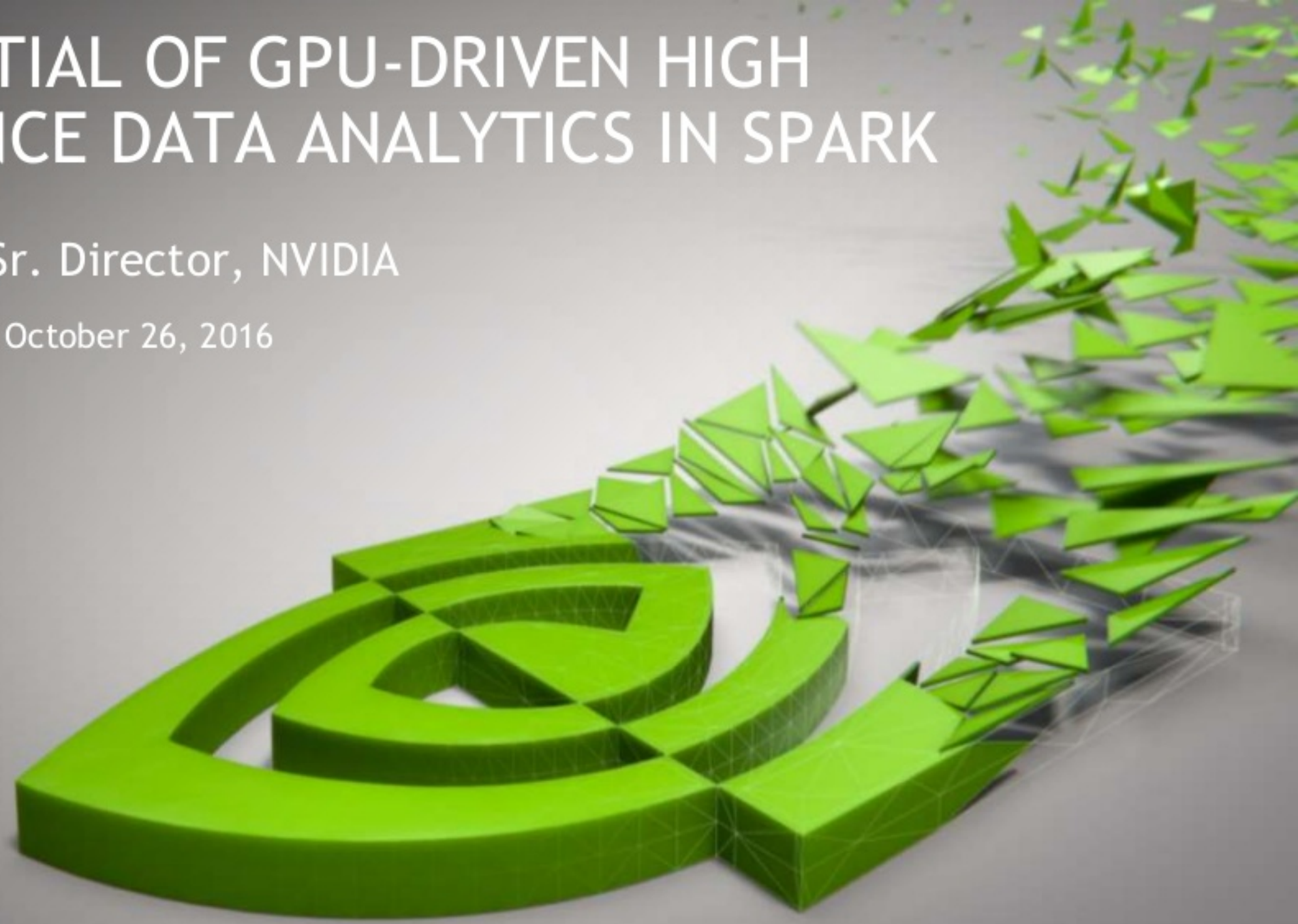
THE POTENTIAL OF GPU-DRIVEN HIGH PERFORMANCE DATA ANALYTICS IN SPARK

Andy Steinbach, Sr. Director, NVIDIA

Spark Summit Brussels, October 26, 2016

NVIDIA



HOW TO SCALE AI & DATA ANALYTICS?

Scale up

Compute intensive

We are headed here

Data intensive

Scale out

HIGH PERFORMANCE DATA ANALYTICS

Scale up

Caffe

Microsoft CNTK

theano

torch

TensorFlow

Deep Learning

Spark + TensorFlow + GPU
Spark + AI framework + GPU

Machine Learning
& DB Query

Spark™

Scale out

DEEP LEARNING - A NEW COMPUTING MODEL

"Training"

ImageNet

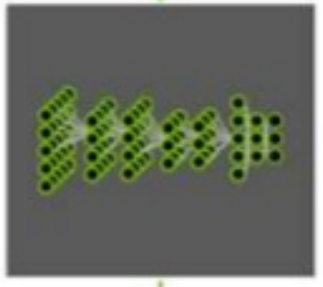
LEARNING ALGORITHM

"millions of trillions of FLOPS"

TB or PB of training data

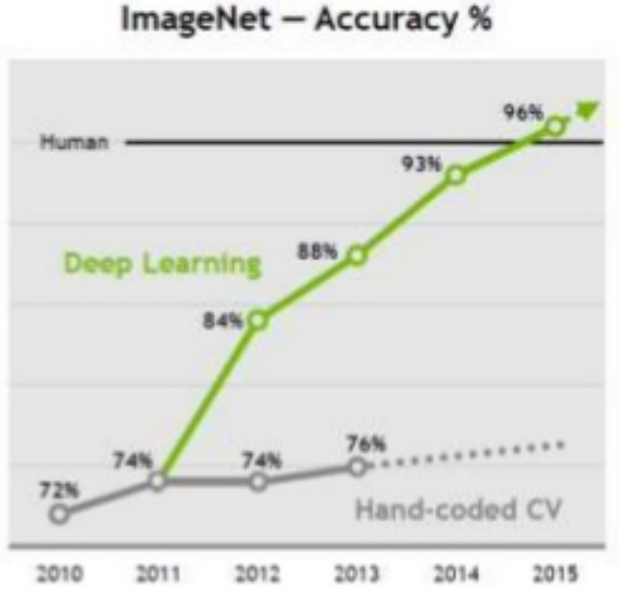
Known Ground Truth Labels

"Inference"



"cat"

"SUPERHUMAN" RESULTS SPARK HYPERSCALE ADOPTION



BEYOND JUST COMPUTER VISION

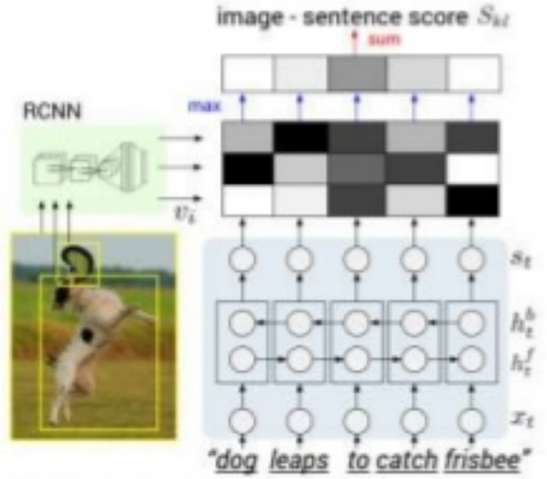


Figure 3. Diagram for evaluating the image-sentence score S_{kl}. Object regions are embedded with a CNN (left). Words (enriched by their context) are embedded in the same multimodal space with a BRNN (right). Pairwise similarities are computed with inner products (magnitudes shown in grayscale) and finally reduced to image-sentence score with Equation 8.

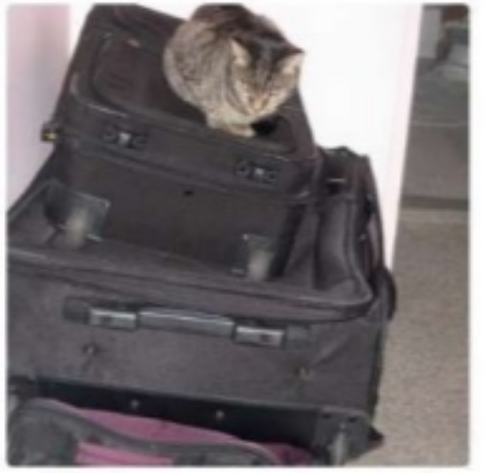
"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

"man in blue wetsuit is surfing on wave."

"baseball player is throwing ball in game."

"woman is holding bunch of bananas."



"black cat is sitting on top of suitcase."

A REVOLUTION IN MEDECINE

Example Data:

CHEST 2V FRONTAL/LATERAL XXXX, XXXX XXXX PM

Comparison: None

Indication: Burmese male has complete TB treatment

Findings: Both lungs are clear and expanded with no infiltrates. Basilar focal atelectasis is present in the lingula. Heart size normal. Calcified right hilar XXXX are present

Impression: No active disease.

MeSH

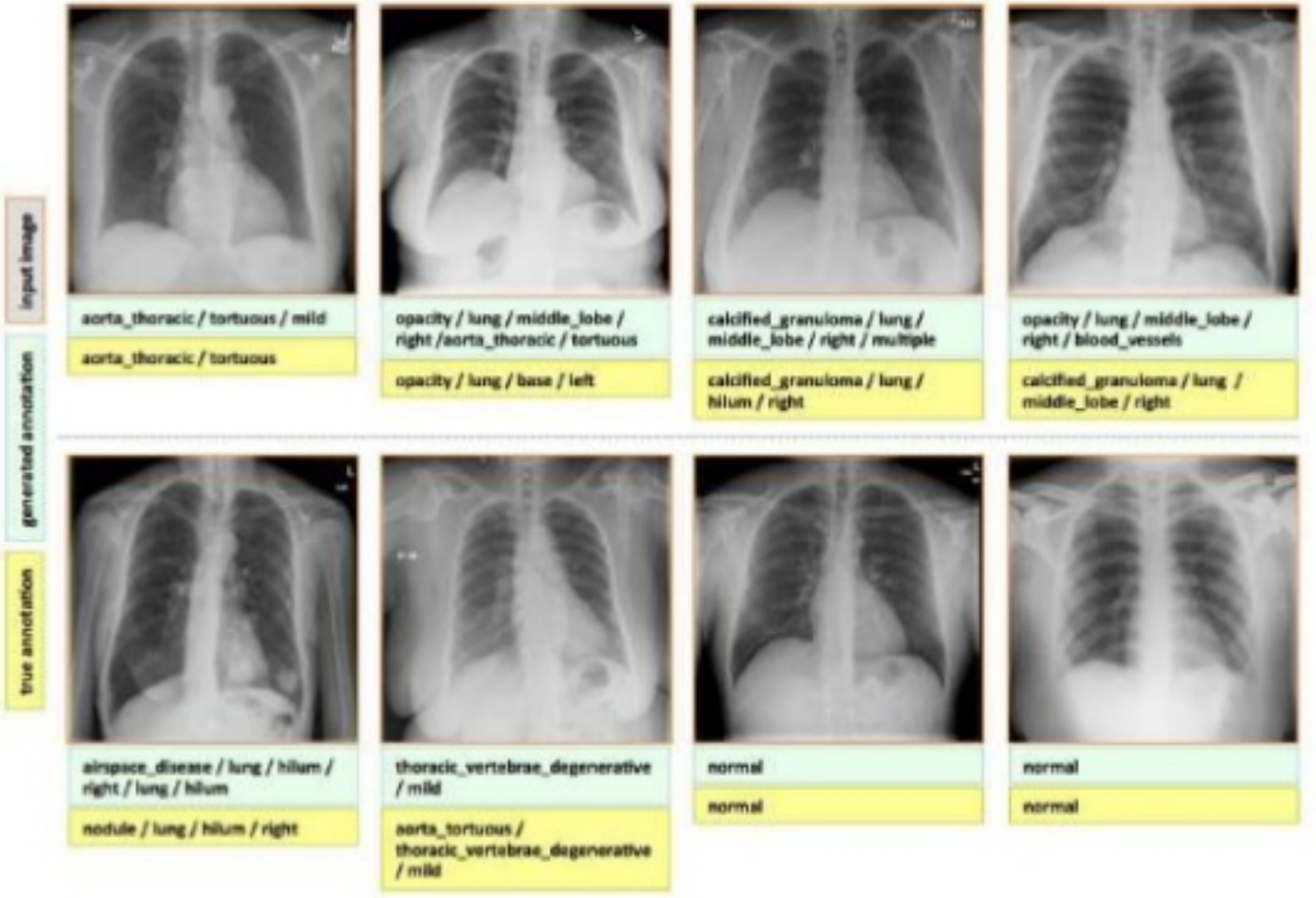
Major

Pulmonary Atelectasis / lingula / focal
Calcinosis / lung / hilum / right

Labelled training examples

Trained model

Generating image annotation:

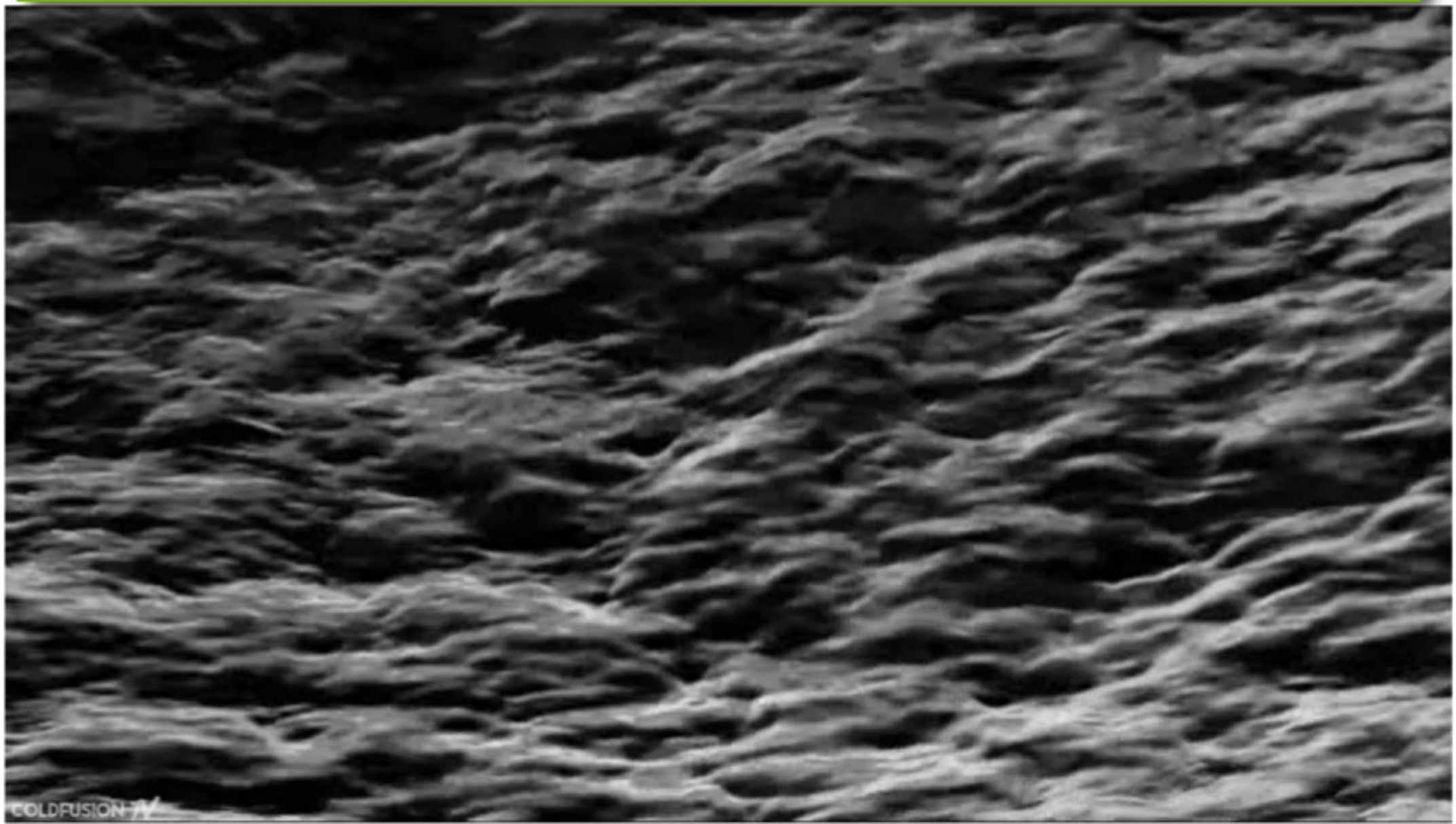


Inference applied to unseen inputs

A REVOLUTION IN ROBOTICS

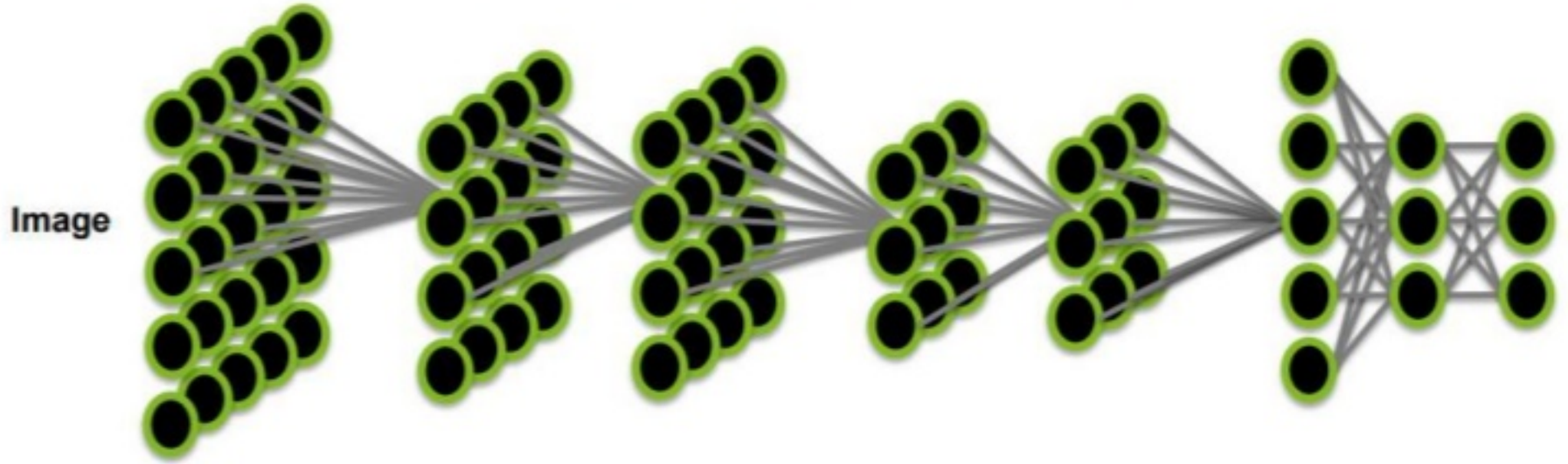
GPU-POWERED SELF-DRIVING CARS

SUPERHUMAN PERFORMANCE



WHAT DOES DEEP LEARNING LEARN?

Input → Feature Representation → Learning Algorithm



Today's Largest Networks

~10 layers

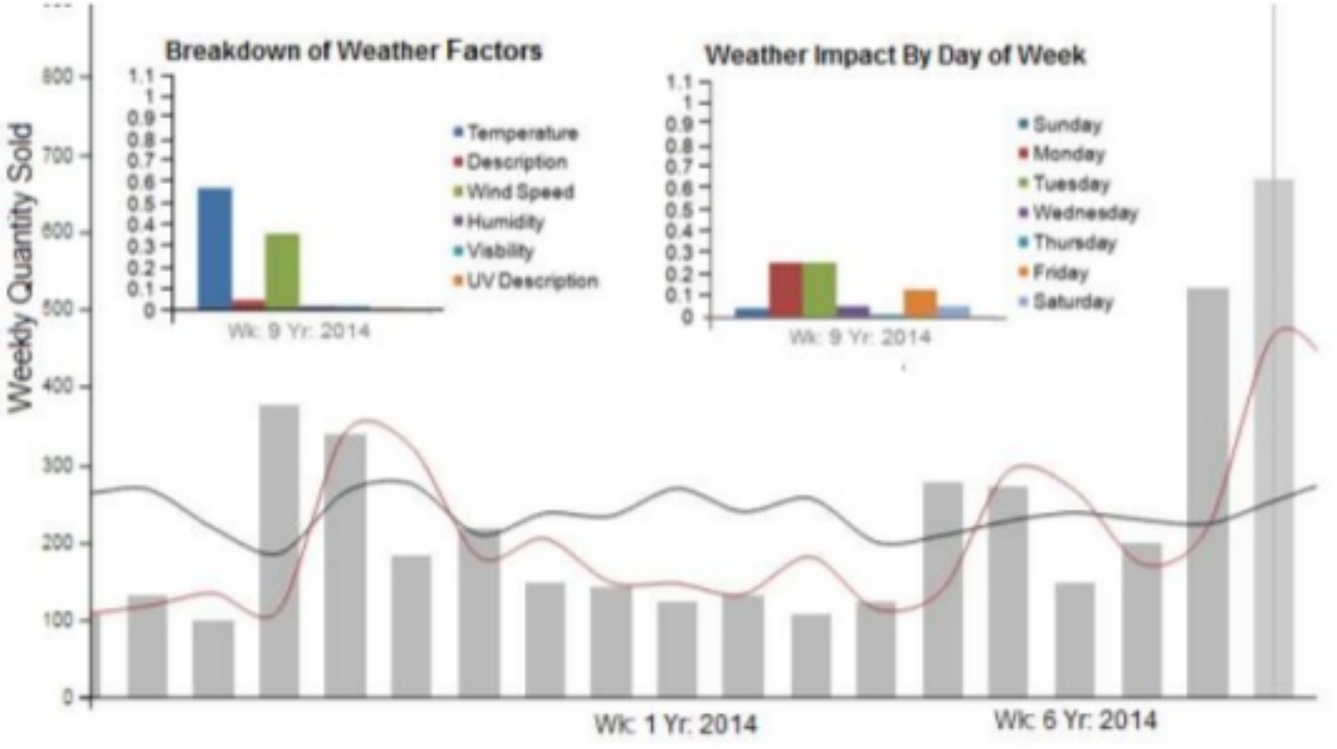
1B parameters

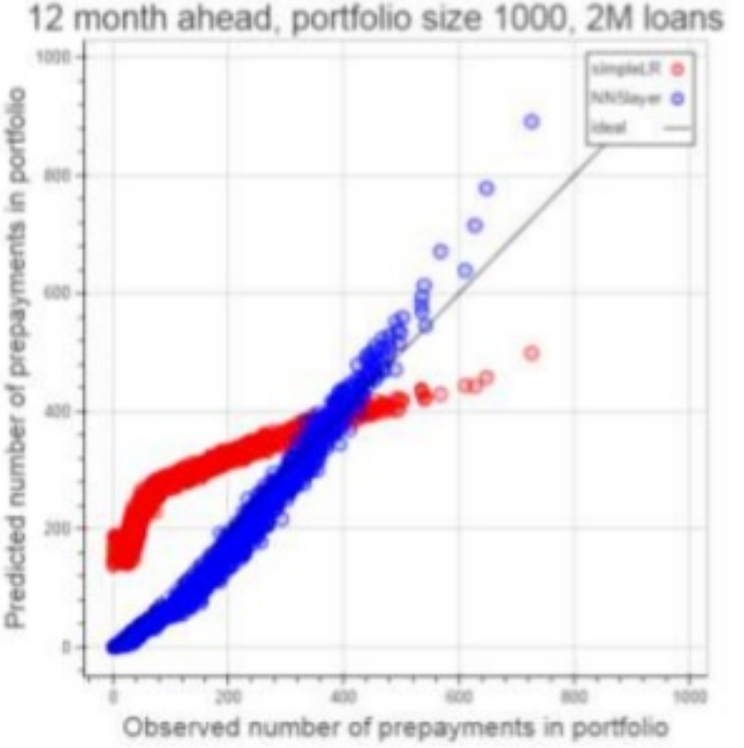
10M images

~30 Exaflops

Days - weeks to train

PREDICTIVE ANALYTICS IS NEXT



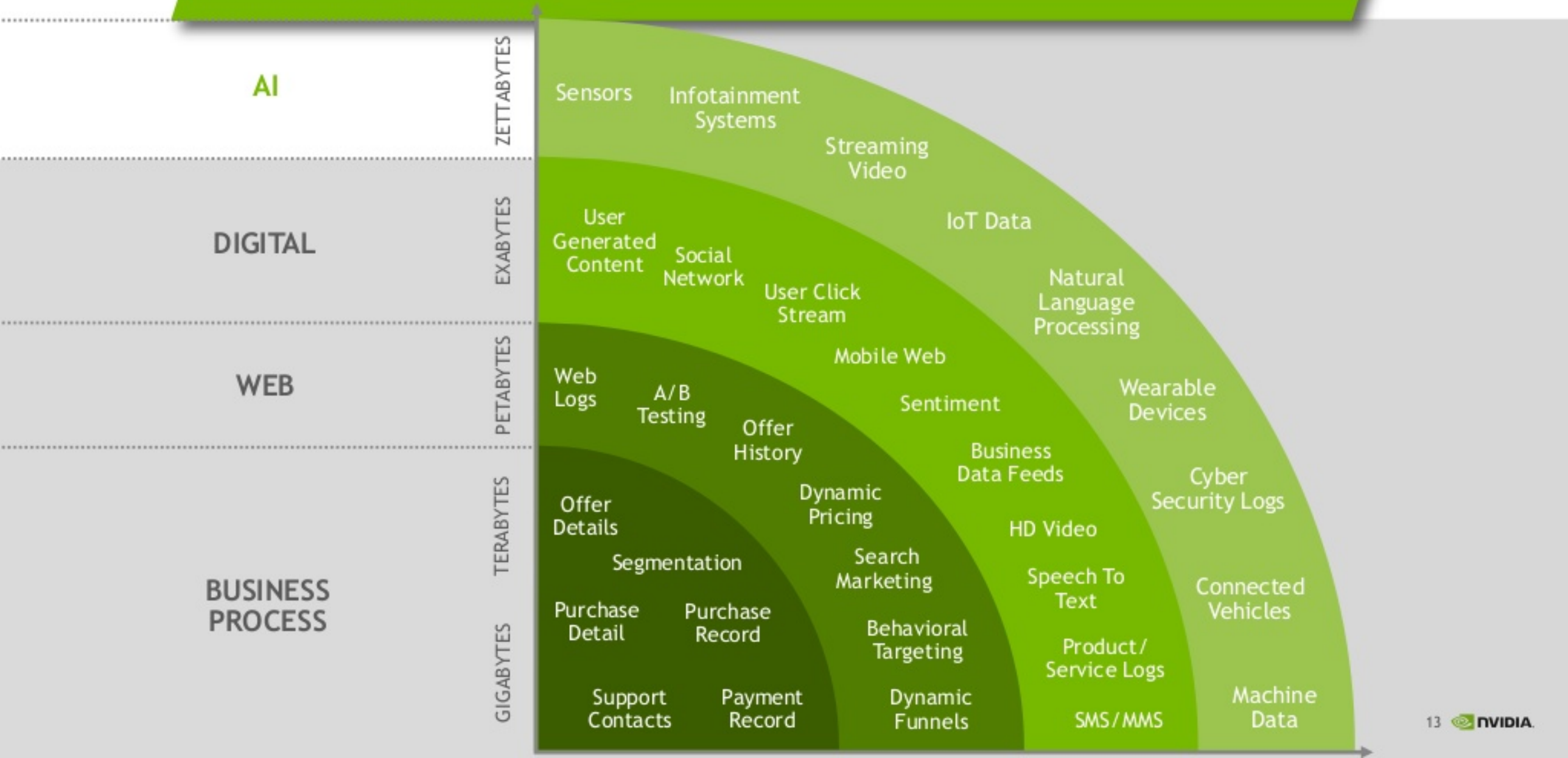


PREDICTIVE ANALYTICS IS NEXT

10,000s of features make up todays fraudulent behavior. AI can detect patterns faster and more accurate than humans

-Hui Wang, Senior Director of Global Risk Sciences, Pay Pal

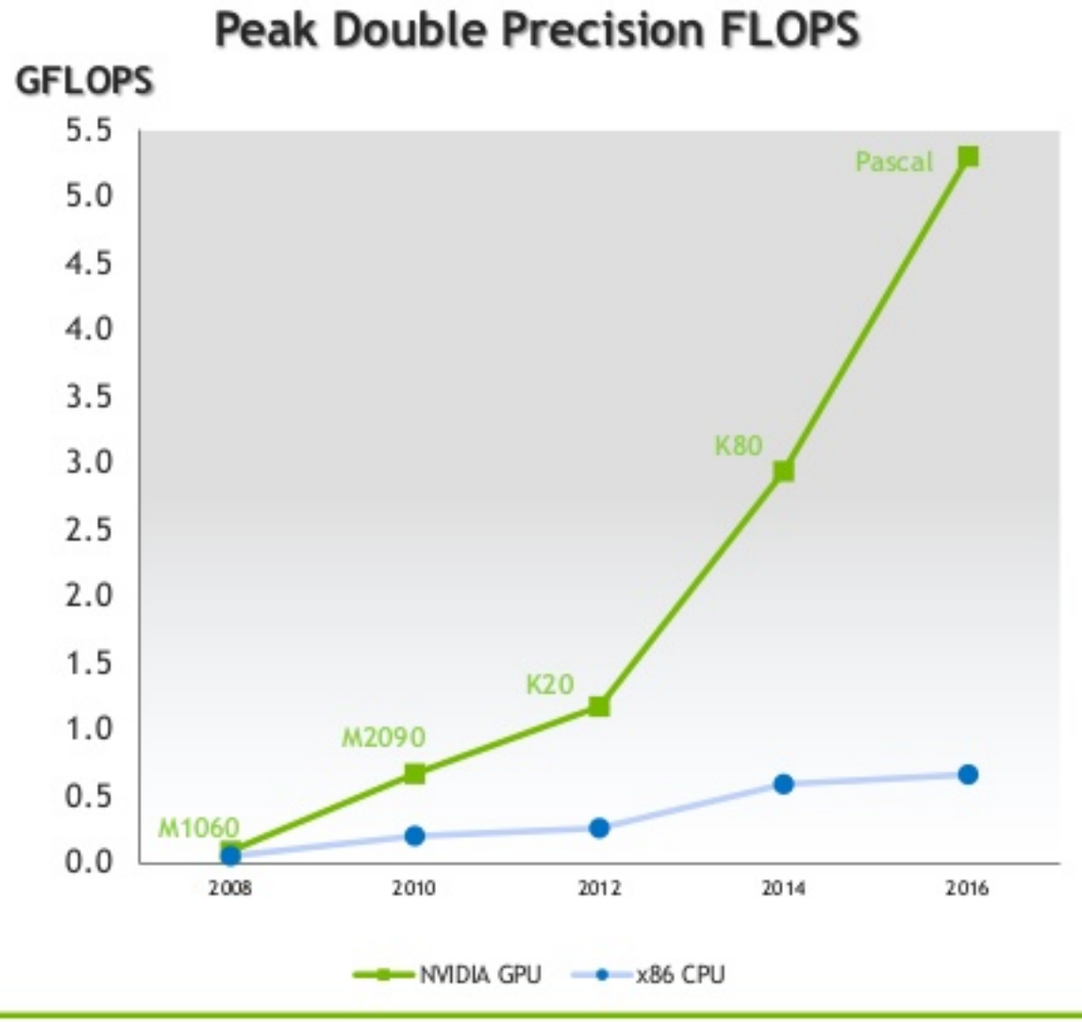
THE NEED TO SCALE UP & OUT IS HUGE

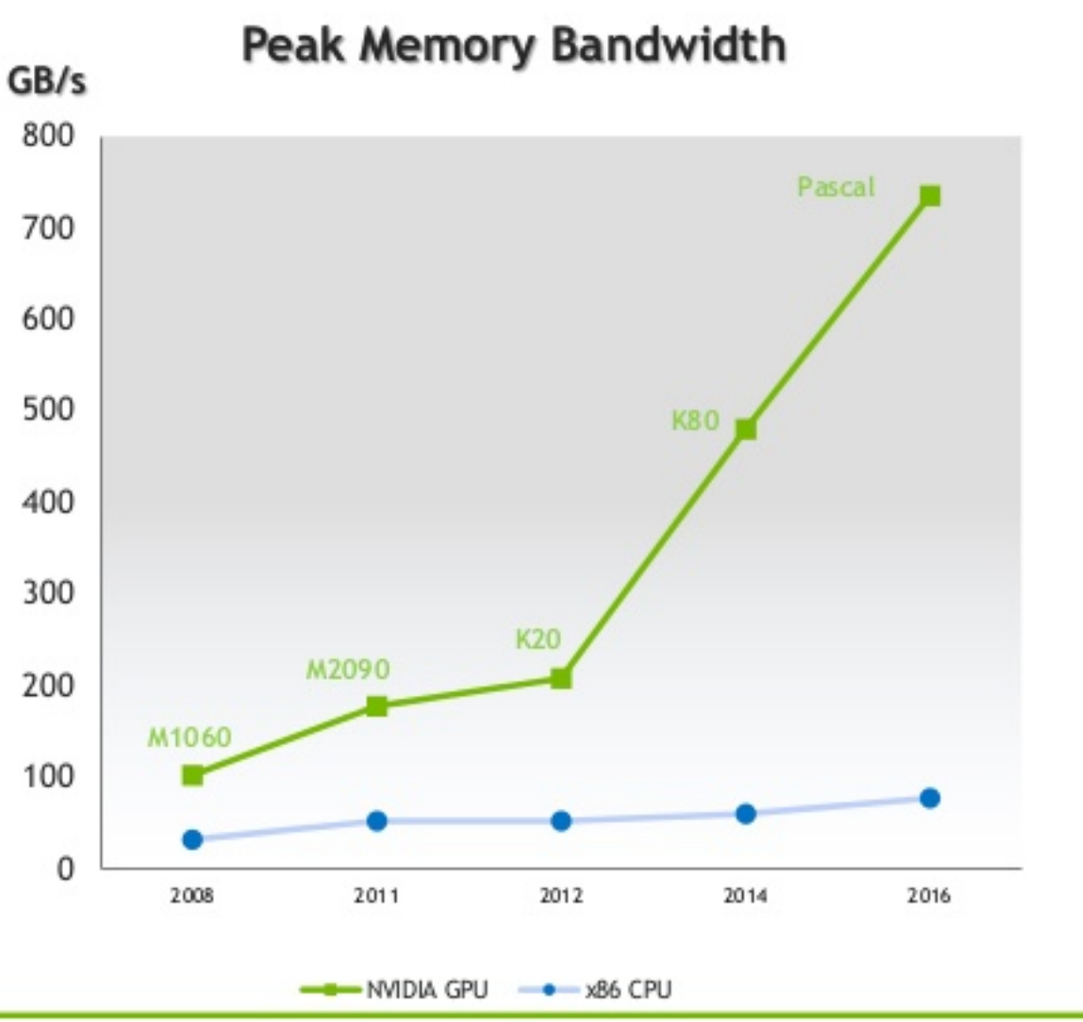


DGX-1 DEEP LEARNING SUPERCOMPUTER

Engineered for deep learning | 170TF FP16 | 8x Tesla P100 in hybrid cube mesh | Accelerates major AI frameworks

PERFORMANCE GAP INCREASES





HOW TO SCALE DATA ANALYTICS?

In practice, compute:

with:

$$\frac{1}{L}\sum_{x} score\,(x)$$

In a nutshell: a complex numerical function

$$score\,(x) = \log\left[\frac{1}{Nb\sqrt{2\pi}}\sum_{k=1}^{N}\exp\left(-\frac{(x-z_k)^2}{2b^2}\right)\right]$$