Teratec - Atelier5
Deep Learning & Algorithms

Benoit PELLETIER

Deep Learning: myths & realities

28-06-2017

Bull
atos technologies

Agenda

- Context
- Deep Learning introduction
- ATOS video analysis platform
- Challenges
- How HPC can help ?

Context

Many major breakthroughs in AI have occurred since 2011

AI Def: ***Creating machines that perform functions that require intelligence when performed by people (Kurzweil, 1990)***

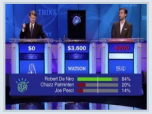
Before 2011

1946: Zuse's *Z3*, first programmable electronic computer

1997: IBM Deep Blue defeats world's chess champion Kasparov

2005: Honda's humanoid robot *Asimo* comes to life

2011 – 2016:

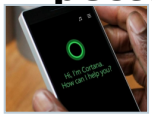


2011: Watson wins *Jeopardy!* against most successful contestants

2014: Alexa, Amazon's intelligent assistant debuts

2016: AlphaGo beats Lee Sedol in a Go match

Expected by 2030+

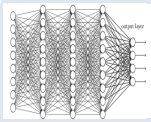


~2020: All-over virtual personal assistants as interface for consumers

~2030: Fully autonomous driving cars become market-ready

20xx: Robots may build robot "children" on their own

Major breakthroughs



Algorithmic advances in **deep learning**

Increasing **computing power**

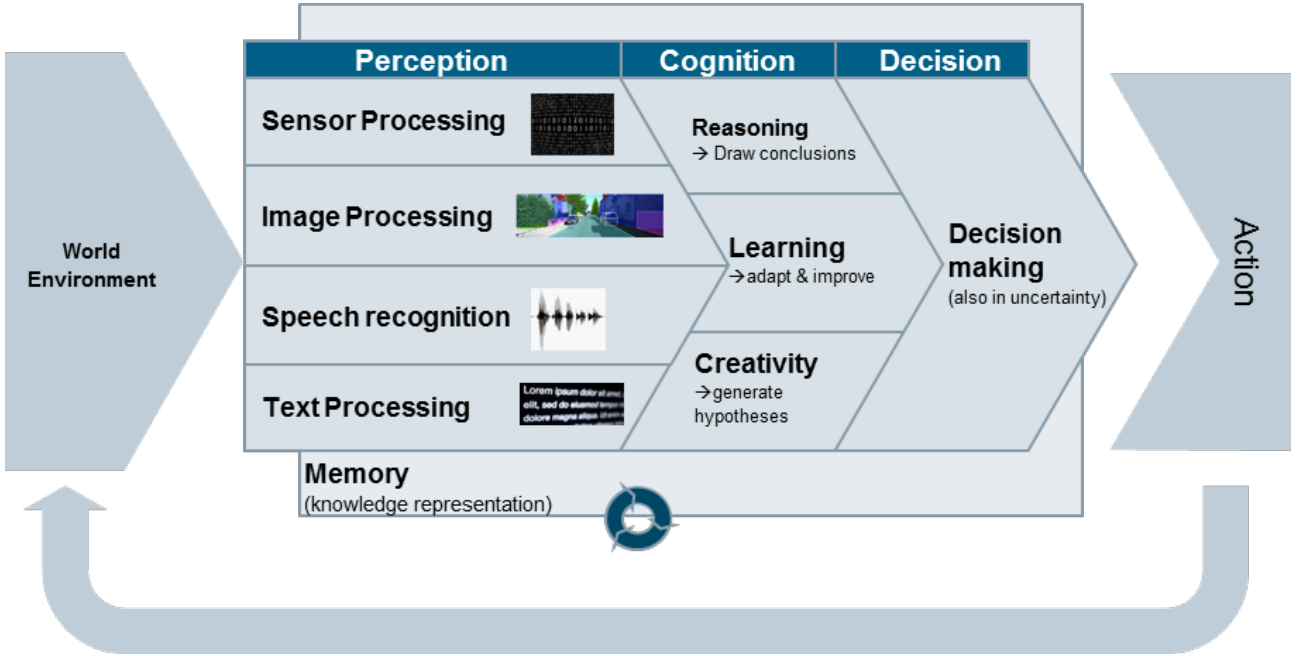
Usage of huge **datasets** leverage full potential of AI

Open platforms and data bases

Gartner – The Arrival of Algorithmic Business
Deloitte – Intelligent automatization entering the business world

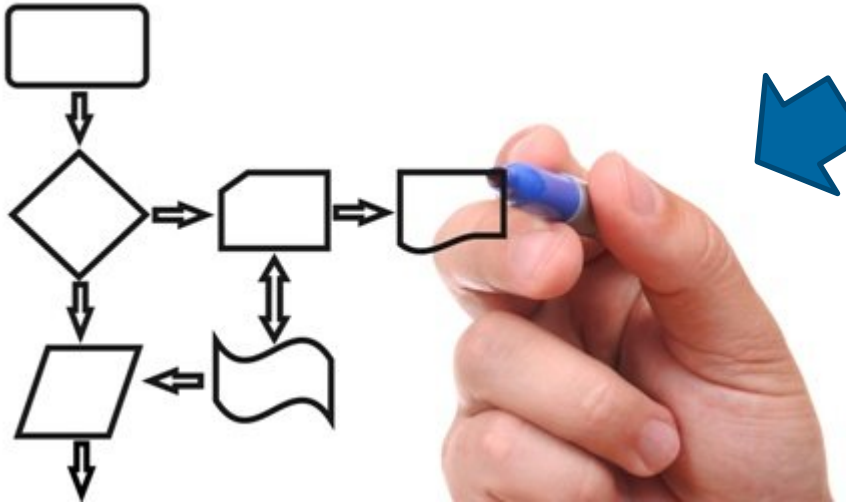
2nd machine age & new industrial revolution

AI Framework



Deep Learning

Traditional software



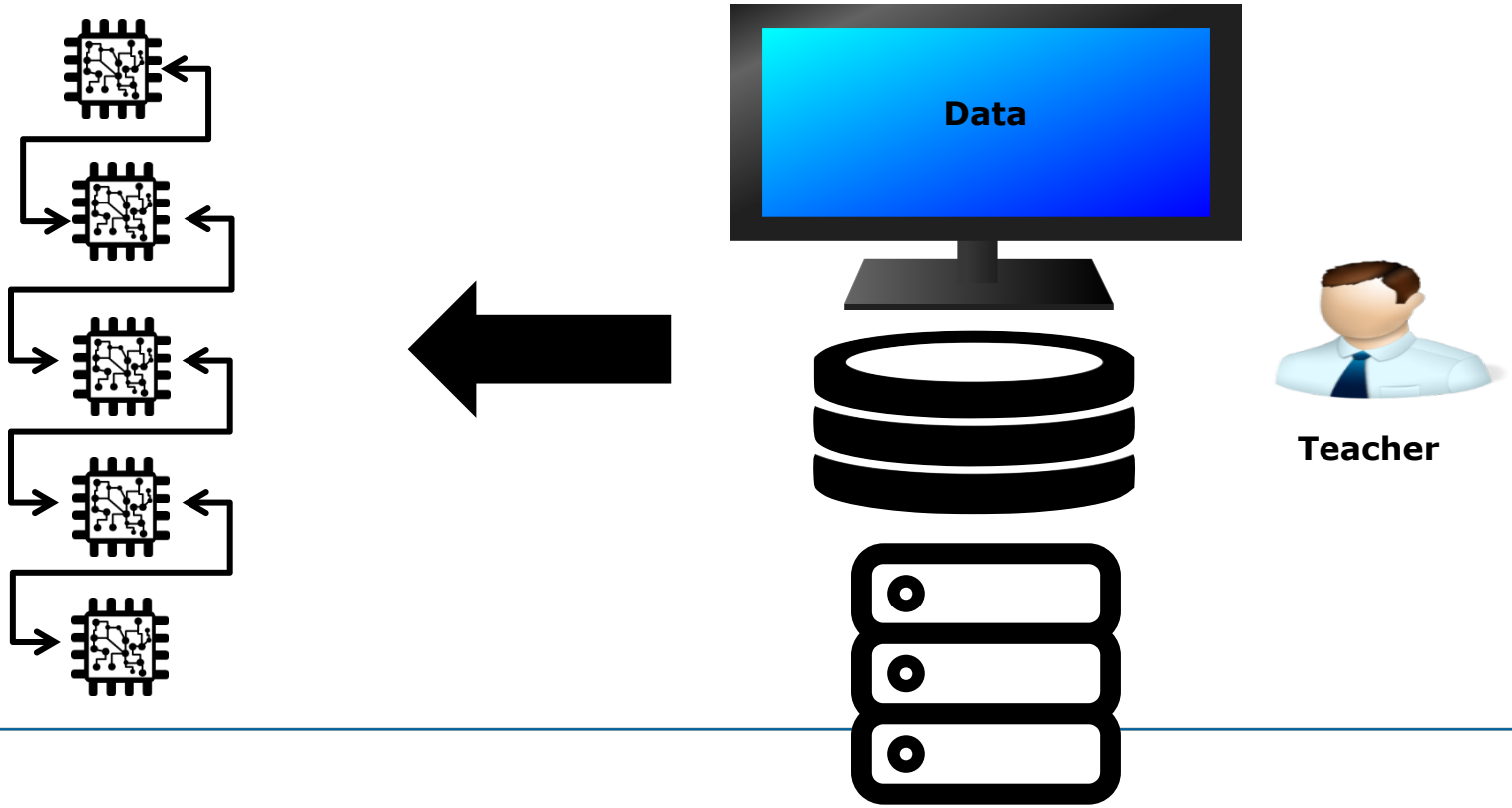
developers

Image Credit : https://www.enisa.europa.eu

Software = mathematical science

Bull
atos technologies

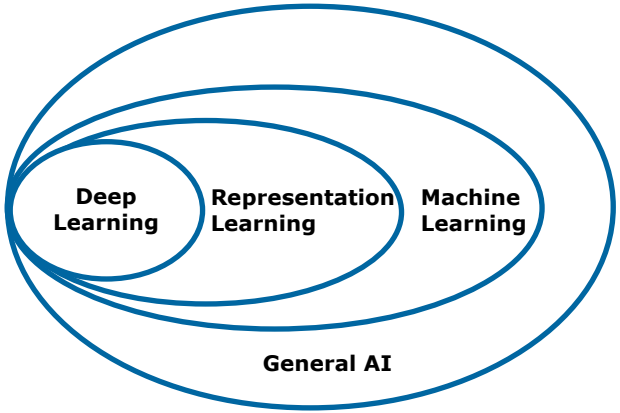
Machine learning



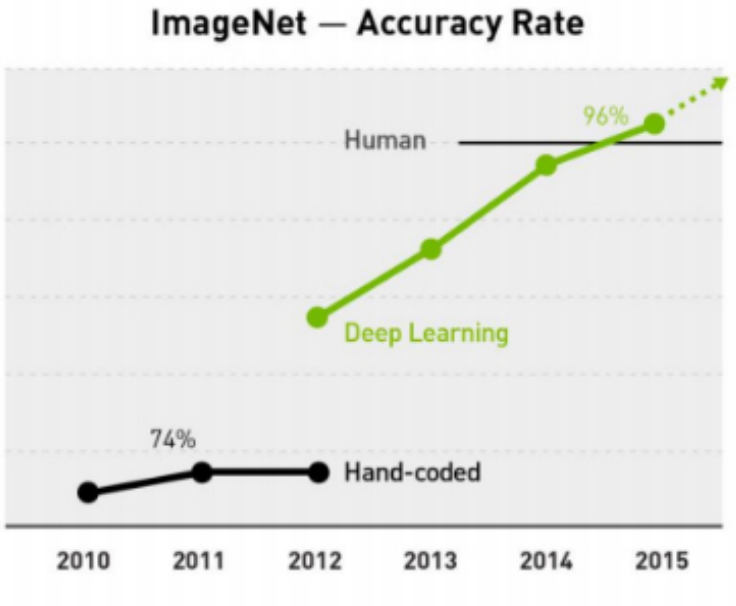
Deep learning

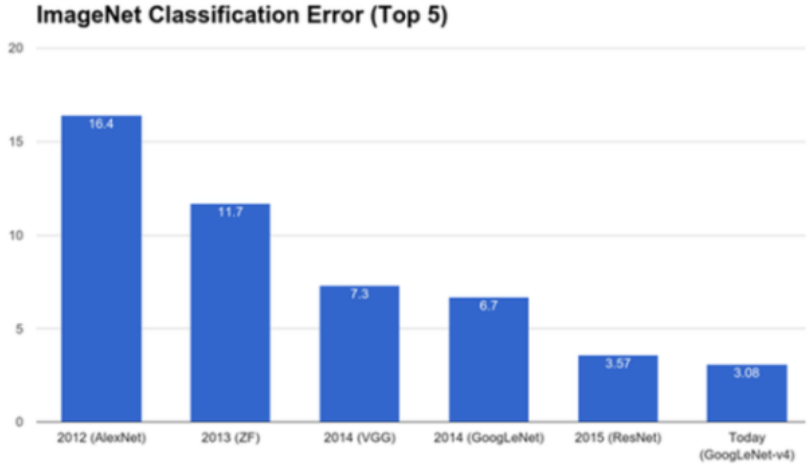
What is DL?

- A **machine learning** technique
- Improved with **experience** and **data**
- Representing the world as a **nested hierarchy** of concepts
- Great **feature representations** capacity
- Can be combined with **supervised**, **unsupervised** or **reinforcement** learning problem…



Deep Learning is the new paradigm!





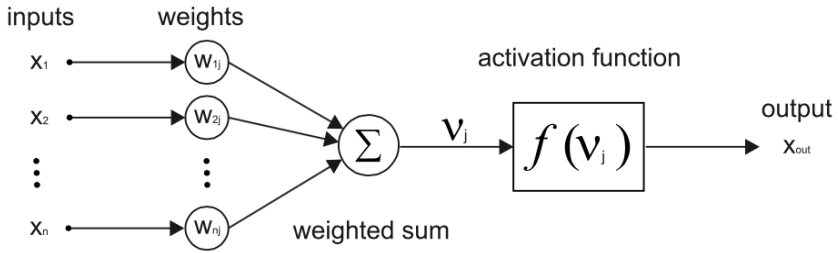
GoogleNet (2014): 22 layers
ResNet (2015) : 152 layers

source: https://www.slideshare.net/MariaChapovalova/tackling-challenges-in-computer-vision-67459618
https://blog.altoros.com/mastering-game-development-with-deep-reinforcement-learning-and-gpus.html

What is a neural network?

How can a model be represented

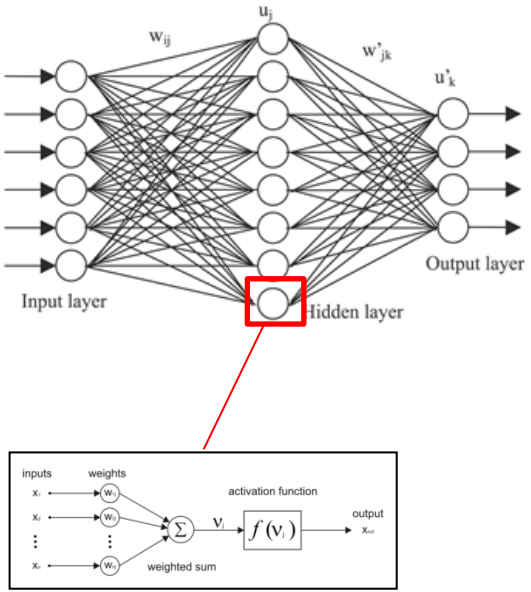
1957: the "Perceptron"



During the learning process, the weights are optimized.

$$X_{out} = f\left(\sum_{i=1}^{n} w_{ij} \times x_i\right)$$

1984-1986: the "Multilayer perceptron"

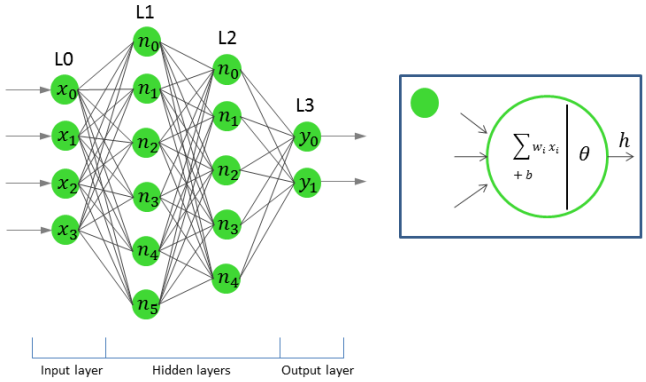


Bull
atos technologies

Many research activities result in various NN architectures

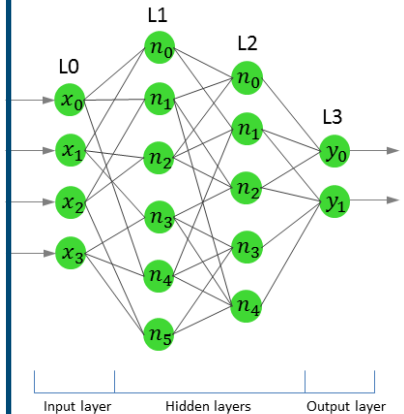
- ✓ Simple classification task
- ✓ Simple regression Problem

Feedforward Neural Network



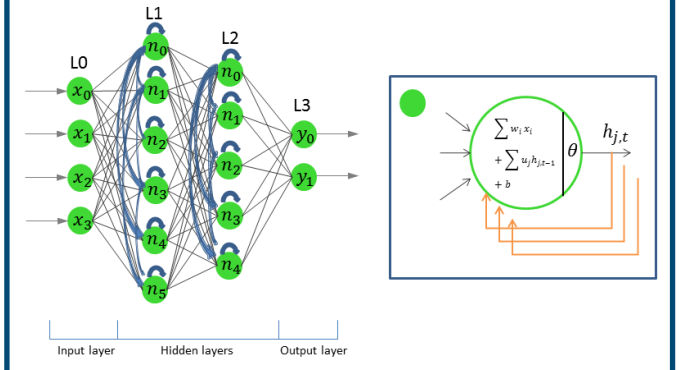
- ✓ Image processing
- ✓ Speech recognition
- ✓ Video recognition
- ✓ Hyperspectral image analysis

Convolutional Neural Network



- ✓ Language prediction, translation
- ✓ Time series forecasting
- ✓ Sequential data
- ✓ Video processing

Recurrent Neural Network

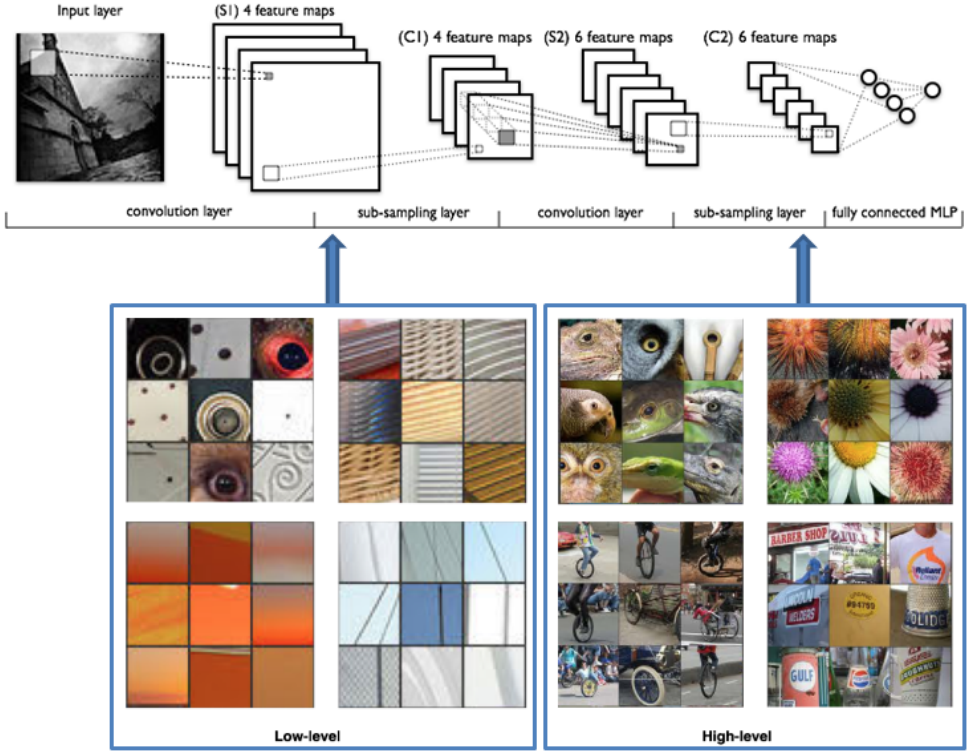


Bull
atos technologies

DL Representation capacity

CNN architecture

- In the special case of CNN, the network weights $\{w_i\}$ can be interpreted as sets of filters
- There are several filters in each layer
- From first layer to final layer: weights represent more general features to more objet-specific features



Large-scale Deep Learning Problem

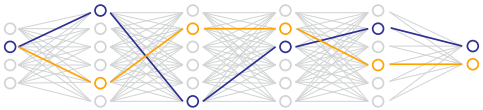
New requirements, but different from training to running

▶ Phase 1 : Model Training

(Data Center – heavy computation cost)

Input data

Creation of multi-layer Neural networks



Production of a trained model

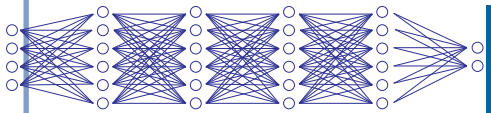
Need a lot of data to train the model; if possible, data labelled by human could be better

Trained model

▶ Phase 2 : Inference

(Local applications or Data Center – Almost Instantaneous)

Different sensors' data, such as camera



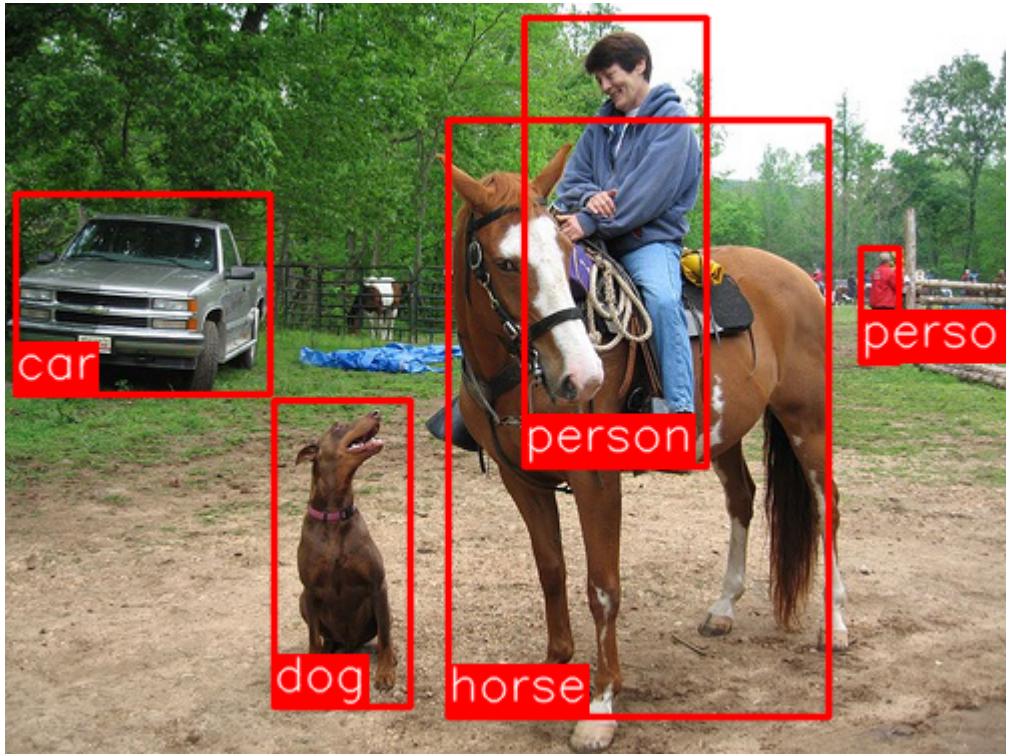
Application of trained model

Pedestrian recognition achieves an accuracy of 97%

Class prediction of new data

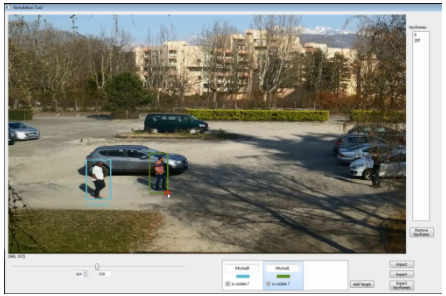
Bull
atos technologies

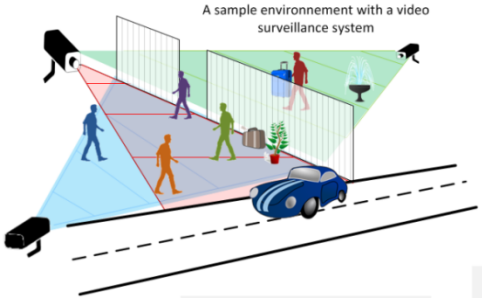
Example of DL: Region localization (RCNN)

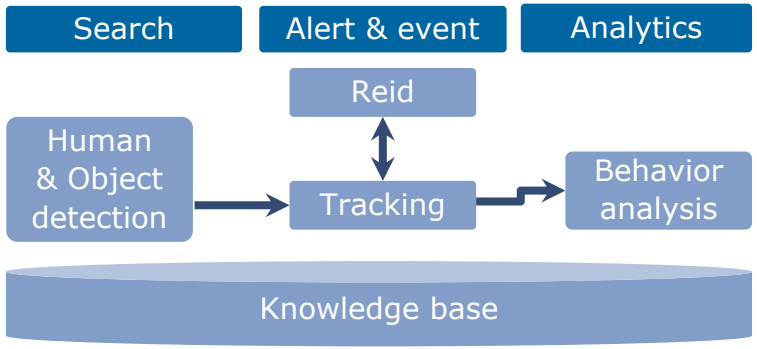


ATOS video analysis platform

Video protection & digital signage use case







- Feature extraction: person, face, moods, clothes, vehicle, bag, gun, behaviors, …
- Video security: crowd movement, scenes of violence, abandoned objects, search a person of interest, search a car plate, …
- Digital signage: optimize commercial spaces, dynamic advertisement, passenger traffic flows well, …
- Tensorflow based

Demo

Bull
atos technologies

Practical example: people re-identification

Sign1

Sign2

Sign3

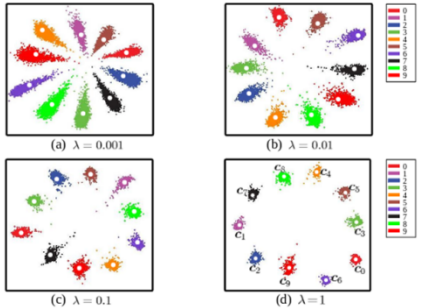
$D(x, y)$ distance
t: threshold

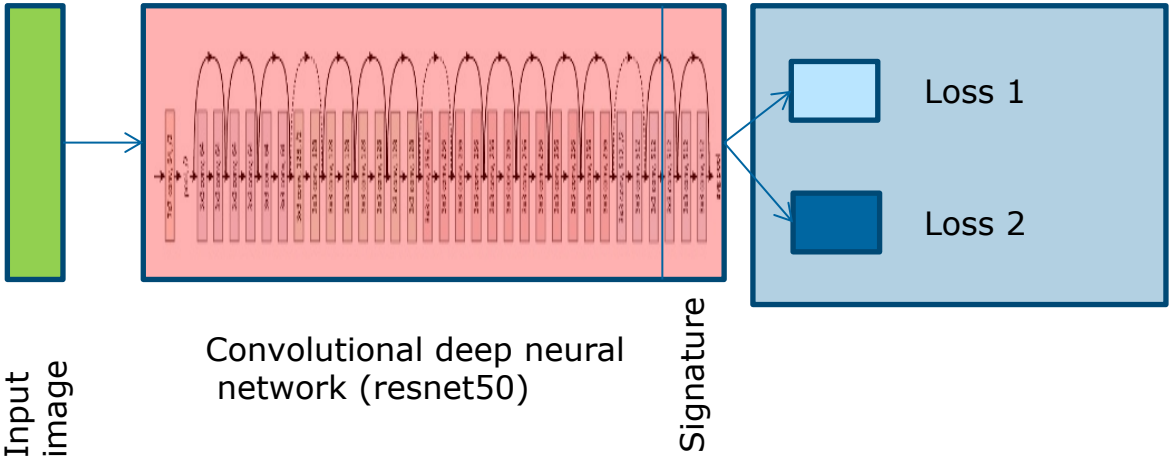
$$D(Sign1, Sign2) \ll t$$
$$and$$
$$D(Sign1, Sign3) \gg t$$
$$and$$
$$D(Sign2, Sign3) \gg t$$

→ Be able to re-identify objects on one video or with multiple cameras

Implementation done in Atos R&D

- Deep neural network for signature extraction
- We implemented two losses: 1 to separate classes, 1 to minimize the intra class variations
- Signature is an embedding of *n* dimensions
- Fine-tuning

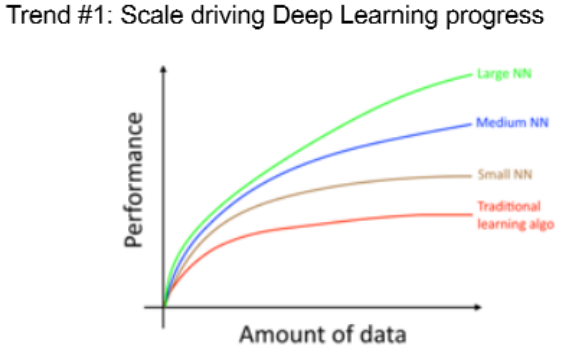




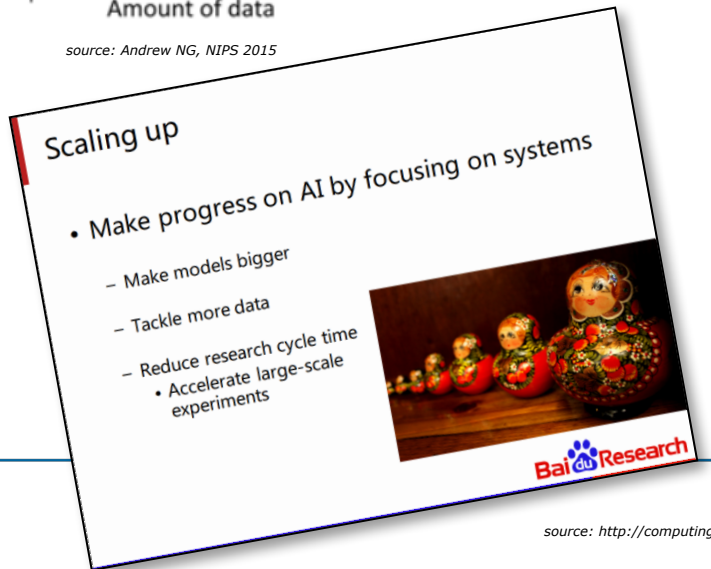
Bull atos technologies

Challenges

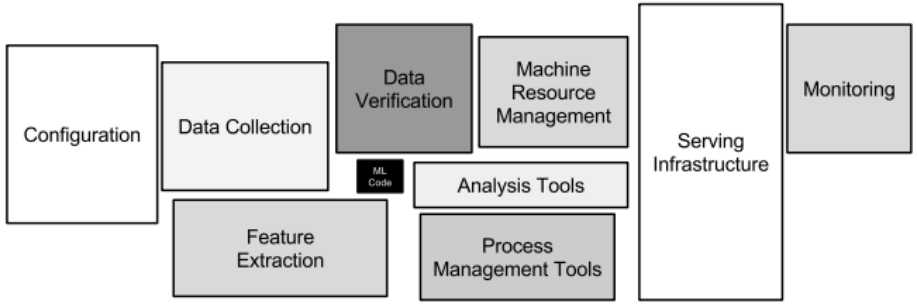
The overall complexity is increasing



source: Andrew NG, NIPS 2015



Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex

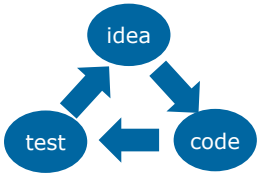


source: Google paper https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

source: http://computing.ornl.gov/workshops/SMC15/docs/bcatanzaro_smcc.pdf

Iterative & empirical process

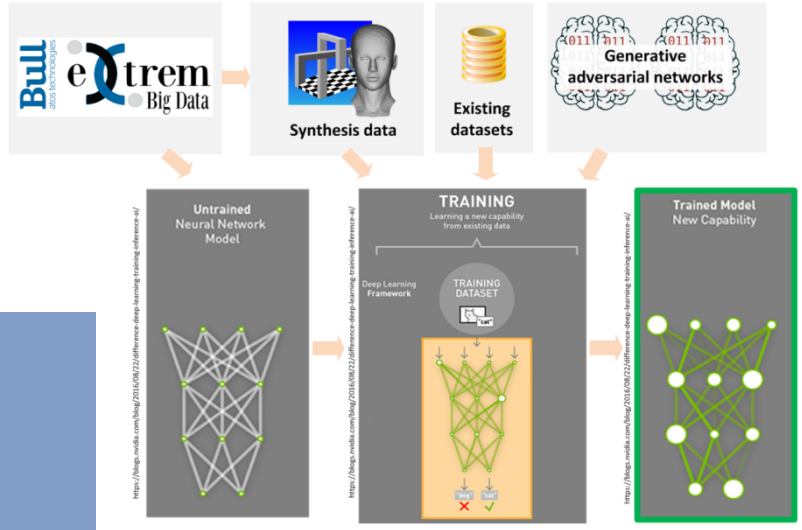
- Our Re identification system
  - Training time: **up to 18 hours** on P100 GPU
  - Many hyperparameters
    - lasts NN layers
    - embedding dimension
    - data augmentation
    - learning rates & optimizers
  - Machine Learning « programmers » design the network structure with experience and by trial and error



Many iterations are required to find out the best architecture & model for our task!
Training time is critical for development productivity!

Dealing with small data, get accurate data

- Real data with good quality are rarely available
- Deep Learning technics might reach or even excel human-level performance in recognition given that the model training is done with a lot of data
- Data is AI bottleneck. Lack of data is slowing down its expansion



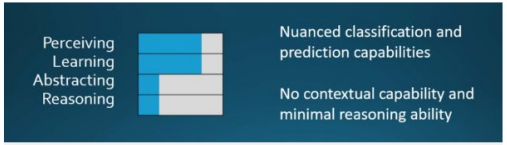
Needs to create our own datasets

1) *do it internally with your datascientists (however they'll want to quit…)*
2) *use amazon Mechanical Turk (however time-consuming, poor quality, …)*
3) *use Active Learning to select the most informative image to build AI with as few images as possible*
4) *develop tooling for generating data*

Bull
atos technologies

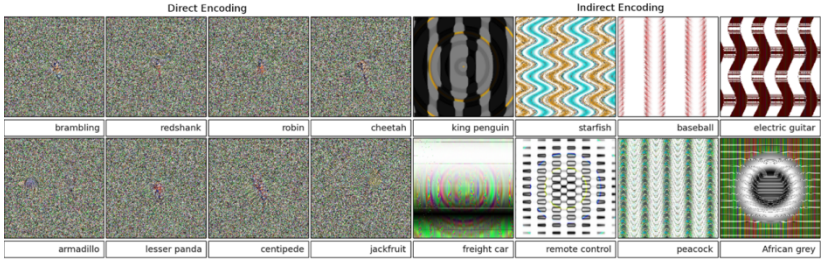
Weak AI, model uncertainty

out of distribution data (not trained for), noisy data, etc

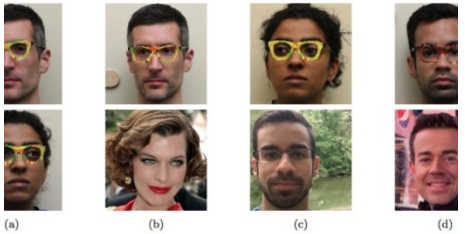


Characteristics of the second wave of AI technology.

"A young boy is holding a baseball bat." Second wave systems are statistically impressive, but individually unreliable.



Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with >= 99.6% certainty to be a familiar object



Eyeglasses to fool State-of-the-art face recognition system
(b) impersonating Milla Jovovich
(d) impersonating Carson Daly

Inherent flaws in second wave systems can be exploited.

99% certainty, that's not a panda, that's a gibbon

Source: https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf
source: https://www.technologyreview.com/s/603803/facebooks-ai-chief-machines-could-learn-common-sense-from-video/
https://machinelearning.technicacuriosa.com/2017/03/19/a-darpa-perspective-on-artificial-intelligence/
http://www.evolvingai.org/fooling

Bull
atos technologies

Getting the common sense (or getting the ability to fill in the blanks)

- Infer the state of the world from partial information
- Infer the future from past & present
- Infer past events from the present

- "The trophy doesn't fit in the suitcase because it's too large/small"
  - (winograd schema)

- Filling in the visual field at the retinal blind spot
- Filling in occluded images, Fill the blanks
- Fillling in missing segments in text, missing words in speech.
- Predicting the consequences of our actions
- Predicting the sequence of actions leading to a result

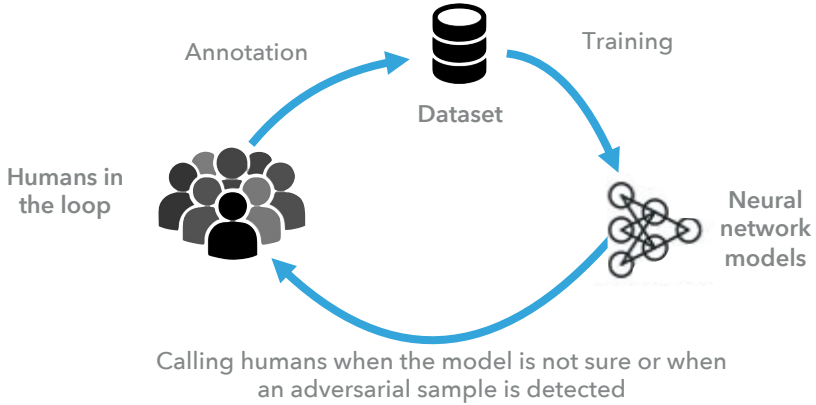
- "Tom picked up his bag and left the room"

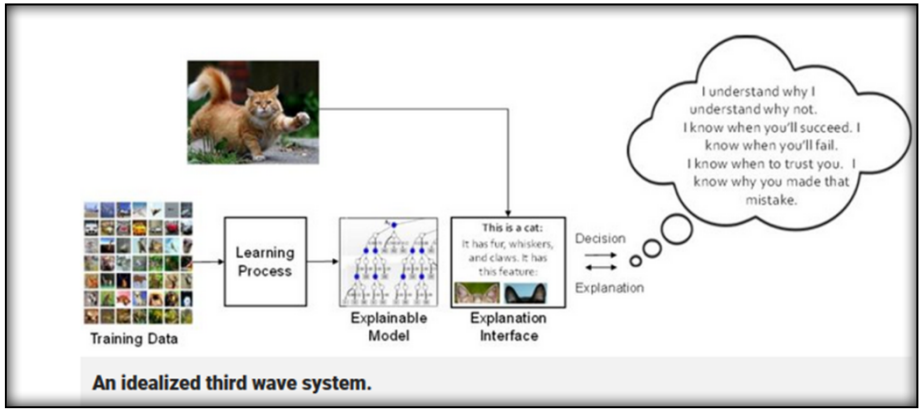
Long term challenge

- Predicting any part of the past, present or future percepts from whatever information is available.

source: Y.Lecun, https://drive.google.com/fil/d/0BxKBnD5y2M8NREZod0tVdW5FLTQ/view

Improve robustness at short term …

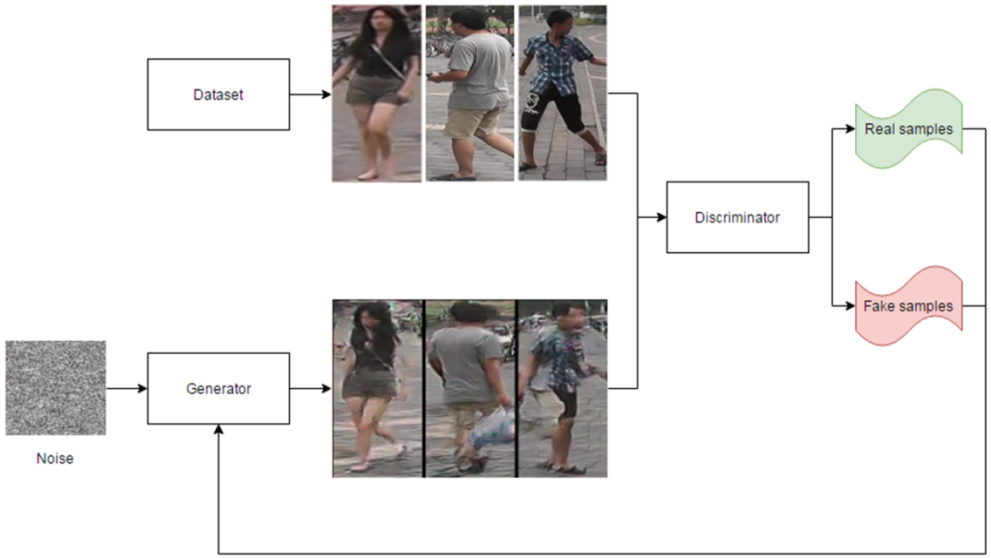




or complete the training dataset!

https://arxiv.org/pdf/1612.07767.pdf

Y.Lecun: « GAN is the most interesting idea in ML of the last 10 years »



Realistic samples

Enrich dataset, augment resolution, …
A lot of applications with GAN, but

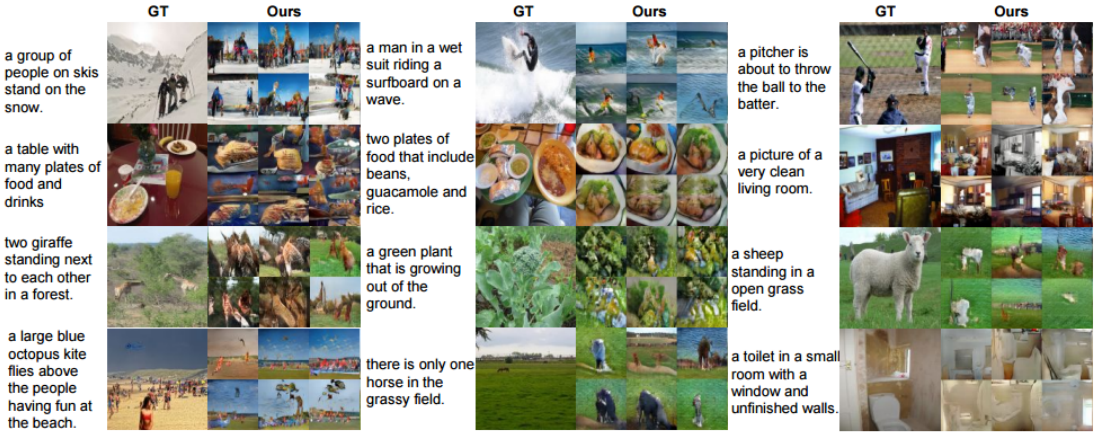
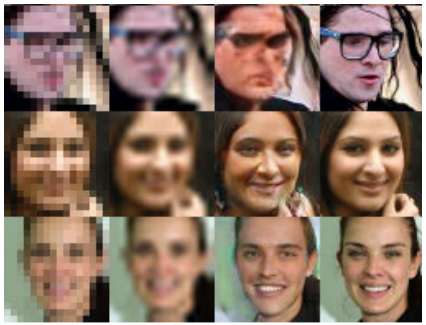


Figure 7. Generating images of general concepts using our GAN-CLS on the MS-COCO validation set. Unlike the case of CUB and Oxford-102, the network must (try to) handle multiple objects and diverse backgrounds.



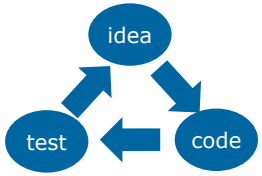
How to control the « creativity » of the network?

source: https://www.researchgate.net/publication/303337197_Generative_Adversarial_Text_to_Image_Synthesis
https://github.com/david-gpu/srez

How HPC can help?

Deep Learning training is HP

- Turnaround is key

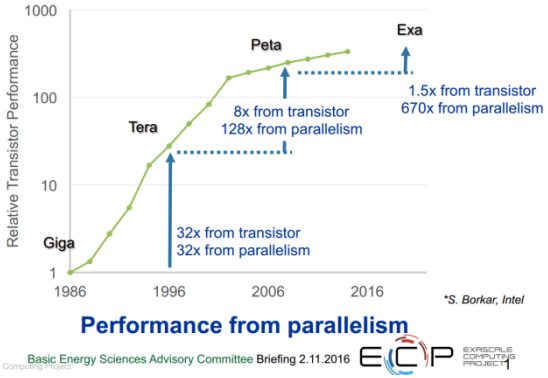


- Scalability issues
  - Compute intensive (GEMM)
  - Data intensive (models & datasets)

- Use most efficient hardware
- Parallel, hetererogenous computing
- Many nodes with fast interconnect

All of that is standard HPC

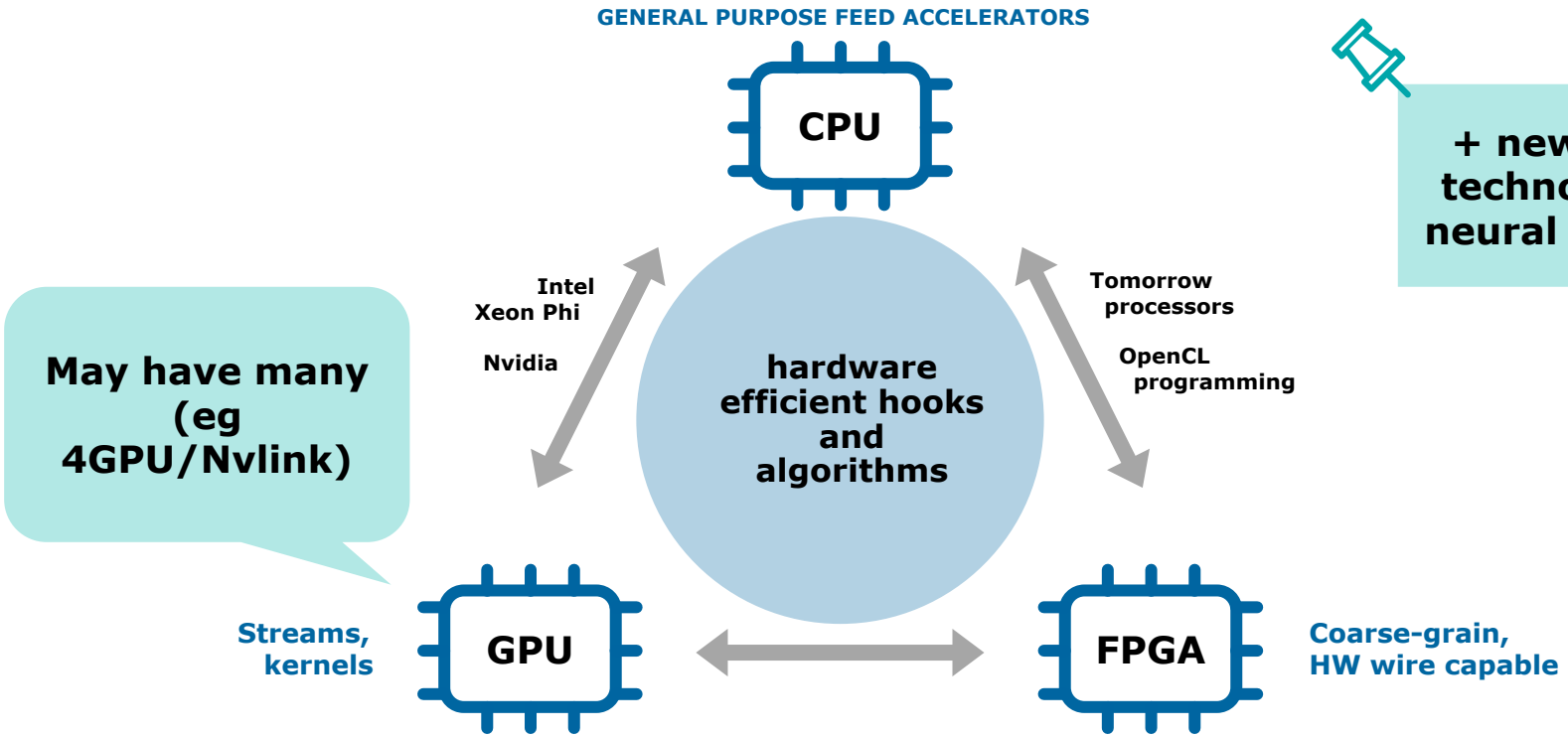
From Giga to Exa, via Tera & Peta*



How to accelerate Deep Learning ?

Compression	Scheduling	Hardware
Reduce precision, pruning	Distribution, async communication, collective	*Co processor unit, High performance network High performance storage*

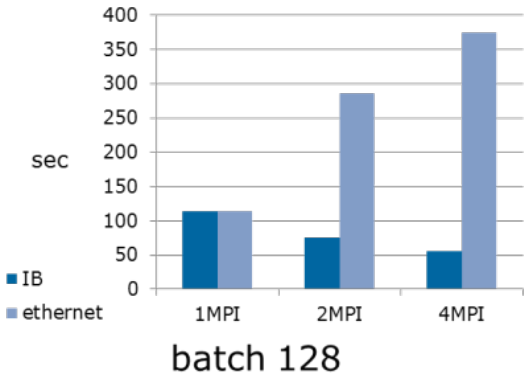
Scale up, select the best compute

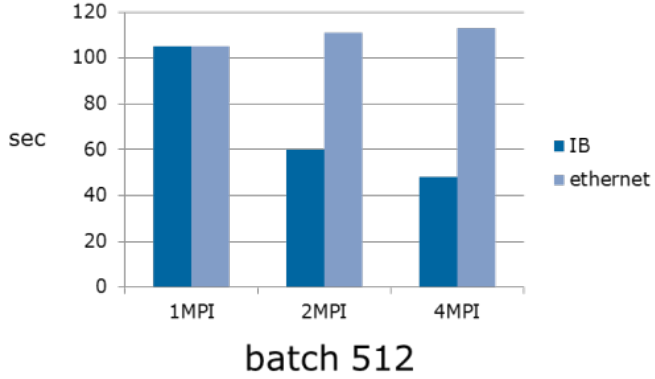


Bull
atos technologies

Scale out, distribute the training

- Network latency impacts the training perf in a distribution mode

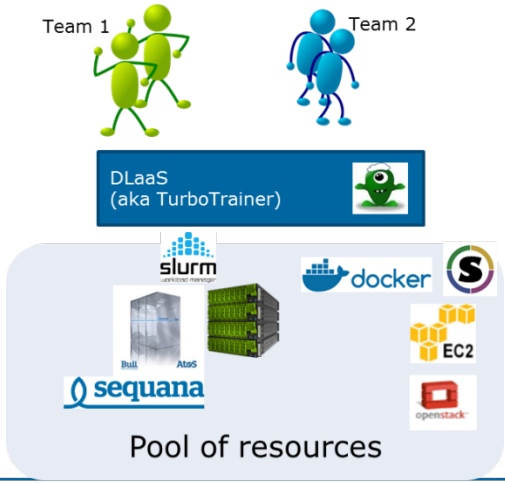




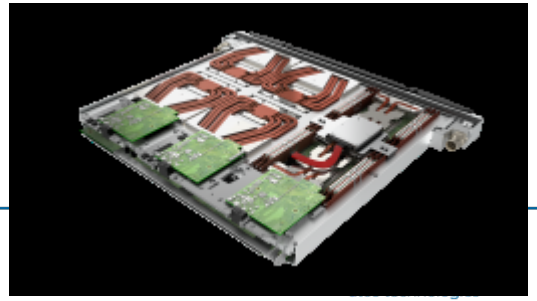
IB vs Ethernet with MPI and CNTK

Enable parallelism

- Having many computing resources
  → one user can launch several experiments in parallel on different nodes to evaluate different hyper-parameters
  → several teams can simultaneously work on different problems



Bull Ystia



Fast data access

- Many images to load
  - Speed is crucial to feed computing resources
  - Data can be loaded in parallel with TensorFlow

We observed that

- a Lustre filesystem (or/and SSD) + interconnect can be efficient

ATOS BDS technologies
Comprehensive solution for AI

Bull Ystia

Bull atos technologies

AI use cases

DLaaS
Turbo trainer

Deep Learning & Computer Vision

TensorFlow

OpenCV

DL Library for embedded

Dataset mngt

blender

GAN-based tools

Real data Annotations tools

HPDA platform

Janus Orchestrator

slurm Workload manager

HPC

SCS

BXI

Bull atos technologies

intel

Atos

New Bullion

NVIDIA

New embedded server for AI

FPGA

GPU

Neural proc

Bull atos technologies

Thanks

For more information please contact:
T+ 33 4 76297270
F+ 33 4 76297607
M+ 33 6 80357914
eric.monchalin@atos.net

Bull
atos technologies