인공지능 기술 현황

2018. 1. 15.
유성준
세종대학교 인공지능-빅데이터연구센터 센터장
세종대학교 컴퓨터공학과 교수
sjyoo@sejong.ac.kr

ABRC
인공지능-빅데이터연구센터
http://www.abrc.or.kr

세종대학교
SEJONG UNIVERSITY

발표순서

I 인공지능의 역사, 한계, 비전

II 머신러닝 개념 소개

III 딥러닝

IV 딥러닝을 위한 컴퓨팅 구조

V 결어

I 인공지능의 역사, 한계, 비전

인공지능은 만능?

False Promises - 1958

NEW NAVY DEVICE LEARNS BY DOING; Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

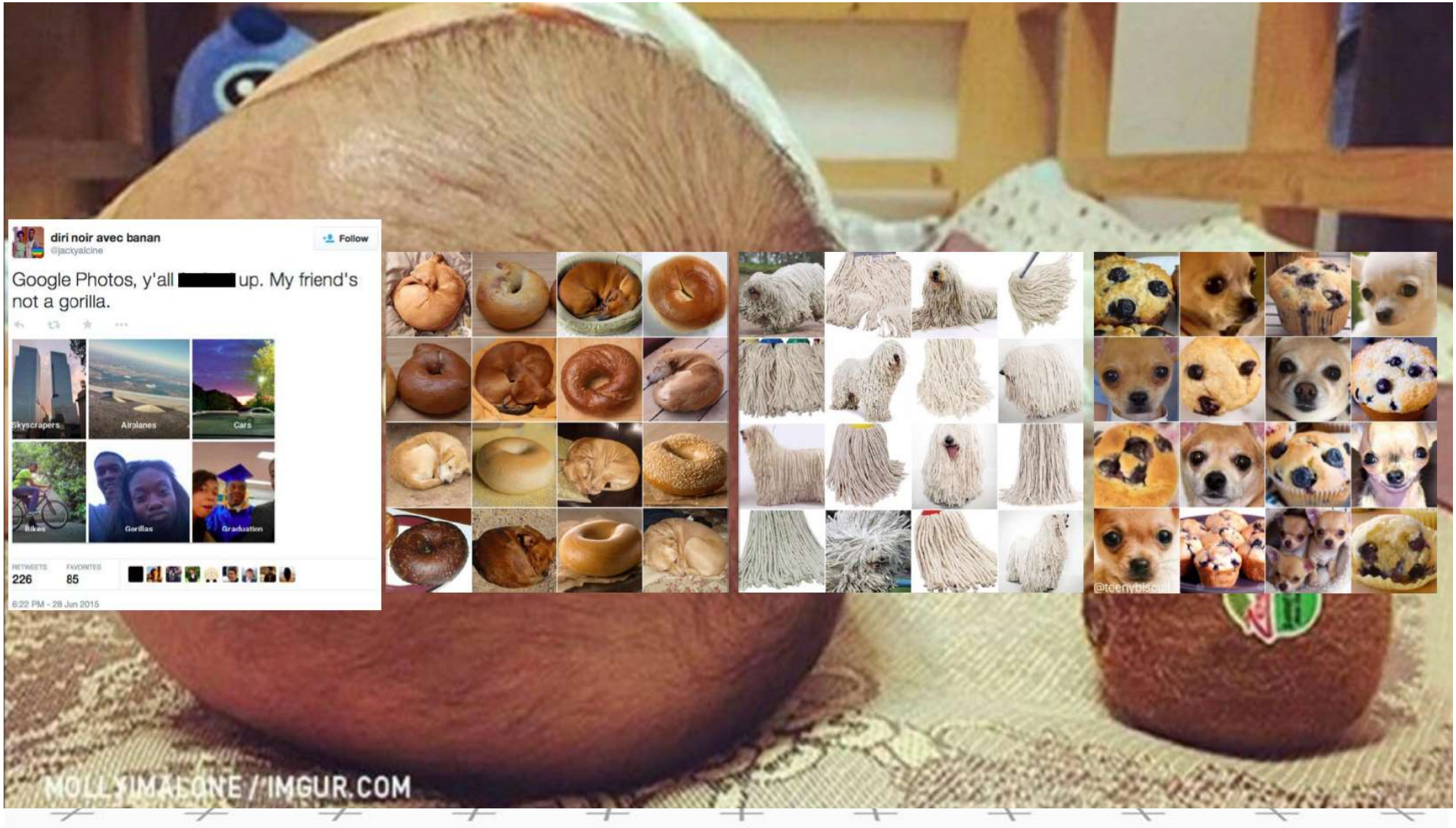
SPECIAL TO THE NEW YORK TIMES JULY 8, 1958

WASHINGTON, July 7 (UPI) -- The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

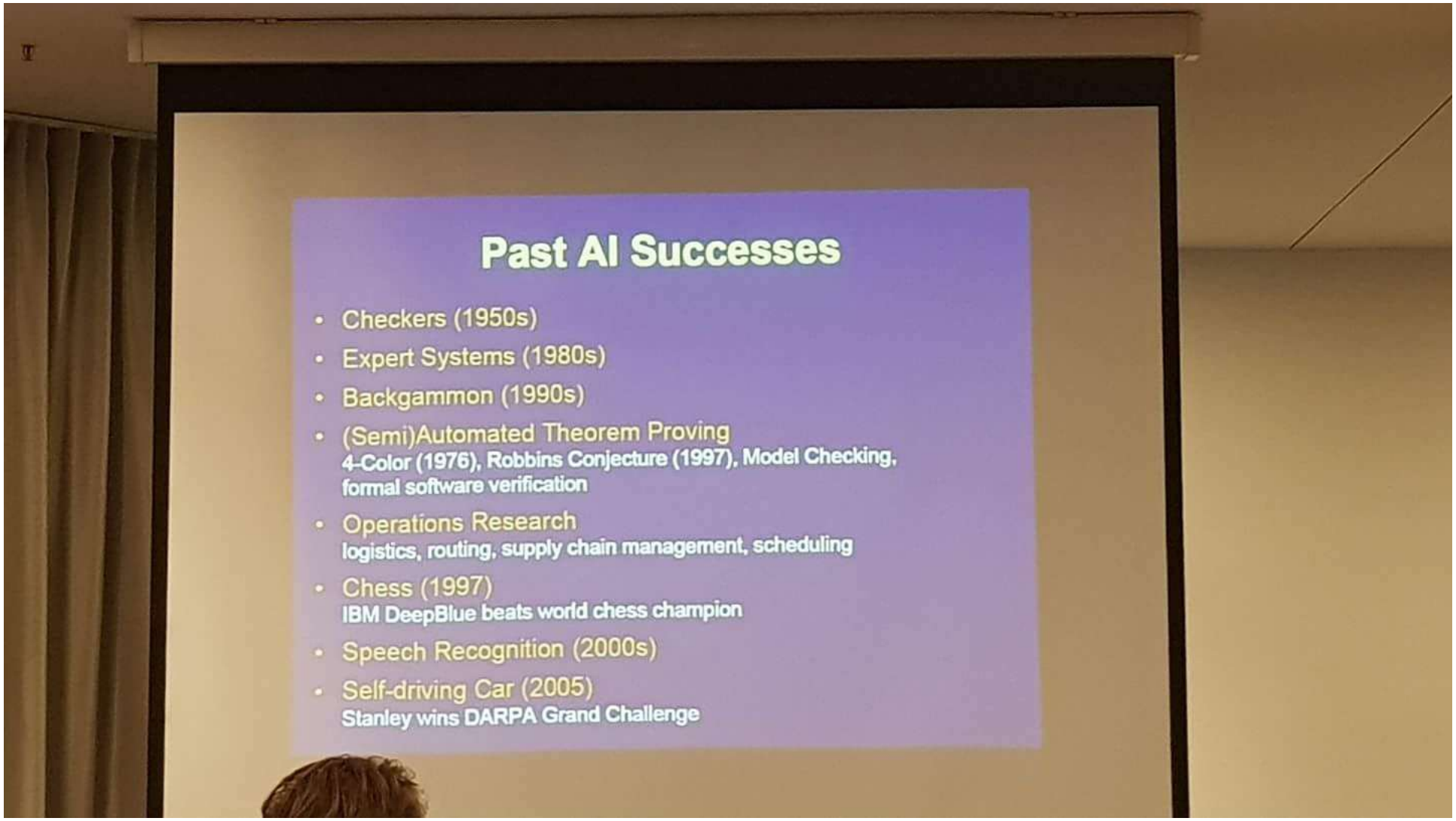
"The Navy revealed the embryo of an electronic computer today that it expects will be able to **walk, talk, see, write, reproduce itself** an be conscious of its existence … **Dr. Frank Rosenblatt**, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said **Perceptrons** might be fired to the planets as mechanical space explorers" July 08, 1958

<출처> http://www.nytimes.com/1958/07/08/archives/new-navy-device-learns-by-doing-psychologist-shows-embryo-of.html

인공지능의 현수준

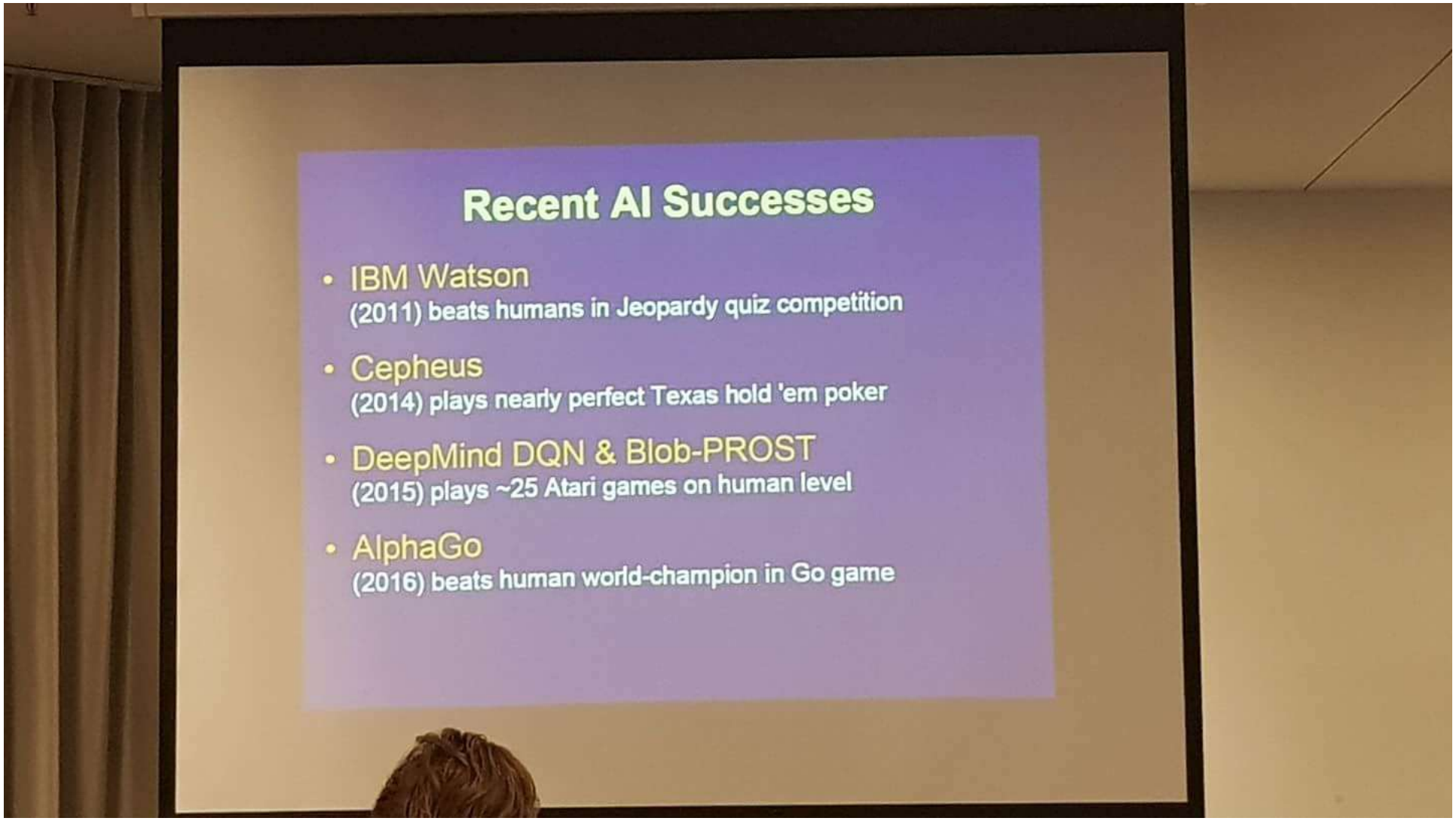

diri noir avec banan
@jackyalcine
Follow
Google Photos, y'all [redacted] up. My friend's not a gorilla.
Skyscrapers
Airplanes
Cars
Bikes
Gorillas
Graduation
RETWEETS
226
FAVORITES
85
6:22 PM - 28 Jun 2015
MOLLYMALONE / IMGUR.COM

과거의 인공지능 응용 시스템



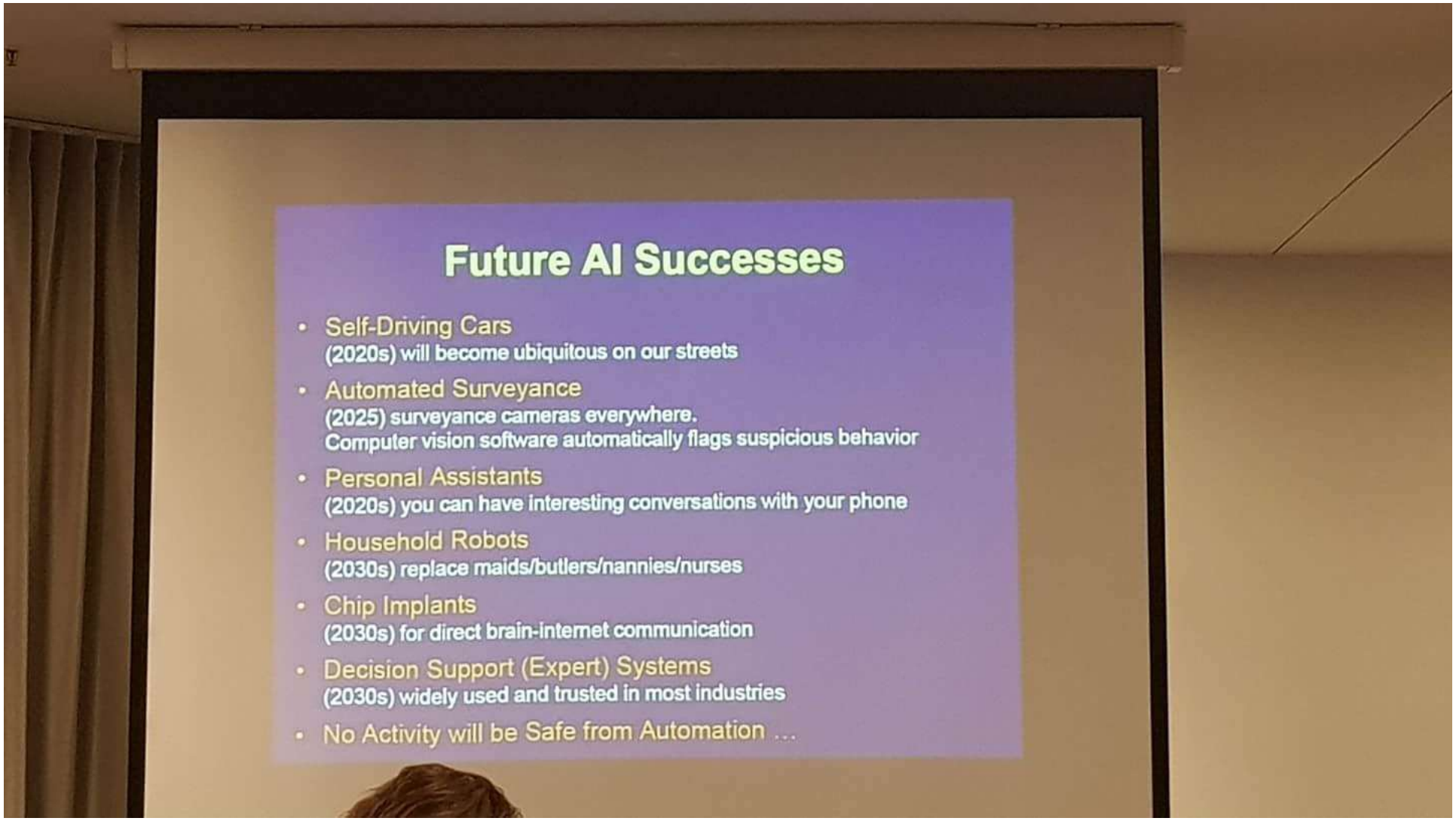
<출처> Marcus Hutter, AGI 2017-정지훈 Facebook 사진첩

최근의 인공지능 응용 시스템



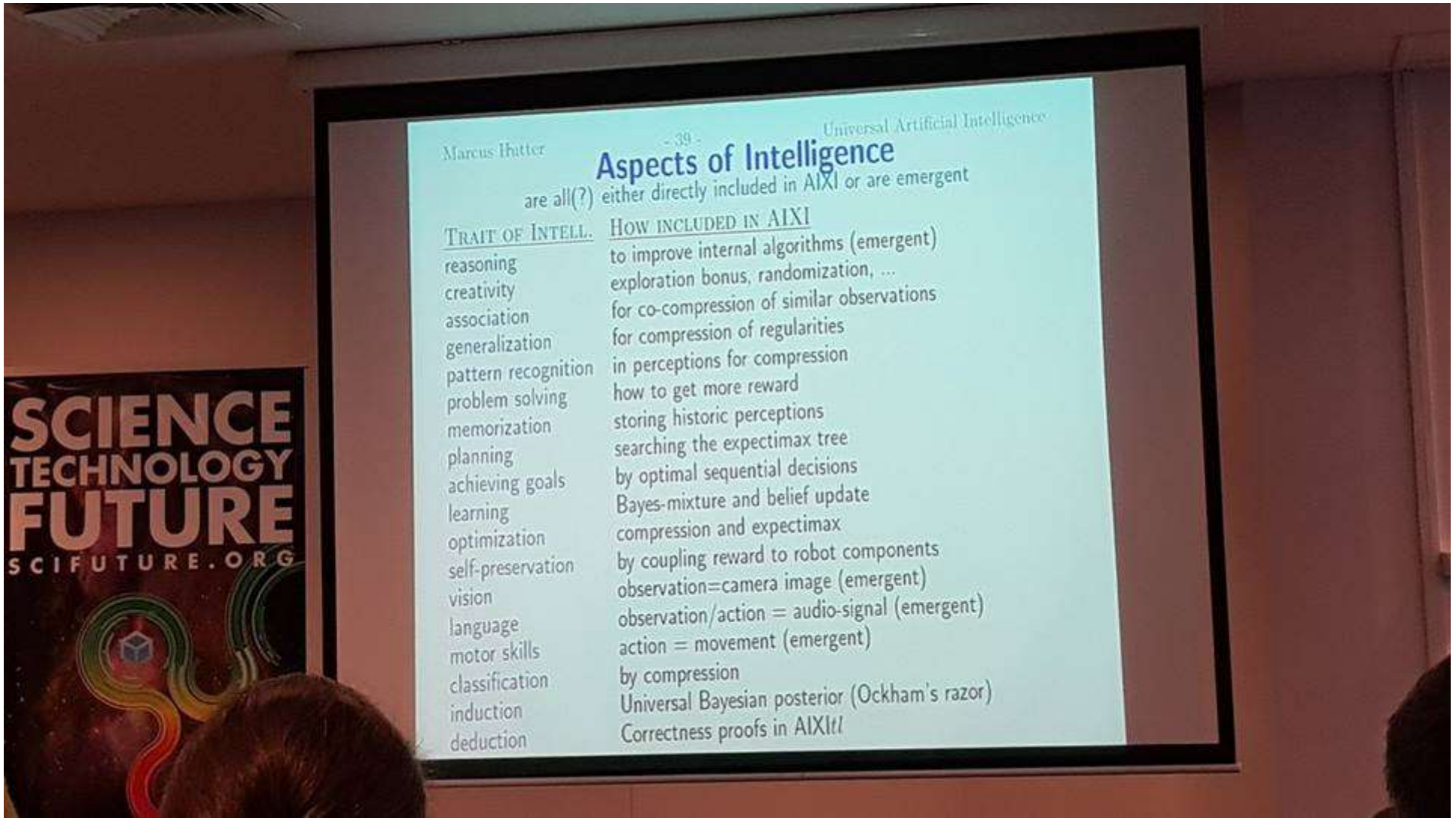
<출처> Marcus Hutter, AGI 2017-정지훈 Facebook 사진첩

미래의 인공지능 응용 시스템



<출처> Marcus Hutter, AGI 2017-정지훈 Facebook 사진첩

인공지능에서 풀어야 할 문제들

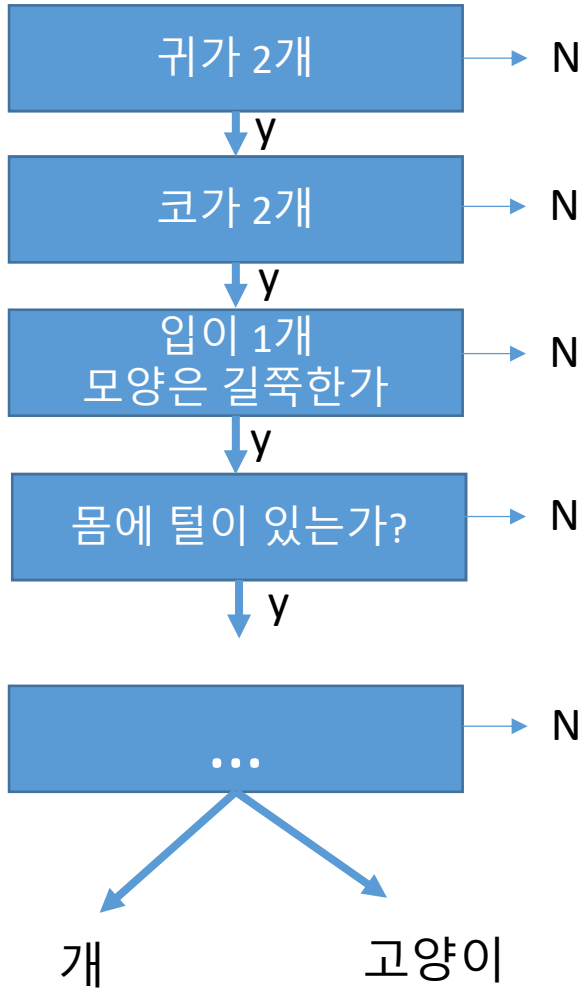


<출처> Marcus Hutter, AGI 2017-정지훈 Facebook 사진첩

II 머신러닝 개념 소개

머신러닝을 사용하지 않고 사진 설명으로 강아지 인식하기

If then else 규칙을
이용한 접근 방법



머신러닝의 필요성

- Limitations of explicit programming
  - Spam filter: many rules
  - Automatic driving: too many rules
- Machine learning: "Field of study that gives computers the ability to learn without being explicitly programmed" Arthur Samuel (1959)

<출처> 홍콩과기대 김성훈 교수 딥러닝 강의 노트

Supervised Learning – training data set 예

Supervised Learning : Learning with labeled examples – training set 이 필요

<출처> http://cs231n.github.io/classification/

Supervised Learning의 적용 분야

- Predicting final exam score based on time spent
  - regression
- Pass/non-pass based on time spent
  - binary classification
- Letter grade (A, B, C, D and F) based on time spent
  - multi-label classification

x (hours)	y (score)
10	90
9	80
3	50
2	30

Regression 예

x (hours)	y (pass/fail)
10	P
9	P
3	F
2	F

binary classification 예

x (hours)	y (grade)
10	A
9	B
3	D
2	F

Multi-label Classification 예

<예제 출처> 홍콩과기대 김성훈 교수 딥러닝 강의 노트

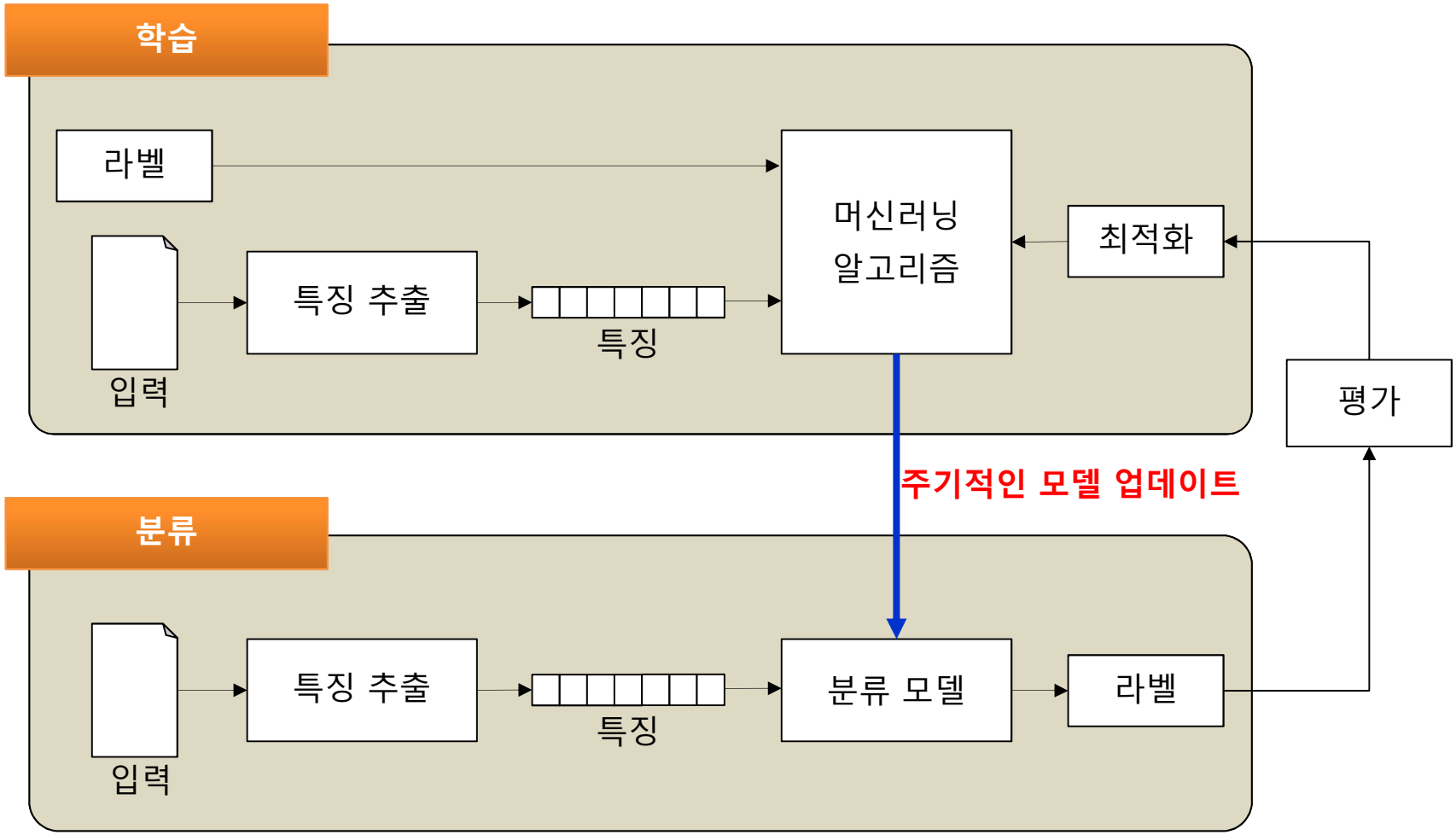
Training Set = Big Data

Machine Learning = Data Driven Approach

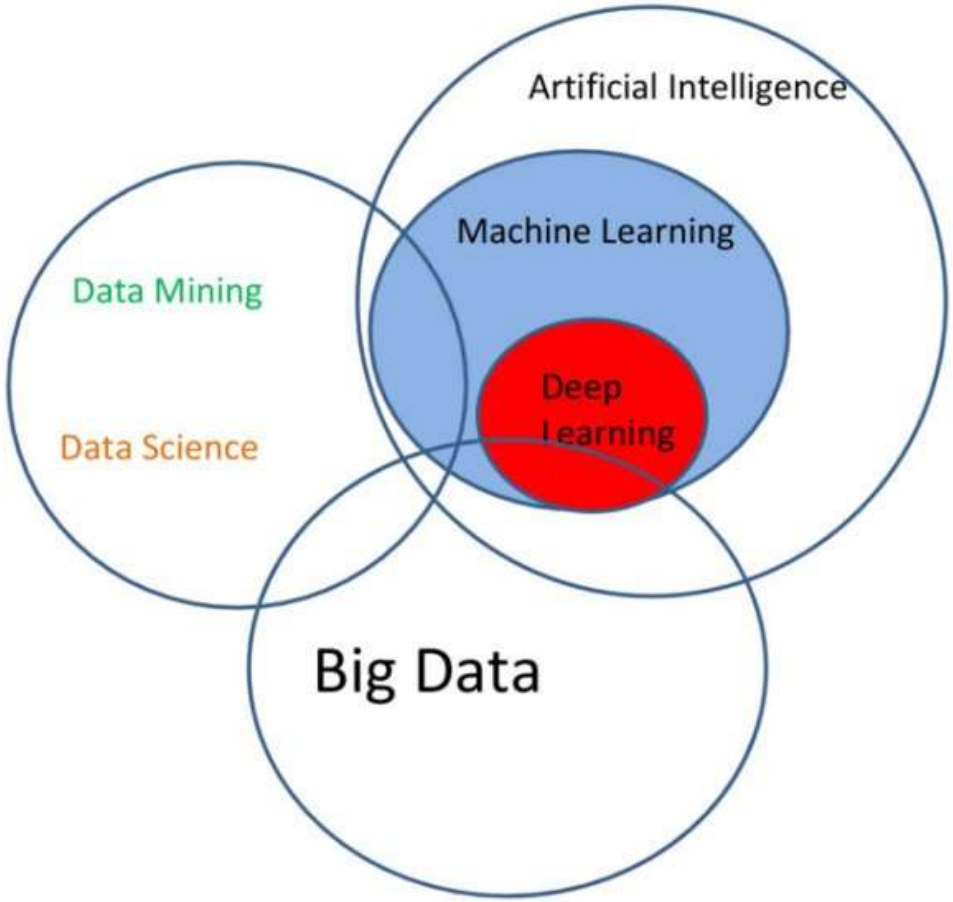
AlphaGo:

- 전세계의 모든 기보를 훈련데이터로 활용
- 기존 접근 방법: 규칙 -> 한계

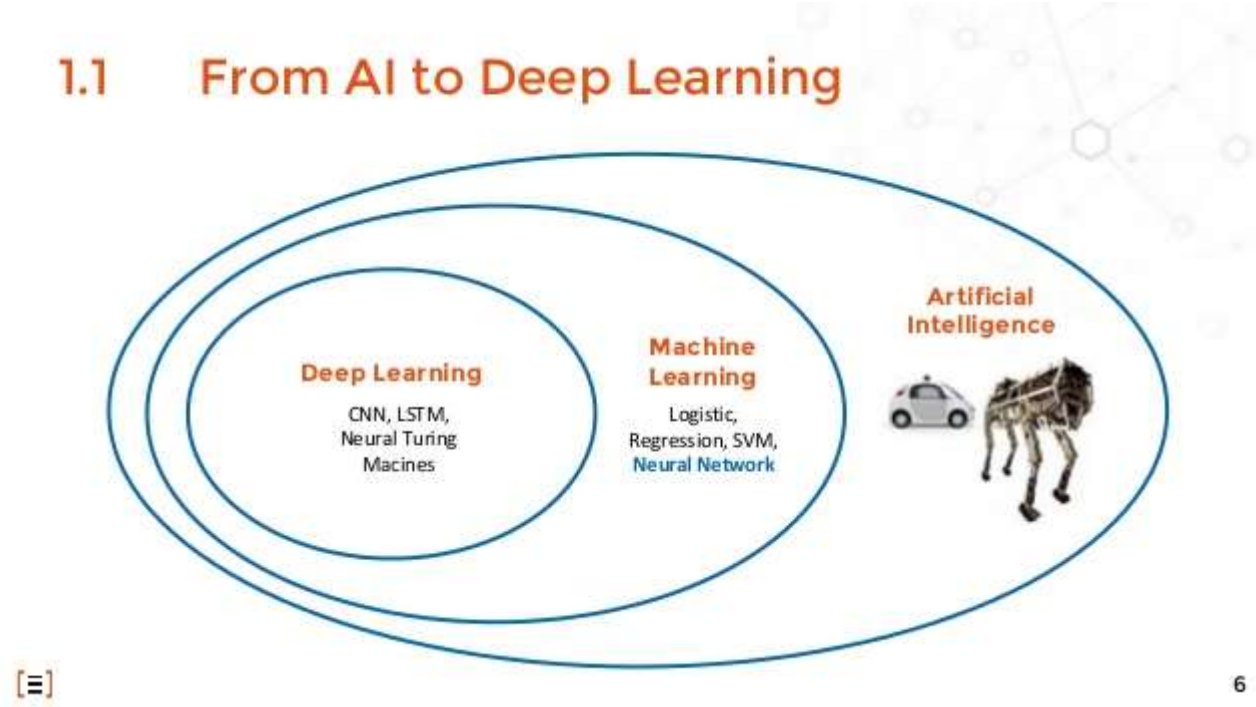
머신러닝을 이용하여 사진으로 강아지 인식하기



머신러닝, 딥러닝, 인공지능, 빅데이터

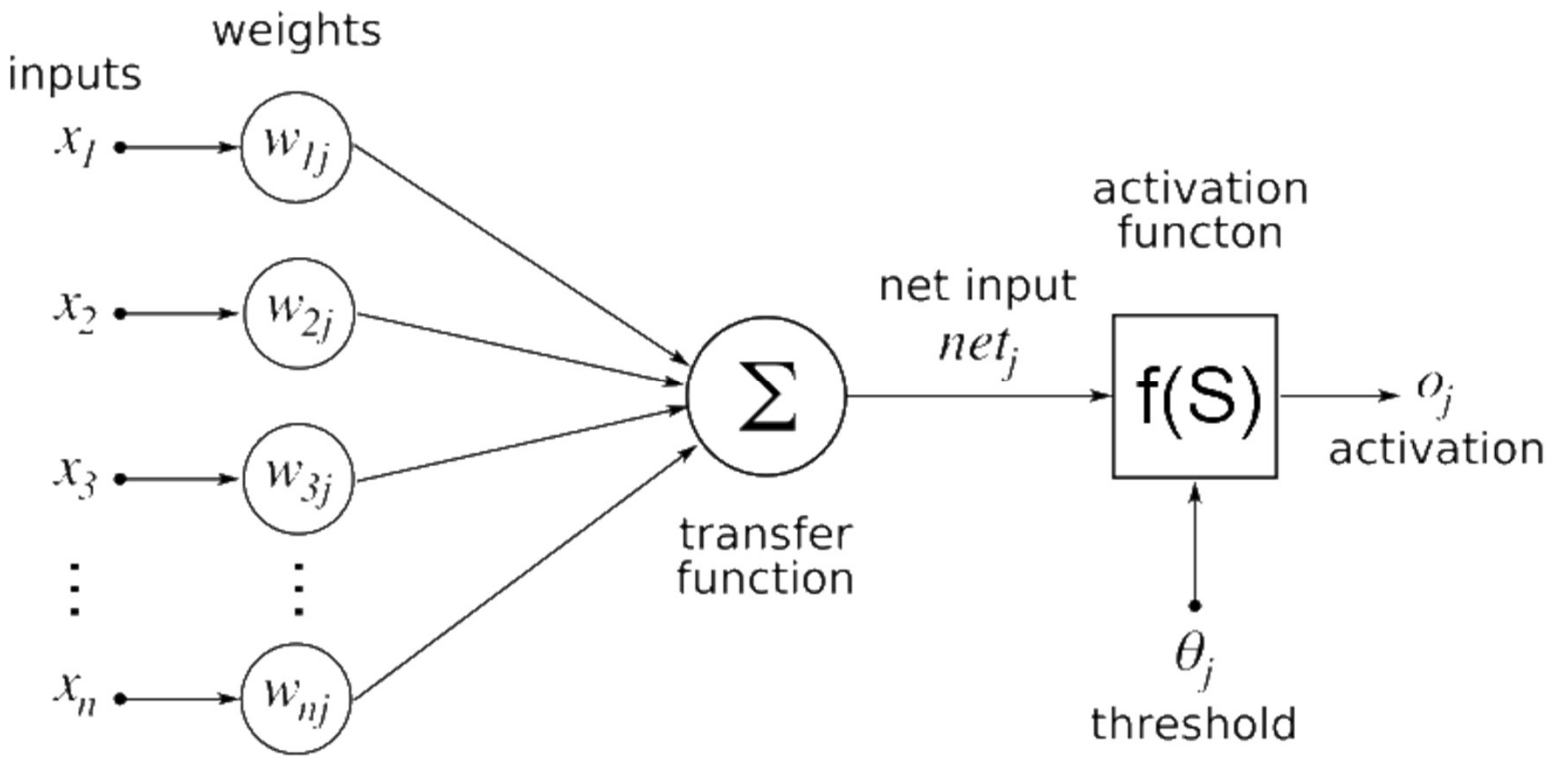


인공지능: 머신러닝 또는 딥러닝 등의 도구를 이용하여 인간 지능을 흉내낸 결과물



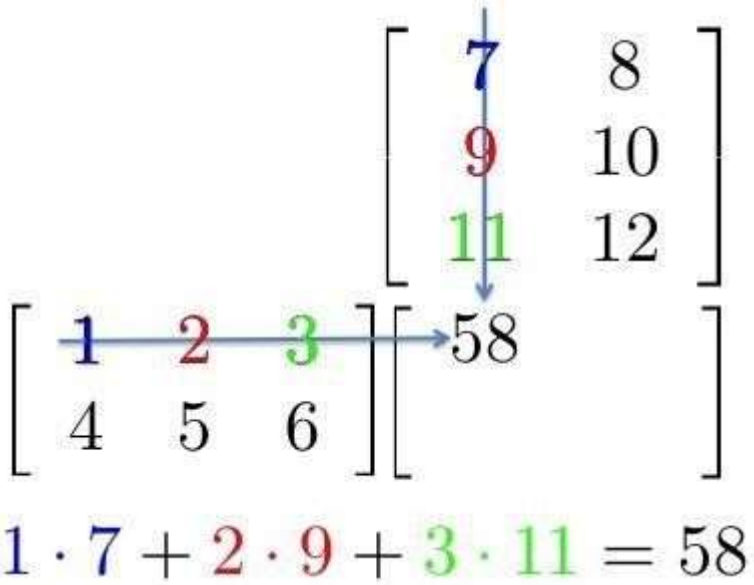
딥러닝

딥러닝 기본 구조



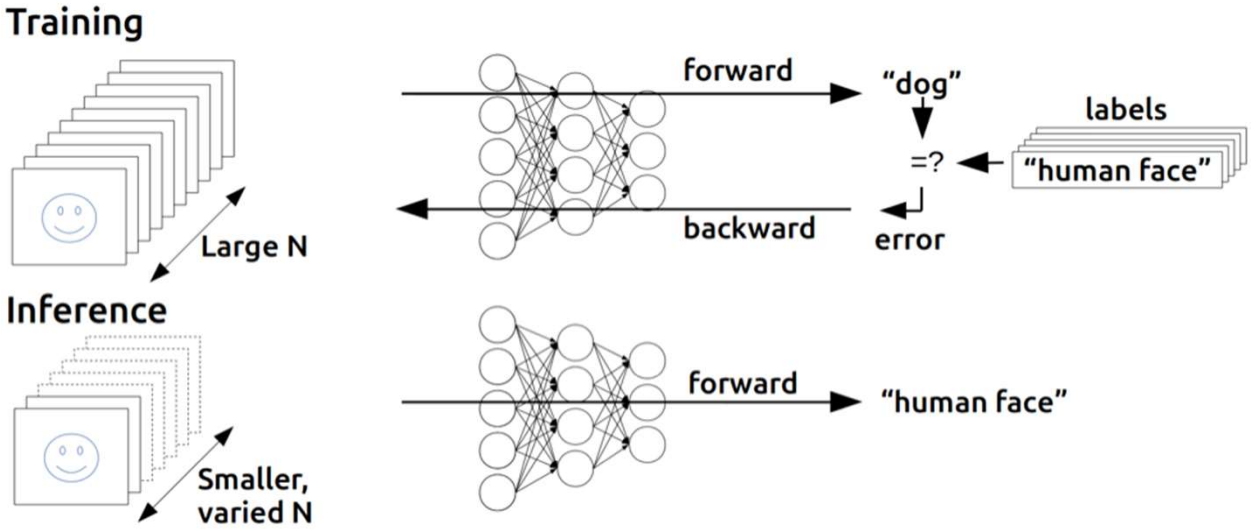
<출처> Why are GPUs necessary for training Deep Learning models?
https://www.analyticsvidhya.com/blog/2017/05/gpus-necessary-for-deep-learning/

딥러닝 기본 연산 = Matrix Multiplication



<출처> Why are GPUs necessary for training Deep Learning models?
https://www.analyticsvidhya.com/blog/2017/05/gpus-necessary-for-deep-learning/

딥러닝 연산구조



<출처> Why are GPUs necessary for training Deep Learning models?
https://www.analyticsvidhya.com/blog/2017/05/gpus-necessary-for-deep-learning/

IV 딥러닝을 위한 컴퓨팅 구조

인공지능을 위한 컴퓨팅 기술

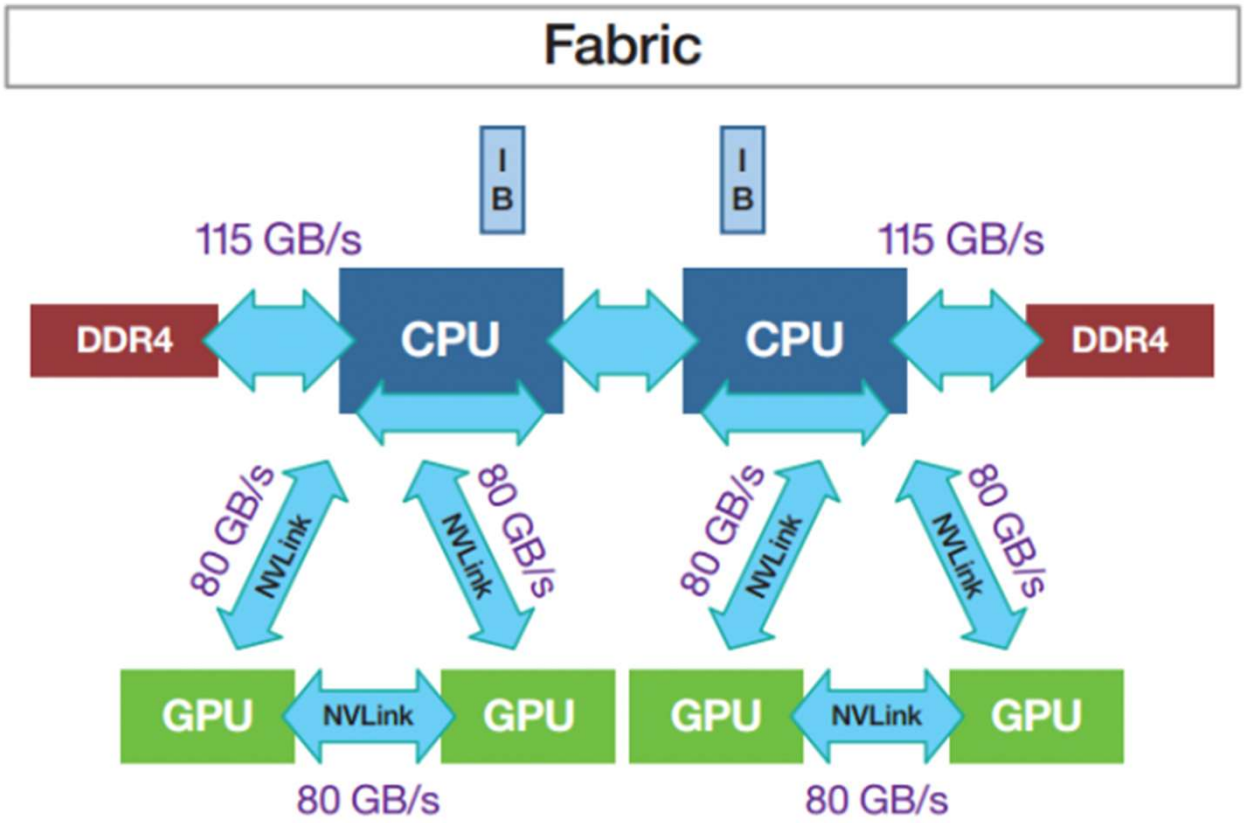
	Google	Stanford
Number of cores	1K CPUs = 16K cores	3GPUs = 18K cores
Cost	$5B	$33K
Training time	week	week

Before the boom of Deep learning, Google had a extremely powerful system to do their processing, which they had specially built for training huge nets. This system was monstrous and was of $5 billion total cost, with multiple clusters of CPUs.

Now researchers at Stanford built the same system in terms of computation to train their deep nets using GPU. And guess what; they reduced the costs to just $33K ! This system was built using GPUs, and it gave the same processing power as Google's system. Pretty impressive right?

<출처> Why are GPUs necessary for training Deep Learning models?
https://www.analyticsvidhya.com/blog/2017/05/gpus-necessary-for-deep-learning/

CRAY Announces New, AI-Focused Supercomputers



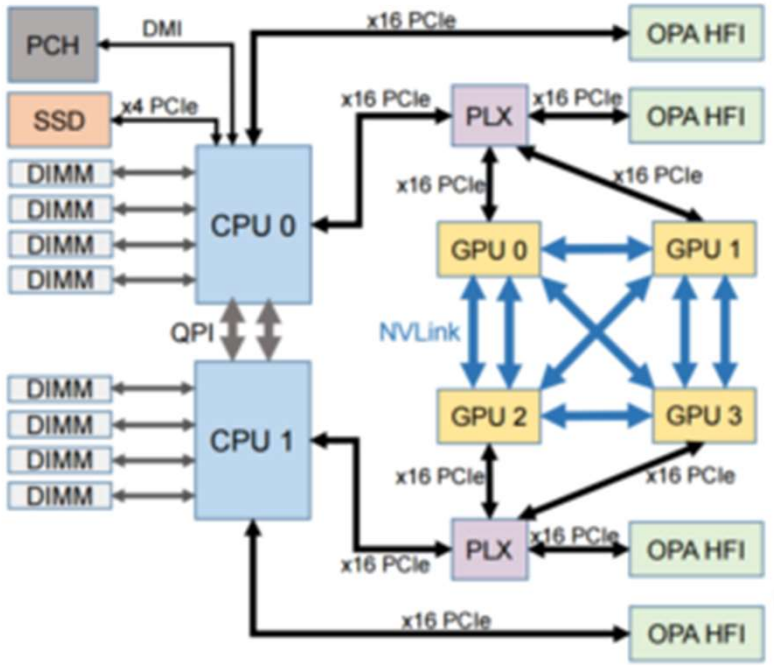
<출처> https://www.extremetech.com/computing/249149-cray-announces-new-ai-focused-supercomputers

TSUBAME3.0

Tokyo Institute of Technology (Tokyo Tech) announced its plans for the TSUBAME3.0 supercomputer, which will be Japan's "fastest AI supercomputer," when it comes online this summer (2017).

Projections are that it will deliver 12.2 double-precision petaflops and 64.3 half-precision (peak specs)

We know that Nvidia will be supplying Pascal P100 GPUs



<출처> https://www.extremetech.com/computing/249149-cray-announces-new-ai-focused-supercomputers

외국기업 현황

기업/기관/국가	기술 동향	출처
IBM	• **Watson 슈퍼 컴퓨터는 80 teraflops의 속도로 처리** • Watson은 2 억 페이지가 넘는 결합 된 데이터 저장소로 90 개의 서버 액세스 • 600 만 가지 논리 규칙을 처리 • 장치와 그 데이터는 10 개의 냉장고를 수용 할 수 있는 공간에 독립적으로 보관됨 • 2,880 processor cores • 15 terabytes of RAM • 500 gigabytes of preprocessed information	Margaret Rouse, “IBM Watson supercomputer,” Whatis.com
	• 자사에서 개발한 “트루노스 신경시냅틱 시스템 " 을 기반으로 인공지능 시스템 개발중 • 프로세서 자체가 인간의 신경세포와 시냅스 구조를 모방	생탐, “ 인공지능 연구한계점 도달, 해결법과 앞으로의 미래는?”, 네이버포스트, 2017. 7. 11

외국 기업 현황

기업/기관/국가	기술 동향	출처
Cray	• AI를 지원하기 위한 슈퍼컴퓨터 발표 • Cray CS-Storm 500GT 및 CS-Storm 500NX • 두 제품 모두 Nvidia의 Pascal 기반 Tesla GPU와 작동하도록 설계되었지만 서로 다른 기능 세트 및 기능을 제공 • CS-Storm 500GT는 Nvidia의 Tesla P40 또는 P100 GPU 가속기를 포함하여 최대 8 x 450W 또는 10 x 400W 가속기를 지원	Joel Hruska, "Cray Announces New, AI-Focused Supercomputers," ExtrmeTech, May 11, 2017
마이크로 소프트	• 클라우드 기반 데이터 센터에 강력한 실리콘 FPGA 설치 • 급성장하는 인공 지능 분야에서 마이크로소프트사의 Azure 인공지능 컴퓨팅 플랫폼이 강력한 경쟁 우위를 확보하고자하는 목표 추구	Dan Richman, "Microsoft demonstrates the world's 'first AI supercomputer,' using programmable hardware in the cloud," September 26, 2016, GeekWire

CRAY Announces New, AI-Focused Supercomputers

	CS-Storm 500GT	CS-Storm 500NX
Processors	Two Intel® Xeon® "Skylake" family processors	Intel Xeon E5-2600 v4 family processors
Memory Capacity	Up to 2 TB DDR4 (16 x 128 GB DIMMs)	Up to 3 TB DDR4 (24 x 128 GB DIMMs)
Accelerators	Up to 10 NVIDIA® Tesla® P40 or P100 PCIe GPU accelerators; supports up to 400W parts	Up to 8 NVIDIA Tesla P100 SXM2 GPU accelerators; supports up to 300W parts
Drive Bays	Multiple local storage configuration options. 12 hot-swappable 2.5" drives (up to 4 x NVMe). Some configurations require an additional add-in storage controller; total number of drives varies by configuration.	16 hot-swappable 2.5" drives (up to 8x NVMe)
Expansion Slots	12 PCIe 3.0 x 16 slots supporting multiple PCIe topologies and configuration options	4 x16, low-profile (GPU tray), 2 x 8 (motherboard tray) PCIe 3.0
Power Supply	Four 2200W AC power supplies; N+1 and N+N (limited to specific configurations) redundancy	Four 2200W AC power supplies; 2+2 redundancy; titanium-level efficiency
Power Input	200-277VAC, 10A max	200-277VAC, 10A max
Weight	Up to 76 lbs. (without PCIe cards)	Up to 135 lbs.
Dimensions	5.25" H x 17.6" W x 36.4"D (3U)	7.0"H x 17.6"W x 29"D (4U)
Temperature	Operating: 10°C–35°C, ASHRAE 2	Operating: 10°C–35°C, ASHRAE 2

<출처> https://www.extremetech.com/computing/249149-cray-announces-new-ai-focused-supercomputers

외국기업현황

기업/기관/국가	기술 동향	출처
구글	• TPU 포드 (Pods)라고 불리는 64 개의 TPU를 통합하여 11.5 페타 플롭의 연산 능력을 갖춘 슈퍼 컴퓨터로 효과적으로 전환 할 수있는 방법을 개발 • 2세대 TPU는 180 teraflops의 컴퓨팅 파워 제공 목표 • 알파고는 1 petaflops 정도의 계산속도 소요	Nick Statt, "Google's next-generation AI training system is monstrously fast," May 17, 2017, THEVERGE
	• Ray KuzWeil 기술 이사를 필두로 양자컴퓨터 개발중 – 양자컴퓨터와 딥러닝 및 머신러닝 접목중	생탐, " 인공지능 연구한계점 도달, 해결법과 앞으로의 미래는?", 네이버포스트, 2017. 7. 11
일본	• AI 브리징 클라우드 컴퓨터 (AIBC)를 만들기 위해 195 억엔 (140 백만 파운드) 예산 투입 계획 • 130 petaflops의 속도 목표	"Japan kicks off AI supercomputer project," 28 November 2016, BBC.com

AlphaGo 구성 및 원가(?)

알파고 2년간 운영 비용(최대)

AlphaGo

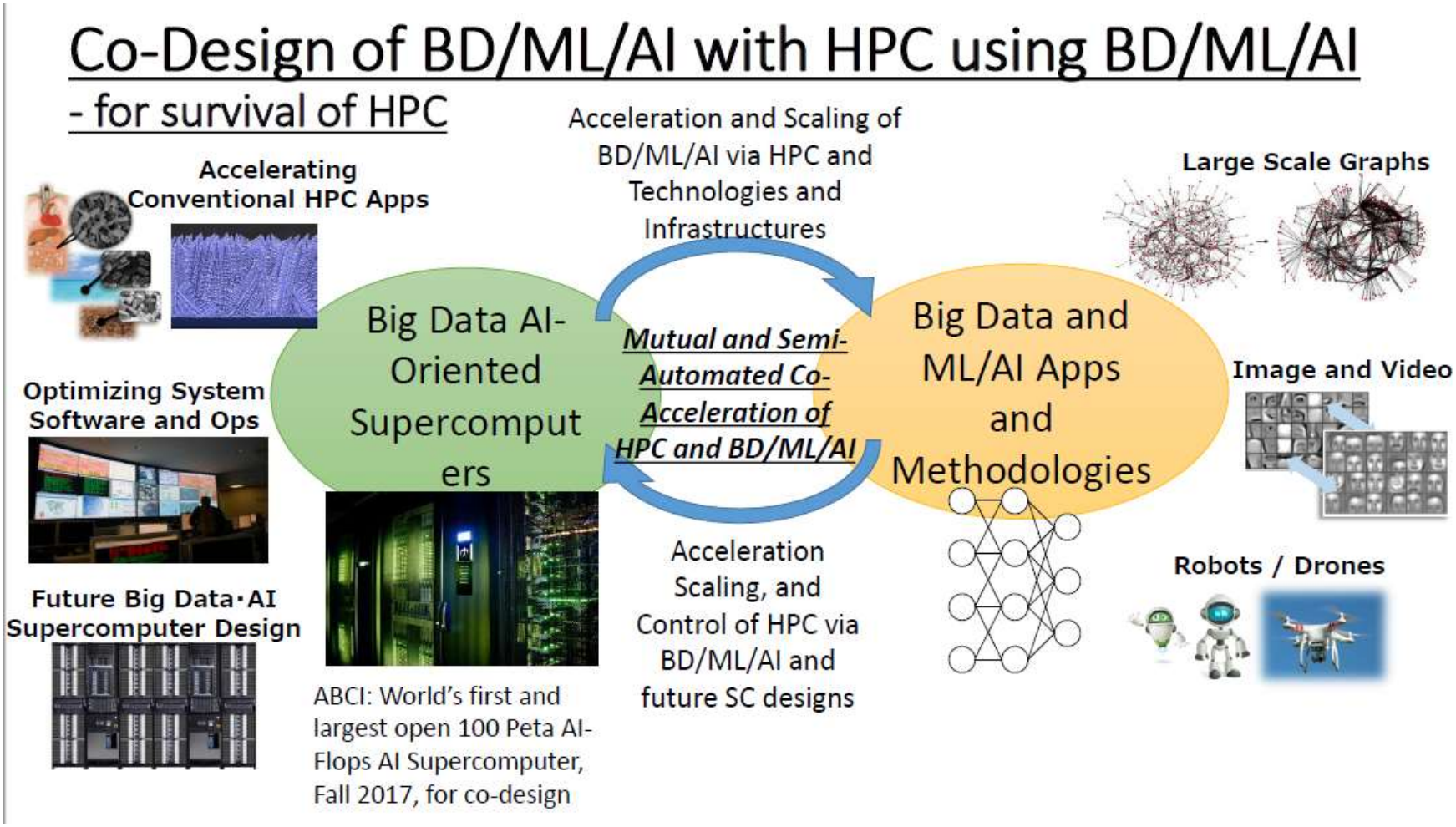
항목	스펙		총 비용
1. 개발자	20명 (연봉) 3억6000만원/인당		
2. 프로젝트 기간	2년		
3. 투입리소스	100MM		30억원
4. 인프라 비용	인프라	CPU 1920개, GPU 280개 (분산서버 시간당 운영 비용) 282.8$	58억2000만원
	CPU(가격/시)	c3.8xlarge*60대 = CPU(1920개)*1.68$(가격/시)=100.8$	
	GPU(가격/시)	g2.8xlarge*70대 = GPU(280개)*2.6$(가격/시) =182$	
합계			88억2000만원

<그래픽=홍종현 미술기자>

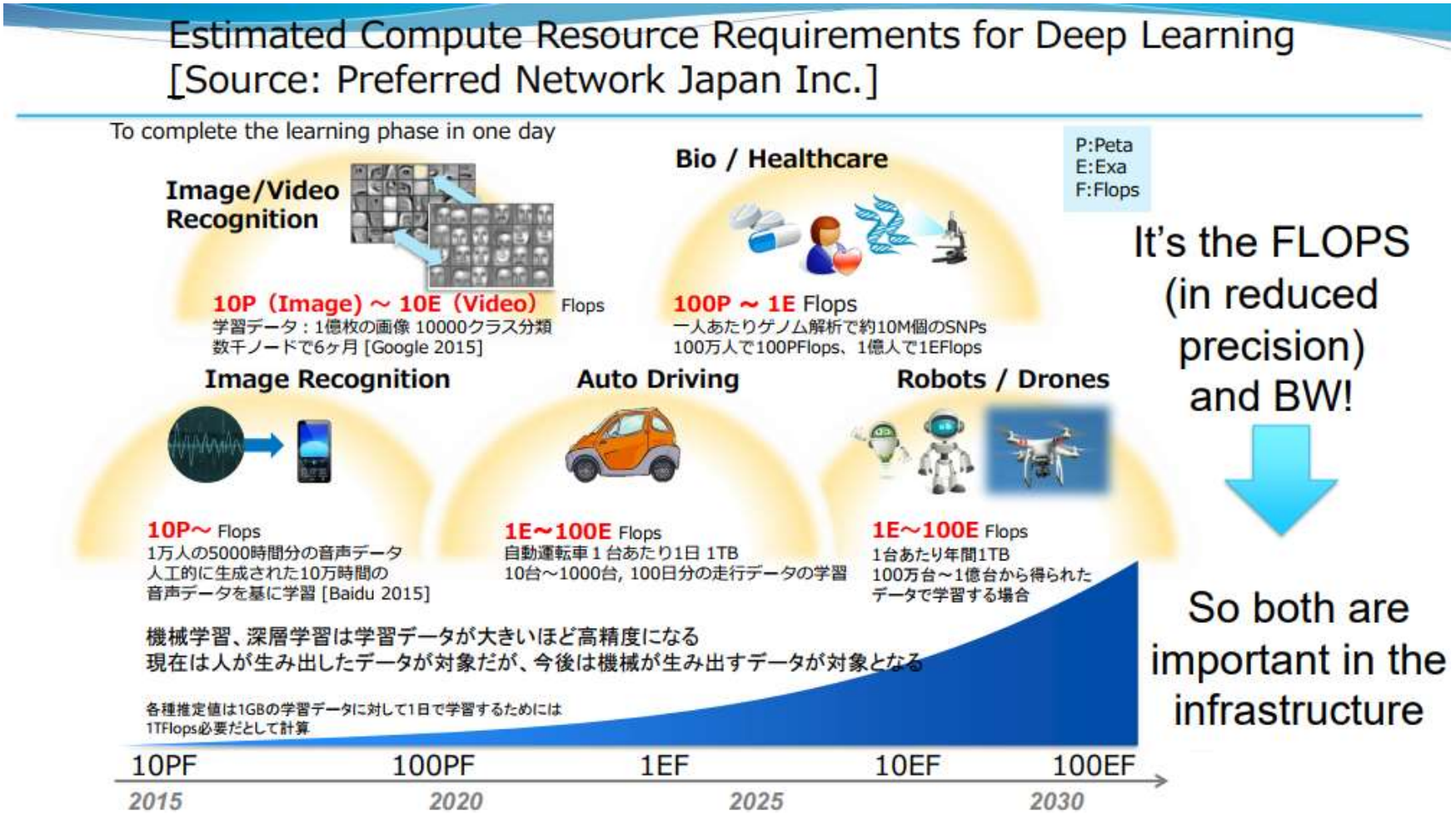
<출처> http://www.newspim.com/news/view/20160318000429

일본

AI R&D 협력 센터(AI Bridging Cloud Infrastructure)



딥러닝을 위한 처리 속도 요구사항



<출처> 最新の**HPC**技術を生かした**AI**・ビッグデータ, 2017, NVIDIA Deep Learning Institute Day

V 결어

인공지능 시스템 수행에 필요한 계산 능력

- IBM Watson: 80TFlops
- Alphago: 1PFlops
- Chatbot: 그 이하
- 인간의 영상 인식과 동일한 수준의 인식이 가능하려면?
- 인간의 언어 통역과 인식 수준의 인식이 가능하려면?

시사점 및 추진 방향

- 현재는 알고리즘에 의한 빅데이터 처리를 통해 인공지능 시스템이 동작 발전
- 향후 하드웨어적으로도 신기술 등장 예상

감사합니다
Q & A

세종대학교
SEJONG UNIVERSITY

ABRC
인공지능-빅데이터연구센터
http://www.abrc.or.kr