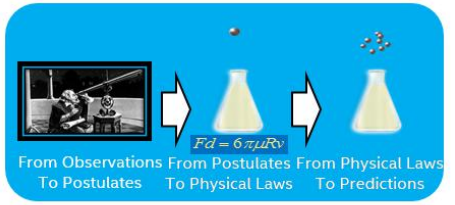
Progress in HPC + AI Convgence

Victor Lee
senior manager for business development and alliances
Intel Corporation

THE NEW CENTER OF POSSIBILITY

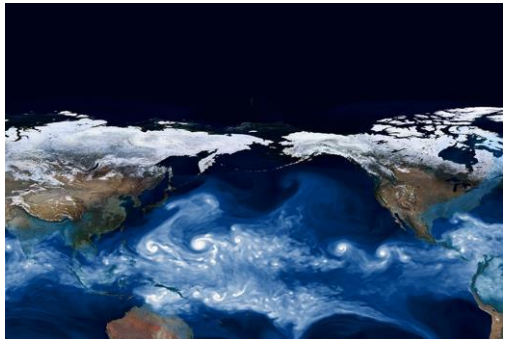
HIGH PERFORMANCE COMPUTING (HPC)



Traditional HPC models and simulates theoretical science through first principle or experimentally. Human is involved in every step.

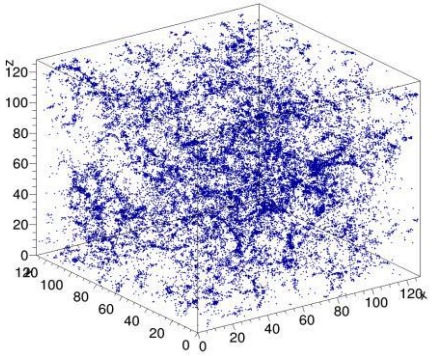
Human Centric

Climate Modeling

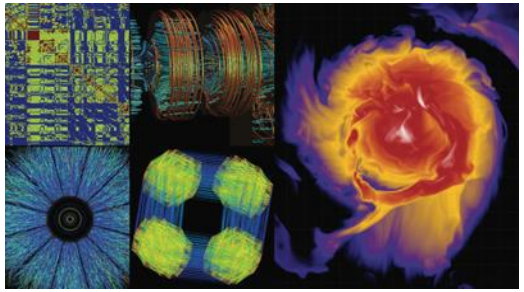


Weather

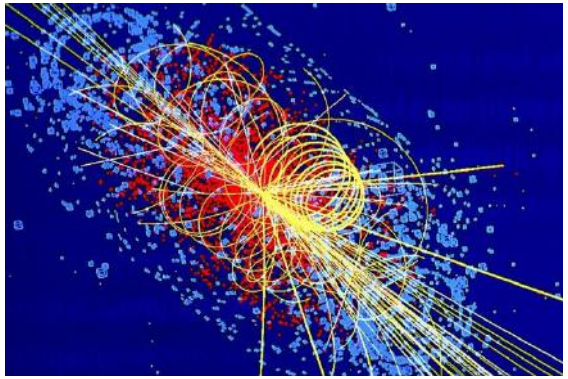
Cosmology



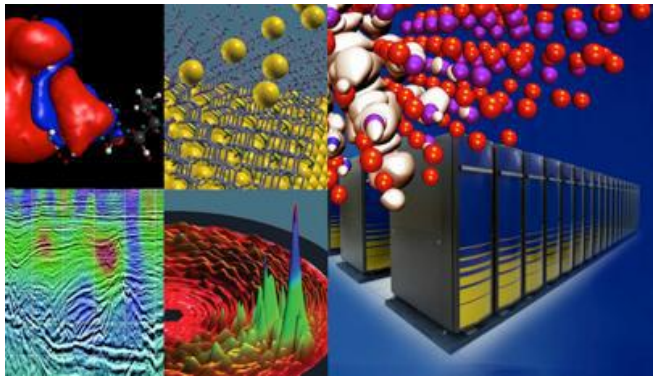
Nuclear Physics



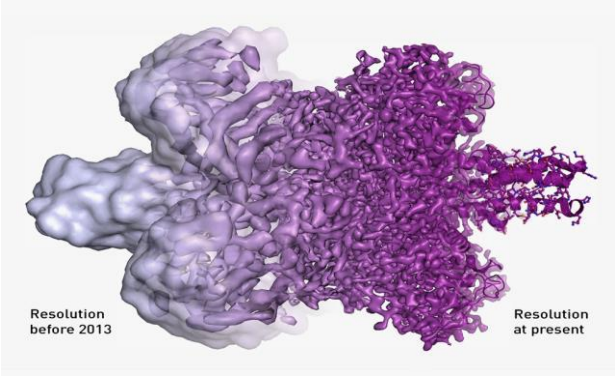
Fundamental Physics



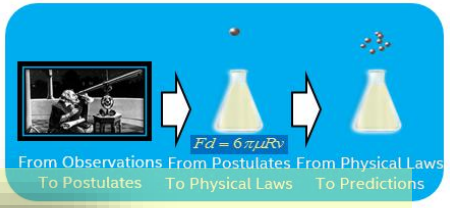
Material Simulations



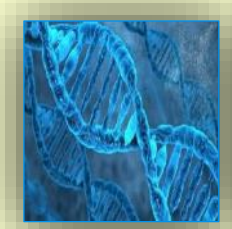
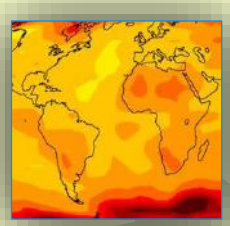
Drug Discovery



EVOLUTION OF COMPUTER AS A TOOL



HPC


Human Centric



Models & Simulations

Supercomputers

Spreadsheet

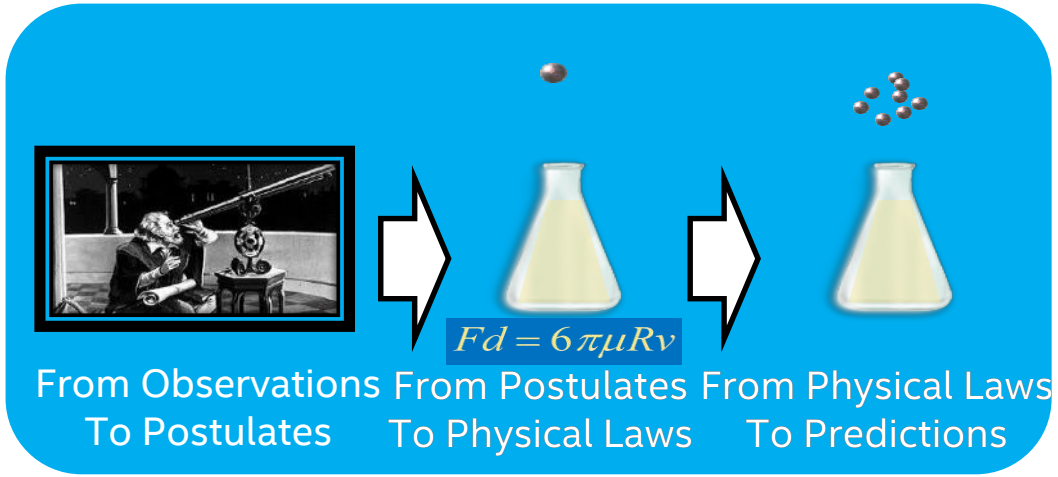
1983 IBM PC XT

Simple Arithmetic

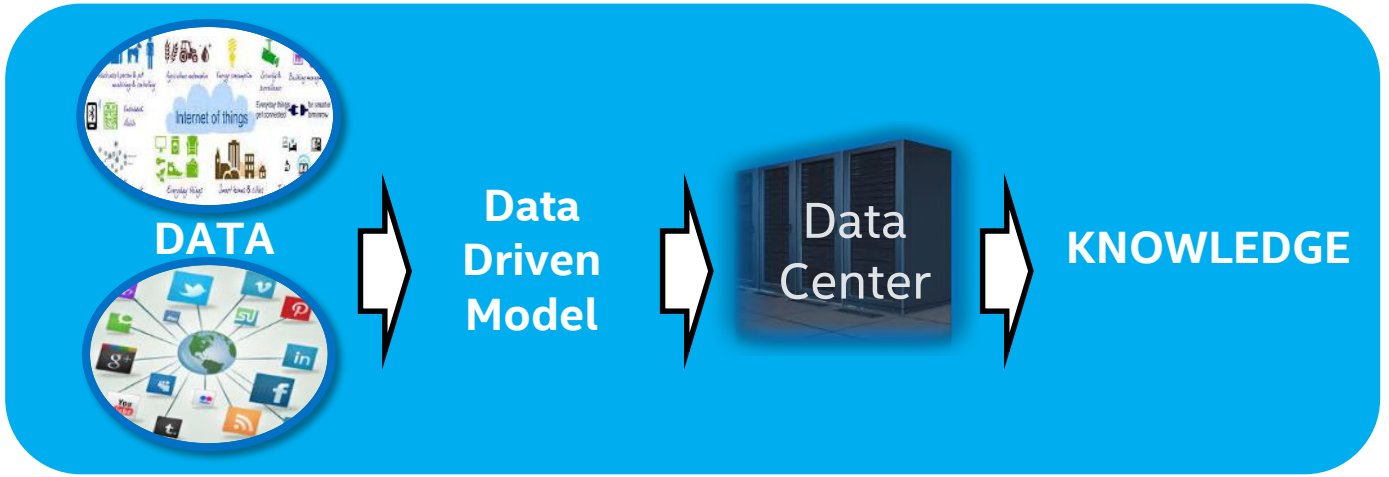
1837 Analytical Engine

DATA ANALYTICS SPEEDS SCIENTIFIC DISCOVERY

Traditional



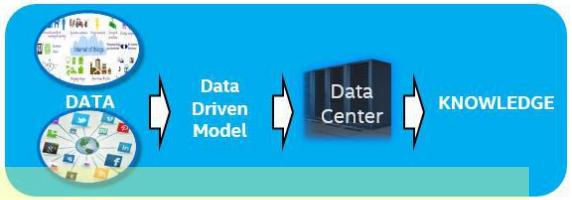
Emerging



Human Centric

Machine Centric

EVOLUTION OF COMPUTER AS A TOOL (NEW FRONTIER)



HPC

AI

Machine Centric

Model
Discovery
Synthesis

Models &
Simulations

Spreadsheet

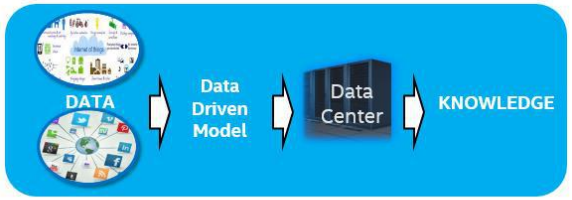
Simple
Arithmetic

Supercomputers

1983 IBM PC XT

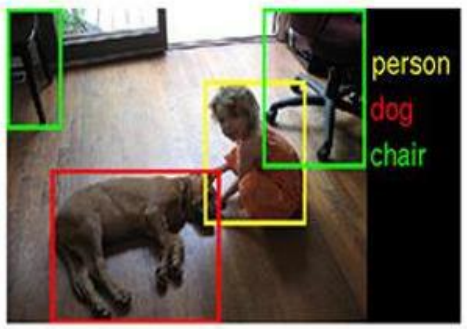
1837 Analytical Engine

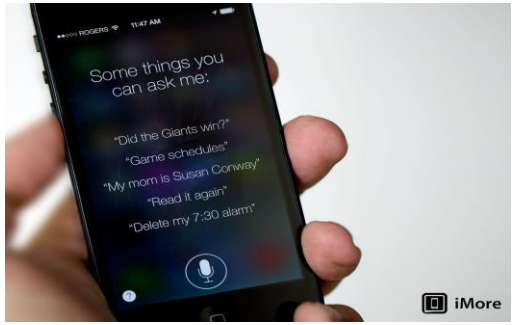
ARTIFICIAL INTELLIGENCE (AI)



Machine Centric

AI (especially in the form of automatic model construction) is being adopted to replace human for tedious manual tasks.

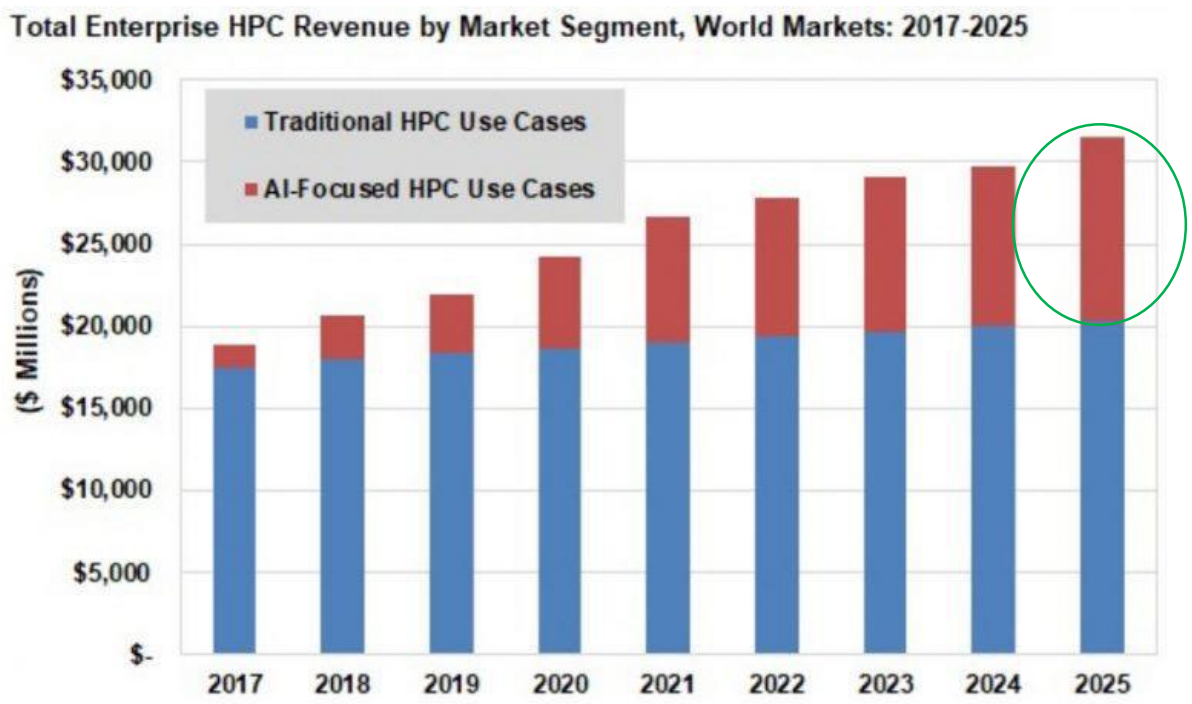




HPDA / HPC + AI = NEW GROWTH FOR HPC

Latest round of AI revolution driven by affordable/ubiquitous compute and vast amount of data.

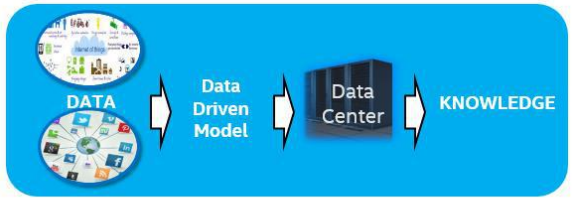
AI is being adopted by all business.



World HPC Revenue (2017-2025) (Source: Tractica)

HPC+AI is becoming the driving-force for HPC growth (~40% HPC market by 2025)

AI FOR SCIENCE



Machine Centric

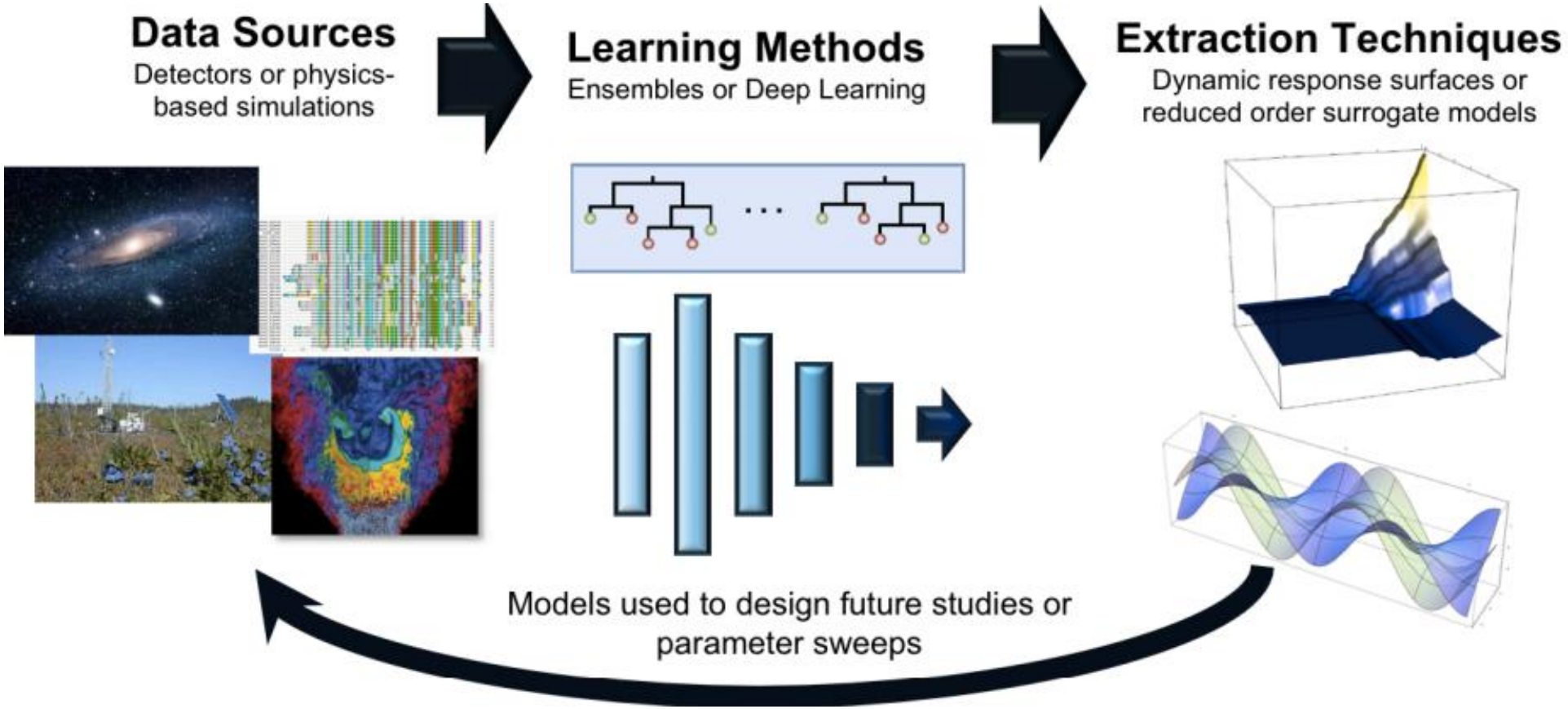
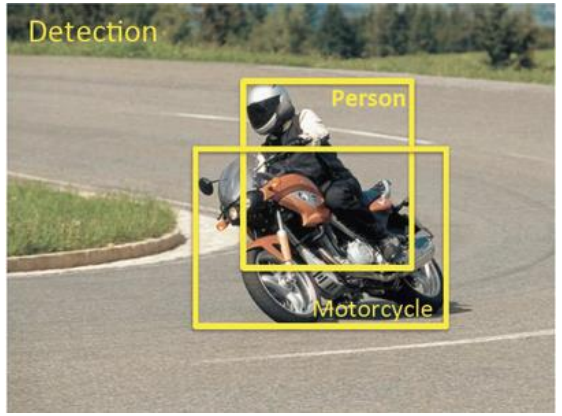


Diagram from J. Brown presentation, AI at Scale in Biology, AI for Science, September 2019

WHAT DOES HPC+AI CONVERGED WORKLOAD LOOK LIKE

+



=

?

TWO PATHS OF HPC + AI: HPDA VS. AI IN MOD/SIM

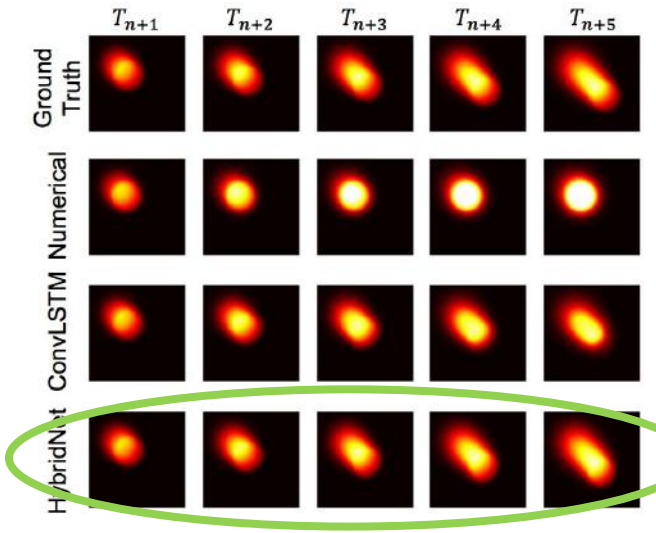
High Perf. Data Analytics:
Harvest patterns / insights from data

Statistical methods /
ML model

Abnormality Detection (e.g. frauds)
Recognition (e.g. face, voice, tumor, terrorist, etc.)
Recommendations (e.g. what movies you may like)

AI in Mod/Sim:
Use of ML models to speedup modeling / simulation

Example: Heat diffusion for manufacturing



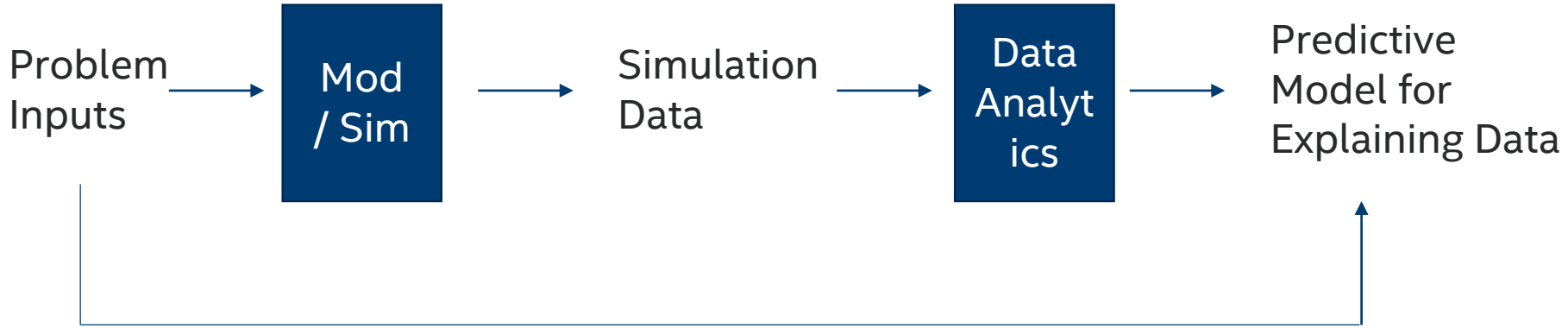
100x speedup with similar accuracy

$$\frac{\partial T_{xy}}{\partial t} = C \cdot (T_{external} - T_{xy}) + K \cdot \Delta T_{xy}$$

* "HybridNet: Integrating Model-based and Data-driven Learning to Predict Evolution of Dynamical Systems " https://arxiv.org/pdf/1806.07439.pdf

* Majority of HPC+AI today

WAYS IN COMBINING HPC AND ML (1: HPDA)



- Machine Learning is applied in TANDEM to model / simulation to IDENTIFY underlying data patterns
  - Mod/Sim and Data Analytics run separately and use different software stacks
- Train on observed / simulation data (e.g. weather, molecular simulation, etc.)
  - Data is stored in IO. Potentially an IO problem.

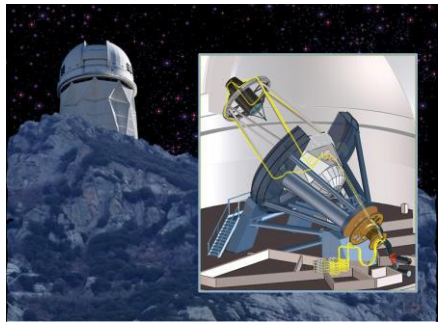
Key Challenges: different management stack, Data Management / IO

HPDA EXAMPLE IN COSMOLOGY (COSMOFLOW @ SC'18)

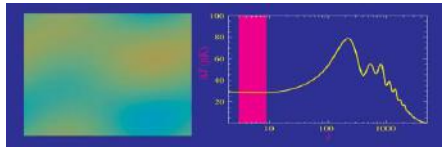
Problem: Finding the parameters that govern the universe expansion from the Big Bang

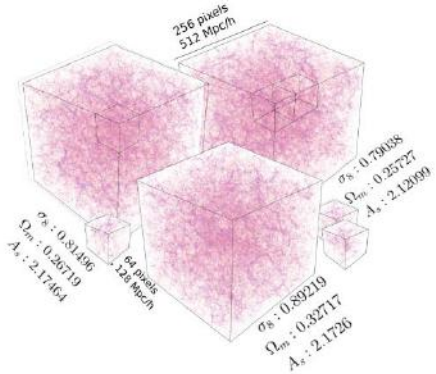
Traditional HPC approach

1) Scientists build universal models.
2) Models are simulated and correlated w/ observations.



Reduced Statistics

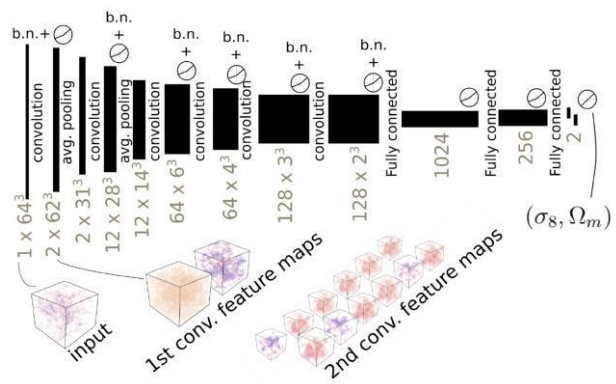


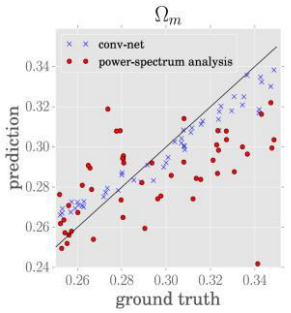


HPDA approach

1) Train 3D CNN to create a model correlates parameters and the generated models.
2) Apply model to infer observations.

Performance:
~1000x speedup on optimized IA framework
~6M times speedup with 8192 nodes
(Reduce training time from 3M to < 10min)

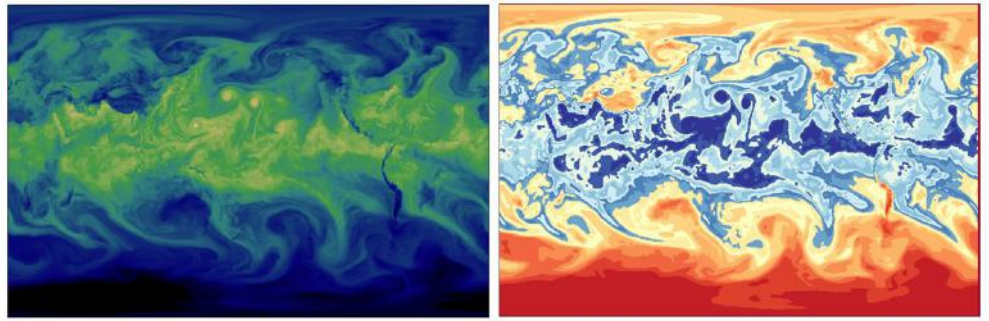




Mathuriya, Amrita, et al. "CosmoFlow: using deep learning to learn the universe at scale." *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2018.

HPDA EXAMPLE IN WEATHER (DISCO @ SC'19)

DisCo – Unsupervised Detection of Spatiotemporal Structure
Collaboration w/ UC Davis and NERSC (2018-2019)

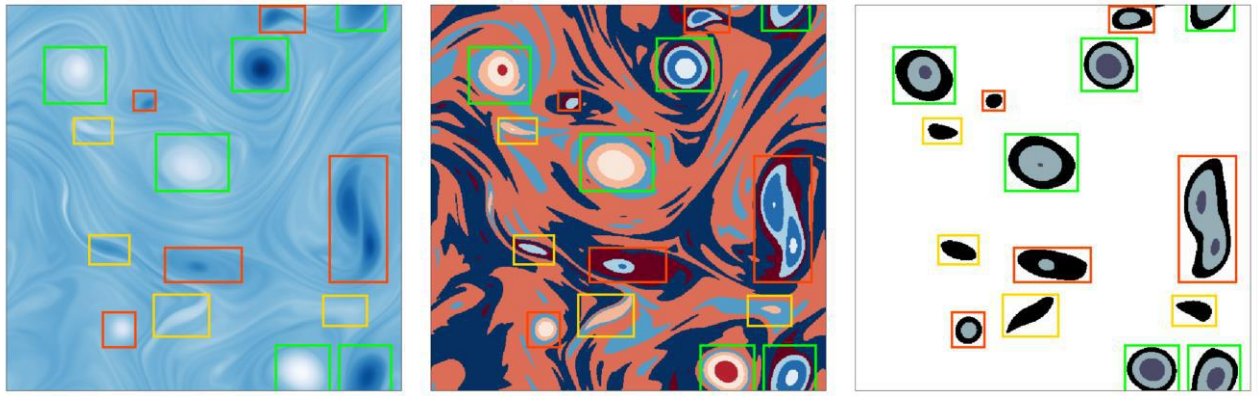


(f) Water vapor field of CAM5.1 climate model simulation (g) Climate local causal state field

- Target spatio-temporal data
- Requires much higher dimensions (10s to 100s)
- First distributed-memory implementation w/ scikit-learn API using daal4py

Performance:

- 30x single node speedup (via. Intel® Data Analytics Acceleration Library (Intel® DAAL))
- 20,000x to 30,000x speedup with 1024-node Cori supercomputer

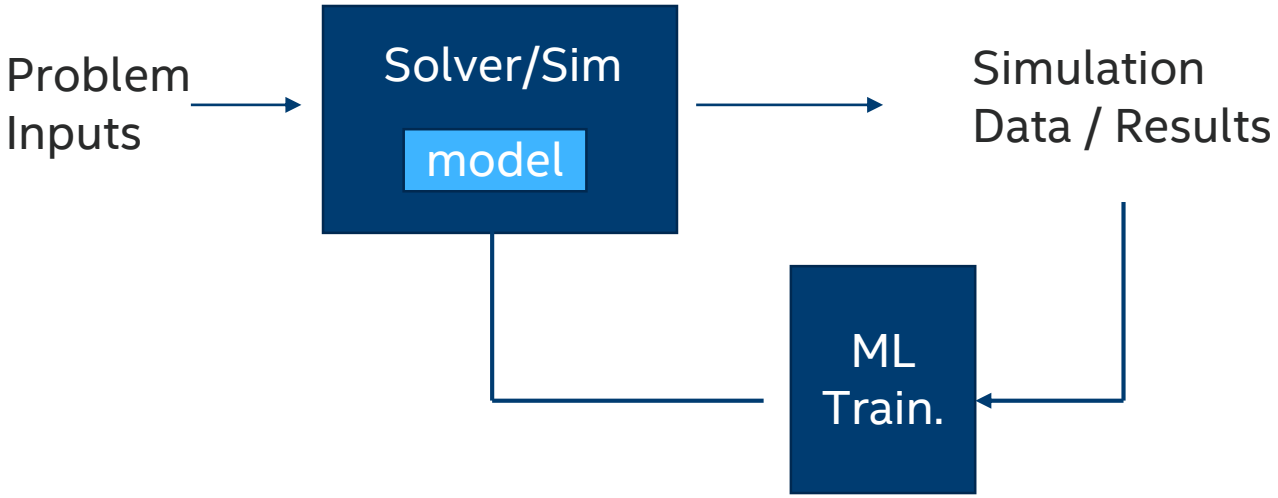


(a) Turbulence vorticity field (b) Turbulence state field, fine structure (c) Turbulence state field, coarse structure

A. Rupe, et. al, "DisCo: Physics-based Unsupervised Discovery of Coherent Structures in Spatiotemporal System", submitted for SC'19 workshop

* Emerging Trend

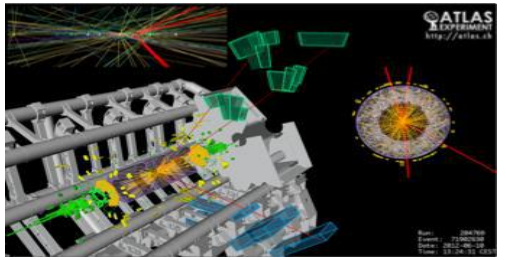
WAYS IN COMBINING HPC AND ML (2: AI IN MOD/SIM)



- Machine learning model can be trained separately, but it is USED with the solver / simulator
  - Implies ML models need to be usable in native solver / simulator programming environment (C/C++ or FORTRAN)
- Data used to train ML models are still pre-generated and stored in IO subsystem

Key Challenge: Data management / IO as well as model integration

HPC + AI IN PARTICLE PHYSICS (ETALUMIS @ SC'19)



Problem: Infer properties of particles and interactions from LHC at CERN

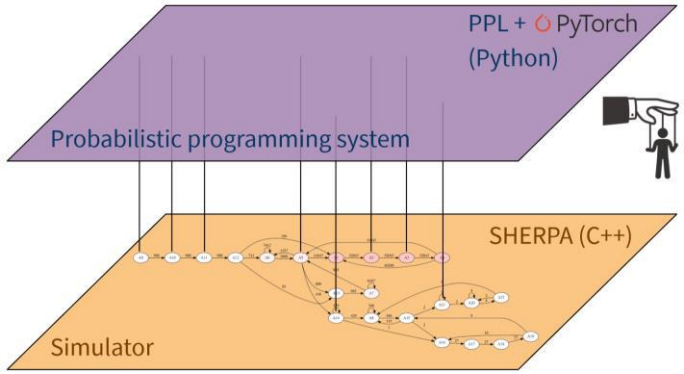
Traditional HPC approach

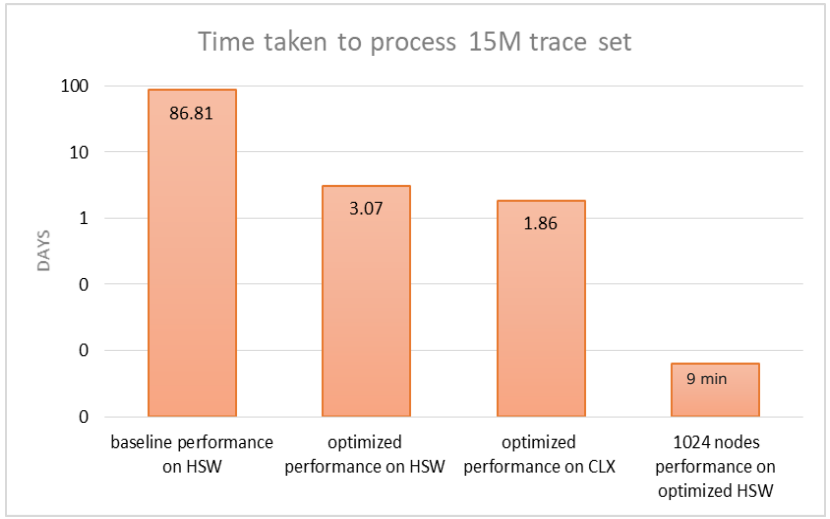
- Particle types are determined by tracing the path of particles inside the detectors. It is then validated using physical model.

HPC + ML approach

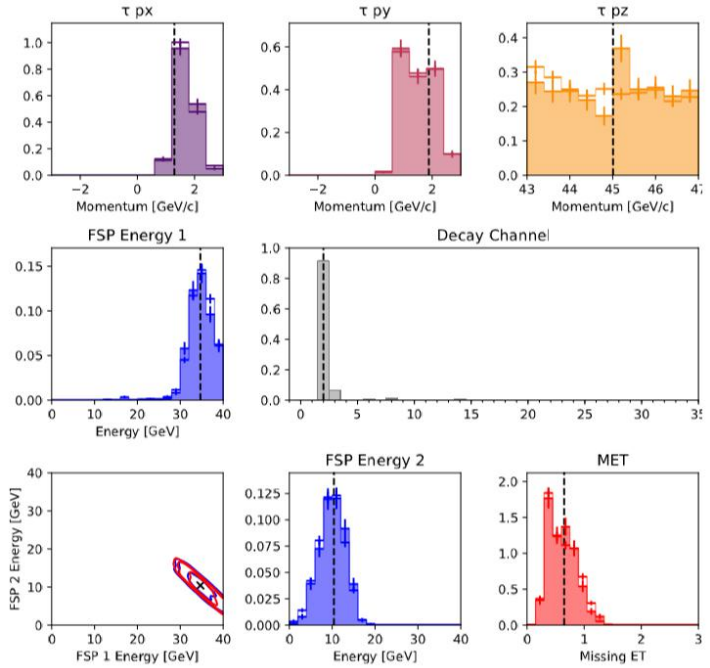
1) Use forward simulation traces to train a Probabilistic Programming model to correlate possible particle types and interactions to detection.
2) Apply model to infer observations directly from detector.

Etalumis
Integration of PPX w/ SIM





Achieved: 28x speedup on single node,
14,000x speedup w/ 1024 nodes;
230x speedup on inference



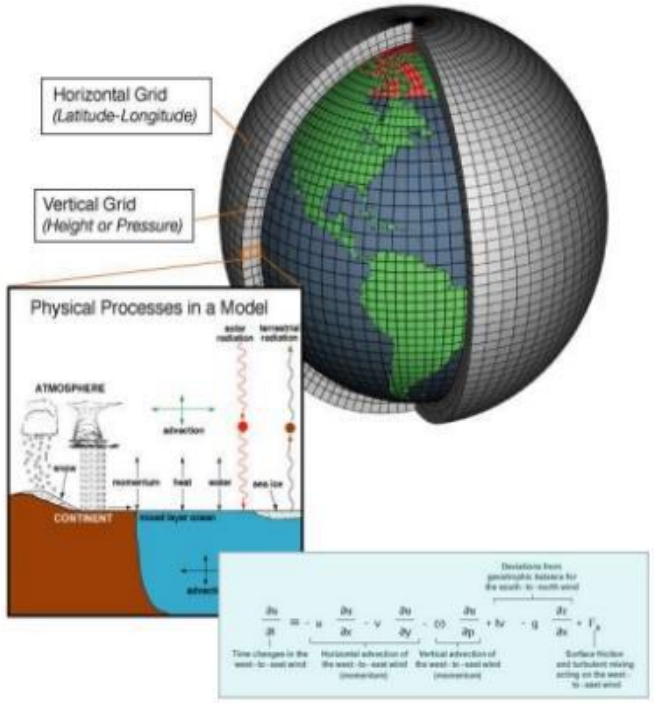
A. Bayden, et. al. "Etalumis: Bringing Probabilistic Programming to Scientific Simulators at Scale", to appear at SC'19

AI IN MOD/SIM EXAMPLE IN CLIMATE / WEATHER FORECAST

Forecast model has an insatiable compute demand (EF is not enough)

- ML model is accurate enough to replace part of the simulation to improve resolution

Global forecast model



Physical laws are presented in a form that a computer can compute the future state of atmosphere from the present state of atmosphere.

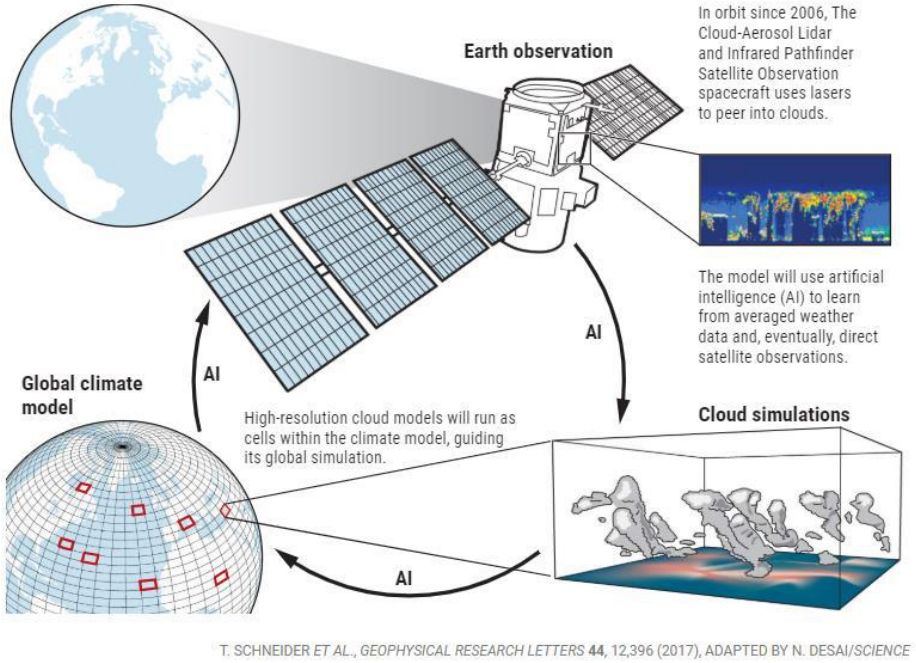
All physical variables (temperature, pressure, humidity, …) are presented in a grid with several layers.

The typical distance between grid points is 3-15 km. The number of vertical levels varies typically between 50 and 150.

Limitations:

- Update cycle (3-) 6 -12 h
- NWP not good in predicting the proper time and place of convective rain storms

ML tackle the limitation (clouds simulation)

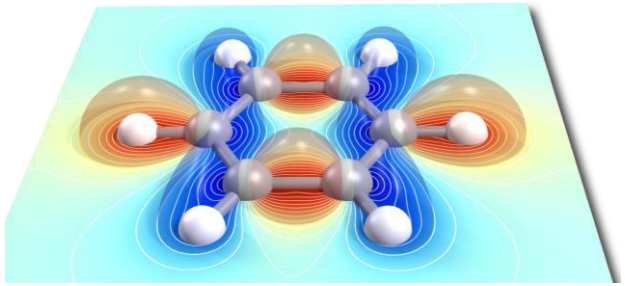


T. SCHNEIDER *ET AL.*, *GEOPHYSICAL RESEARCH LETTERS* **44**, 12,396 (2017), ADAPTED BY N. DESAI/*SCIENCE*

Methods described in https://www.sciencemag.org/news/2018/07/science-insurgents-plot-climate-model-driven-artificial-intelligence

AI IN MOD/SIM EXAMPLE DL GUIDED MONTE CARLO SIMULATION

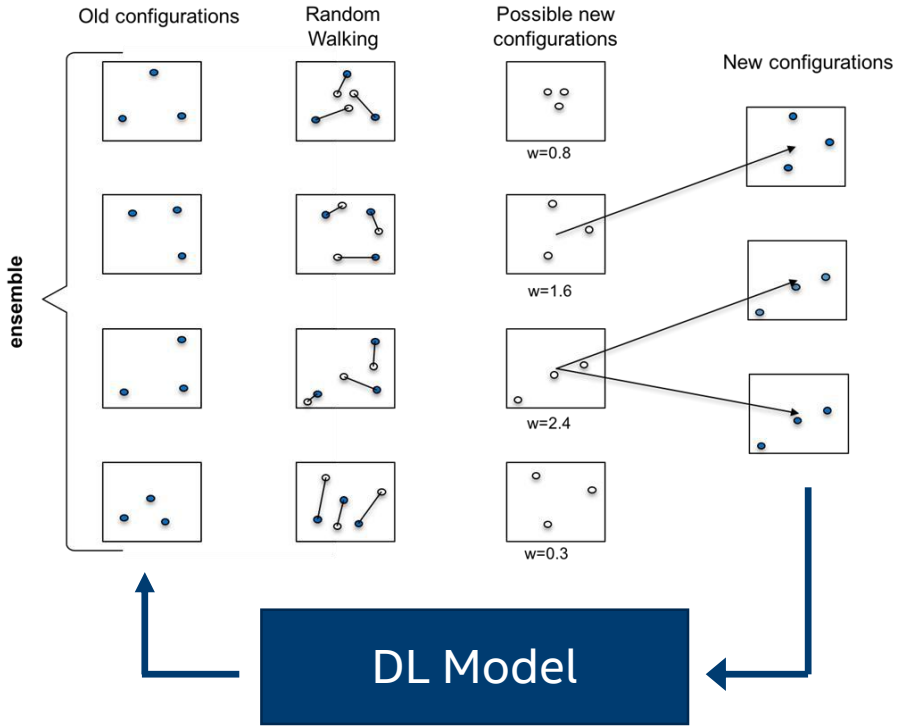
Quantum Molecular Simulations (QMS) are commonly used to study material properties



Goal: determine the best wave functions to describe the molecule energy

Traditional method is iterative and the amount of computation grows exponentially with number of atoms. One can quickly require EF of compute with ~1000 atoms.

QMCPACK: One of DoE Exascale Projects



Deep learning model is used to guide selection of "New Configurations" and allow skipping of critical simulation steps.

HPC + AI CONVERGENCE CHALLENGES

Significant increase in compute demands

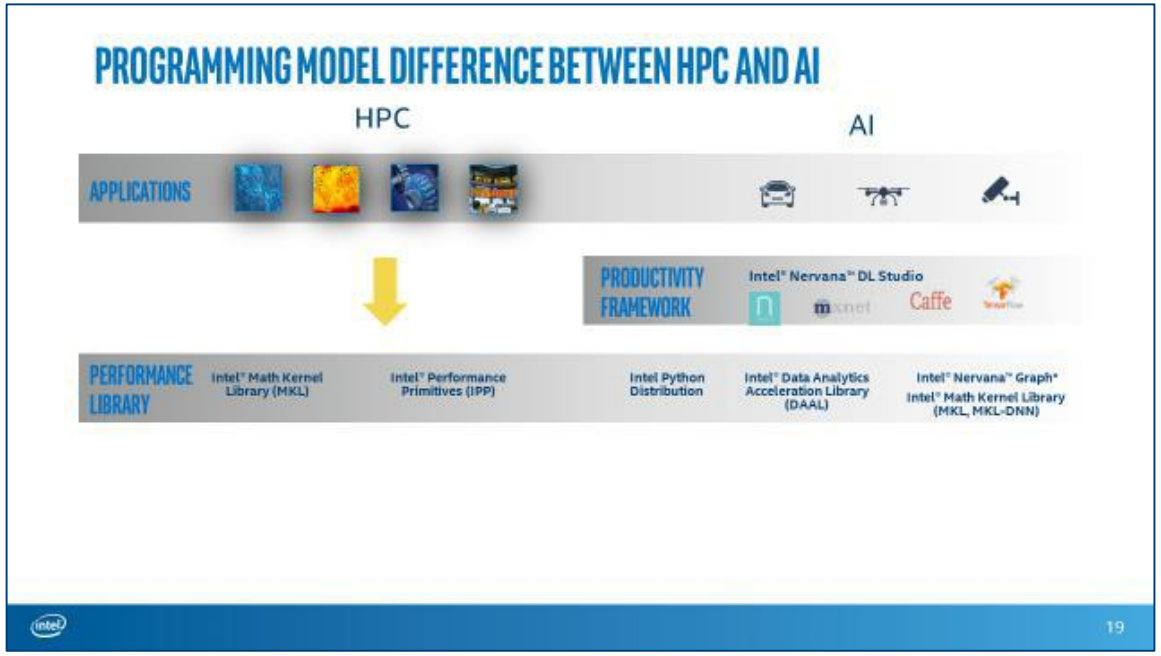
- 300,000x increase in ML training demand (compare to only 10x increase in HPC demands)
- Few optimized networks outside of image / speech domains.

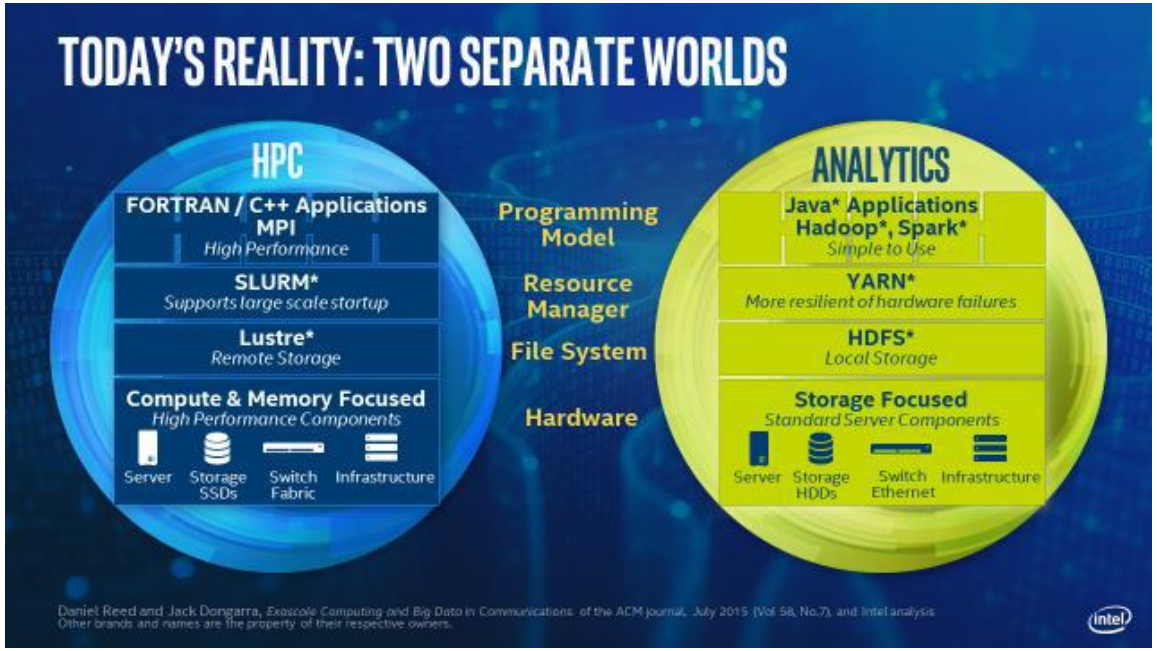
Not enough labelled data. Currently available labeled datasets:

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

HPC + AI CONVERGENCE CHALLENGES

- Different programming models / tools
- Different resource management stacks
- Different data formats, file systems





GAPS

Integration of ML model to mod/sim code:

- ML models are described and coded at higher level frameworks (TF, Pytorch, etc.)
- Trend: Community is working on more interchangeable format (ONNX, etc.)
- Still far from plug-n-play into HPC model / simulation codes
- **Need:** framework to embed ML model to model / simulation codes

Custom layers:

- Most commercial customers embed their IPs in customer layers/ops
- These layers are often not optimal for CPU, GPU, ACC, etc.
- **Need:** optimizer to generate high performance code for CPU, GPU, ACC, etc.

SUMMARY

HPC + AI is one of the most critical emerging workloads

Challenges exist for HPC + AI convergence:

- Programming model difference
- Resource management difference
- Data management requirements / file system differences

Solutions are WIP to address resource management / IO issues

Gaps exist in lack of tools to (1) integrate ML models with Mod/Sim code as well as (2) generate optimal implementation for custom layers

NOTICES AND DISCLAIMERS

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

No product or component can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks .

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/benchmarks .

Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

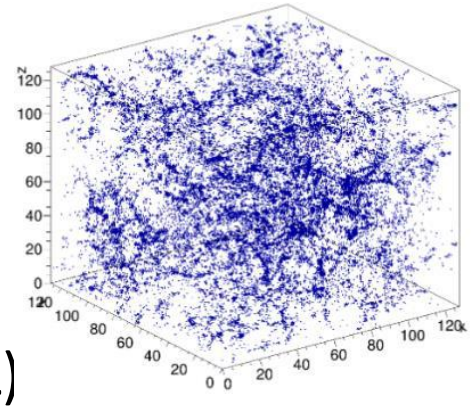
Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel, the Intel logo, and Intel Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as property of others.

THANK YOU!

Cosmology Application



- Cosmological parameters estimation using Tensorflow
  - Achieved million times speedup on Cori (using 8192 nodes of KNL)
    - Single node performance improved > 1000x (~ 2x performance of Nvidia P100)
    - >80% scaling efficiency on 8192 nodes KNL on Cori
  - Reduced time to train from 3-month to < 10mins.
  - Enable scientists to explore new science (e.g. 3 parameters estimation)