



Overview of AI at Fermilab

Nhan Tran

February 23, 2022

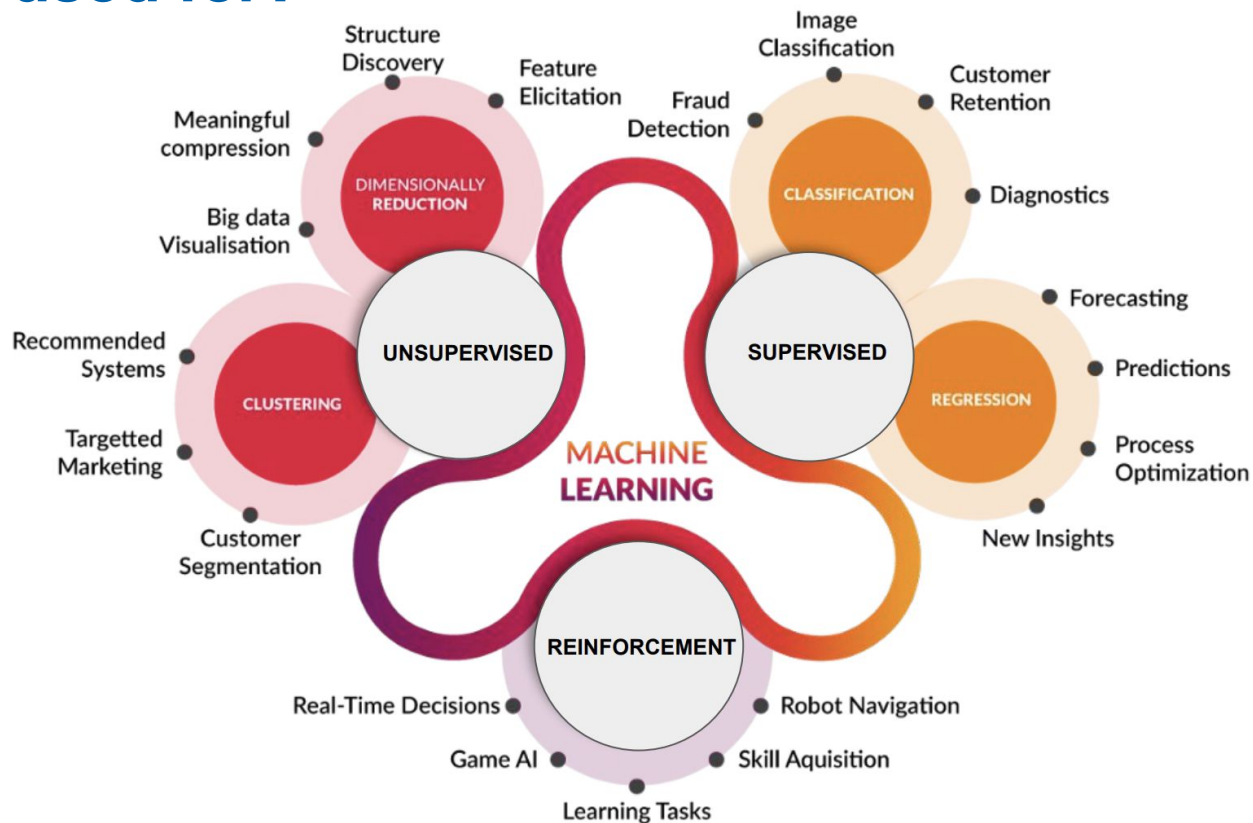
All-Engineers Retreat 2022

Contents

AI Strategy at Fermilab

AI Highlights

What is it used for?



Artificial Intelligence

AI for Physics, Physics for AI



At Fermilab, we are committed to artificial intelligence (AI) research and development to enhance the scientific mission of particle physics.

The unique challenges of high-energy physics research present opportunities for advancing AI technologies. From massive and rich data sets to building and operating some of the world's most complex detector and accelerator systems, the technologies we are developing have potential connections to a broad domain of cutting-edge AI

research.

Fermilab's Artificial Intelligence Project aims to

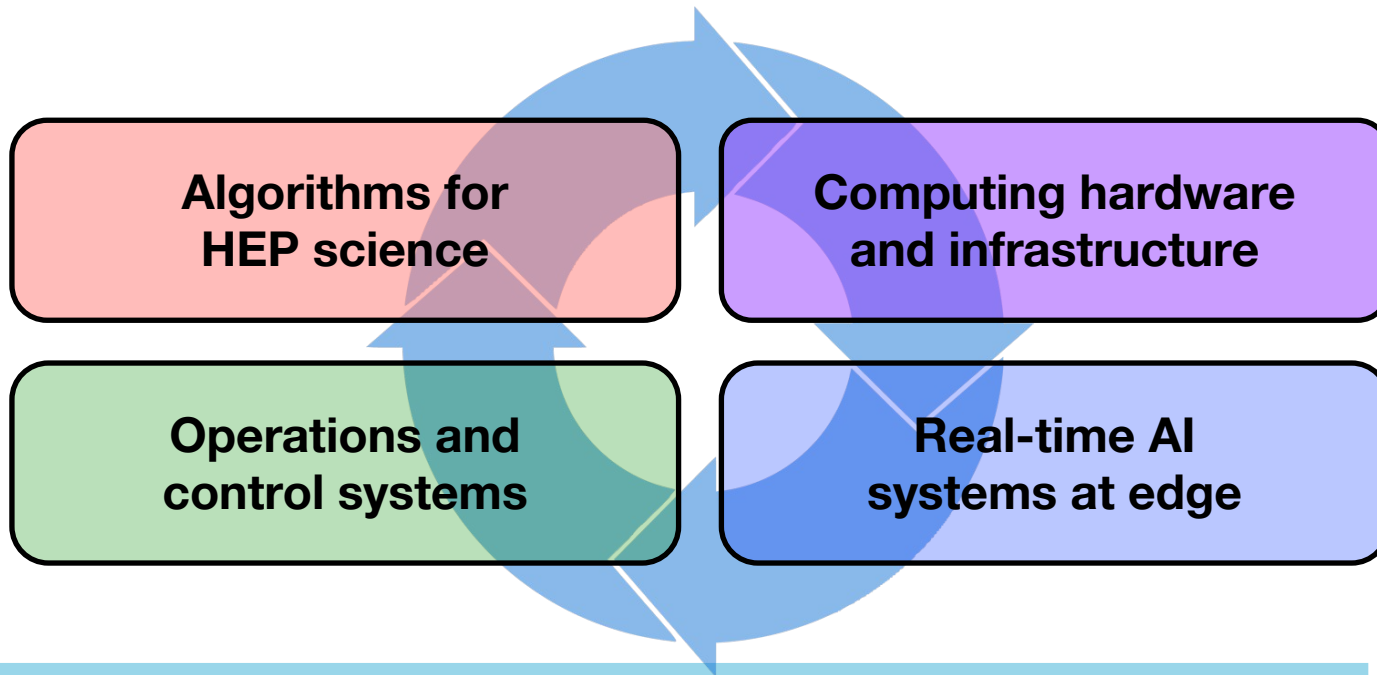
- Accelerate high energy physics research with the goal of solving the mysteries of matter, energy, space and time
- Develop AI capabilities within the national ecosystem that build on high-energy physics challenges and technologies
- Build community around cross-cutting problems in order to share the work of Fermilab and the high-energy physics community's AI work with the world

Mission statement

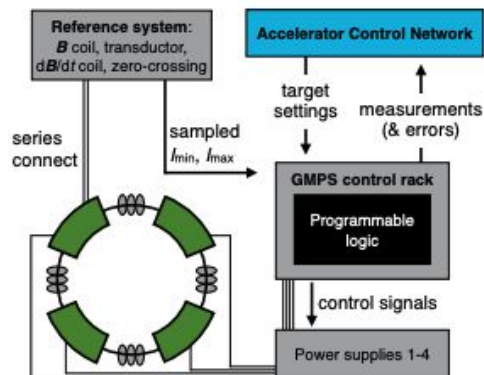
DOE HEP builds and operates among the hardest and biggest projects with the most complex devices in science -- accelerators and detectors. Our priority is using AI for real-time controls, operations, and data processing to **accelerate HEP science**.

Mission statement

DOE HEP builds and operates among the hardest and biggest projects with the most complex devices in science -- accelerators and detectors. Our priority is using AI for real-time controls, operations, and data processing to **accelerate HEP science**.



Artificial Intelligence - Quad Chart



Key Current Activities

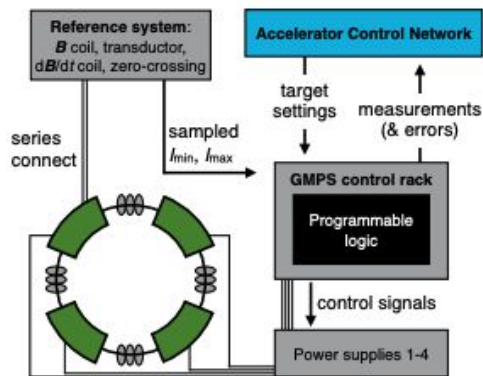
New and on-going Accelerator AI projects

Intelligent real-time systems, hardware, and co-design

AI methods for HEP (UQ, representations, generative)

AI integration into HEP workflows

Artificial Intelligence - Quad Chart



Key Current Activities

New and on-going Accelerator AI projects

Intelligent real-time systems, hardware, and co-design

AI methods for HEP (UQ, representations, generative)

AI integration into HEP workflows

Future Activities / Roadmap

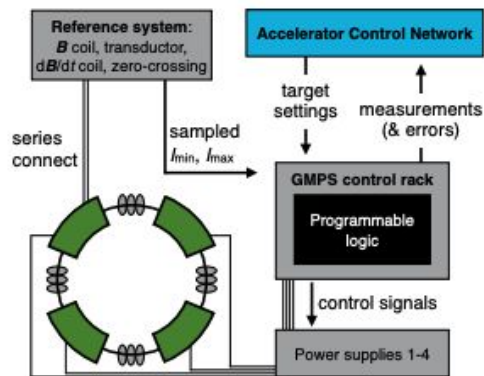
Integration w/accelerator projects (PIP-II, ACORN,...)

Build hub for intelligent hardware systems

Foster community to develop novel AI methods in HEP

Scalable AI computing infrastructure

Artificial Intelligence - Quad Chart



Key Current Activities

New and on-going Accelerator AI projects

Intelligent real-time systems, hardware, and co-design

AI methods for HEP (UQ, representations, generative)

AI integration into HEP workflows

Future Activities / Roadmap

Integration w/accelerator projects (PIP-II, ACORN,...)

Build hub for intelligent hardware systems

Foster community to develop novel AI methods in HEP

Scalable AI computing infrastructure

Prioritized Needs

Hire and train AI algo & operations experts

Engage diverse Fermilab and broader AI community

Develop computing support for mid-sized AI projects

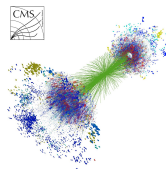
Seed resources to dev, build, demo AI capabilities

Aligning with National AI strategic plan

THE NATIONAL
ARTIFICIAL INTELLIGENCE
RESEARCH AND DEVELOPMENT
STRATEGIC PLAN: 2019 UPDATE

Make Long-Term Investments in AI Research

- Learn new representations of physics data, applying novel AI structures at-scale
- High performance and high throughput computing for massive data challenges
- Advancing automated design tools and hardware for real-time AI



Develop Effective Methods for Human-AI collaboration

- Building operations and controls systems to aid operators in complex, low latency systems



Understand and Address the Ethical, Legal, and Societal Implications of AI

- Techniques to decorrelate observables and reduce systematic biases

Ensure the Safety and Security of AI Systems

- Uncertainty quantification for real science; understanding/visualizing what the AI is learning

Develop Shared Public Datasets and Environments for AI Training/Testing

- Open-source tools for AI-on-chip adopted outside of particle physics; public data sets being adopted by other domains



Better Understand the National AI R&D Workforce Needs

Expand Public-Private Partnerships to Accelerate Advances in AI

Contents

AI Strategy at Fermilab

AI Highlights

**Algorithms for
HEP science**

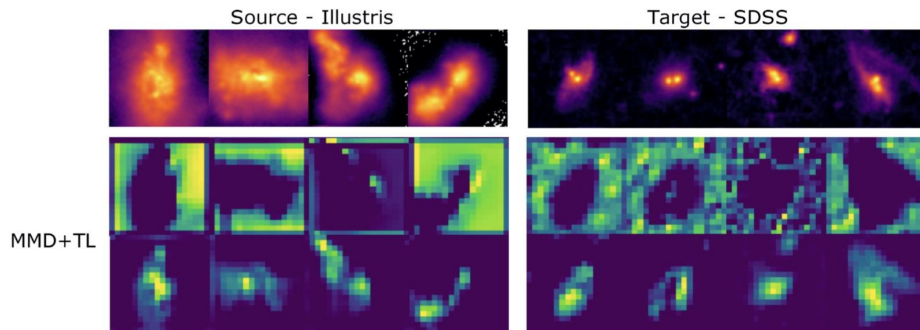
**Computing hardware
and infrastructure**

**Operations and
control systems**

**Real-time AI
systems at edge**

Cutting-edge AI for HEP

A lot of amazing work happening here! Hard to capture in briefly - here's a couple examples:

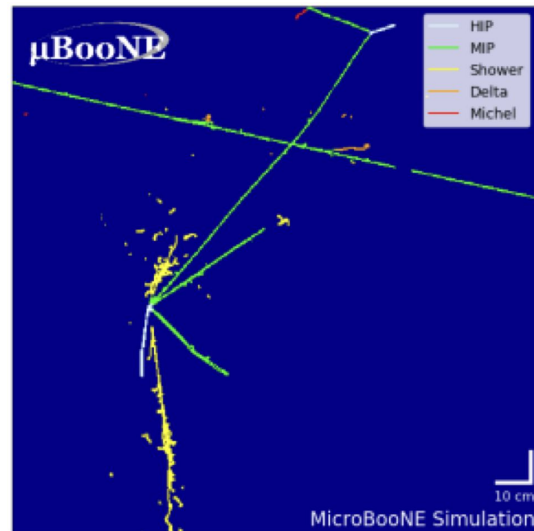


Domain adaptation for training galaxy merger models with data vs. simulation differences

A. Ciprianovic, D. Kafkes et al, arxiv:2103.01373

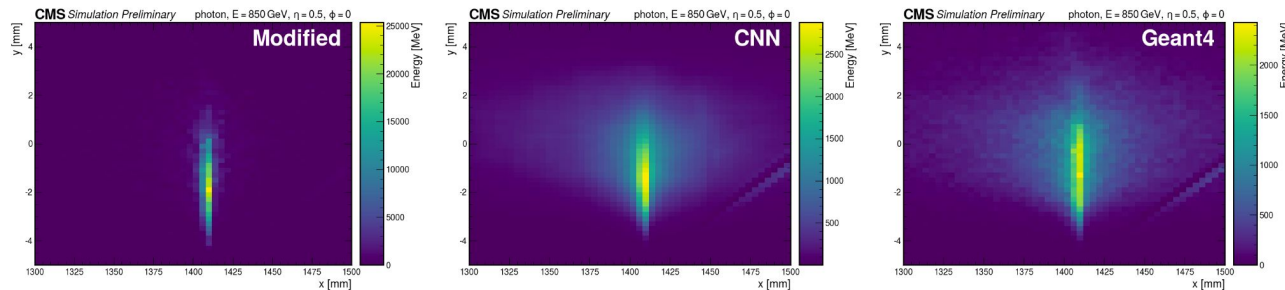
Semantic segmentation with submanifold sparse convolutional networks

<https://arxiv.org/abs/2012.08513>



Cutting-edge AI for HEP

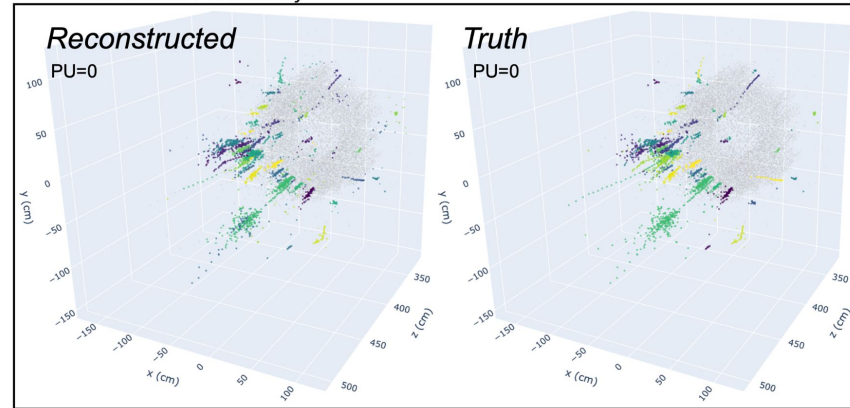
A lot of amazing work happening here! Hard to capture in briefly - here's a couple examples:



Denoising for accelerated simulation at the LHC
K. Pedro et al.

Point cloud clustering with Graph NNs for CMS HGCal
L. Gray, T. Klijnsma, K. Pedro, G. Pradhan

CMS Simulation Preliminary



AI for (accelerator) operations

Recent accelerator operations mini-workshop:

<https://indico.fnal.gov/event/52417/>

High level vision

Real-time autonomous edge and
centralized beam controls

Intelligent robotics for
automated monitoring

Large scale simulation of
accelerator (digital twin)

Complex-wide monitoring for
predictive maintenance and fault
detection

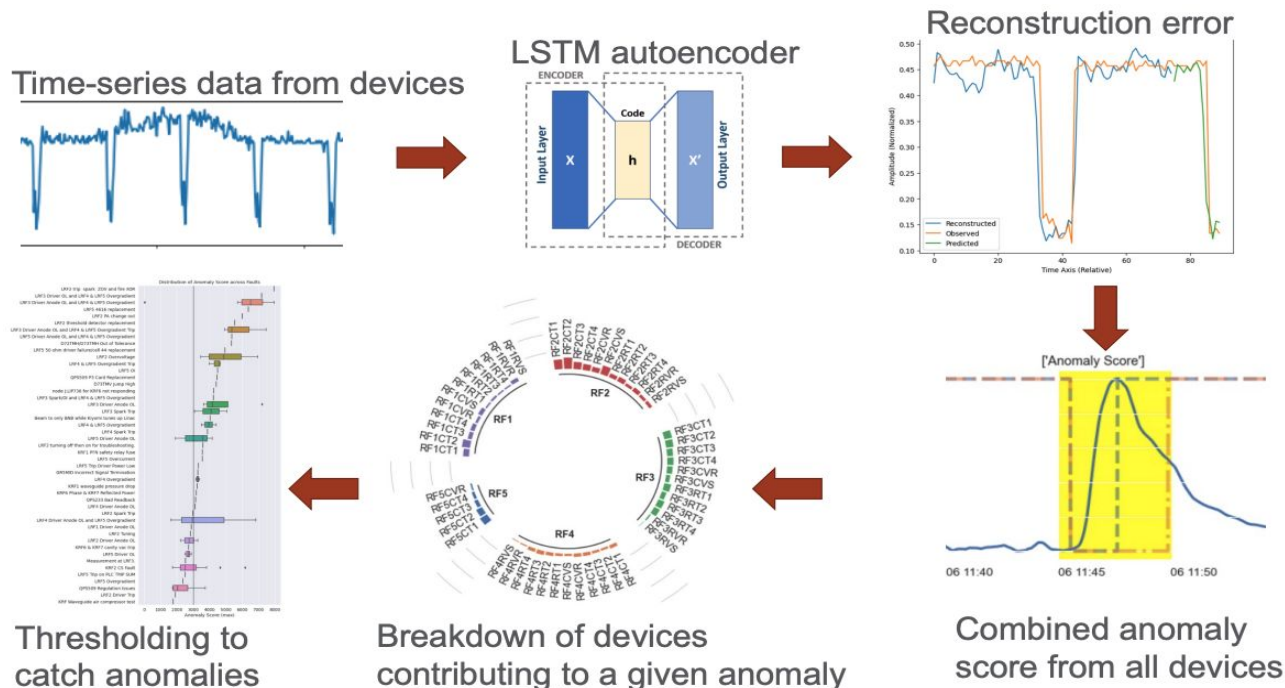
Explainable algorithms including
uncertainties and robust
safeguards for operators

Example: big data analytics

[More details, see slides here](#)

Recent accelerator operations mini-workshop:

<https://indico.fnal.gov/event/52417/>

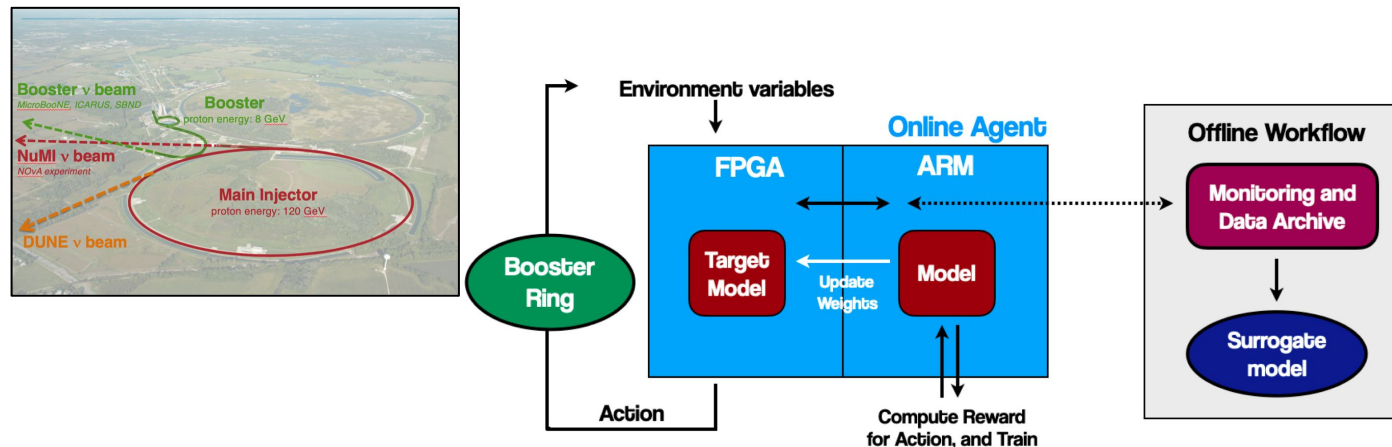


Example: real-time controls

[More details, see slides here](#)

Recent accelerator operations mini-workshop:

<https://indico.fnal.gov/event/52417/>

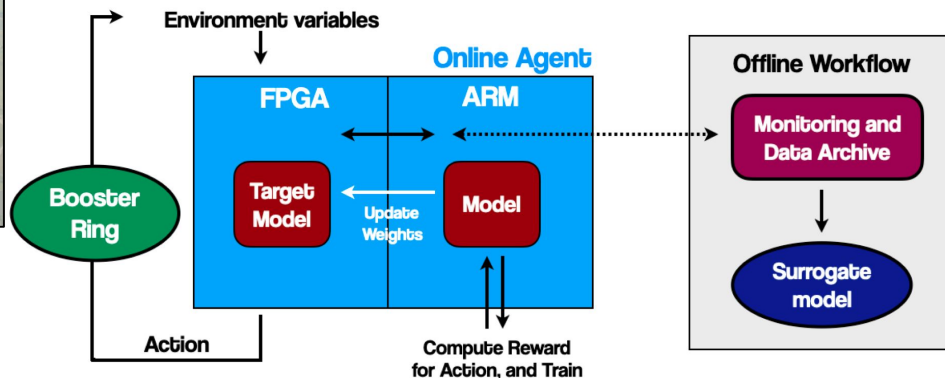
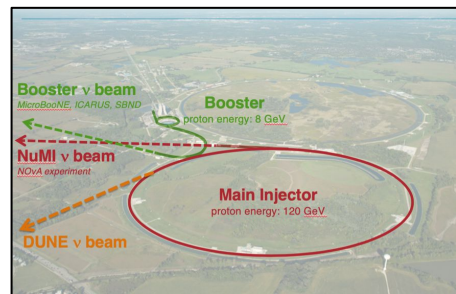


Example: real-time controls

[More details, see slides here](#)

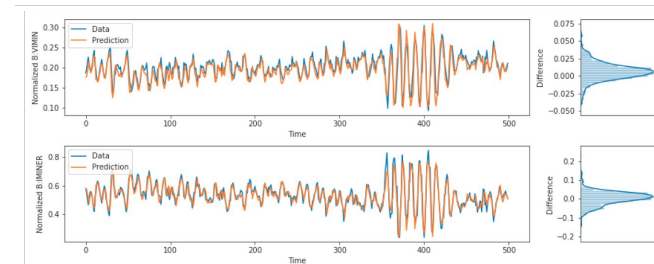
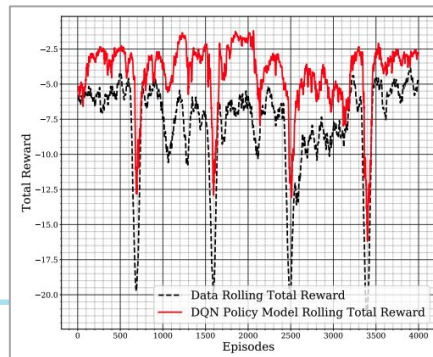
Recent accelerator operations mini-workshop:

<https://indico.fnal.gov/event/52417/>



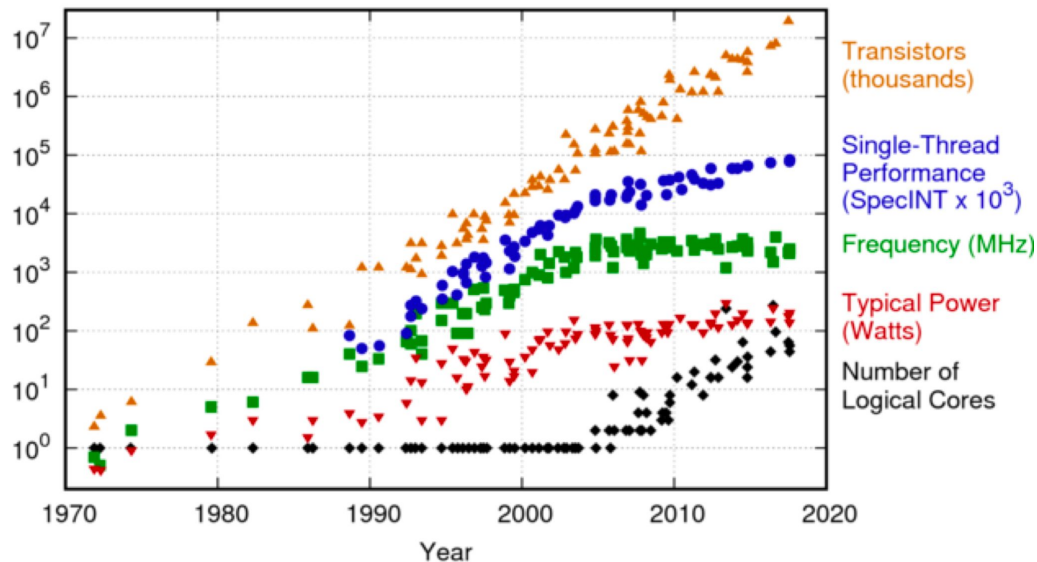
Digital twin

online
model
reward



Advances in computing

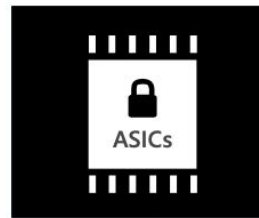
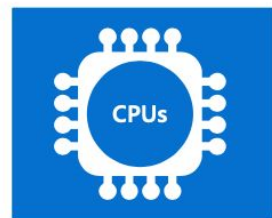
42 Years of Microprocessor Trend Data



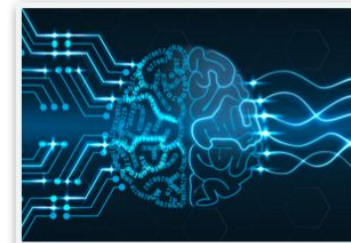
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
 New plot and data collected for 2010-2017 by K. Rupp



Advances in computing




Advances in
heterogeneous
computing driven by
machine learning



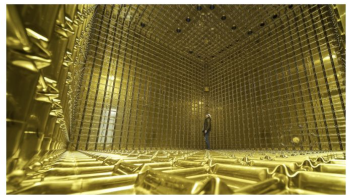
Advances in computing

[Nvidia link](#)
[Microsoft link](#)


NVIDIA. DEVELOPER

[HOME](#)
[BLOG](#)
[FORUMS](#)
[DOCS](#)
[DOWNLOADS](#)
[TRAINING](#)

DEVELOPER BLOG



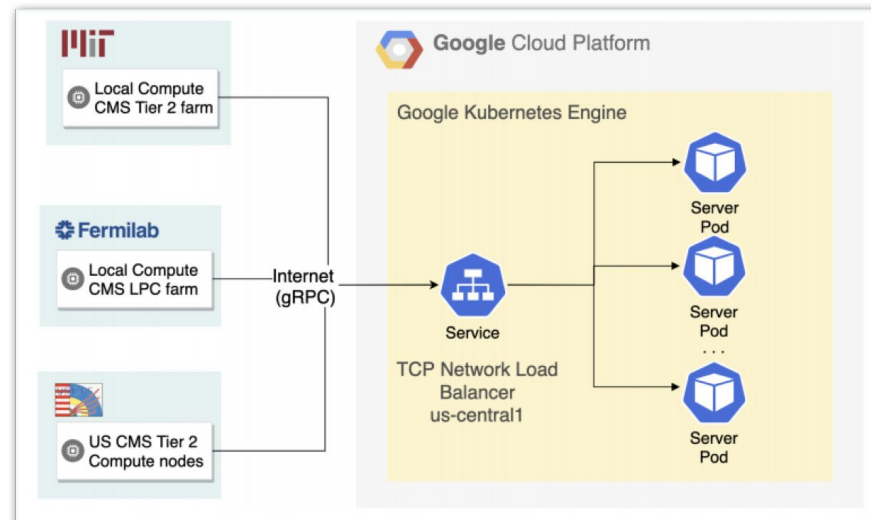
AI / DEEP LEARNING | HPC

Apr 30, 2021

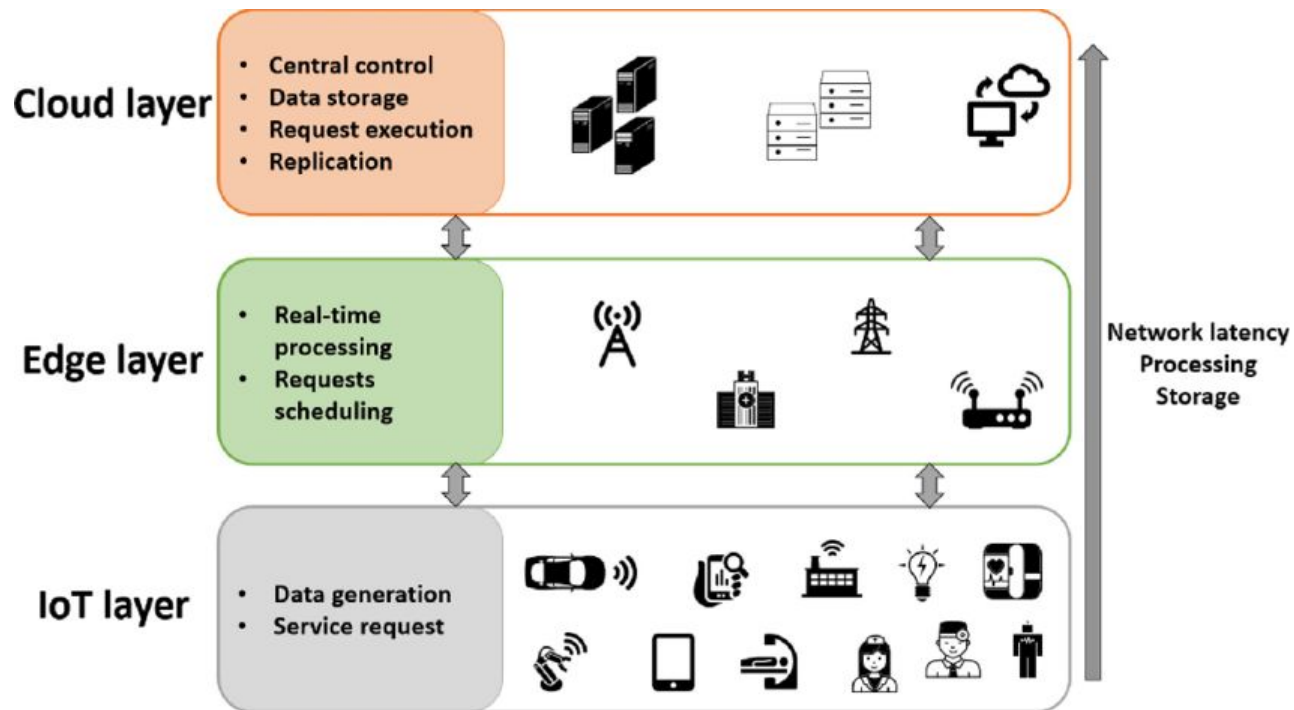
Scaling Inference in High Energy Particle Physics at Fermilab Using NVIDIA Triton Inference Server

By Shankar Chandrasekaran, Lindsey Gray, Farah Hariri, Kevin Pedro, Vartika Singh, Nhan Tran, Mike Wang and Tingjun Yang

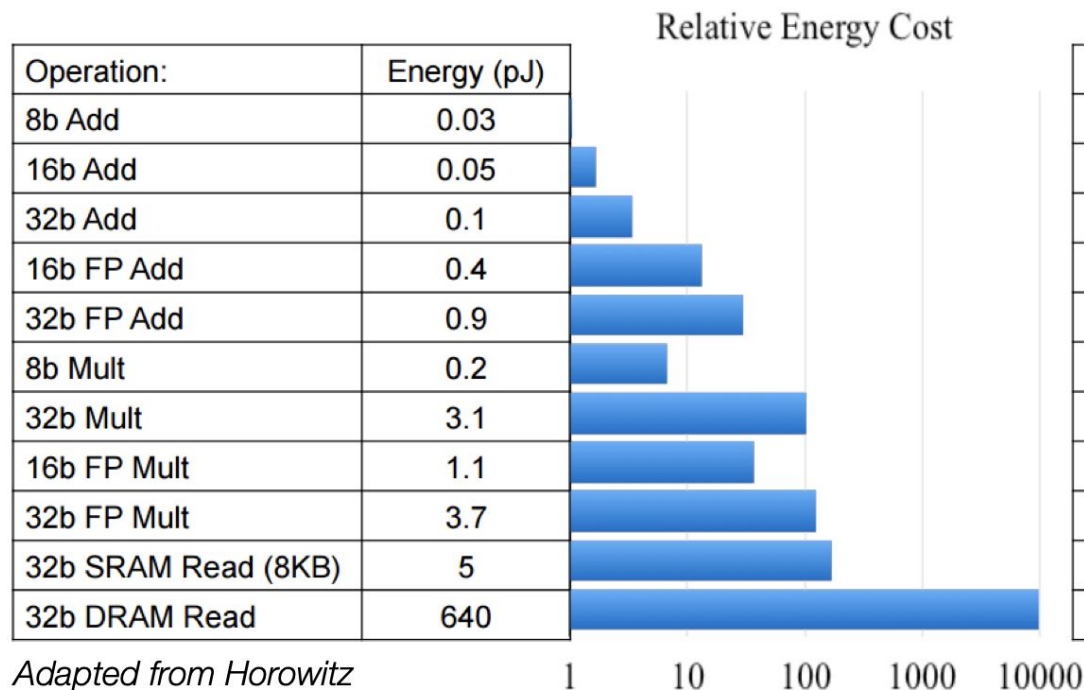
Tags: featured, kubernetes, NGC, physics, Triton



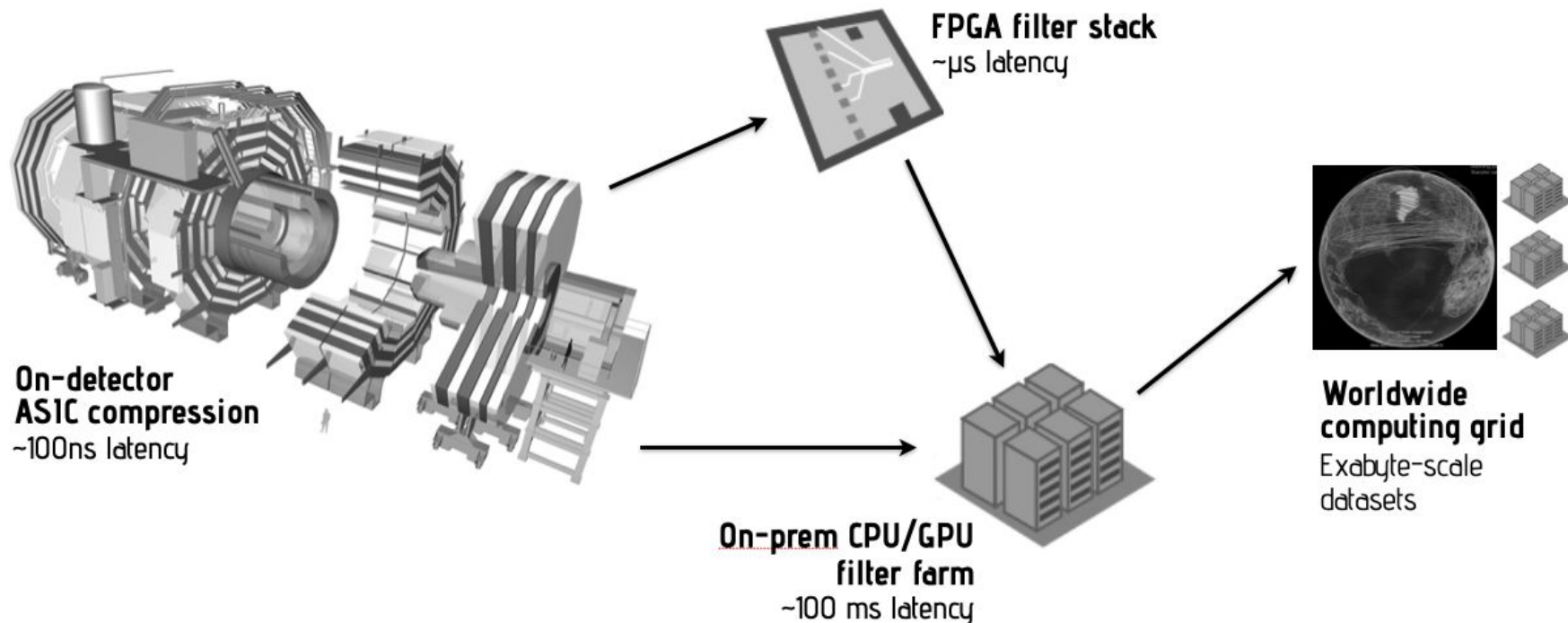
Intelligent devices and the Internet of Things



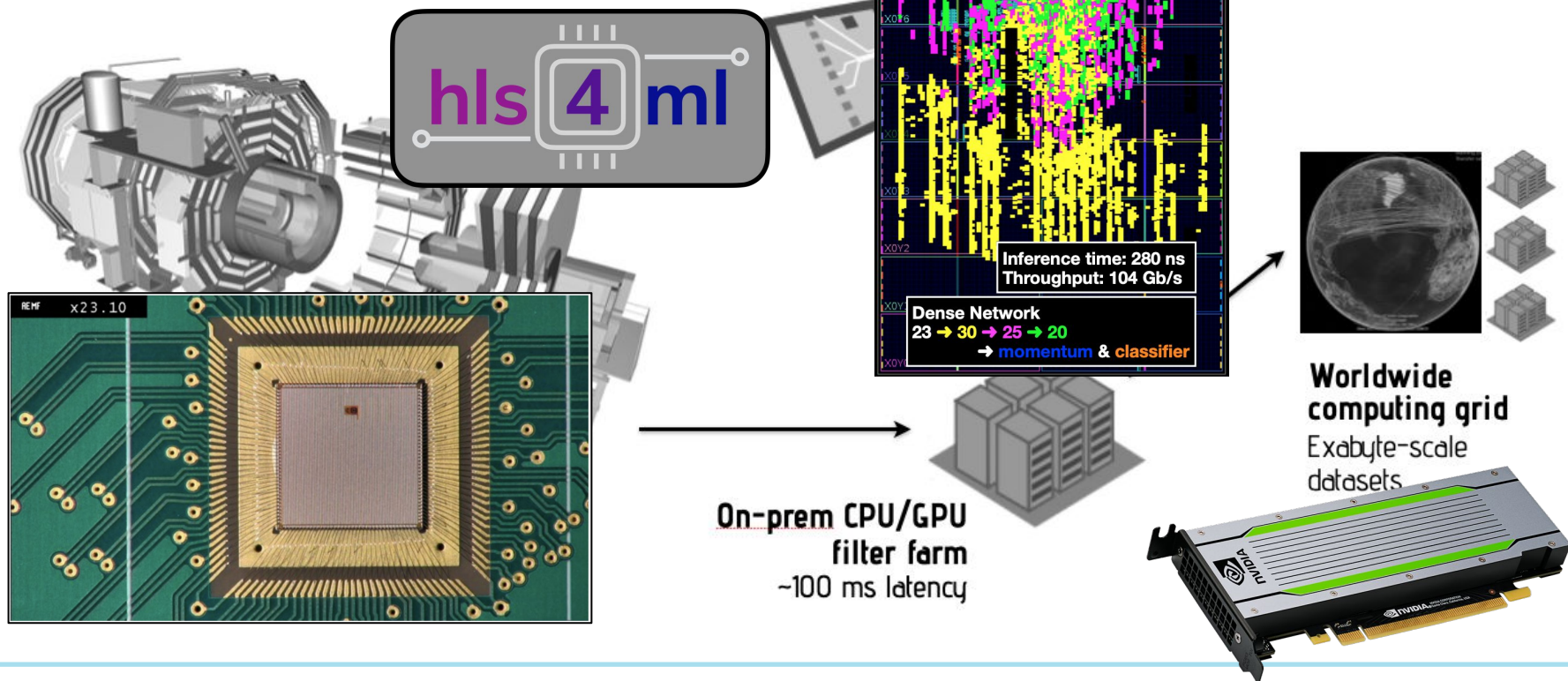
Intelligent devices and the Internet of Things



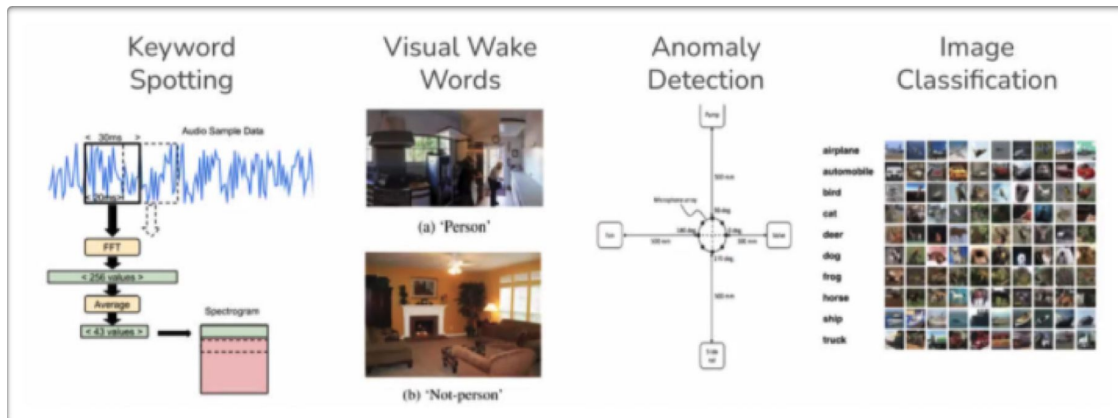
Intelligent devices



Intelligent devices



Intelligent devices



MLCommons launches machine learning benchmark for devices like smartwatches and voice assistants

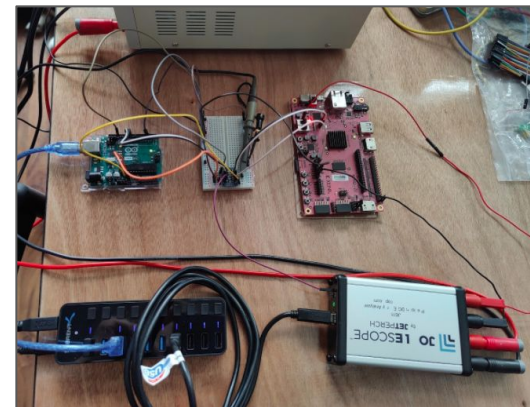
by Ben Wodecki 6/16/2021



With experts from Qualcomm, Fermilab, and Google aiding in its development

MLCommons, the open engineering consortium behind the MLPerf benchmark test, has launched a new measurement suite aimed at 'tiny' devices like smartwatches and voice assistants.

MLPerf Tiny Inference is designed to compare performance of embedded devices and models with a footprint of 100kB or less, by measuring



hls4ml only "open division" submission, competitive results on a Pynq-Z2 SoC platform!

Next submission with Xilinx Research Labs collaboration

Outlook

Outlook

Fermilab AI program has grown significantly in past 2 years

Using AI to advance our scientific mission

Fermilab unique AI capabilities can contribute to greater AI community

If you're interested...

It's been difficult to meet in person during the pandemic but we would like to hear from you - email at ai@fnal.gov or visit ai.fnal.gov and contact one of our liaisons

We have had some events in the past like [Al jamboree](#) and [tutorials](#), looking to restart seminars and other get togethers - let us know if you have ideas!



R&D Priority Areas and Practices

Align with national AI strategy

Semiconductors

- **Real-time AI** for FPGAs, ASICs, and beyond CMOS (neuromorphic, photonics,...)

Support R&D for future high performance computing

- Leading role in **heterogeneous compute** for accelerating AI inference

Build, strengthen, and expand strategic multi-sector partnerships

- **Collaborations across many sectors**, physics and outside (computer science, electrical engineering,...)
Strong ties with Northwestern (ECE) and UChicago/TTIC (CS); many collaborations with other labs and universities; engaging with industry partners (Microsoft, Xilinx, etc.)

HEP Overlap with White House FY 2021 R&D Priority Areas and Practices

- **Semiconductors: Working in collaboration with industry and academic partners, where appropriate**
 - Prioritize investments that will enable whole of government access to trusted and assured microelectronics for future computing and storage paradigms
- **Artificial Intelligence, Quantum Information Science, and Computing:**
 - Prioritize basic and applied research investments that are consistent with 2019 Executive Order on Maintaining American Leadership in Artificial Intelligence and the 5 strategies detailed in 2019 update of the National Artificial Intelligence Research and Development Strategic Plan
 - Prioritize R&D advancing fundamental QIS, building and strengthening the workforce, engaging industry, and providing infrastructure supporting QIS while coordinating relevant activities to ensure intelligence, defense, and civilian efforts grow synergistically
 - Explore new applications in and support R&D for high performance future computing paradigms, fabrication, devices, and architectures alongside sustainable and interoperable software; data maintenance and curation; and appropriate security.
- **Build and Leverage a Diverse, Highly Skilled American Workforce**
 - Prioritize efforts to build strong foundations for STEM literacy, to increase diversity, equity, and inclusion, and to prepare the STEM workforce, including college-educated STEM workers and those working in skilled trades that do not require a four-year degree
 - Build R&D capacity at institutions that serve high proportions of underrepresented or underserved groups
- **Support Transformative Research of High Risk and Potentially High Reward**
 - Support risk taking in their R&D investments and within the communities they support, and they should ensure that review processes fully consider the possible rewards, risks, and benefits of failure for potentially transformative research.
- **Build, Strengthen, and Expand Strategic Multisector Partnerships**
 - Partnerships with academic institutions, established and startup businesses nonprofit institutions, and others involved in the U.S. S&T enterprise are instrumental to building and leveraging our Nation's innovation capacity and lie at the core of success for the Second Bold Era of S&T.
 - Prioritize investments and policies that facilitate or strengthen multisector partnerships, including partnerships that engage institutions seeking to build S&T capacity

U.S. DEPARTMENT OF ENERGY Office of Science 21 November 2019 HEP Budget Planning and Execution 49

Support transformative research of high risk and high reward

Build and leverage diverse and skilled American workforce