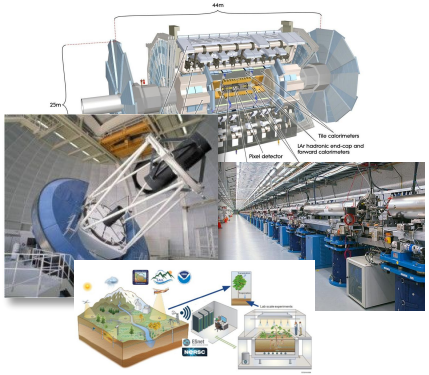
Directions in Data at NERSC

Wahid Bhimji

Data Day 2024

Why are we here? Science

Observed data



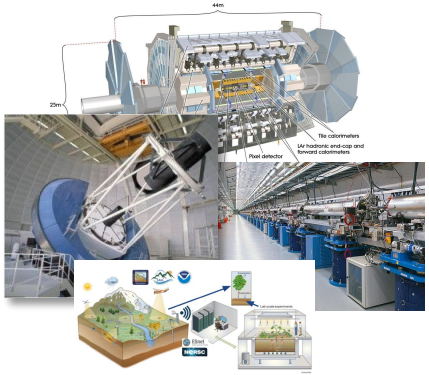
Secrets of the universe

Ω_c
Ω_m
σ_8
θ_{12}
m_H ….

Data Acquisition; Availability; Access; Analysis; AI; Archiving

Science data presents challenges

Observed data → Secrets of the universe



Ω_c Ω_m σ_8 θ_{12} m_H ….

Data Acquisition; Availability; Access; Analysis; AI; Archiving

SCIENCE LOST

- Inefficient Filtering
- Storage capacity limitations
- Bad data quality
- Human steering

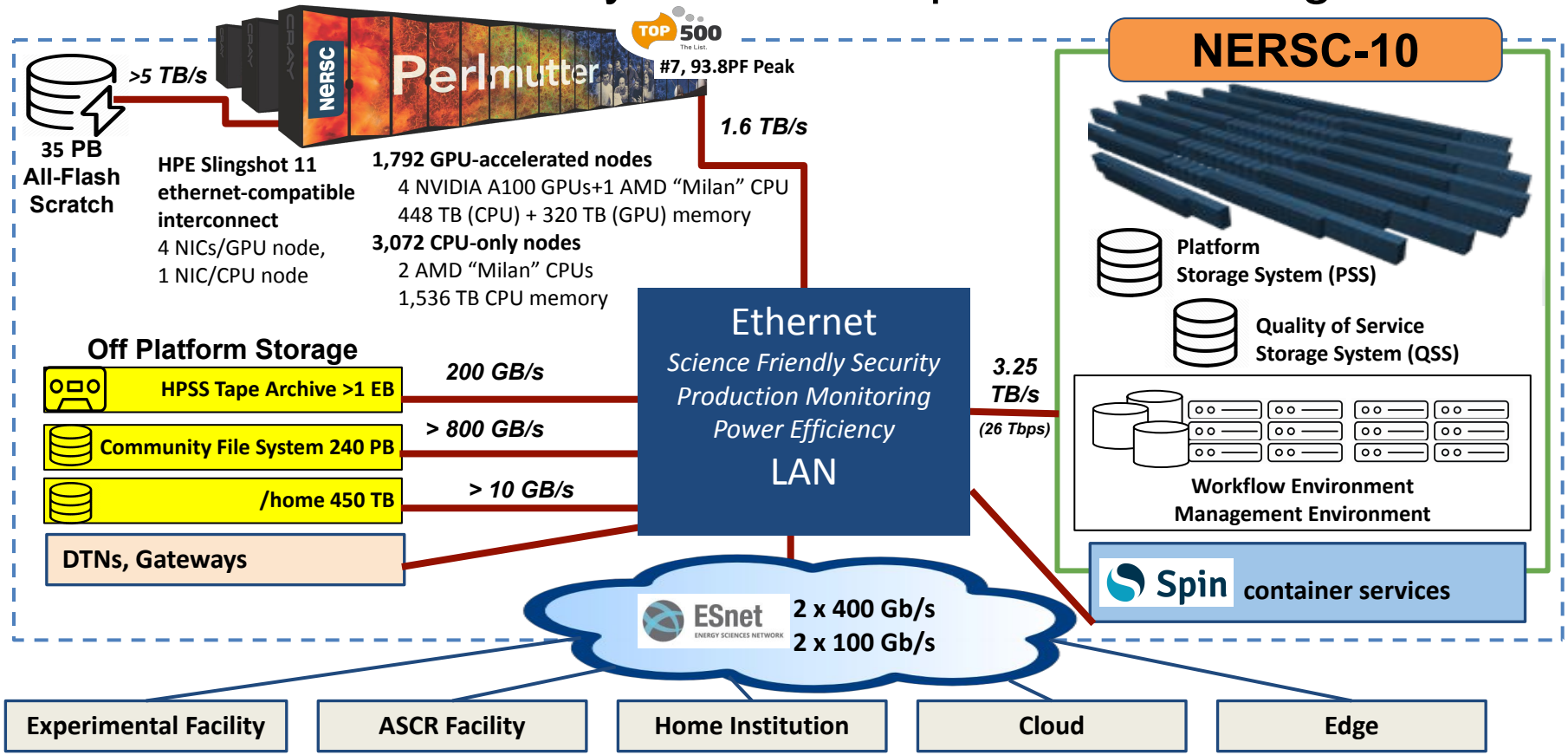
- Inadequate data transfer
- Poor data management tools

- Limited user interfaces
- Unproductive software
- I/O Bottlenecks

- Insensitive analysis techniques

- Irreproducible science
- Poor data preservation

NERSC has rich data systems to help these challanges



NERSC also has rich data software

Data Transfer and Access

Data Management

Data Analytics

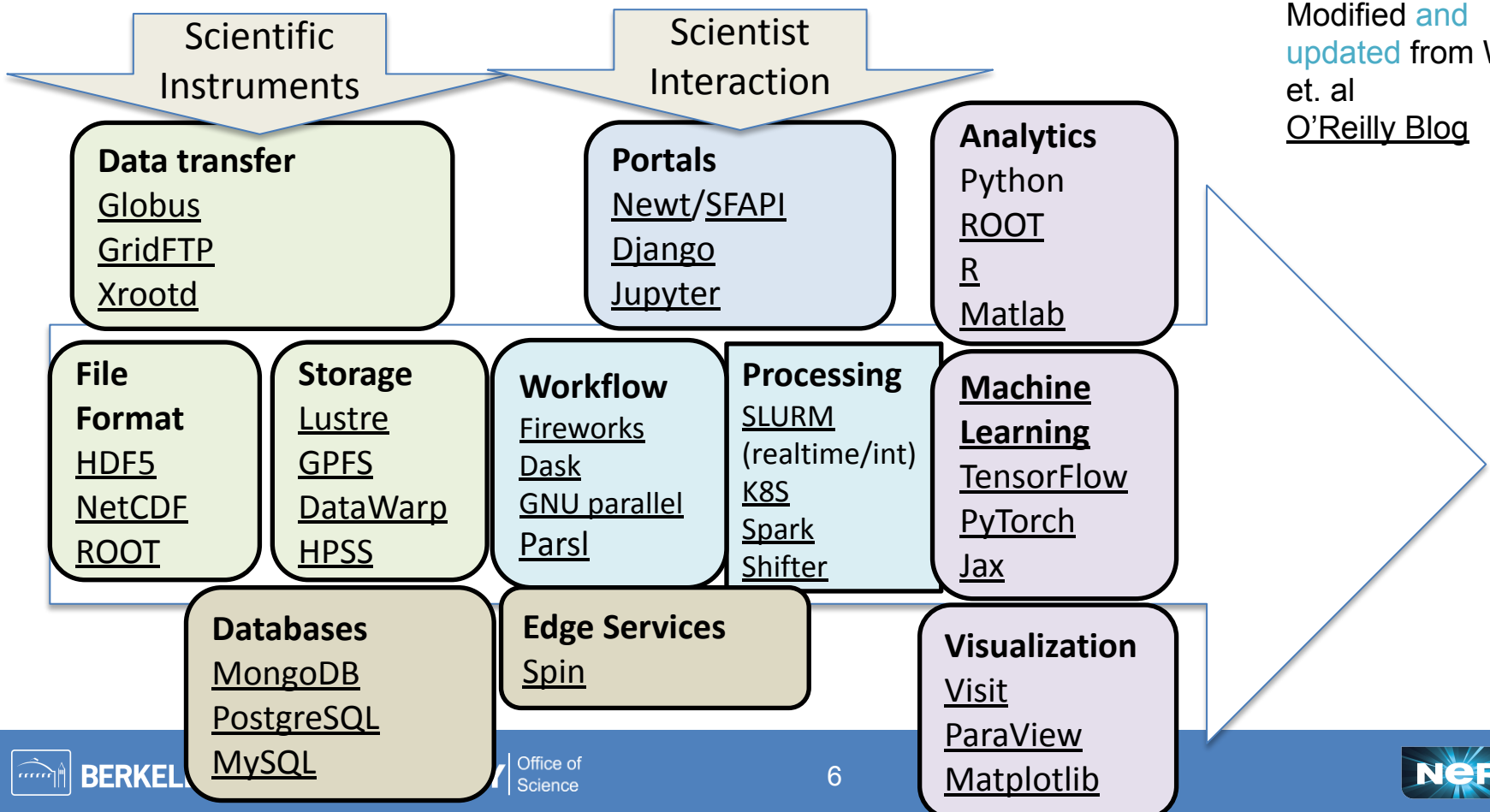
AI / Machine Learning

Visualization

Containers/Cloud-tech

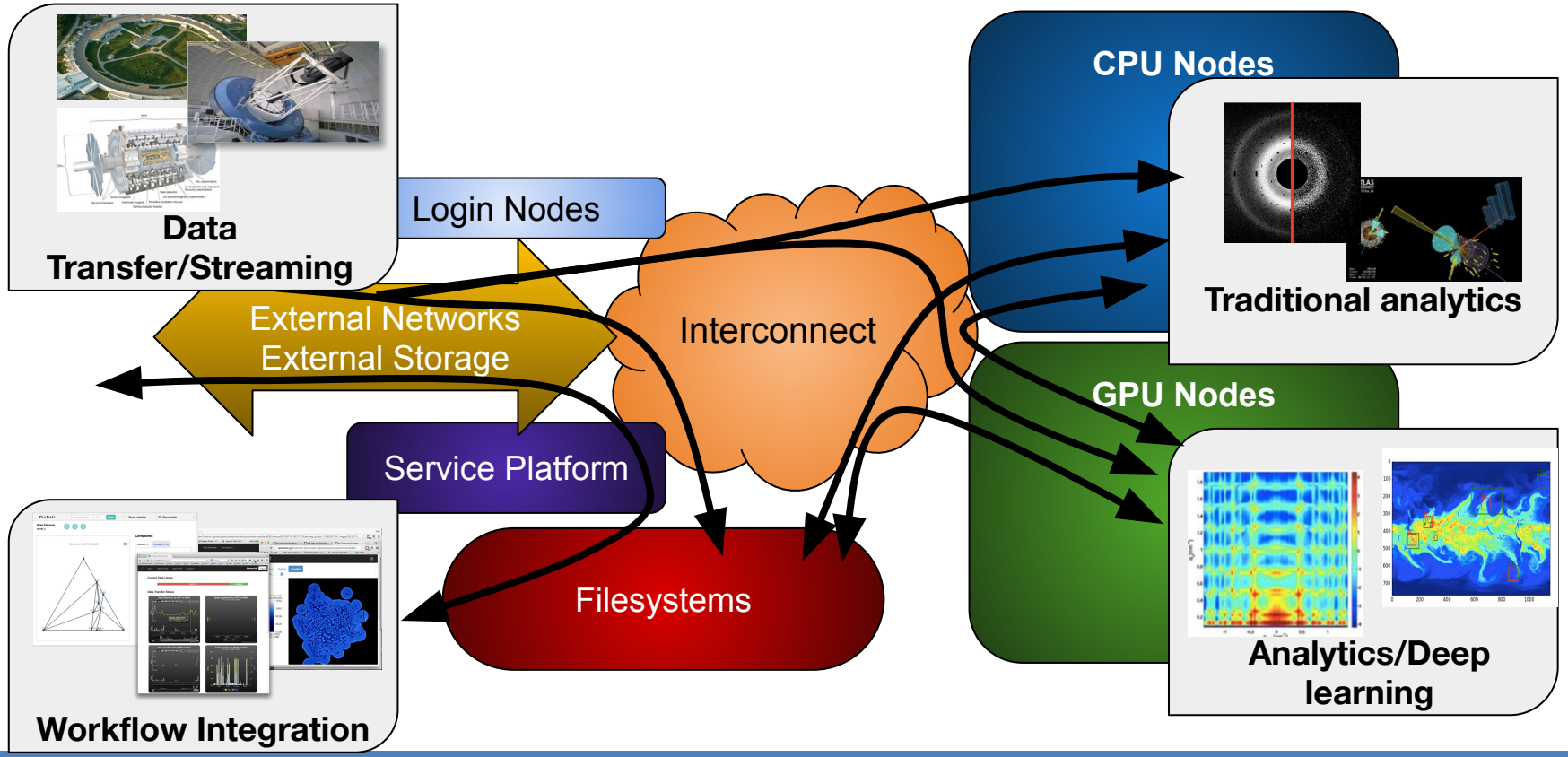
Workflows

NERSC provides a rich data *eco*system

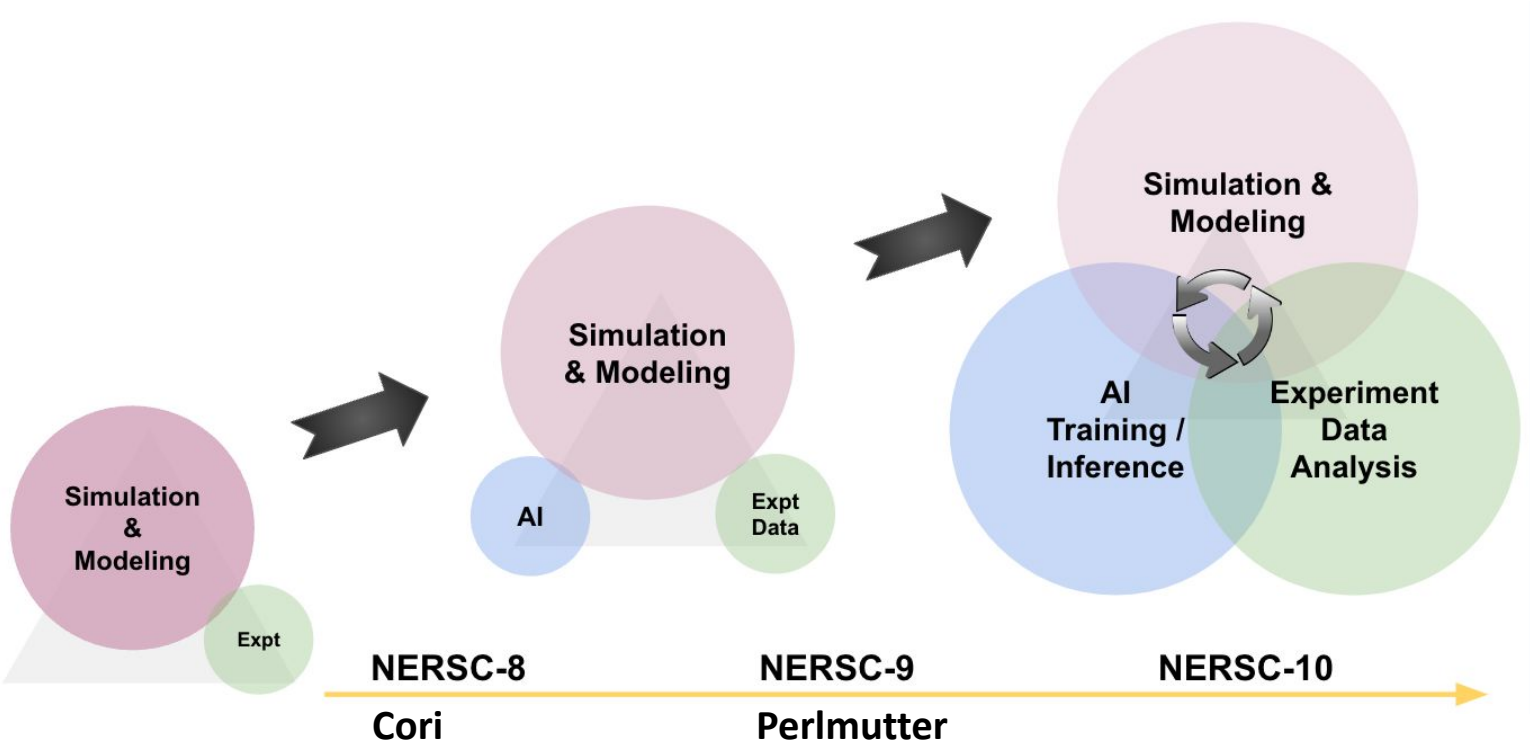


Modified and updated from WB et. al
O'Reilly Blog

Data ecosystem impacts entire workflow

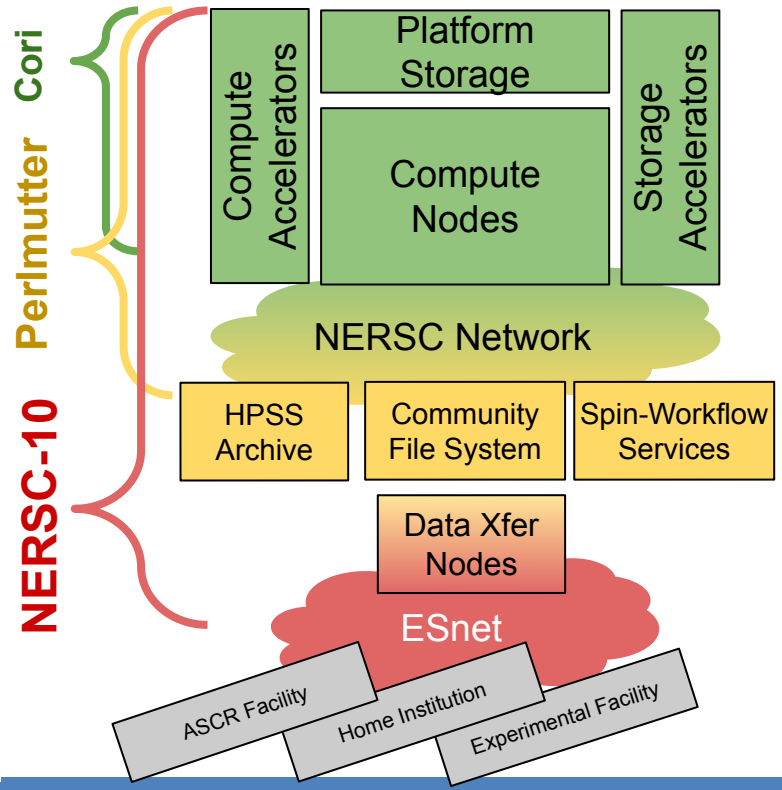


Data Directions: NERSC-10 - integrated scientific workflows



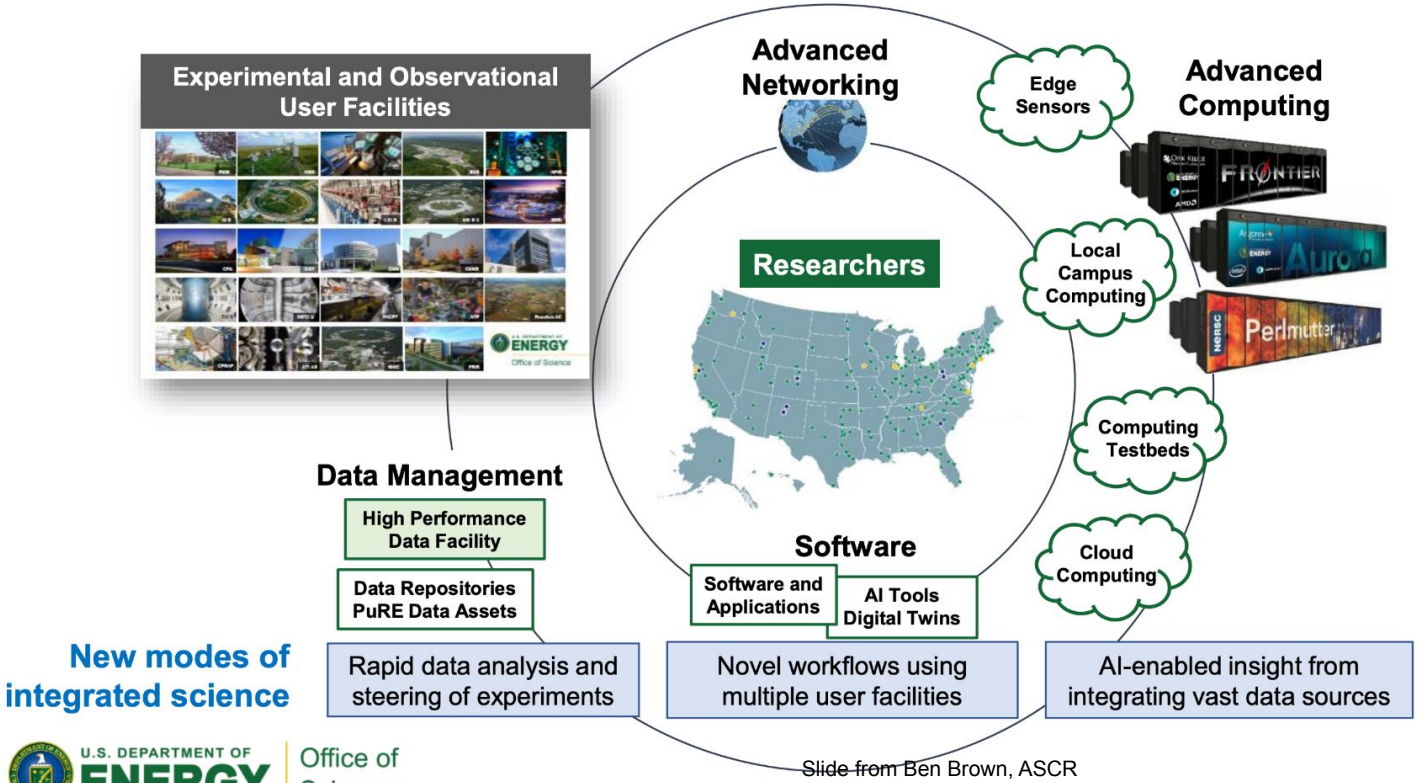
Data Directions: NERSC-10 workflows extend beyond system

- **Quality of Service** – computation, storage and networking designed to emphasize response-time plus throughput/utilization.
- **Seamlessness** – tight integration of system components to enable high performance across workflow steps.
- **Portability** – Modular workflow execution across heterogeneous HPC, edge and cloud.
- **Programmability** – APIs to manage data, execute distributed code, and interact with system resources.
- **Orchestration** – coordinate resource management across different resource domains.
- **Security** – authentication, authorization and auditing (e.g., identify proofing, access/privacy control, records of transactions).



Data Directions: IRI workflows extend beyond system

DOE's Integrated Research Infrastructure (IRI) Vision:
To empower researchers to meld DOE's world-class research tools, infrastructure, and user facilities seamlessly and securely in novel ways to radically accelerate discovery and innovation



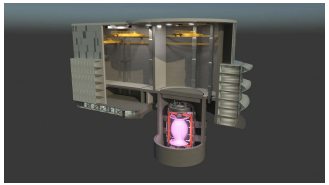
U.S. DEPARTMENT OF ENERGY | Office of Science

Slide from Ben Brown, ASCR
(https://science.osti.gov/-/media/ber/berac/pdf/202304/Brown_IRI_for_BERAC_202304.pdf)

Data Directions: Saving the science

Future circular collider, CERN

Commercial fusion reactor, UKAEA



Secrets of the universe

42+/- 6

Data Acquisition; Availability; Access; Analysis; AI; Archiving

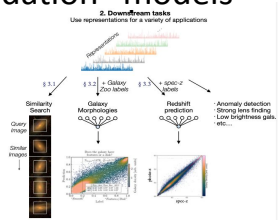
SCIENCE ENABLED ↑

- Extremely low latency data transfer and access and/or compute workflow across edge and HPC
- Compose reusable, resilient automated services and workflows
- Granular, automatically discoverable, data
- Experiment with seamless, smart, human computing interfaces
- Leverage reusable, fast AI-inference with rich science models and uncertainties
- Curate with re-interpretation and smart archiving

Data Directions at NERSC - example projects

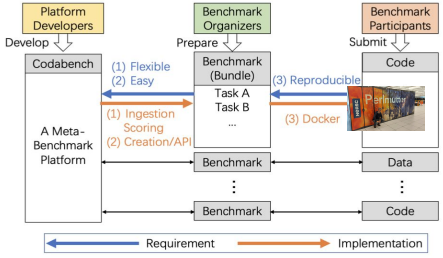
Pervasive AIScience

Large multi-purpose, "foundation" models



Sky survey self-supervised learning

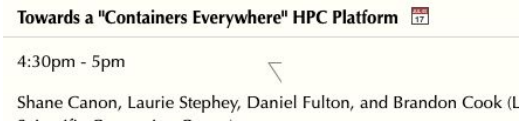
Platforms for scientific AI



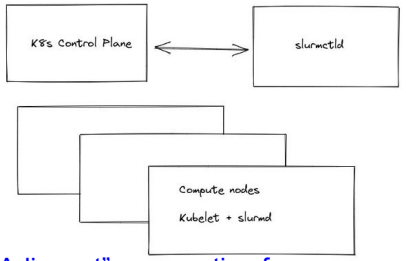
"Fair Universe"

Converged HPC+Cloud

"Containers Everywhere"

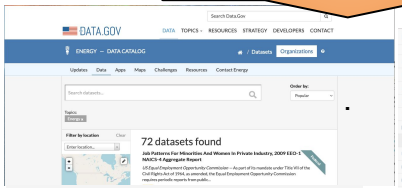


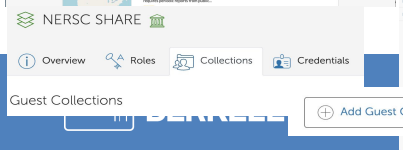
Slurm and k8s orchestration

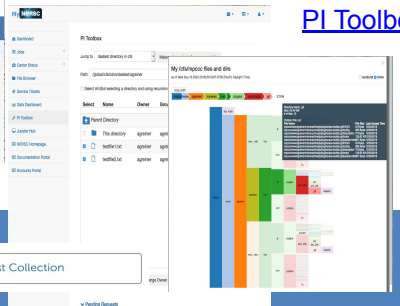


"Adjacent" perspective from SC22 SchedMD talk

FAIR Data Services







PI Toolbox

Scalable analytics

Project dragon

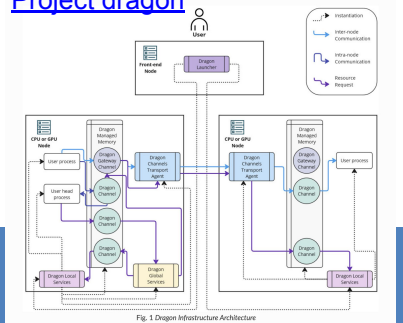
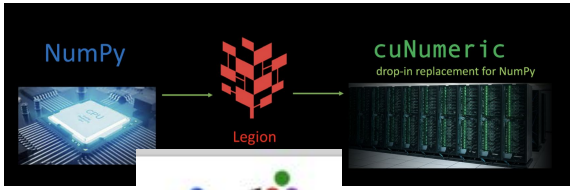


Fig. 1 Dragon Infrastructure Architecture



julia

NERSC

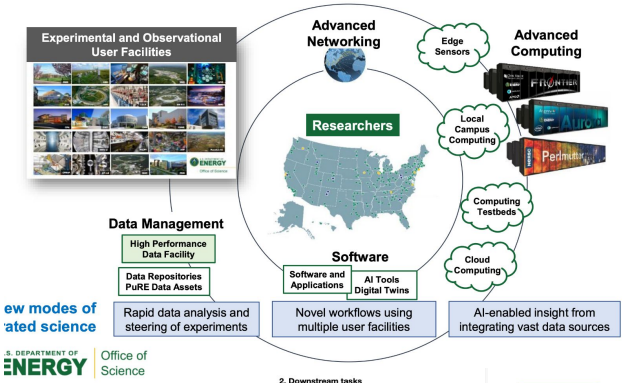
Directions in data services

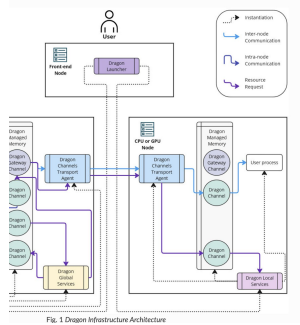
Compose services and compute seamlessly across distributed facilities

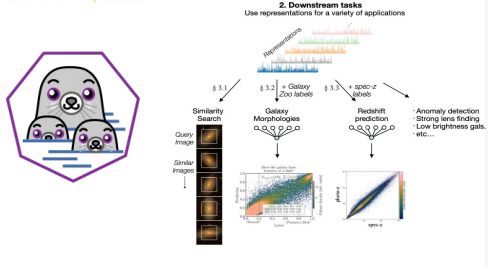
Experiment with and apply performant, productive analytics at scale

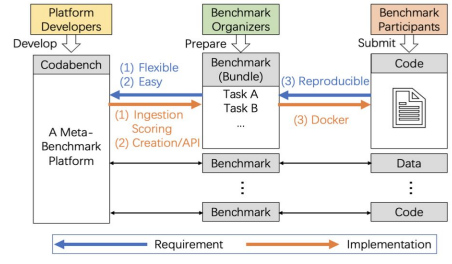
Leverage large AI models, fine-tune to new problems, apply to new data pipelines

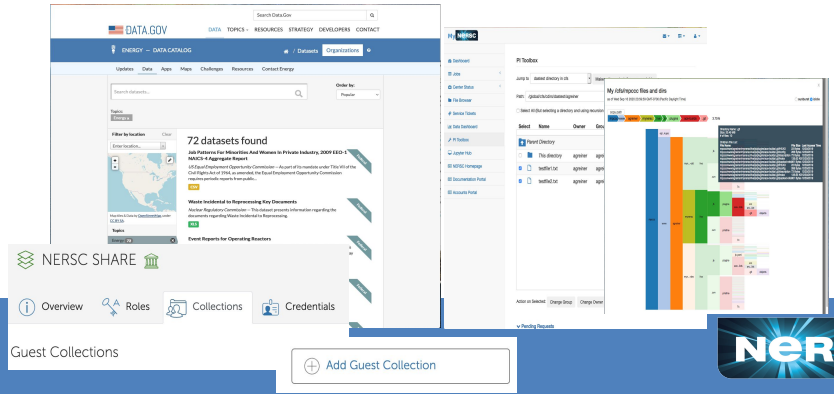
Curate and re-use data through FAIR management services











Directions in data services

Compose services and compute seamlessly across distributed facilities

Experiment with and apply performant, productive analytics at scale

Leverage large AI models, fine-tune to new problems, apply to new data pipelines

Curate and re-use data through FAIR management services

Day 1

Time (PDT)	Topic	Presenters
9:00 - 9:10 a.m.	Introduction/Welcome	
9:10 - 9:30 a.m.	Directions in Data at NERSC	Wahid Bhimji
9:30 - 10:00 a.m.	HPC Containers, Podman-HPC, Shifter	Adam Lavely
10:00 - 10:20 a.m.	Science Applications: Checkpoint/Restart in Containers	Madan Timalsina
10:20 - 10:30 a.m.	Break	
10:30 - 11:00 a.m.	Choosing the right storage for your data	Steve Leak, Ravi Cheema
11:00 - 11:30 a.m.	Distributed Python at NERSC	Daniel Margala
11:30 - 12:00 a.m.	Julia	Johannes Blaschke
12:00 - 1:30 p.m.	Lunch	
1:30 - 2:00 p.m.	Deep Learning at Scale	Shashank Subramanian
2:00 - 2:30 p.m.	Science Applications in AI	Jared Willard
2:30 - 3:00 p.m.	Nvidia Triton Demo: Incorporating AI inference into workflows	Andrew Naylor

Day 2

Time (PDT)	Topic	Presenters
9:00 - 9:30 a.m.	Integrated Research Infrastructure and N10	Debbie Bard
9:30 - 10:00 a.m.	Superfacility API	Bjoern Enders
10:00 - 10:30 a.m.	Science Applications: Data Streaming to NCEM	Sam Welborn, Peter Ercius
10:30 - 11:00 a.m.	Science Applications: EJFAT/JIRIAF	Vardan Gyuriyan, Jeng-Yuan Tsai
11:00 - 11:30 a.m.	Science Applications: JAWS	Daniela Cassol
11:30 - 12 p.m.	Data Transfers and the Globus	Nick Tyler

Questions?
Collaboration?

Wahid Bhimji
wbhimji@lbl.gov
https://docs.nersc.gov/analytics/analytics/
https://docs.nersc.gov/machinelearning/
https://docs.nersc.gov/services/

And questions for you:
What are you hoping to get out of this data day?
What would you have liked to have on the agenda that you don't see?

Questions

What are you hoping to get out of this data day?

What would you have liked to have on the agenda that you don't see?