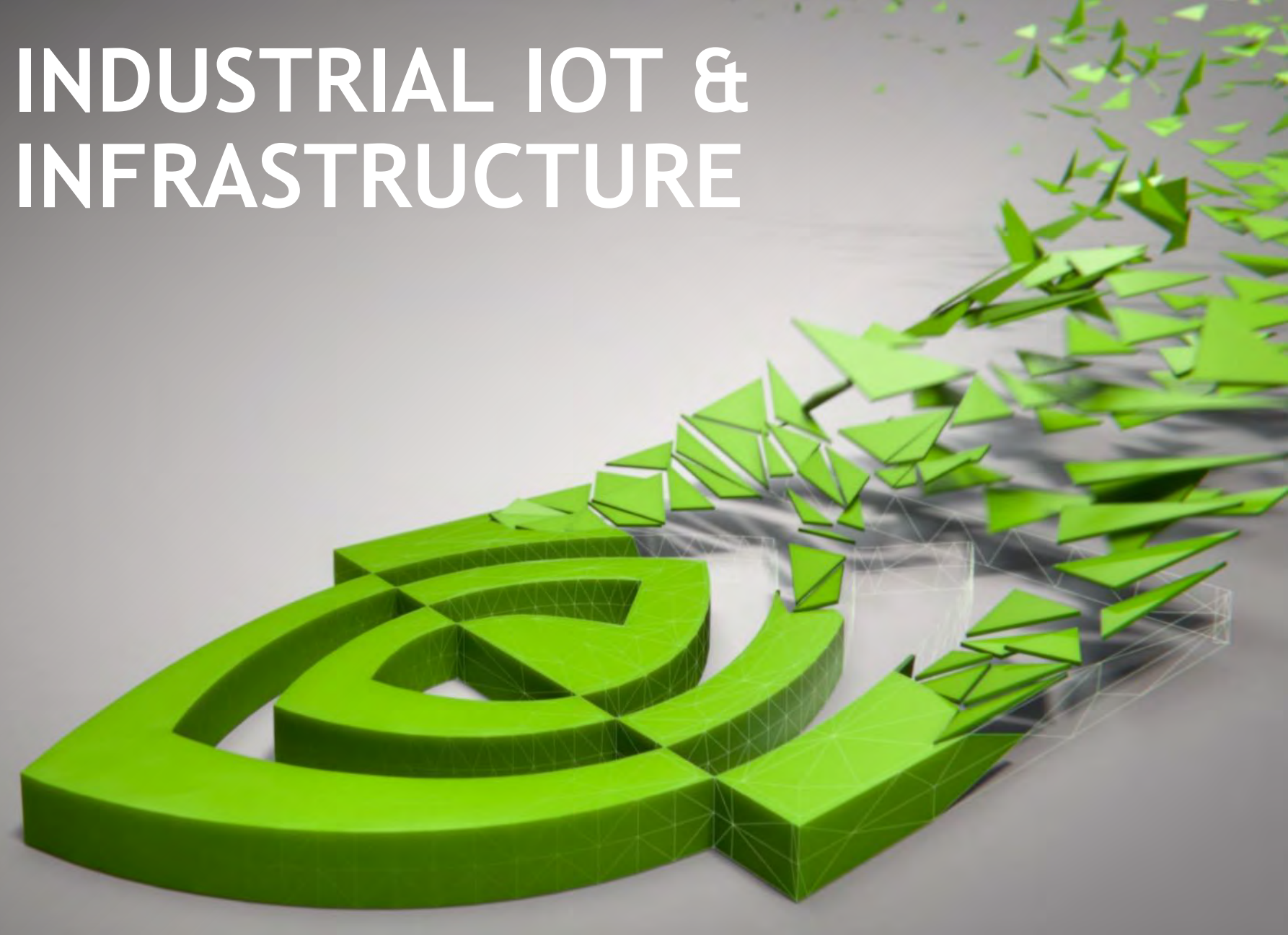
AI FOR INDUSTRIAL IOT & SMART INFRASTRUCTURE

APRIL 2017

Piyush Modi, PhD

NVIDIA

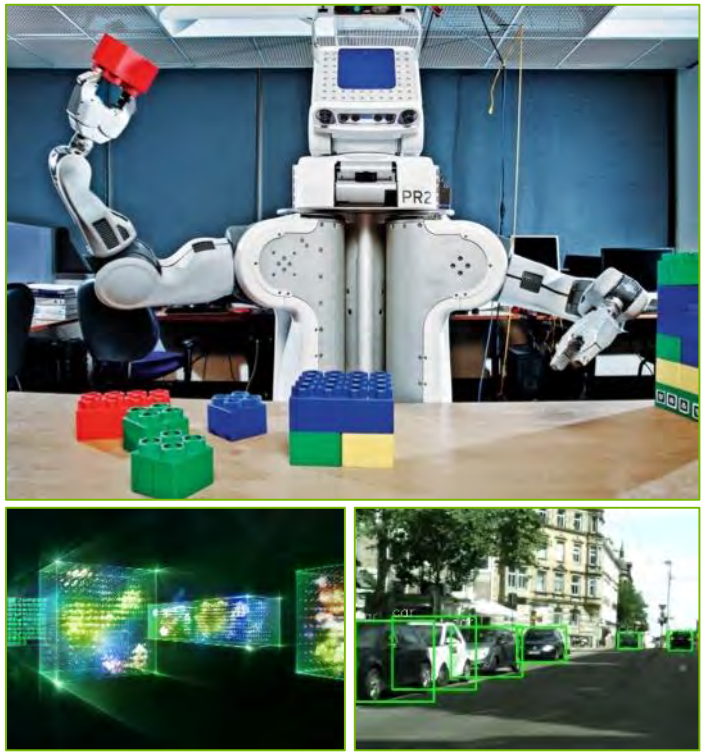


NVIDIA “THE AI COMPUTING COMPANY”

Pioneered GPU Computing | Founded 1993 | $6.9B | 10,300 Employees

Computer Graphics

GPU Computing



Artificial Intelligence

AGENDA

Artificial Intelligence (AI) & Deep Learning (DL)

Industrial IOT (IIOT) & Smart Infrastructure

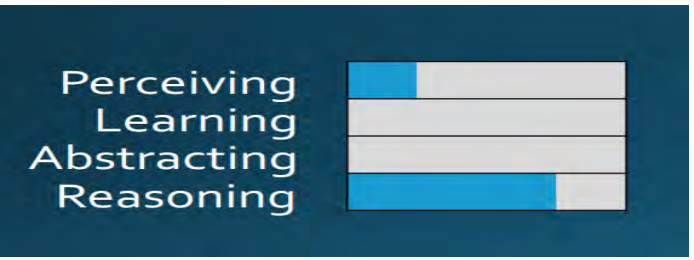
AI for IIOT & AI Cities – Use Cases

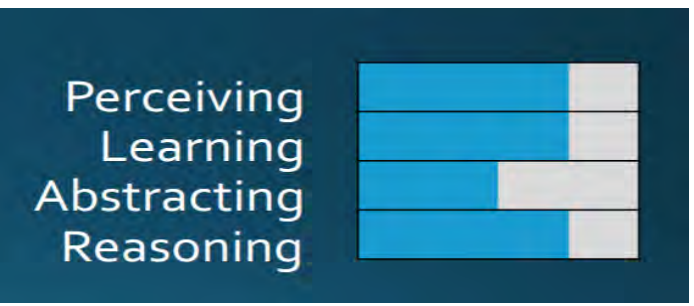
Q&A

AI & DEEP LEARNING

THREE WAVES OF AI: DARPA PERSPECTIVE

Artificial intelligence is a programmed ability to process information





Hand Crafted Knowledge

Emphasis on Reasoning and no learning (e.g. Expert Systems)

Statistical Learning

Self learning features, representational and transferable learning

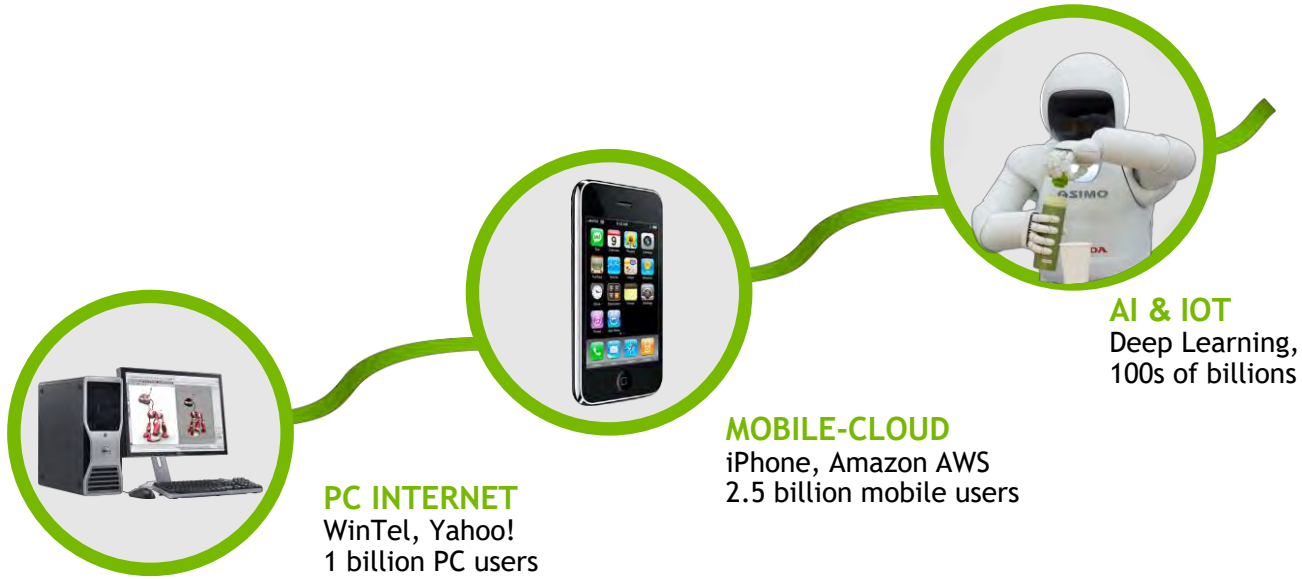
Contextual Adaptation

Systems construct contextual explanatory models for classes of real world phenomena

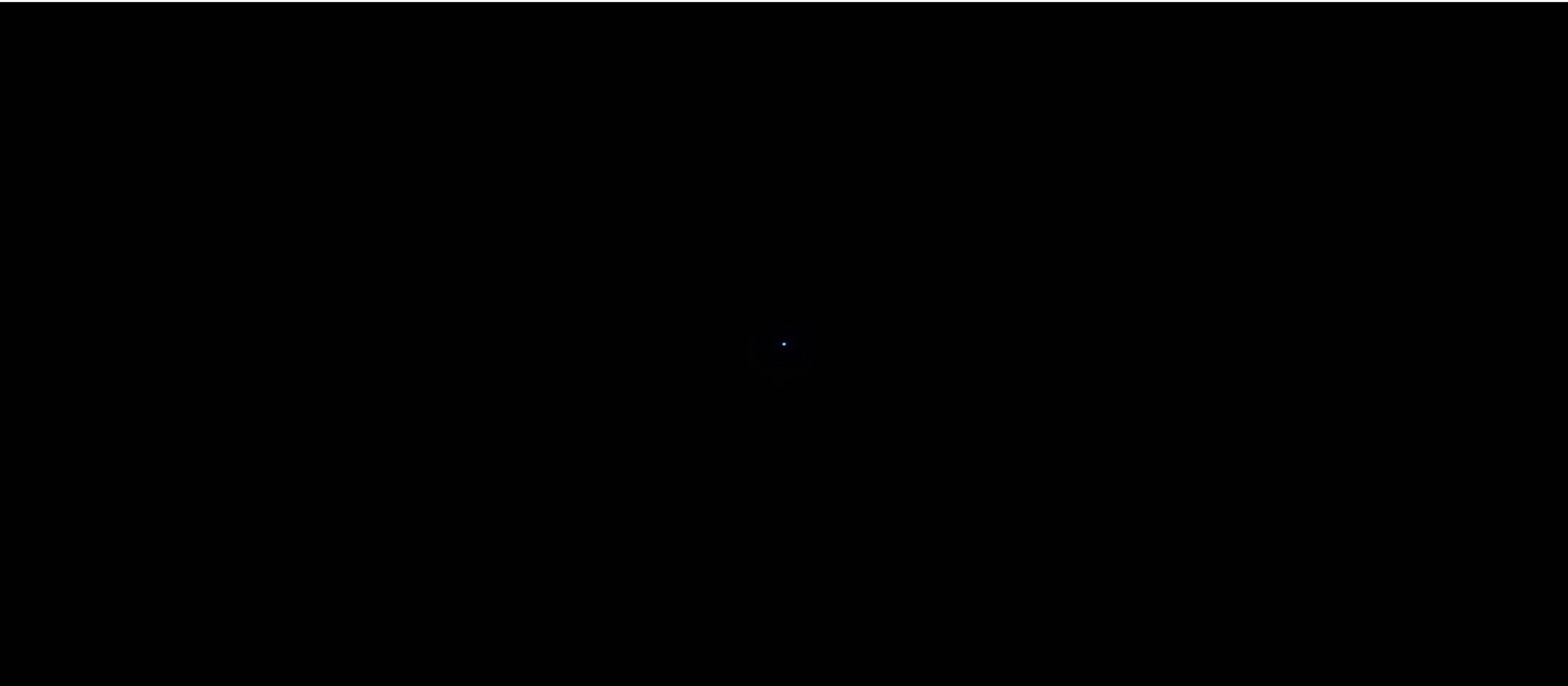
Ability to explain Why? and Why Not?

A NEW ERA OF COMPUTING

> “ It’s clear we’re moving from a mobile first to an AI-first world ”
>
> Sundar Pichai, Google CEO

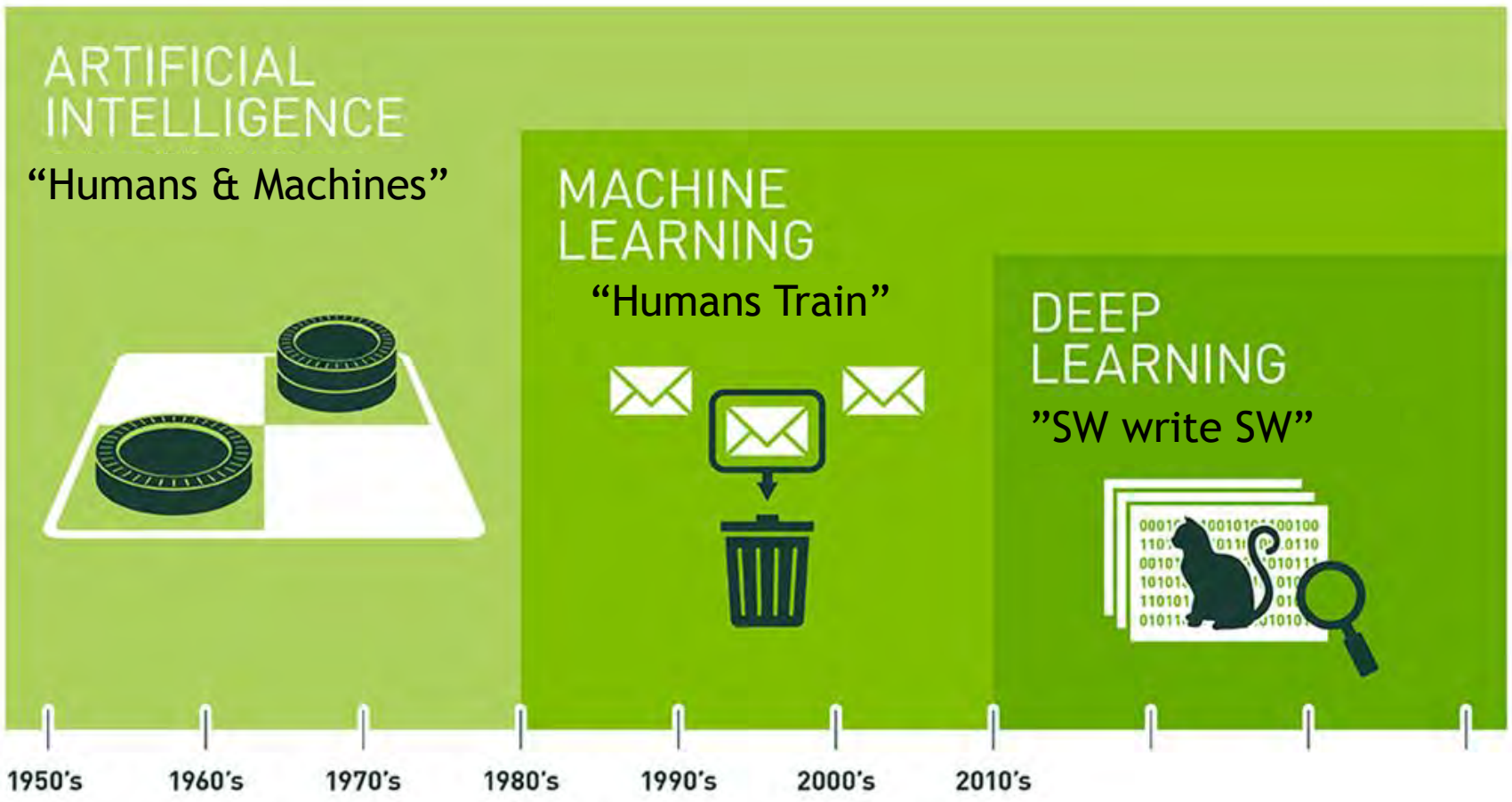


DEEP LEARNING / AI EVERYWHERE



AI: MACHINE LEARNING & DEEP LEARNING

(ML) (DL)



Source : Michael Copeland, Journalist for WIRED, Fortune, and Business 2.0

WHY *"NOW?"*

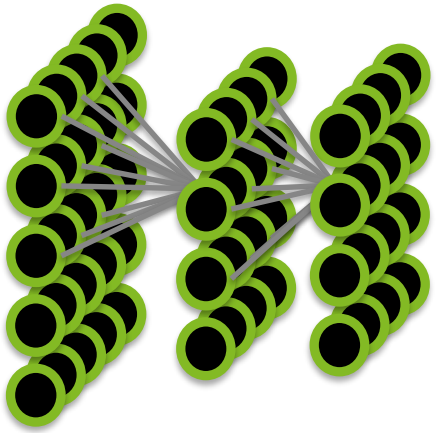
Big Data

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



WHY *"NOW?"*

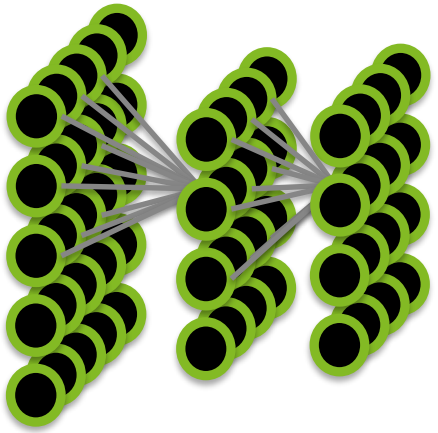
Big Data

facebook | 350 million images uploaded per day

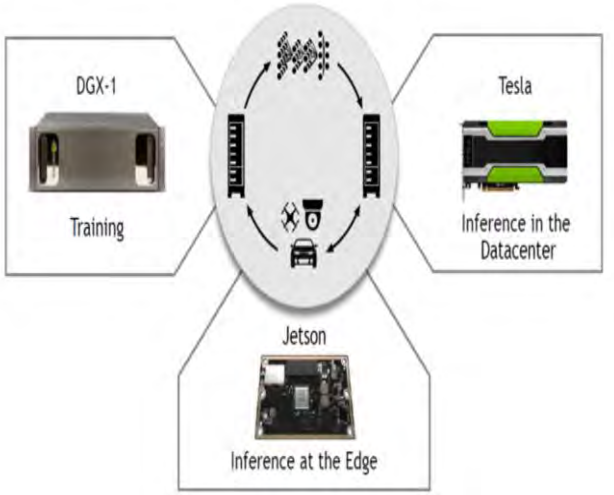
Walmart | 2.5 Petabytes of customer data hourly

YouTube | 300 hours of video uploaded every minute

Better Algorithms

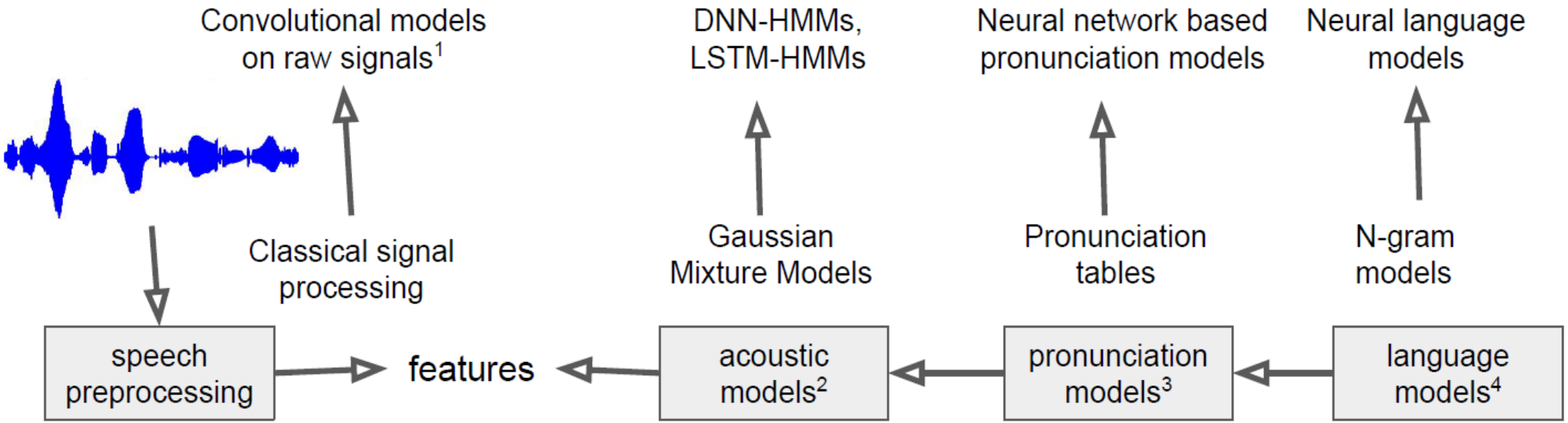


GPU Acceleration



Speech Recognition -- the neural network invasion

- Each of the components seems to be better off with a neural network



1. Jaitly, Navdeep, and Geoffrey Hinton. "Learning a better representation of speech soundwaves using restricted boltzmann machines." *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011.
2. Hinton, Geoffrey, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." *IEEE Signal Processing Magazine* 29.6 (2012): 82-97.
3. Rao, Kanishka, et al. "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks." *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015.
4. Mikolov, Tomas, et al. "Recurrent neural network based language model." *Interspeech*. Vol. 2. 2010.

DEEP LEARNING FOR SPEECH RECOGNITION

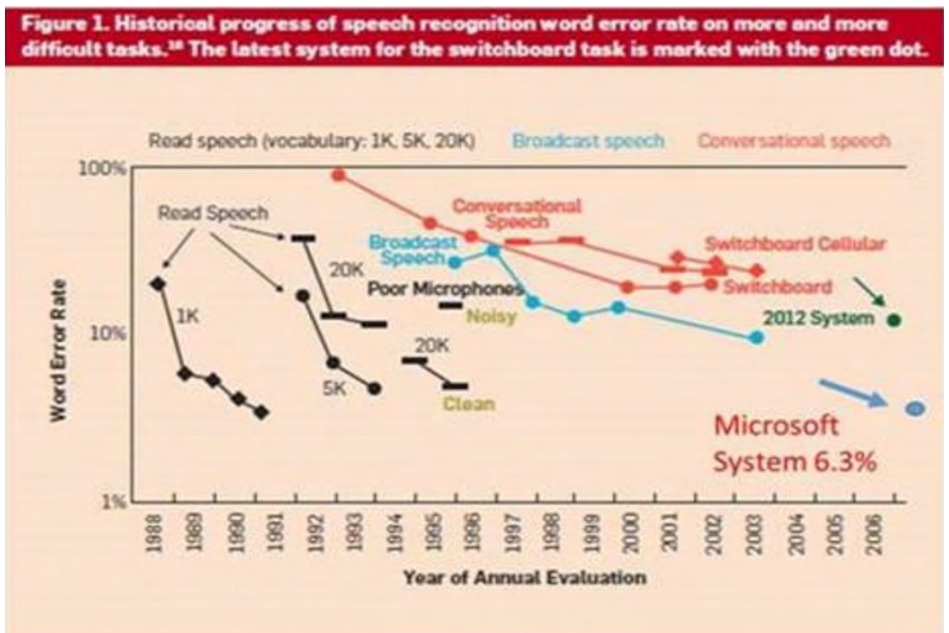
Improved WER and Better ability to Learn with more Data

A comparison of the Percentage Word Error Rates using DNN-HMMs and GMM-HMMs on five different large vocabulary tasks.

task	hours of training data	DNN-HMM	GMM-HMM with same data	GMM-HMM with more data
Switchboard (test set 1)	309	18.5	27.4	18.6 (2000 hrs)
Switchboard (test set 2)	309	16.1	23.6	17.1 (2000 hrs)
English Broadcast News	50	17.5	18.8	
Bing Voice Search (Sentence error rates)	24	30.4	36.2	
Google Voice Input	5,870	12.3		16.0 (>>5,870hrs)
Youtube	1,400	47.6	52.3	

Deep Neural Networks for Acoustic Modeling in Speech Recognition , Google 2012

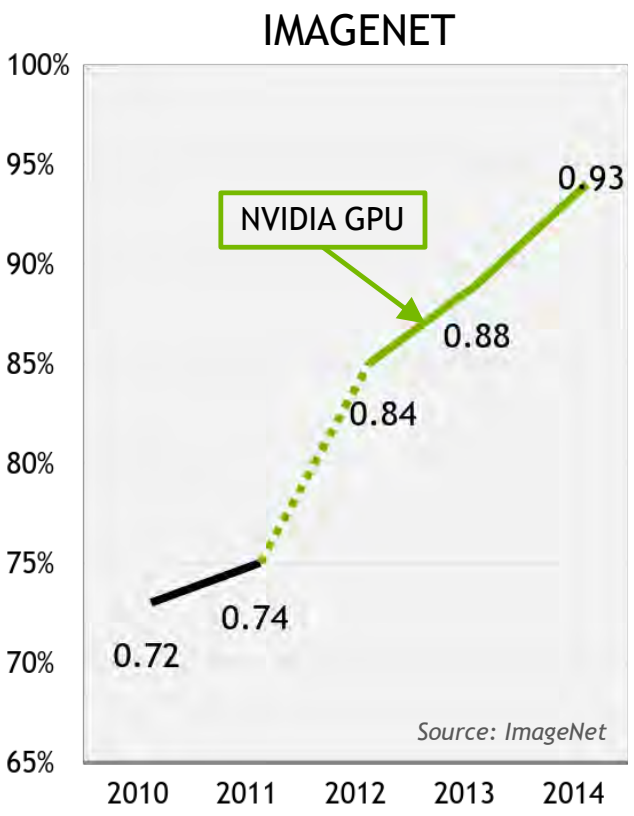
Microsoft 2016



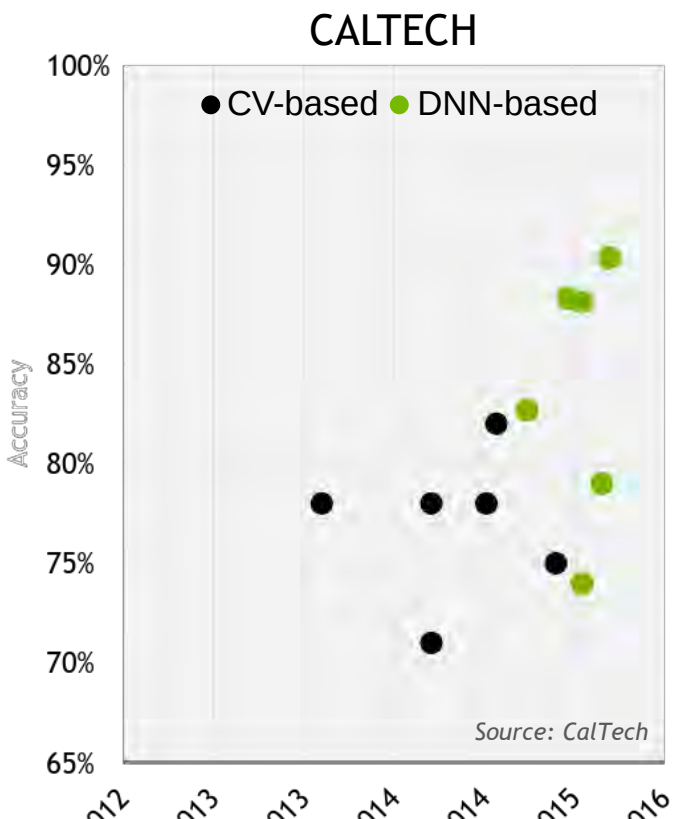
2017 Switch Board Task WER: IBM 5.5 ; % MSFT 5.6 % ; Human as per IBM 5.1%

DEEP LEARNING FOR VISUAL ANALYTICS

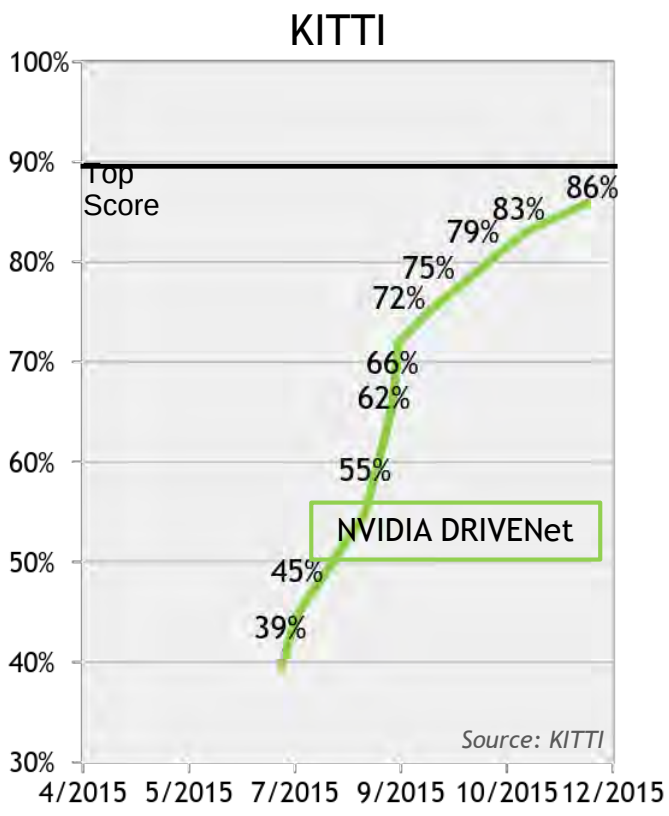
Image Recognition



Pedestrian Detection



Object Detection



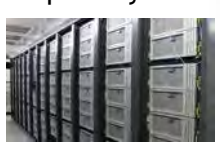
THE MOST COMPLEX CHALLENGES

2006

CUDA Launched

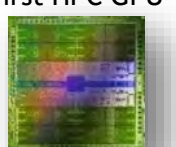
2008

World's First GPU Top500 System



2010

Fermi: World's First HPC GPU



Oak Ridge Deploys World's Fastest Supercomputer w/ GPUs

2012

Discovered How H1N1 Mutates to Resist Drugs

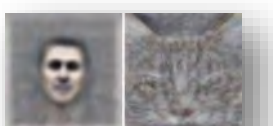
AlexNet beats expert code by huge margin using GPUs

IMAGENET

2014

World's First Atomic Model of HIV Capsid

Stanford Builds AI Machine using GPUs



World's First 3-D Mapping of Human Genome

GPU-Trained AI Machine Beats World Champion in Go

Google Outperforms Humans in ImageNet

2016

POWERING DEEP LEARNING

Every major DL framework leverages NVIDIA SDKs

COMPUTER VISION

OBJECT DETECTION

IMAGE CLASSIFICATION

SPEECH & AUDIO

VOICE RECOGNITION

LANGUAGE TRANSLATION

NATURAL LANGUAGE PROCESSING

RECOMMENDATION ENGINES

SENTIMENT ANALYSIS

TIME SERIES

ANOMALY DETECTION & CLASSIFICATION

REINFORCEMENT LEARNING

DEEP LEARNING FRAMEWORKS

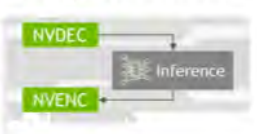
Caffe | Microsoft CNTK | mxnet | TensorFlow | theano | torch

NVIDIA DEEP LEARNING SDK

cuDNN

TensorRT

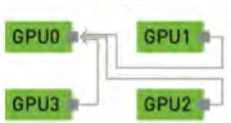
DeepStream SDK



cuBLAS

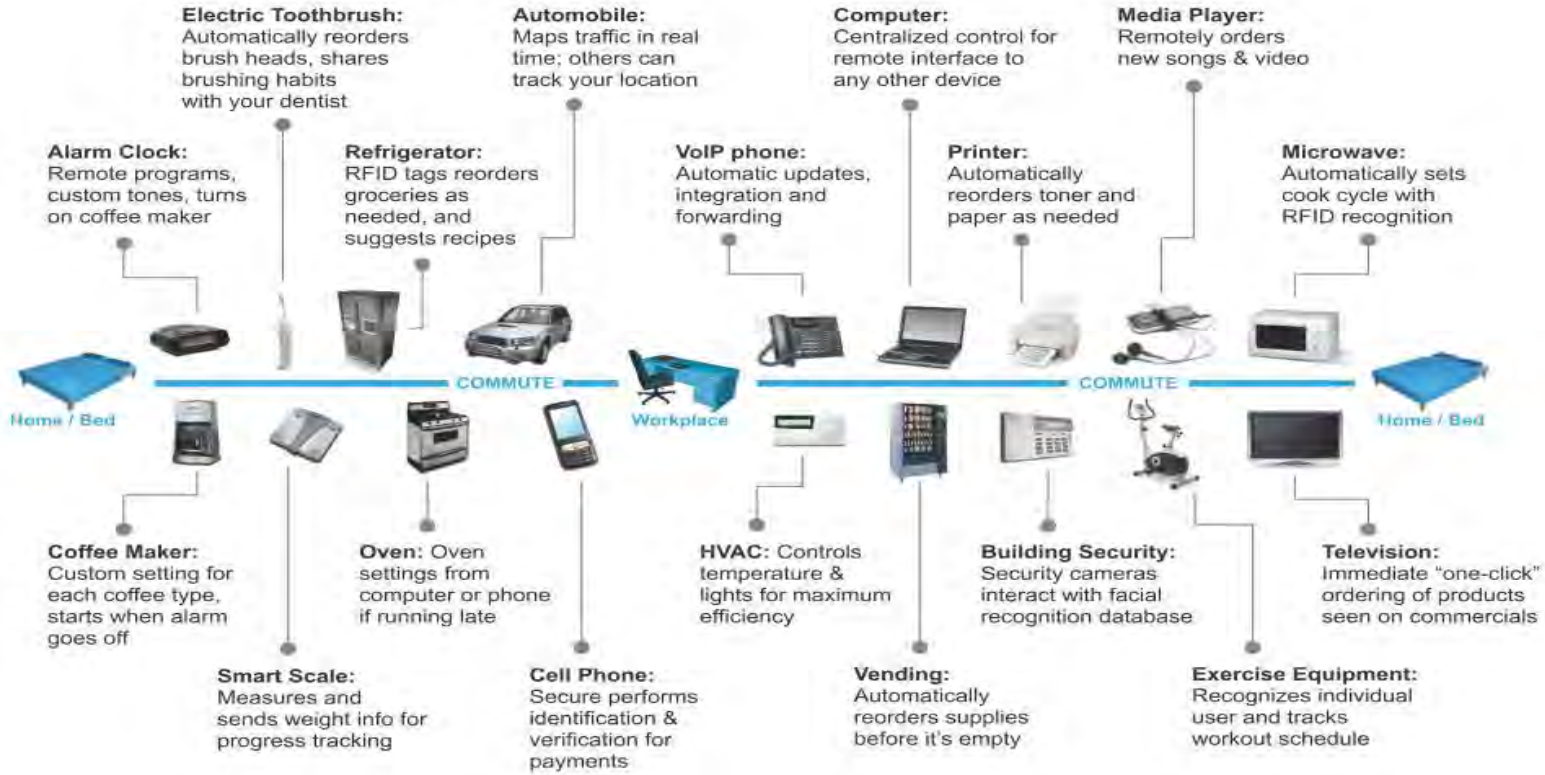
cuSPARSE

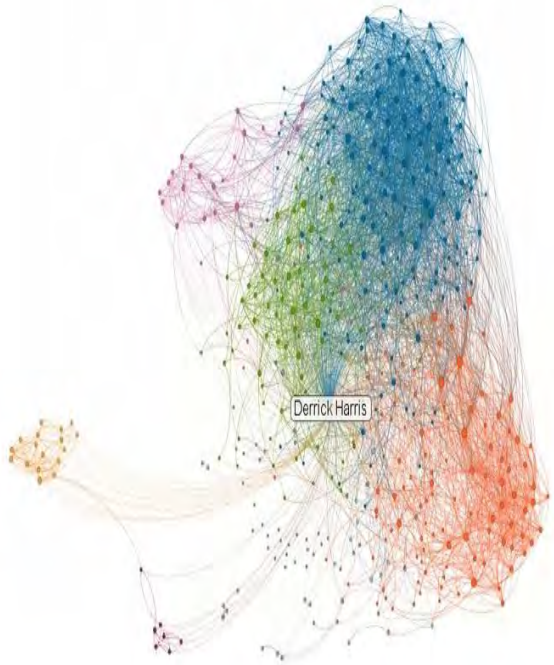
NCCL



THE EXPANDING UNIVERSE OF MODERN AI

INDUSTRIAL IOT AND SMART INFRASTRUCTURE (AI CITIES)





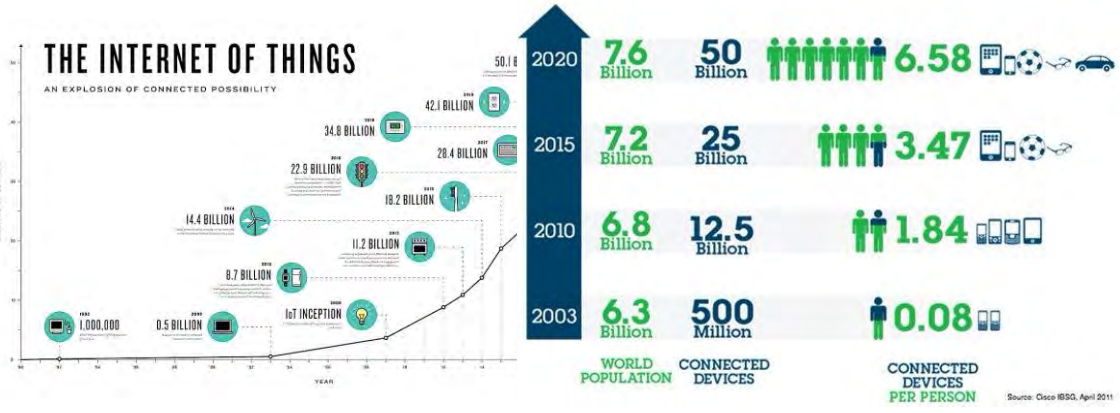
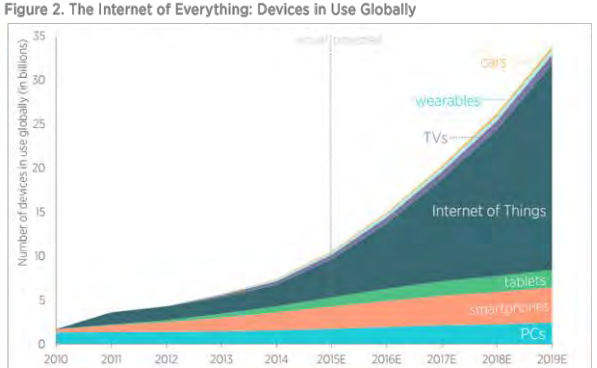
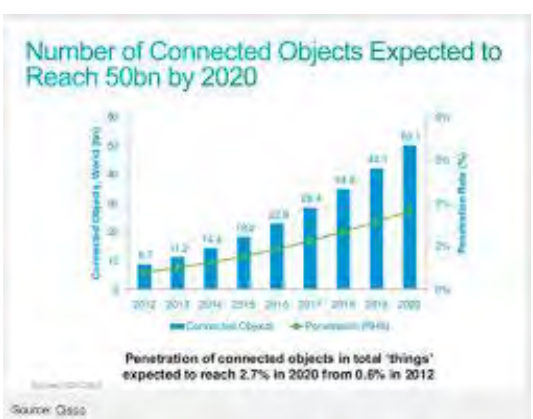


Figure 2. The Internet of Everything: Devices in Use Globally



Source: John Greenough, "The Internet of Everything 2015," *Business Insider Intelligence*. Produced by Adam Thierer and Andrea Castillo, Mercatus Center at George Mason University, 2015.

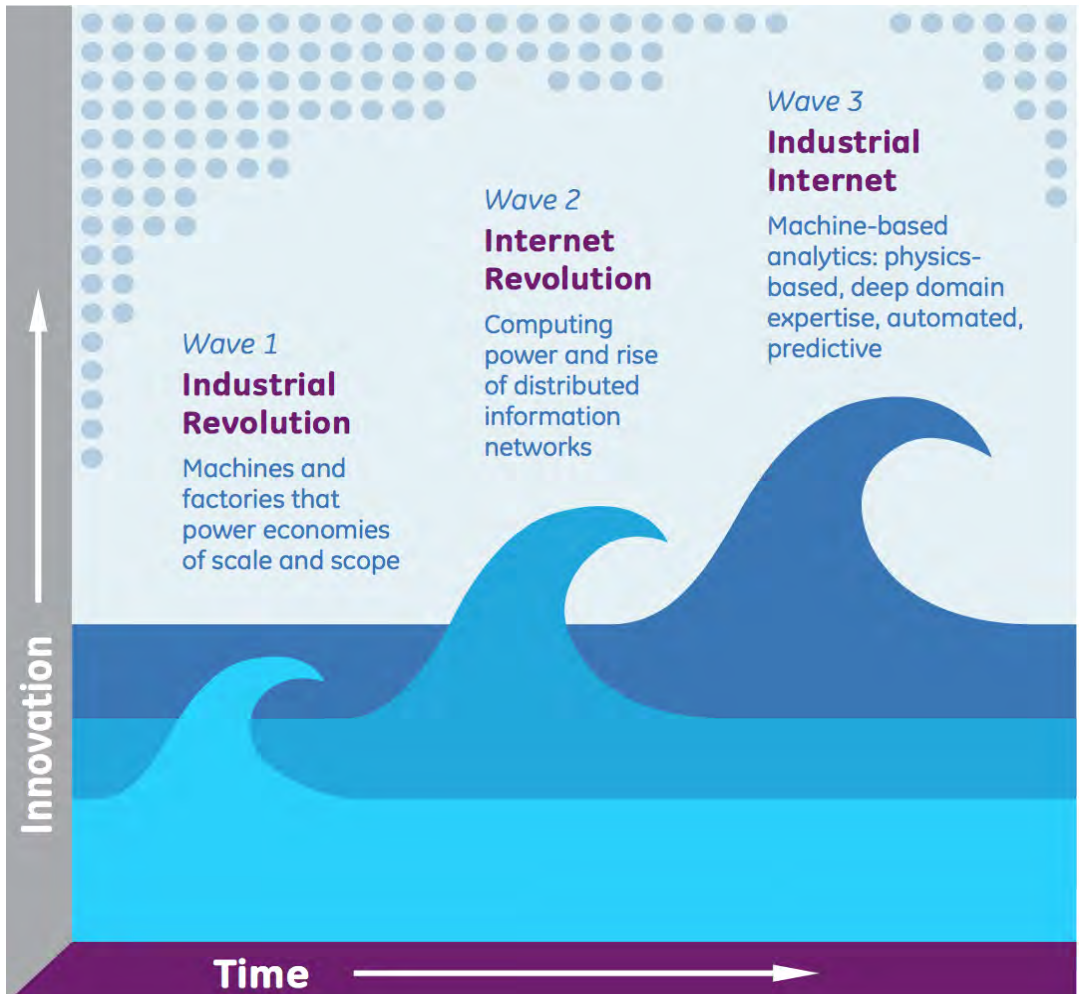


INDUSTRIAL “THINGS”






THE NEXT INDUSTRIAL REVOLUTION



"mobile computing, inexpensive sensors collecting terabytes of data, rise of AI to use that data will fundamentally change the way the global economy is organized."

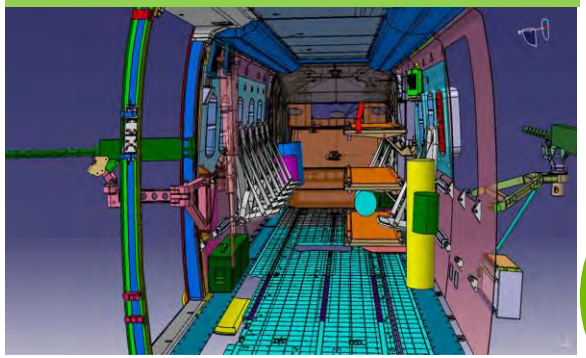
— Fortune CEO – The Revolution is Coming March 8, 2016

Industrial Internet: Pushing the Boundaries of Minds and Machines, P. Evans & Marco Annunziata, GE Nov 26, 2012

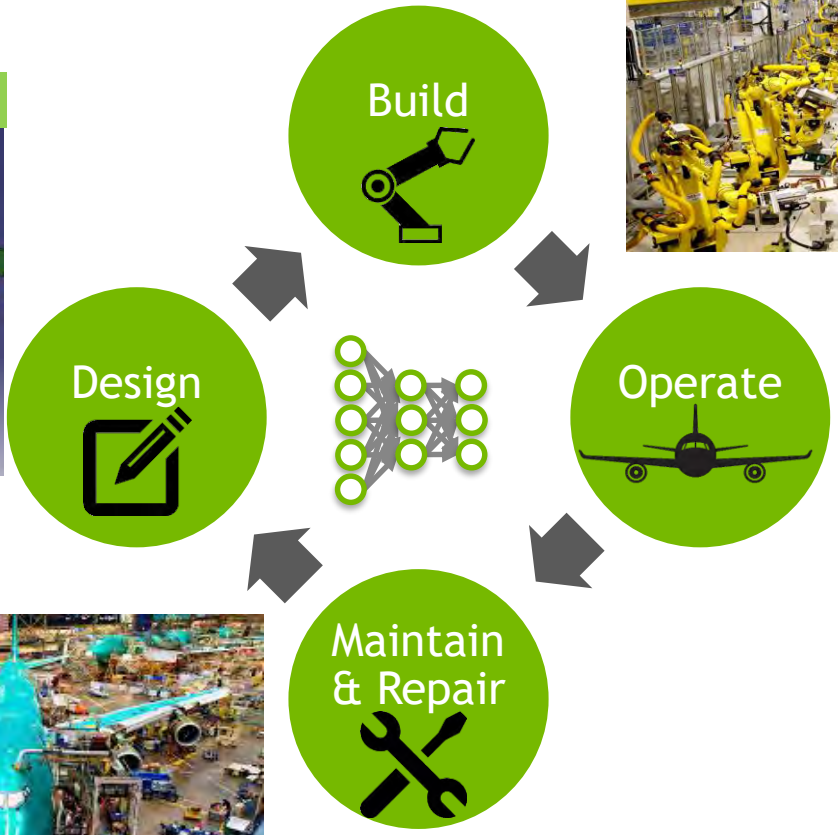
DEEP LEARNING / AI EVERYWHERE

AI product development value cycle

Generative and Predictive Design



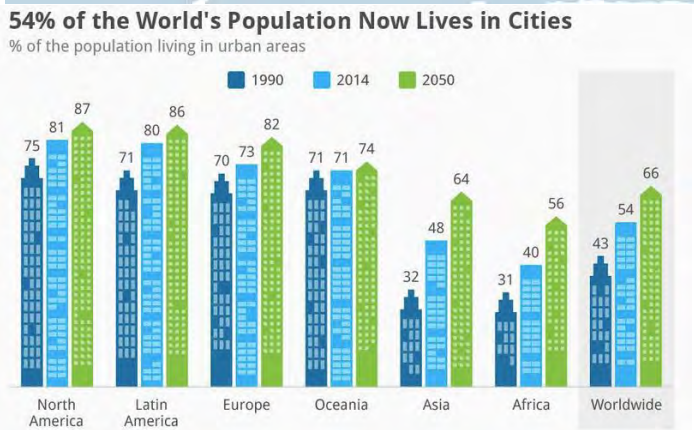
Robots & Smart Manufacturing



Uptime,Efficiency Inspection

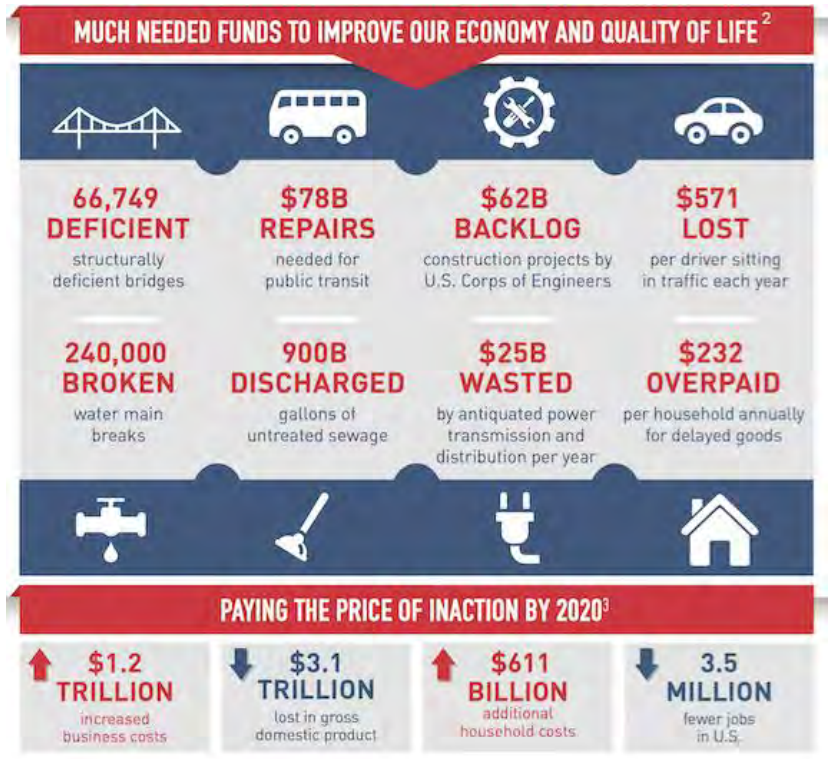
Product Improvement Predictive Maintenance

AI CITIES: NEED FOR INTELLIGENCE



11,774 Terror Attacks Worldwide in 2015; 28,328 casualties. 1.3M casualties due to road accidents worldwide every year.

Digitization making the invisible visible

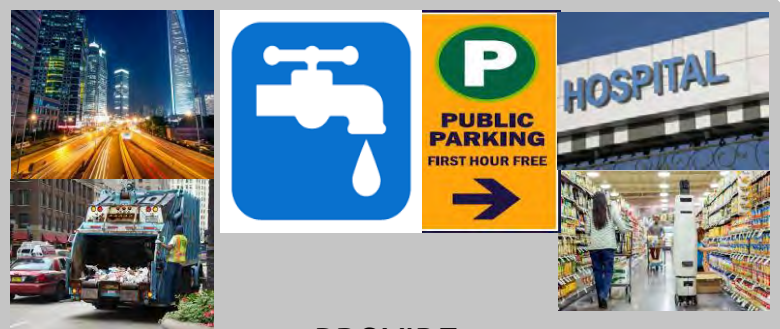


"Global Logistics Market to US$4 Trillion in 2015"

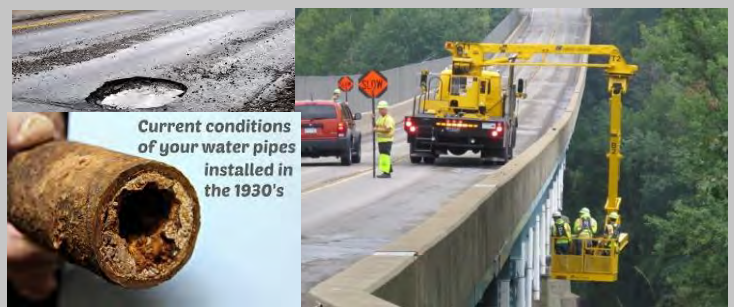
Sources: McKinssey, Bipartisan Policy Center, Population Reference Bureau.

WHAT COULD AN AI CITY DO?

SENSE, LEARN, THINK AND ACT AS THE DISTRIBUTED BRAIN

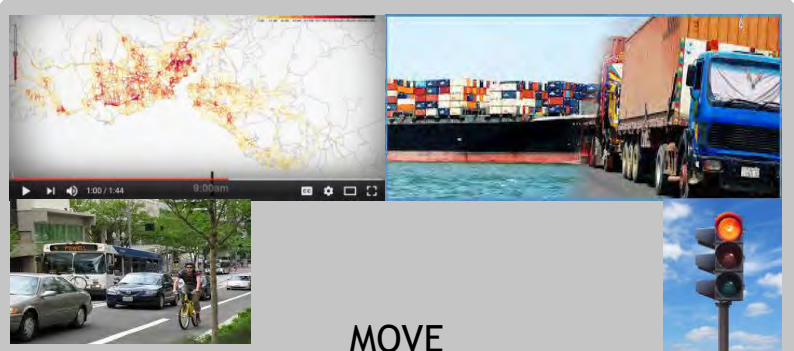


PROVIDE
Necessary resources efficiently

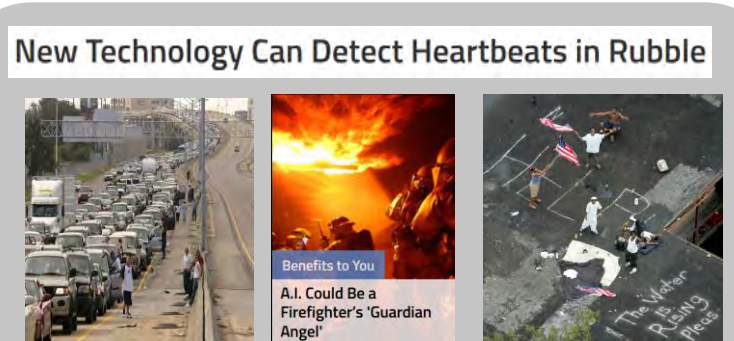


MONITOR & PREDICT
Its asset health & maintenance

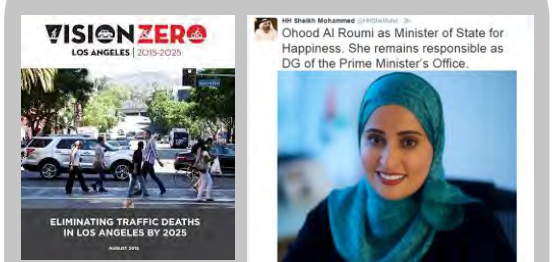
SAVE & PREEMPT
With timely action



MOVE
People & Goods optimally



BE RESILIENT
Prepare, Detect, Evacuate, Search, Rescue



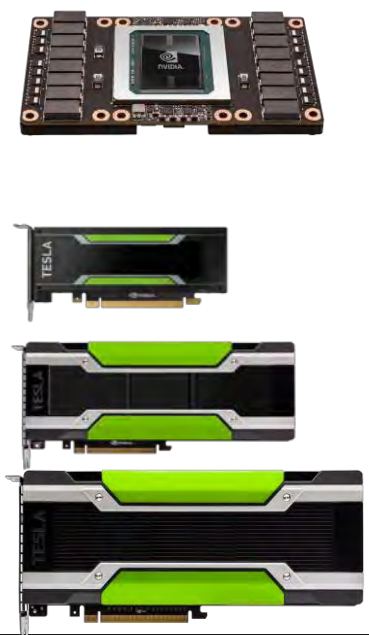
EMPOWER
Vision to become reality

INDUSTRIAL TALKS AT GTC 2017

- Digital Twin, AI, & Industrial Internet of Things, GE
- INDUSTRIAL STRENGTH AI & IMAGING ANALYTICS
- Deep Representation and Reinforcement Learning for Anomaly Detection and Control in Multi-modal Aerospace Applications, United Technologies
- Deep Learning for the IoT: Leveraging Representation Learning, Bosch AI Research
- Deep Learning Applications for Embedded Avionics on the Jetson Platform, Boeing
- Approach to Practical Application of Deep Learning in Manufacturer's Production Line, Fujikura
- AI and the Battle for Cyber Security, Cylance
- High-performance Deep Learning for Embedded Devices, Amazon Web Services
- Deep Learning for Building Energy Intelligence, Verdigris Technologies
- Utilizing GPUs for Avionics Maintenance and Safety, Analatom Inc,
- Leveraging Deep Learning and Drones in Capital Projects Monitoring: Case Study, PwC
- Deep Learning for 3D Design and Making, Autodesk
- Deep Learning for Predictive Maintenance, Reliability Solutions
- High-speed Robotic Weeding, Blue River Technology
- Industrial Perspective on Next Generation of Social Consumer Robots, SoftBank Robotics
- Collision Avoidance for Indoor Navigation of Mobile Robots via RL, University of Washington
- Real-time Anomaly Detection on Video and SCADA with Unsupervised ML Engine, Giant Gray
- GPU Accelerated Deep Learning Framework for Cyber-enabled Manufacturing, Iowa State
- Adaptive 3D Printing Using Multi-agent Systems and Deep Learning, UCL
- Real-time Vertical Relief Profile and Free Space Estimation for Low-Altitude UAV-Sprayer, Kray Technologies

ONE ARCHITECTURE

DEEP LEARNING, TRAINING, INFERENCE & HPC



DEEP LEARNING SUPERCOMPUTER

DGX-1

VISUALIZATION/ AR/ VR

QUADRO

EDGE COMPUTING

JETSON

POWERING DEEP LEARNING

Every major DL framework leverages NVIDIA SDKs

COMPUTER VISION

OBJECT DETECTION

IMAGE CLASSIFICATION

SPEECH & AUDIO

VOICE RECOGNITION

LANGUAGE TRANSLATION

NATURAL LANGUAGE PROCESSING

RECOMMENDATION ENGINES

SENTIMENT ANALYSIS

TIME SERIES

ANOMALY DETECTION & CLASSIFICATION

REINFORCEMENT LEARNING

DEEP LEARNING FRAMEWORKS

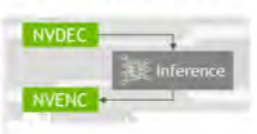
Caffe | Microsoft CNTK | mxnet | TensorFlow | theano | torch

NVIDIA DEEP LEARNING SDK

cuDNN

TensorRT

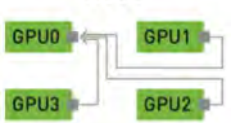
DeepStream SDK



cuBLAS

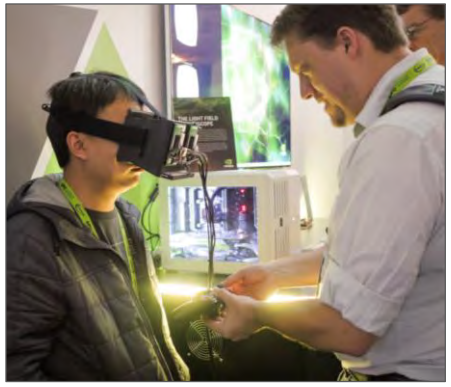
cuSPARSE

NCCL



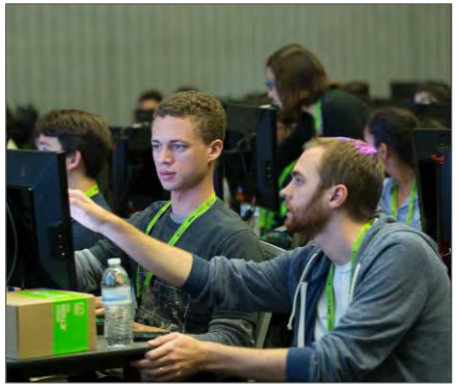
GPU TECHNOLOGY CONFERENCE

May 8 - 11, 2017 | Silicon Valley | #GTC17
www.gputechconf.com



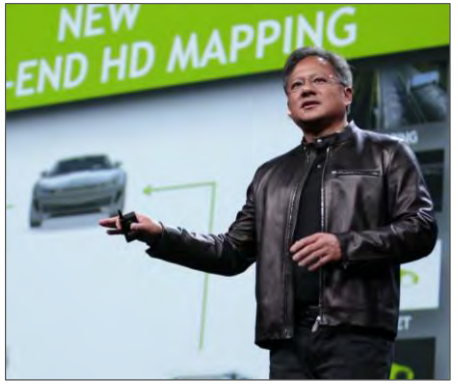
CONNECT

Connect with technology experts from NVIDIA and other leading organizations



LEARN

Gain insight and valuable hands-on training through hundreds of sessions and research posters



DISCOVER

See how GPUs are creating amazing breakthroughs in important fields such as deep learning and AI

INNOVATE

Hear about disruptive innovations from startups

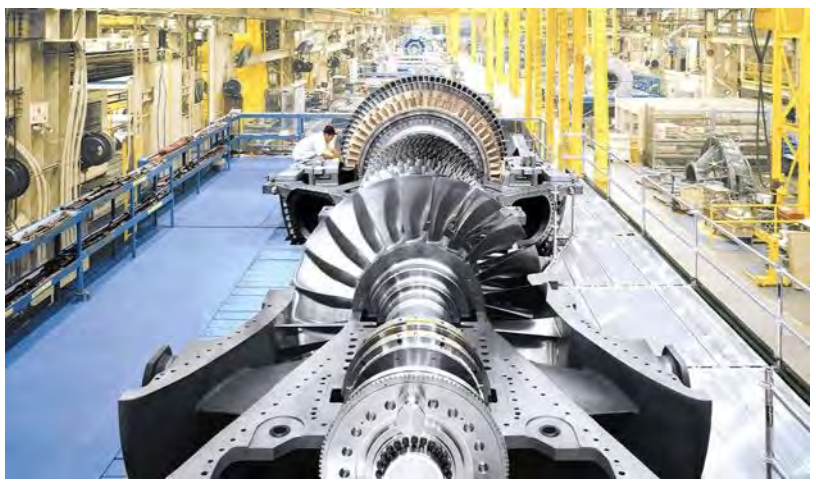
REGISTER EARLY: SAVE UP TO $240 AT WWW.GPUTECHCONF.COM

Don’t miss the world’s most important event for GPU developers
May 8 – 11, 2017 in Silicon Valley

INDUSTRIAL IOT & AI CITIES: USE CASES

GENERAL ELECTRIC GLOBAL RESEARCH

Application: Gas Turbine Combustors

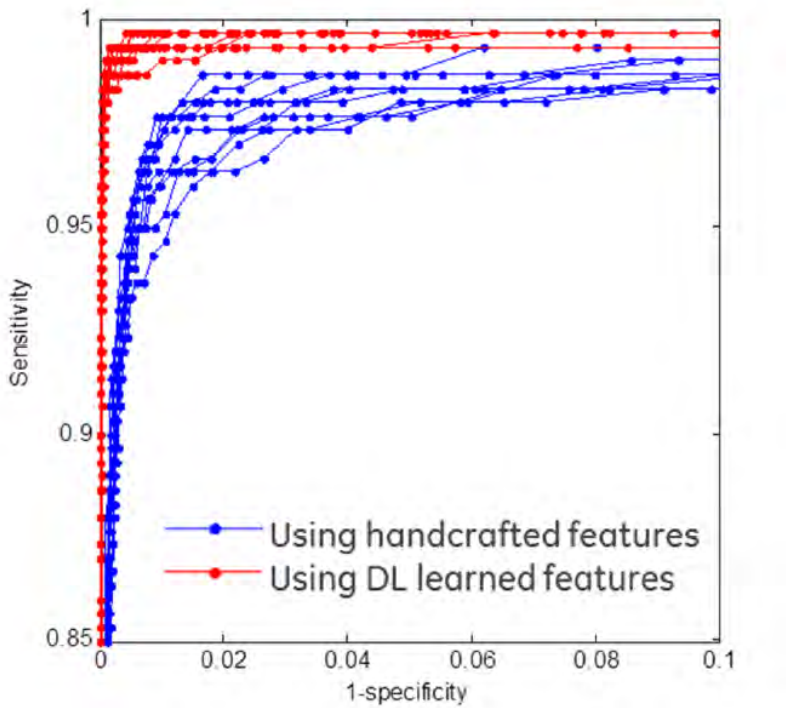


- Anomaly detection from exhaust temperature profile
- Preventative early detection of catastrophic failures
- Challenges: complex system, multiple dependencies (e.g. machine type & configuration, fuel, ambient conditions, aging equipment)

Accurate and Reliable Anomaly Detection for Gas Turbine Combustors: A Deep Learning Approach, W. Yan et al, General Electric Global Research Center

GENERAL ELECTRIC GLOBAL RESEARCH

Predictive Maintenance: Gas Turbine Combustors

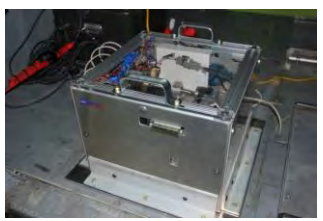


- Deep learning outperforms handcrafted features
- Maximum sensitivity (true positive) with maximum specificity (true negative)
- Deep learning also less problem specific, more scalable

Accurate and Reliable Anomaly Detection for Gas Turbine Combustors: A Deep Learning Approach, W. Yan et al, General Electric Global Research Center

UNITED TECHNOLOGIES RESEARCH CENTER

Anomaly Detection and Fault Classification in NASA Flight Data



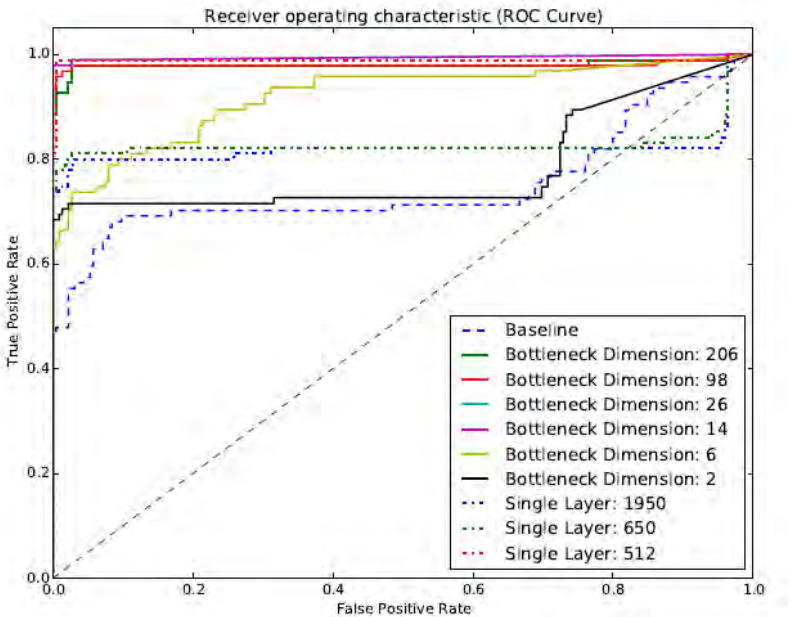
https://c3.nasa.gov/dashlink/, Balaban et al

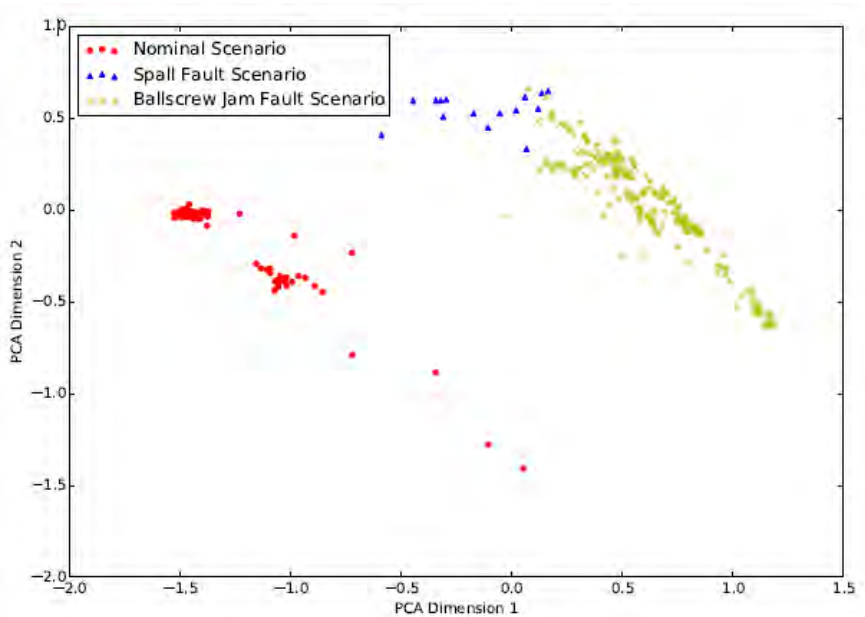
- Trained on raw time series from heterogeneous sensors
- Real data collected from multi-sensor electromechanical actuators for aircraft operating scenarios
- Deep auto-encoder
- Fine-tuned with back-propagation to minimize reconstruction error
- No hand-crafted features

Anomaly Detection and Fault Disambiguation in Large Flight Data: A Multi-modal Deep Auto-encoder Approach, K. Reddy et al, United Technologies Research Center (PHM16)

UNITED TECHNOLOGIES RESEARCH CENTER

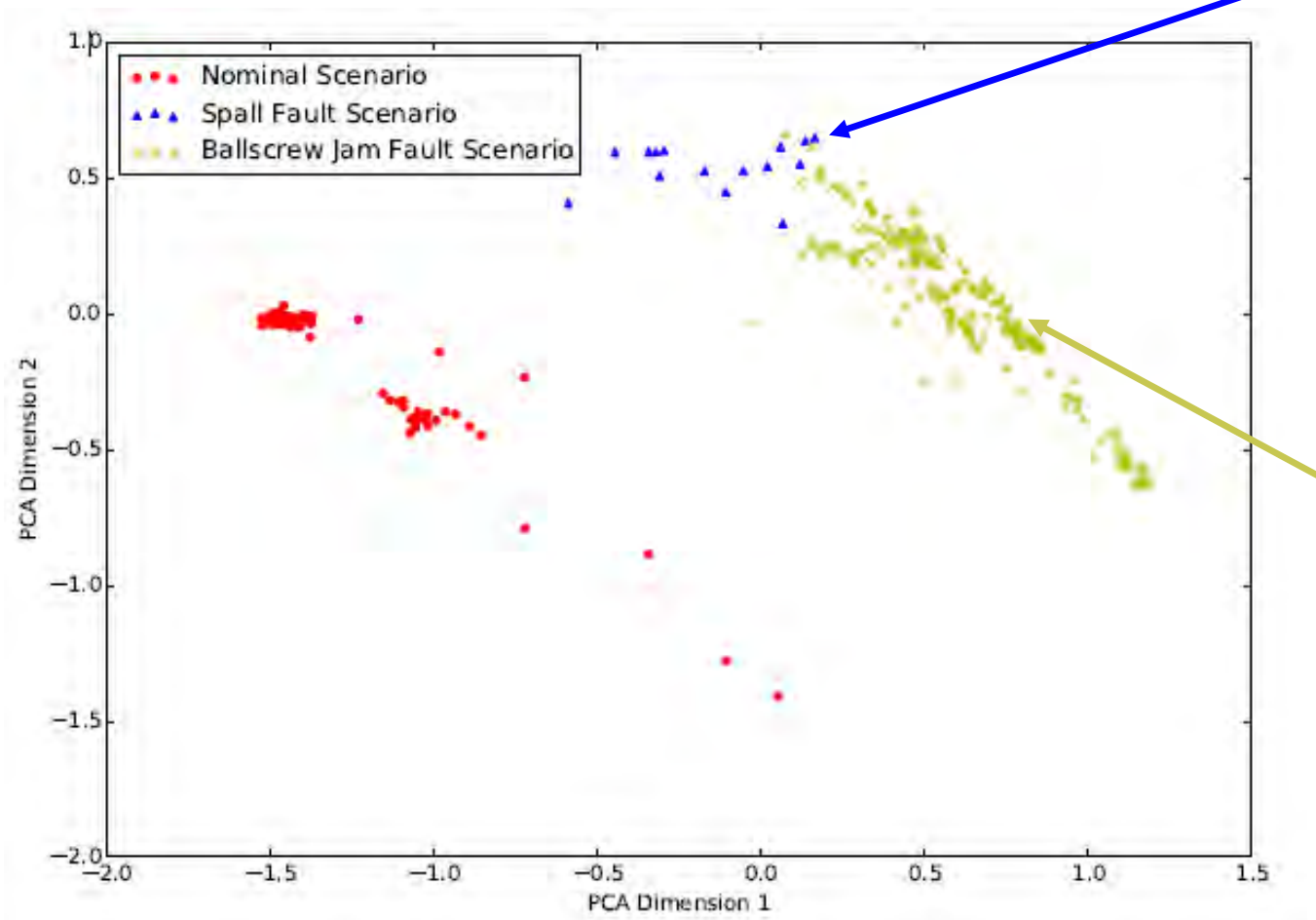
Fault Detection & classification

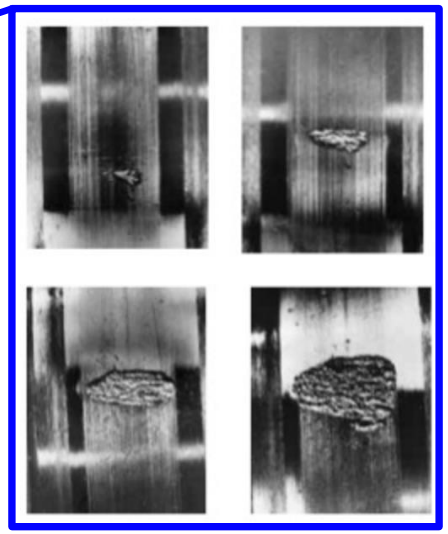


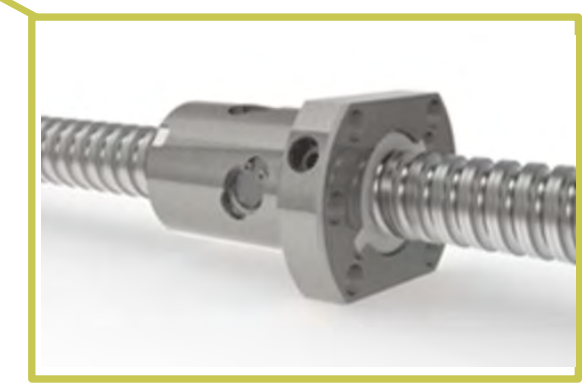


- 11-layer 14-dimensional bottleneck DAE yields 97.8% true positive detection rate with 0.0% false alarm

Anomaly Detection and Fault Disambiguation in Large Flight Data: A Multi-modal Deep Auto-encoder Approach, K. Reddy et al, United Technologies Research Center (PHM16)



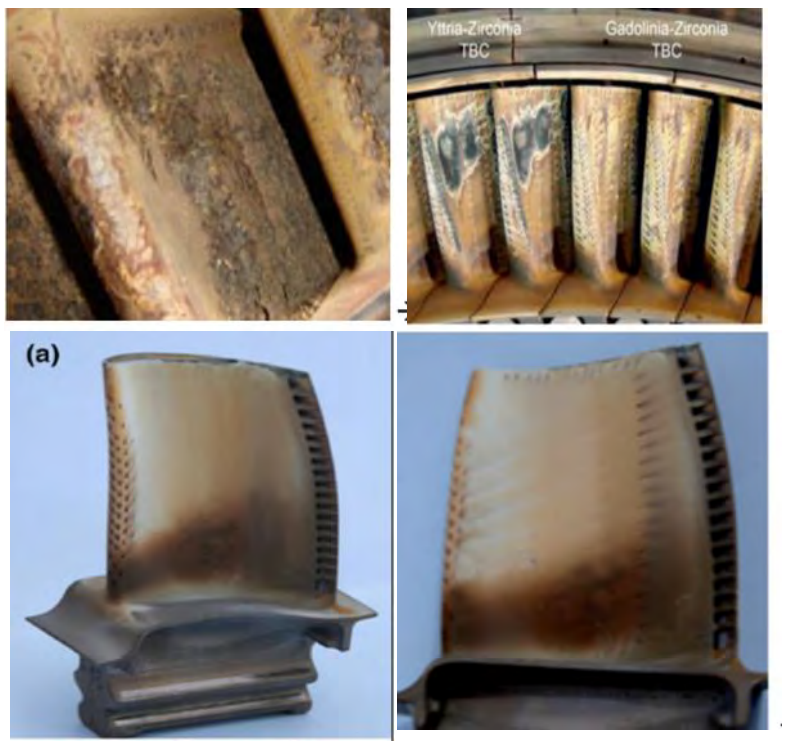




Anomaly Detection and Fault Disambiguation in Large Flight Data: A Multi-modal Deep Auto-encoder Approach, K. Reddy et al, United Technologies Research Center (PHM16)

NON DESTRUCTIVE TESTING

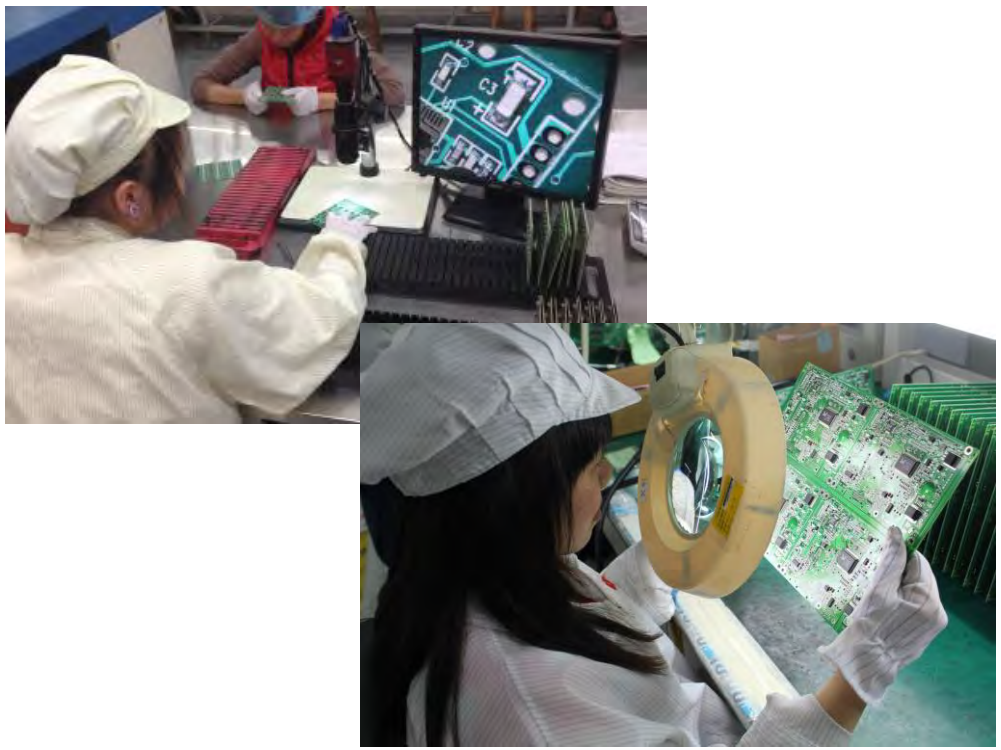
Defect identification, Life Prediction



- Classify defects
  - Discoloration,
  - Coating/material loss,
  - Corrosion
- Predict defect progression
- Multi-modal (radiography, computed tomography, remote visual inspection, ultrasound, Electromagnetic) image co-relations

APPLICATION: INDUSTRIAL INSPECTION

Foxconn | Test Research, Inc. | NVIDIA

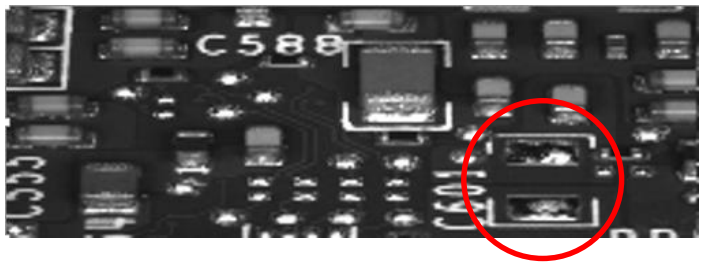


- **Industry**
  Electronics manufacturing
- **Use case**
  Component assembly, circuit boards
  Surface-Mount Technology (SMT)
- **Problem**
  Quality control inspection
  Manual
  Labor intensive

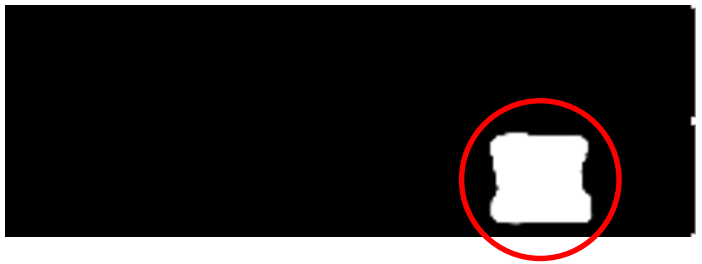
APPLICATION: INDUSTRIAL INSPECTION

Foxconn | Test Research, Inc. | NVIDIA

Reference example
No missing components



Missing component example
One or more components missing



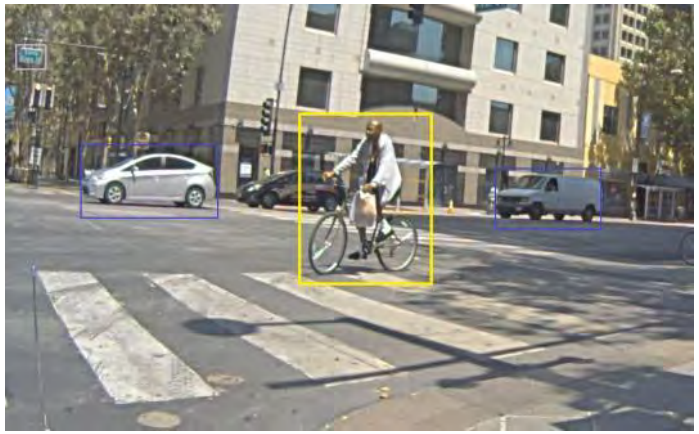
Fault localization

APPLICATION: SENSOR ESTIMATION

United Technologies Research Center

- **Industries**
  Aerospace & Building Management Systems
- **Use cases**
  Sensor estimation & prediction
  Virtual sensors

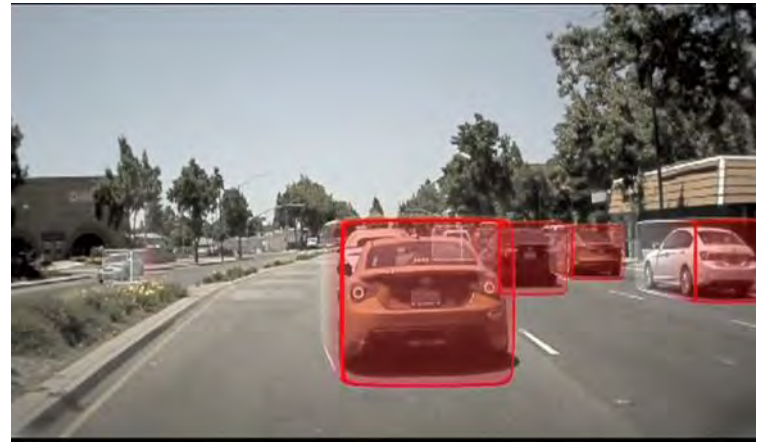
DEEP LEARNING FOR SELF DRIVING CARS



Multi-class detection (DriveNet)

OpenRoadNet

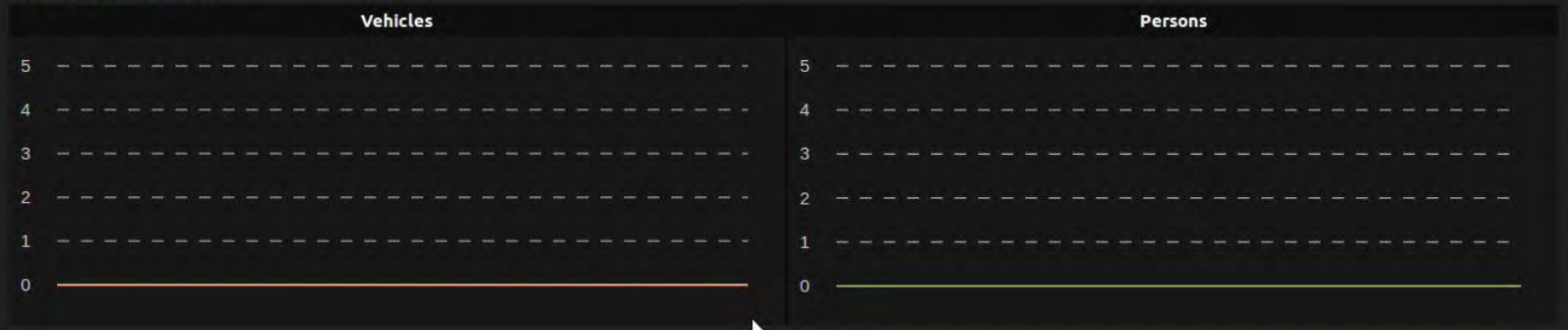
LaneNet



3D Bounding Boxes

Camera 1 18:44:47 03/30/2017

Detection Trends:



NOTE: Objects displayed are subject to a configurable detection duration

NVIDIA

AI FOR DRONES

AERIALTRONICS
REMOTELY PILOTED AIRCRAFT SYSTEMS

Problem

Aerial inspection is

- Imprecise: often needs multiple flights
- Time consuming: manual review of footage
- Dangerous: drone crashes into subject or operator

Solution

Automate the process

- Vision-enabled navigation
- On-board verification
- On-board fault classification

AI FOR SMART BUILDINGS

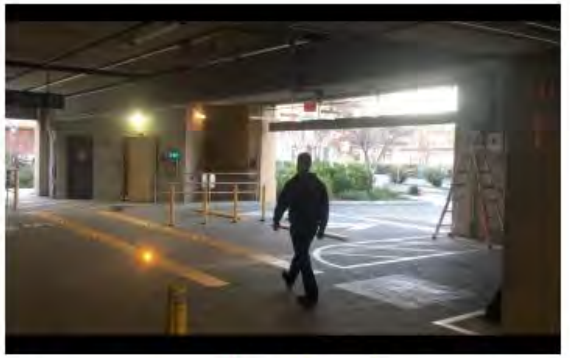
VIMOC

Examples of fielded solutions

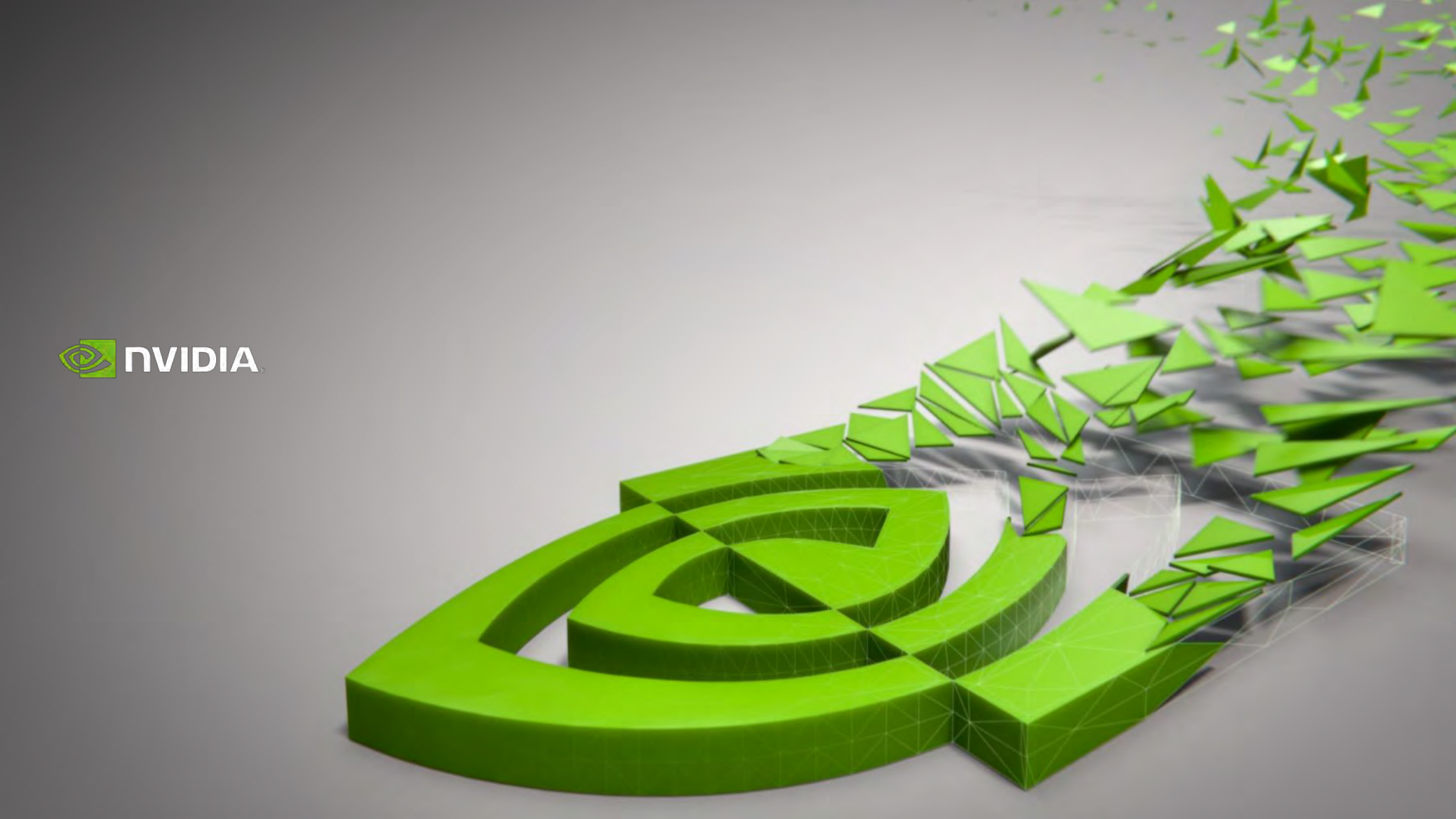
"End to end" (see video links by clicking on images)

Bicycle and Pedestrian Detection

Garage Occupancy Optimization



Crosswalk Pedestrian Safety Detection and Alert System


NVIDIA

NVIDIA DGX-1

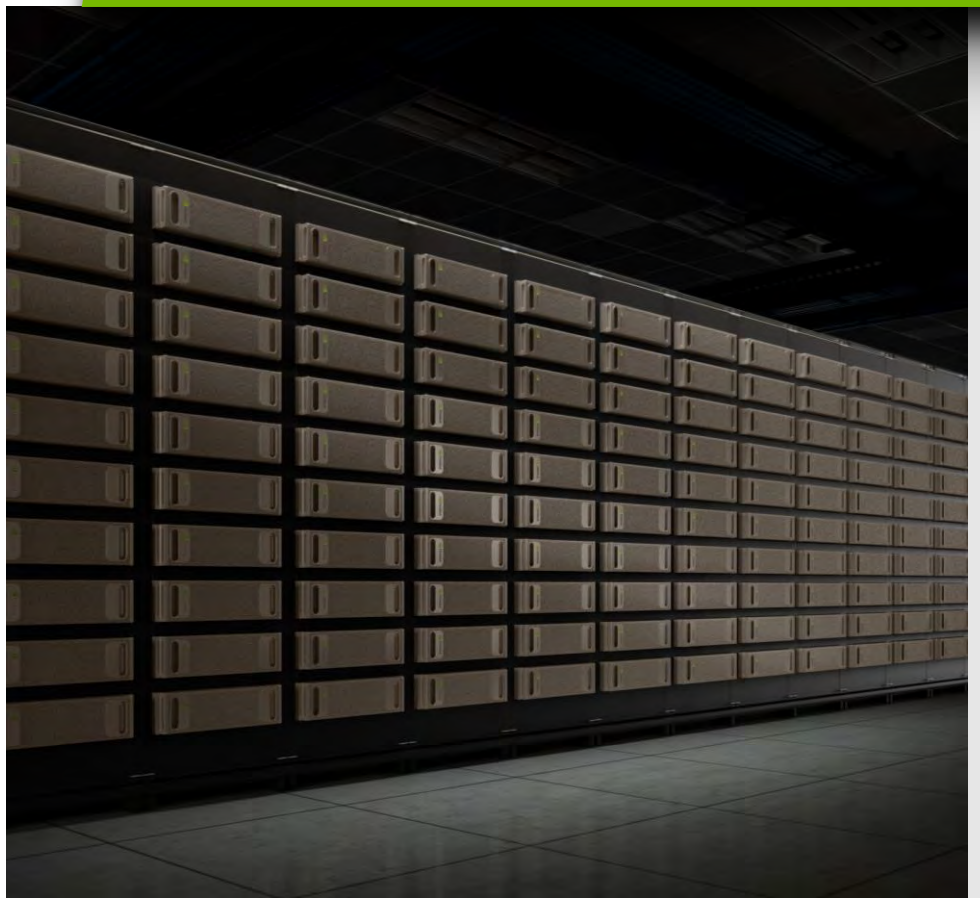
AI Supercomputer-in-a-Box

170 TFLOPS | 8x Tesla P100 16GB @ 732GB/s each | NVLink Hybrid Cube Mesh

2x Xeon | 8 TB RAID 0 | Quad IB 100Gbps, Dual 10GbE | 3U — 3200W

NVIDIA 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

FACEBOOK BIG SUR SERVER DEPLOYMENT

Pineville Data Center, Circa 2016

- Four 8-GPU servers per rack
- Scale-out data center's limited power and cooling prevents denser configurations

Project Olympus HGX-1

Hyperscale GPU Accelerator

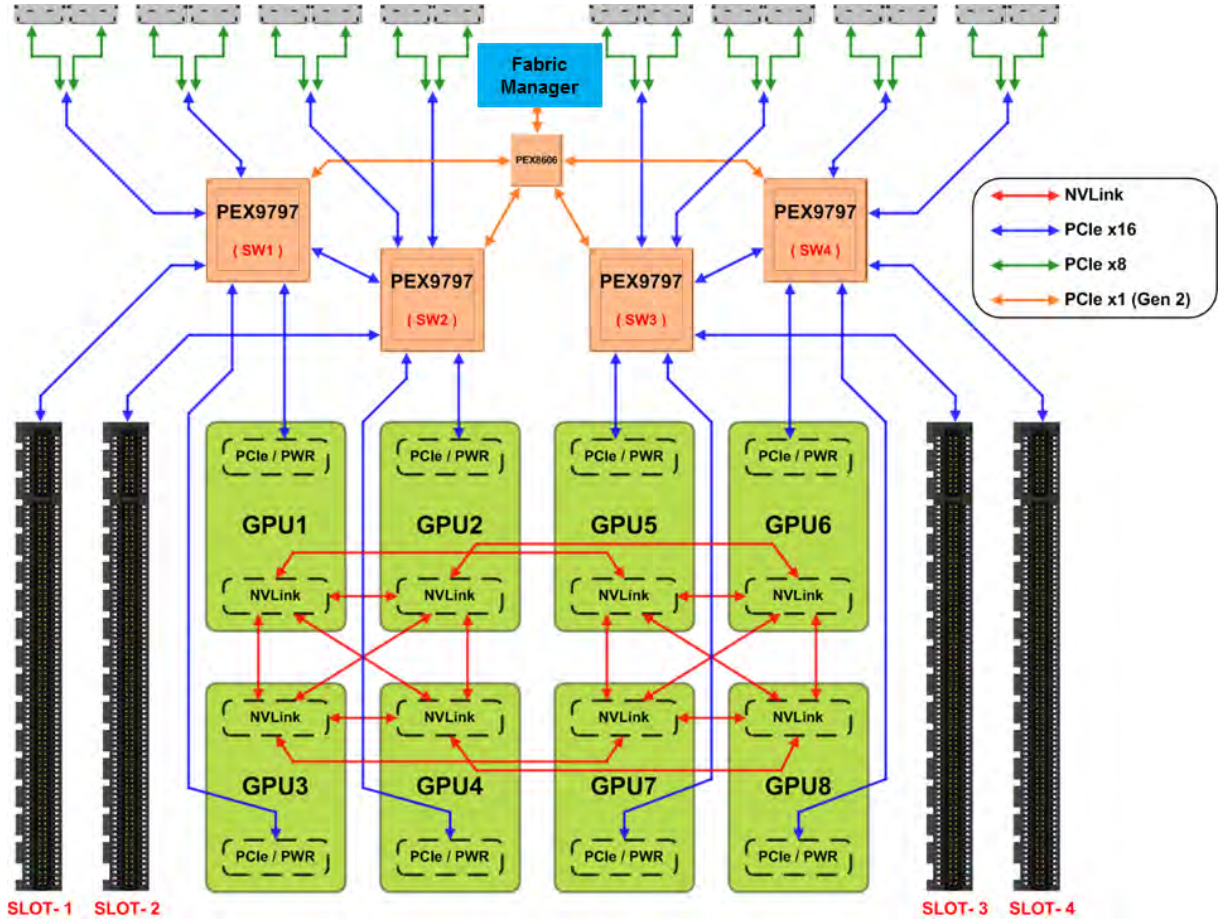
Configurable PCIe Cable to host + Expansion slots

NVIDIA P100 GPU

NVLink Hybrid Cube Mesh Fabric

20 Gbyte/sec per link Duplex

Adapters for other GPUs



PROJECT OLYMPUS HGX-1

Industry Standard Hyperscale GPU Accelerator