Machine Learning in Science: Applications, Algorithms, and Architectures

Kathy Yelick

Associate Dean for Research, Division of Computing, Data Science, and Society
Professor of Electrical Engineering and Computer Sciences
University of California, Berkeley

Senior Advisor on Computing, Lawrence Berkeley National Laboratory

Science + HPC

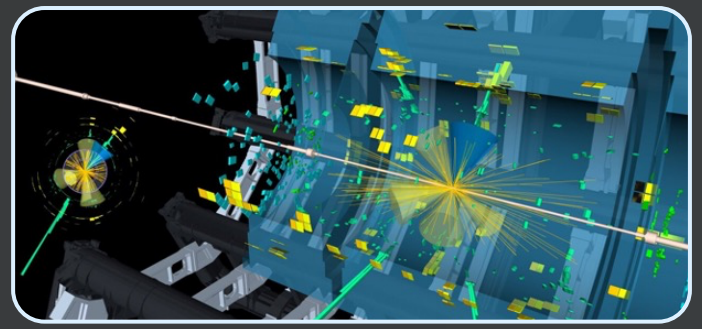
AI + Cloud

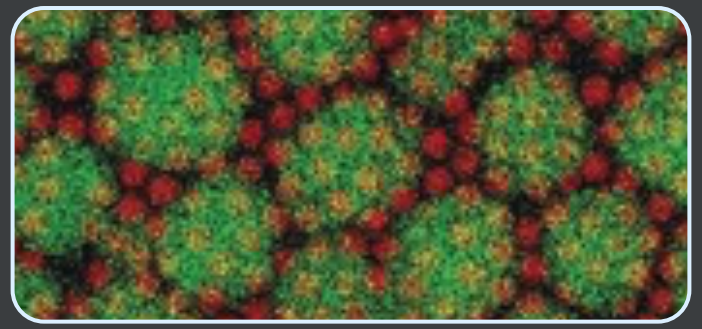
Applications

Where is ML being used in Science?

And what is different?

Opportunities for ML in Science





Analyze

- Classify
- Regression
- Cluster and denoise
- Extract features

Accelerate

- Design
- Surrogate models
- Inverse problems
- Generative models

Automate

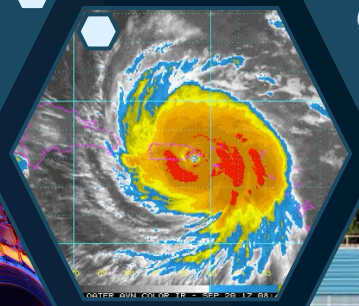
- Self-driving lab
- Instrument control
- Smart infrastructure
- Robotics

Cross-cutting themes for ML in Scienc

Physics-aware Learning

Uncertainty Quantifi-cation

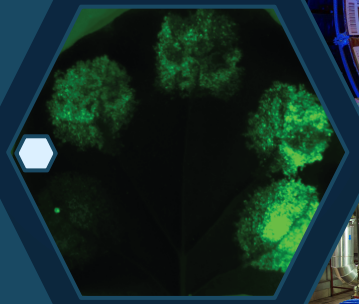
Learning across scales



Control of experiments

Inverse Design

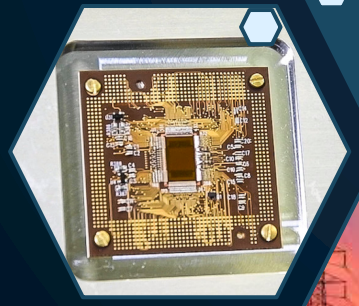
Interpret-ability



Complex, 3D+, sparse data

Transfer learning across instruments

Fairness



Federated learning on sensors

Thanks to Ben Nachman for several examples

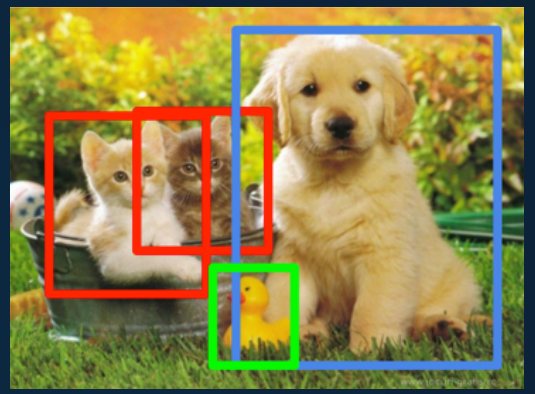
Data Analytics via Supervised Learning

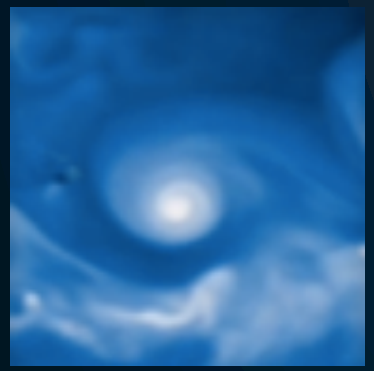
Classification

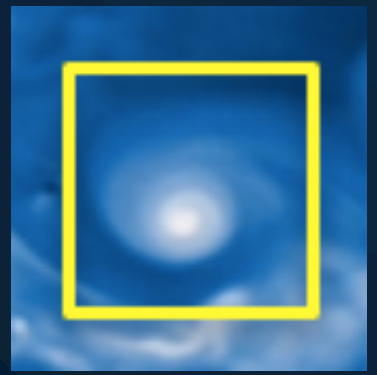
Classification + Localization

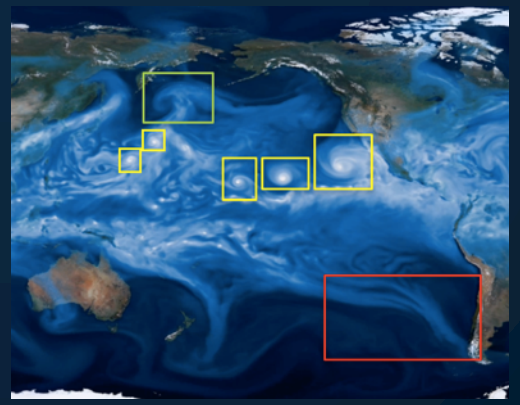
Object Detection

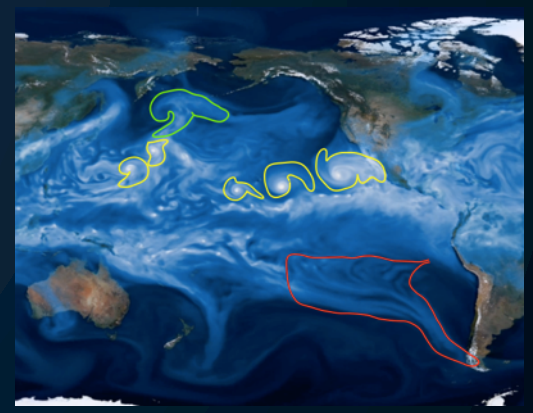
Instance Segmentation









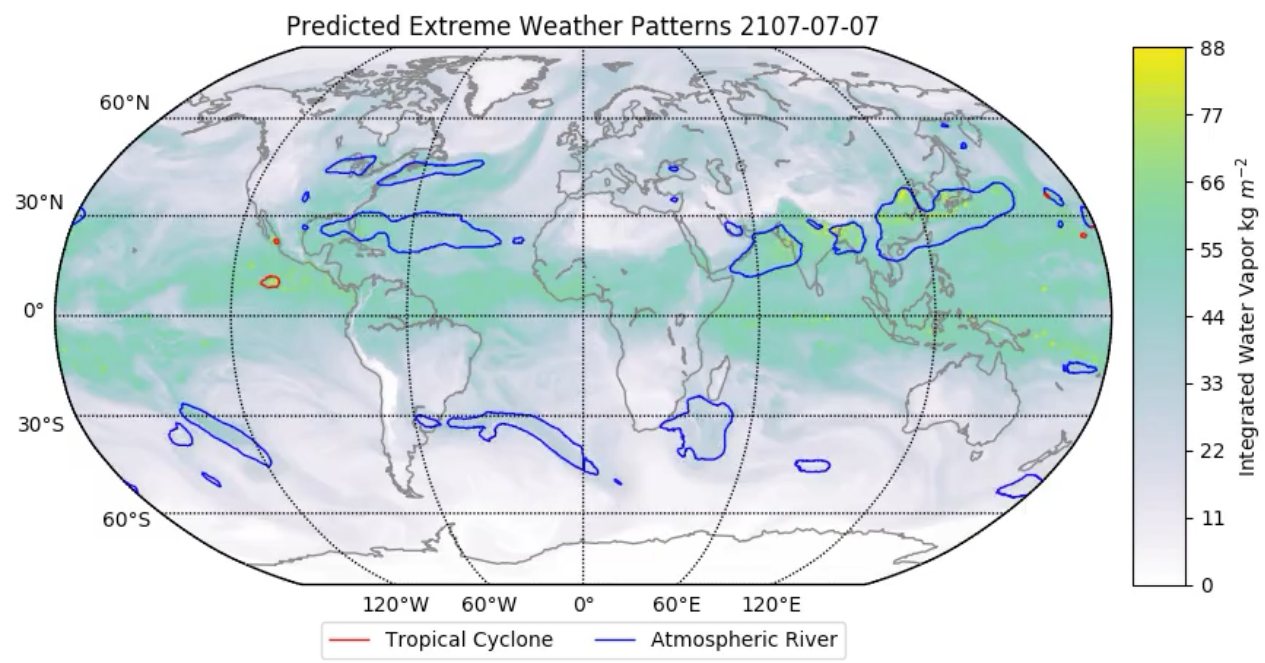


Extending image-based methods to complex, 3D, scientific data sets is non-trivial!

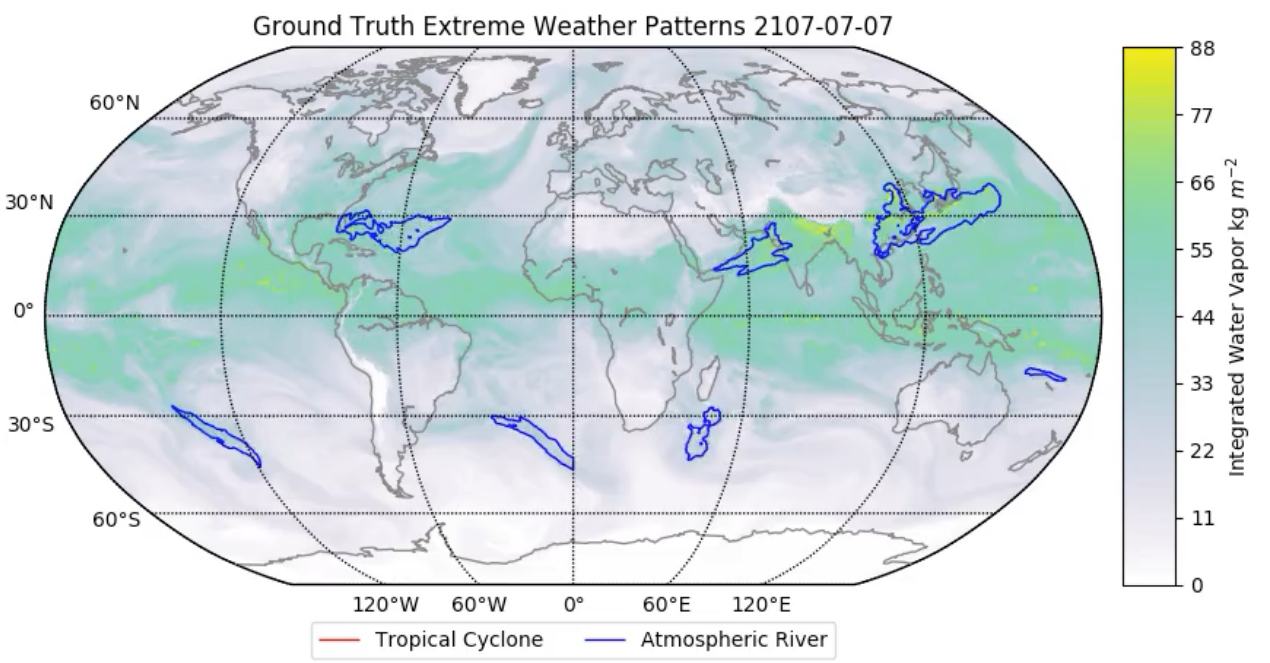
Slide source Prabhat, LBNL

CNNs for 3D Climate Simulation Data on HPC

Predicted Extreme Weather



Ground Truth Extreme Weather

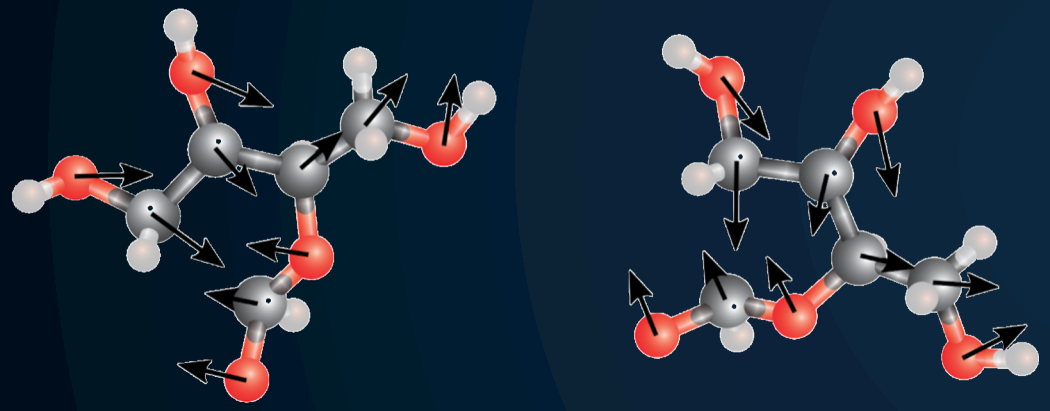


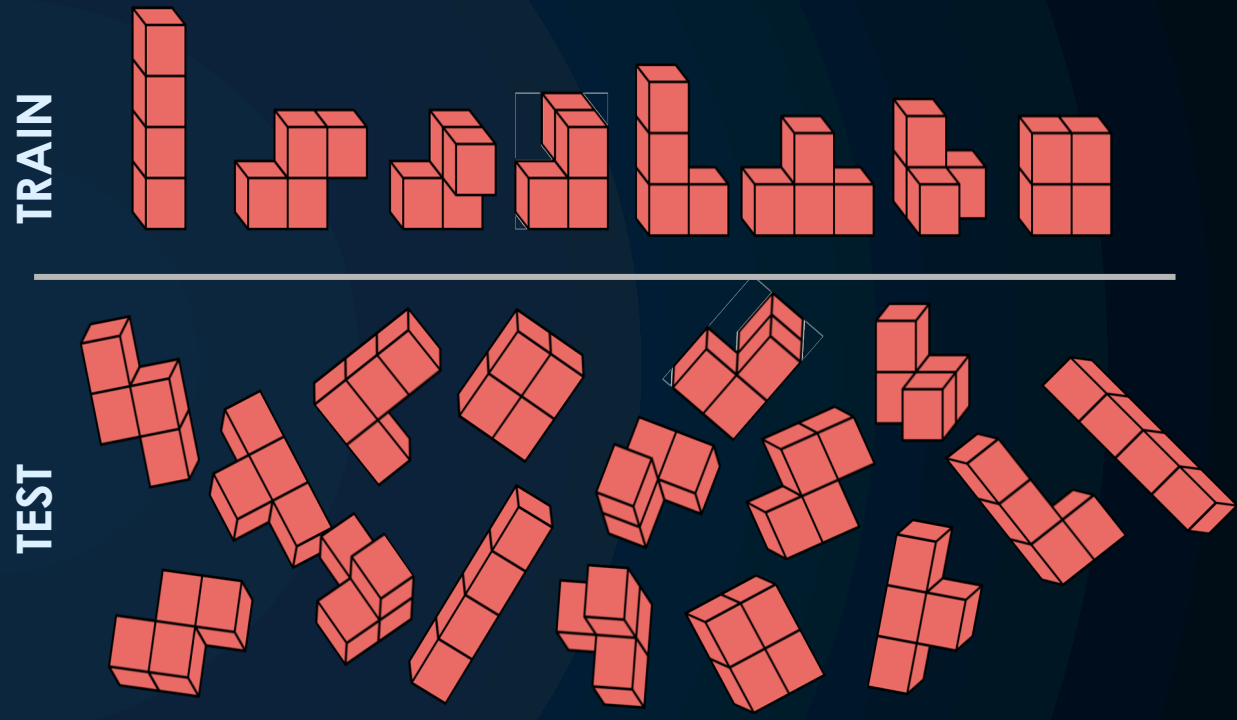
- **Deep learning results are smoother than heuristic labels**
- **Achieved over 1 EF peak on OLCF Summit: Gordon Bell Prize in 2018**

Thorsten Kurth, Sean Treichler, Joshua Romero, Mayur Mudigonda, Nathan Luehr, Everett Phillips, Ankur Mahesh, Michael Matheson, Jack Deslippe, Massimiliano Fatica, Prabhat, Michael Houston

CNNs for Materials with Physical Laws

Physics-aware learning



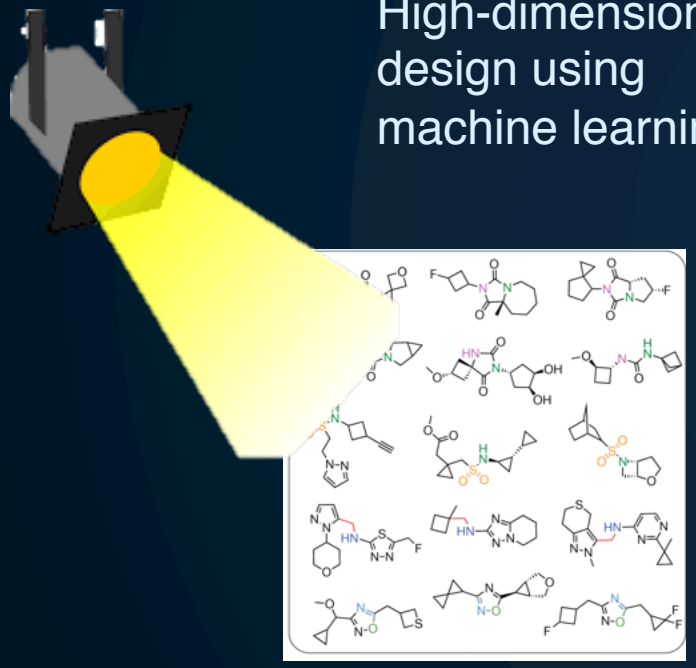


A network with 3D translation- and 3D rotation-equivariance

Slides from Tess Smidt and Risi Condor; E.g., 2018 paper by Thomas, Smidt, Kearnes, Yang, Li, Kohlhoff, Riley

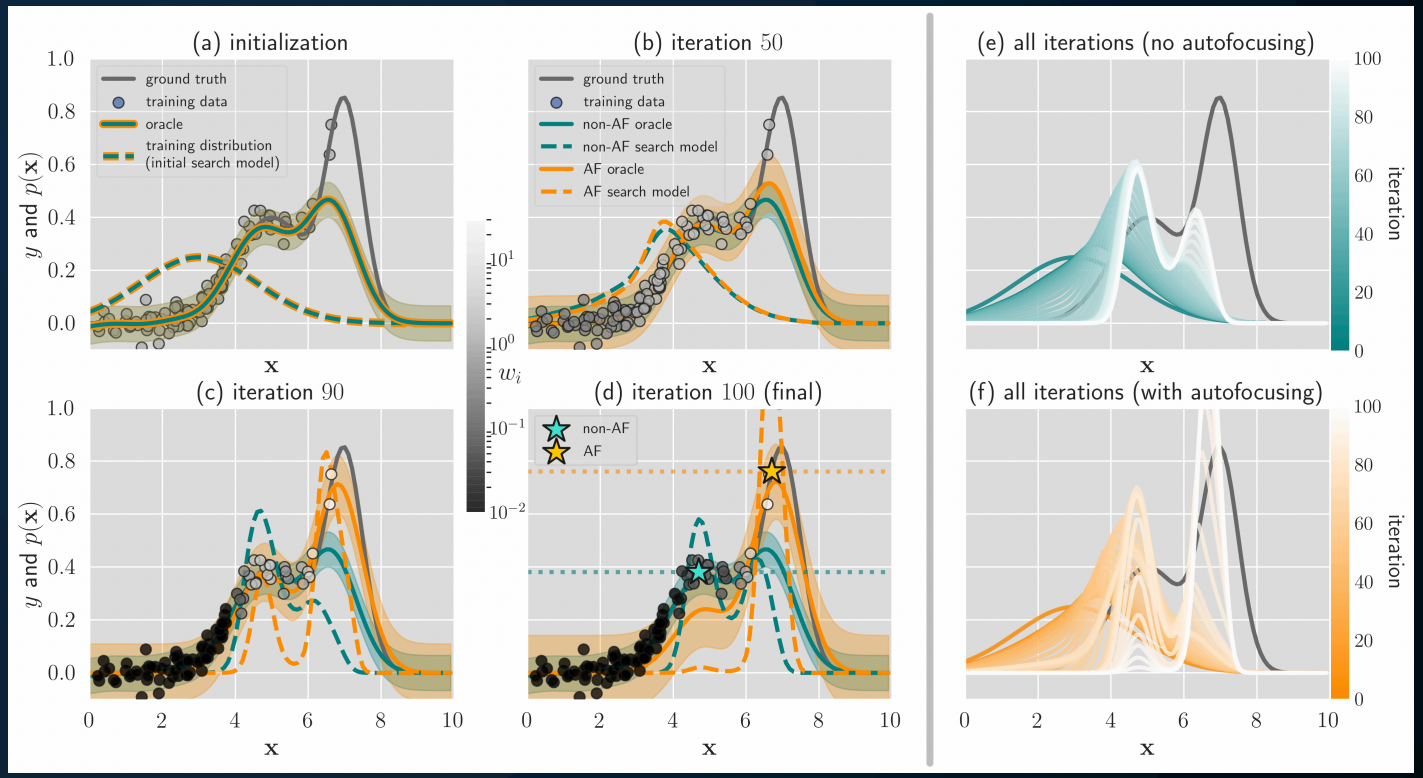
Inverse Design with ML

Designing materials, proteins, and small molecules with ML



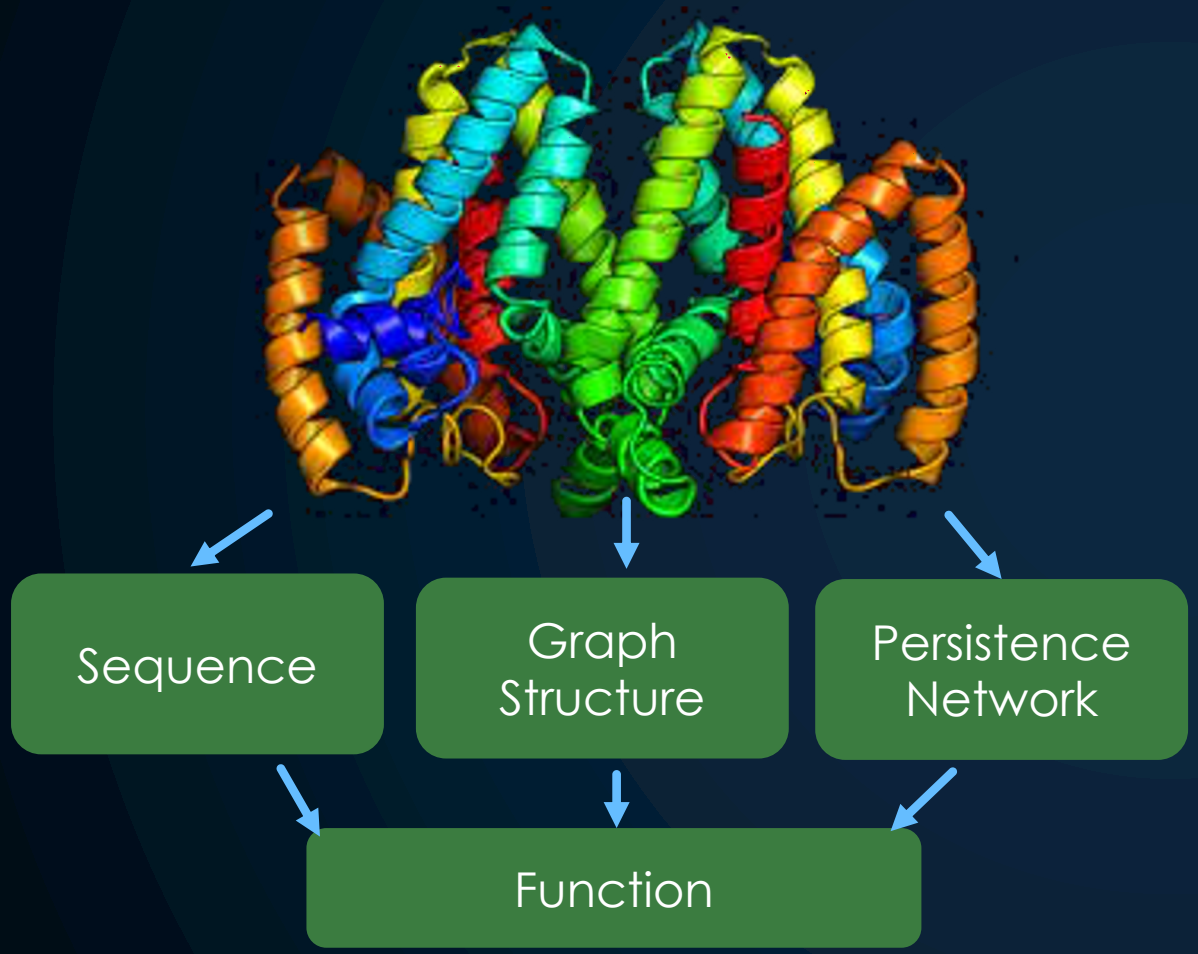
High-dimensional design using machine learning

Search for a molecules using an autofocusing generative model: moves around the design space, guided by an oracle

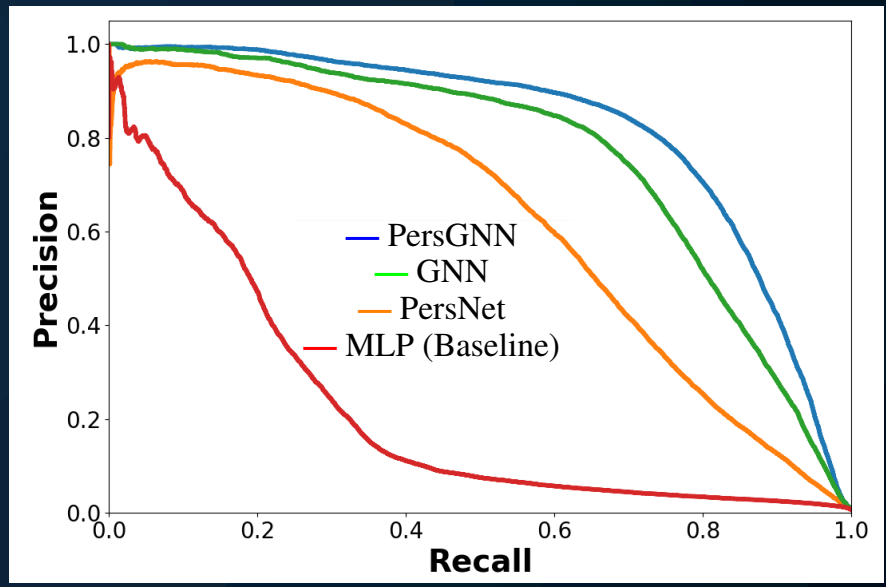


Clara Fannjiang and Jennifer Listgarten at NeurIPS '20

Learning from sequence + graph structure



Which proteins are good catalysts, bind to small molecules, etc.



Aditi S Krishnapriyan, Nicolas Swenson, Dmitriy Morozov, Katherine Yelick, Aydin Buluc

Graph NNs for Neutrino classification

- Apply graph convolutional techniques to irregular, 3D detector grid
- Increase sensitivity of IceCube detector: 6.3x more events
- And improve Signal-to-Noise ratio by 3x


17th IEEE International Conference on Machine Learning and Applications
IEEE December 17-20, 2018, Orlando, FL, USA AMLA

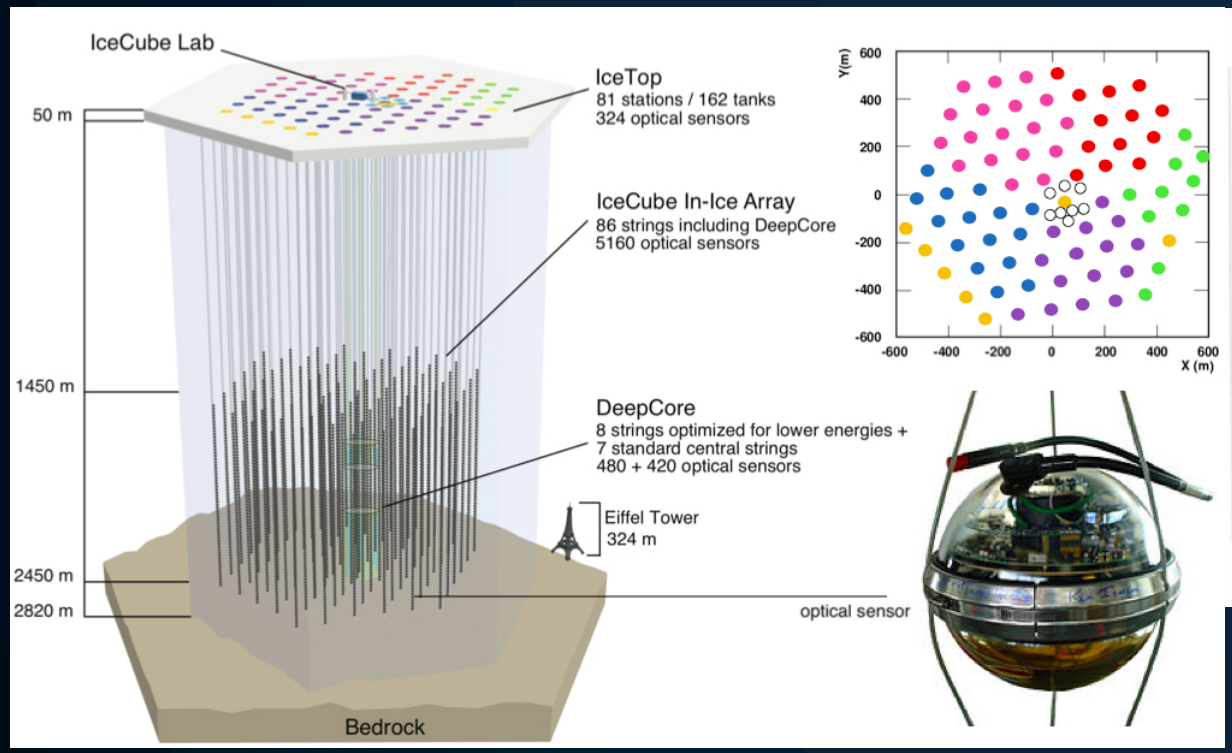
BEST PAPER AWARD

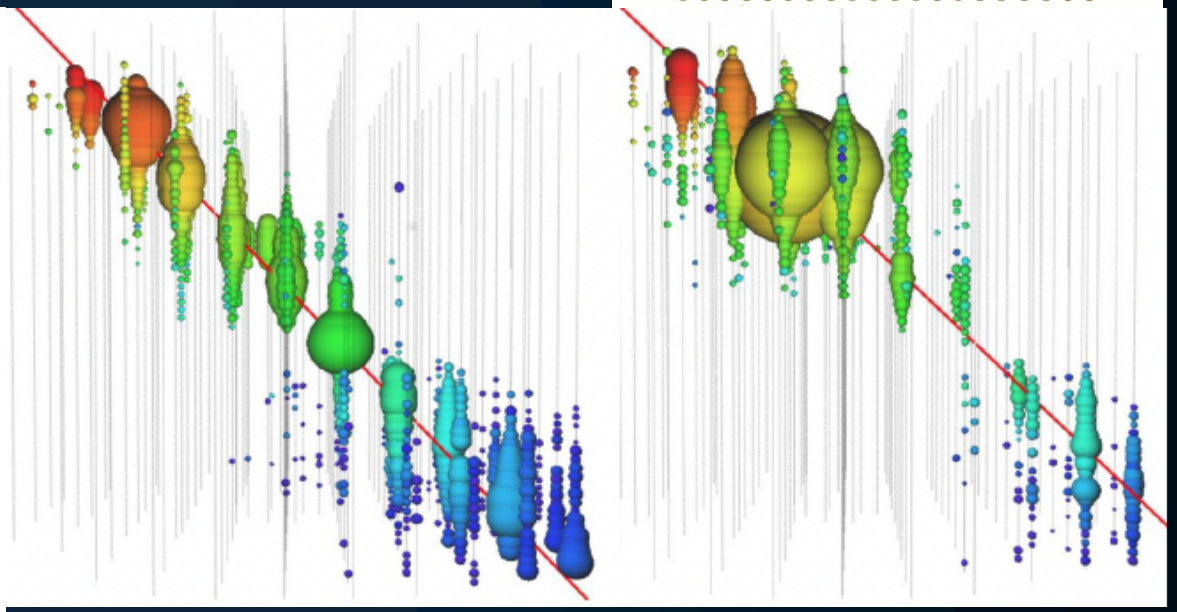
Graph Neural Networks for IceCube Signal Classification

Nicholas J Choma (New York University); Federico Monti (Università della Svizzera italiana); Lisa Gerhardt (Lawrence Berkeley National Laboratory); Tomasz Palczewski (University of California, Berkeley); Zahra Ronaghi (Lawrence Berkeley National Laboratory); Prabhat (Lawrence Berkeley National Laboratory); Wahid Bhimji (Lawrence Berkeley National Lab); Michael Bronstein (Imperial College); Spencer R. Klein(University of California, Berkeley); Joan Bruna (Courant Institute of Mathematical Sciences, NYU, USA)

Mehmed Kantardzic, Conference Co-Chair

Moamar Sayed-Mouchaweh, Conference Co-Chair



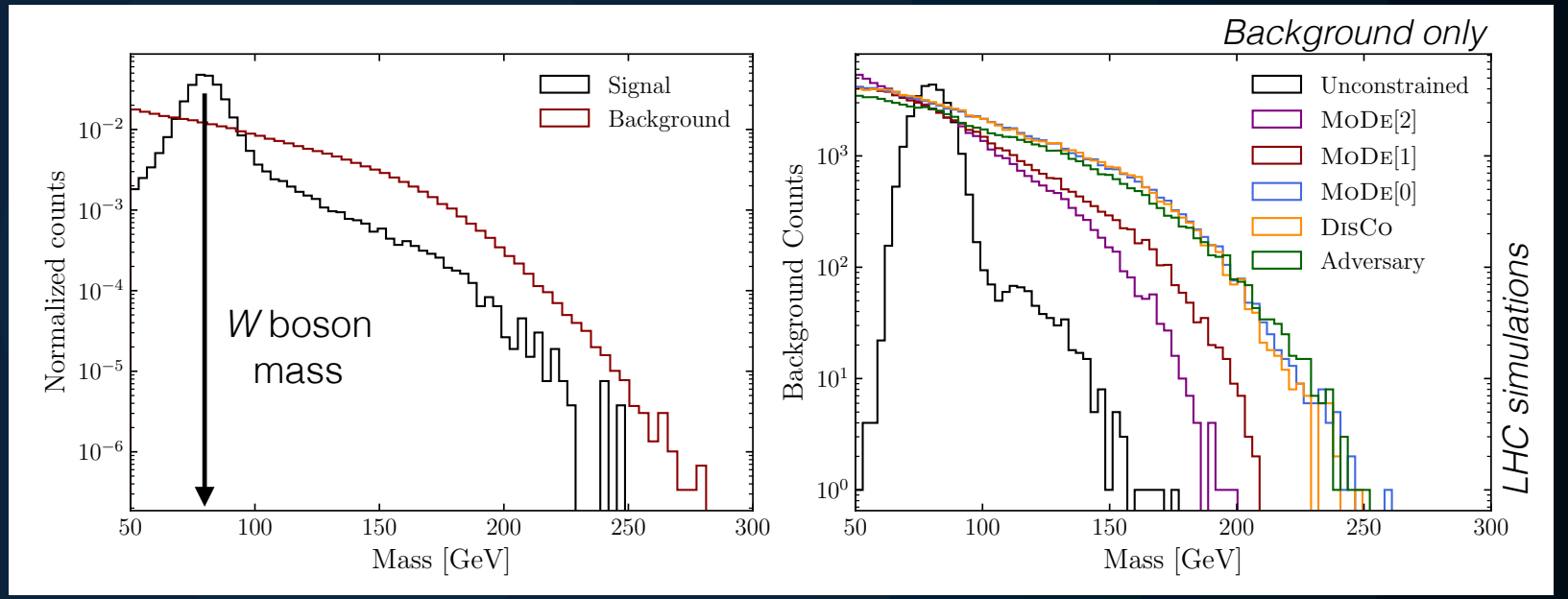


Pattern of light deposition for muons bundles (left) and a high-energy single muon (right).

Contributors: Nick Choma, Joan Bruna, Federico Monti, Michael Bronstein, Spencer Klein, Tomasz Palczewski, Lisa Gerhardt, Wahid Bhimji and IceCube collaboration

Fairness in Physics

Separating signal from noise in the search for Lorentz-boosted *W* bosons at Large Hadron Collider



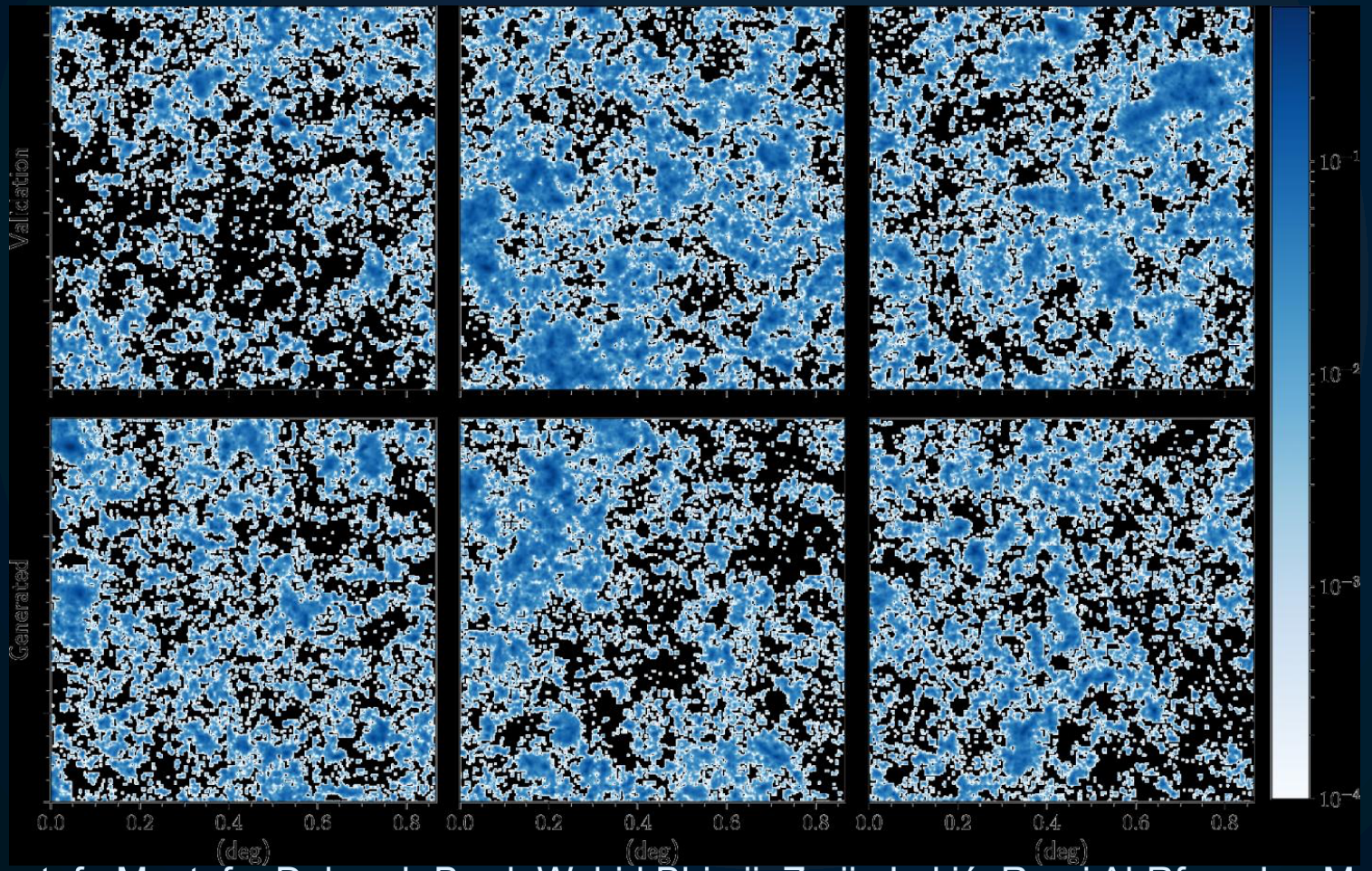
Signal and background events without selection.

Back-ground distributions at 50% signal efficiency (true positive rate) for different classifiers.

O. Kitouni, B. Nachman, C. Weisser, M. Williams, 2010.09745

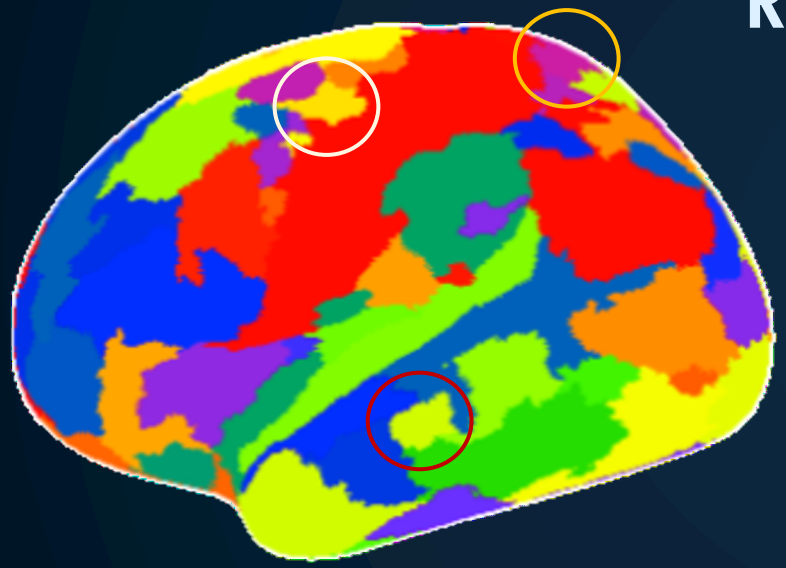
GANs to Build Scientific Data

Generate convergence maps of weak gravitational lensing, to help in understanding the physical laws governing the universe.



CosmoGAN: Mustafa Mustafa, Deborah Bard, Wahid Bhimji, Zarija Lukić, Rami Al-Rfou, Jan M. Kratochvil

Learning Relationships with Graphical Models



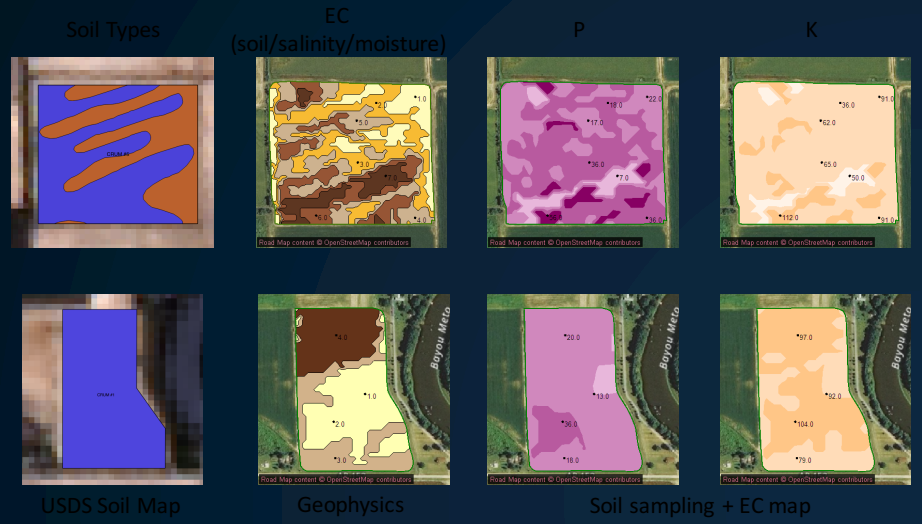
Discovering Regions and Co-Regions of Brain Activity from fMRI

91K x 91K Sample Covariance matrix

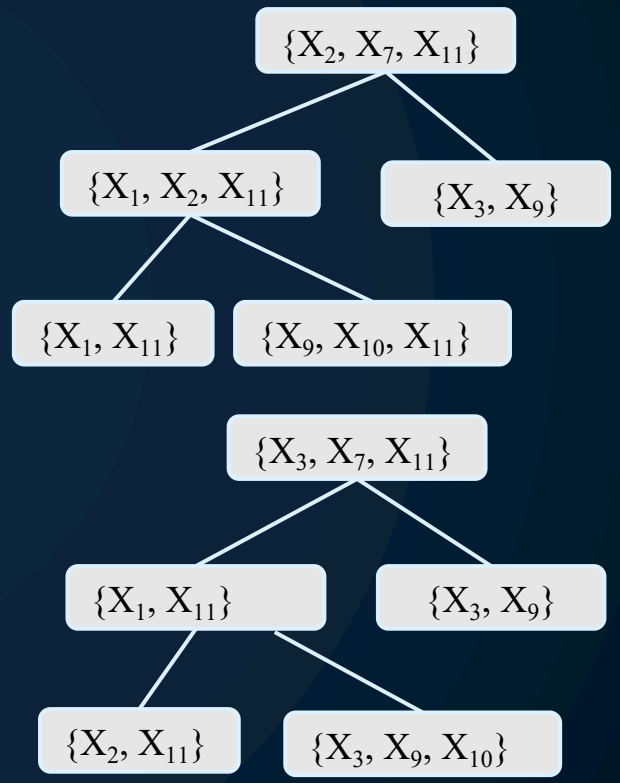
- 91K data points (each $2mm^3$)
- 5K time points (0.7 secs for 2 hrs)
- Averaged over 1,200 subjects

Koanantakool, Buluc, Morozov, Oliker, Yelick, Oh, AISTAT 2018.

Learning Mechanistic Models



Iterative Random Forest
High dimensional, sparse data



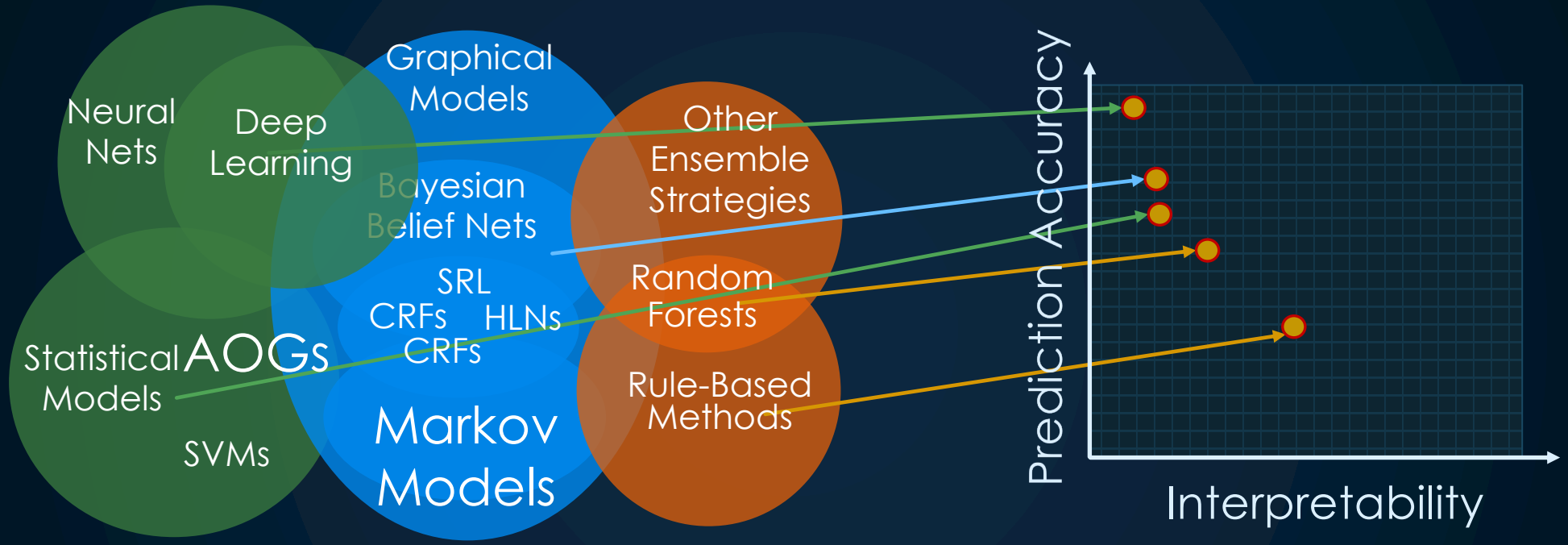
4D Virtual Farmland

- Hyperspectral imaging
- Environmental sensors
- Microbiome sequencing

- **Design microbial amendments**

Schaettle, K. B.; Falco, N.; Ulrich, C.; Dafflon, B.; Wainwright, H. M.; Brown, J. B.

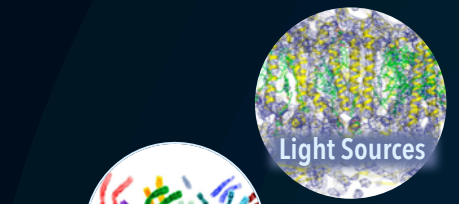
Robust, interpretable, explainable methods

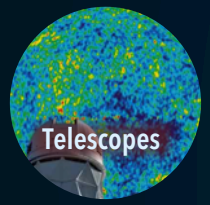


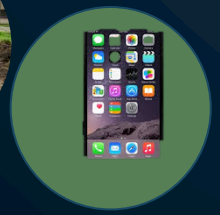
- Our goal for interpretability is methods that are useful for science and engineering applications

Graph based on image from: David Gunning, DARPA Explainable AI (XAI) Program

Edge Computing and Automation in Science







Edge Computing for Science

ESnet

Computing and Data Facilities

Experimental Facilities

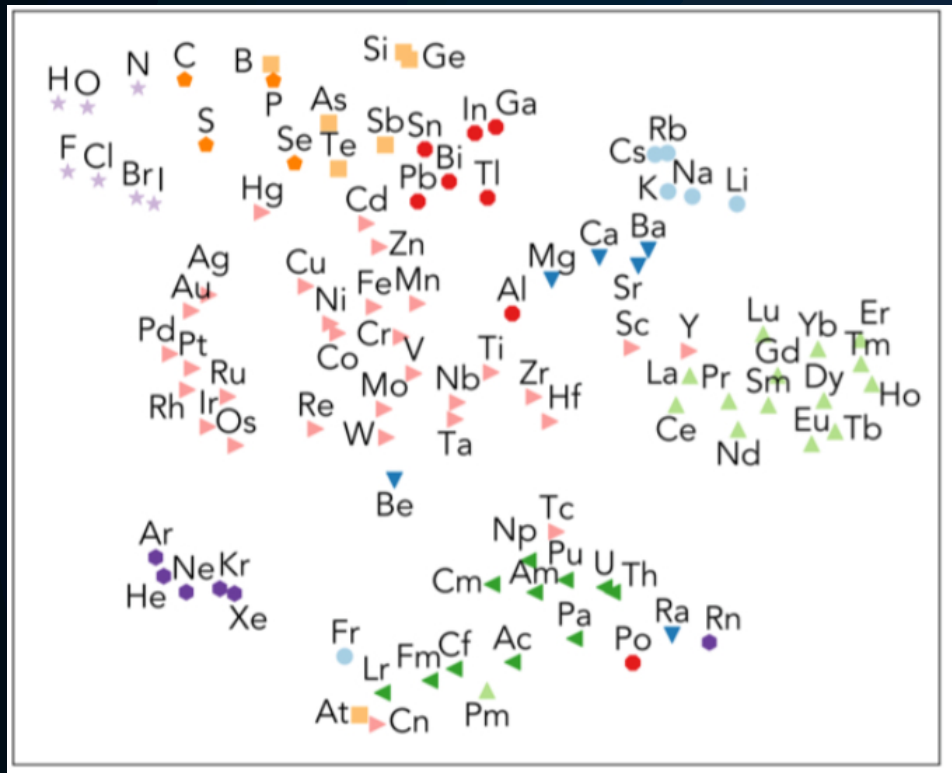
User Community

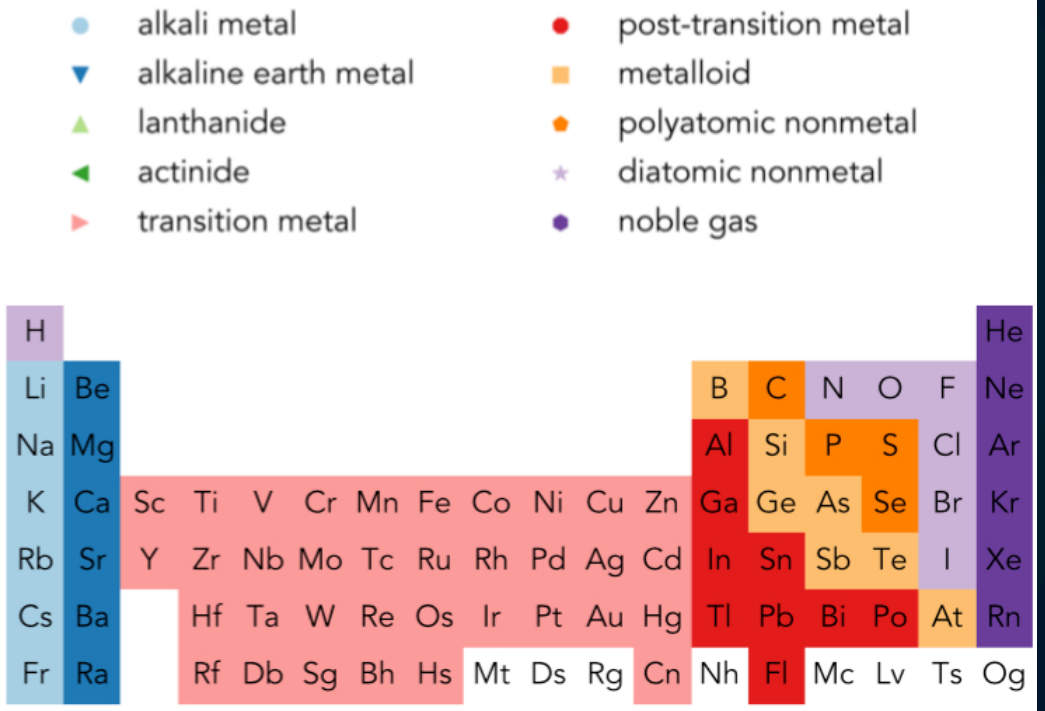
Embedded Sensors

Interconnected facilities where data is acquired, stored, analyzed and served

Using NLP on scientific publications

Analyze 3.3 million abstracts from materials science papers

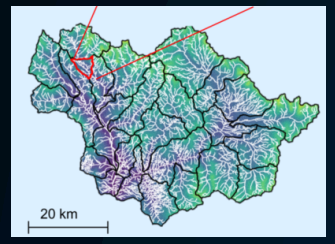




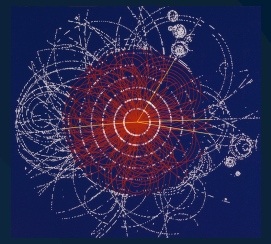
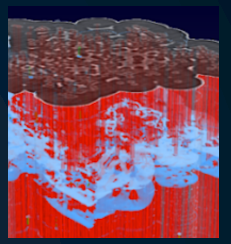
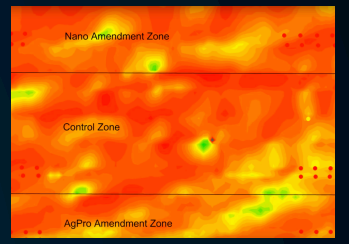
Word2vec's representation of the elements, projected onto two dimensions

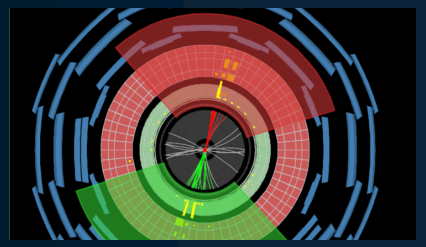
Vahe Tshitoyan, Leigh Weston, John Dagdelen, Anubhav Jain

Machine Learning at Berkeley Lab

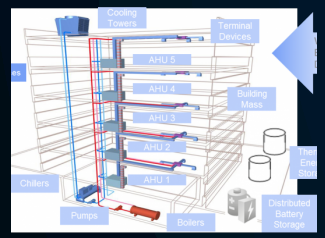


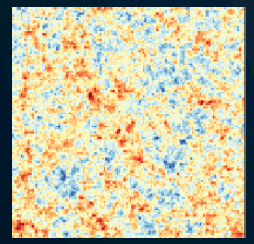


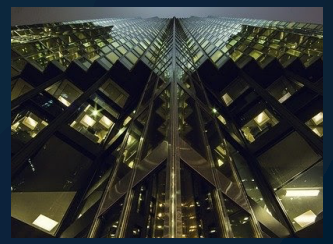
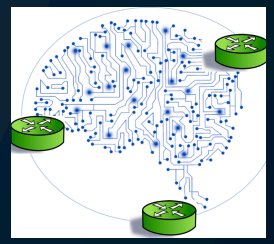
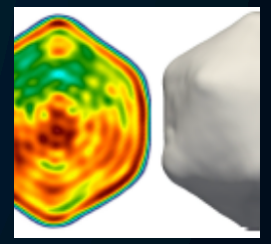





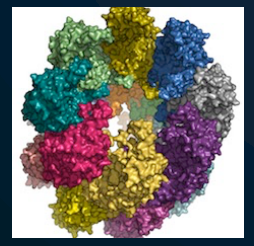
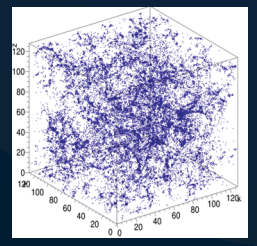



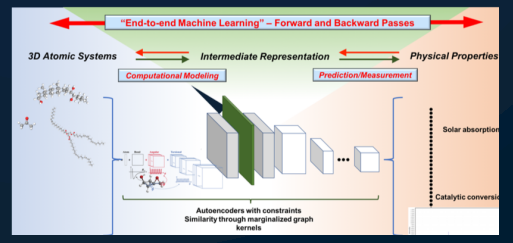


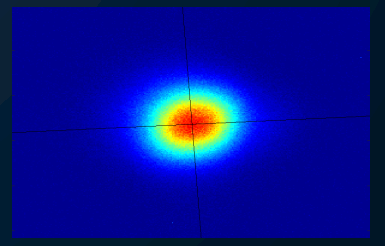





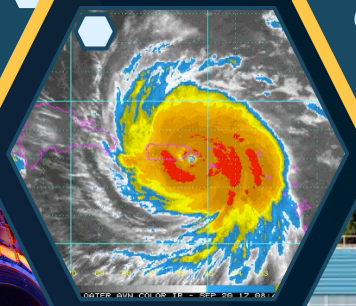



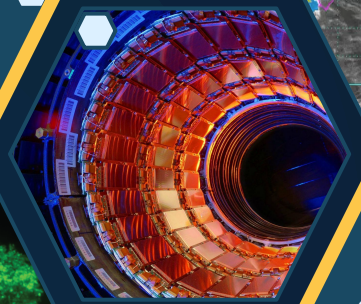


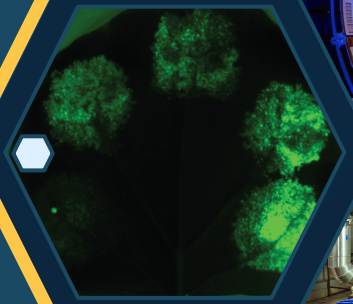


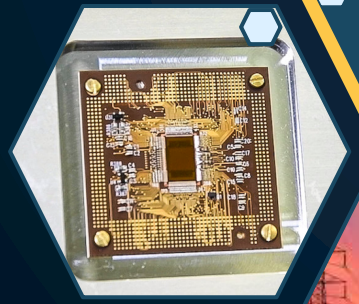
https://ml4sci.lbl.gov

Cross-cutting application themes


Physics-aware Learning
Uncertainty Quantifi-cation
Learning across scales

Control of experiments

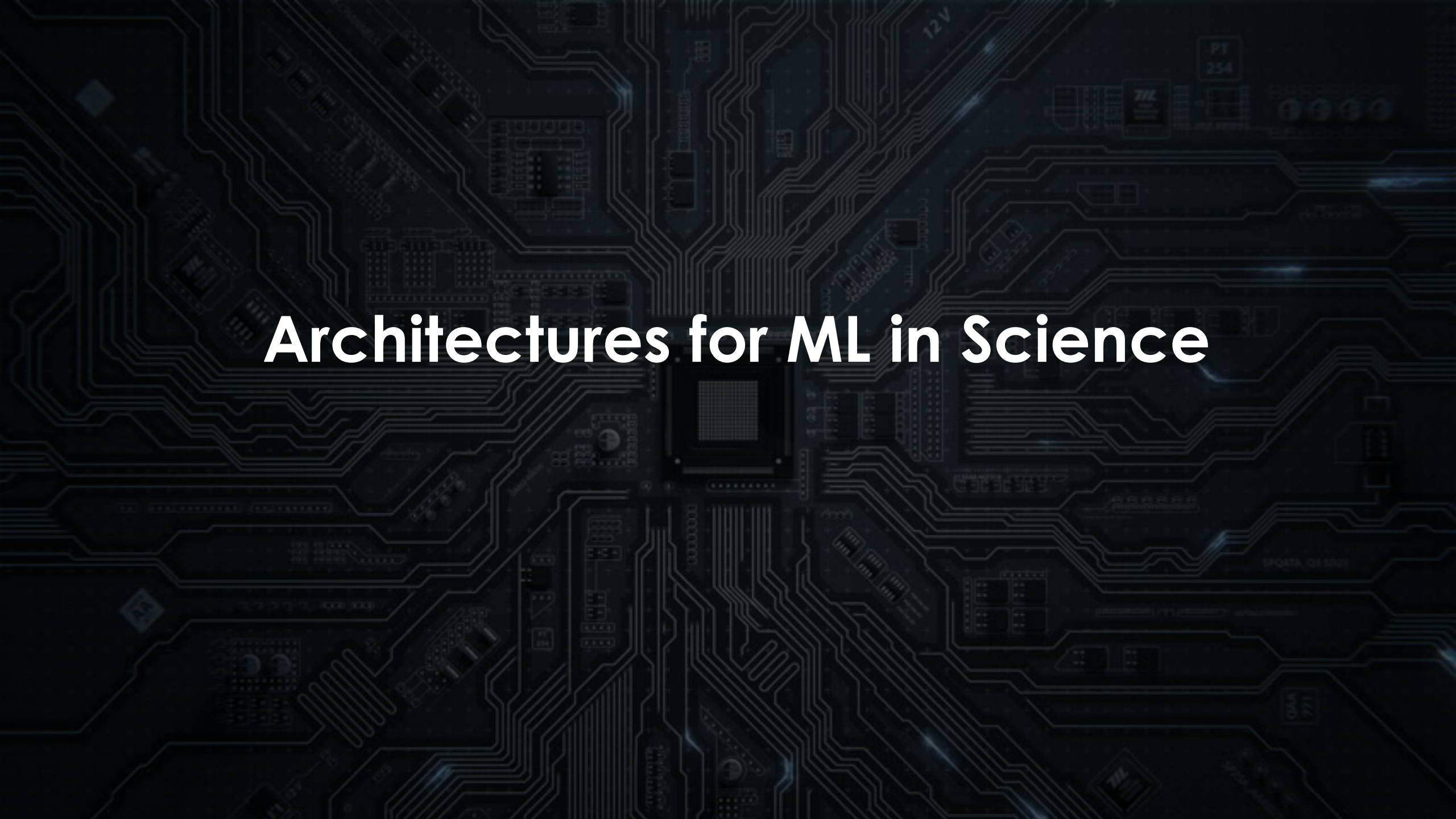
Inverse Design


Interpret-ability

Complex, 3D+, sparse data
Transfer learning

Fairness


Federated learning on sensors


Thanks to Ben Nachman for several examples

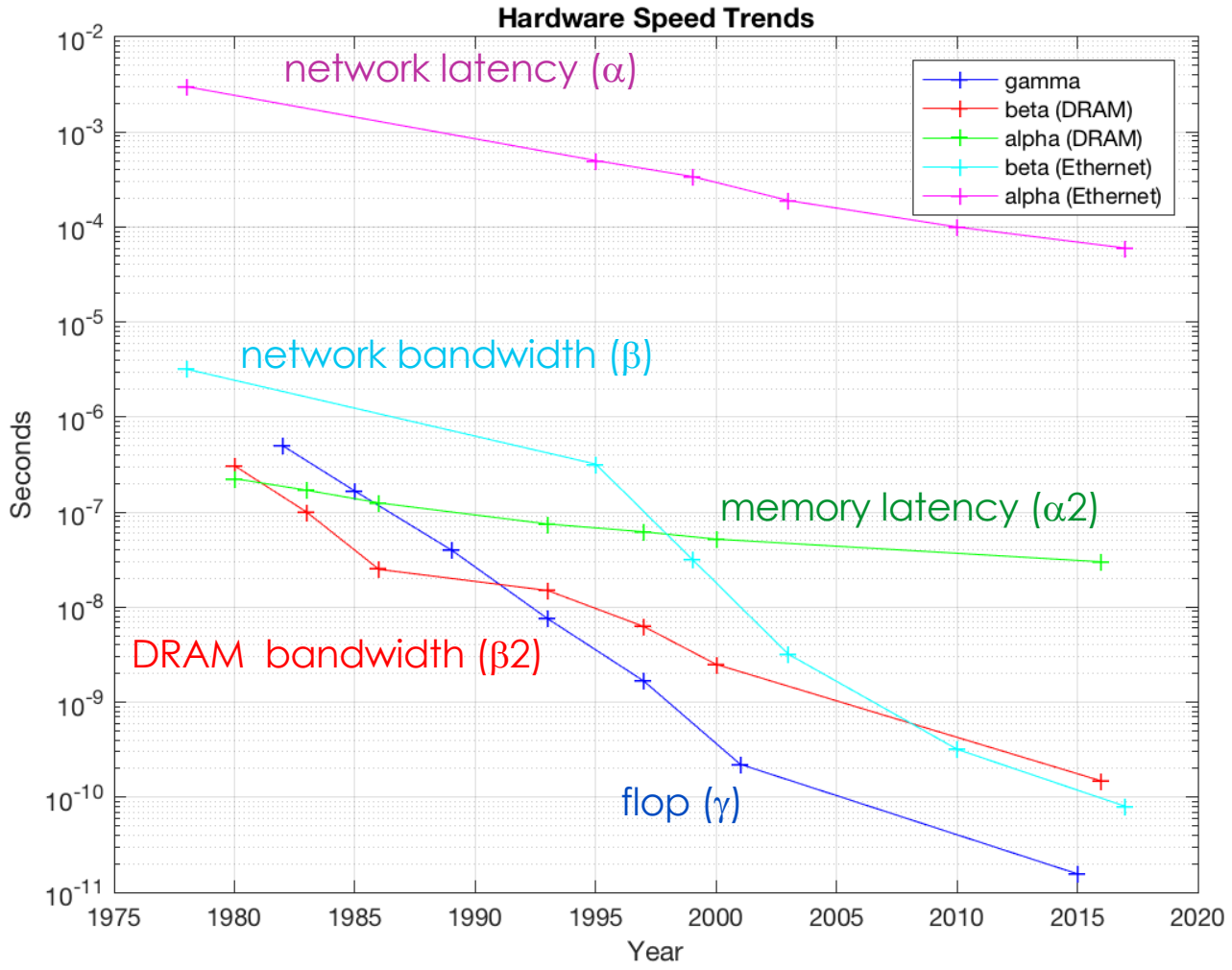

Architectures for ML in Science

Moore's Law

It's hard to think exponentially

But it's also hard to stop

Communication Dominates



Time =

flops * γ +

message * α +
bytes comm * β +

diff memory locs * α2 +
memory words * β2

Data from Hennessy / Patterson, Graph from Demmel

Data Movement is Expensive

Hierarchical energy costs.

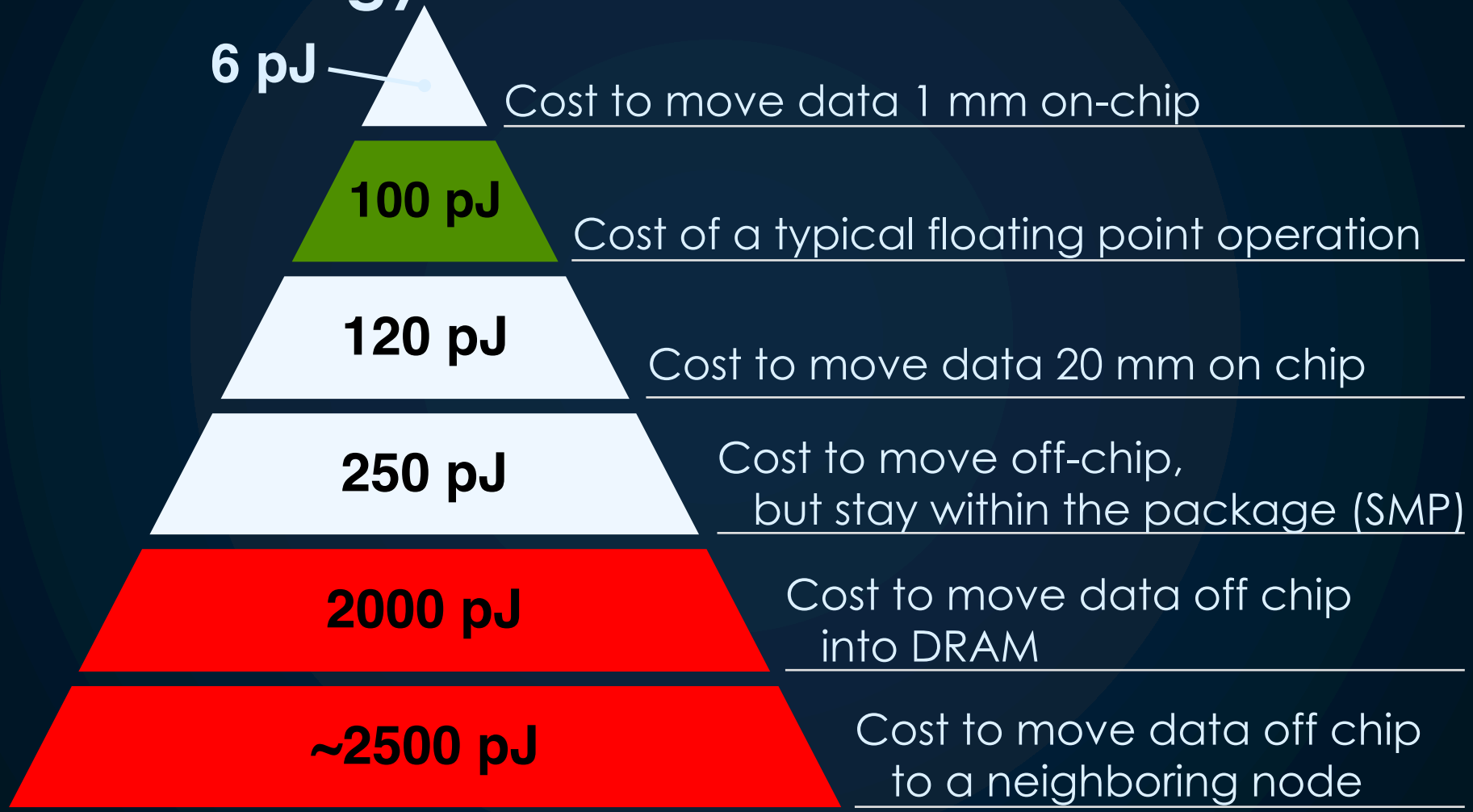
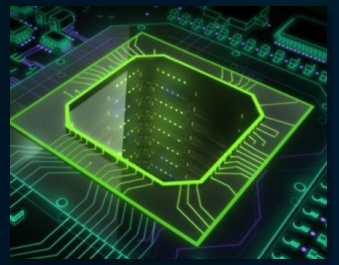


Image: http://slideplayer.com/slide/7541288/

Trend Toward Specialization

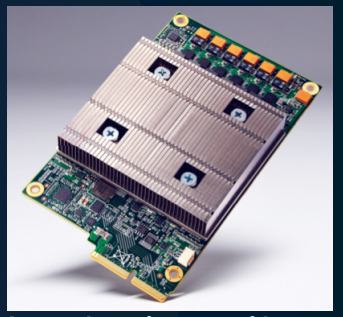


NVIDIA builds deep learning appliance with P100 Tesla's

FPGAs in Microsoft cloud

RISC-V is an open hardware platform

Intel buys deep learning startup, Nervana



Google designs its own Tensor Processing Unit (TPU)

Specialization Spectrum

Full Custom | *Open ISA* | *FPGA* | *FPGA + standard ops* | *Old GPU* | *GPGPUs* | *Simple cores* | *High end cores*

China (Sunway), Japan (ARM), and Europe/Barcelona (RISC-V) are doing this in HPC

AI Chip Landscape

More on https://basicmi.github.io/AI-Chip/

Tech Giants/Systems

Google
Microsoft
aws
IBM
facebook
Apple
Tesla
HUAWEI
Baidu 百度
Alibaba Group 阿里巴巴集团
FUJITSU
NOKIA
TOSHIBA
Hewlett Packard Enterprise
DELL

IC Vender/Fabless

intel
SAMSUNG
NVIDIA
QUALCOMM
AMD
NXP
ST
XILINX
MEDIATEK
BROADCOM
MARVELL
Rockchip 瑞芯微电子

IP/Design Service

arm
SYNOPSYS
Imagination
cadence
CEVA
VeriSilicon
SiFive
ARTERIS IP
Alchip
GUC
FARADAY
eSilicon

Startup in China

Cambricon 寒武纪科技
地平线 Horizon Robotics
BITMAIN
intellifusion 云天励飞
ChipIntelli
Think Force
Canaan
云知声 Unisound
Rokid
AISPEECH 思必驰 专注人性化的智能语音
肇观电子 NextVPU
Enflame
亿智科技
清微智能 TSING MICRO

Startup Worldwide

cerebras
WAVE COMPUTING
Graphcore
habana
thinci
SambaNova SYSTEMS
KALRAY
LIGHTELLIGENCE
HAILO Empowering Intelligence
Esperanto TECHNOLOGIES
Tenstorrent
MYTHIC
Preferred Networks
brainchip
PEZY Computing
GREENWAVES TECHNOLOGIES
AIMOTIVE
KONIKU
Tachyum
flexlogix
SYNTIANT
gyrfalcon technology
NOVUMIND

扫码访问AI芯片文章

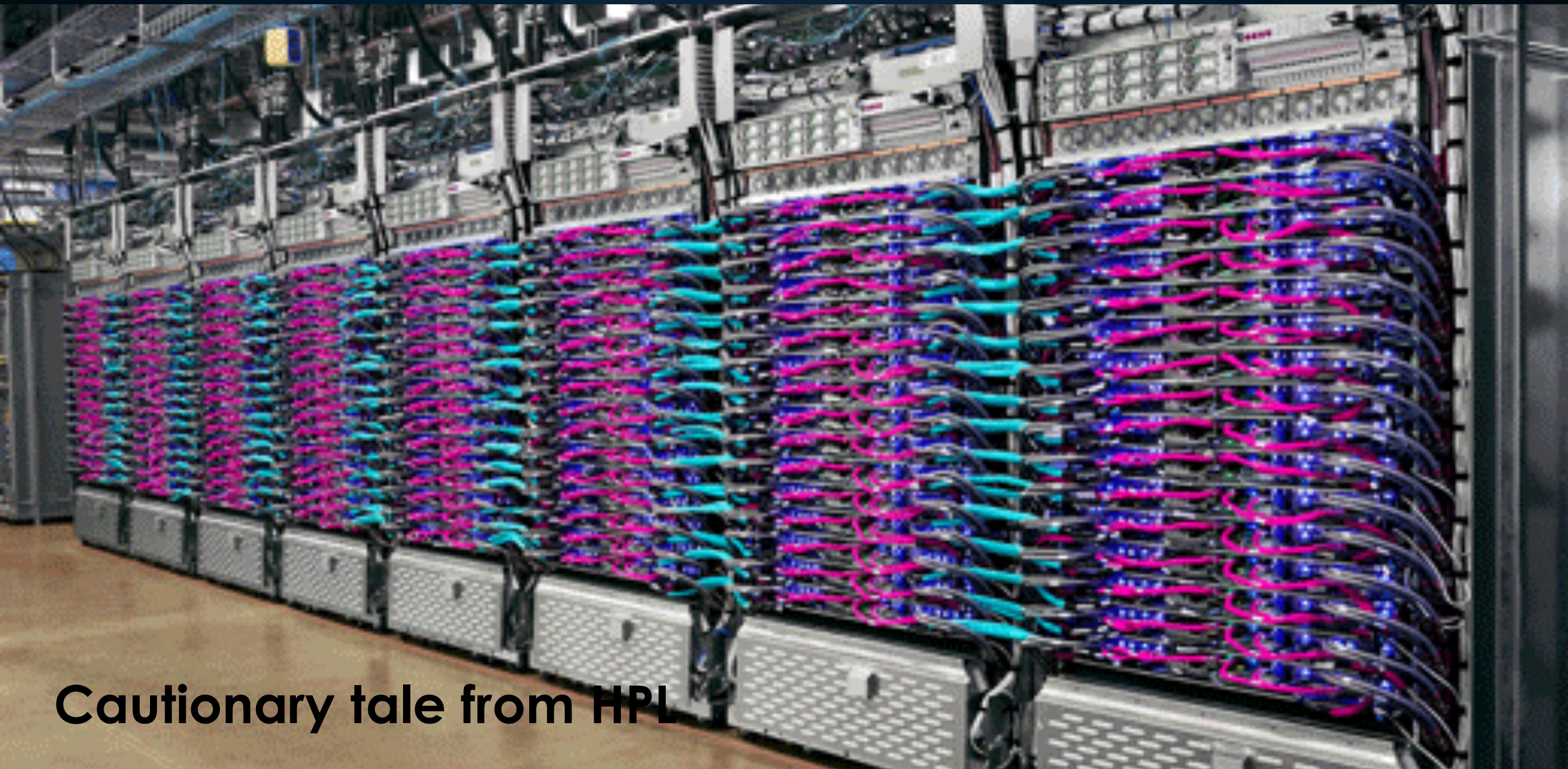
Compiler

XLA
GLOW
tvm
NVIDIA TensorRT
ONNC
nGraph Compiler stack (Beta)
plaidML

Benchmarks

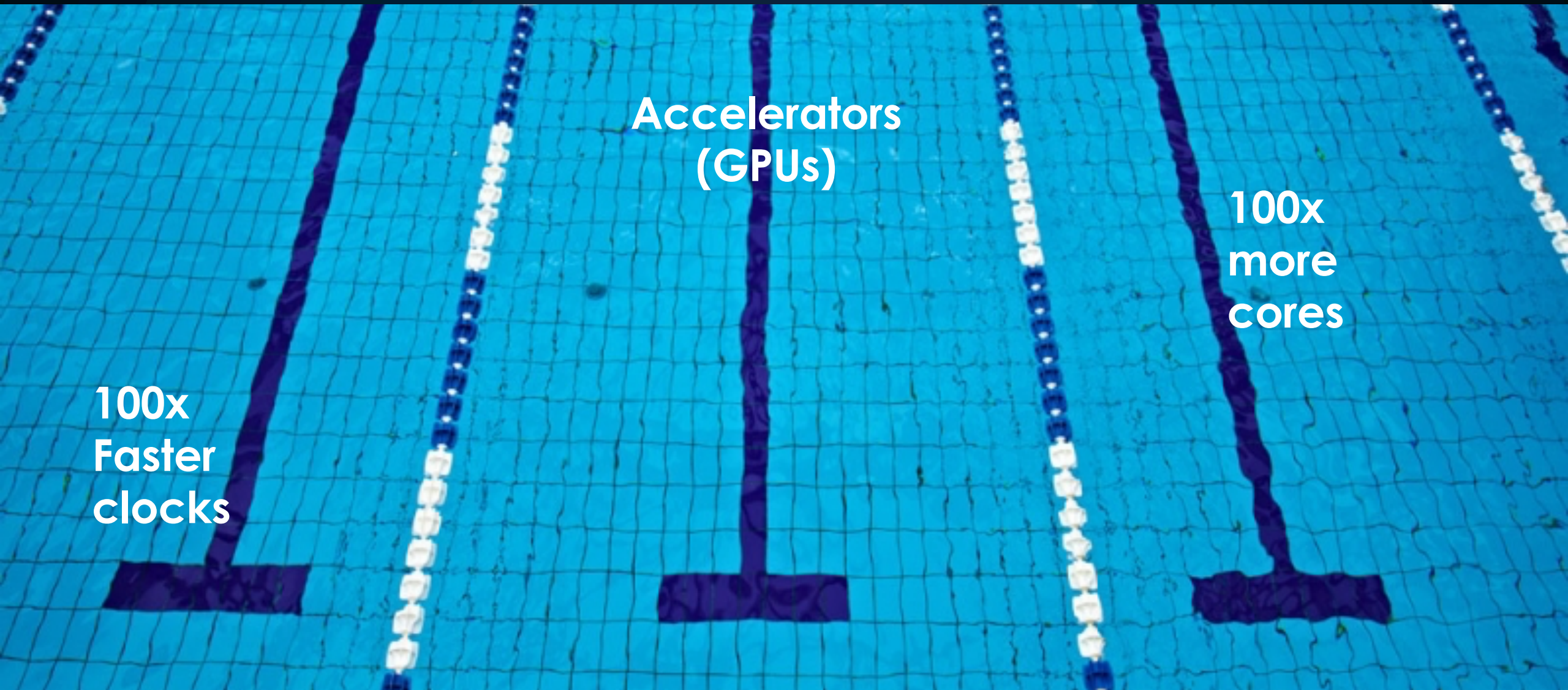
MLPerf
AI - Benchmark
AI Matrix.

Are CNNs the only application?



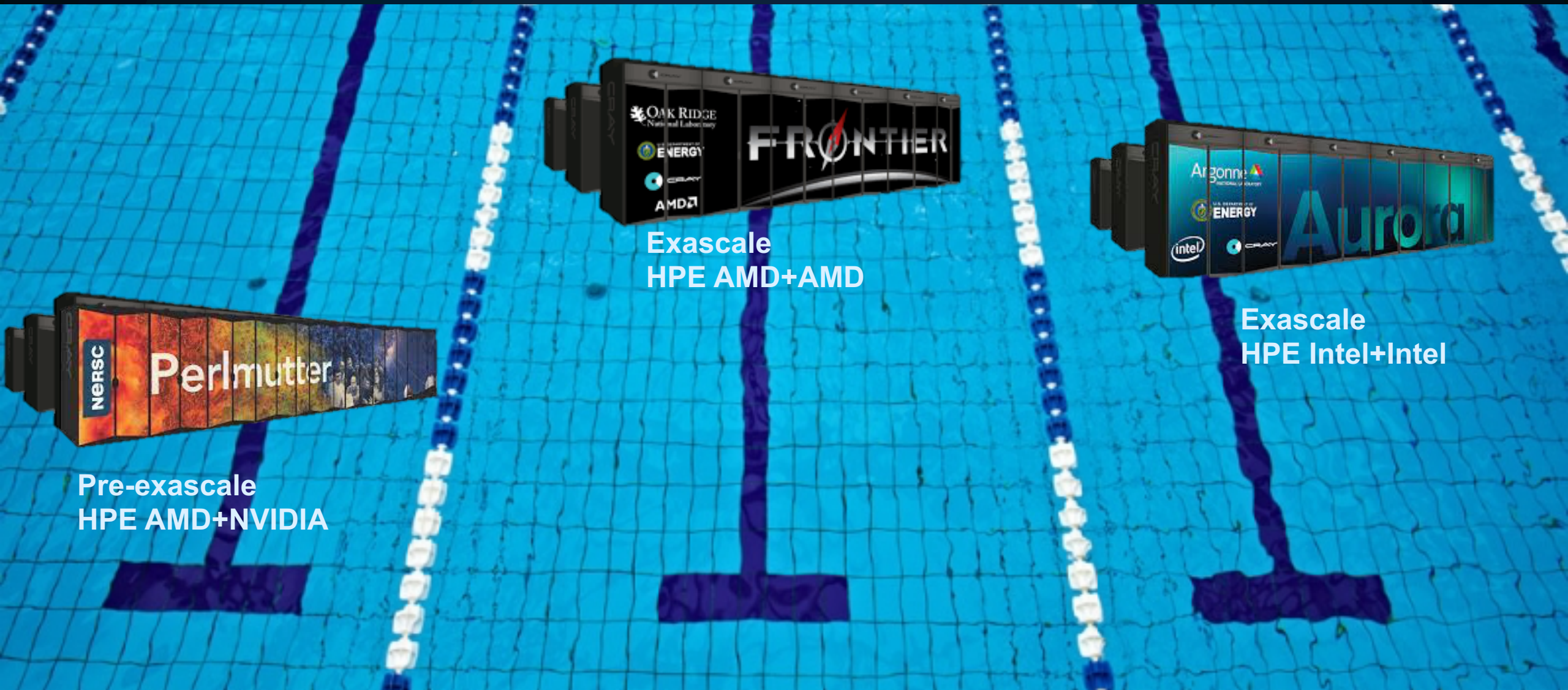
Cautionary tale from HPL

Exascale Architecture Plans (2008)



Exascale Architecture Plans (2021)

US DOE Office of Science Systems

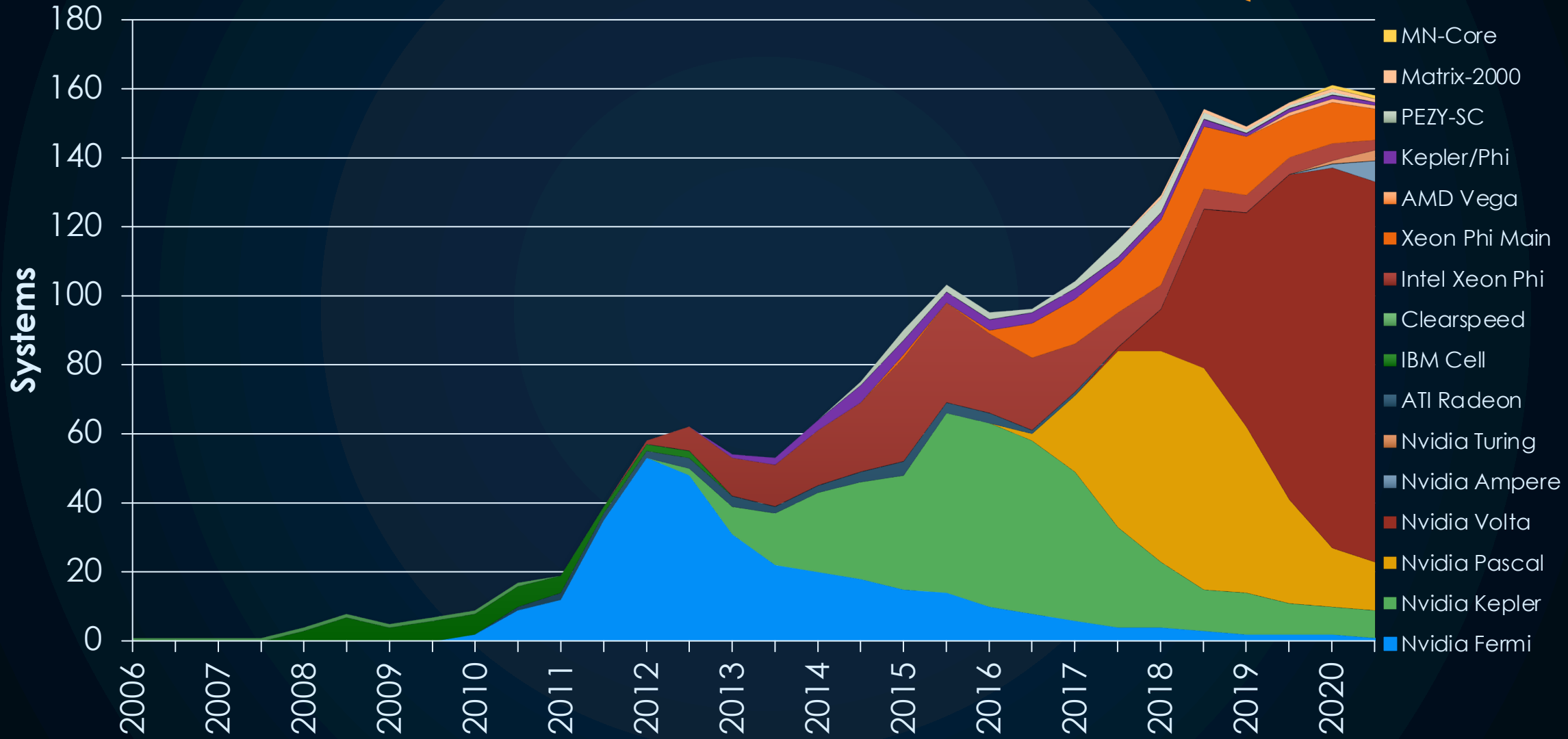


Pre-exascale
HPE AMD+NVIDIA

Exascale
HPE AMD+AMD

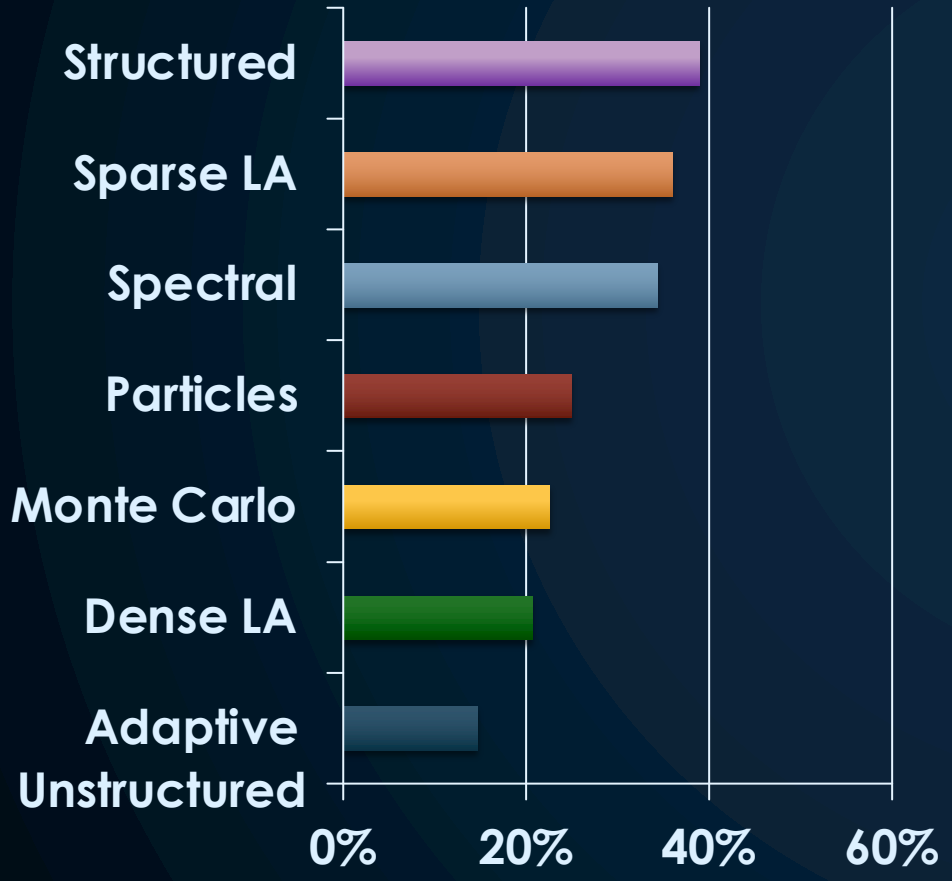
Exascale
HPE Intel+Intel

Accelerators in the Top500



Domain-Specific Code Generators
for specialized hardware

NERSC survey: what motifs do they use?

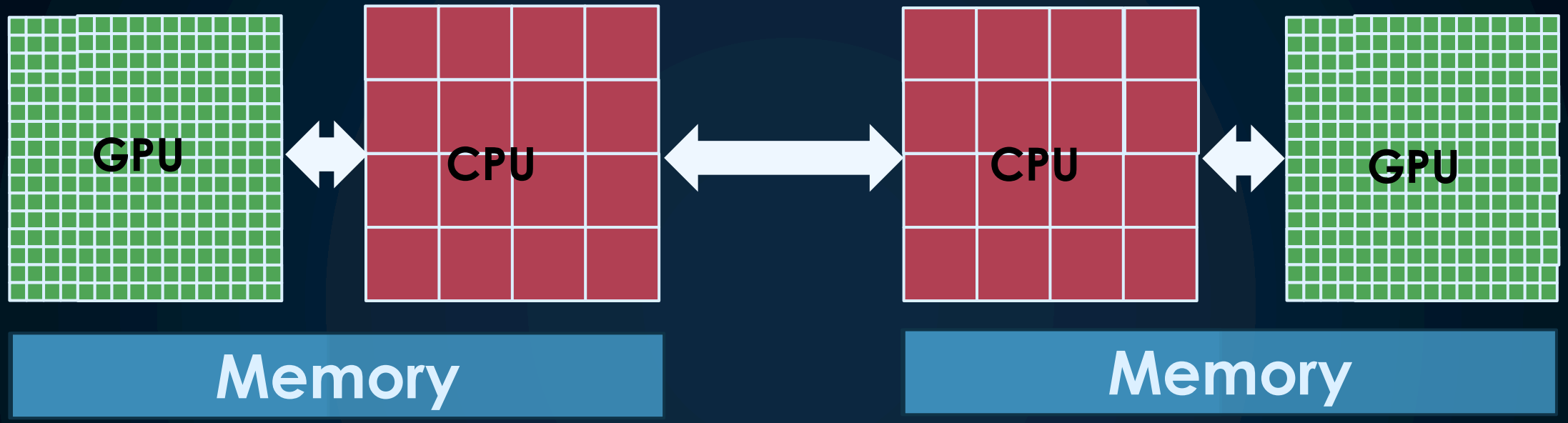


What code generators do we have?

Dense Linear Algebra	Atlas, Magma
Spectral Algorithms	FFTW, Spiral
Sparse Linear Algebra	OSKI, Bernoulli
Structured Grids	Halide, Orio, Snowflake,
Unstructured Grids	GraphIT
Particle Methods	
Alignment	SALSA

And many more

Specialization, Yes Accelerators, No!



More cores

More data parallelism

Narrow data types

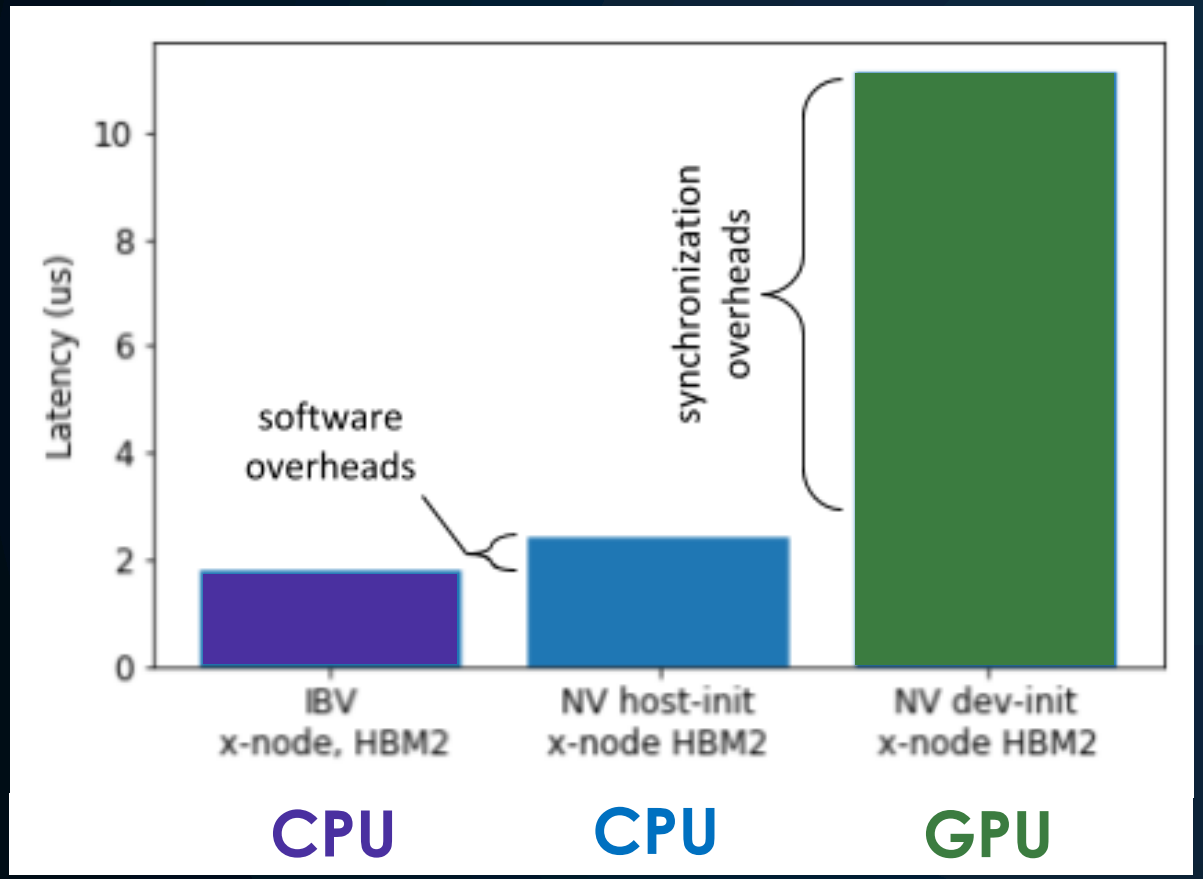
More memory spaces

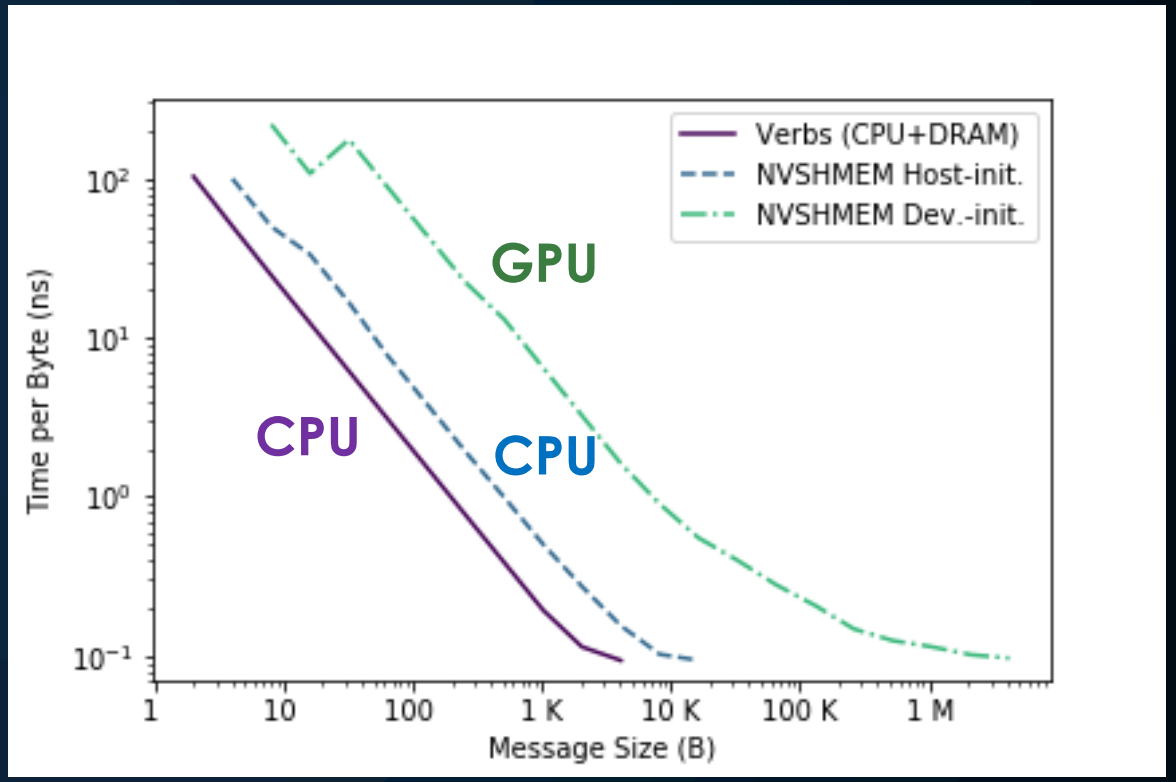
CPUs in control

CPUs communicate

Put Accelerators in Charge of Communication

Architecture and software are not yet structured for accelerated-initiated communication (Summit with NVLink between Power9 CPUs and NVIDIA GPUs)





Taylor Groves et al

Science + HPC

AI + Cloud

Google 1997

NERSC 1996

Cray T3E 900: 460 Gflop/s

850 GB of disk

106 TB Data archive

Cloud vs HPC

Cloud	HPC
~~Focus on storage~~	~~Focus on computing (flop/s)~~
~~Cheap commodity components~~	~~High end components (some specialization)~~
~~Commodity networks~~	~~High performance networks~~
On-node disks (air cooled)	Separate storage (compute liquid cooled)
Resilience, replicated SW	Integrated & efficient SW
Pay as you go	Purchased for mission; pay in "hours"
< 50% utilization	> 90% utilization
On-demand access	Large jobs wait in queues
Multiple jobs per node	Dedicated nodes
Evergreen procurement	~4 year procurement cycles

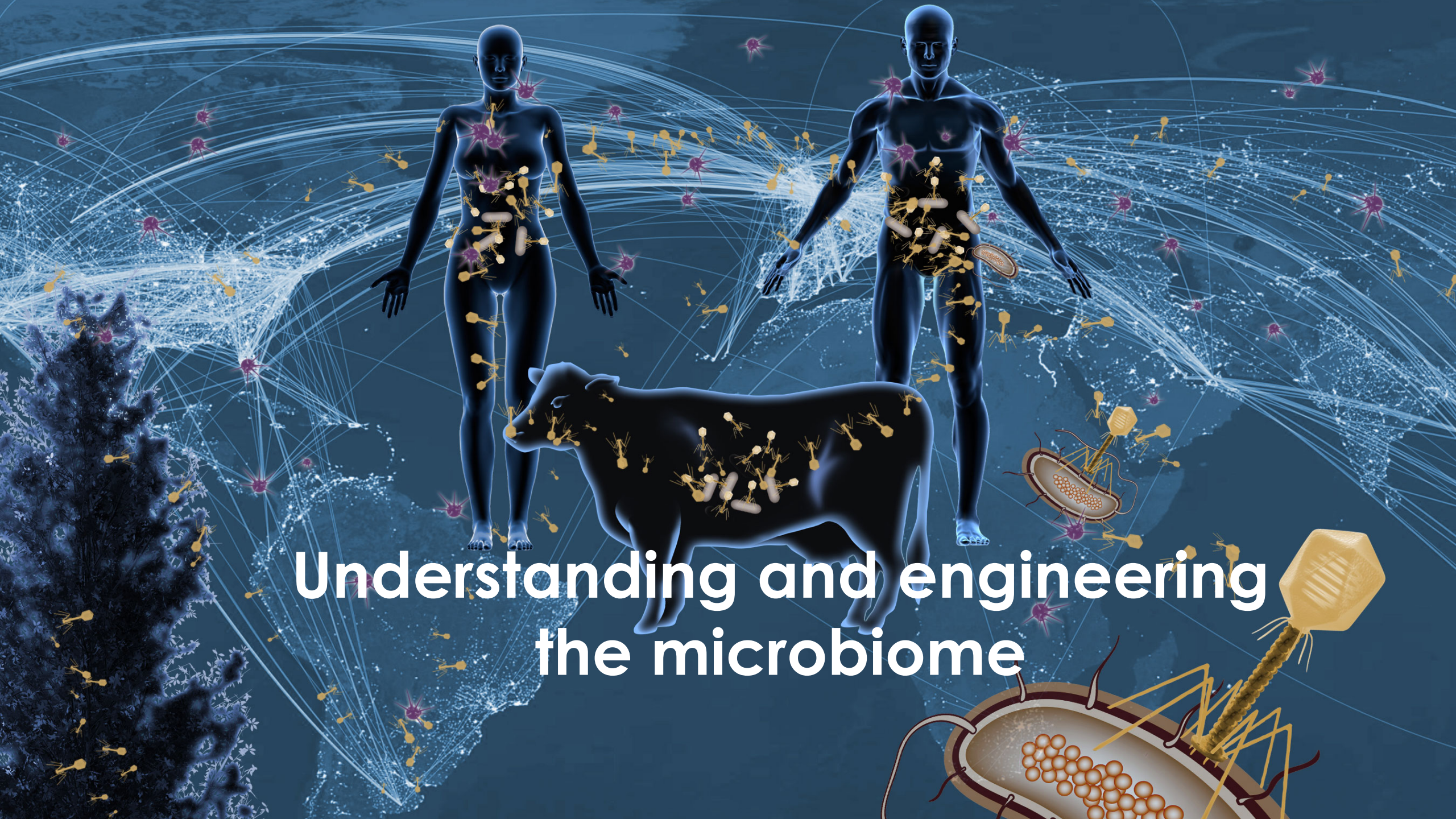
Policy and business model

E.g., see G. Guidi, M. Ellis, A. Buluc, K. Yelick, D. Culler, 2021

Algorithms for ML in Science

A dive into microbial science

Parallelism Motifs



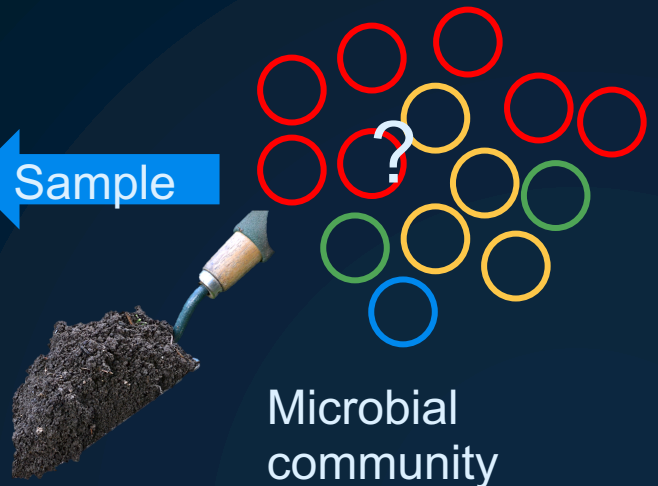
Understanding and engineering the microbiome

ExaBiome: Exascale Microbiome Analysis

Who, what, why, how?

Microbiome analysis: metagenome

Sample

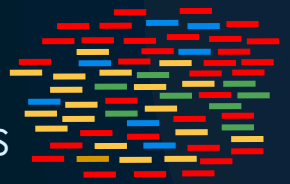


Microbial community

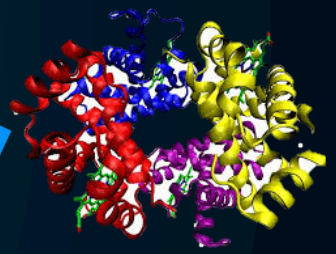
Biological Pathways

Sequence

Reads



Structure & function



Contigs

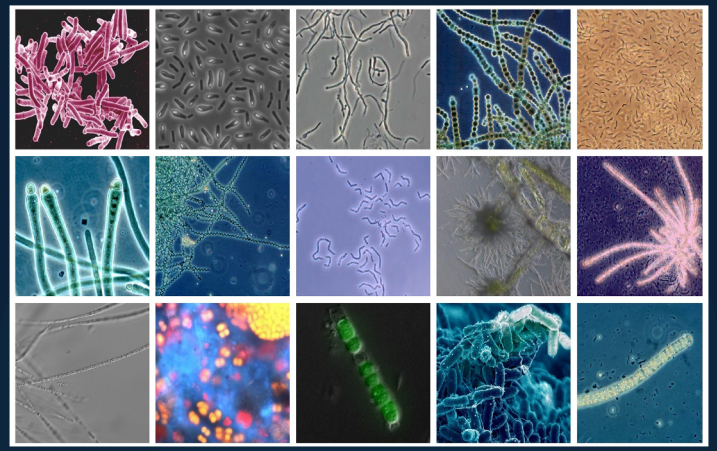
Assemble

Binning

Proteins

Terabyte to Petabyte Metaegnomes

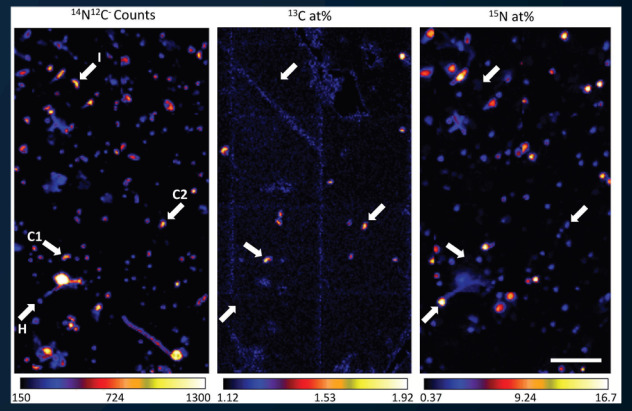
What happens to microbes after a wildfire? (1.5TB)



What are the microbial dynamics of soil carbon cycling? (3.3 TB)

How do microbes affect disease and growth of switchgrass for biofuels (4TB)

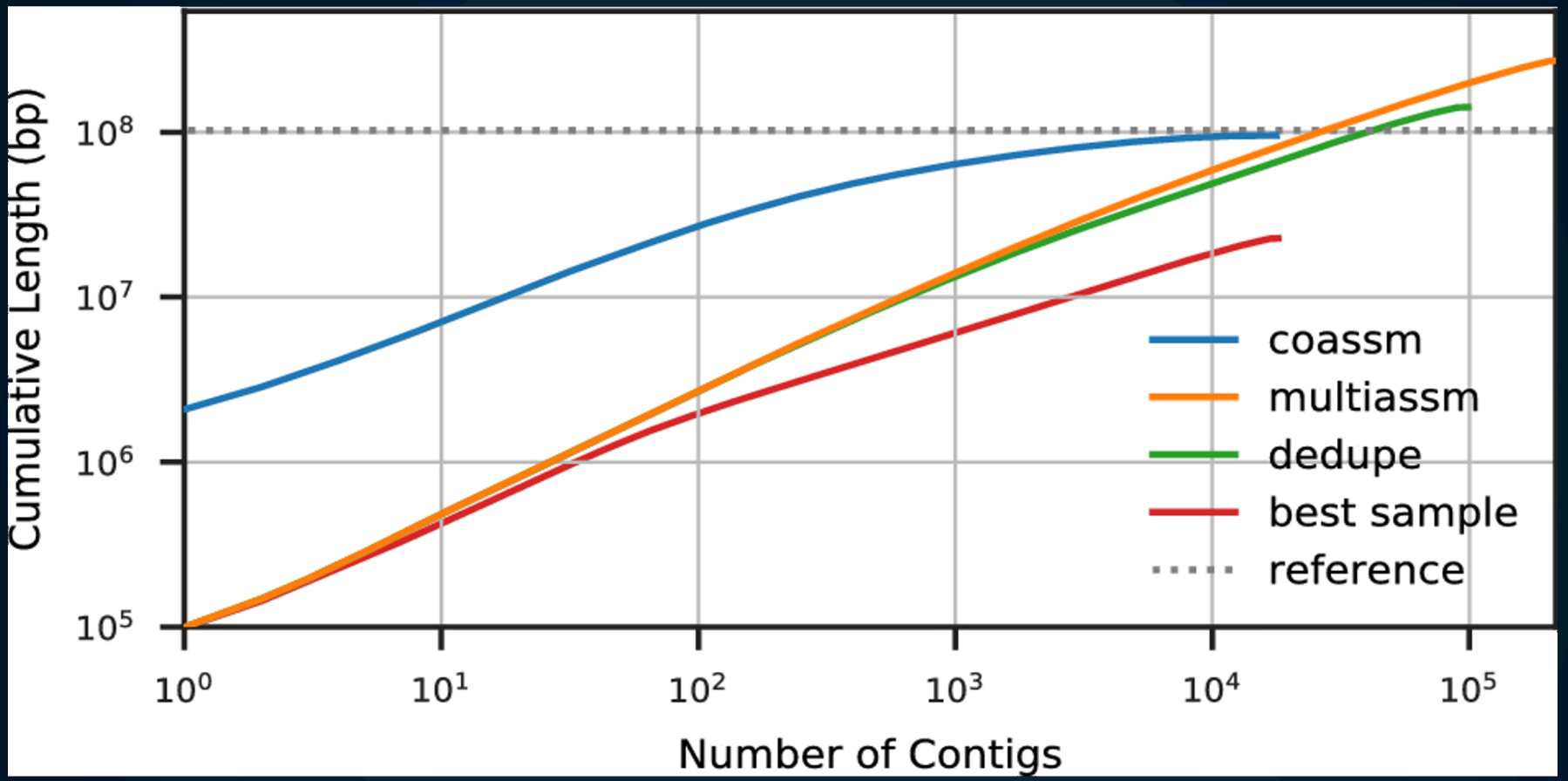
What at the seasonal fluctuations in a wetland mangrove? (1.6 TB)



Combine genomics with isotope tracing methods for improved functional understanding (8TB)

JGI-NERSC-KBase FICUS projects

Big Data, Big Iron → Better Science



Multiassembly: assembling many samples separately

Coassembly: assembling many samples together

S. Hofmeyr, R. Egan, E. Georganas, A. Copeland, R. Riley, A. Clum, E. Eloe-Fadrosh, S. Roux, E. Goltsman, A. Buluç, D. Rokhsar, L. Oliker, K. Yelick, 2020

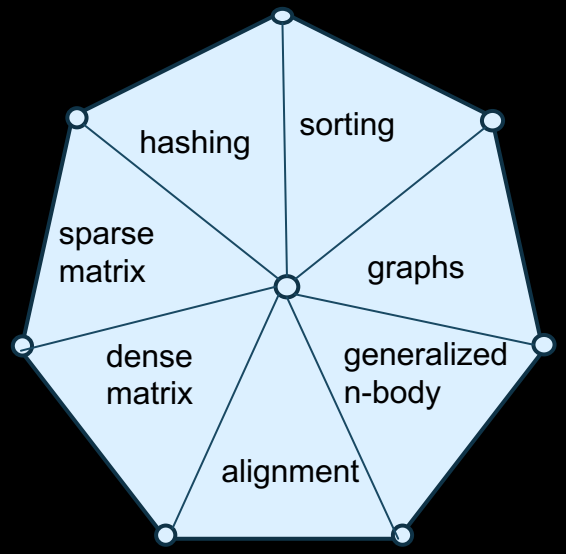
Analytics vs. Simulation Kernels:

7 Dwarfs of Simulation	7 Giants of Big Data
Particle methods	Generalized N-Body
Unstructured meshes	Graph-theory
Dense Linear Algebra	Linear algebra
Sparse Linear Algebra	Sorting
Spectral methods	Hashing
Structured Meshes	Alignment
Monte Carlo methods	Basic Statistics

Phil Colella

NRC Report + our paper

Yelick, et al. "The Parallelism Motifs of Genomic Data Analysis", Philosophical Transactions A, 2020


hashing
sorting
sparse
matrix
graphs
dense
matrix
generalized
n-body
alignment

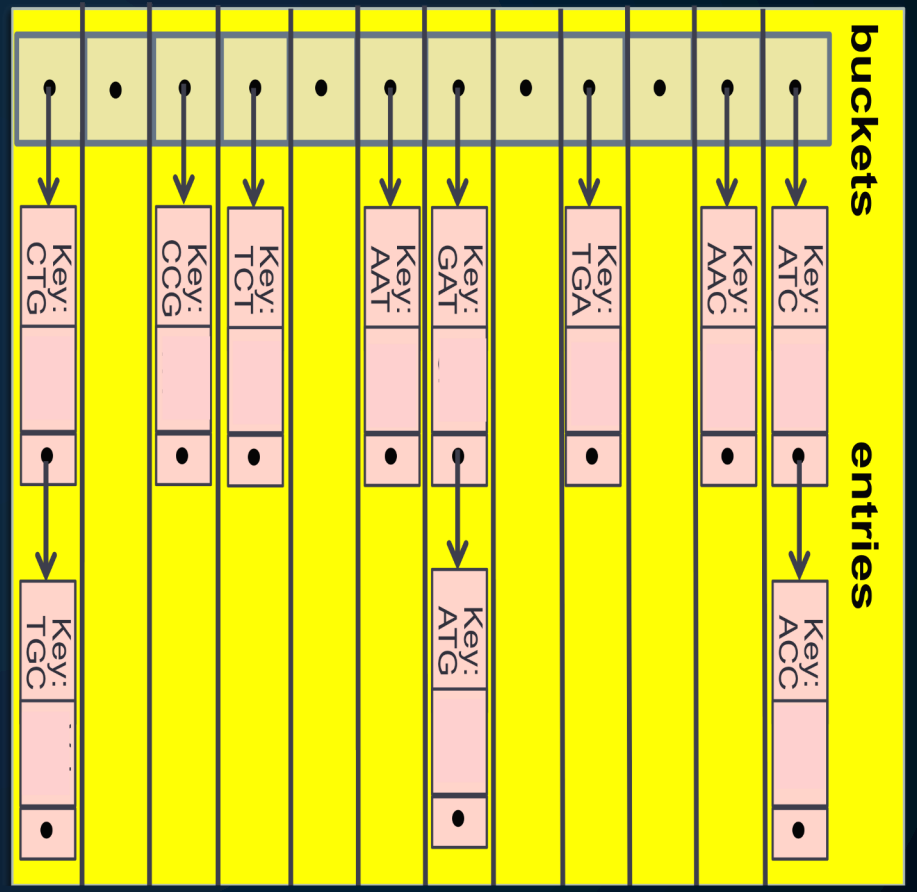
Hashing

Universally useful: Hash Tables of K-Mers

Make hash table of k-mers

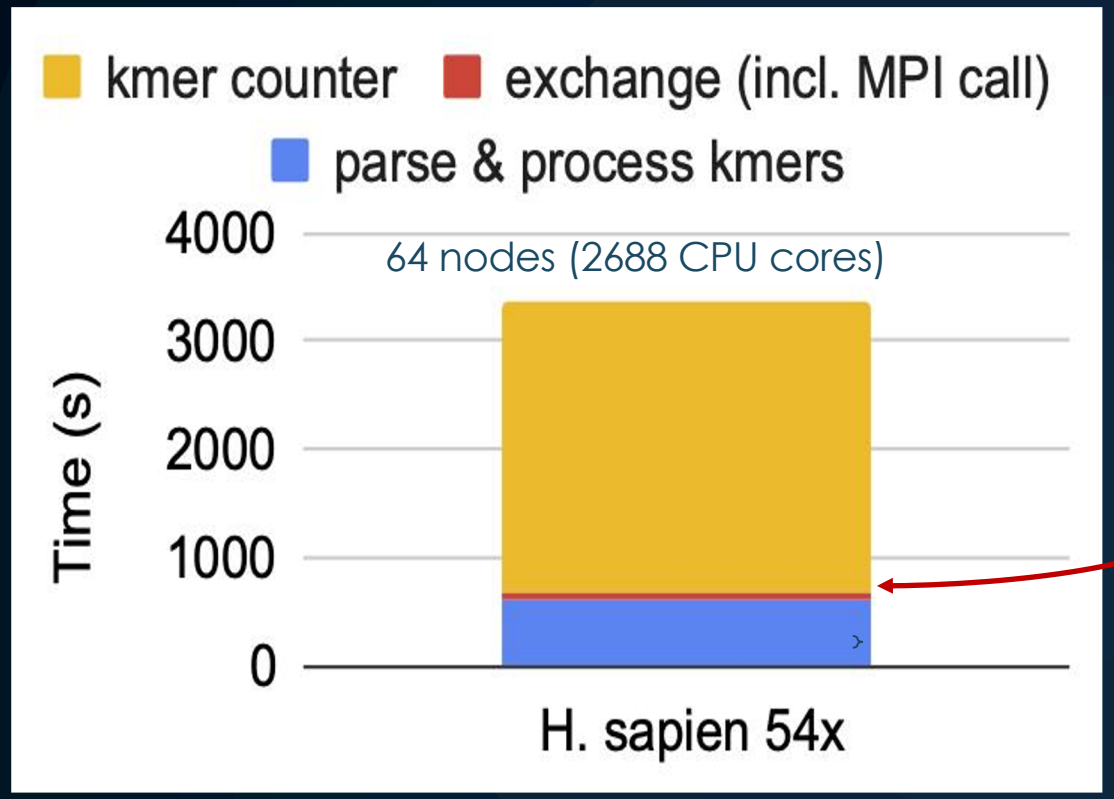
```
AAC        TGA        CCG
 ACC        GAT        CGT
  CCT        ATT        GTC
```

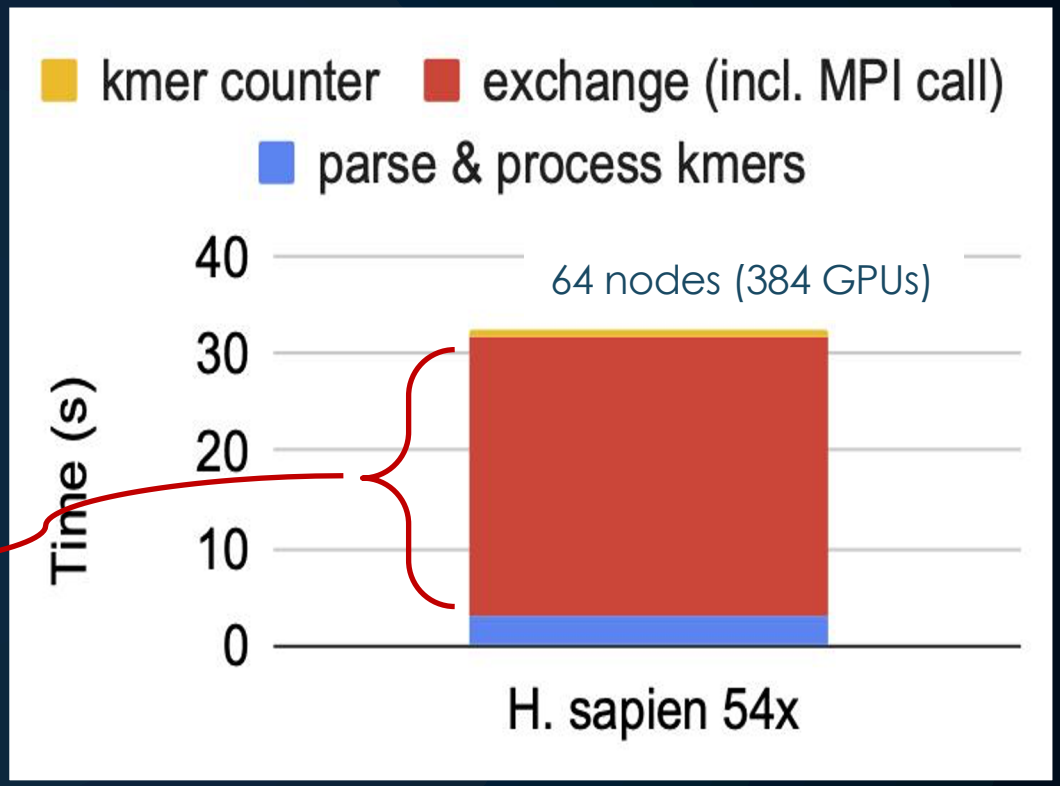
1-sided comm or irregular
all-to-all + memory



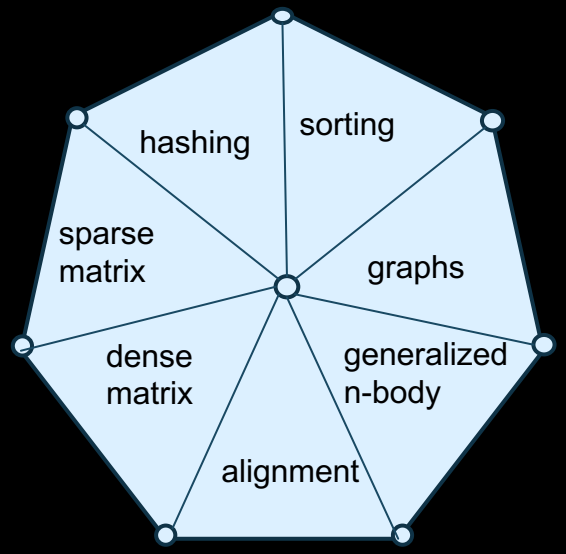
K-mer Counting on GPU HPC

Summit (with NVIDIA V100)





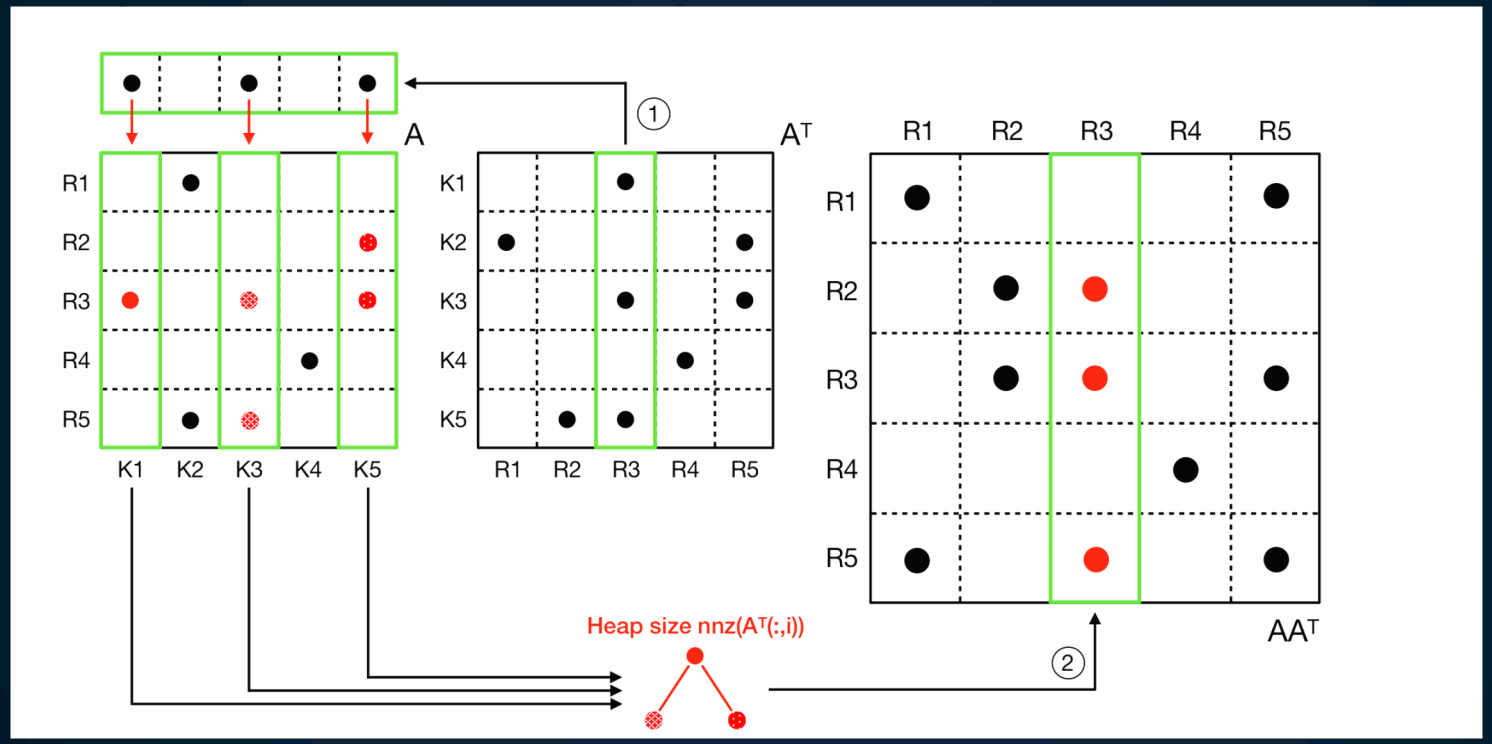
- Now communication-bound
- Use clever hashing (minimizers) and aggregation (supermers) reduce number of messages (latency) and volume (bandwidth)



Generalized N-Body

Set Alignment is a Sparse All-to-All

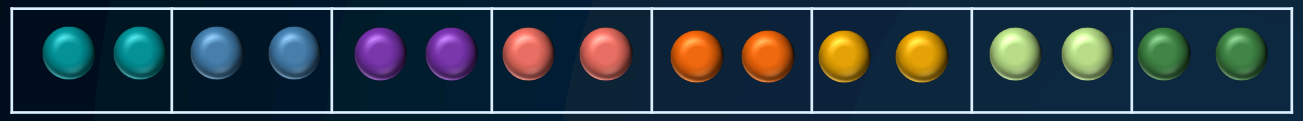
Run expensive alignment on all pairs with a common k-mer



Avoid Communication, Maximize Parallelism

Compute on all pairs of particles or strings, or…

Obvious solution

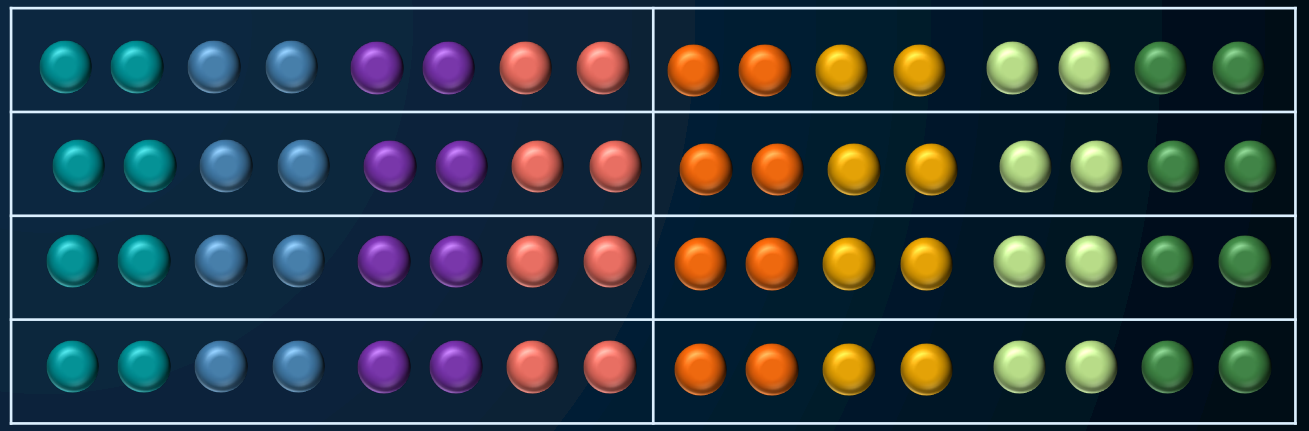


16 particles on 8 processors
Pass all particles around (p steps)

Decreases

- #messages by factor $\mathbf{c^2}$,
- #volume sent by factor **c**

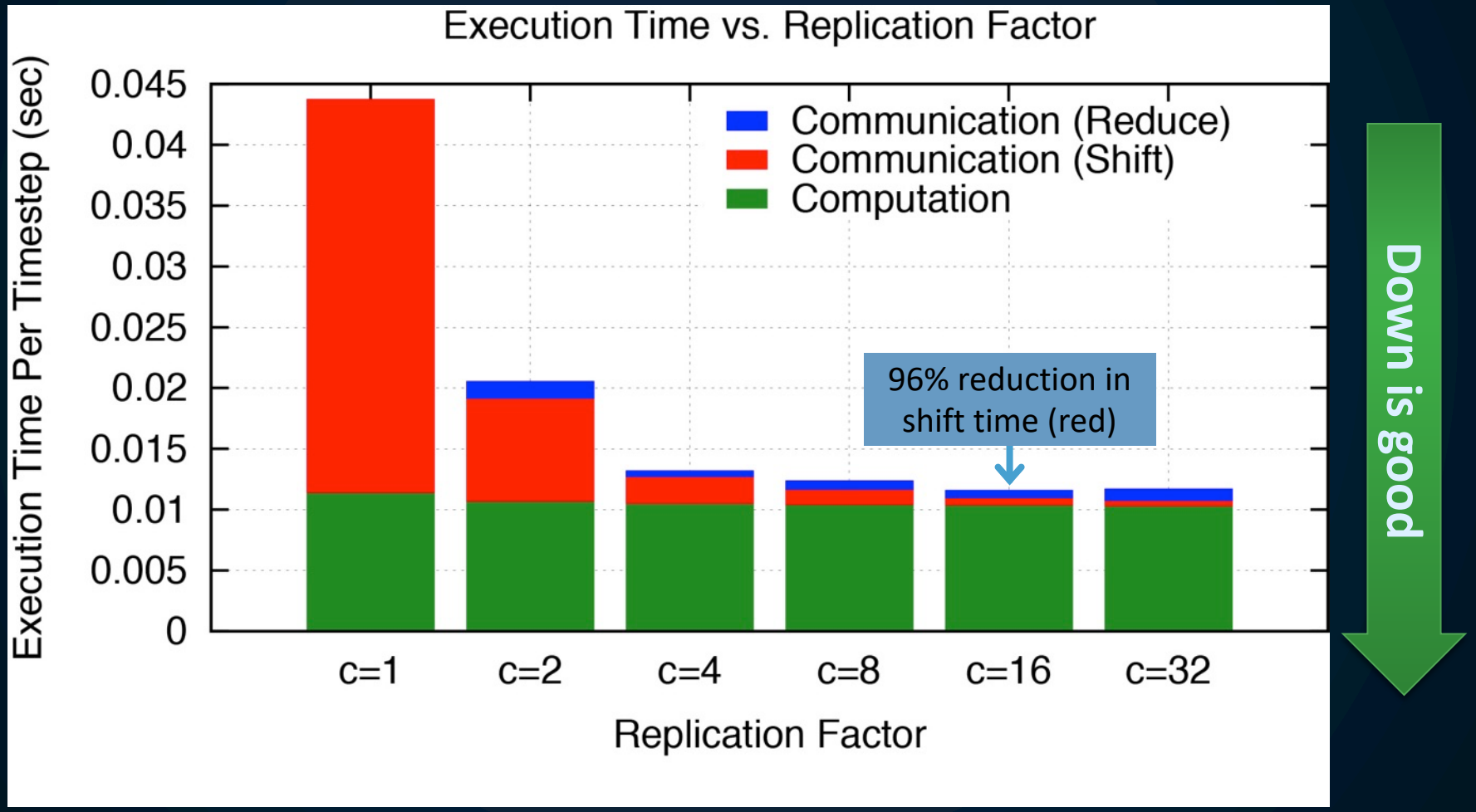
Better solution



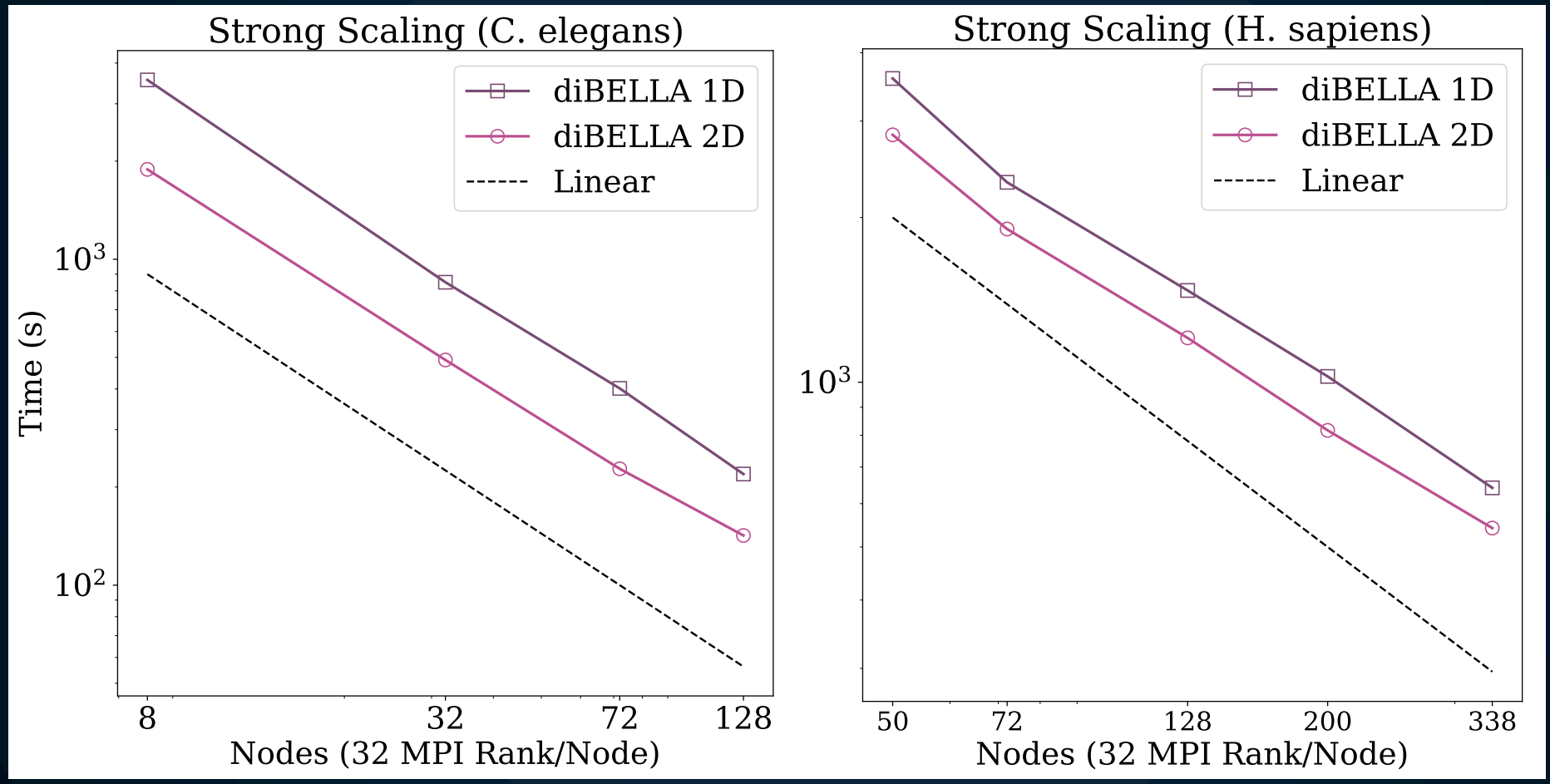
c = 4 copies of particles
8 particles on each

Less Communication..

Cray XE6; n=24K particles, p=6K cores

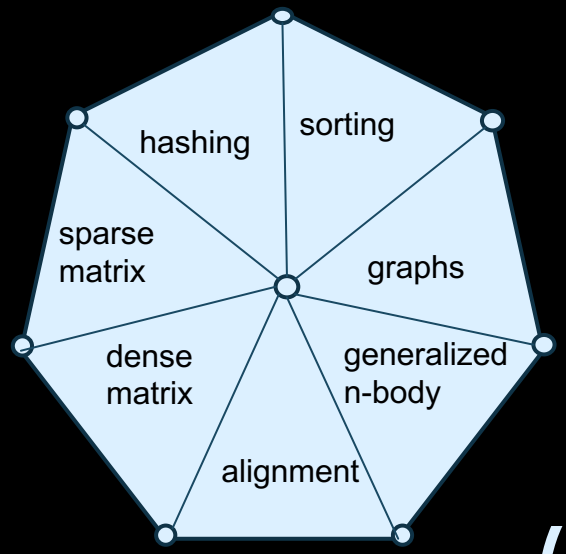


1D vs 2D Algorithm on DNA “overlap”



Lower is Better

G. Guidi, O. Selvitopi†, M. Ellis, L. Oliker, Y, A. Buluc

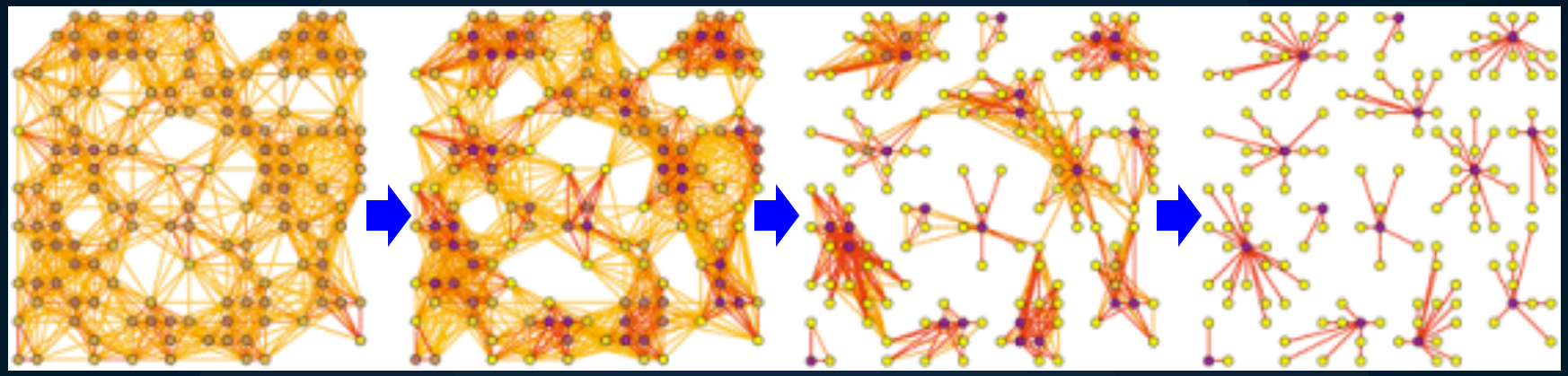


Graphs and Sparse Matrices (unsupervised learning)

Protein Clustering with Sparse Matrices

Input: Adjacency matrix A (sparse)

Image source: http://micans.org/mcl/



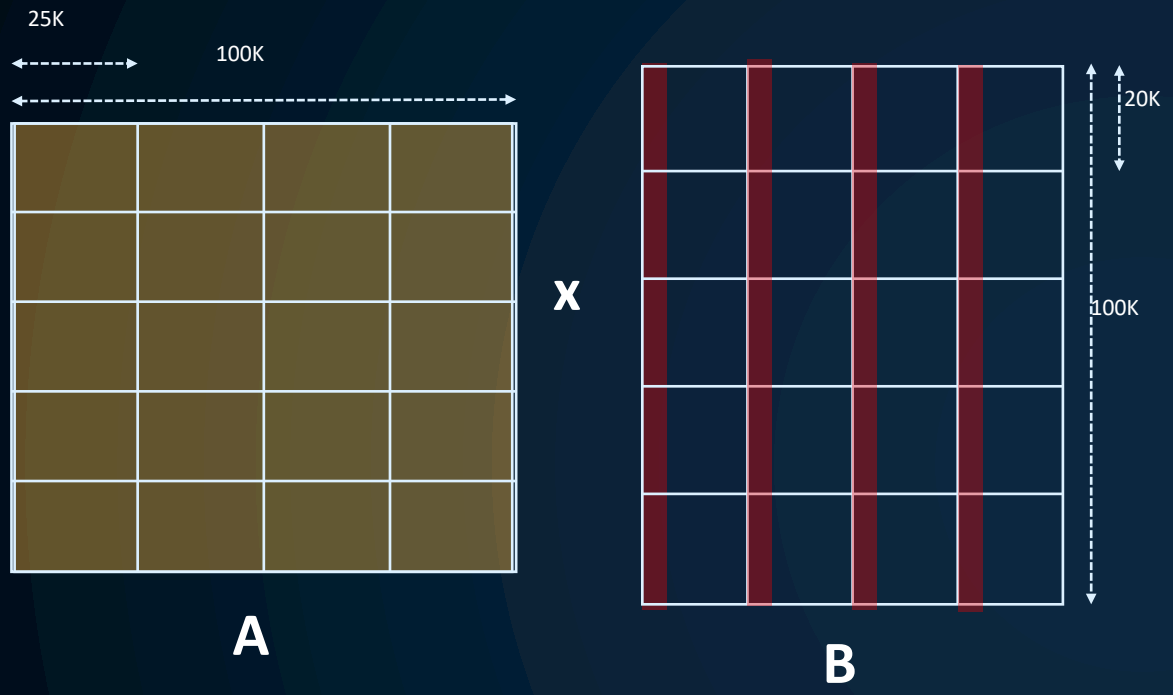
Initial network | Iteration 1 | Iteration 2 | Iteration 3

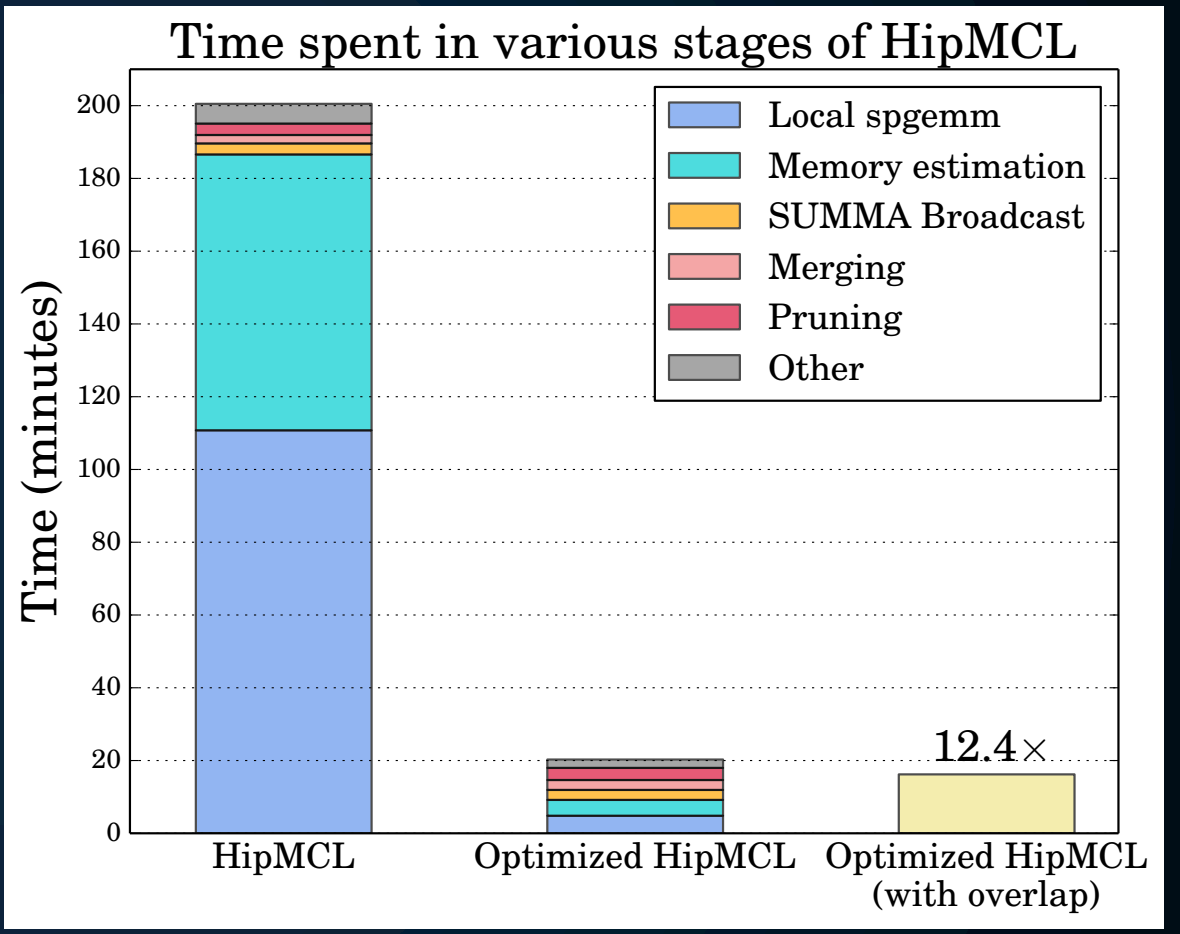
- **Similarity Matrix**: “Many-to-many” protein alignment
- **Expansion**: Square matrix, pruning small entries, dense columns
- **Inflation**: element-wise powers

PASTIS + HipMCL

Oguz Selvitopi; Md Taufique Hussain; Ariful Azad; Aydın Buluç

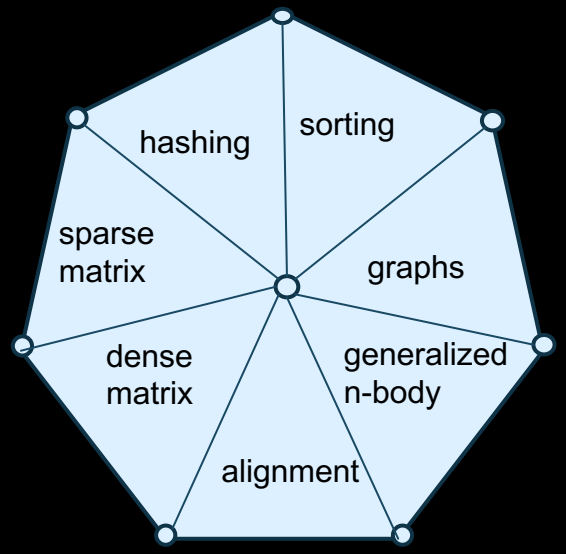
Sparse Matrix Algorithms





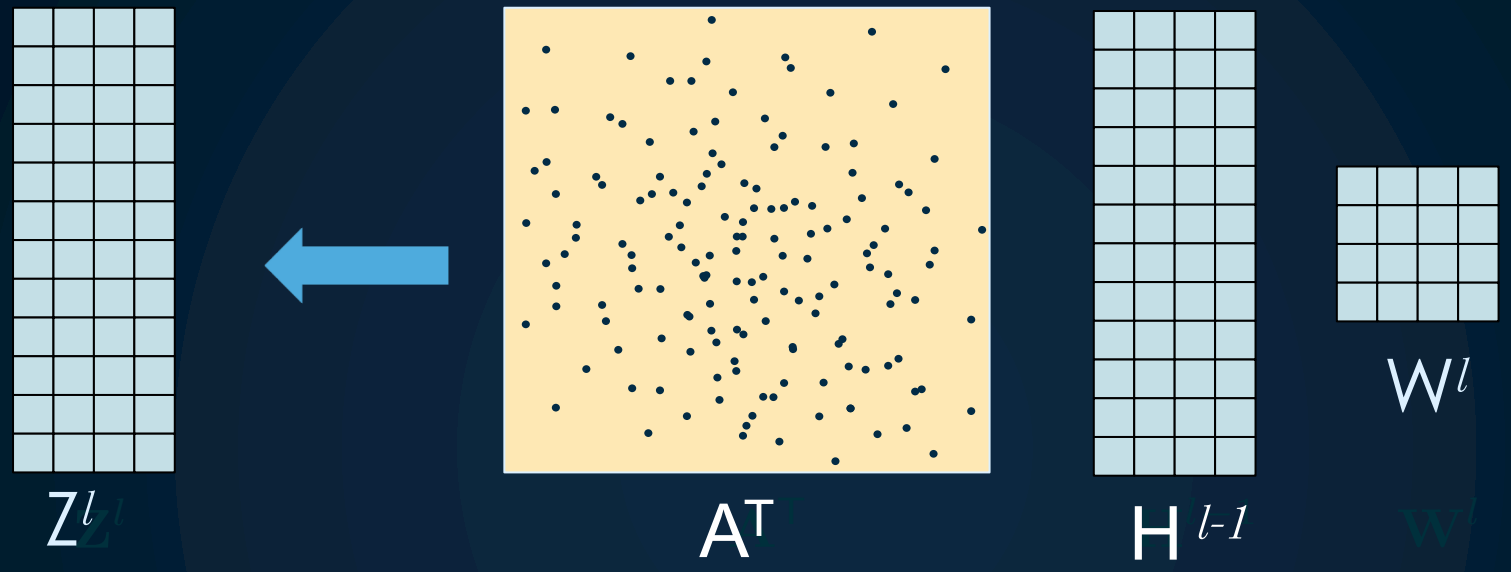
Distributed memory enabled first-of-kind science

12.4× faster with GPUs



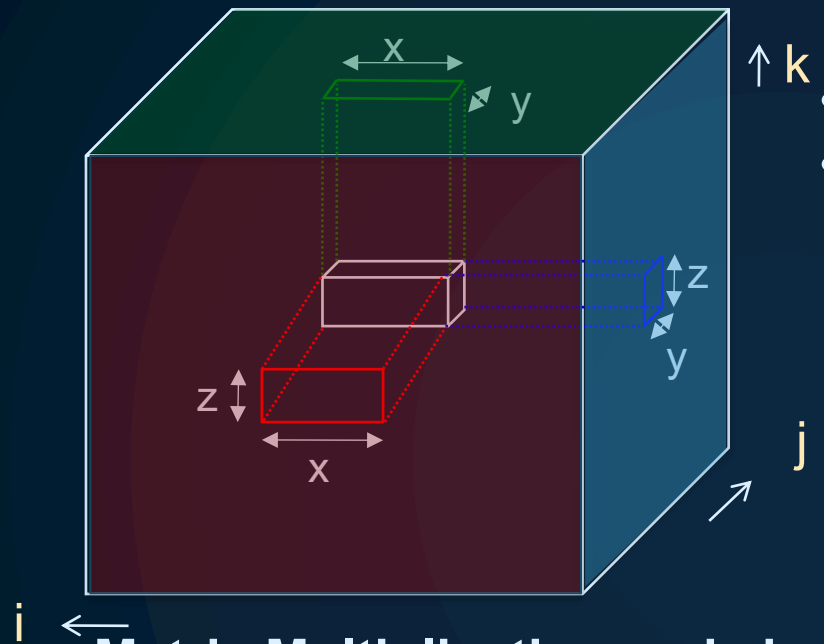
Graphs and Sparse and Dense Matrices (supervised learning)

Bottleneck in GNN Training



- $A^T H^{l-1}$ sparse-dense matmul (SpMM)
- $(A^T H^{l-1})\ W^l$ dense-dense matmul (DGEMM)
- **SpMM is the bottleneck, not DGEMM!**

Communication-Avoiding Matrix Multiply

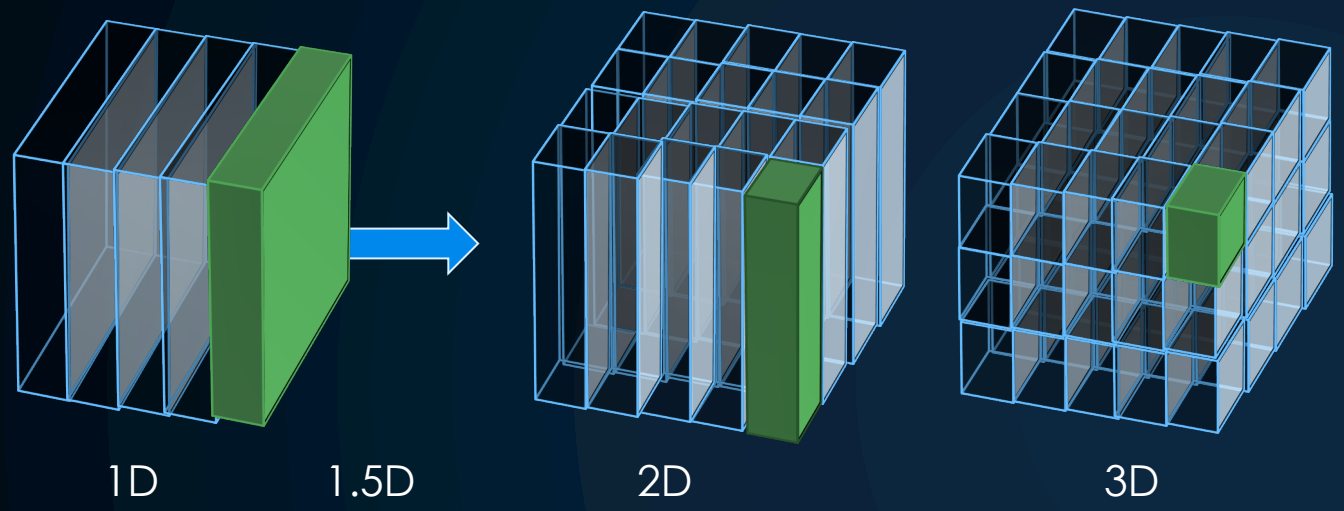


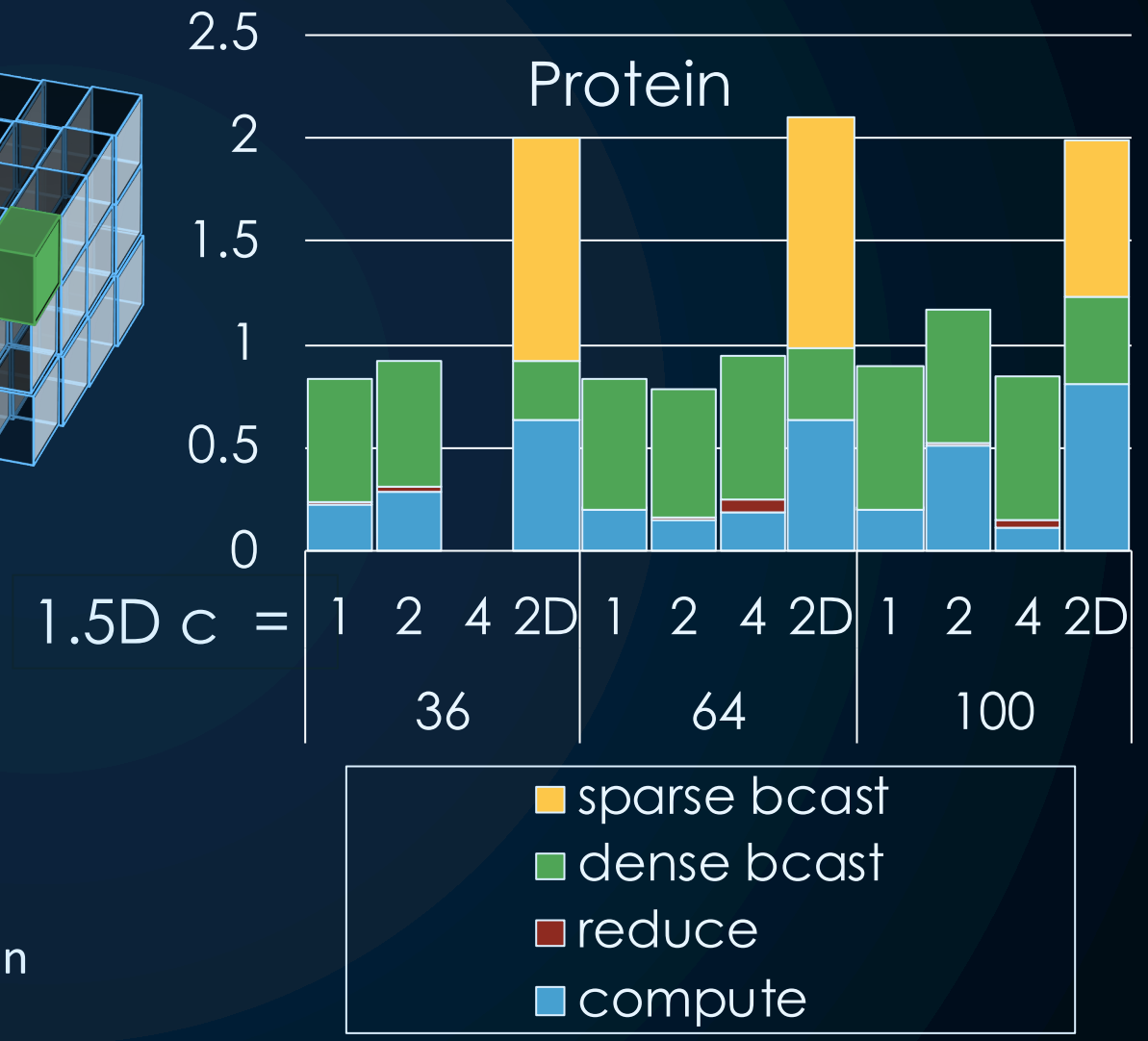
- 2D algorithm: never chop k dim
- 3D: Assume + is associative; chop k, which is → replication of C matrix

Matrix Multiplication code has a 3D iteration space
Each point in the space is a constant computation (*/+)

```
for i
  for j
    for k
          C[i,j] …  A[i,k] …   B[k,j]  …
```

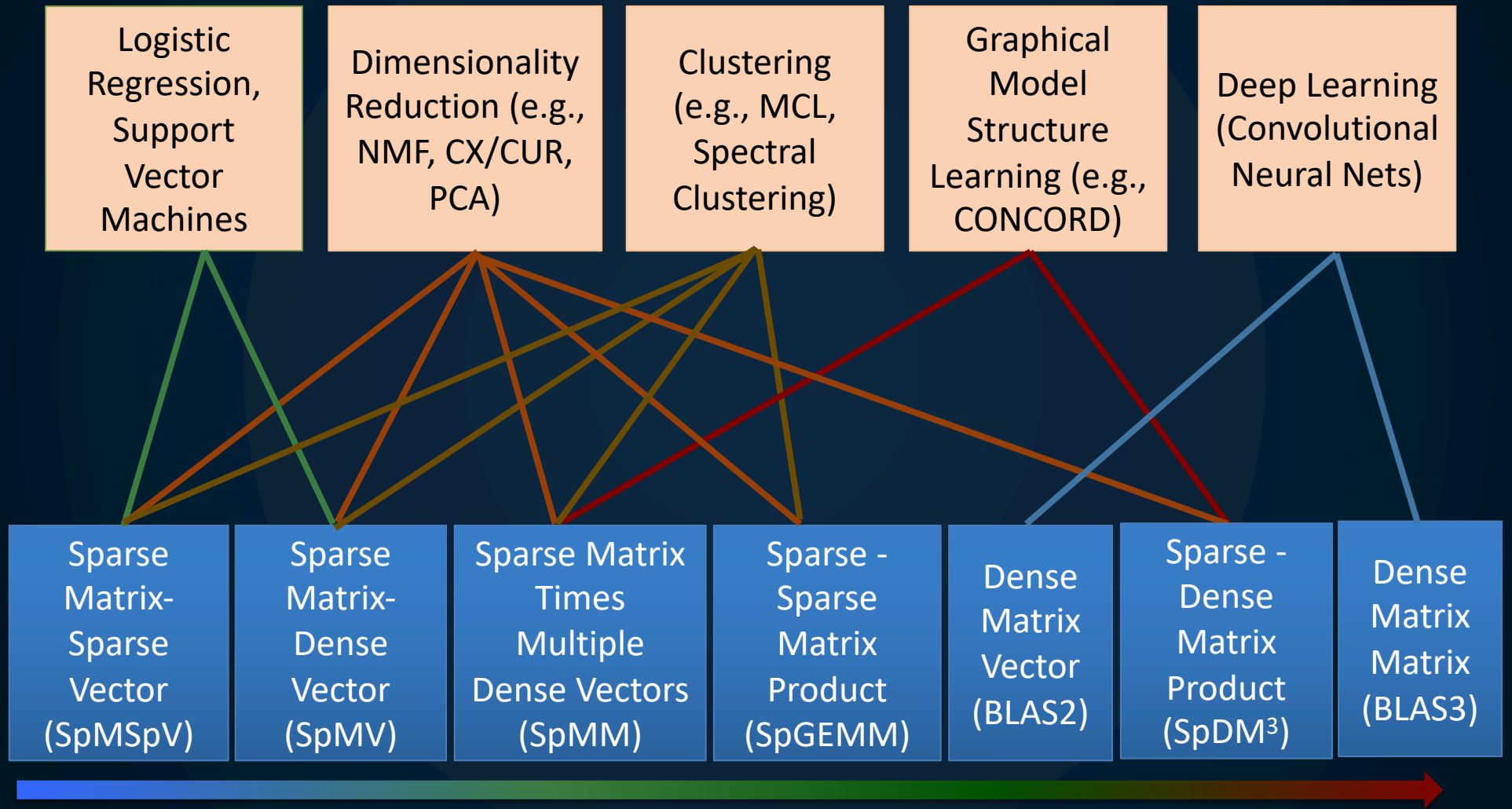
Avoiding Communication in GNNs





Tripathy, Yelick, Buluc, Reducing Communication in Graph Neural Network Training, SC'20

Machine Learning Mapping to Linear Algebra



Aydin Buluc, Sang Oh, John Gilbert, Kathy Yelick

Take-Aways

- Applications
  - Every domain of science
  - Analysis, Acceleration, Automation
  - Science emphasizes quantifying uncertainties, interpretability, etc.
- Architectures
  - Specialization will be increasingly important
  - Communication will dominate; Need better integration, lower overheads
  - Cloud and HPC differ on the business model
- Algorithms
  - Irregular, fine-grained problems in data and ML, not just simulation
  - Avoid communication to match hardware