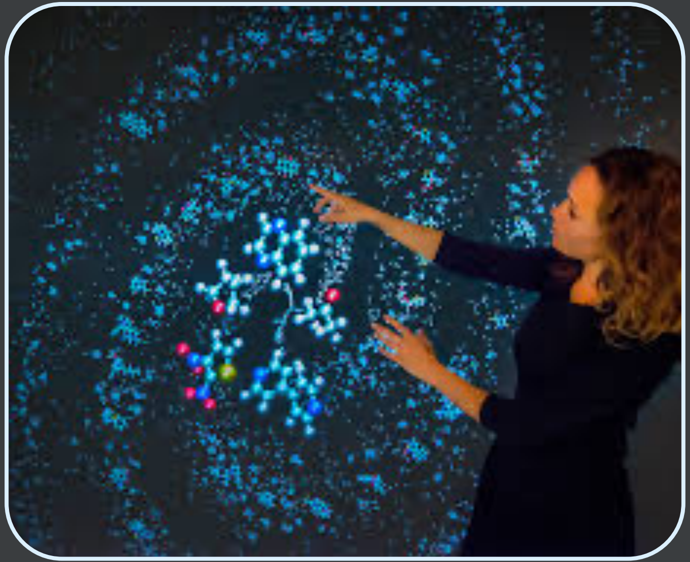
Automating Science: Applications, Algorithms, and Architectures

Kathy Yelick
Vice Chancellor for Research
Robert S. Pepper Distinguished Professor of Electrical Engineering and Computer Sciences
University of California, Berkeley

Senior Faculty Scientist, Lawrence Berkeley National Laboratory

Opportunities in Science

Analyze



Explore

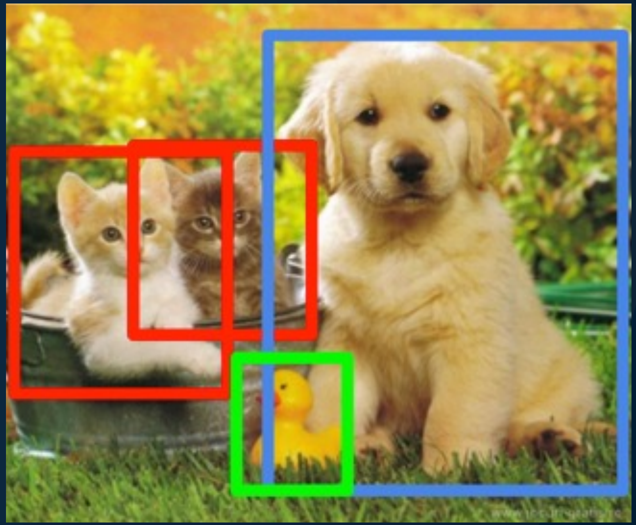
Automate

Analyze Images to Find Cats

Classification

Localization

Detection

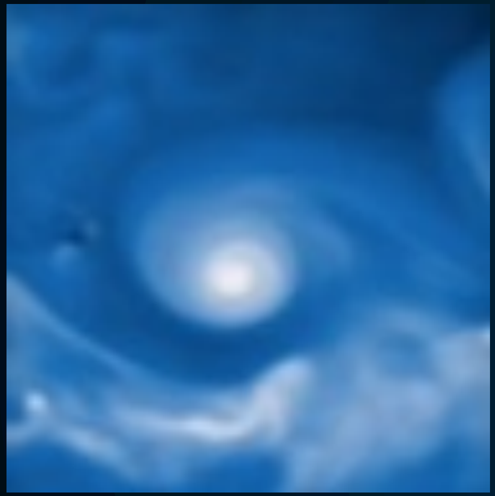


Segmentation

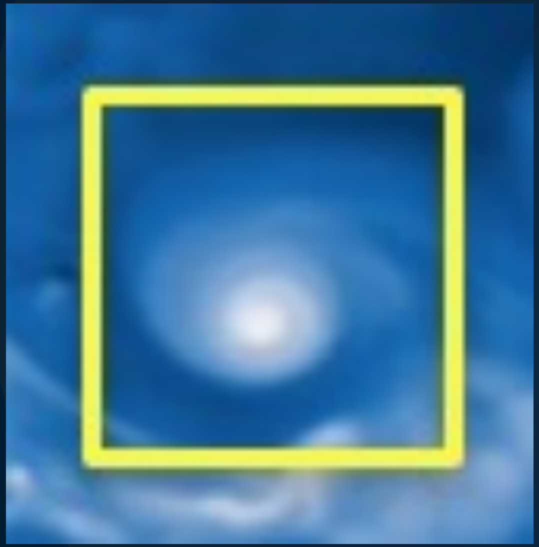
Source: Prabhat

Analyze Simulations to Find Hurricanes

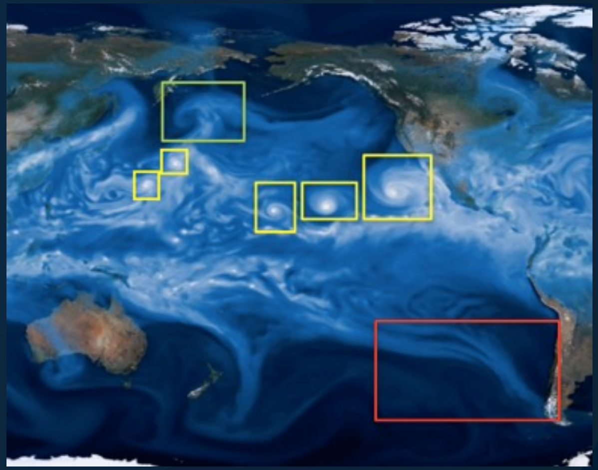
Classification



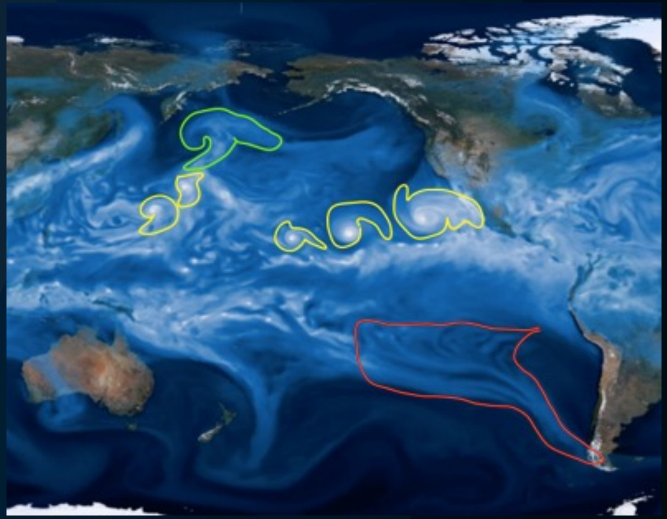
Localization



Detection



Segmentation

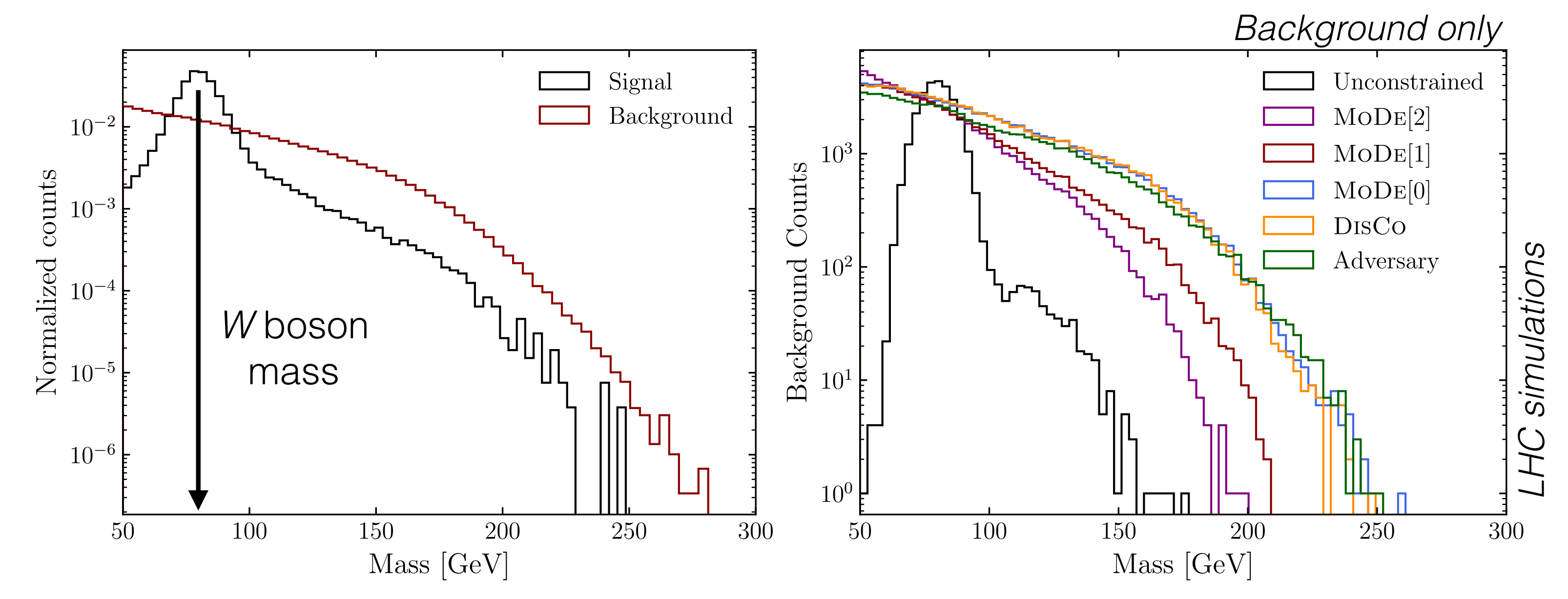


Extending image-based methods to complex, 3D, scientific data sets is non-trivial!

Source: Prabhat

Fairness in Physics

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

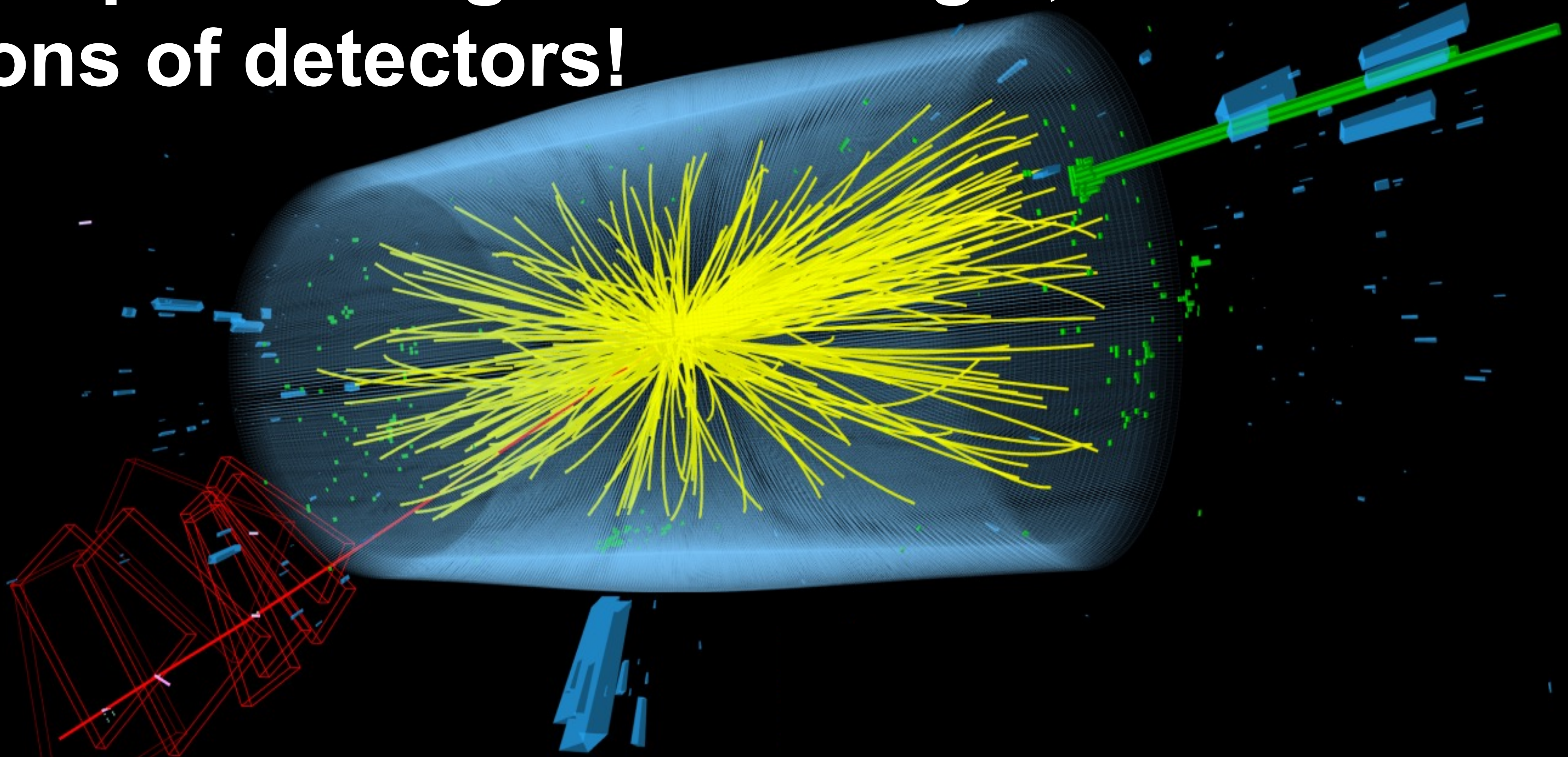


Signal and background events without selection.

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

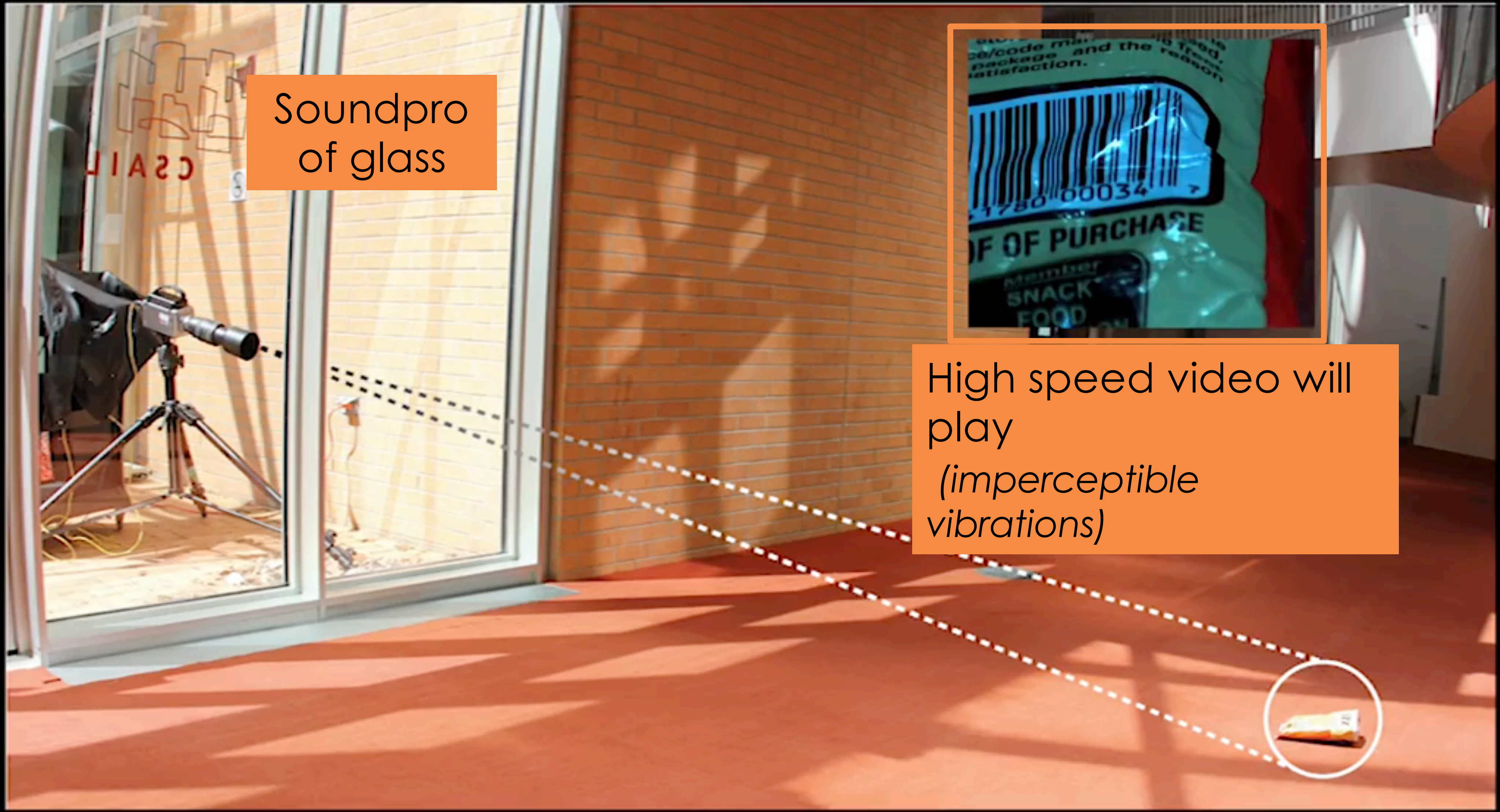
O. Kitouni, B. Nachman, C. Weisser, M. Williams, 2010.09745

Deep Learning: like adding 4,000 extra tons of detectors!

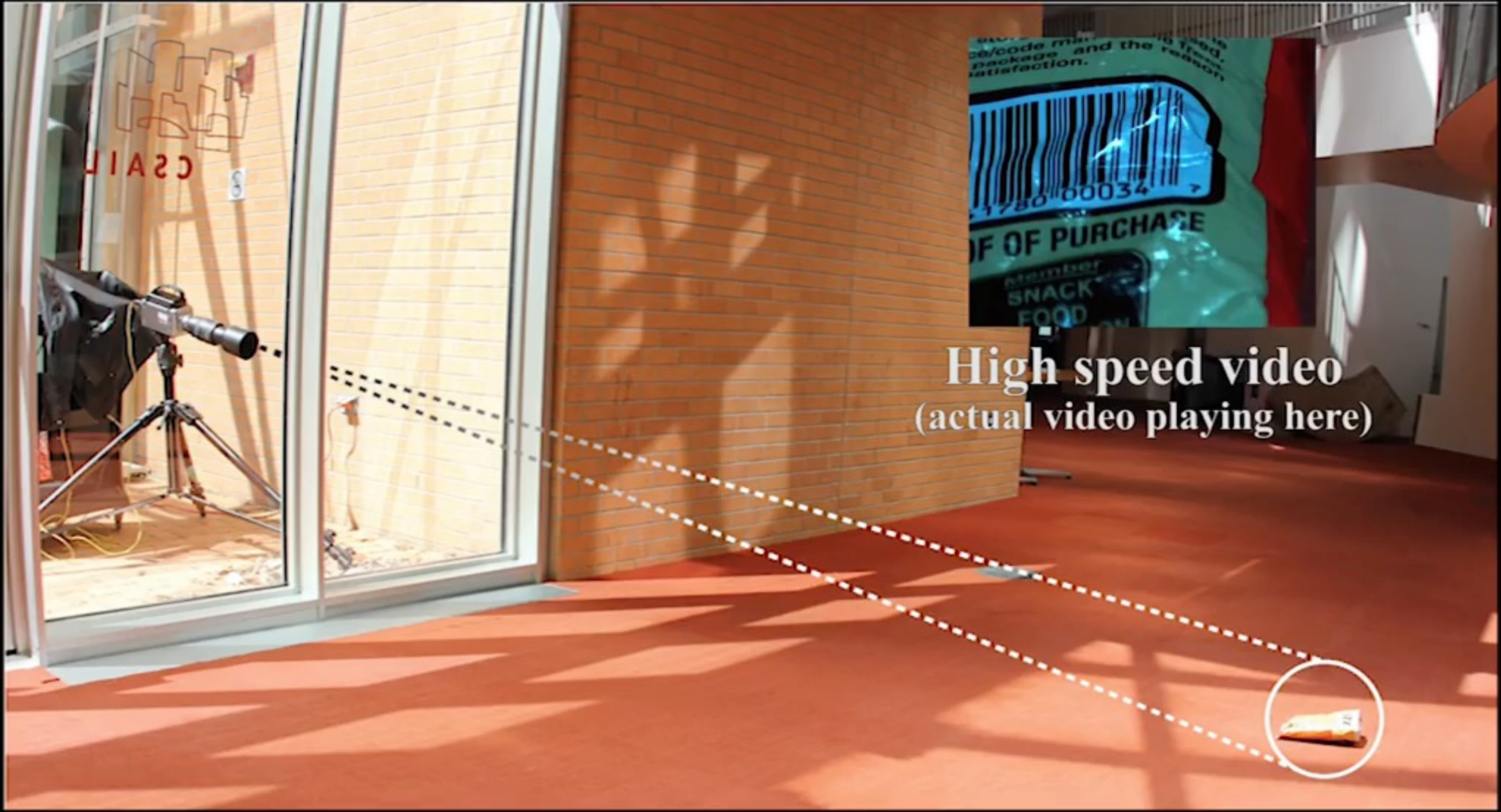


Based on 8/12/2016 slide by Joe Lykken at Fermilab

Extracting signals from noisy data: "Visual Microphone"



Abe Davis, M Rubinstein, N Wadhwa, GJ Mysore, F Durand, WT Freeman, MIT


High speed video
(actual video playing here)

First Image of a Black Hole

This is not replicating human vision

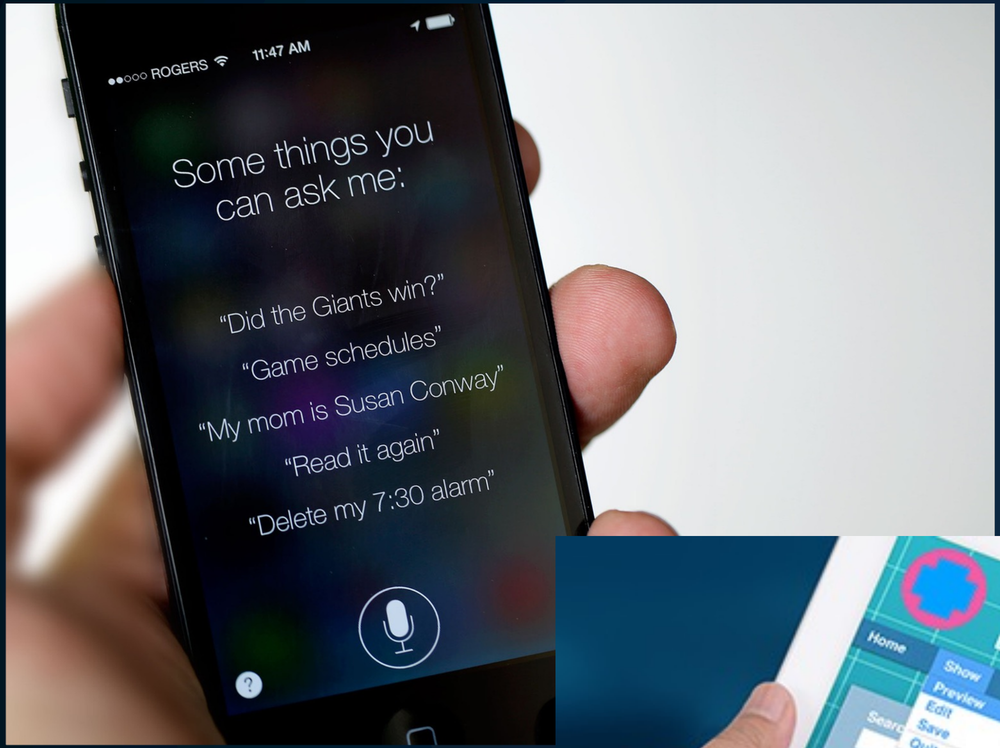
Filtering, De-Noise and Curating Data

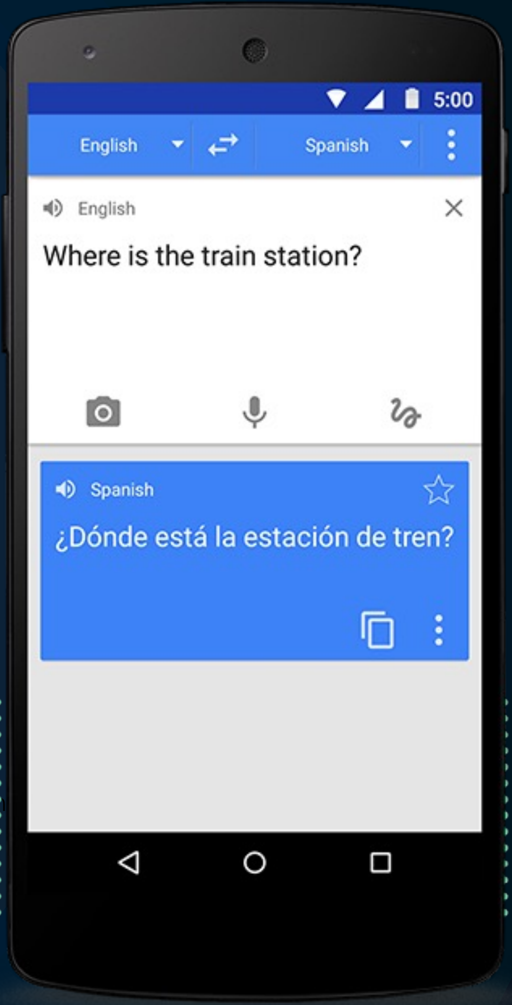
AmeriFlux & FLUXNET: 750 users access carbon sensor data from 960 carbon flux data years; Developing ML to denoise data.

Arno Penzias and Robert Wilson discover Cosmic Microwave Background in 1965

Gilberto Z. Pastorello, Dario Papale, Housen Chu, Carlo Trotta, Deb A. Agarwal, Eleonora Canfora, Dennis D. Baldocchi, M. S. Torn

AI for Natural Language Processing (NLP)

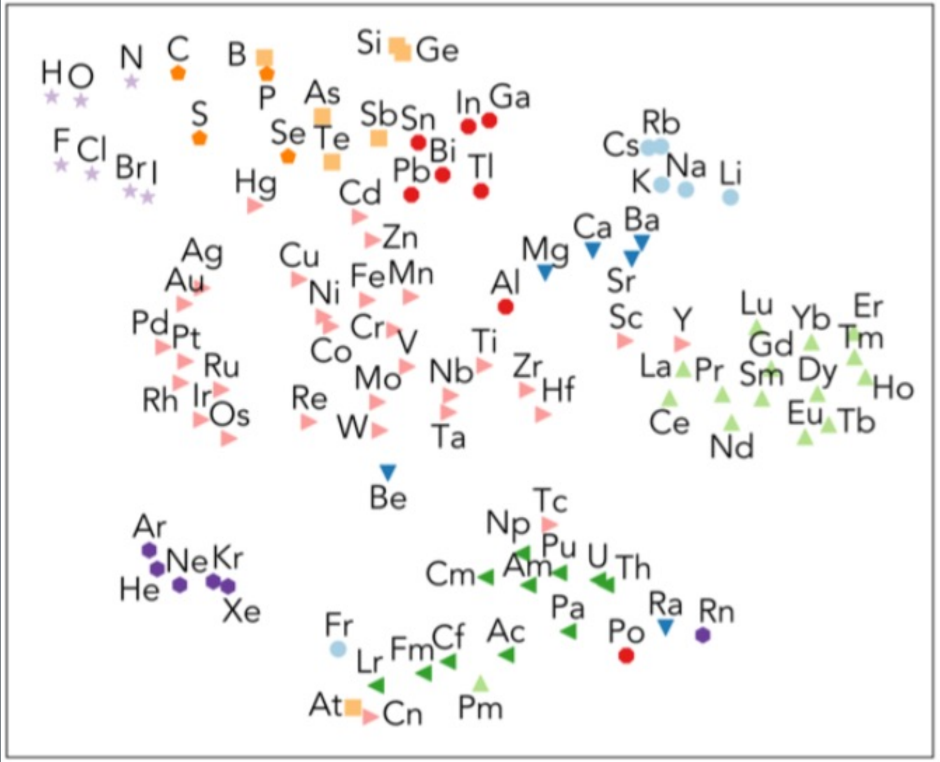


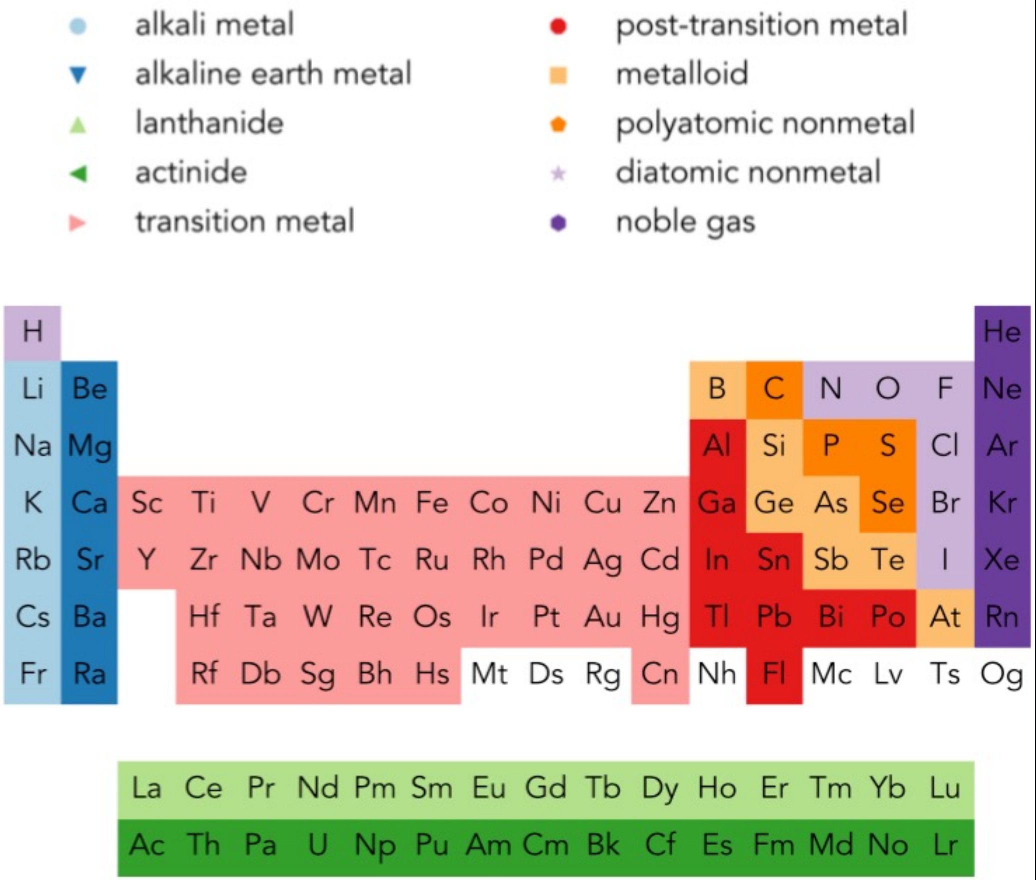


Slide source: Steve Farrell

Using NLP on scientific publications

Analyze 3.3 million abstracts from materials science papers

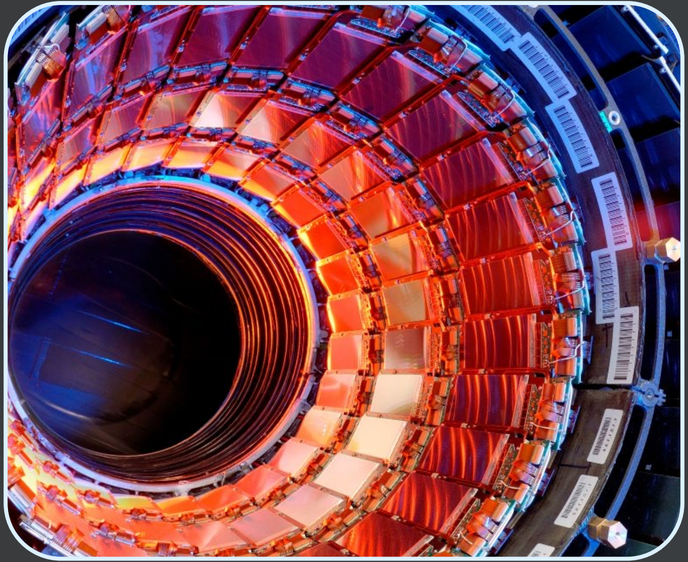




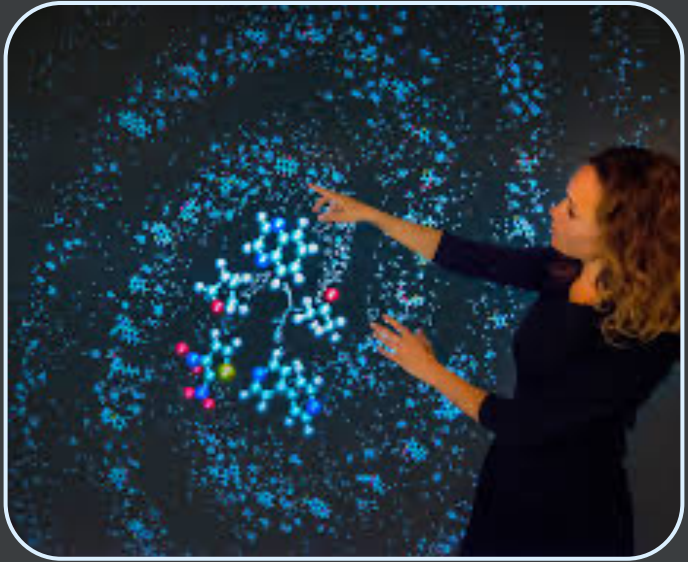
Word2vec's representation of the elements, projected onto two dimensions

Vahe Tshitoyan, Leigh Weston, John Dagdelen, Anubhav Jain

Opportunities in Science



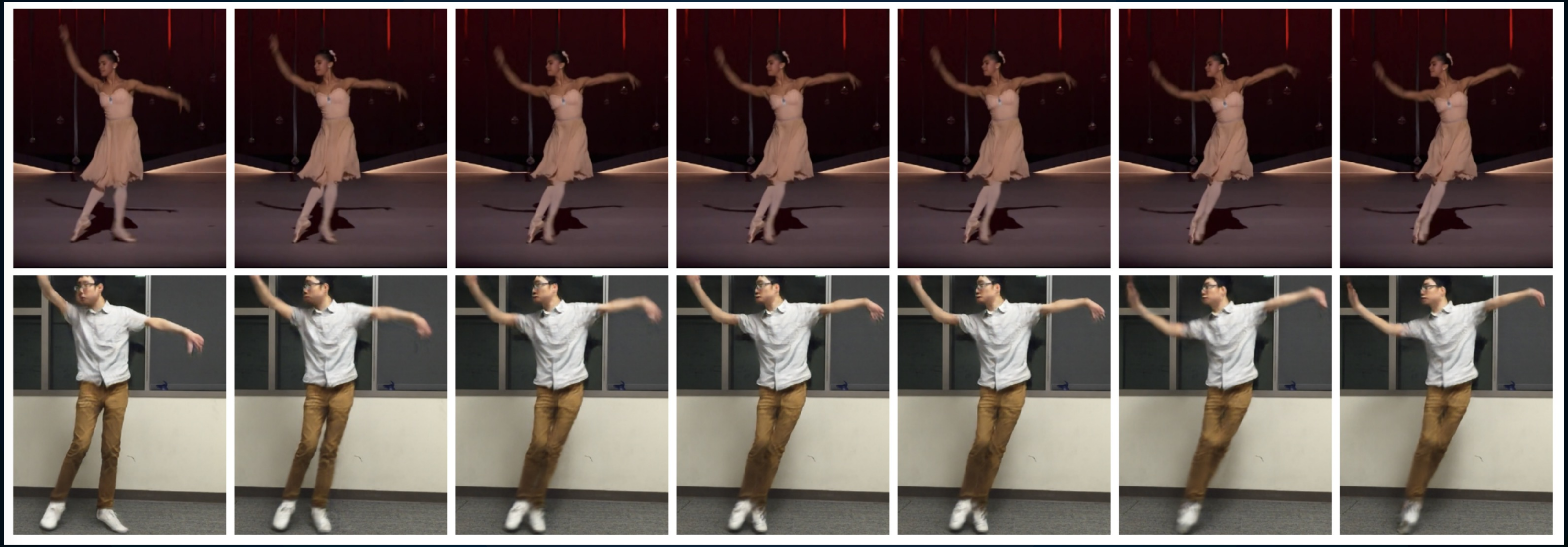
Analyze



Explore

Automate

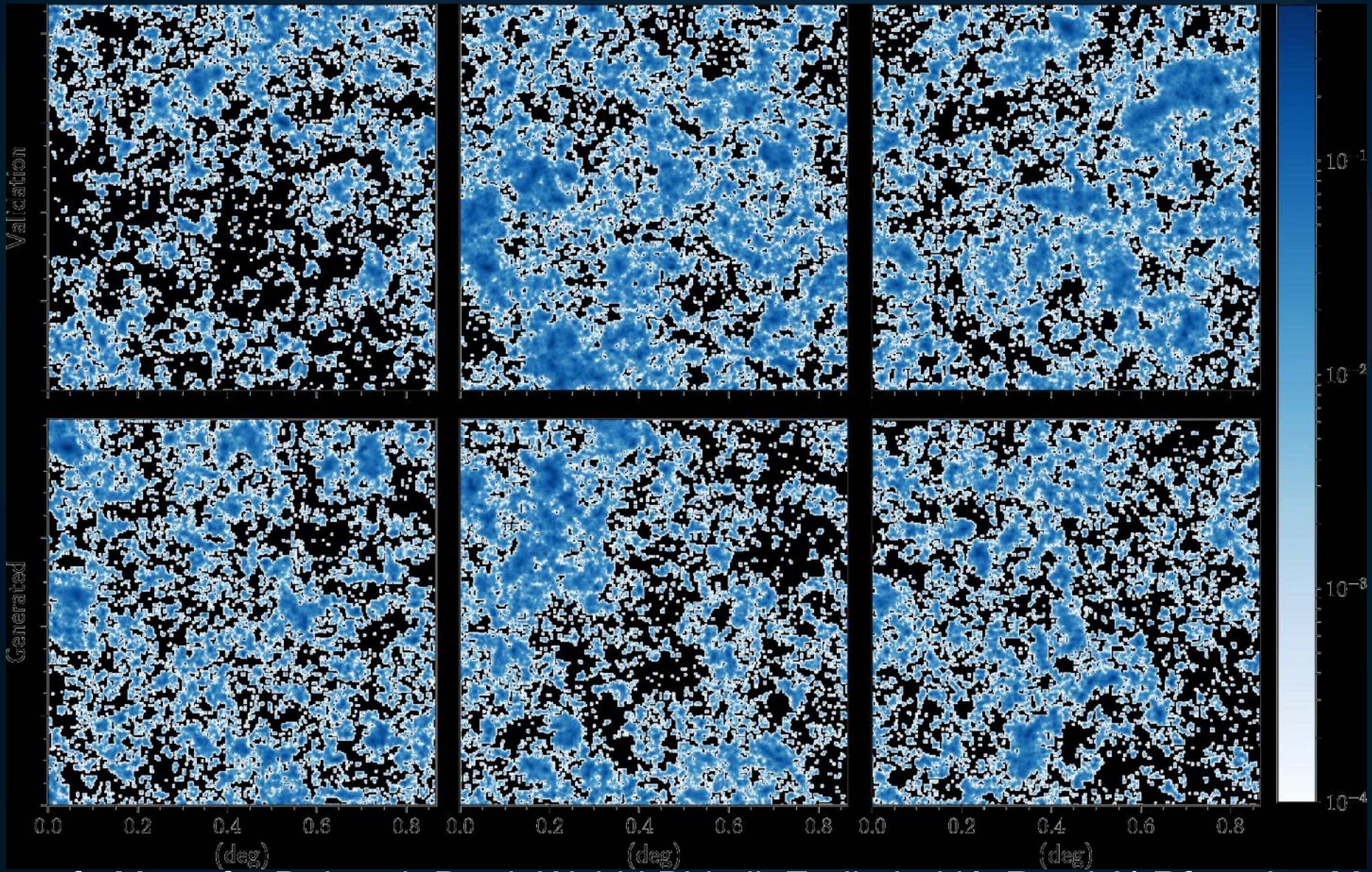
Generate Videos



Caroline Chan, Shiry Ginosar, Tinghui Zhou, Alexei A. Efros, UC Berkeley

Generate Data from Expensive Experiments

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

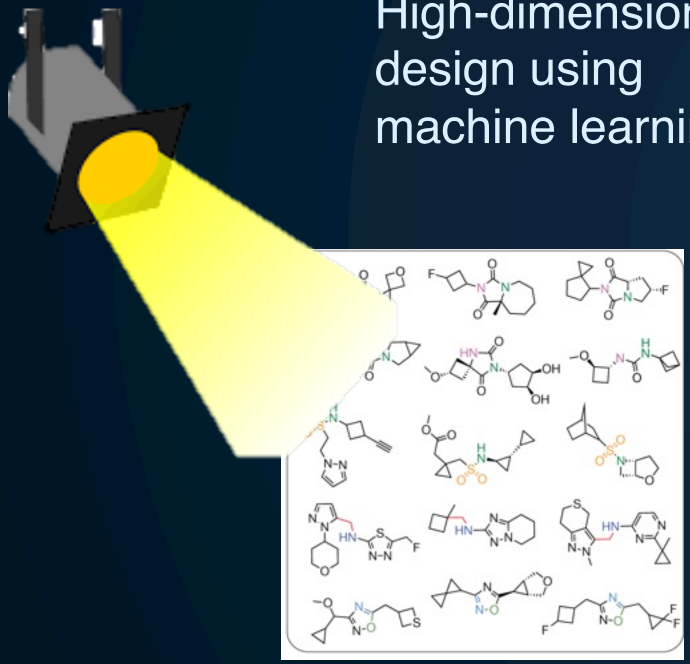


CosmoGAN: Mustafa Mustafa, Deborah Bard, Wahid Bhimji, Zarija Lukić, Rami Al-Rfou, Jan M. Kratochvil

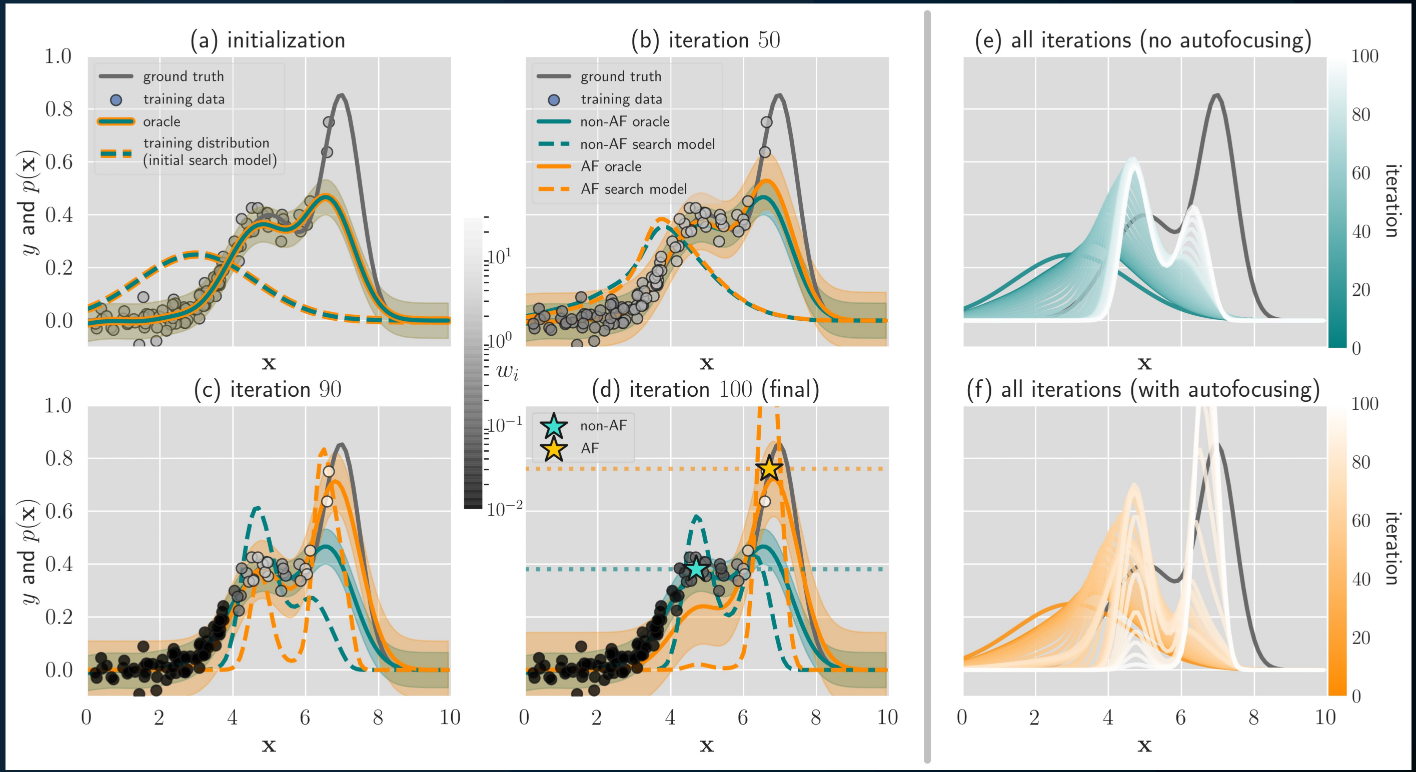
Inverse Design with ML

Designing materials, proteins, and small molecules with ML

High-dimensional design using machine learning



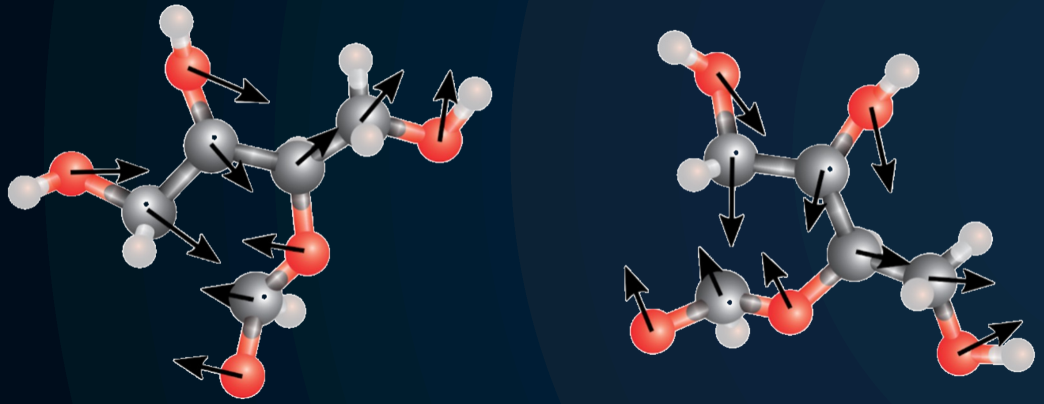
Search for a molecules using an autofocusing generative model: moves around the design space, guided by an oracle

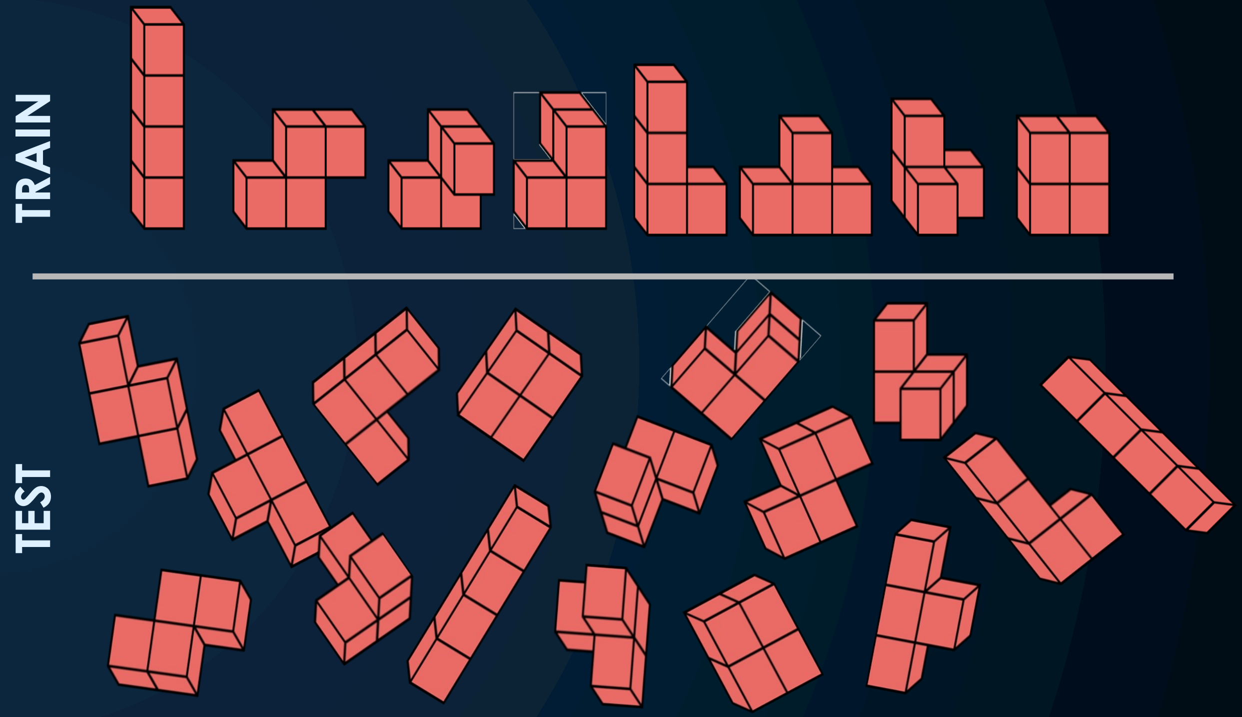


Clara Fannjiang and Jennifer Listgarten at NeurIPS '20

CNNs for Materials with Physical Laws

Physics-aware learning



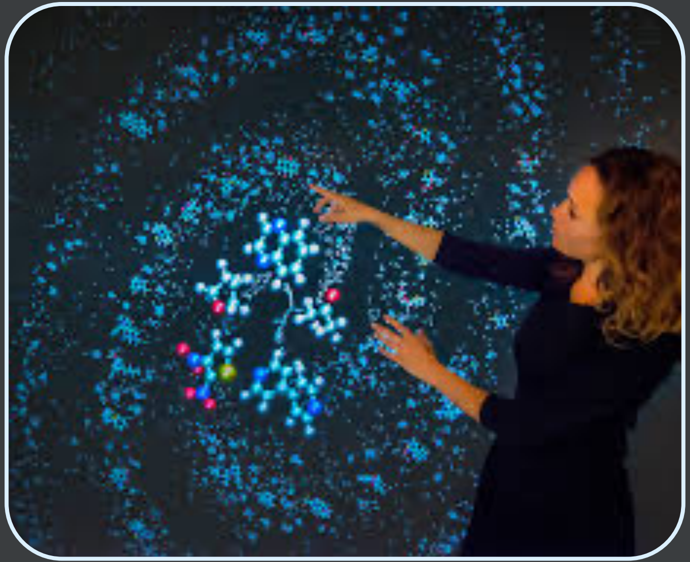


A network with 3D translation- and 3D rotation-equivariance

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

Opportunities in Science

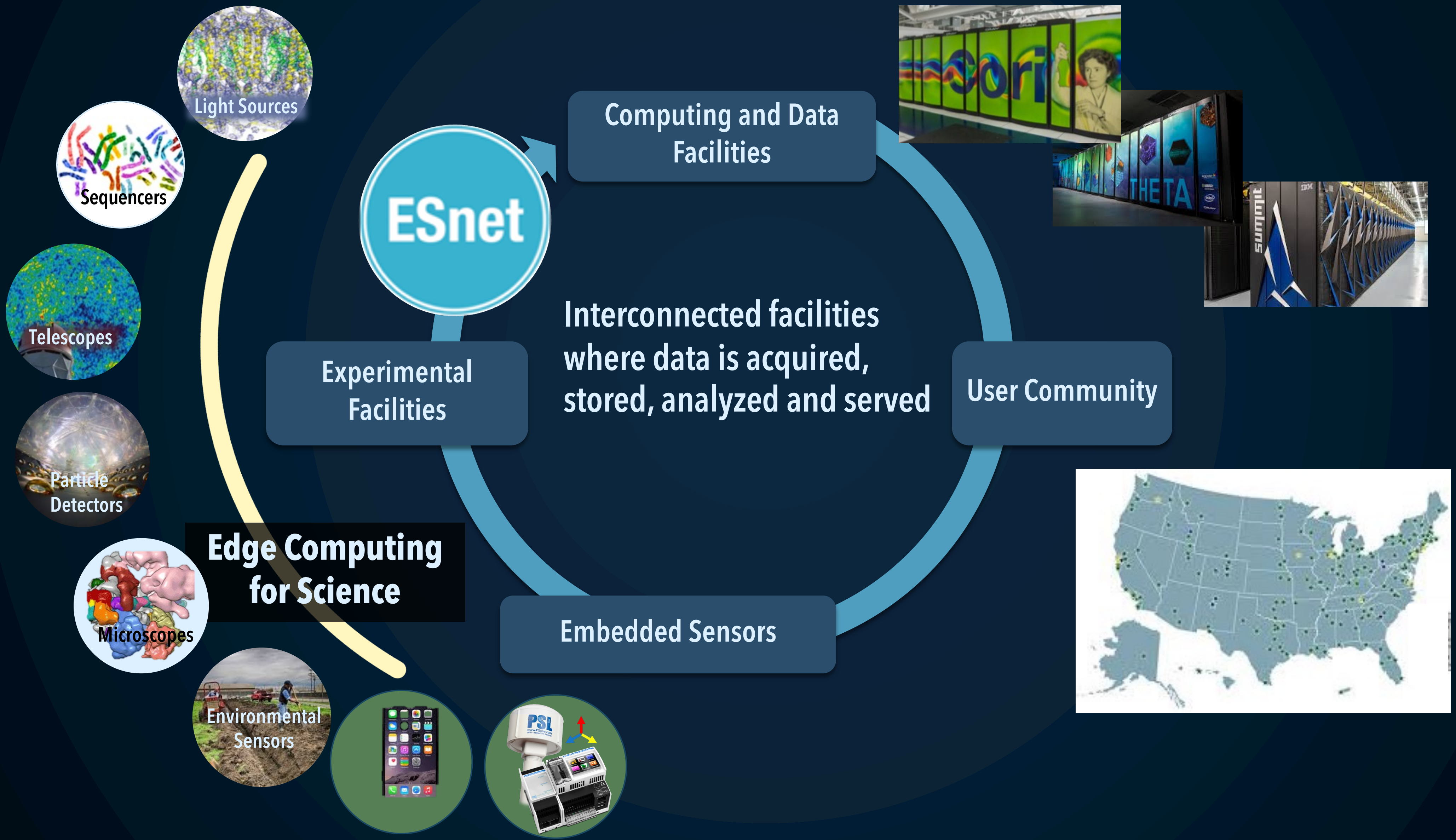
Analyze



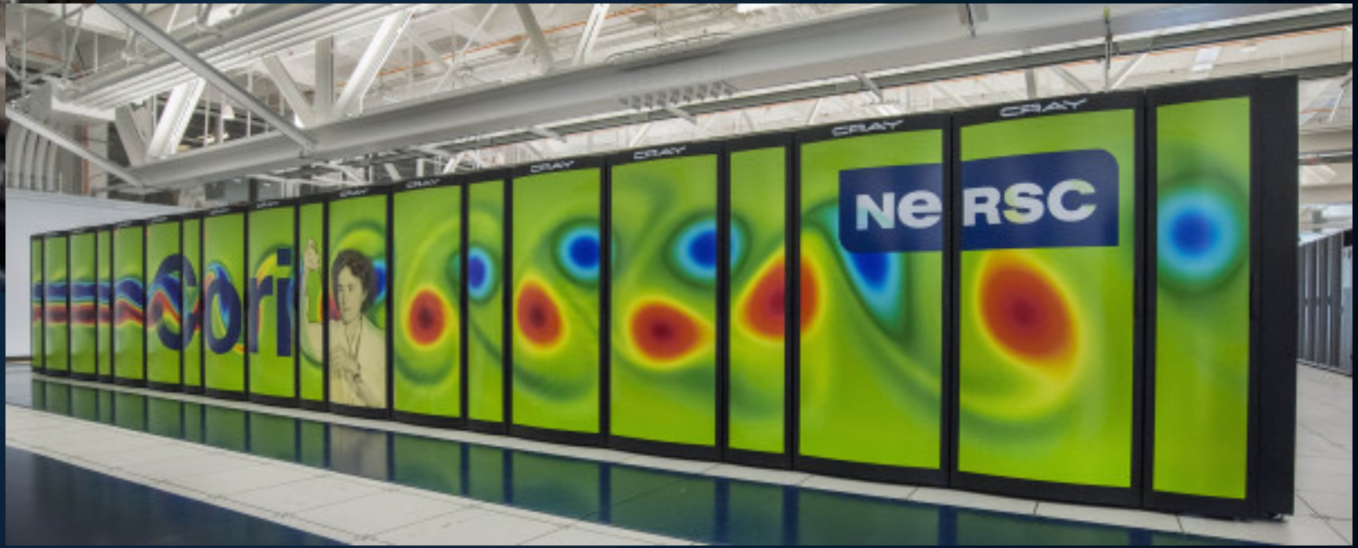
Explore

Automate

Edge Computing and Automation in Science

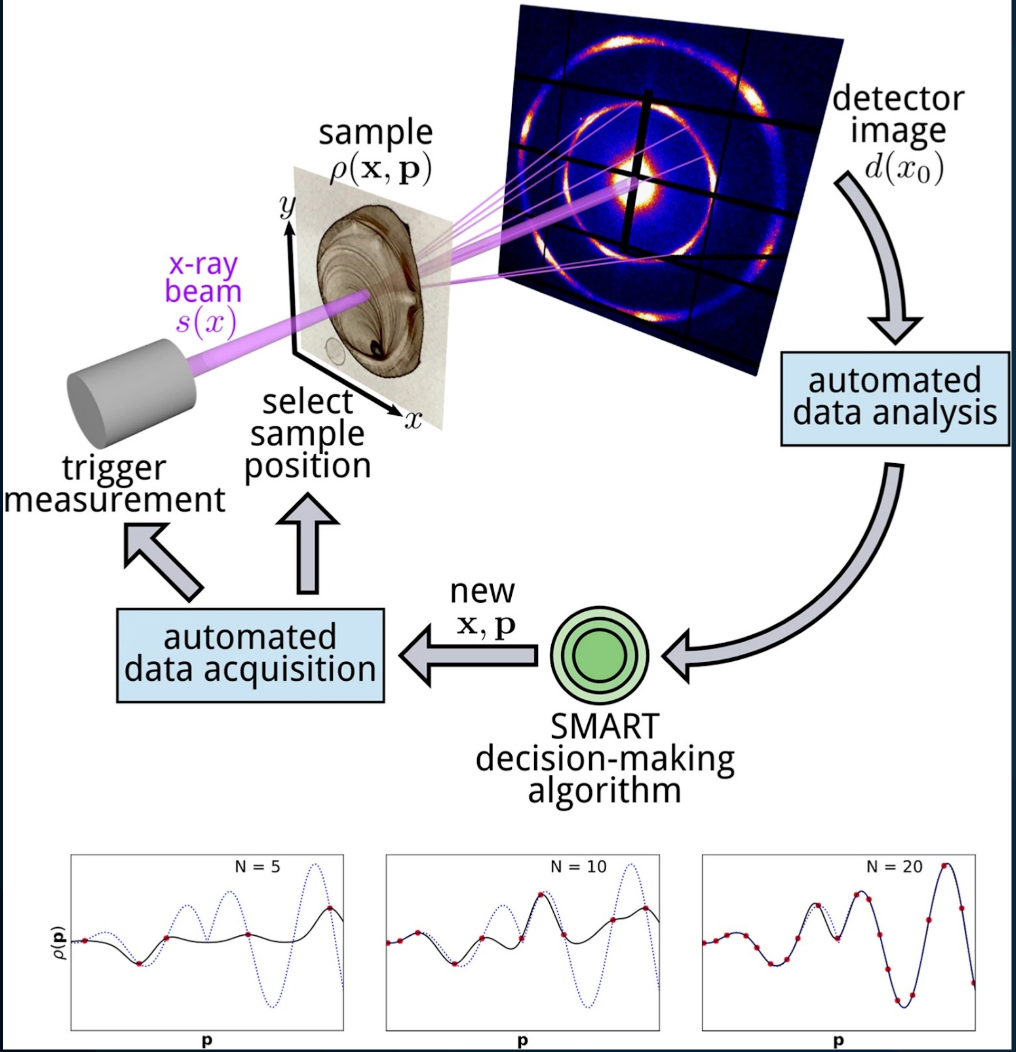


Streaming Experimental Data



Researchers from Turkey working at the Linac Coherent Light Source at SLAC have used X-ray crystallography to capture detailed images of the structure of the SARS-CoV-2 virus.

Automated experiments

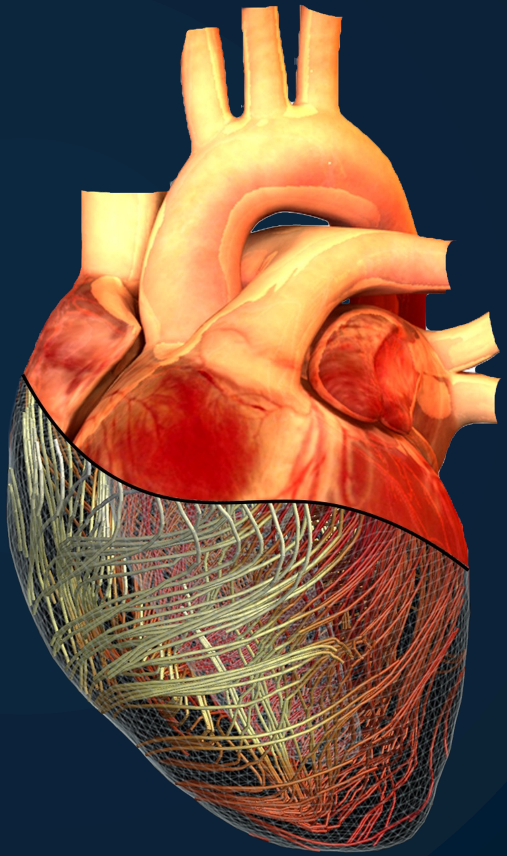


Utilization and robustness

- AI-based autonomous discovery
- Decisions based on small datasets
- Uncertainty estimates

Source: CAMERA Project, PI James Sethian
Slide input: Lavanya Ramakrishna

Digital Twins

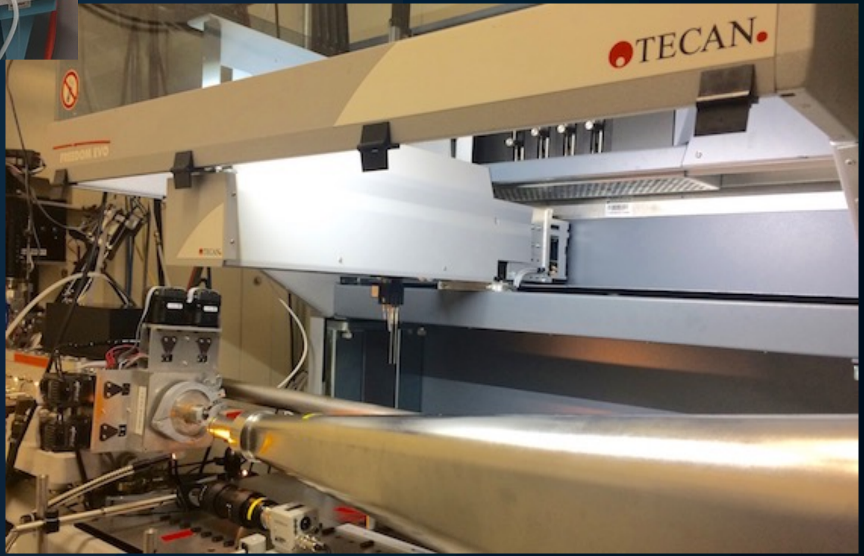


- Simulations
- Sensors / data
- Multi-level
- Real-time

Robotics and precision control in science

MassSpec robot at JGI

Nanoparticle Robot at the Molecular Foundry



Robot at SYBLIS beamline at ALS

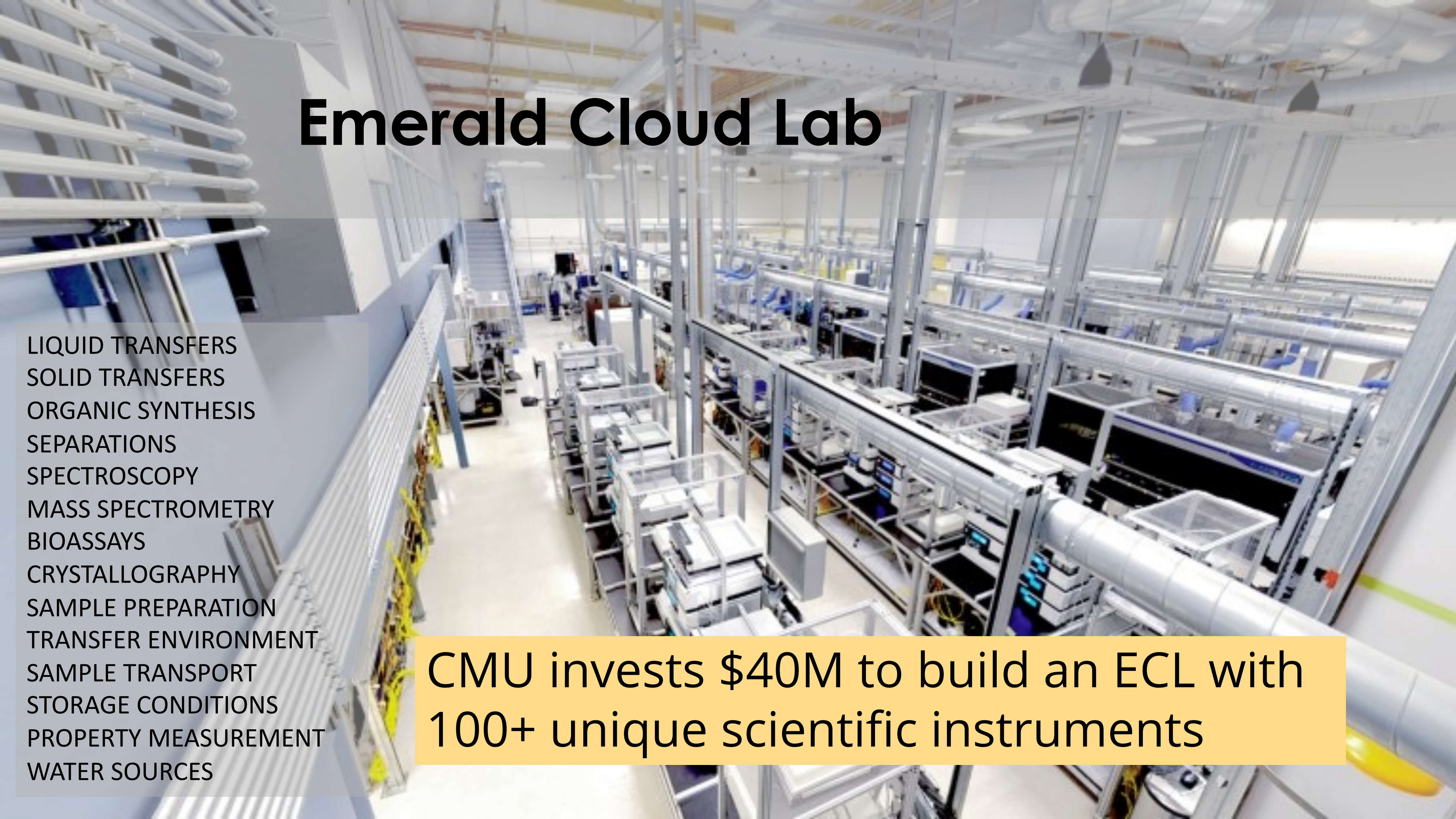
Self-Driving Cars

Self-Driving Laboratories

Automated COVID-19 Testing at the Innovative Genomics Institute at Berkeley

Strateos Cloud Lab

10 YEARS
14K SQUARE FEET
200+ INSTRUMENTS



Emerald Cloud Lab

LIQUID TRANSFERS
SOLID TRANSFERS
ORGANIC SYNTHESIS
SEPARATIONS
SPECTROSCOPY
MASS SPECTROMETRY
BIOASSAYS
CRYSTALLOGRAPHY
SAMPLE PREPARATION
TRANSFER ENVIRONMENT
SAMPLE TRANSPORT
STORAGE CONDITIONS
PROPERTY MEASUREMENT
WATER SOURCES

CMU invests $40M to build an ECL with 100+ unique scientific instruments

Why Cloud Lab?

Source: Emerald Cloud Lab

Setting up

"Plugging an experiment into a browser forces researchers to translate the exact details of every step into unambiguous code"

https://www.theguardian.com/

ML in Science

Economics of Science


CloudBank

This is not just about replicating human capabilities

Is there an ML Advantage in science?

2018 ACM Turing Award for Deep Learning

Yoshua Bengio
Photo: Facebook

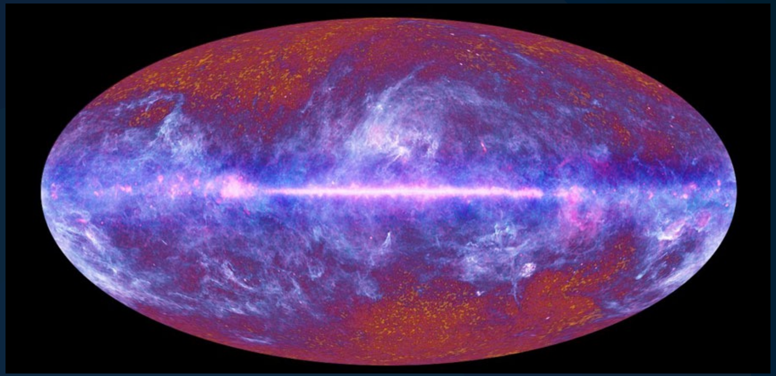
Yann LeCun
Photo: Google

Geoffrey Hinton
Photo: Botler AI

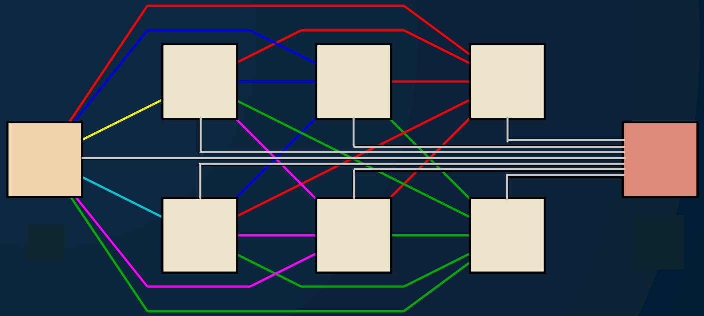
Hinton's Turing Lecture:
"So I think a lot of the credit for deep learning really goes to the people who collected the big databases like Fei Fei Li and the people who made the computers go fast like David Patterson and others."

Where can data+compute yield breakthroughs?

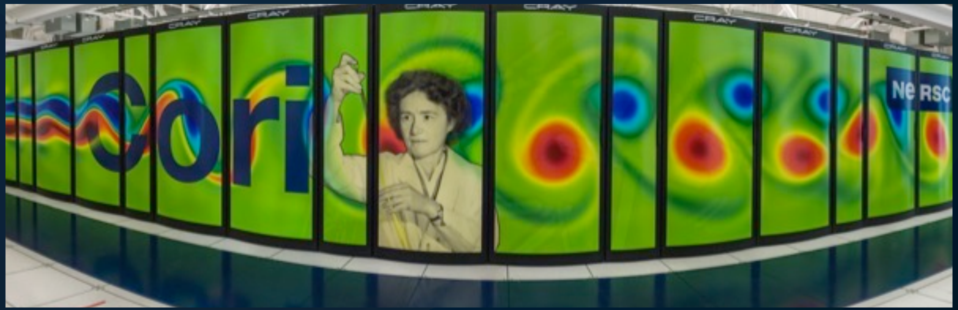
Big Data



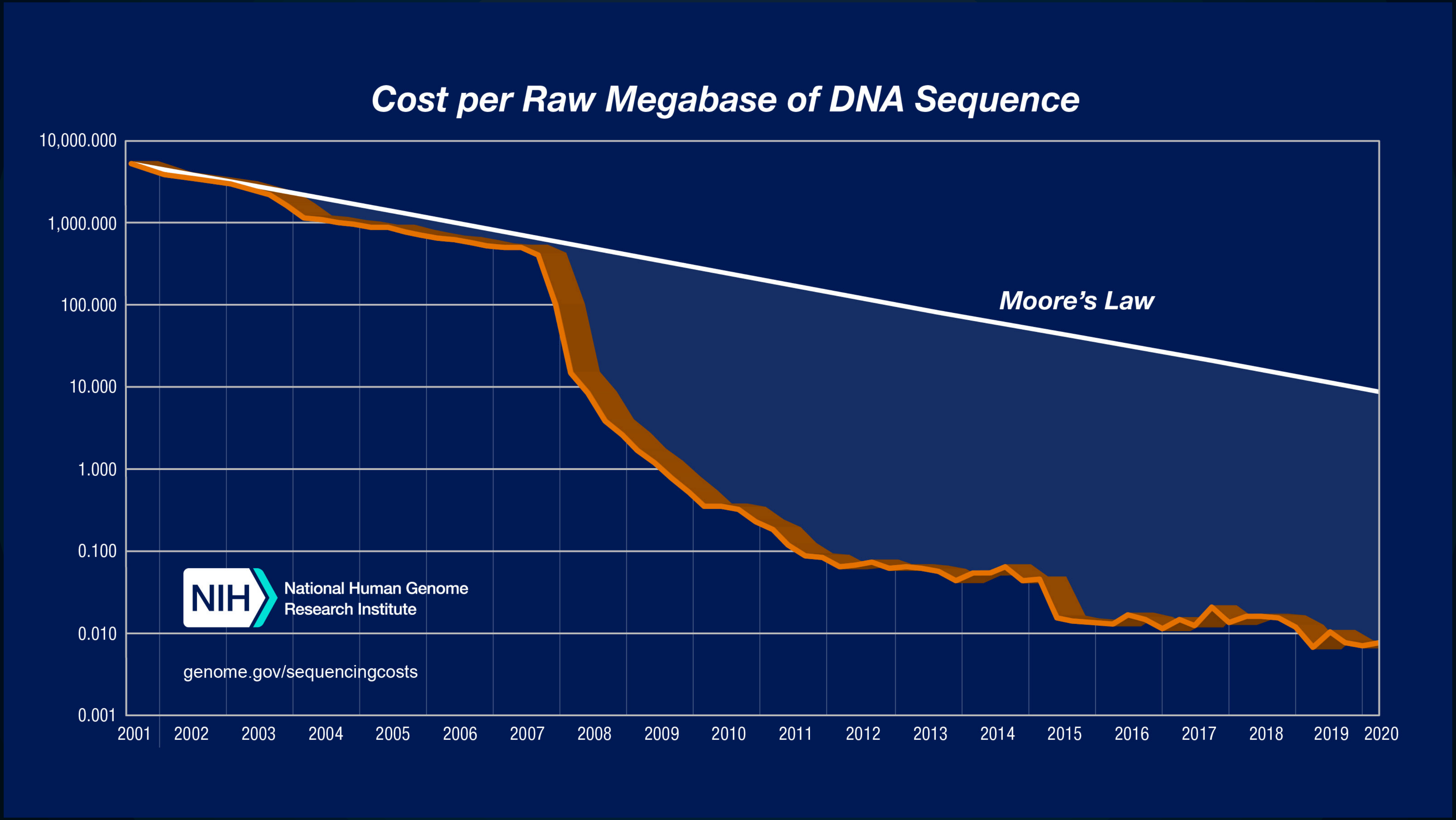
Scalable Algorithms



Big Iron



Sequencing continues to improve in cost and quality



De Novo Metagenome Assembly is Hard

How do microbes change across 17 years? (25TB)

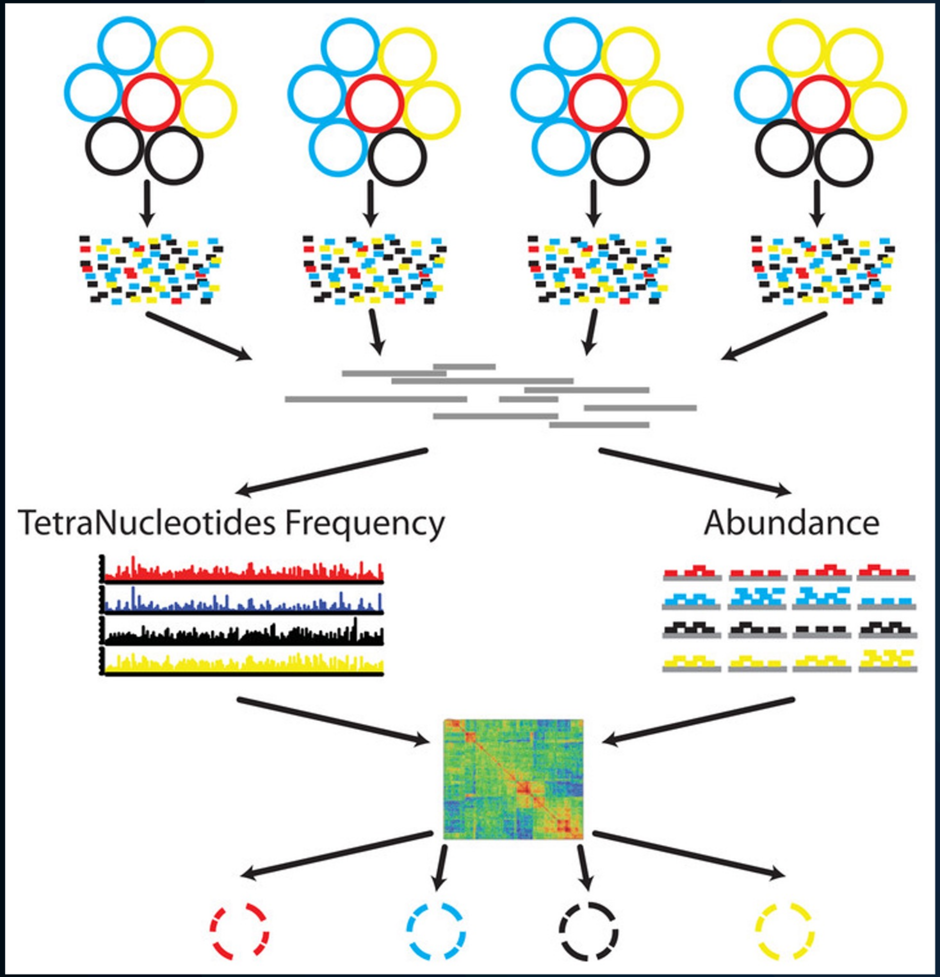
(Non-automated science)

<u>Tara Oceans Assembly</u>

Microbial data from all oceans, collected from 2009-13

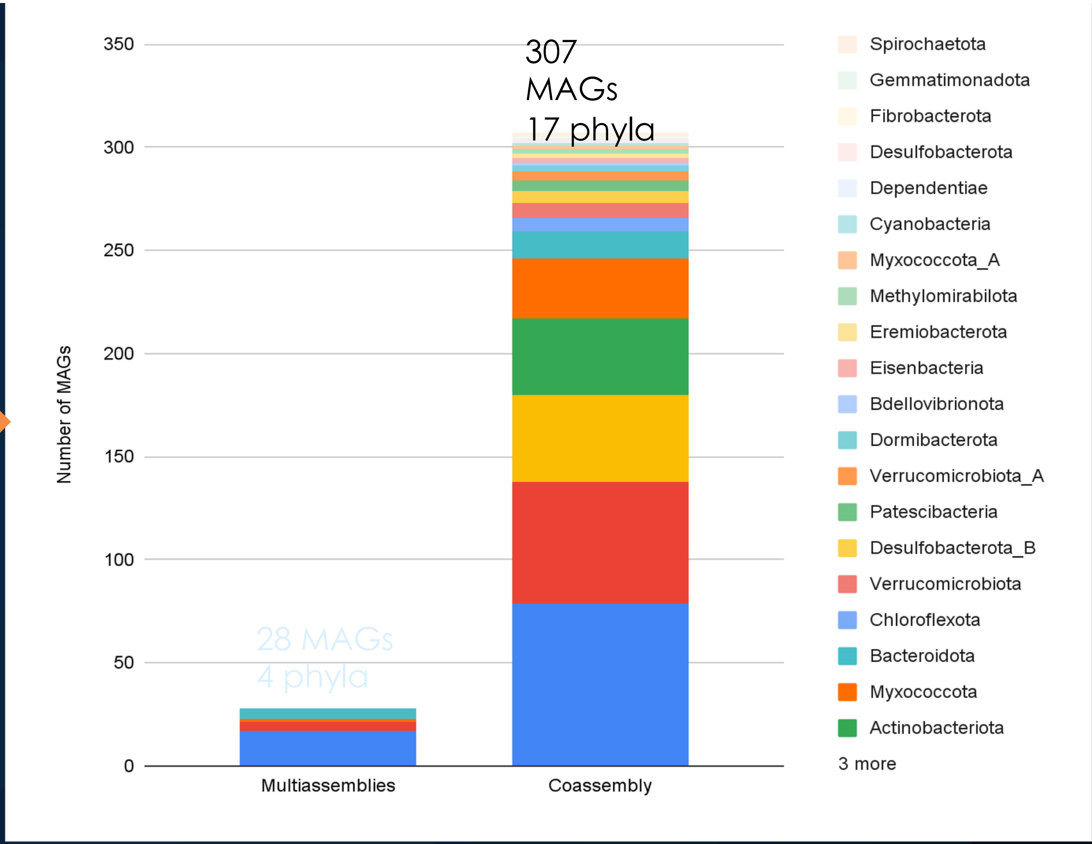
84 Terabytes, never before co-assembled

Terascale Data + HPC Reveals more Genomes and Diversity



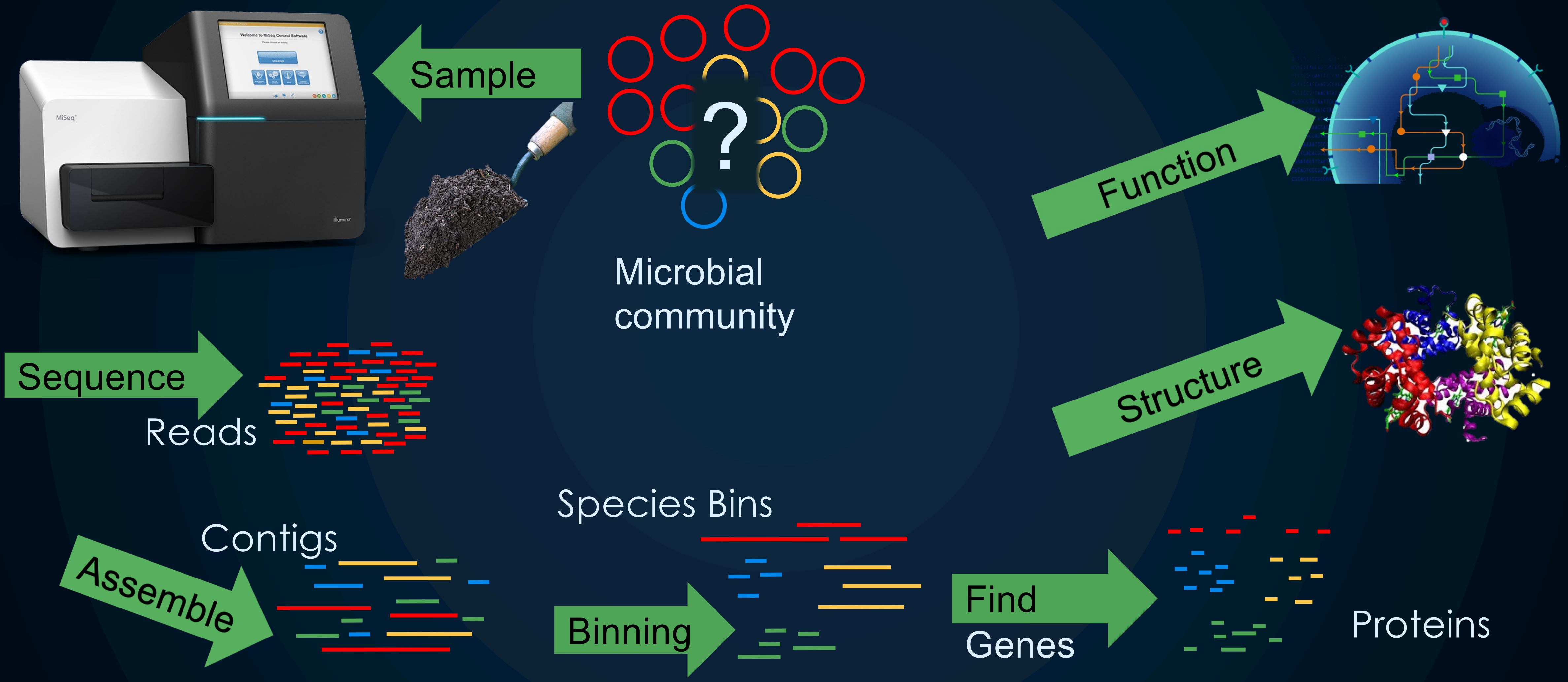
Metagenome-assembled genomes (MAGs)

More Genomes (MAGs), more phyla, coassembly



Co-Assembly of large environmental studies require an HPC metagenome assembler, MetaHipMer

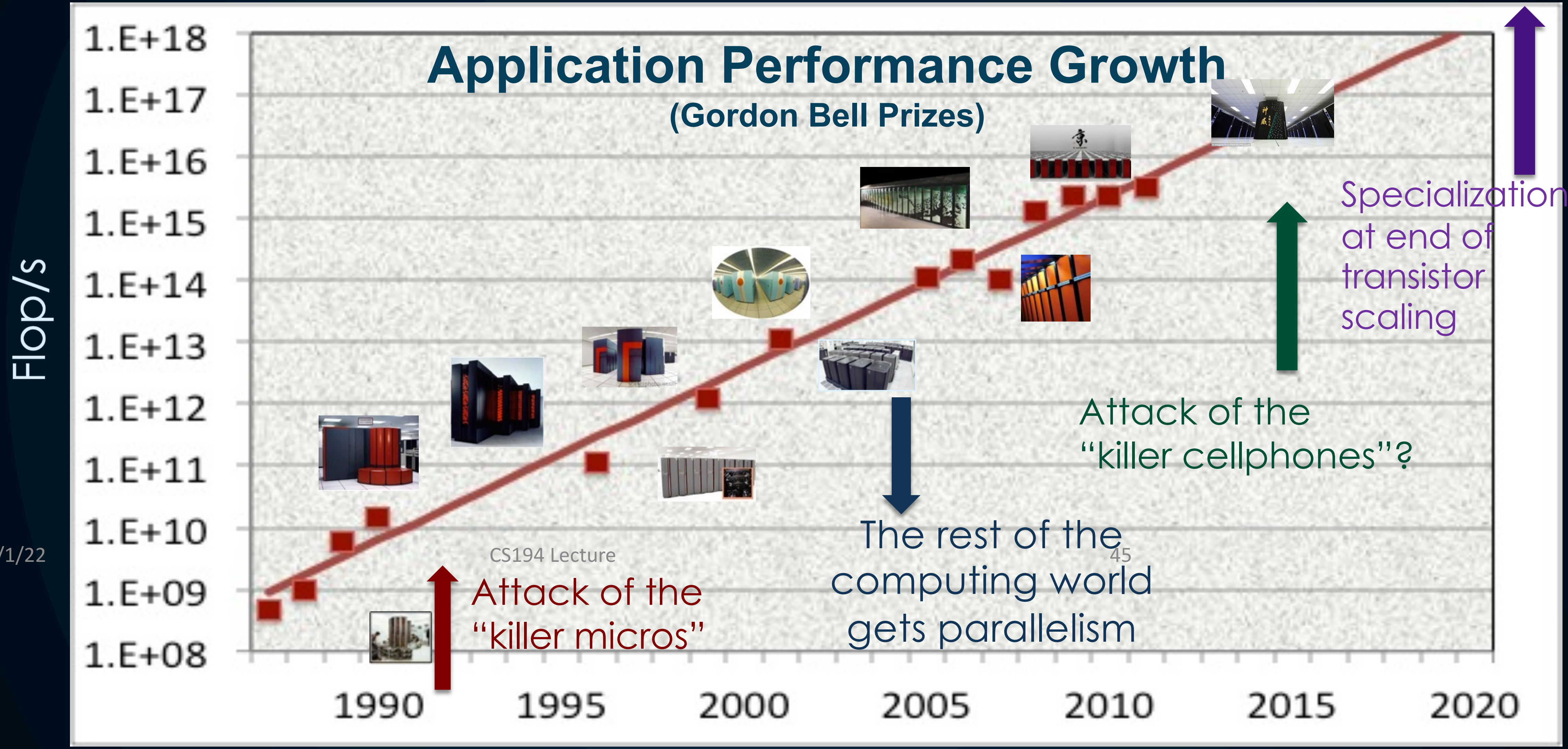
Microbiome analysis: Machine Learning Options



ML is everywhere!

Hardware (and Software and Algorithms)

Technology Transitions



AI Chip Landscape

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

Tech Giants/Systems

Google
Microsoft
aws
IBM
facebook
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

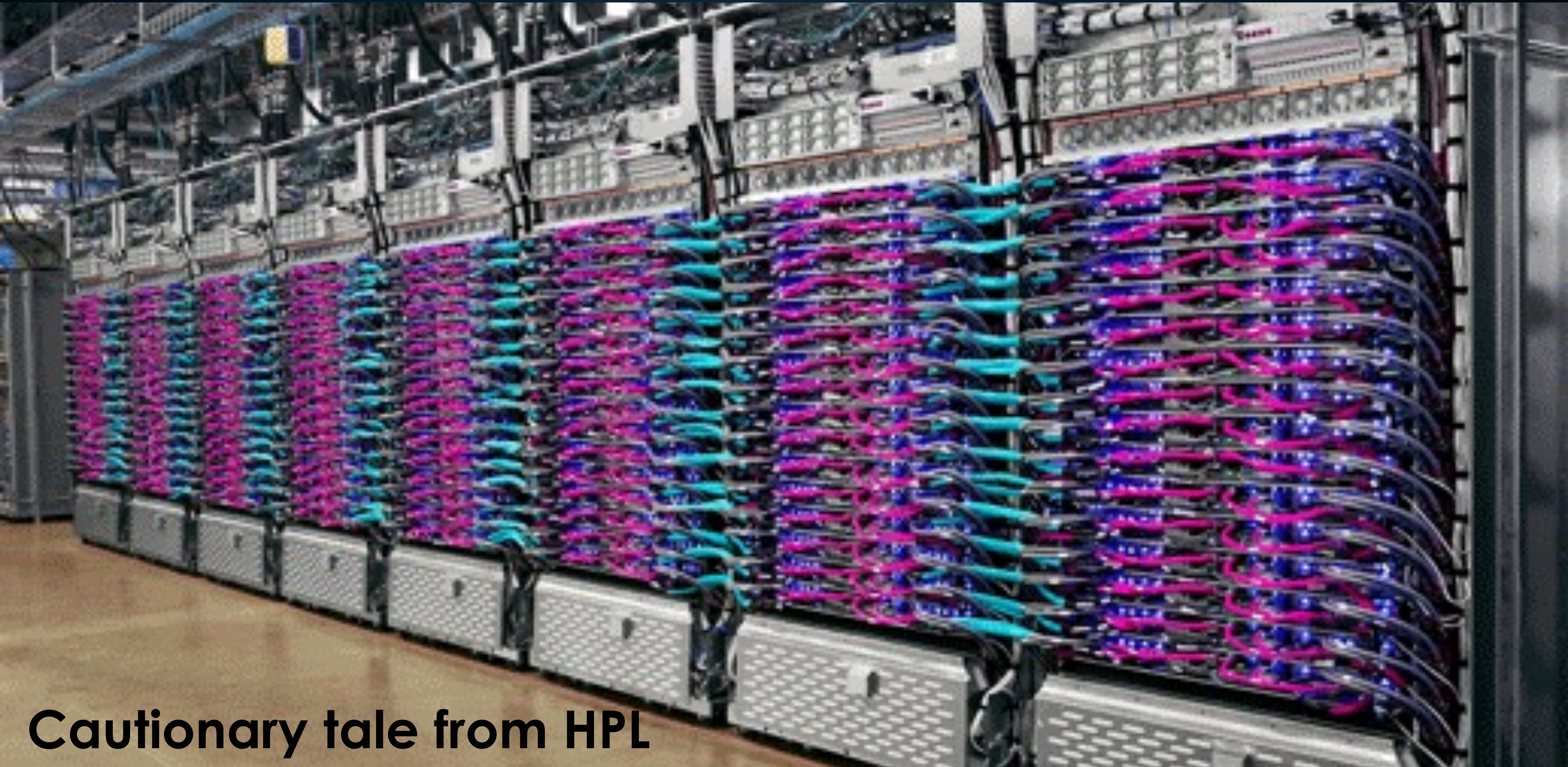
Compiler

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

Benchmarks

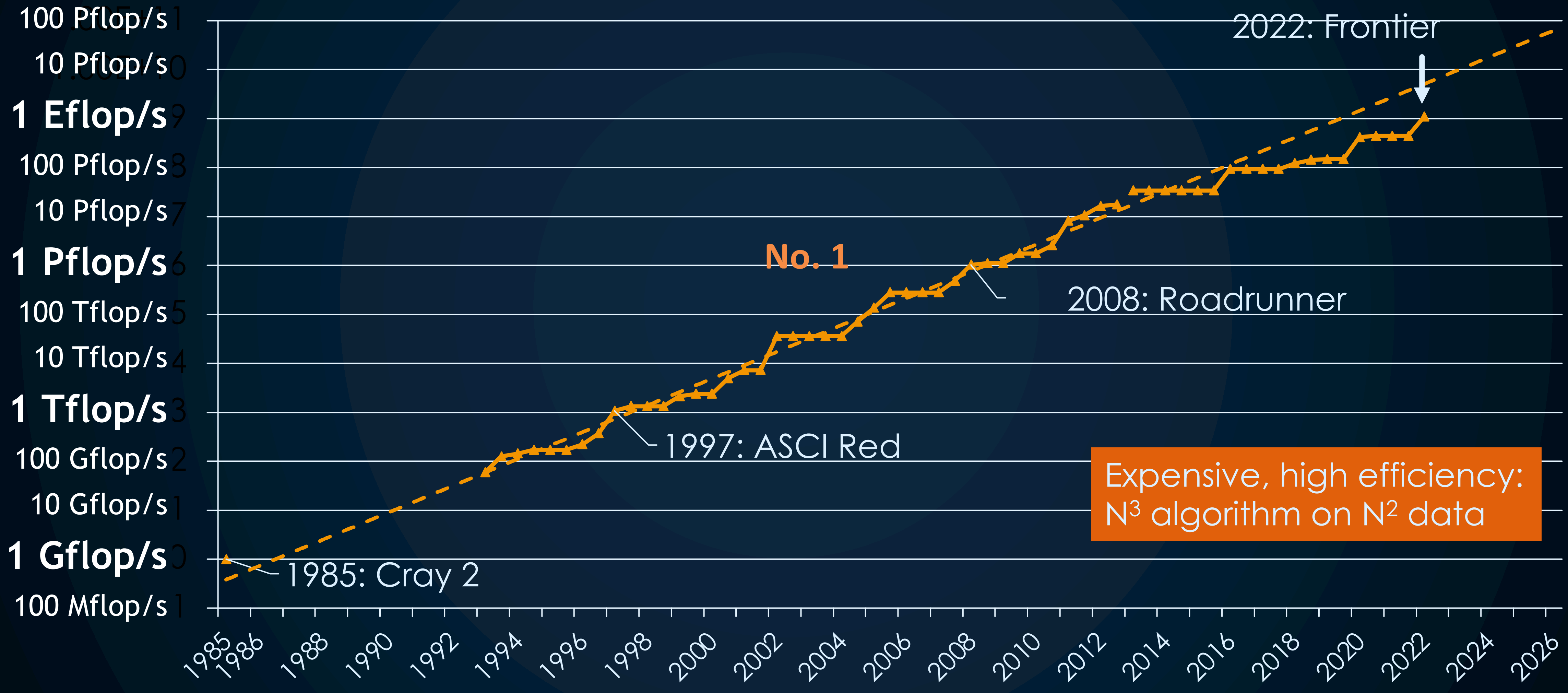
MLPerf
AI - Benchmark
AI Matrix.

Are CNNs the only application?



Cautionary tale from HPL

Top500: Linpack Benchmark

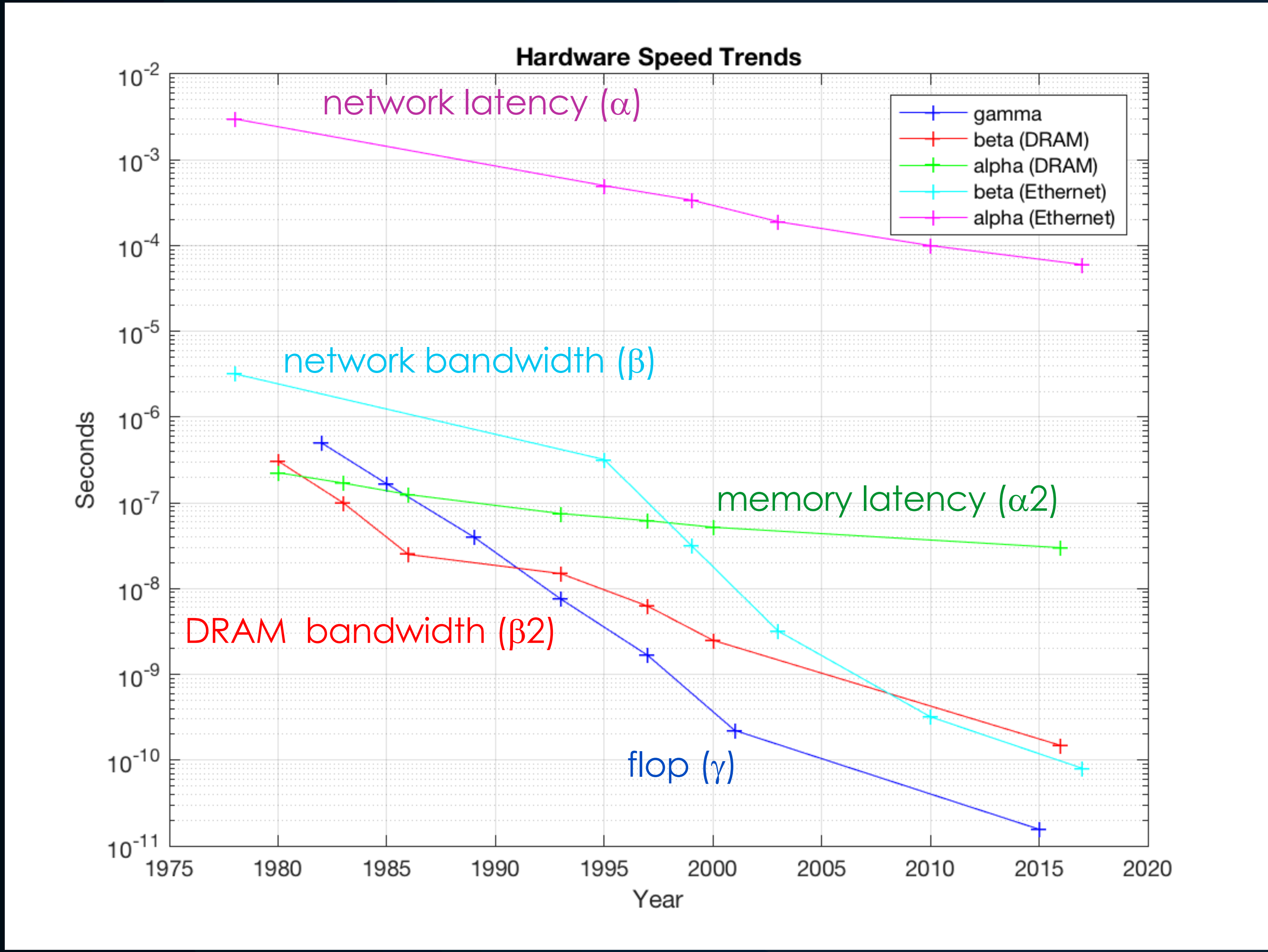


Expensive, high efficiency:
N^3 algorithm on N^2 data

Response: sparsity, hierarchy, etc.

Improve runtime
Worse hardware utilization (% peak)

Communication Dominates



Time =

flops * γ +

message * α +
bytes comm * β +

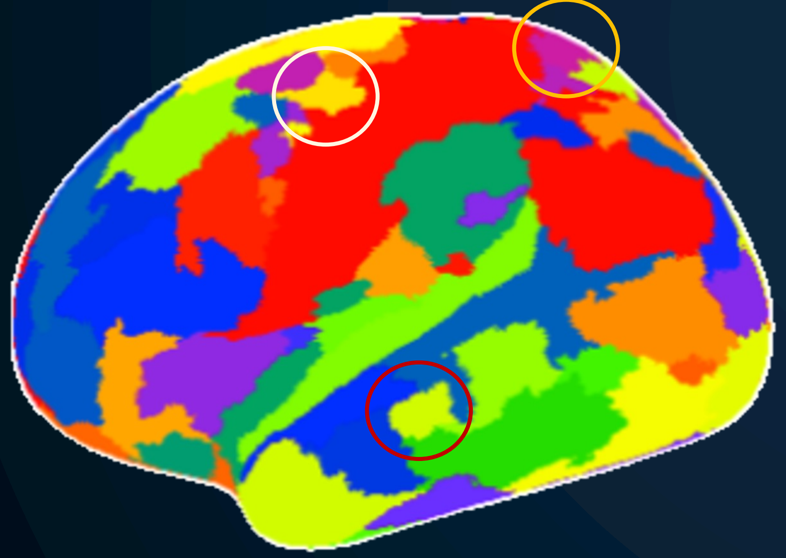
diff memory locs * $\alpha 2$ +
memory words * $\beta 2$

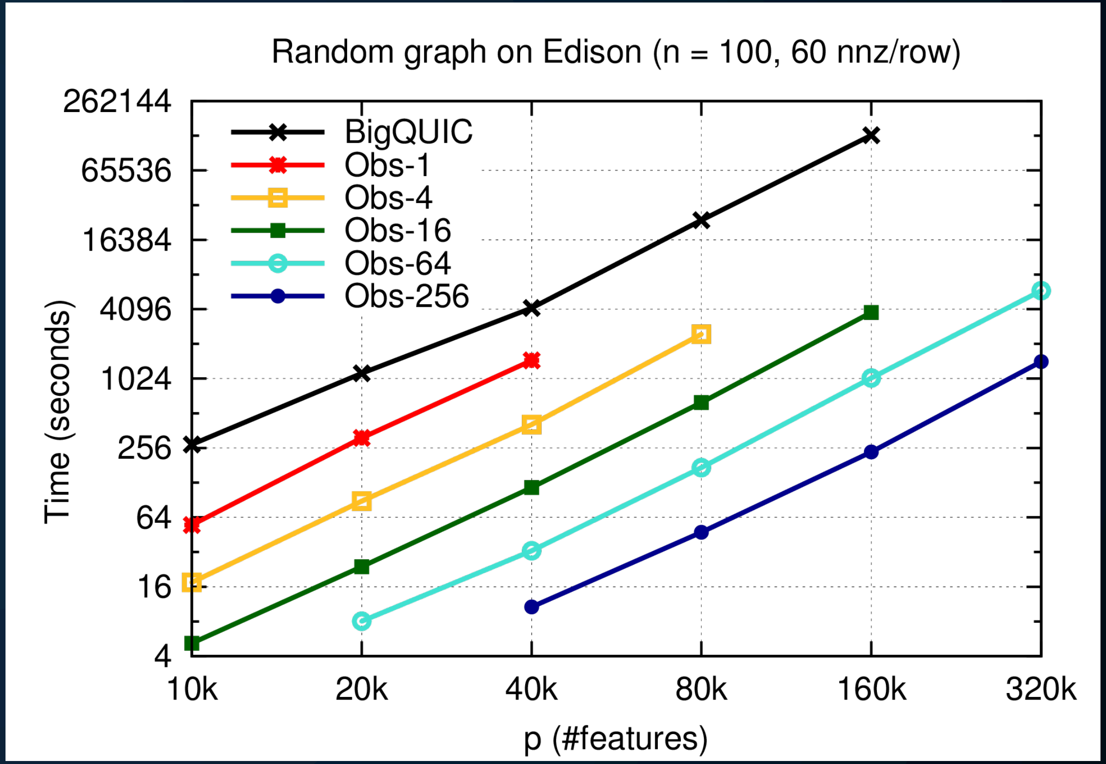
Data from Hennessy / Patterson, Graph from Demmel

Learning Relationships with Graphical Models

Discovering Regions and Co-Regions of Brain Activity from fMRI

91K x 91K Sample Covariance matrix

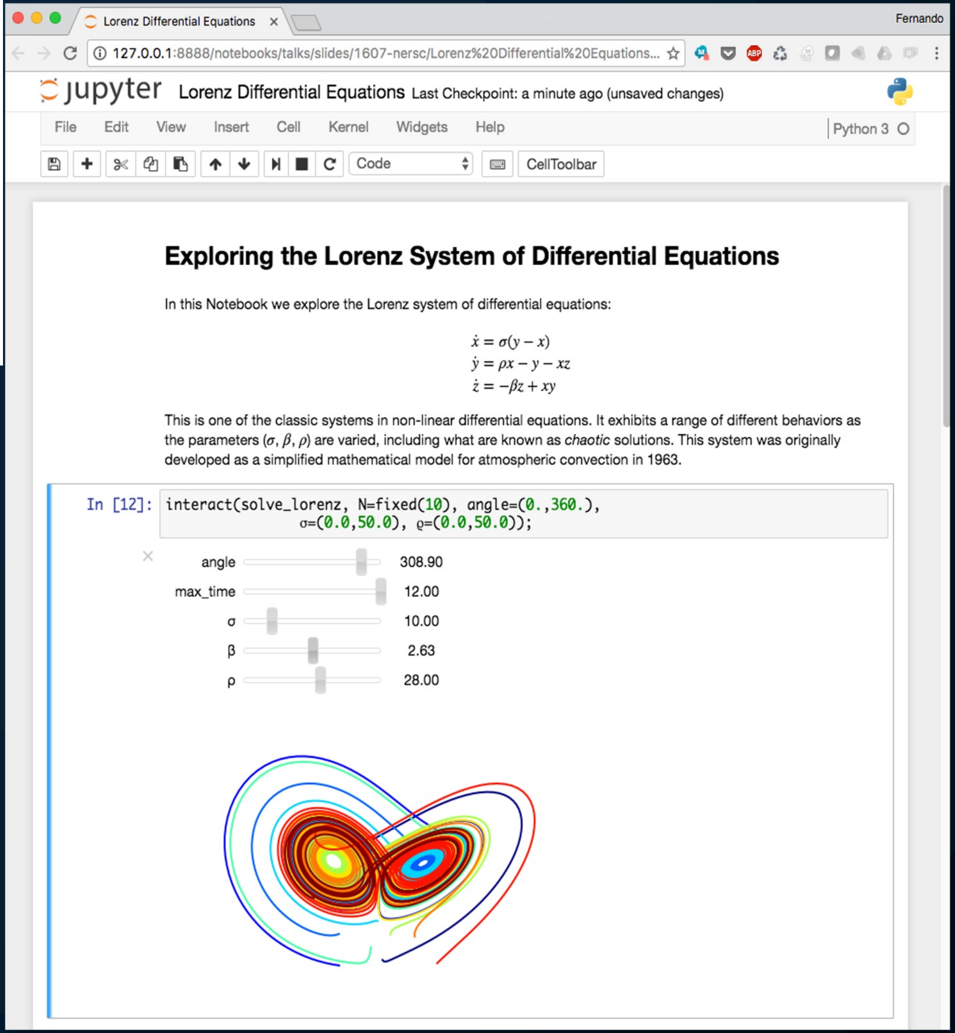




Koanantakool, Buluc, Morozov, Oliker, Yelick, Oh, AISTAT 2018.

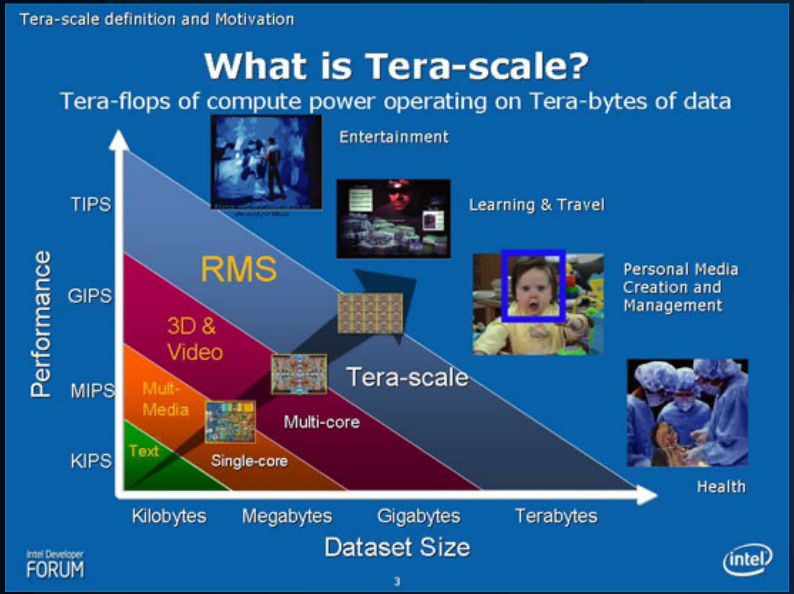
The IPython/Jupyter Notebook

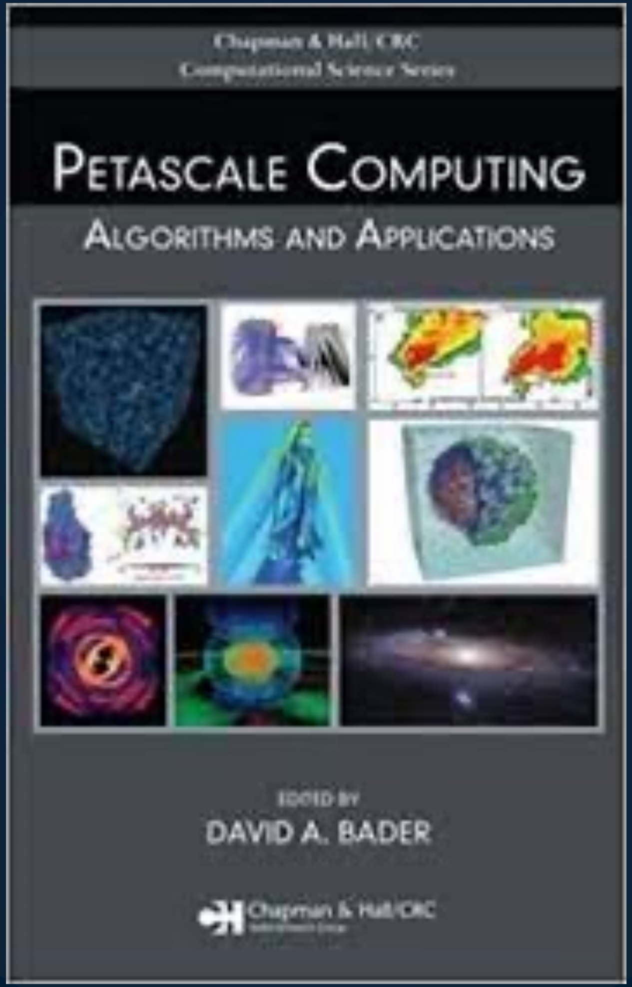
- Rich web client
- Text & math
- Code
- Results
- Share, reproduce.

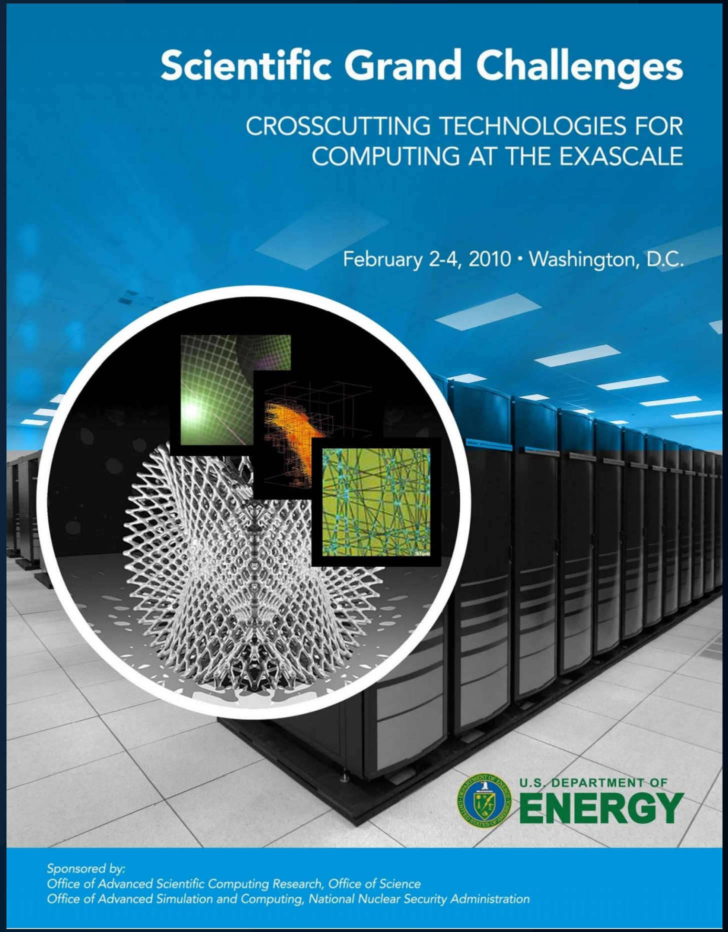


Transform publishing, research, teaching!

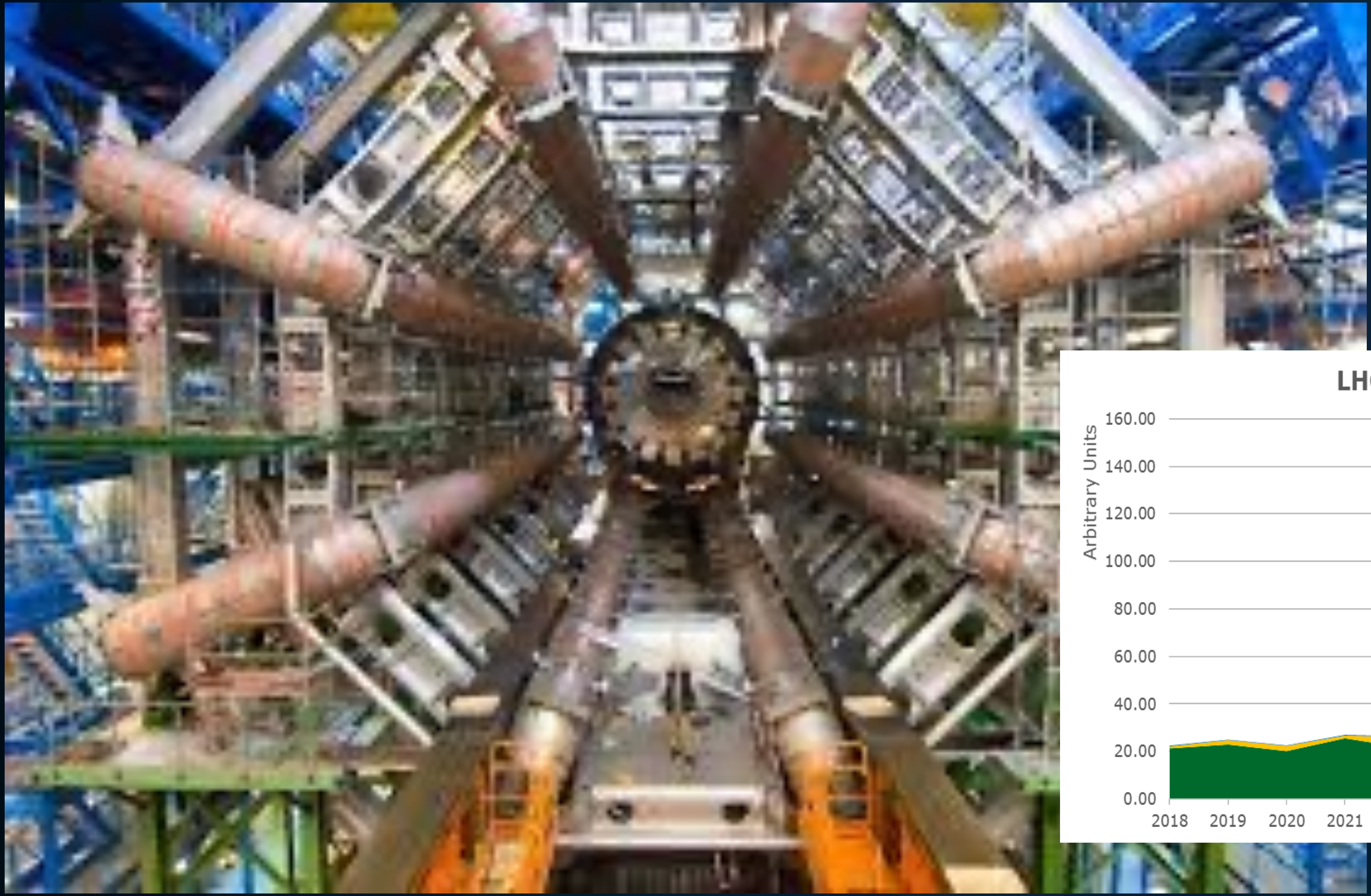
It’s hard to think exponentially

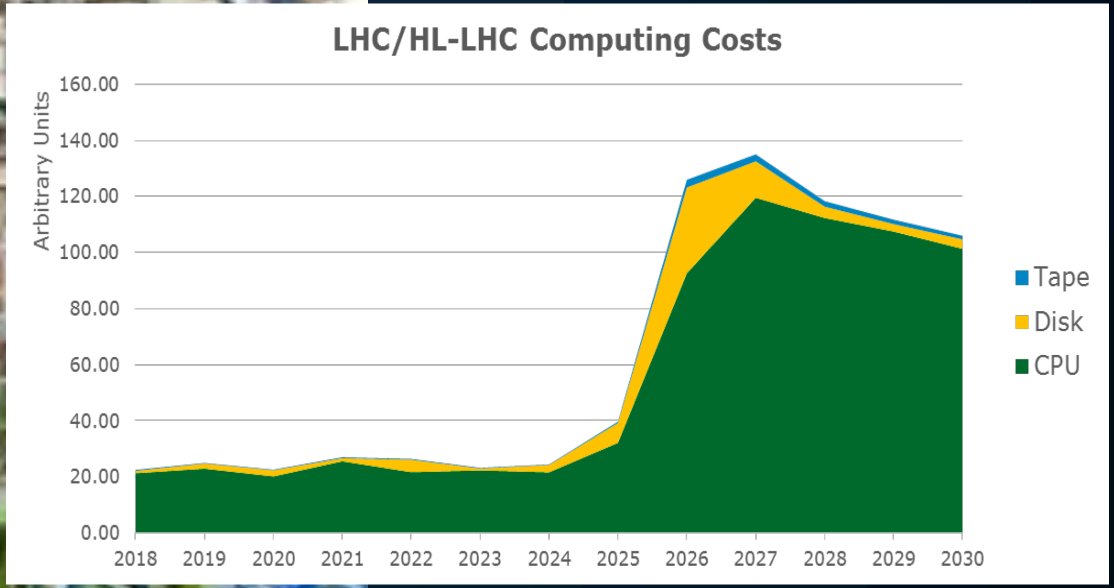






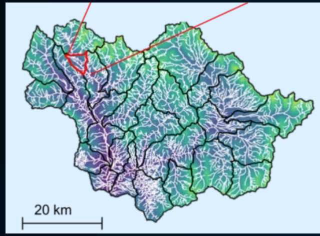
Prediction of Atlas computing +$1B



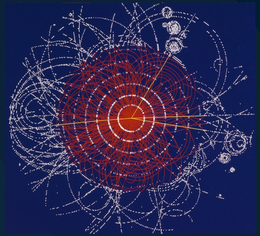
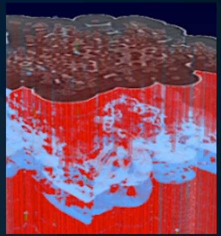
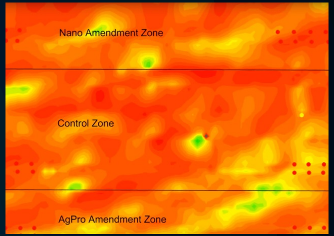


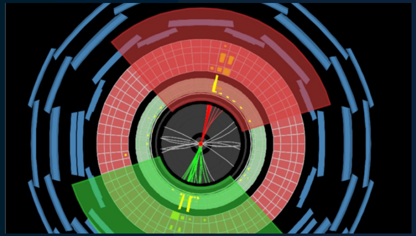
Thanks, Moore!

What Applications and is Science Different?

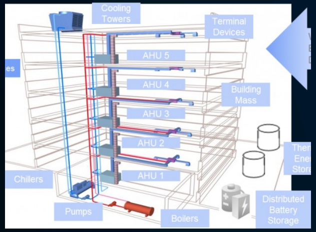


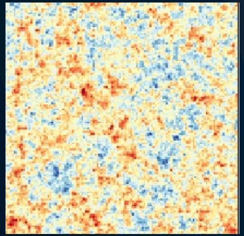


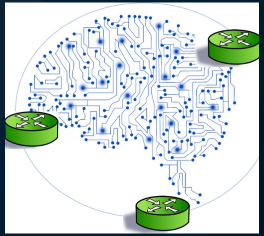
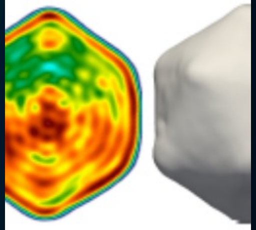



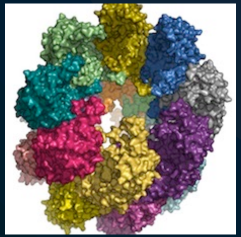
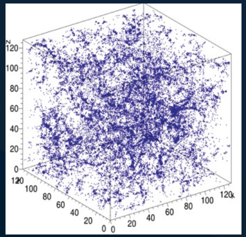
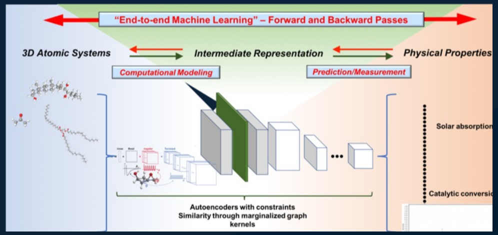


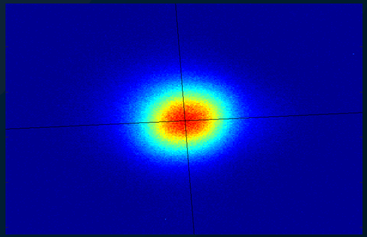









https://ml4sci.lbl.gov

Superfacility in Practice

Facility	Instrument	Location	Users	Compute	Data/Year	Bandwidth	Timeframe
ALS	Lightsource	Berkeley	100s	50M	600TB	10Gb/sec	2025 Upgrade
DESC	Telescope	France	100s	150M	2000TB		2024
DESI	Telescope	Arizona	100s	200M	500TB	~10GB/night	2020
JGI	Genomics	Berkeley	100s	75M	self		Continuously
KSTAR	Tokamak	Korea	10s	145M	20TB	10GB/hour	1-2 per year
LCLS	Lightsource	Stanford	100s	12M	1000TB	100 Gb/sec	~bimonthly
LZ	Dark Matter	South Dakota	100s	20M	1000TB	1GB/hour	2021, 24/7
NCEM	Electron Microscope	Berkeley	10s	1M	600TB	100Gb/sec	2021

Exascale Architecture Plans (2008)

Exascale Architecture Plans (2021)

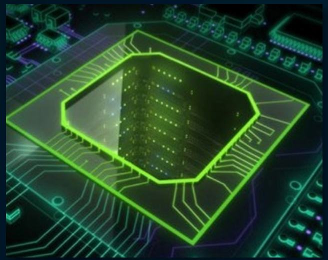
US DOE Office of Science Systems

Exascale
HPE AMD+AMD

Exascale
HPE Intel+Intel

Pre-exascale
HPE AMD+NVIDIA

Trend Toward Specialization

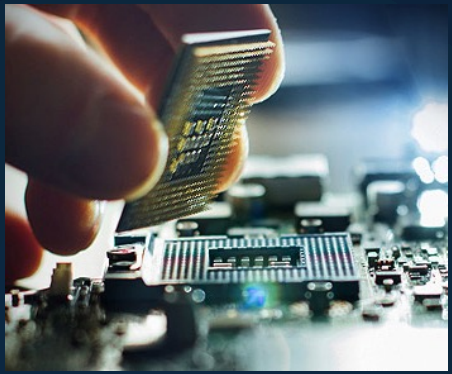


NVIDIA builds deep learning appliance with P100 Tesla's

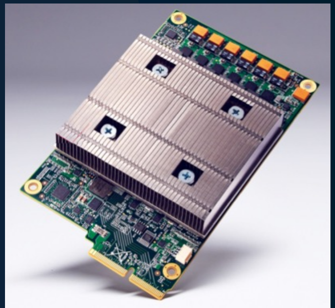
FPGAs in Microsoft cloud

RISC-V is an open hardware platform

RISC-V



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

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

Thanks!

2 Parallelism Models

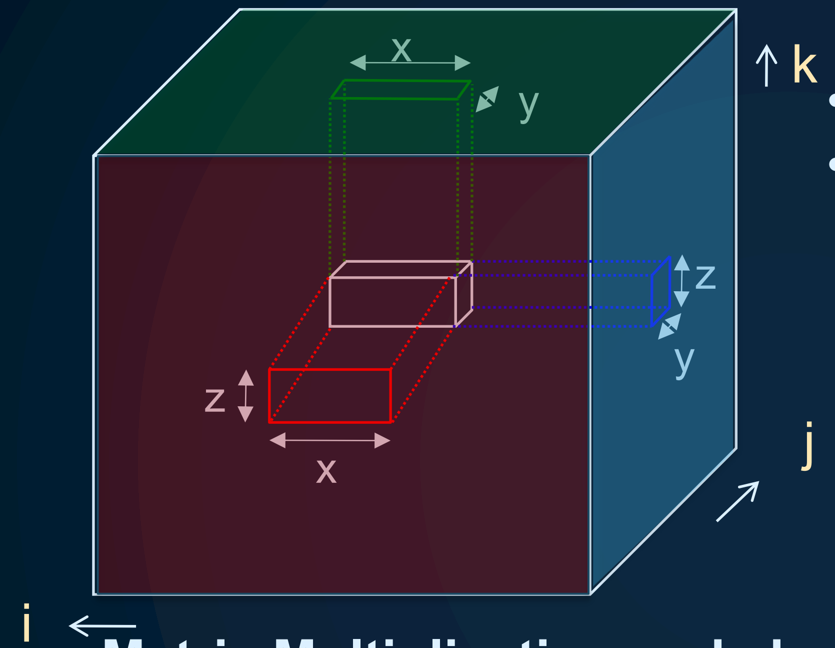
Bulk synchronous

- Latency: reduce span
  - Log time algorithms
- Bandwidth: reduce volume
  - Iteration space tiling
- (Sparse) matrix abstraction
  - For general semiring

Asynchronous

- Latency: hide cost
  - Overlap and minimize overhead
- Bandwidth: maximize utilization
  - "All the wires all the time"
- Partitioned Global Address Space
  - Application-specific optimizations

Communication-Avoiding Matrix Multiply



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

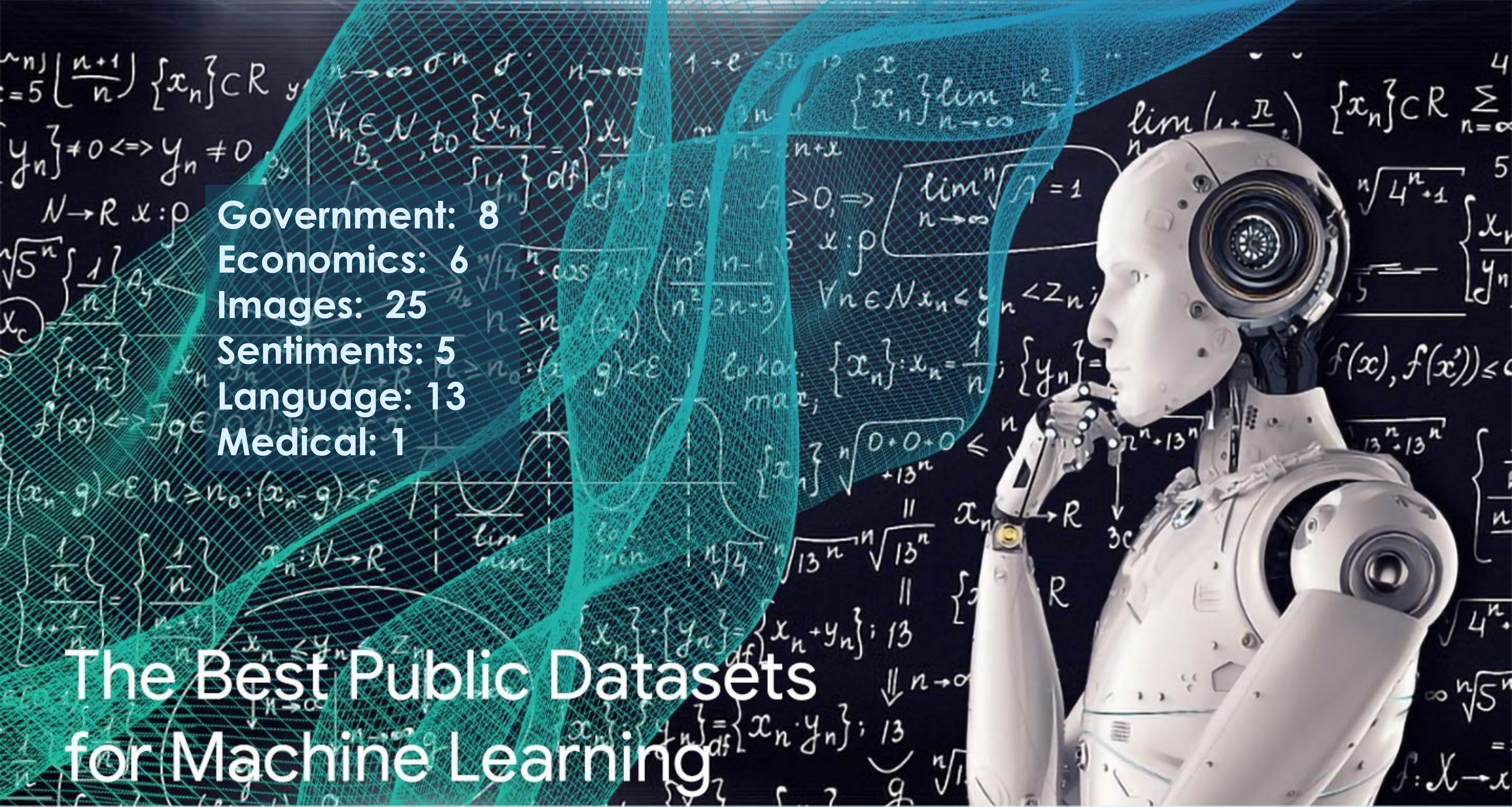
Matrix Multiplication code has a 3D iteration space
Each point in the space is a constant computation (*/+)

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

Be smart about price vs. cost

Factor	HPC Center	Commercial Cloud
Utilization (30% private, 90% HPC, 60%? Cloud); Note: trades off against wait times, elasticity	++	
Cost of people, largest machines lowest people costs/core		+
Cost of scientific consulting	++	
Cost of power, advantage for placement of center, bulk		++
Energy efficiency (PUE, 1.1-1.3 is possible; 1.8 typical)		
Cost of specialized hardware (interconnect)		+
Cost of commodity hardware		+
Profit	+++	

Sophisticated users who spend a lot of money on computing, use commercial clouds only when the spot pricing is very low; otherwise it's too expensive



Government: 8
Economics: 6
Images: 25
Sentiments: 5
Language: 13
Medical: 1

https://medium.com/towards-artificial-intelligence/the-50-best-public-datasets-for-machine-learning-d80e9f030279

Google Computing Platform 1997

NERSC Scientific Computing Center 1996

NERSC 2022

Google 2022

Over 8 years needed 300,000 times more computing to do machine learning!


Supercomputers
NERSC
Oakland
Alameda
Daly City
Pacifica
San Francisco Bay
San Leandro
Moss Beach
Princeton
Foster City
San Mateo
Half Moon Bay
Palo Alto
Woodside
Los Altos
Greenland Sea
Norwegian Sea
ATLANTIC OCEAN
PACIFIC OCEAN
INDIAN OCEAN

Learning from sequence + graph structure

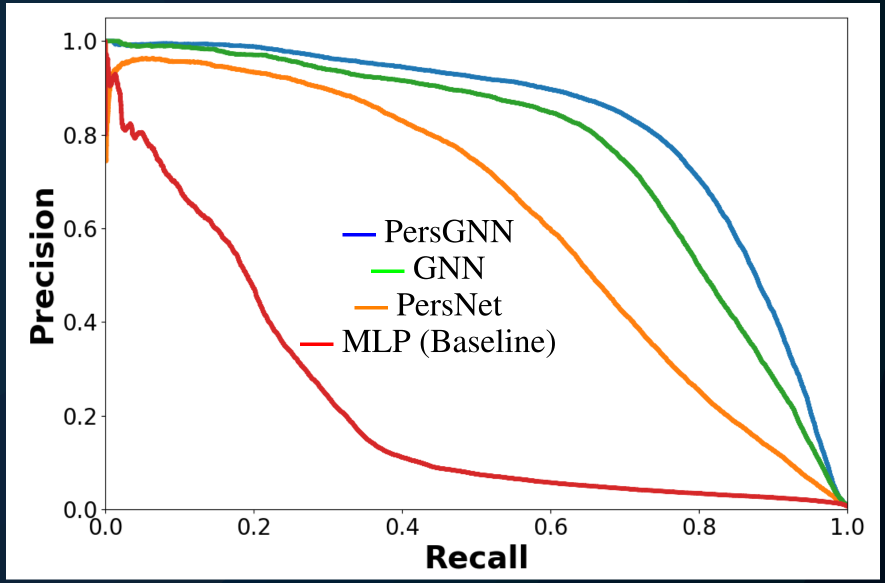
Sequence

Graph Structure

Persistence Network

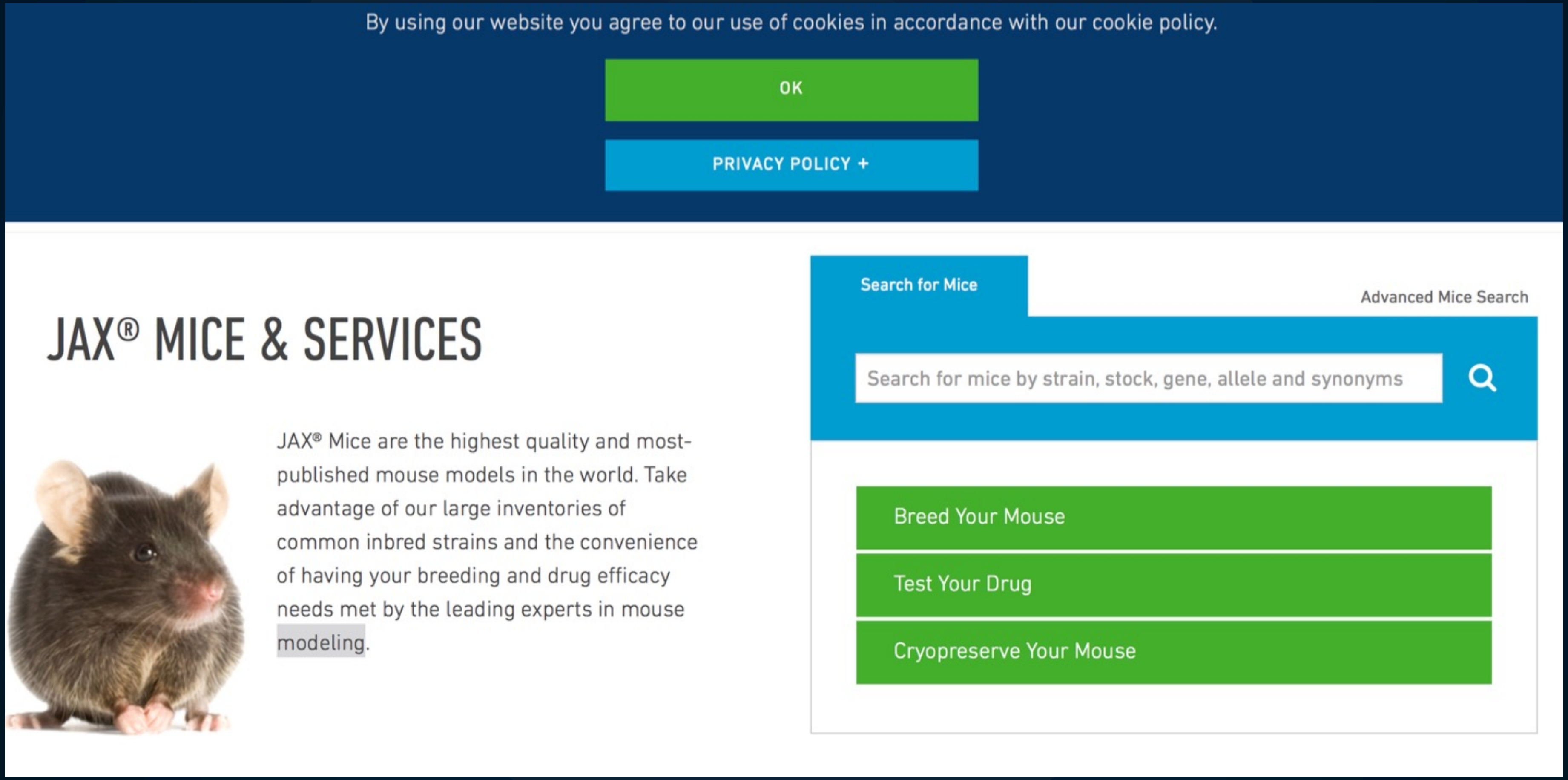
Function

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



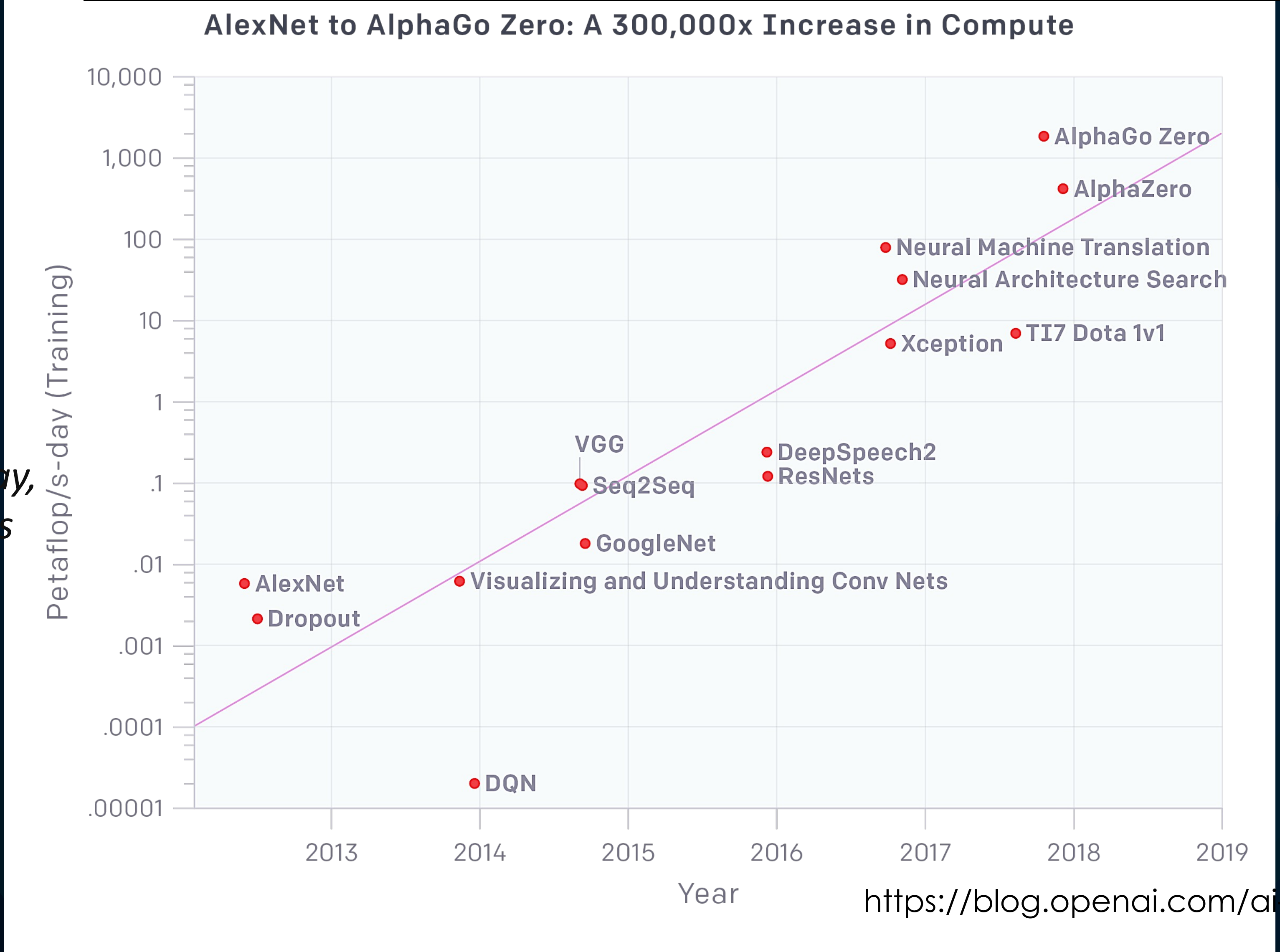
Aditi S Krishnapriyan, Nicolas Swenson, Dmitriy Morozov, Y, Aydin Buluc

Experimental Science is Changing



Why HPC for Learning?

300,000x increase from 2011 (AlexNet) to 2018 (AlphaGoZero)



A petaflop/s-day = 10^{15} neural net operations per second for one day, ~= 10^{20} operations

From 2011-2017 the fastest Top500 machine grew < 10x

https://blog.openai.com/ai-and-compute/