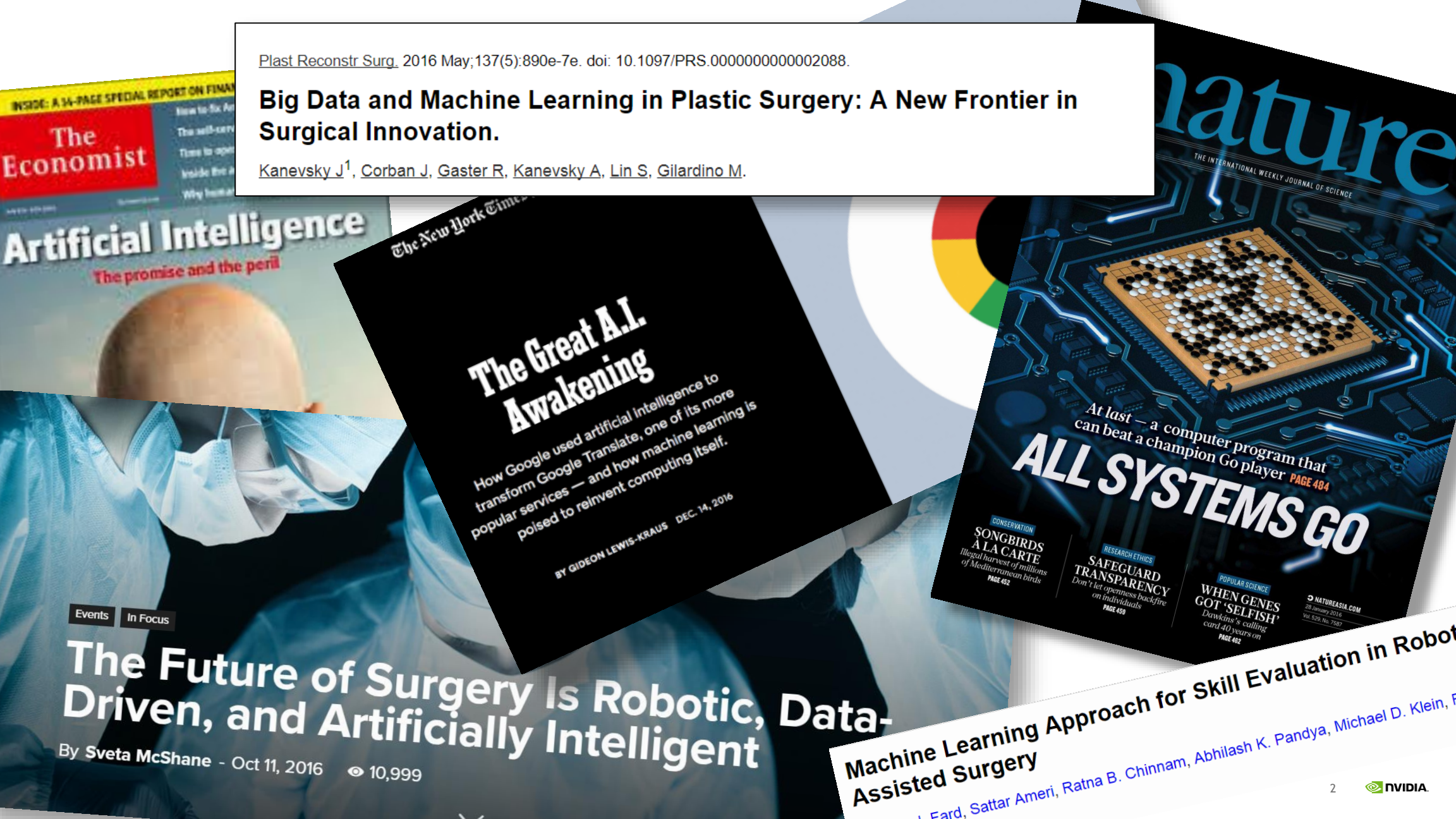
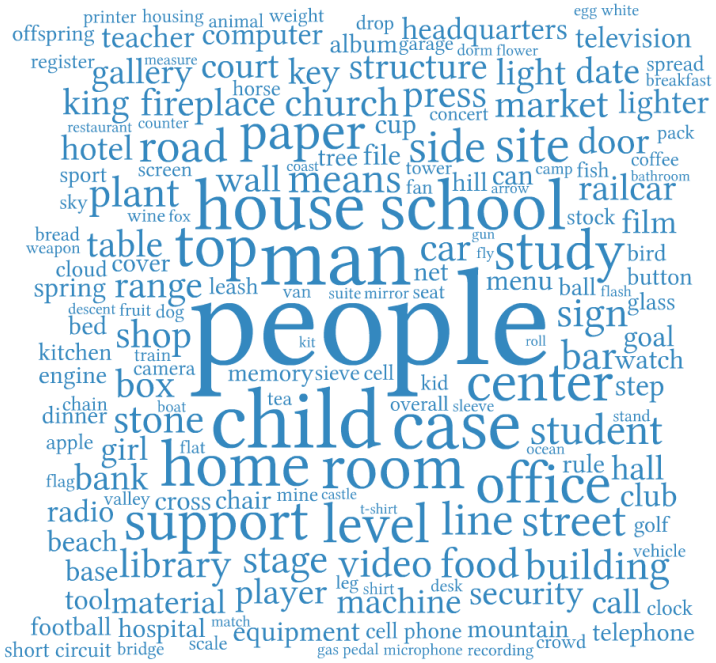
AI FOR COMPUTATIONAL SCIENCE

Peter Messmer, 6/15/2018

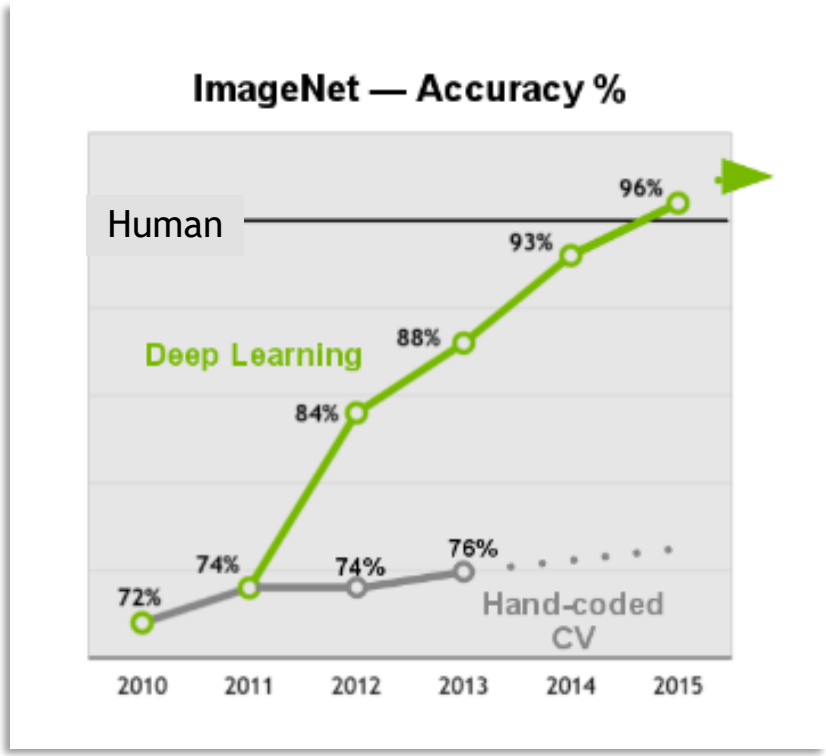

Plast Reconstr Surg. 2016 May;137(5):890e-7e. doi: 10.1097/PRS.0000000000002088.
Big Data and Machine Learning in Plastic Surgery: A New Frontier in Surgical Innovation.
Kanevsky J[1], Corban J, Gaster R, Kanevsky A, Lin S, Gilardino M.
The Economist
Artificial Intelligence
The promise and the peril
The New York Times
The Great A.I. Awakening
How Google used artificial intelligence to transform Google Translate, one of its more popular services — and how machine learning is poised to reinvent computing itself.
BY GIDEON LEWIS-KRAUS DEC. 14, 2016
nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
At last — a computer program that can beat a champion Go player PAGE 484
ALL SYSTEMS GO
CONSERVATION
SONGBIRDS À LA CARTE
Illegal harvest of millions of Mediterranean birds
PAGE 452
RESEARCH ETHICS
SAFEGUARD TRANSPARENCY
Don't let openness backfire on individuals
PAGE 459
POPULAR SCIENCE
WHEN GENES GOT 'SELFISH'
Dawkins's calling card 40 years on
PAGE 462
NATUREASIA.COM
28 January 2016
Vol. 529 No. 7587
Events
In Focus
The Future of Surgery Is Robotic, Data-Driven, and Artificially Intelligent
By Sveta McShane - Oct 11, 2016
10,999
Machine Learning Approach for Skill Evaluation in Robot-Assisted Surgery
Sattar Ameri, Ratna B. Chinnam, Abhilash K. Pandya, Michael D. Klein,
2
NVIDIA

DEEP LEARNING LEADS TO SUPER-HUMAN CAPABILITIES



IMAGENET

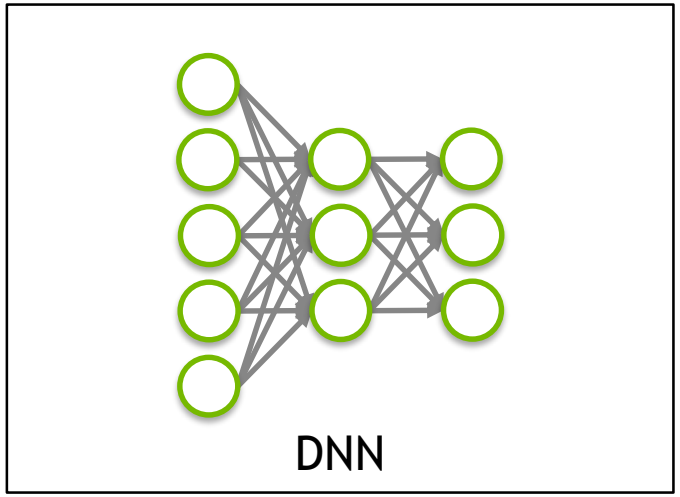
14,197,122 images, 21841 synsets indexed



Machine outperforms humans at classification accuracy

NVIDIA

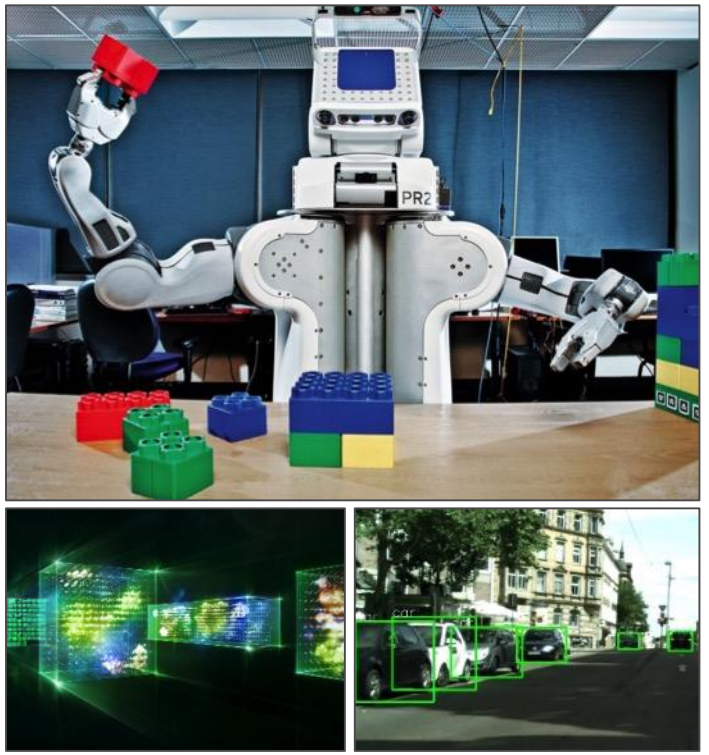
THE BIG BANG IN MACHINE LEARNING



"Google's AI engine also reflects how the world of computer hardware is changing. (It) depends on machines equipped with GPUs… And it depends on these chips more than the larger tech universe realizes."

WIRED

NVIDIA - AI COMPUTING COMPANY



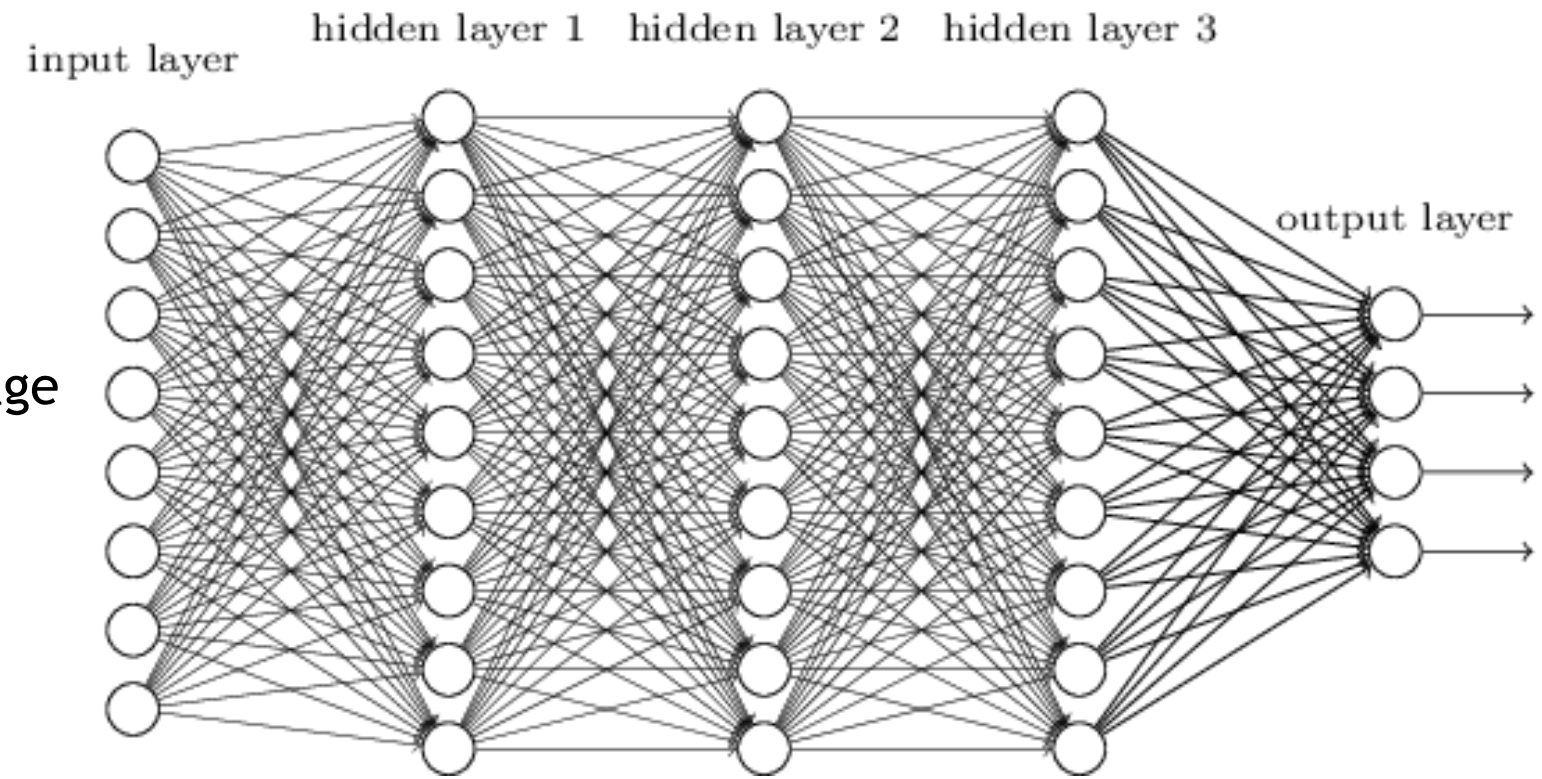
Computer Graphics

GPU Computing

Artificial Intelligence

ANATOMY OF A DEEP NEURAL NETW

Image



Probability of classes

$$y = f(\sum_{k=0}^{n} x^k a^k)$$

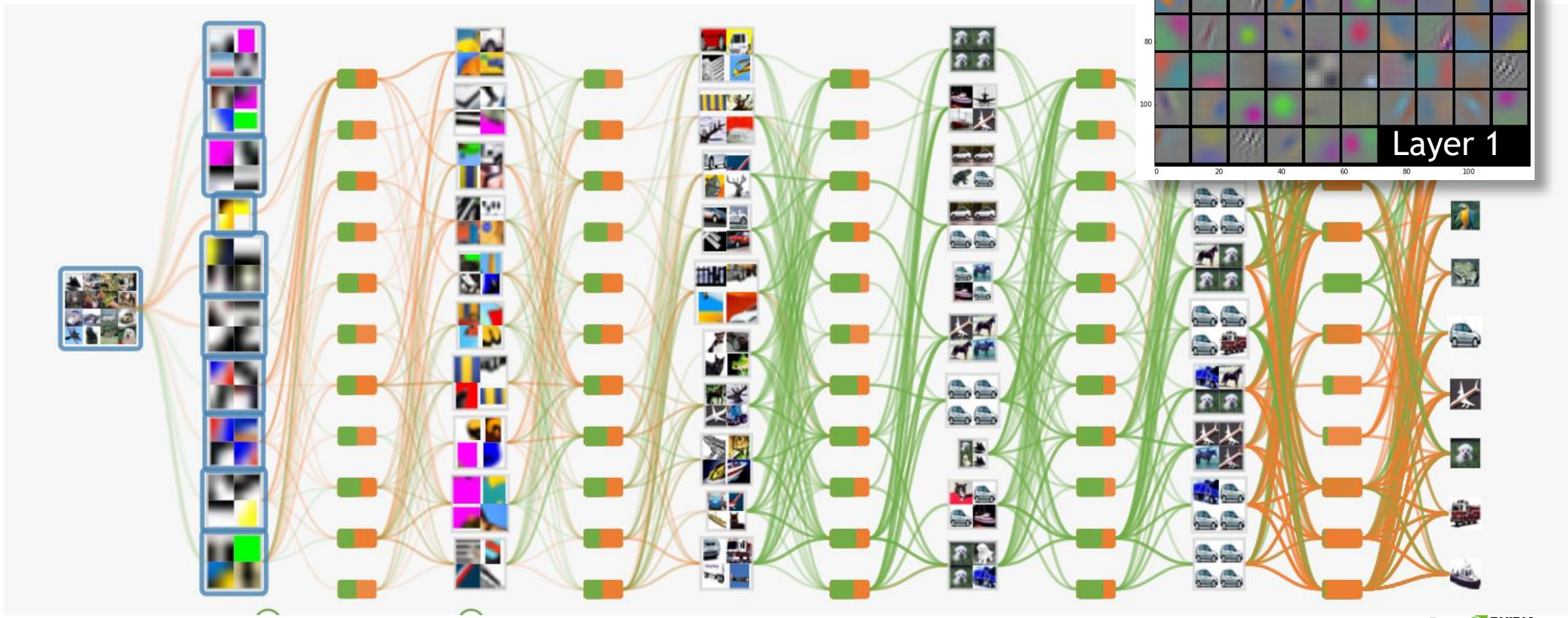
What happens in the layers?
How to chose the weights?

Google Inception Network

NVIDIA.

LOOKING INSIDE A NEURAL

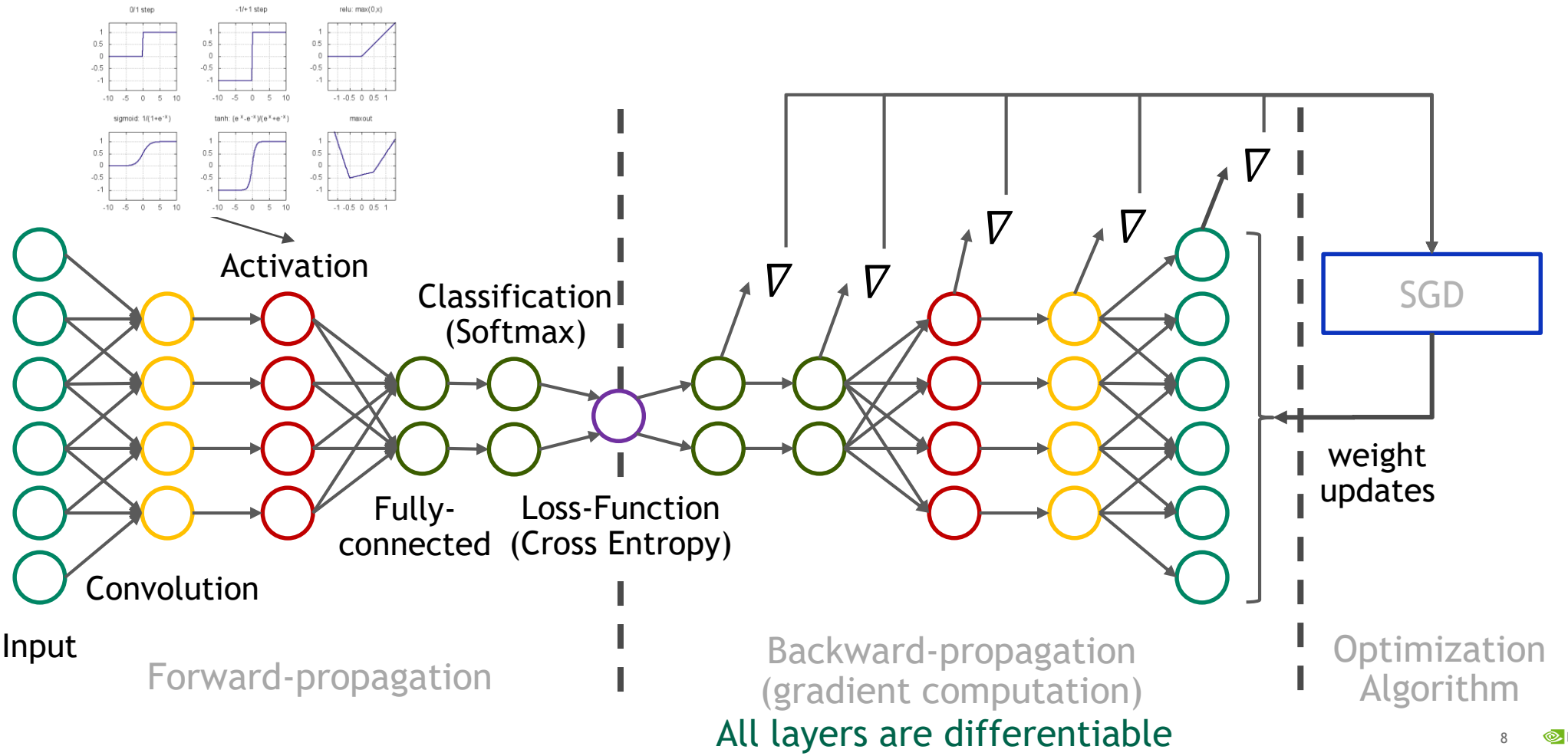
Different layers are sensitive to different fe



NVIDIA.

1-SLIDE INTRO TO CONVOLUTIONAL NEURAL NETS

Forward/Backward Propagation



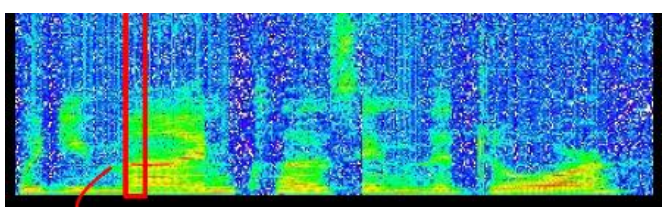
All layers are differentiable

NVIDIA.

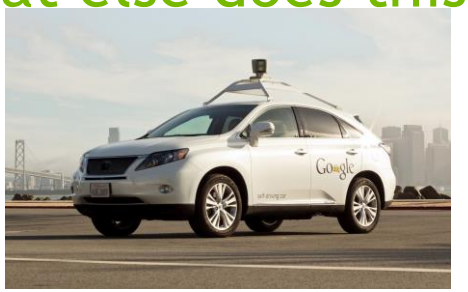
OK, MACHINES ARE GREAT AT IMAGE RECOGNITION..

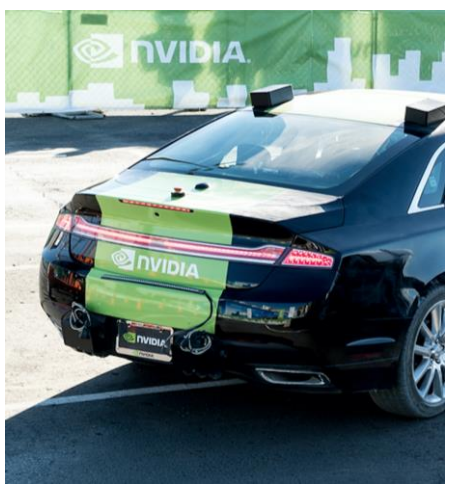
.. what else does this enable?

Medical Imaging, Segmentation

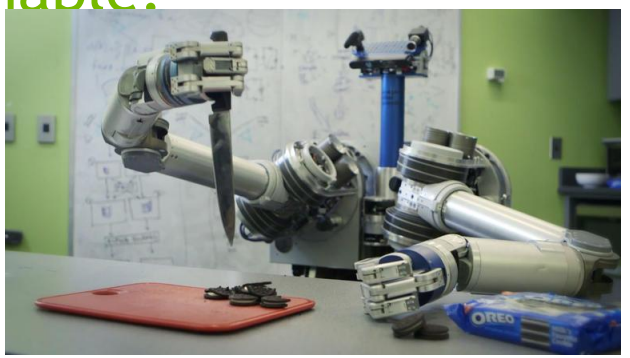


Voice Recognition





Autonomous Vehicles
Collision avoiding



Robots: Controlled collisions

Generative network

"Map your problem to images and you're in business"

HOW TO FIND THE TRAINING DATA?

Human in the loop

Manual Labelling

Controlled driving

Robot programming

NVIDIA.

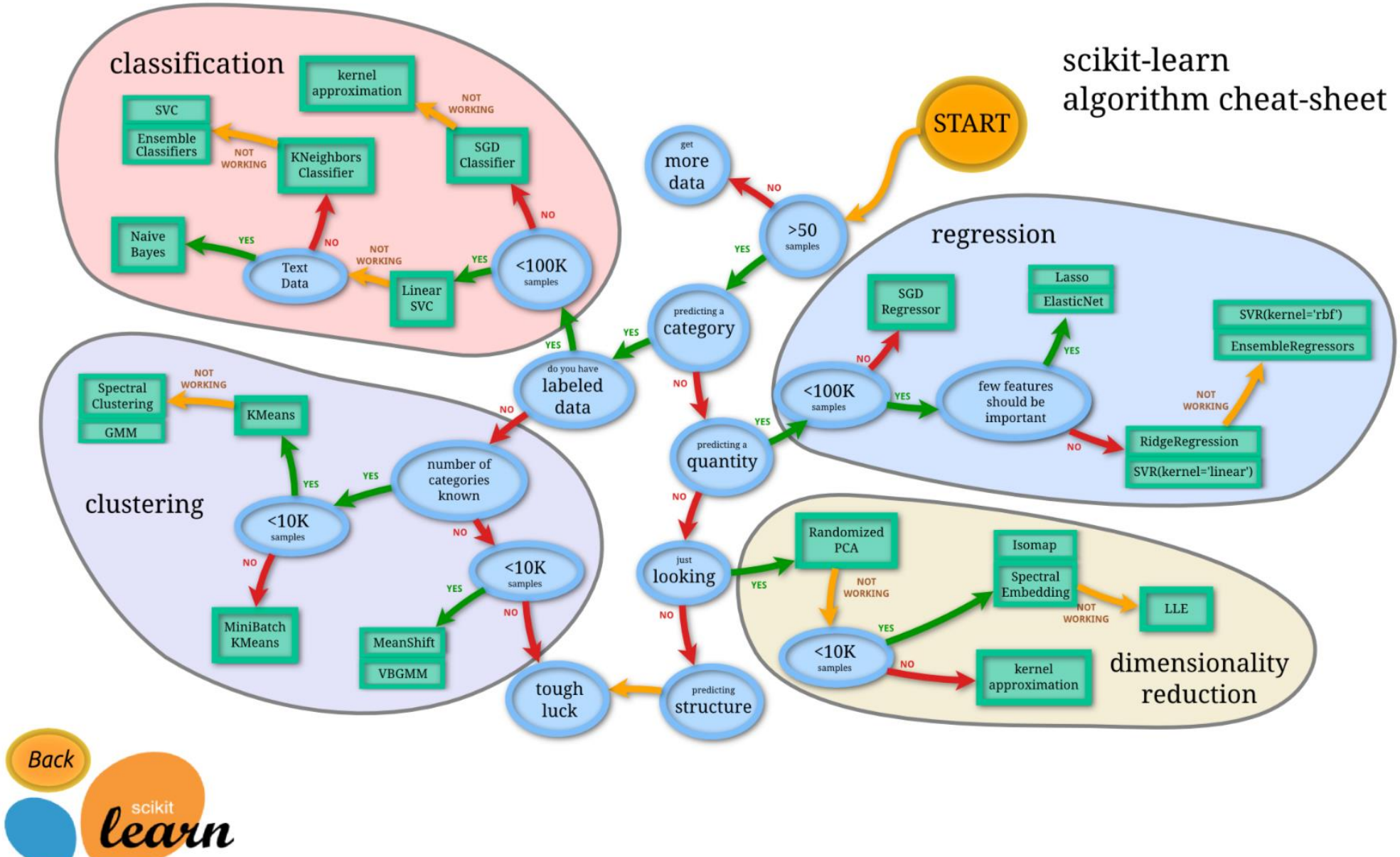
Train your self-driving car AI in *Grand Theft Auto V* – what could possibly go wrong?

Elon Musk-backed OpenAI reveals Universe – a universal training ground for computers

Reinforcement learning environment hosts virtualized video games and more

By Thomas Claburn in San Francisco 5 Dec 2016 at 06:00 22 SHARE

How to train a robot?


scikit-learn
algorithm cheat-sheet
START
get more data
>50 samples
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
classification
kernel approximation
SVC
Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Text Data
Linear SVC
<100K samples
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
<100K samples
few features should be important
RidgeRegression
SVR(kernel='linear')
clustering
Spectral Clustering
GMM
KMeans
number of categories known
<10K samples
MiniBatch KMeans
<10K samples
MeanShift
VBGMM
dimensionality reduction
Randomized PCA
Isomap
Spectral Embedding
LLE
<10K samples
kernel approximation
YES
NO
NOT WORKING
Back
scikit learn

DOMAIN CLASSIFICATION

Computational Mechanics	Earth Sciences	Life Sciences	Computational Physics	Computational Chemistry
Computational Fluid Mechanics	Climate Modeling	Genomics	Particle Science	Quantum Chemistry
Computational Solid Mechanics	Weather Modeling	Proteomics	Astrophysics	Molecular Dynamics
	Ocean Modeling			
	Seismic Interpretation			

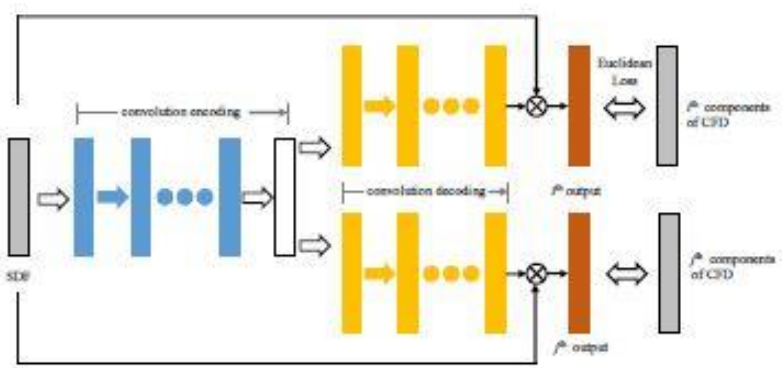

COMPUTATIONAL
MECHANICS

CONVOLUTIONAL NEURAL NETWORKS FOR STEADY FLOW APPROXIMATION

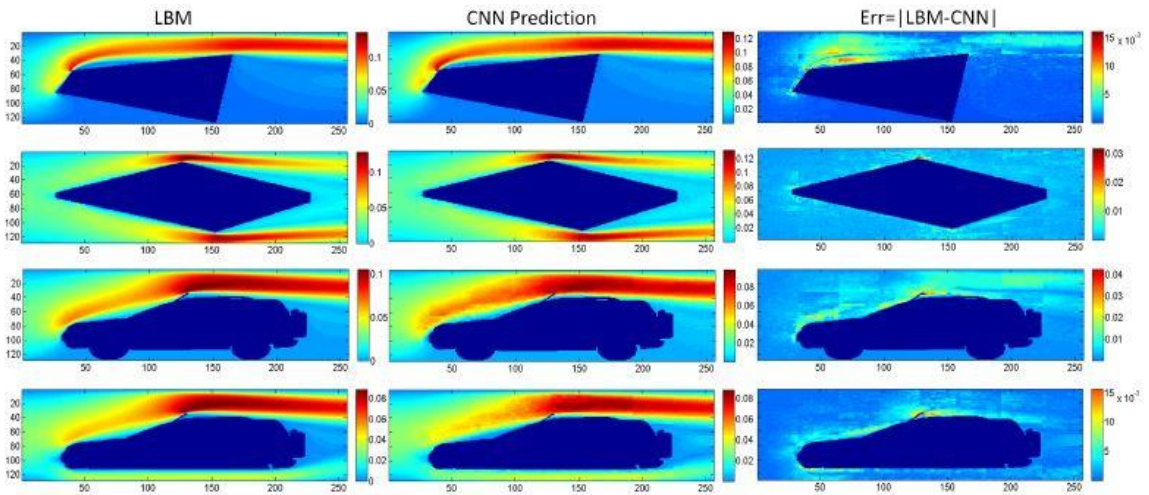
A quick general CNN-based approximation model for predicting the velocity field of non-uniform steady laminar flow by Guo, et al. (2016)

CNN-based approximation model trained by LBM simulation results

SFD data is used as import and error is used as loss function to train the convolutional neural networks.



CNN based CFD surrogate model architecture



Results comparison between LBM model and CNN based surrogate model

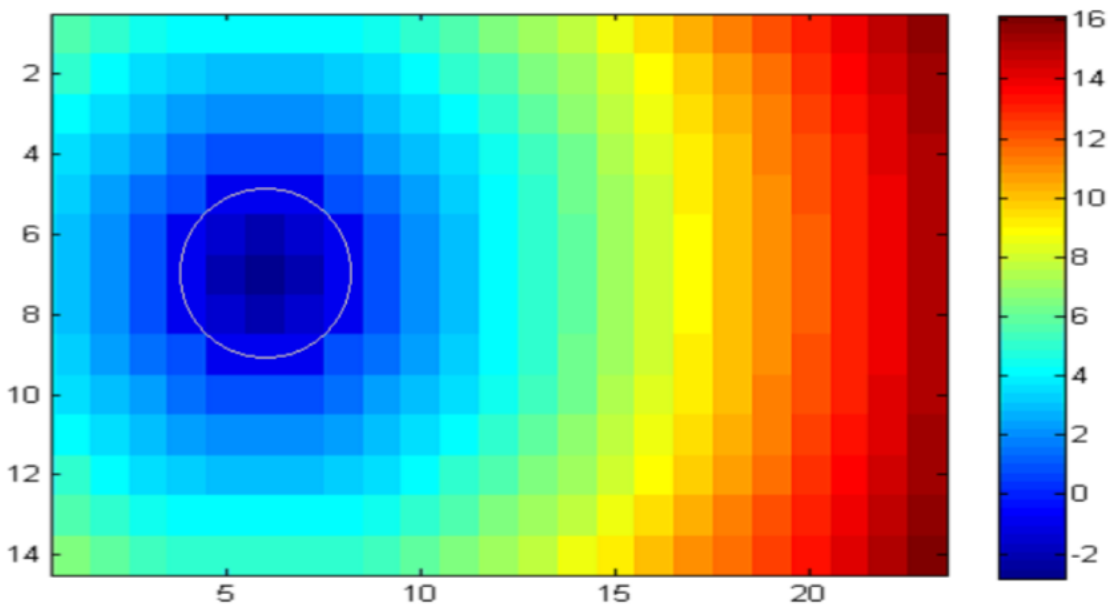
http://www.kdd.org/kdd2016/papers/files/adp1175-guoA.pdf

HOW TO ENCODE GEOMETRY?

Signed distance function $$D(i,j) = \min_{(i',j') \in Z} \left|(i,j) - (i',j')\right| \operatorname{sign}(f(i,j))$$

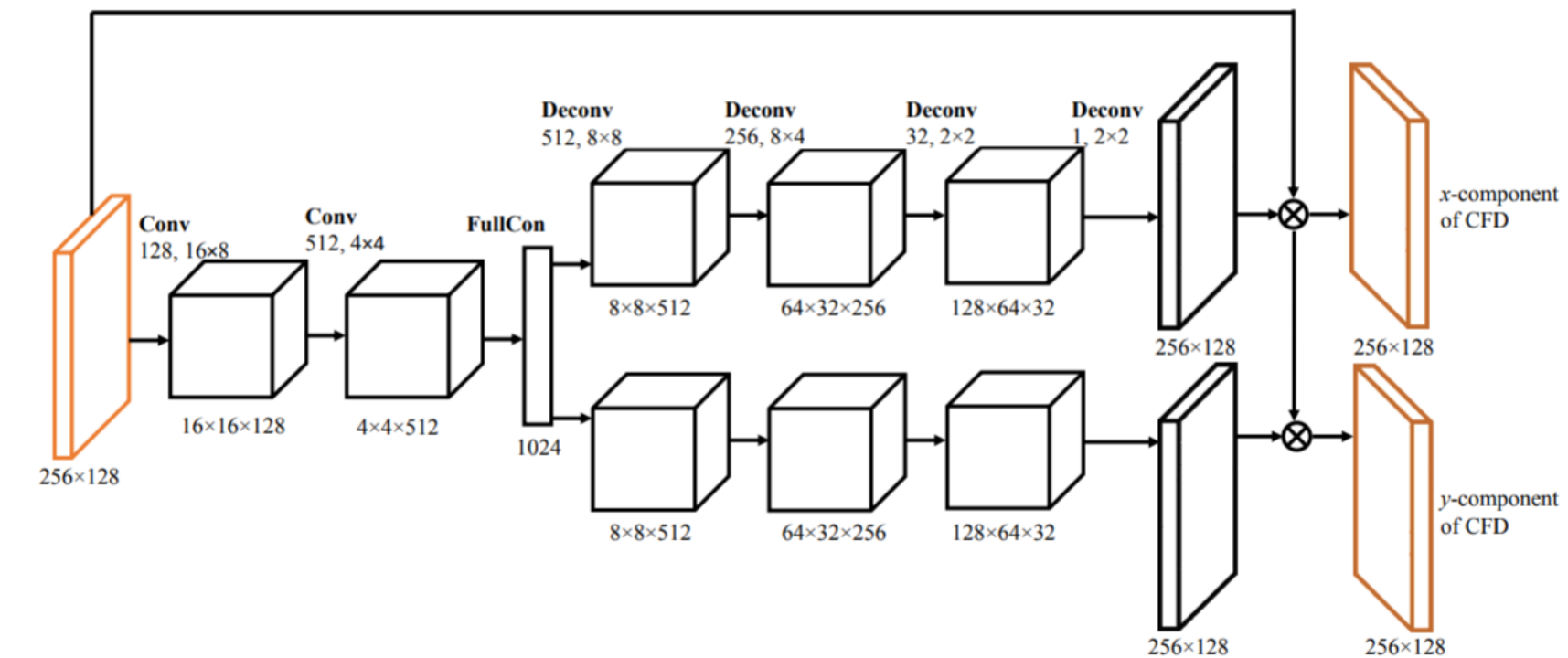
With zero level set $$Z = \left\{(i,j) \in R^2 : f(i,j) = 0\right\}$$

Discretized representation of geometry
Suitable for convolutions with pattern detection filters



Images from http://www.kdd.org/kdd2016/papers/files/adp1175guoA.pdf

NEURAL NETWORK ARCHITECTURE



CNN based CFD surrogate model architecture

Images from http://www.kdd.org/kdd2016/papers/files/adp1175guoA.pdf

HOW TO GENERATE TEST DATA?

LBM simulations for simple parametric shapes:
Triangles, quadrilaterals, pentagons, hexagons, dodecagons
256x128 point grid

100’000 samples training data set
10’000 samples validation dataset

Batch Size	Speedup (CPU)	Speedup (GPU)
1	5699	139
10	10732	262
100	11977	292

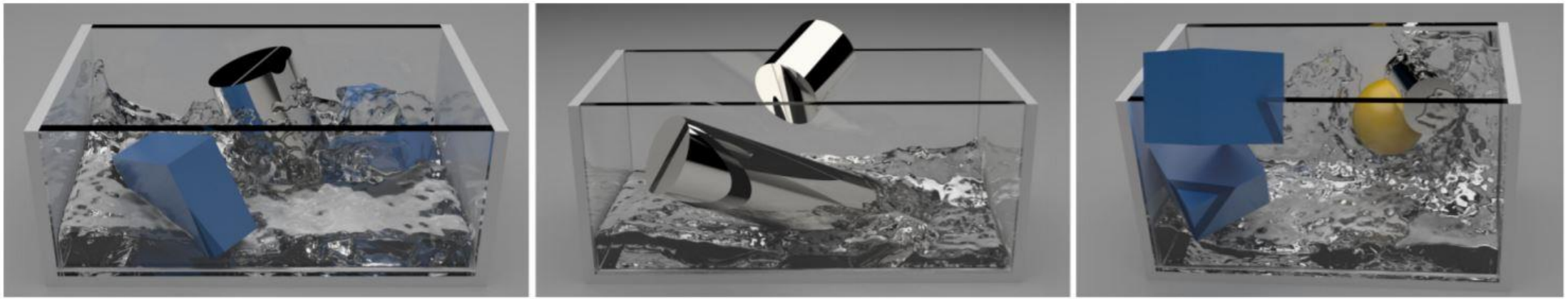
2D car test data set

Images from http://www.kdd.org/kdd2016/papers/files/adp1175guoA.pdf

INTERACTIVE FLUID SIMULATION WITH REGRESSION FOREST

Fluid Simulation with Trained Regression Forest [Ladicky et al, 2015]

- Regression Forested model trained with data generated with SPH method
- Realtime simulation generated by trained regression forest with GPU acceleration



Data driven fluid simulation using regression forests

EULERIAN FLUID SIMULATION WITH NEURAL-NETWORK

Accelerating Eulerian Fluid Simulation with Neuro-Networks

Acceleration of traditional Eulerian Fluid Simulation with Neuro-Network has been attempted by some researchers

The most computing costly pressure projection step is replaced with trained neuron-network

Convolutional Network has been tested and shown positive acceleration within reasonable error in the most recent publications.

Data Driven projection method in fluid simulation
[Cheng Yang et al. 2016]

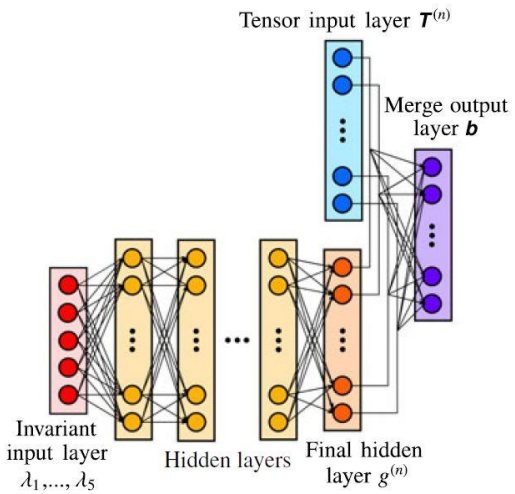
Accelerating Fluid Simulation with Convolutional Network [Tompson et al. ICML 2017]

TURBULENCE MODELING WITH MACHINE LEARNING TECHNIQUES

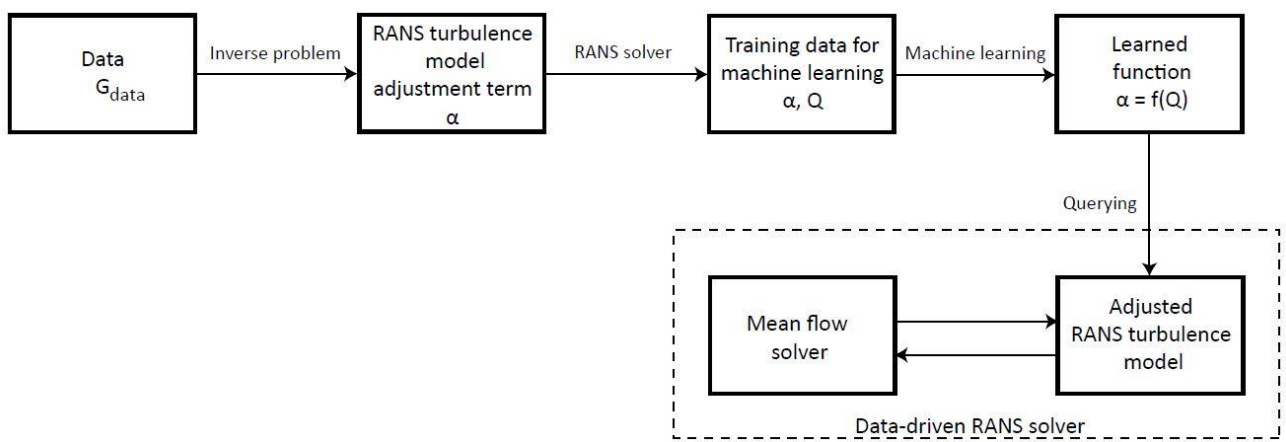
RANS method couple with machine learning techniques has been new frontier for turbulence modeling

The idea is to use machine learning techniques to learn from data generated by computational expensive DNS and add the term into RANS model to improve the accuracy of turbulence modeling

RANS results are used as import and DNS results are used as label to update the model.



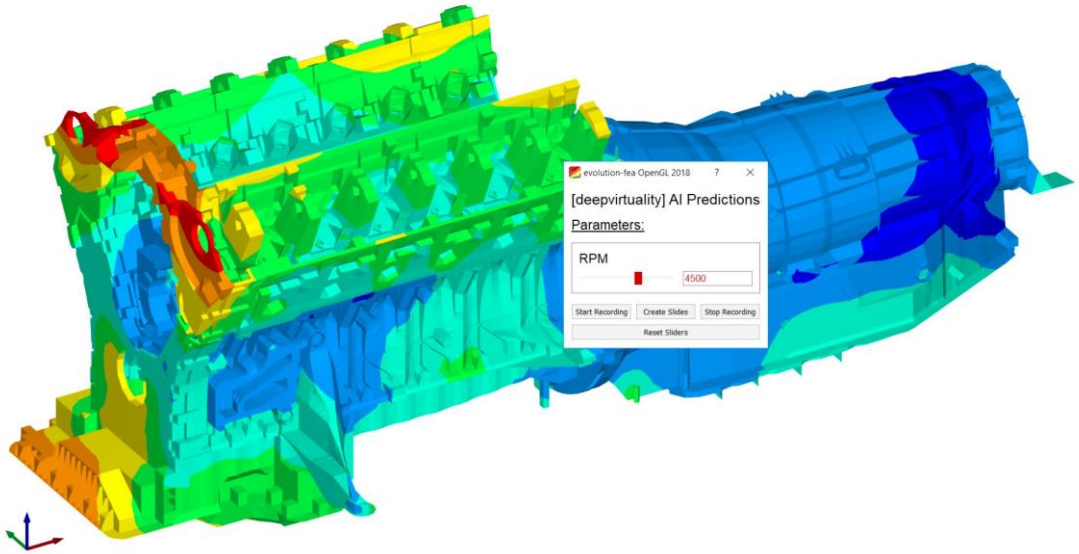
Tensor Basis Neural Network(TBNN) is proposed by Julia Ling and et al. (J.Fluid Mech 2016)



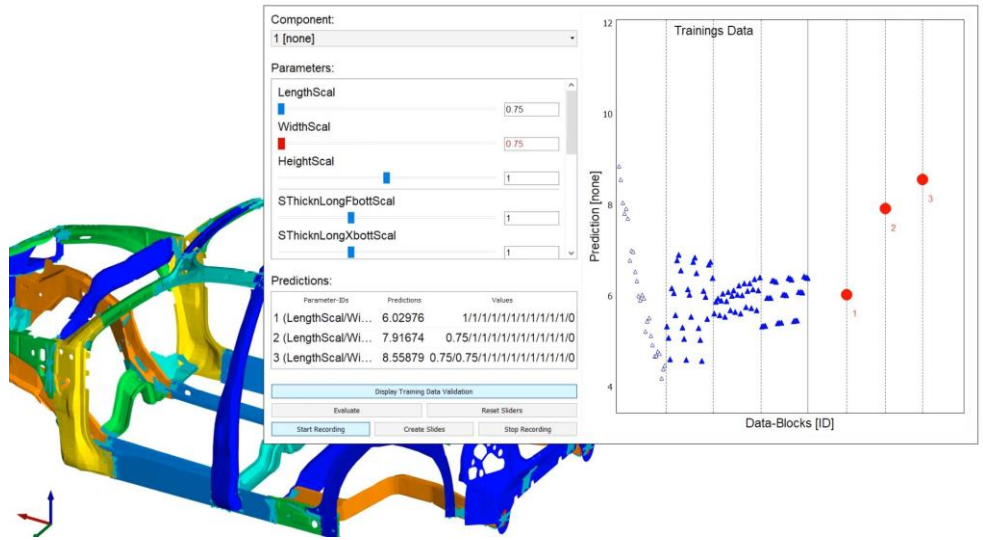
Inverse Modeling Framework propsed by Universrity Michigan from “Machine Learning Methods for Data-driven turbulence modeling”, Zhang and Duraisamy (2015)

FEA UPDATED WITH NEURAL NETWORK

FEA trained deep neural network for surrogate modelling of estimated stress distribution. Deepvirtuality, a spinoff from Volkswagen Data:Lab under Nvidia Inception Program has demonstrate with their software aimed for a quicker prediction of structural data.



An demonstration of Structure Born Noise of a V12 Engine with Deepvirtuality



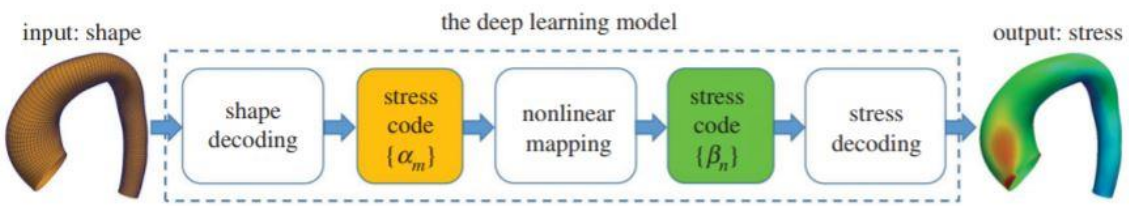
Torsional Frequencies of a Car Body by Deepvirtuallity

FEA UPDATED WITH NEURAL NETWORK IN BIO-TISSUE

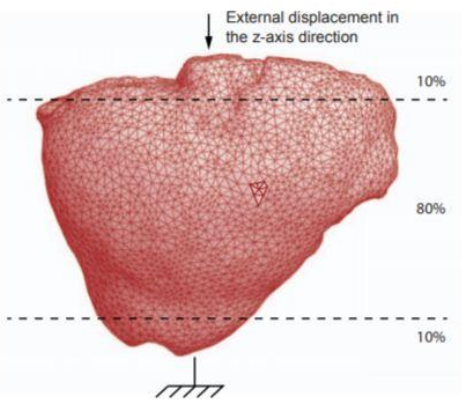
FEA trained deep neural network for surrogate modelling of estimated stress distribution. Traditional machine learning method has been used before, now deep learning techniques has been attempted for such model.

FEA generated stress distribution data is feed into neural network to train the neural network for fast stress distribution estimation. (Liang et al, 2018)

Ensembled decision tree model has also been applied for FEA update in “Machine Learning for modeling the biomechanical behavior of human soft tissue”. Data driven simulation has been done on Liver and Breast tissue. (Martin-Guerrereo, 2016)



Neural Network used for stress mapping by Liang Liang et al (2018)

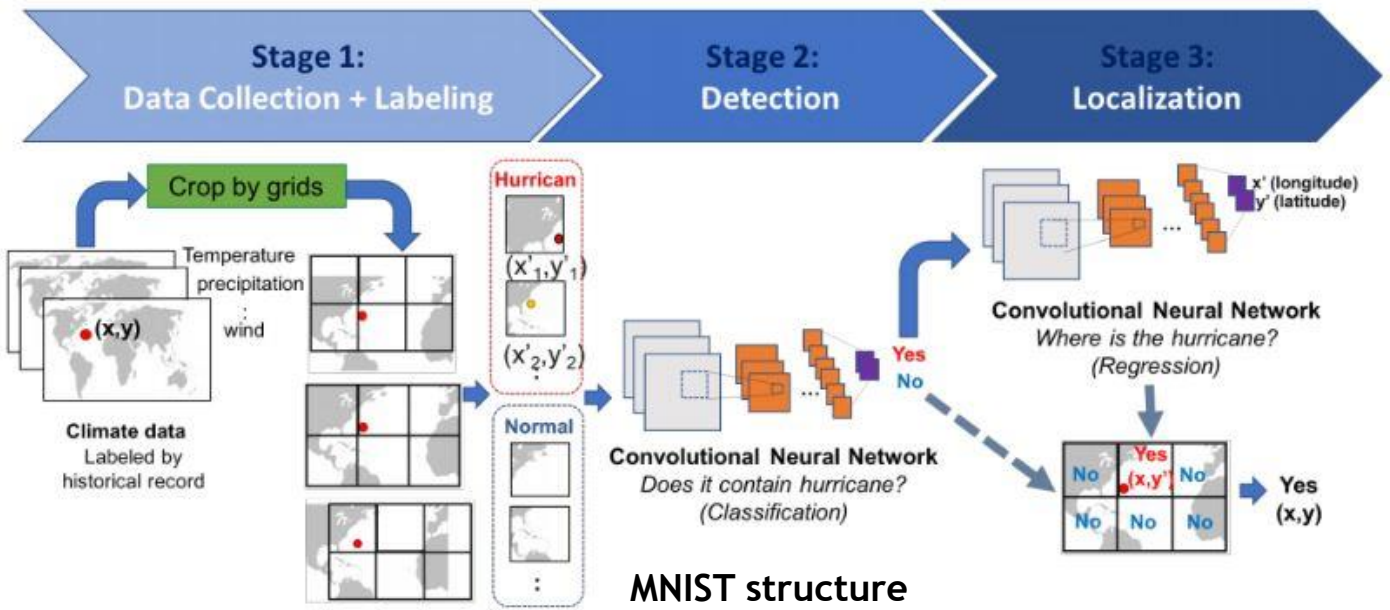


FEA model for Liver from “Machine Learning for modelling the biomechanical behavior of human soft tissue” (Martin-Guerrereo, 2016)

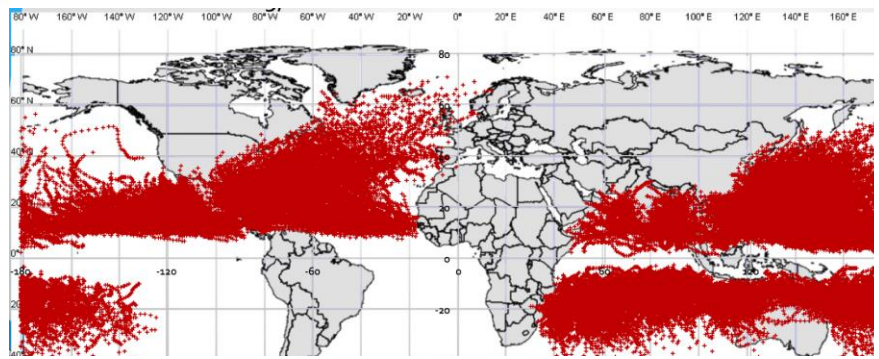

EARTH SCIENCES

ANOMALY DETECTION IN CLIMATE DATA

Identifying “extreme” weather events in multi-decadal datasets with 5-layered Convolutional Neural Network. Reaching 99.98% of detection accuracy. (Kim et al, 2017)



Systemic framework for detection and localization of extreme climate event

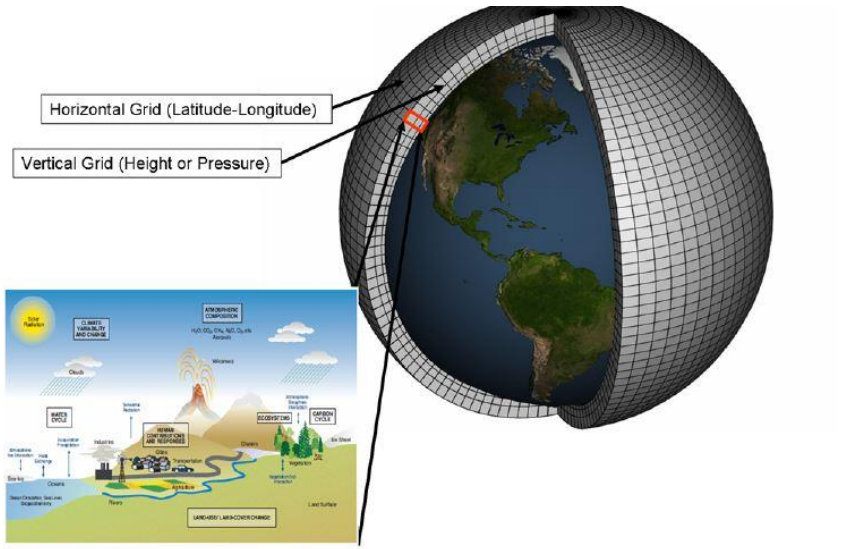


Dataset: Visualization of historic cyclones from JWTC hurricane report from 1979 to 2016

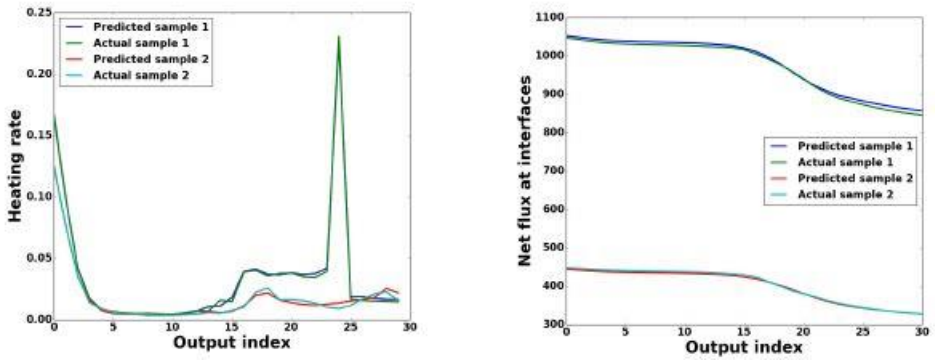
DOI 10.1109/ICDMW.2017.47

EMULATING RRTMG WITH DEEP NEURAL NETWORKS FOR THE ENERGY EXASCALE EARTH SYSTEM MODEL

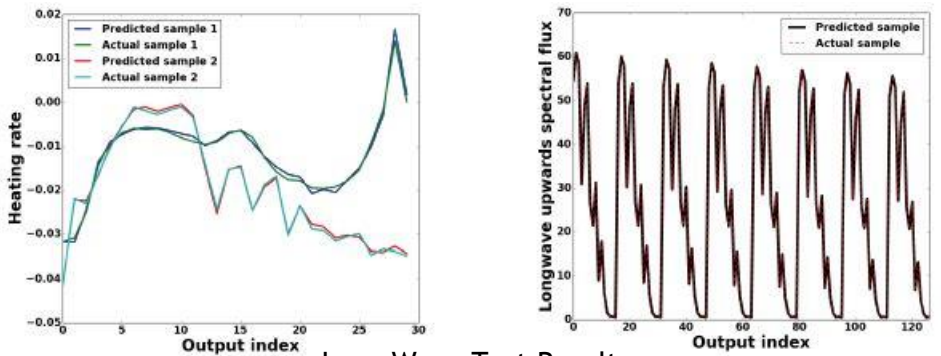
- Rapid Radiation Transfer Model for GCMs(RRTMG) is the most time consuming component of General Circulation Models(GCMs)
- Oak Ridge National Laboratory made use of Deep Neural Network to learn from RRTMG model.



GCM for climate modeling



Short Wave Test Results



Long Wave Test Results

EXAMPLE: FOG PREDICTION AT ZURICH AIRPORT

New data products via DL post-processing

Fog has operational impact

Not simulated by operational weather model

Require significantly higher resolution

Use deep learning to correlate fog conditions from weather simulation data

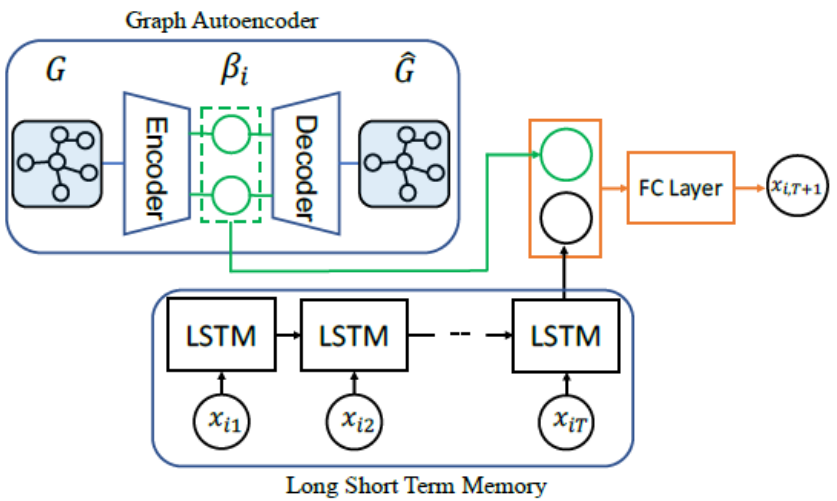
Historic weather forecast t+24hrs, correlation with observations

NVIDIA.

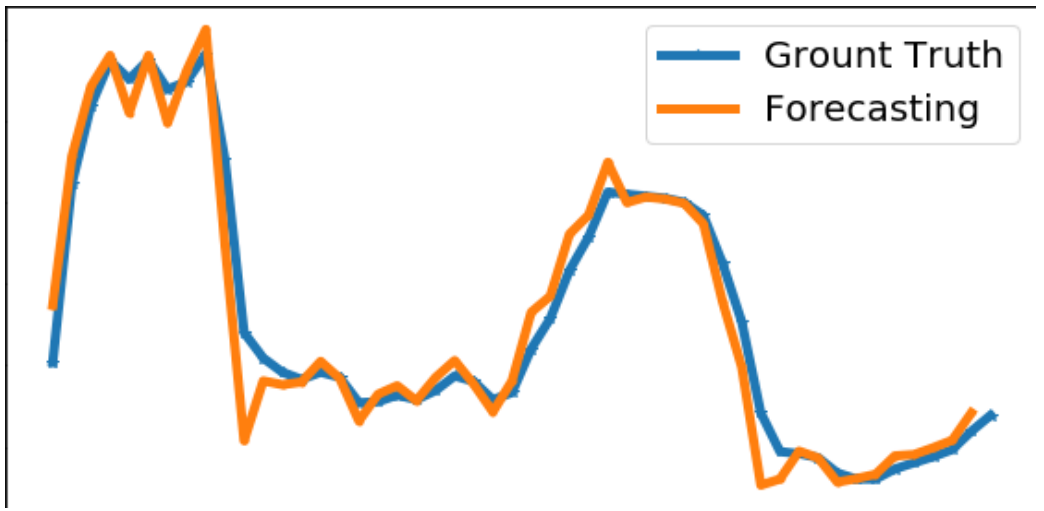
GAE + RNN FOR IMPROVED SPATIOTEMPORAL DATA

Multiple climate sets covering different places and time: Combining them is a huge challenge (Seo et al, 2017)

New network handles both spatial and temporal properties together to solve this problem.



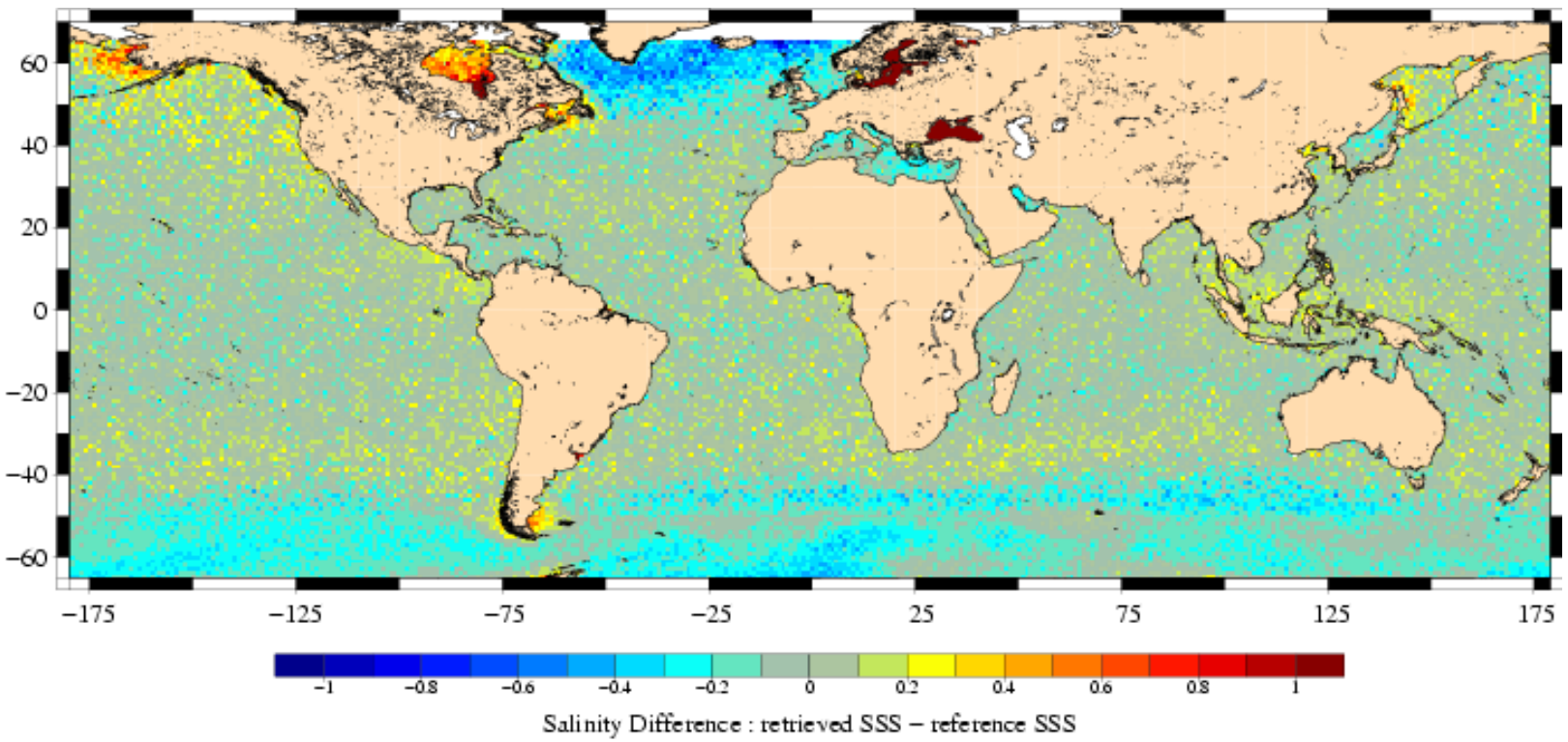
GAE-RNN model architecture



Forecasting of temperature

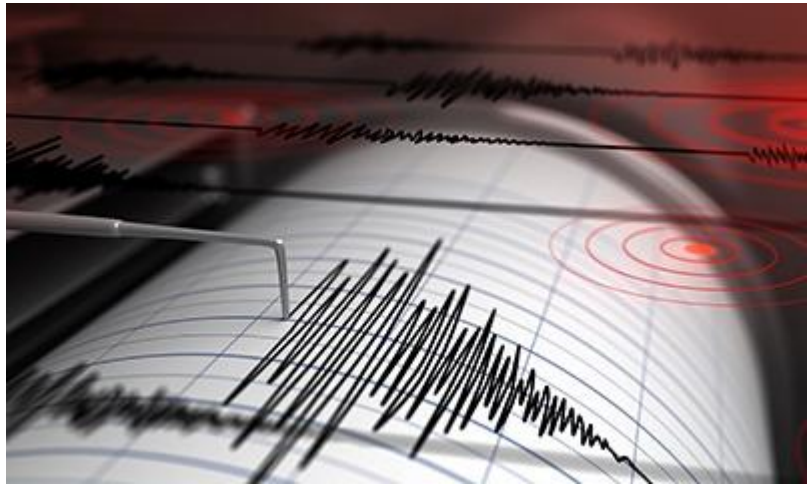
PREDICTING OCEAN SALINITY

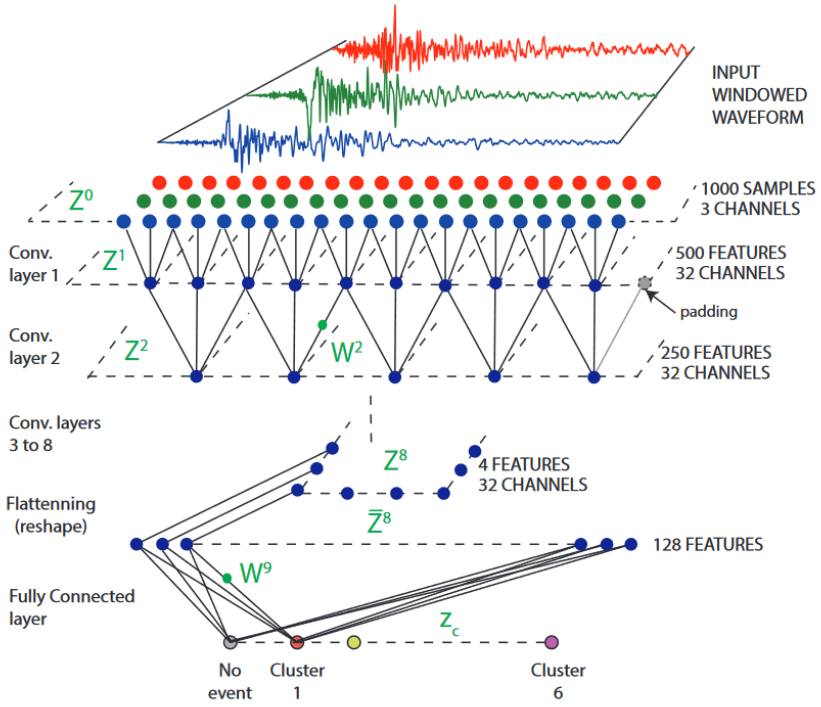
- Bharaskin, et al using a simple MLP with low resolution data
- Later work by Amman, et al uses surface brightness from Satellite images and ensembles to combine multiple DNNS - achieved > 97% accuracy



DEEP LEARNING FOR SEISMIC EVENTS

- Detecting earthquakes from seismic data [Perol, et al]
- 20x improvement in detection vs manual.
- Orders of magnitude faster






COMPUTATIONAL
PHYSICS

DEEP LEARNING IN HIGH ENERGY PHYSICS - CERN

Challenges:

- HL-LHC (High-Luminosity Large Hadron Collider) project, the ever increasing event complexity
- Model Independent Searches

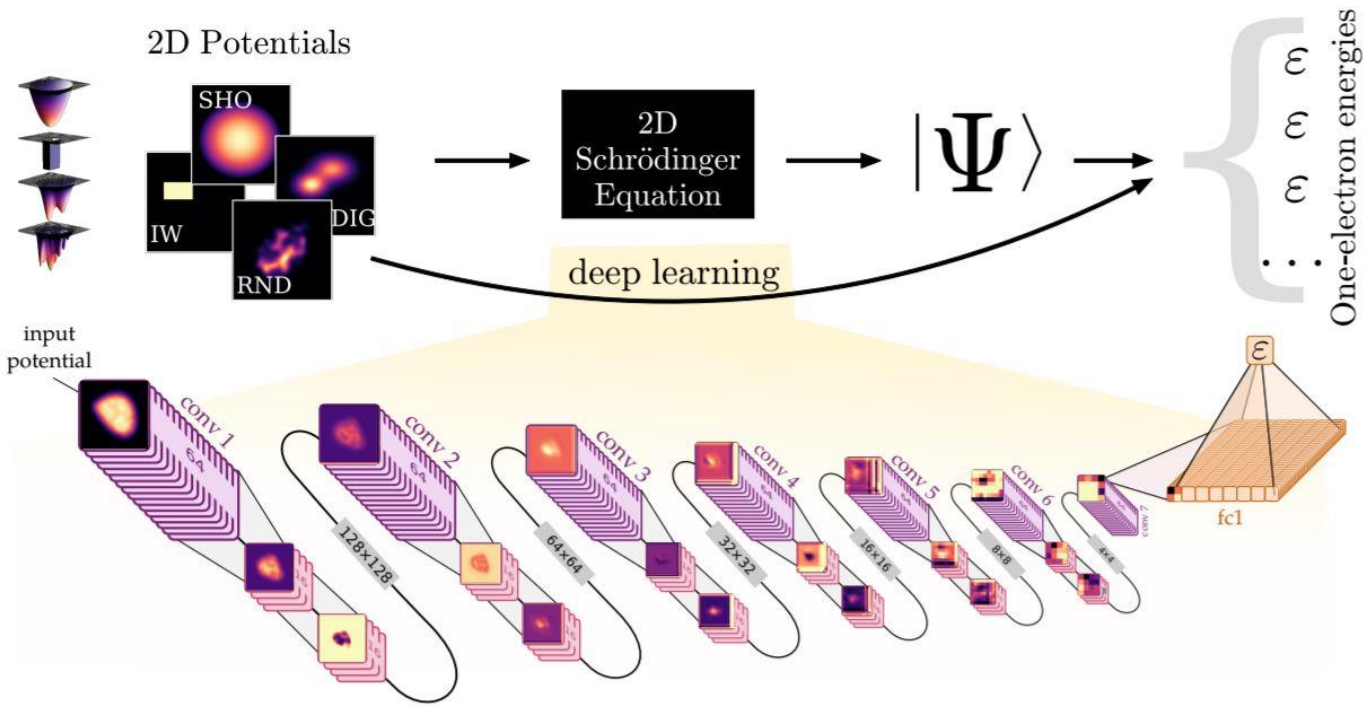
Deep Learning Solutions for:

- Triggering on rare signals
- Faster data processing and simulation
- Pattern recognition to extract physics content
- Lower Energy Computation
- Unsupervised Learning to Search for New Physics

DEEP LEARNING AND THE SCHRODINGER EQUATION

ConvNN is used to be trained for solving Schrodinger equation.

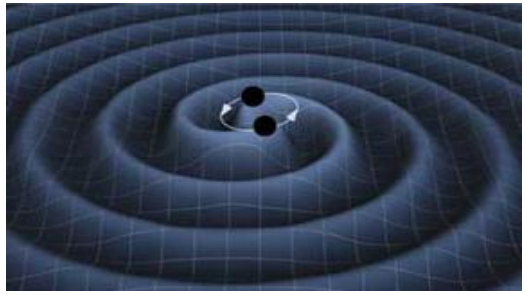
ConvNN is trained with simulation data to predict the ground-state energy of an electron in four classes of confining two-dimentional electrostatic potential



Deep learning and the Schrodinger equation

DEEP LEARNING FOR GRAVITATIONAL WAVE DETECTION

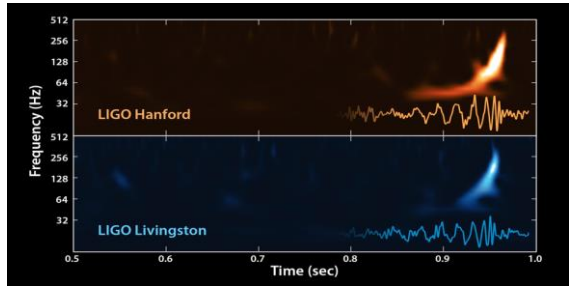
Deep learning method named deep filtering was used in the first detection of gravitational wave. Numerical simulated data was used for training deep filtering, a convolutional neural network to replace matched filtering. It provided 20X speed up on single core and potential to be accelerated further with GPU.



Gravitational wave due to black hole collide and merge

To be observed

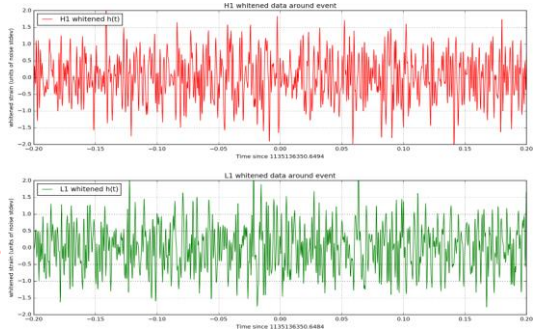
LIGO facility



Actual Signal Caused by Gravitational Wave

How to find The signal???

Deep Learning

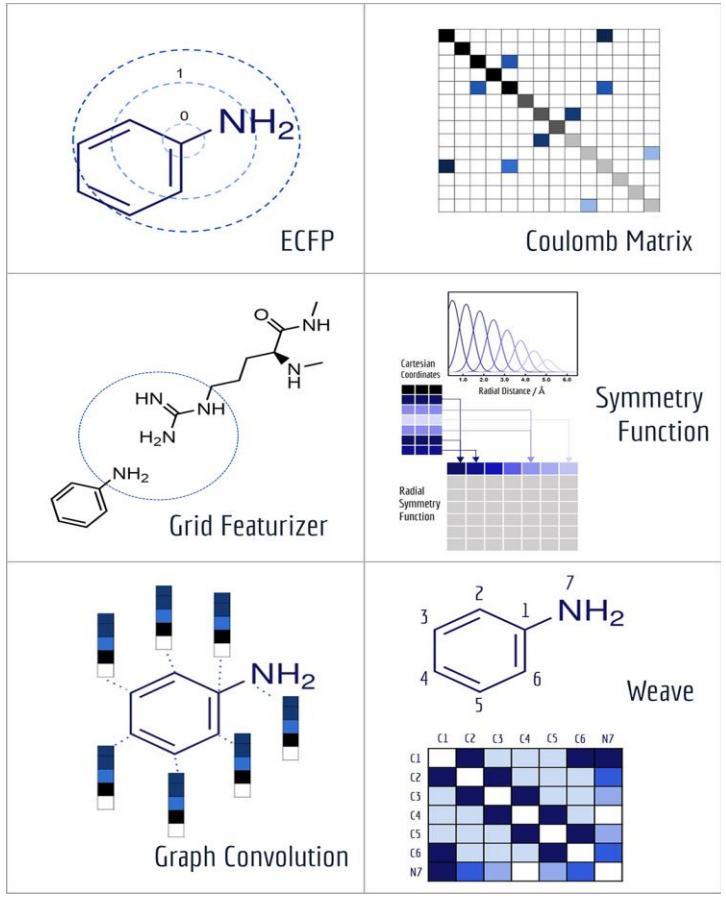


Actual observed data


COMPUTATIONAL
CHEMISTRY

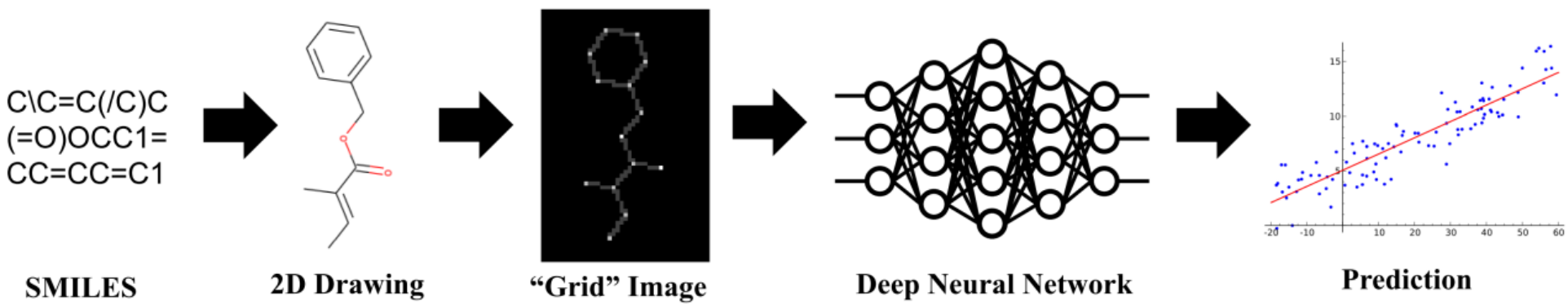
SHARING AND USING DATASETS FOR DRUG DISCOVERY

- DeepChem: Open Source framework for drug discovery using deep learning
- MoleculeNet: System for using/benchmarking using DeepChem – The “ImageNet” of Chemistry
  - “Smart” splitters for training/validation/testing
  - 17 curated datasets containing > 700,000 compounds
  - Selection of featurizers and models



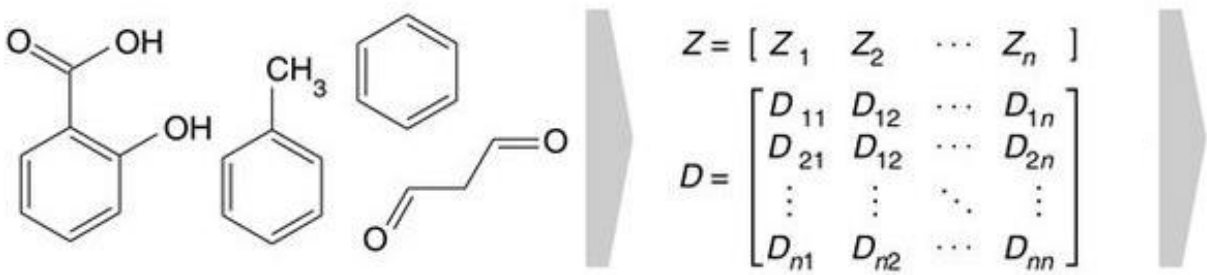
CHEMCEPTION: THE QC VERSION OF INCEPTION

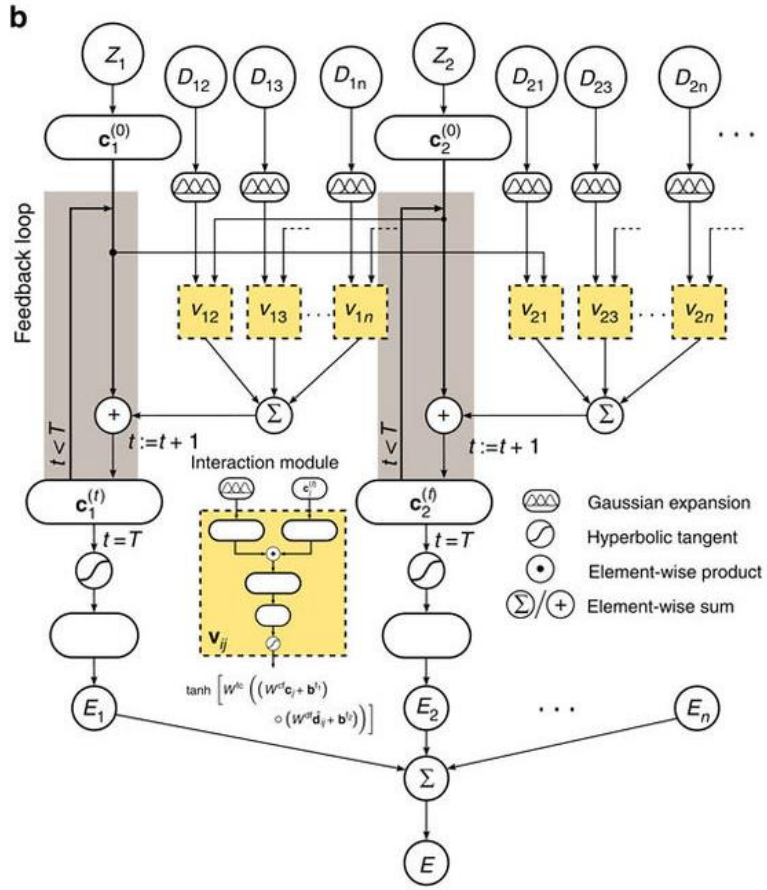
- Goh, et al. devised a CNN called Chemception that can perform all predictive requirements (toxicity, activity, solvation) as well as current expert QSAR/QSPR for a complex molecule (HIV) after being trained for only 24 hours on a single GTX 1080



PREDICTING MD ENERGIES (1)

- Schutt, et al. devised a DTNN that can calculate the chemical space for a medium-sized molecule with an max error of 1 kcal/mol.

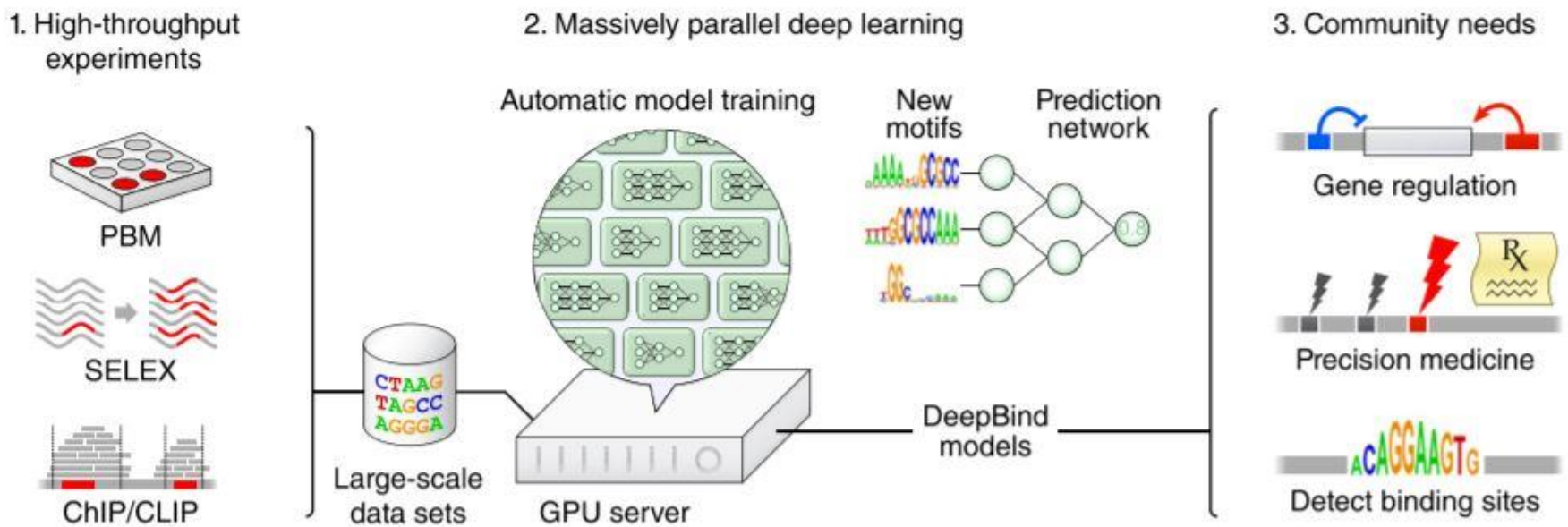





LIFE SCIENCES

PREDICTING OF SEQUENCE SPECIFIES OF DNA- AND RNA BINDING PROTEINS

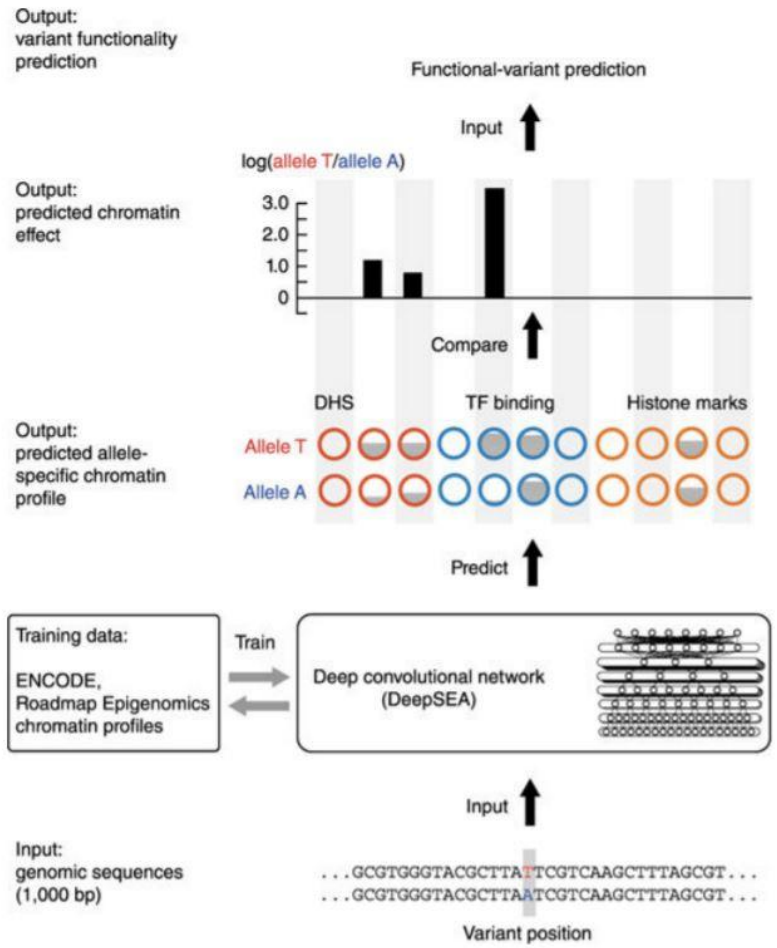
- DeepBind was proposed and build by [Babak et al, Nature Biotechnology] for predicting of sequence specifies of DNA- and RNA-binding proteins.



DeepBind's input data, training procedure and applications

CREATING A UNIVERSAL SNP AND SMALL INDEL VARIANT CALLER WITH DEEP NEURAL NETWORKS

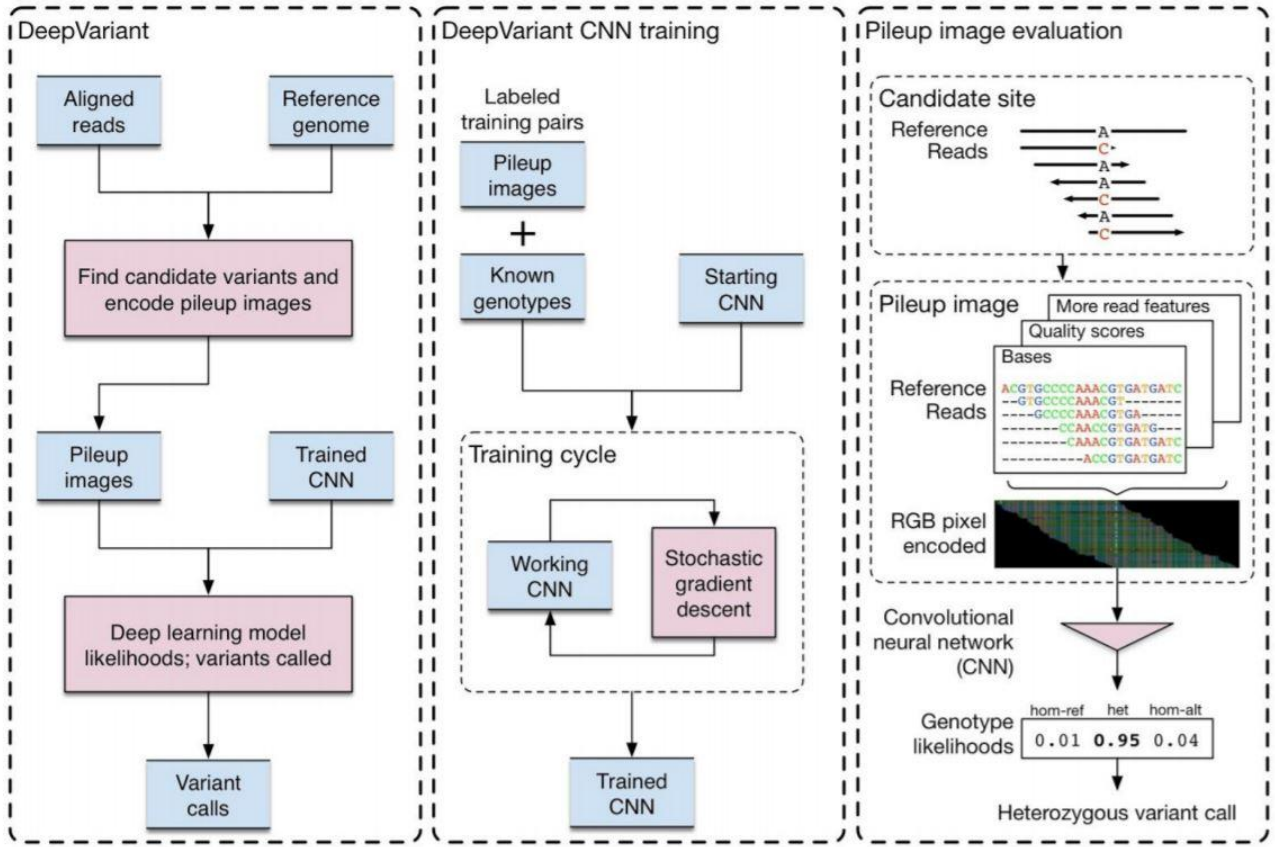
- [Zhou, et al, Nature Methods] proposed another deep learning based model for predicting effects of noncoding variants named DeepSEA.
- The core of DeepSEA is a typical convolutional neural network trained with ENCODE, Roadmap Epigenomics chromatin profiles



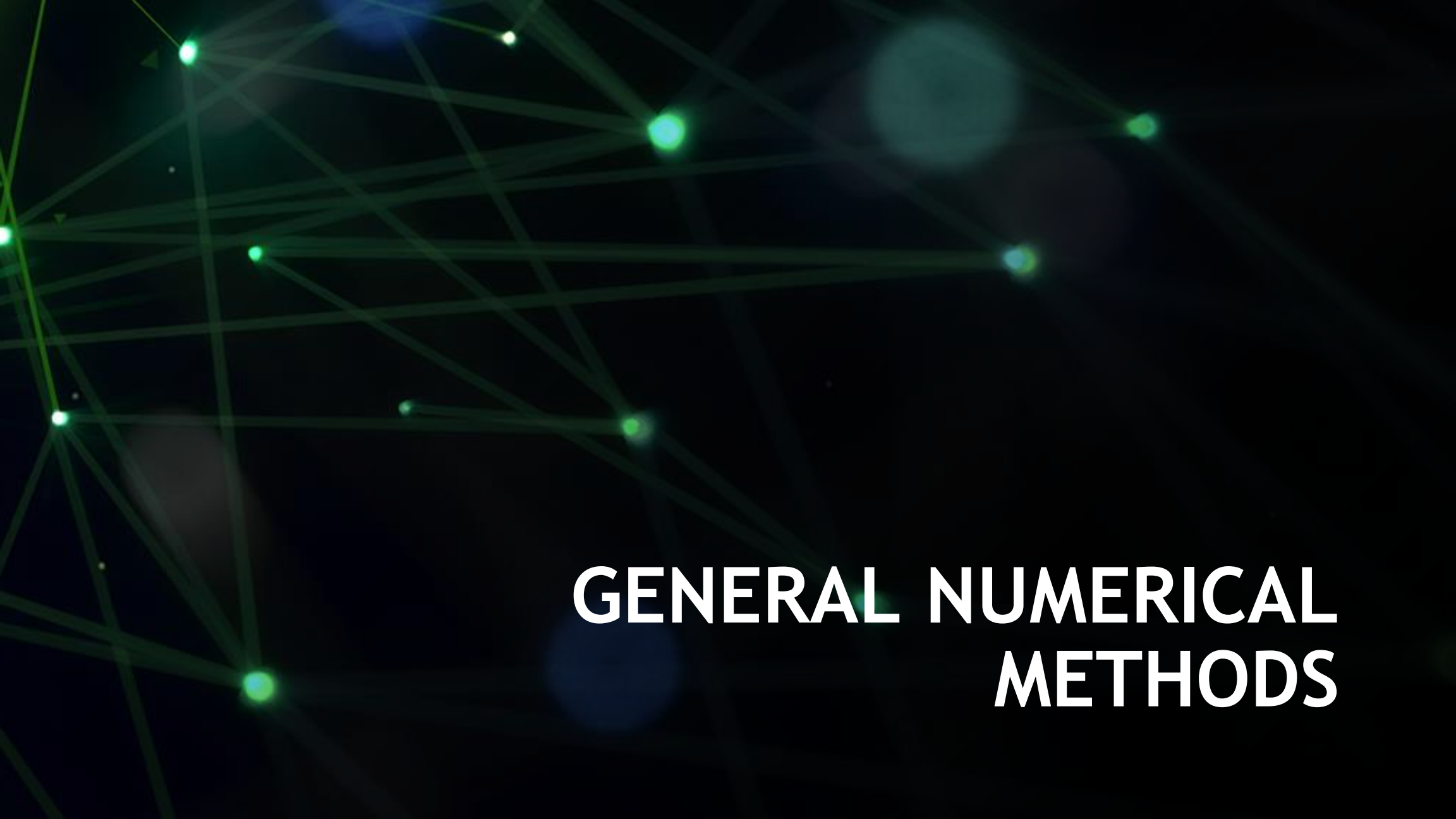
DeepSEA framework

CREATING A UNIVERSAL SNP AND SMALL INDEL VARIANT CALLER WITH DEEP NEURAL NETWORKS

- DeepVariant is proposed and build by Ryan, et al for rapid determination of genetic variants in an individual's genome with billions of short and errorful sequence reads.
- It out performed statistical models handcrafted by thousands of researchers in decades.



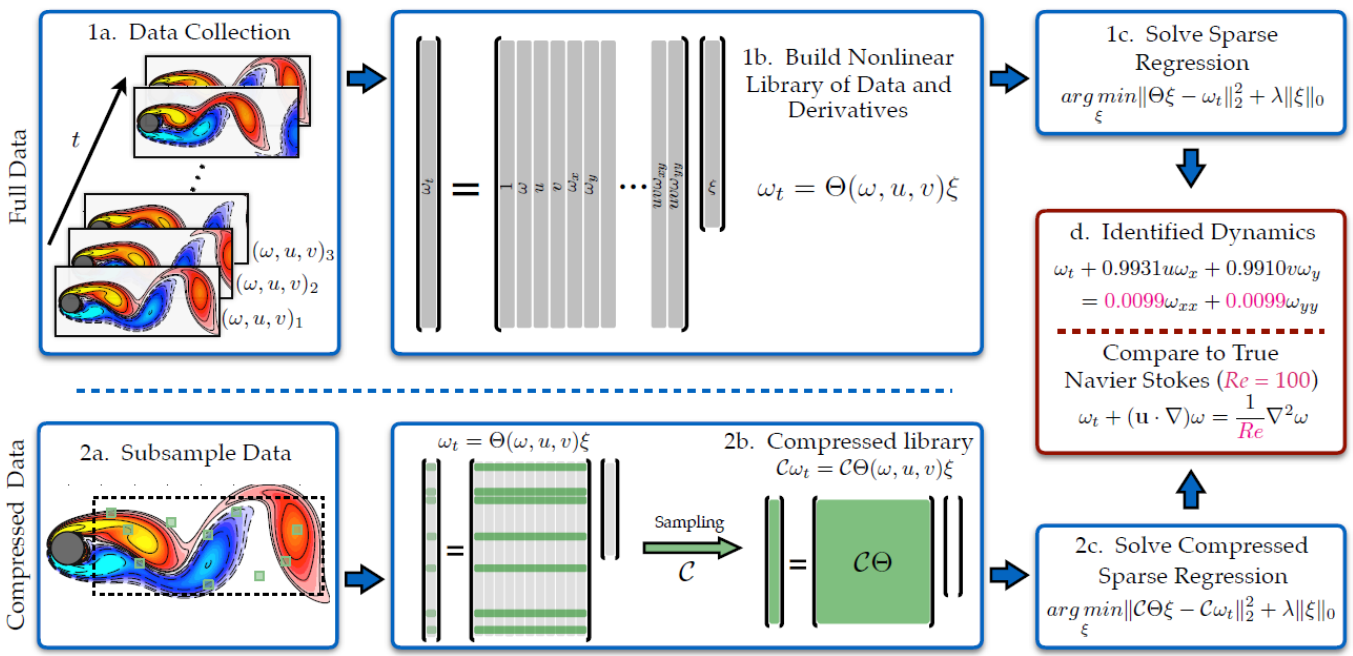
DeepVariant Framework



GENERAL NUMERICAL METHODS

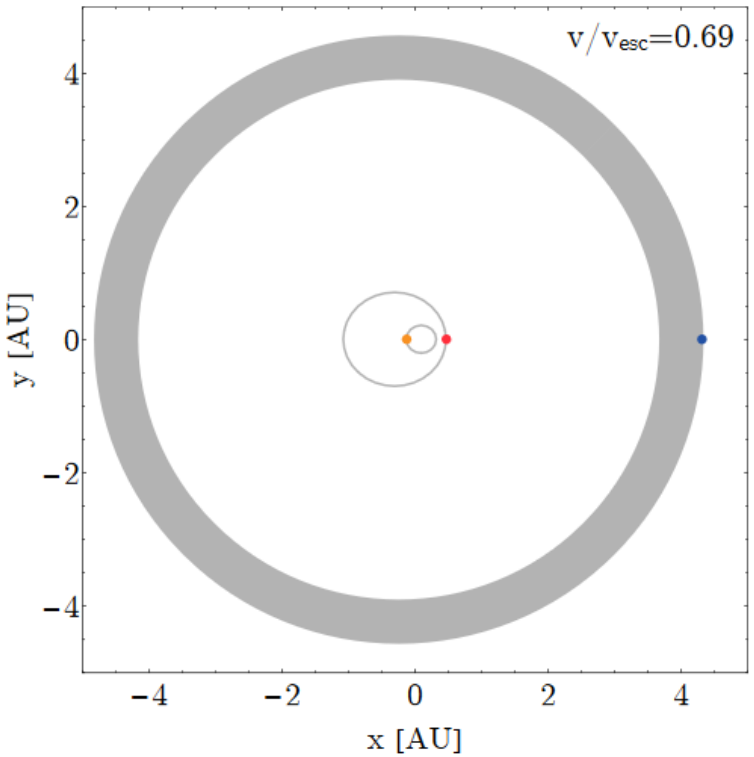
PDE-FIND: DATA-DRIVEN DISCOVERY OF PARTIAL DIFFERENTIAL EQUATIONS

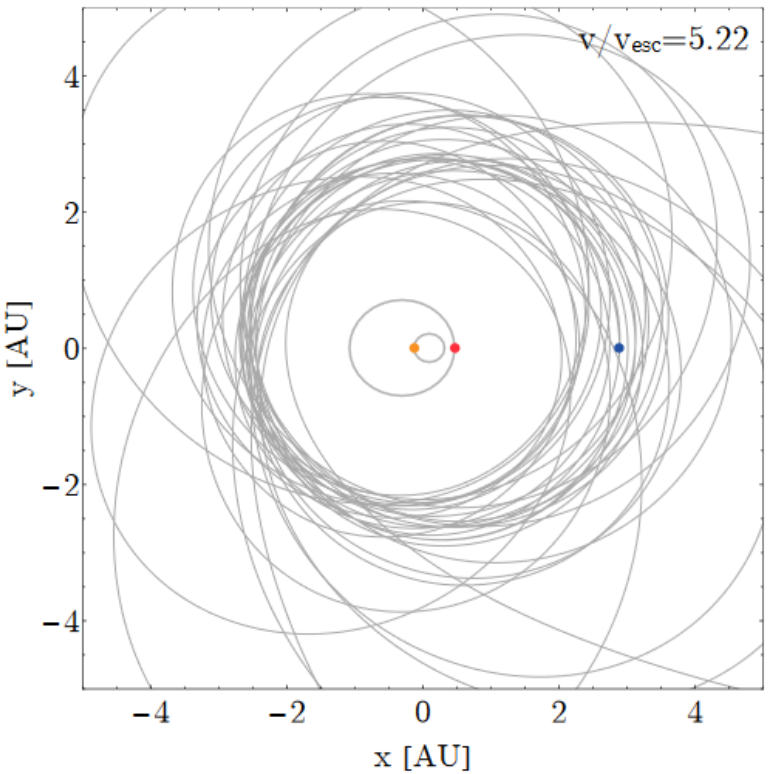
- Rudy, et al. Create a large internal library from data and derivatives. Select an active subset from the library and apply sparse regression techniques to solve the PDE.



DEEP LEARNING FOR N-BODY SIMULATION

- Calculating stability predictions for complex orbits [Lam & Kipping]



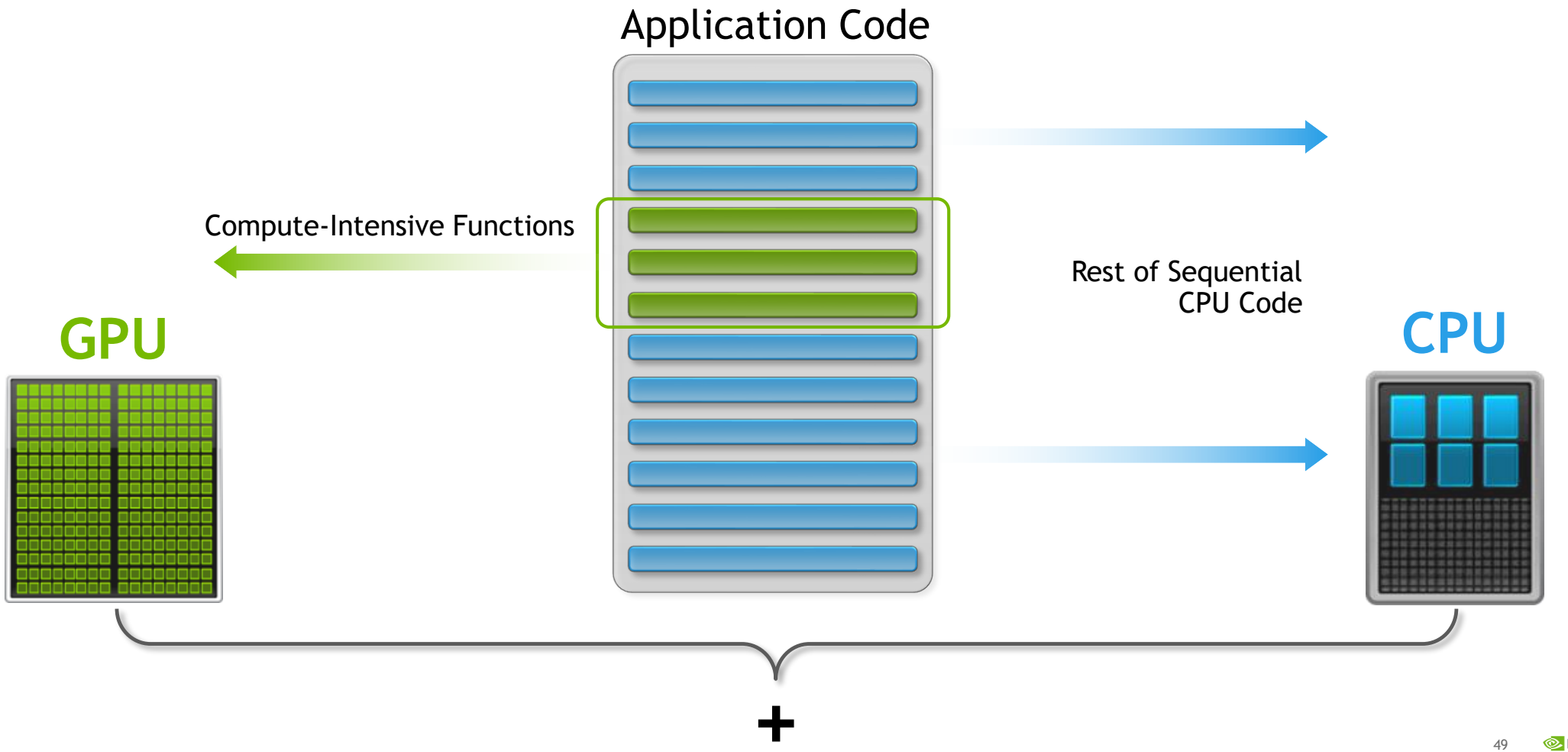


SUMMARY

- Broad range of problem domains tackled with DL
- CNN are fantastic at image “understanding”
- Many science problems map to these patterns
- Successful approaches often map to the problem to relatively simple representaions

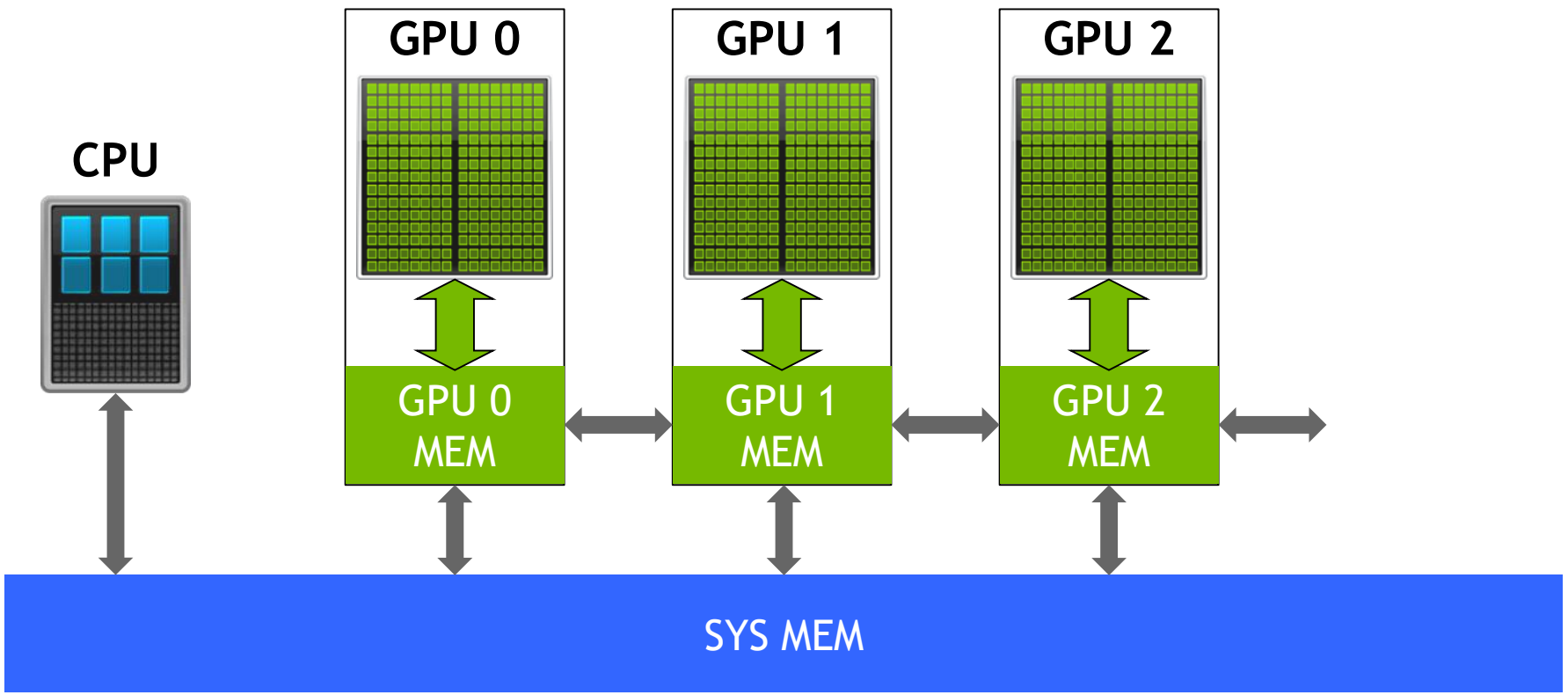
GPU ARCHITECTURE

HOW GPU ACCELERATION WORKS



NVIDIA.

HETEROGENEOUS ARCHITECTURES

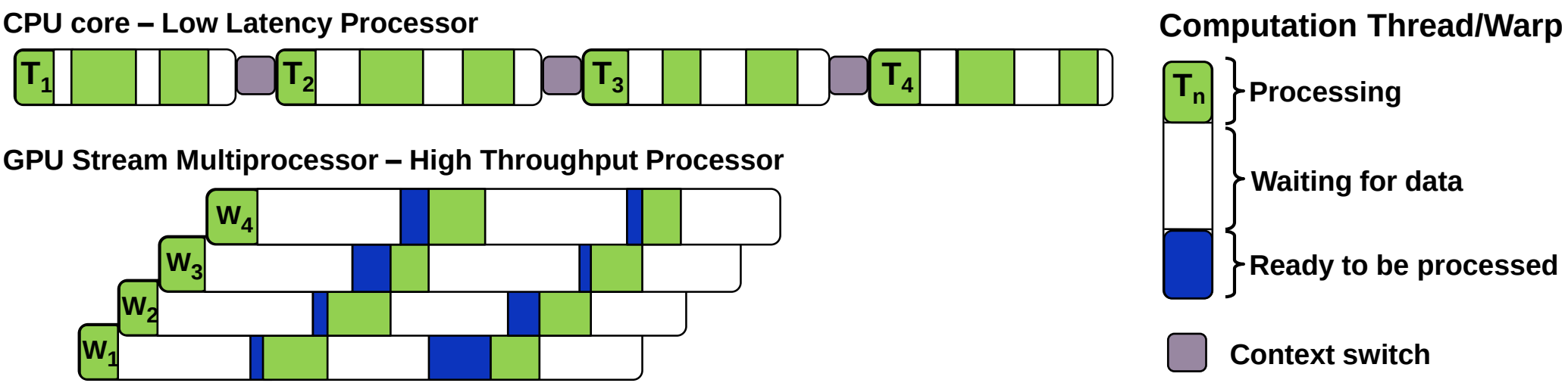


NVIDIA.

LOW LATENCY OF HIGH THROUGHPUT?

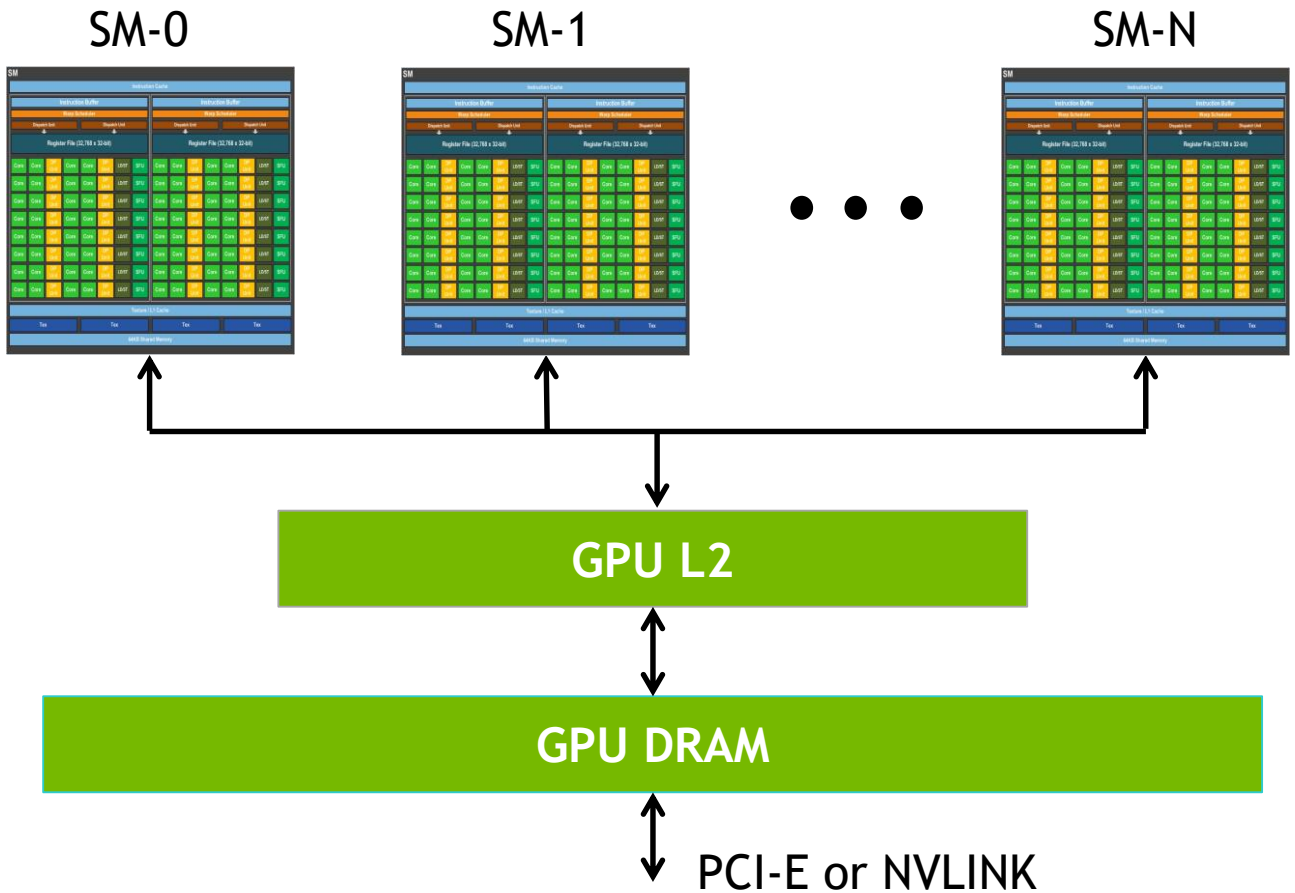
CPU architecture must **minimize latency** within each thread

GPU architecture **hides latency** with computation from other threads (warps)



NVIDIA.

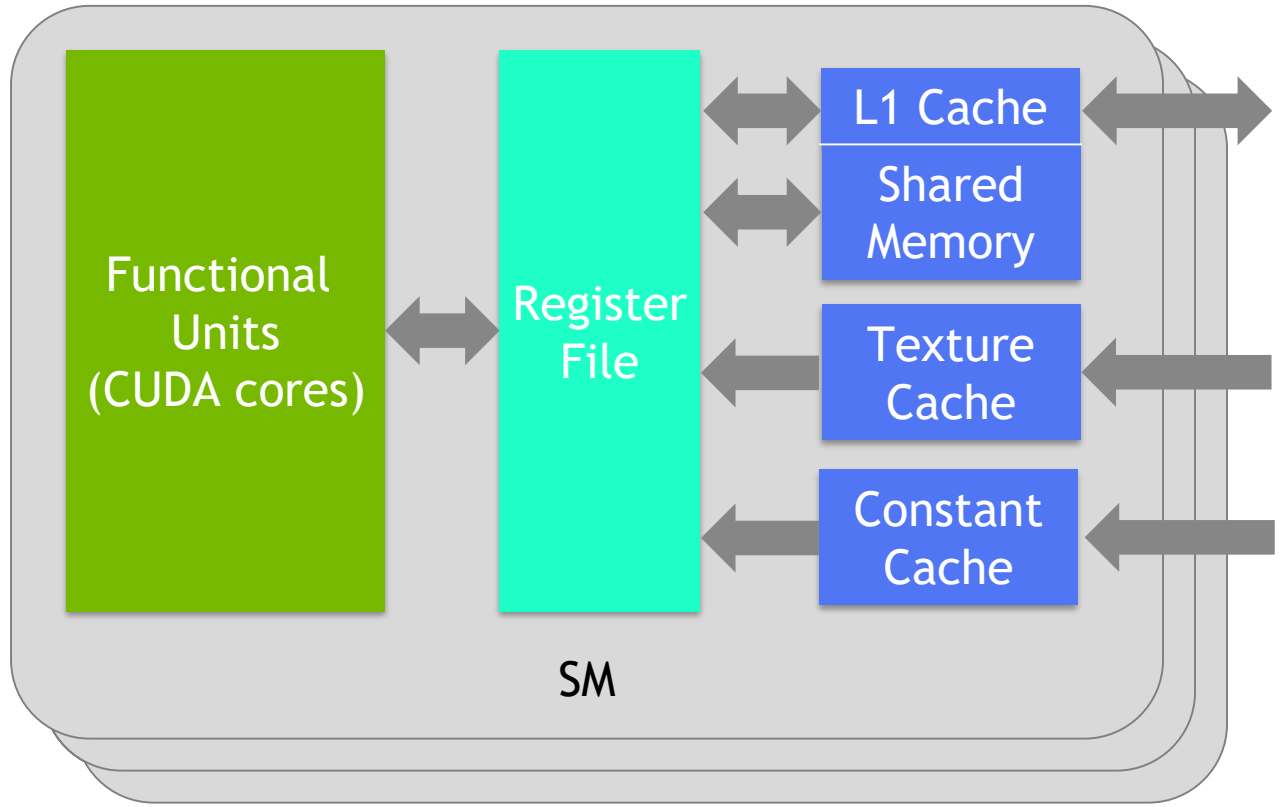
GPU ARCHITECTURE



GPU SM ARCHITECTURE

Kepler SM

	GK110
FP32 Cores	192
FP64 Cores	64
Register File	256 KB
Shared Memory	16/32/48 KB



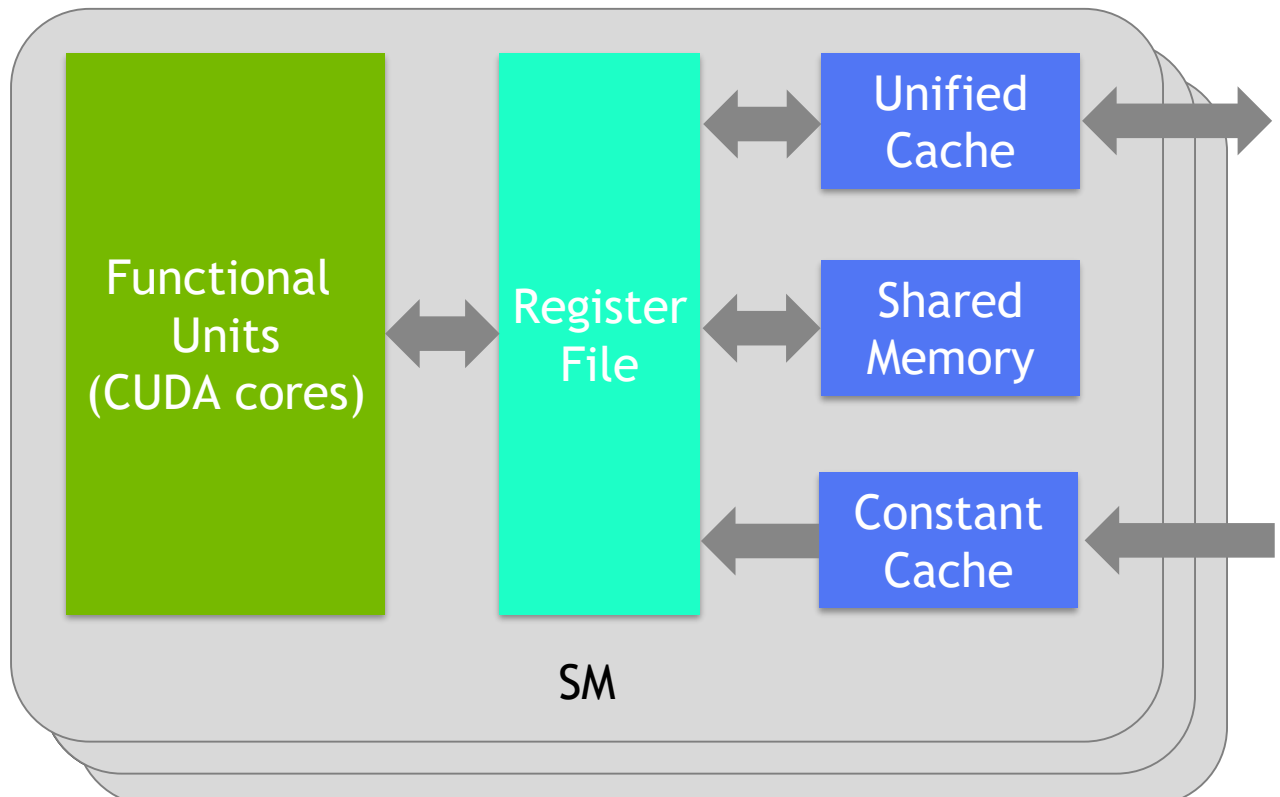
15 SMs on Tesla K40

NVIDIA

GPU SM ARCHITECTURE

Pascal SM

	GP100
FP32 Cores	64
FP64 Cores	32
Register File	256 KB
Shared Memory	64 KB



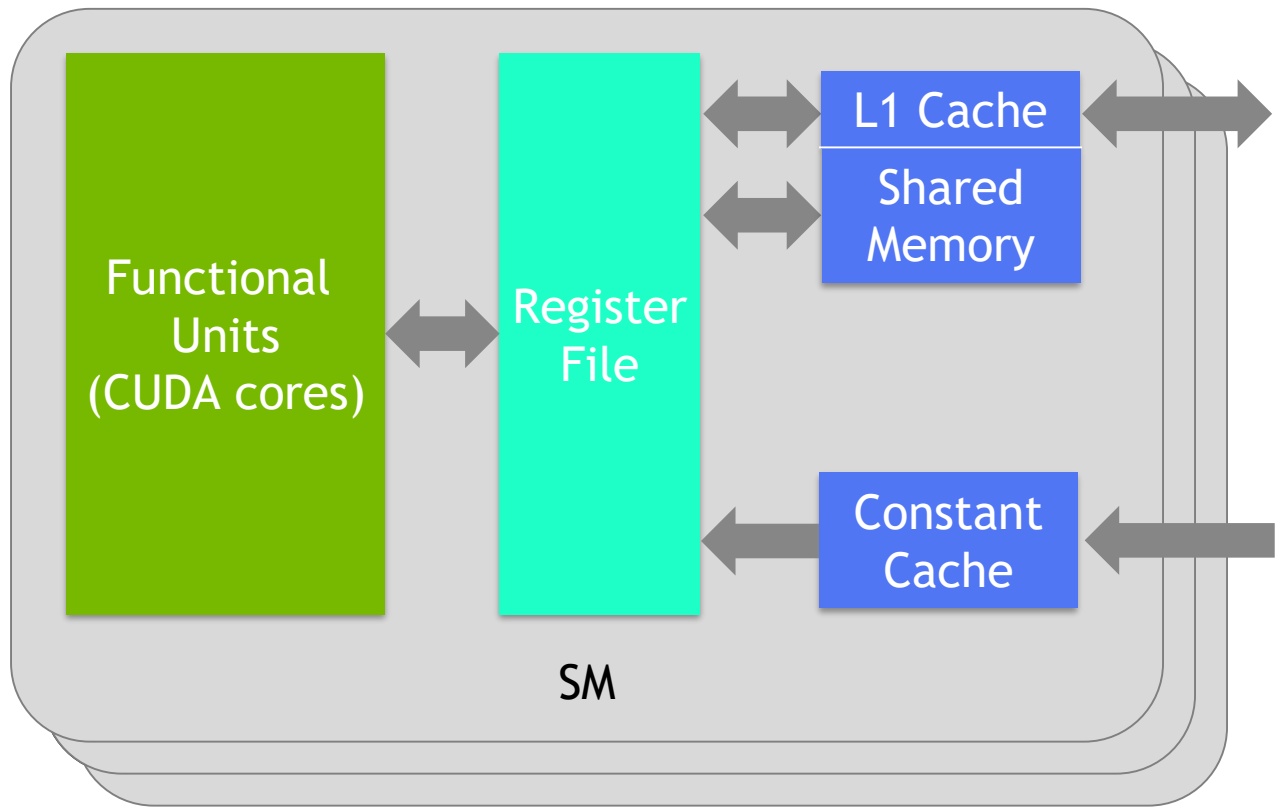
56 SMs on Tesla P100

NVIDIA

GPU SM ARCHITECTURE

Volta SM

	GV100
FP32 Cores	64
FP64 Cores	32
Tensor Cores	8
Register File	256 KB
Shared Memory	up to 96 KB



80 SMs on Tesla V100

NVIDIA.

TESLA FAMILY

GPU comparison (boost clocks)

	Tesla K40	Tesla P100	Tesla V100
Peak FP32 (TFLOP/s)	5.04	10.6	15
Peak FP64 (TFLOP/s)	1.68	5.3	7.5
Peak Tensor Core (TFLOP/s)	N/A	N/A	120
Memory Size (GB)	12	16	16
Memory Bandwidth (GB/s)	288	732	900

NVIDIA.

NVIDIA TESLA V100

- 21B transistors
  815 mm^2, 12nm FFN
- 80 SM
  5120 CUDA Cores
  640 Tensor Cores
- 7.8 FP64 TFLOPS
- 15.6 FP32 TFLOPS
- 125 Tensor TFLOPS
- 16 GB HBM2
  900 GB/s memory bandwidth
- 300 GB/s NVLink bandwidth

*full GV100 chip contains 84 SMs

NVIDIA.

TENSOR CORE

Mixed Precision Matrix Math - 4x4 matrices

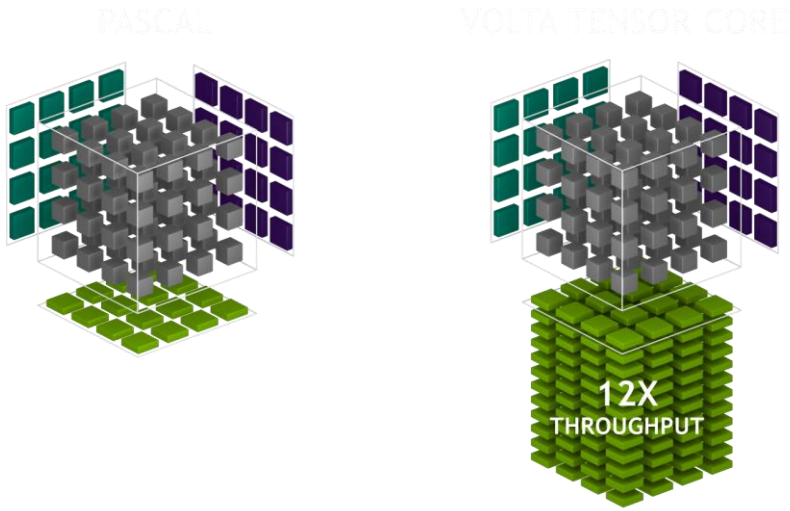
New CUDA TensorOp instructions
& data formats

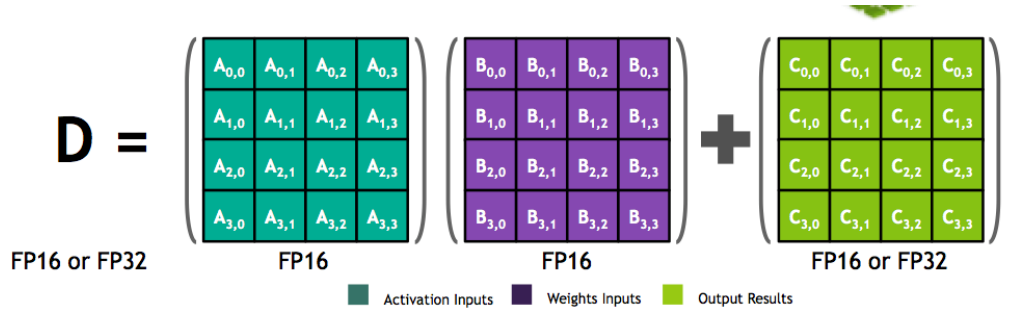
4x4 matrix processing array

D[FP32] = A[FP16] * B[FP16] + C[FP32]

Using Tensor cores via

- Volta optimized frameworks and libraries (cuDNN, CuBLAS, TensorRT, ..)
- CUDA C++ Warp Level Matrix Operations

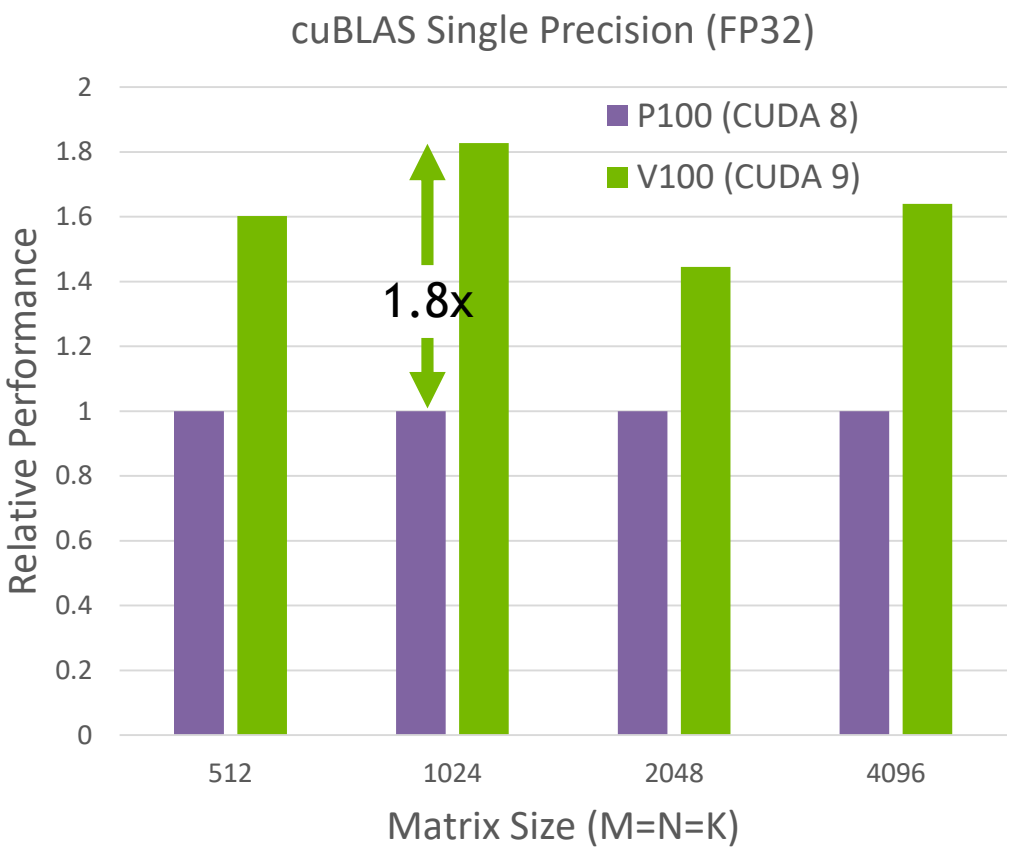


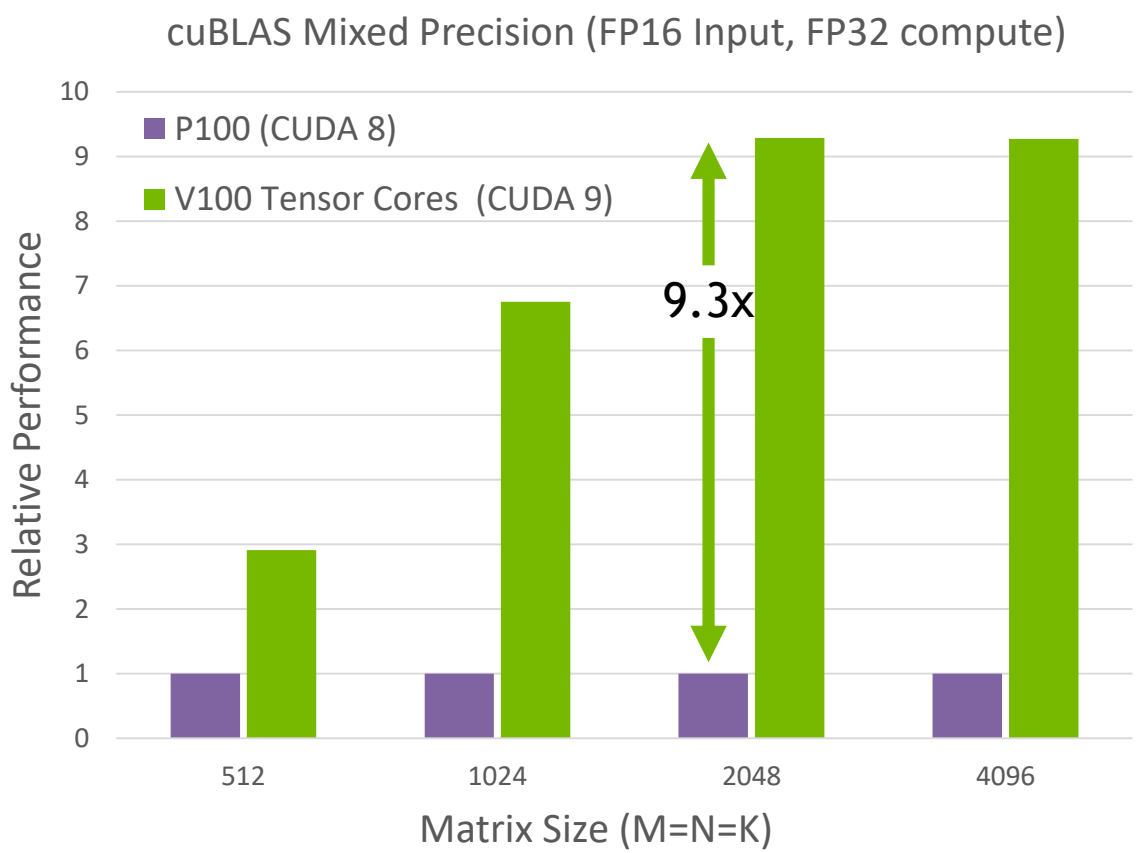


NVIDIA

cuBLAS GEMMS FOR DEEP LEARNING

V100 Tensor Cores + CUDA 9: over 9x Faster Matrix-Matrix Multiply

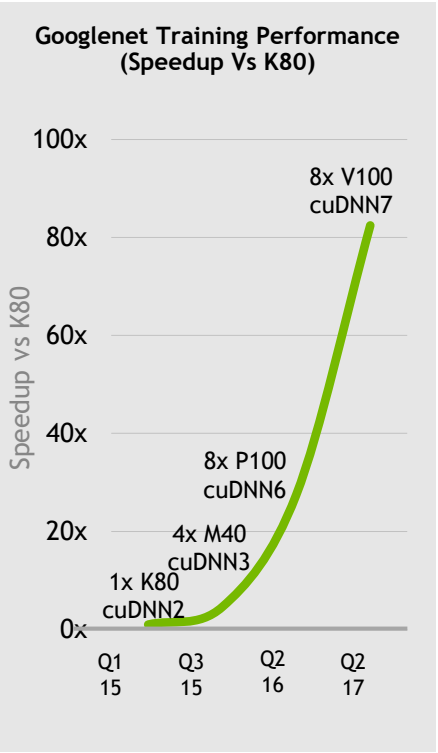




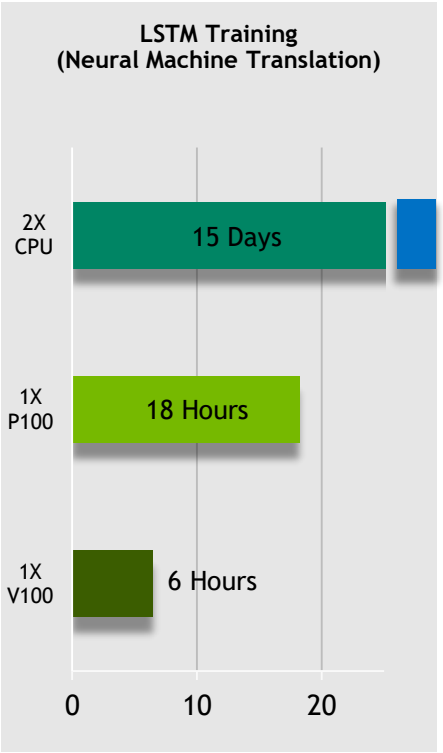
Note: pre-production Tesla V100 and pre-release CUDA 9. CUDA 8 GA release.

AI PERFORMANCE ON VOLTA

3X Faster DL Training Performance

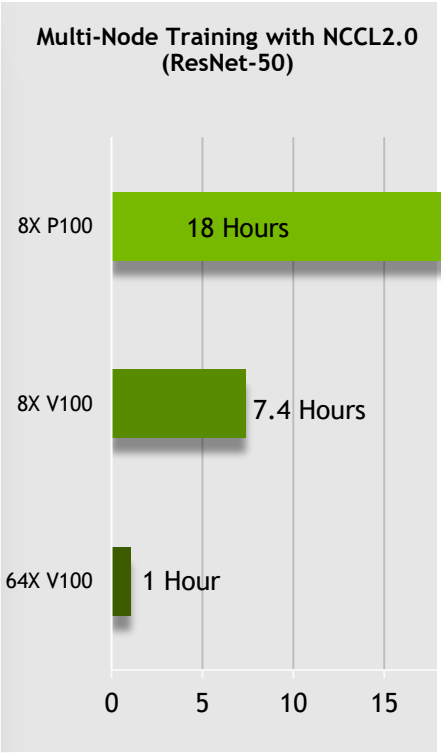


Over 80x DL Training Performance in 3 Years



3X Reduction in Time to Train Over P100

Neural Machine Translation Training for 13 Epochs |German ->English, WMT15 subset | CPU = 2x Xeon E5 2699 V4 | V100 performance measured on pre-production hardware.



85% Scale-Out Efficiency Scales to 64 GPUs with Microsoft Cognitive Toolkit

ResNet50 Training for 90 Epochs with 1.28M images dataset | Cognitive Toolkit with NCCL 2.0 | V100 performance measured on pre-production hardware.

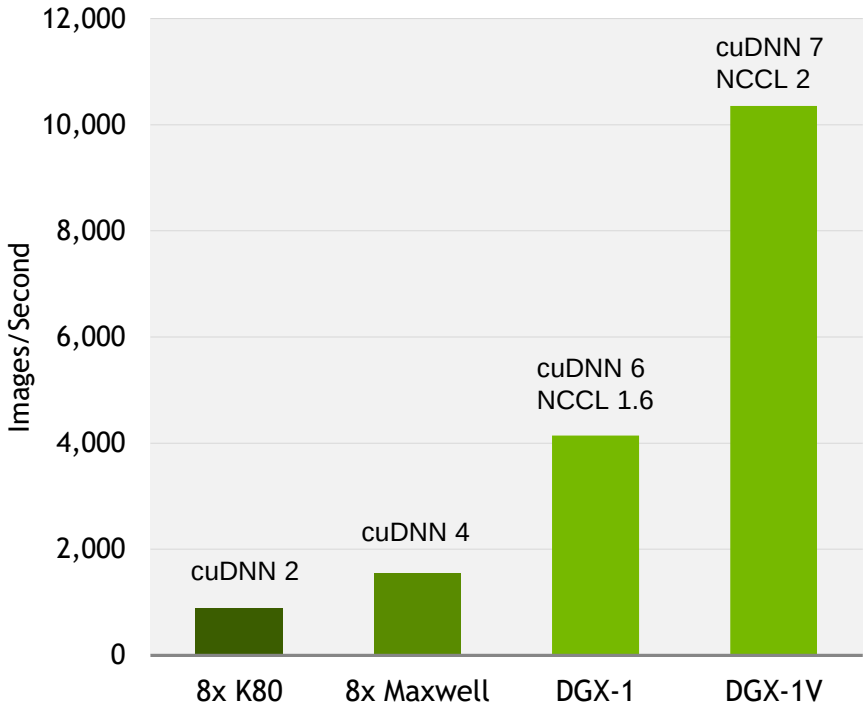
Deep Learning Primitives

NVIDIA cuDNN 7

High performance building blocks for deep learning frameworks

- Drop-in acceleration for widely used deep learning frameworks such as Caffe2, Microsoft Cognitive Toolkit, PyTorch, Tensorflow, Theano and others
- Accelerates industry vetted deep learning algorithms, such as convolutions, LSTM RNNs, fully connected, and pooling layers
- Fast deep learning training performance tuned for NVIDIA GPUs

Deep Learning Training Performance

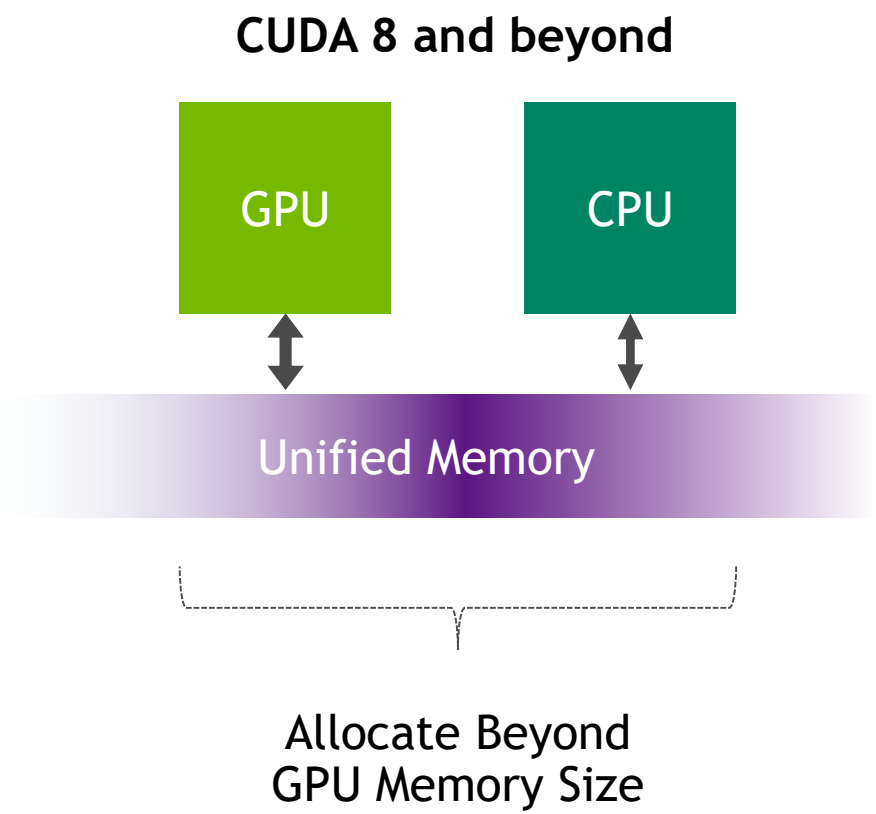


"NVIDIA has improved the speed of cuDNN with each release while extending the interface to more operations and devices at the same time."

— Evan Shelhamer, Lead Caffe Developer, UC Berkeley

UNIFIED MEMORY

Large datasets, simple programming, High Performance

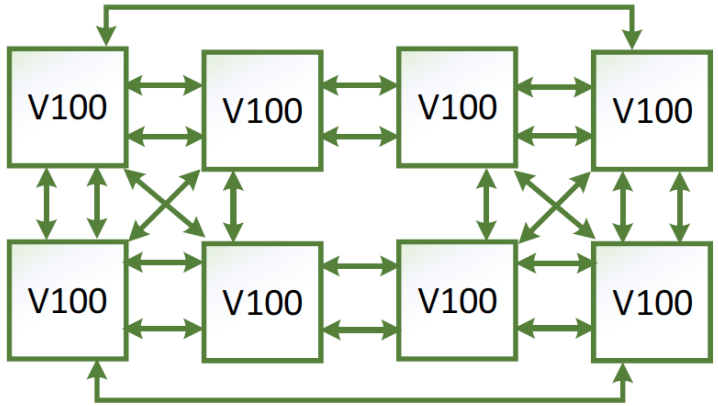


Enable Large Data Models	Oversubscribe GPU memory Allocate up to system memory size
Tune Unified Memory Performance	Usage hints via cudaMemAdvise API Explicit prefetching API
Simpler Data Access	CPU/GPU Data coherence Unified memory atomic operations

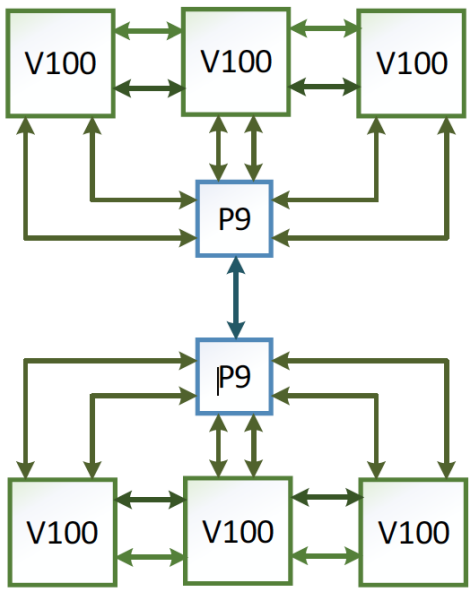
NVIDIA

VOLTA NVLINK

- 6 NVLINKS @ 50 GB/s bidirectional
- Reduce number of lanes for lightly loaded link (Power savings)
- Coherence features for NVLINK enabled CPUs



Hybrid cube mesh
(eg. DGX1V)



POWER9 based node

NVIDIA.

NVIDIA DGX-1

AI supercomputer-appliance-in-a-box

8x Tesla V100 connected via NVLINK

(120 TFLOPS FP32, 960 Tensor TFLOPS)

Dual Xeon CPU, 512 GB Memory

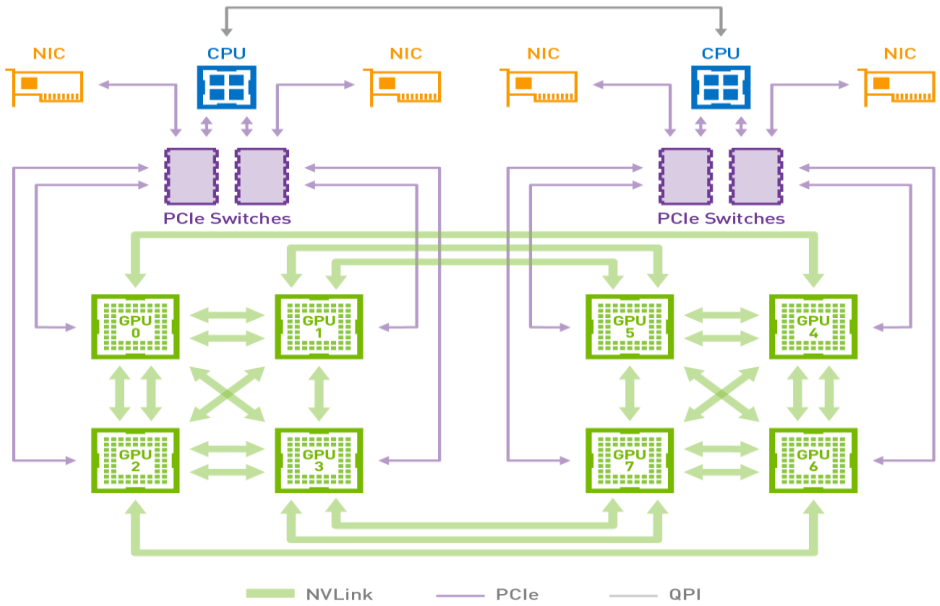
7 TB SSD Deep Learning Cache

Dual 10GbE, Quad IB 100Gb

3RU – 3200W

Optimized Deep Learning Software across the entire stack

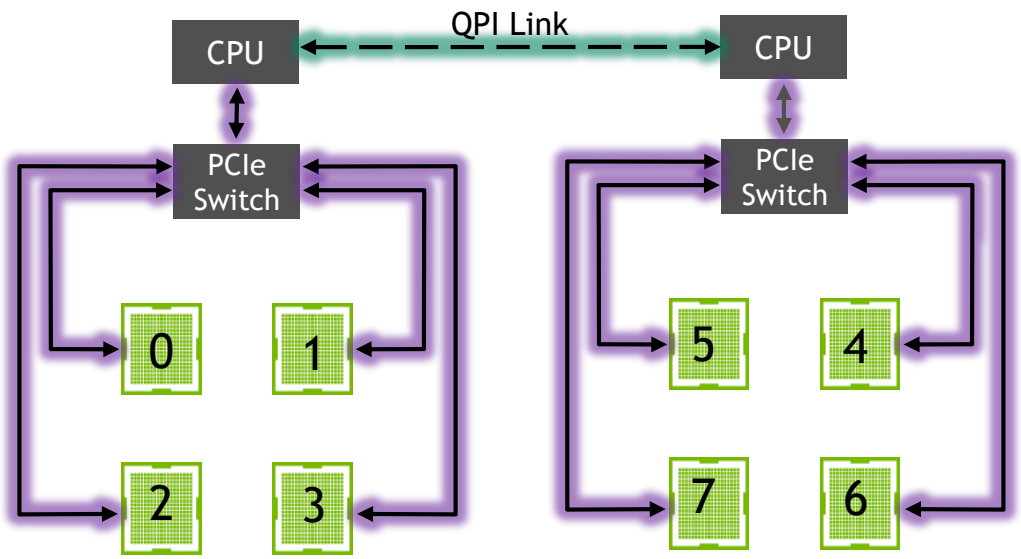
- Containerized frameworks
- Always up-to-date via the cloud



NVLINK AND MULTI-GPU SCALING

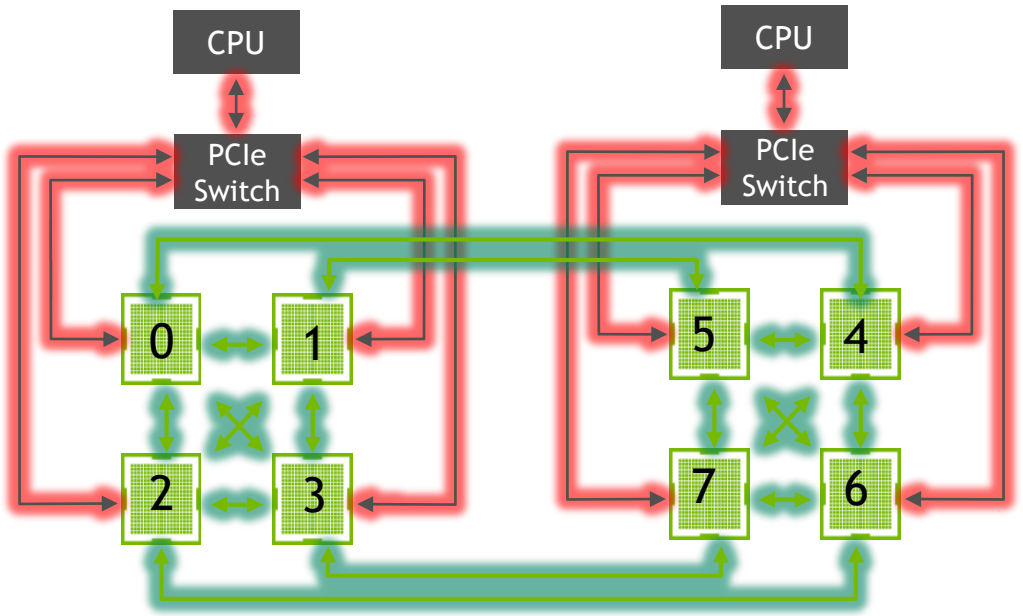
For Data Parallel Training

PCIe based system



- Data loading over PCIe
- Gradient averaging over PCIe and QPI
- Data loading and gradient averaging share communication resources: Congestion

NVLINK based system



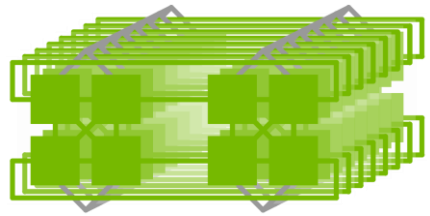
- Data loading over PCIe (red)
- Gradient averaging over NVLink (blue)
- No sharing of communication resources: No congestion

NVIDIA.

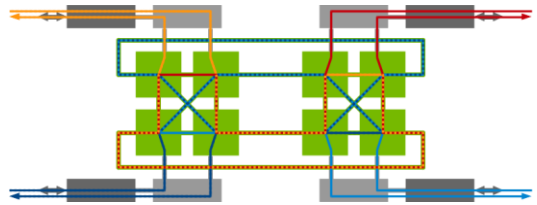
NVIDIA Collective Communications Library (NCCL) 2

Multi-GPU and multi-node collective communication primitives

Multi-GPU:
NVLink
PCIe



Multi-Node:
InfiniBand verbs
IP Sockets



Automatic
Topology
Detection

High-performance multi-GPU and multi-node collective communication primitives optimized for NVIDIA GPUs

- Fast routines for multi-GPU multi-node acceleration that maximizes inter-GPU bandwidth utilization
- Easy to integrate and MPI compatible. Uses automatic topology detection to scale HPC and deep learning applications over PCIe and NVink
- Accelerates leading deep learning frameworks such as Caffe2, Microsoft

developer.nvidia.com/nccl

WHAT'S NEW IN NCCL 2

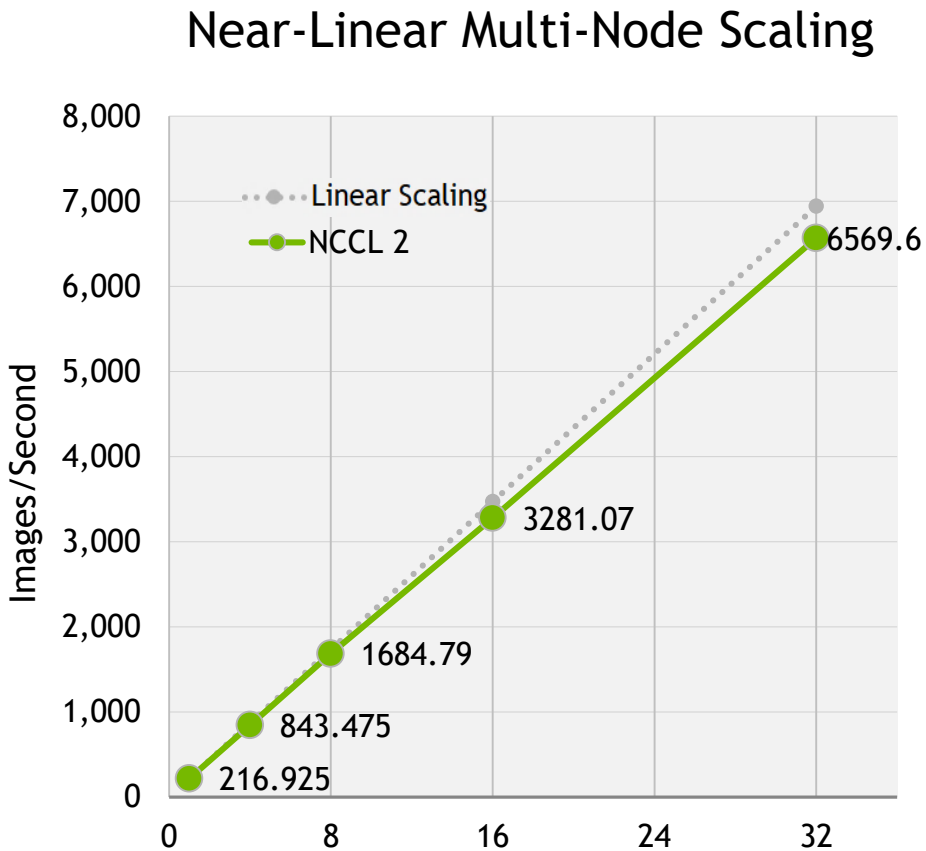
Performance

- Delivers over 90% multi-node scaling efficiency using up to eight GPU-accelerated servers

New Features

- Multi-node, multi-GPU communication collectives
- Automatic topology detection to determine optimal communication path
- Optimized to achieve high bandwidth over PCIe and NVink high-speed interconnect

Available now as a free download to members of NVIDIA Developer Program



Microsoft Cognitive Toolkit multi-node scaling performance (images/sec), NVIDIA DGX-1 + cuDNN 6 (FP32), ResNet50, Batch size: 64


NVIDIA