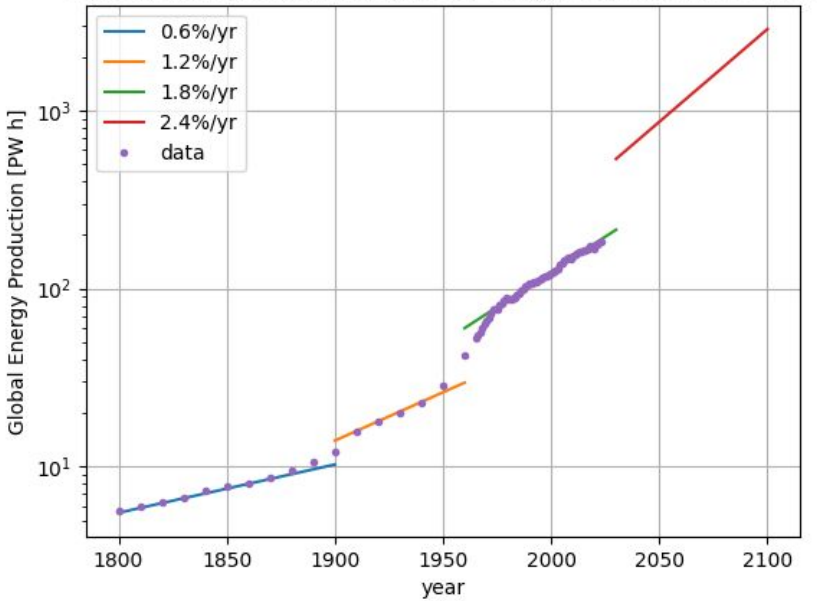
Get ready…

Big changes are a comin’



Data points from: ourworldindata.org/grapher/global-energy-consumption-source

FES Workflows

Common patterns at DIII-D and LCLS

Ryan N Coffee / Sr. Scientist / LCLS-TID-PULSE

June 27th 2025

SLAC NATIONAL ACCELERATOR LABORATORY

Stanford University

U.S. DEPARTMENT of ENERGY

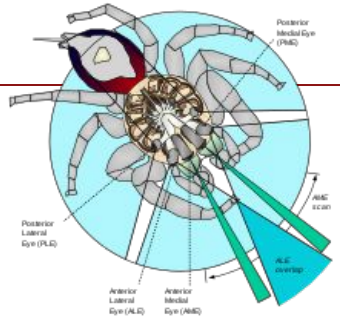
Codesign is unavoidable

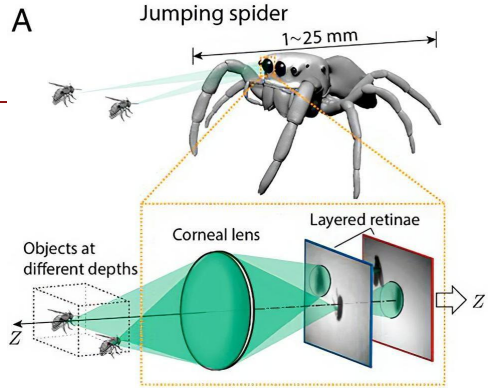
Hardware and wetware work in unison

ML in Science is predominantly acceleration of known interpretation

Let's design for Jumping Spider Specificity and efficiency…

… We need an AI Pit Crew



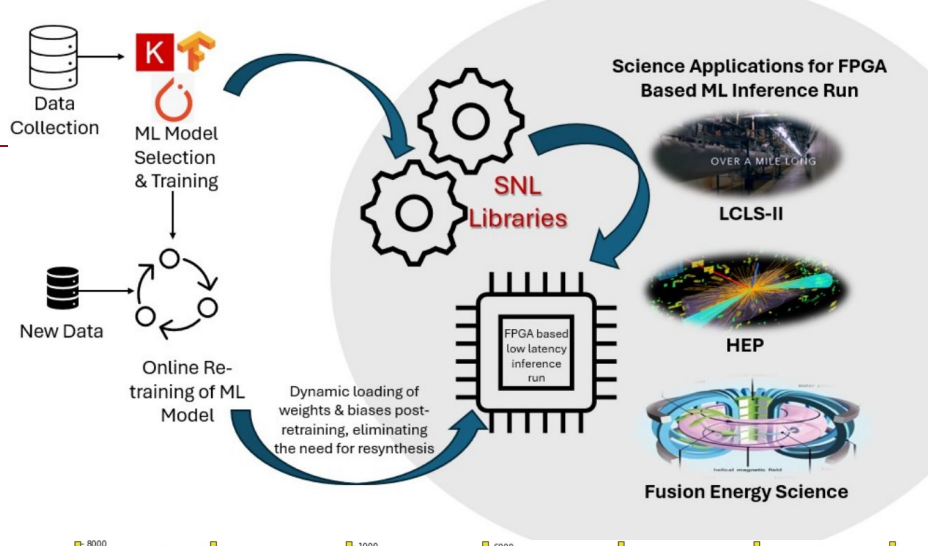


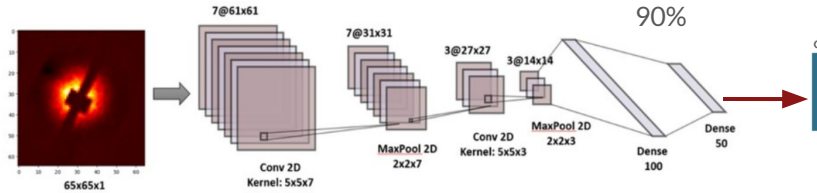
EdgeAI is Heterogeneous

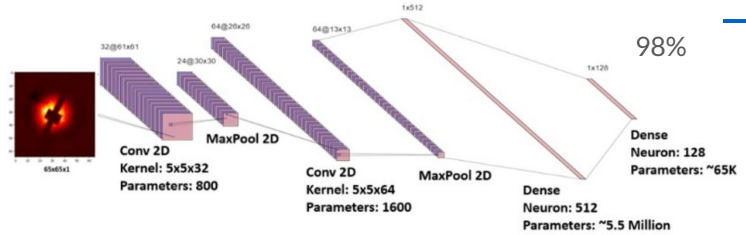
Data Rates are superhuman

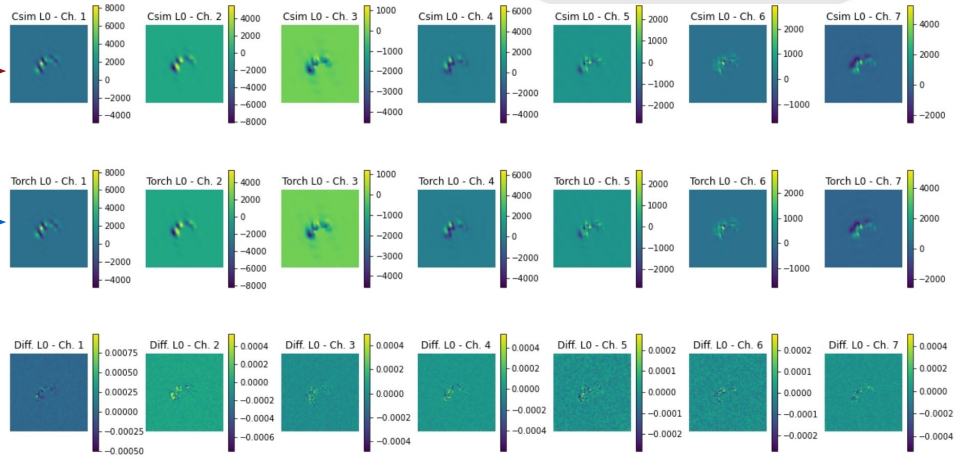
Beamline optimization happens at human timescales

The firehose is filling drives with $#!+









A Dave, *et al.*, "FPGA-Accelerated SpeckleNN with SNL…," arXiv 2502.19734

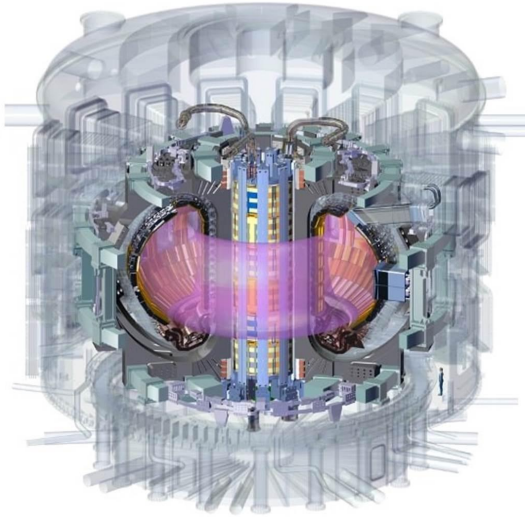
EdgeAI is Unavoidable

At the light source

- 1MPix at 1MSps rates
- 1Byte/sample = 1TBps

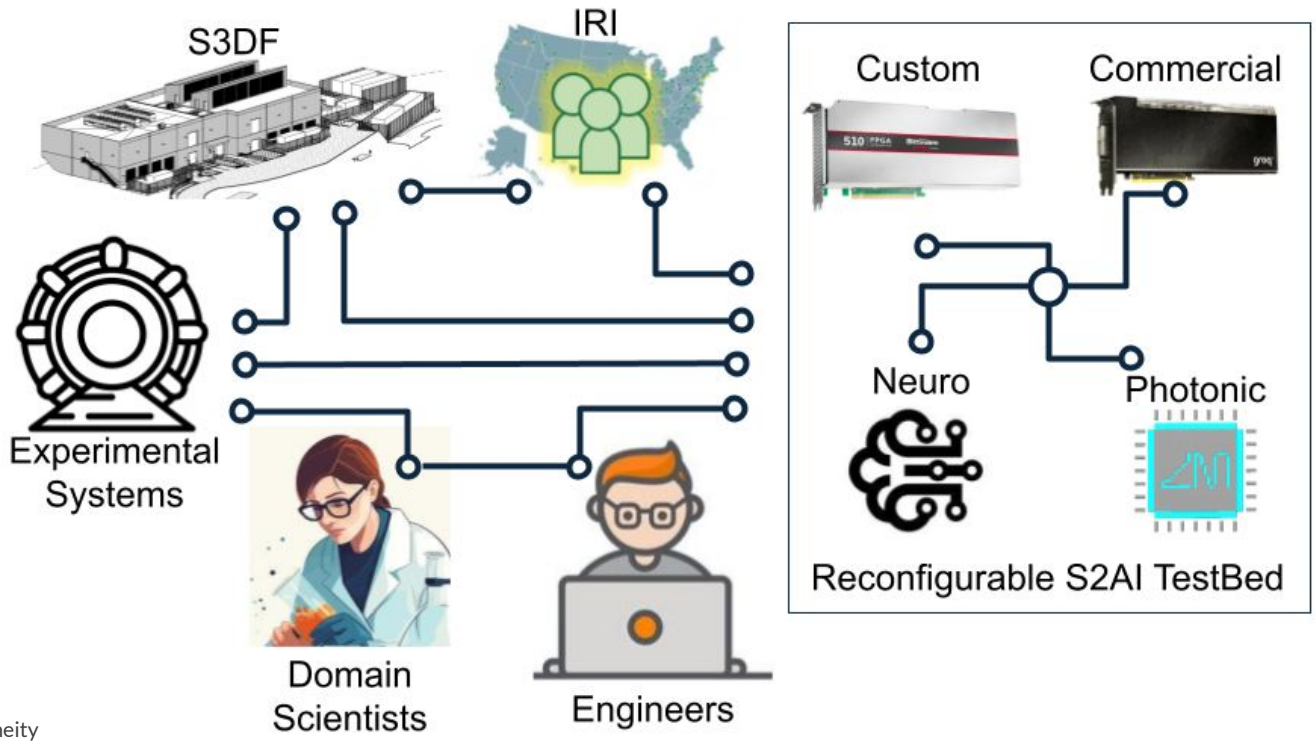
At the tokamak

- 1k channels at GSps rates
- 1Byte/sample = 1TBps



Heterogeneity of models = Heterogeneity of Hardware

Know the "Why", need the "How" **… S3AI to support hetero-stream exploration**



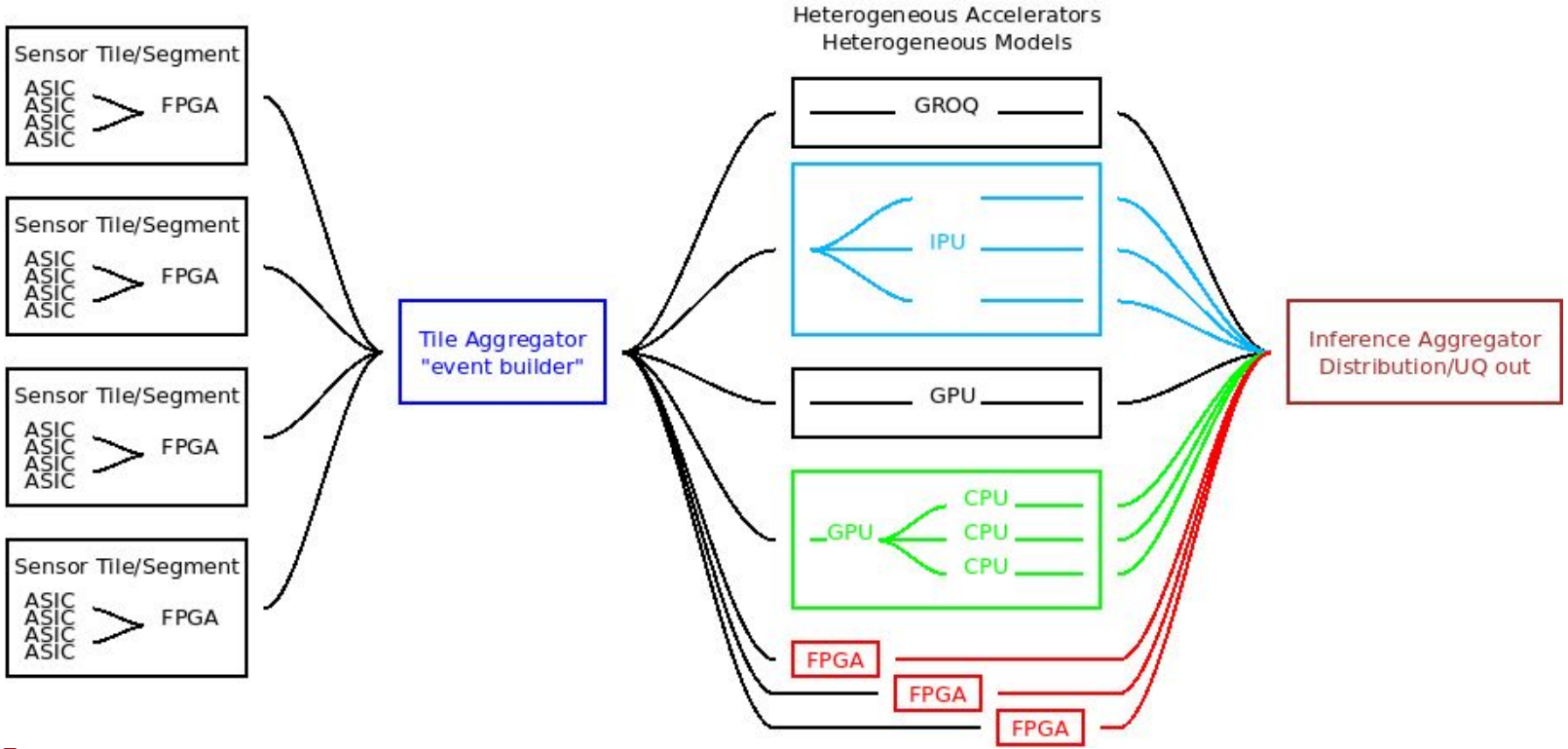
Heterogeneity of models = Heterogeneity of Hardware

Know the "Why", need the "How" **… S3AI to support hetero-stream exploration**

- Tight coupling between sensors → bleeding edge → near edge → local HPC → and remote LCF
- S3AI → S3DF → OLCF
- Multi-scale heterogeneous computational ecosystem

Heterogeneity of models = Heterogeneity of Hardware

Know the "Why", need the "How" **… S3AI to support hetero-stream exploration**

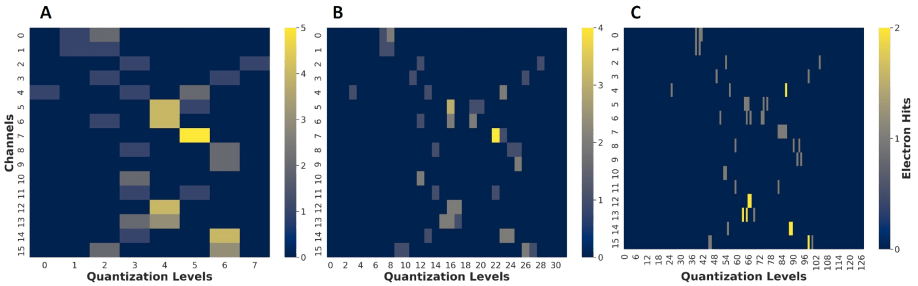


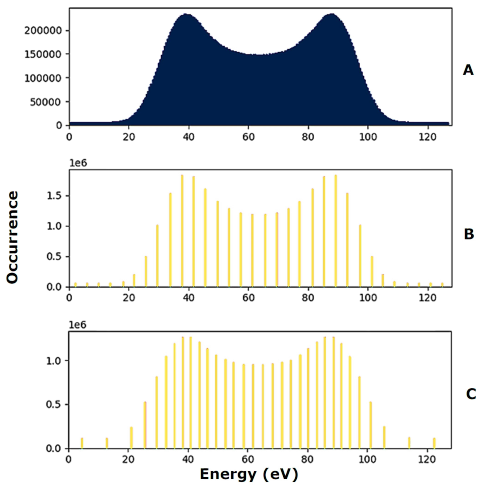
Quantization is Unavoidable

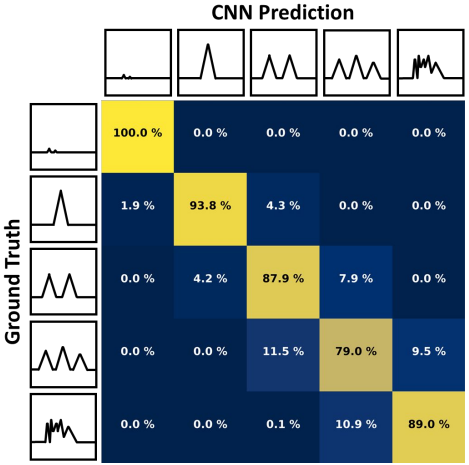
Optimizes information per bit

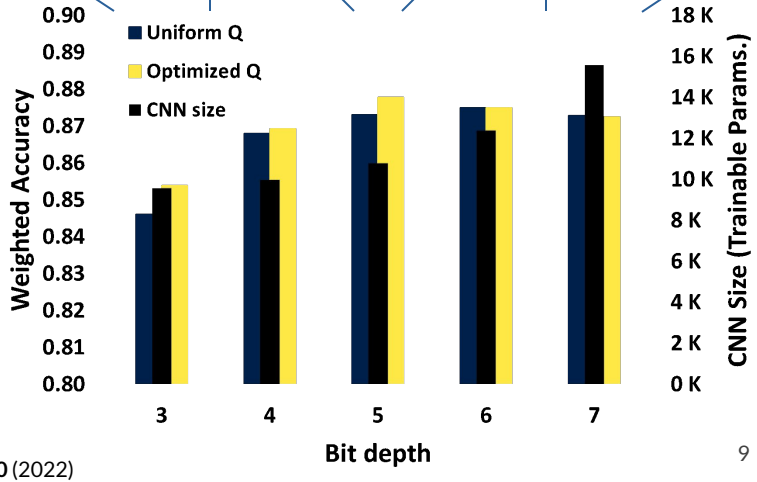
Reduces input token dimensionality

Puts metadata to work









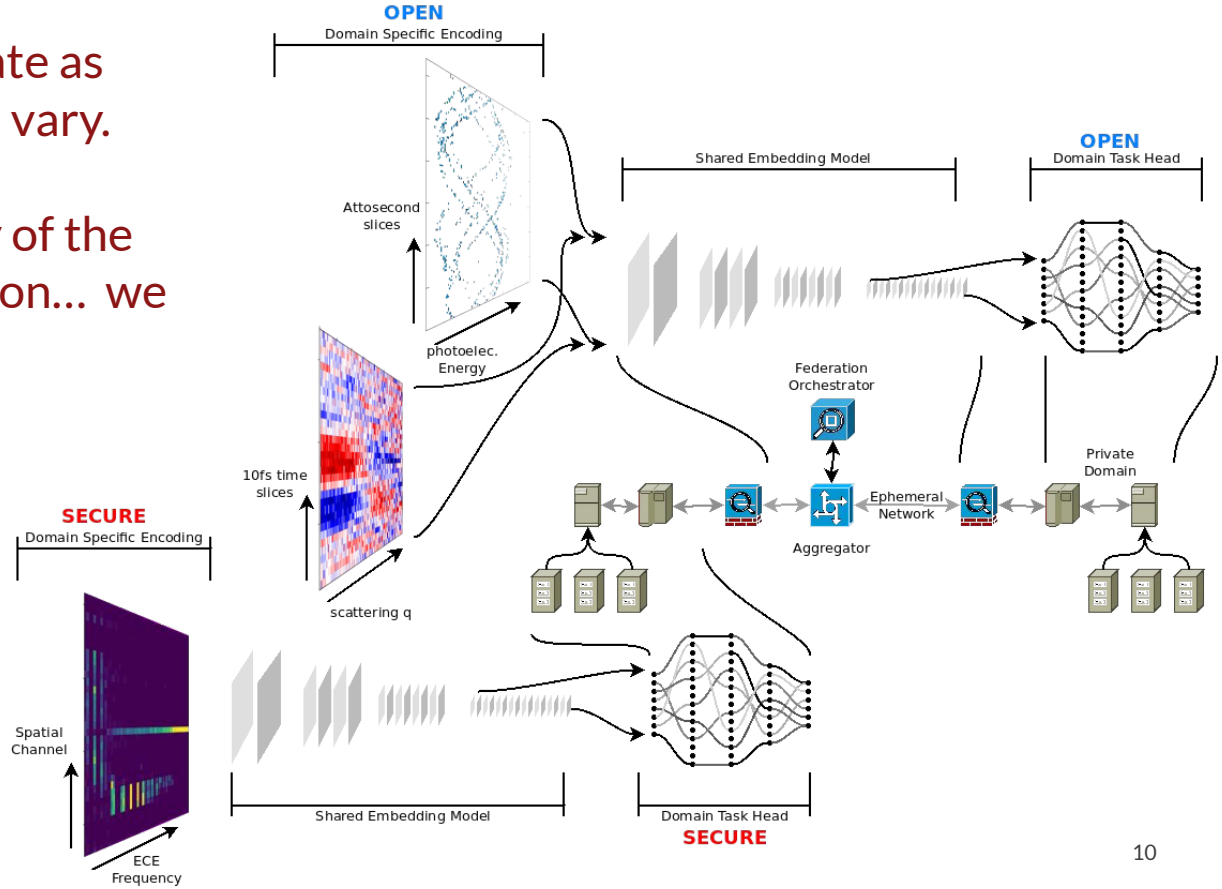
Gouin-Ferland, Coffee and Therrien, Front. Phys. **10** (2022)

Is embedding the equalizer?

Basis functions will update as the samples and sources vary.

If we maintain geometry of the embedding representation… we still share pre-trained foundation models?

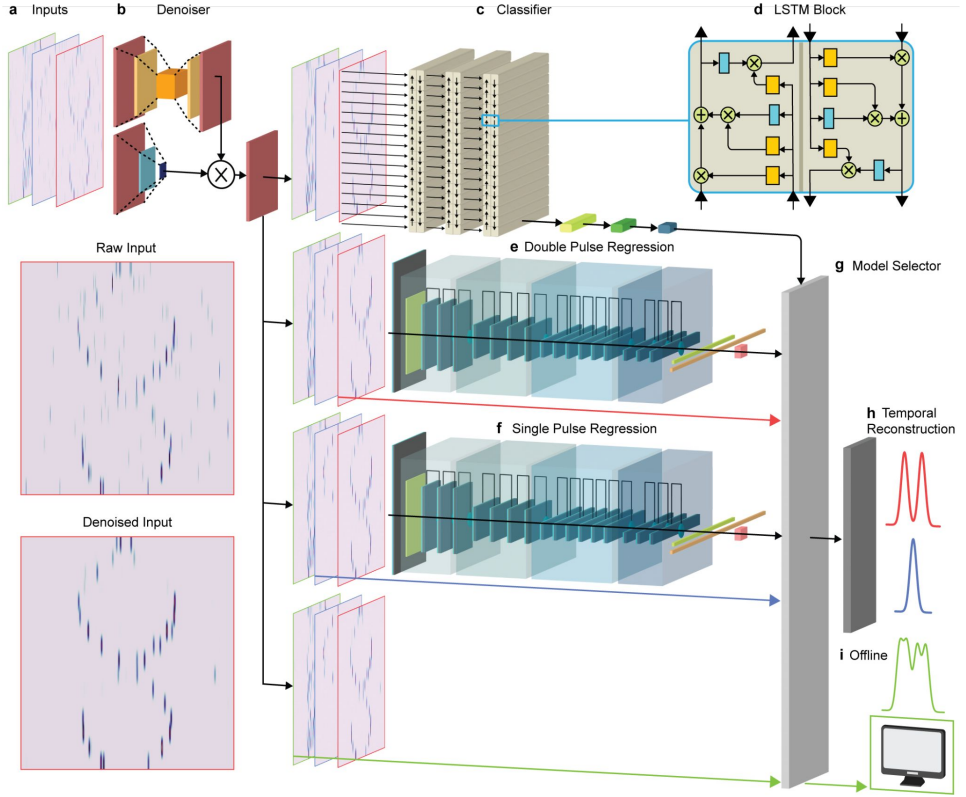
Open foundation…
… closed embedding for sensitive data



Heterogeneity of models

Streaming may flow broad and shallow rather than narrow and deep.

What does this look like for Fusion?



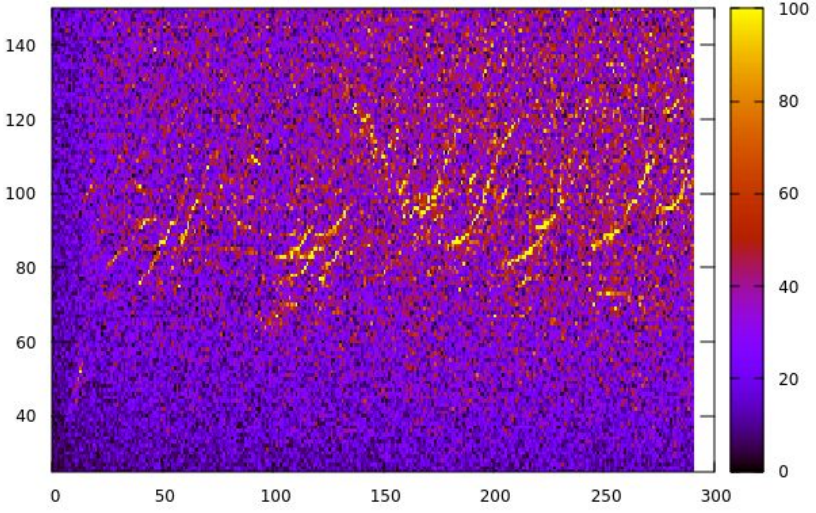
Hirschman *et al.*, "A Hybrid Neural Network for High-Throughput Attosecond Resolution Single-shot X-ray Pulse Characterization", (2025) arXiv 2502.16141

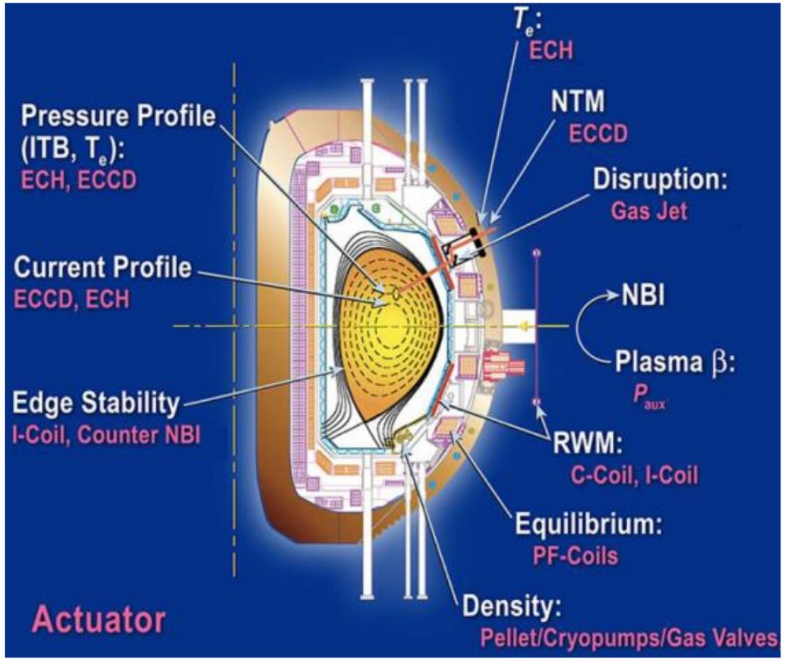
Sensors at DIII-D

Streaming may flow broad and shallow rather than narrow and deep.

What does this look like for Fusion?

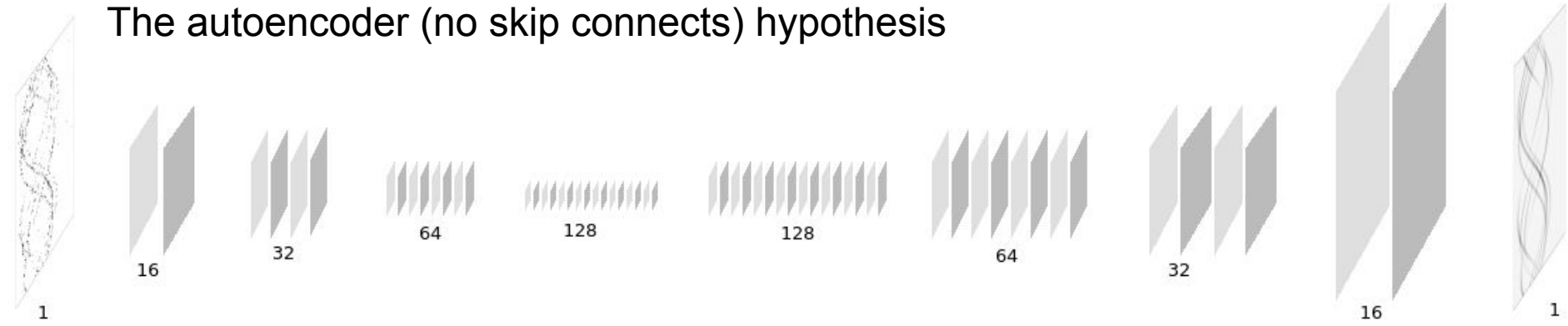
One of 32 channels of ECE spectrograms

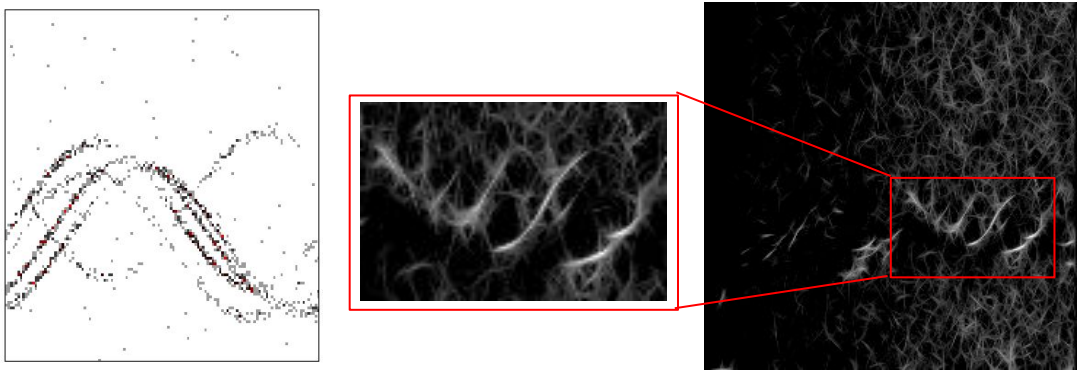




The structure of information

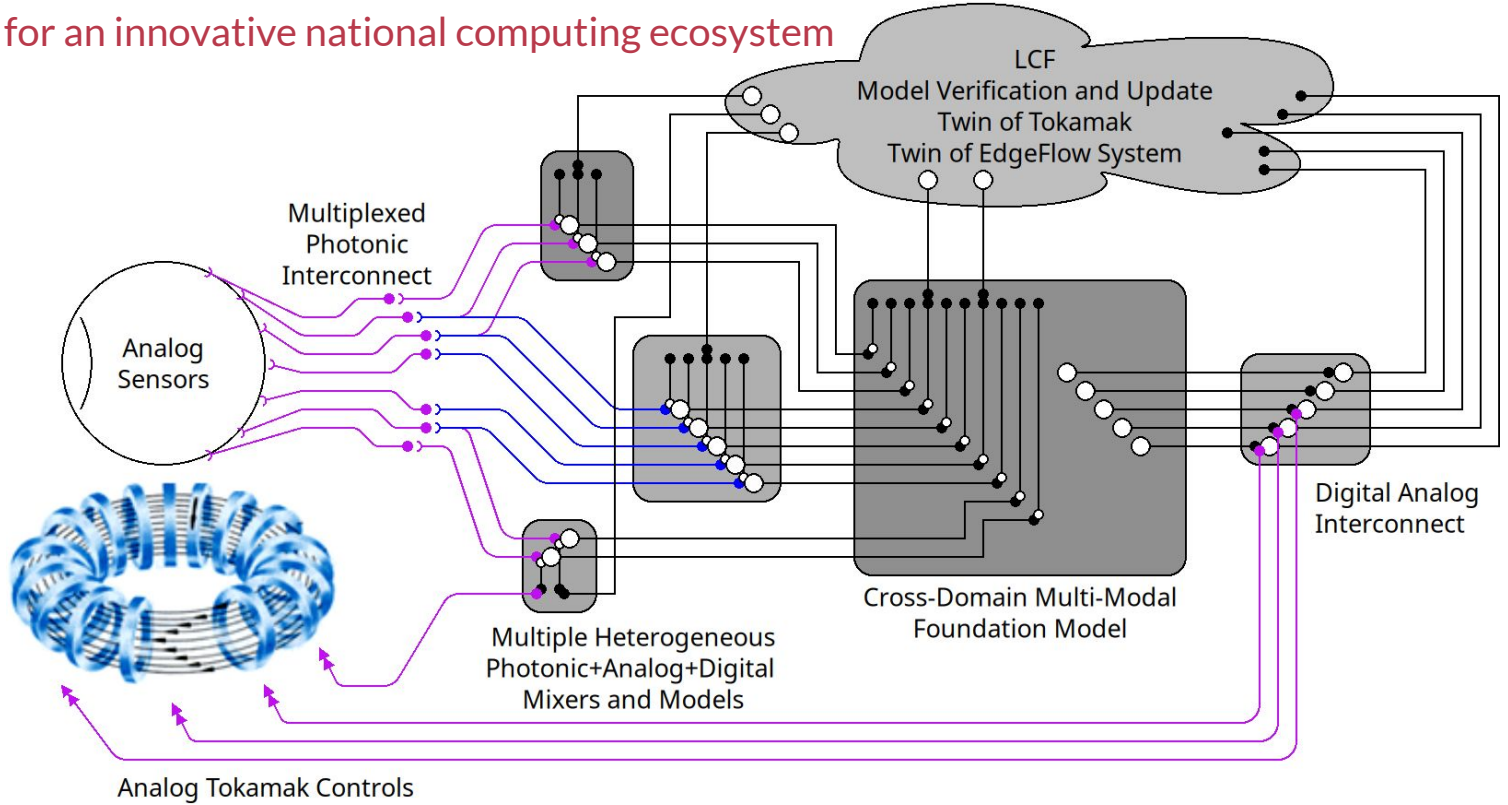
The autoencoder (no skip connects) hypothesis



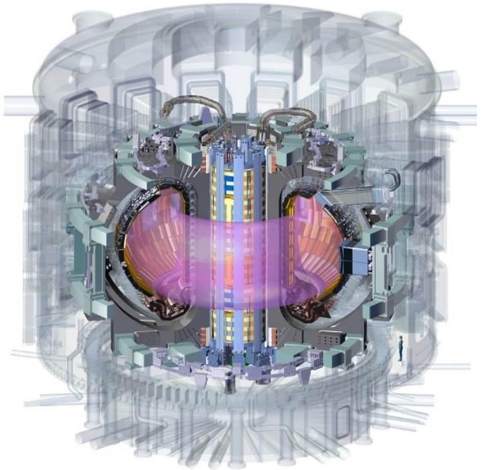


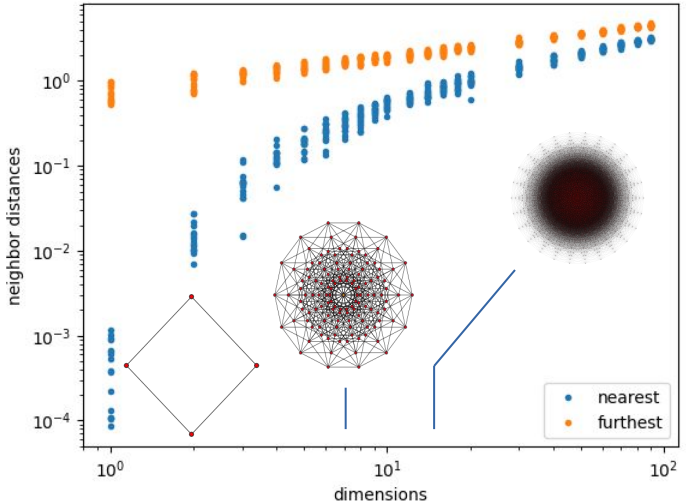
Multi-hardware, Multi-site, Multi-loop

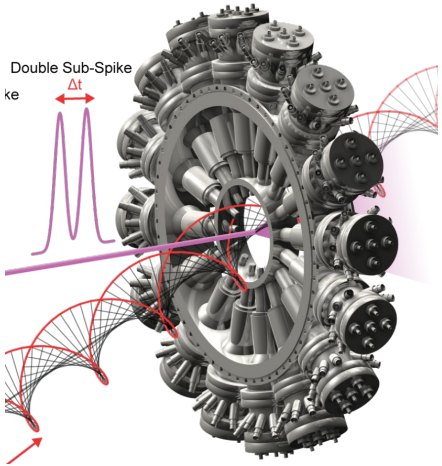
Design for an innovative national computing ecosystem



Hallucination and Dimensionality… the big “gotcha”







If it takes 1-10T tokens of high quality to train modern LLMs, what would 1T tokens of mean for a **tokamak like the DIII-D?**

1T = 512freq x 32ch per 1 ms of data.

200k shots at 5000ms each = 1B ... **0.1% of the minimum**

Most corners in parameter hypercube are unoccupied.

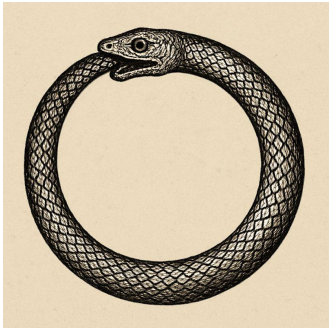
At LCLS-II, at 1MHz, each shot gives a “tokenizable” spectrum in the CookieBox of 20ch x 512bins.

Beamtimes are unique, and roughly 48 hours of useful beamtime (typically really 24 is good)

48hr * 3.6e9 shots/hr = 173B ... **< 20% of the minimum**

Even at the data firehose, **we need dimensional reduction**

What does hallucination look like?



Internet Death

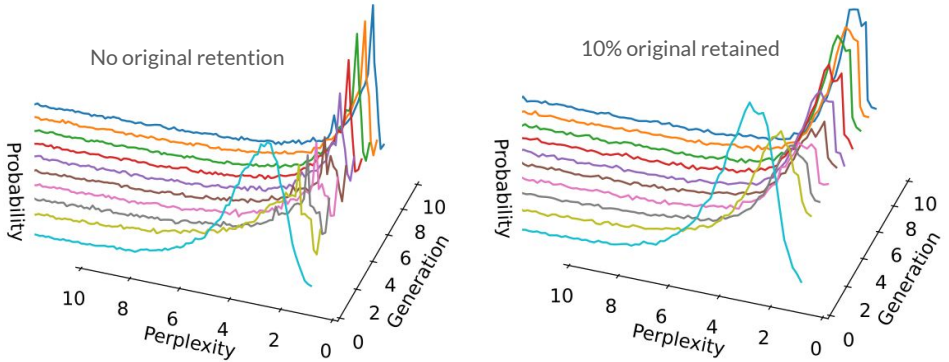
Synthetic data is flooding faster than human-authored content. Humans create slow. AI creates fast, millions of new pages daily....

In the long term, it could **destroy the very foundation the AI economy depends on: real knowledge.**

Models that once reflected the world will soon reflect only themselves. **And when knowledge eats itself, originality and truth disappear.**

Stephen Klein, Curiouser.AI

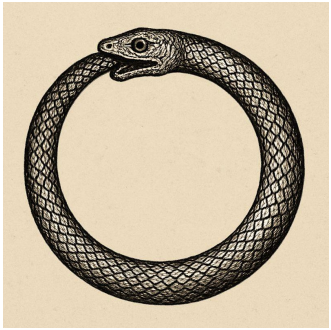
We need distribution metrics that can trigger compensatory outlier acquisitions



"Over the generations, models tend to produce samples that the original model trained with real data is more likely to produce. At the same time, a much longer tail appears for later generations. Later generations start producing samples that would never be produced by the original model, that is, **they start misperceiving reality based on errors introduced by their ancestors.**"

Ilia Shumailov, *et al.*, "AI models collapse when trained on recursively generated data," Nature **631**, 755–759 (2024)

Mass hallucination feedback



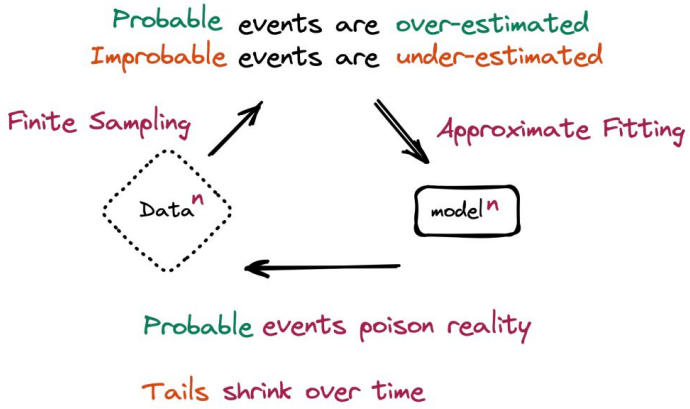
Internet Death

Synthetic data is flooding faster than human-authored content. Humans create slow. AI creates fast, millions of new pages daily....

In the long term, it could **destroy the very foundation the AI economy depends on: real knowledge.**

Models that once reflected the world will soon reflect only themselves. **And when knowledge eats itself, originality and truth disappear.**

Stephen Klein, Curiouser.AI



"We note that **access to the original data distribution is crucial**: in learning where the tails of the underlying distribution matter, one needs access to real human-produced data. In other words, the use of LLMs at scale to publish content on the Internet will pollute the collection of data to train them: **data about human interactions with LLMs will be increasingly valuable**."

Ilia Shumailov, *et al.*, "The Curse of Recursion: Training on Generated Data Makes Models Forget," arXiv 2305.17493

Audience Participation