Research Findings: HPC-enabled AI

Steve Conway
sconway@hyperionres.com
September 2019

Recent Worldwide Studies/Projects for U.S. Federal Agencies

- The Evolution of AI Hardware and Software Ecosystems
- The Evolution of Field Competencies in Machine/Deep Learning and Resultant Industries
- AI Primer for Senior Decision-Makers
- AI Hardware Technology, Vendor Status and Trends


Converging HPC-Enterprise Market

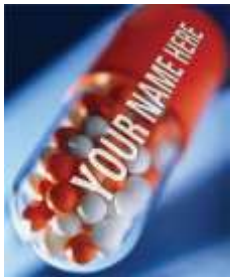
HPDA Data-Intensive Simulation

- Existing HPC users
  - Larger problem sizes
  - Higher resolution
  - Iterative methods
  - EP jobs to the cloud (Novartis)
- New commercial users
  - E.g., SMEs

HPDA Data-Intensive Analytics

- Existing HPC users
  - Intelligence community, FSI
  - Data-driven science/ engineering (e.g., biology)
  - Knowledge discovery
  - ML/DL, cognitive, AI
- New commercial users
  - Fraud/anomaly detection
  - Business intelligence
  - Affinity marketing
  - Personalized medicine
  - Smart cities
  - IoT

Most Economically Important HPDA-AI Use Cases



Precision Medicine

Automated Driving Systems

Fraud and anomaly detection

Affinity Marketing

Business Intelligence

Cyber Security

Edge/IoT/Smart Cities

High Growth Areas: HPDA-AI (May 2019)

- HPDA is growing faster than overall HPC market.
- AI subset is growing faster than all HPDA.

Table 1

Forecast: Worldwide HPC-Based AI Revenues vs Total HPDA Revenues ($ Millions)

	2018	2019	2020	2021	2022	2023	CAGR 18-23
WW HPC Server Revenues	13,706	14,495.000	15,780	17,376	18,983	19.947	7.8%
Total WW HPDA Server Revenues	3,153	3,598	3,932	4,737	5,467	6,450	15.4%
Total HPC-Based AI (ML, DL, and Other)	747	938	1,094	1,399	1,810	2,725	29.5%

Source: Hyperion Research 2019

Table 2

Forecast: Worldwide ML, DL & Other AI HPC-Based Revenues ($ Millions)

	2018	2019	2020	2021	2022	2023	CAGR 18-23
ML in HPC	532	675	875	1130	1479	1940	29.5%
DL in HPC	177	216	301	392	510	665	30.3%
Other AI in HPC	38	47	66	80	95	120	25.9%
Total	747	938	1,094	1,399	1,810	2,725	29.5%

Source: Hyperion Research 2019

Where Do You Run HPC workloads? (Choose ALL that apply)

On-premise HPC data center	67.2%
On-premise enterprise data center (business operations)	36.2%
More than one external cloud (e.g., Amazon, Google, Microsoft)	29.3%
On-premise private or hybrid cloud	19.0%
One external cloud (e.g., Amazon, Google, Microsoft)	19.0%
Not sure/don't know	1.7%
Other	3.5%

Machine Learning Goes Back At Least to the 1950s

Thomas Bayes
Statistician
1702-1761

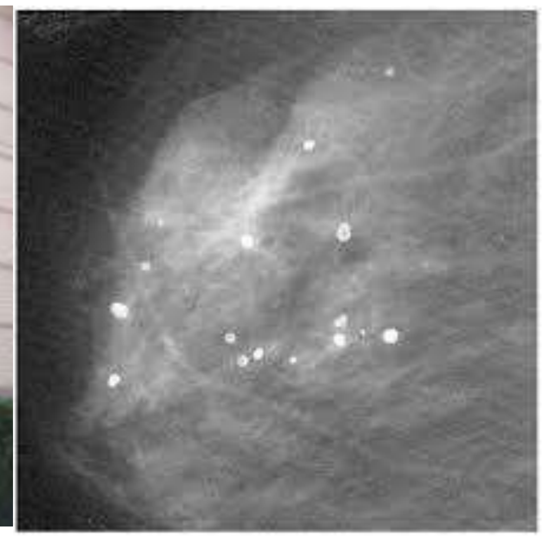
Decade	Summary
<1950s	Statistical methods are discovered and refined.
1950s	Pioneering machine learning research is conducted using simple algorithms.
1960s	Bayesian methods are introduced for probabilistic inference in machine learning.[1]
1970s	'AI Winter' caused by pessimism about machine learning effectiveness.
1980s	Rediscovery of backpropagation causes a resurgence in machine learning research.
1990s	Work on machine learning shifts from a knowledge-driven approach to a data-driven approach. Scientists begin creating programs for computers to analyze large amounts of data and draw conclusions – or "learn" – from the results.[2] Support vector machines (SVMs) and [3]recurrent neural networks (RNNs) become popular. The fields of [4] computational complexity via neural networks and super-Turing computation started.
2000s	Support Vector Clustering [5] and other Kernel methods [6] and unsupervised machine learning methods become widespread.[7]
2010s	Deep learning becomes feasible, which leads to machine learning becoming integral to many widely used software services and applications.

Source: Wikipedia

1990s: HPC-Enabled Machine Learning

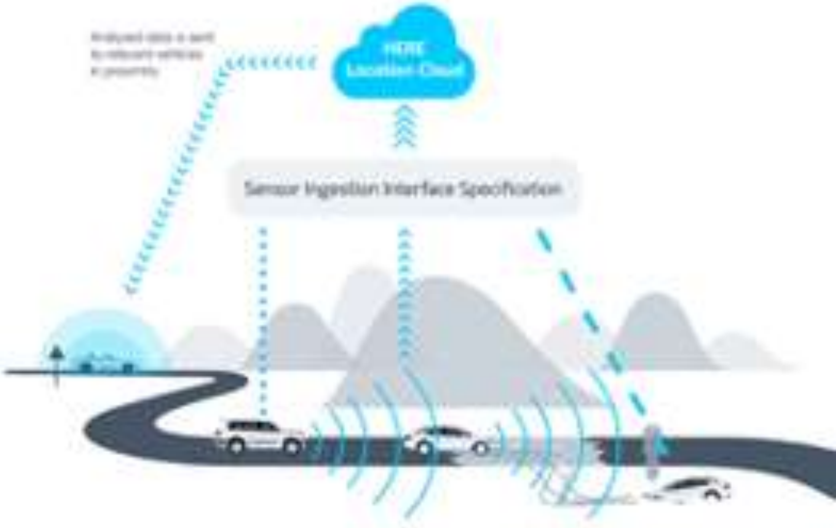
- By the early 1990s, George Washington University Hospital (Washington, DC) was routinely using a Cray supercomputer to help detect breast cancer after training it to identify early indicators, called microcalcifications, on X-ray films with better-than-human ability.

1993 Finalist
Computerworld
Smithsonian
Awards



Coupled Environments

- **Automated Driving Systems**
  - Embedded processor for local control (car-car, car-environment)
  - Private cloud for citywide and beyond (“air traffic control”)
- **Healthcare/Precision Medicine**
  - Healthcare systems are already private cloud-based.
  - Future: couple in-office HPC decision-support engine to private cloud.
- **5G Will Reduce Local-Cloud Latency Issue**



AI Is Still Near the Start

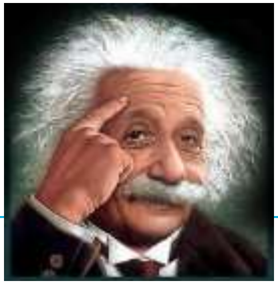
Low IQ (Weak Inferencing) → High IQ (Strong Inferencing)

Today: Bounded Problems

- Many observations but few choices to make
- "One trick dogs": 10 AI solutions in a box to solve 10 problems
- Already very useful:
  - Image & voice recognition
  - Advanced driver assistance
  - Reading an MRI

Future: Unbounded, Too

- Many observations, many choices to make
- Versatile decision-makers capable of serious experiential learning
- Examples:
  - Discerning human motivation
  - Fully automated driving
  - Diagnosing/"curing" a cancer



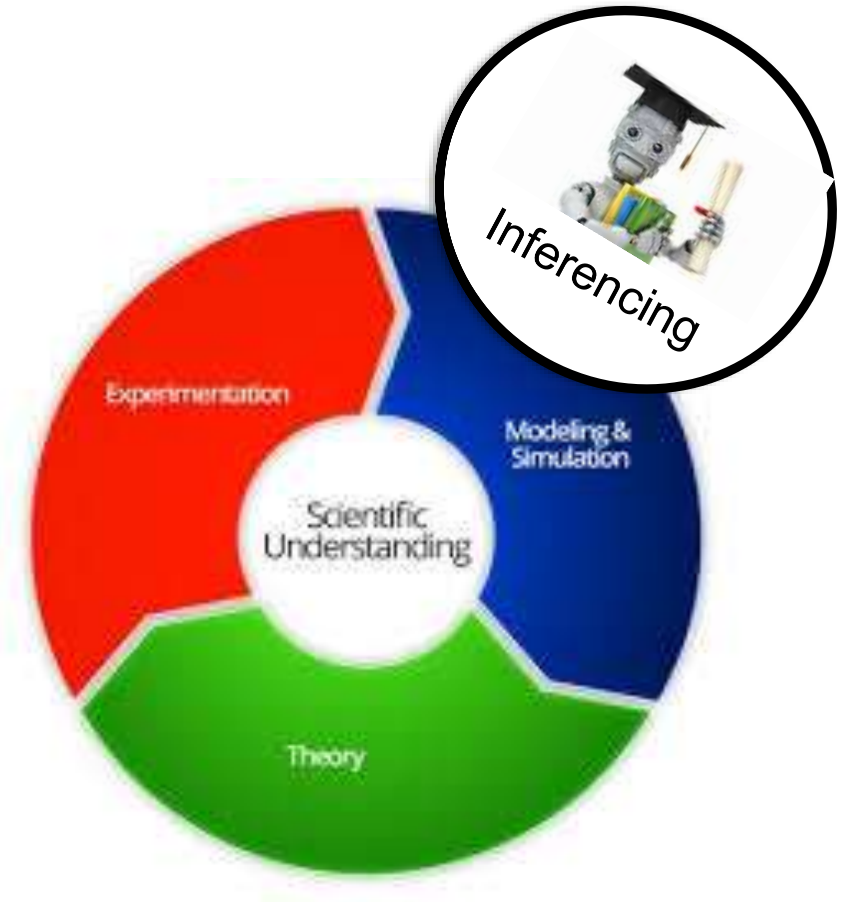
Hyperion Research Worldwide Survey of Leading AI Experts

Importance of HPC for Advancing AI

Importance	Number	Percent
Extremely important	47	87
Somewhat Important	7	13
Not very important	0	0
Not sure/don't know	0	0
Total	54	100

Intelligent Simulation

- AI adds fourth branch to the scientific method, ***inferencing.*** Complements theory, experiments & established simulation methods.
- Inferencing is the ability to guess, based on incomplete information
- Simulation is becoming much more data-intensive—esp. iterative methods.
- When inferencing is applied to data-intensive simulation, the result is ***intelligent simulation.***

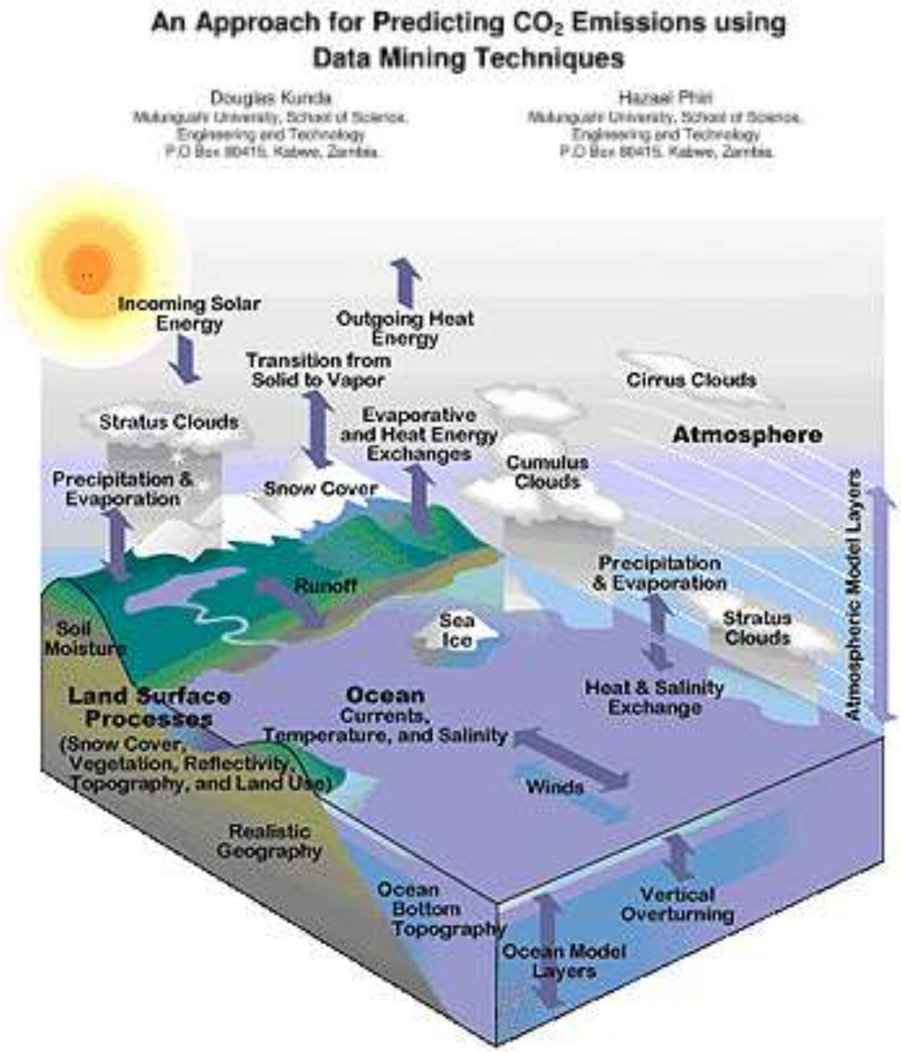


Building Consumer Trust in ADS

- RAND Corp. estimates 8.8 billion miles of physical testing would be needed to attain 95% consumer trust in self-driving vehicles. This would take 400 years.
- Adding HPC simulation can reduce time frame to 5-10 years.

Climate Knowledge Discovery

- Climate research is inherently data-intensive.
- Ensemble models & adding new factors (e.g., carbon cycle) have made it much more so.
- Climate knowledge discovery algorithms add a data analytics approach.
- The first IEEE workshop on this topic (2008), was called "Data Mining for Climate Change and Impacts."

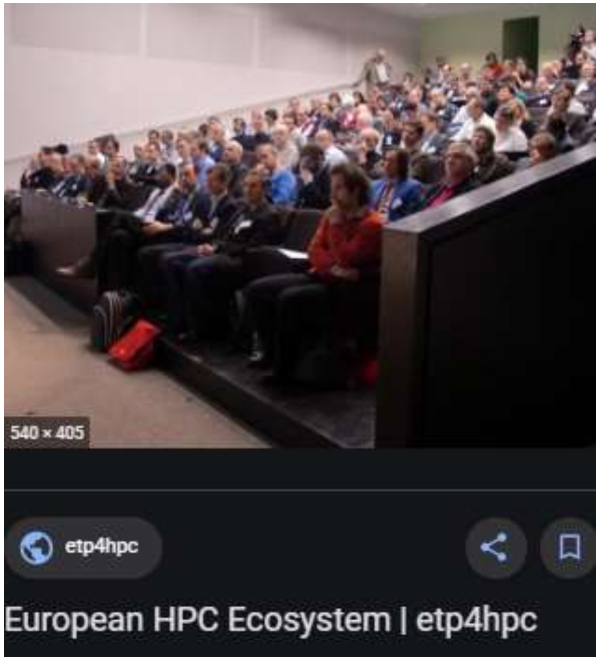


One Swim Lane in the Future

- Today: Mostly simulation and analytics on same compute-centric HPC system (budget reasons)
- In 2 years: More buyers will acquire separate system for analytics-AI.
  - But orthogonal findings of simulation & analytics runs will still need to be combined in the researcher's brain.
- Farther ahead: same system efficiently performs concurrent simulation & analytics runs – and integrates the orthogonal results.
  - Finally, economically efficient!

An Exploding Ecosystem

- The AI ecosystem has been growing quickly
- Targeted problems are more complex
- One result: new, AI-specific hardware
  - From companies large and small
  - China is very active
- The categories of processors and technologies continue to grow as well, and now include:
  - GPUs
  - TPUs
  - FPGAs
  - ASICs & eASICs
  - Neuromorphic Chips
  - IPUs
  - Inference Chips
  - Training Chips
  - Dataflow processors
  - Vector processors
  - 3D stacking
  - Optical interconnects



Co-Design

- AI chips will be centered on co-design, with specific tasks in mind. Examples:
  - Low-power ASICs at the edge
  - Custom AI chips in hyperscale data centers or the cloud
- GPUs will remain important but not for all AI workloads.
- Software and model-designed hardware is the direction forward.

Power

- Power consumption is critical to chip design for AI workloads.
  - Low power chips can be placed closer to the edge.
  - Latency for near real time training and inference require the compute to be next to the stored data.
- Processing and memory also need to be closer together.
  - With faster interconnect/fabric speeds

Cloud Companies Joining the Processor Development Party

- Google uses tensor cores to accelerate machine learning workloads.
  - Only available on Google cloud for now
  - Google announced the third generation TPU last year.
- Amazon, at their re:Invent conference in November of 2018, announced their inference chip, Inferentia.
  - Designed to accelerate machine learning, especially inferencing.