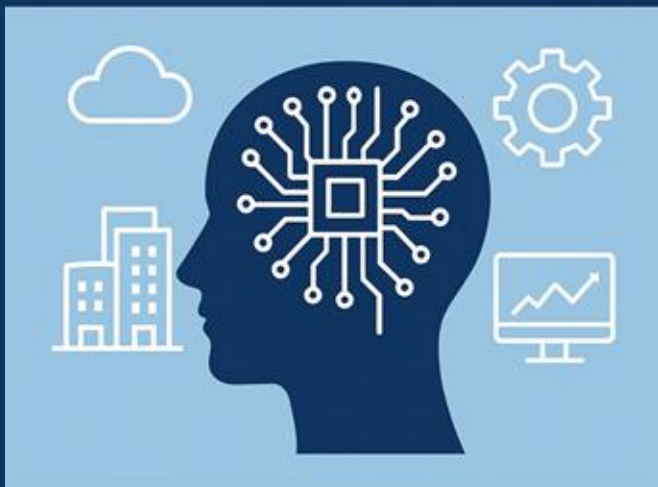


VIRTUAL WORKSHOP

AI Meets CI

Intelligent Infrastructure
for Major & Midscale
Facilities



NSF CI Compass 2026 Virtual Workshop: AI meets CI

Panel I: Adoption of AI in MF: AI methods

January 12, 2026

Panel I: Adoption of AI in MF: AI methods

Discuss **current and future plans for AI adoption in NSF Research Infrastructures** (e.g. IceCube, Rubin Observatory, CMS, NHERI) to enhance discovery, operations, data management, and decision-making.

Goal: Gain insights into current adoption strategies, highlight adoption challenges specific to large-scale scientific infrastructure, and identify opportunities for cross-facility collaboration in building intelligent cyberinfrastructure ecosystems.

Panel I: Adoption of AI in MF: AI methods



Benedikt Riedel
(IceCube)



Leanne Guy
(Rubin/LSST)



Javier Duarte
(CMS)



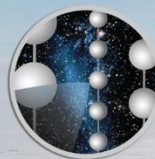
Krishna Kumar
(TACC, NHERI)

AI from the Edge of the World to the Edge of the Universe

Benedikt Riedel
UW-Madison

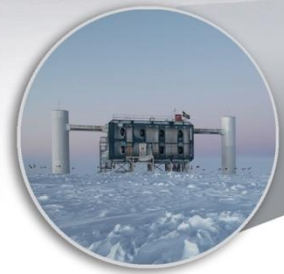
CI Compass
12 January 2026





ICECUBE

SOUTH POLE NEUTRINO OBSERVATORY



IceCube Laboratory
Data is collected here and sent by satellite to the data warehouse at UW–Madison



Digital Optical Module (DOM)
5,160 DOMs deployed in the ice

50 m

1450 m

2450 m

IceTop

86 strings of DOMs,
set 125 meters apart

IceCube
detector

DeepCore

Antarctic bedrock

Amundsen–Scott South Pole Station, Antarctica
A National Science Foundation-managed research facility

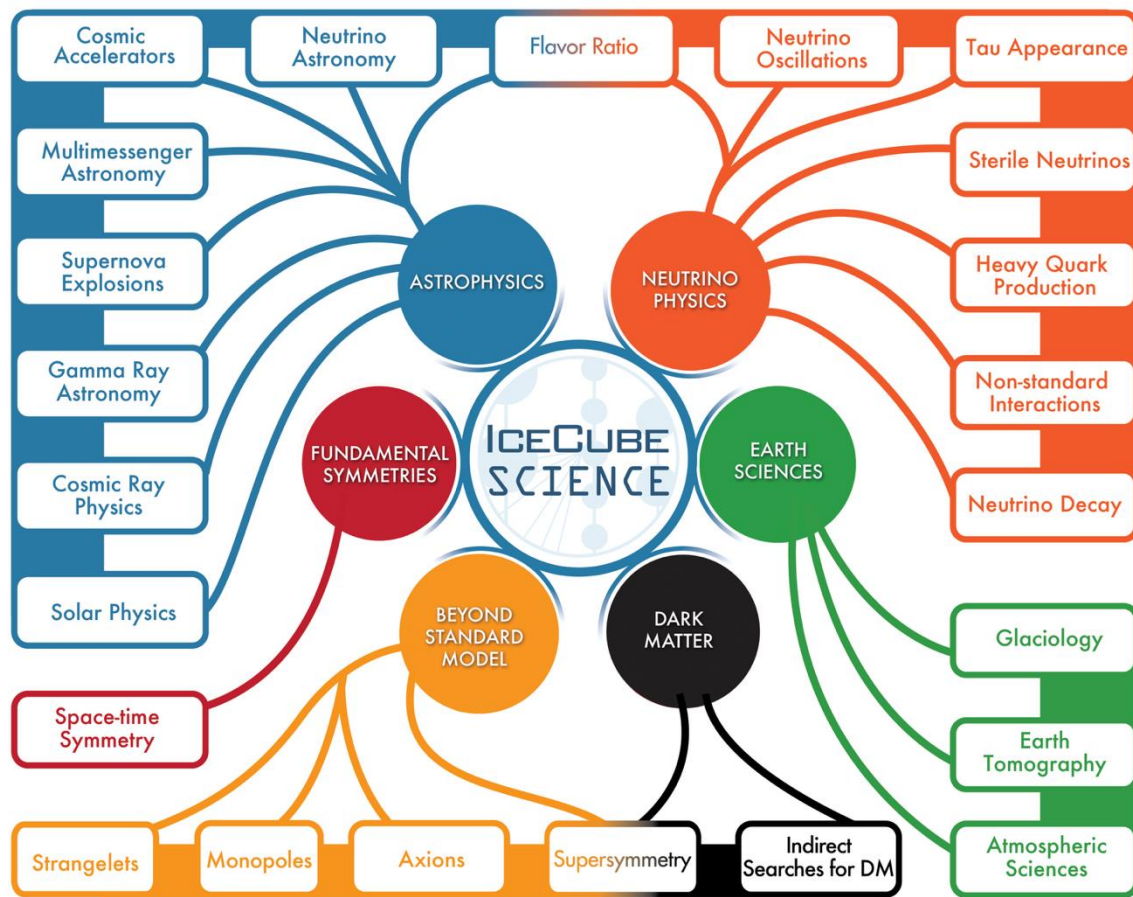


60 DOMs
on each
string

DOMs
are 17
meters
apart

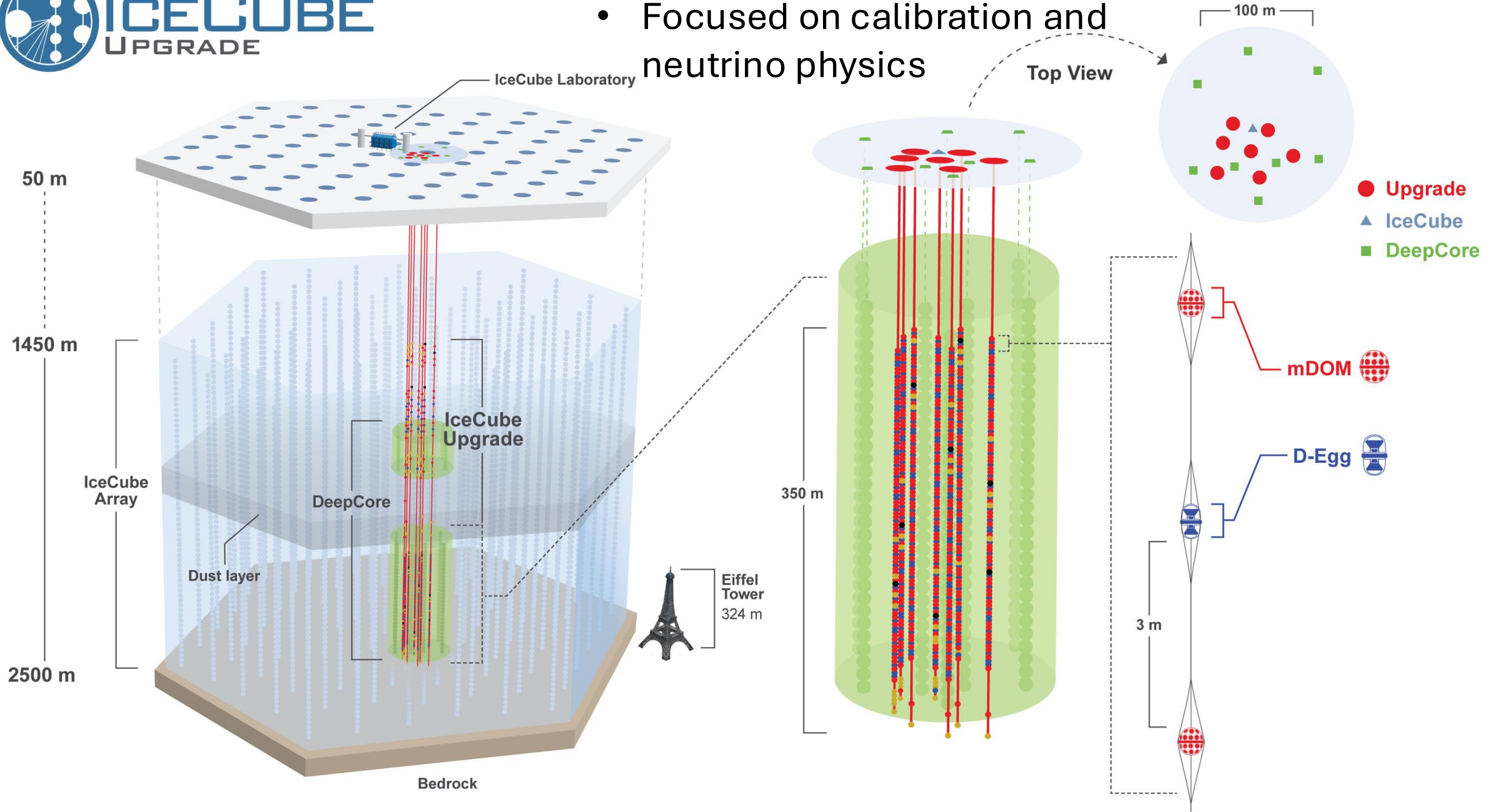


IceCube Science



- Novel instrument in multiple fields
- Broad science focus – MeV to EeV energies
 - Particle Astrophysics: Neutrino Astronomy, Multi-Messenger Astrophysics, CCSN
 - Particle Physics: Neutrino Properties
- High uptime: 99% for over 15 years
- Funded by NSF through MREFC program, Office of Polar Programs, and MPS

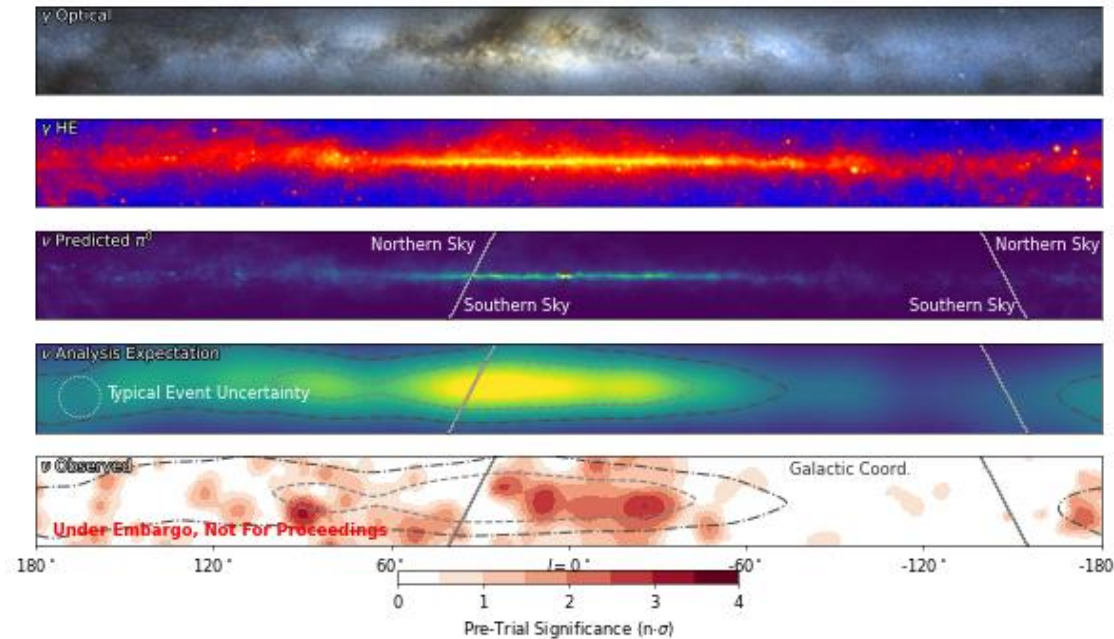
- Deploying right now
- Focused on calibration and neutrino physics



AI from the Scientist Lens

- AI Training & Inference – AI for everything all the time
 - Any new method in development (particular for Upgrade) is either AI driven or a hybrid – “Classical” methods are reaching a cost vs. benefit threshold
- Many free tools available (at least in the US)
 - Github Copilot
 - Gemini
 - Microsoft Copilot
- Great tool to accelerate your code development – “Give me a code snippet to plot X”

A new View of our Galaxy



- Using Cascade events for “large” (e.g. Milky Way) or isolated sources a good option for Southern Sky point source search – Clear differentiation from background
- Maximum-Likelihood + parametrized encoding method for cascade pointing insufficient to find a source
- Hybrid approach with ML encoding information for Maximum-Likelihood method improved resolution – Simulation-based inference
- Third High Energy Neutrino Point Source! – Galactic Plane

AI from the Hardware CI Lens

- Big difference between Training and Inference
 - Training – Essentially your traditional HPC application
 - Inference – High throughput oriented (LLM inference has HPC-like hardware)
- In-house GPU resources focused on simulation needs (FP32 compute)
 - Resources are largely a decade old, handful of more modern cards
 - High-end “AI” GPUs are not cost effective for FP32 compute – Demand for multi-GPU or multi-node training is limited
 - High-end “AI” GPUs through allocation or institutional resources
- GPU utilization is a big issue – Effectively using expensive resources
 - Significant software/workflow effort to fully utilize training resources
 - AI Inference as a Service/SONIC/IaaS as part of A3D3 to address utilization
- AI Inference at the South Pole – How to integrate into DAQ, on-site processing at remote power constraint site

AI from the Software CI Lens

- Great tool to accelerate your code development
 - “Write me a workflow management observation system for Pegasus.”
 - Faster iteration through features, software upgrades, etc.
- How good, safe, user-friendly, etc. is the AI code?
- A lot of possibilities: Can we feed it our logs to make sense of them? Can we create AI Agents that act on the logs?
- Do we have time or funding to explore the options?