# Defense Data Generation in Distributed Deep Learning System

Se-Yoon Oh / ADD-IDAR

2017. 10. 31

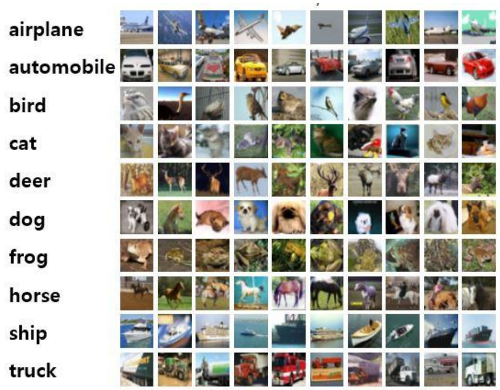
syoh@add.re.kr

# Overview

1. Introduction
2. Data Generation Synthesis
3. Distributed Deep Learning
4. Conclusions

# • ML Process (Supervised Learning, DNN)

① **Labeled Training Images**



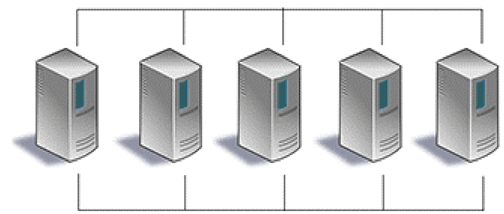

② **Neural Network Framework**

Caffe, Chainer, DL4J, KERAS, CNTK, MatConvNet, MINERVA, mxnet, Purine, TensorFlow, theano, torch

③ **NN Model Definition**

```
model.add(Convolution2D(128, 3, 3)
```

**Train Model on GPU Accelerated Cluster**



④

**Trained Network**

input

conv

conv

pool

fullycon

fullycon

softmax

**CAT**

• **ML Process** **(Supervised Learning, DNN)**

① **Labeled Training Images**


**Benchmark DB ?**

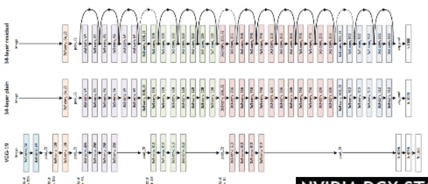
**Transfer learning ?**

**One-shot learning ?**

② **Neural Network Framework**

③ **NN Model Definition**

```
model.add(Convolution2D(128, 3, 3)
```

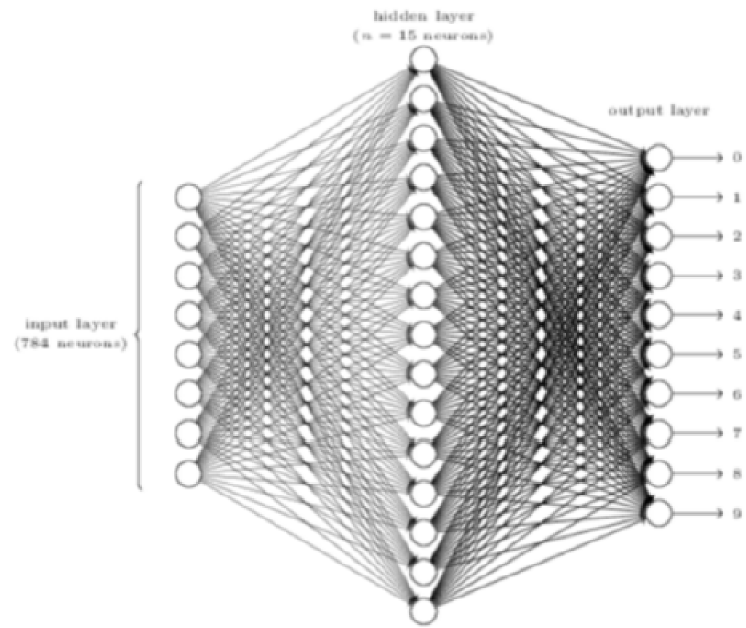


④ **Train Model on GPU Accelerated Cluster**


# • NN Hyper-Parameter

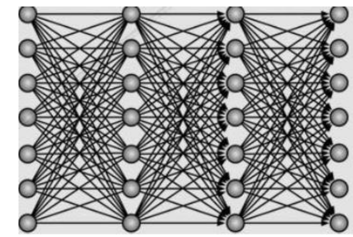
## Hyper-Parameters

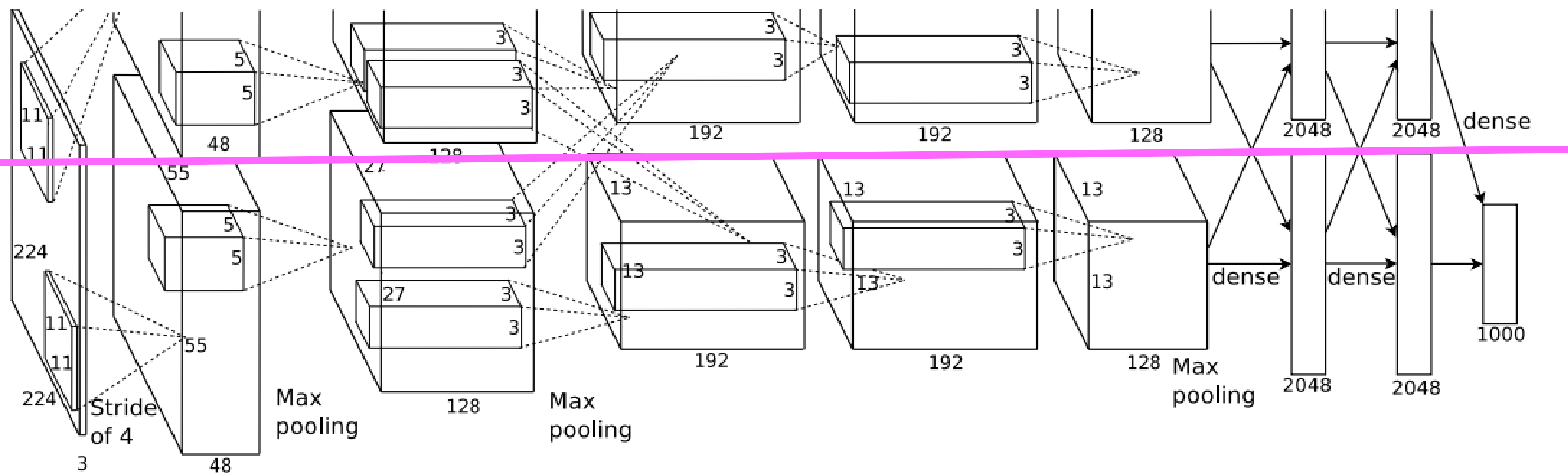
- Learning rates
- Cost function
- Regularization parameter
- Batch size
- Training epochs
- No. of layers
- Convolutional filters
- Convolutional kernel size
- Weight initialization
- Dropout fractions
- ...



## ImageNet Classification with Deep Convolutional Neural Networks (Alex Krizhevsky,2012)

- 650,000 neurons, 60,000,000 parameters
- 1.2 M images(224*224*3), 1,000 classes
- 7 Hidden layers, ReLU, DropOut, 2 GPU(GTX580)/week
- Overlapped max pooling, Augmentation



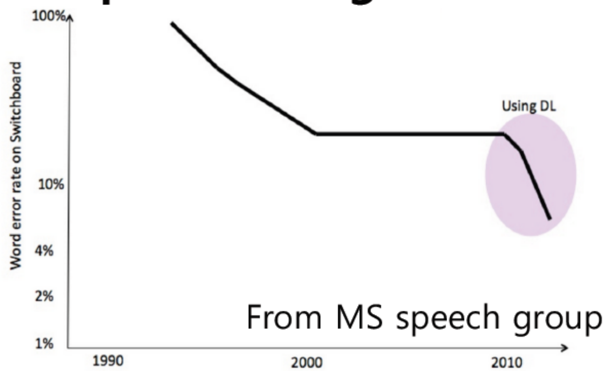



1st Layer
- 290,400개 (N)
- ((11*11*3)+1) * 96 = 34,944개 (P)
- 290,400 * 364 = 105,750,600개 (C)

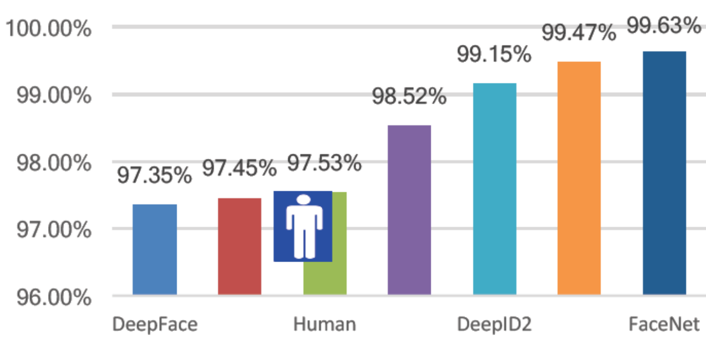
ConV L : 90~95% 연산량 (5% Para)
Full L : 5~10% 연산량 (95% Para)

# • Tech. Trends

## Speech recognition




## Face verification (LFW)




## Object Classification(ImageNet)

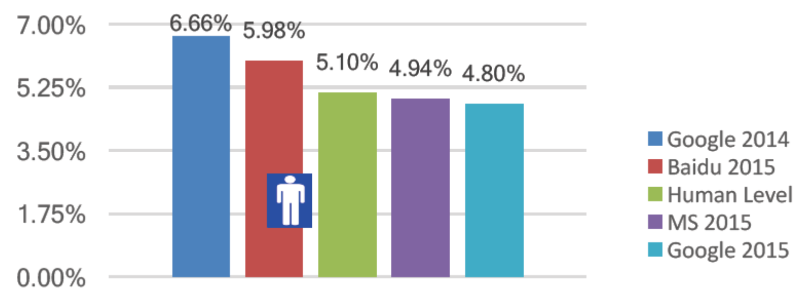
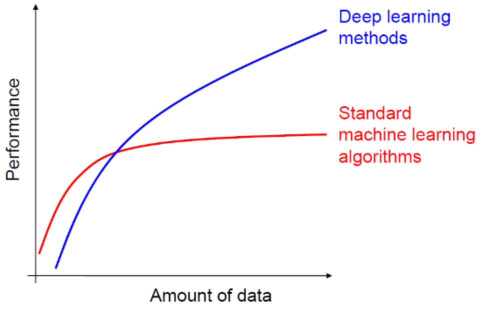



*Table 3.* ILSVRC classification task top-5 performance (with provided training data only).

| Team | Time | Place | Top-5 error |
|---|---|---|---|
| SuperVision | 2012 | 1 | 16.42% |
| ISI | 2012 | 2 | 26.17% |
| VGG | 2012 | 3 | 26.98% |
| Clarifai | 2013 | 1 | 11.74% |
| NUS | 2013 | 2 | 12.95% |
| ZF | 2013 | 3 | 13.51% |
| GoogLeNet | 2014 | 1 | 6.66% |
| VGG [3] | 2014 | 2 | 7.32% |
| MSRA | 2014 | 3 | 8.06% |
| Andrew Howard | 2014 | 4 | 8.11% |
| DeeperVision | 2014 | 5 | 9.51% |
| MSRA PReLU-nets | 2015.2 | - | 4.94% |
| BN-Inception | 2015.2 | - | 4.82% |
| Deep Image | 2015.5 | - | **4.58%** |

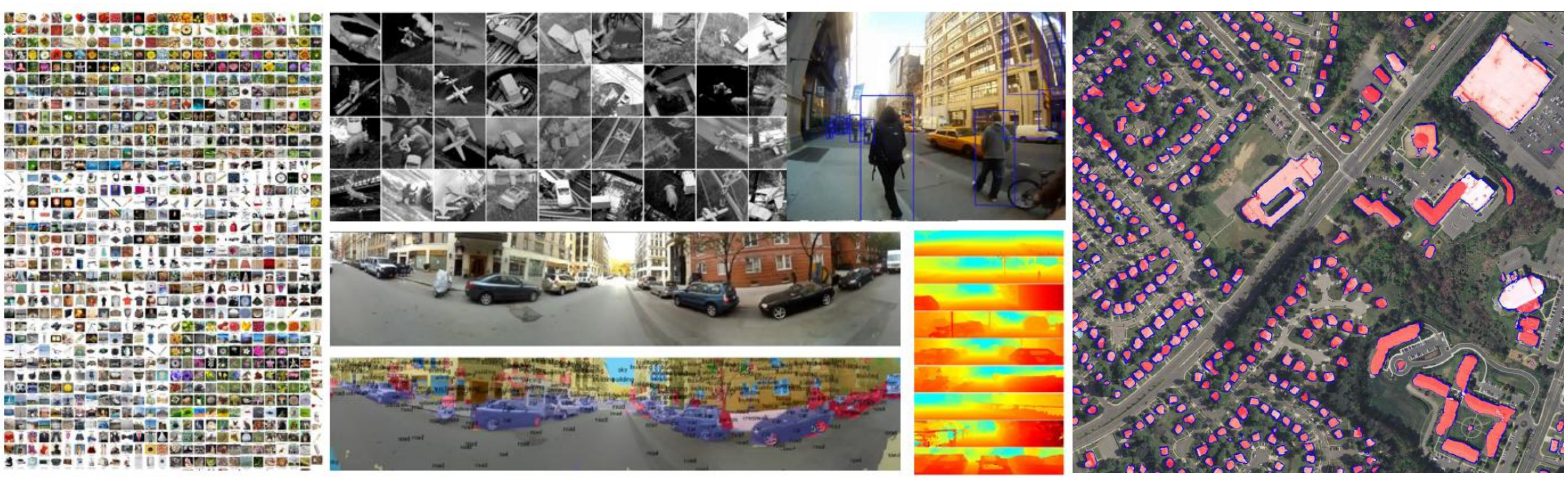
- Supervised learning(SL)
  - Standard in the current state of ML
  - Performed well on many task
  - Ground-truth annotation/label is required in training
- Unsupervised learning(UL)
  - How most learning is done in the real world
  - Impractical to accurately annotate data
  - UL seems to be the key to real artificial intelligence
- Transfer learning & fine-tuning
  - CNN -> Training with millions of unlabeled images(UL) -> Deep understanding in images
  - Quickly learn the new task with much smaller labelled images (…?)

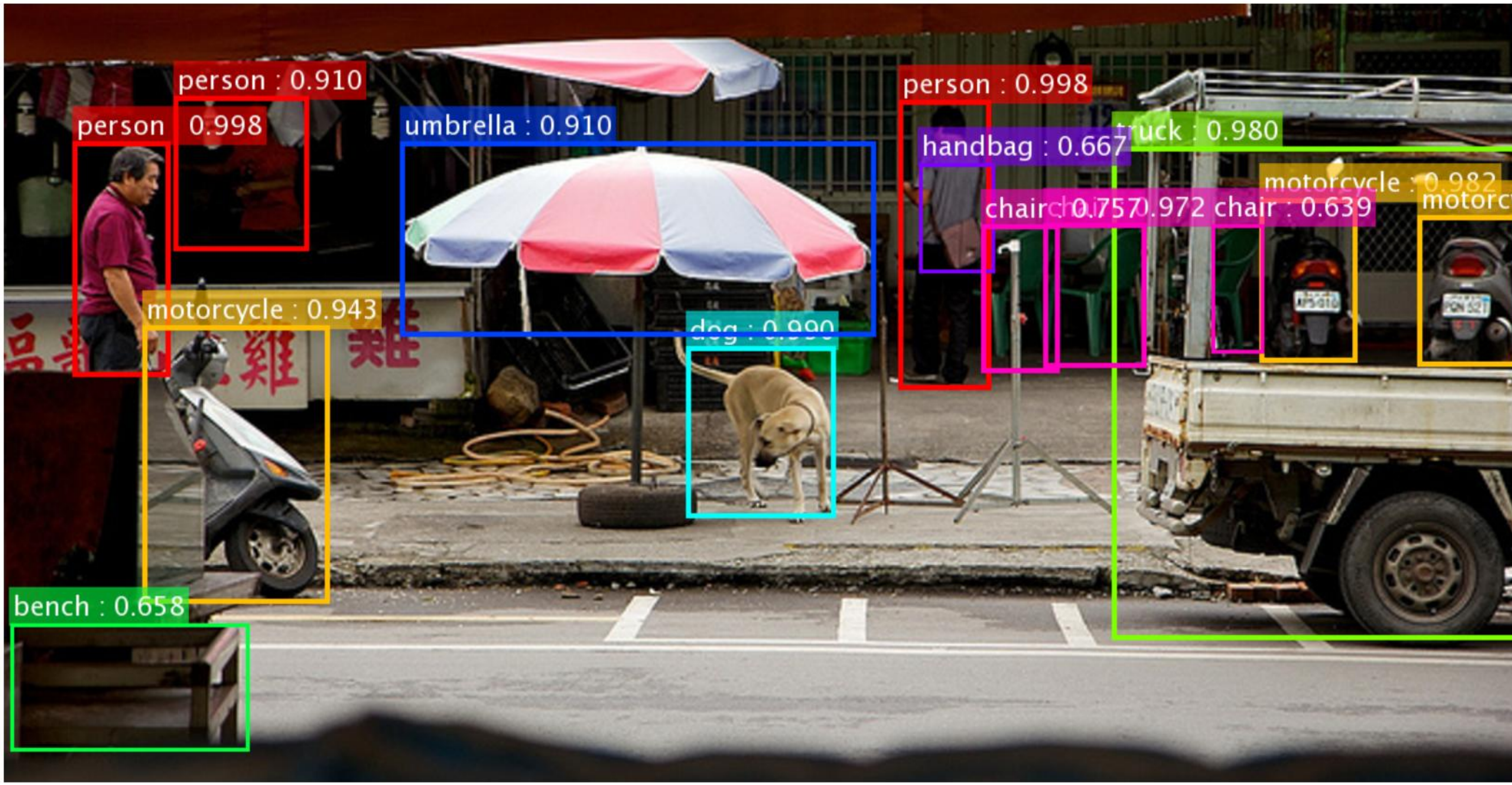
- **Two Things Considered**
  - **Labeled data**
    - Few labeled data in defense
    - Classification/object detection/segmentation/tracking/autonomy/…
  - **Computing power**
    - ML/DL based on HPC & GPU
    - Big data hub & data analytics

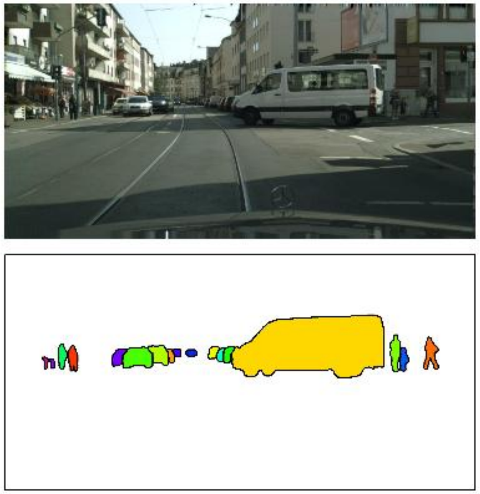


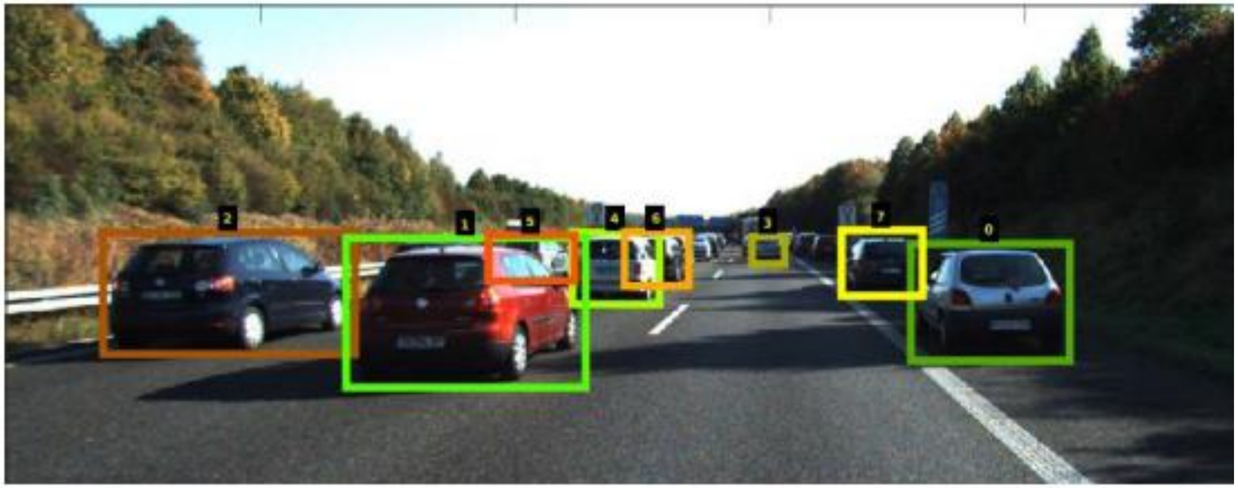
# Overview

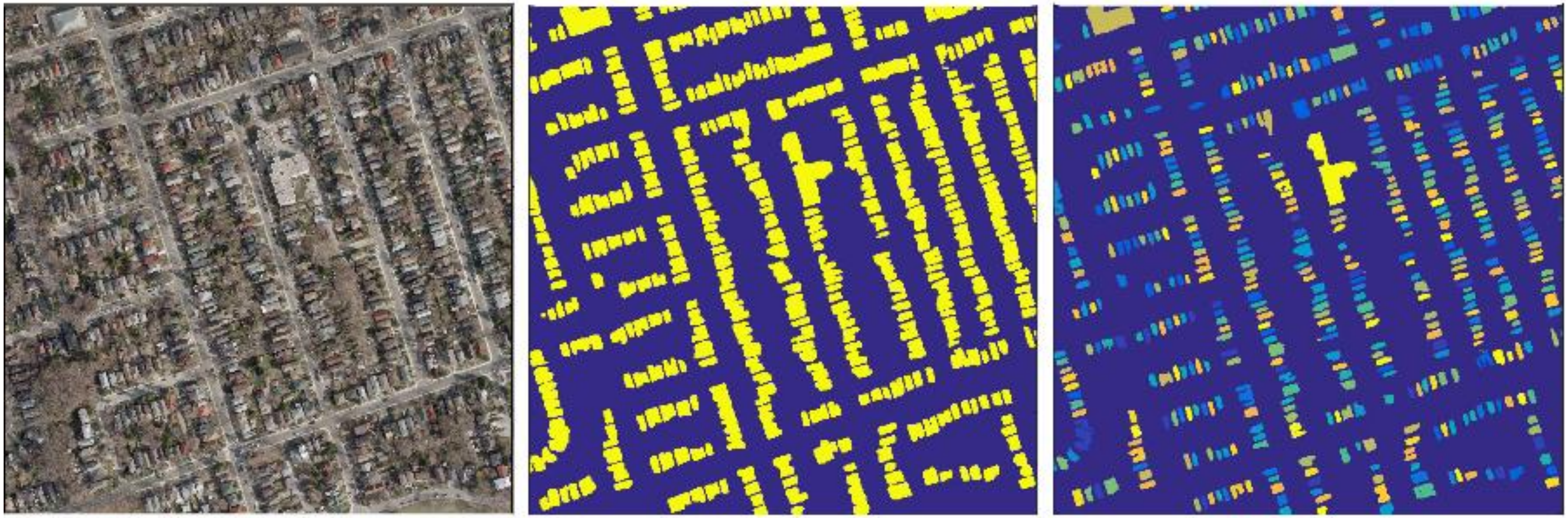
- # Problem Statement
  - ML demands access to very large sets of training data
  - Significant problem may arise in real-world defense ML applications
  - To overcome problems of small-data, novel synthetic data generation frameworks for massive ML training system are required

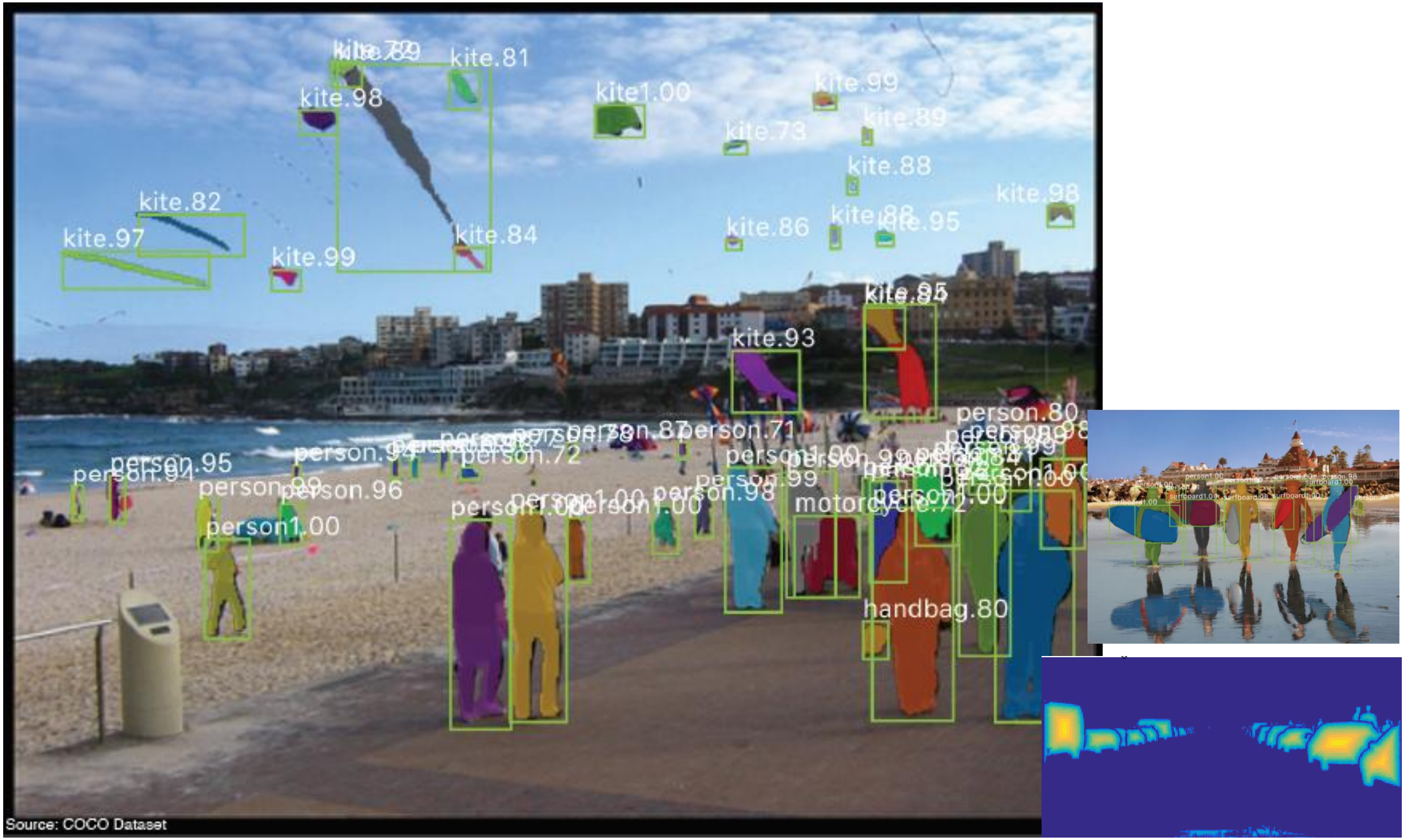



person : 0.910
person
0.998
umbrella : 0.910
person : 0.998
handbag : 0.667
truck : 0.980
motorcycle : 0.982
motorcy
chair : 0.639
motorcycle : 0.943
dog : 0.990
bench : 0.658





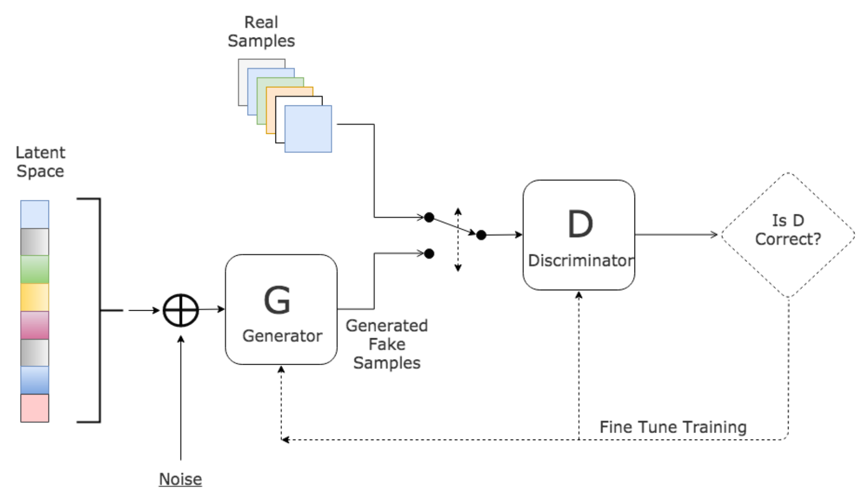



kite.81
kite.98
kite1.00
kite.99
kite.89
kite.73
kite.88
kite.98
kite.82
kite.97
kite.84
kite.99
kite.86
kite.95
kite.93
person.80
person.71
person.95
person.96
person1.00
person.72
person.99
person.98
motorcycle.72
handbag.80
Source: COCO Dataset

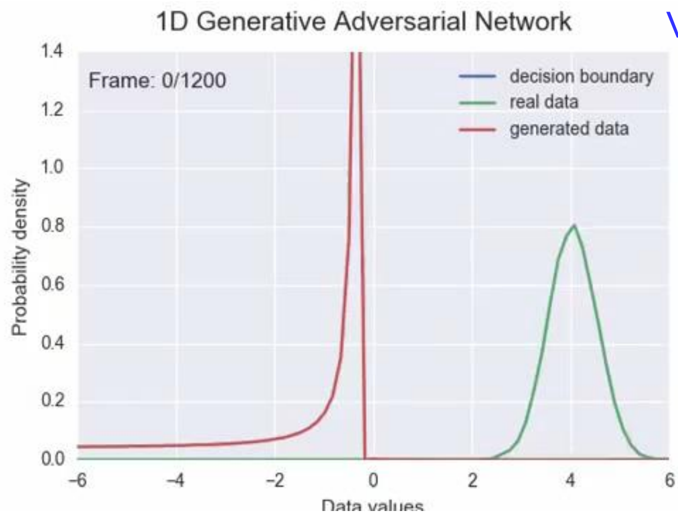
- **Objectives & Motivation**
  - Develop a massive ML training dataset generation/synthesis system
    1) Automatic Data Augmentation framework
    2) 3D CAD modeling based synthetic/virtual image datasets generation framework
    3) Game Development Engine based virtual scenes/military env. creation framework
    4) Generative Models(GANs) framework for photorealistic images tuning

- **Generative Adversarial Networks(GANs)**
  - **Data** : a dataset as samples from a true data distribution
  - **Generator** : Random noise as input, and transforms it, outputting a sample of data
  - **Discriminator** : Classifies sample input data as real(from true data distribution) or fake(from the generator), x = g(z)

**Play a song on guitar**
- Song (Data)
- Hands/brain (G)
- Ears/brain (D)

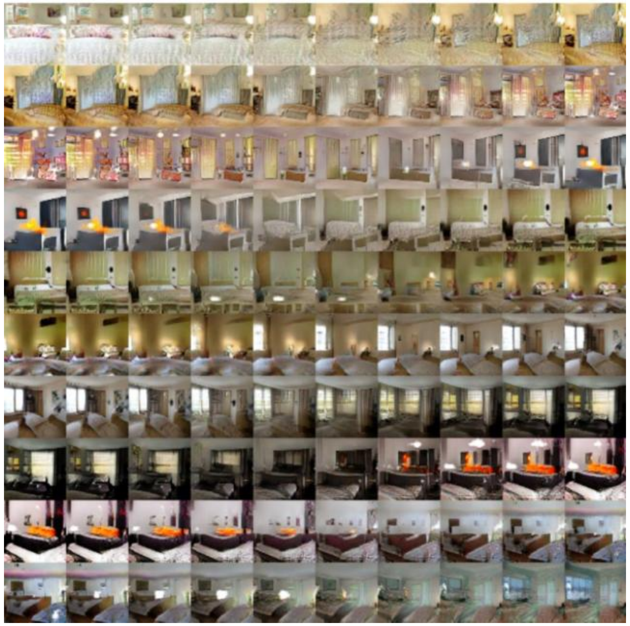


- **GANs**




AYLIEN 2016

Latent Space Variation



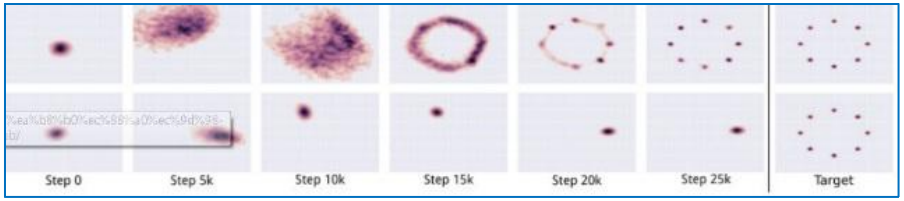
volcano

junco

PPGN 2016 / Evolving AI Lab


Yann Le Cun 2017.3 China

- **Difficulties**
  - **Mode collapsing**
  - **Oscillating during training**
  - **No learning**



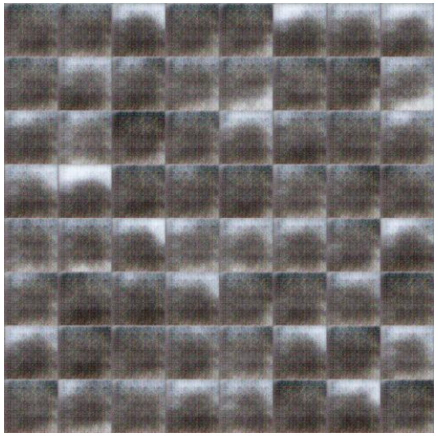

CycleGAN
Ian Goodfellow
GTC 2017
(Zhu et al., 2017)

- **GANs-based Image Generation**
  - Dataset: 1,300 images of Army-tank (ILSVRC2012)
  - Model: DCGAN

Real army-tank training images(ImageNet DB)

Generated synthetic army-tank images (Epoch=189)

- **GANs-based Image Generation**
  - **Changes of generated synthetic images along with iterations**



**Epoch=5**

**Epoch=26**

**Epoch=99**

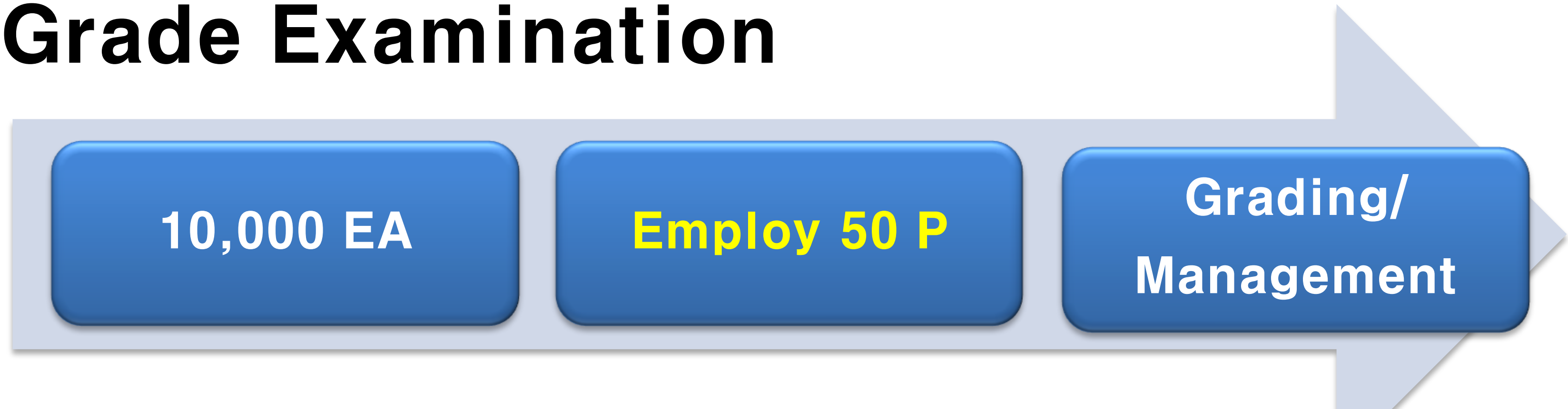
**Epoch=189**

# Overview

1. Introduction
2. Data Generation Synthesis
3. Distributed Deep Learning
4. Conclusions

- **Grade Examination**

- Grade Examination




  - Employee Mgt. : absence/medical/missing/…
- General Management
  - Management/processing/copying/data missing/summary/reporting/…
  - Distributed Computing
    - Big Data Computing : Apache-Hadoop S/W

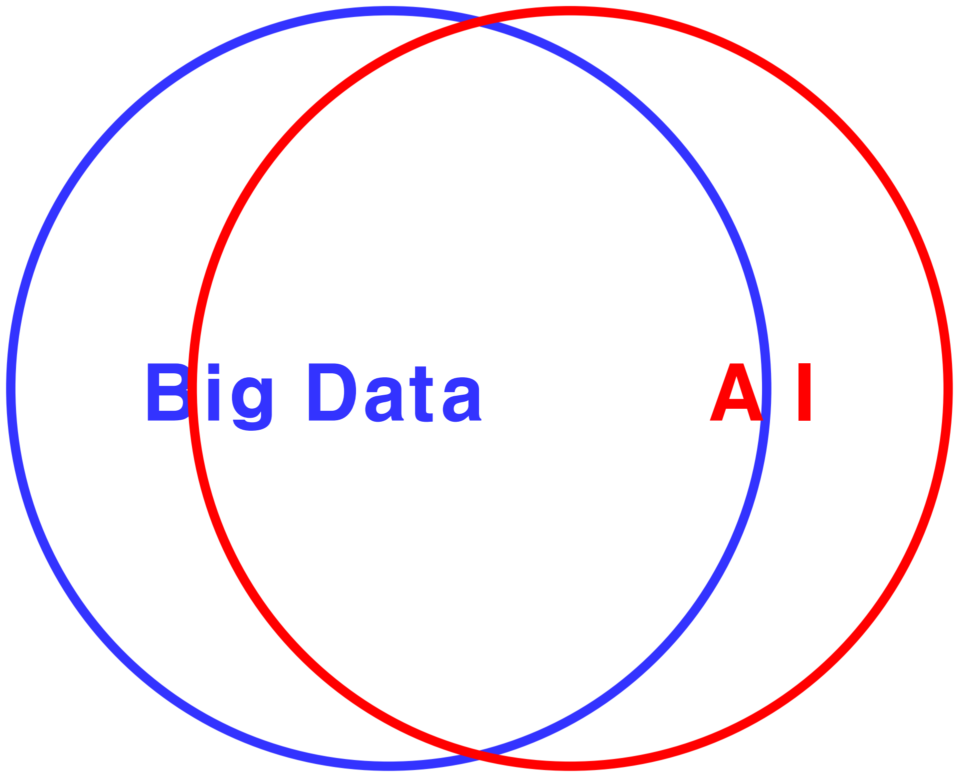
- # Sizes of data units

ZB/2yr/2배/40ZB(2020) 57*Sand

| Name | Equals to | | Size in bytes |
|---|---|---|---|
| Bit | 1 bit | | 1/8 |
| Nibble | 4 bits | | 1/2 |
| Byte | 8 bits | | 1 |
| Kilobyte | 1024 bytes | | 1024 |
| Megabyte | 1024 kilobytes | | 1,048,576 |
| Gigabyte | 1024 megabytes | | 1,073,741,824 |
| Terrabyte | 1024 gigabytes | | 1,099,511,627,776 |
| Petabyte | 1024 terrabytes | 24PB Google/Day/2012 | 1,125,899,906,842,624 |
| Exabyte | 1024 petabytes | | 1,152,921,504,606,846,976 |
| Zettabyte | 1024 exabytes | 1.2ZB/2%/2013 | 1,180,591,620,717,411,303,424 |
| Yottabyte | 1024 zettabytes | | 1,208,925,819,614,629,174,706,176 |

– **1TB (@ 80 Mb/s)**
  - **1 disk – 3.4 hr**
  - **100 disks – 2 min**
  - **1000 disks – 12 sec**

- **Big Data / AI**

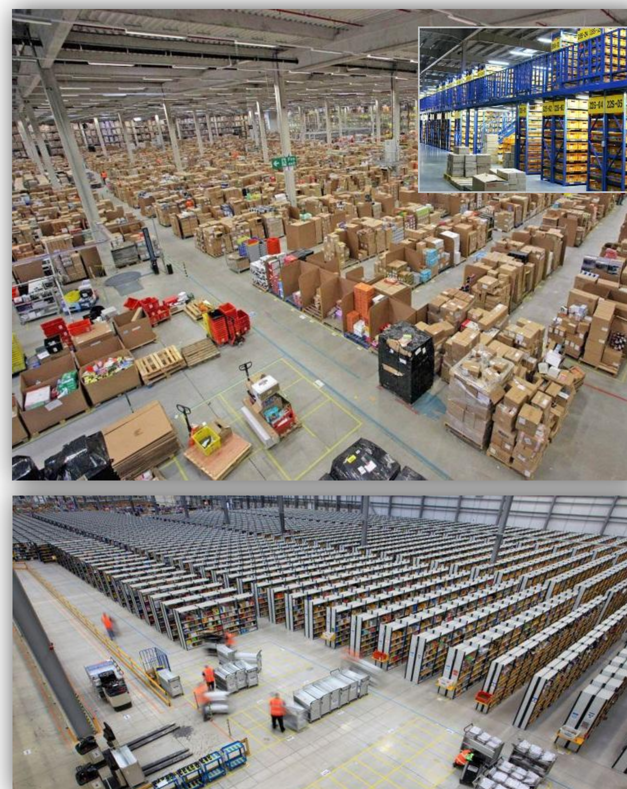


RDBMS → 1Terabyte/200$, Hadoop → 1Terabyte/5$

- RDB

Library

- Big Data DB

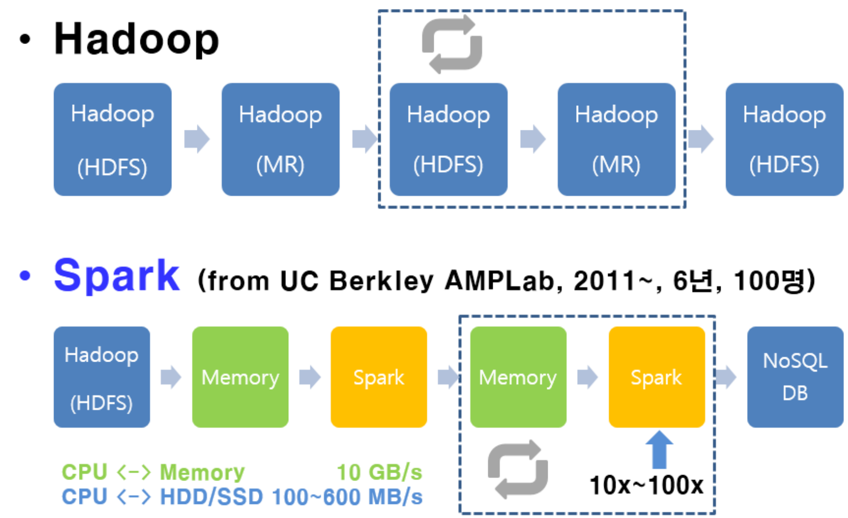


Logistics

- # Big Data Component (Hadoop)

Infomatica/Talend/InfoShere/DataStage/Apache.Hadoop/Spark/Avro/Ambari/Chukwa/Mahout/Tez/Yarn/Zookeeper/Pig/Hive/MapReduce/Hcatalog/Impalra/Kafka/Hbase/SINGA/Stinger/MapR/Aster/Splunk/Cassandra/Mongo/Riak/Vormic/Hana/Esper/Oracle/Tibco/R/SAS/SPSS/Tableau/Spotfire/D3/Qliktech/MicroStrategy/Exadata/Greenplum/Netizza/Radian6/SMA/CloudBigdata/Amazon/Softlayer/Rackspac/Cloudant/Dynamo/Teradata cloud/ …

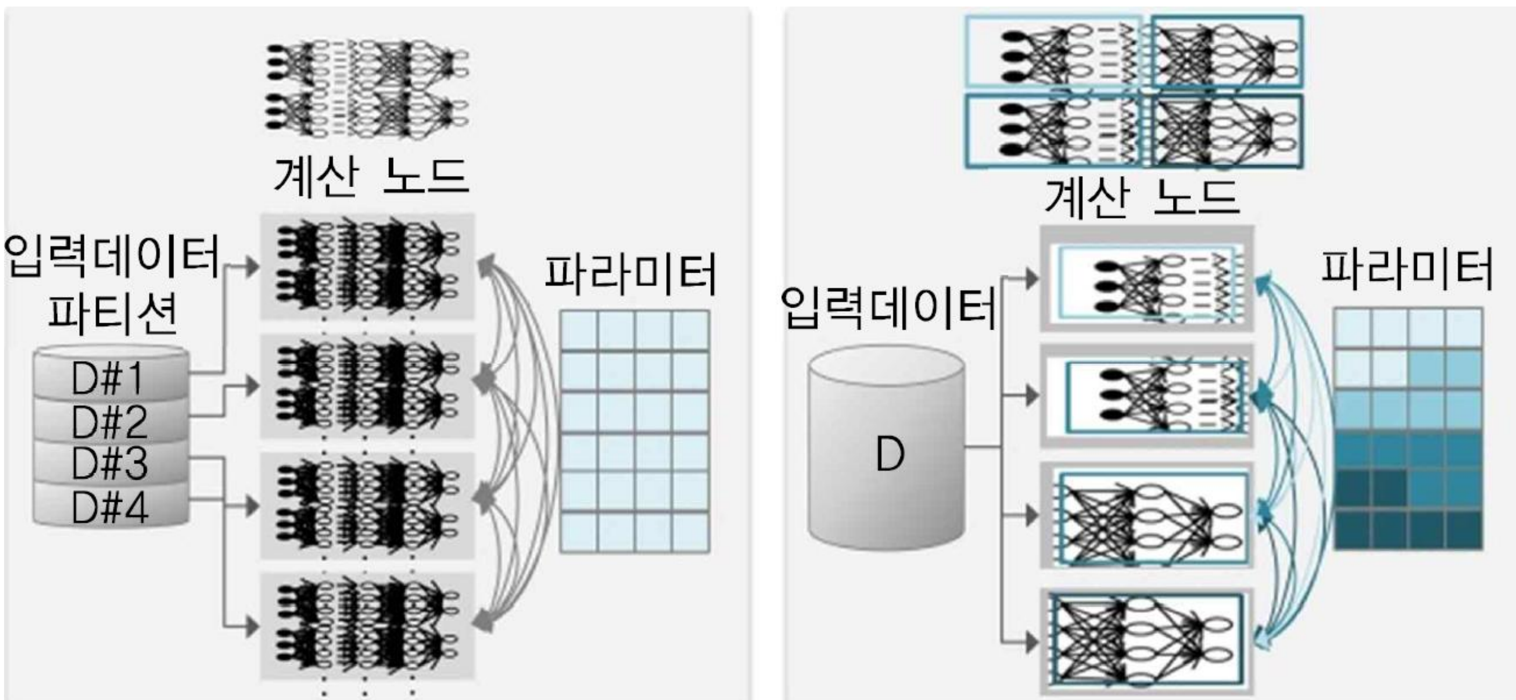
- Apache Spark
  - 2012.06 : V.0.5.1
  - …
  - 2014.05 : V.1.0.2
  - 2015.09 : V.1.5.0
  - 2015.10 : V.1.5.1
  - 2015.11 : V.1.5.2
  - 2016.01 : V.1.6.0
  - 2016.03 : V.1.6.1
  - 2016.07 : V.2.0.0
  - 2017.10 : V.2.1.2




Google TensorFlow ('15.11)

- **Distributed Computing**
  - **Data/Model parallelism**
  - **Full-mesh/Star topology**

ETRI E/T Trends
(안신영 et al. 2016)



- 1,200,000 Image
- 1,000 Classes

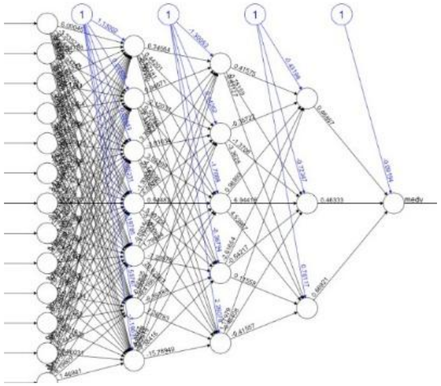
[Image Big Data : ImageNet DB]

**Trained *w, b***

**Image(x)**

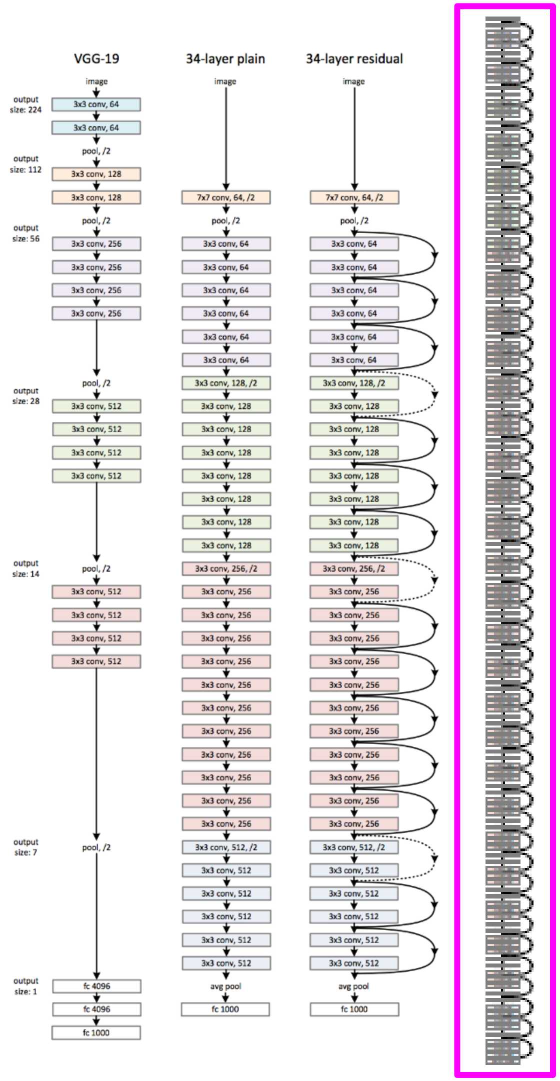
**Label(y)**

**Computer (Training)**



**NN Model**

# DNN Error w.r.t Hidden Layer




200 Layers

| DNN | Layer | Error | Year | Remark |
|---|---|---|---|---|
| - | - | 25.80% | 2011 | |
| AlexNet | 8 | 16.40% | 2012 | |
| VGG | 19 | 7.30% | 2014 | |
| GoogleNet | 22 | 6.70% | 2014 | |
| | 152 | 5.50% | 2015 | |
| **ResNet** | **200** | **4.80%** | **2016** | ◎ |
| | 1,000 | 4.62% | 2016 | |

## Benchmark DNN

- ResNet-200 (2016)
- ImageNet (1.2M, 1,000Class)

# Computing Time w.r.t GPU

| | | GPU | Petaflops (FP32) | Time | Remark |
|---|---|---|---|---|---|
| Case 1 | 2016 | 40 | 0.4 | 4.2 Days | |
| Case 2 | – | 2,000 | 20.0 | 2 hrs | |
| AIST-AIRC (J) | 2015 | 8 | 0.1 | 16.8 Days | 0.7 PF * 90(P) |
| | 2018 | 6,500 | 65~100 | 37 min | |

Tsubame 3(J)
HPE 540 Node
Tesla P100 : 2,160

- HPC with 8 GPU → 3 Weeks (ResNet 200 Layer)
- ImageNet 1.2M RGB Images with 1,000 classes

200,000 USD for 8-GPU HPC

## • BD/DL System : Open Source Platform

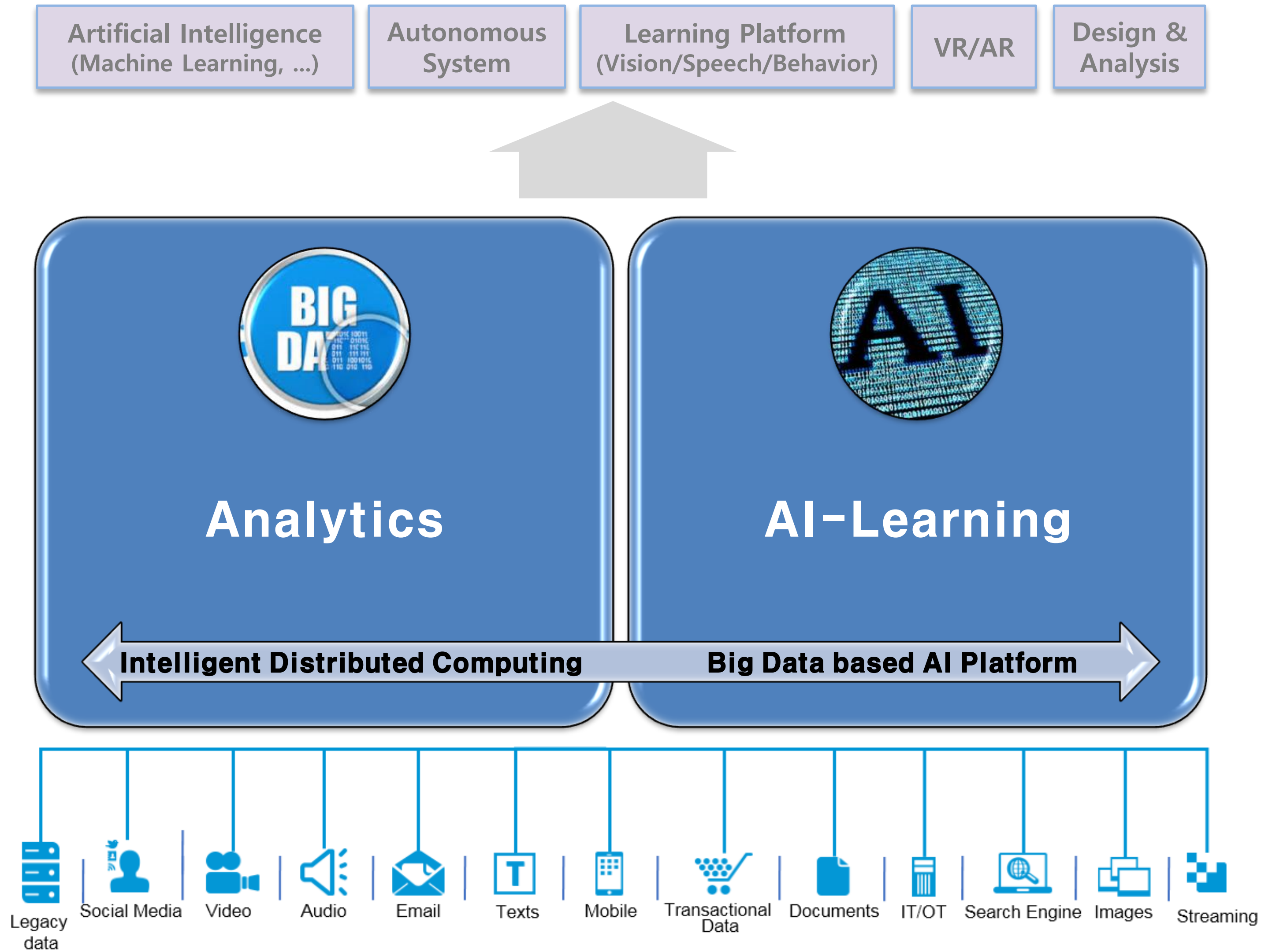
| Architecture | • Component |
| --- | --- |
| **Applications** | • Vision / Speech / Behavior / … |
| DL Framework | • Caffe, Torch, Theano, DL4J, CNTK, Pylearn2, TensorFlow, KERAS, … |
| Libraries(SDK) | • cuDNN, NCCL, IMSL Lib., GPU VSIPL, cuBLAS, cuSPARSE, cuFFT |
| System Software | • Apache Hadoop, Spark, Ambari, HBase, Yarn, Drivers, CUDA, … |
| **Hardware** | • Clustered Computer (Linux)<br>• X86, GP-GPU |

# Overview

- Two Things Considered
  - Labeled data
    - Few labeled data in defense
    - Classification/object detection/ segmentation/tracking/autonomy/…

    → Synthetic/virtual data generation
  - Computing power
    - ML/DL based on HPC & GPU
    - Big data hub  & data analytics

    → Distributed deep learning


Artificial Intelligence (Machine Learning, …)
Autonomous System
Learning Platform (Vision/Speech/Behavior)
VR/AR
Design & Analysis
BIG DATA
Analytics
AI
AI-Learning
Intelligent Distributed Computing
Big Data based AI Platform
Legacy data
Social Media
Video
Audio
Email
Texts
Mobile
Transactional Data
Documents
IT/OT
Search Engine
Images
Streaming