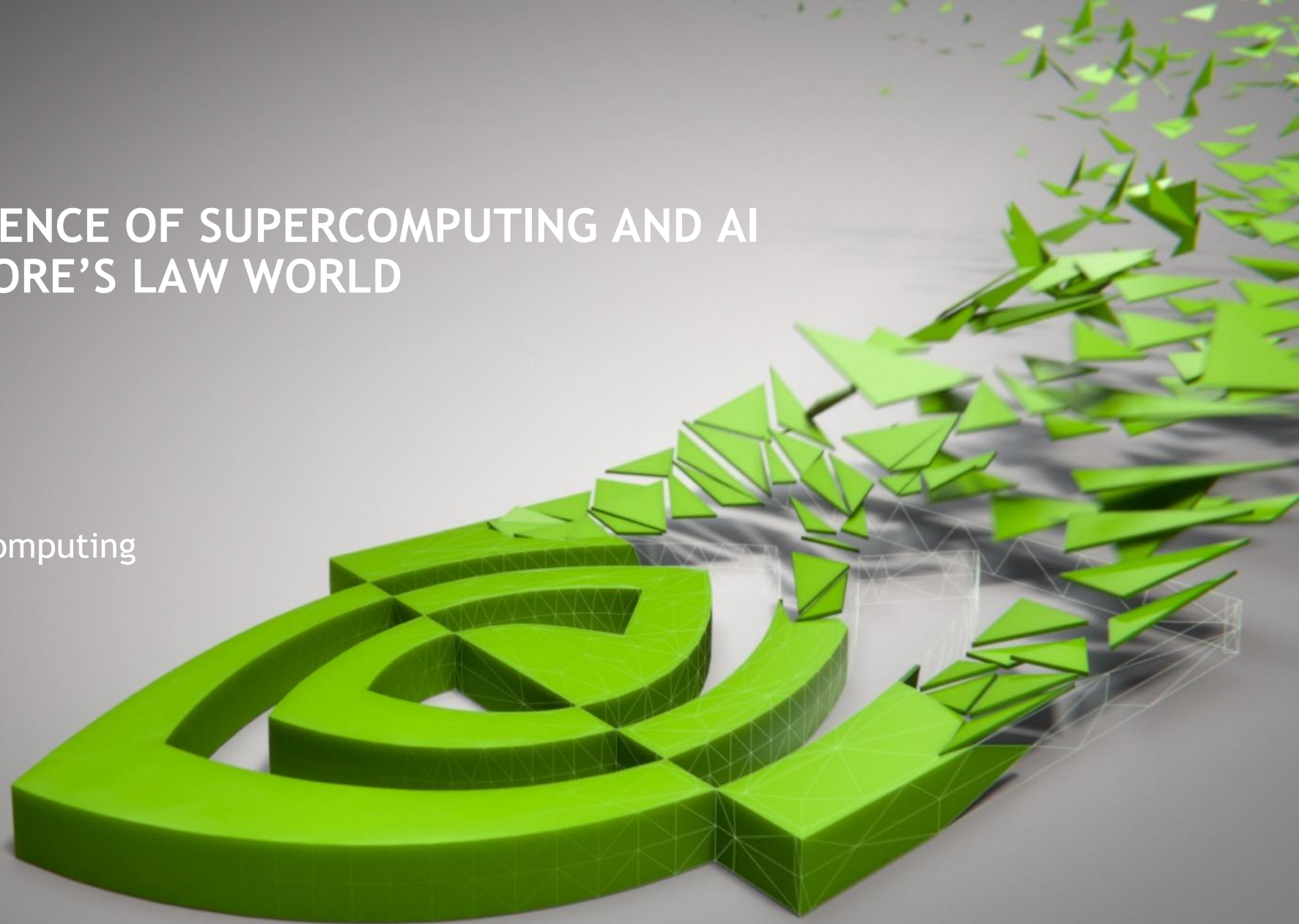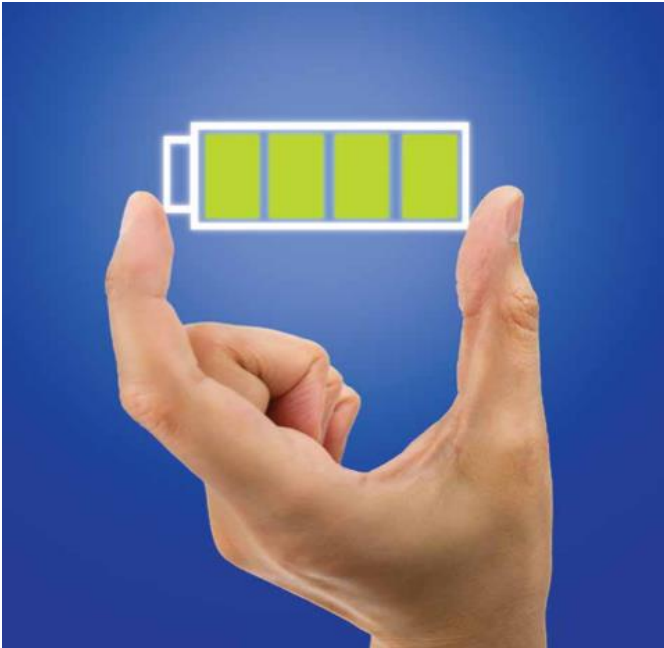# HPC + AI

THE CONVERGENCE OF SUPERCOMPUTING AND AI IN A POST-MOORE'S LAW WORLD

Piero Altoe'

IBD, Accelerated Computing
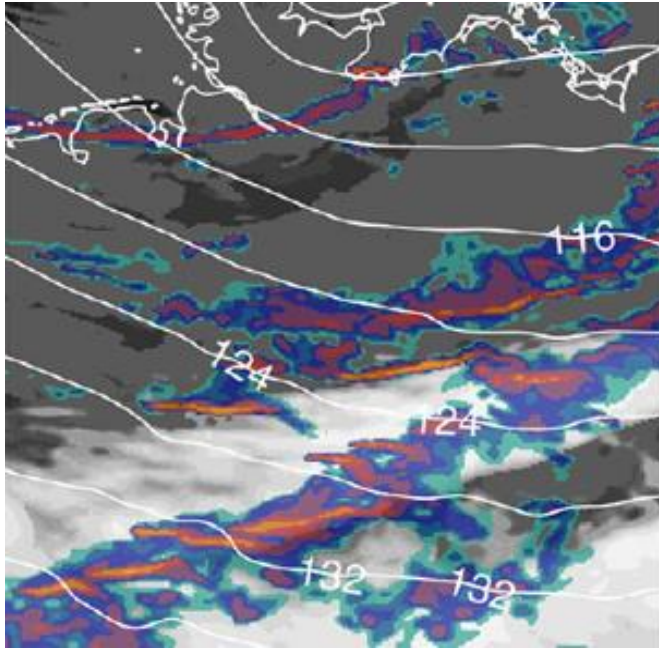
nVIDIA.
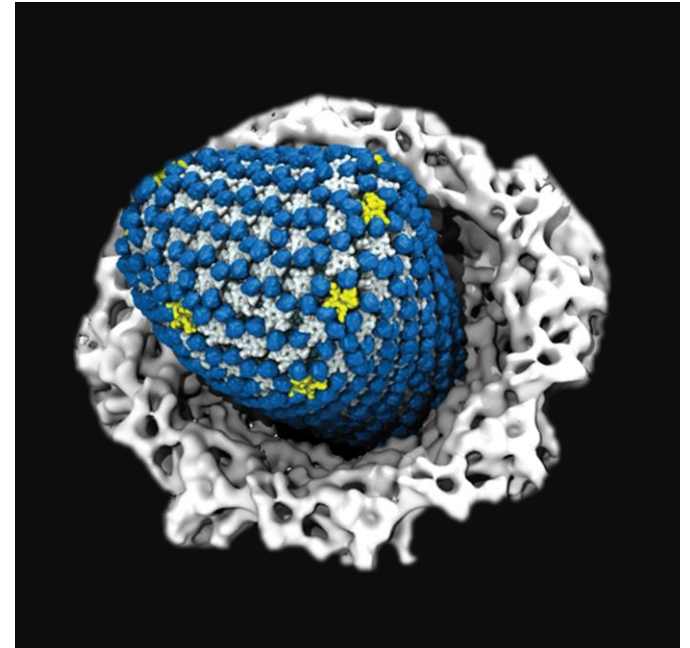
GRAND CHALLENGES REQUIRE MASSIVE COMPUTING

REINVENTING THE LI-ION BATTERY

3M Node Hours | 7 Days on Titan

CLOUD RESOLVING CLIMATE SIMULATIONS
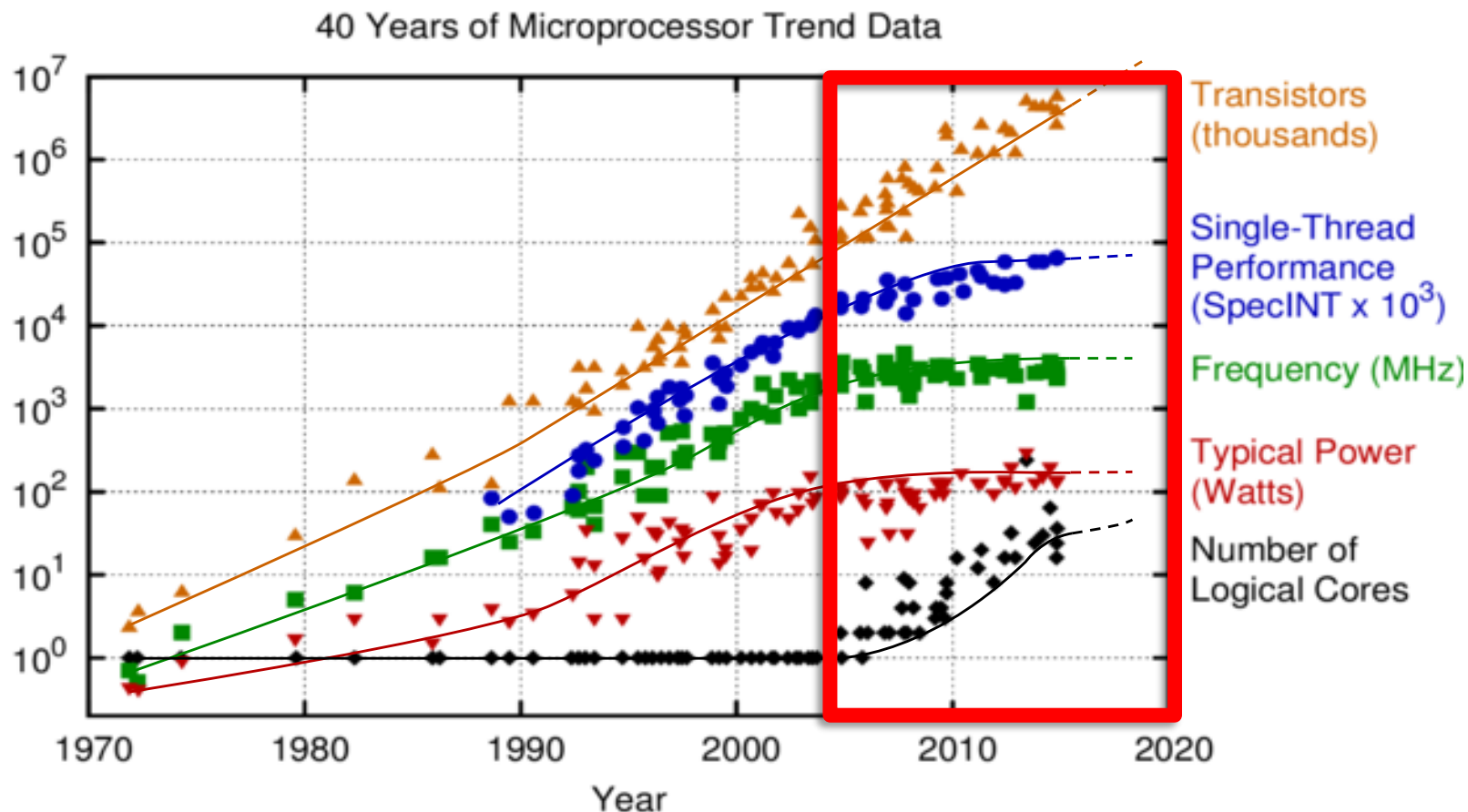
100M Node Hours | 840 Days on Piz Daint

UNDERSTANDING HIV'S STRUCTURE

10M node Hours | 16 Days on BlueWaters

BAD TIMING
THE SLOW DEATH OF MOORE'S LAW

40 Years of Microprocessor Trend Data

Transistors (thousands)

Single-Thread Performance (SpecINT x 10^3)

Frequency (MHz)

Typical Power (Watts)

Number of Logical Cores

Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

NVIDIA.

TOP500 EFFECTS

⬢ nVIDIA.
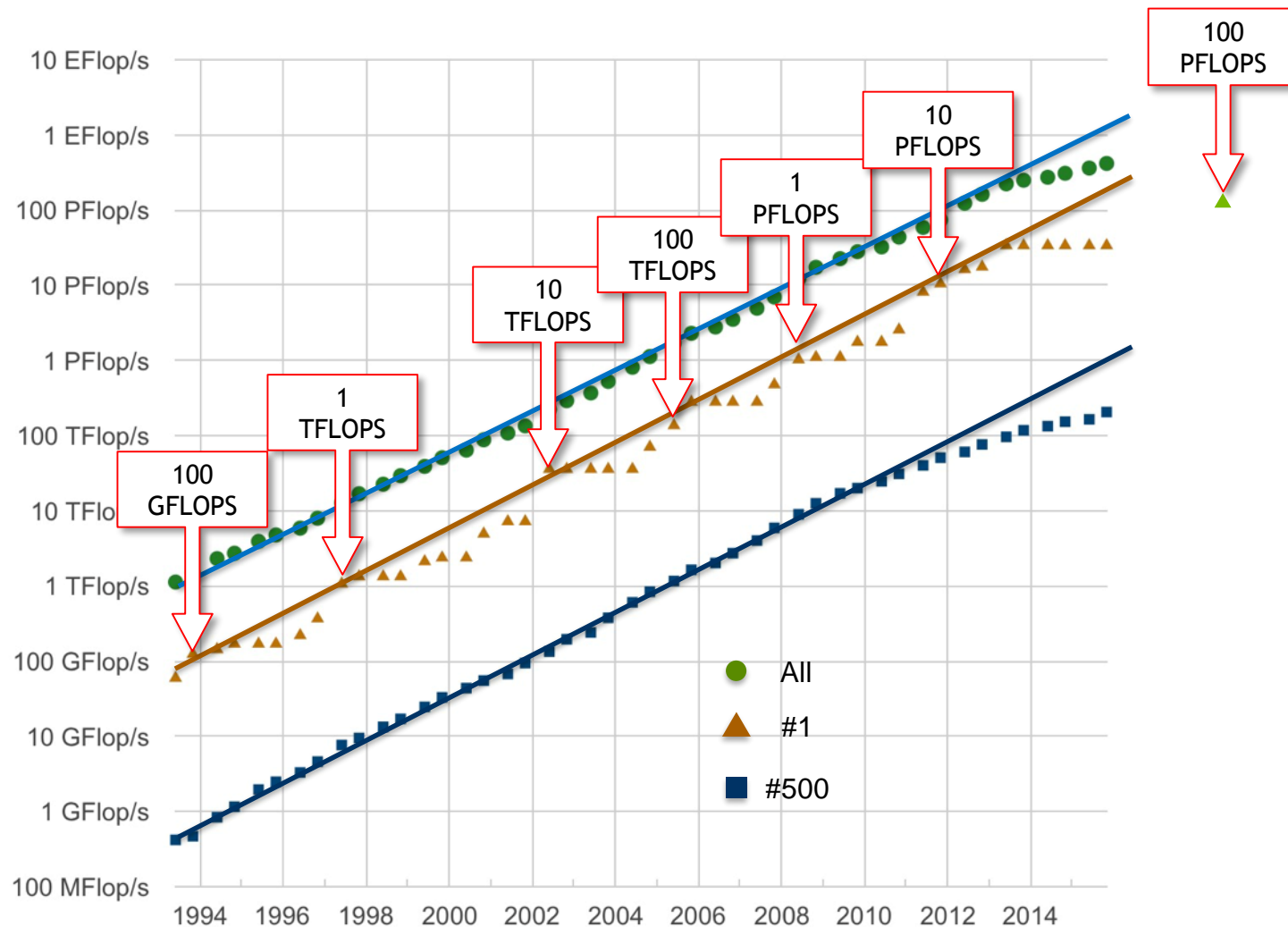
SOMETHING NEW:
AI + HPC = *REVOLUTION*
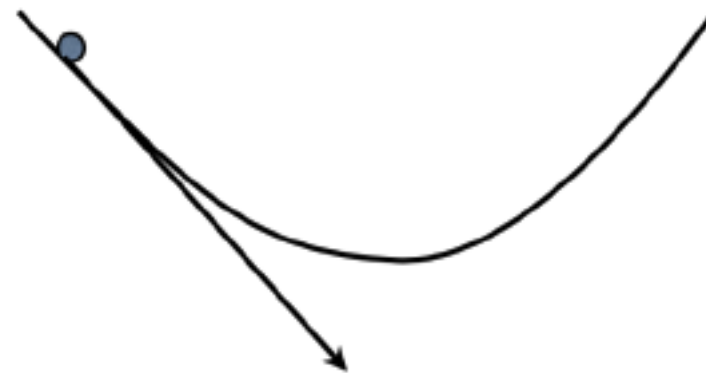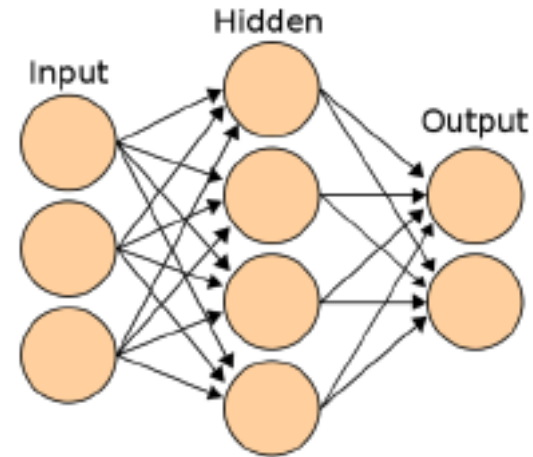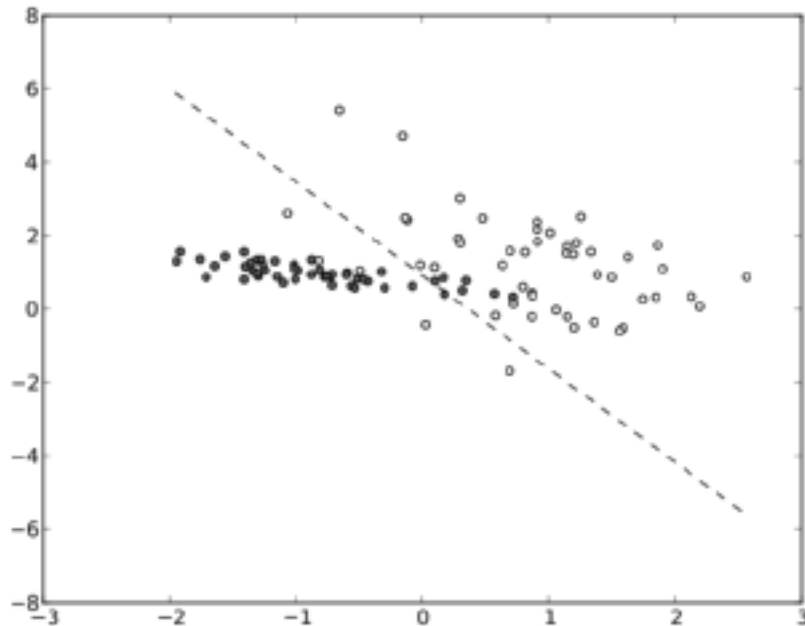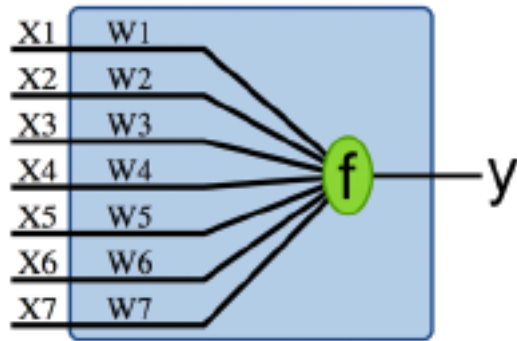
INGREDIENTS: BIG DATA

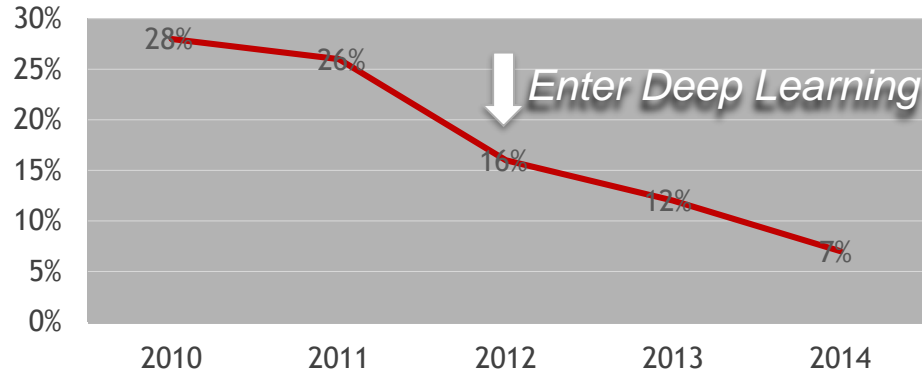(We are the sensors for our cloud service providers)
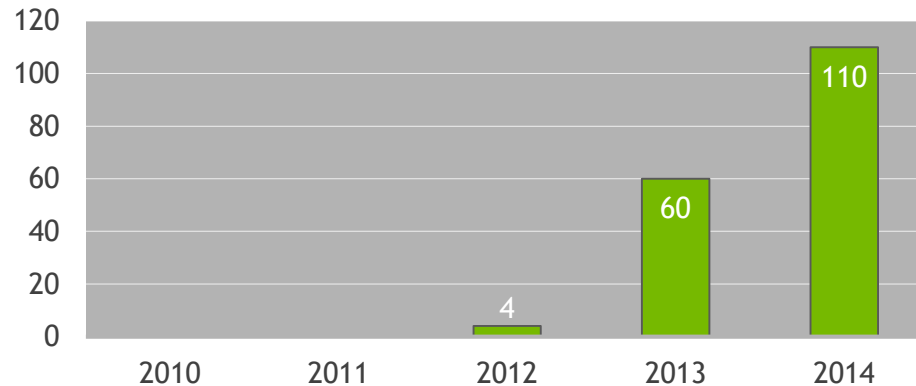
INGREDIENTS: BIG DATA
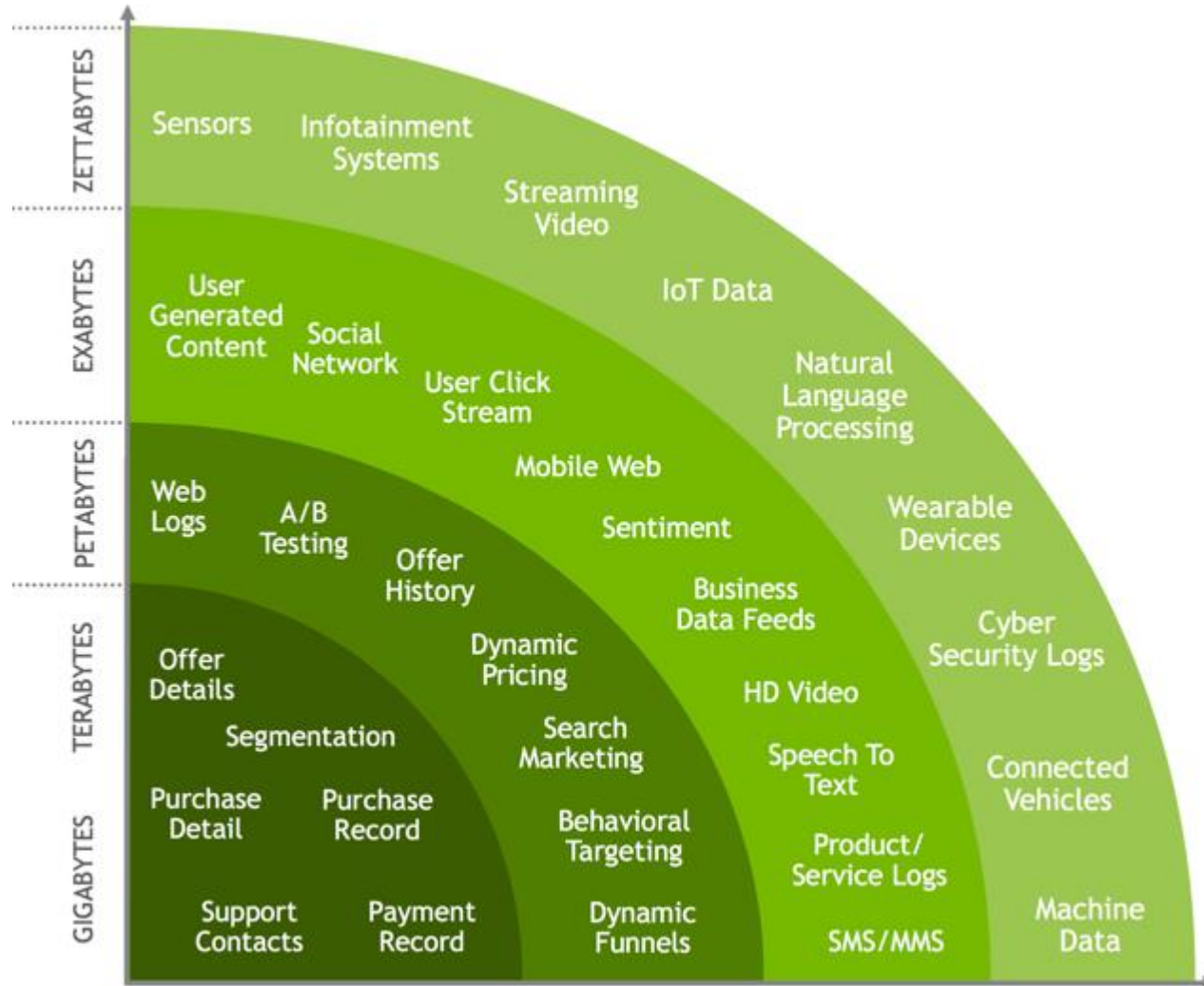
INGREDIENTS: AI ALGORITHMS

NOW, JUST ADD HPC AND STIR...

Classification Error Rates

- 28% (2010)
- 26% (2011)

Enter Deep Learning

- 16% (2012)
- 12% (2013)
- 7% (2014)

GPU Entries

- 4 (2012)
- 60 (2013)
- 110 (2014)

Classification Task:
1.2M images • 1000 object categories

Team	Date	Top-5 Test Error
GoogLeNet	2014	6.66%
Baidu Deep Image	01/12/2015	5.98%
Baidu Deep Image	02/05/2015	5.33%
Microsoft	02/05/2015	4.94%
Google	03/02/2015	4.82%
Baidu Deep Image	03/17/2015	4.83%

Trained Human Performance: 5.1%

IMAGENET

NVIDIA.

ALGORITHMS + BIG DATA + GPUS =
THE BIG BANG OF MODERN AI

Google Photo

Arterys FDA Approved

Reinforcement Learning

AlphaGo

Super Resolution

Deep Voice

Transfer Learning

ImageNet

Stanford & NVIDIA Large-scale DNN on GPU

NVIDIA BB8

Captioning

BRETT

Auto Encoders

Style Transfer

NMT

IDSIA CNN on GPU

U Toronto AlexNet on GPU

Baidu DuLight

LSTM

Superhuman ASR

GAN
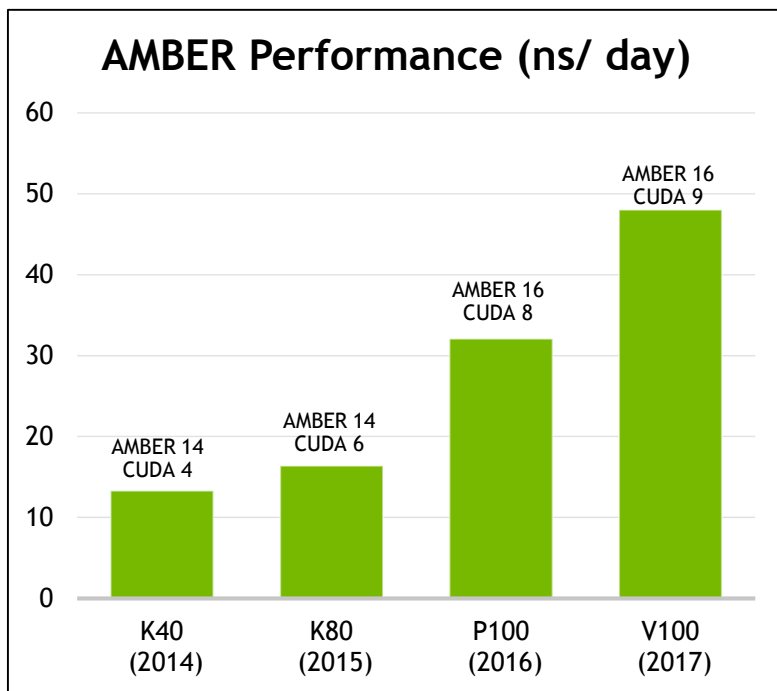
recognition/classification -> recursion/time series -> generative

AI: A NEW COMPUTING PARADIGM
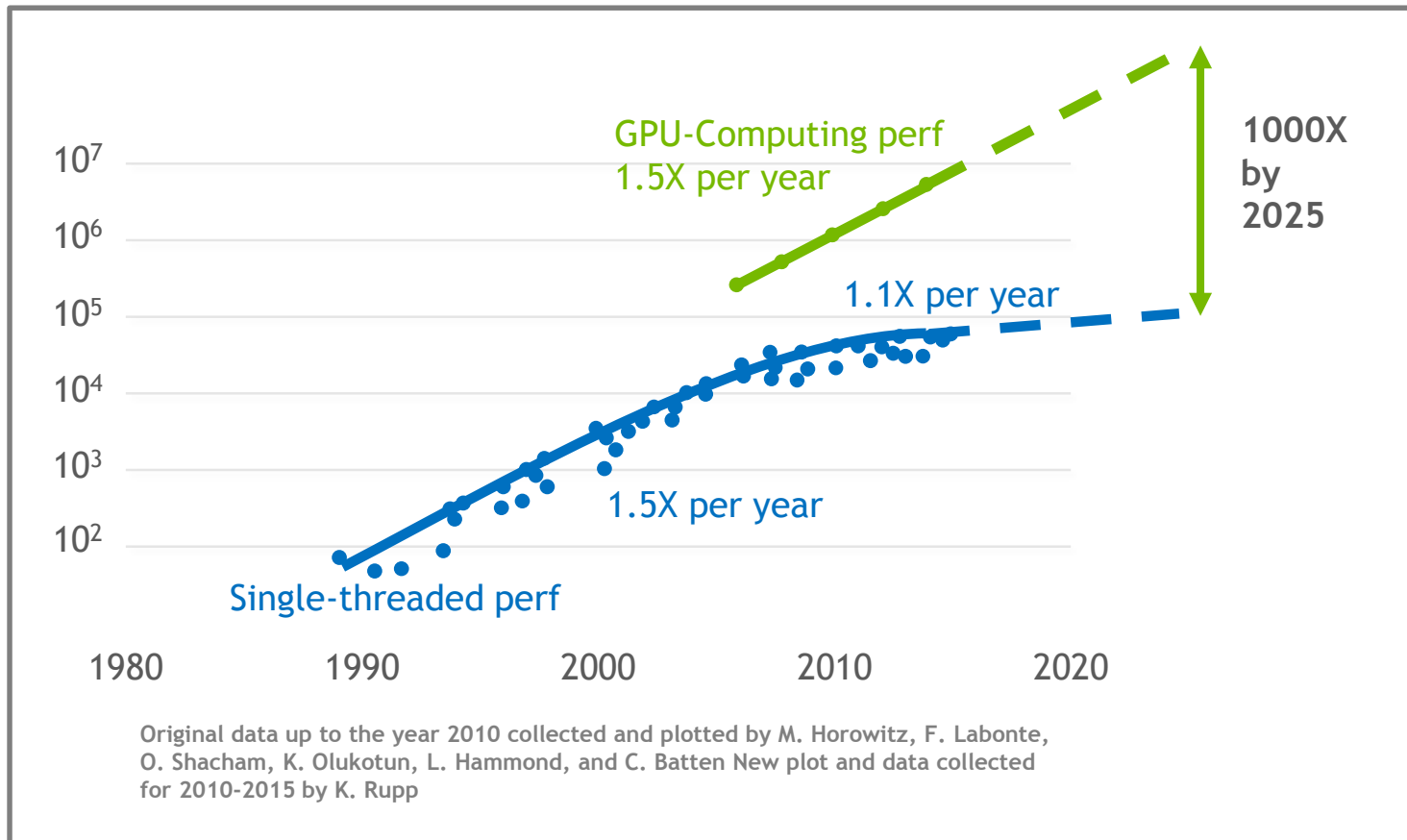
GPU COMPUTING PERFORMANCE OVER TIME

AI on a super-Moore's Law progression

AMBER Performance (ns/ day)

- K40 (2014): AMBER 14 CUDA 4
- K80 (2015): AMBER 14 CUDA 6
- P100 (2016): AMBER 16 CUDA 8
- V100 (2017): AMBER 16 CUDA 9

4x in 3 years

GoogleNet Performance (i/s)

- 8X K80 (2014): cuDNN 2 CUDA 6
- 8X MAXWELL (2015): cuDNN 4 CUDA 7
- DGX-1 (2016): cuDNN 6 CUDA 8 NCCL 1.6
- DGX-1V (2017): cuDNN 7 CUDA 9 NCCL 2
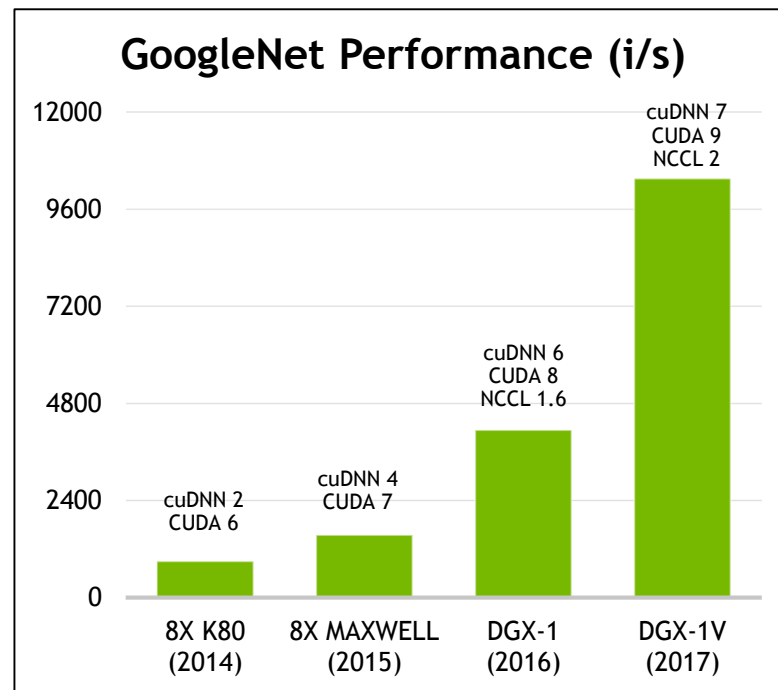
12x in 3 years
(65x in 5 years)

Amber dataset: Cellulose NVE; GoogLeNet dataset: Imagenet

2018: 10X AI GAIN IN ONE YEAR

DGX-1, SEP'17

DGX-2, Q3'18

15 days

1.5 days

| | DGX-1V | DGX-2 |
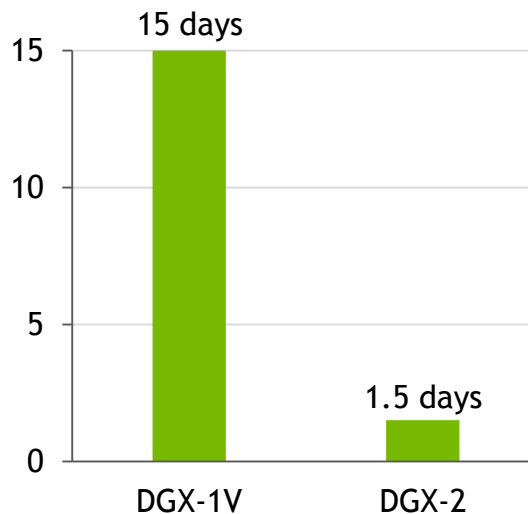
PyTorch Stack: Time to Train FAIRSEQ

software improvements across the stack including NCCL, cuDNN, etc.
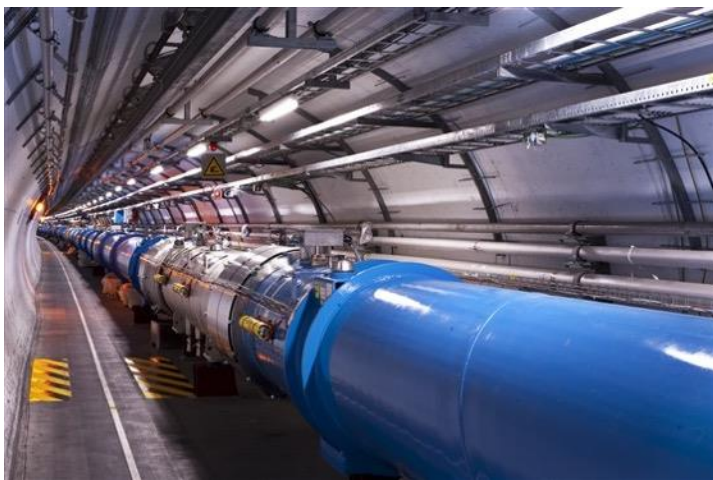
NVIDIA.

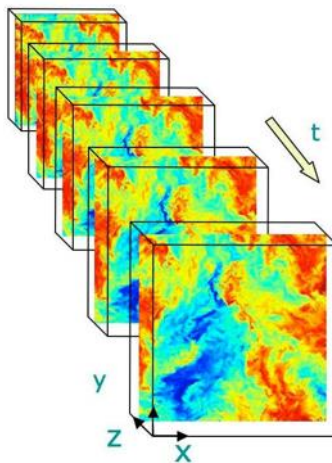# WHAT DOES THIS HAVE TO DO WITH *SCIENCE?*

(HPC + AI = ?)

BIG DATA IN SCIENCE

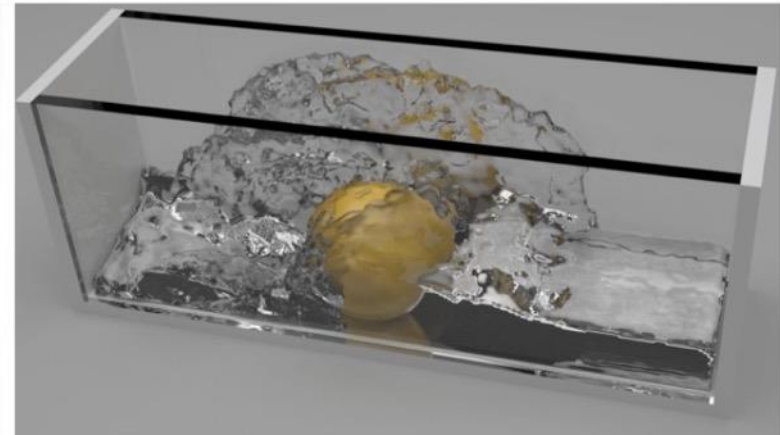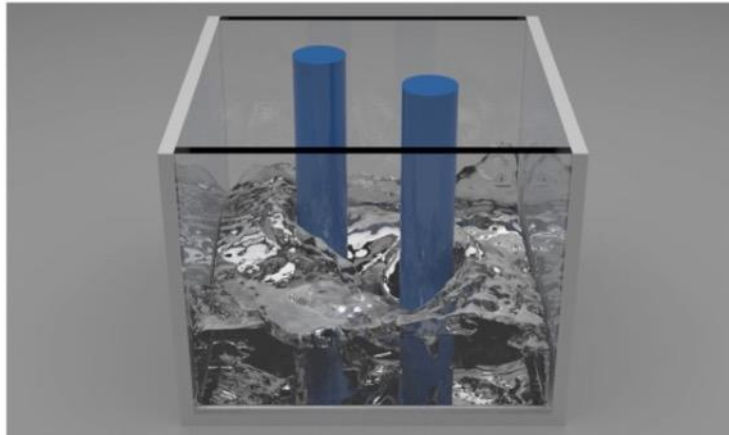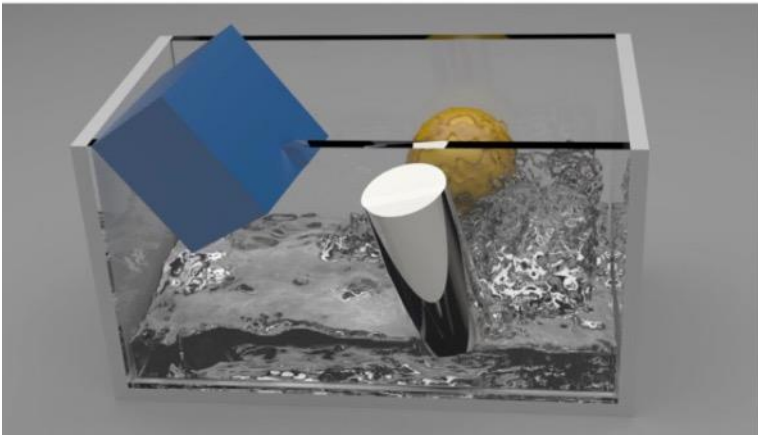## Big Science ingests/outputs Big Data
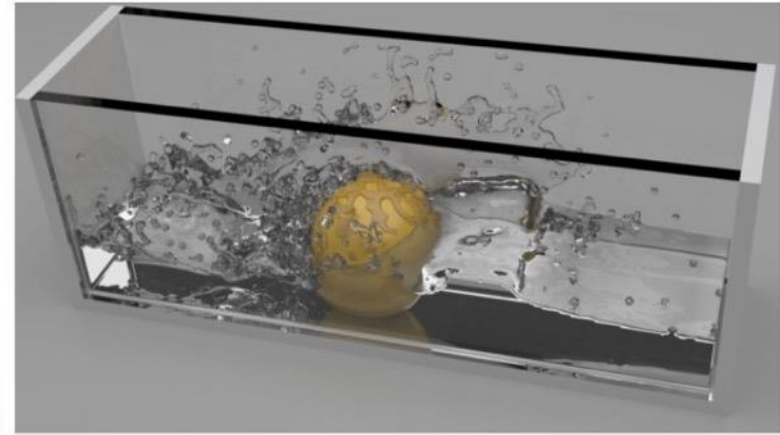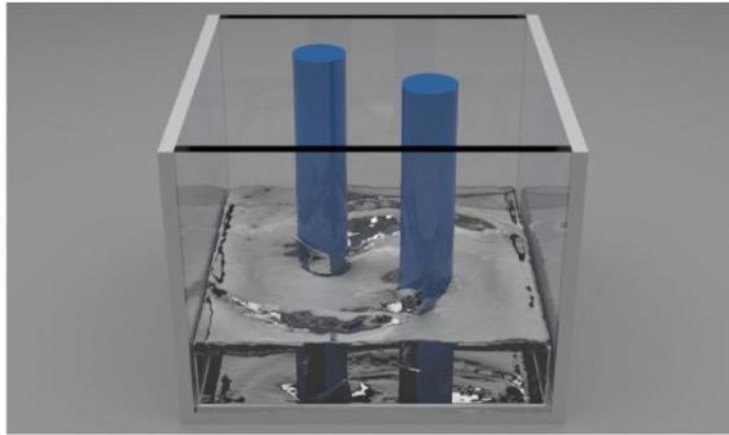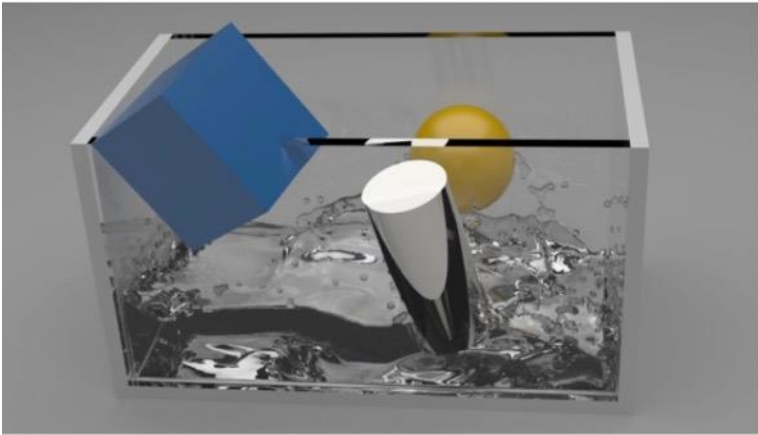

Large Hadron Collider
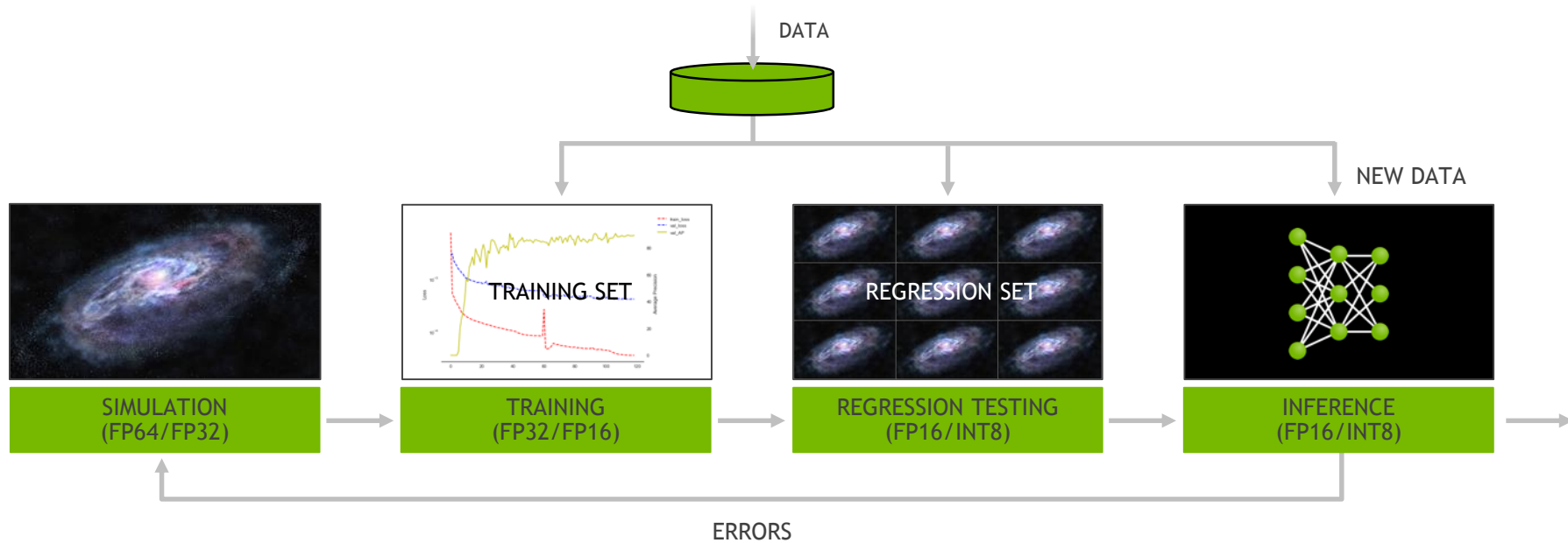
Johns Hopkins
Turbulence Database

Square Kilometer Array

2015: USING NUMERIC SIMULATIONS TO TRAIN AI
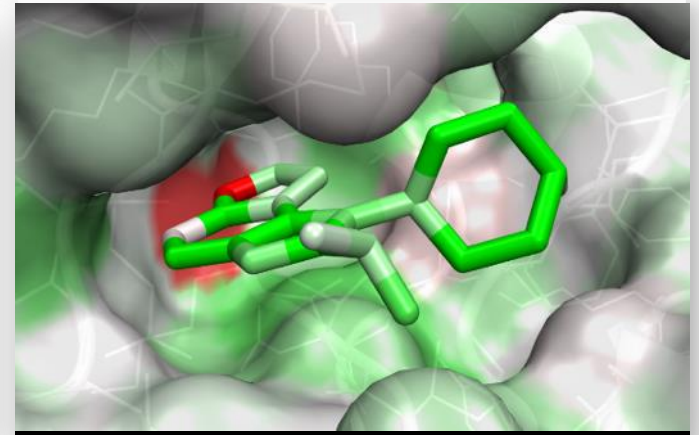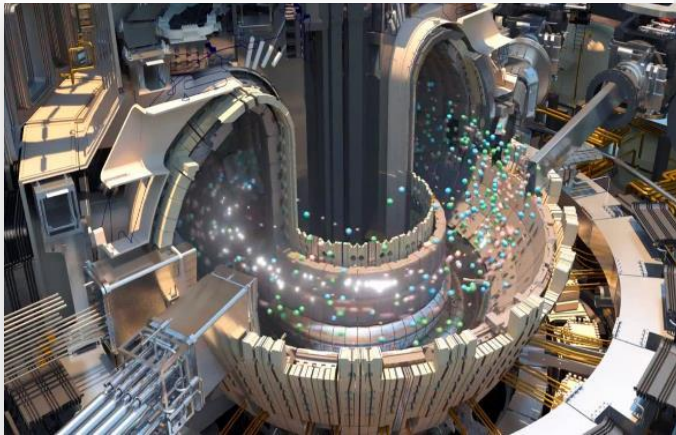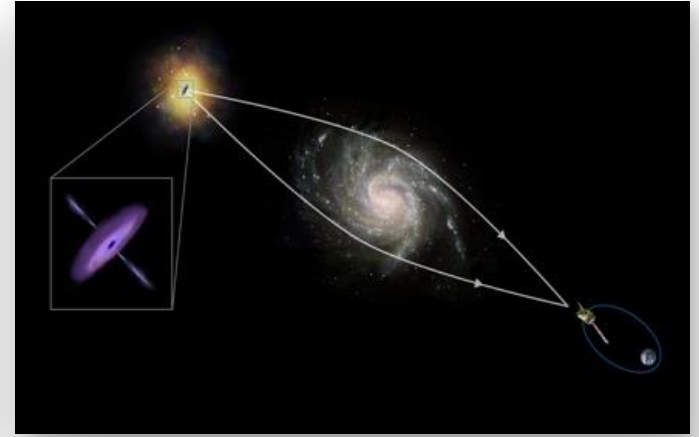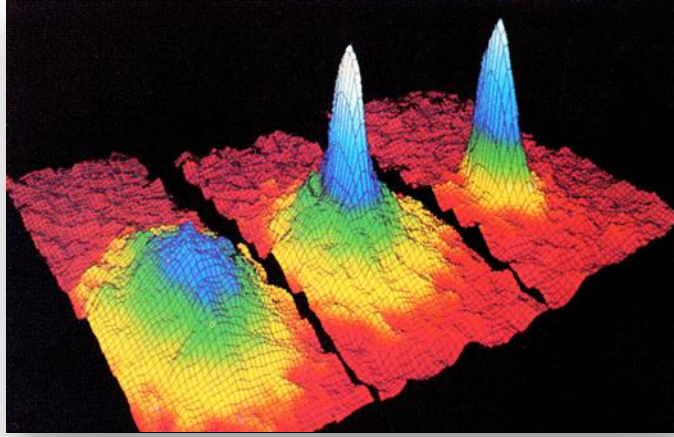
2015: USING NUMERIC SIMULATIONS TO TRAIN AI

AI WORKFLOW FOR HPC

DATA

NEW DATA

TRAINING SET

REGRESSION SET

SIMULATION (FP64/FP32)	TRAINING (FP32/FP16)	REGRESSION TESTING (FP16/INT8)	INFERENCE (FP16/INT8)

ERRORS

DEEP LEARNING COMES TO HPC

APPROXIMATING QUANTUM CHEMISTRY

Background
Developing a new drug costs $2.5B and takes 10-15 years. Quantum chemistry (QC) simulations are important to accurately screen millions of potential drugs to a few most promising drug candidates.

Challenge
QC simulation is computationally expensive so researchers use approximations, compromising on accuracy. To screen 10M drug candidates, it takes 5 years to compute on CPUs.

Solution
Researchers at the University of Florida and the University of North Carolina leveraged GPU deep learning to develop ANAKIN-ME, to reproduce molecular energy surfaces with super speed (microseconds versus several minutes), extremely high (DFT) accuracy, and at 5-6 orders of magnitude lower cost.
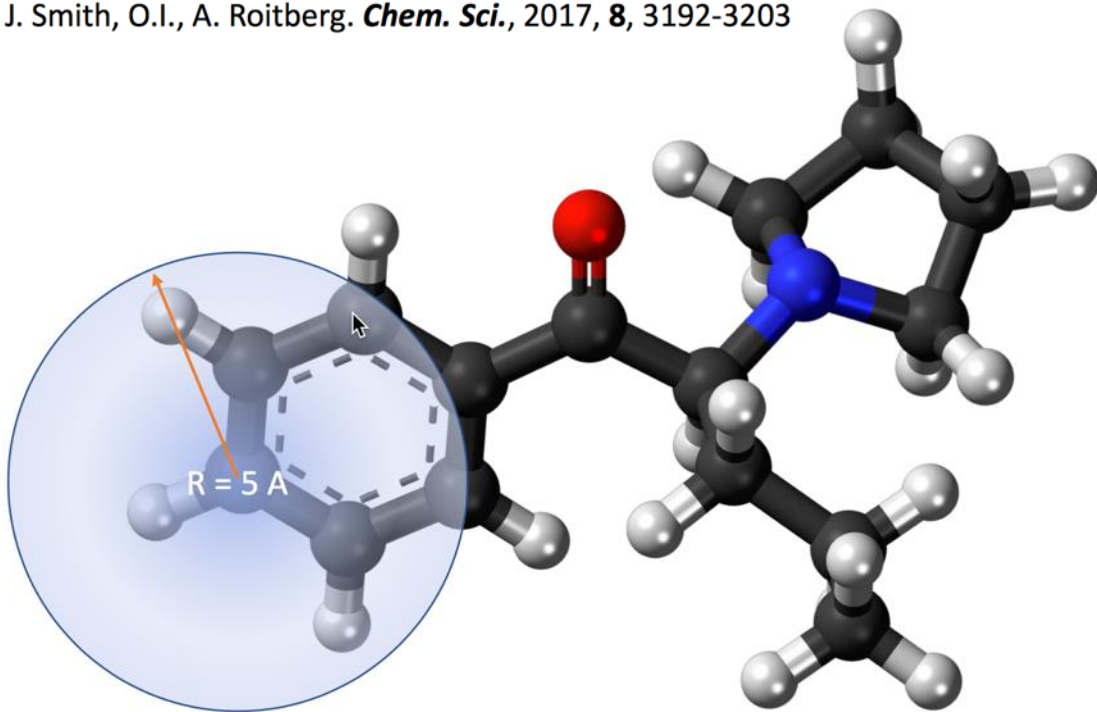
Impact
Faster, more accurate screening at far lower cost
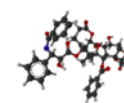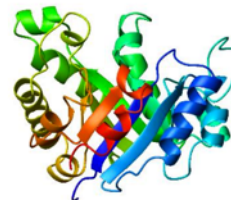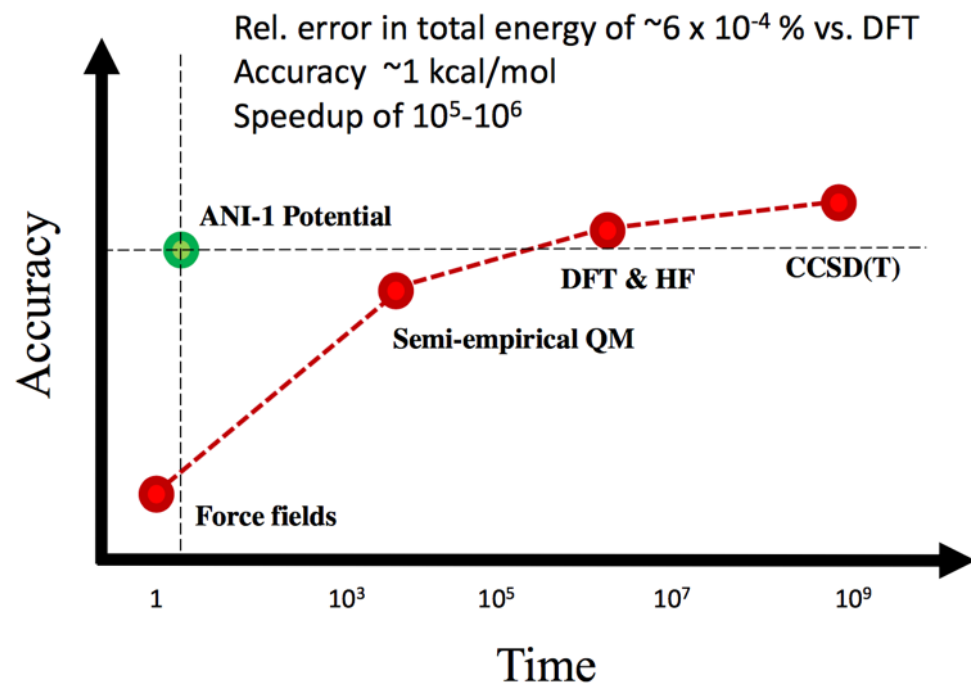
UF | UNIVERSITY of FLORIDA

NEURAL NETWORK MODEL APPROACH

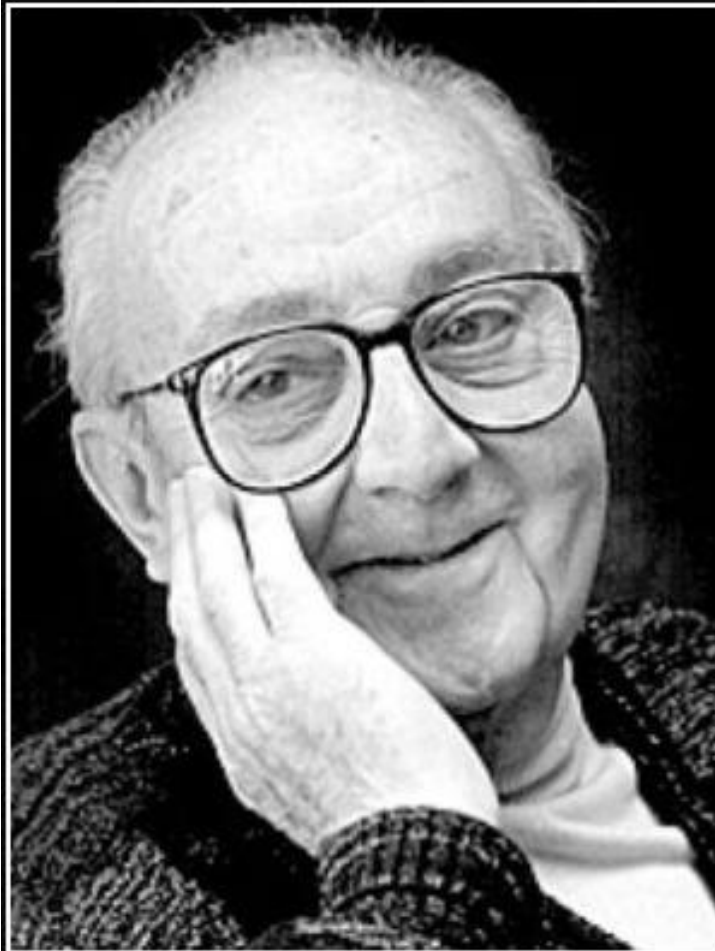J. Smith, O.I., A. Roitberg. *Chem. Sci.*, 2017, **8**, 3192-3203

R = 5 A

Rel. error in total energy of ~6 x 10^{-4} % vs. DFT
Accuracy ~1 kcal/mol
Speedup of 10^5-10^6

ANI-1 Potential

Accuracy

DFT & HF

CCSD(T)

Semi-empirical QM

Force fields

1   10^3 10^5 10^7 10^9

Time

Training set: ~20M DFT data points.
Molecules with 1 to 8 atoms from GDB database

Accessible molecular systems
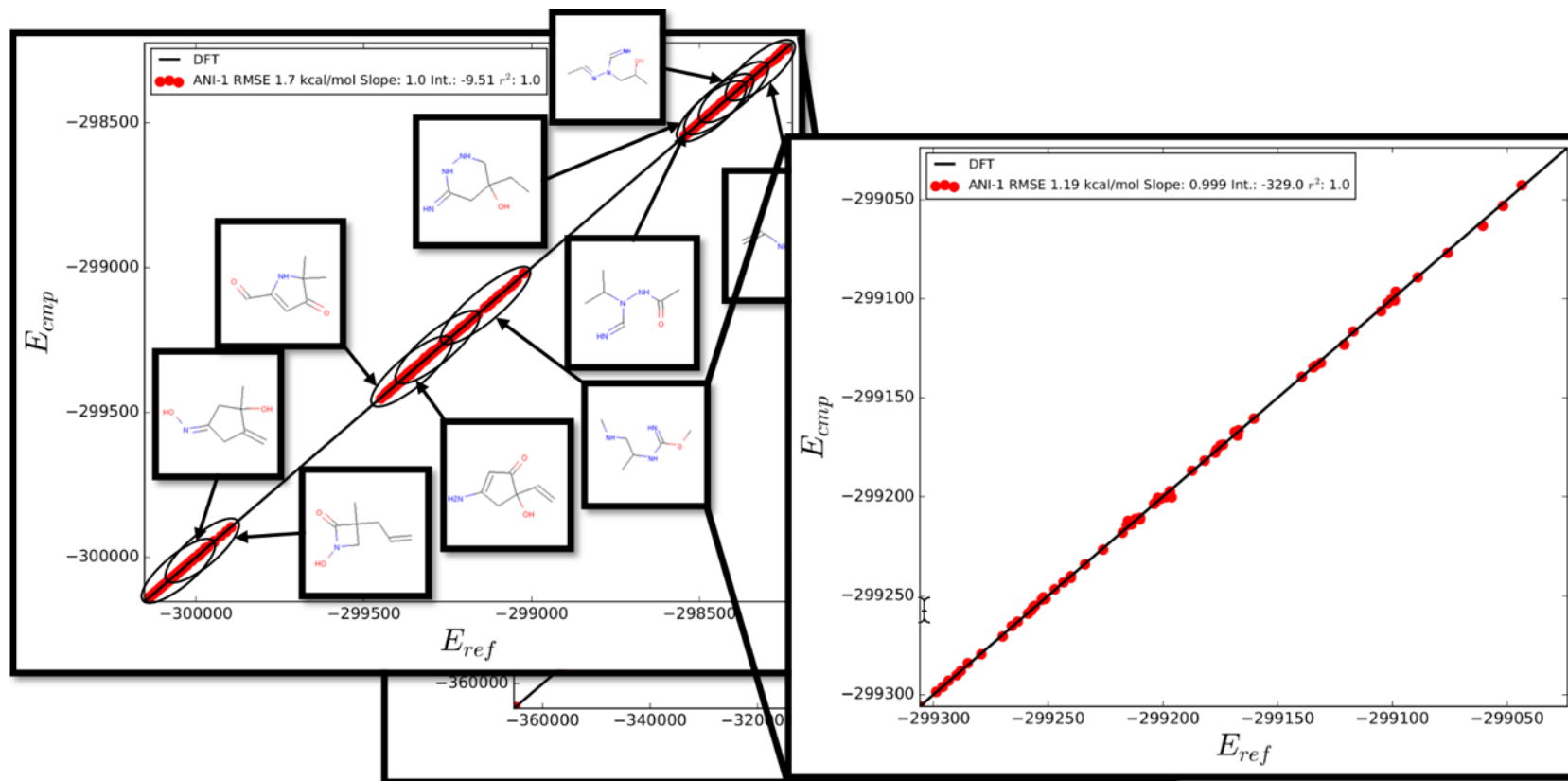
CORRELATION VALIDATION

All models are wrong, but some are useful.

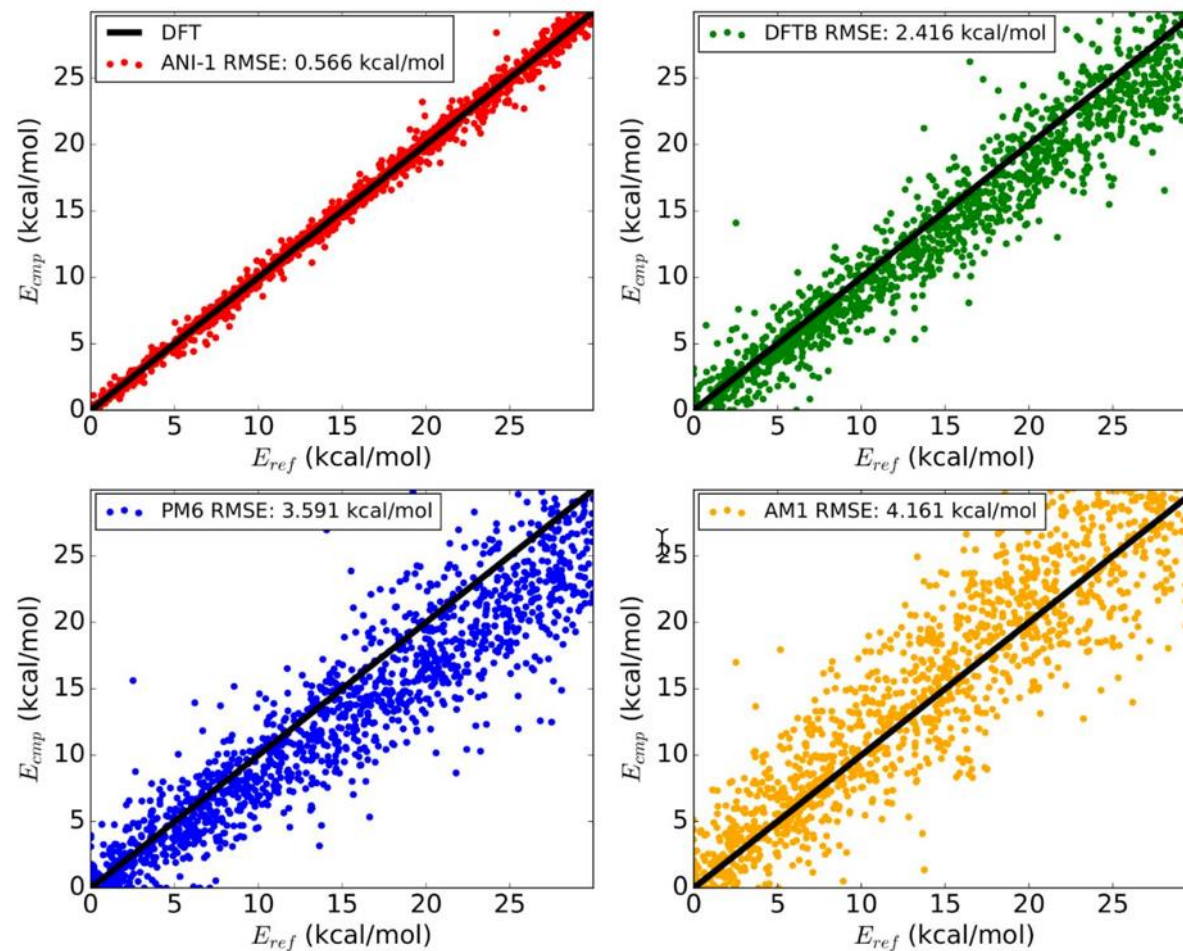— *George E. P. Box* —

AZ QUOTES

CORRELATION VALIDATION

Total energy correlation
ANI-1 vs. DFT
(131 molecules with 10 heavy atoms, 8200 total molecules + conformations) [units: kcal/mol]

J. Smith, O.I., A. Roitberg. *Chem. Sci.*, 2017, **8**, 3192-3203

CORRELATION VALIDATION

J. Smith, O.I., A. Roitberg. *Chem. Sci.*, 2017, **8**, 3192-3203

THE SHAPE OF AI SUPERCOMPUTING

DESIGNED FOR AI SUPERCOMPUTING

Piz Daint

Europe's Fastest Supercomputer

19.6 PetaFLOPS

10.4 GF/W

Powered by Tesla P100

Largest simulation of Universe along with ORNL Titan

Set DL Scaling record with Microsoft Cognitive Toolkit

DESIGNED FOR AI SUPERCOMPUTING

Tsubame 3

#1 Green500 System, June 2017

Powered by 2160 P100s
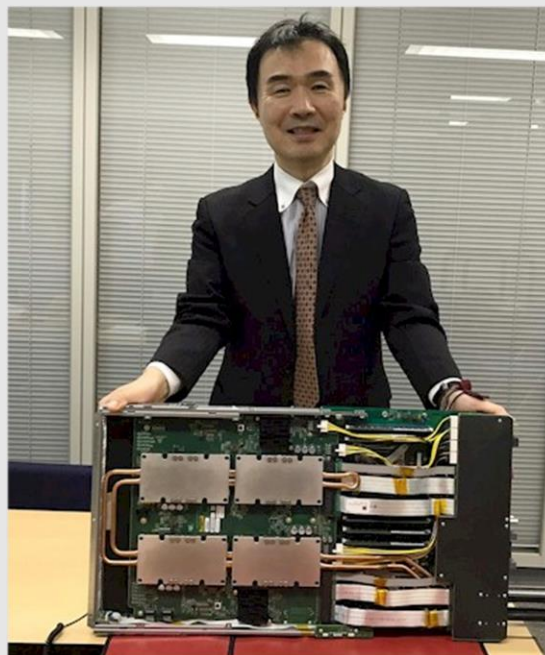
Tokyo Tech's Tsubame 3 will be AI/HPC hybrid

20 February 2017 | By Peter Judge

DatacenterDynamics
The Business of Data Centers

JAPAN KEEPS ACCELERATING WITH TSUBAME 3.0 AI SUPERCOMPUTER

February 17, 2017 | Timothy Prickett Morgan

THENEXTPLATFORM

Next-Generation TSUBAME Will Be Petascale Supercomputer for AI

Michael Feldman | February 18, 2017 00:04 CET

TOP500
The List

"NVIDIA's broad AI ecosystem will enable Tokyo Tech to begin training TSUBAME3.0 immediately to help us more quickly solve some of the world's once unsolvable problems."

- Satoshi Matsuoka, Prof Computer Science, TiTech & Project lead Tsubame 3

DESIGNED FOR AI SUPERCOMPUTING
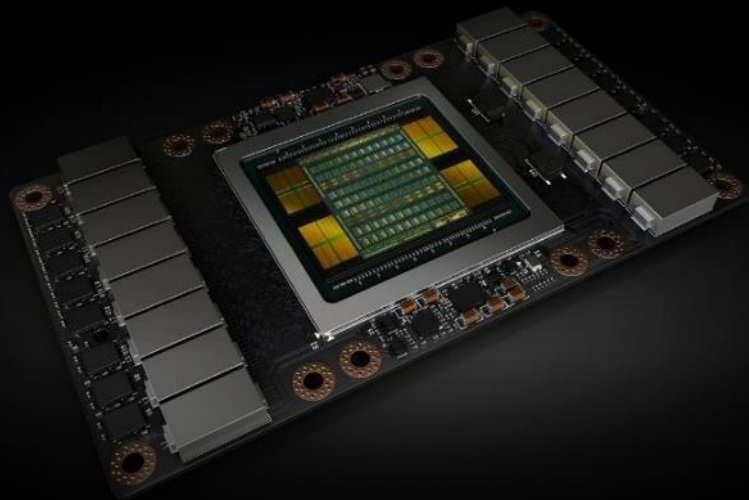
Most Powerful AI Supercomputer in Japan

4,352 Tesla V100 GPUs

37 PetaFLOPS FP64 HPC Performance

0.55 ExaFLOPS AI Performance

ABCI SUPERCOMPUTER TO BE
INSTALLED AT THE KASHIWA II CAMPUS OF
THE UNIVERSITY OF TOKYO

NVIDIA.

VOLTA TENSOR CORE GPU FUELS WORLD'S FASTEST SUPERCOMPUTER

Fused HPC and AI Computing In a Unified Platform

Genomics (CoMet)
World's First Exascale Run
Finding Genes-to-disease Connection
Same accuracy as FP64 w/ Tensor Core

150X
Over Titan

Quantum Chemistry (QMCPack)
Simulate New Materials
High-Temperature Semiconductors

50X
Over Titan

Summit Supercomputer   AI: 3 Exaflops
Oakridge National Labs   HPC: 122 Petaflops

Measured performance: Summit node vs Titan node

AI SUPERCOMPUTING IS HERE

Extending The Reach of HPC By Combining Computational & Data Science
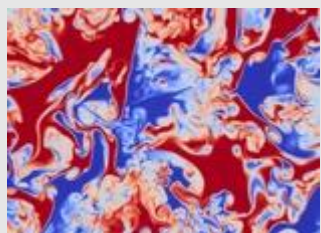
Turbulent Flow
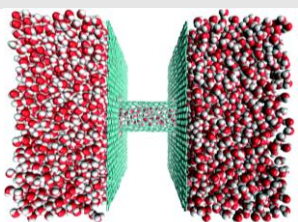
Molecular Dynamics
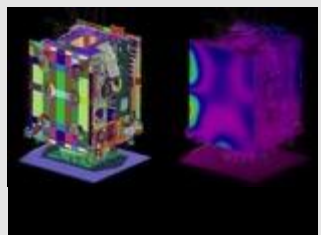
Structural Analysis

N-body Simulation

COMPUTATIONAL SCIENCE

"What's happening?"
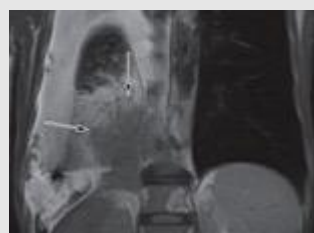
"Is there cancer?"

"Next move?"

"What does she mean?"

DATA SCIENCE

Drug Discovery

Clean Energy

Understanding Universe

Monitoring Climate Change

COMPUTATIONAL & DATA SCIENCE

GTC EUROPE
9-11 October MUNICH
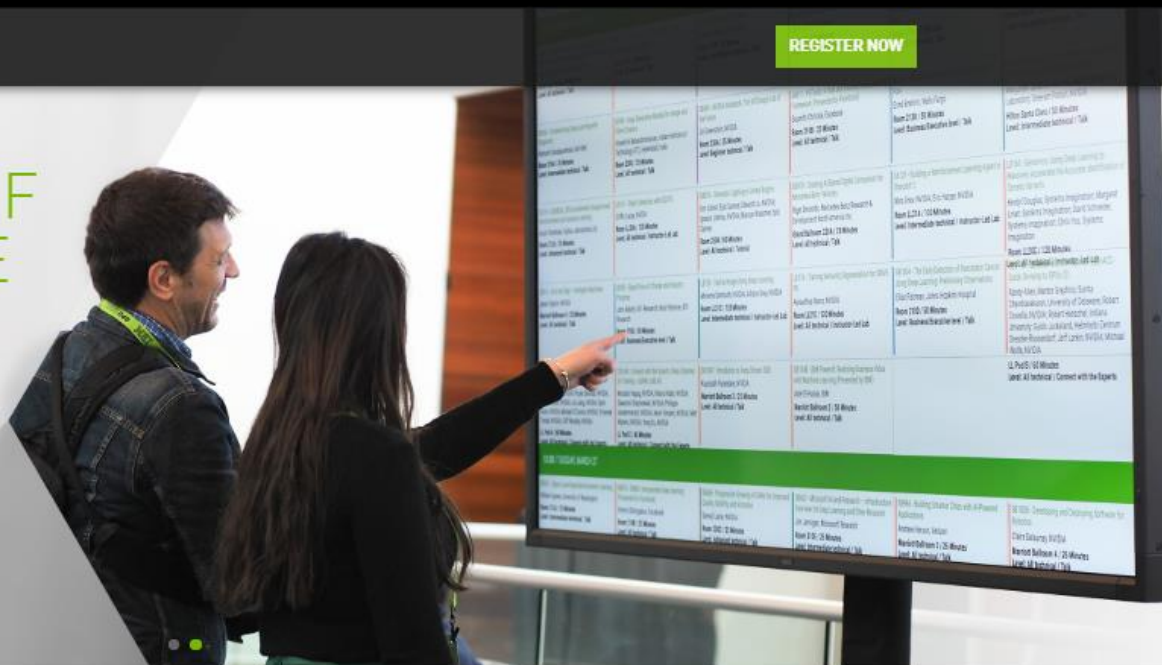
DEEP LEARNING INSTITUTE

NVIDIA.

NVIDIA RESEARCH

THIS IS ONLY THE BEGINNING...

THANK YOU