NVIDIA ACCELERATEDANALYTICS

May 2017: PVAMU MCBDA Conference
Bob Crovella

NVIDIA — THE AI COMPUTING COMPANY

Gaming

VR / AR / MR

Data Center

Self-Driving Cars

Visual & AI Computing

GPU

AGENDA

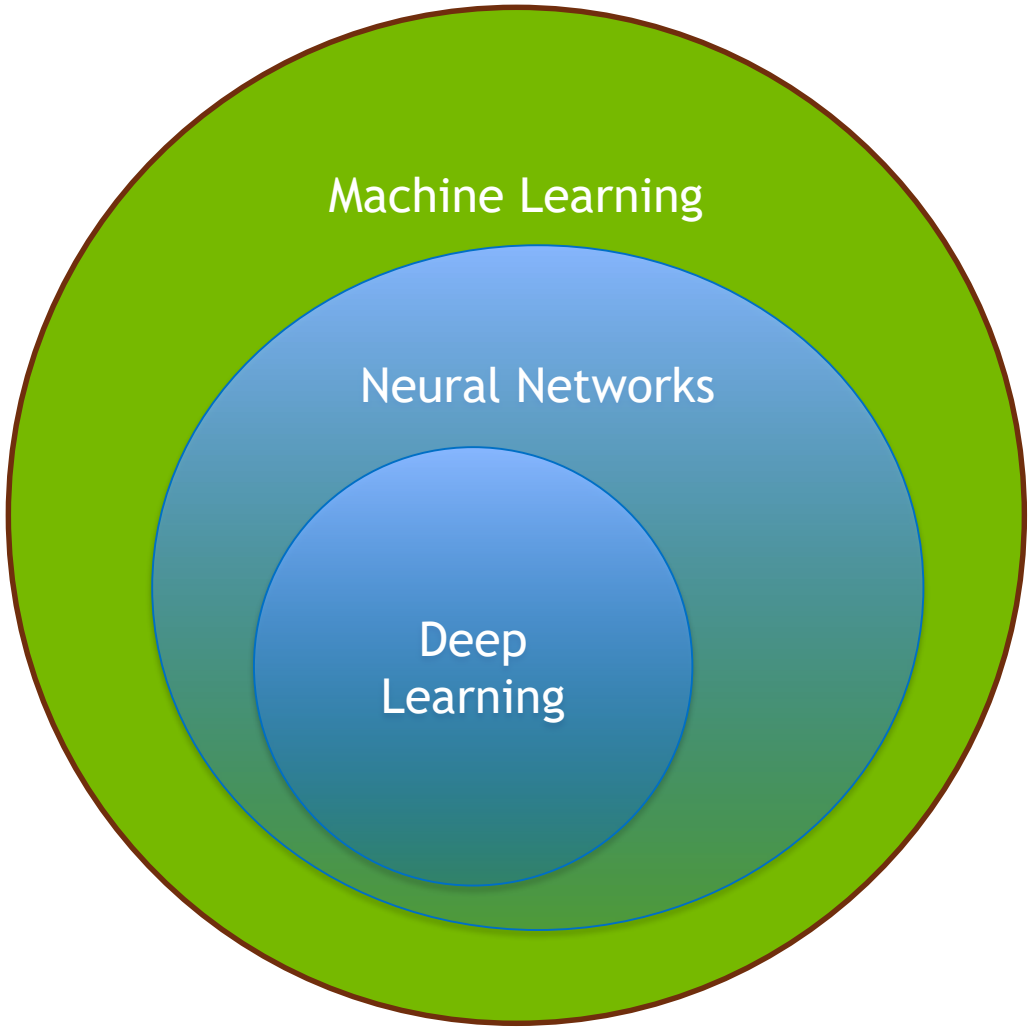
What is Deep Learning?

Example Use Cases (Healthcare emphasis)

GPU Accelerated Analytics

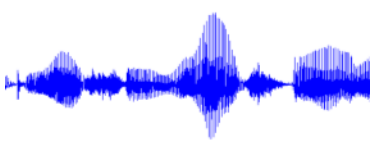
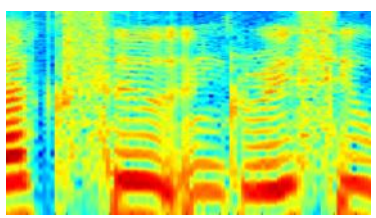
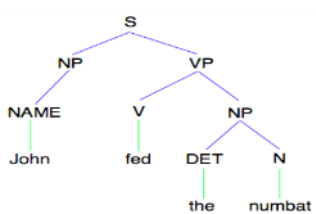
- DL Analytics
- Graph Analytics

What is Deep Learning?


Machine Learning
Neural Networks
Deep
Learning

Traditional machine perception

Hand crafted feature extractors

Raw data	Feature extraction	Classifier/ detector	Result
		SVM, shallow neural net, …	
		HMM, shallow neural net, …	Speaker ID, speech transcription, …
		Clustering, HMM, LDA, LSA …	Topic classification, machine translation, sentiment analysis…

Machine learning approach

Train:

Errors

MODEL

Dog
Cat
Raccoon

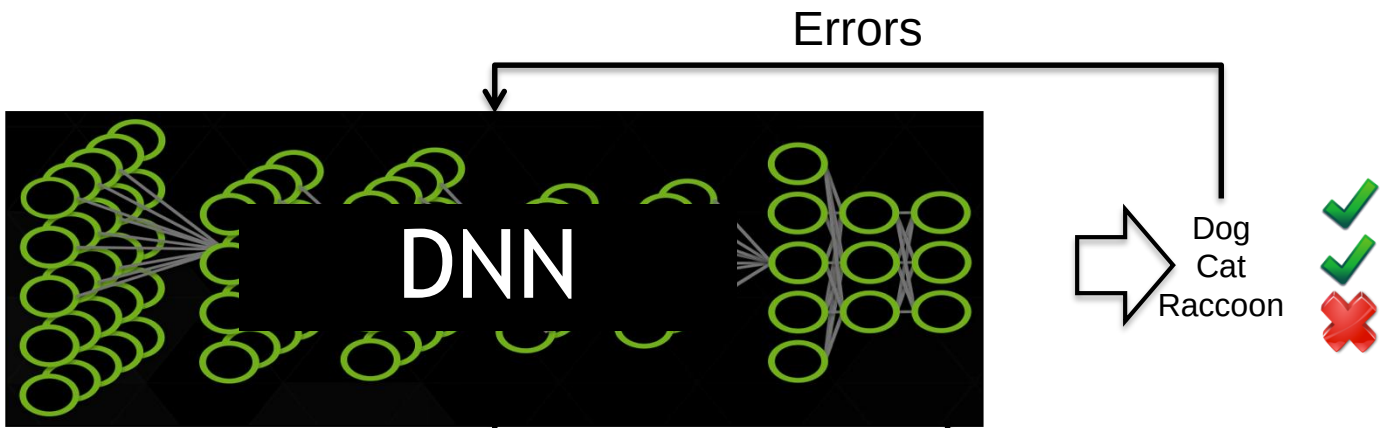
Deploy:

MODEL

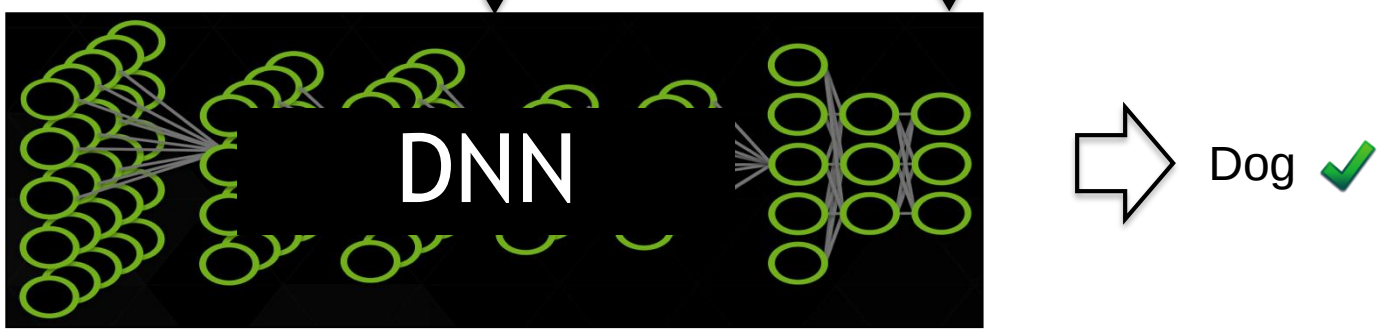
Dog

Deep learning approach

Train:



Deploy:

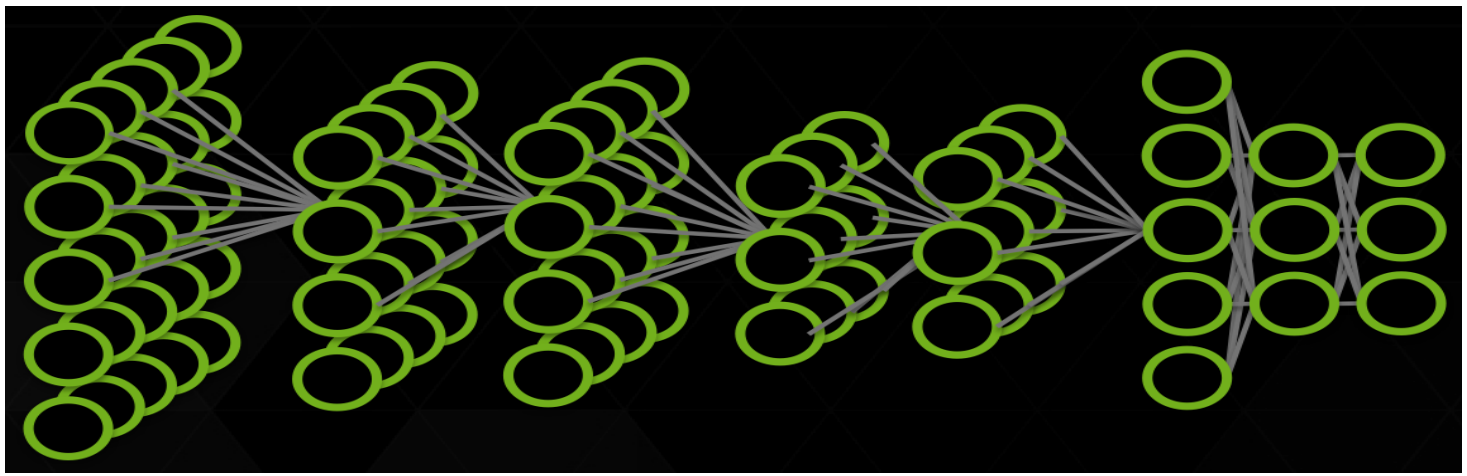


Artificial neural network

A collection of simple, trainable mathematical units that collectively learn complex functions

Hidden layers

Input layer

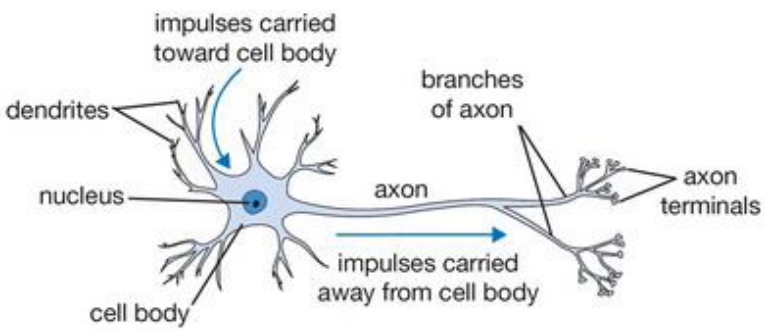


Output layer

Given sufficient training data an artificial neural network can approximate very complex functions mapping raw data to output decisions

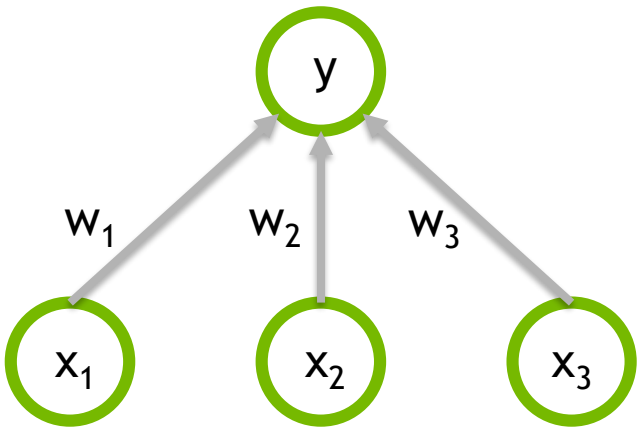
Artificial neurons

Biological neuron



From Stanford cs231n lecture notes

Artificial neuron

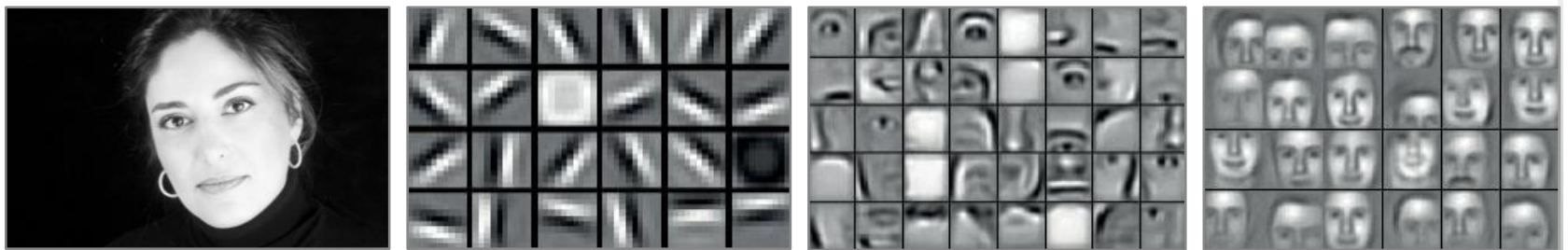


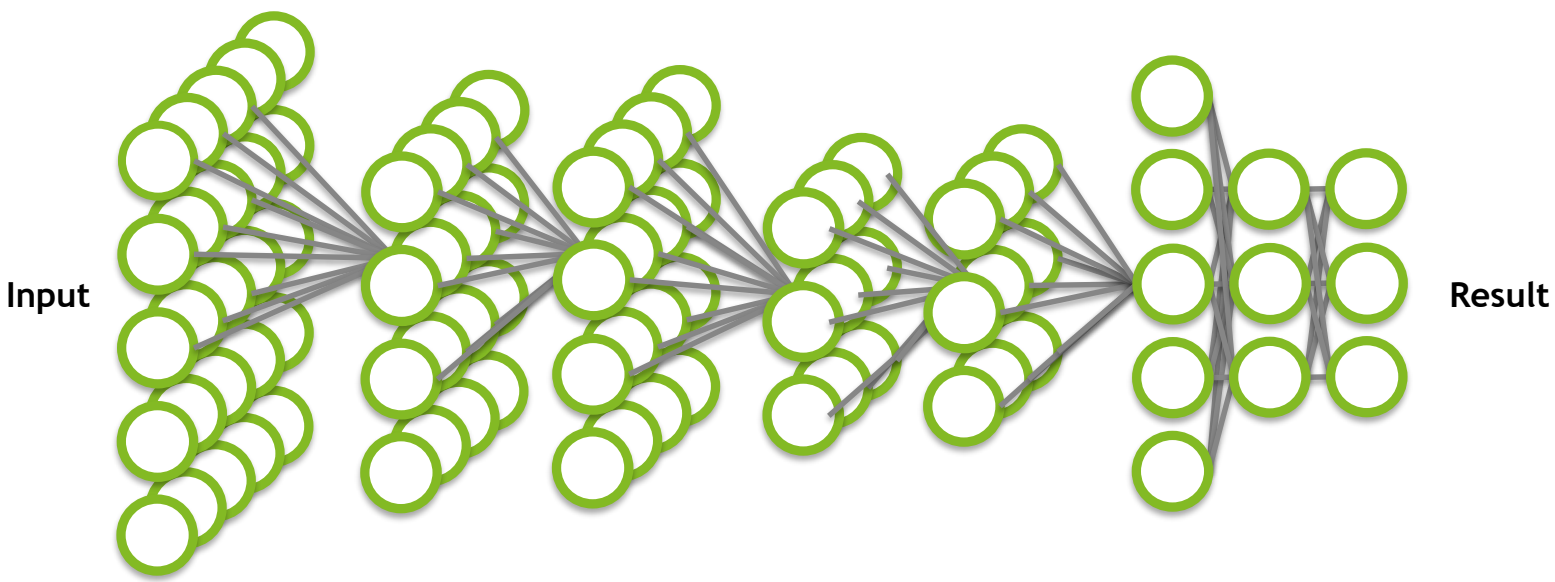
$$y=F(w_1x_1+w_2x_2+w_3x_3)$$

$$F(x)=\max(0,x)$$

Deep neural network (DNN)

Raw data | Low-level features | Mid-level features | High-level features





Application components:

Task objective
- e.g. Identify face

Training data
- 10-100M images

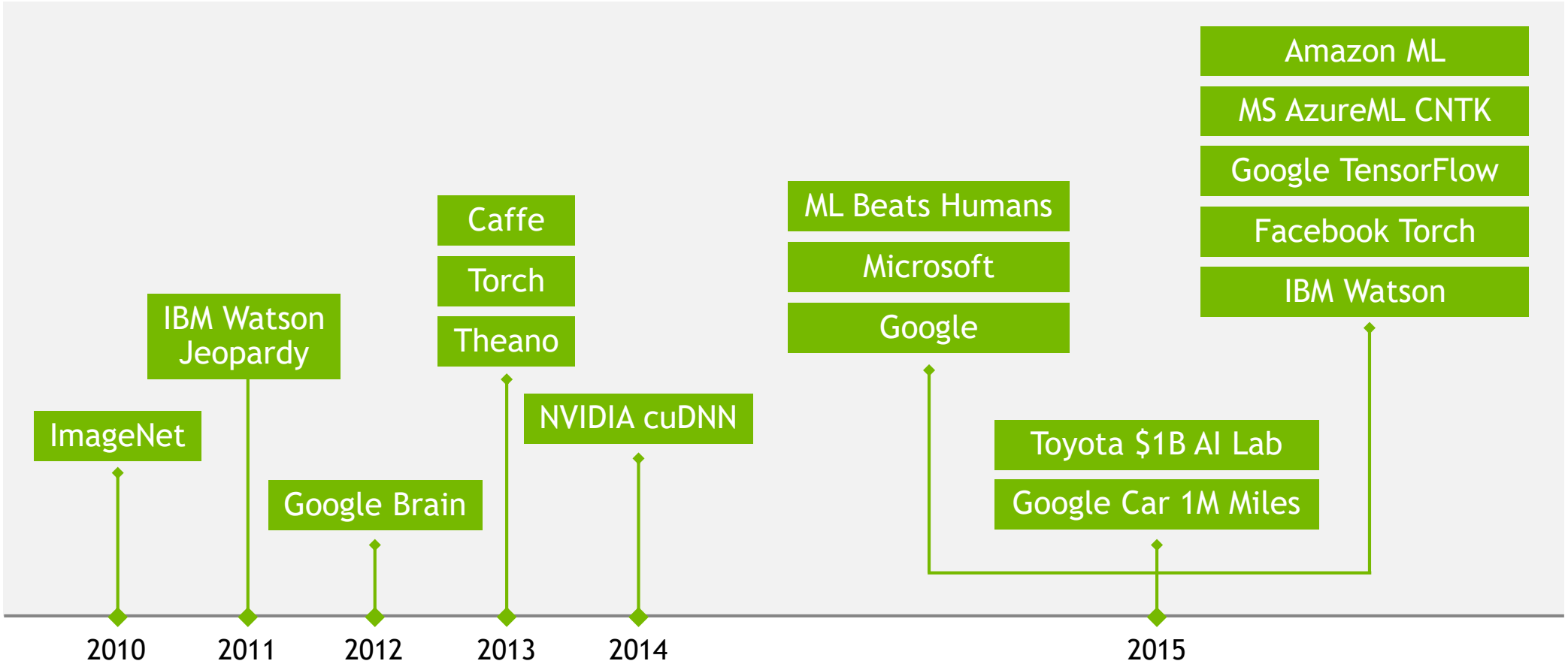
Network architecture
- ~10s-100s of layers
- 1B parameters

Learning algorithm
- ~30 Exaflops
- 1-30 GPU days

Deep learning benefits

- Robust
  - No need to design the features ahead of time – features are automatically learned to be optimal for the task at hand
  - Robustness to natural variations in the data is automatically learned
- Generalizable
  - The same neural net approach can be used for many different applications and data types
- Scalable
  - Performance improves with more data, method is massively parallelizable

The AI race is on



AlphaGo

First Computer Program to Beat a Human Go Professional

Training DNNs: 3 weeks, 340 million training steps on 50 GPUs

Play: Asynchronous multi-threaded search

- Simulations on CPUs, policy and value DNNs in parallel on GPUs
- Single machine: 40 search threads, 48 CPUs, and 8 GPUs
- Distributed version: 40 search threads, 1202 CPUs and 176 GPUs

Outcome: Beat both European and World Go champions in best of 5 matches

http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html
http://deepmind.com/alpha-go.html

Baidu Deep Speech 2

End-to-end Deep Learning for English and Mandarin Speech Recognition

English and Mandarin speech recognition

Transition from English to Mandarin made simpler by end-to-end DL

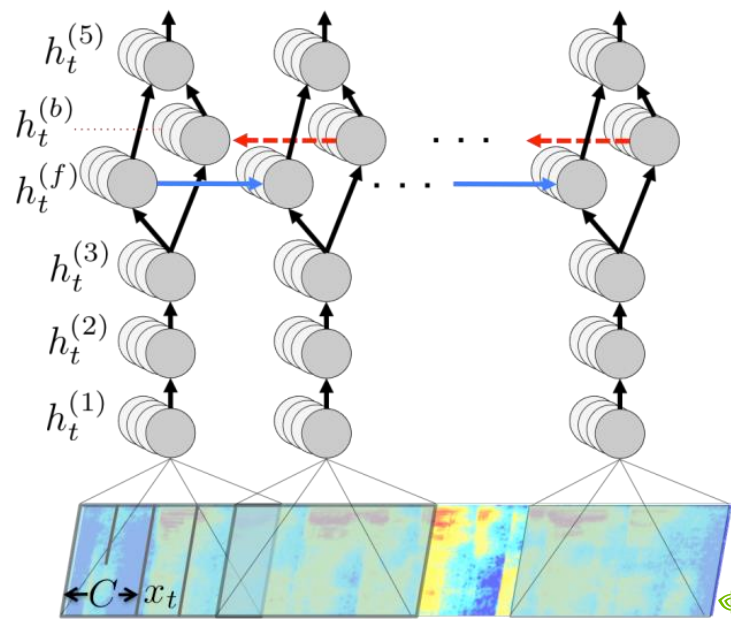
- No feature engineering or Mandarin-specifics required

More accurate than humans

- Error rate 3.7% vs. 4% for human tests

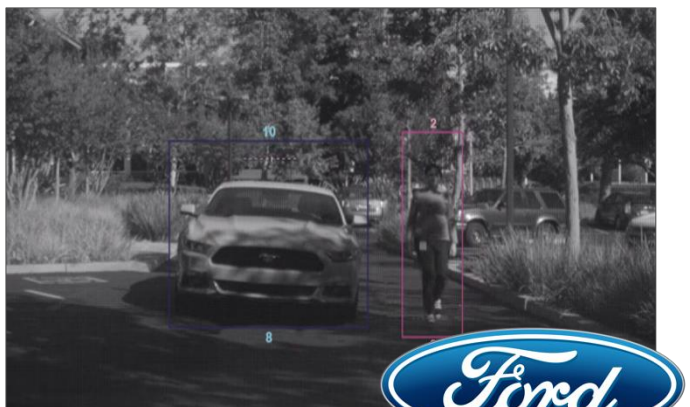
http://svail.github.io/mandarin/

http://arxiv.org/abs/1512.02595

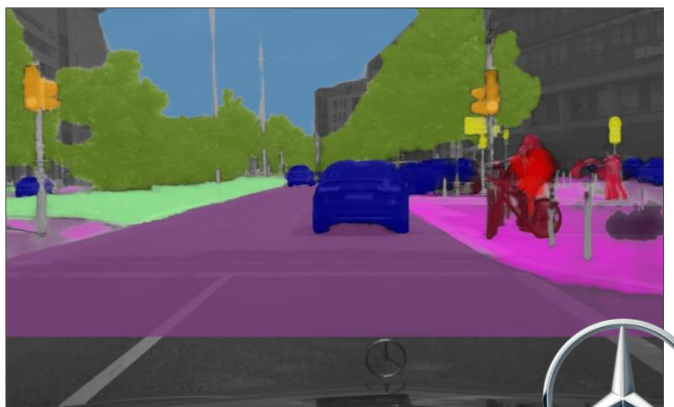


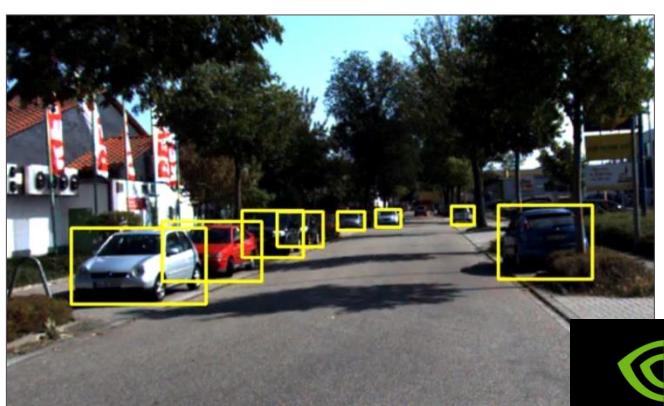
Deep Learning for Autonomous vehicles

Audi



Ford



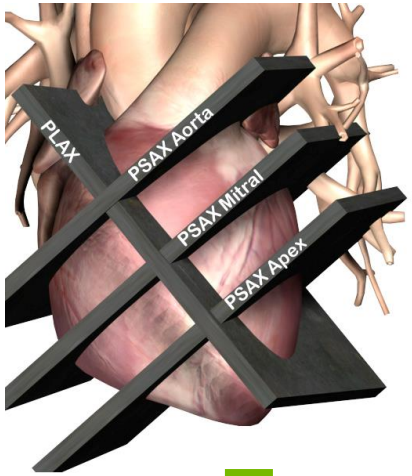


NVIDIA

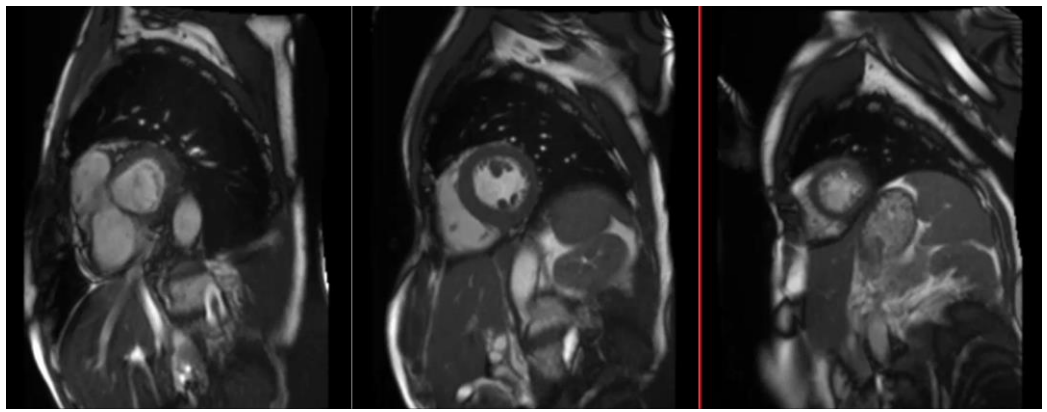
Audi

Automating Cardiac MRI analysis

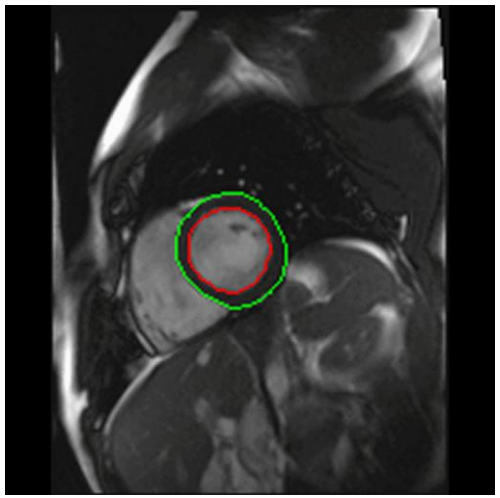
DL performance matches expert cardiologist at computing ejection fraction – a key indicator of heart disease



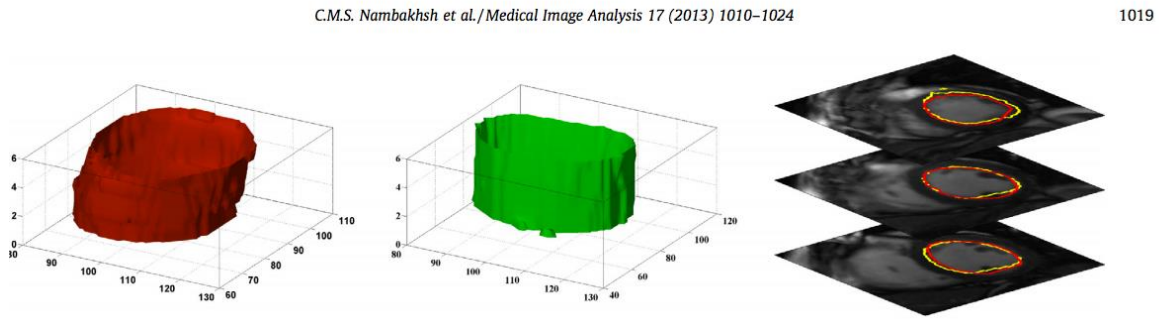
MRI imaging



Manual annotation



Software volume estimate



Icahn School of Medicine at Mount Sinai

Safeguarding patients' health through enhanced preventative medicine

- 'Deep Patient' analyzes electronic health records to predict 78 diseases, up to one year prior to onset
- Neural network trained on 100,000's records using NVIDIA® Tesla® K80 GPU and CUDA® programming model.

"For most diseases, prevention is easier than reversal. Deep Patient could have a huge impact on people's health."

-Joel T. Dudley, Assistant Professor of Genetics, Genomic Sciences Director of Biomedical Informatics

GPUs and Analytics

USE MORE PROCESSORS TO GO FASTER

DATA & ANALYTICS USE CASES

AUTOMOTIVE
Auto sensors reporting location, problems

COMMUNICATIONS
Location-based advertising

CONSUMER PACKAGED GOODS
Sentiment analysis of what's hot, problems

FINANCIAL SERVICES
Risk & portfolio analysis
New products

EDUCATION & RESEARCH
Experiment sensor analysis

HIGH TECHNOLOGY / INDUSTRIAL MFG.
Mfg. quality
Warranty analysis

LIFE SCIENCES
Clinical trials

MEDIA/ENTERTAINMENT
Viewers / advertising effectiveness

ON-LINE SERVICES / SOCIAL MEDIA
People & career matching

HEALTH CARE
Patient sensors, monitoring, EHRs

OIL & GAS
Drilling exploration sensor analysis

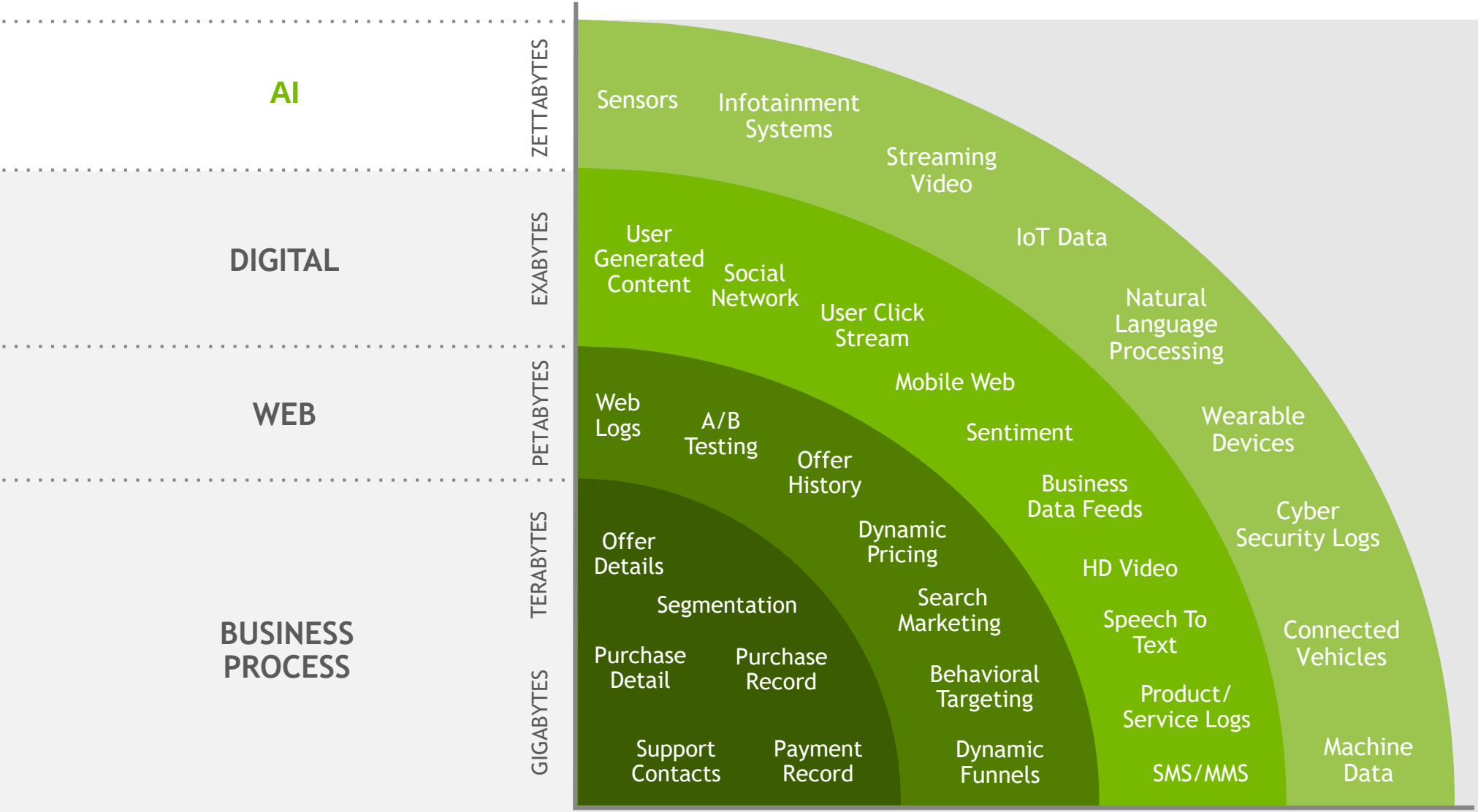
RETAIL
Consumer sentiment

TRAVEL & TRANSPORTATION
Sensor analysis for optimal traffic flows

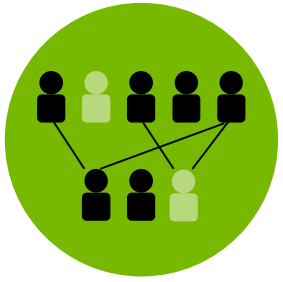
UTILITIES
Smart Meter analysis for network capacity,

LAW ENFORCEMENT & DEFENSE
Threat analysis - social media monitoring, photo analysis

DATA DELUGE TO DATA HUNGRY

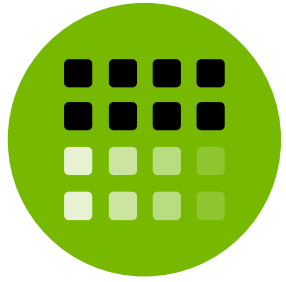


WORKAROUNDS ARE NOT THE ANSWERS



Sampling misses the whole picture

EXPLORE THE OUTLIERS AND LONG-TAIL EVENTS



Pre-aggregation struggles at scale

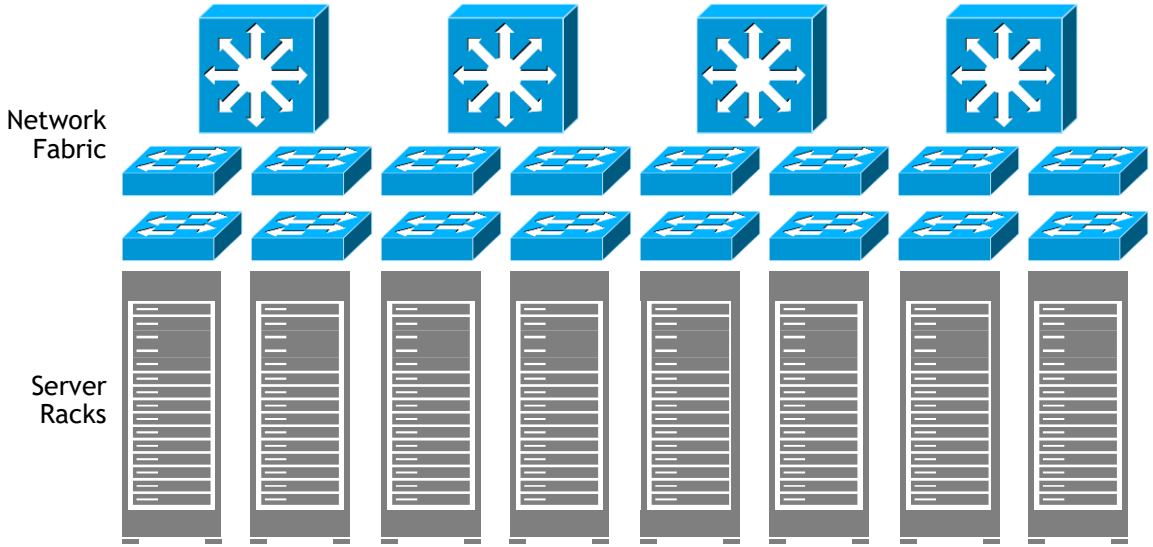
RELY ON ACCURATE DATA

Scale out on CPU infrastructure has tremendous hidden costs

SCALE WITH A ROI

SCALE OUT

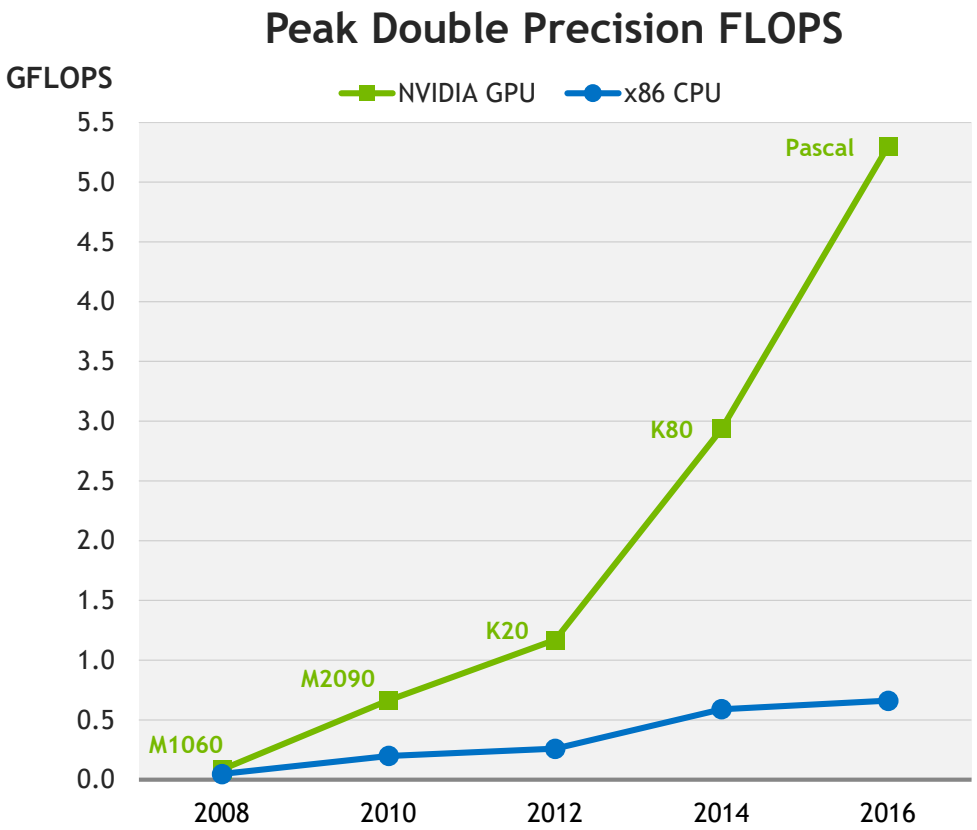
Lots of nodes Interconnected with vast network overhead

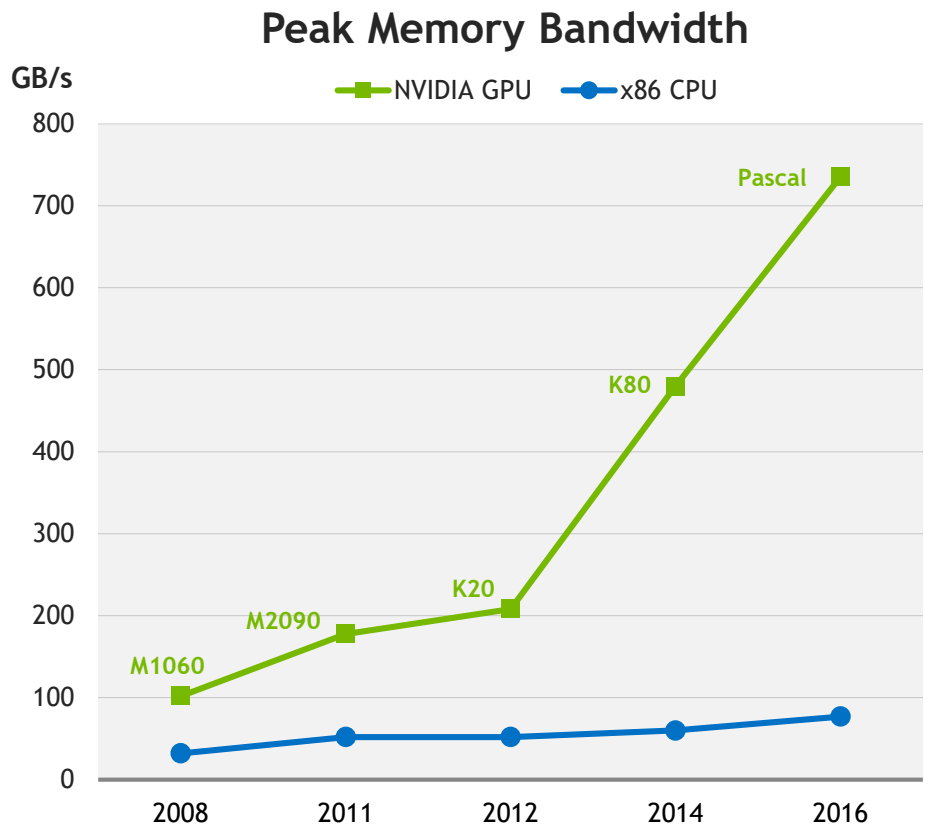


STRONG SCALE

Few lightning-fast nodes with performance of hundreds of weak nodes

THE ADVANTAGES OF GPU-ACCELERATED DATA CENTER





TOWARD REAL TIME BIG DATA ANALYTICS

GPUs enable the next generation of in-memory processing

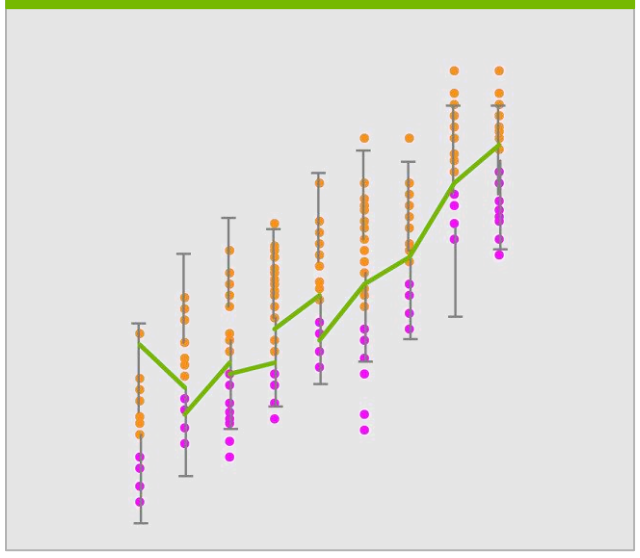
	DUAL BROADWELL SERVER	NVIDIA DGX-1 SERVER	GPU PERFORMANCE INCREASE
Aggregate Memory Bandwidth	150 GB/s	5760 GB/s	38 X
Aggregate SP FLOPS	4 TF	85 TF	21 X

Single DGX-1 server provides the compute capability of dozens of dual-cpu servers

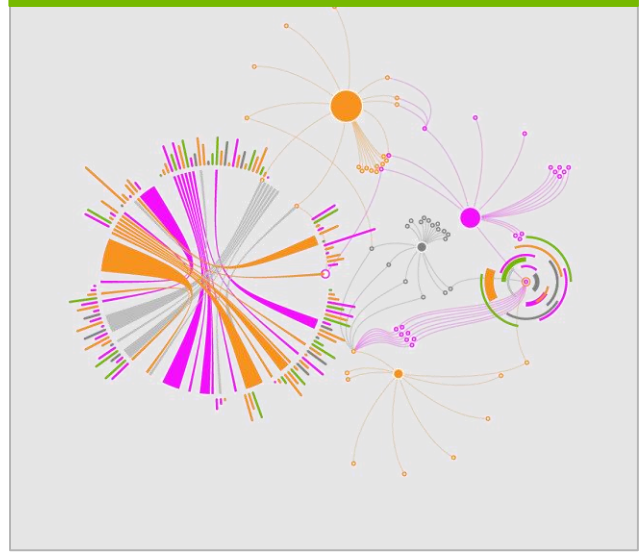
NVIDIA ACCELERATED ANALYTICS

GPUs in the Data Center

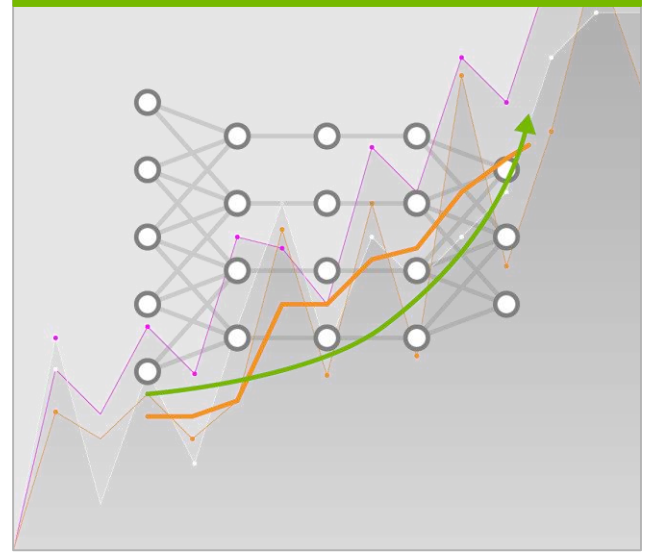
ANALYZE



VISUALIZE



AI-ACCELERATE



ANALYTICS ECOSYSTEM

VISUALIZATION	tableau, kibana
DATABASES	cassandra, SAP HANA, ORACLE, hadoop
CORE TECHNOLOGIES	Spark, hadoop HDFS
TRADITIONAL DATA CENTER	hp, IBM, Quanta Computer, DELL, Lenovo, CRAY, CISCO

DEEP LEARNING	Caffe, TensorFlow, torch, theano, Microsoft CNTK, DEEPLEARNING4J
ACCELERATED VISUALIZATION	MAPD, graphistry, kinetica
ACCELERATED DATABASES	MAPD, SQREAM TECHNOLOGIES, blazegraph, kinetica, Blazing
CORE TECHNOLOGIES	BIDMACH Spark Scheduler, Spark Mesos
GPU-ACCELERATED DATA CENTER	NVIDIA Tesla GPUs, NVIDIA DGX Products, Cloud

GPU-ACCELERATION ENGINES

MapD

MapD is 55x to 1,000x faster than comparable CPU databases on billion+ row datasets

Kinetica

Hardware costs that are $^{1}/_{10}$ that of standard in-memory databases

BlazeGraph

200-300x speed-up

Graphistry

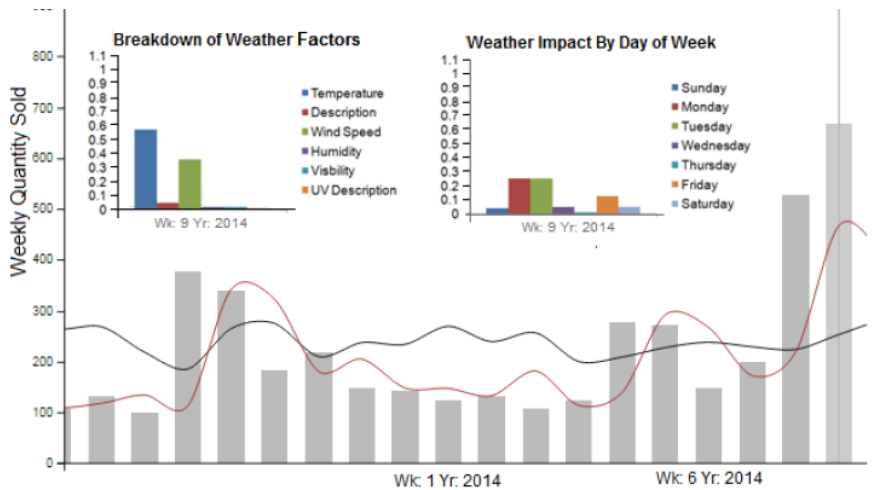
See 100x more data at millisecond speed

SQream

The supercomputing powers of the GPU combined with SQream's patented technology, results in up to 100 times faster analytics performance on terabyte-petabyte scale data sets

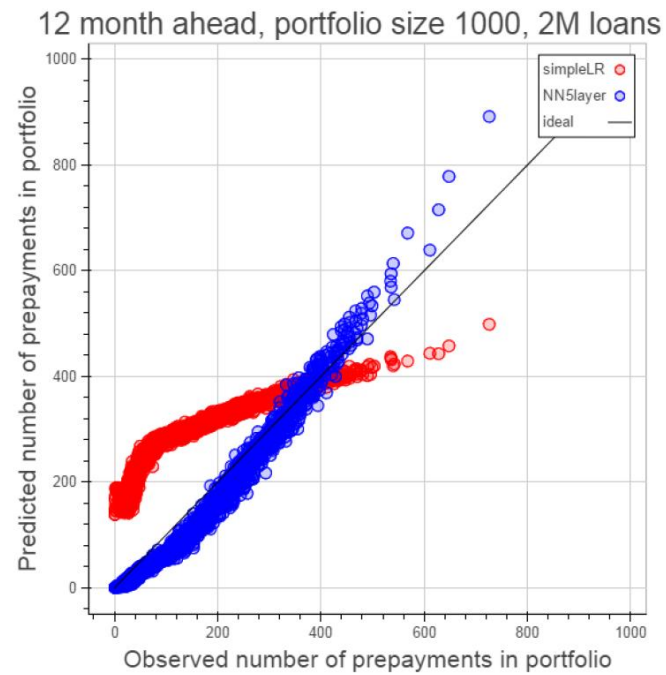
PREDICTIVE ANALYTICS AI

INVENTORY MANAGEMENT



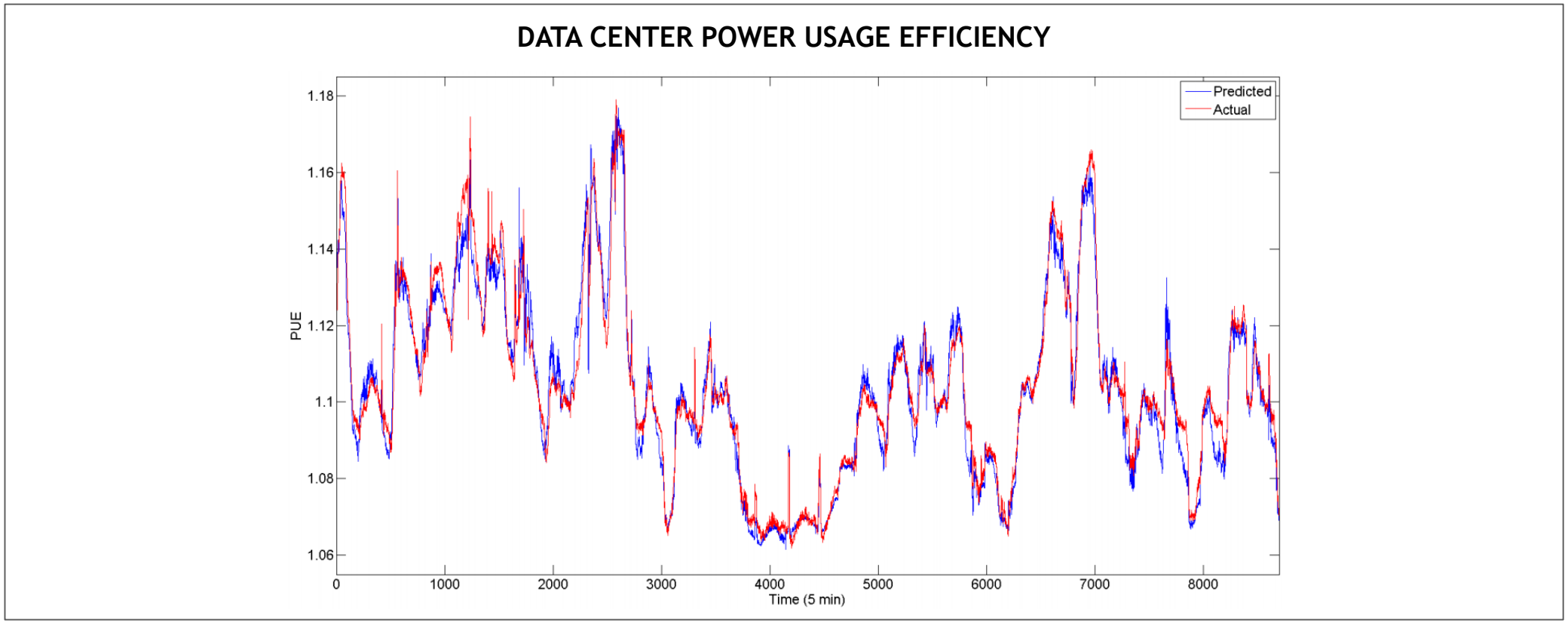
Source: IBM, 2016, Riemer et. al.

MORTGAGE RISK



Source: https://arxiv.org/abs/1607.02470

A GAME CHANGER FOR INDUSTRIAL IOT



Source: Jim Gao, Google

KEY DRIVERS

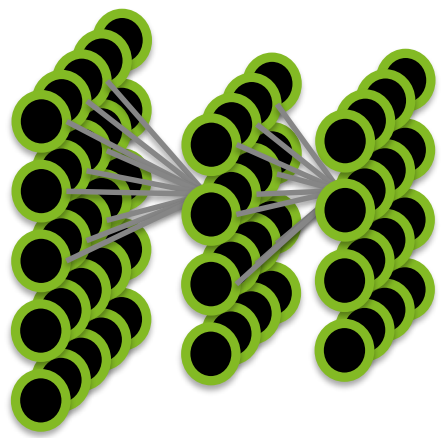
BIG DATA

facebook — 350 million images uploaded per day

Walmart — 2.5 Petabytes of customer data hourly

YouTube — 300 hours of video uploaded every minute

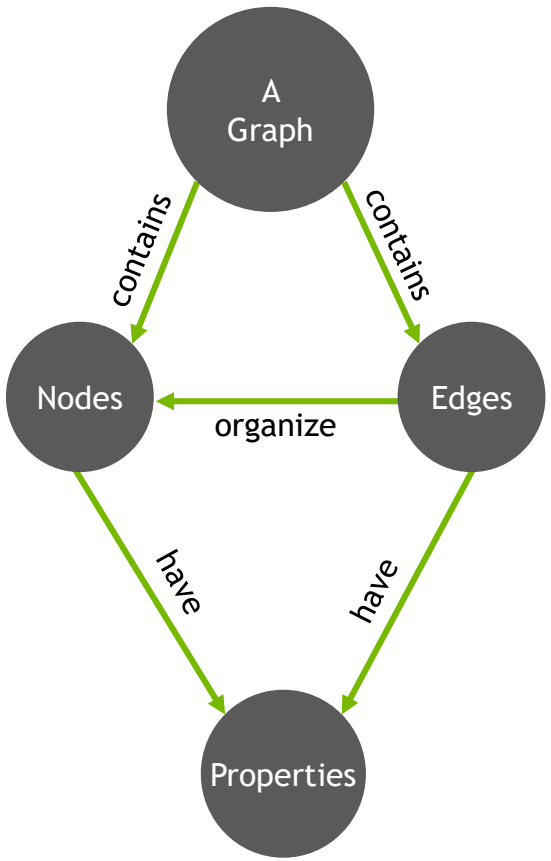
BETTER ALGORITHMS



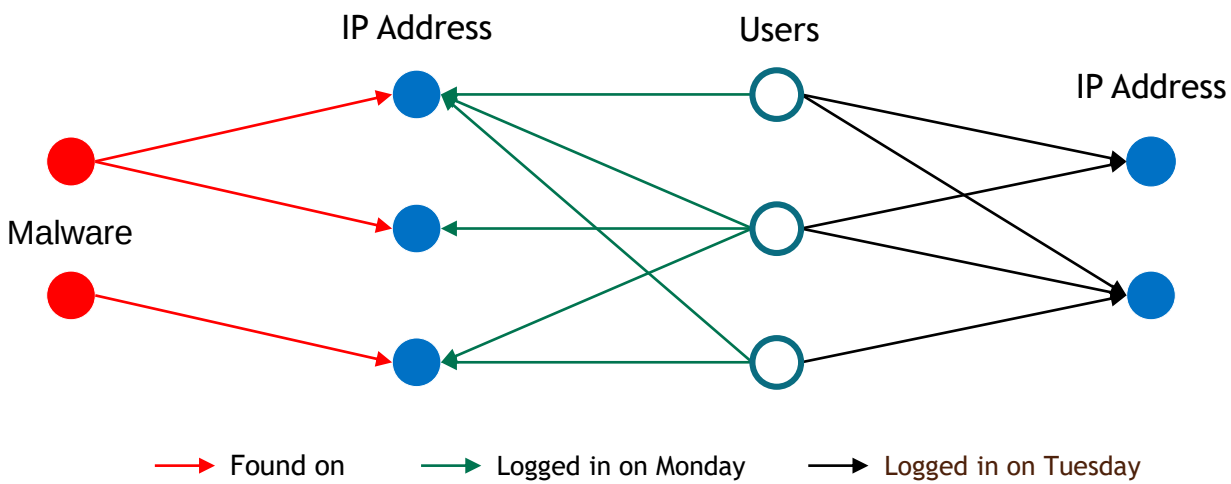
GPU ACCELERATION

“ The three breakthroughs that have finally unleashed A.I. on the world.” WIRED

WHAT IS A GRAPH?



HOW DOES MALWARE SPREAD?

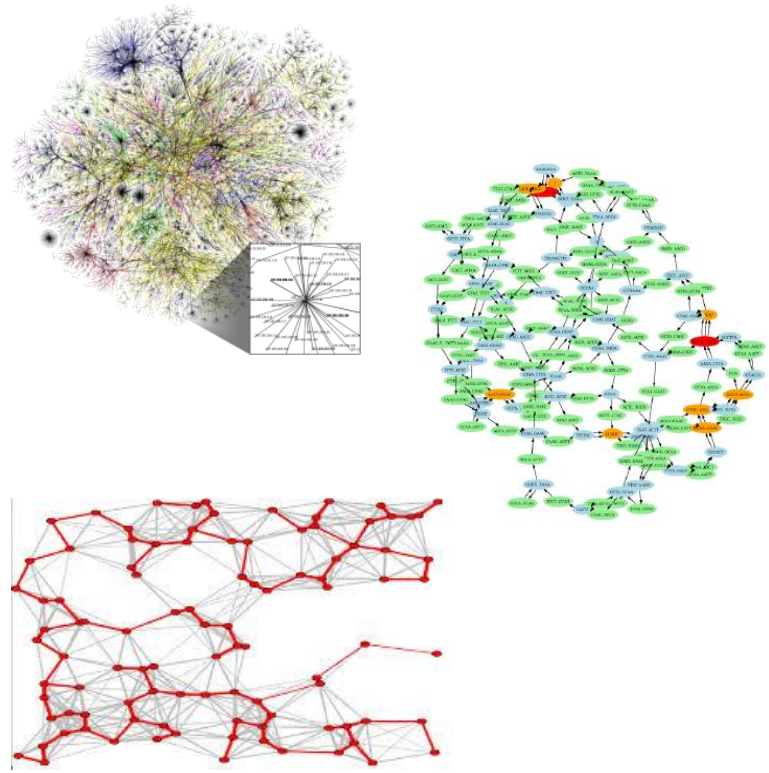


Nodes and edges can represent different things
Edges are to graphs as joins are to SQL

nvGRAPH

Easy onramp to GPU accelerated graph analytics

- GPU optimized algorithms
- Reduced cost and increased performance
- Standard formats and primitives
  Semi-rings, load-balancing
- Performance constantly improving



nvGRAPH

Accelerated graph analytics

nvGRAPH for high performance graph analytics

Deliver results up to 3x faster than CPU-only

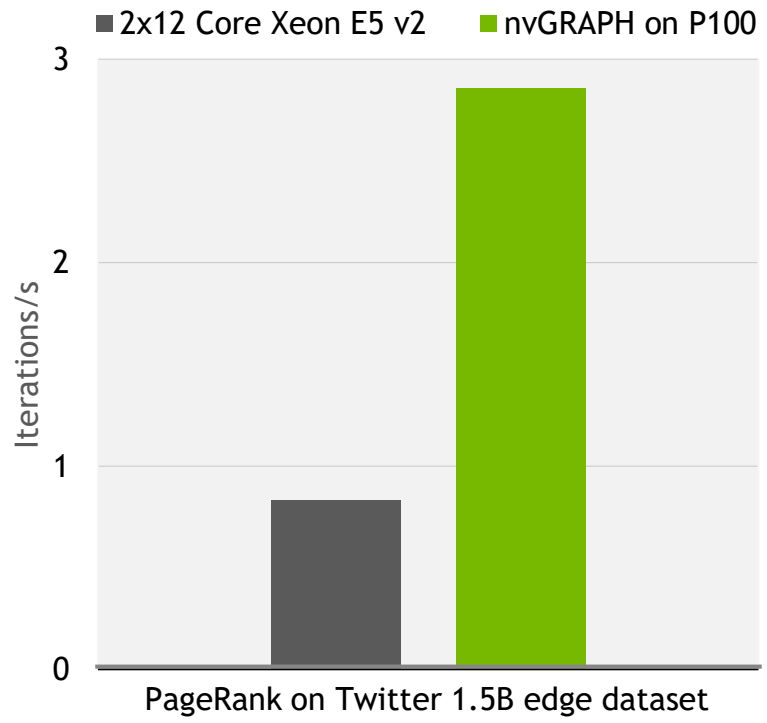
Solve graphs with up to 2 Billion edges on a single GPU

Accelerates a wide range of graph analytics applications:

PAGERANK	SINGLE SOURCE SHORTEST PATH	SINGLE SOURCE WIDEST PATH
Search	Robotic Path Planning	IP Routing
Recommendation Engines	Power Network Planning	Chip Design / EDA
Social Ad Placement	Logistics & Supply Chain Planning	Traffic sensitive routing

developer.nvidia.com/nvgraph

NVGRAPH: 3.4X SPEEDUP



nvGraph on P100

GraphMat on 2 socket 12-core Xeon E5-2697 v2 CPU,@ 2.70 GHz

COMING SOON

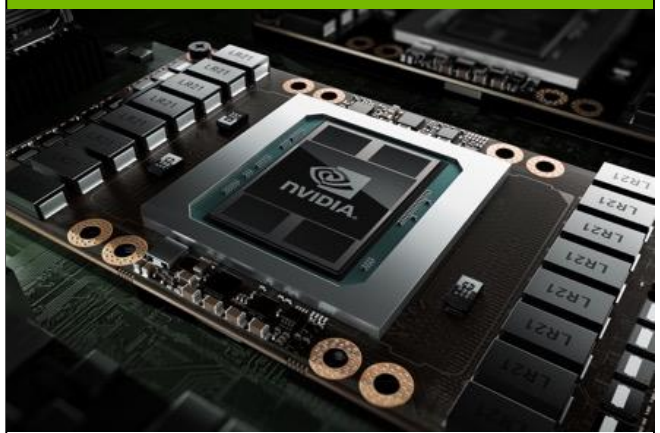
Features in next release

PARTITIONING	Min edge cut
CLUSTERING	Maximum modularity, Jaccard
BFS	Direction optimizing
GRAPH CONTRACTION	Visualization, hierarchical clustering

GPU ACCELERATED ANALYTICS

As data continues to grow at a continually accelerating pace, the limits of CPU based systems are being realized. GPUs are accelerating storage, processing, analytics, and visualization

TESLA
Servers in every shape and size



Hewlett Packard Enterprise · IBM · Quanta Computer · DELL · Lenovo · CRAY · cisco

DGX-1
The accelerated analytics supercomputer for instant productivity

NVIDIA

CLOUD
Everywhere

Alibaba.com · amazon · Baidu 百度 · Google · IBM · Microsoft

THANK YOU!