

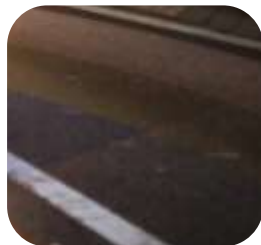
CRAY



Perspective on HPC-enabled AI

Tim Barr

September 7, 2017



AI is Everywhere



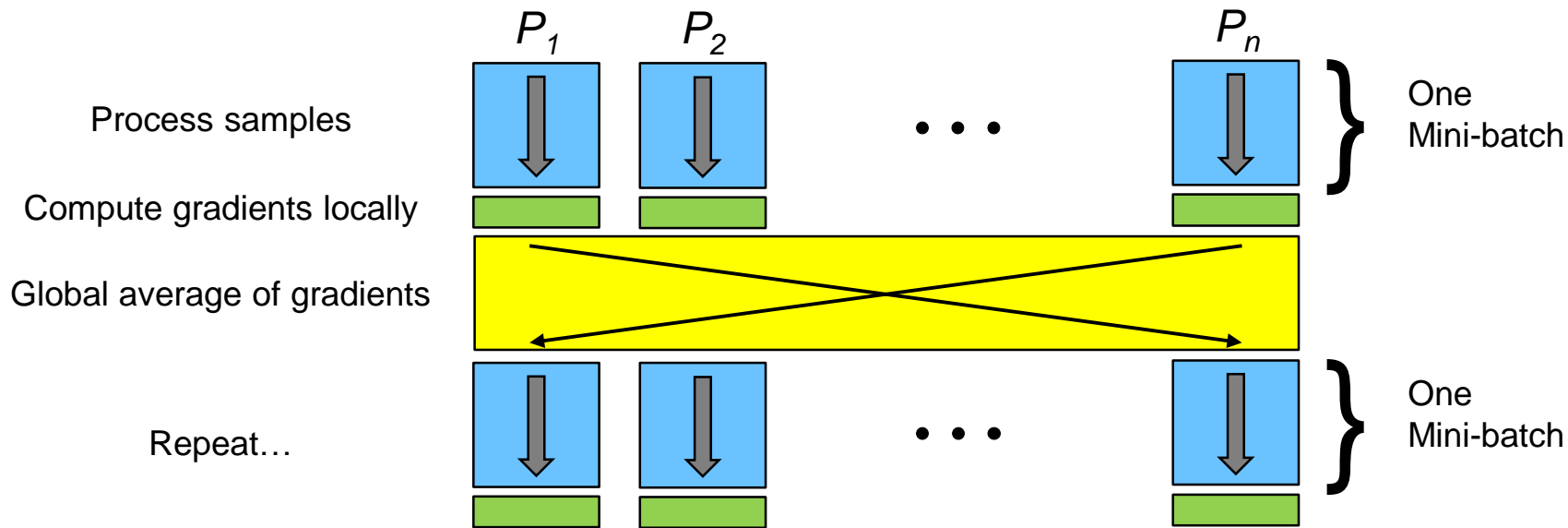
Deep Learning Component of AI

The punchline: Deep Learning is a High Performance Computing problem

- Delivers benefits similar to HPC in other disciplines
 - The value is in the decisions that are enabled
- Characterized by the same underlying factors
 - Large amount of computation
 - Large amount of data motion (I/O and network)
- The same methods work
 - HPC Technology and HPC Best Practice apply directly to DL

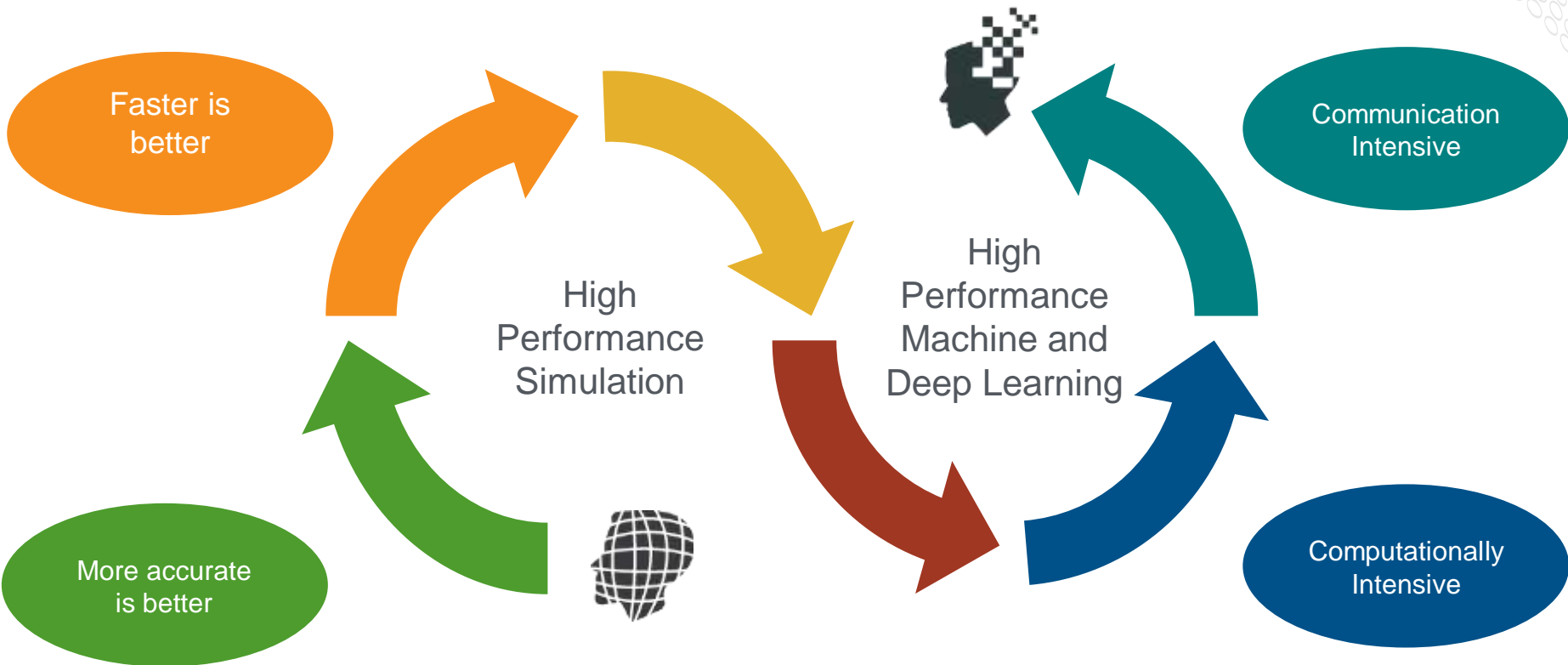
Deep Learning Training: Behind the Scenes

Computationally-intensive training phase

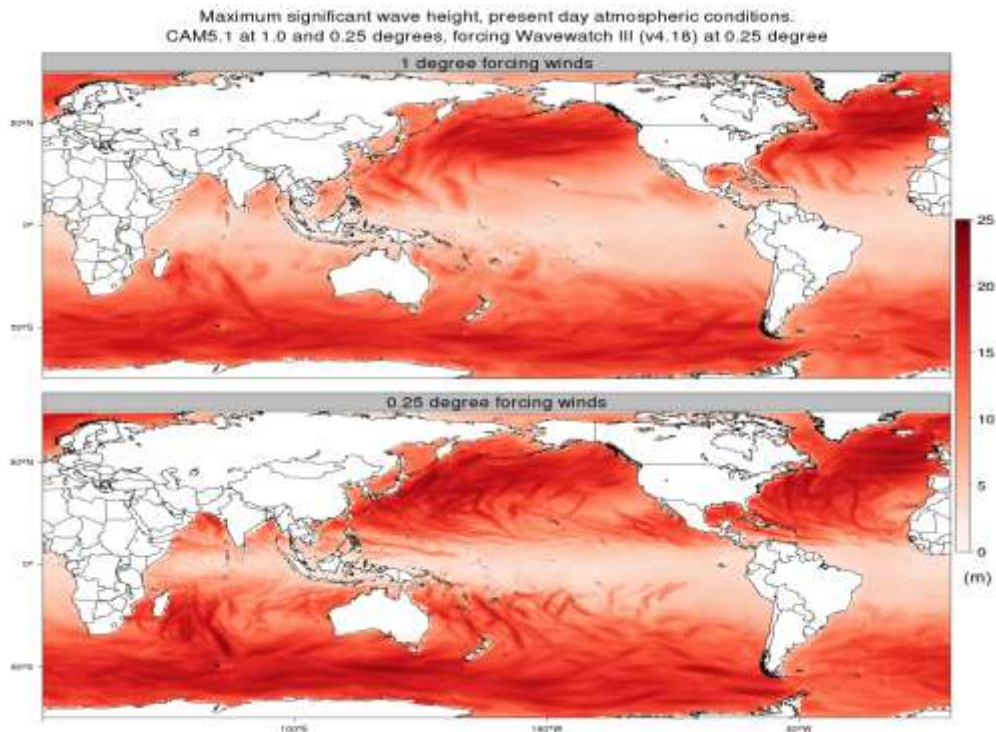


Deploying lots of computational power requires lots of communication.

Why Are We Here?



Let's Use Weather As An Example

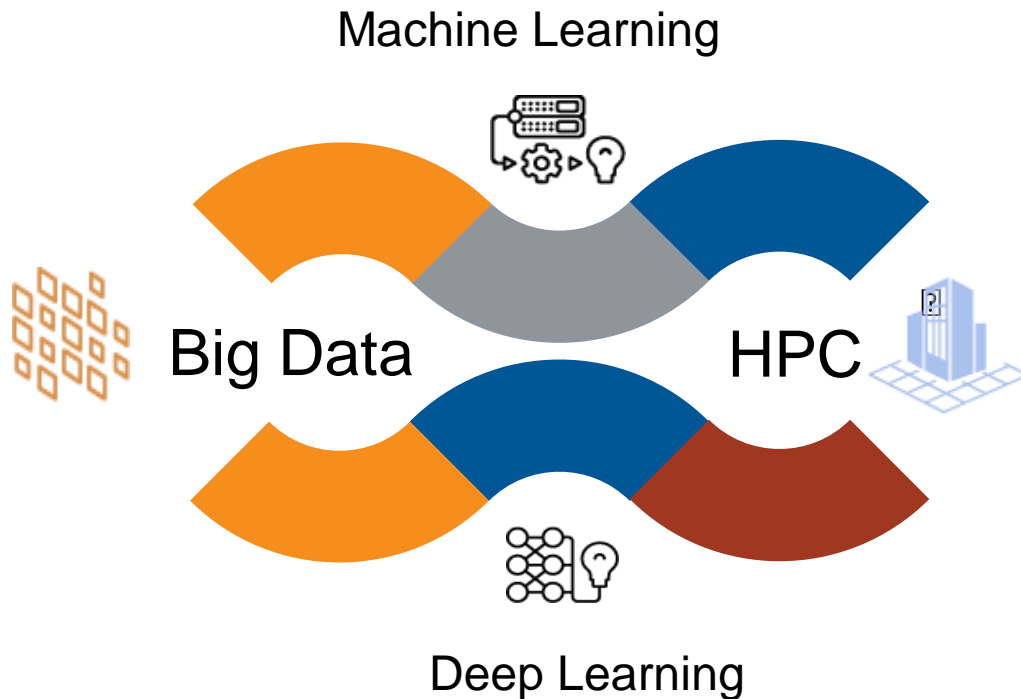


- **More Accurate is Better**
 - At 100km (top) and 25km (bottom)
 - Missed tropical cyclones and big waves up to 30 meters high
- **Faster is Better**
 - Higher resolution simulation requires 64X more computation

<http://www.nersc.gov/news-publications/nersc-news/science-news/2017-2/researchers-catch-extreme-waves-with-high-resolution-modeling>

HPC and AI Will Converge

2x Digital data is doubling in size every two years, and by 2020 the digital universe will reach 44 zettabytes²



28% believe HPC will allow them to scale computationally to build deep learning algorithms that can take advantage of high volumes of data¹

40% Reduction in error rates when 10x more data is being used in coordination with AI in speech recognition¹

1. "Are AI/Machine Learning/Deep Learning in Your Company's Future?", insideBigData + NVIDIA

2. EMC Digital Universe with Research & Analysis by IDC



What is Deep Learning ?



ARTIFICIAL INTELLIGENCE

Design of intelligent systems that augments human productivity. Systems that help decision makers do what they do best; leveraging computers doing what they do best



Sense		Comprehend		Predict		Act and Adapt	
ANALYTICS				 MACHINE LEARNING			
Search for the what, when, where and why				Learn patterns from the past to predict future			
Leverage domain and data science to query datasets for insights:				Unsupervised Group, cluster and organize content with domain-specific heuristic models		Supervised Train mathematical predictive models with labelled data	
Descriptive	What happened?			 DEEP LEARNING		Train and use neural networks as a predictive model	
Diagnostic	Why did it happen?						
Predictive	What will happen?			Vision		Speech	
Prescriptive	How to make it happen?						

Performance will be an AI Innovation and Adoption Driver

“AI and machine learning have reached a critical tipping point and will increasingly augment and extend virtually every technology enabled service, thing or application.”

“The combination of ***extensive parallel processing power, advanced algorithms and massive data sets to feed the algorithms has unleashed this new era.***”

Gartner’s Top 10 Strategic Technology Trends for 2017

“***Fast data is just as important as big data.*** In 2016, we’ll witness the emergence of a new class of real-time applications in e-commerce and financial technology services ***powered by super-speedy data analytics.*** ‘Fast data’ is the second iteration of big data, and it will create a lot of value.”

Fortune Magazine, December 2015

In a competitive international economy, advanced AI combined with supercomputing are essential ingredients for:

- Solution of strategically important problems
- Maintaining global leadership in industry, government and academia
- Creating next generation technologies, products and services

Deep Learning Will Require Supercomputing

- *An AI Revolution Started For Courageous Enterprises*
 - Yes, Deep Learning Warrants All The Fuss
 - Expect To Need Thousands Of Cores

FOR APPLICATION DEVELOPMENT & DELIVERY PROFESSIONALS

Deep Learning: An AI Revolution Started For Courageous Enterprises

AD&D Pros Can Develop Applications That Can See, Understand, Talk, And Learn

by Mike Gualtieri, Diego Lo Giudice, and Brandon Purcell
May 12, 2017

Why Read This Report

Deep learning is a revolution started. A revolution because it allows enterprises to create predictive models with uncanny accuracy on previously hard-to-analyze data such as images, voice, and natural language. A revolution because the internet giants have all embraced deep learning as their go-forward AI strategy. And, finally, a revolution because it has only just begun. Once a revolution gets big enough, it disrupts. That's the opportunity for application development and delivery (AD&D) professionals who build enterprise and customer applications.

Key Takeaways

Yes, Deep Learning Warrants All The Fuss

A branch of machine learning, deep learning focuses on the creation of artificial neural networks that represent knowledge and can learn from new data.

Teach Your Apps To Be Smart Like Humans

The most prominent and successful use cases for deep learning are computer vision, voice recognition, and natural language processing.

Expect To Need Thousands Of Cores

The vector computations required to train deep-learning models are orders of magnitude greater than traditional machine learning. Enterprises wishing to use deep learning must acquire new hardware or use specialized cloud instances such as graphics processing units.

FORRESTER.COM

Deep Learning with Supercomputers

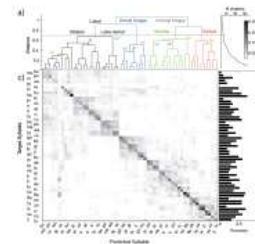
NERSC – Deep Learning in Science



Modeling galaxy shapes



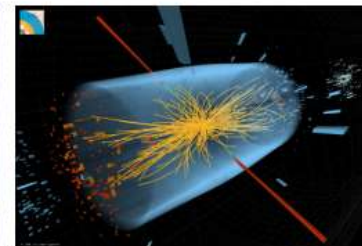
Clustering Daya Bay events



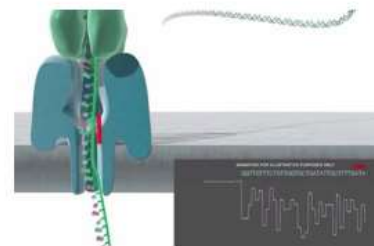
Decoding speech from ECoG



Detecting extreme weather



Classifying LHC events



Oxford Nanopore sequencing

Opportunities to apply DL widely in support of classic HPC simulation and modelling

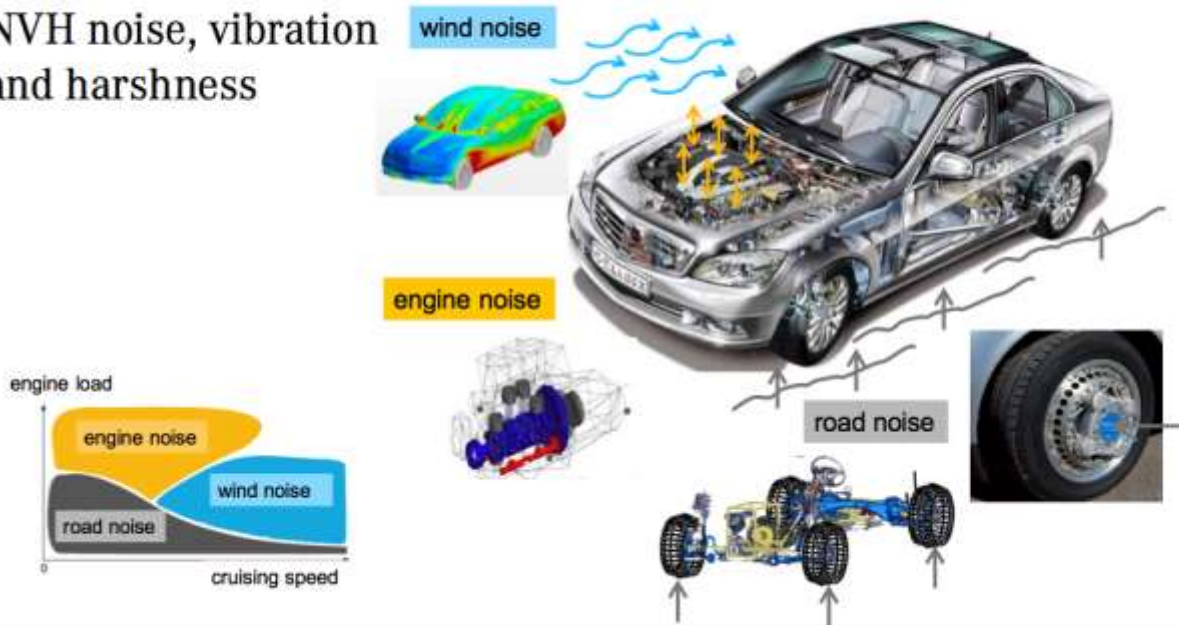
Deep Learning in Automotive

Noise, Vibration and Harshness at Daimler

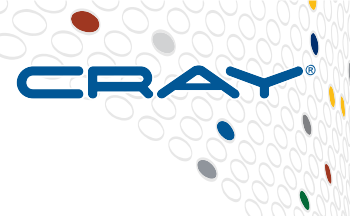


- Noise, Vibration and Harshness is a traditional HPC application used in automotive and aerospace
- Deep Learning has the potential to do an automatic evaluation of results in complex, multi-component, non-linear applications

Deep learning and Data Analytics in CAE
NVH noise, vibration and harshness



Deep Learning Examples in Manufacturing



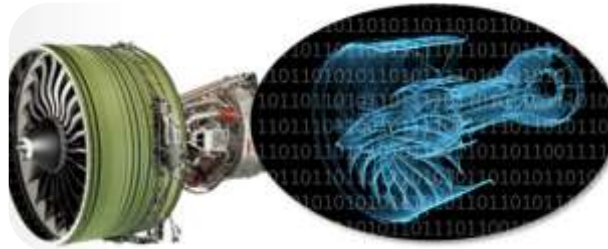
Aerospace Drones

10-fold increase in the commercial drone fleet by 2021...FAA, 2017



Digital Twin

"Top 10 technologies for 2017",
Gartner



Autonomous Vehicle

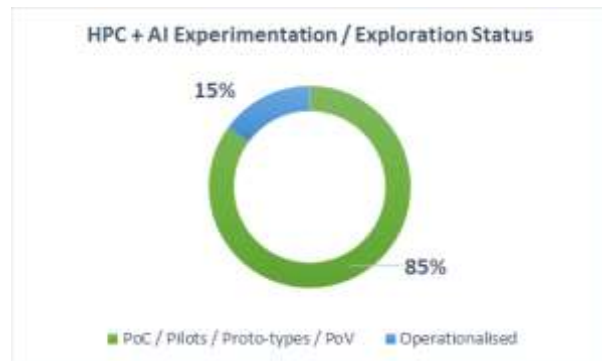
OEMs will invest \$7 billion in development...Frost &Sullivan, 2016



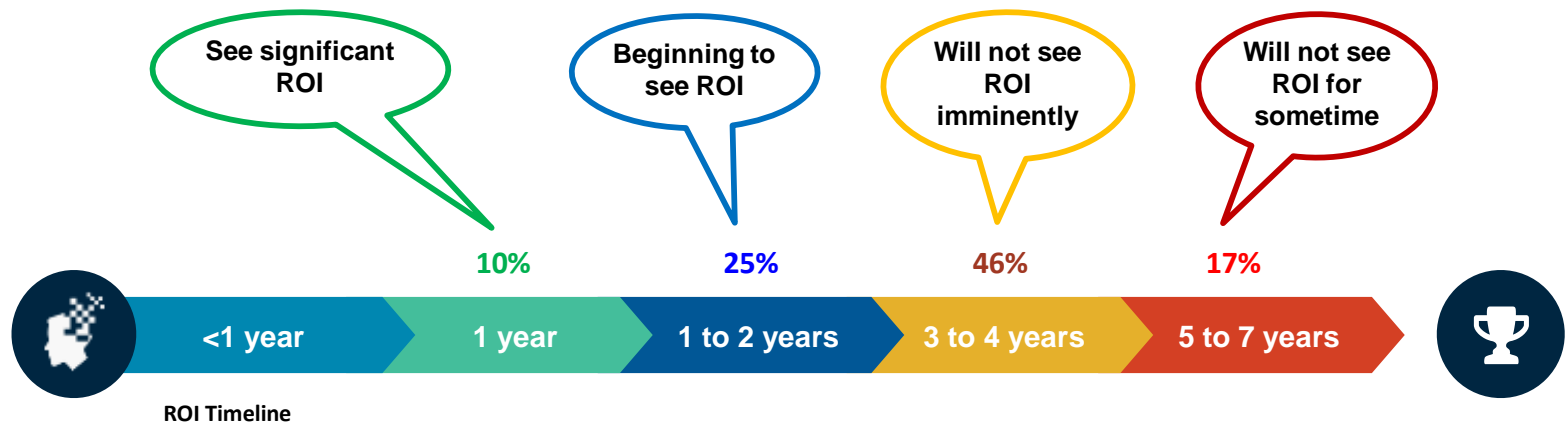
Leveraging data analytics and deep learning between engineering disciplines and across the enterprise has great potential for product quality and innovation

When Should You Start?

A Sample from the Financial Services Sector



- ROI payoff will be 1 – 2 years
- Time to begin experimentation is now



Why Deep Learning Now?



Electronic brain

Perceptron

ADALINE

XOR

Backpropagation

SVM

Deep Learning



S. McCulloch - W. Pitts



F. Rosenblatt



B. Widrow - M. Hoff



M. Minsky - S. Papert



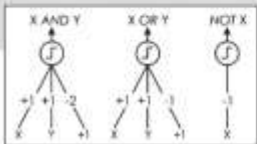
D. Rumelhart - G. Hinton - R. Williams



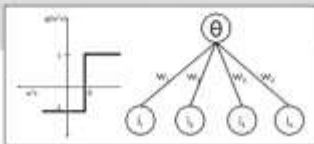
V. Vapnik - C. Cortes



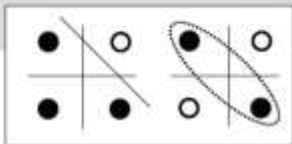
G. Hinton - S. Ruslan



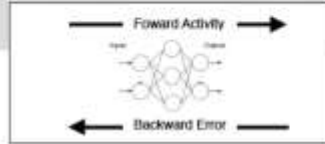
Adjustable weights
Weights are not learned



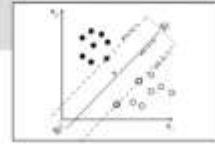
Learnable weights and threshold



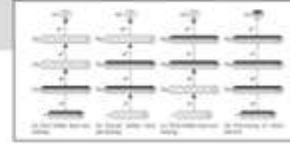
XOR Problem



Solution to nonlinearly separable problems
Big computation, local optima/overfitting



Limitations of learning prior
Kernel function:
Human intervention



Hierarchical feature learning

Deep Learning Challenges

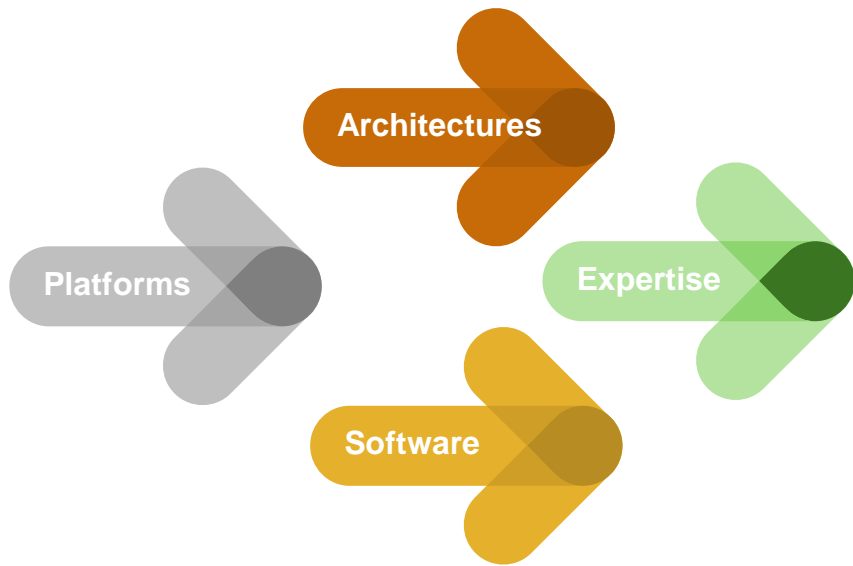


“AI systems still demand considered design, knowledge engineering and model building”, Forrester AI

TechRadar Q1 2017

- A lot to learn for practitioners and end-users:
 - Large, complex workflows
 - Different Toolkits + Data Movement + Network
 - Defining the value returned to the business
- Training times grow with data sizes and complexity:
 - Days to Weeks
 - Compounded with hyper parameter optimization (O(1000) is not unrealistic)

Enabling resource intensive training by delivering performance efficiencies and scalability

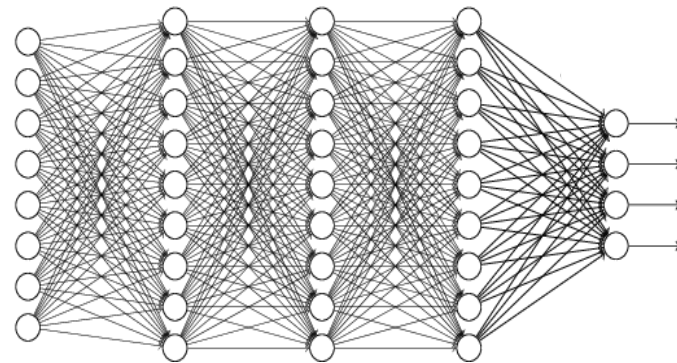


- Deep Learning Platforms - dense GPU to scalable platforms with optimized software stacks
- Apply HPC best practices and expertise to improve deep learning frameworks and core algorithms

Reduce Total Workflow Time

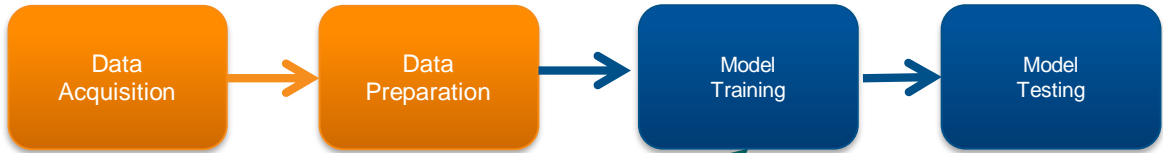
Why? The Deep Neural Net Training Problem

- **DNN model with weights on all connections**
 - Largest models now hundreds of layers, and millions (to billions) of nodes
- **Large set of labeled training data**
- **Idealized training algorithm:**
 - For every *minibatch* of training samples:
 - run samples forward through the model
 - compute the error vs. the training data
 - back-propagate error through the NN to update the weights (gradient descent)
- **After all data processed, iteratively optimize *hyperparameters* until required accuracy is achieved**



A (not particularly deep) neural net

Reduce Total Workflow Time



Apply HPC best practices and expertise to improve deep learning frameworks and core algorithms

- **Minutes, Hours:**
 - Interactive research!
Instant gratification!
 - **1-4 days**
 - Tolerable
 - Interactivity replaced by running many experiments in parallel
 - **1-4 weeks:**
 - High value experiments only
 - Progress stalls
 - **>1 month**
 - Don't even try
- Source: Large-Scale Deep Learning for Intelligent Computer Systems, Jeff Dean, Google

Cray Focus: Deep Learning Training at Scale

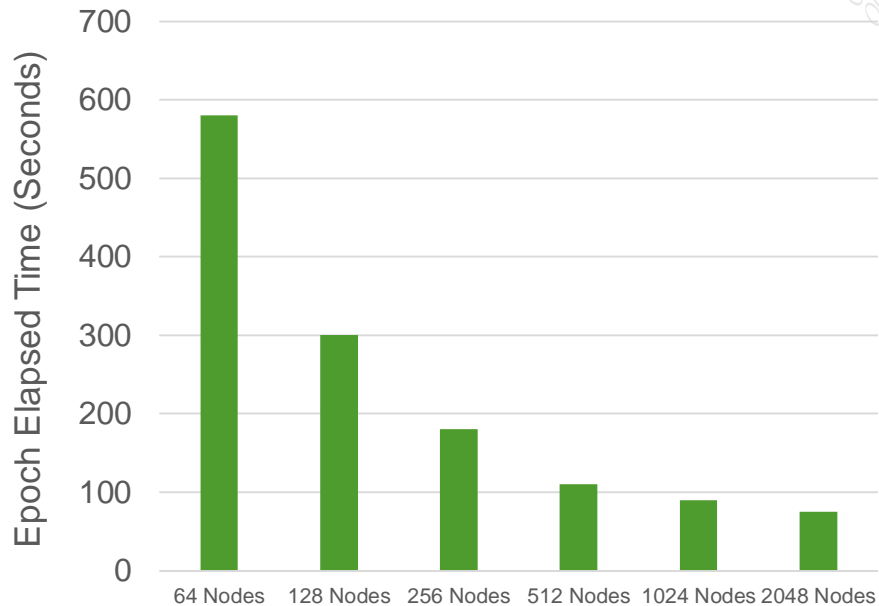
CNTK: Distributed Version vs Cray MPI Parallel Implementation



- Apply HPC Best Practices and Cray Expertise to improve DL systems and core algorithms with real-world use cases
- Collaborations across Cray customers and other stakeholders
- Currently optimizing different toolkits:
 - CNTK
 - TensorFlow
 - MXNet

“Applying a supercomputing approach to optimize deep learning workloads represents a powerful breakthrough for training and evaluating deep learning algorithms at scale. Our collaboration with Cray and CSCS has demonstrated how the Microsoft Cognitive Toolkit can be used to push the boundaries of deep learning.”

- Dr. Xuedong Huang, distinguished engineer, Microsoft AI and Research



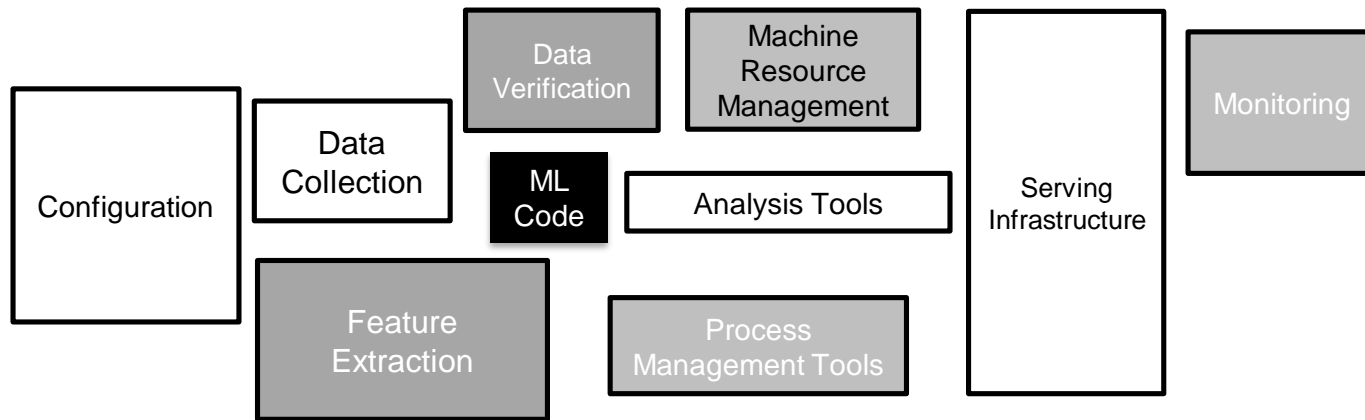
Microsoft Cognitive Toolkit



CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre



HPC Focus: Comprehensive Systems



“Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.”

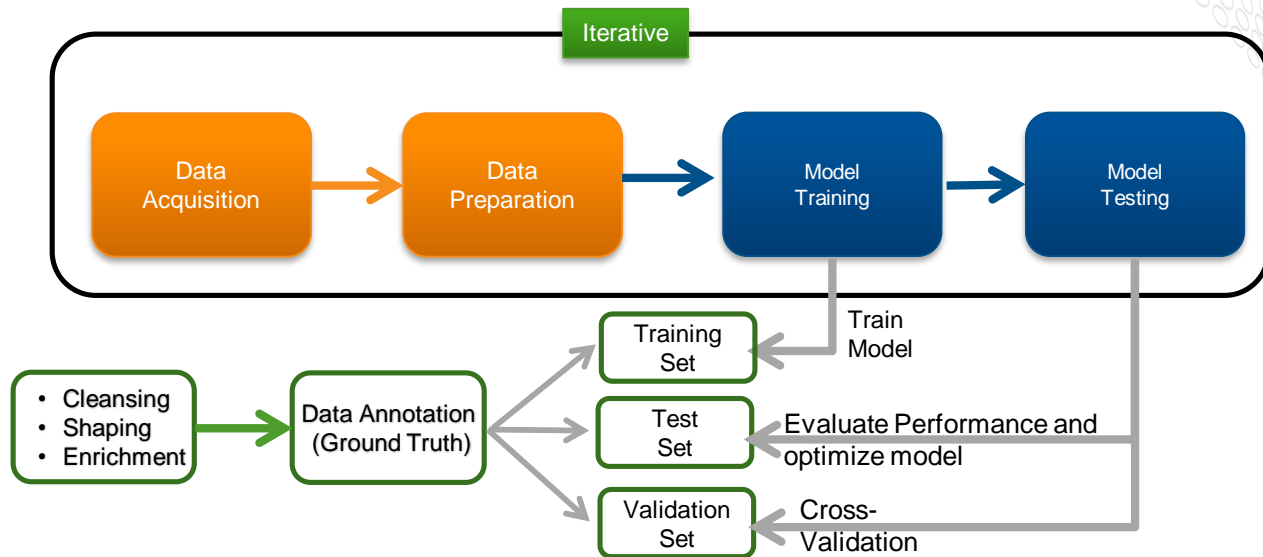
-Adapted from *Hidden Technical Debt in Machine Learning Systems*, Sculley et. al., NIPS '15

HPC Supports the Entire AI Workflow



Deep Learning workflows are not limited to training.

- Similar to other HPC and analytics workloads, significant portions of DL jobs are devoted to data collection, preparation and management.



AI is everywhere... Even the grocery store



Trending: [Seattle Seahawks players talk about their experience wearing new high-tech Vicis helmet](#)

Whole Foods offers Amazon Echo as 'Farm Fresh Pick of the Season' as tech giant takes over upscale grocer

BY NAT LEVY on August 28, 2017 at 8:00 am



CRAY



Thank You

