INTRO TO DEEP LEARNING

Carlo Nardone, Sr. Solution Architect, EMEA

February 2017

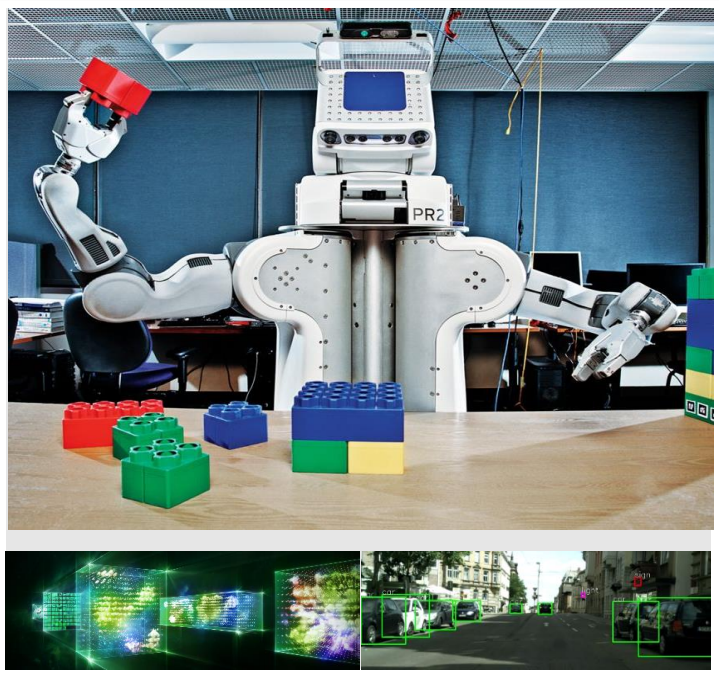
NVIDIA

HPC ADMINTECH 2017

NVIDIA
“THE AI COMPUTING COMPANY”

Computer Graphics

GPU Computing



Artificial Intelligence

THE EXPANDING UNIVERSE OF MODERN AI

" THE BIG BANG "

Big Data
GPU
Algorithms

RESEARCH

Berkeley
Carnegie Mellon University
DEEPMIND
Massachusetts Institute of Technology
NYU
OpenAI
Université de Montréal
UNIVERSITY OF OXFORD
UNIVERSITY OF TORONTO

CORE TECHNOLOGY / FRAMEWORKS

facebook
torch
Preferred Networks
Chainer
Université de Montréal
theano
Google
TensorFlow
Berkeley
Caffe
Microsoft
CNTK
UNIVERSITY OF OXFORD
MatconvNet
NVIDIA
cuDNN

AI-as-a-PLATFORM

amazon web services
IBM Watson
Google
Microsoft Azure

START-UPS

api.ai
Personal Assistants
conversational interface

BLUE RIVER TECHNOLOGY
Agriculture
crop-yield optimization

clarifai
Tech
visual recognition platform

deep genomics
Genomics
genetic interpretation

drive.ai
Automotive
computer vision

MetaMind
eCommerce & Medical
recommendation engines

Morpho
Tech
computer vision

Orbital Insight
Geospatial
predictions from images

nervana
Tech
AI-as-a-service

SADAKO
Waste Management
sorting robots

SocialEyes
Medical
diabetic retinopathy

HOW ARE YOU
Education
teaching robots

1,000+ AI START-UPS
$5B IN FUNDING
Source: Venture Scanner

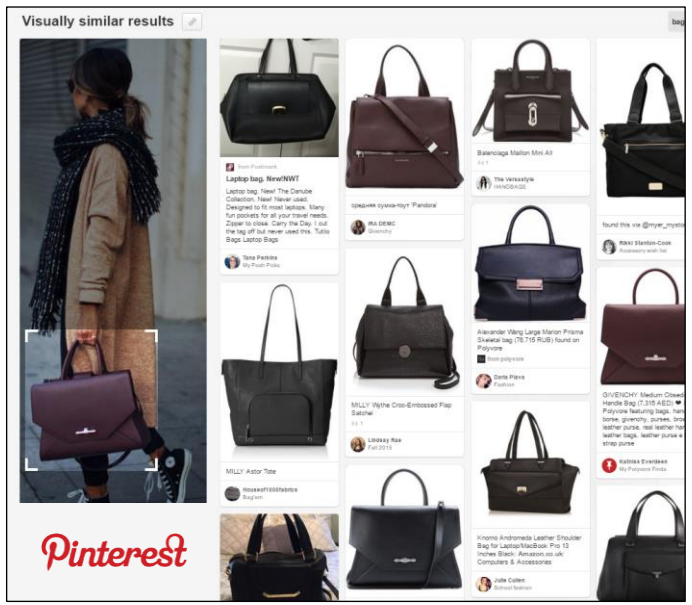
INDUSTRY LEADERS

Alibaba.com
AstraZeneca
Audi
Baidu 百度
Bloomberg
BMW
charles SCHWAB
CISCO
ebay
FANUC ROBOTICS
Ford
GE
gsk
THE HOME DEPOT
MASSACHUSETTS GENERAL HOSPITAL
Mercedes-Benz
MERCK
Pinterest
Schlumberger
Shell
SIEMENS
TARGET
TESLA
TOYOTA
UBER
VOLVO
Walmart
YAHOO
Yandex
yelp

© 2016 NVIDIA Corporation. All rights reserved. All other trademarks are property of their respective owners.

AI IS EVERYWHERE

"Find where I parked my car"



"Find the bag I just saw in this magazine"

"What movie should I watch next?"

FUELING ALL INDUSTRIES

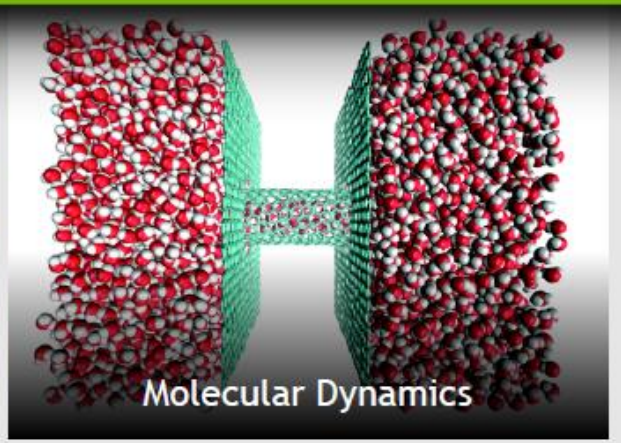
Increasing public safety with smart video surveillance at airports & malls

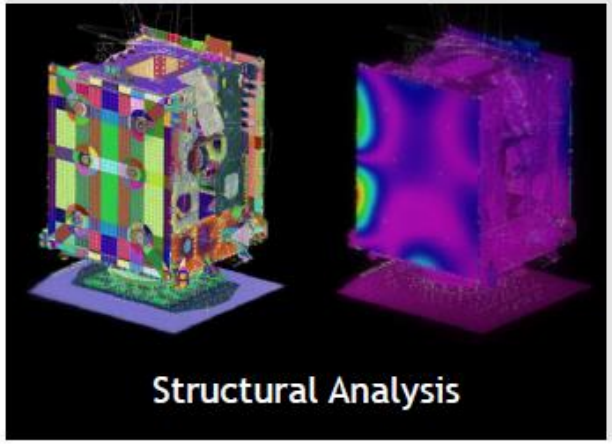
Providing intelligent services in hotels, banks and stores

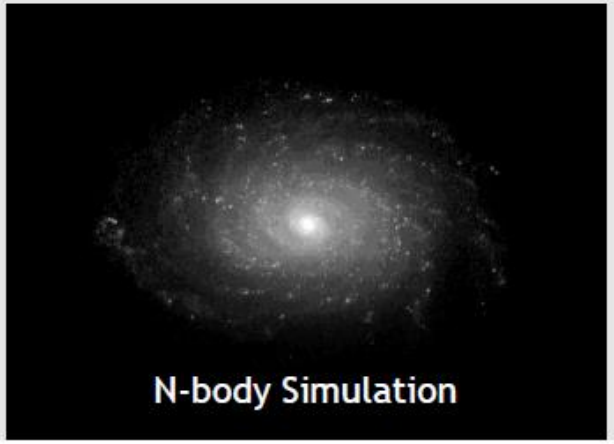
Separating weeds as it harvests, reduces chemical usage by 90%

AI & DEEP LEARNING: THE HPC «KILLER APP»

COMPUTATIONAL SCIENCE





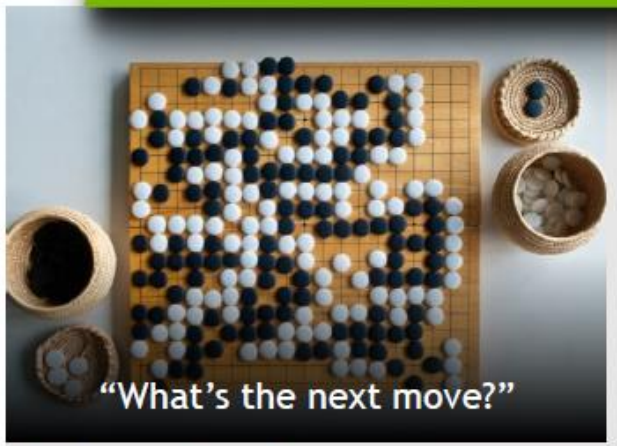


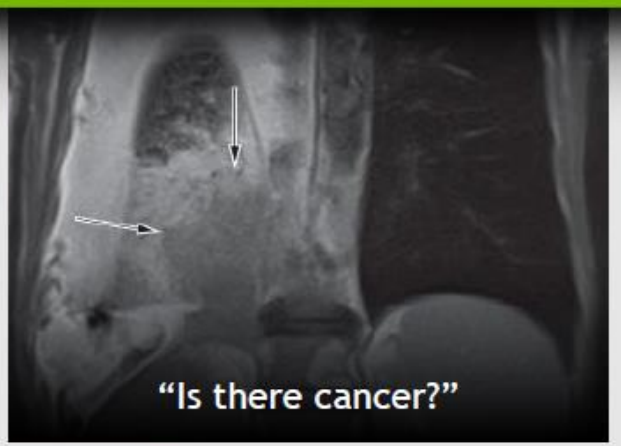
Numerical models to understand and predict physical and biological behavior

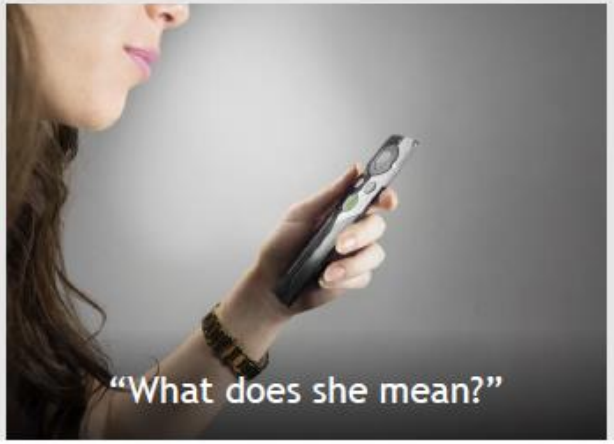
Based on the laws of physics — motion, gravity, mass-energy, thermodynamics, electrostatics

Computational methods like PDE, FEM, MC, LA

DATA SCIENCE







Combinatorial explosion

Incomplete information

No laws-of-physics equations exist

Deep learning extracts multi-dimensional features from data

Breakthrough for AI

DEEP LEARNING IS A SUPERCOMPUTING CHALLENGE

DATA PIPELINE

PROCESS
AUGMENT
AUTO LABEL
MANUAL LABEL
CURATE

PETABYTES OF DATA
PETAFLOPS MACHINE

TRAINING PIPELINE

MODEL CONFIGURATION
HYPERPARAMETER TUNING
MODEL TRAINING

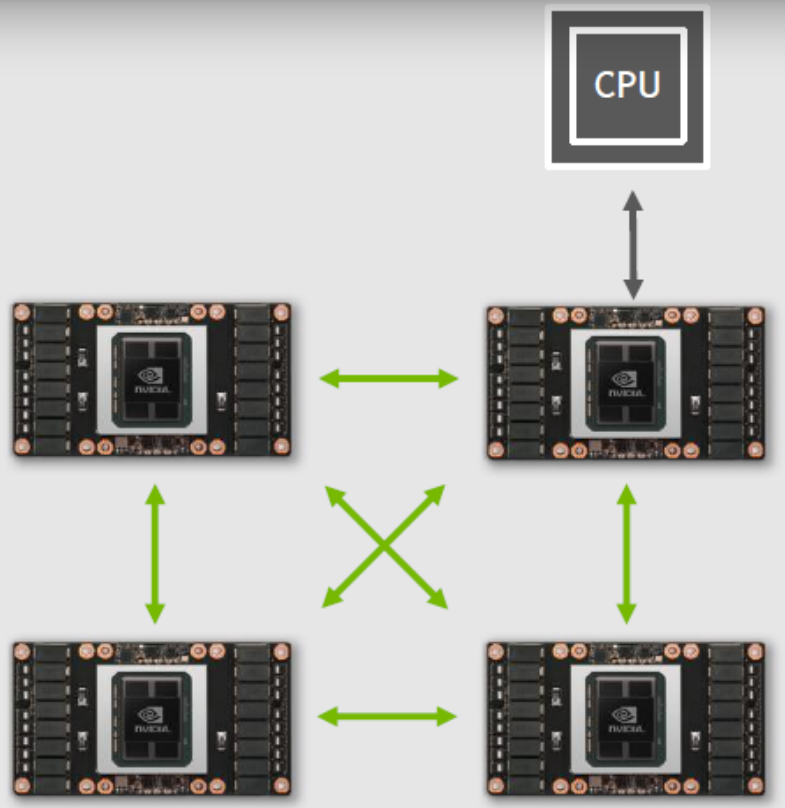
100's OF PETAFLOPS TO
EXAFLOPS MACHINES

INFERENCING

RECOGNIZE
CLASSIFY
PREDICT
GENERATE

100's OF GIGAFLOPS
TO TERAFLOPS

THE ENGINE FOR AI SUPERCOMPUTING

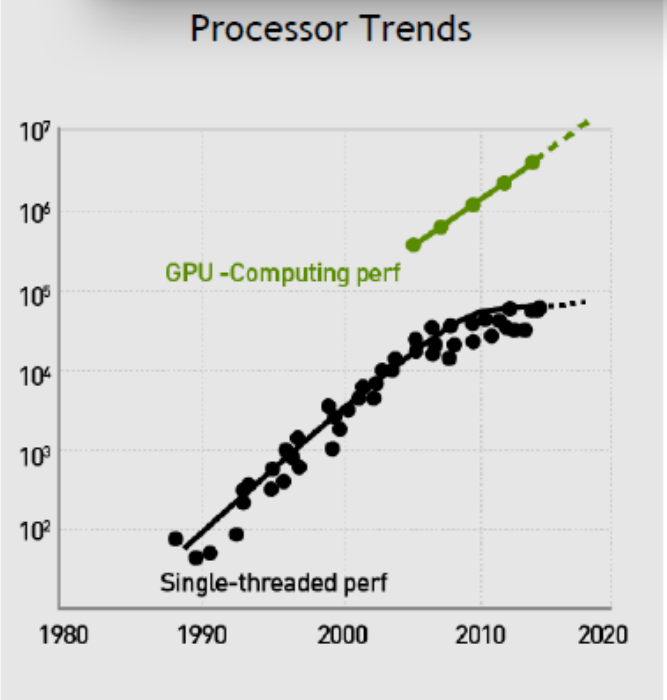


Future supercomputers designed for computational and data science

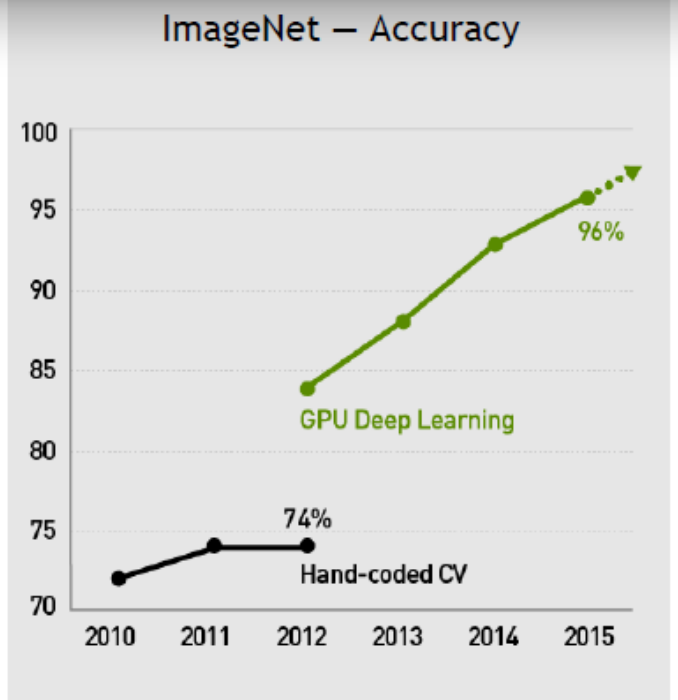
Strong CPU - Variable Precision Computation - High-Speed Links

4X	5.3 TF	FP64
4X	10.6 TF	FP32
4X	21.2 TF	FP16
	640GB/s	NVLink

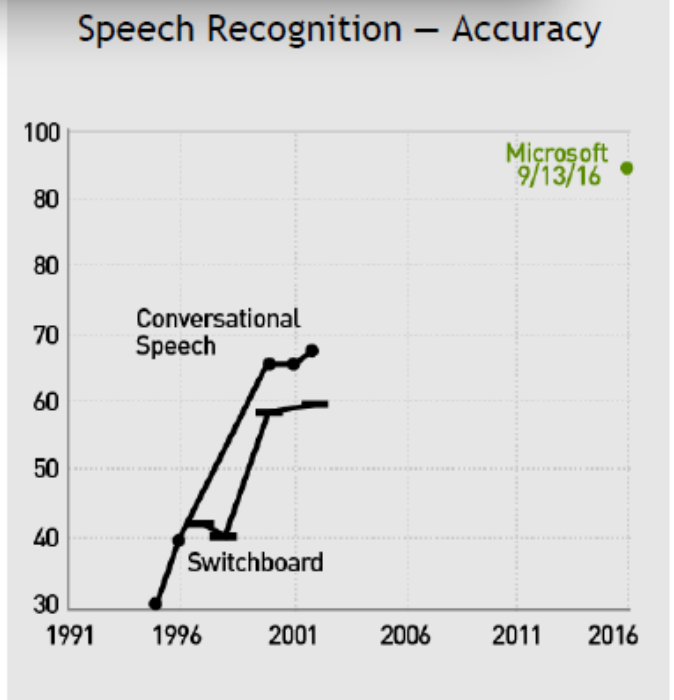
AI IS THE PATH TO EXASCALE



GPU Boosts HPC

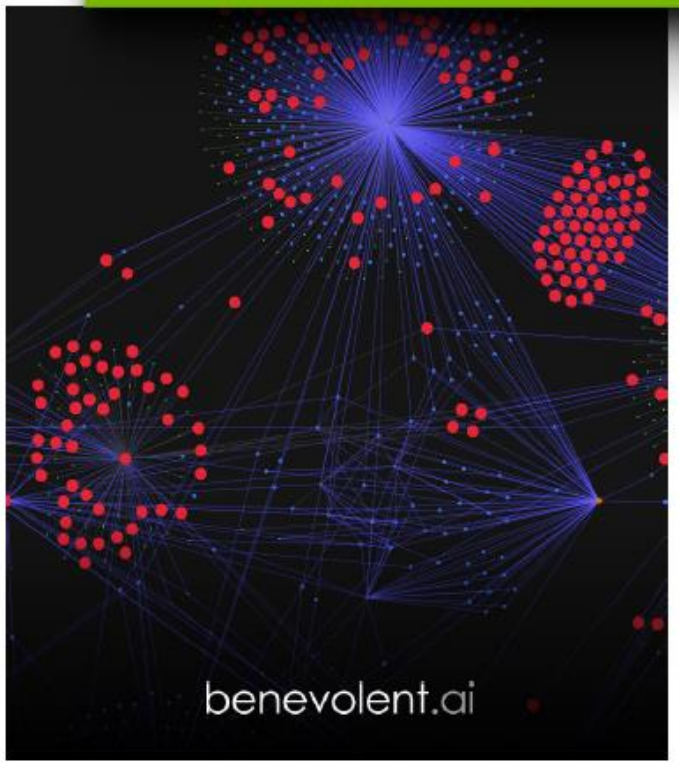


GPU Boosts AI

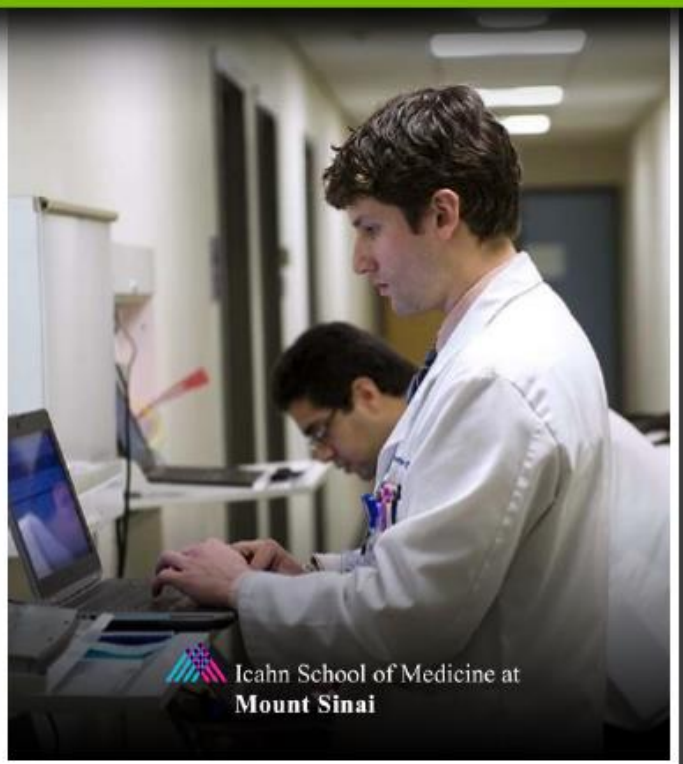


GPU Boosts AI

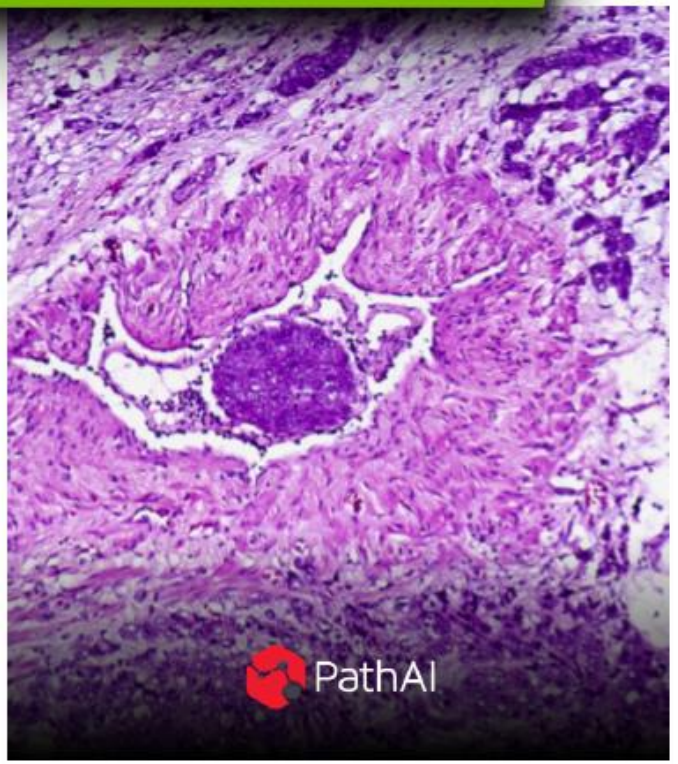
AI IS REVOLUTIONIZING HEALTHCARE



Accelerating Targeted Drug Development



Predicting Disease from Medical Records



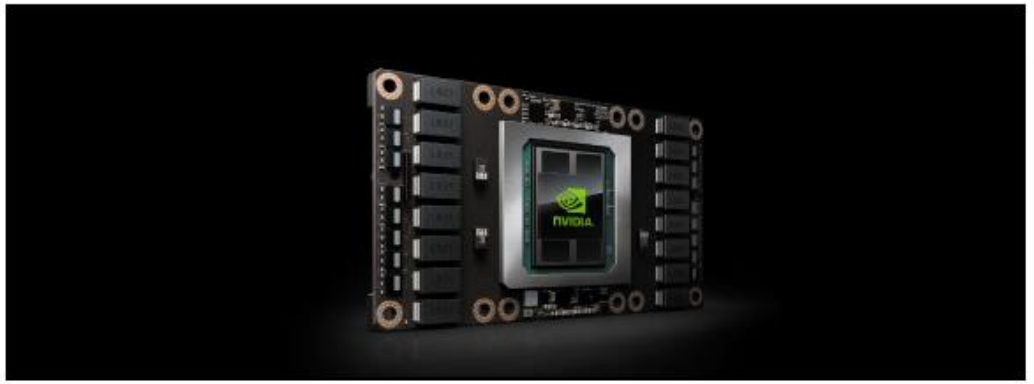
Reducing Cancer Diagnosis Error Rate by 85%

NVIDIA AI COMPUTING PLATFORM

One Architecture Everywhere

Alibaba Cloud aliyun.com
amazon web services
IBM Cloud
Microsoft Azure

GPU DL-as-a-Service



Advance GPU Deep Learning

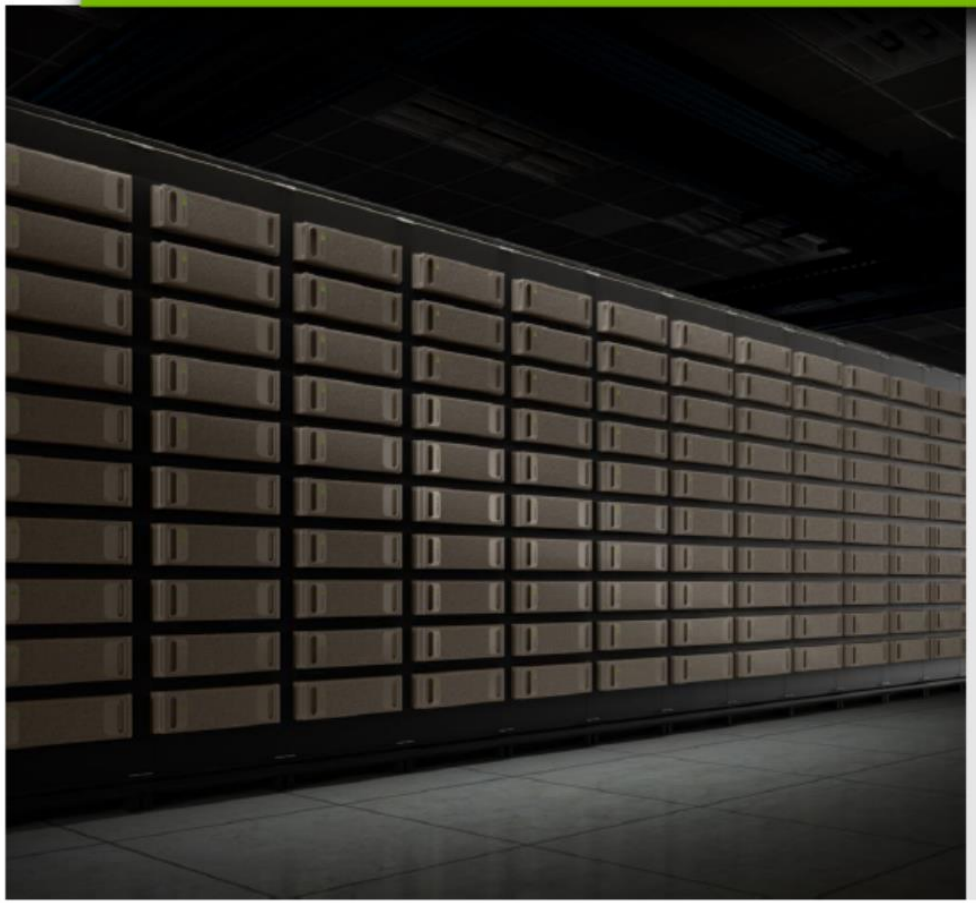
Caffe
Chainer
Microsoft CNTK
mxnet
PaddlePaddle Baidu Deep Learning
TensorFlow
theano
torch

Accelerate Every Framework

GTC, DLI, Inception

INTRODUCING DGX SATURNV

124 NVIDIA DGX-1 "Rocket for Cancer Moonshot"



Fastest AI Supercomputer in TOP500

4.9 Petaflops Peak FP64
19.6 Petaflops Peak FP16

Most Energy Efficient Supercomputer

#1 Green500
9.5 GFLOPS per Watt

Rocket for Cancer Moonshot

CANDLE Development Platform
Common platform with DOE labs — ANL, LLNL, ORNL, LANL

NVIDIA AI COMPUTING FOR EVERY INDUSTRY

AI Transportation

AI Factory

CANDLE

Argonne NATIONAL LABORATORY
Lawrence Livermore National Laboratory
Los Alamos NATIONAL LABORATORY EST. 1943
NATIONAL CANCER INSTITUTE
OAK RIDGE National Laboratory
U.S. DEPARTMENT OF ENERGY

AI Healthcare

Microsoft Azure

The Microsoft Cognitive Toolkit

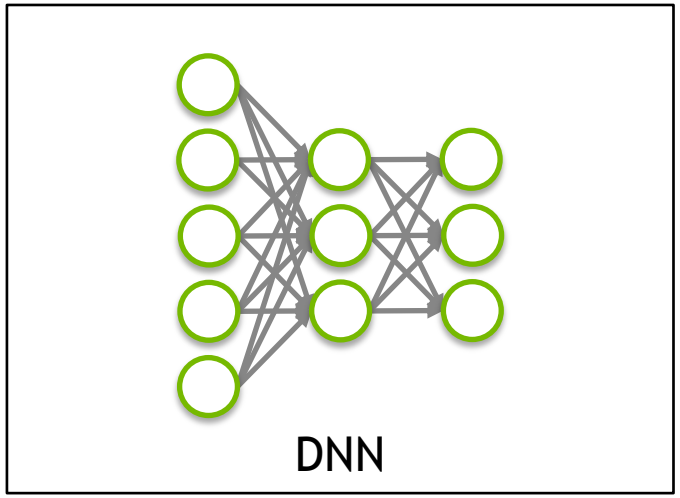
IBM

PowerAI Toolkit
Minsky

AI Enterprise

MACHINE LEARNING & DEEP LEARNING

THE BIG BANG IN MACHINE LEARNING

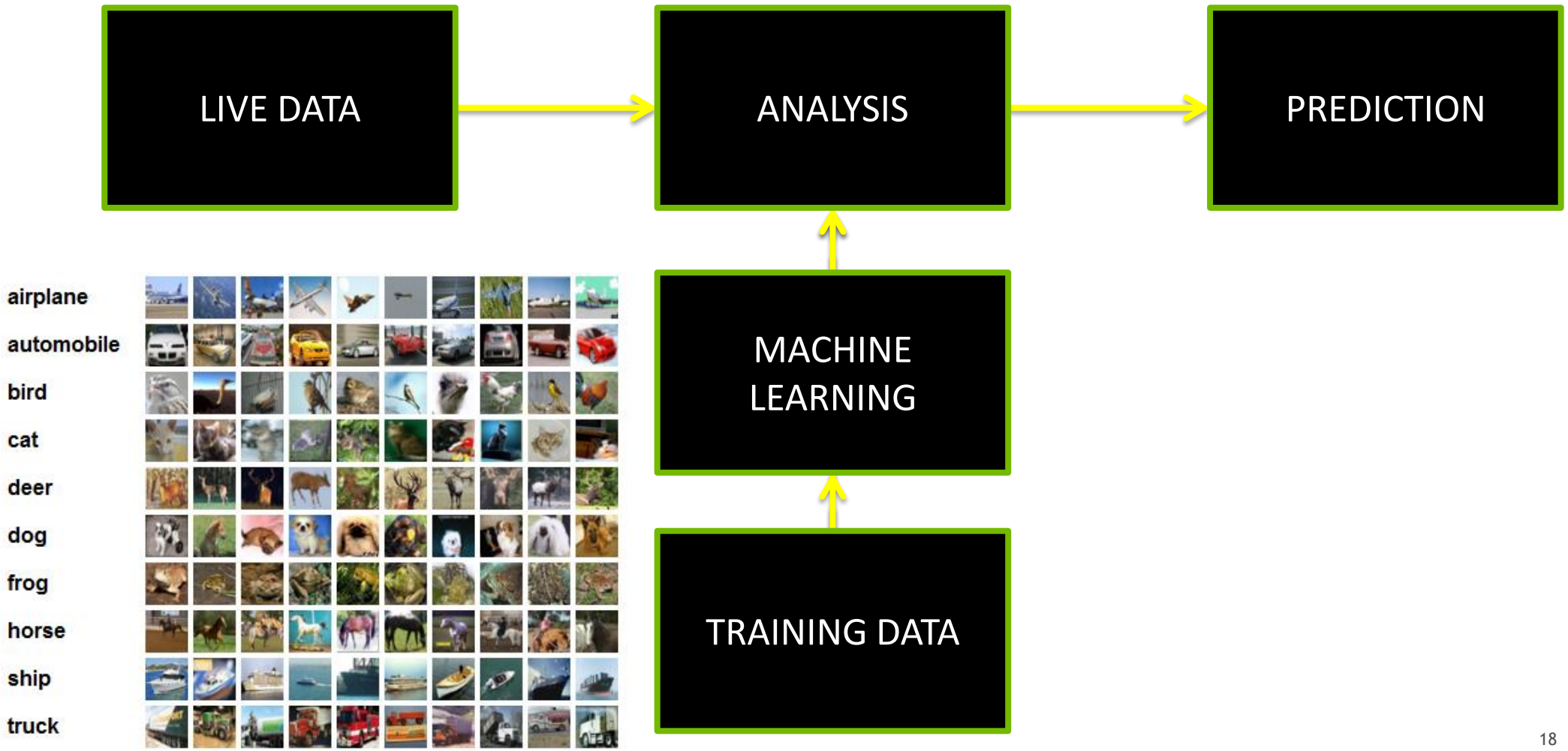


"The GPU is the workhorse of modern A.I."

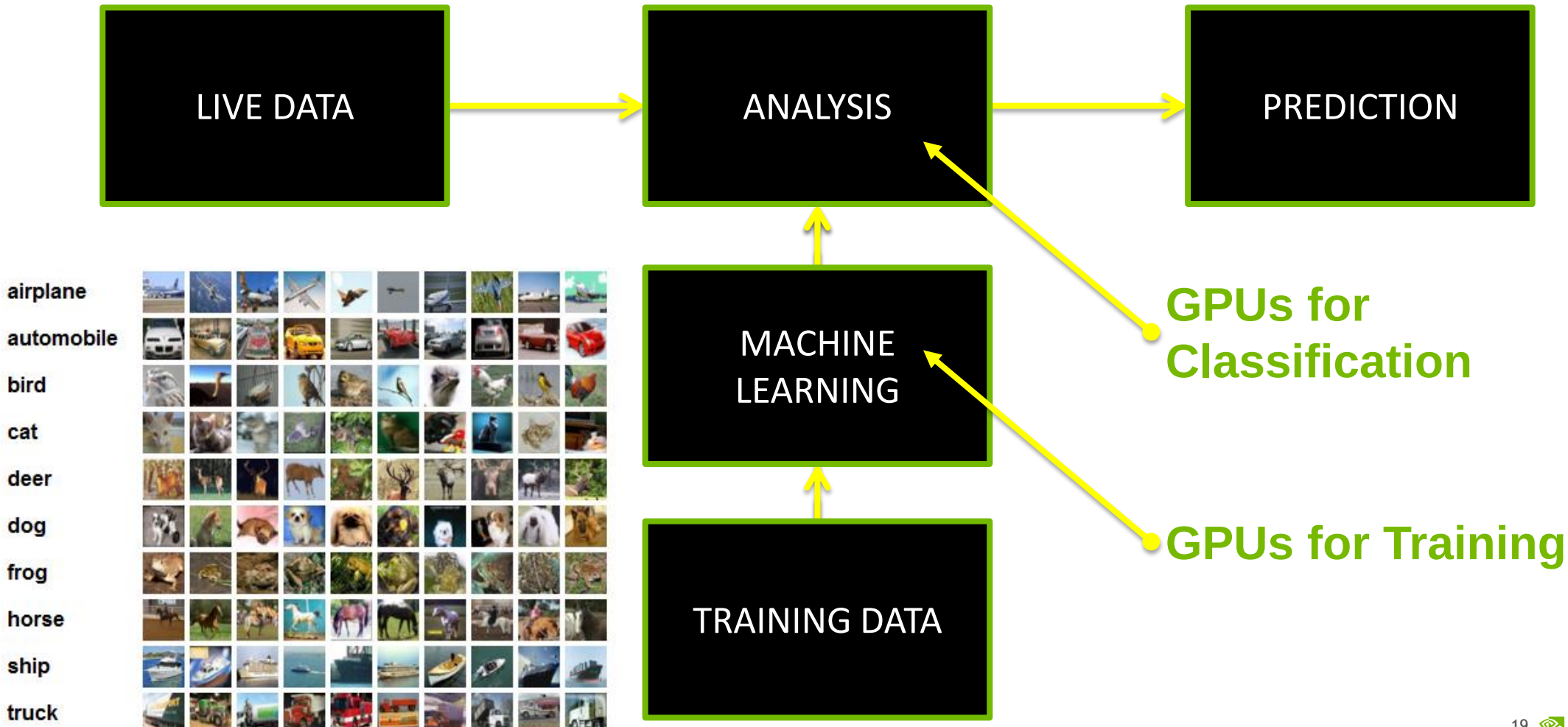
POPULAR SCIENCE

Provided under Data:Lab Not for distribute
Do not copy

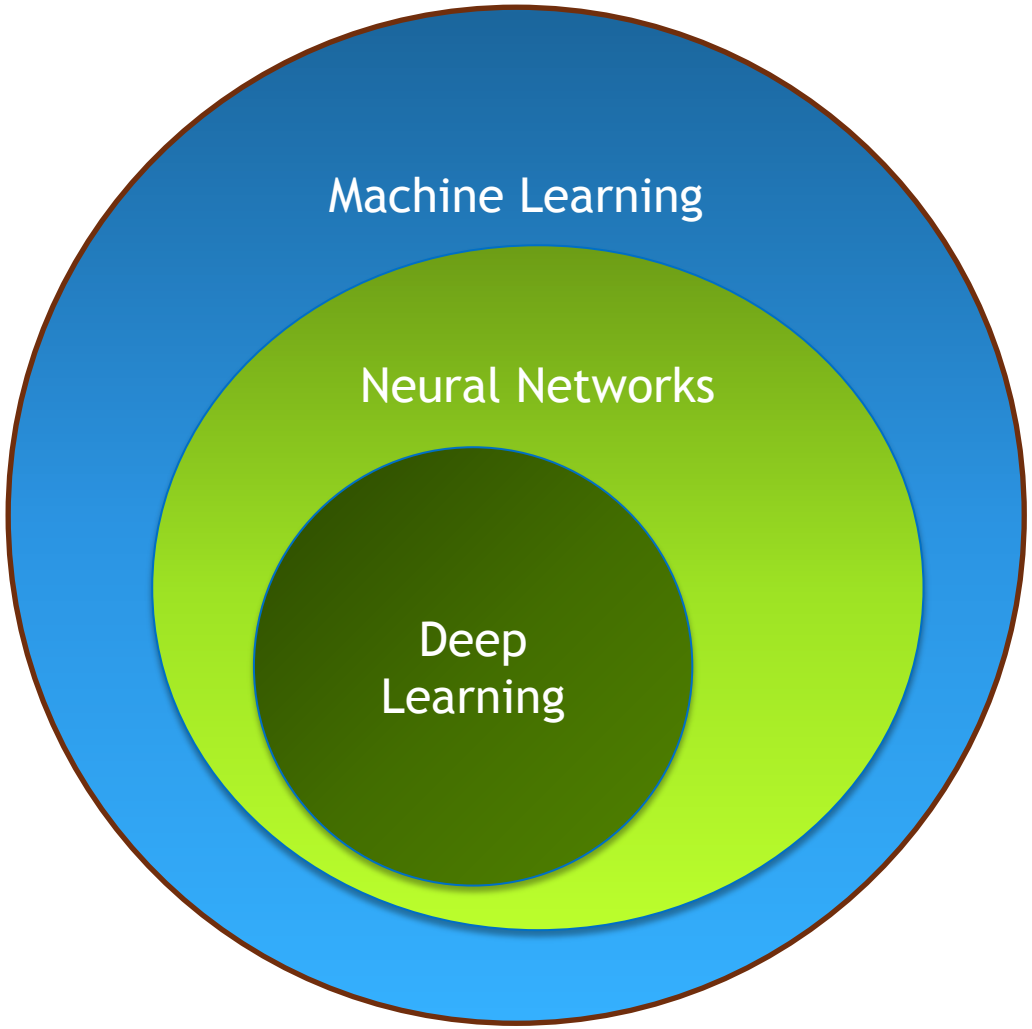
MACHINE LEARNING IN A NUTSHELL



MACHINE LEARNING IN A NUTSHELL

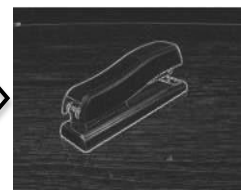
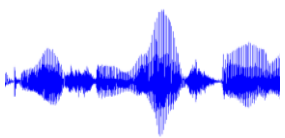
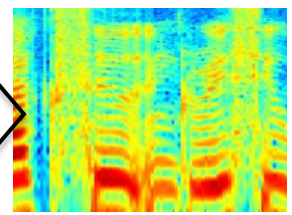
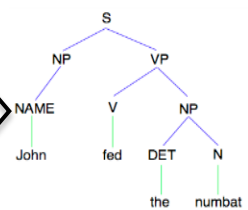


MACHINE LEARNING & DEEP LEARNING



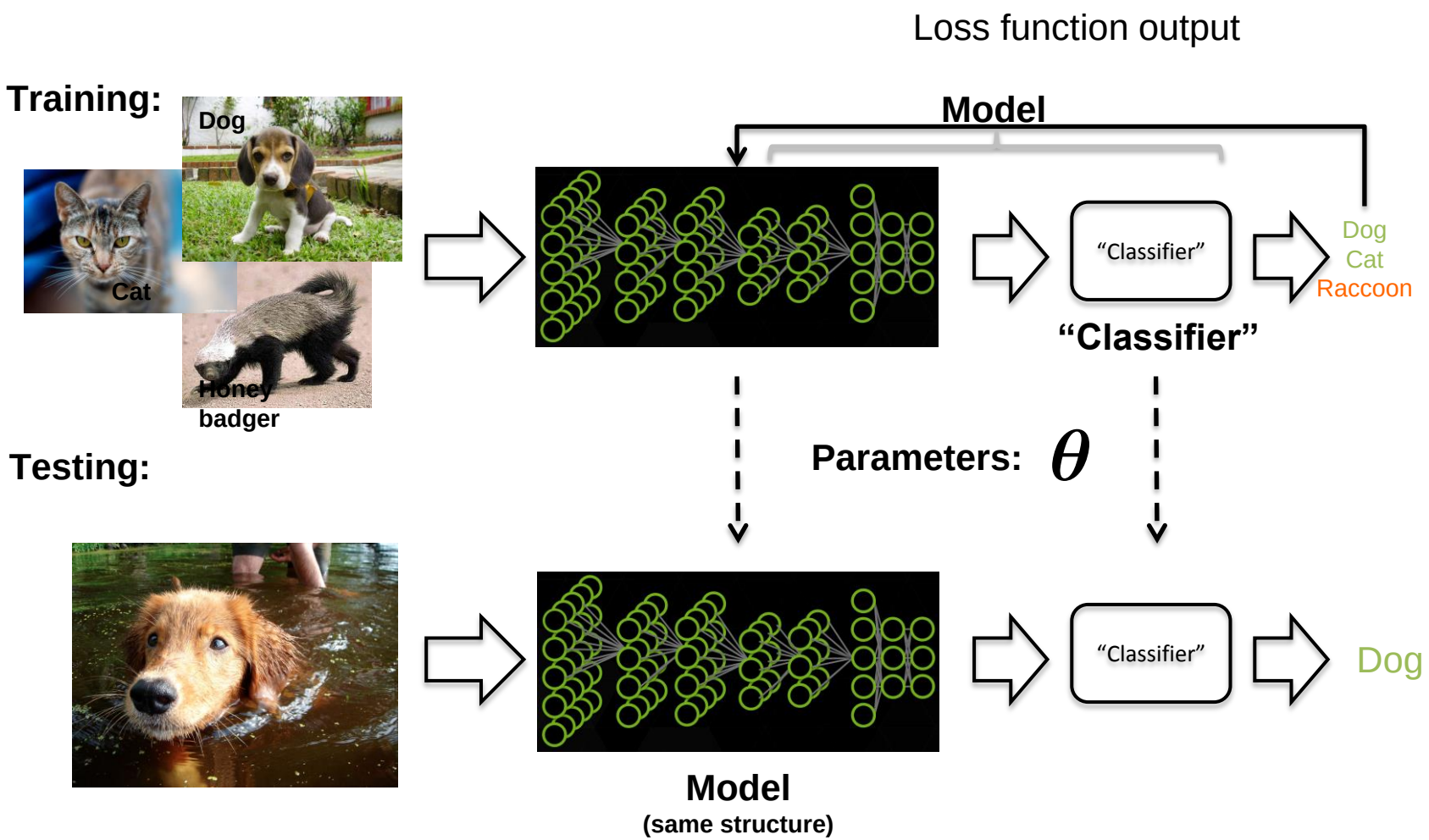
TRADITIONAL MACHINE PERCEPTION

Hand-tuned features

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

Source: courtesy of Andrew Ng, Stanford U.

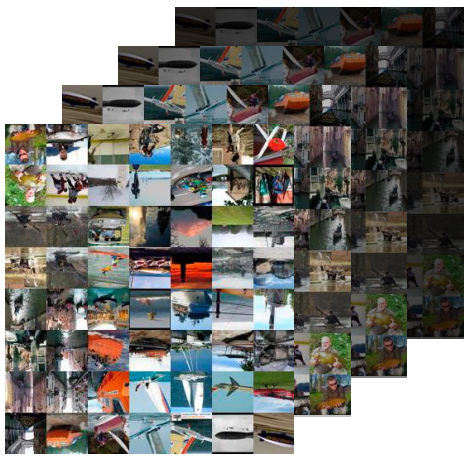
NEURAL NETWORKS APPROACH



Source: courtesy of Andrew Ng, Stanford U.

DEEP LEARNING — A NEW COMPUTING MODEL

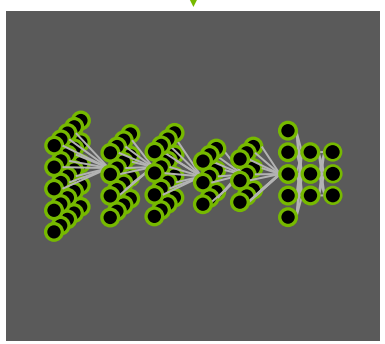
“Software that writes software”



(A lot of) Labeled Data

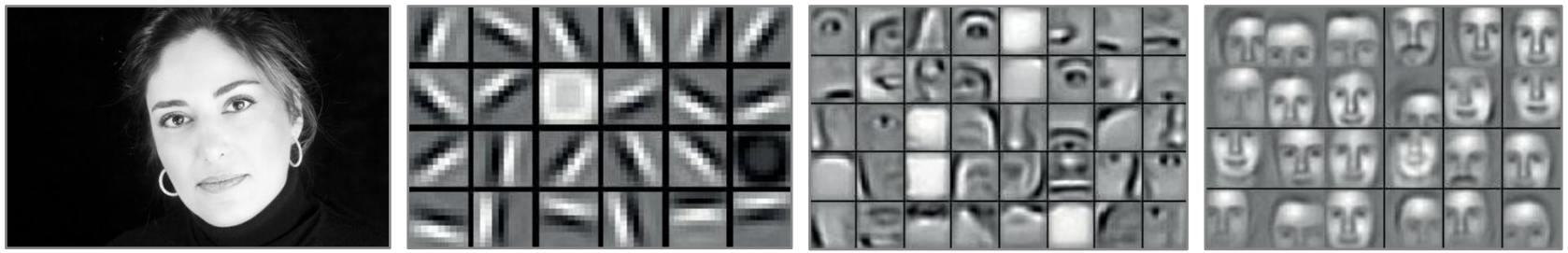
LEARNING ALGORITHM

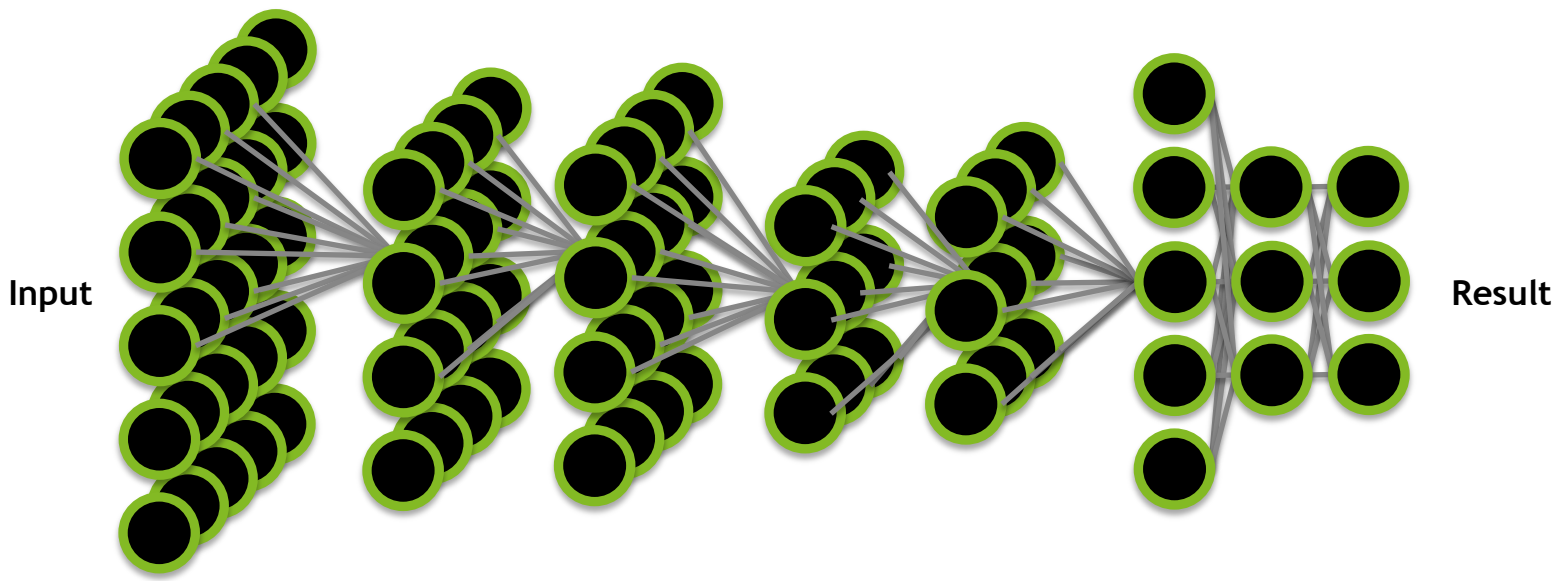
“millions of trillions of FLOPS”



“little girl is eating piece of cake"

DEEP NEURAL NETWORK





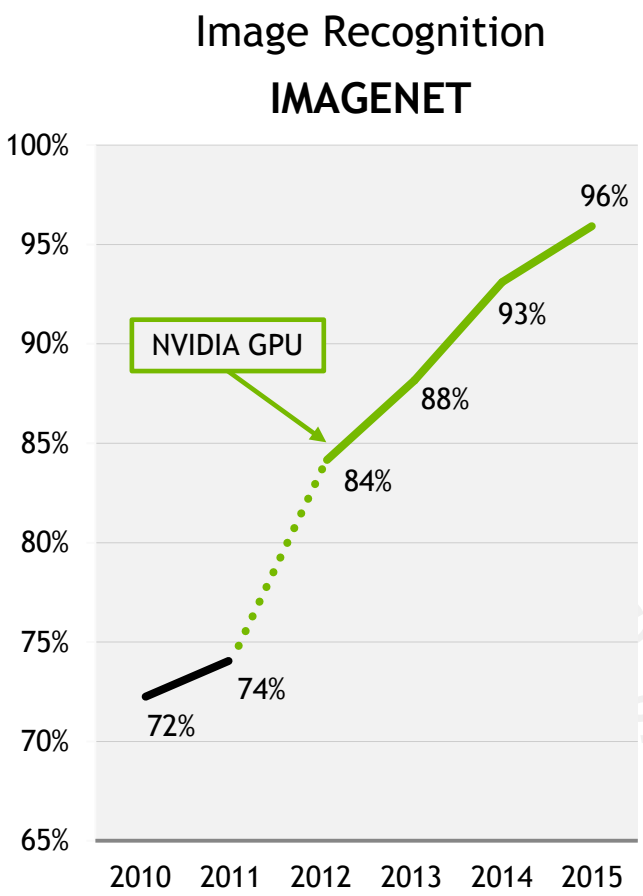
Today's Largest Networks

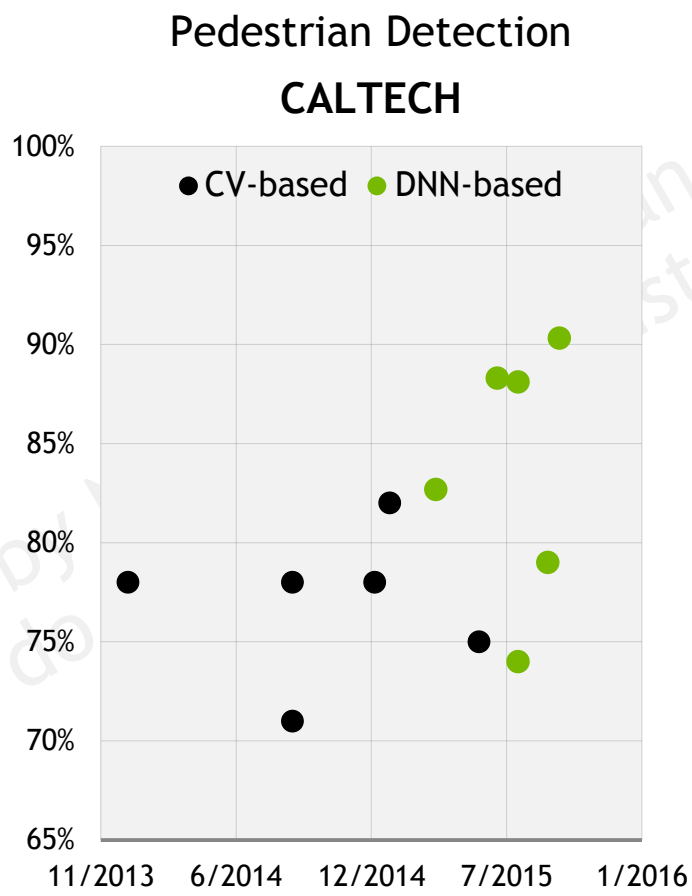
~10 layers
1B parameters
10M images
~30 Exaflops
~30 GPU days

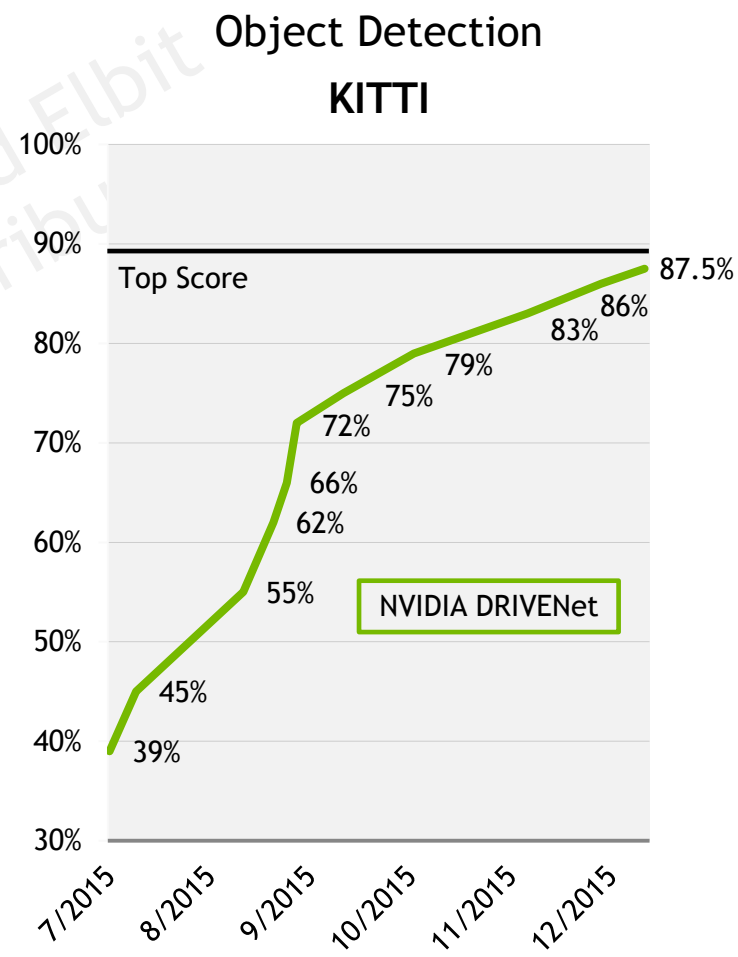
Human brain has trillions of parameters – 1,000 more.

Image source: "Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks" ICML 2009 & Comm. ACM 2011. Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Ng.

AMAZING RATE OF IMPROVEMENT

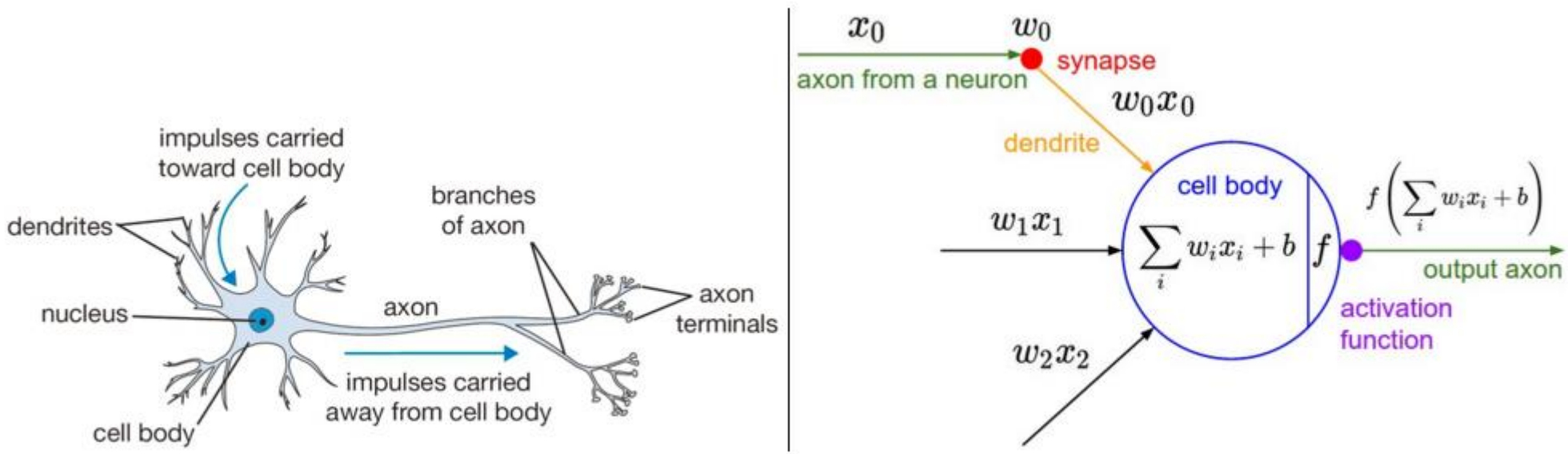






DEEP NEURAL NETWORKS 101

BIOLOGICALLY INSPIRED "NEURON"



A cartoon drawing of a biological neuron (left) and its mathematical model (right).

Source: @karpathy U.Stanford course CS231n http://cs231n.github.io/

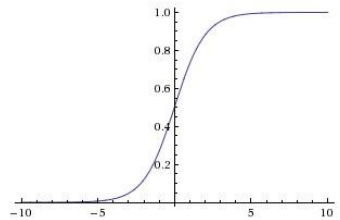
ACTIVATION FUNCTION

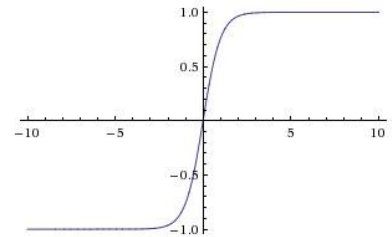
Sigmoid – squashes to range [0,1] $\sigma(x) = \frac{1}{(1+e^{-x})}$

Tanh - squashes to range [-1, 1].

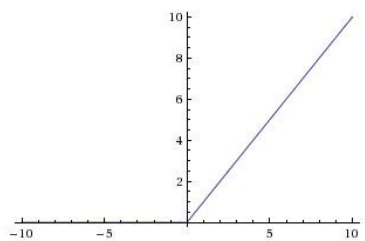
ReLU – squashes to range [0,1] $f(x) = max(0, x)$

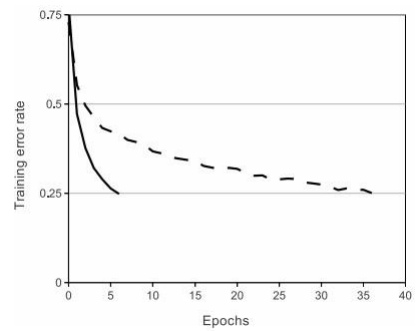
… many more in literature





Left: *Sigmoid non-linearity squashes real numbers to range between [0,1]* ***Right:*** *The tanh non-linearity squashes real numbers to range between [-1,1].*





Left: *Rectified Linear Unit (ReLU) activation function, which is zero when x < 0 and then linear with slope 1 when x > 0.* ***Right:*** *A plot from Krizhevsky et al. (pdf) paper indicating the 6x improvement in convergence with the ReLU unit compared to the tanh unit.*

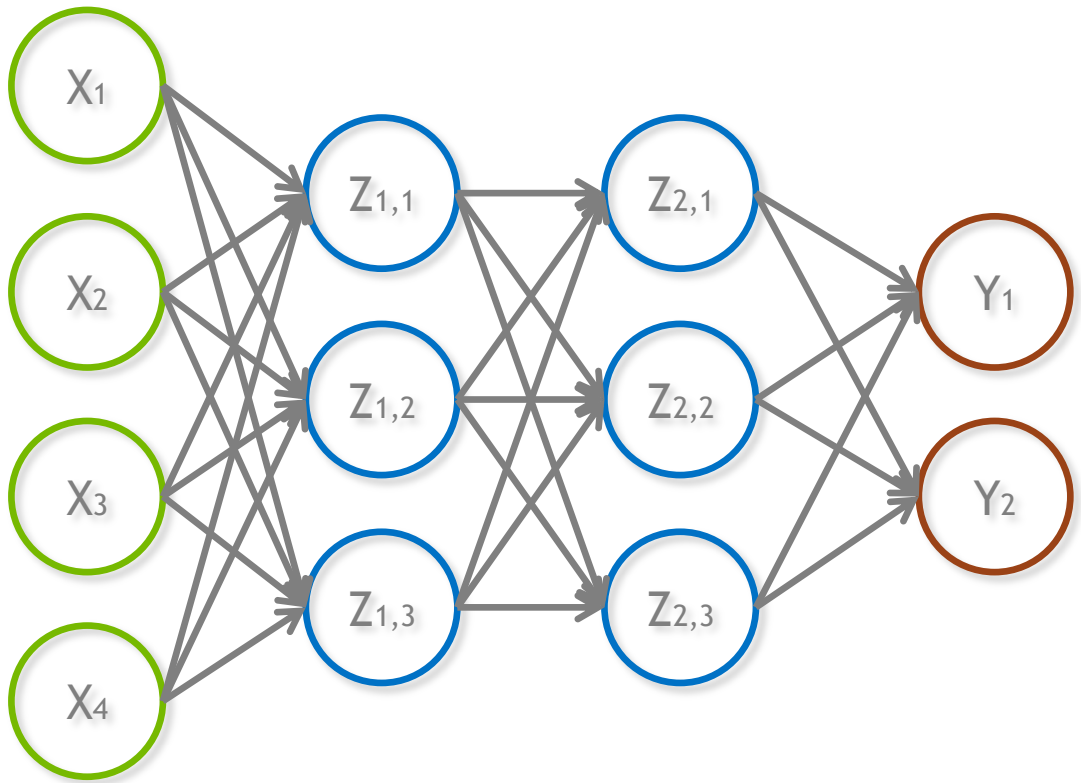
FEED-FORWARD NEURAL NETWORKS

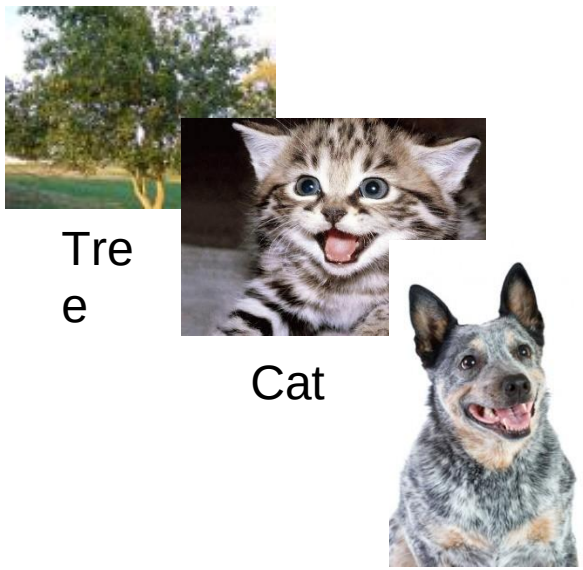
Multilayer architecture

Deep nets: with multiple hidden layers

Efficient training enabled with backpropagation, in 1974

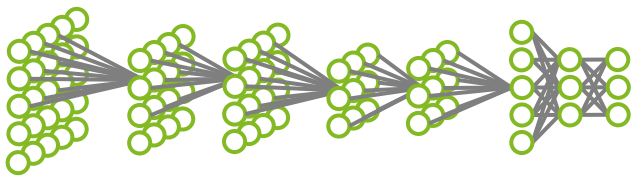
Each neuron and each weight parameter is just a single floating point number





Machine Learning Software

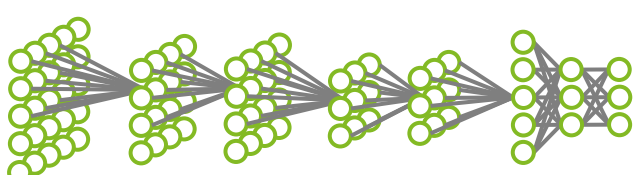
Forward Propagation

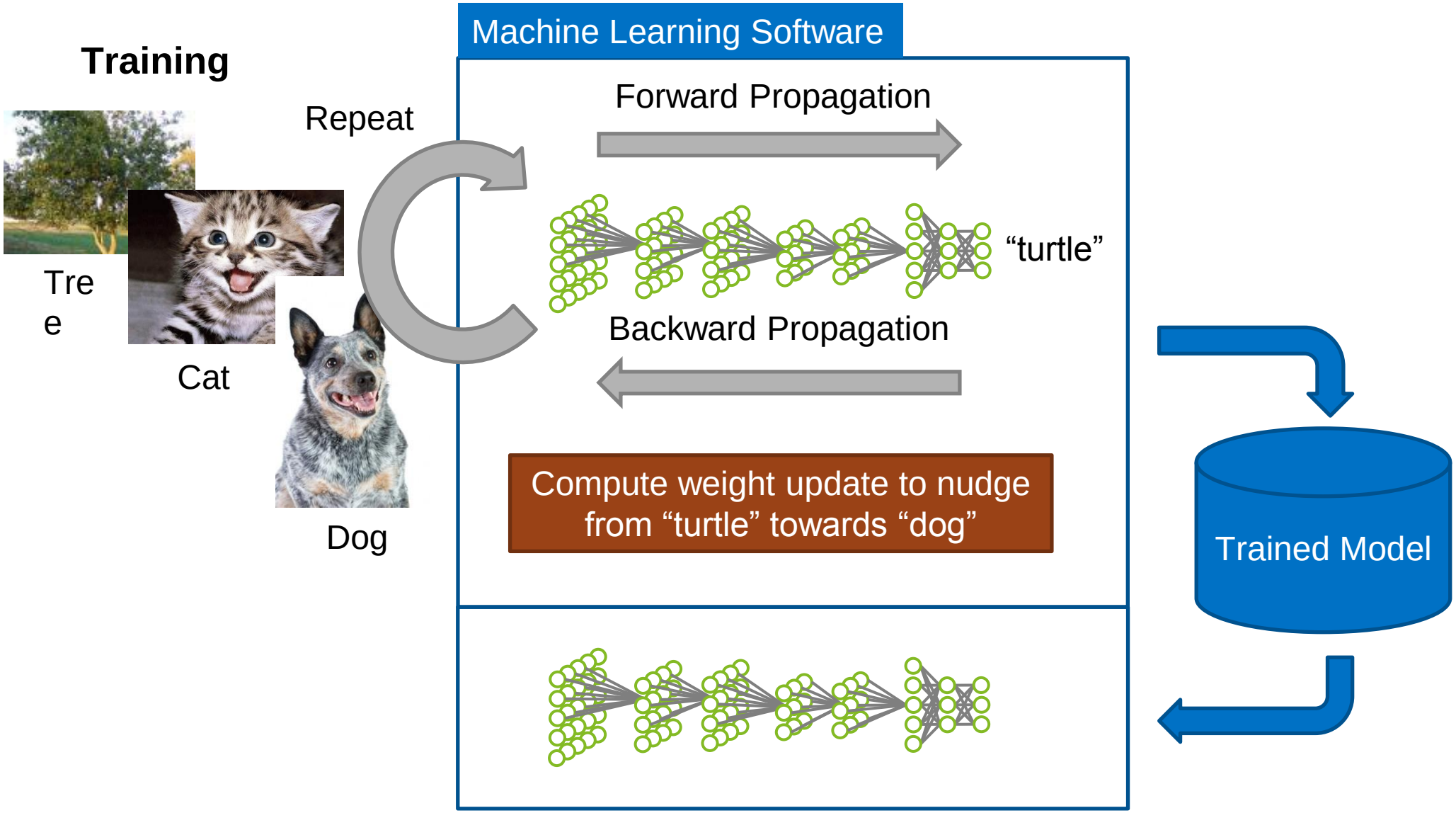


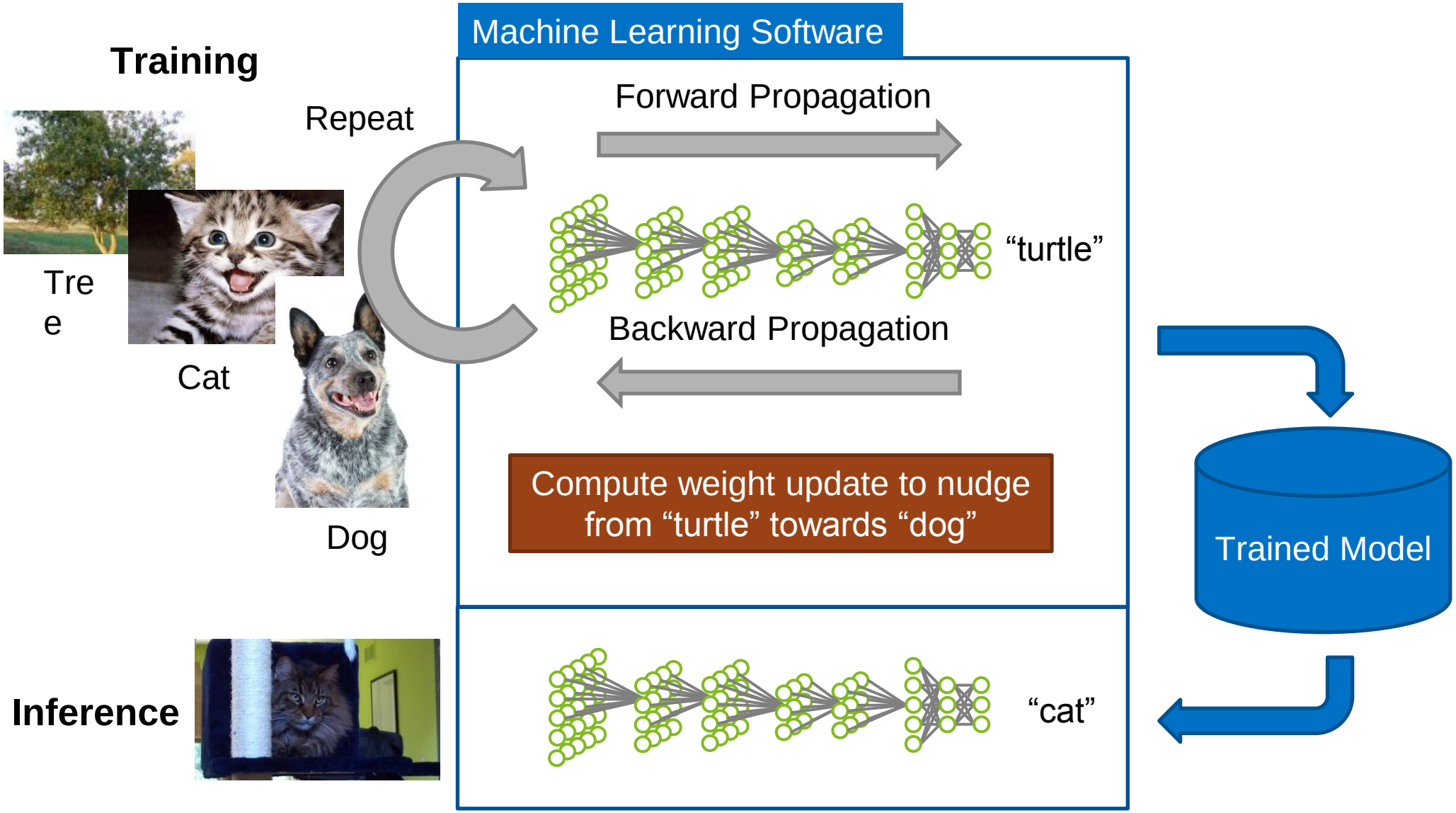
“turtle”

Backward Propagation

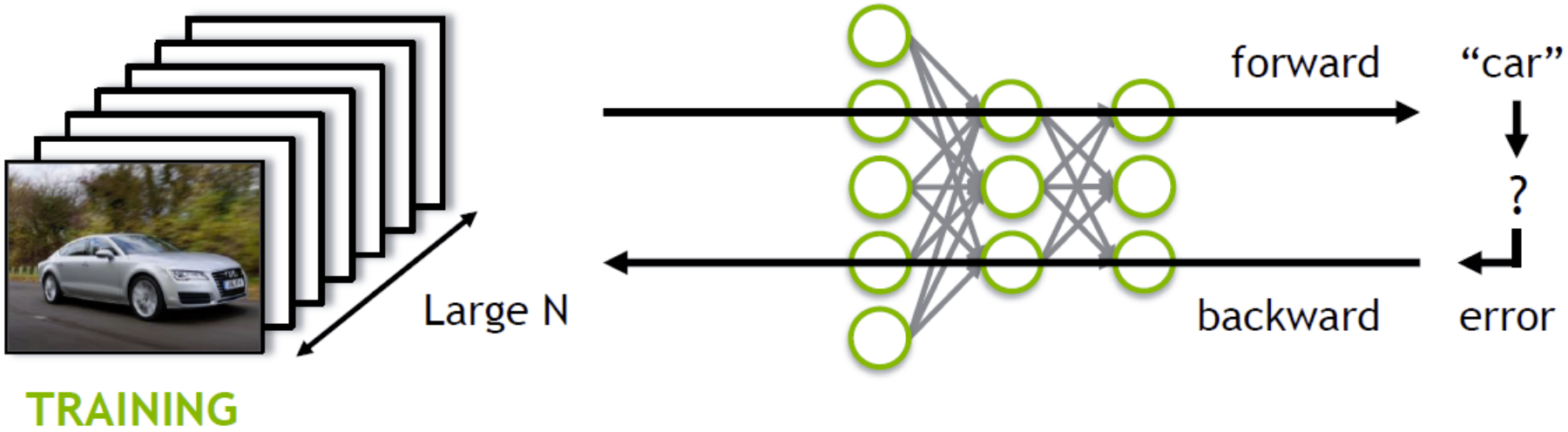
Compute weight update to nudge from “turtle” towards “dog”

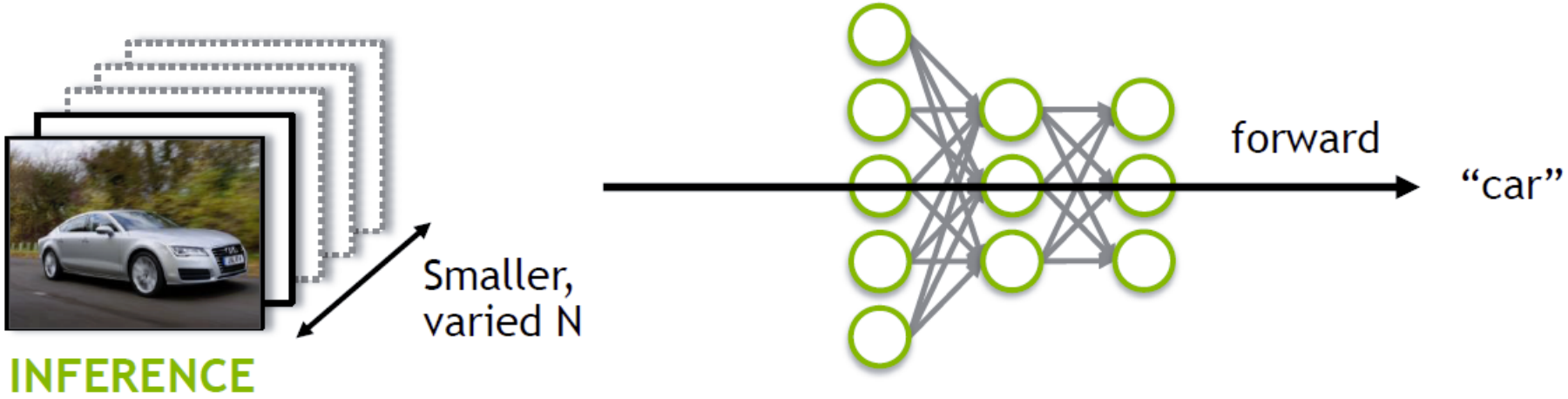



Training
Repeat
Tre
e
Cat
Dog
Machine Learning Software
Forward Propagation
"turtle"
Backward Propagation
Compute weight update to nudge from "turtle" towards "dog"
Trained Model


Training
Repeat
Tre
e
Cat
Dog
Machine Learning Software
Forward Propagation
"turtle"
Backward Propagation
Compute weight update to nudge from "turtle" towards "dog"
Trained Model
Inference
"cat"

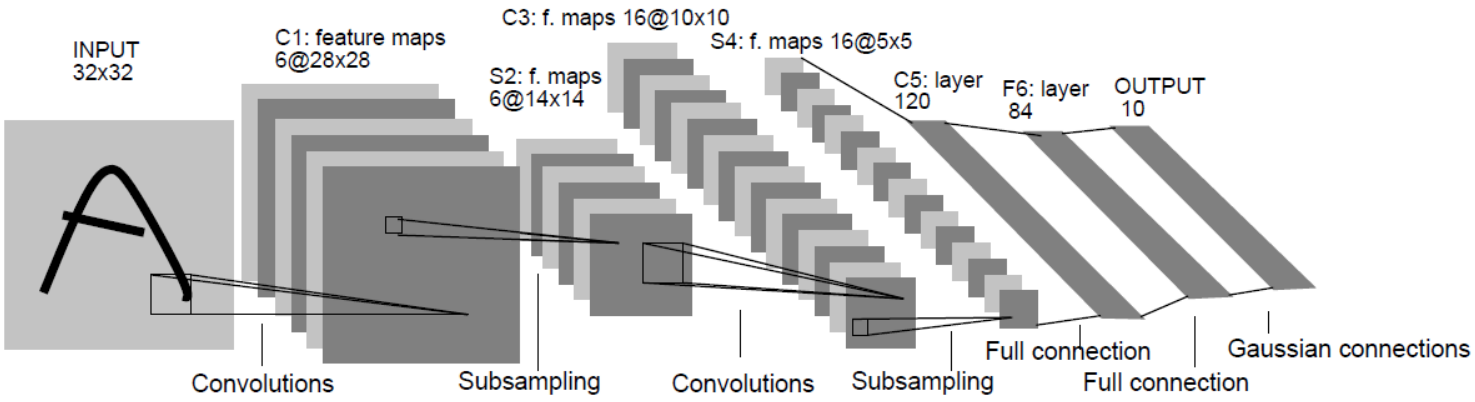
TRAINING VS INFERENCE

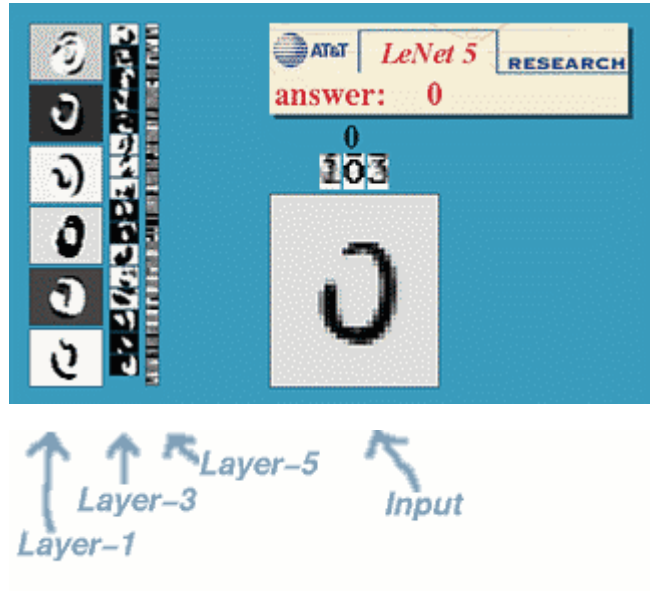




CONVOLUTIONAL NEURAL NETWORKS

Local receptive field + weight sharing



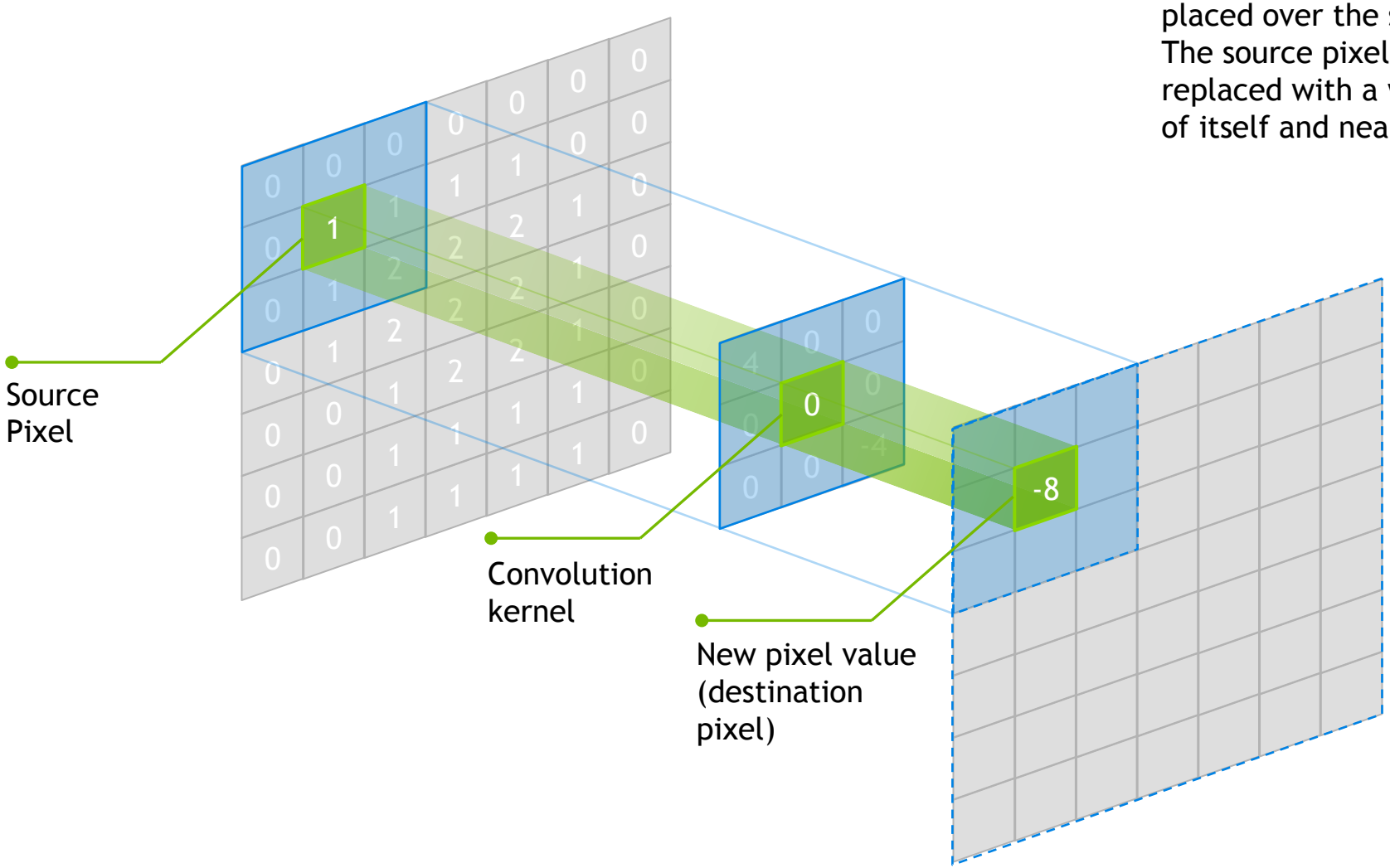


Introduced by Yann LeCun et al in 1998

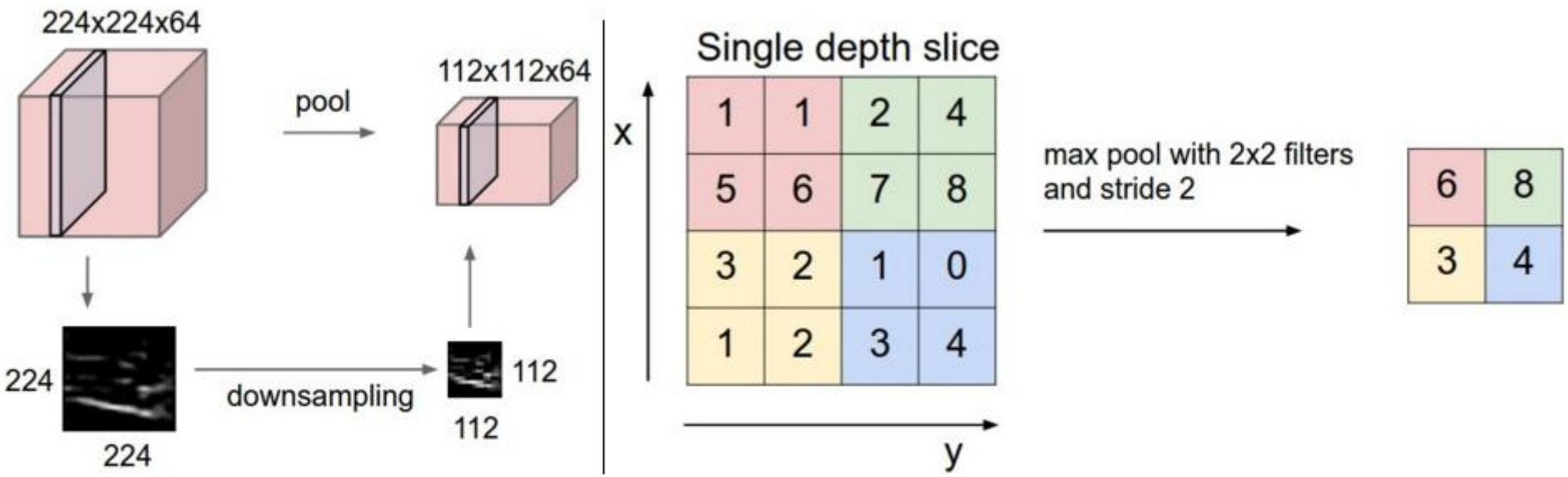
MNIST: 0.7% error rate

CONVOLUTION

Center element of the kernel is placed over the source pixel. The source pixel is then replaced with a weighted sum of itself and nearby pixels.



POOLING LAYER (SUBSAMPLING)

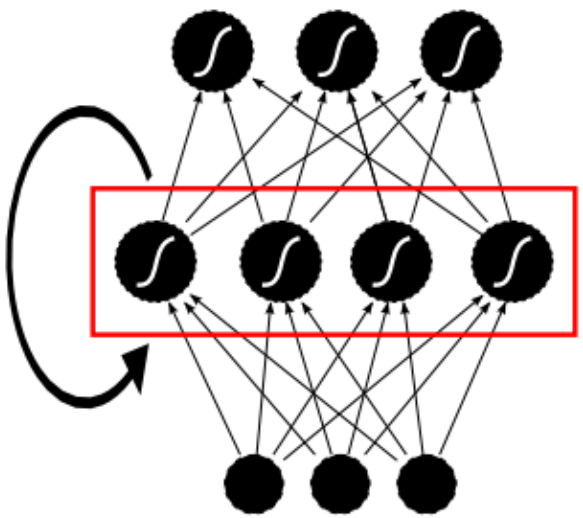


Pooling layer downsamples the volume spatially, independently in each depth slice of the input volume. **Left:** In this example, the input volume of size [224x224x64] is pooled with filter size 2, stride 2 into output volume of size [112x112x64]. Notice that the volume depth is preserved. **Right:** The most common downsampling operation is max, giving rise to **max pooling**, here shown with a stride of 2. That is, each max is taken over 4 numbers (little 2x2 square).

Source: @karpathy U.Stanford course CS231n http://cs231n.github.io/

RECURRENT NEURAL NETWORKS

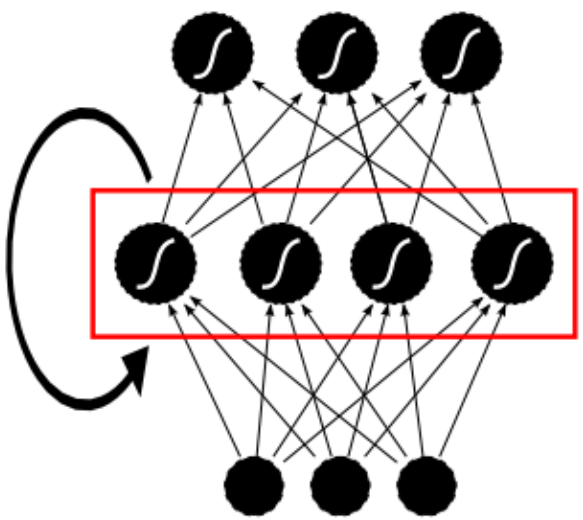
Time-Dependent or Sequence-Based Data

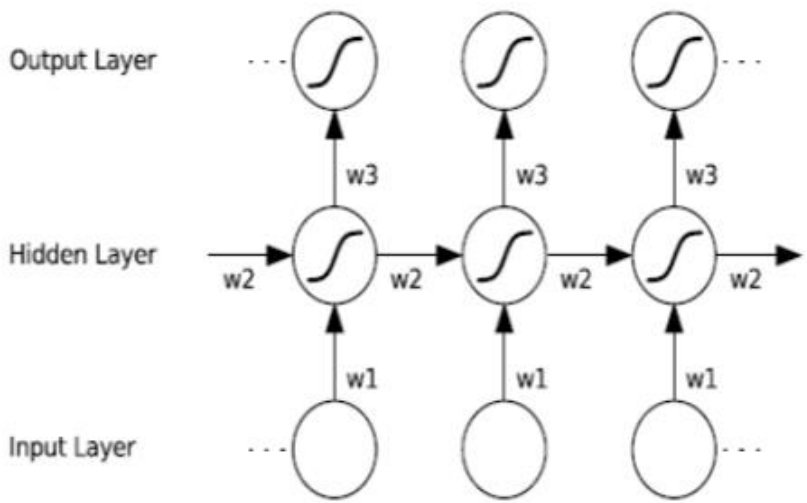


Source: courtesy of Alex Graves, TUM

RECURRENT NEURAL NETWORKS

Time-Dependent or Sequence-Based Data



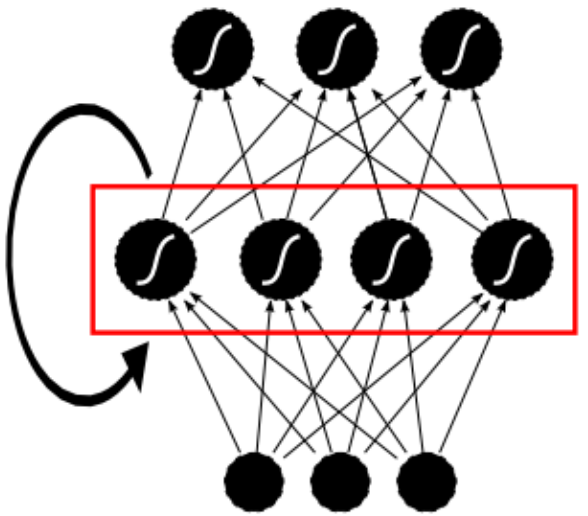


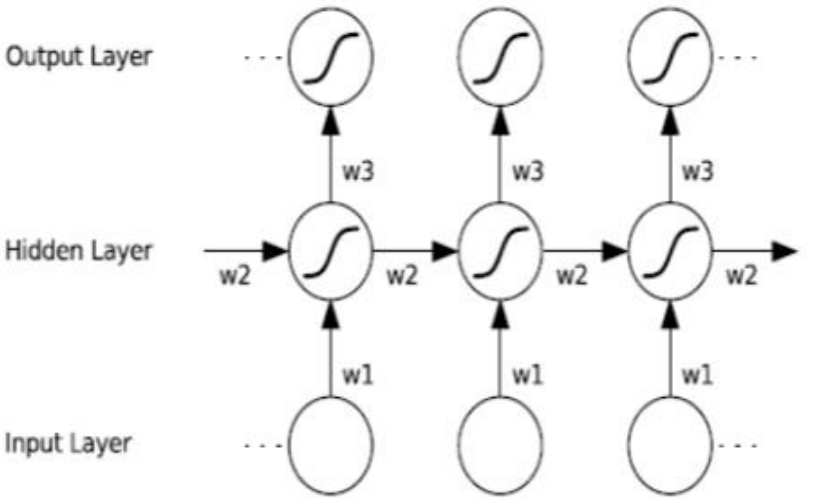
RNN Backpropagation: Rumelhart, Hinton, Williams, 1996

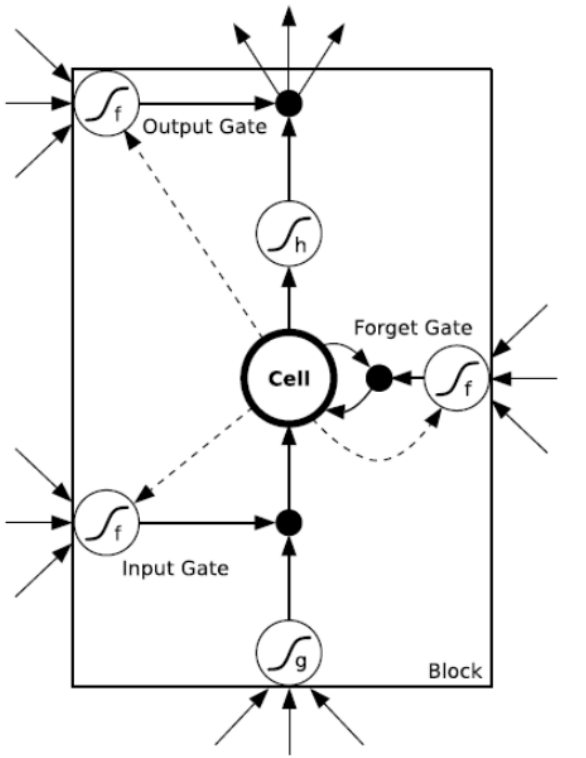
Source: courtesy of Alex Graves, TUM and Boris Ginzburg, NVIDIA

RECURRENT NEURAL NETWORKS

Time-Dependent or Sequence-Based Data







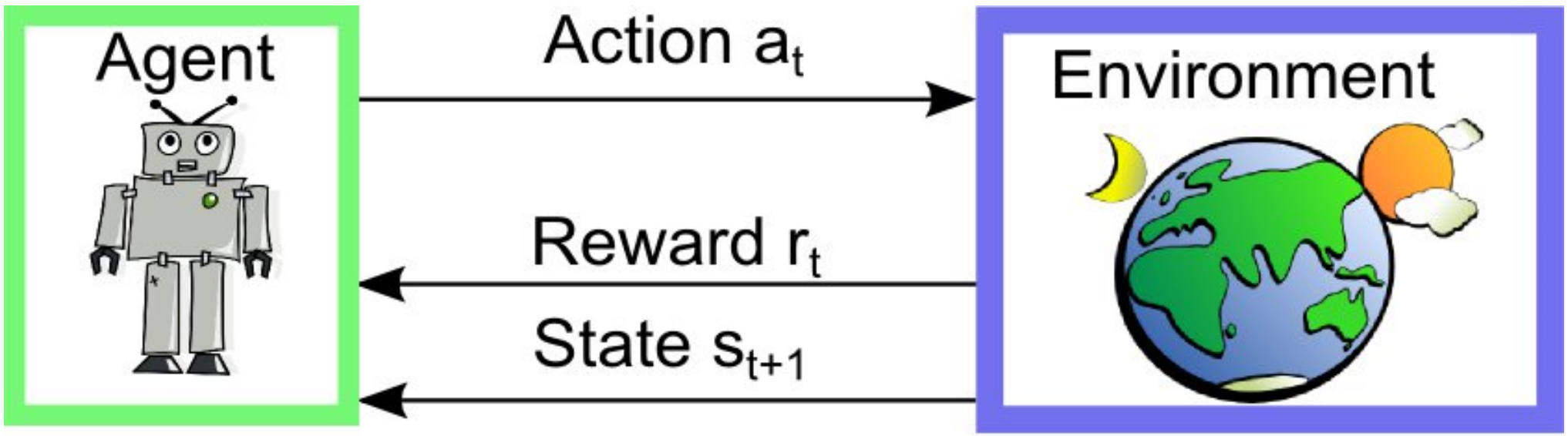
RNN Backpropagation: Rumelhart, Hinton, Williams, 1996

LSTM Cell: Hochreiter and Schmidhuber, 1997

Source: courtesy of Alex Graves, TUM and Boris Ginzburg, NVIDIA

REINFORCEMENT LEARNING

Prediction becomes action

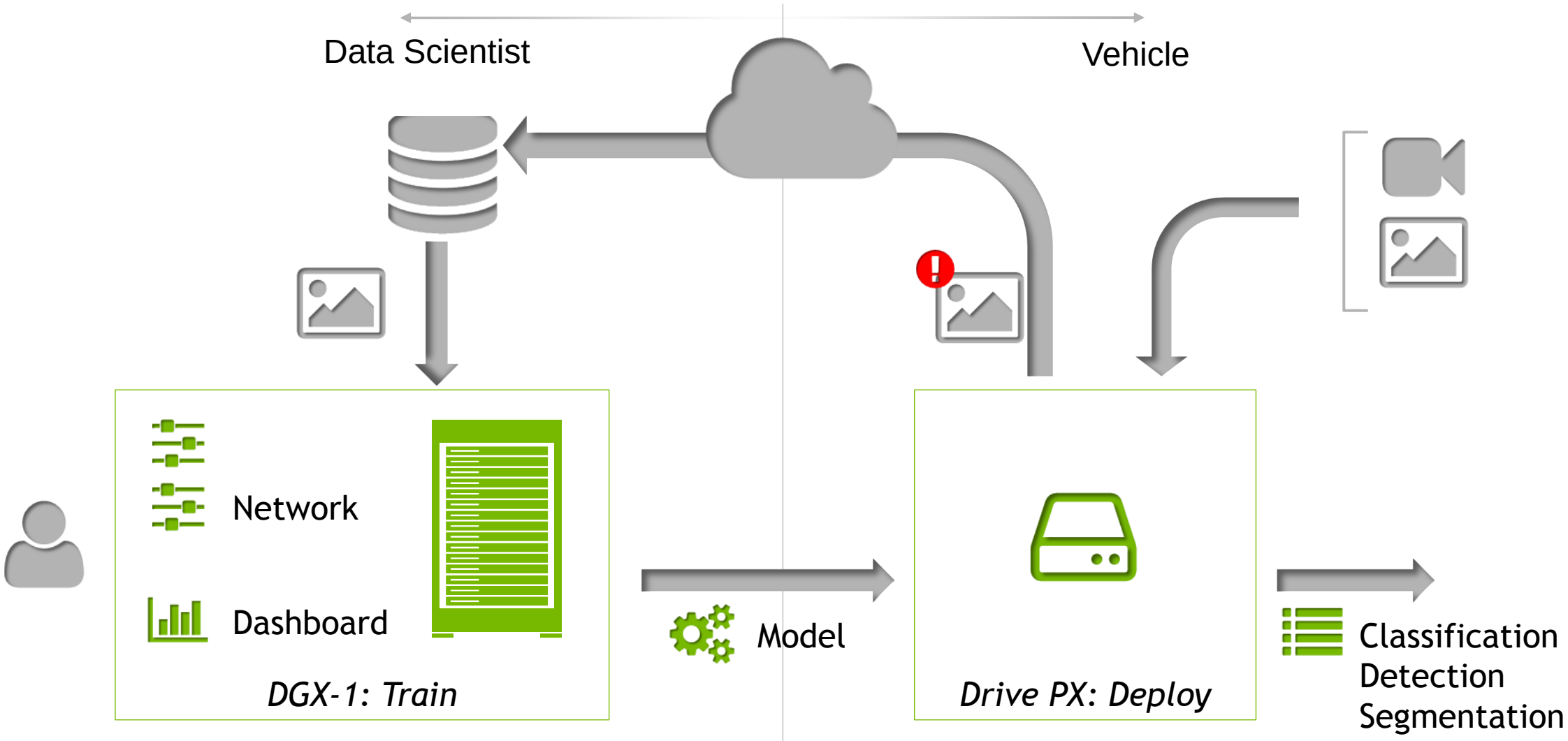


Reinforcement Learning Setup

http://www.ausy.tu-darmstadt.de/

GOOGLE DEEPMIND ALPHAGO CHALLENGE

ACTIVE LEARNING



GPU DEEP LEARNING

NEWS FEATURE

nature

THE LEARNING MACHINES

Using massive amounts of data to recognize photos and speech, deep-learning computers are taking a big step towards true artificial intelligence.

BY NICOLA JONES

Three years ago, researchers at the secretive Google X lab in Mountain View, California, extracted some 10 million still images from YouTube videos and fed them into Google Brain — a network of 1,000 computers programmed to soak up the world much as a human toddler does. After three days looking for recurring patterns, Google Brain decided, all on its own, that there were certain repeating categories it could identify: human faces, human bodies and ... cats[1].

Google Brain's discovery that the Internet is full of cat videos provoked a flurry of jokes from journalists. But it was also a landmark in the resurgence of deep learning: a three-decade-old technique in which massive amounts of data and processing power help computers to crack messy problems that humans solve almost intuitively, from recognizing faces to understanding language.

Deep learning itself is a revival of an even older idea for computing: neural networks. These systems, loosely inspired by the densely interconnected neurons of the brain, mimic human learning by changing the strength of simulated neural connections on the basis of experience. Google Brain, with about 1 million simulated neurons and 1 billion simulated connections, was ten times larger than any deep neural network before it. Project founder Andrew Ng, now director of the Artificial Intelligence Laboratory at Stanford University in California, has gone on to make deep-learning systems ten times larger again.

Such advances make for exciting times in

BRUCE ROLFF/ISTOCKPHOTO

146 | NATURE | VOL 505 | 9 JANUARY 2014

© 2014 Macmillan Publishers Limited. All rights reserved

TESLA REVOLUTIONIZES DEEP LEARNING

GOOGLE BRAIN APPLICATION

	BEFORE TESLA	AFTER TESLA
Cost	$5,000K	$200K
Servers	1,000 Servers	16 Tesla Servers
Energy	600 KW	4 KW
Performance	1x	6x

GPUS AND DEEP LEARNING

	NEURAL NETWORKS	GPUS
Inherently Parallel	✓	✓
Matrix Operations	✓	✓
FLOPS	✓	✓
Bandwidth	✓	✓

GPUs deliver --

- *same or better prediction accuracy*
- *faster results*
- *smaller footprint*
- *lower power*

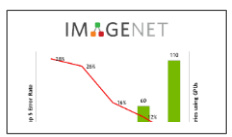
POWERING THE DEEP LEARNING ECOSYSTEM

NVIDIA SDK accelerates every major framework

COMPUTER VISION

OBJECT DETECTION

IMAGE CLASSIFICATION



SPEECH & AUDIO

VOICE RECOGNITION

LANGUAGE TRANSLATION

NATURAL LANGUAGE PROCESSING

RECOMMENDATION ENGINES

SENTIMENT ANALYSIS

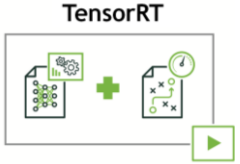
DEEP LEARNING FRAMEWORKS

Caffe
Chainer
DL4J
Deeplearning4j
Mocha.jl
julia
KERAS
MatConvNet
Microsoft CNTK
MINERVA
mxnet
OpenDeep
Purine
Pylearn2
TensorFlow
torch
theano

NVIDIA DEEP LEARNING SDK

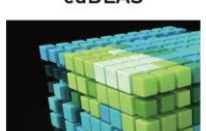
cuDNN

TensorRT



DeepStream SDK

cuBLAS



cuSPARSE

NCCL

developer.nvidia.com/deep-learning-software

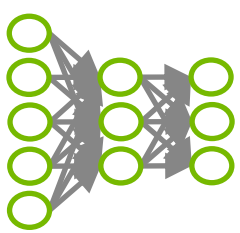
EVERY DEEP LEARNING FRAMEWORK IS GPU-ACCELERATED

TORCH	NYU, facebook
CAFFE	Berkeley University of California
THEANO	Université de Montréal
MATCONVNET	University of Oxford
MOCHA.JL	Massachusetts Institute of Technology
PURINE	NUS National University of Singapore, Shanghai Jiao Tong University
MINERVA	NYU, Peking University, Microsoft
MXNET*	University of Washington, Carnegie Mellon University
BIG SUR	facebook
TENSORFLOW	Google
WATSON	IBM
CNTK	Microsoft

cuDNN

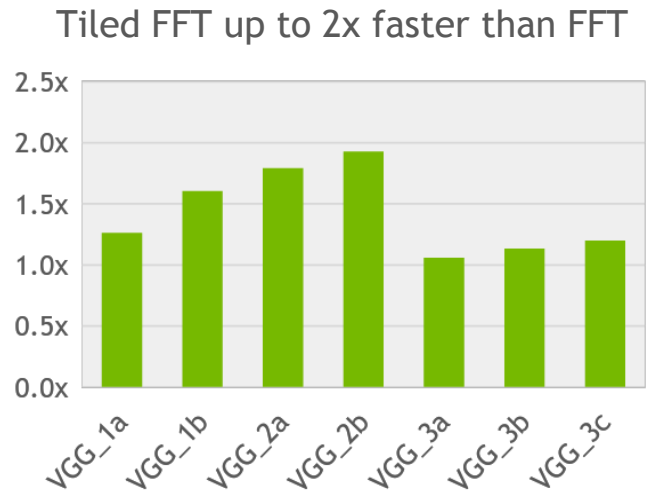
Deep Learning Primitives

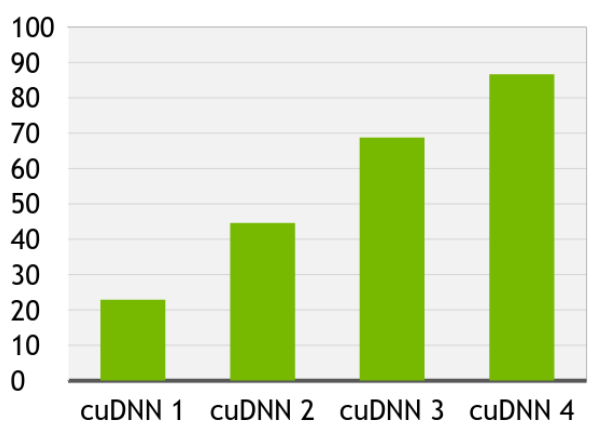
IGNITING ARTIFICIAL INTELLIGENCE



developer.nvidia.com/cudnn

- GPU-accelerated Deep Learning subroutines
- High performance neural network training
- Accelerates Major Deep Learning frameworks: Caffe, Theano, Torch
- Up to 3.5x faster AlexNet training in Caffe than baseline GPU





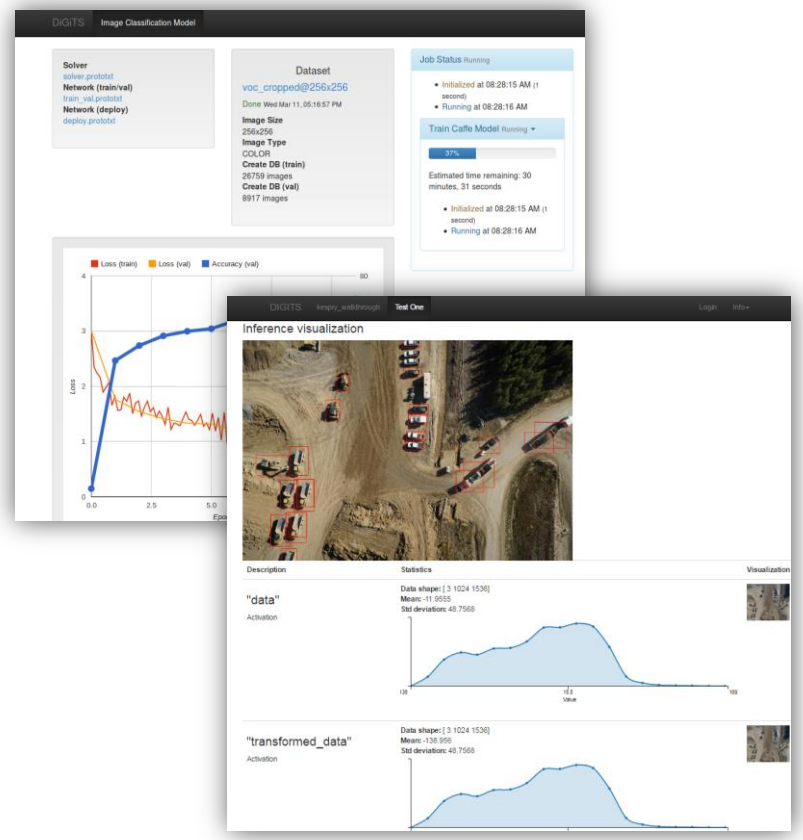
TensorFlow
torch
Caffe
Microsoft CNTK
Purine
KERAS
theano
Chainer
MINERVA
DL4J Deeplearning4j
MatConvNet

NVIDIA DIGITS

Interactive Deep Learning GPU Training System

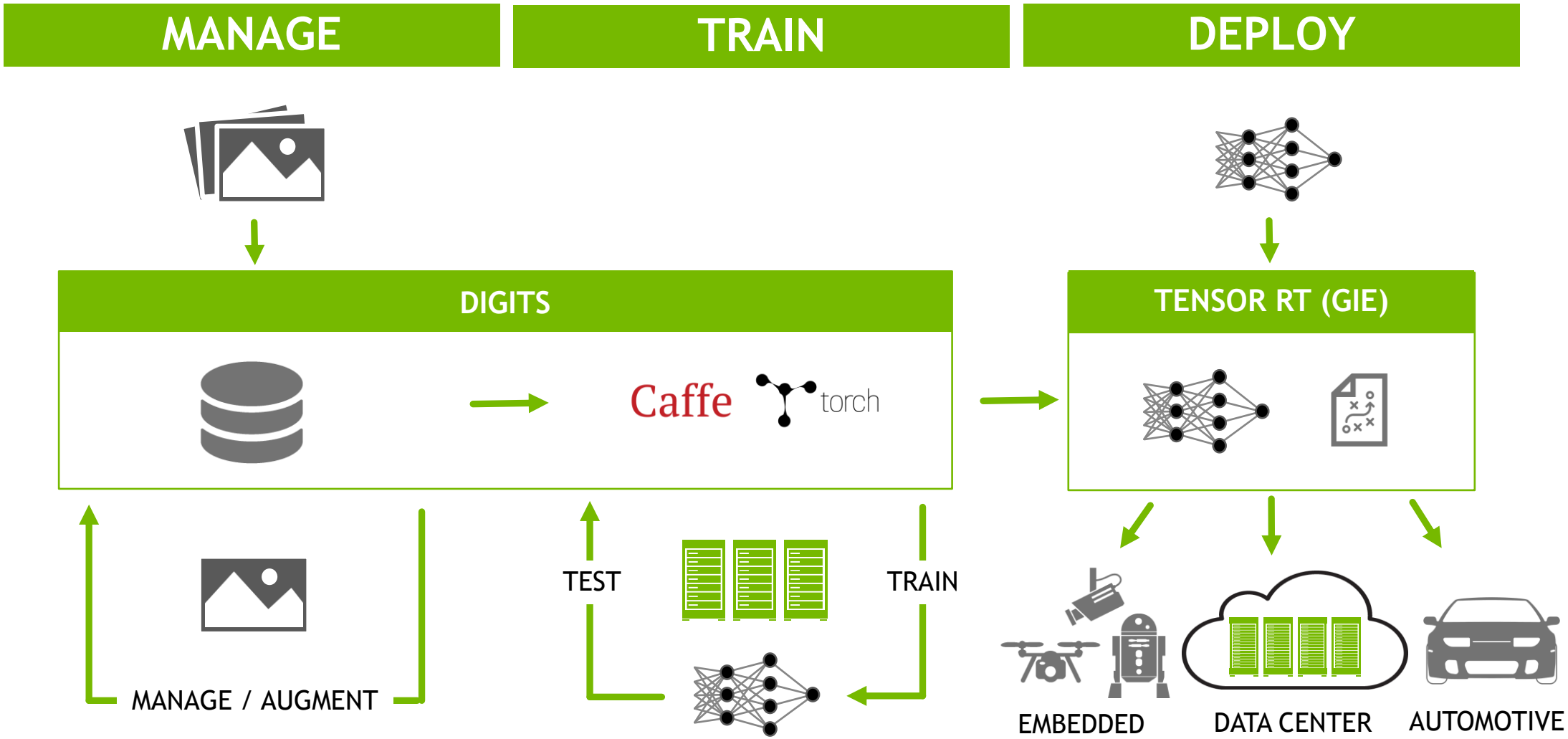
Interactive deep neural network development environment for image classification and object detection

- Schedule, monitor, and manage neural network training jobs
- Analyze accuracy and loss in real time
- Track datasets, results, and trained neural networks
- Scale training jobs across multiple GPUs automatically

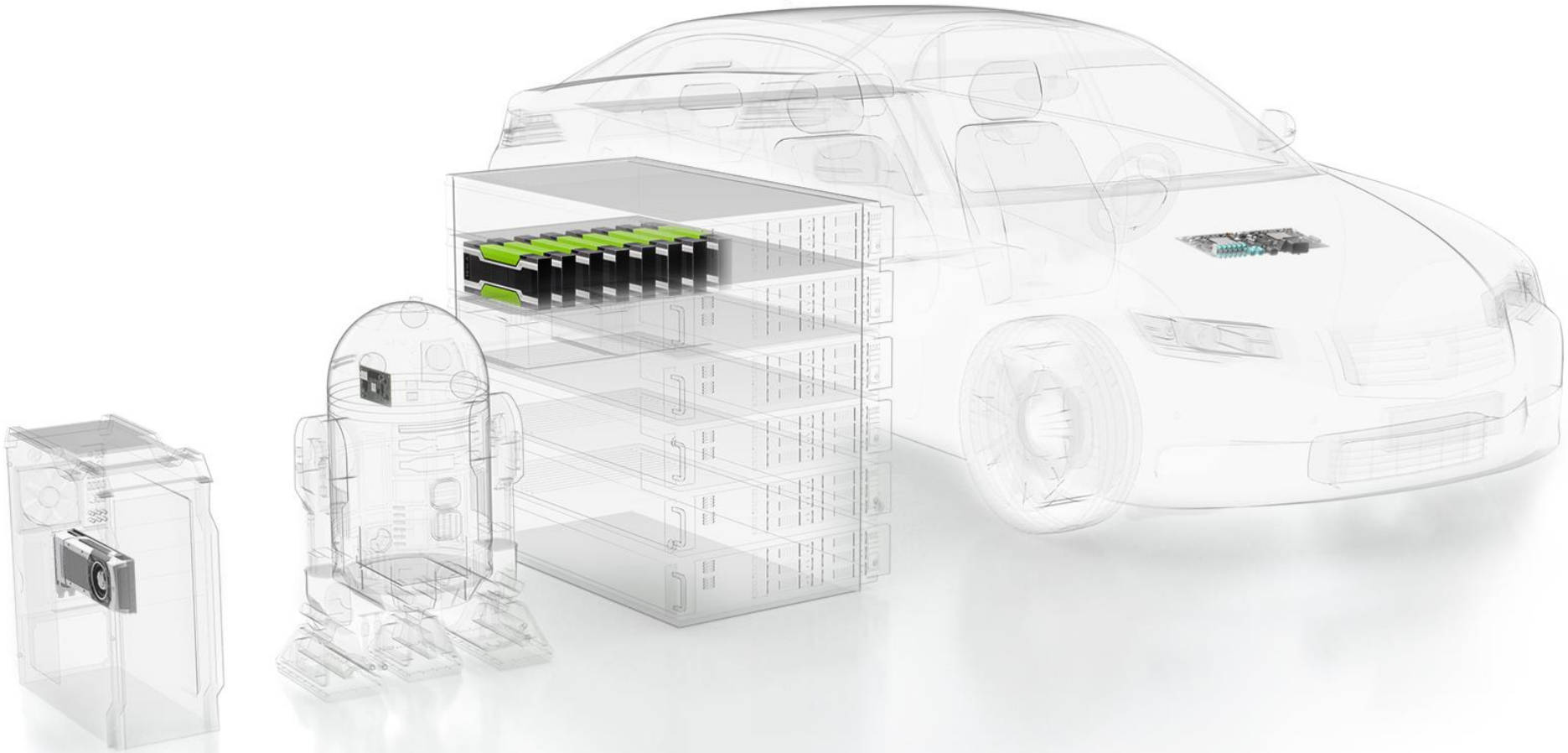


developer.nvidia.com/digits

A COMPLETE DEEP LEARNING PLATFORM



DEEP LEARNING EVERYWHERE



THANK YOU!

cnardone@nvidia.com

developer.nvidia.com/deep-learning