Powering AI Robots with Deep Learning

NVIDIA Deep Learning Day 2017

COEX Intercontinental Hotel, Seoul, Korea, Tuesday, October 31, 2017

Byoung-Tak Zhang

Seoul National University
School of Computer Science and Engineering &
Surromind Robotics, Inc.
http://bi.snu.ac.kr/

Outline

1. AI Robots Come of Age

Humans (NI) and Machines (AI)

- Introspectionism - 1920

Psyche

- Behaviorism ←→ Cybernetics 1920 - 1950

Mind (= Computer)

- Cognitivism ←→ Symbolic AI 1950 - 1980

Brain

- Connectionism ←→ Neural Nets (ML) 1980 – 2010

Body

- Action Science ←→ Autonomous Robots 2010 –

Environment

Embodied Mind | Mind Machine (= Smart Machine)

Early AI Robots
(Autonomous Mobile Robots)

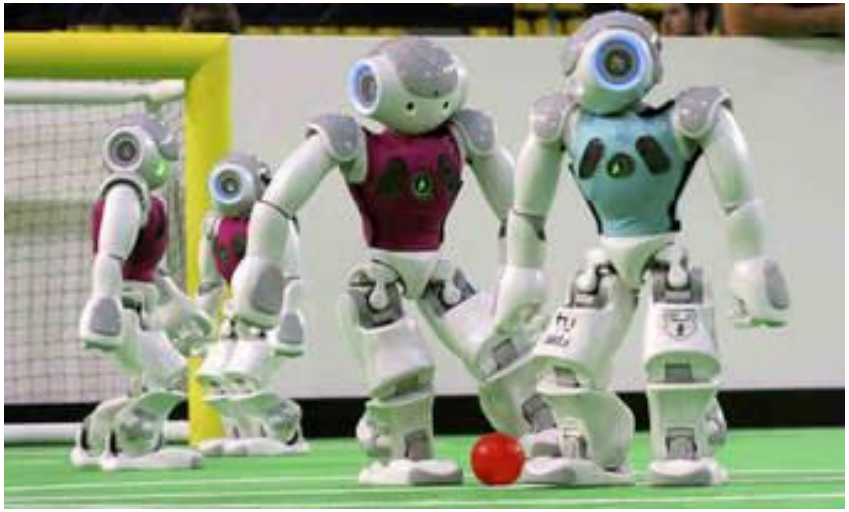
Shakey (SRI)

Cart (Stanford Univ.)

RHINO (U. Bonn)

CoBot (CMU)

RoboCup (1997~)



Home Robots

PR2 Fetches Beer (Willow Garage)

PR2 Making Popcorns (TU Munich)

Dash at Hotel (Sevioke)

SpotMini (Boston Dynamics)

Human-Like Robots

Humanoid Robot Nao (Aldebaran)

Emotion Robot Pepper (SoftBank)

Life-Like Robots (Hanson Robotics)

Atlas (Google Boston Dynamics)

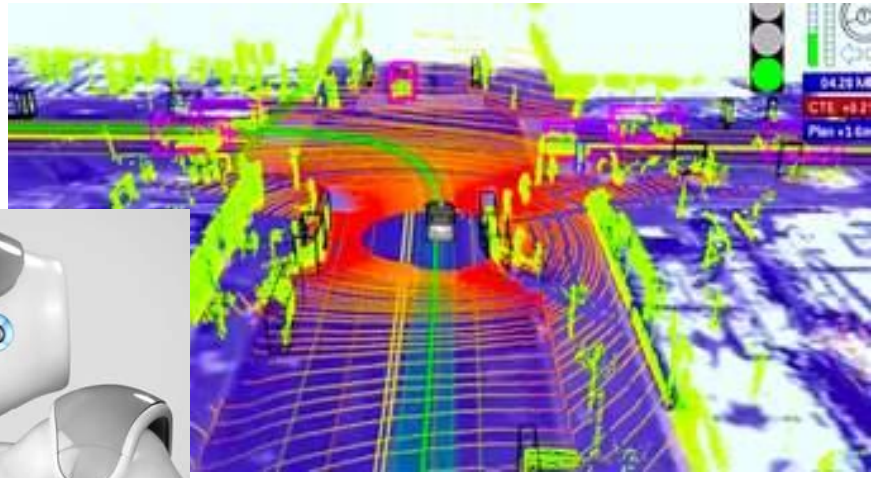
Robot Life in a City

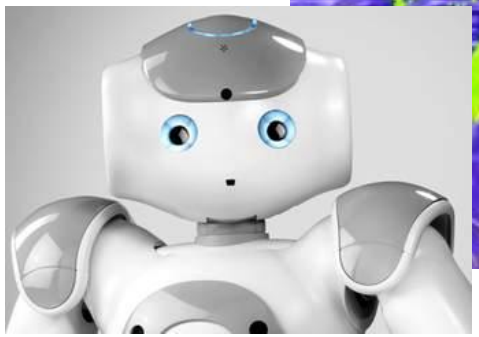
Obelix (University of Freiburg, Germany)

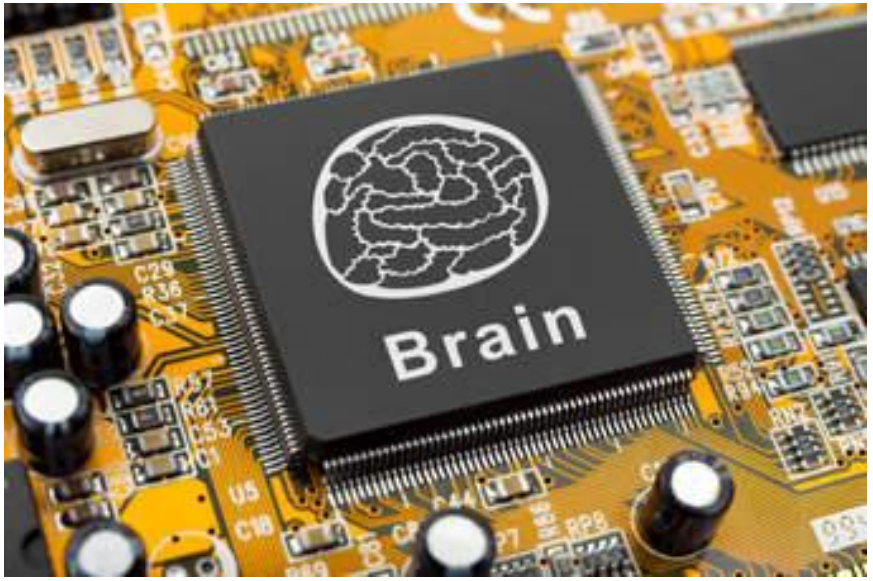
AI Robots for the 4th Industrial Revolution

Enabling Technologies for AI Robots

- **Perception**
  - Object recognition
  - Person tracking
- **Control**
  - Manipulation
  - Action control
- **Navigation**
  - Obstacle avoidance
  - Map building & localization
- **Interaction**
  - Vision and voice
  - Multimodal interaction
- **Computing Power**
  - Cloud computing
  - GPUs, parallel computing
  - Neural processors

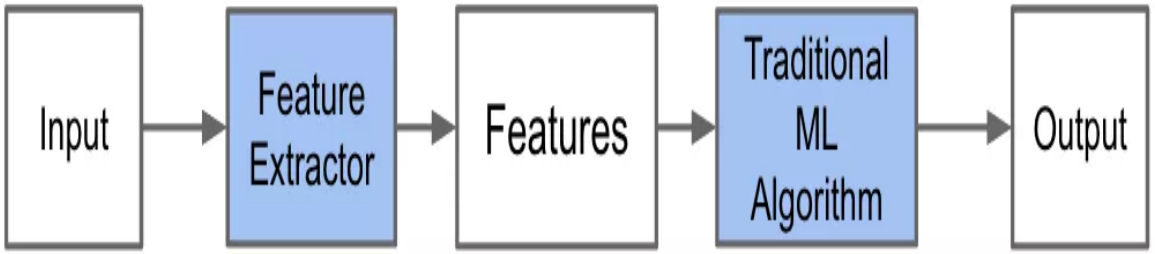




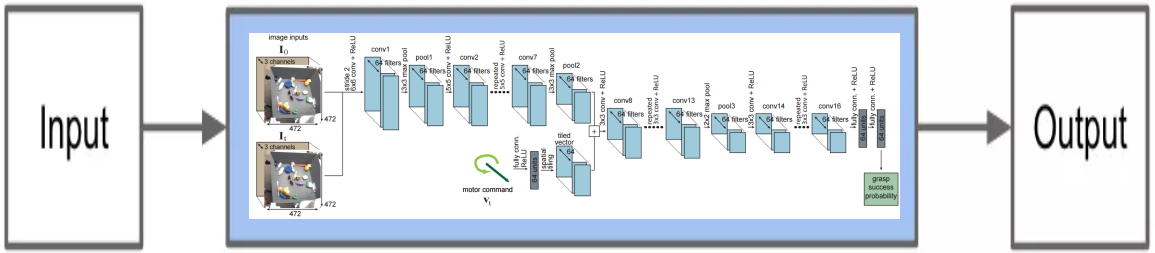


2. Deep Learning for AI Robots

Traditional Machine Learning vs. Deep Learning



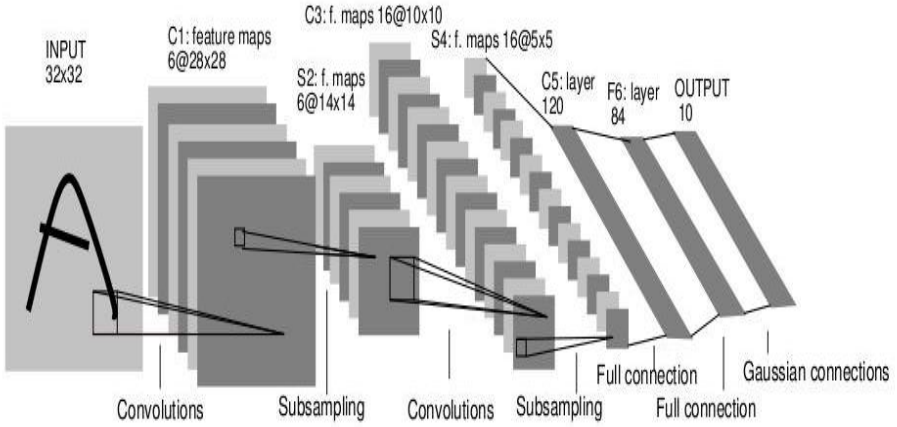
Traditional Machine Learning Flow

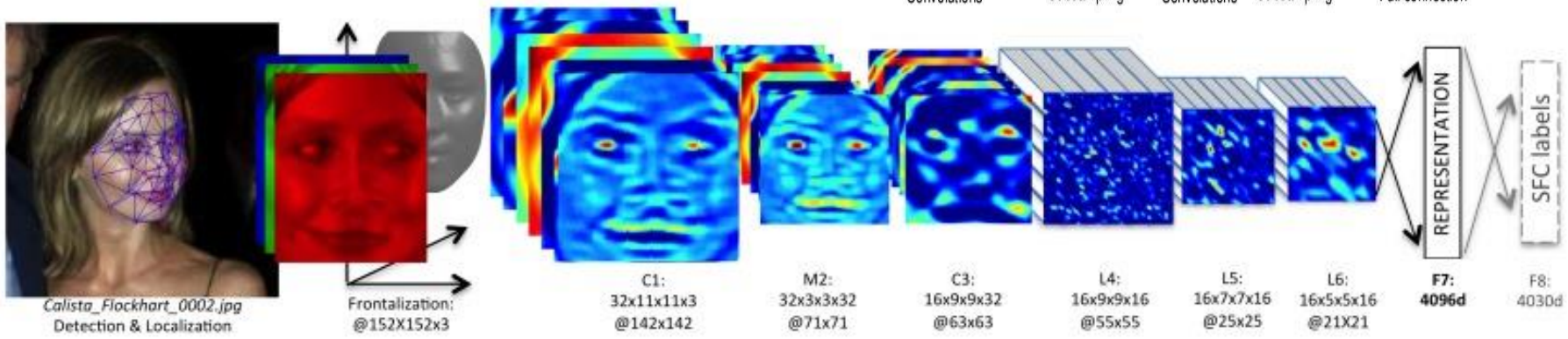


Deep Learning Flow

Deep Learning Revolution

- Big Data + Parallel Computing + Deep Learning
- From programming to learning
- Automated- or self-programming
- Paradigm shift in S/W
- Self-improving systems
- Intelligence explosion

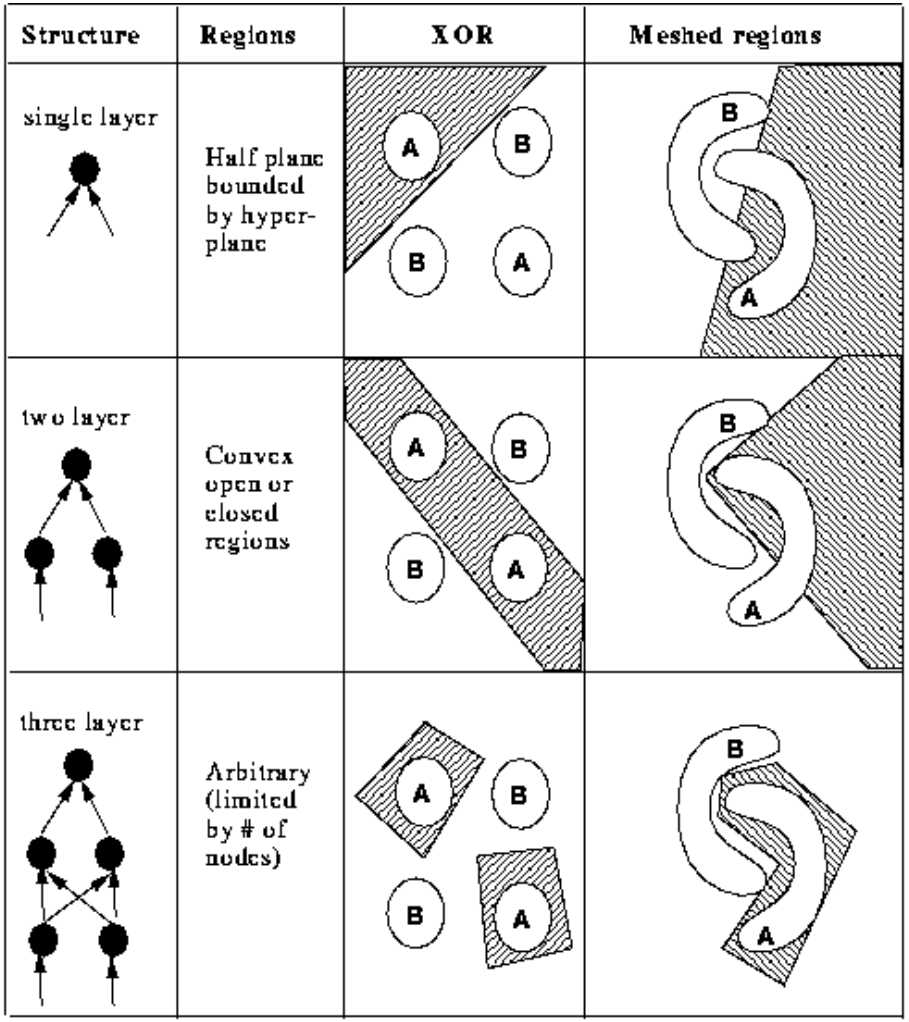




Power of Deep Learning

- Multiple boundaries are n eeded (e.g. XOR problem)
  → Multiple Units

- More complex regions are needed (e.g. Polygons)
  → Multiple Layers

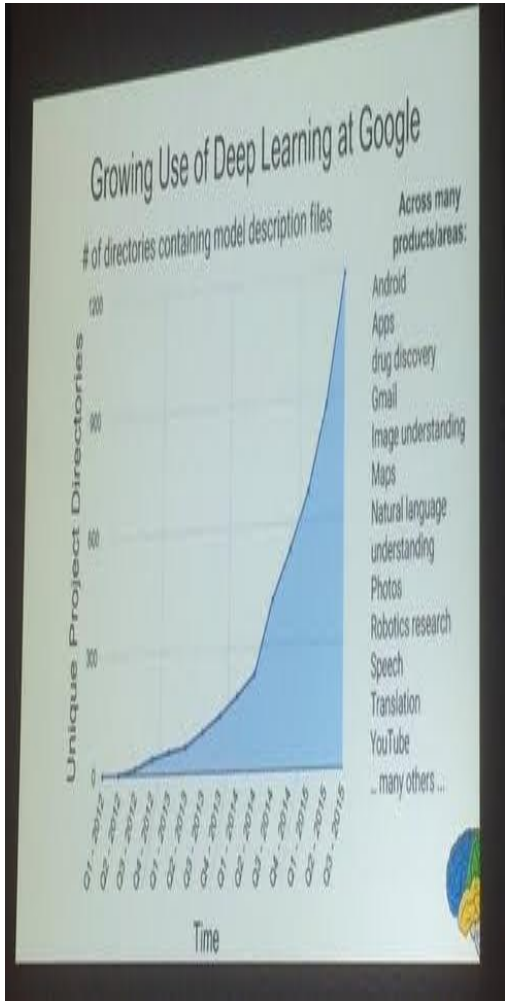


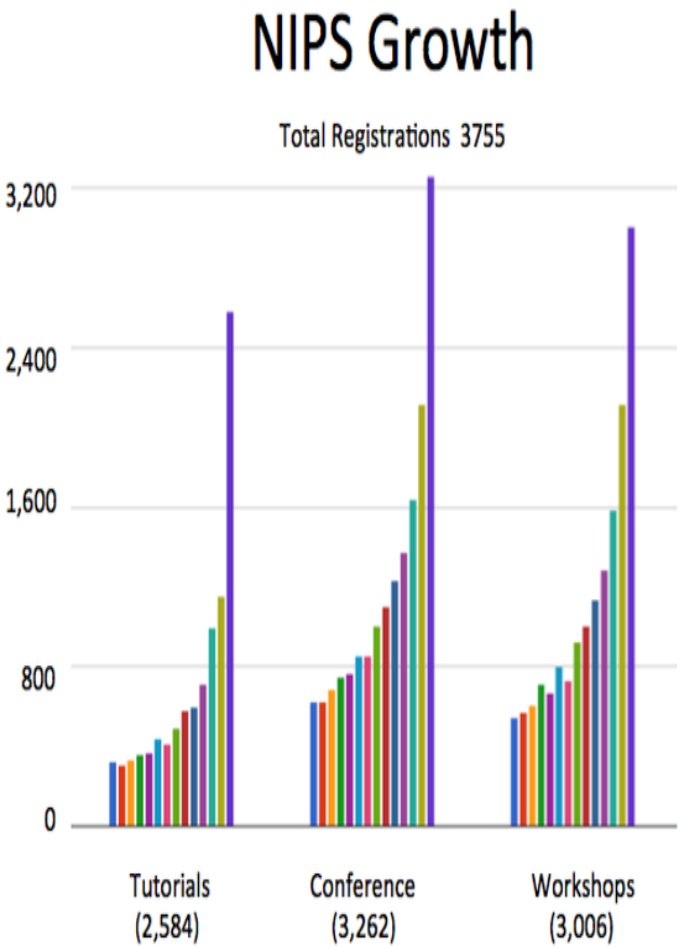
Big Data + Deep Learning

=>

Automatic Programming

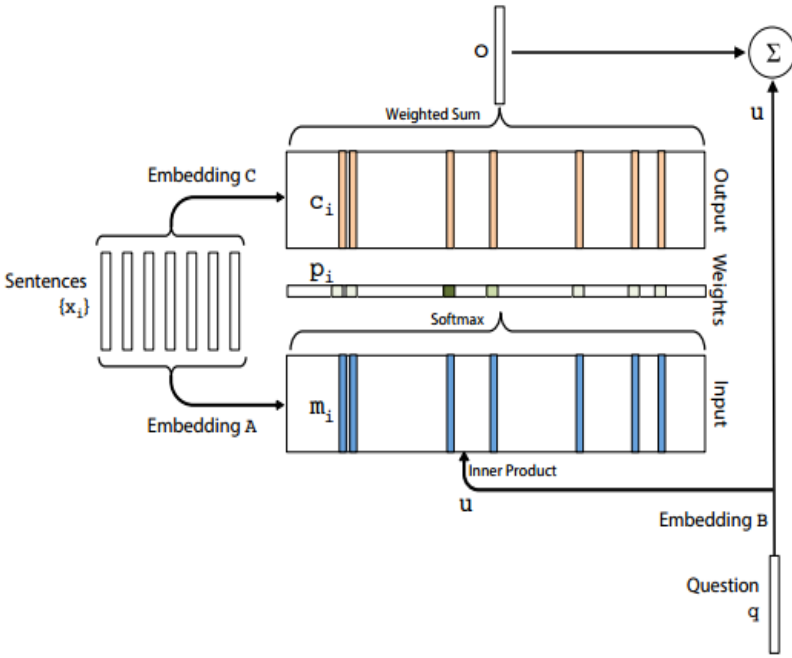
AI / Deep Learning Growth

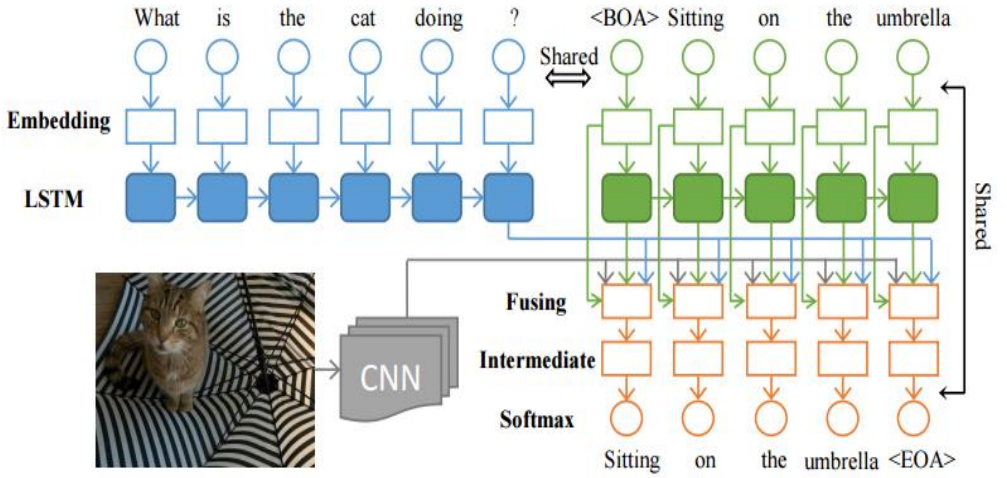




Deep Learning for Voice and Dialogue

- Speech LSTM-RNN (Recurrent Neural Networks)
- End-to-End Memory Networks (N2N MemNet)
- CNN + RNN for Question Answering





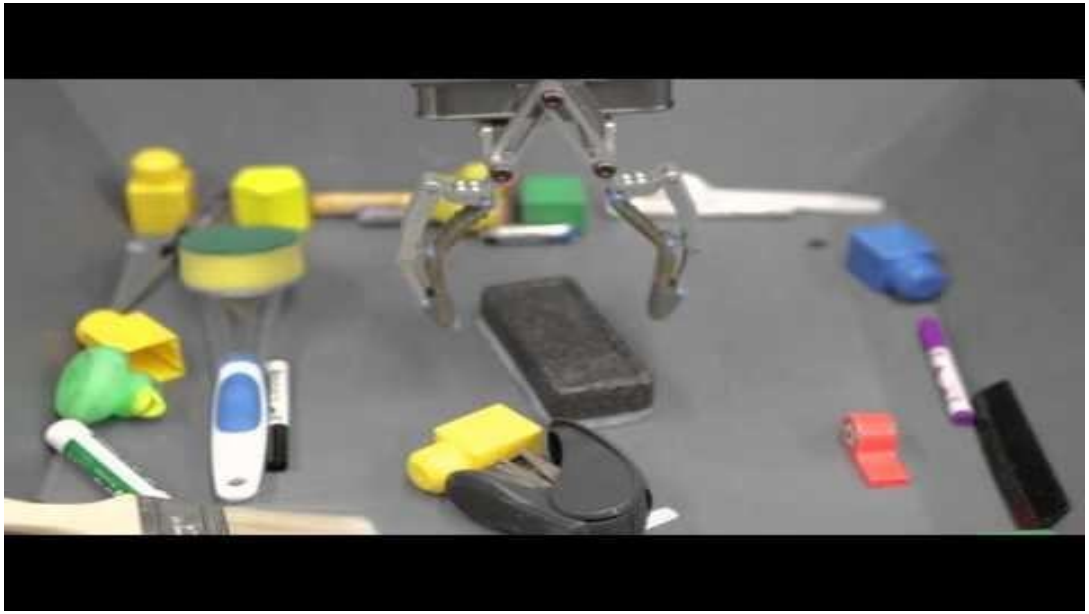
Sukhbaatar, Sainbayar, Jason Weston, and Rob Fergus. "End-to-end memory networks." *Advances in Neural Information Processing Systems*. 2015.

Gao, Haoyuan, et al. "Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question." *Advances in Neural Information Processing Systems*. 2015.

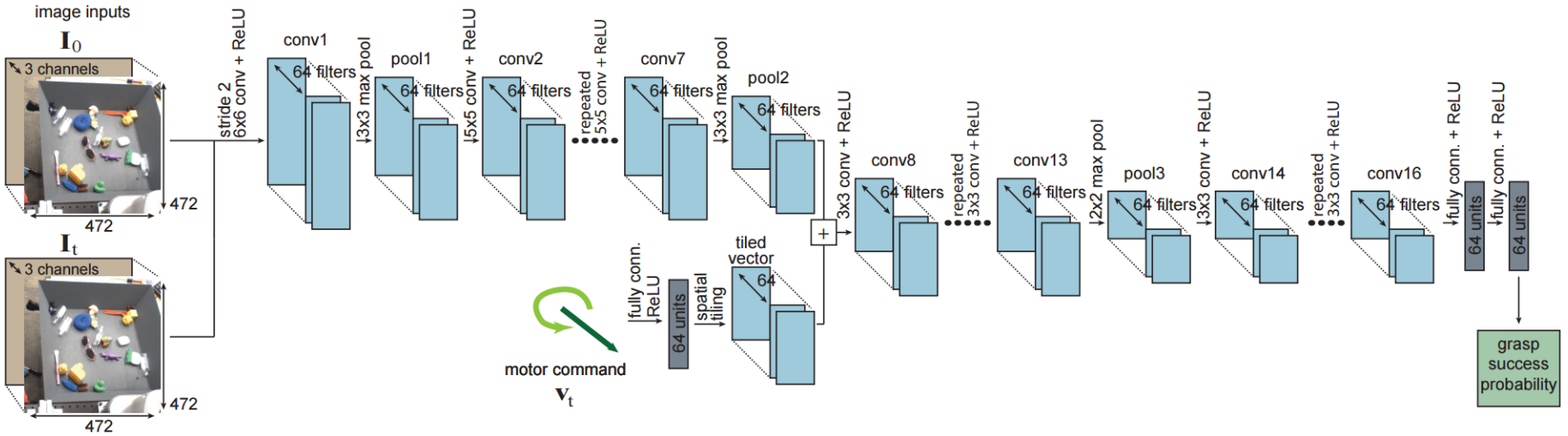
Interaction: Conversational Interface

Amazon Echo	Google Home	SKT Nugu

Deep Learning for Robotic Grasping



(Levine et al, 2016)

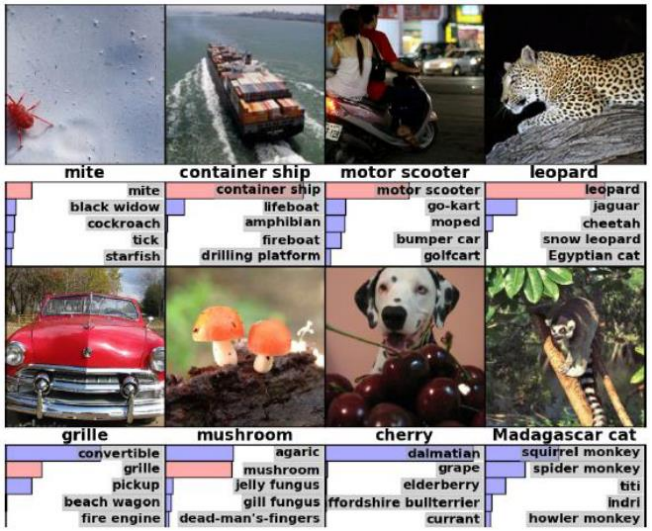


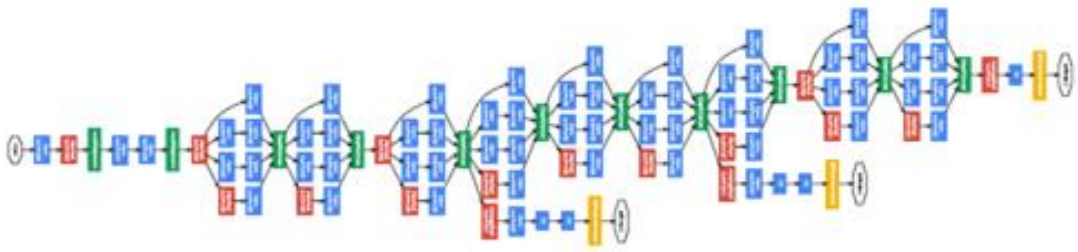
Deep Reinforcement Learning for Action Control

BRETT (Univ. of California, Berkeley)

Deep Learning for Perception

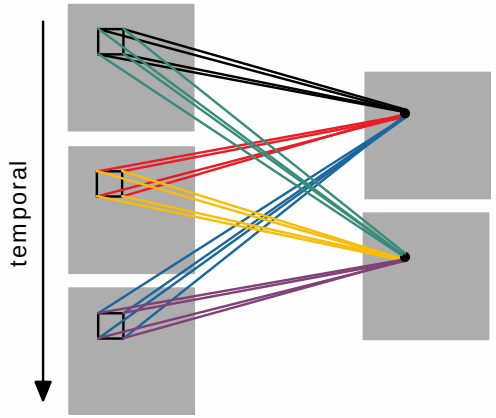
- **ImageNet Large-Scale Visual Recognition Challenge**
  - Image Classification/Localization
  - 1.2M labeled images, 1000 classes
  - Convolutional Neural Networks (CNNs) has been dominating the contest since..
    - 2012 non-CNN: 26.2% (top-5 error)
    - 2012: (Hinton, AlexNet)15.3% **(Using GPUs)**
    - 2013: (Clarifai) 11.2%
    - 2014: (Google, GoogLeNet) 6.7%
    - (pre-2015): (Google) 4.9%
      - Beyond human-level performance

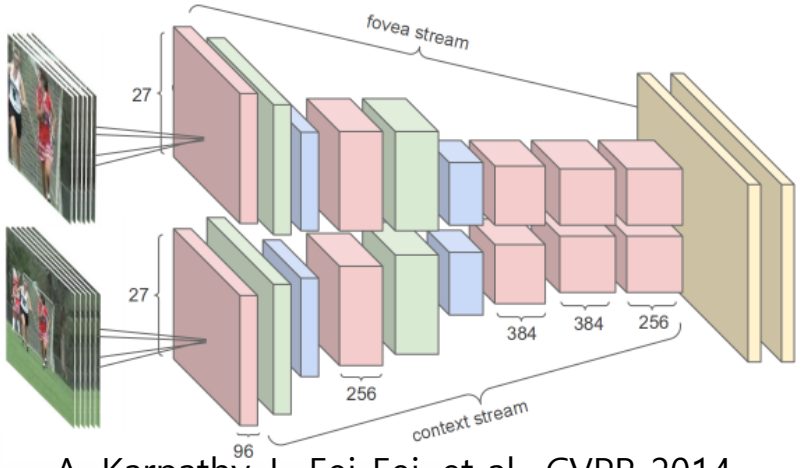




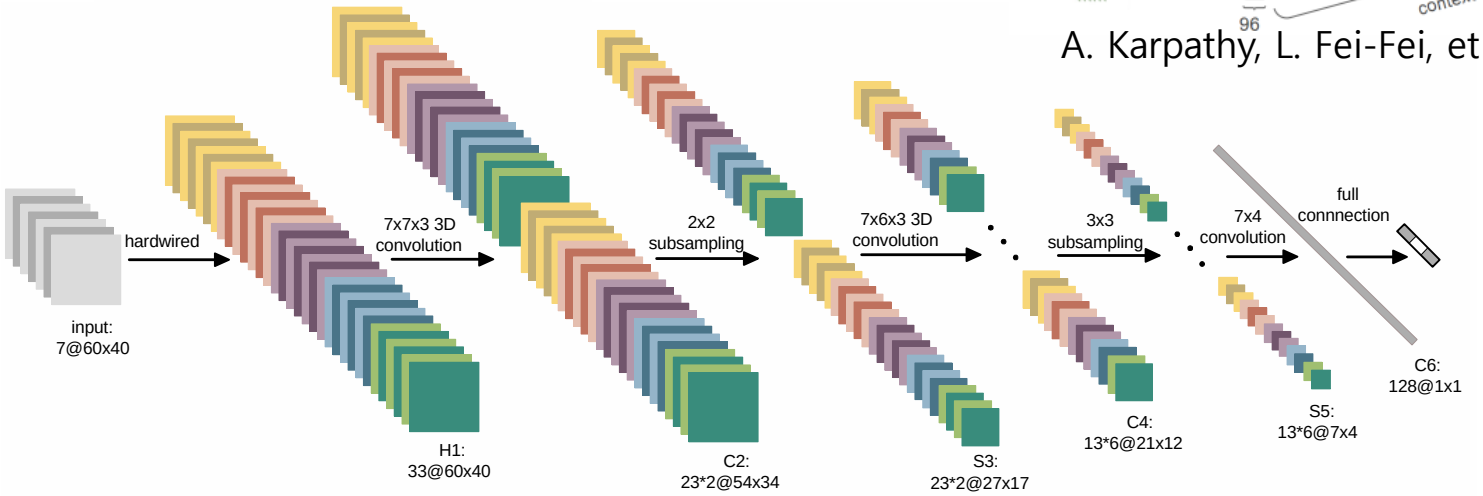
Deep Learning for Video Analysis

- **Use 3D CNNs to model the temporal patterns as well as the spatial patterns**





A. Karpathy, L. Fei-Fei, et al., CVPR 2014



S. Ji, K. Yu, et al., PAMI, 2013

Deep Learning for Autonomous Driving (NVIDIA)

VQA Challenge (2016)

- **Visual Question Answering (VQA) is a new dataset containing open-ended questions about images. These questions require an understanding of vision, language and common sense to answer.**
  - 254,721 images (**MSCOCO** and **abstract scenes**)
  - 3 questions per image (764,163 total)
  - 10 ground truth answers per question
  - 3 plausible (but likely incorrect) answers per question
  - Open-ended and multiple-choice answering tasks
- **Winner (UC Berkeley & Sony) 66.9% accuracy on real-image open-ended QA. Naver, Samsung, SNU, Postech (현재 SNU 1등 중)**

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Cambot (SNU)

[Kim et al., NIPS-2016]

Deep Hypernets for Visual Dialogue

- **Deep hypernetworks with hierarchical concept structure are used as knowledge base for Q&A**

Hierarchical formulation

$$P(\mathrm{x}) = \sum_{\mathrm{h}_n} \cdots \sum_{\mathrm{h}_1} P(\mathrm{h}_n \mid \mathrm{h}_{n-1}) \cdots P(\mathrm{h}_2 \mid \mathrm{h}_1) P(\mathrm{h}_1 \mid \mathrm{x}) P(\mathrm{x})$$

Joint probability of hidden variables $\mathbf{h}_i^{(s)}$ in the s^{th} layer

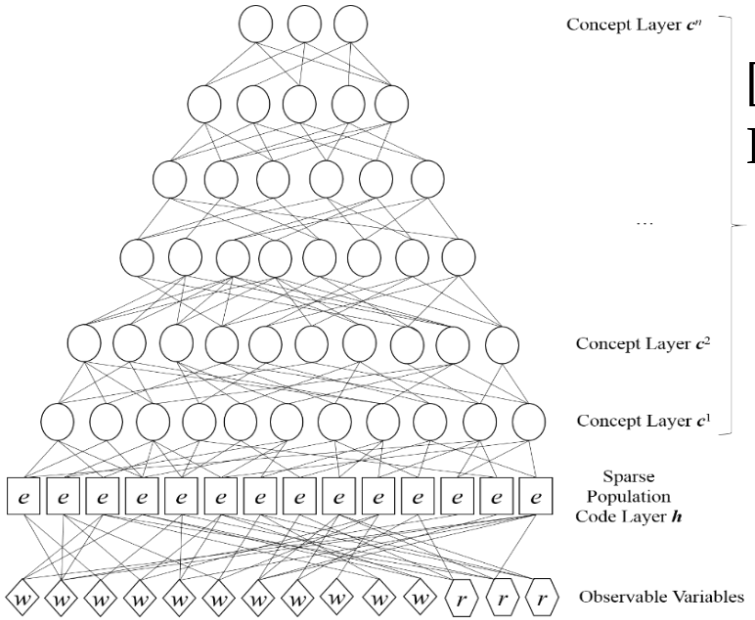
$$P(\mathbf{h}_s) = \frac{\exp(-E(\mathbf{h}_s))}{\sum_j \exp(-E(\mathbf{h}_j))}$$

$$E(\mathbf{h}_j) = h(s(\mathbf{h}_j))$$

$$s(\mathbf{h}_j) = \sum_{i1} w_{i1}^{(j)} h_{i1}^{(j)} + \sum_{i1,i2} w_{i1i2}^{(j)} h_{i1}^{(j)} h_{i2}^{(j)} + \ldots + \sum_{i1,i2,\ldots ik} w_{i1i2}^{(j)} h_{i1}^{(j)} \cdots h_{ik}^{(j)}$$

Learning is done by adjusting $s(h_j)$ towards maximizing likelihood $P(x|W)$

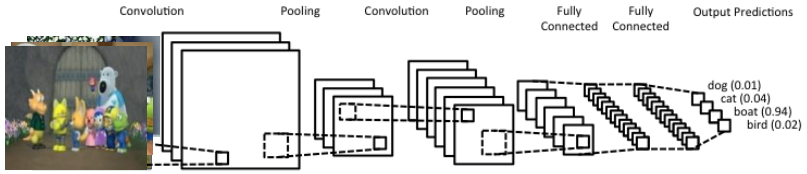
$$\frac{\nabla}{\nabla w_{i_1,i_2,\ldots,i_k}^{(k)}} \ln \prod_{n=1}^{N} P(\boldsymbol{x}^{(n)} \mid W) = N \left\{ \left[x_{i_1} x_{i_2} \ldots x_{i_k} \right]_{Data} - \left[x_{i_1} x_{i_2} \ldots x_{i_k} \right]_{P(x|W)} \right\}$$



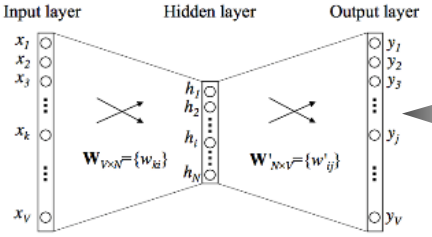
[Kim et al., IJCAI-2017]

Knowledge Base Construction

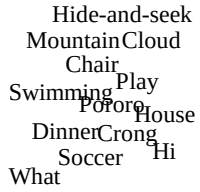
Fc7 Feature of Convolutional Neural Networks



Neural Word Embedding Using Word2vec



Preprocessing

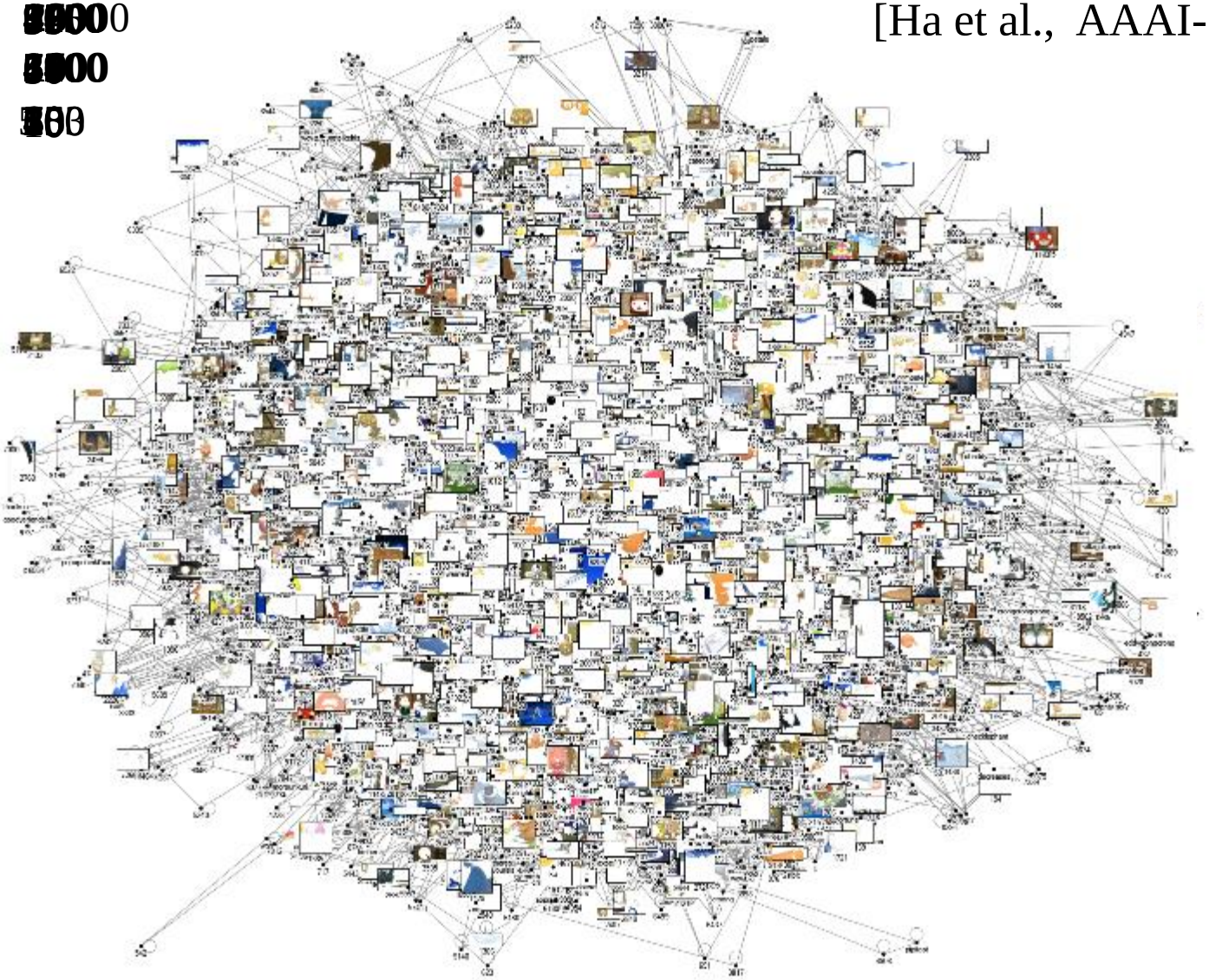


1. K.-M. Kim, C.-J. Nan, J.-W. Ha, Y.-J. Heo, and B.-T. Zhang, "Pororobot: A Deep Learning Robot That Plays Video Q&A Games", *AAAI 2015 Fall Symposium on AI for Human-Robot Interaction* (*AI-HRI 2015*), 2015.
2. J.-W. Ha, K.-M. Kim, B.-T. Zhang, Automated Visual-Linguisitc Knowledge Construction via Concept Learning from Cartoon Videos," In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (*AAAI 2015*), 2015.
3. B.-T. Zhang, J.-W. Ha, M. Kang, Sparse Population Code Models of Word Learning in Concept Drift, In *Proceedings of Annual Meeting of the Cognitive Science Society* (*Cogsci*), 2012.

Learning from Cartoon Videos

Image 개수 : [illegible]0
Word 개수 : [illegible]
Episode 개수 : [illegible]

[Ha et al., AAAI-2015]



Pororobot (SNU)

[Ha et al., AAAI-2015]

AUPAIR: Autonomous Personal AI Robot

AUPAIR (SNU & Surromind Robotics)

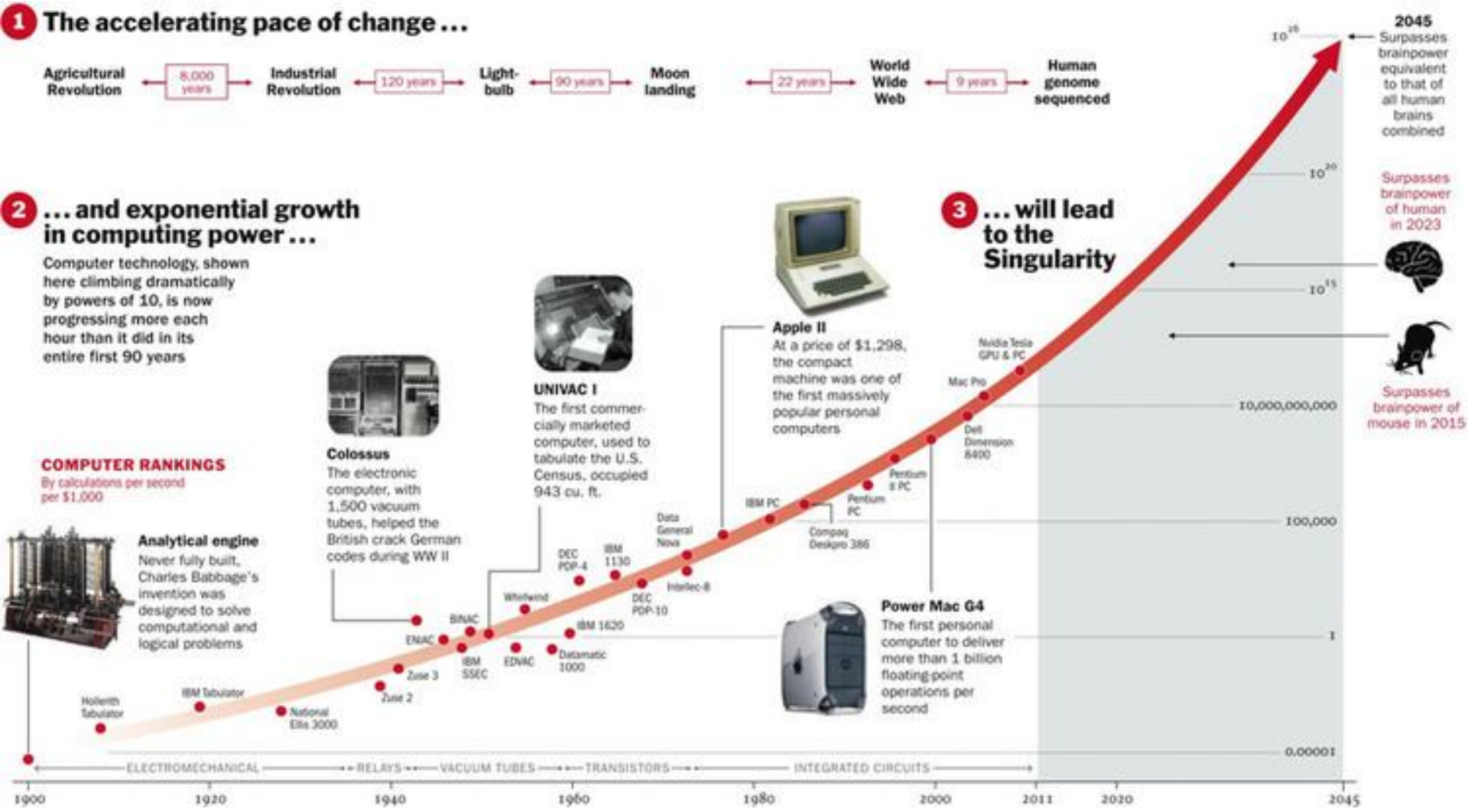
Winning the RoboCup@Home 2017

https://www.youtube.com/watch?v=a2ZJTpbMWsQ

http://mnews.joins.com/article/21823070#home

3. New AI

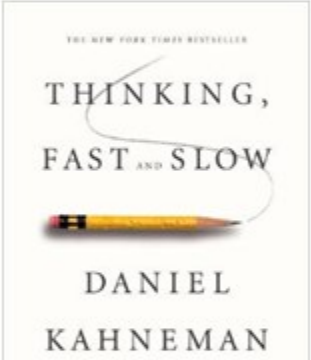
Human Intelligence and Artificial Intelligence

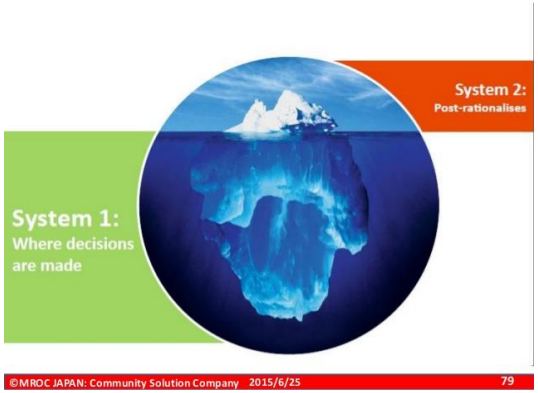


Dual Process Theories of Mind

Kahneman, *Thinking, Fast and Slow* (2011)

- System 1
  - Automatic, Fast, Unconscious
  - Heuristic, "Hot"
    - Emotions, Stereotypes
- System 2
  - Controlled, Slow, Conscious
  - Algorithmic, "Cold"
    - Logical, Systematic
- System 1 Usually Wins the Race





New AI (System 1) and Old AI (System 2)

System 1	System 2
Unconscious reasoning **New AI**	Conscious reasoning **Old AI**
Judgments based on intuition	Judgments based on critical examination
Processes information quickly	Processes information slowly
Hypothetical reasoning	Logical reasoning
Large capacity	Small capacity
Prominent in animals and humans	Prominent only in humans
Unrelated to working memory	Related to working memory
Operates effortlessly and automatically	Operates with effort and control
Unintentional thinking	Intentional thinking
Influenced by experiences, emotions, and memories	Influenced by facts, logic, and evidence
Can be overridden by System 2	Used when System 1 fails to form a logical/acceptable conclusion
Prominent since human origins	Developed over time
Includes recognition, perception, orientation, etc.	Includes rule following, comparisons, weighing of options, etc.

Humans (NI) and Machines (AI)

- Introspectionism - 1920

Psyche

- Behaviorism ←→ Cybernetics 1920 - 1950

Mind (= Computer)

- Cognitivism ←→ Symbolic AI 1950 - 1980

Brain

- Connectionism ←→ Neural Nets (ML) 1980 – 2010

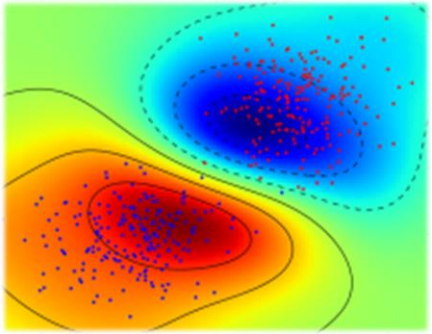
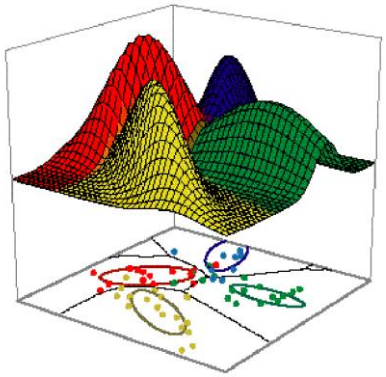
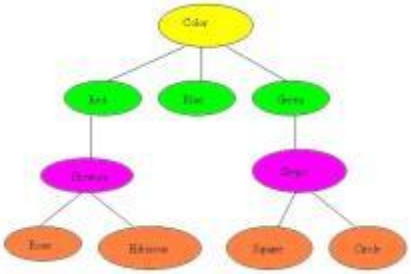
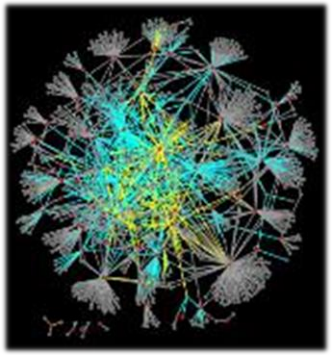
Body

- Action Science ←→ Autonomous Robots 2010 –

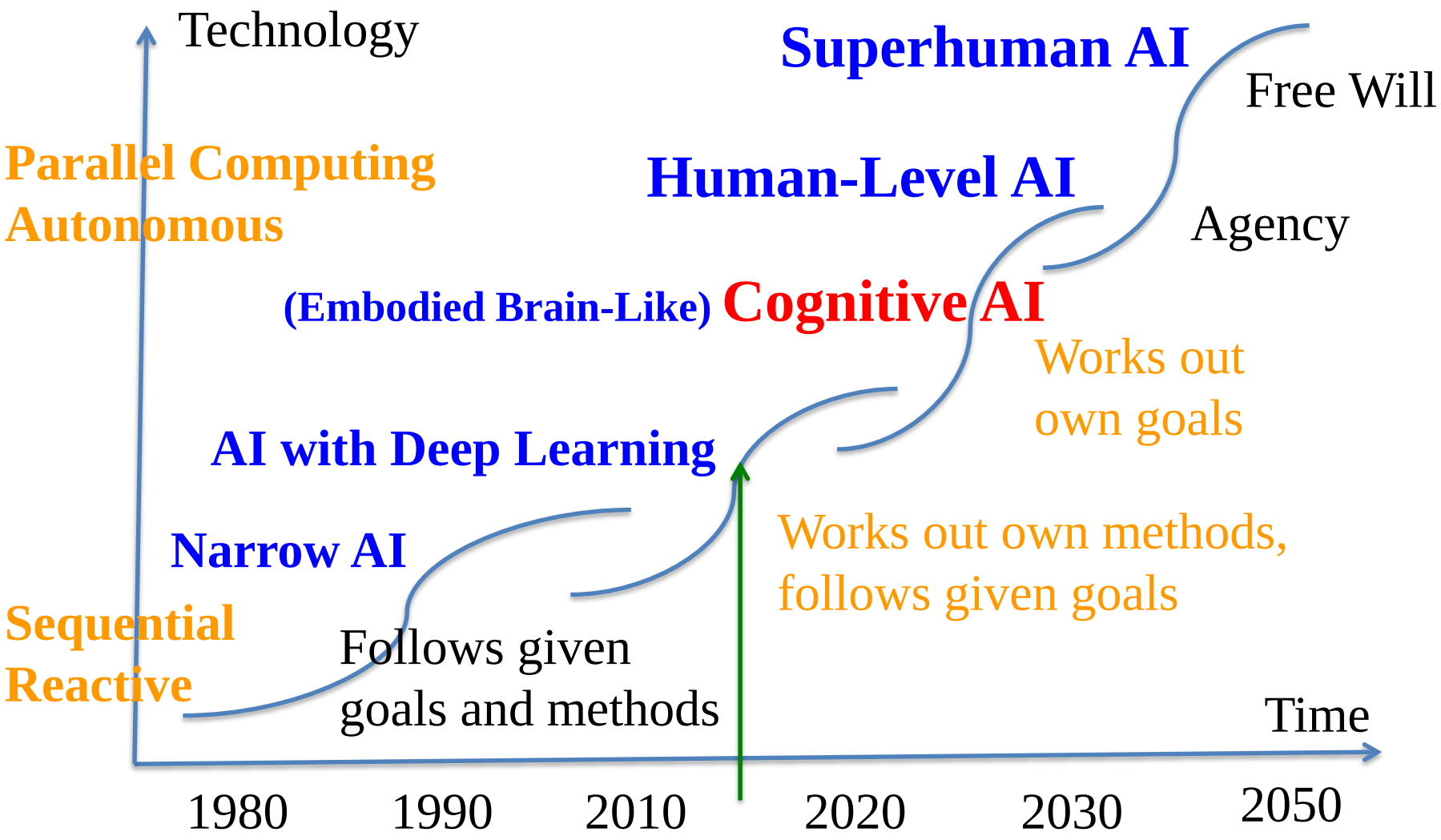
Environment

Embodied Mind | Mind Machine (= Smart Machine)

Autonomous Machine Learning

1G: Supervised Learning (1980s~2000)	2G: Unsupervised Learning (2000~Present)	3G: Autonomous Learning (Next Generation)
		
		
• Decision Trees • Kernel Methods • Multilayer Perceptrons	• Deep Networks • Markov Networks • Bayesian Networks	• Complex Adaptive Systems • Perception-Action Cycle • Lifelong Continual Learning

Future of AI



Modified from Eliezer Yudkowsky & David Wood