Introduction

Lecture 1

Hyung-Sin Kim

SNU Graduate School of Data Science


WELCOME

About This Bootcamp & Competition

Lecture 1-1

Hyung-Sin Kim

SNU Graduate School of Data Science

Instructors

hyungkim@snu.ac.kr https://aiot.snu.ac.kr/

서울대학교
SEOUL NATIONAL UNIVERSITY

Berkeley
UNIVERSITY OF CALIFORNIA

Google

Course Structure

- 2 review lectures
  - Video
- 1 intro session
  - Hybrid
- 7 main lectures
  - Video
- 7 practice session
  - Hybrid
- Group Project

		Type	Content	HW & Activity
Before 8/8	Self-study (Video provided)	**Lecture 0**	ML/DL review (2 hours)	Jupytor notebook, TensorFlow installation
2022. 8. 8 (Mon)	14:00 ~ 15:30 (942-302, hybrid)	**Lecture 1**	Introduction to Ambient AI	
	15:30 ~ 17:00 (942-302, hybrid)	**Practice 1**	TensorFlow (1) - Introduction to TensorFlow and Implementation of an MLP	**HW #1**: Training an MLP with TensorFlow (~8/10)
2022. 8. 10 (Wed)	Self-study (Video provided)	**Lecture 2**	CNN for 2D object classification	
	14:00 ~ 15:30 (942-302, hybrid)	**Practice 2**	TensorFlow (2) - Implementation of a CNN	**HW #2**: Training a CNN with TensorFlow (~8/12)
2022. 8. 12 (Fri)	Self-study (Video provided)	**Lecture 3**	Lightweight CNN for 2D object classification	
	14:00 ~ 15:30 (942-302, hybrid)	**Practice 3**	TensorFlow (3) - Implementation of a more complex CNN	**HW #3**: Training a more complex CNN with TensorFlow (~8/19)
2022. 8. 17 (Wed)	Self-study (Video provided)	**Lecture 4**	Transfer Learning, Two-stage 2D object detectors	
		Lecture 5	One-stage 2D object detectors	
2022. 8. 19 (Fri)	14:00 ~ 17:00 (942-302, hybrid)	**Practice 4**	TensorFlow (4) - Implementation of an Object Detection Model	
2022. 8. 22 (Mon)	Self-study (Video provided)	**Lecture 6**	DNN Quantization (1)	
		Lecture 7	DNN Quantization (2)	
		Lecture 8	DNN Pruning	
2022. 8. 24 (Wed)	14:00 ~ 15:30 (942-302, hybrid)	**Practice 5**	TensorFlow Lite - Quantization and pruning of a heavy DNN	
	15:30 ~ 17:00 (942-302, hybrid)	**Practice 6**	Coral and edge TPU (Inference on the edge)	
2022. 8. 26 (Fri)	14:00 ~ 17:00 (942-302, hybrid)	**Practice 7**	Biggest face detection on Coral	
2022. 9. 16 (Fri)	14:00	**Project Presentation**		

Tools

TensorFlow

TensorFlow Lite

Group-based Study, If You Want

- https://forms.gle/ZLdZX6meQobbhq9r7

Data Science

Lecture 1-2

Hyung-Sin Kim


SNU Graduate School of Data Science

Let’s demystify Data Science!

- SNU GSDS view -

We Want Insight

SNULife

에브리타임
서울대

Useful insight!

Data Science – From Data to Insight

Big, Diverse Data

	A	B	C	D
1	Date	Apples	Oranges	Total Fruit
2	6/1/2012	125	75	200
3	6/2/2012	118	84	202
4	6/3/2012	164	72	236
5	6/4/2012	114	65	179
6	6/5/2012	98	96	194
7	6/6/2012	172	82	254

Useful insight!

Data Science

Data Science – From Data to Insight

Big, Diverse, **Dirty** Data

	A	B	C	D
1	Date	Apples	Oranges	Total Fruit
2	6/1/2012	125	75	200
3	6/2/2012	118	84	202
4	6/3/2012	164	72	236
5	6/4/2012	114	65	179
6	6/5/2012	98	96	194
7	6/6/2012	172	82	254

Useful insight!

Data Science

Data Science – From Data to Insight

Big, Diverse, **Dirty** Data

	A	B	C	D
1	Date	Apples	Oranges	Total Fruit
2	6/1/2012	125	75	200
3	6/2/2012	118	84	202
4	6/3/2012	164	72	236
5	6/4/2012	114	65	179
6	6/5/2012	98	96	194
7	6/6/2012	172	82	254

Useful insight!

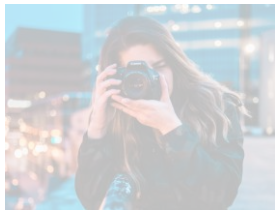
Storing and Managing – Data Lake

Cleansing and Pre-processing

Analysis

Data Science – Four Pillars (ABC+D)

Big, Diverse, Dirty Data

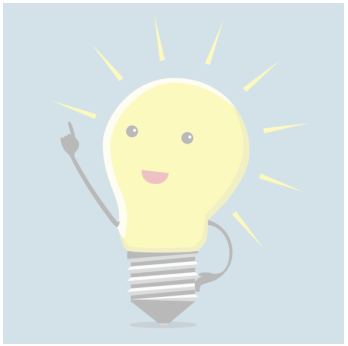


	A	B	C	D
1	Date	Apples	Oranges	Total Fruit
2	6/1/2012	125	75	200
3	6/2/2012	118	84	202
4	6/3/2012	164	72	236
5	6/4/2012	114	65	179
6	6/5/2012	98	96	194
7	6/6/2012	172	82	254

Domain (application)

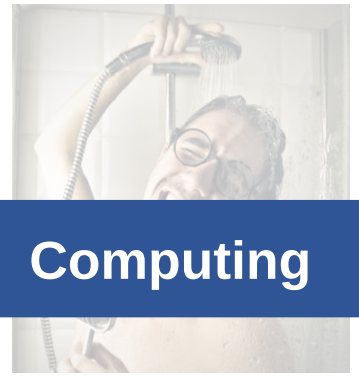
Where data comes from…
Where insights are applied to…

Useful insight!

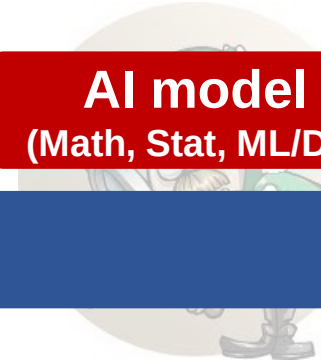


Big Data

Storing and Managing – Data Lake



Cleansing and Pre-processing

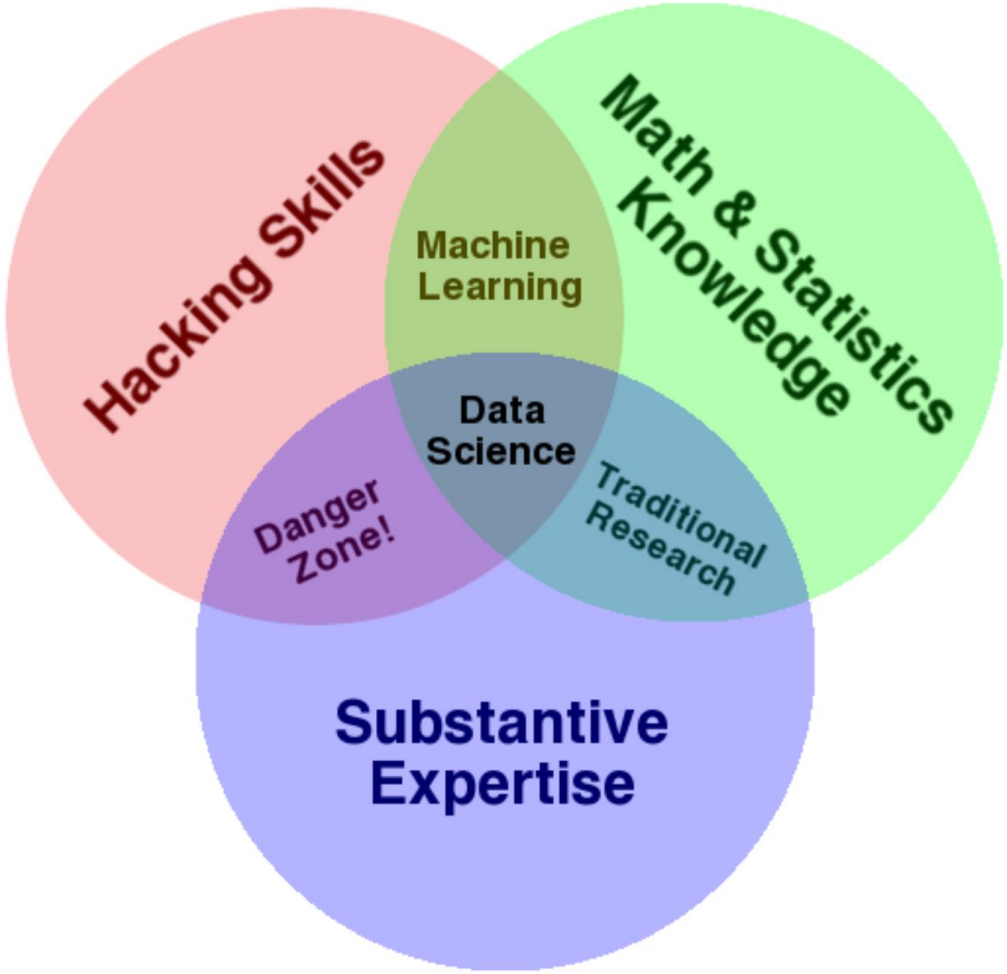


AI model
(Math, Stat, ML/DL)

Analysis

Computing

Drew Conway's Venn Diagram (2010)



[The image is from http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram]

Ullman's Venn Diagram (2021)

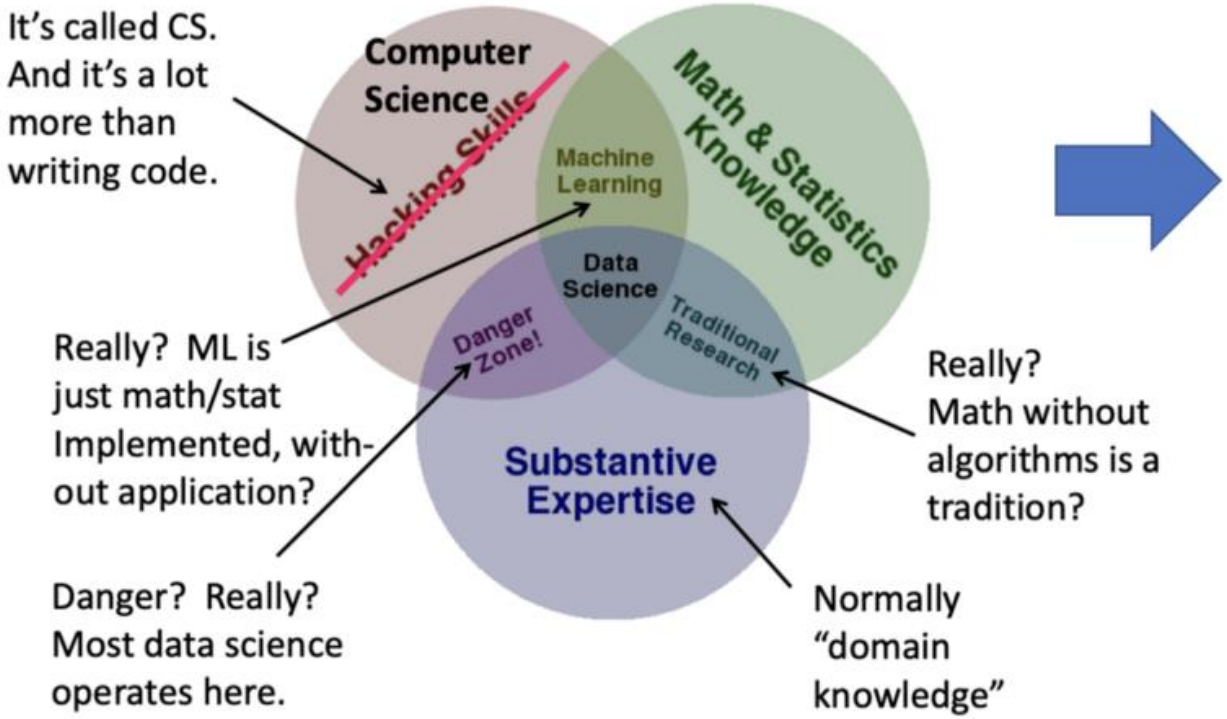
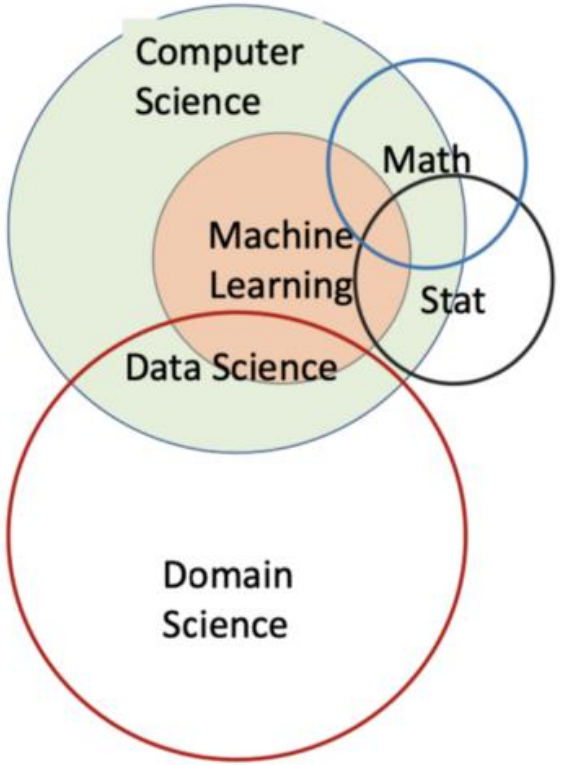
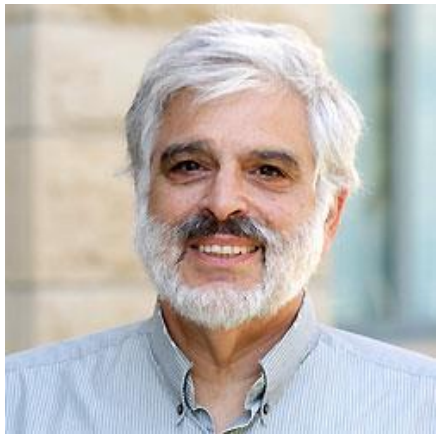


Figure 1: The Conway Venn diagram for data science

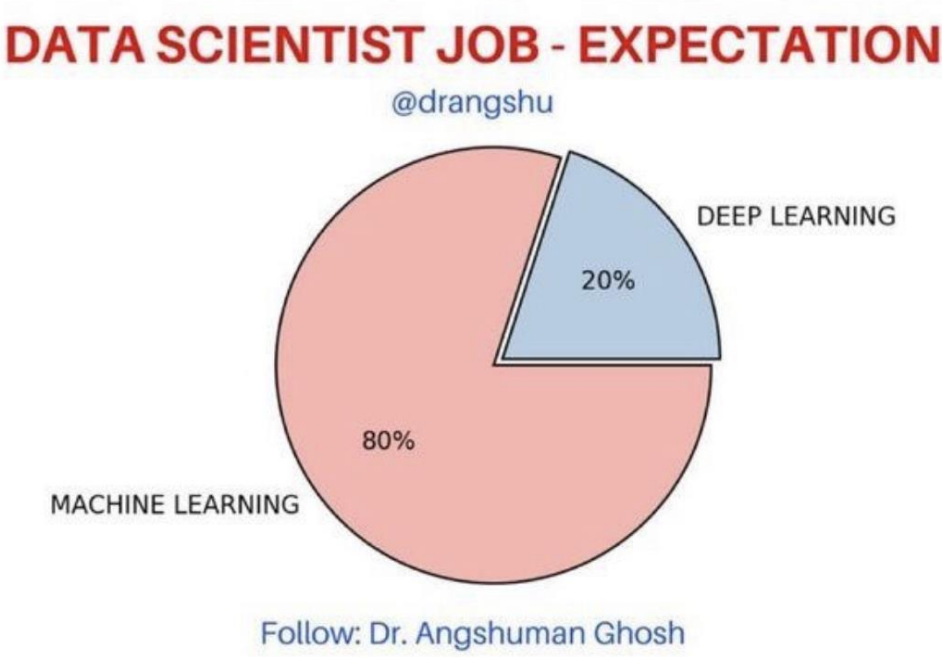


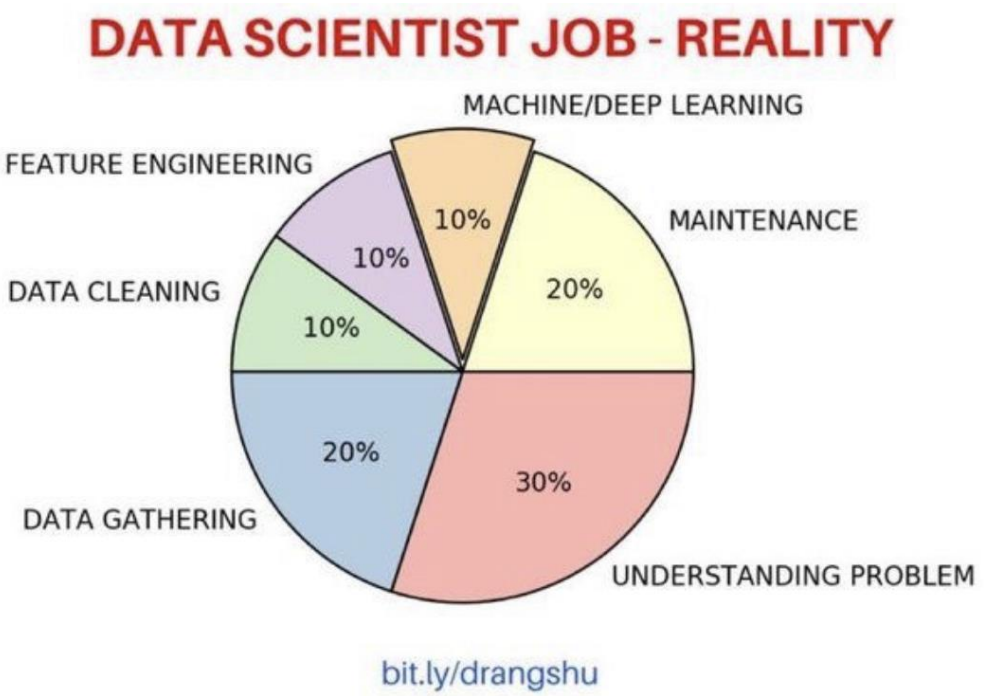
Bubbles not to scale



[The figures are from "The Battle for Data Science," http://sites.computer.org/debull/A20june/p8.pdf]

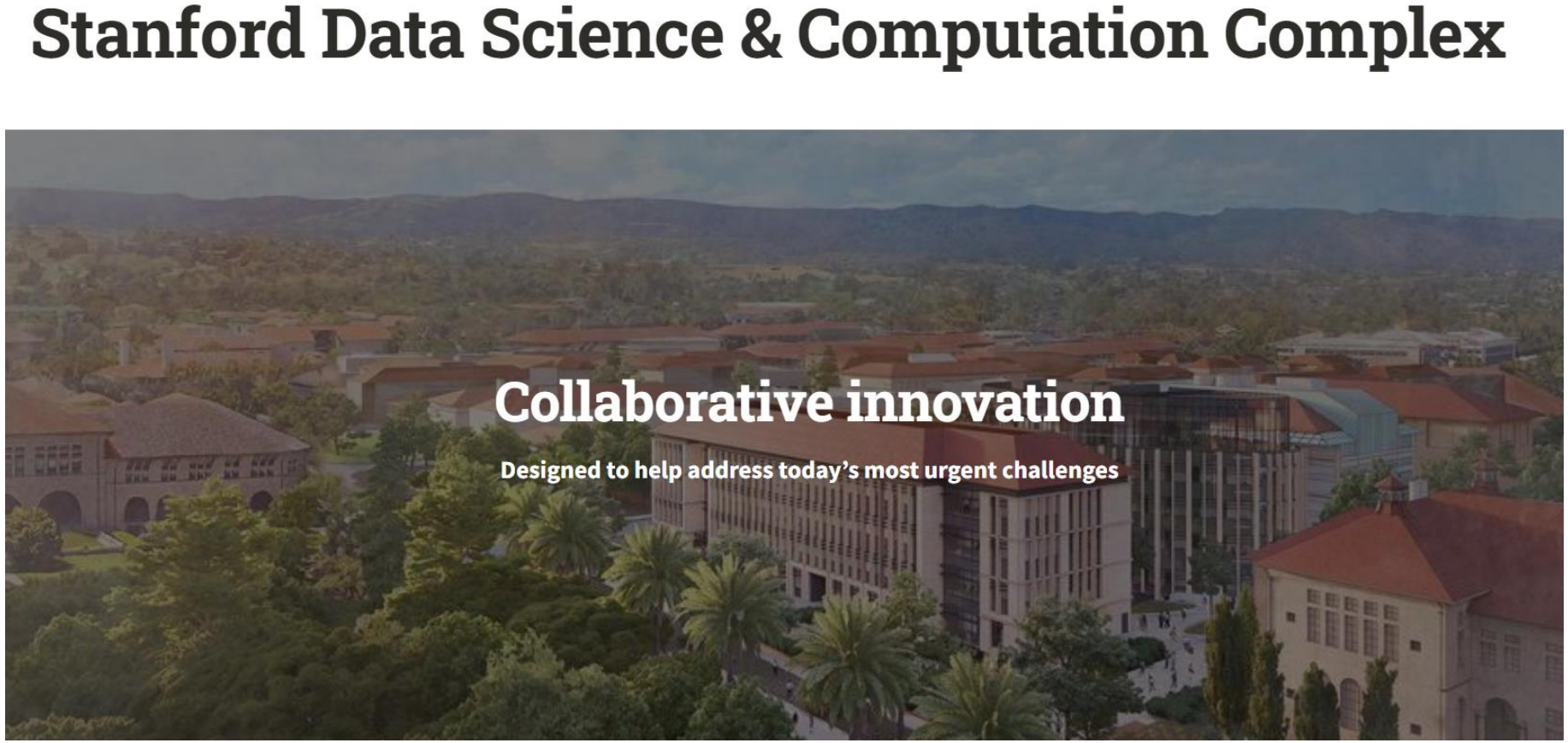
Data Scientist Job …





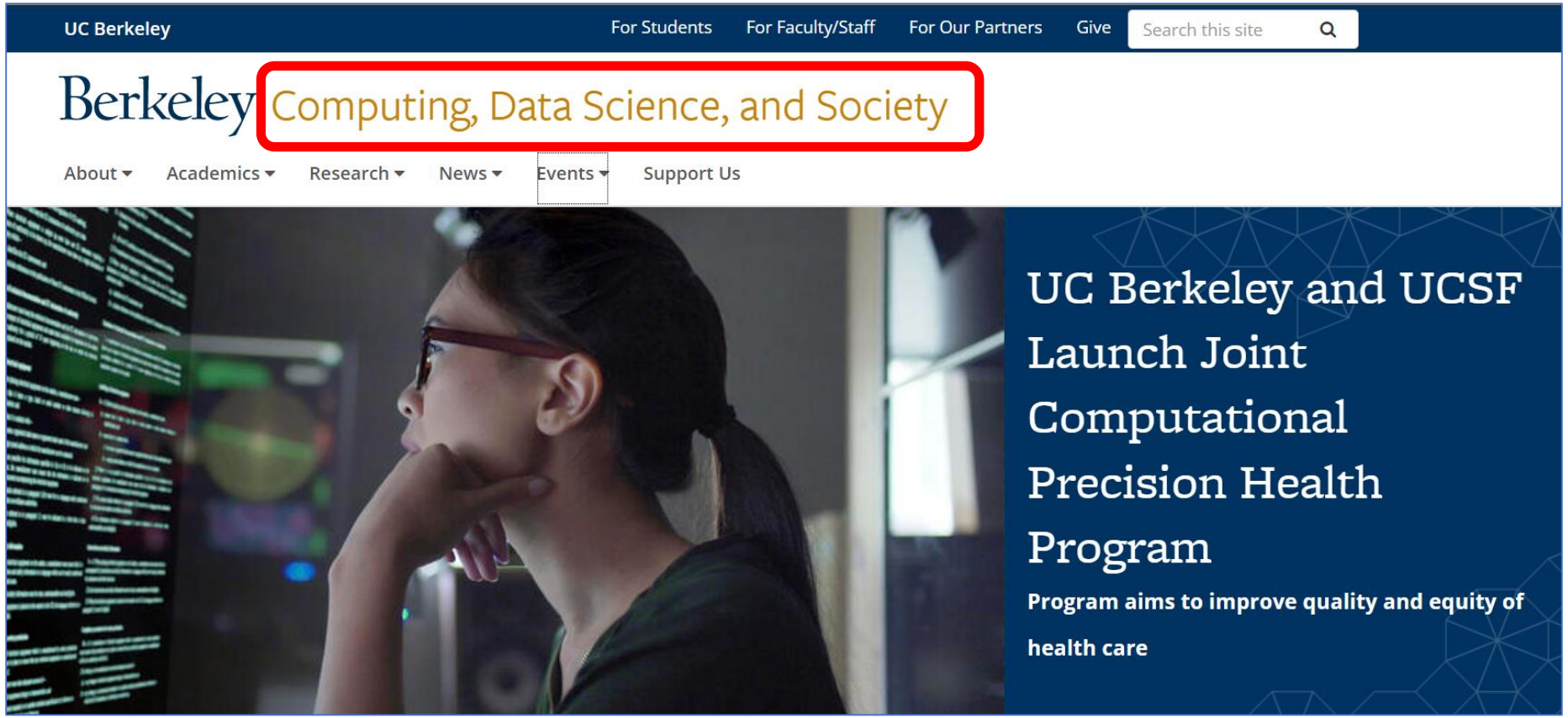
Data Science, a Global Megatrend

- https://www.youtube.com/watch?v=2ZopVhw6t3Y



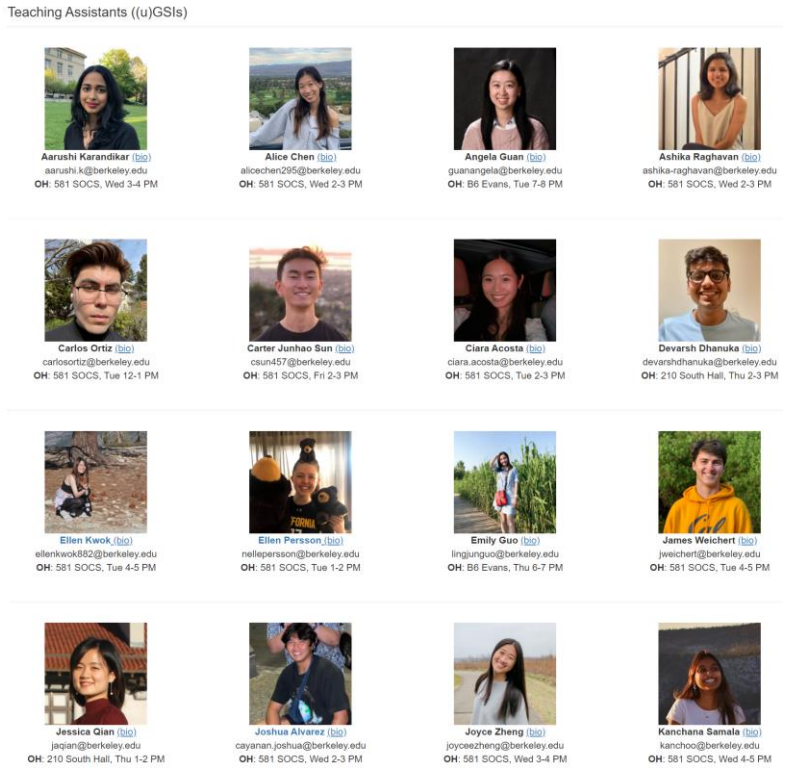
Data Science, a Global Megatrend

- https://www.youtube.com/watch?v=9Hx5FppPVso



Data Science, a Global Megatrend

- In UC Berkeley
  - 6,000 undergraduate students take DS courses per year
  - ~2,000 students and 76 TAs for a single course (Data 8, Spring 2022)



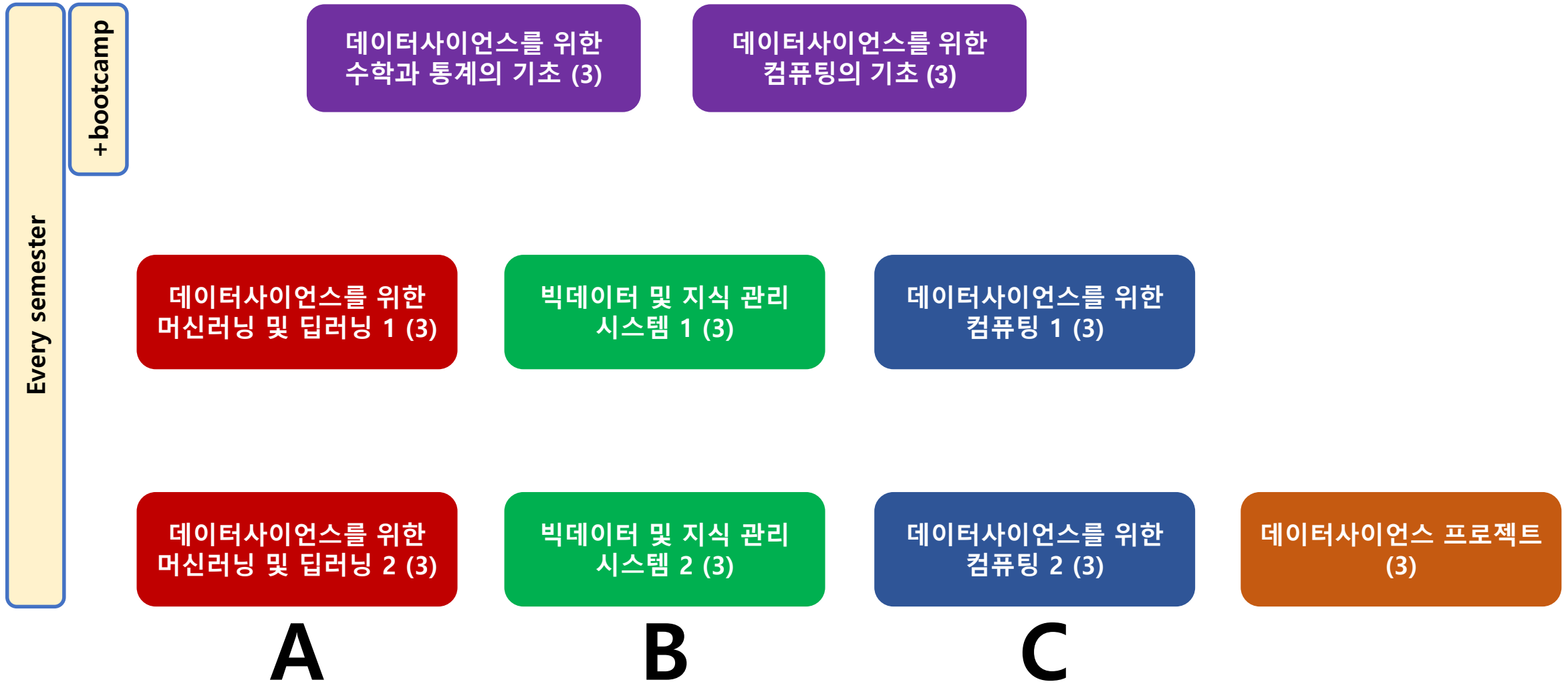
Graduate School of Data Science

GSDS – Vision and Mission

- Make students from **various backgrounds (D)** dive into **core principles (ABC)** of data science and let these **ambidexters** lead data-driven **innovation** in various fields
  - Globally unique mission… challenging of course…

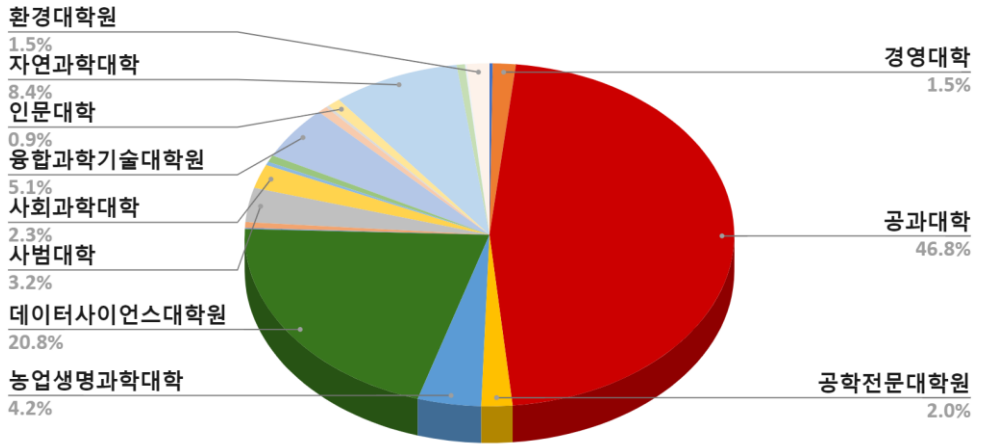
- Methodologies
  - Not an undergraduate department as a silo but a **graduate** school as a **hub**
  - Not an MBA-ish but a **hardcore** program to change students' DNA
  - Not only advanced but also **basic** courses that are open to all students outside of GSDS (no number limit)
  - Intensive labor of faculty members… ☺

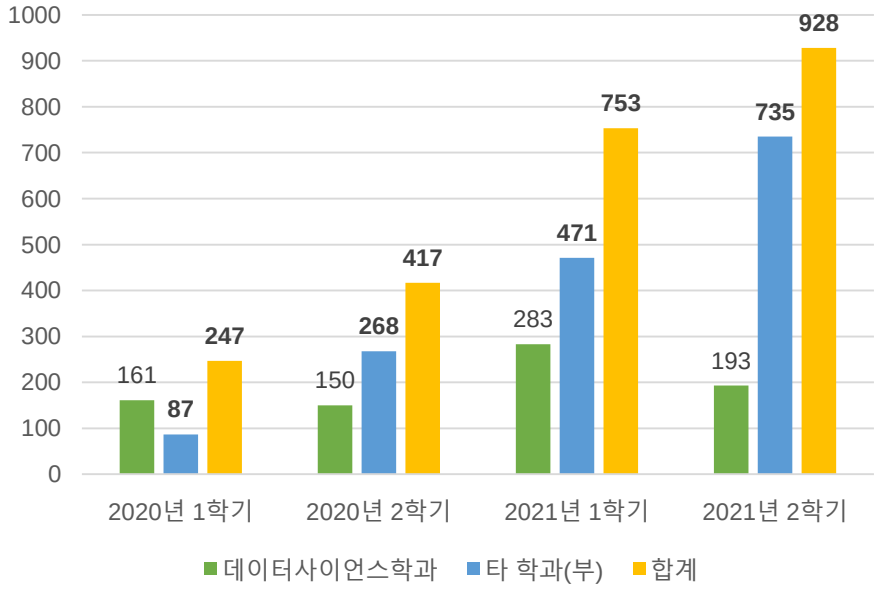
GSDS – Curriculum



GSDS – Growth

- 7 new faculty members have joined
  - Google, Amazon, US faculty…
- 40 MS/15 PhD to 80 MS/30 PhD per year
- Being recognized gradually…





Ambient AI – What and Why

Lecture 1-3

Hyung-Sin Kim

SNU Graduate School of Data Science

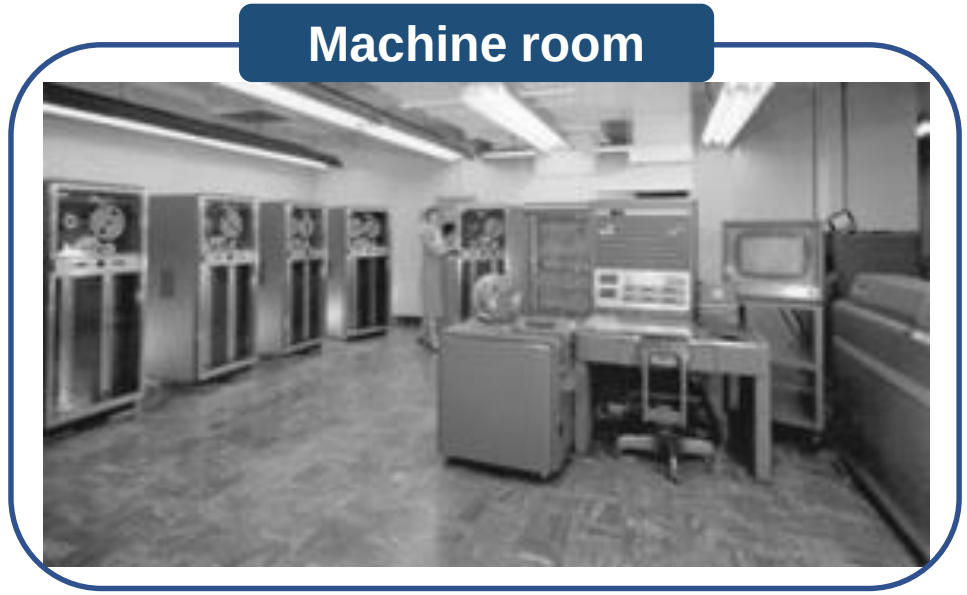
History of Computing Paradigm

History of Computing Paradigm – 1960 ~ 1980

A huge workstation

Intra-net

I/O (typing and printing)



History of Computing Paradigm – 1980 ~ 2000

A huge workstation

Machine room

Personal Computer

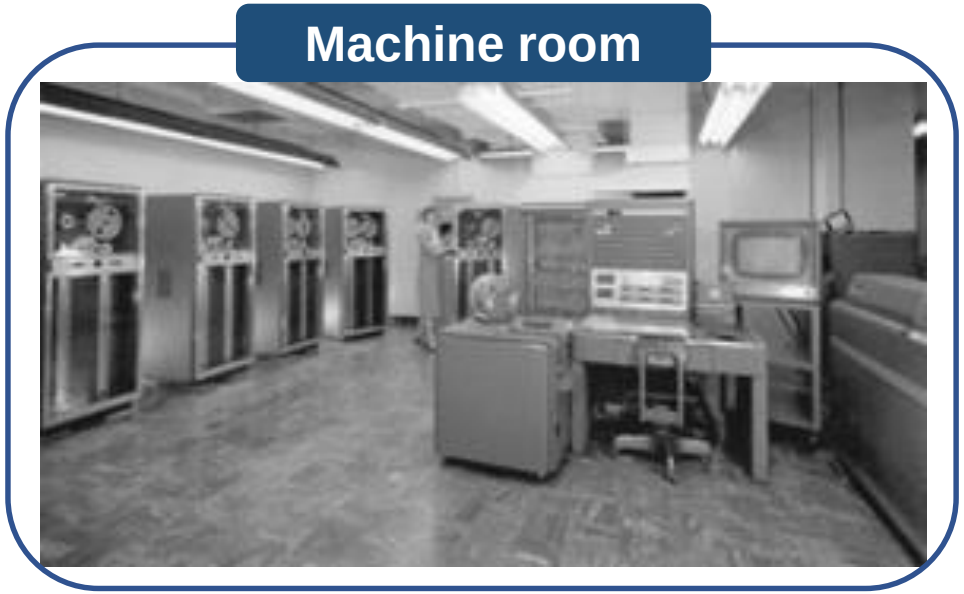
History of Computing Paradigm – 2000 ~ 2010

A huge workstation

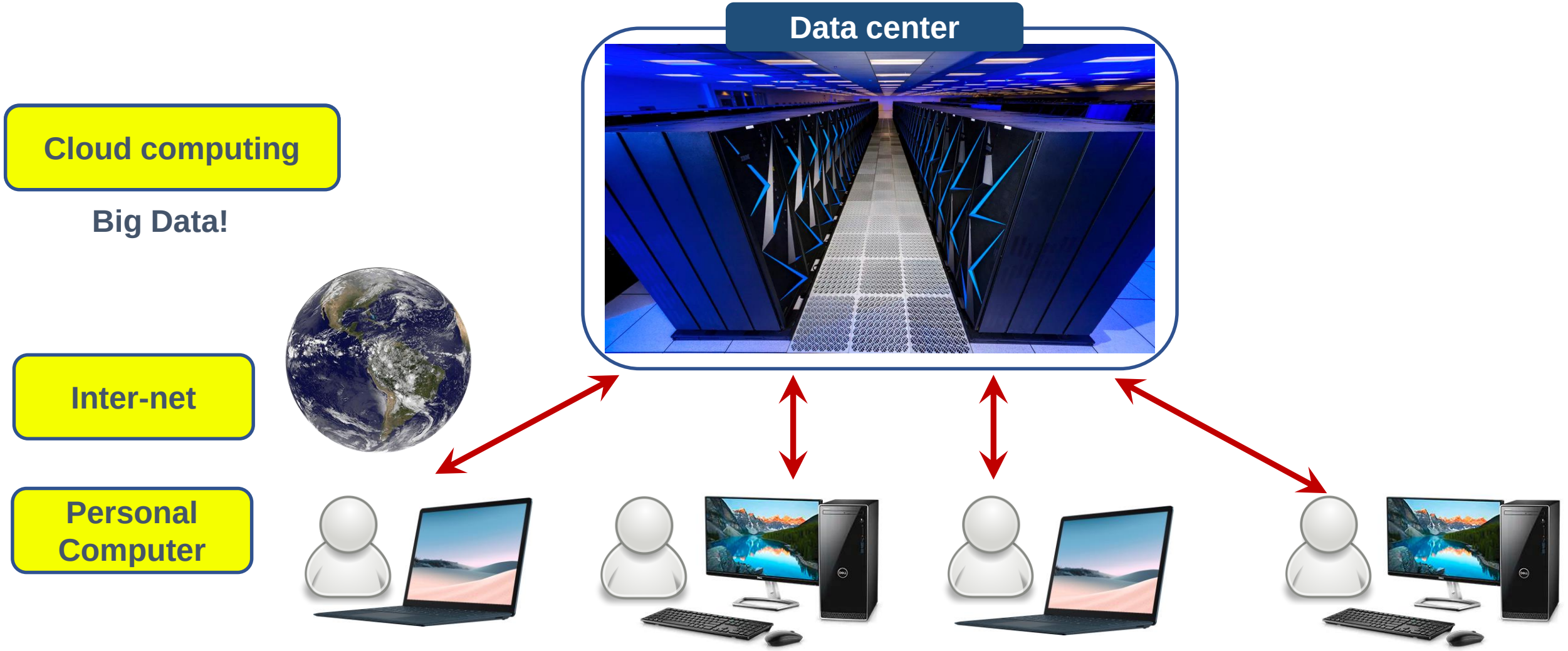
Big Data!

Inter-net

Personal Computer



History of Computing Paradigm – 2000 ~ 2010

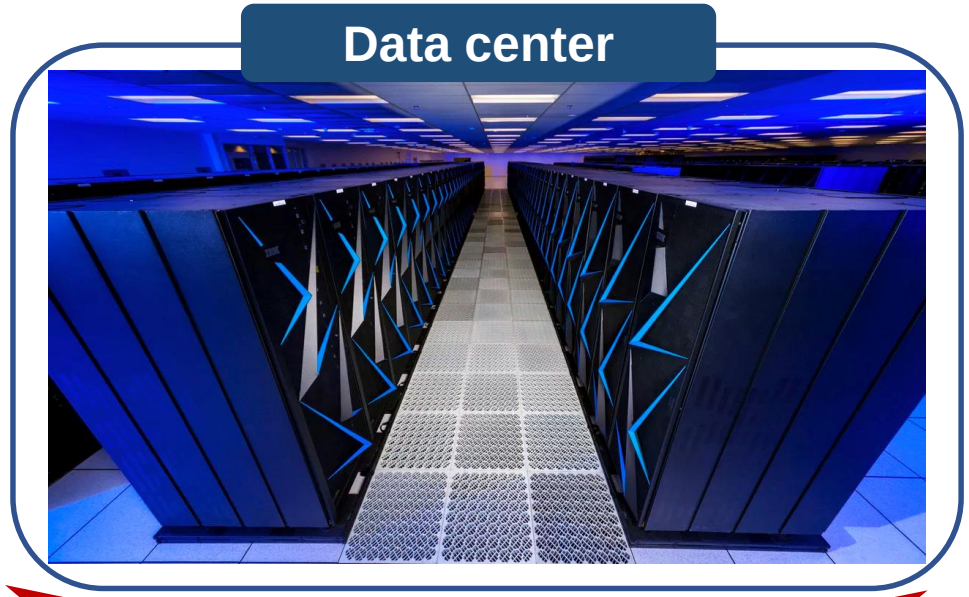


History of Computing Paradigm – 2010 ~ Present

Cloud computing

1) Privacy
2) Latency
3) Connectivity

Big Data!



Inter-net

Edge computing

Server (1960 – 1980)

Edge (1980 – 2000)

Server (2000 – 2010)

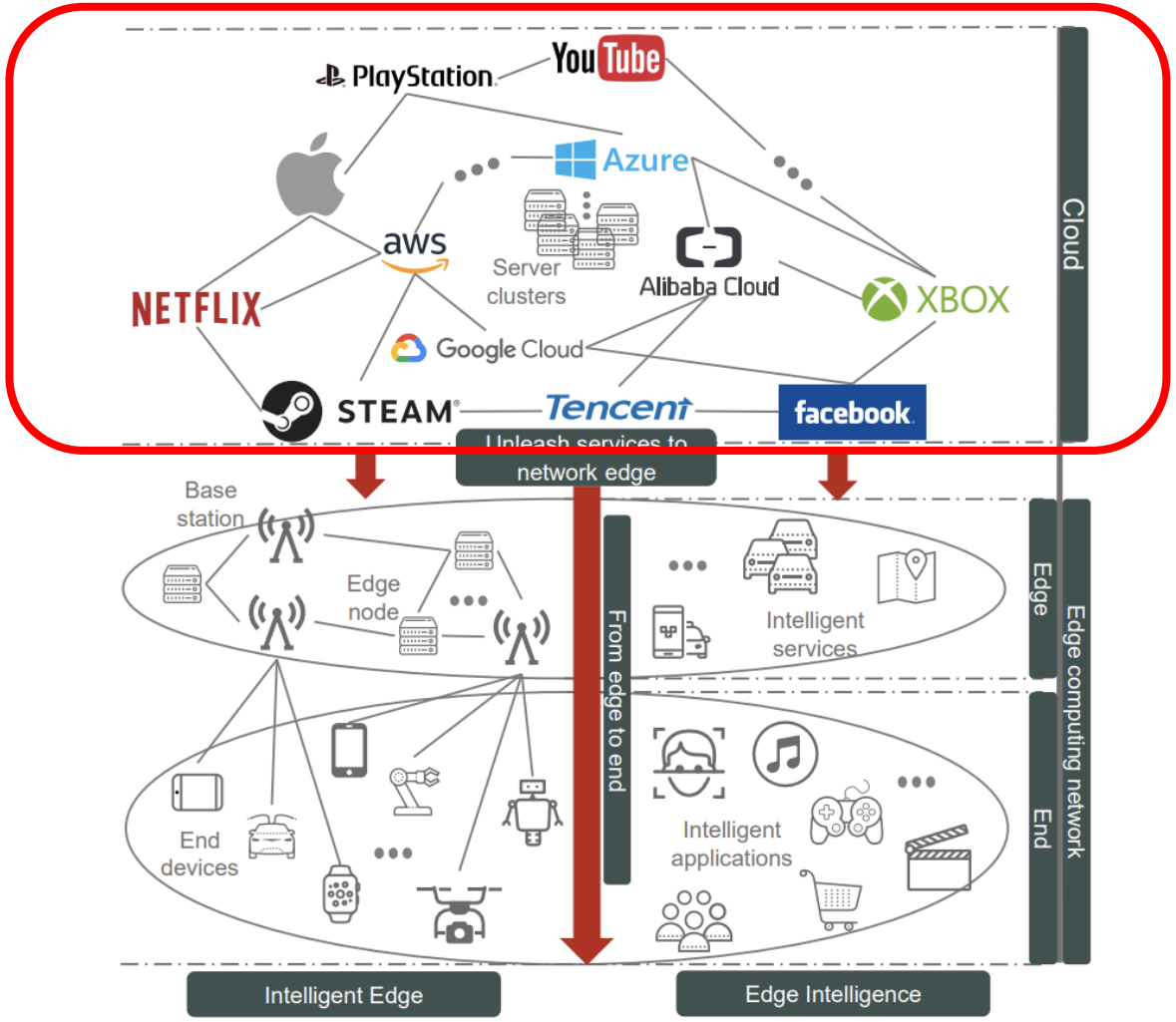
Server + Edge (2010 – Present)

Computing technology directly impacts AI paradigm

Ambient AI – Background (ML/DL)

ML / DL

2009~



Ambient AI – Background (ML/DL Algorithm)

- LeNet-5 in 1998
  - https://www.youtube.com/watch?v=FwFduRA_L6Q&ab_channel=YannLeCun

Gradient-Based Learning Applied to Document Recognition

YANN LECUN, MEMBER, IEEE, LÉON BOTTOU, YOSHUA BENGIO, AND PATRICK HAFFNER

Invited Paper

Multilayer neural networks trained with the back-propagation algorithm constitute the best example of a successful gradient-based learning technique. Given an appropriate network architecture, gradient-based learning algorithms can be used to synthesize a complex decision surface that can classify high-dimensional patterns, such as handwritten characters, with minimal preprocessing. This paper reviews various methods applied to handwritten character recognition and compares them on a standard handwritten digit recognition task. Convolutional neural networks, which are specifically designed to deal with the variability of two dimensional (2-D) shapes, are shown to outperform all other techniques.

Real-life document recognition systems are composed of multiple modules including field extraction, segmentation, recognition, and language modeling. A new learning paradigm, called graph transformer networks (GTN's), allows such multimodule systems to be trained globally using gradient-based methods so as to minimize an overall performance measure.

Two systems for online handwriting recognition are described. Experiments demonstrate the advantage of global training, and the flexibility of graph transformer networks.

A graph transformer network for reading a bank check is also described. It uses convolutional neural network character recognizers combined with global training techniques to provide record accuracy on business and personal checks. It is deployed commercially and reads several million checks per day.

Keywords— *Convolutional neural networks, document recognition, finite state transducers, gradient-based learning, graph transformer networks, machine learning, neural networks, optical character recognition (OCR).*

NOMENCLATURE

GT	Graph transformer.
GTN	Graph transformer network.
HMM	Hidden Markov model.
HOS	Heuristic oversegmentation.
K-NN	K-nearest neighbor.
NN	Neural network.
OCR	Optical character recognition.
PCA	Principal component analysis.
RBF	Radial basis function.
RS-SVM	Reduced-set support vector method.
SDNN	Space displacement neural network.
SVM	Support vector method.
TDNN	Time delay neural network.
V-SVM	Virtual support vector method.

Manuscript received November 1, 1997; revised April 17, 1998.

Y. LeCun, L. Bottou, and P. Haffner are with the Speech and Image Processing Services Research Laboratory, AT&T Labs-Research, Red Bank, NJ 07701 USA.

Y. Bengio is with the Département d'Informatique et de Recherche Opérationelle, Université de Montréal, Montréal, Québec H3C 3J7 Canada.

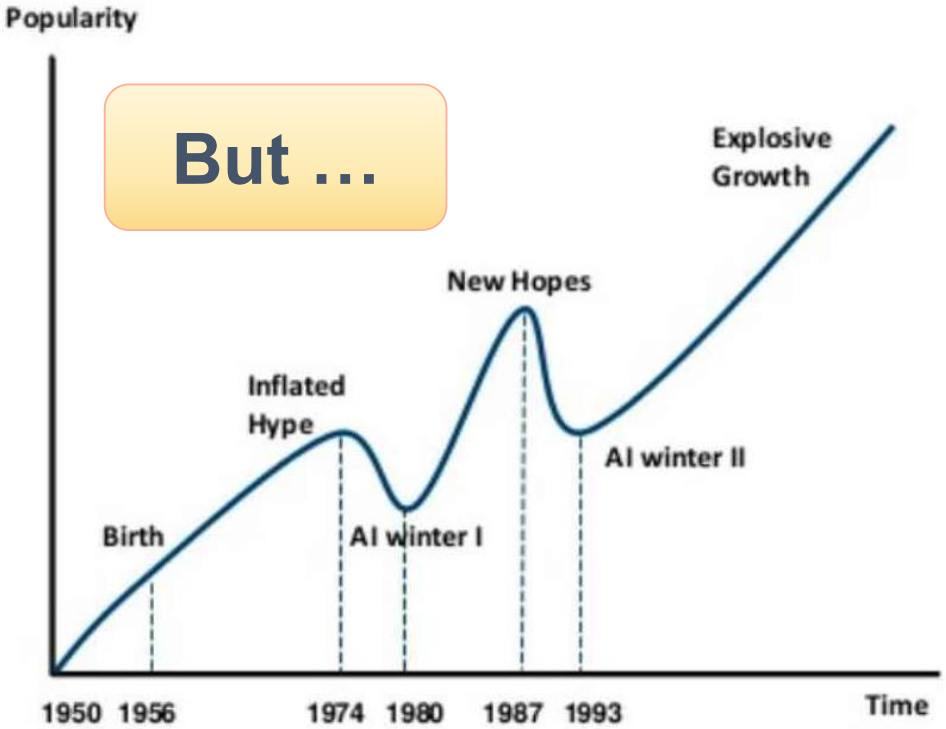
Publisher Item Identifier S 0018-9219(98)07863-3.

I. INTRODUCTION

Over the last several years, machine learning techniques, particularly when applied to NN's, have played an increasingly important role in the design of pattern recognition systems. In fact, it could be argued that the availability of learning techniques has been a crucial factor in the recent success of pattern recognition applications such as continuous speech recognition and handwriting recognition.

The main message of this paper is that better pattern recognition systems can be built by relying more on automatic learning and less on hand-designed heuristics. This is made possible by recent progress in machine learning and computer technology. Using character recognition as a case study, we show that hand-crafted feature extraction can be advantageously replaced by carefully designed learning machines that operate directly on pixel images. Using document understanding as a case study, we show that the traditional way of building recognition systems by manually integrating individually designed modules can be replaced by a unified and well-principled design paradigm, called GTN's, which allows training all the modules to optimize a global performance criterion.

Since the early days of pattern recognition it has been known that the variability and richness of natural data, be it speech, glyphs, or other types of patterns, make it almost impossible to build an accurate recognition system entirely by hand. Consequently, most pattern recognition systems are built using a combination of automatic learning techniques and hand-crafted algorithms. The usual method



Ambient AI – Background (ML/DL Computing)

Infrastructure to store and process Big Data

Google

NVIDIA

PyTorch

Google TensorFlow

aws

Microsoft Azure

Google Cloud

Ambient AI – Background (ML/DL Data)

- ImageNet in 2009

Benchmark Big dataset for anyone to play with

ImageNet: A Large-Scale Hierarchical Image Database

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei
Dept. of Computer Science, Princeton University, USA
{jiadeng, wdong, rsocher, jial, li, feifeili}@cs.princeton.edu

Abstract

The explosion of image data on the Internet has the potential to foster more sophisticated and robust models and algorithms to index, retrieve, organize and interact with images and multimedia data. But exactly how such data can be harnessed and organized remains a critical problem. We introduce here a new database called "ImageNet", a large-scale ontology of images built upon the backbone of the WordNet structure. ImageNet aims to populate the majority of the 80,000 synsets of WordNet with an average of 500-1000 clean and full resolution images. This will result in tens of millions of annotated images organized by the semantic hierarchy of WordNet. This paper offers a detailed analysis of ImageNet in its current state: 12 subtrees with 5247 synsets and 3.2 million images in total. We show that ImageNet is much larger in scale and diversity and much more accurate than the current image datasets. Constructing such a large-scale database is a challenging task. We describe the data collection scheme with Amazon Mechanical Turk. Lastly, we illustrate the usefulness of ImageNet through three simple applications in object recognition, image classification and automatic object clustering. We hope that the scale, accuracy, diversity and hierarchical structure of ImageNet can offer unparalleled opportunities to researchers in the computer vision community and beyond.

1. Introduction

The digital era has brought with it an enormous explosion of data. The latest estimations put a number of more than 3 billion photos on Flickr, a similar number of video clips on YouTube and an even larger number for images in the Google Image Search database. More sophisticated and robust models and algorithms can be proposed by exploiting these images, resulting in better applications for users to index, retrieve, organize and interact with these data. But exactly how such data can be utilized and organized is a problem yet to be solved. In this paper, we introduce a new image database called "ImageNet", a large-scale ontology of images. We believe that *a large-scale ontology of images is a critical resource for developing advanced, large-scale content-based image search and image understanding algorithms, as well as for providing critical training and benchmarking data for such algorithms.*

ImageNet uses the hierarchical structure of WordNet [9]. Each meaningful concept in WordNet, possibly described by multiple words or word phrases, is called a "synonym set" or "synset". There are around 80,000 noun synsets in WordNet. In ImageNet, we aim to provide on average 500-1000 images to illustrate each synset. Images of each concept are quality-controlled and human-annotated as described in Sec. 3.2. ImageNet, therefore, will offer tens of millions of cleanly sorted images. In this paper, we report the current version of ImageNet, consisting of 12 "subtrees": *mammal, bird, fish, reptile, amphibian, vehicle, furniture, musical instrument, geological formation, tool, flower, fruit*. These subtrees contain 5247 synsets and 3.2 million images. Fig. 1 shows a snapshot of two branches of the mammal and vehicle subtrees. The database is publicly available at `http://www.image-net.org`.

The rest of the paper is organized as follows: *We first show that ImageNet is a large-scale, accurate and diverse image database* (Section 2). In Section 4, we present a few simple application examples by exploiting the current ImageNet, mostly the mammal and vehicle subtrees. Our goal is to show that *ImageNet can serve as a useful resource for visual recognition applications such as object recognition, image classification and object localization.* In addition, the construction of such a large-scale and high-quality database can no longer rely on traditional data collection methods. Sec. 3 describes how ImageNet is constructed by leveraging Amazon Mechanical Turk.

2. Properties of ImageNet

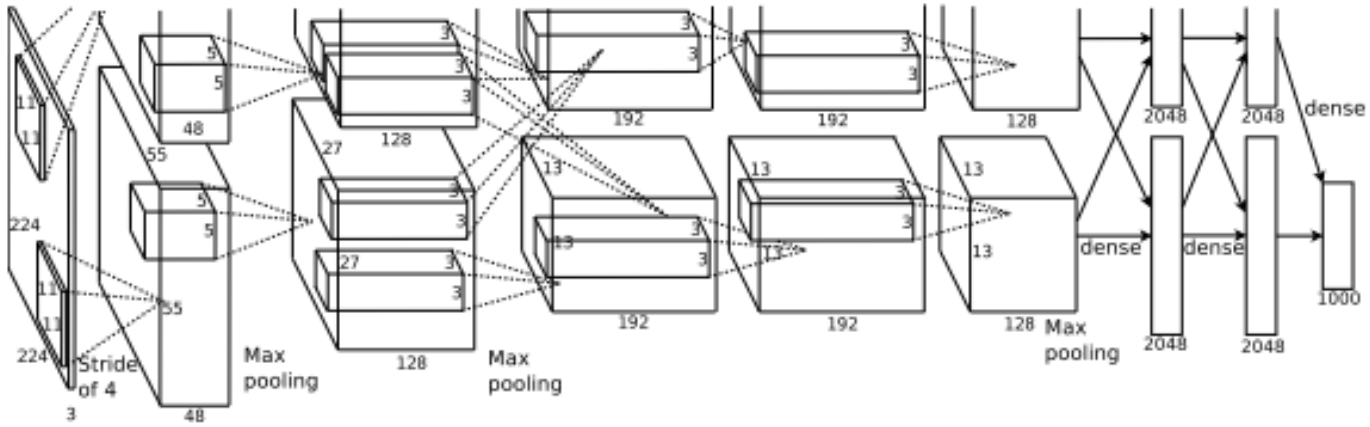
ImageNet is built upon the hierarchical structure provided by WordNet. In its completion, ImageNet aims to contain in the order of 50 million cleanly labeled full resolution images (500-1000 per synset). At the time this paper is written, ImageNet consists of 12 subtrees. Most analysis will be based on the mammal and vehicle subtrees.

Scale ImageNet aims to provide the most comprehensive and diverse coverage of the image world. The current 12 subtrees consist of a total of 3.2 million cleanly annotated

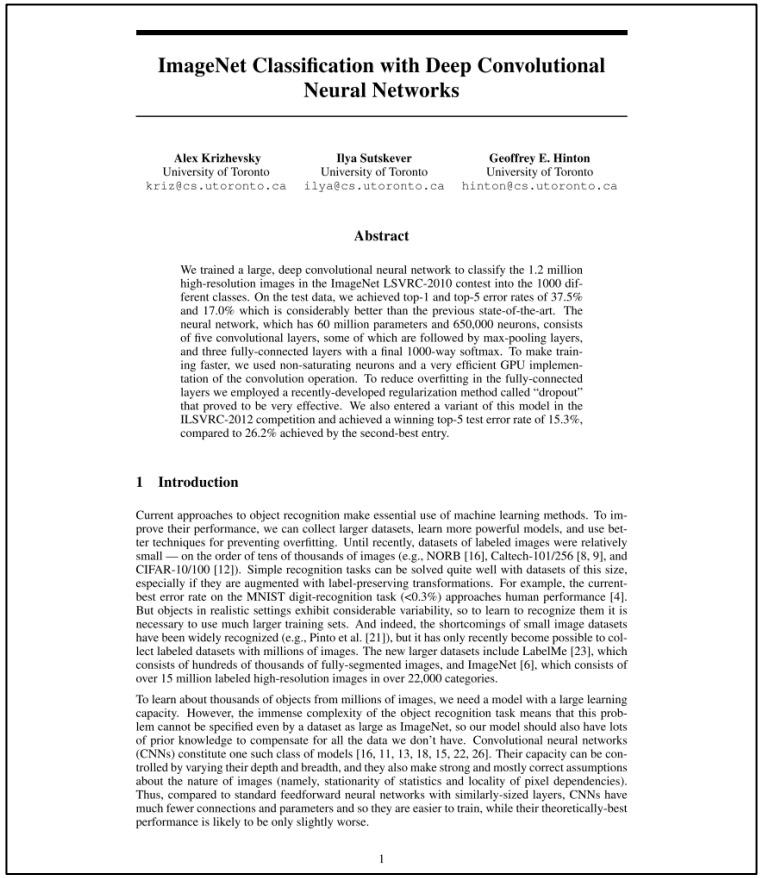
978-1-4244-3991-1/09/$25.00 ©2009 IEEE

Ambient AI – Background (ML/DL Algorithm)

- AlexNet in 2012
  - Convolutional Neural Network (CNN) works very well on ImageNet!
    - http://www.image-net.org/challenges/LSVRC/
  - DNN started receiving significant attention!
  - Cited more than 84k times…



Deep neural networks outperform feature extraction-based ML



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky University of Toronto kriz@cs.utoronto.ca
Ilya Sutskever University of Toronto ilya@cs.utoronto.ca
Geoffrey E. Hinton University of Toronto hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

1 Introduction

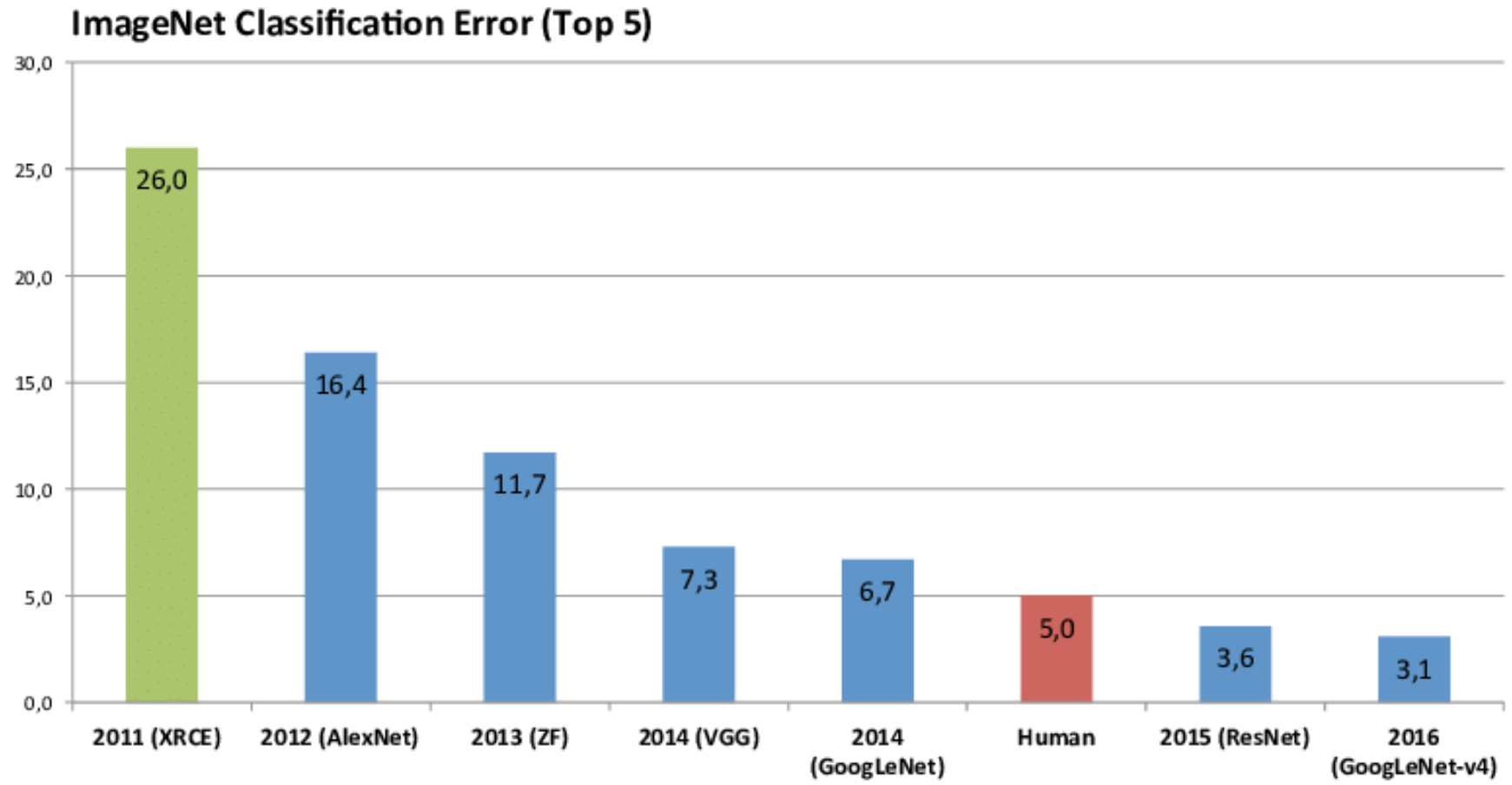
Current approaches to object recognition make essential use of machine learning methods. To improve their performance, we can collect larger datasets, learn more powerful models, and use better techniques for preventing overfitting. Until recently, datasets of labeled images were relatively small — on the order of tens of thousands of images (e.g., NORB [16], Caltech-101/256 [8, 9], and CIFAR-10/100 [12]). Simple recognition tasks can be solved quite well with datasets of this size, especially if they are augmented with label-preserving transformations. For example, the current-best error rate on the MNIST digit-recognition task (<0.3%) approaches human performance [4]. But objects in realistic settings exhibit considerable variability, so to learn to recognize them it is necessary to use much larger training sets. And indeed, the shortcomings of small image datasets have been widely recognized (e.g., Pinto et al. [21]), but it has only recently become possible to collect labeled datasets with millions of images. The new larger datasets include LabelMe [23], which consists of hundreds of thousands of fully-segmented images, and ImageNet [6], which consists of over 15 million labeled high-resolution images in over 22,000 categories.

To learn about thousands of objects from millions of images, we need a model with a large learning capacity. However, the immense complexity of the object recognition task means that this problem cannot be specified even by a dataset as large as ImageNet, so our model should also have lots of prior knowledge to compensate for all the data we don't have. Convolutional neural networks (CNNs) constitute one such class of models [16, 11, 13, 18, 15, 22, 26]. Their capacity can be controlled by varying their depth and breadth, and they also make strong and mostly correct assumptions about the nature of images (namely, stationarity of statistics and locality of pixel dependencies). Thus, compared to standard feedforward neural networks with similarly-sized layers, CNNs have much fewer connections and parameters and so they are easier to train, while their theoretically-best performance is likely to be only slightly worse.

1

Ambient AI – Background (ML/DL Algorithm)

- For some **specific** tasks, deep neural networks are better than human



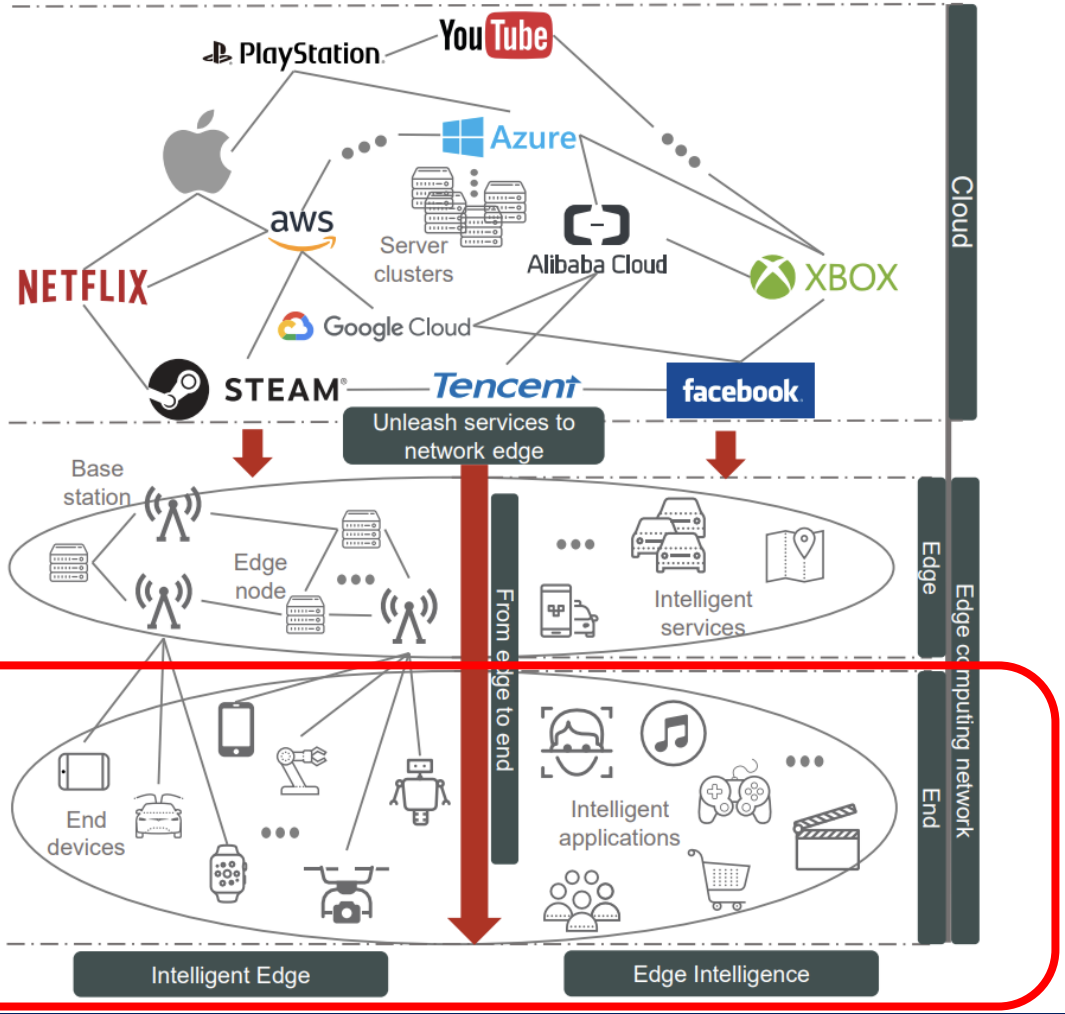
Ambient AI – Background (IoT)

ML / DL

2009~

Cyber-physical System (IoT)

1999~



Ambient AI – Background (IoT)

- Smart Dust in 1999
  - Computing, sensing, and networking are all possible in a small device!
  - 8-bit/4 MHz CPU, 0.5 kB RAM, 19.2 kbps

Next Century Challenges: Mobile Networking for "Smart Dust"

J. M. Kahn, R. H. Katz (ACM Fellow), K. S. J. Pister
Department of Electrical Engineering and Computer Sciences, University of California, Berkeley
{jmk, randy, pister}@eecs.berkeley.edu

Abstract

Large-scale networks of wireless sensors are becoming an active topic of research. Advances in hardware technology and engineering design have led to dramatic reductions in size, power consumption and cost for digital circuitry, wireless communications and Micro ElectroMechanical Systems (MEMS). This has enabled very compact, autonomous and mobile nodes, each containing one or more sensors, computation and communication capabilities, and a power supply. The missing ingredient is the networking and applications layers needed to harness this revolutionary capability into a complete system. We review the key elements of the emergent technology of "Smart Dust" and outline the research challenges they present to the mobile networking and systems community, which must provide coherent connectivity to large numbers of mobile network nodes co-located within a small volume.

1 Introduction

As the research community searches for the processing platform beyond the personal computer, networks of wireless sensors have become quite interesting as a new environment in which to seek research challenges. These have been enabled by the rapid convergence of three key technologies: digital circuitry, wireless communications, and Micro ElectroMechanical Systems (MEMS). In each area, advances in hardware technology and engineering design have led to reductions in size, power consumption, and cost. This has enabled remarkably compact, autonomous nodes, each containing one or more sensors, computation and communication capabilities, and a power supply.

Berkeley's *Smart Dust* project, led by Professors Pister and Kahn, explores the limits on size and power consumption in autonomous sensor nodes. Size reduction is paramount, to make the nodes as inexpensive and easy-to-deploy as possible. The research team is confident that they can incorporate the requisite sensing, communication, and computing hardware, along with a power supply, in a volume no more than a few cubic millimeters, while still achieving impressive performance in terms of sensor functionality and communications capability. These millimeter-scale nodes are called "Smart Dust." It is certainly within the realm of possibility that future prototypes of Smart Dust could be small enough to remain suspended in air, buoyed by air currents, sensing and communicating for hours or days on end. At least one popular science fiction book has articulated just such a vision [12].

In this paper, we are concerned with the networking and applications challenges presented by this radical new technology. These kinds of networking nodes must consume extremely low power, communicate at bit rates measured in kilobits per second, and potentially need to operate in high volumetric densities. These requirements dictate the need for novel ad hoc routing and media access solutions. Smart dust will enable an unusual range of applications, from sensor-rich "smart spaces" to self-identification and history tracking for virtually any kind of physical object.

The study of "Smart Dust systems" is very new. The main purpose of this paper is to present some of the technological opportunities and challenges, with the goal of getting more systems-level researchers interested in this critical area. The remainder of this paper is organized as follows. Section 2 presents an overview of the technology that underlies Smart Dust. Section 3 outlines the key networking challenges presented by this technology. In Section 4, we describe some of the potential applications of Smart Dust and the challenges they pose. Section 5 discusses related projects from the research community. Section 6 presents our summary and conclusions.

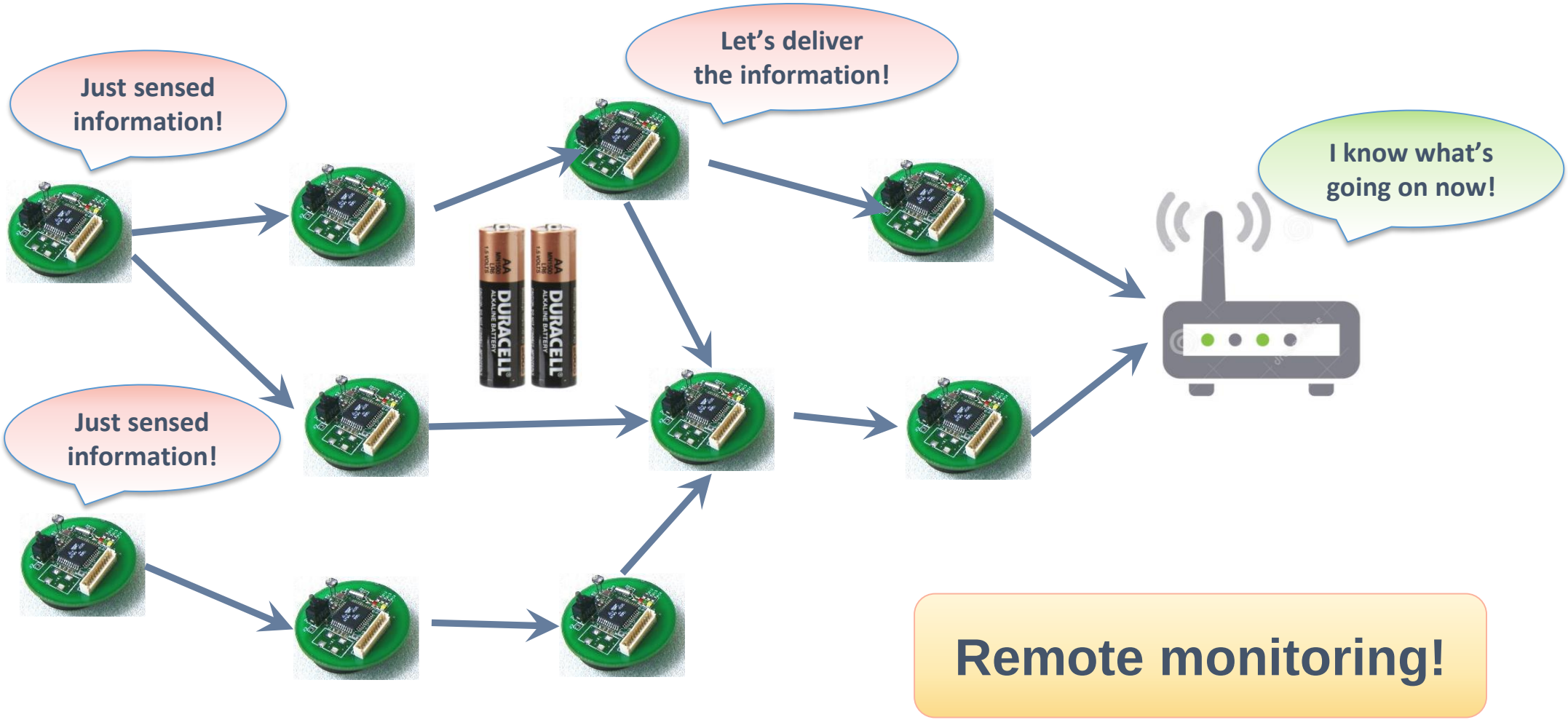
2 Smart Dust Technology

A Smart Dust *mote* is illustrated in Figure 1. Integrated into a single package are MEMS sensors, a semiconductor laser diode and MEMS beam-steering mirror for active optical transmission, a MEMS corner-cube retroreflector for passive optical transmission, an optical receiver, signal-processing and control circuitry, and a power source based on thick-film batteries and solar cells. This remarkable package has the ability to sense and communicate, and is self-powered!

A major challenge is to incorporate all these functions while maintaining very low power consumption, thereby maximizing operating life given the limited volume available for energy storage. Within the design goal of a cubic millimeter volume, using the best available battery technology, the total stored energy is on the order of 1 Joule. If this energy is consumed continuously over a day, the dust mote power consumption cannot exceed roughly 10 microwatts. The functionality envisioned for Smart Dust can be achieved only if the total power consumption of a dust mote is limited to microwatt levels, and if careful power management strategies are utilized (i.e., the various parts of the dust mote are

Mobicom '99 Seattle Washington USA

Ambient AI – Background (IoT)

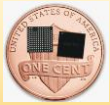


Ambient AI – Background (IoT)

- Go and Deploy!

Ambient AI – Background (IoT)

Hardware evolution, including low-power AI accelerators

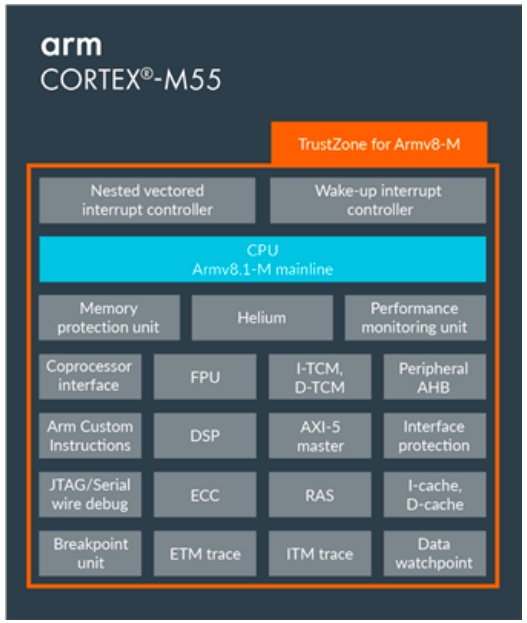



Google

TensorFlow Lite

NVIDIA

ASUS
Inspiring Innovation • Persistent Perfection



ARM

Wait…

We have “smart”phones, “smart” speakers, “smart” watches…

Why don’t we use them more smartly?

Ambient AI – Background (Edge Computing)

ML / DL

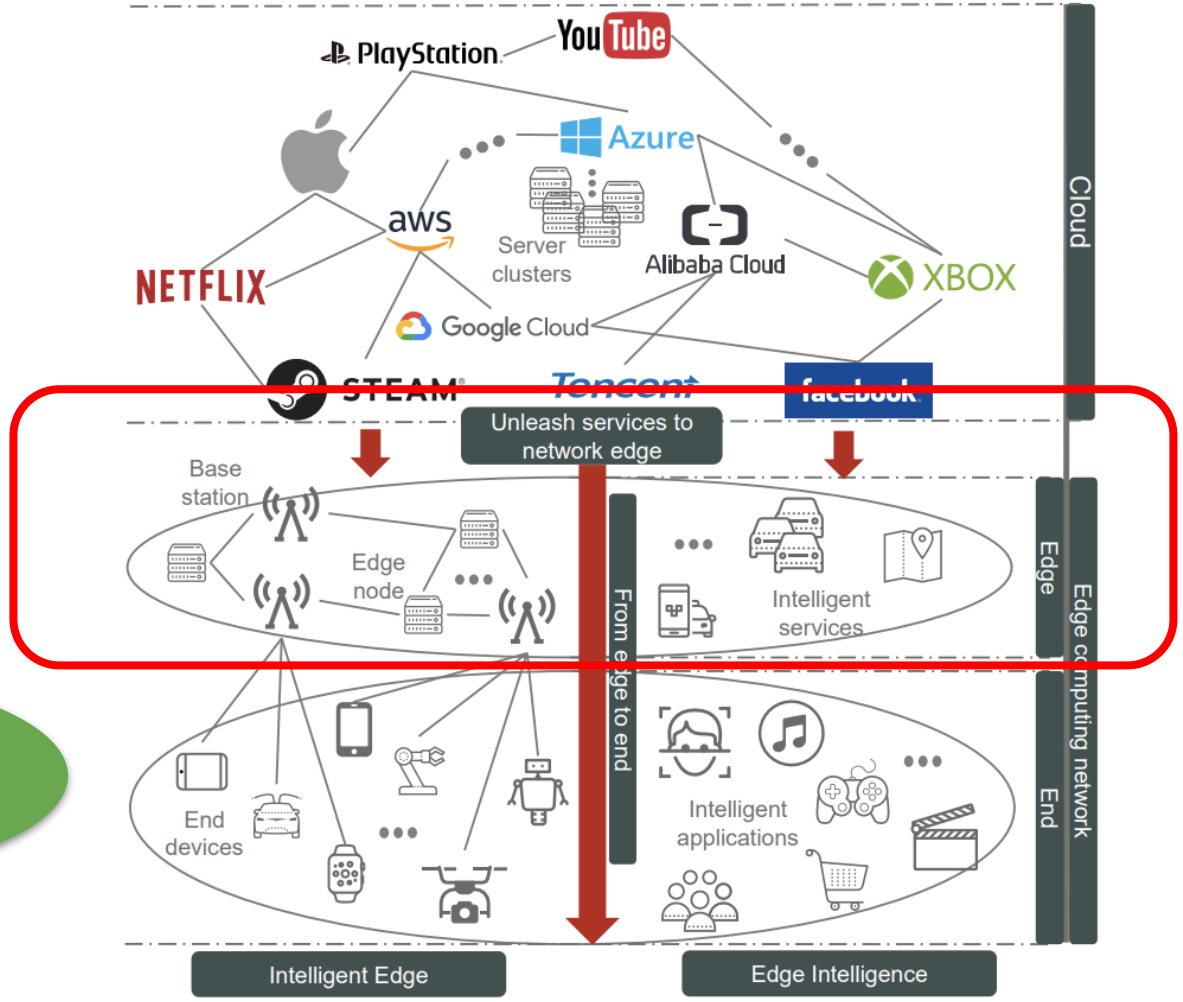
2009~

Edge Computing

2009~

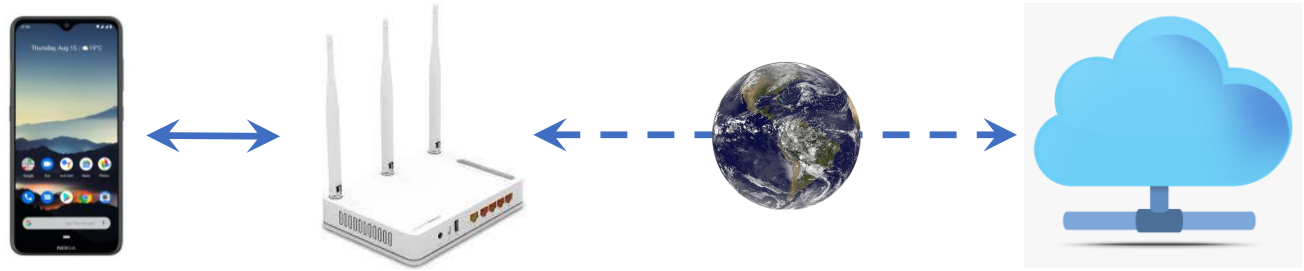
Cyber-physical System (IoT)

1999~



Ambient AI – Background (Edge Computing)

- Cloudlets in 2009
  - When a mobile application needs intensive computing, use nearby Wi-Fi gateways (cloudlets) instead of the cloud



Faster response!

- Fog, Dew, Mist computing…

VIRTUAL MACHINES

The Case for VM-Based Cloudlets in Mobile Computing

A new vision of mobile computing liberates mobile devices from severe resource constraints by enabling resource-intensive applications to leverage cloud computing free of WAN delays, jitter, congestion, and failures.

Mahadev Satyanarayanan
Carnegie Mellon University

Paramvir Bahl
Microsoft Research

Ramón Cáceres
AT&T Research

Nigel Davies
Lancaster University

Mobile computing is at a fork in the road. After two decades of sustained effort by many researchers, we've finally developed the core concepts, techniques, and mechanisms to provide a solid foundation for this still fast-growing area. The vision of "information at my fingertips at any time and place" was just a dream in the mid 1990s; today, ubiquitous email and Web access is a reality that millions of users worldwide experience through BlackBerries, iPhones, Windows Mobile, and other mobile devices. On one path of the fork, mobile Web-based services and location-aware advertising opportunities have begun to appear, and companies are making large investments in anticipation of major profits.

Yet, this path also leads mobile computing away from its true potential. Awaiting discovery on the other path is an entirely new world in which mobile computing seamlessly augments users' cognitive abilities via compute-intensive capabilities such as speech recognition, natural language processing, computer vision and graphics, machine learning, augmented reality, planning, and decision making. By thus empowering mobile users, we could transform many areas of human activity (see the sidebar for an example).

This article discusses the technical obstacles to this transformation and proposes a new architecture for overcoming them. In this architecture, a mobile user exploits virtual machine (VM) technology to rapidly instantiate customized service software on a nearby *cloudlet* and then uses that service over a wireless LAN; the mobile device typically functions as a thin client with respect to the service. A cloudlet is a trusted, resource-rich computer or cluster of computers that's well-connected to the Internet and available for use by nearby mobile devices.

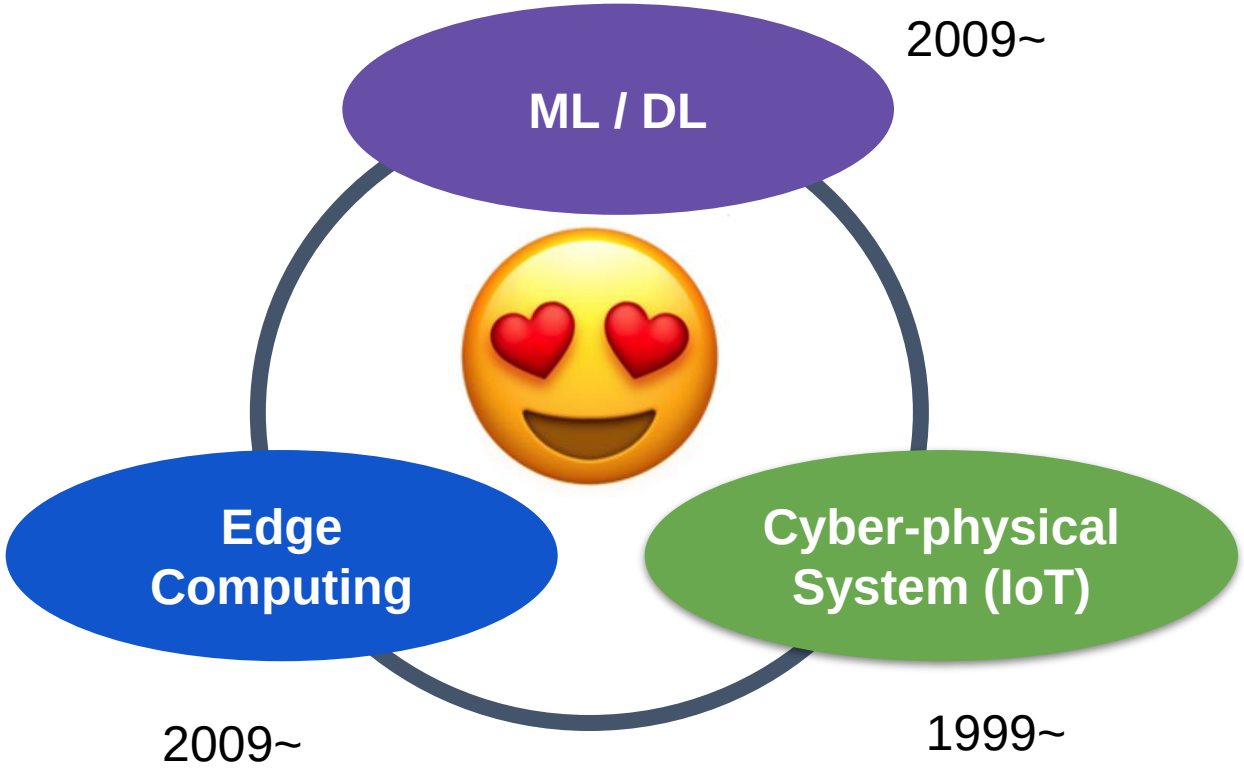
Our strategy of leveraging transiently customized proximate infrastructure as a mobile device moves with its user through the physical world is called *cloudlet-based, resource-rich, mobile computing*. Crisp interactive response, which is essential for seamless augmentation of human cognition, is easily achieved in this architecture because of the cloudlet's physical proximity and one-hop network latency. Using a cloudlet also simplifies the challenge of meeting the peak bandwidth demand of multiple users interactively generating and receiving media such as high-definition video and high-resolution images. Rapid customization of infrastructure for diverse applications emerges as a critical requirement, and our results from a proof-of-concept prototype suggest that VM technology can indeed help meet this requirement.

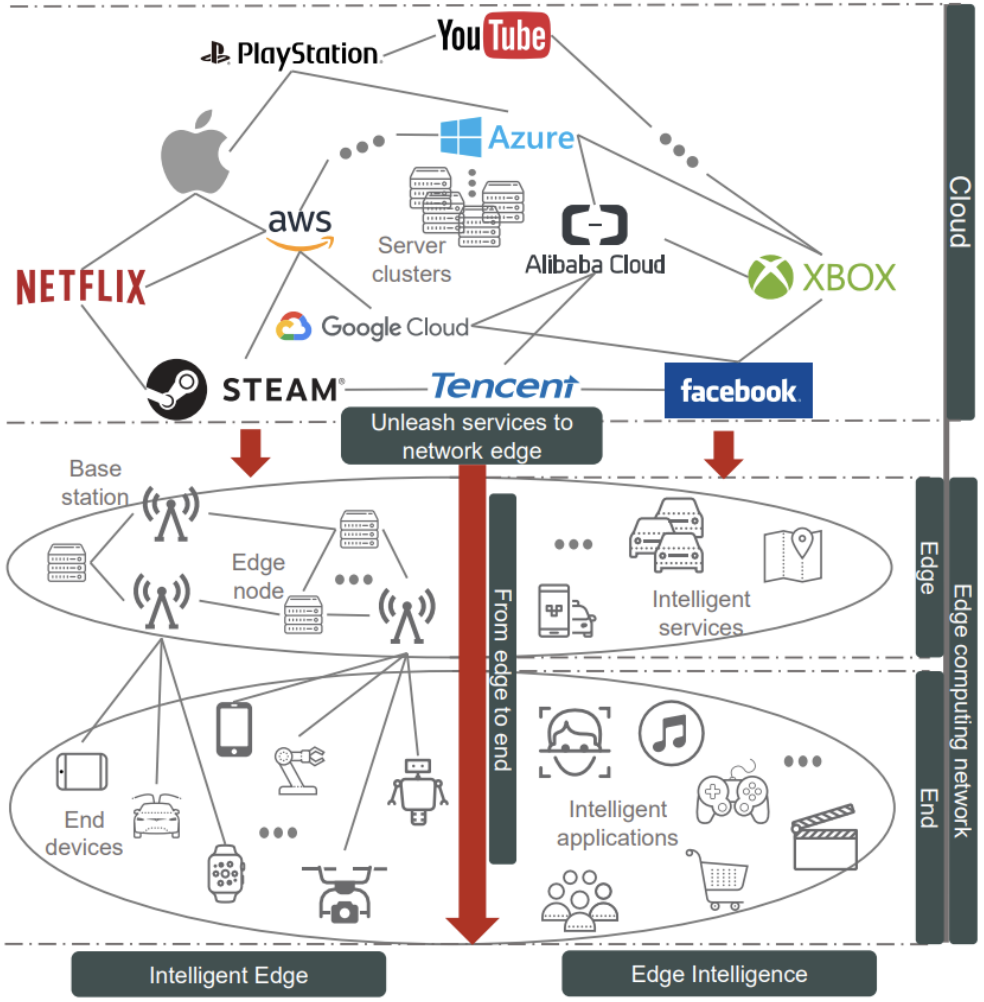
Resource-Poor Mobile Hardware

The phrase "resource-rich mobile computing" seems like an oxymoron at first glance. Researchers have long recognized that mobile hardware is necessarily resource-poor relative

Published by the IEEE CS ■ 1536-1268/09/$26.00 © 2009 IEEE

Ambient AI – Background

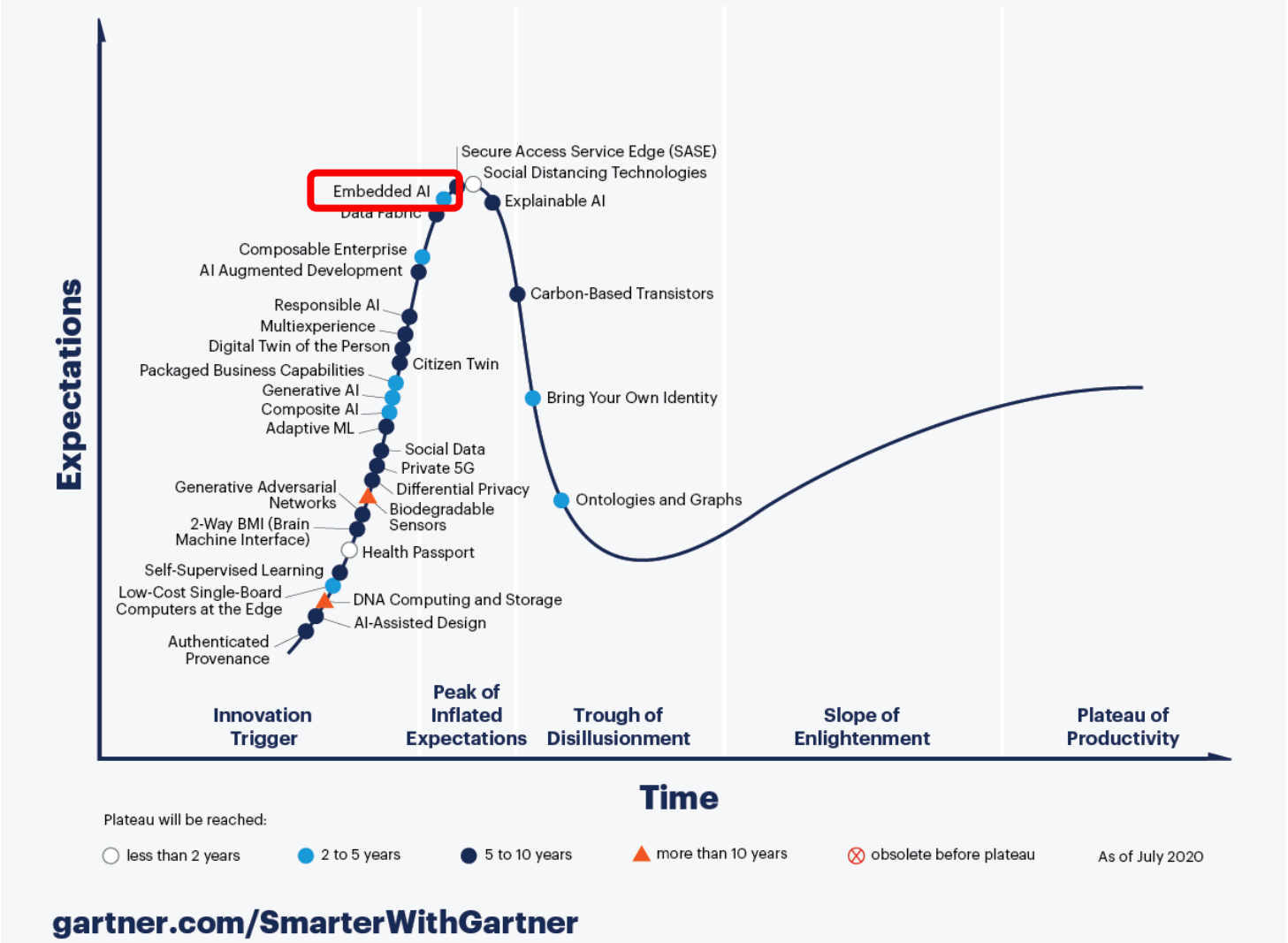




Ambient AI – Why?

- For next-generation AI applications!
- Inference on the edge directly
  - Fast interaction with users (real-time service, such as AR)
  - Robust regardless of Internet connectivity
- Not sending raw data to the cloud
  - Better user Privacy
  - Saving network bandwidth

Ambient AI – Hype



Ambient AI – Research and Applications

Lecture 1-4

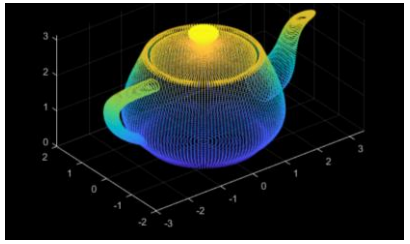
Hyung-Sin Kim

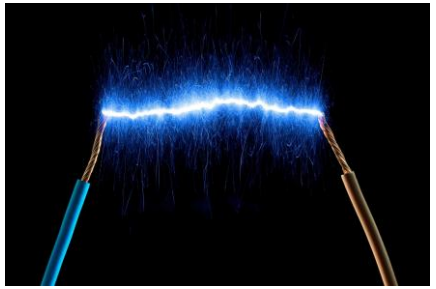
SNU Graduate School of Data Science

Research topics

Application-oriented Systems

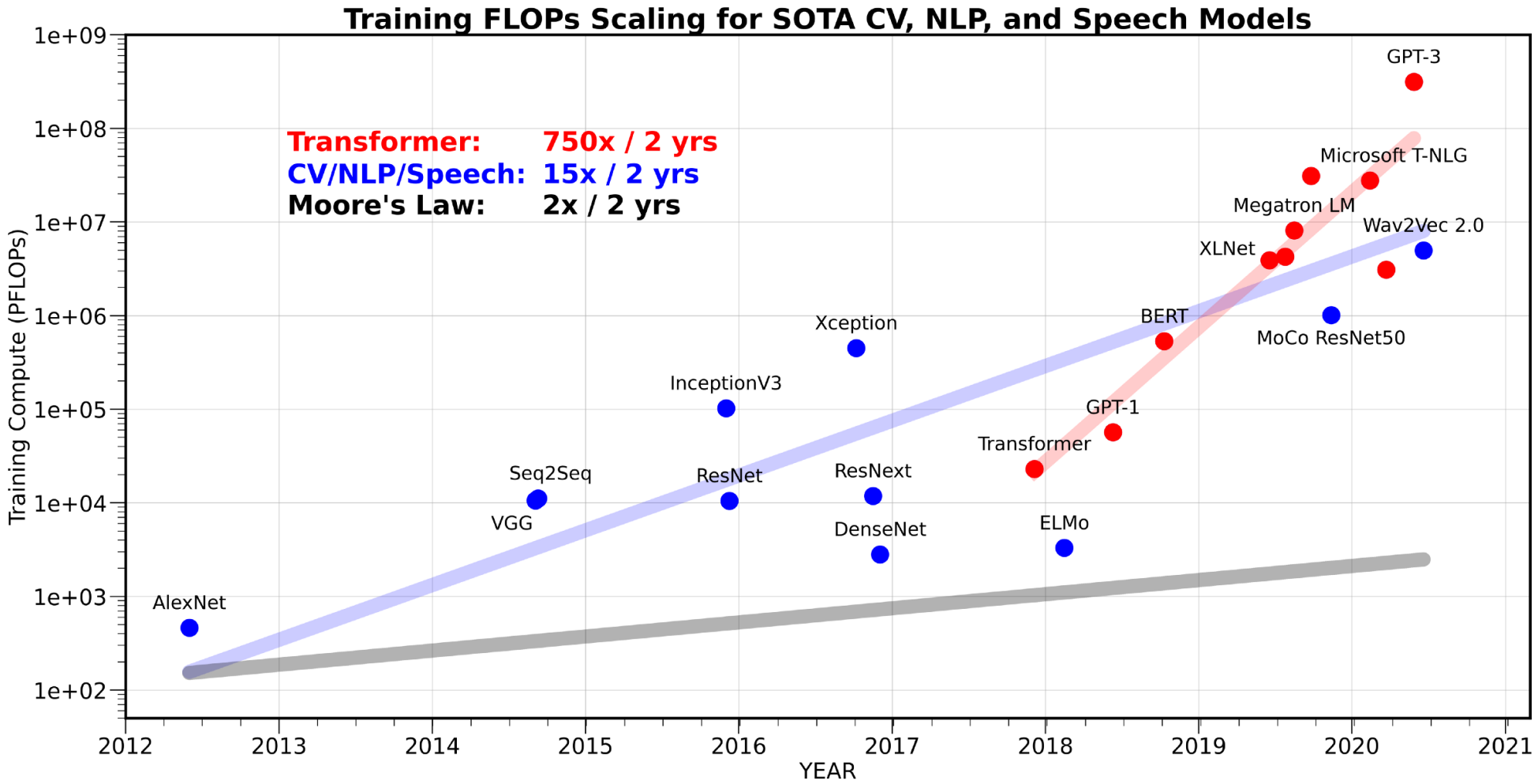
- Given a target application, what kind of information is good for the target?
  - **2D vision (RGB camera)**
  - **3D (RGB-D, LiDAR, RADAR)**
  - Motion (vibration)
  - **Sound**
  - Electric current
  - Wireless signal





We have a variety of IoT sensors!

Deep Neural Network Lightening



[The figure comes from RISELab blog post at https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8]

Deep Neural Network Lightening

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the

Google fires prominent AI ethicist Timnit Gebru

Gebru says the decision came from Google's head of AI, Jeff Dean

By Zoe Schiffer | @ZoeSchiffer | Dec 3, 2020, 1:11pm EST

Listen to this article

SHARE

Photo by Kimberly White / Getty Images for TechCrunch

Google fires top ethical AI expert Margaret Mitchell

The tech giant claims Mitchell violated staff codes of conduct.

Google has fired the co-lead of the company's ethical AI unit, Margaret Mitchell, on the heels of the removal of Timnit Gebru.

Mitchell, an ethical artificial intelligence (AI) expert who has previously worked on machine learning bias, race and gender diversity, and language models for image capture, was hired by Google to co-lead the firm's Ethical AI team with Gebru -- a post that has lasted roughly two years, as noted by Reuters.

Margaret Mitchell [right], was fired on the heels of the removal of Timnit Gebru.

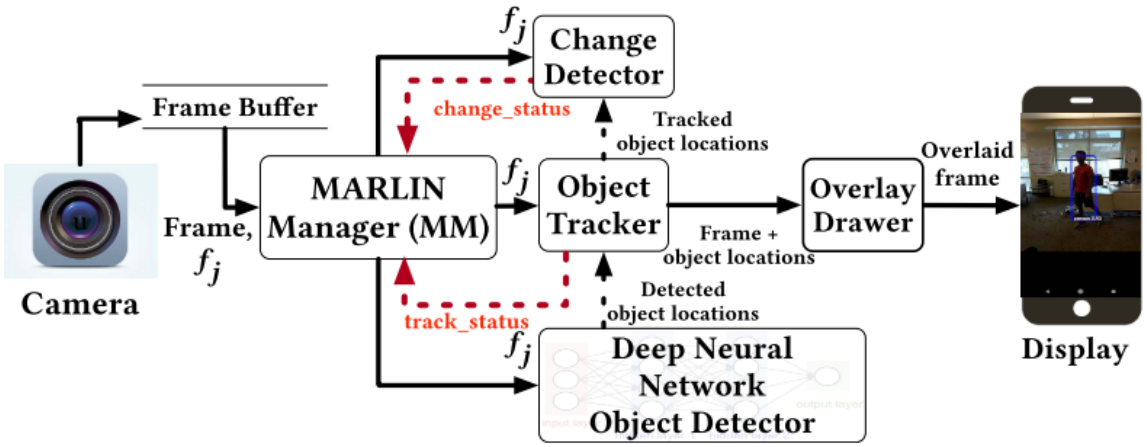
Deep Neural Network Lightening

- Quantization
  - Real numbers to integers, both for weights and activations
- Pruning
  - Cutting trivial edges to reduce model complexity
- Knowledge distillation
  - A small DNN can distill knowledge from a heavy ensemble model (even Mobile**BERT**)

AI models are too **heavy** for us to process… ☹

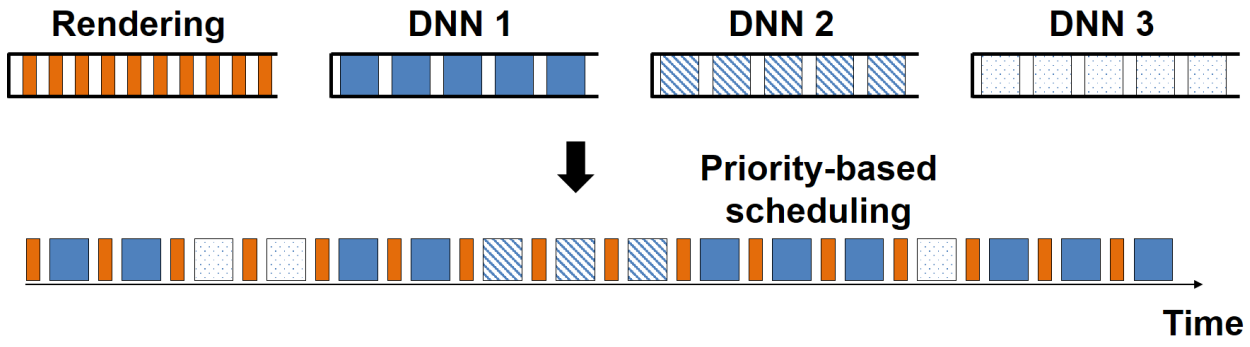
Lightweight Execution

- Decision for when to execute DNNs
  - Should we execute a **heavy** object detection model for every single image of a video even when the object does not move at all?
  - We might be able to use **lightweight** signal processing instead of DNNs many times…
  - The control logic should be simple



Resource Utilization

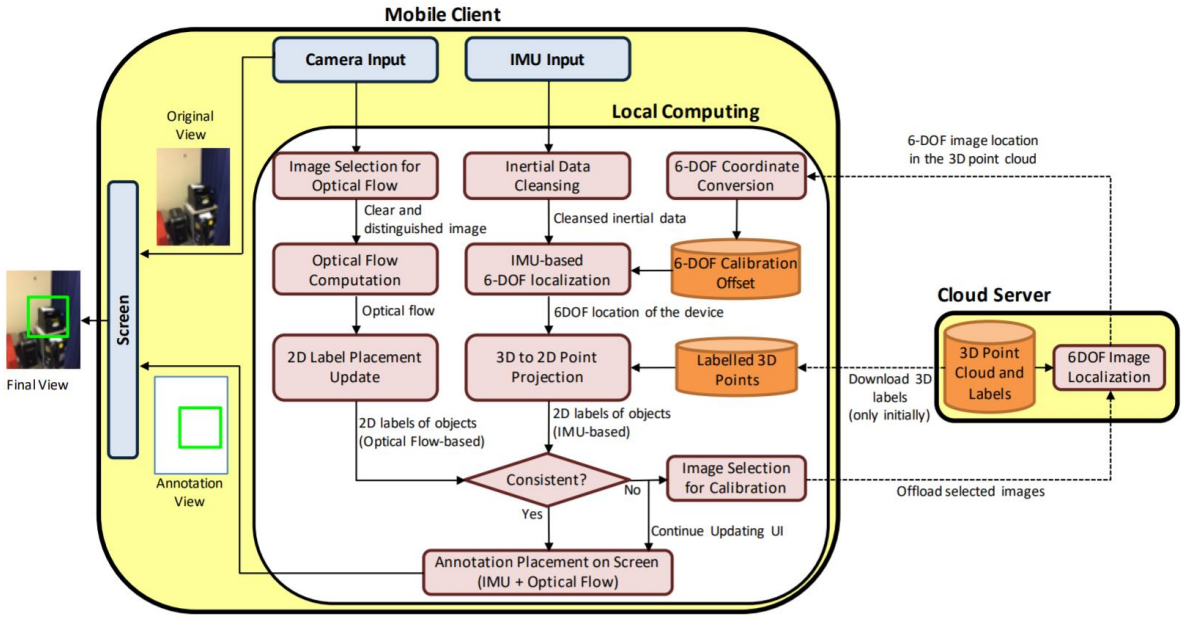
- Efficient utilization of edge computing resources
  - An edge device might need to run multiple DNNs
  - When to use What resource (CPU/GPU/TPU) for What task?



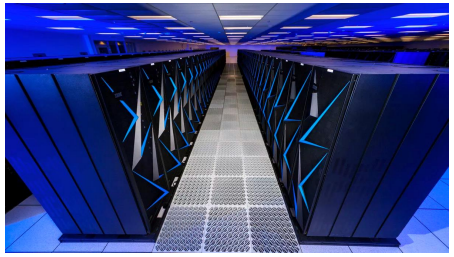
We do have **various** computing resources!

Resource Utilization

- Utilization of cloud and edge synergistically
  - Light, latency-sensitive tasks on the edge
  - Heavy, latency-tolerant tasks on the cloud



I'm here!

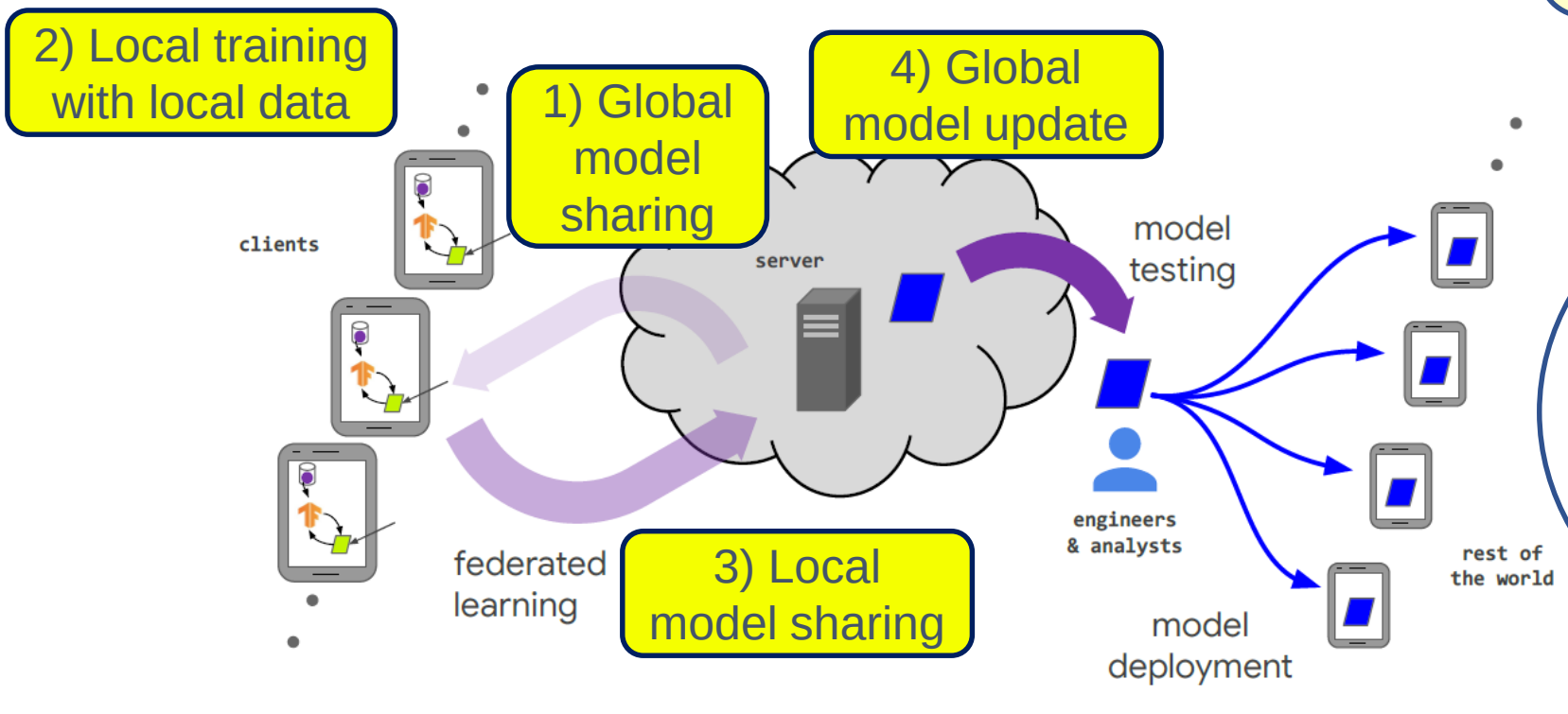


We do have **various** computing resources!

On-device Distributed Learning

- Federated learning
  - Data lives and dies locally
  - Local training and weight sharing

We don't want to share our data with the server even for training!



Adaptable Machine Learning

- Can we train a **meta** (adaptable) model that is quickly personalized by using few-shot data?

meta model

training data

Braque | Cezanne

Fine-tuned model

test datapoint

By Braque or Cezanne?

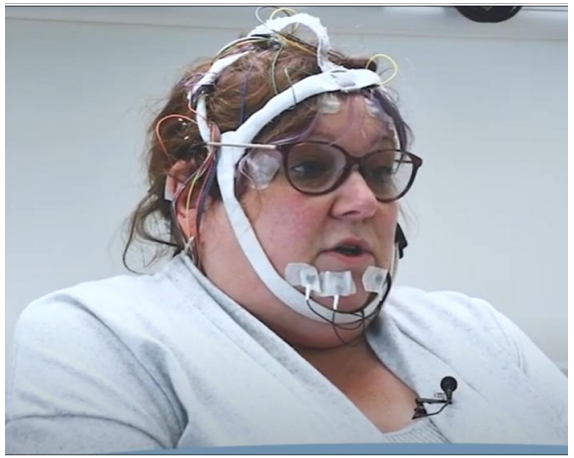
The pre-trained model doe not perform well for our own tasks… But we only have **small data**

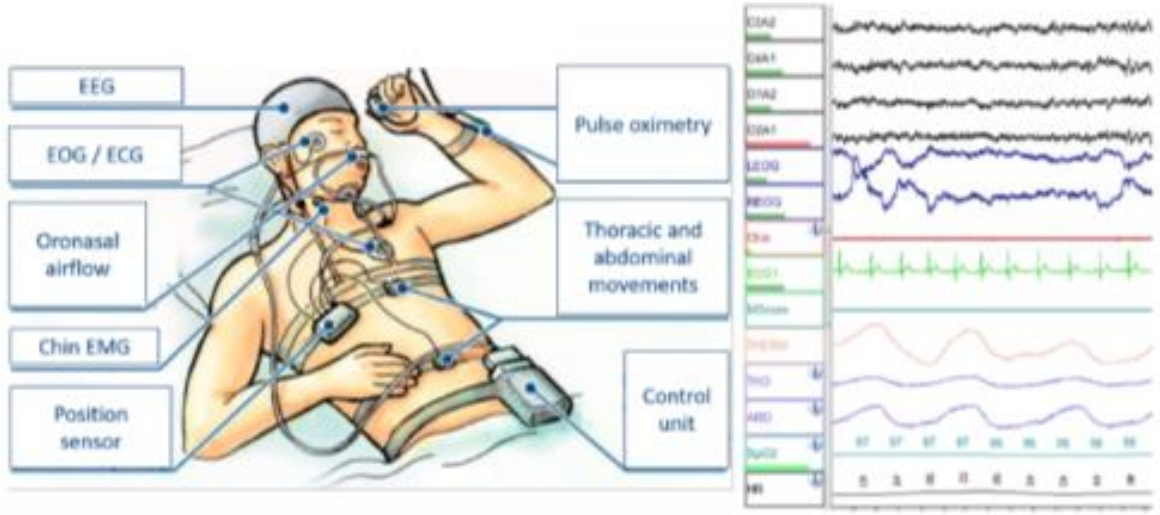
Applications

Healthcare (Sleep AI)

- Limitations of PSG
  - Numerous equipment must be attached to the body
  - It takes a long time to read the results
  - Recording is performed in an unfamiliar environment (night to night variability)
  - Space and manpower must be secured - expansive

➡ **Decreasing its practicality for utility** in various research objectives, including at home monitoring

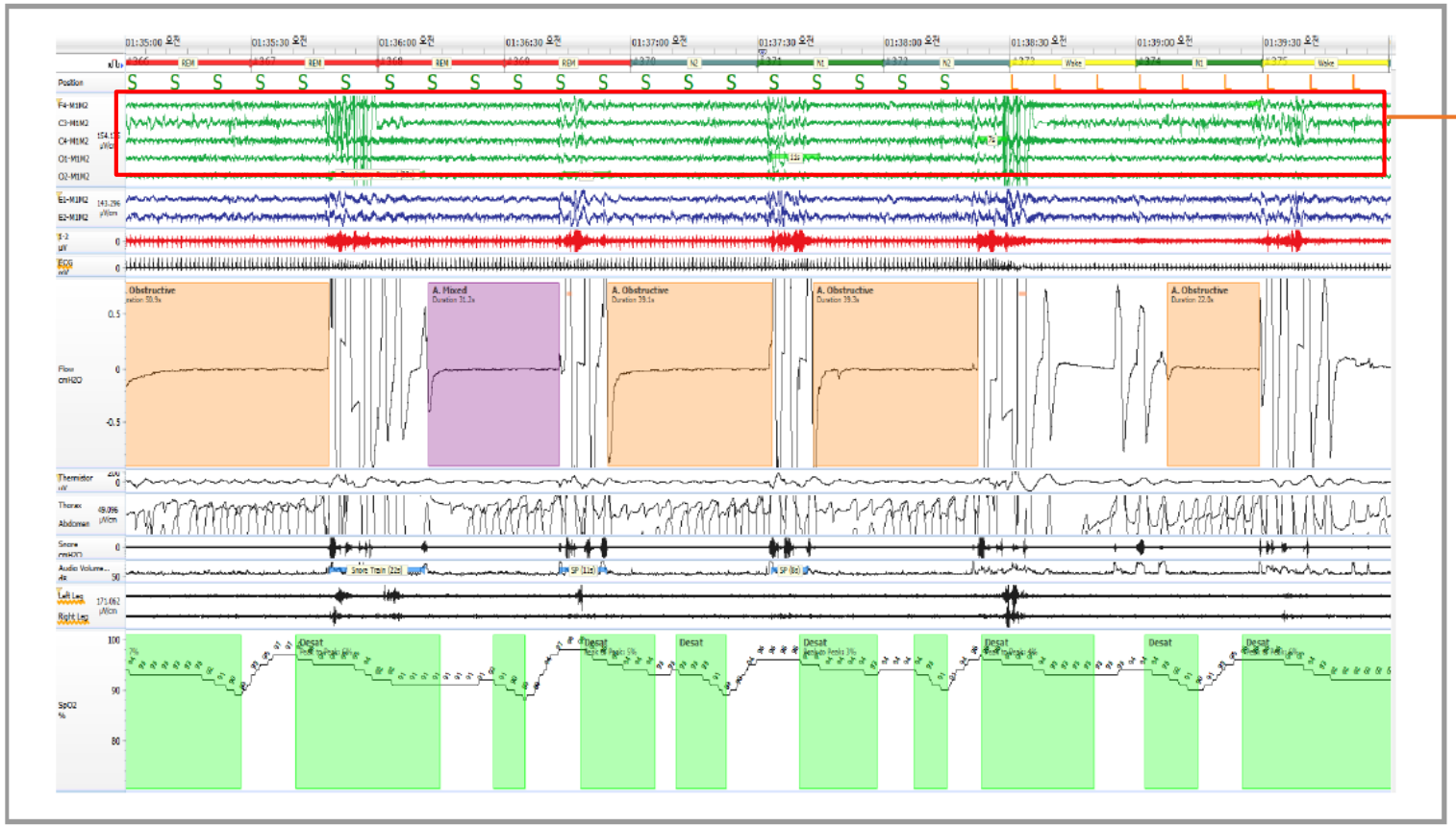




This figure is from "The meaning of sleep quality: A survey of available technologies", *IEEE Access*, vol. 7, 2019

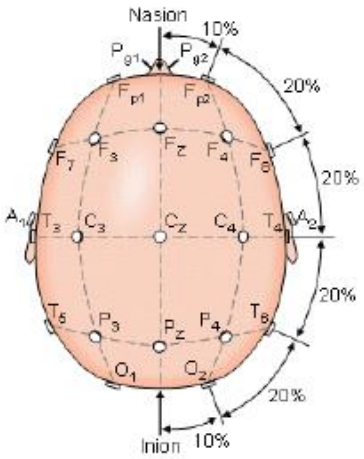
Healthcare (Sleep AI) – Single-Channel EEG

- 수면다원검사 중 기록한 23개 채널 신호 중 1개의 EEG 신호만으로 수면 단계 분류 → 수면 모니터링 간소화 + 다른 의료기기 연동



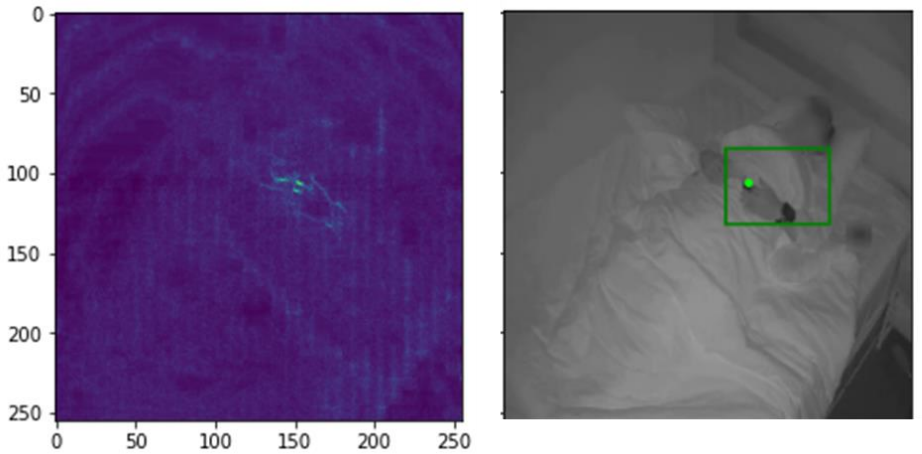
EEG signal (택 1)

- Wake, N1, N2, N3, REM 5단계 분류

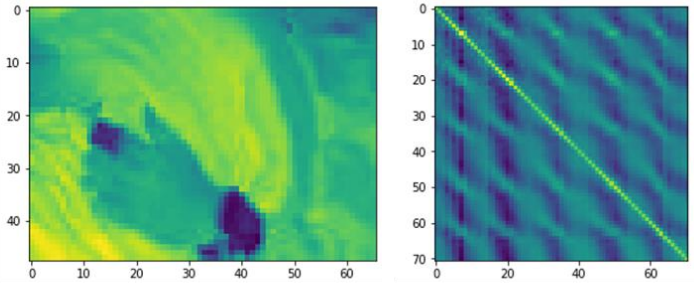


Healthcare (Sleep AI) – Infrared Video

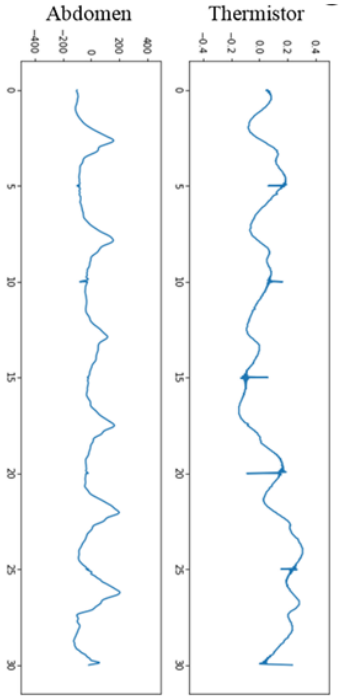
- 수면 적외선 영상을 이용한 비접촉식 생체신호(호흡패턴) 모니터링
  - 일상 생활에서 지속적인 수면 모니터링 → 환자 관리 및 진단에 활용
  - 중환자실 환자, 신생아 생체신호 모니터링 → 접촉식 센서 사용으로 인한 불편 및 노동부담 완화, 감염 예방
  - 수면다원검사 간소화
  - 실시간 모니터링이 가능하도록 하여 의료장비(인공호흡기 등)와 연결, 치료 보조에 활용



호흡과 관련된 가슴 복부 움직임 포착



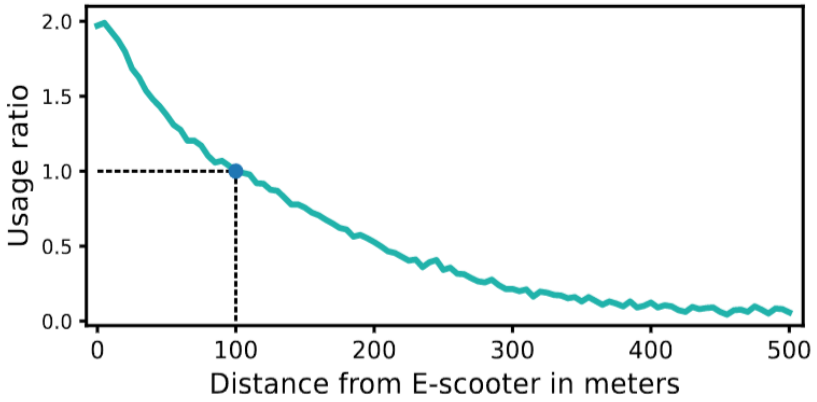
주기적인 호흡 패턴 추출

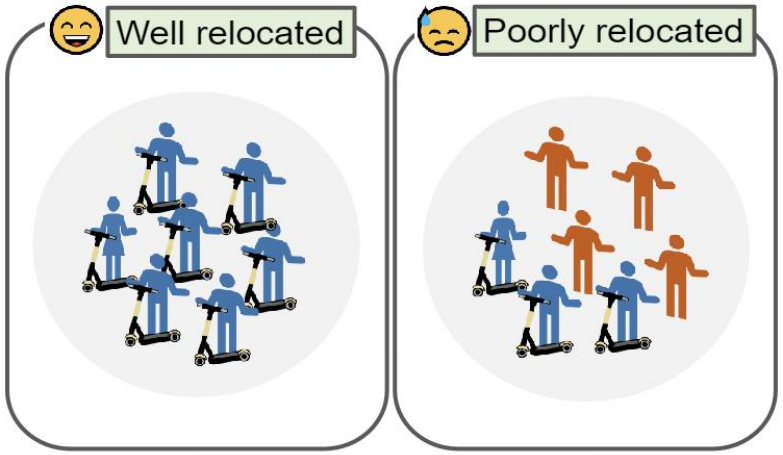


Respiratory Pattern Estimation

Micro Mobility – Motivation

- E-scooter relocation

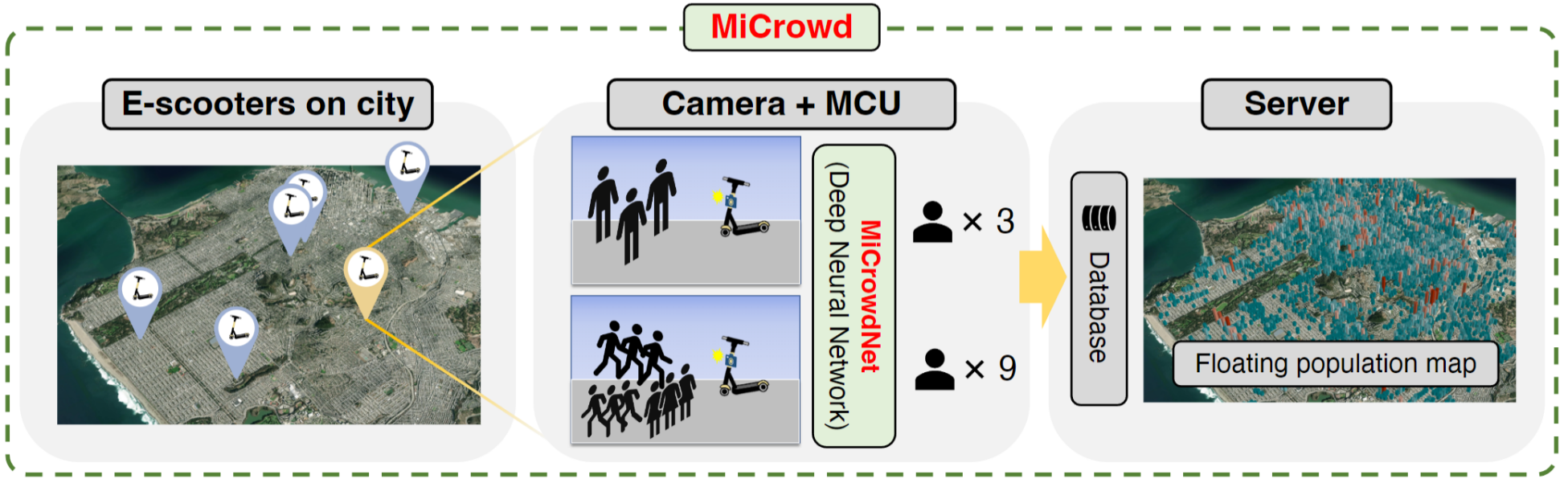




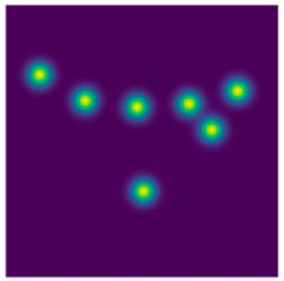
- Illegal parking detection

Micro Mobility – Approach

- Image-based On-device Crowd counting



(a) Image (input data)



(b) Density map (label)

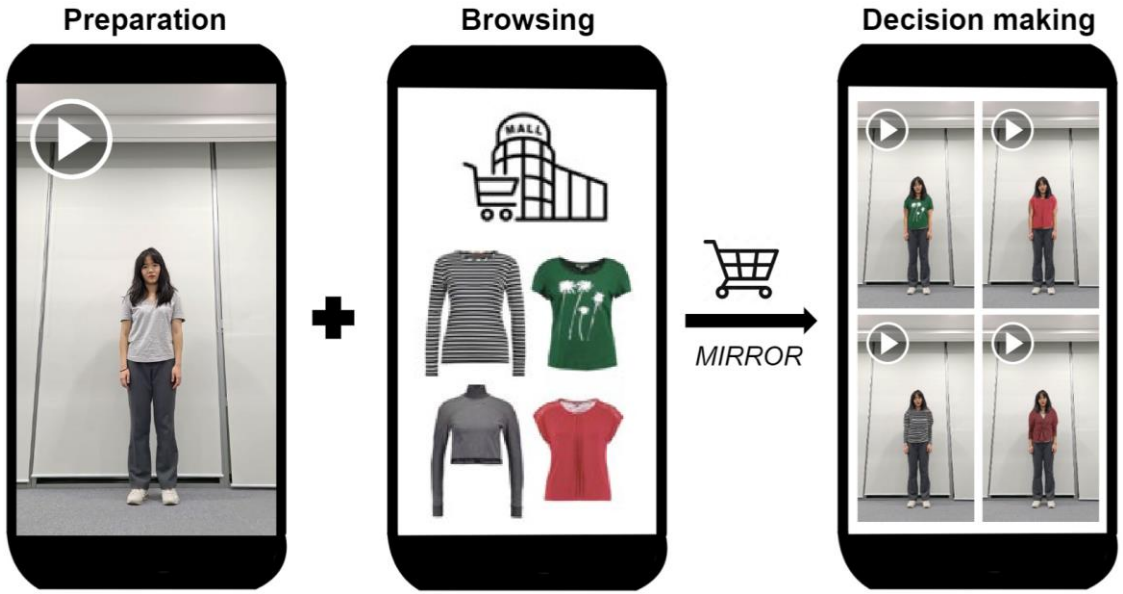
Fashion – Motivation

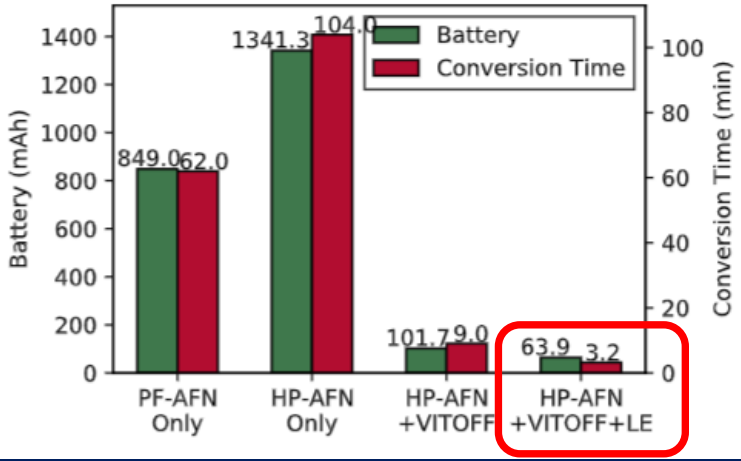
- One missing part in mobile cloths shopping: **No fitting room**

- Frustration and returns… ☹

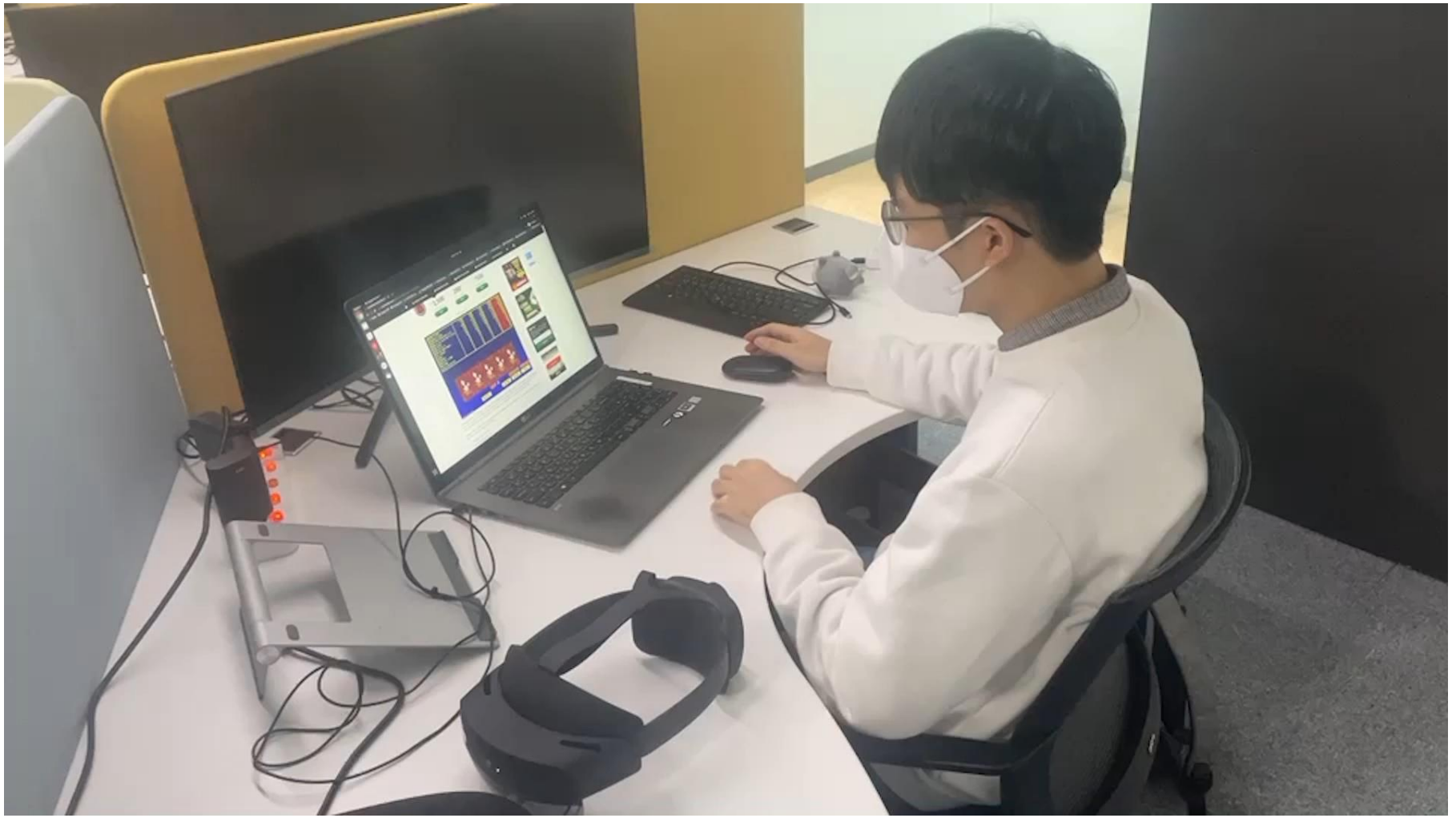
How about providing a **virtual fitting room** on a smartphone?

Fashion – Approach





AR Poker Guide (Class Project)



Animal Detection (Class Project)

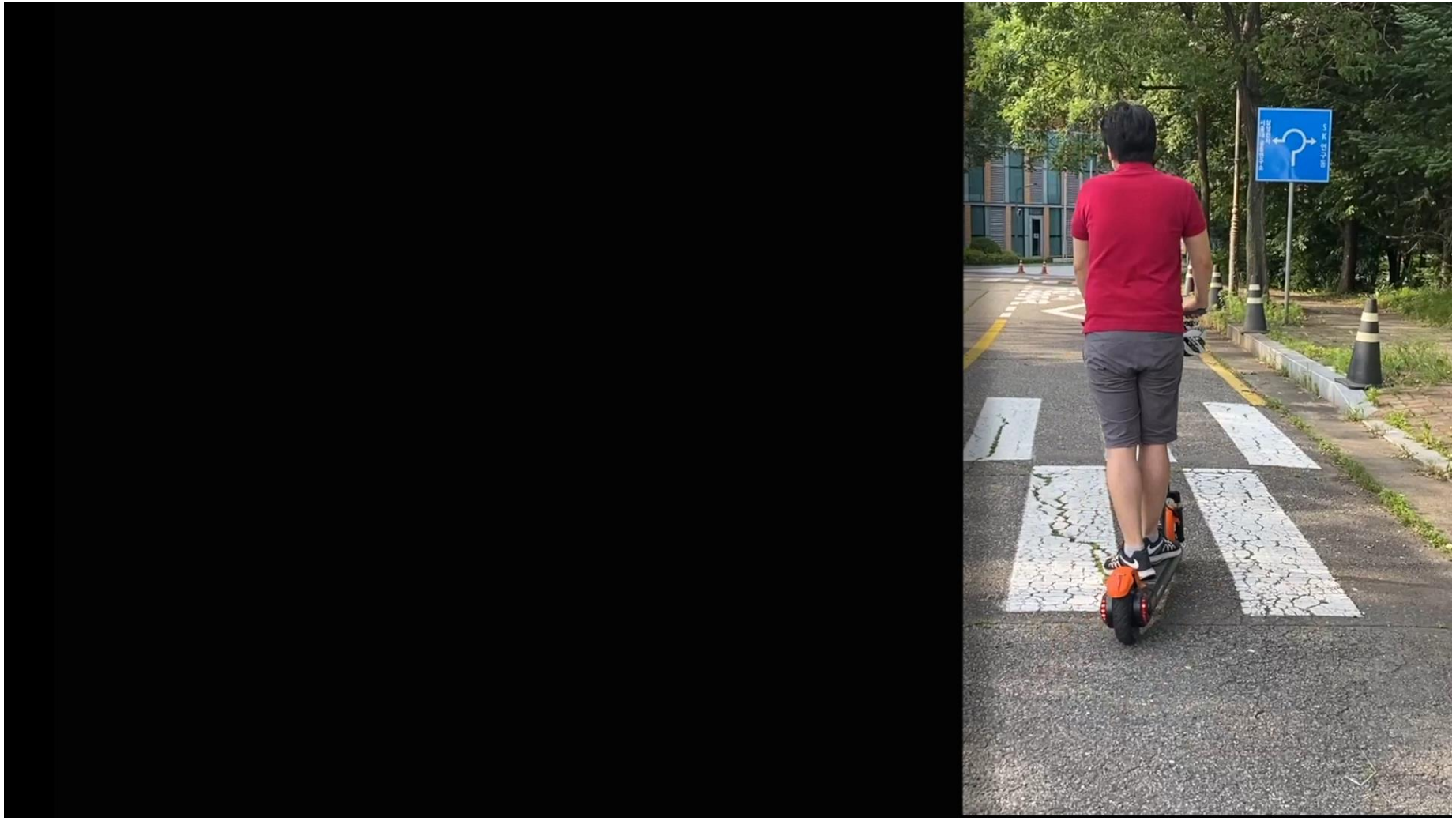
Crowd Density Detection (Class Project)

Smart BlackBox (Class Project)

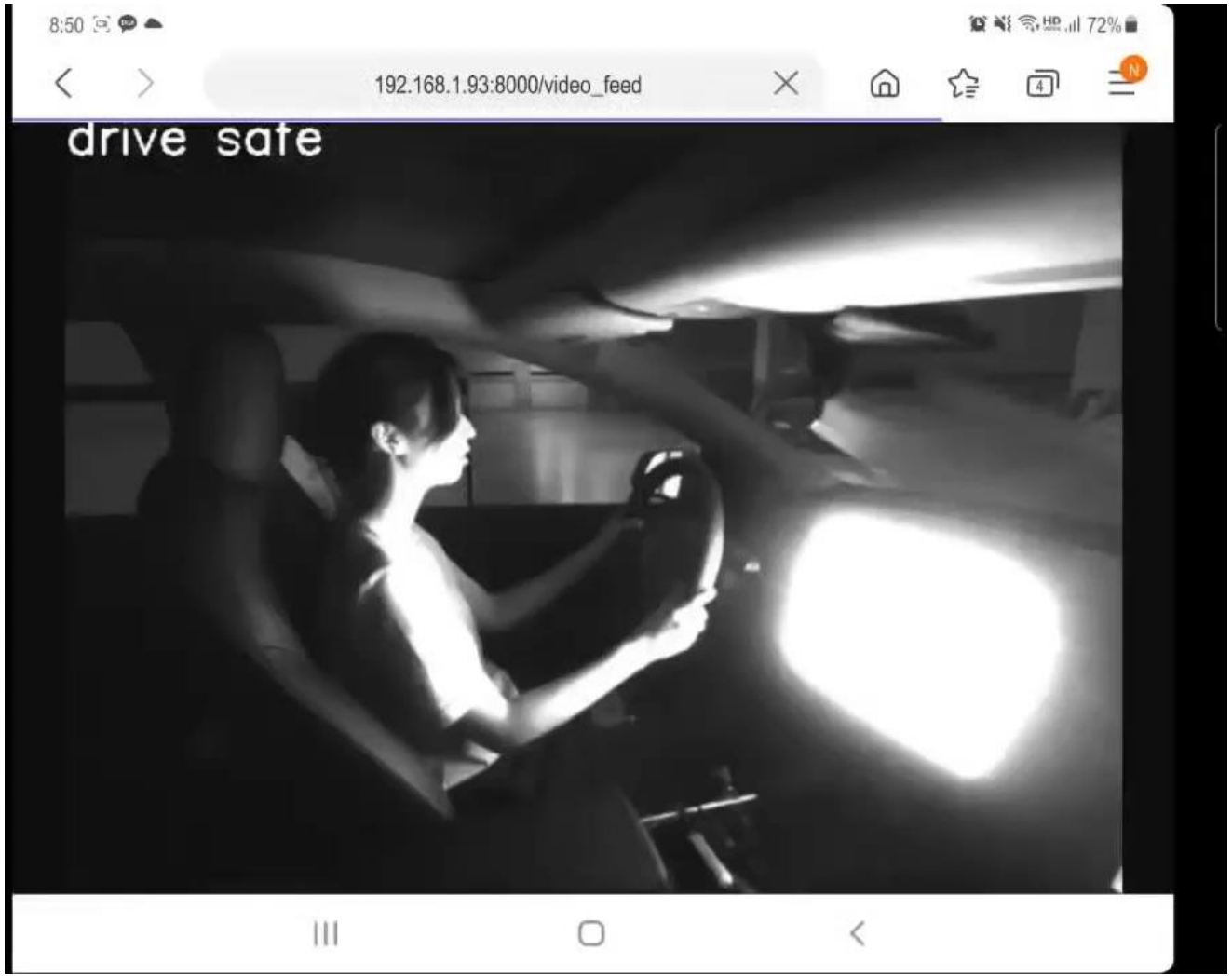
Face Filtering (Class Project)

Child Climbing Detection (Class Project)

Safe Driving Detection (Class Project)



Driver Distraction Detection (Class Project)



Thanks!