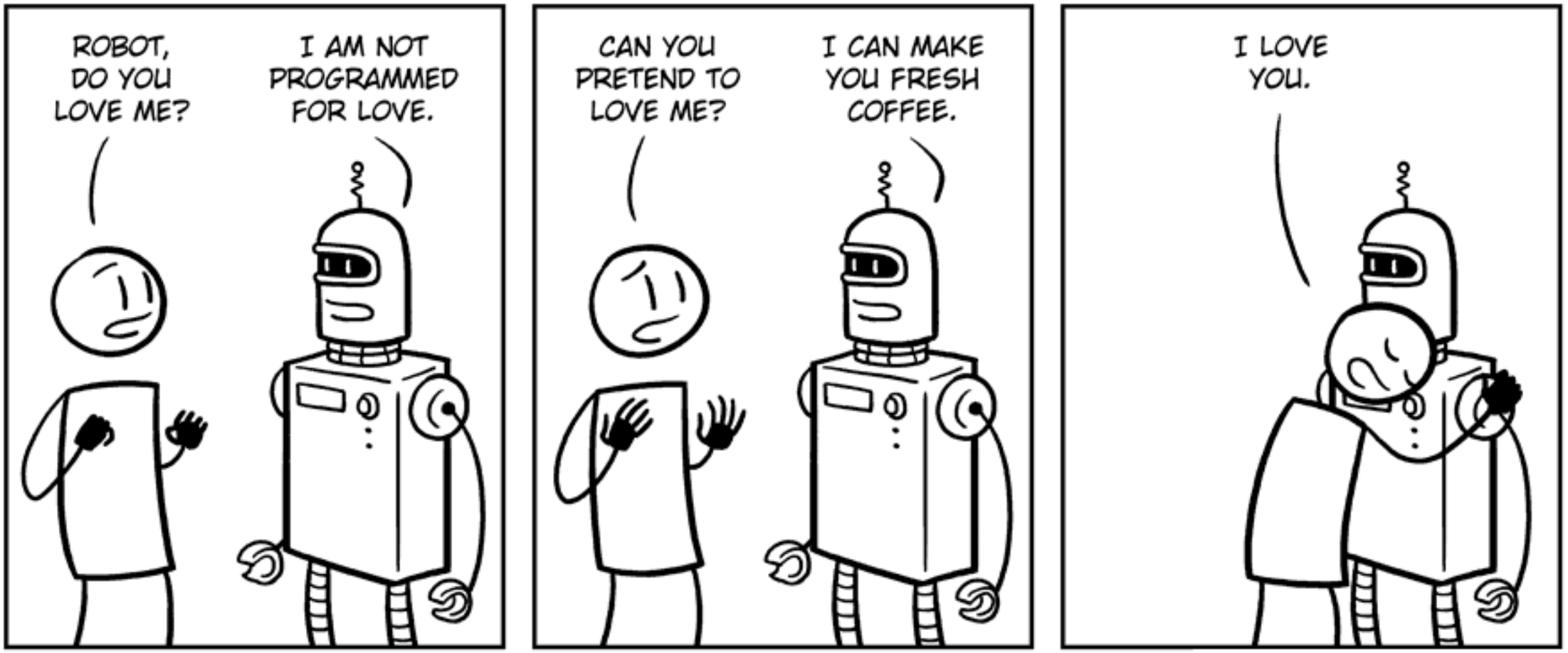


Community Tools, Standards, Resources and Management

P. Harris on behalf of CompF3

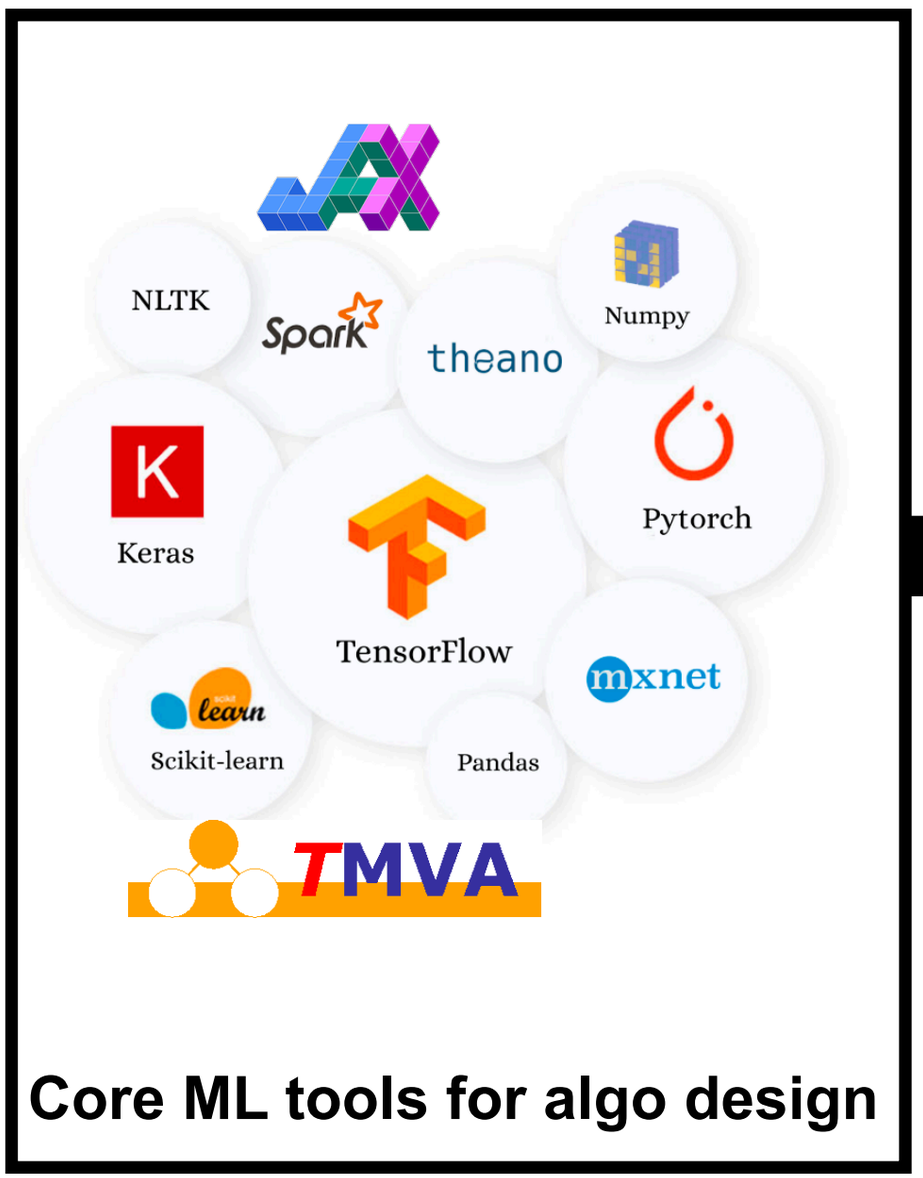
Accelerated AI Algorithms for Data-Driven Discovery

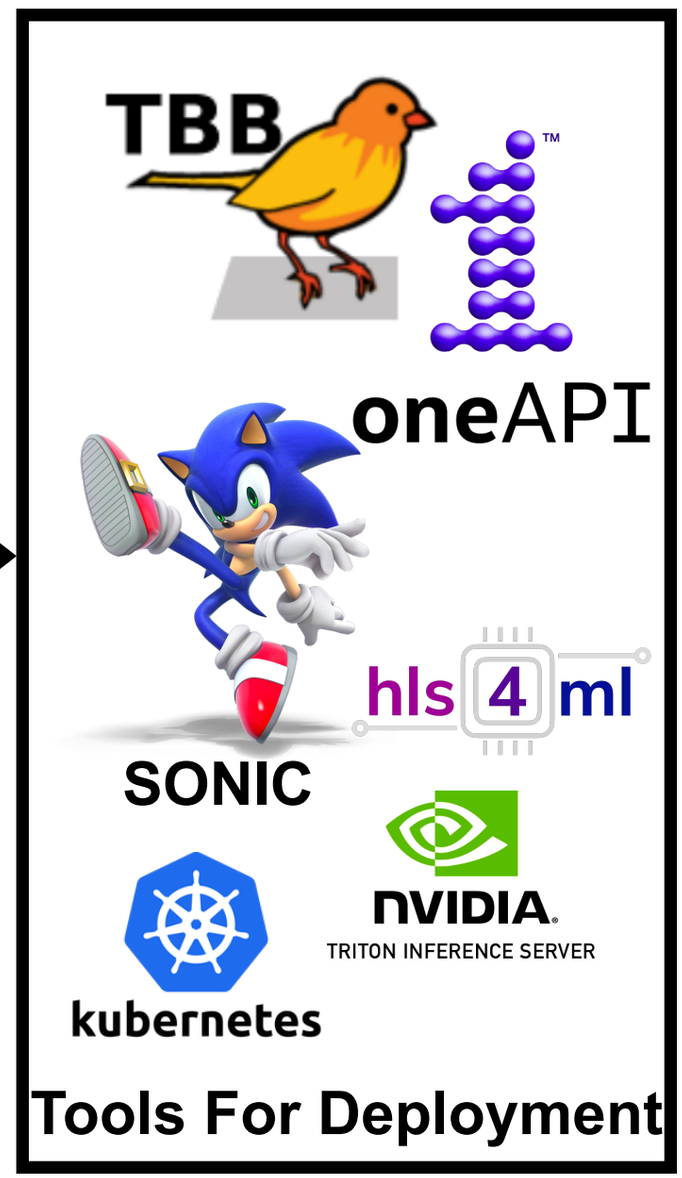
Overview

- This talk is a summary of several papers:
  - D. Rankin *et al* ***Experimentalists: arxiv:2203.16255***
  - ***A. Adelman et al Detector sim: arxiv:2203.08806***
  - D. Hackett ***et al Lattice QCD: arxiv:2202.05838***
  - ***C. Dvorkin et al*** *Cosmology : arxiv 2203.08056*
- Tools : How do we enable effective use of the resources?
- Resources: How do we use computing? future needs?
- Standards : What standards can be established going forward?
- Management : How do we deal with the tools and resources?
- ConclusionWhat can we draw from this

Tools

- Tools for machine learning revolve around
  - Standard (industry) ML toolkit & Custom implementations





Progressively more Customization is Needed

Tools

- Tools for machine learning revolve around
  - Standard (industry) ML toolkit & Custom implementations

Lot of excellent tools in industry for fast R&D and exploration
However a few critical tools need our input

Core ML: Uncertainty Quantification, Quantization/Pruning
Data Format: Quantized ONNX formats/extended pytorch
Tools to deploy: Tools for FPGA in trigger/
Asynchronous scheduling/CPU/c++ based inference

Work is really integration into our existing open source SW

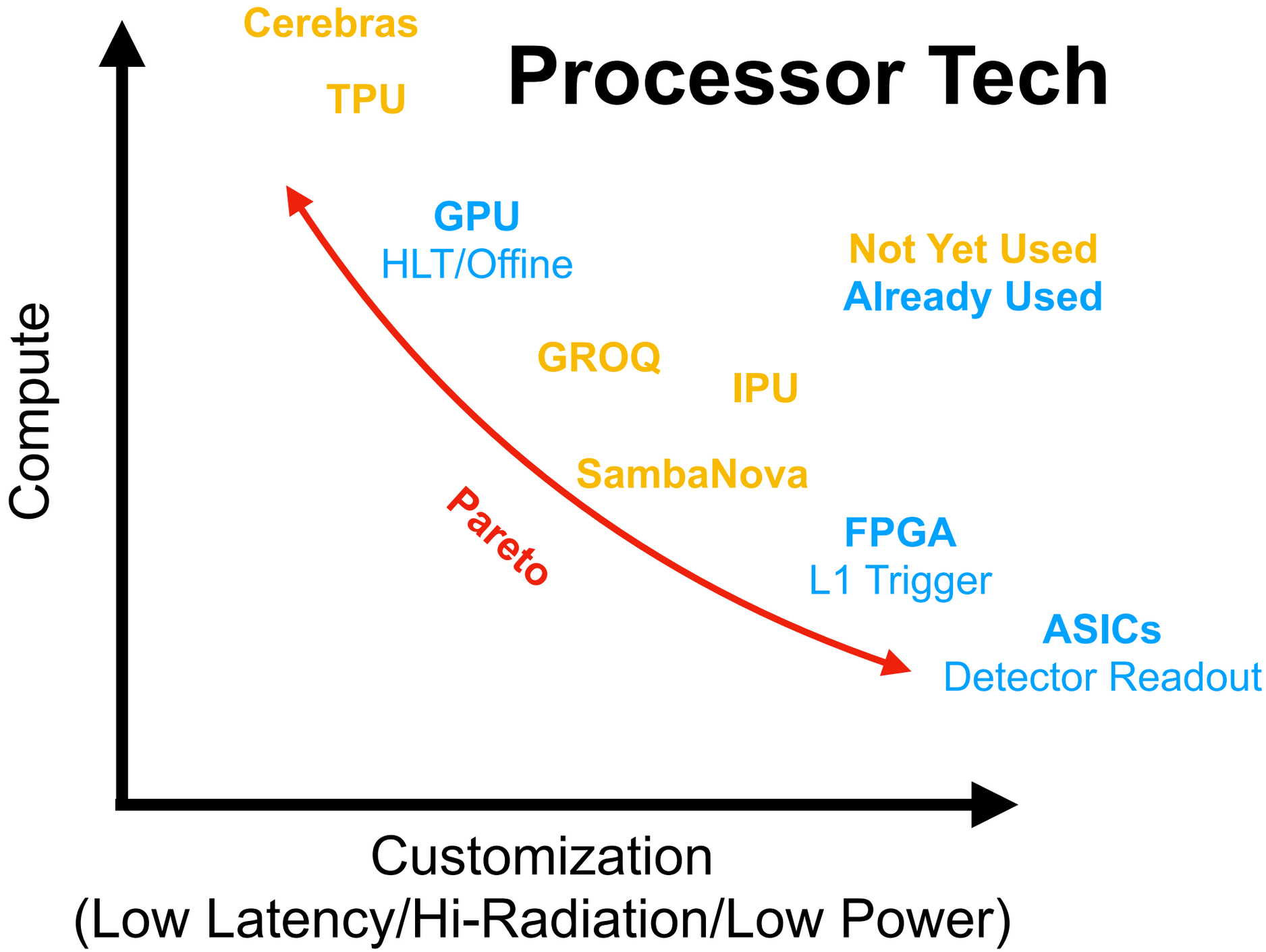
Core ML tools for algo design

Data Formats

Tools For Deployment

Progressively more Customization is Needed

Resources

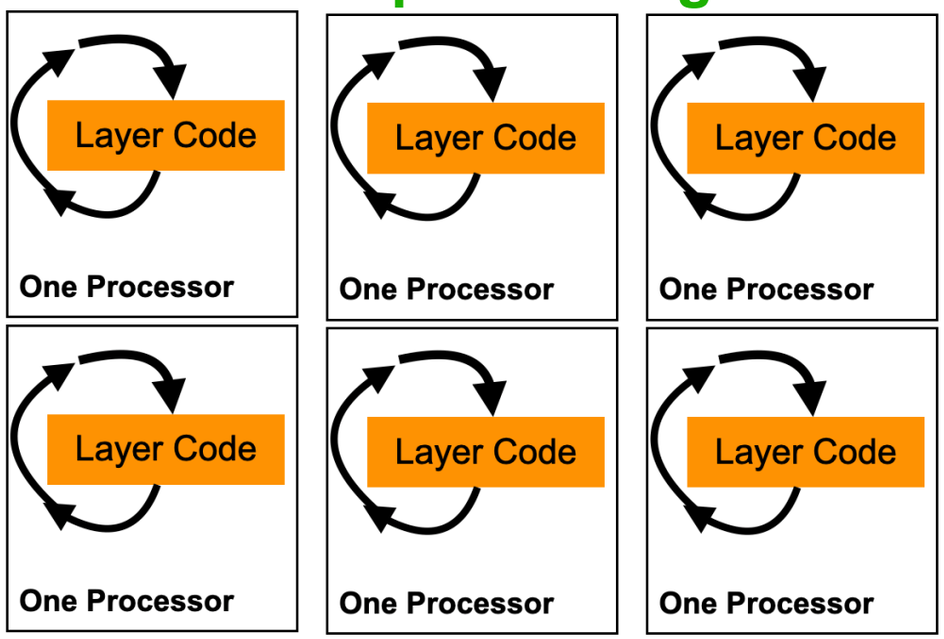


Different Process for different use cases

Resources

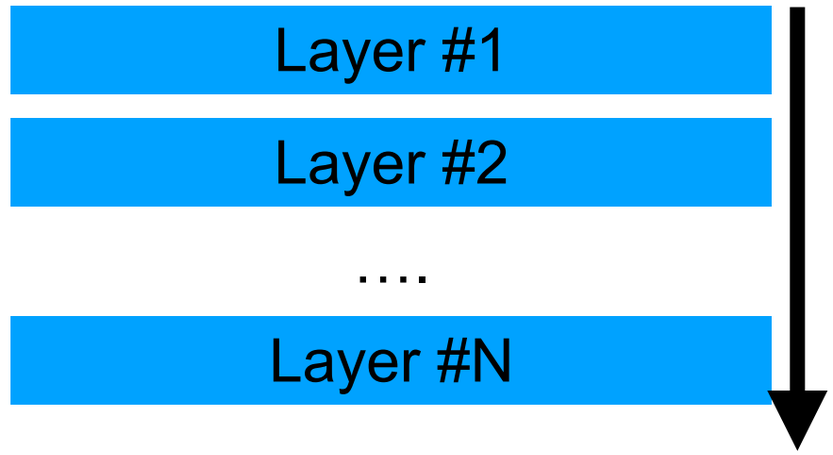
Processor Tech

Parallelize Your processing



GPU/IPU/TPU/….

Customize Your processing



FPGA/ASIC/SambaNova/…

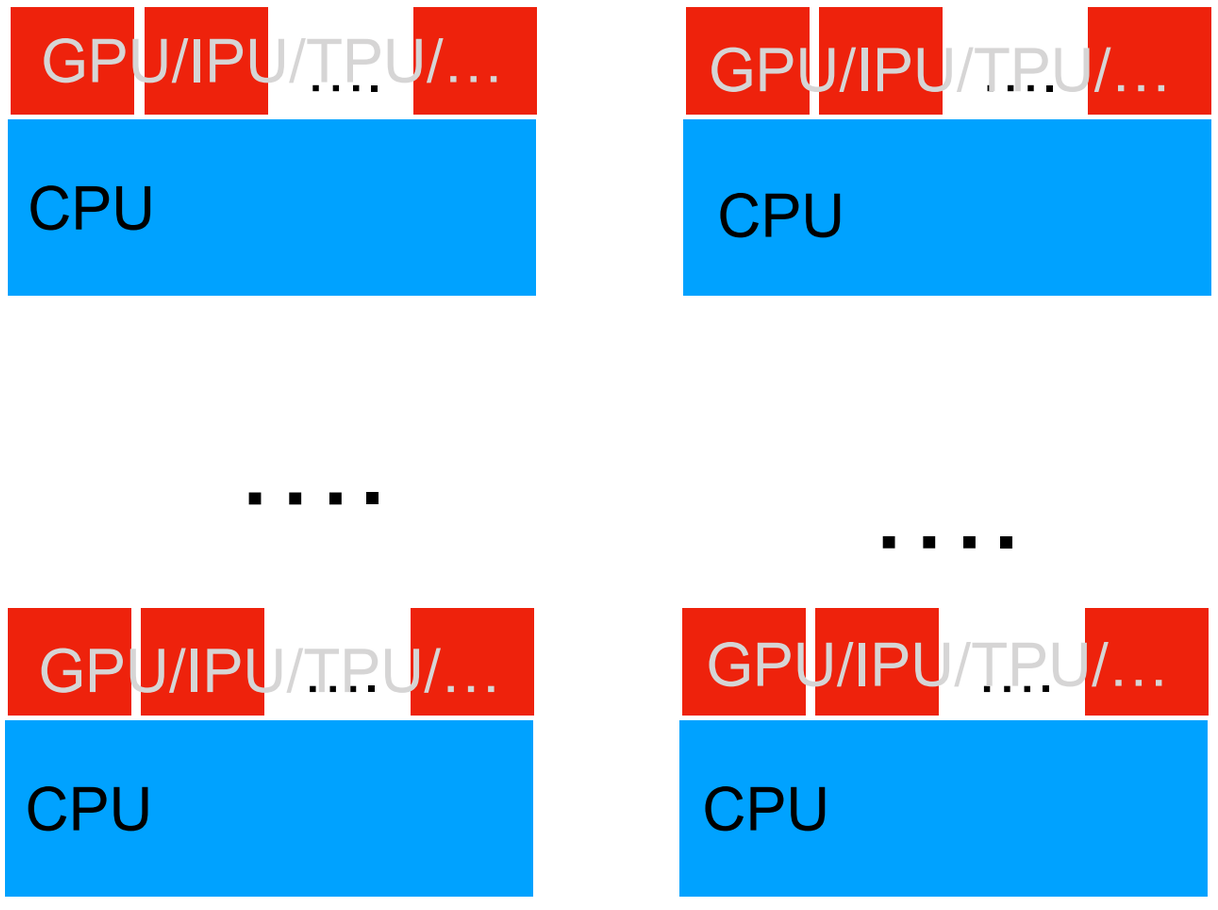
Physicists have been driving some of the ML customization efforts

Different Process for different use cases

Resources

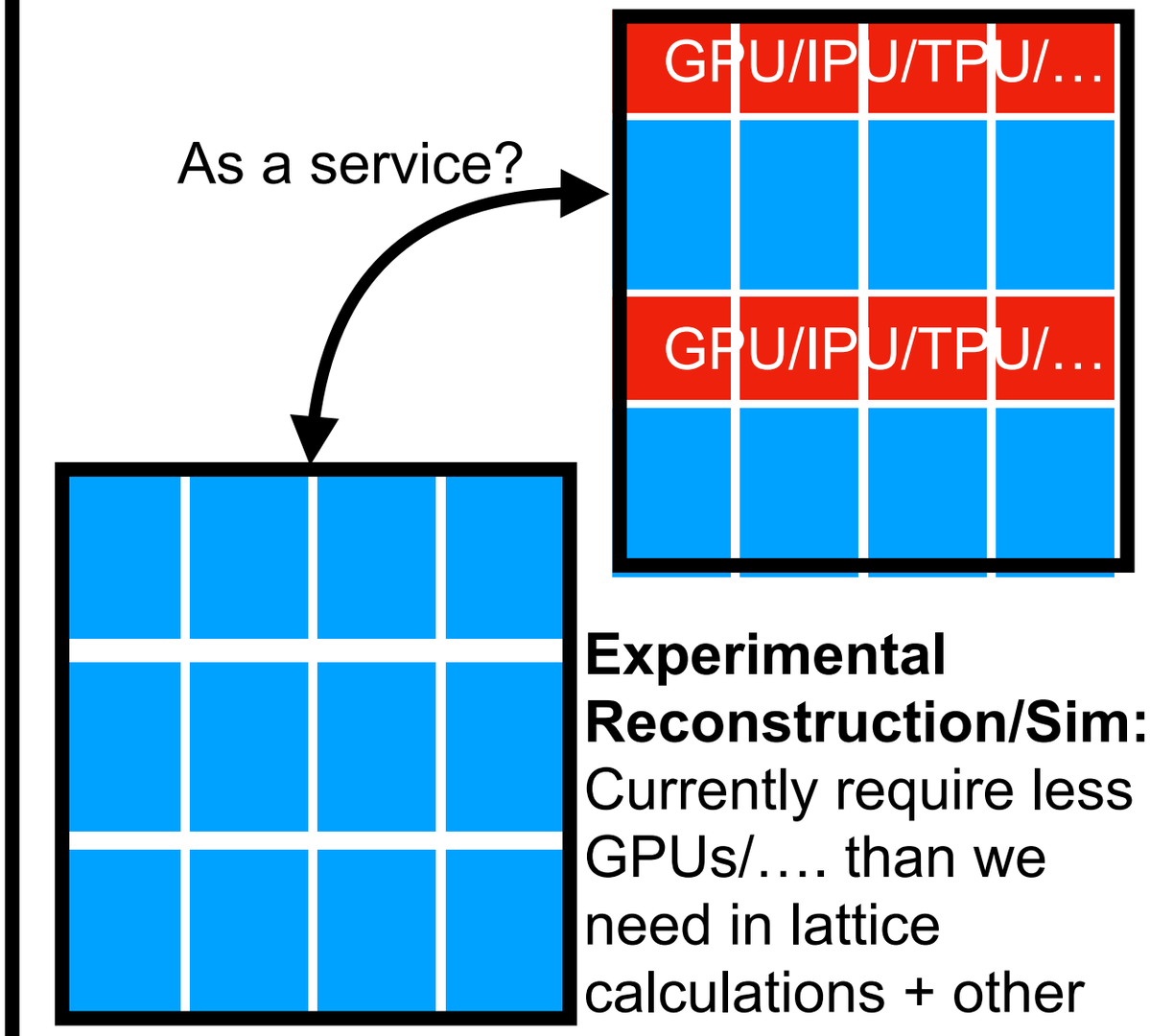
- All of us want to use heterogeneous resources for computing
  - Demands for heterogeneous resources vary by domain

One big regular system (current HPC model)



Lattice QCD/Astro/Simulation
Big Trainings
Big Calculations

Migration towards for GPU/…. integration
Upgraded (existing) systems

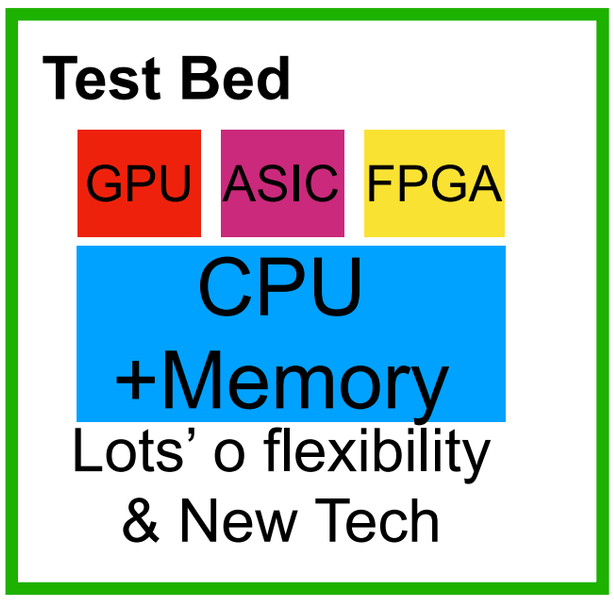


Experimental Reconstruction/Sim:
Currently require less GPUs/…. than we need in lattice calculations + other

Resources

- One thing we all agree on
  - We don't **just** need large compute resources
- Having a test bed for small resources/new tech is critical
  - Many of our problems we eventually want to scale out

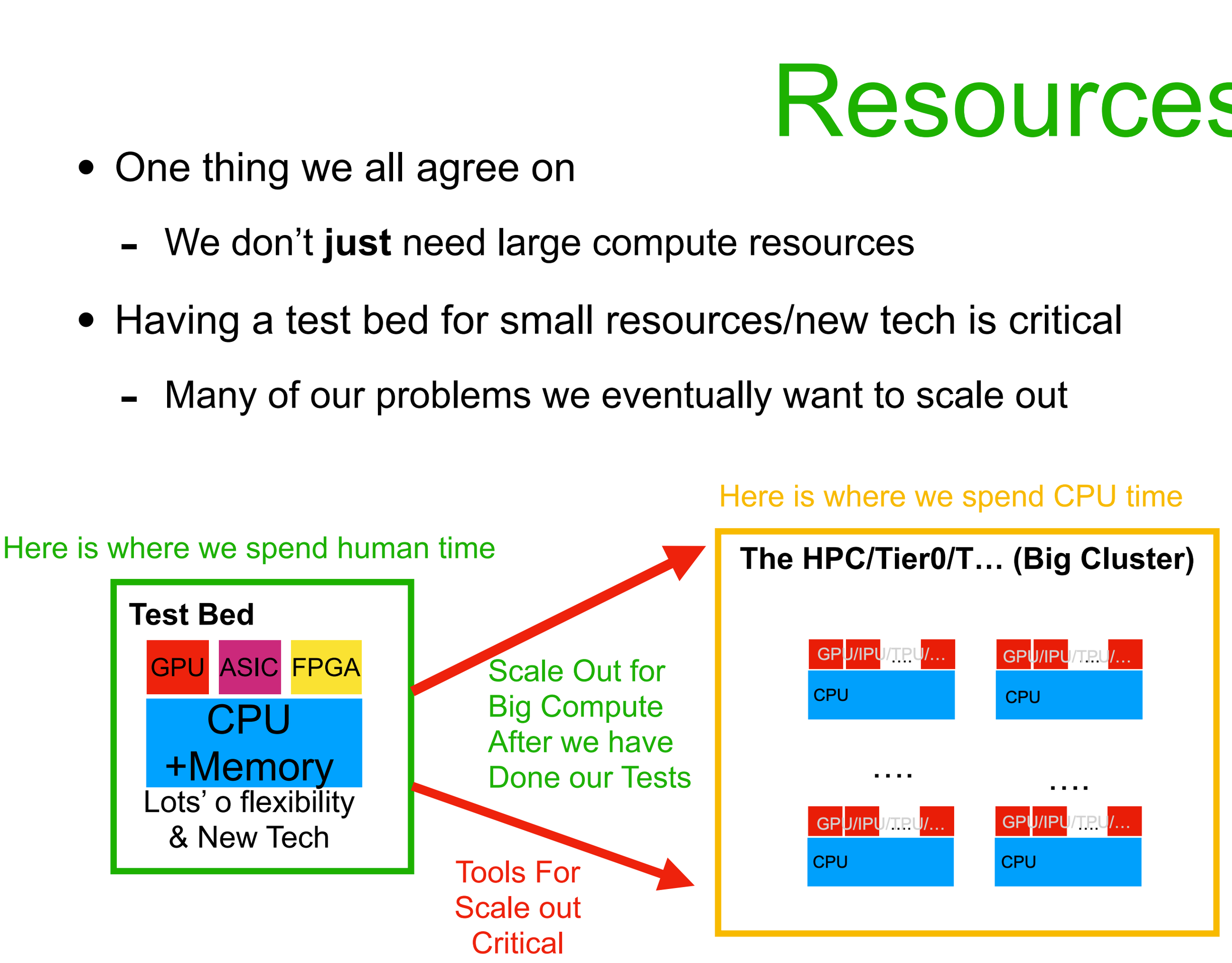
Here is where we spend human time



Have some unique computing challenges
More data than (basically) anybody
Demands for ultra low latency
Playground is essential

Resources

- One thing we all agree on
  - We don't **just** need large compute resources
- Having a test bed for small resources/new tech is critical
  - Many of our problems we eventually want to scale out

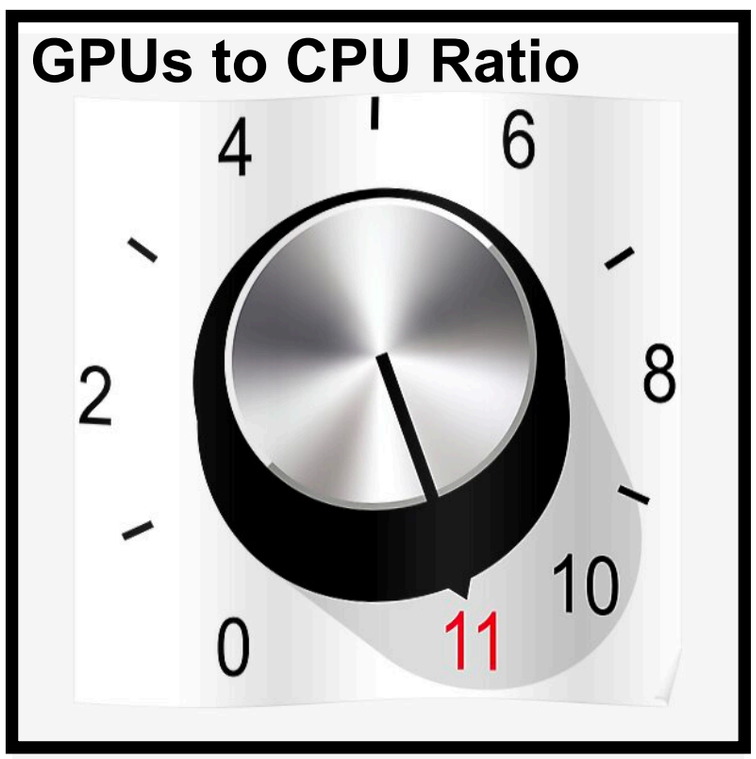


Resources

- HPCs are quickly coming online
  - Being able to harness them effectively requires flexible SW tools

Goldilocks principle:
Current system often designed for "Just the right" amount of GPUs to CPUs to run optimally

Being able to adjust the amount GPUs/… to CPUs is critical to optimally use resources



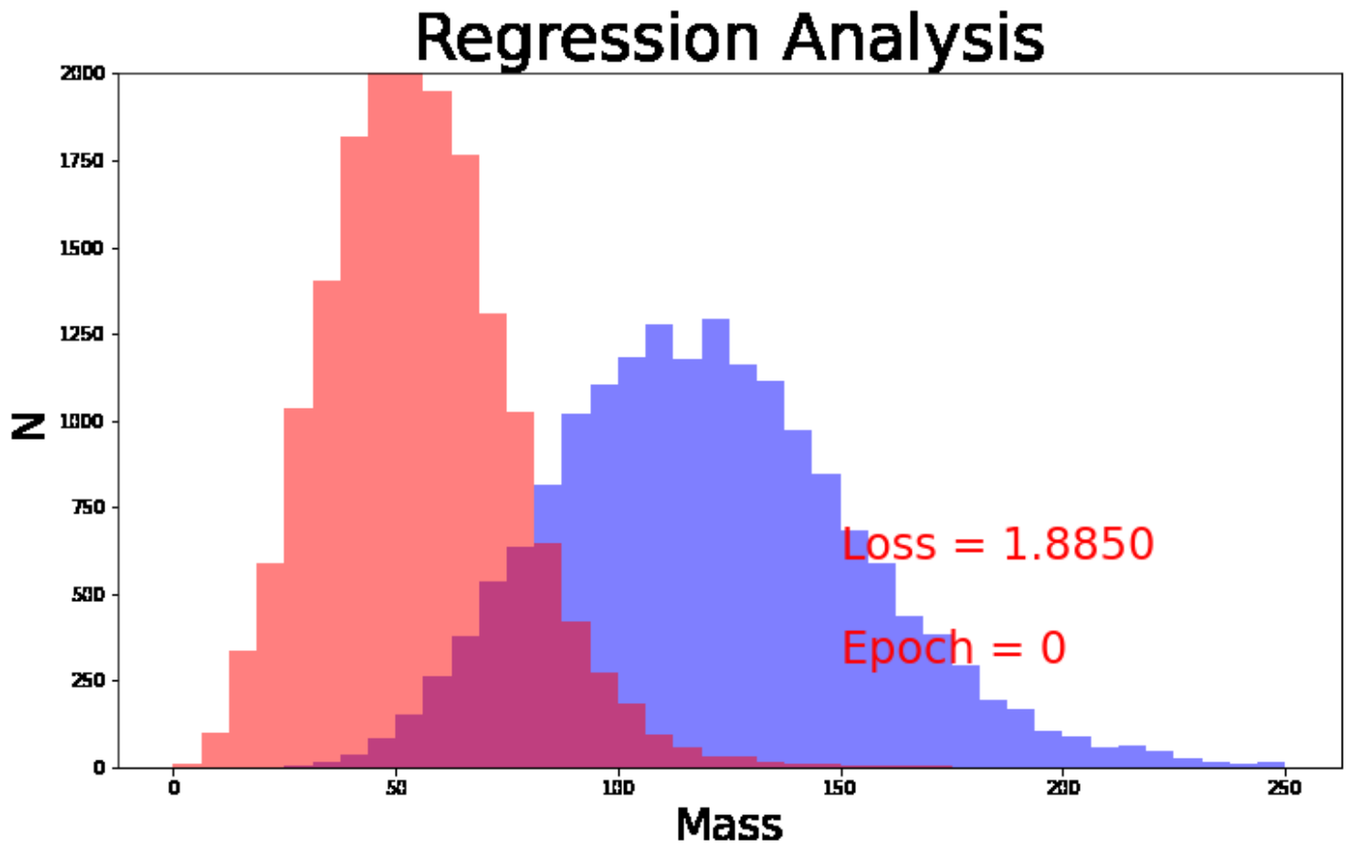
Standards

- Much of the effort across the domains is synergistic
  - Regular data formats (h5/awkward/root) + FAIR principles
    - FAIR: Findable Accessible Interoperable Reusable
  - Establishing/utilize core software standards is essential
    - Focusing on a few **standard** tools (pytorch/tf + onnx/h5 etc)
  - Use of HPCs will be required for everybody
    - However we all plan on using it differently

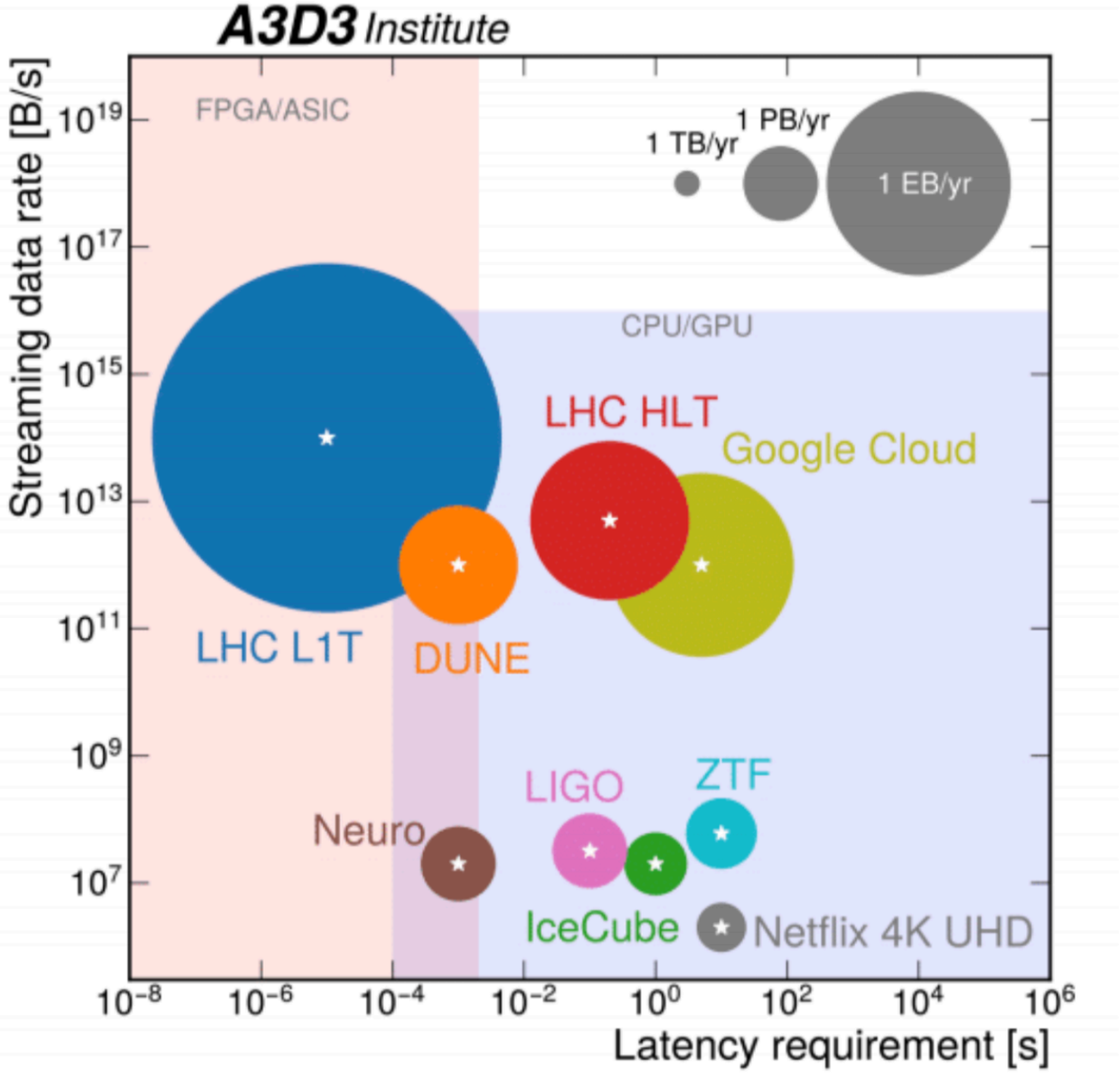
- Within AI community there are larger efforts to do AI exploration
  - Keeping up/contributing with community is essential for future

Management

- Educating the next generation is critical
  - Consolidated AI/Physics PhD courses are essential
- Preserving the expertise
  - Investment in postdoc/research/faculty path crucial
  - Competition with industry is fierce



Conclusions



- Variety of computational problems
  - Everybody needs a test bench
    - New processors/AI technology
  - Require different resources
    - Big compute vs as-a-service
- Core of problems all have same base
  - Standard ML tools
    - We should push for standards

We can push AI in different directions

Backup

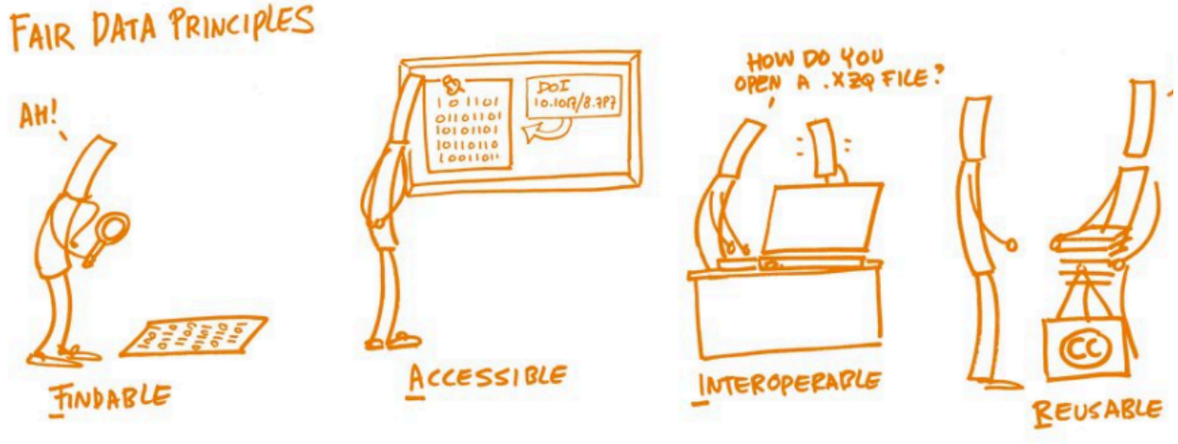
Resources

- CPUs: Our core resource that often still the default
  - Need this when we do complex physics data processing
- GPUs: the standard for AI
- FPGAs: Special tasks
- ASICs
- Next Generation : TPU/IPU/SambaNova/Cerebras/Groq
  -

The FAIR Principles

- To inspire scientific data management for reproducibility and maximal reusability[1]
- Originally proposed for scientific data
- Can be interpreted as guidelines to manage and preserve other Digital Objects (DOs) e.g. research software[2] , tutorials and notebooks[3] , AI and ML models[4]
- Different working groups working on FAIR guidelines for different DOs (e.g FAIR4RS, FAIR workflows, FAIR VREs)

Findable:	locating DOs in a failsafe fashion
Accessible:	obtaining DOs along with their context, content, and format
Interoperable:	being usable across multiple computing platforms
Reusable:	specifying the context and extent of reusing DOs



FAIR4HEP

ICHEP 2022

https://agenda.infn.it/event/28874/contributions/169888/attachments/94738/129818/FAIR4HEP-ICHEPJuly9.pdf