimal Control** of the experimental facility. Optimal and stable control of the instruments are a serious practical challenge, especially in tolerant-sensitive contexts.

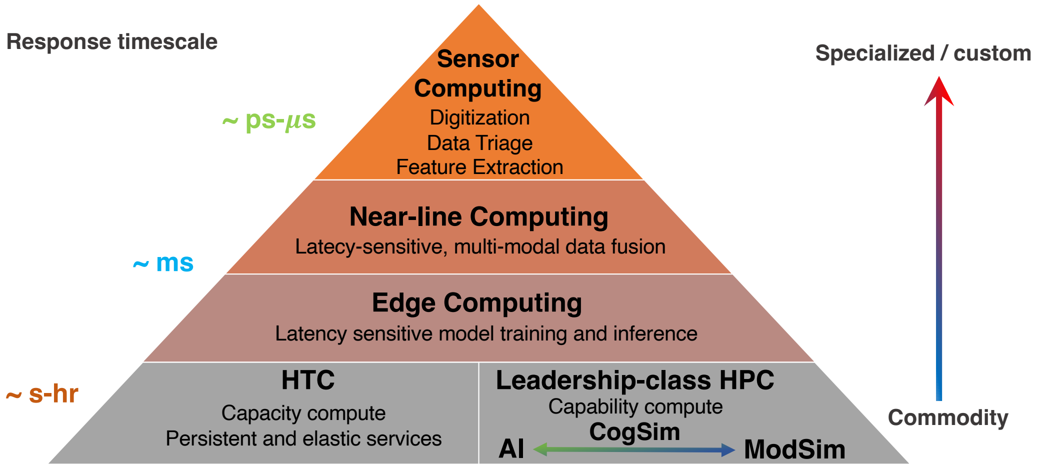
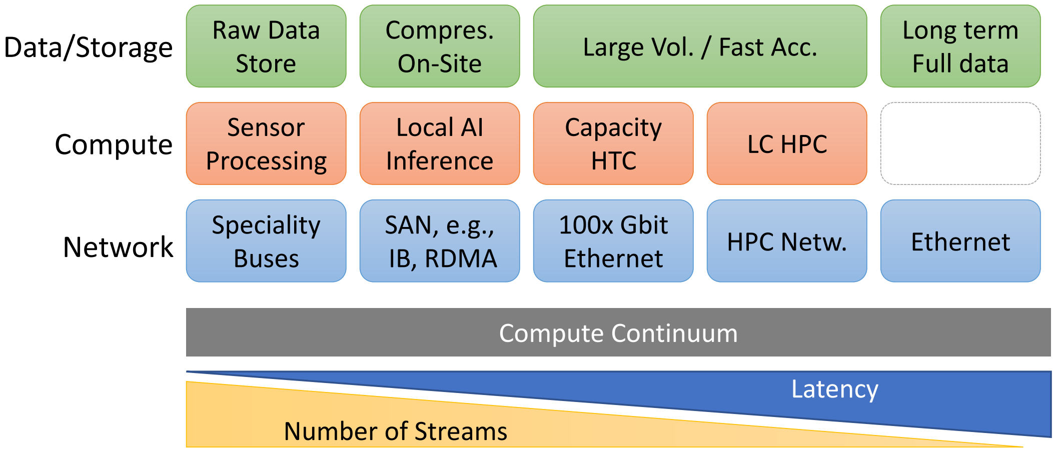


Figure 4 Current State-of-the-Art Hardware supporting Data Intensive Facilities.

5.5 Section 4: Hardware

Brian Spears (LLNL – Livermore, US)
Martin Schulz (TU München – Garching, DE)
Andrea Santamaria Garcia (KIT – Karlsruher Institut für Technologie, DE)



Data/Storage	Raw Data Store	Compres. On-Site	Large Vol. / Fast Acc.		Long term Full data
Compute	Sensor Processing	Local AI Inference	Capacity HTC	LC HPC	
Network	Speciality Buses	SAN, e.g., IB, RDMA	100x Gbit Ethernet	HPC Netw.	Ethernet

Figure 5 Current state-of-the-art of hardware systems supporting the design and operation of AI augmented facilities. It features a clearly segmented technology stack separated in compute, storage and network (horizontal) and technologies at different locations (vertical).

On the hardware level we are currently seeing three trends that will radically impact AI augmented facilities in how they are built, operated, and used:

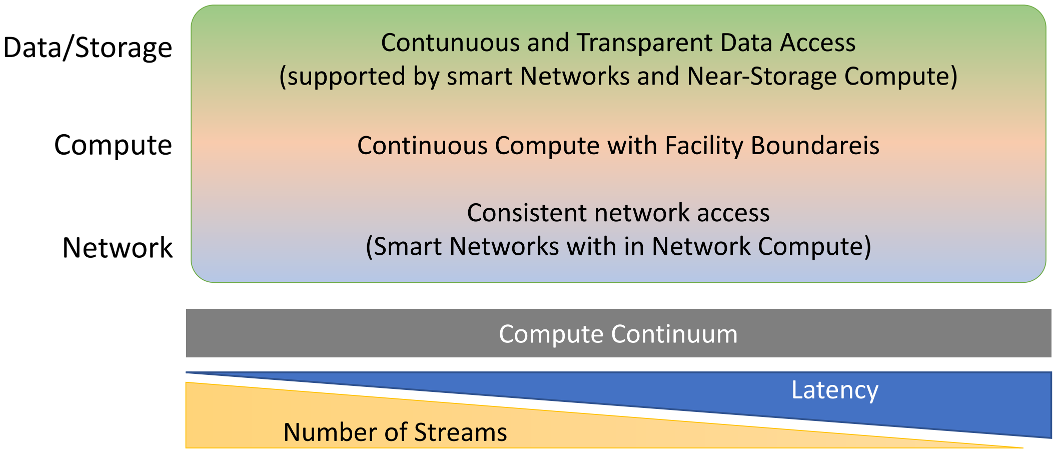

Data/Storage
Contunuous and Transparent Data Access
(supported by smart Networks and Near-Storage Compute)
Compute
Continuous Compute with Facility Boundareis
Network
Consistent network access
(Smart Networks with in Network Compute)
Compute Continuum
Latency
Number of Streams

Figure 6 Future hardware vision: integrated systems and APIs making storage and compute available independent of location and hardware system, supporting fused augmented facilities.

- Modern AI accelerators are moving from data-center-only use towards facility / on-premise systems making both inference and learning capabilities available closer to the data source.
- Convergence of compute and network in the form of SmartSwitches and SmartNICs, which enables processing on the fly and independent of actual location in a decentralized fashion
- Convergence of compute and storage in the form of Near or In Storage Compute, enabling low-latency data processing at the various and distributed storage location.

Figure 4 illustrates the hierarchy of needs for an AI-enhanced, integrated, experimental and compute facility.

- At the top of the hierarchy is the set of primary data inputs from the experiments represented as a set of sensors. The hardware needs here are extremely low-latency, working with single or small number of samples, high data capture rates, and most importantly engagement with a set of custom / bespoke / specialized sensors that have unique demands such as custom analog signaling that requires high-speed analog to digital converters (ADCs), proprietary formats, etc. Some of the key demands for AI at this level are **feature extraction, data processing / conditioning, local controletc. Primarily inference tasks**. Another class of sensors are those that are deployed in a Size, Weight, and Power (SWaP) constrained environment / facility. Typical examples of these are battery powered and may be geographically distributed (e.g. drones, buoys, ...).
- The next level of the hierarchy is the latency-optimized AI-accelerator that is designed for data-fusion of heterogenous, multi-modal data near the input sensors, forming a “near-line” compute capability. These accelerators need to be optimized to process “single samples” of data (i.e. all of the data from a single event / shot / experiment), and cannot afford to batch multiple samples together prior to processing. Another way to think about this is that they need to be optimized for streaming data sets, where decisions and analysis have to be produced for each sample in real-time. [keywords: streaming, real-time, latency-optimized] Primarily focused on AI inference tasks.
- Integrating the output of the real-time sensor streams output from the near-line accelerators are the near-experiment computing resources. These resources are now expected to work with batches of data samples, and start to blend workloads that include both inference of more complex / larger models as well as the training or fine-tuning of existing or new models

- The bottom of the compute hierarchy is anchored by two classes of computing needs / ecosystems: Leadership-class High Performance Computing (LC HPC) aka capability computing and capacity-based High Throughput Computing (HTC) with persistent and elastic services.
  - HPC resources are intended for full-system, monolithic, modeling and simulation (ModSim) jobs and training of large neural networks (e.g. foundation models). They are designed with highly-interconnected accelerators (e.g. GPUs) using proprietary network architectures (InfiniBand, Slingshot, etc.) Traditionally, these are run in a batch-scheduled manner. Additionally, there is a need to establish new HPC system architecture designs that are optimized for traditional modeling and simulation codes (ModSim), AI training and inference workloads, as well as hybrid cognitive simulations (CogSim) workflows. Optimizations for ModSim include accelerators with high-precision data types (64b arithmetic), high cross-section bandwidth networking, and high-bandwidth write-optimized parallel file systems. HPC optimized for AI training and inference will leverage low-precision accelerators, with read-optimized parallel file systems and support for advanced object-stores. CogSim HPC systems will require a blend of both capabilities.
  - HTC resources are designed to serve the needs of many users, models, or data streams. Persistent and elastic services are designed to capture experimental data in real-time with the ability to meet data surges from the experiment. Additionally, they provide the ability to update / refine / fine-tune AI models in the background as new data is captured.

Fundamental shifts needed in the AI-accelerator architecture space:

1. Engagement with customized sensors requires AI-accelerators to interface with a myriad of signaling protocols, formats, and data types (including varying precisions). The current state of the practice is to develop custom accelerator sub-systems that are able to funnel the outputs of these sensors into standard AI-accelerators such as GPUs or a specializable systems such as a blended FPGA / AI-accelerator (such as a Xilinx Versal) through a series of customized ADC capture cards, FPGAs and ASICs. These data capture sub-systems can become quite complex and are as unique as the sensors themselves. The emergence of chiplet-based AI accelerator design would enable a fundamental shift in how research teams would be able to interface state of the art AI accelerators with their custom sensors. Chiplet-based architectures offer the promise of allowing multiple vendors to integrate proprietary hardware into unique hardware processes in a timely and affordable manufacturing process. Advances in this field would enable sensor manufacturers and research teams to more easily and efficiently couple SOTA AI accelerator hardware directly (or closely) to their sensors and thus unleash new capabilities that can be exploited at the sensor.
2. AI accelerators in the near-line computing regime will need to be tuned for real-time, streaming workloads, where they need to produce inference results on only a single sample of data that contains a large volume of heterogeneous data fields. Frequently these accelerators will need to be able to meet hard latency limits and execute neural network models (inference) with predictable performance characteristics.
3. AI-accelerators close to the sensor may need to meet the demands of a Size, Weight, and Power (SWaP) constrained environments. For these systems, the operational cost in terms of inference requests per Watt will be a key design metric.

5.6 Section 5: Facilities Upgrades/Actions

Kyle Chard (University of Chicago, US)
Derek Mariscal (LLNL – Livermore, US)
Rafael Ferreira da Silva (Oak Ridge National Laboratory, US)

Cross-cutting Requirements. There are several cross-cutting requirements that span experimental and computational facilities, broadly spanning the computing continuum from edge to data center.

Connectivity. Central to integration is the need for unimpeded access between experimental and computational facilities, such that experiment results can be rapidly available to AI models running nearby or at compute facilities and that AI informed decisions can be enacted at experiment facilities. Ideally, we require both in-bound and out-bound network access between participating entities, with minimal restrictions (e.g., firewalls, NAT). In many cases, regulations and policies will limit the degree to which connectivity is permitted; however, this represents an opportunity to explore approaches to reduce friction. We refer to the success of the Science DMZ model in science as a way of providing minimally impeded access to scientific data and resources.

Security. Traditionally, compute and experimental facilities have rigid security models that prevent automation. To attain semi/full automation it is necessary that facilities implement flexible and interoperable security models that enable remote and automated actions, while ensuring accountability for actions. As we move towards automated and AI-based techniques, there is a need for delegatable access, via which humans may permit operations to be performed within some bound and for some period of time. We further require methods to audit operations, restrict the scope of access, and revoke access rapidly.

Data. AI is dependent on data and as such movement between experiment and/or compute facilities is integral in enabling AI-augmented experiments. Important data movement characteristics may different significantly between use cases, for example, to deliver streams of experiment data for analysis, to move trained models to the edge, or to move huge amounts of simulation data to available compute resources. We require common interfaces (e.g., Globus, HTTPS, common messaging protocols) and methods to securely access, move, and share data between participants.

Storage. The plethora of data storage options spanning experiments to HPC presents obstacles for accessing and preserving important data. Methods are required to a) access storage that is physically and logically distributed; and b) optimize storage for different tiers of data (e.g., ephemeral storage of intermediate data vs long-term preservation of experiment data). Data access requires a consistent view of data irrespective of storage, and interfaces via which data can be accessed by new and traditional methods (e.g., using HTTP for integrated data visualization in existing tools).

Compute. AI-enabled methods will require that computation be fluid, such that it may be executed where it makes the most sense (e.g., near to a detector, where simulation data reside). Enabling such fluidity requires a) common interfaces to execute tasks (e.g., simulation, model training, inference) across the computing continuum, from edge to HPC; and b) portable packaging such that codes may be easily executed without herculean efforts to configure and contextualize environments. Community efforts such as Superfacility and funcX lay the groundwork to provide common APIs, while work in container technologies such as Shifter will support the need for portable codes.

Policies. Perhaps the most challenging problem to address is the need to support the varied policies across participating organizations, between users, and in consideration of AI-managed actions. Understanding of the requirements for defining a trustability model between facilities is a first step towards enabling an AI augmented automated facility.

Experimental Facilities. Experimental facilities will in many cases require updates to enable AI-augmentation. This includes diagnostics with integrated edge computing resources for rapid data distillation, remote communication with and precise control over machine inputs, robust pipelines for two-way communications with high performance computing, and new tools for monitoring the system/diagnostics health and calibration.

Detectors/measurements. Detector measurements and sensors collect data that will inform AI about the real world. Crucially, the detectors must be able to deliver electronic signals (i.e. the information must be readily digitized). For diagnostics without direct AI-ready interfaces, edge hardware must be deployed to provide for real-time high-accuracy data processing, localized storage (depending on network pipeline resources and latency needs), and access to diagnostic controls. Such hardware can include GPUs, FPGAs, or other specialized hardware that can perform complex calculations quickly (discussed in HW). By processing data at the edge, AI will be able to consume heterogenous data to drive the experimental process while retaining full-fidelity data for validation.

Controls access. In AI augmented experimental facilities, it will be necessary to interface with machine and diagnostic controls in ways that are often inaccessible to users. Common, open-source control interfaces such as EPICS [ref], BlueSky [ref] are examples for enabling control and data acquisition from heterogeneous hardware[ref] that could be readily accessed via AI while retaining proprietary interfaces. New operation policies for experimental facilities will need to be developed to allow control of high-value systems (accelerators, robotics, lasers, etc.) with AI with robust limitations that prioritize machine safety while allowing transformative capability for fine control.

Robot control. AI augmented experimental and computational facilities often use robots to automate tasks such as sample preparation, data collection, and analysis. To control these robots, it is essential to have a common interface that allows for arbitrary control of the robotics. This can be achieved by developing custom software that integrates with the robot's control system and is again amenable to autonomous control.

Asynchronous, event-based operation. Keeping pace with high experiment throughputs and rapid AI-based decisions will require movement to a more asynchronous model of communication and control. Event-based models have been transformative in industry as a way of decoupling monolithic applications and improving performance. Moving to event-based models will require common methods to a) detect events; b) propagate events across distributed systems; and c) make decisions based on complex heterogeneous event streams. For many applications, it will be important that events include timestamps such that they can be ordered. Consistency across event generators is application specific.

Compute Resources. Computational facilities often target the deployment of large-scale computing systems to solve complex problems from a diverse set of computational science domains. As the use of AI solutions has significantly increased in these domains, there is a need to provide specialized support to heterogeneous computing architectures and flexible policies that address the different set of requirements.

Flexible and available allocation. Reservations and dedicated resources are crucial for ensuring that researchers have access to the resources they need when they need them (i.e., urgent computing). To this end, there is a need for the development of policies that allow different levels of reservations based on the urgency of the application. On the other hand, it is also necessary to support pre-emptive mechanisms to meet unforeseen requirements or impromptu workloads.

Hardware. Specialized hardware (e.g., GPUs, TPUs, and upcoming DPUs) is used to accelerate compute- and data-intensive tasks including data analysis, machine learning training, and AI inference. Specialized inference hardware is used for real-time analysis and processing of experimental data. Enabling support for these resources requires specialized software for bridging these resources to traditional HPC and storage resources.

Policies (access). Policies are important to ensure that resources are used fairly and efficiently. In an AI augmented facility, policies will have to be extended to encompass the heterogeneous set of resources and requirements from near real-time to long-term campaigns that might involve both HPC and specialized hardware. Additionally, policies will have to be flexible regarding experiment/access management, i.e. in order to enable steering across facilities there might be needed to provide support to generic/global accounts or provide specialized APIs that can manage/access external resources from/to computing nodes.

5.7 Section 6: First Steps

Marina Ganeva (Forschungszentrum Jülich, DE)
Tom Gibbs (NVIDIA Corp. – Santa Clara, US)

The first steps include programmatic proof of concept (POC) starter projects that are co-led between the Data Center and Selected Experimental Facilities, community outreach with workshops and centers of excellence along with training and development events such as hackathons.

The POC projects will develop and demonstrate the new workflows that connect the Science Data center with the Experimental facility, where AI is used as part of the workflow that connects them.. An example of a program is the SuperFacility Project at NERSC[1]. Each POC project must have scope that is sufficient to demonstrate the potential for improved science and identify features and requirements that can't be met with the current infrastructure.

A key feature of the new workflows are AI algorithms is that once trained at the HPC Data Center are fast enough and accurate to be deployed at the experiments for control and/or improvement of experiment operation. Another feature is Active Learning that may be executed at the Data Center and/or the experimental facility that can be used in conjunction with experiment operation. Simulation workflows are critical for explainability and validation of AI approaches as well as for generation of synthetic training datasets.

Examples of projects include improving the control of a Fusion Experiment, Multi Messenger Astrophysics, Improving the process for drug discovery with a biology lab or data analysis/evaluation on the fly as a neutron/x-ray/electron/light scattering experiment executes. Key features have already been identified that will need to be acted on in the near term to allow for the new workflow include the following:

[1] https://www.nersc.gov/research-and-development/superfacility/

- The Data Center Facility will need to include features that allow interactive usage that is interrupt driven along with the current batch oriented usage model for long running jobs.
- Both the Data Center and Experiment will need to be able to support the bandwidth and latency for data transfer. Reconfigurable storage will also be needed to support the experiment as it executes, which acts as a workflow buffer between the storage at the experiment and persistent platform storage at the data center.
- The Experiment hardware/software should allow for automated control. Curated data in the form that it can be used by AI will also be required.
- Funding agencies will need to initiate projects to pursue the new use cases for a sufficient period of time to develop a working example.

Workshops that cross domains as well as specific to a given domain will be needed to promote the new methods, socialize new concepts, identify key requirements and avoid redundancy. Where possible Centers of Excellence, Data Hubs and other vehicles should be developed for concentrated effort on specific projects within a community. The workshops and COEs should be balanced with Hackathons and bootcamps that can serve to help train users on the new tools and approaches. Additionally, hackathons and bootcamps allow for collective efforts to address challenging problems relevant for multiple facilities (inverse problem-related issues, uncertainty quantification, etc.) that result in code that can be evaluated.

5.8 Section X: Figures

Peer-Timo Bremer (LLNL – Livermore, US)
Annika Eichler (DESY – Hamburg, DE)

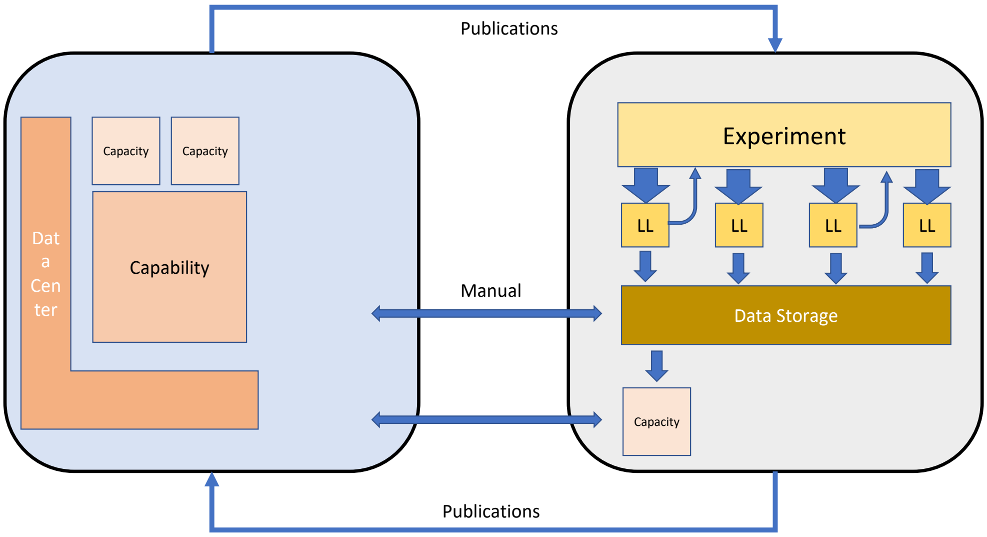


Figure 7 Conceptual drawing of the current state of the art in which computer centers (left) and experimental facilities are connected primarily through laborious manual data transfers and coordinated largely through publications.

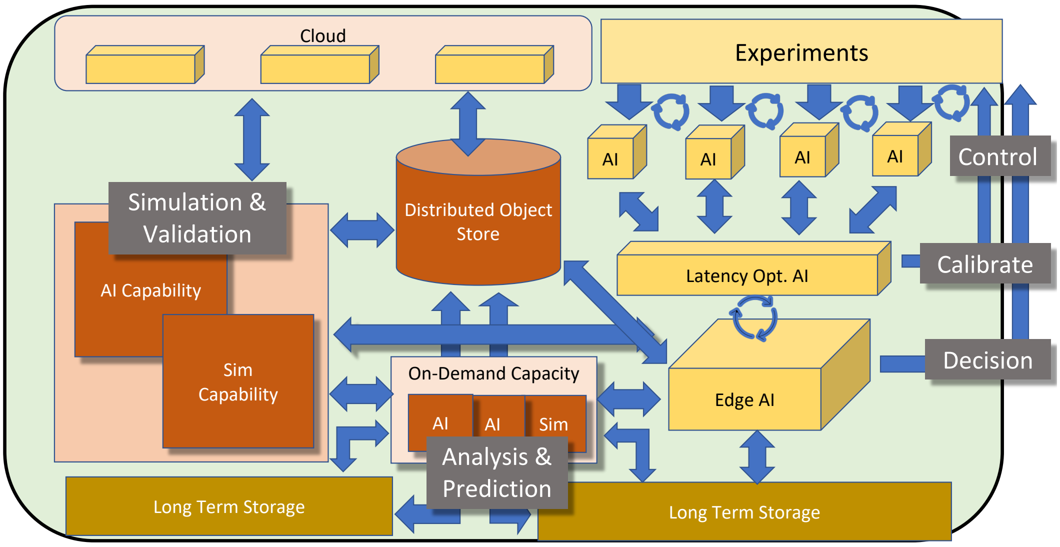


Figure 8 Conceptual drawing of future AI augmented facilities in which AI informs and supports all aspect of an experiment from real-time control to on premise compute, on demand processing, and large-scale simulations capabilties.

6 Acknowledgement

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, LLNL-TR-854217; MJVS acknowledges support from the Royal Society URF-R1221874; RHF contribution was funded by Department of Energy's National Nuclear Security Administration under Award Number(s) DE-NA0004104; Further support from the U.S. Department of Energy, Office of Science, under Lawrence Berkeley National Laboratory contract number DE-AC02-456 05CH11231; Partially funded by the Center of Advanced Systems Understanding (CASUS), which is financed by Germany's Federal Ministry of Education and Research (BMBF) and by the Saxon Ministry for Science, Culture and Tourism (SMWK) with tax funds on the basis of the budget approved by the Saxon State Parliament.

Participants

- Niko Bier
DLR – Braunschweig, DE
- Martin Boehm
Institut Laue-Langevin – Grenoble, FR
- Peer-Timo Bremer
LLNL – Livermore, US
- Michael Bussmann
Helmholtz-Zentrum Dresden-Rossendorf – Görlitz, DE
- Sunita Chandrasekaran
University of Delaware – Newark, US
- Kyle Chard
University of Chicago, US
- Annika Eichler
DESY – Hamburg, DE
- Rafael Ferreira da Silva
Oak Ridge National Laboratory, US
- Roger French
Case Western Reserve University – Cleveland, US
- Marina Ganeva
Forschungszentrum Jülich, DE
- Tom Gibbs
NVIDIA Corp. – Santa Clara, US
- Maria Girone
CERN – Geneva, CH
- Shantenu Jha
Brookhaven National Lab – Upton, US & Rutgers University – Piscataway, US
- Thomas Kühne
Universität Paderborn, DE
- Ravi Madduri
Argonne National Laboratory – Lemont, US
- Derek Mariscal
LLNL – Livermore, US
- Arvind Ramanathan
Argonne National Laboratory – Lemont, US
- Andrea Santamaria Garcia
KIT – Karlsruher Institut für Technologie, DE
- Martin Schulz
TU München – Garching, DE
- Brian Spears
LLNL – Livermore, US
- Matthew Streeter
Queen's University of Belfast, GB
- Jeyan Thiyagalingam
Rutherford Appleton Lab. – Didcot, GB
- Brian Van Essen
LLNL – Livermore, US
- Jean-Luc Vay
Lawrence Berkeley National Laboratory, US