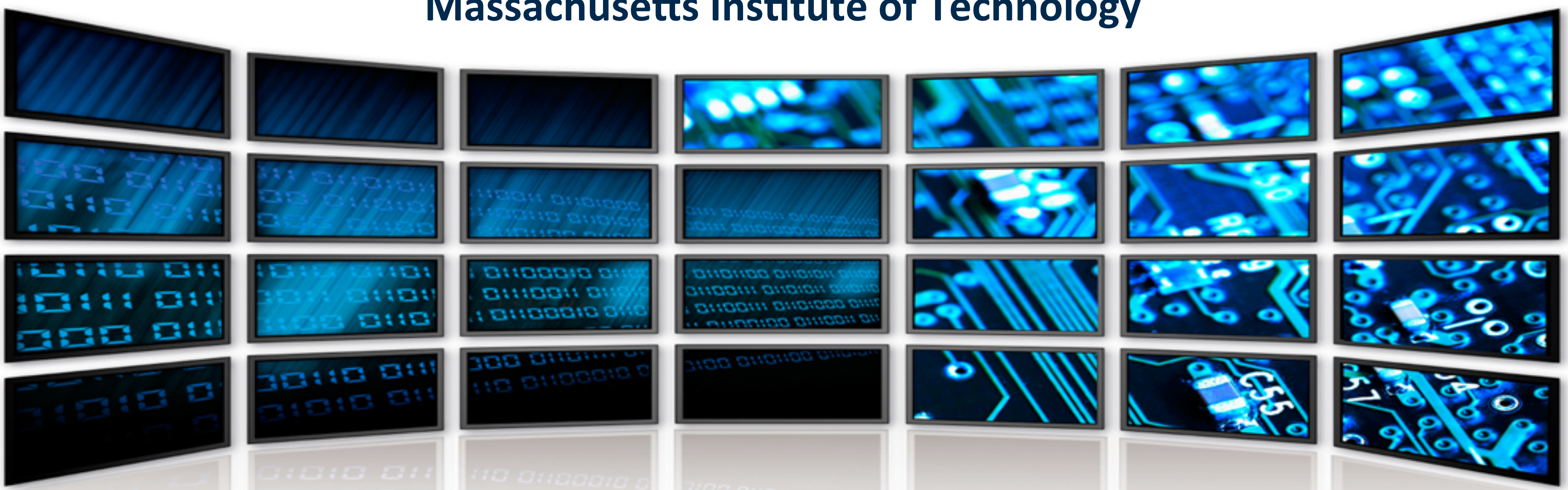
What If Your Smart Phone Didn't Need the Cloud?

Vivienne Sze

Massachusetts Institute of Technology

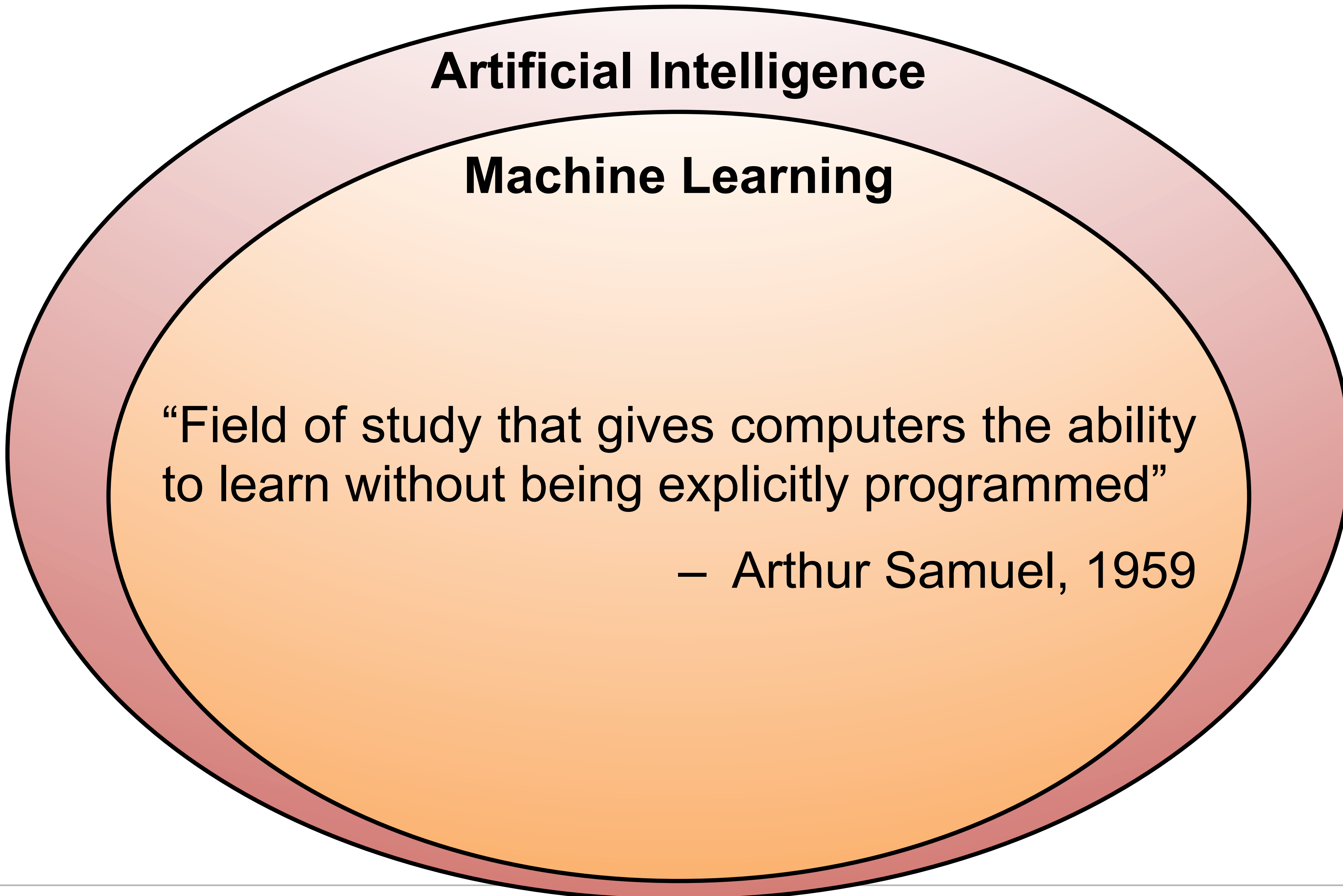


Contact Info
email: sze@mit.edu
website: www.rle.mit.edu/eems

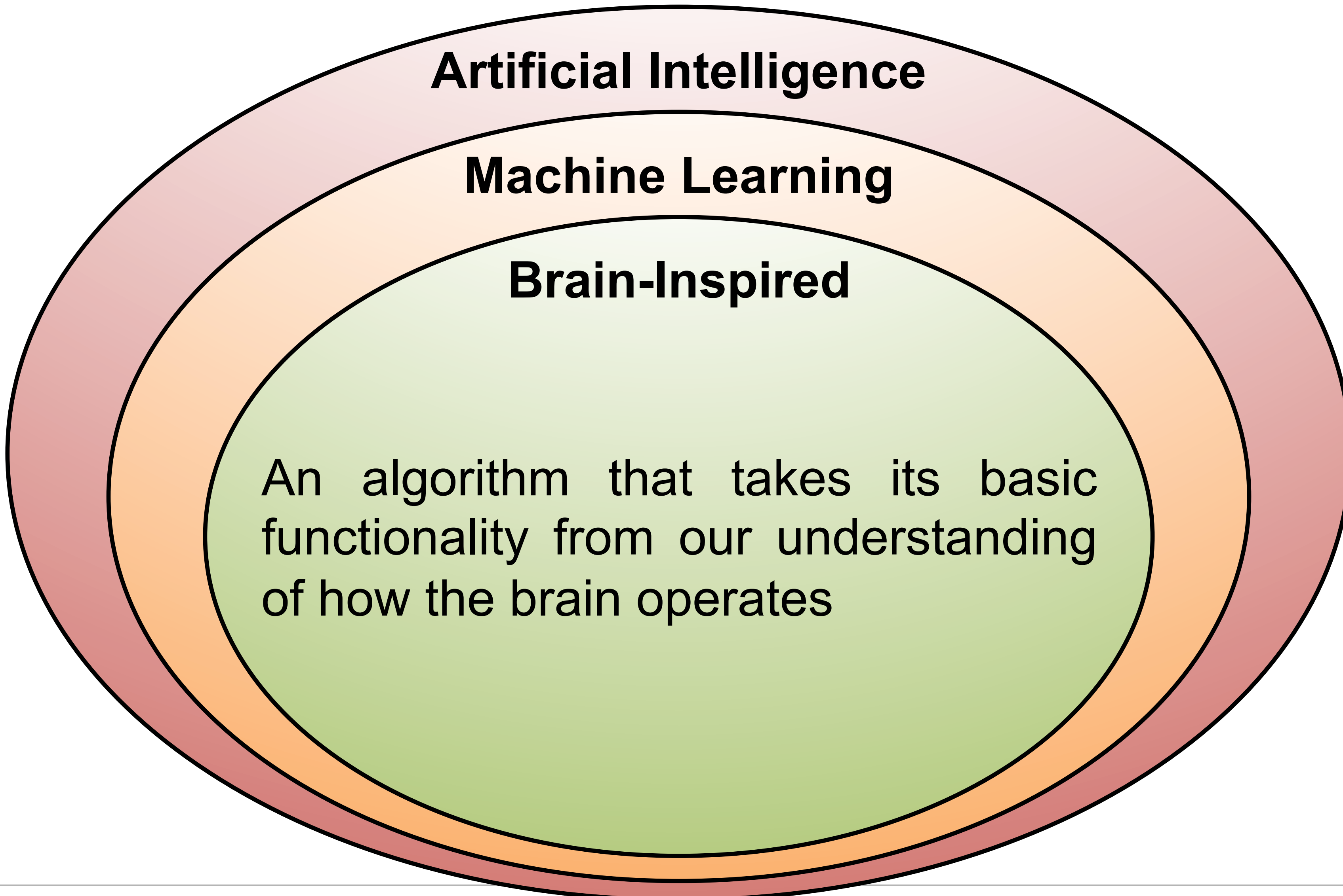
Outline

- What is Deep Learning?
- How is Deep Learning being used?
- Why is Edge Computing important?
- How can we enable Deep Learning at the Edge?

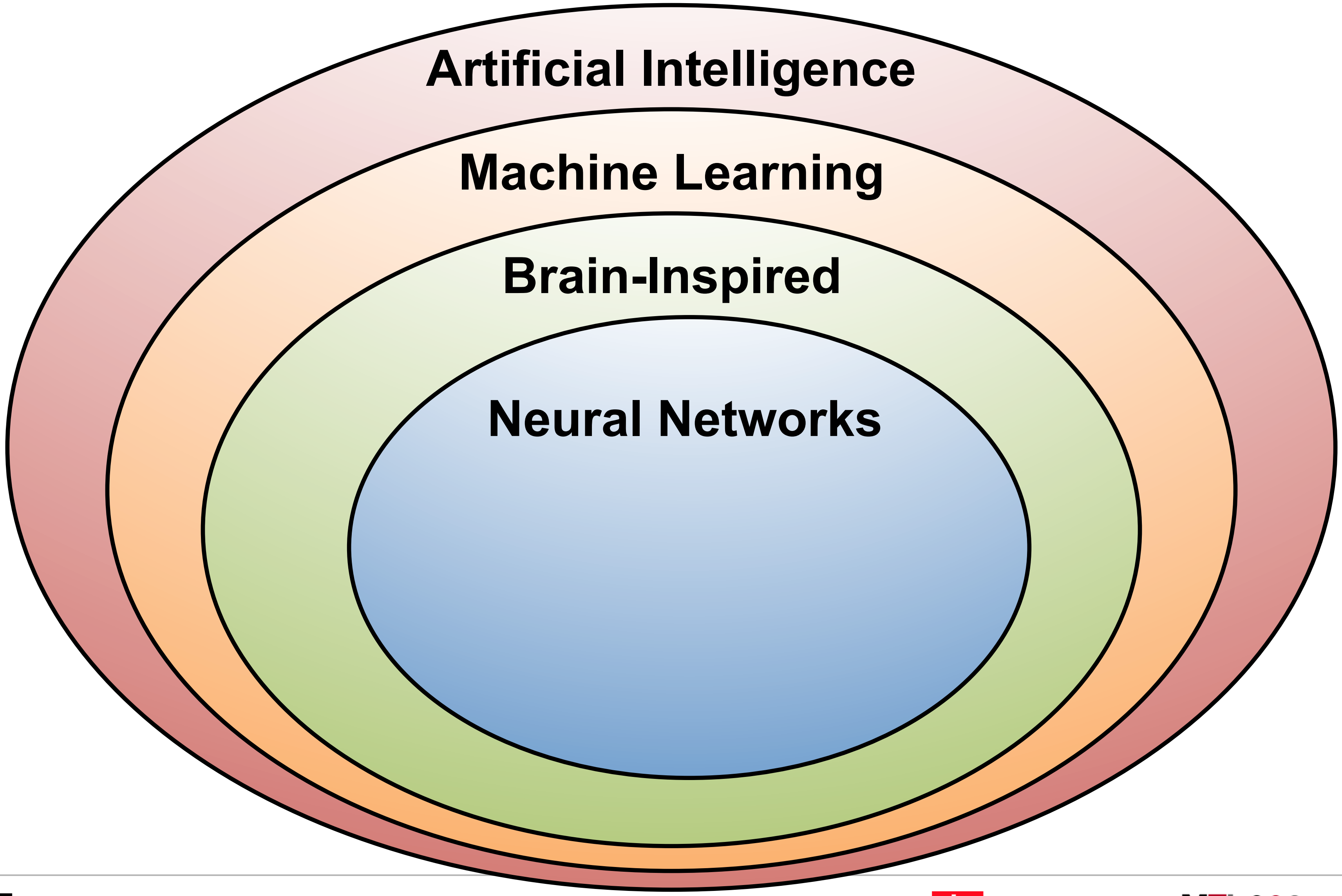
AI and Machine Learning



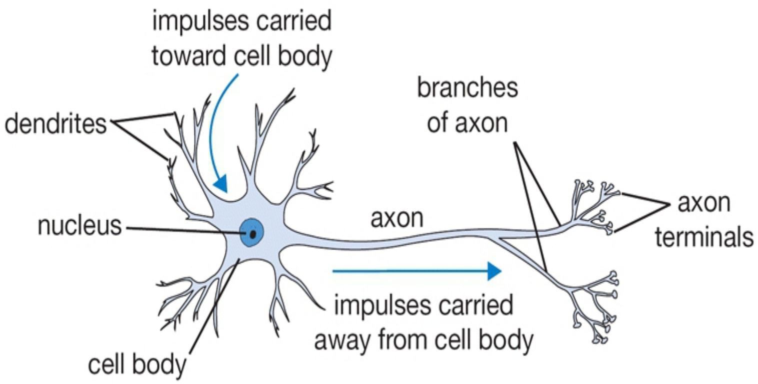
Brain-Inspired Machine Learning



Neural Networks



Neural Networks: Weighted Sum



The brain contains
~10^{11} neurons connected with
~10^{14} – 10^{15} synapses

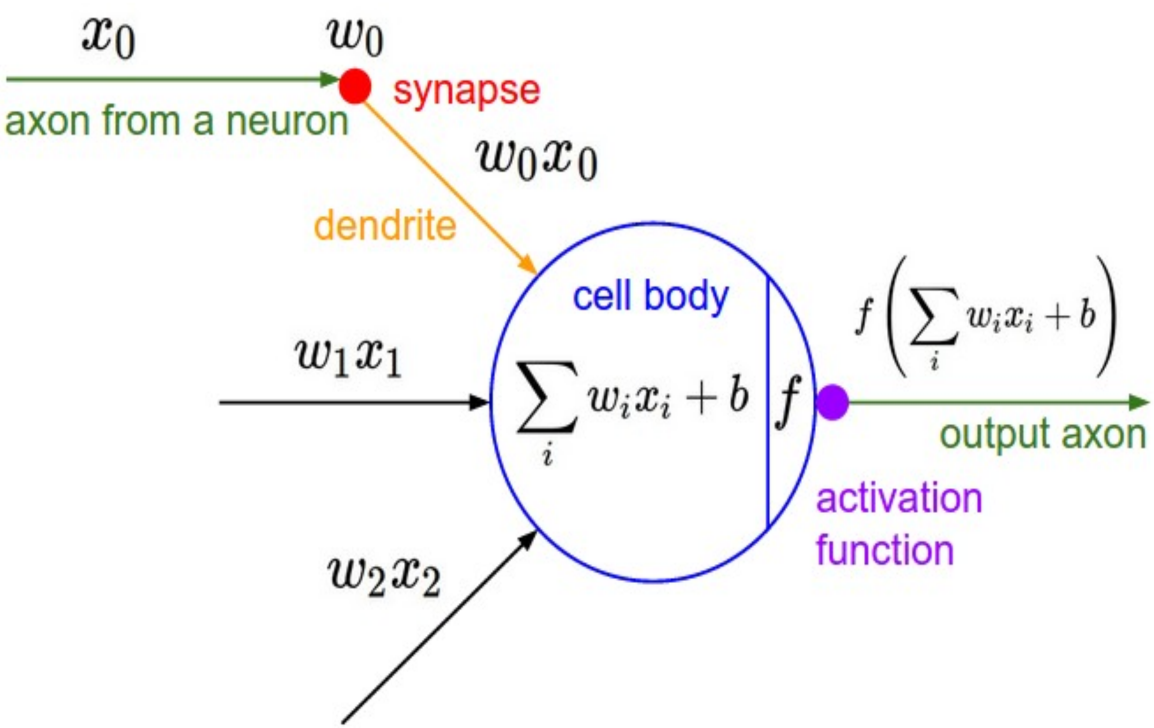


Image Source: Stanford

Many Weighted Sums

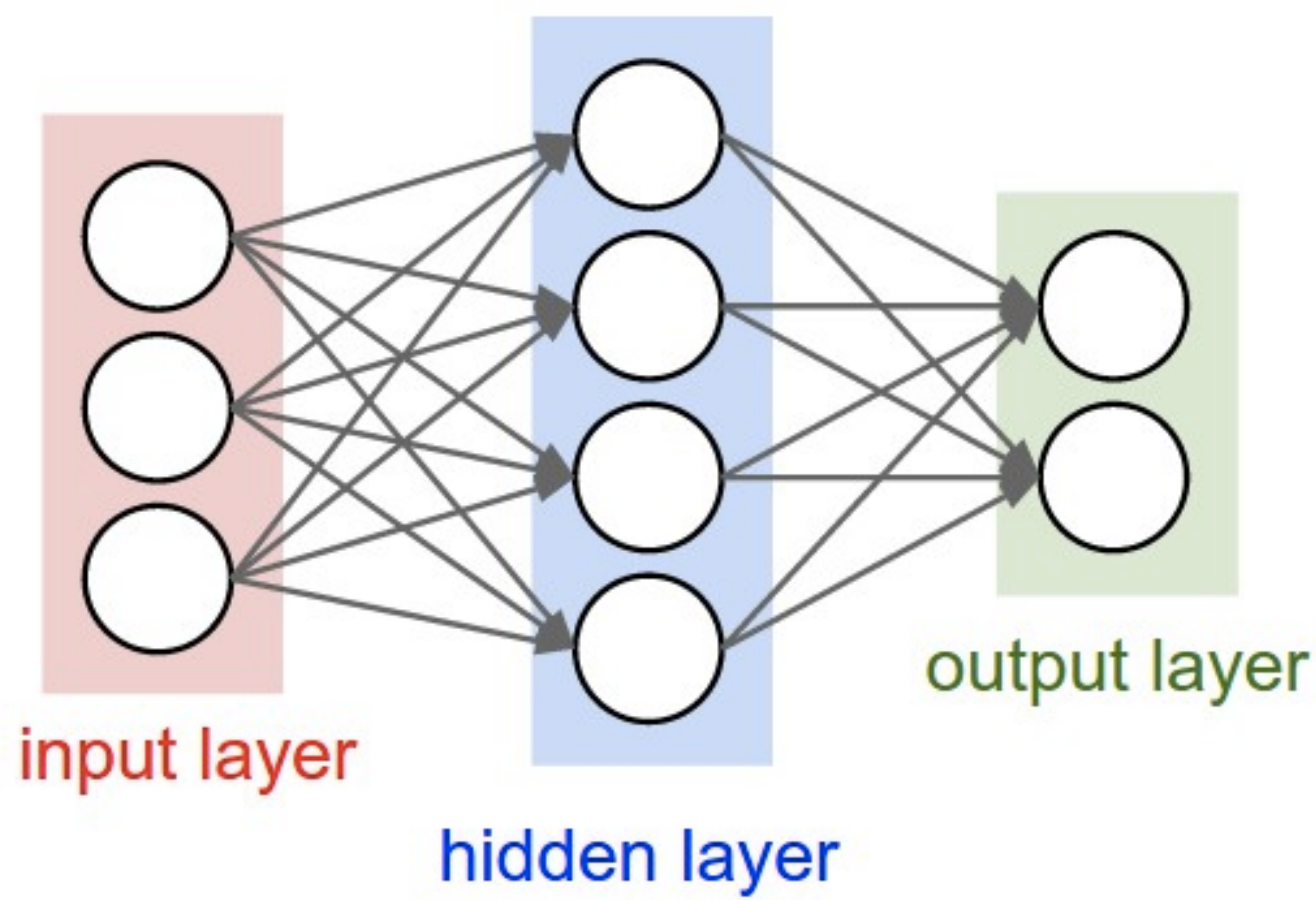
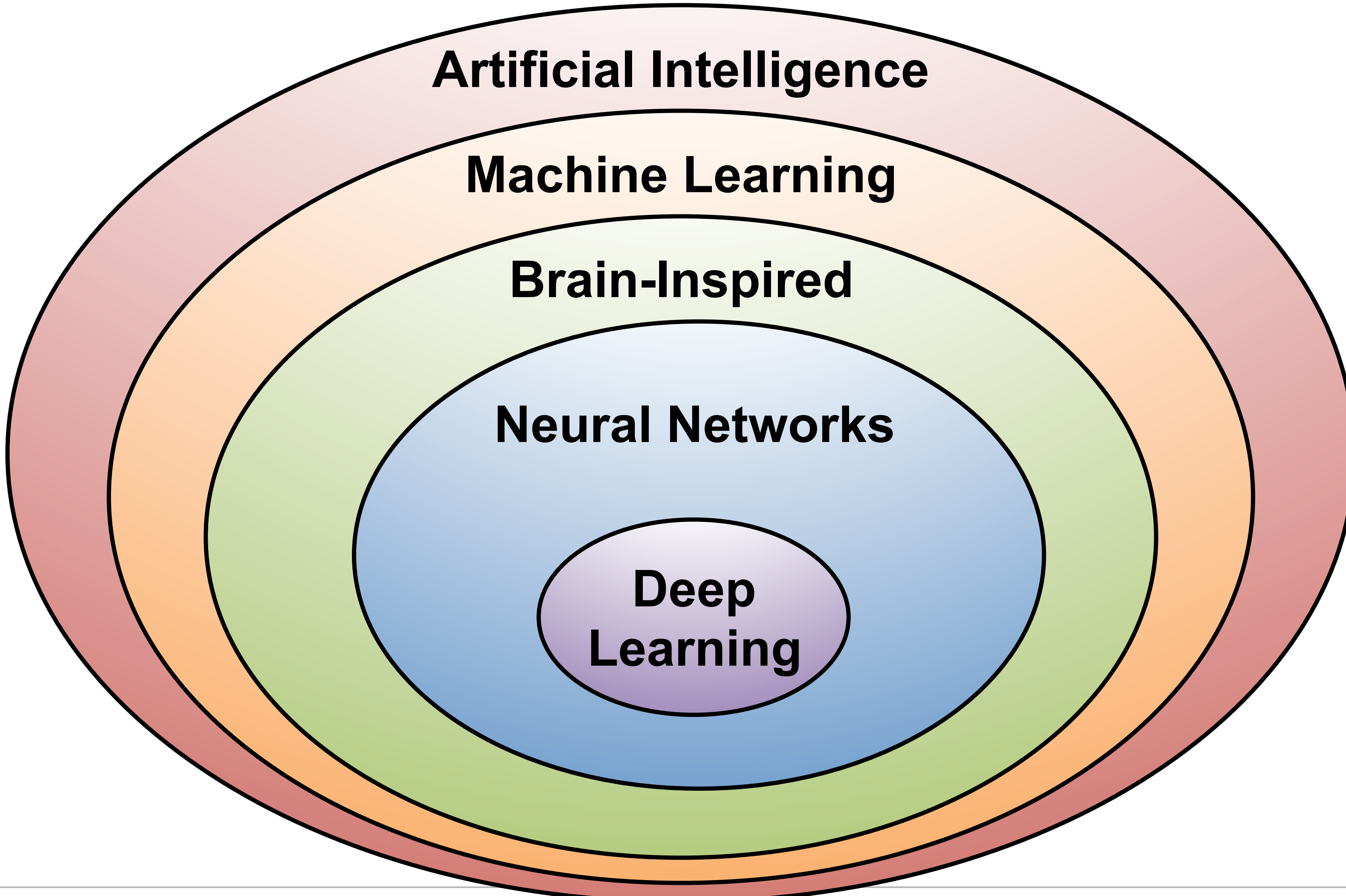


Image Source: Stanford

Deep Learning



What is Deep Learning?

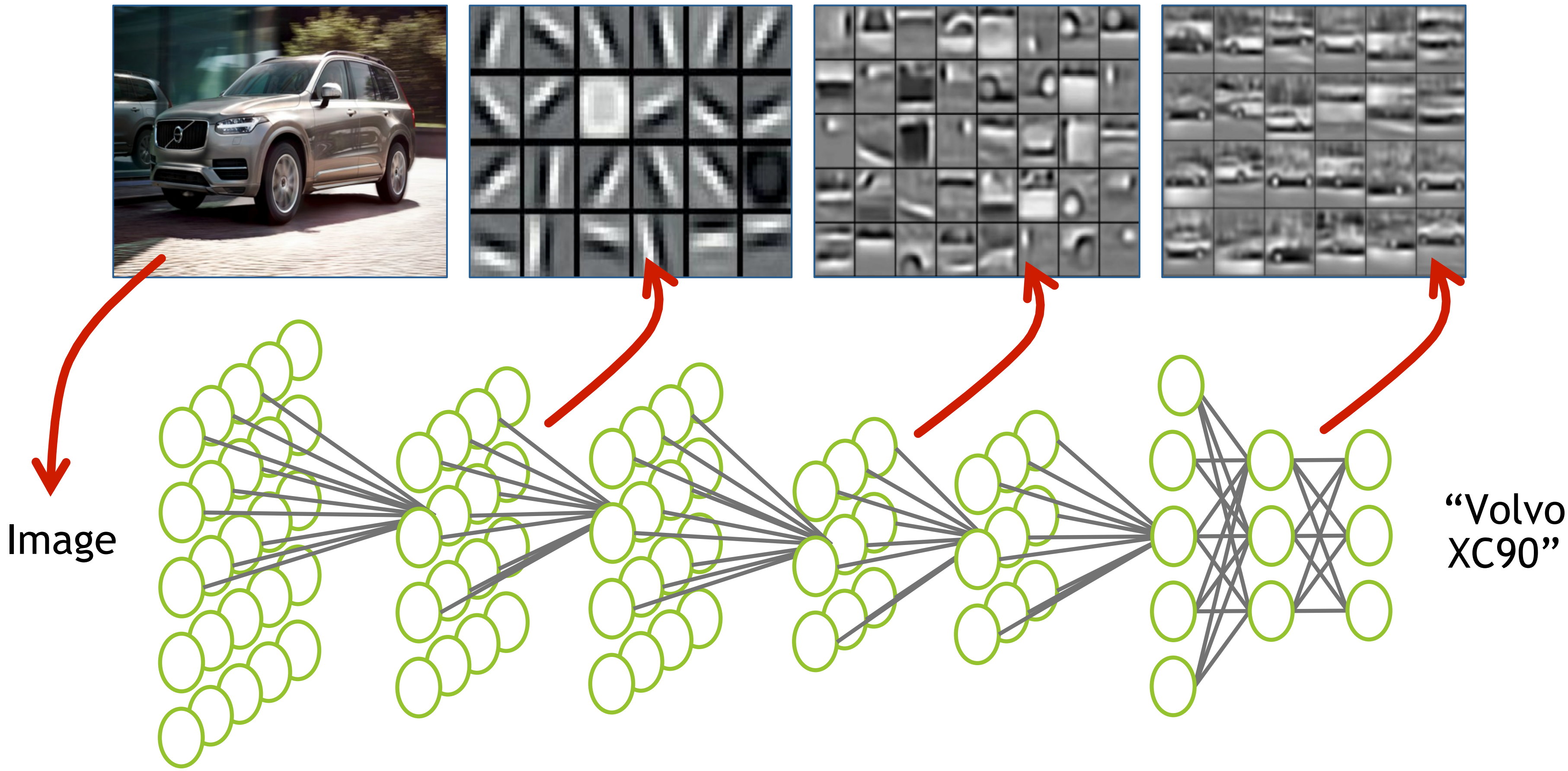


Image Source: [Lee et al., Comm. ACM 2011]

Why is Deep Learning Hot Now?

Big Data Availability

GPU Acceleration

New ML Techniques

facebook — **350M** images uploaded per day

Walmart — **2.5 Petabytes** of customer data hourly

YouTube — **300 hours** of video uploaded every minute

ImageNet Challenge

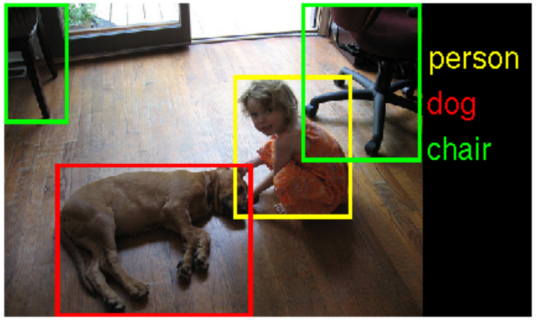
IMAGENET

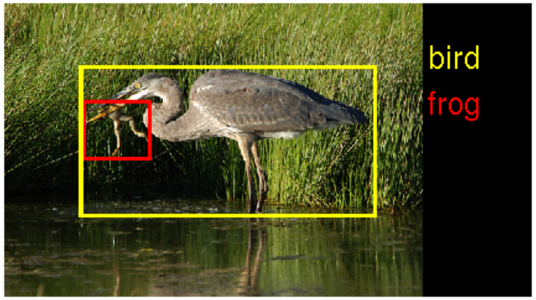
Image Classification Task:

1.2M training images • 1000 object categories

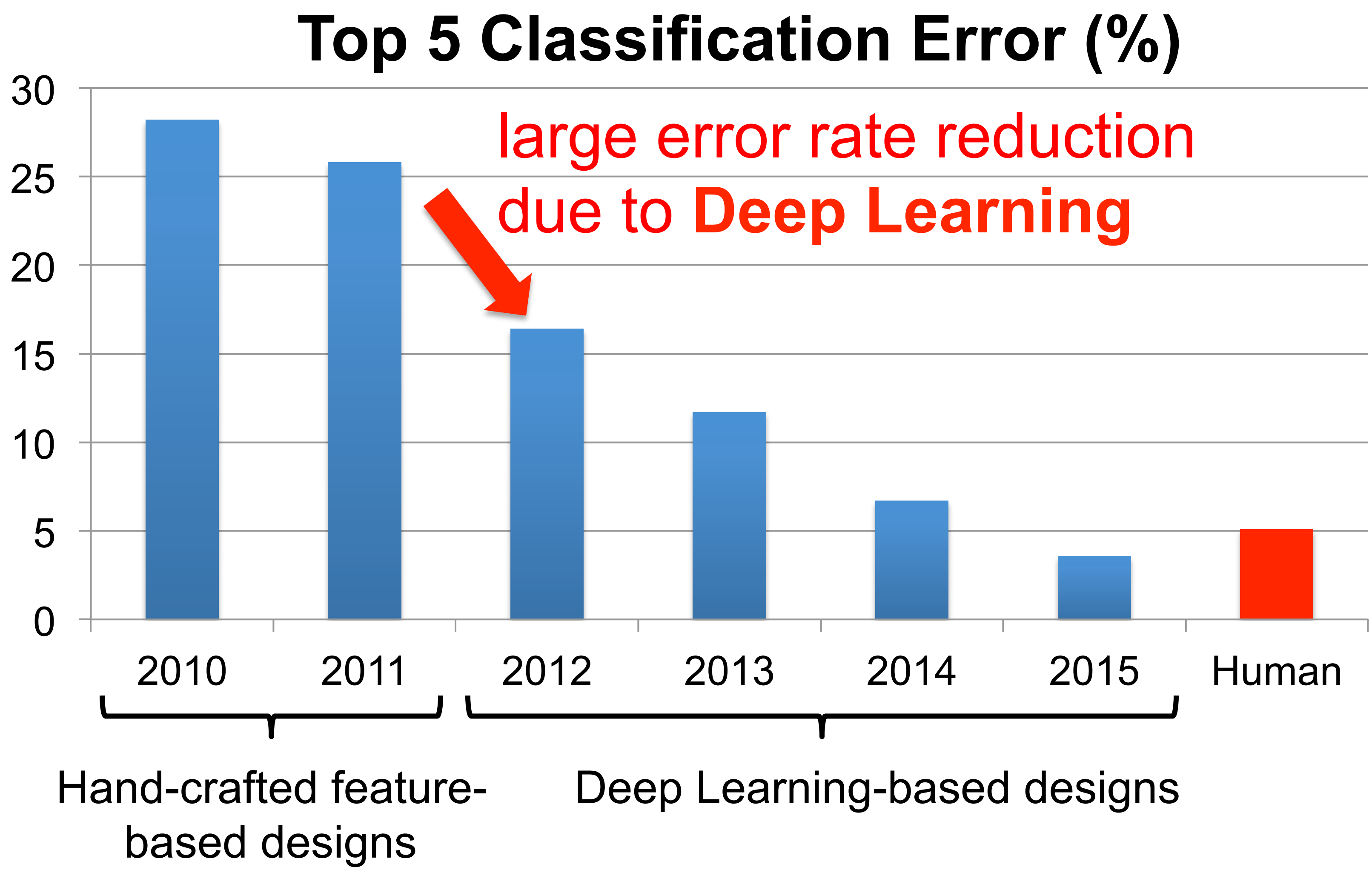
Object Detection Task:

456k training images • *200 object categories*



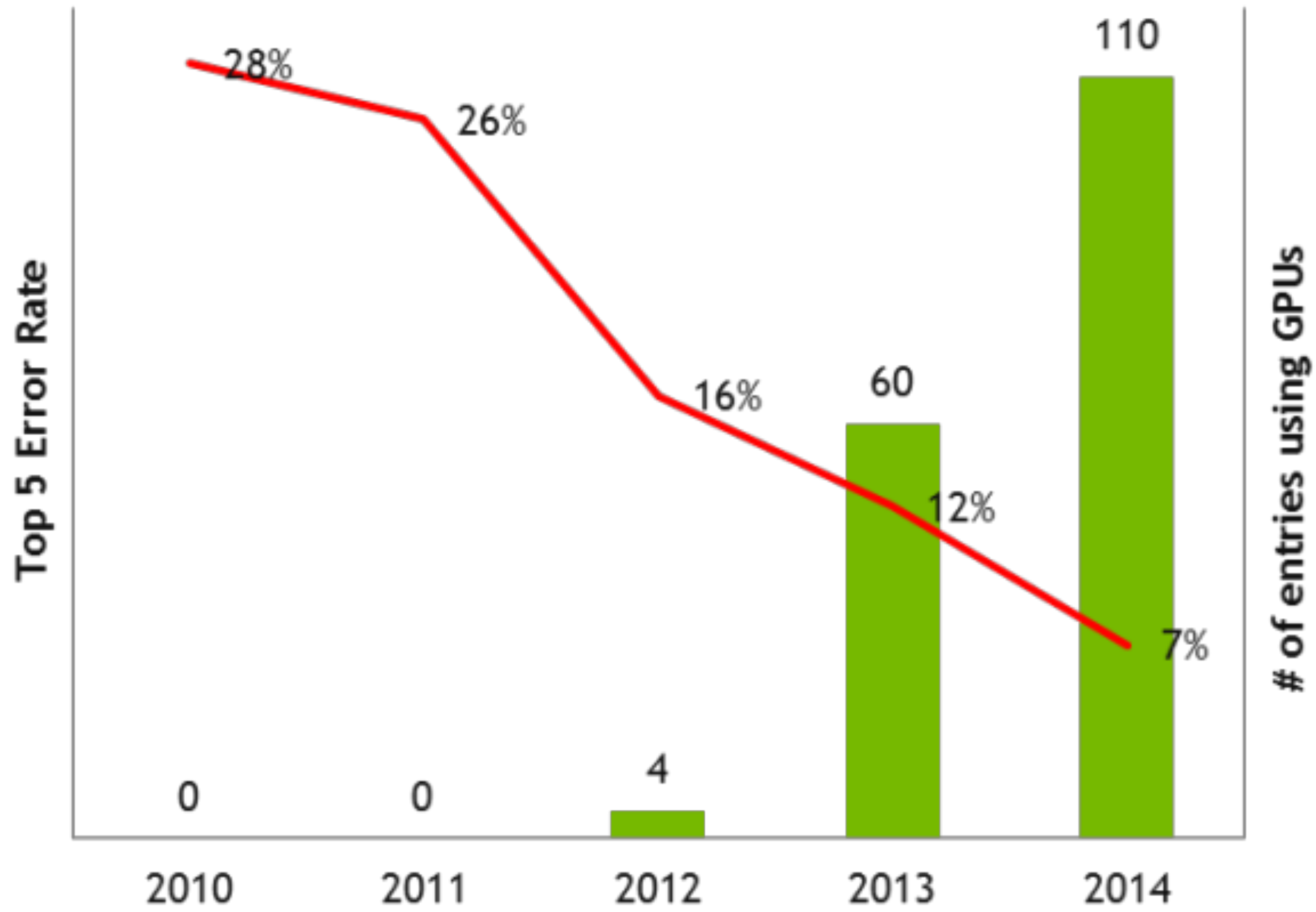


ImageNet: Image Classification Task



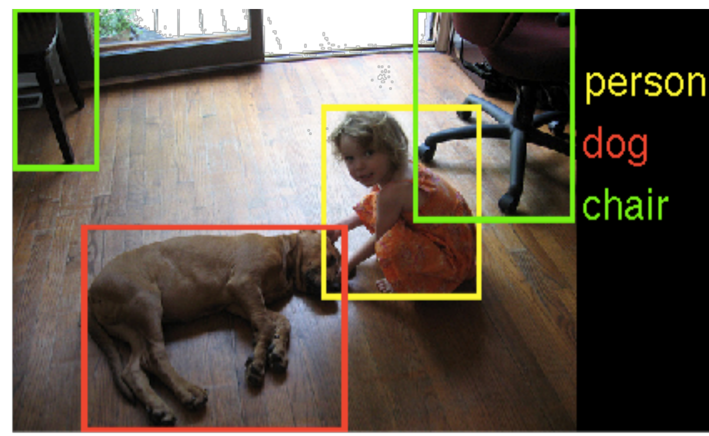
[Russakovsky et al., IJCV 2015]

GPU Usage for ImageNet Challenge



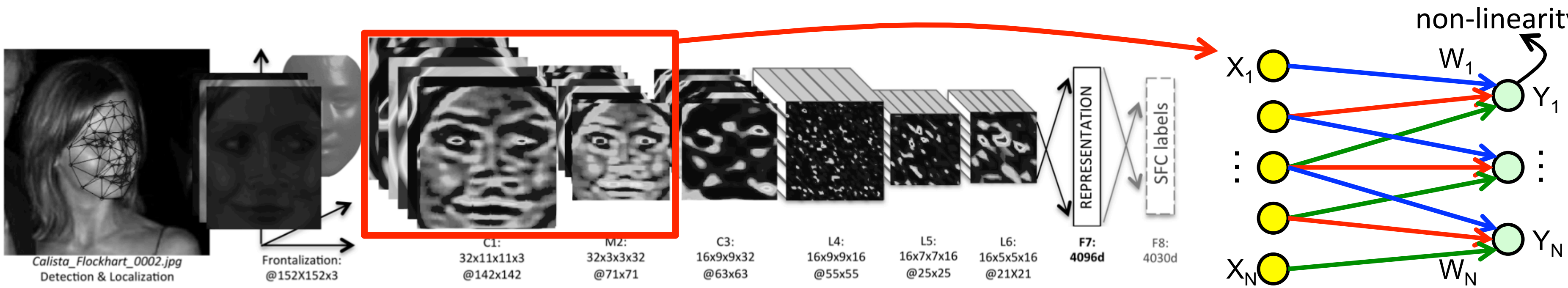
Deep Learning on Images

- Image Classification
- Object Localization
- Object Detection
- Image Segmentation
- Action Recognition
- Image Generation



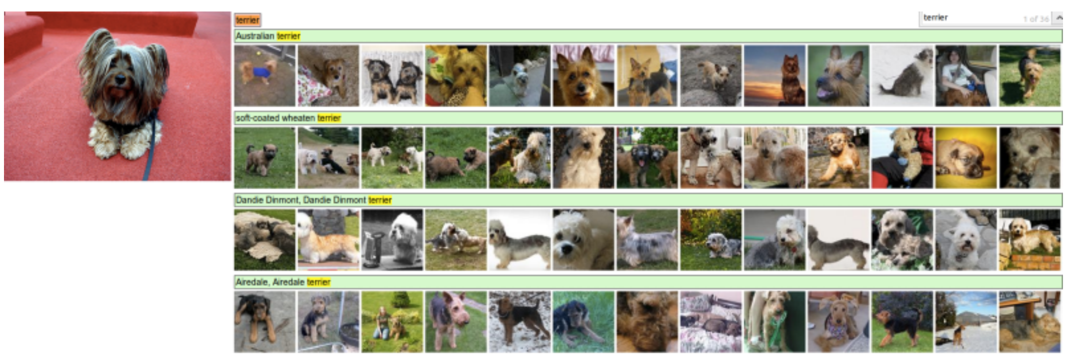
Human or *Superhuman* Accuracy Level

- Face recognition
  - Deep learning accuracy (97.25%) vs. Human accuracy (97.53%)



[Yaniv et al., CVPR 2014]

- Fine grained category recognition (e.g. dogs, monkeys, snakes, birds)
  - Deep learning errors: 7 vs. Human errors: 28

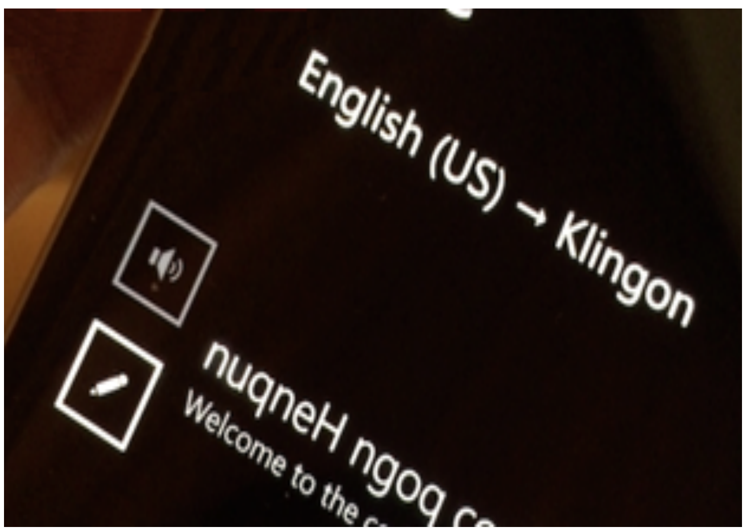


120 species of dogs

[O. Russakovsky et al., IJCV 2015]

Deep Learning for Speech

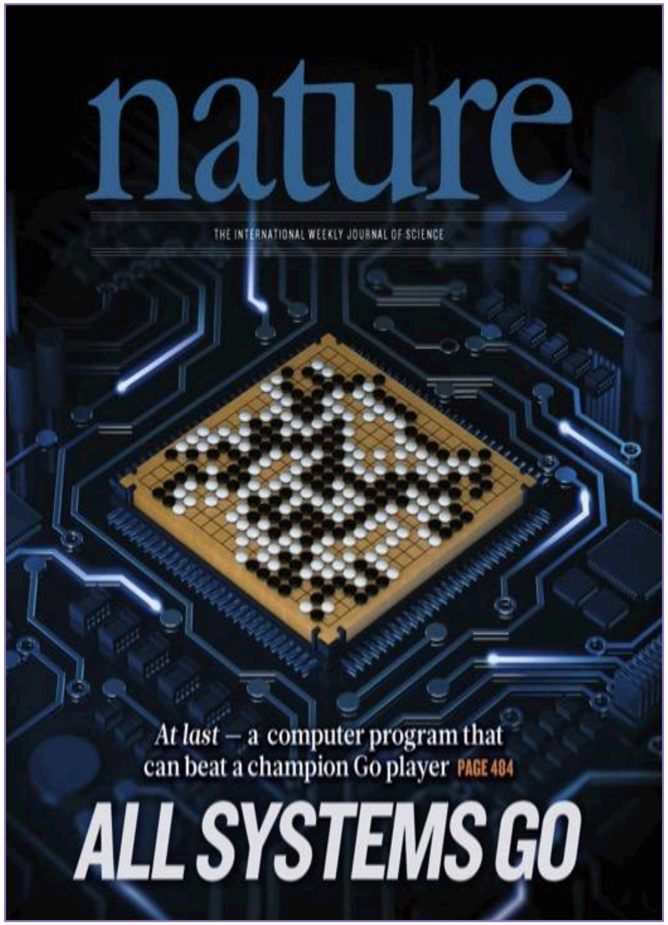
- **Speech Recognition**
- **Natural Language Processing**
- **Speech Translation**
- **Audio Generation**



Deep Learning on Games

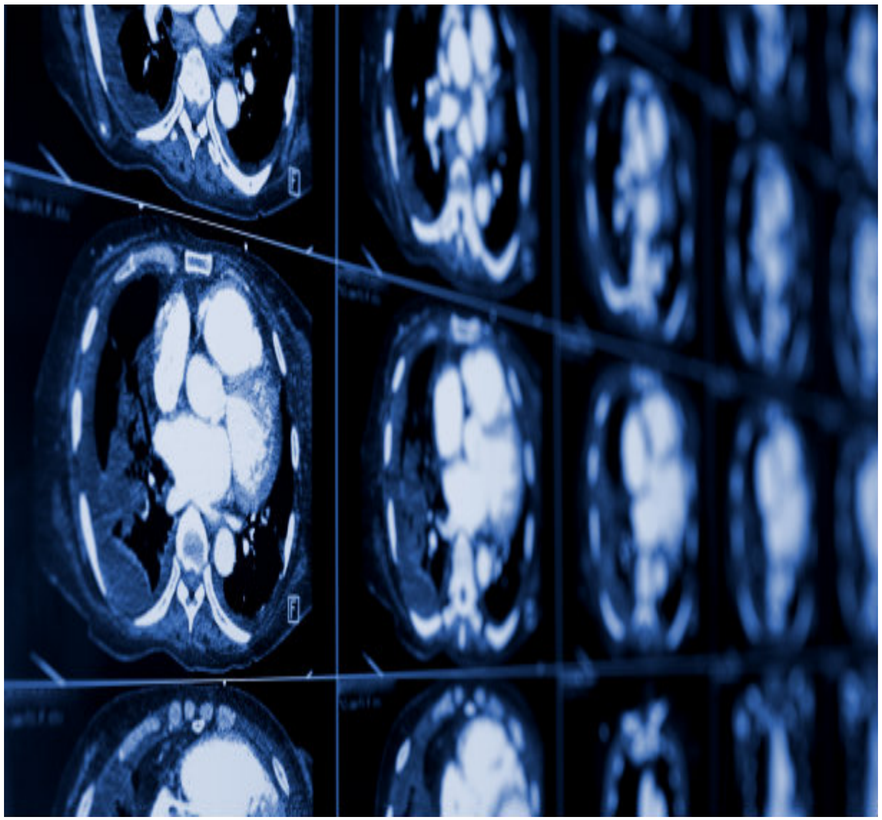
Google DeepMind AlphaGo

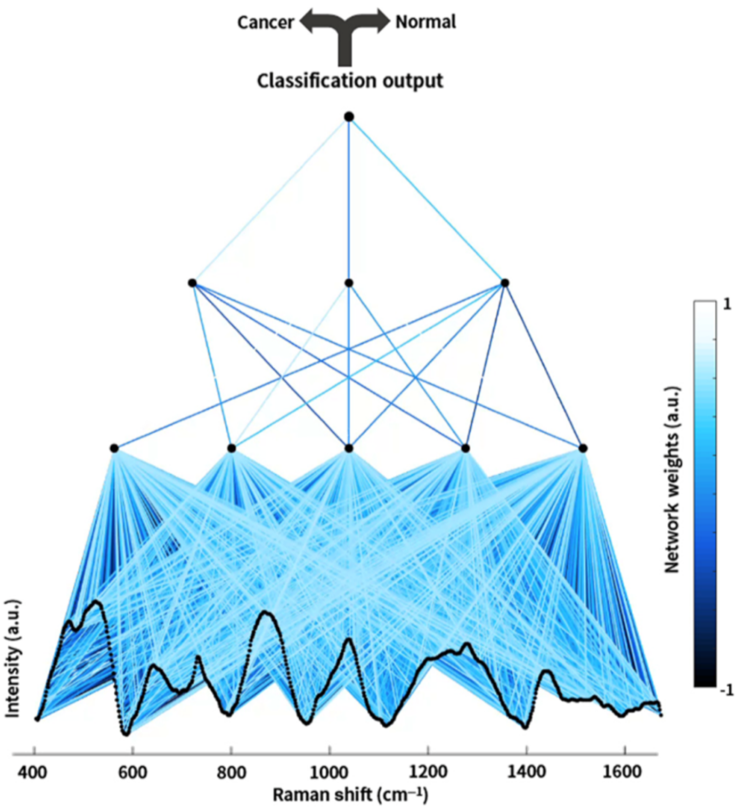
Go is exponentially more complex than chess (10^{170} legal positions)



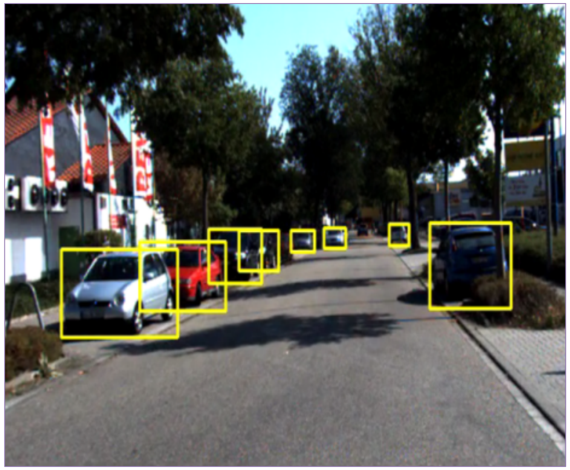
Medical Applications of Deep Learning

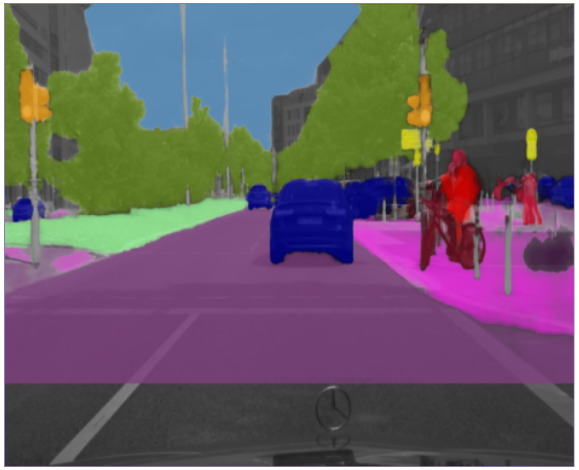
- **Brain Cancer Detection**

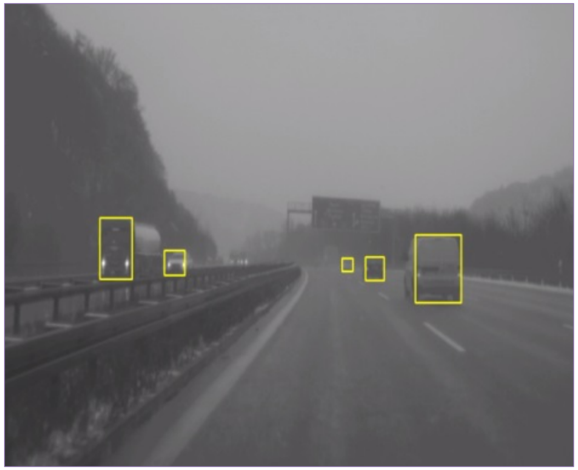




Deep Learning for Self-driving Cars







Other Emerging Applications

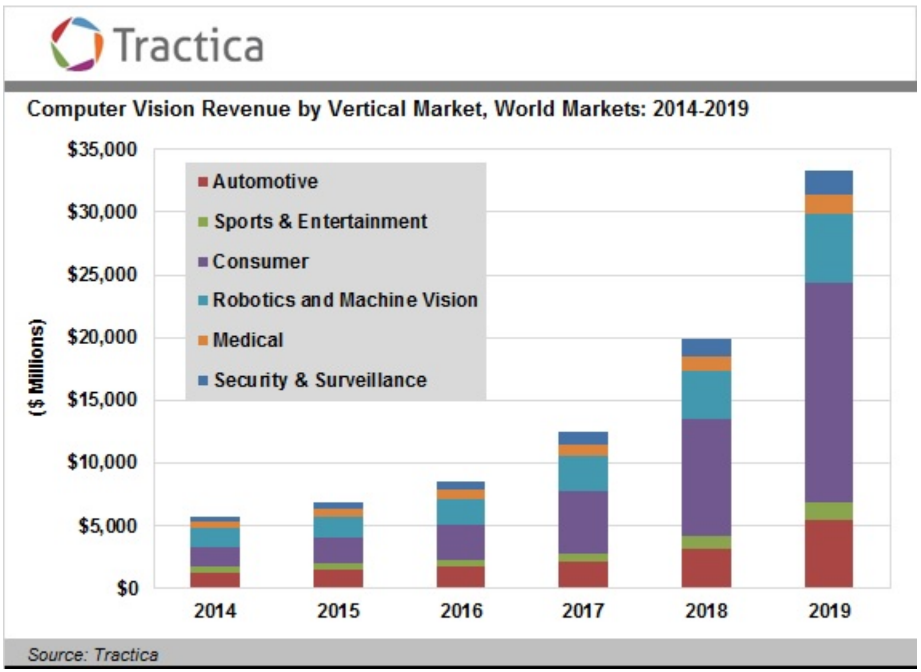
- **Medical** (Cancer Detection, Pre-Natal)
- **Finance** (Trading, Energy Forecasting, Risk)
- **Infrastructure** (Structure Safety and Traffic)
- Weather Forecasting and Event Detection

This talk will focus on image classification

http://www.nextplatform.com/2016/09/14/next-wave-deep-learning-applications/

Opportunities

$500B Market over 10 Years!



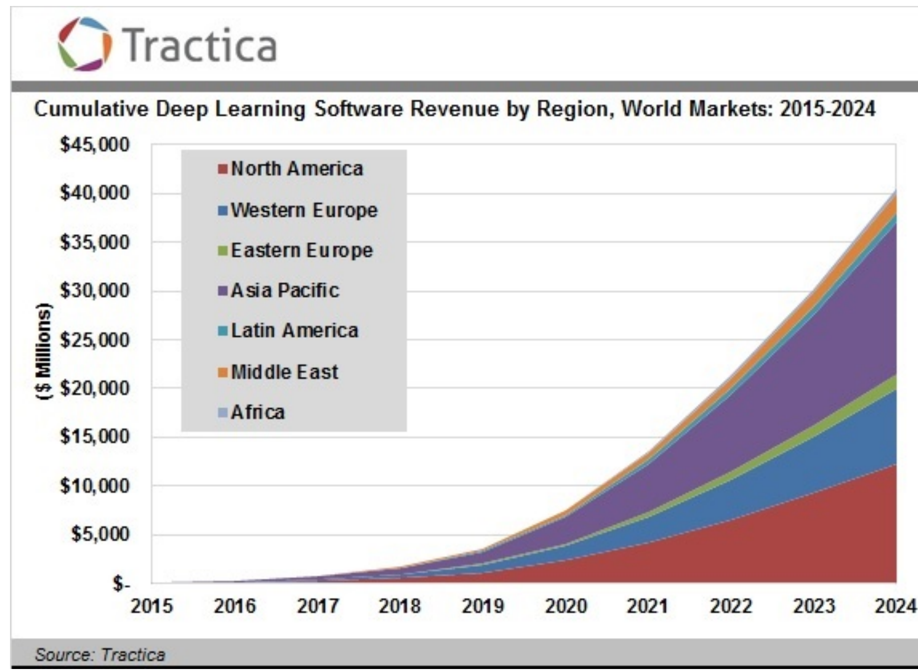


Image Source: Tractica

Opportunities

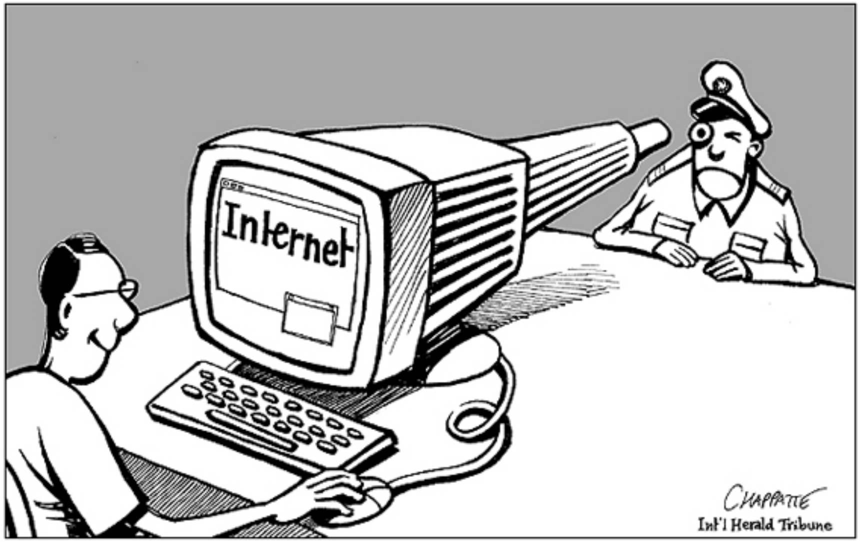
From EE Times – September 27, 2016

”Today the job of training machine learning models is limited by compute, if we had faster processors we'd run bigger models…in practice we train on a reasonable subset of data that can finish in a matter of months. We could use improvements of several orders of magnitude – 100x or greater.”

– Greg Diamos, Senior Researcher, SVAIL, Baidu

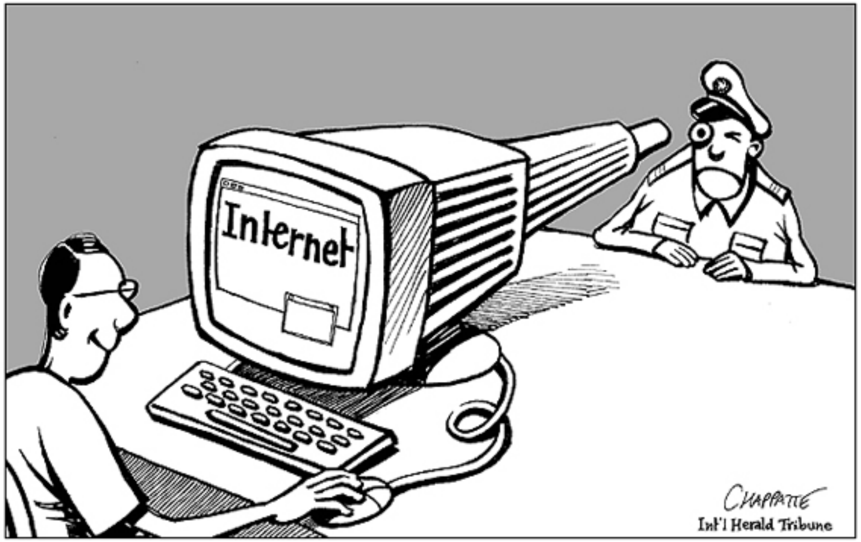
Processing at "Edge" instead of the "Cloud"

Privacy



Processing at "Edge" instead of the "Cloud"

Privacy



Latency

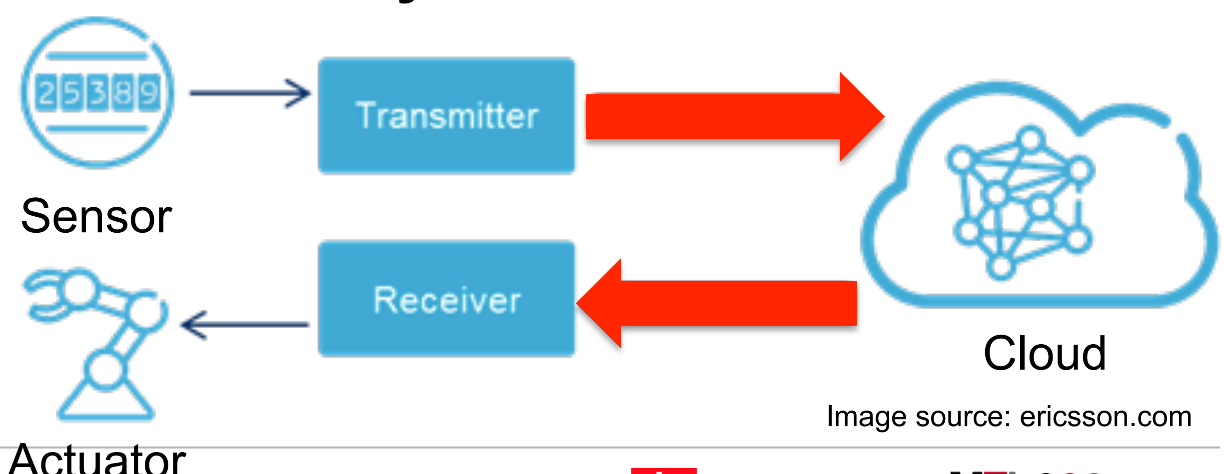
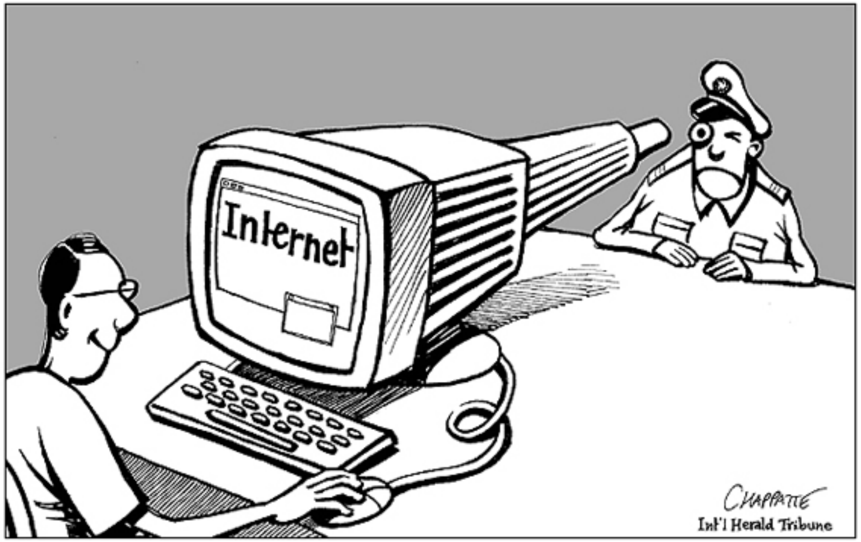


Image source: ericsson.com

Processing at “Edge” instead of the “Cloud”

Privacy



Communication

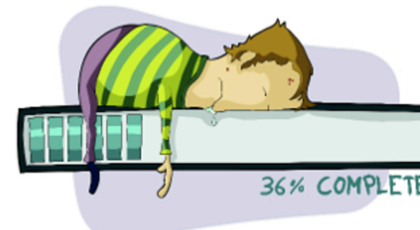


Image source:
www.theregister.co.uk

Latency

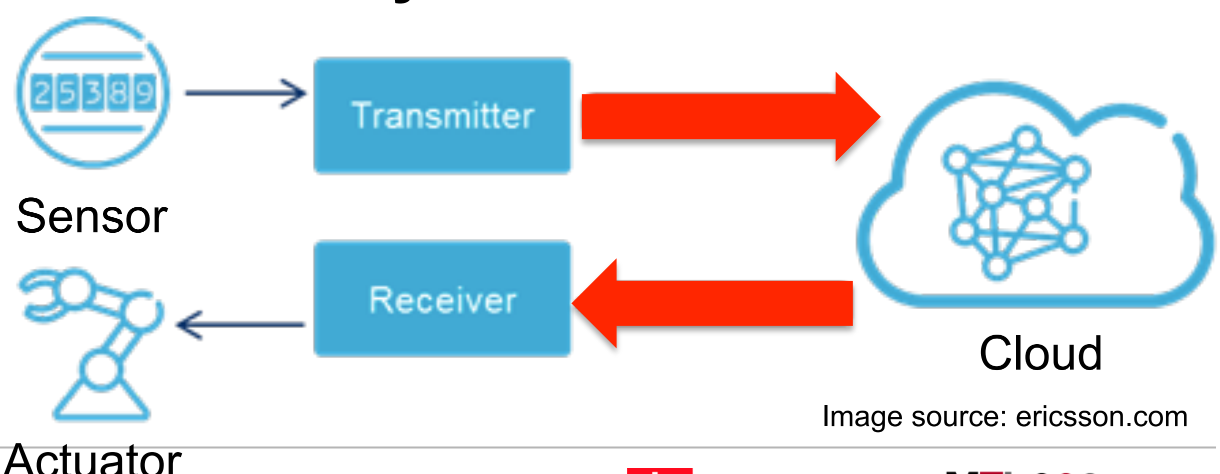


Image source: ericsson.com

Video is the Biggest Big Data

Over 70% of today's Internet traffic is video
Over 300 hours of video uploaded to YouTube **every minute**
Over 500 million hours of video surveillance collected **every day**

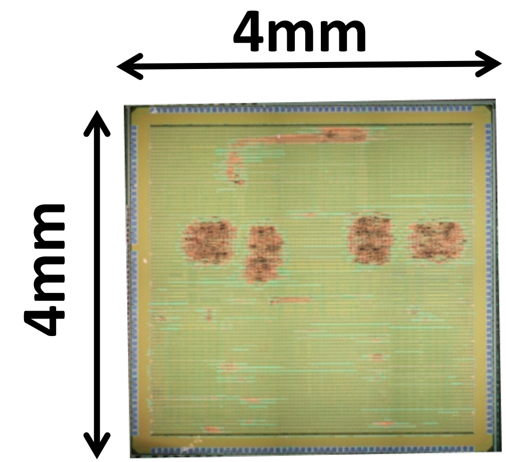
Energy limited due to battery capacity

Power limited due to heat dissipation

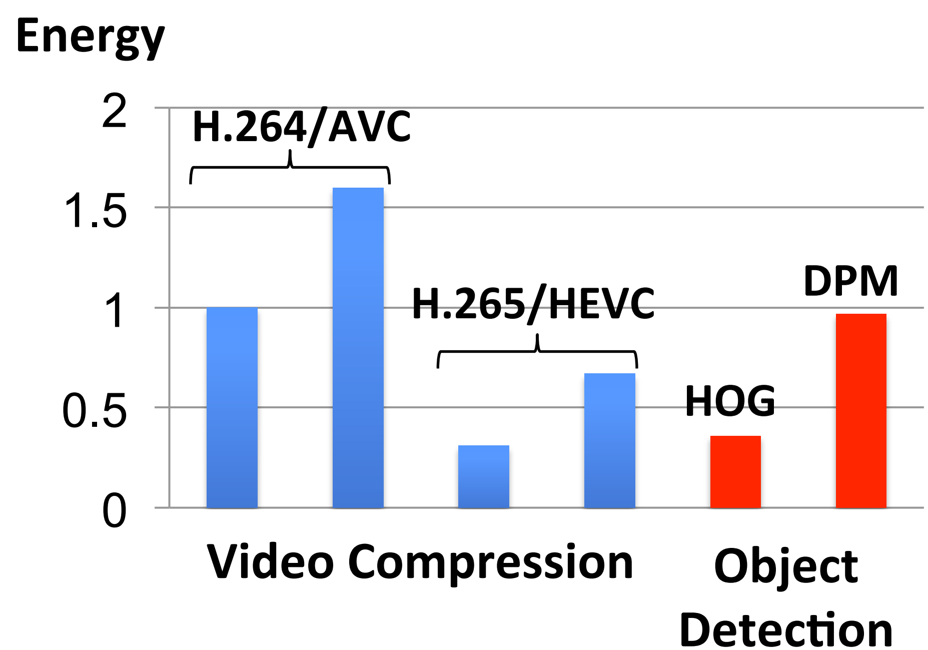
Need energy-efficient pixel processing!

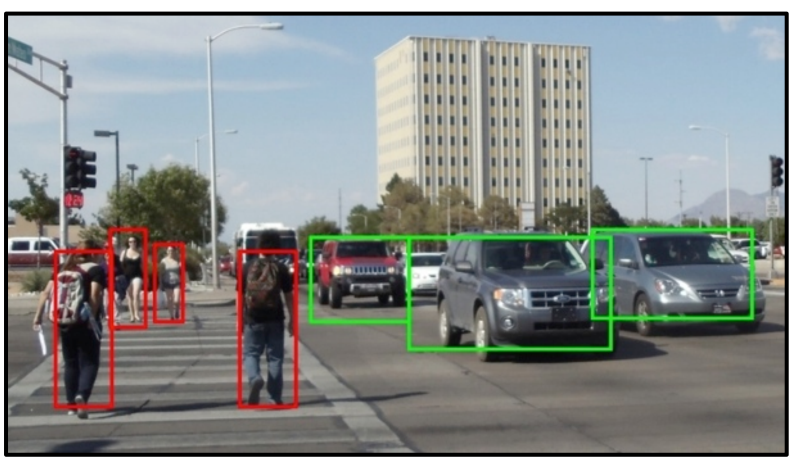
Typical Constraints on Video Coding

- **Area cost**
  - Memory Size 100-500kB
- **Power budget**
  - < 1W for smartphones
- **Throughput**
  - Real-time 30 fps
- **Energy**
  - ~1nJ/pixel



MIT Object Detection Chip [VLSI 2016]



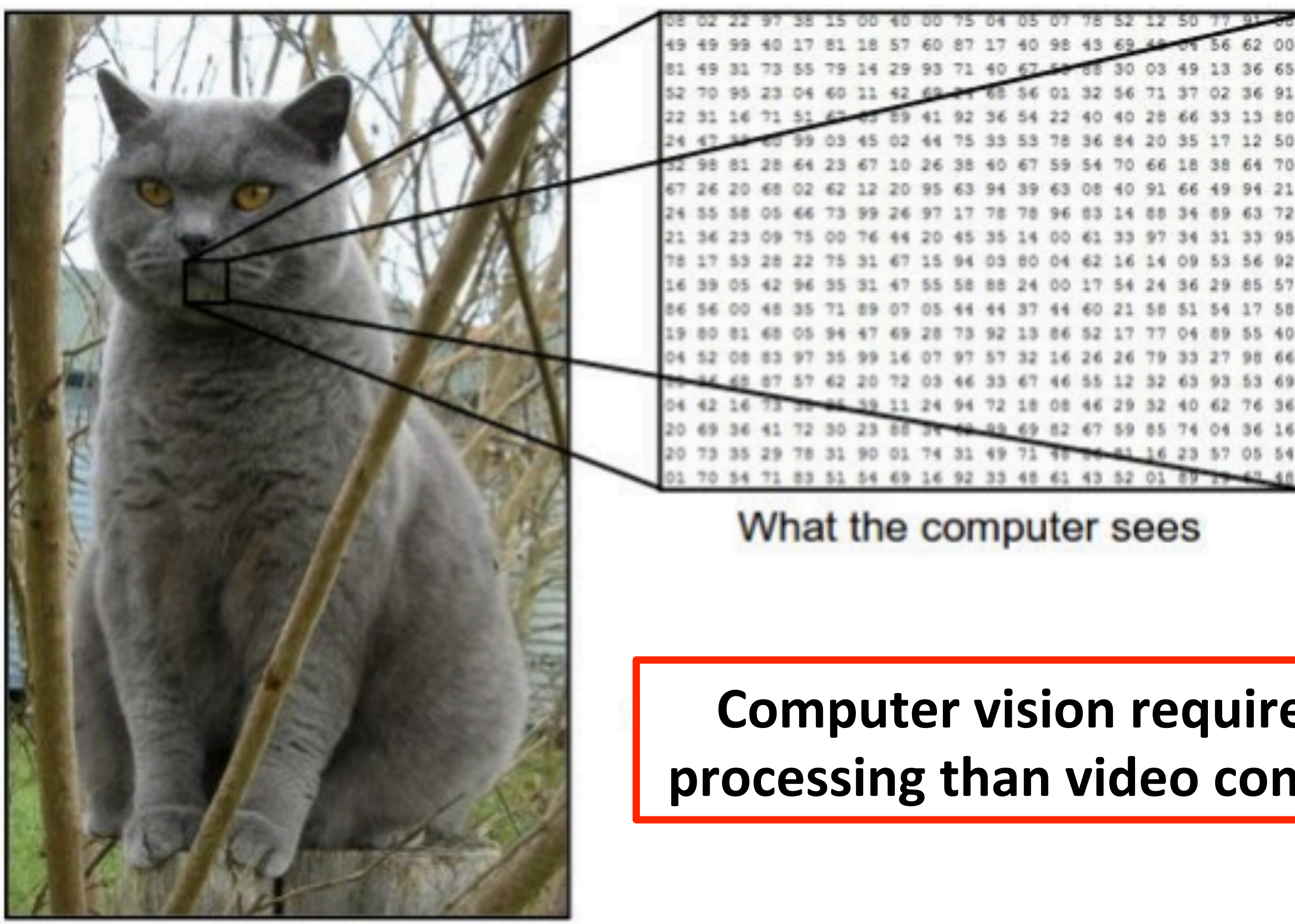


Why is Vision Difficult?

Cat

Image Source: Stanford cs231n

Why is Vision Difficult?



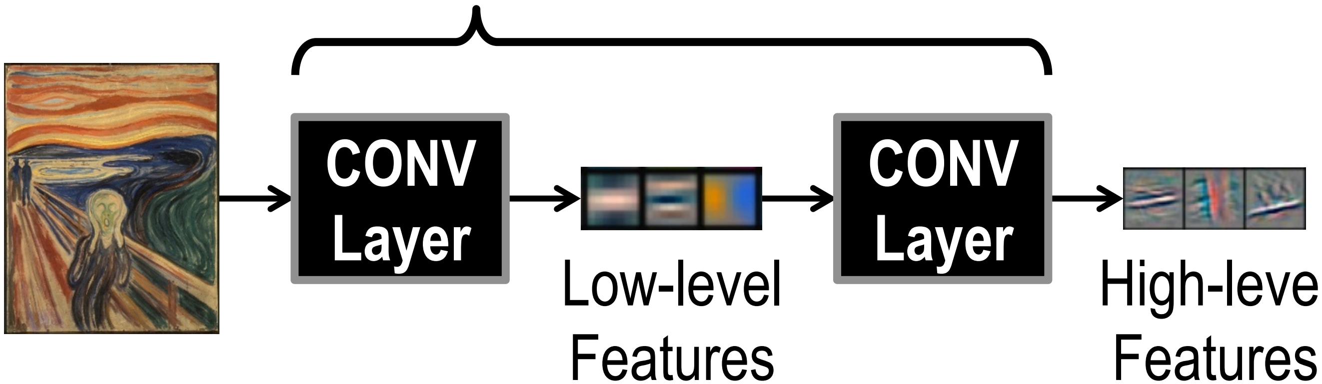
Computer vision requires more processing than video compression

Eyeriss: Energy-Efficient Hardware for DCNNs

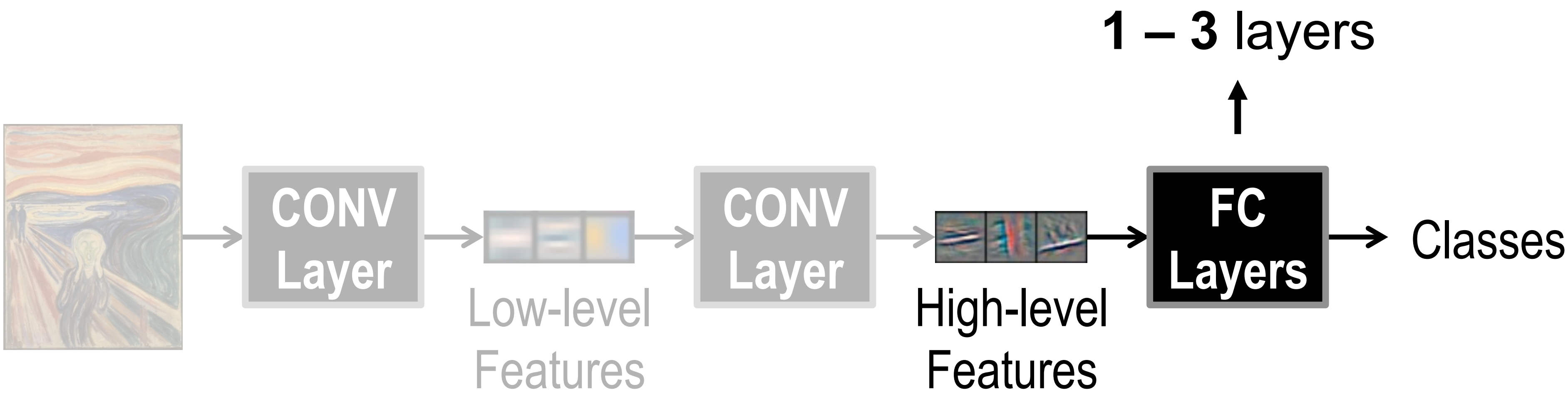
Yu-Hsin Chen, Tushar Krishna, Joel Emer, Vivienne Sze, ISSCC 2016 [paper] / ISCA 2016 [paper]

Deep Convolutional Neural Networks

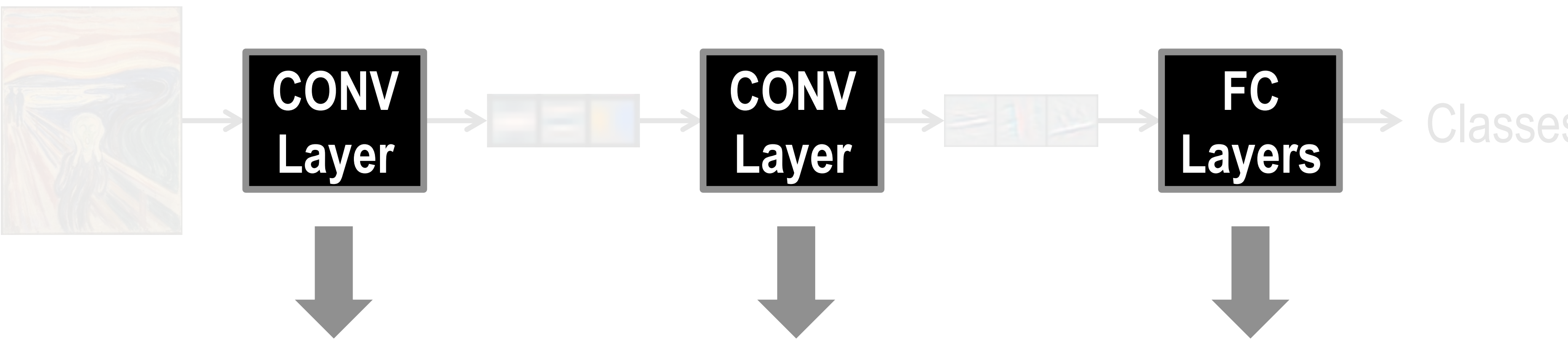
Modern ***deep*** CNN: up to **1000** CONV layers

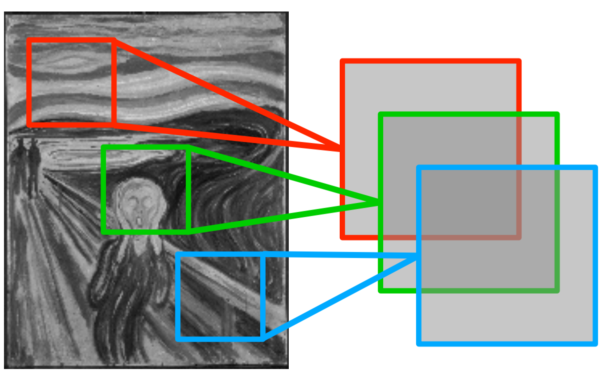


Deep Convolutional Neural Networks



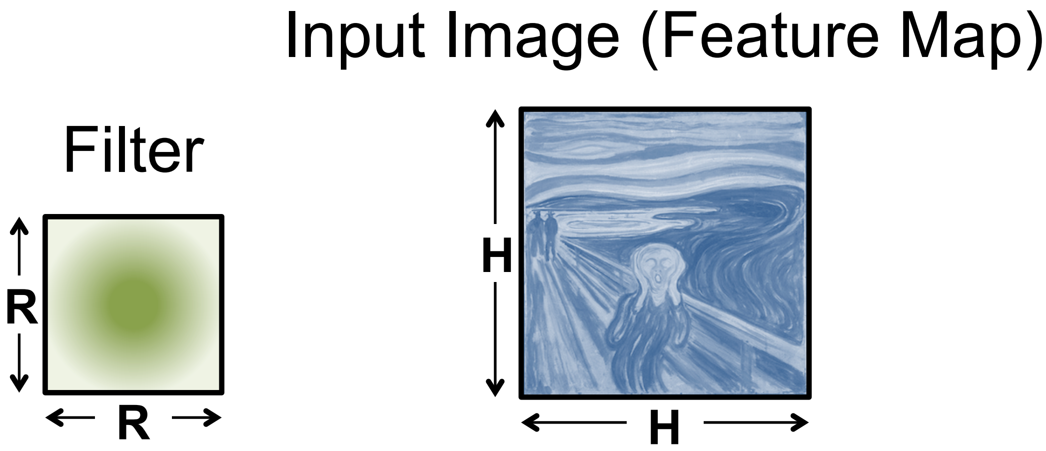
Deep Convolutional Neural Networks



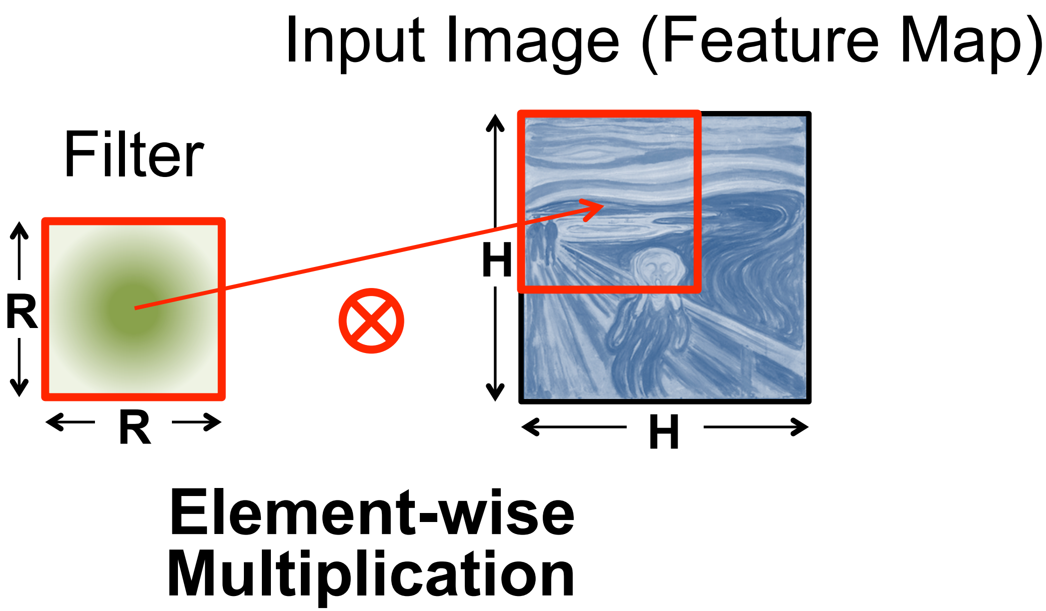


Convolutions account for more than 90% of overall computation, dominating **runtime** and **energy consumption**

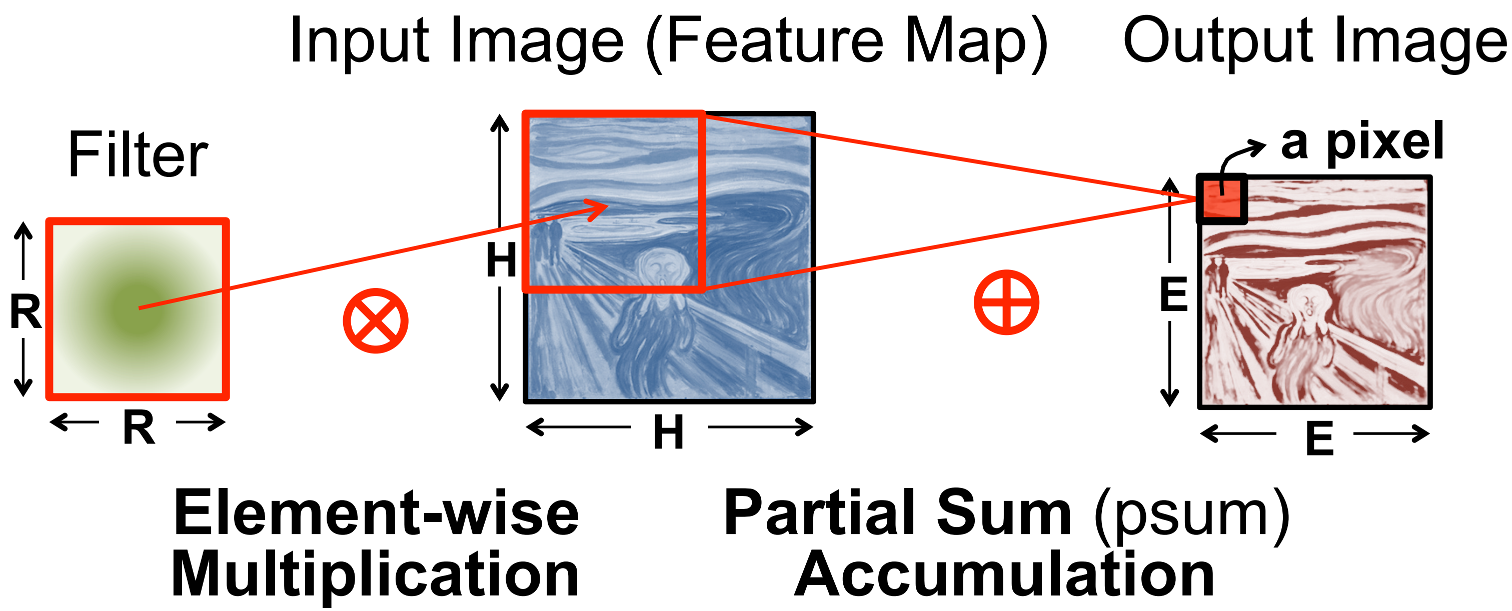
High-Dimensional CNN Convolution



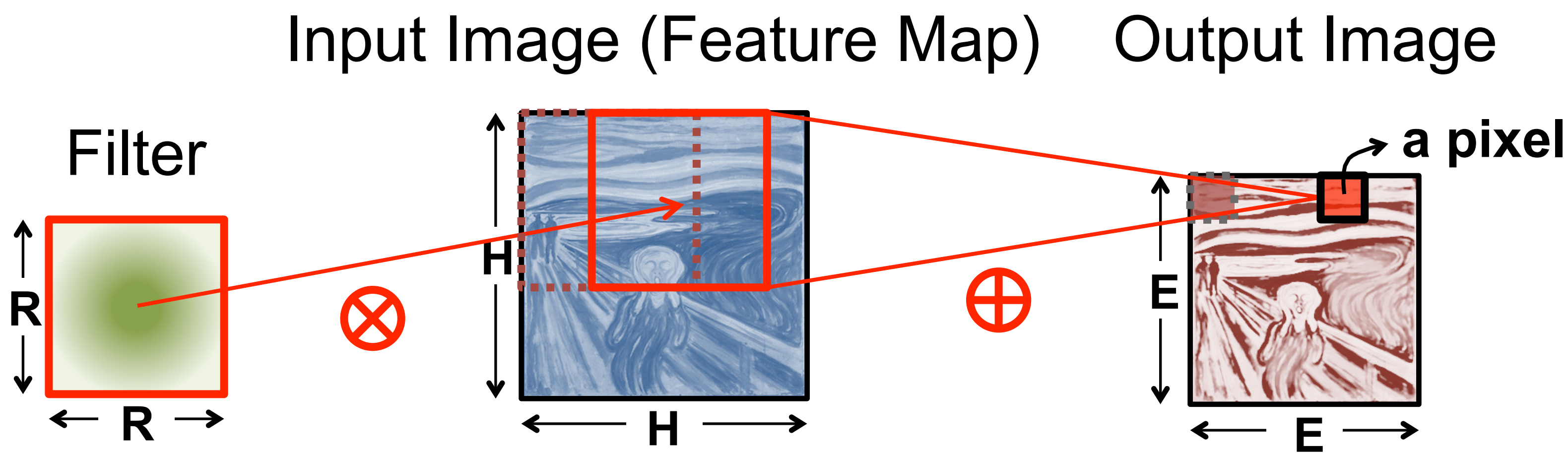
High-Dimensional CNN Convolution



High-Dimensional CNN Convolution

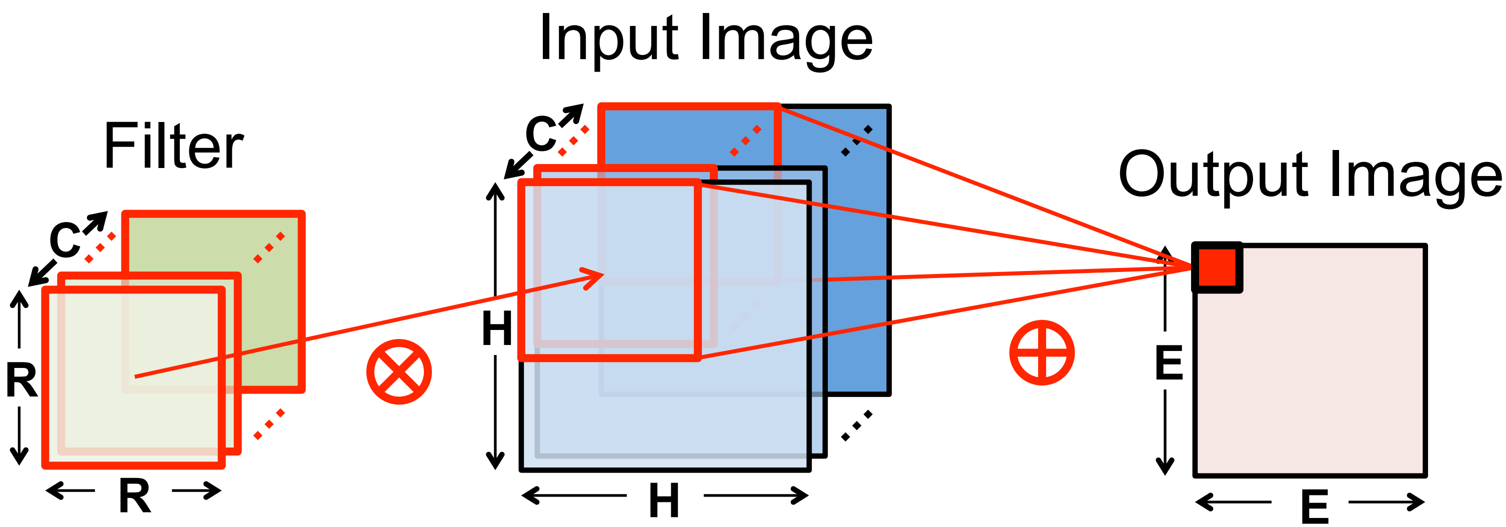


High-Dimensional CNN Convolution



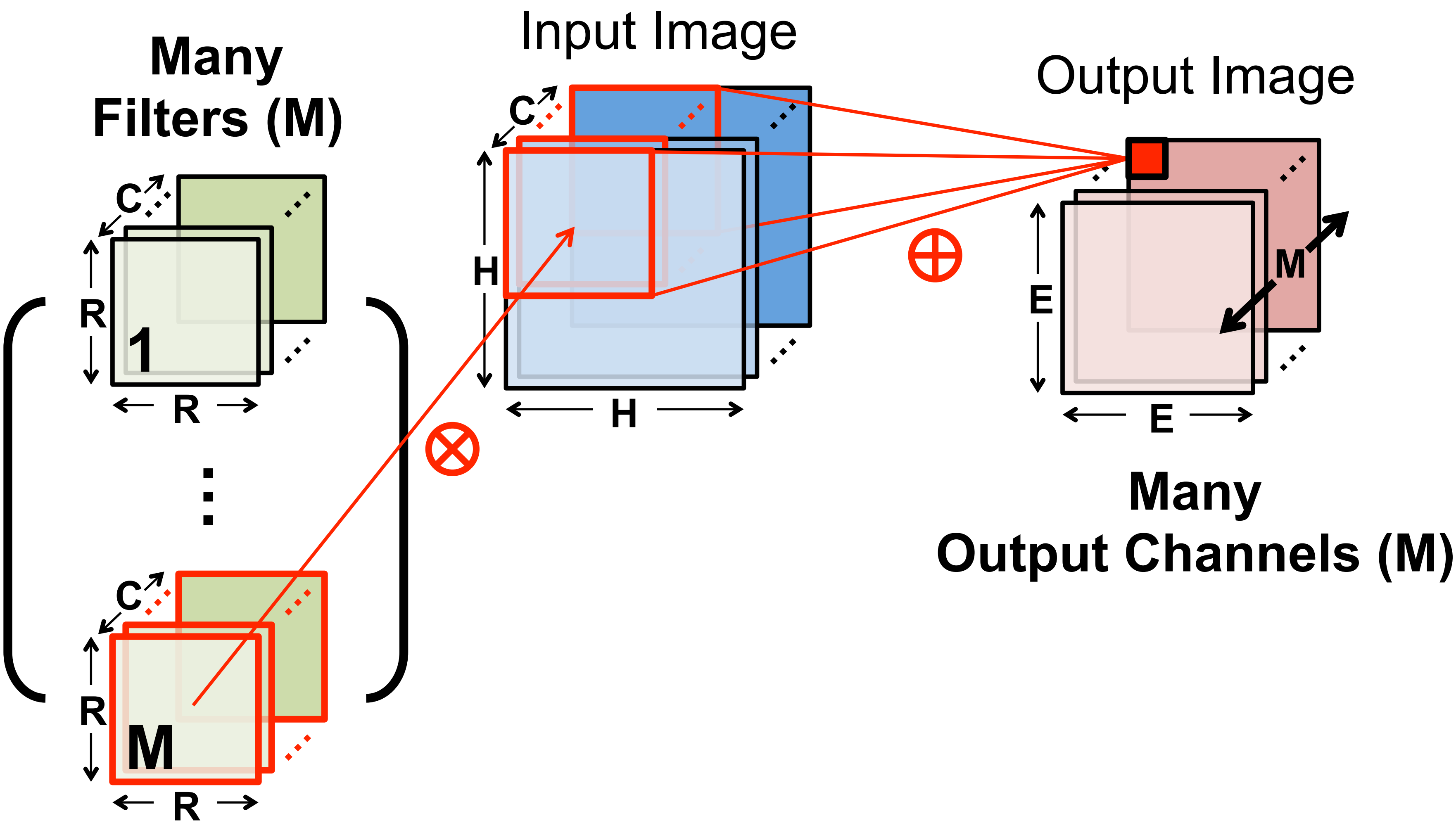
Sliding Window Processing

High-Dimensional CNN Convolution



Many Input Channels (C)

High-Dimensional CNN Convolution



AlexNet: 96 – 384 Filters (M)

High-Dimensional CNN Convolution

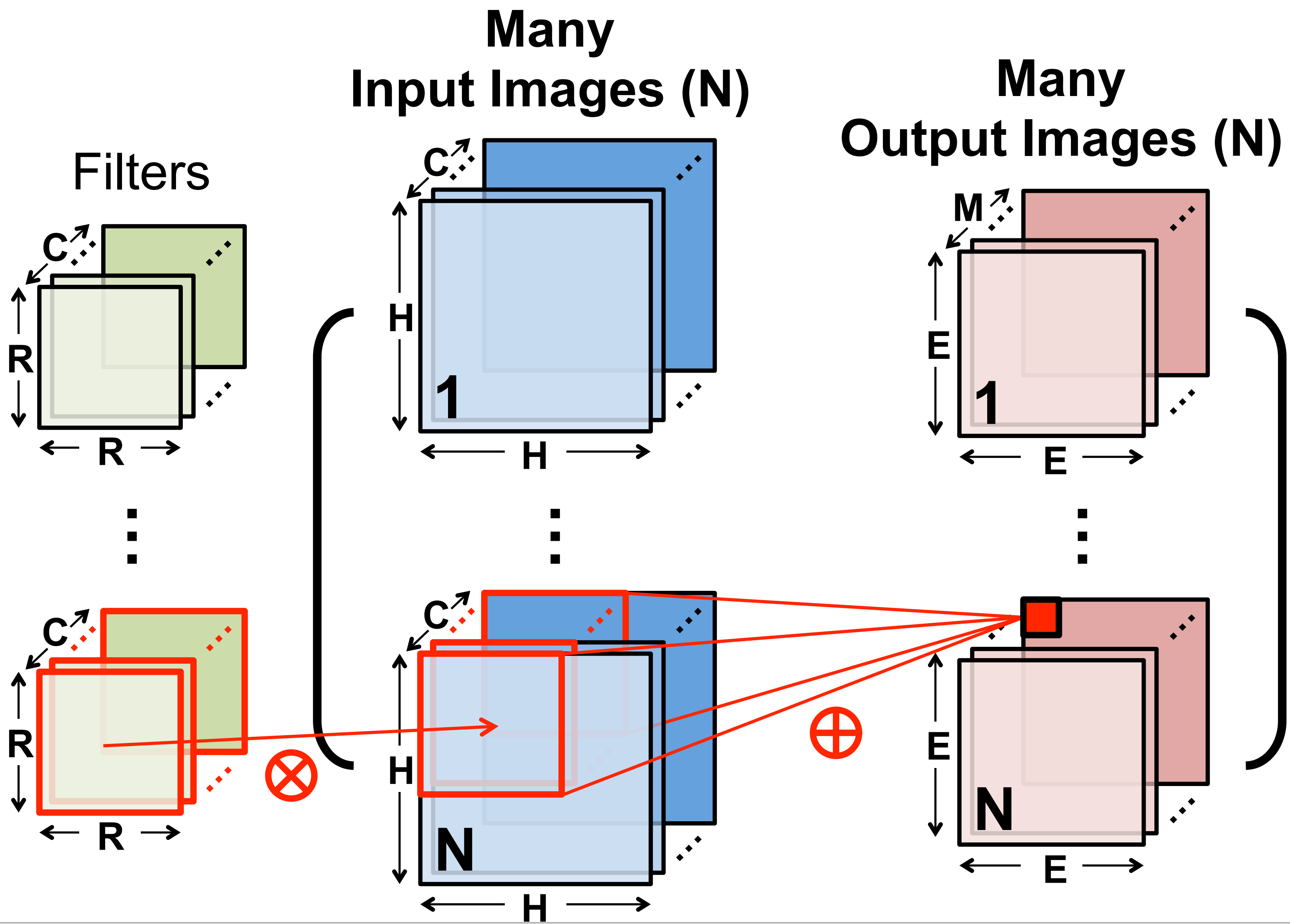
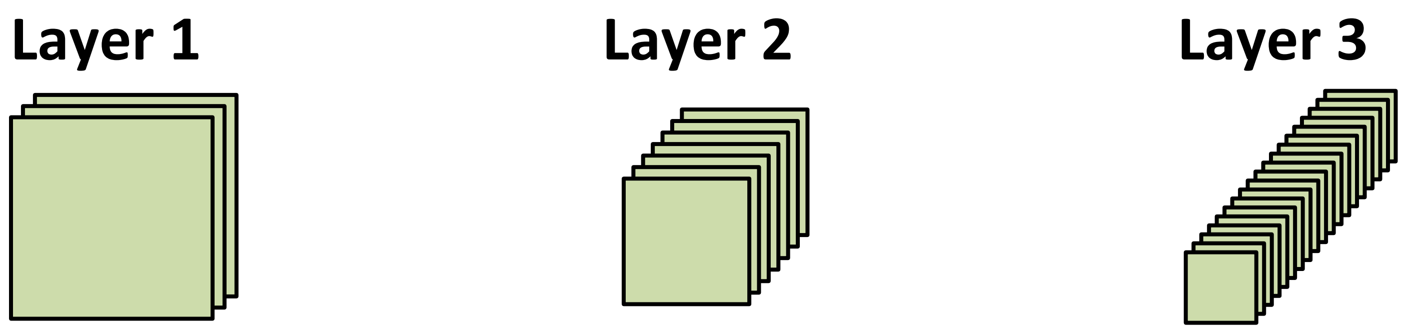


Image batch size: 1 – 256 (N)

Large Sizes with Varying Shapes

AlexNet[1] Convolutional Layer Configurations

Layer	Filter Size (R)	# Filters (M)	# Channels (C)	Stride
1	11x11	96	3	4
2	5x5	256	48	1
3	3x3	384	256	1
4	3x3	384	192	1
5	3x3	256	192	1



34k Params
105M MACs

307k Params
224M MACs

885k Params
150M MACs

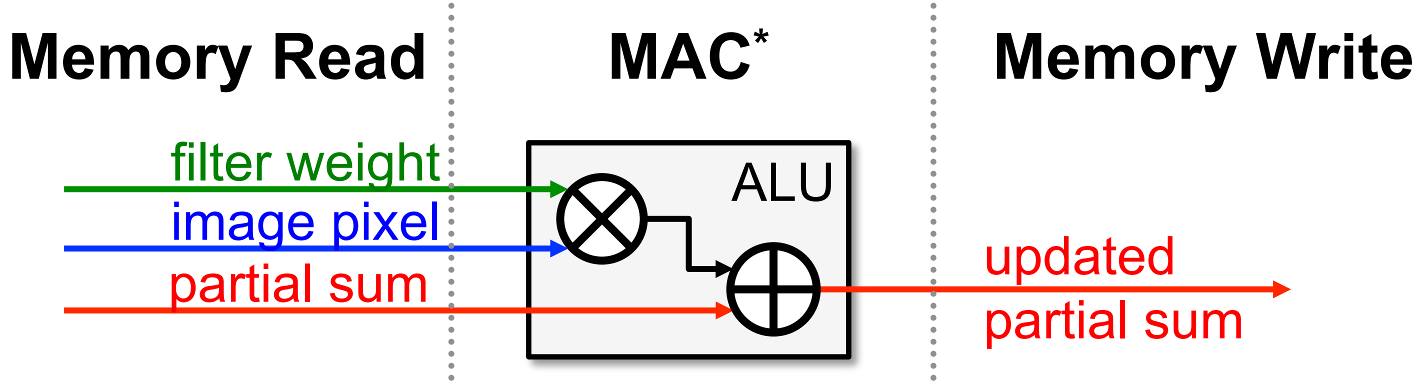
1. [Krizhevsky, NIPS 2012]

Properties We Can Leverage

- Operations exhibit **high parallelism**
  → **high throughput** possible

Properties We Can Leverage

- Operations exhibit **high parallelism** → **high throughput** possible
- Memory Access is the Bottleneck

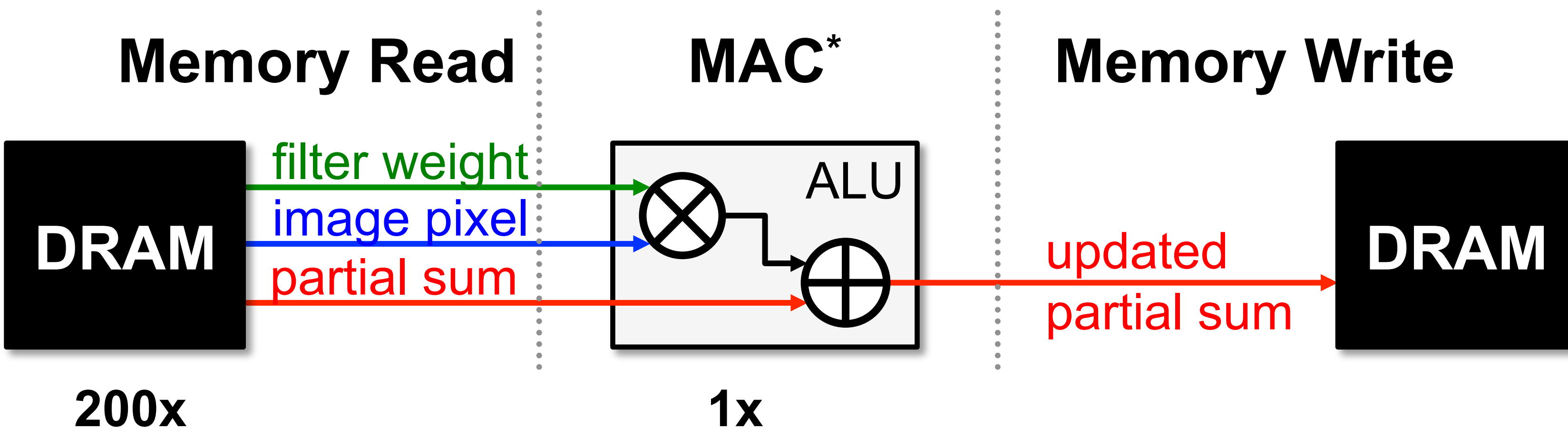


* multiply-and-accumulate

Properties We Can Leverage

- Operations exhibit **high parallelism**
  → **high throughput** possible

- Memory Access is the Bottleneck

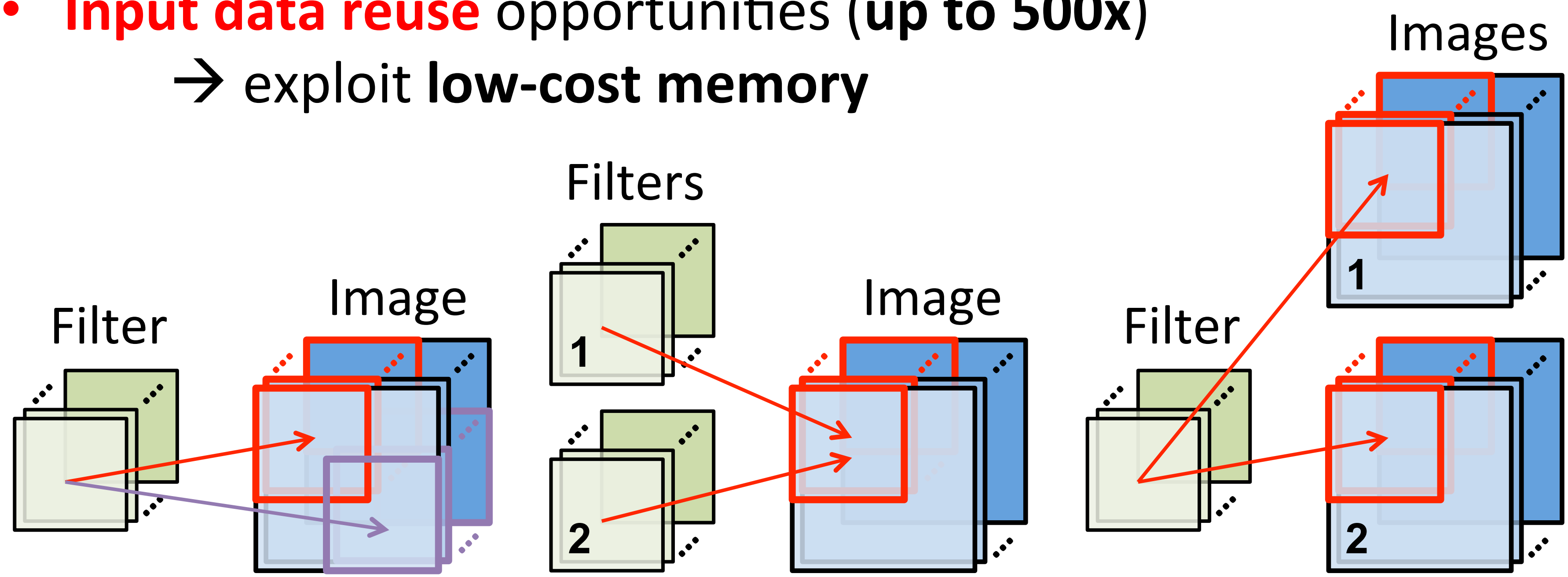


<u>Worst Case</u>: all memory R/W are **DRAM** accesses

- Example: AlexNet [NIPS 2012] has **724M** MACs
  → **2896M** DRAM accesses required

Properties We Can Leverage

- Operations exhibit **high parallelism**
  → **high throughput** possible
- **Input data reuse** opportunities (**up to 500x**)
  → exploit **low-cost memory**



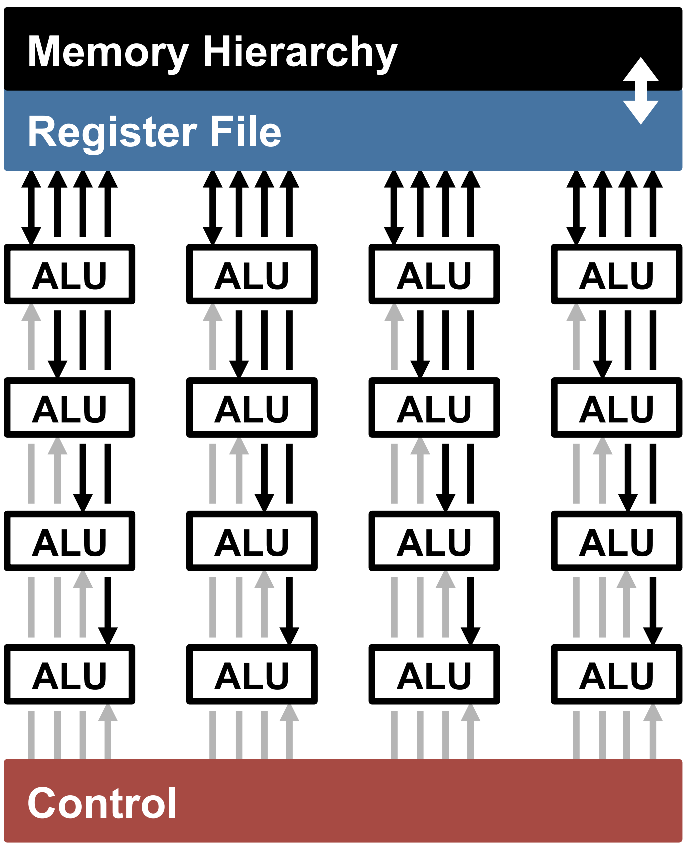
Convolutional Reuse (pixels, weights)

Image Reuse (pixels)

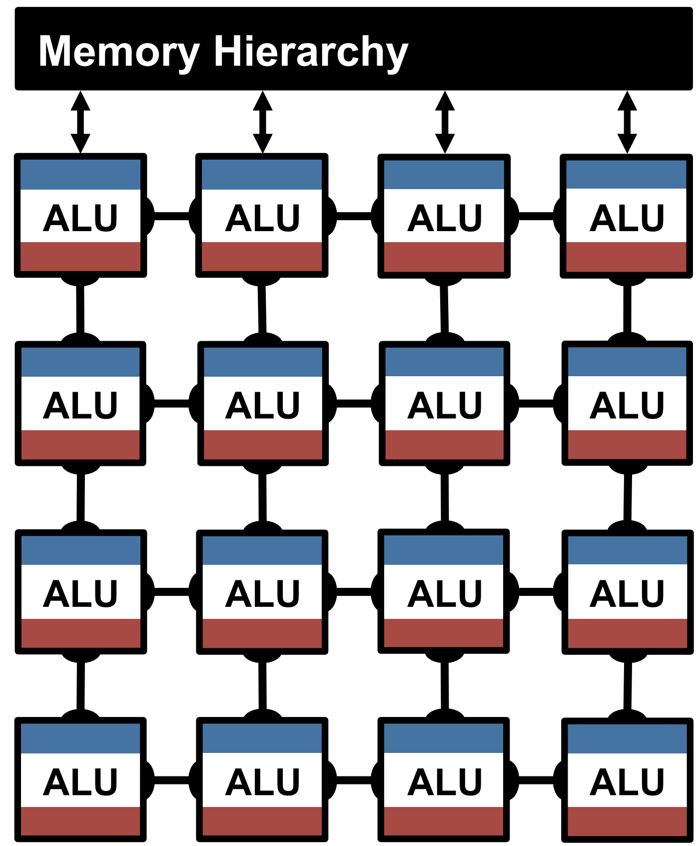
Filter Reuse (weights)

Highly-Parallel Compute Paradigms

Temporal Architecture (SIMD/SIMT)



Spatial Architecture (Dataflow Processing)



Advantages of Spatial Architecture

Temporal Architecture (SIMD/SIMT)

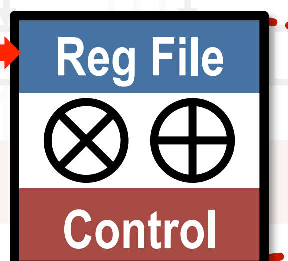
Spatial Architecture (Dataflow Processing)

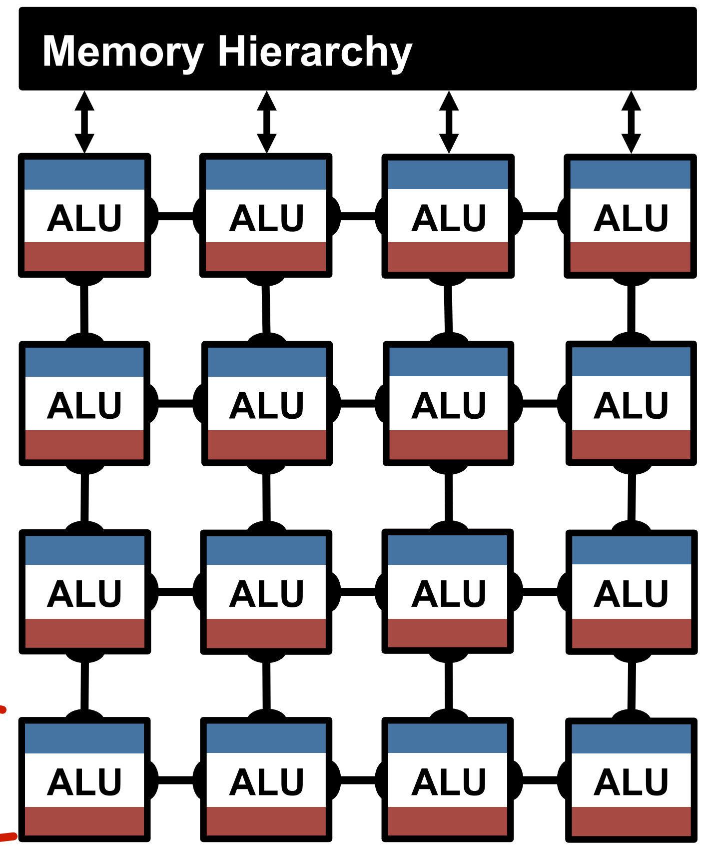
Efficient Data Reuse
Distributed local storage (RF)

Inter-PE Communication
Sharing among regions of PEs

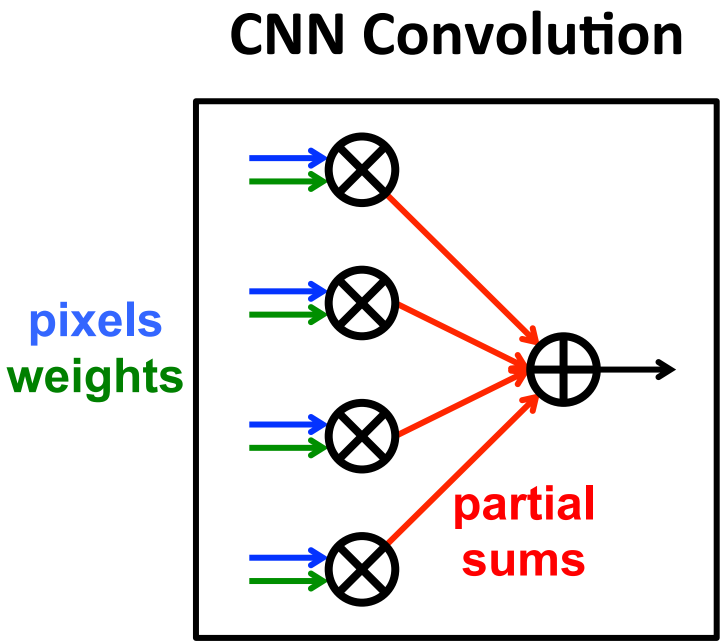
Processing Element (PE)

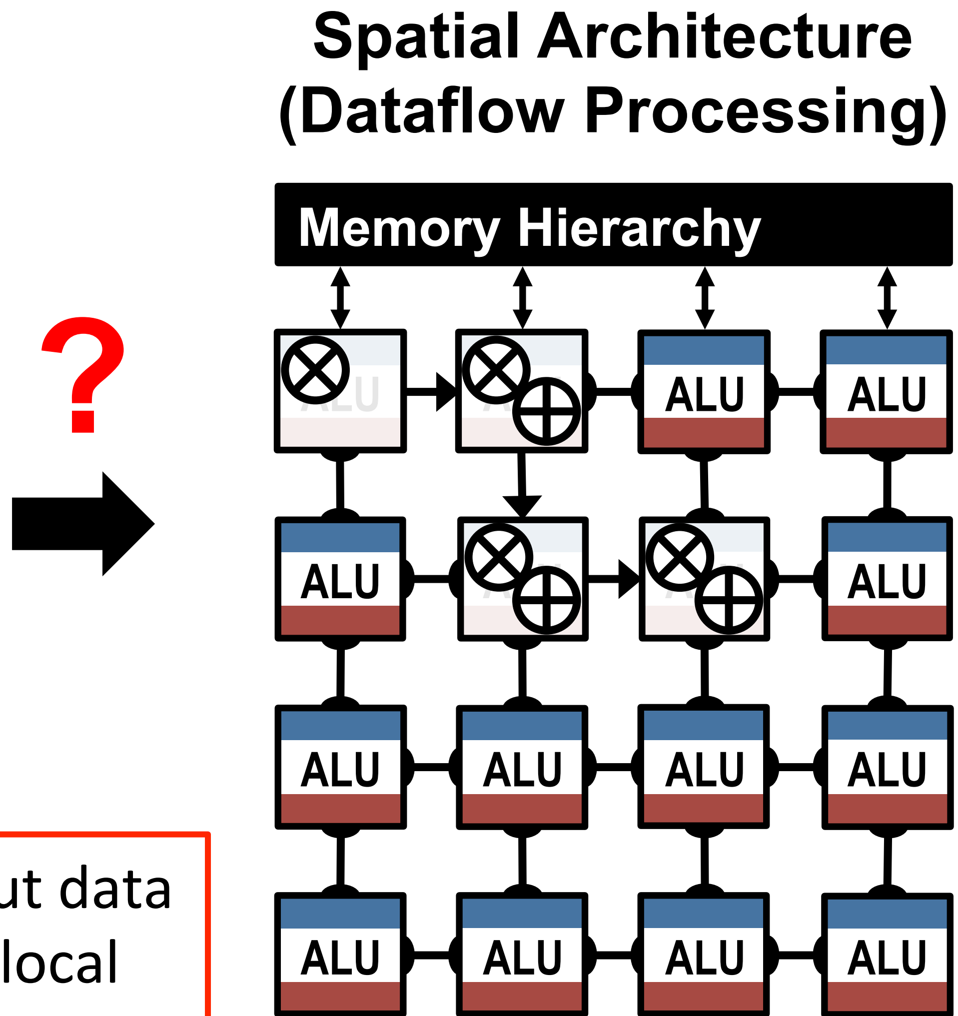
0.5 – 1.0 kB





How to Map the Dataflow?





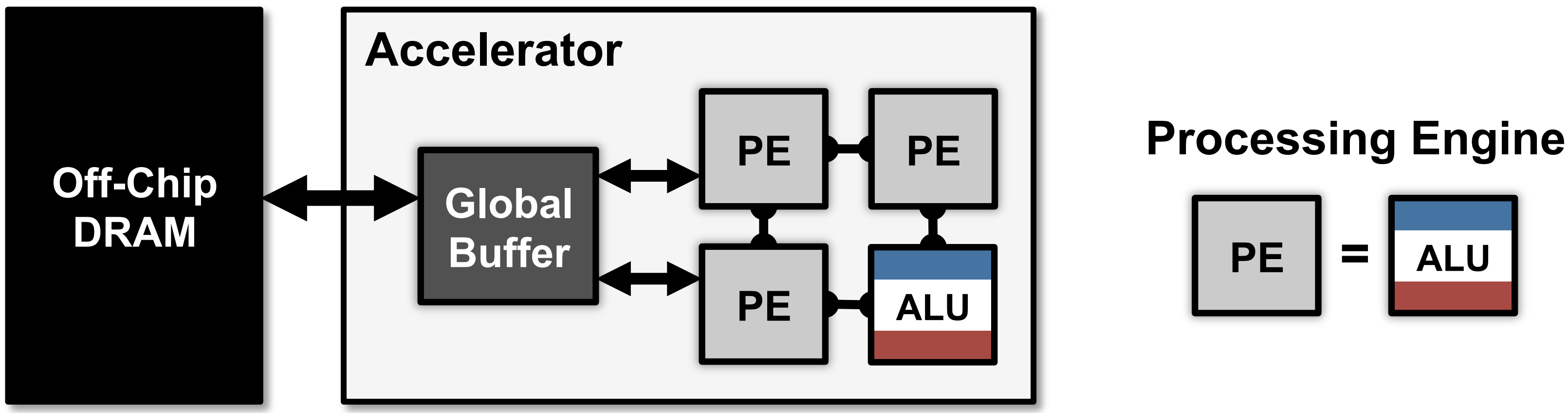
Goal: Increase reuse of input data (**weights** and **pixels**) and local **partial sums** accumulation

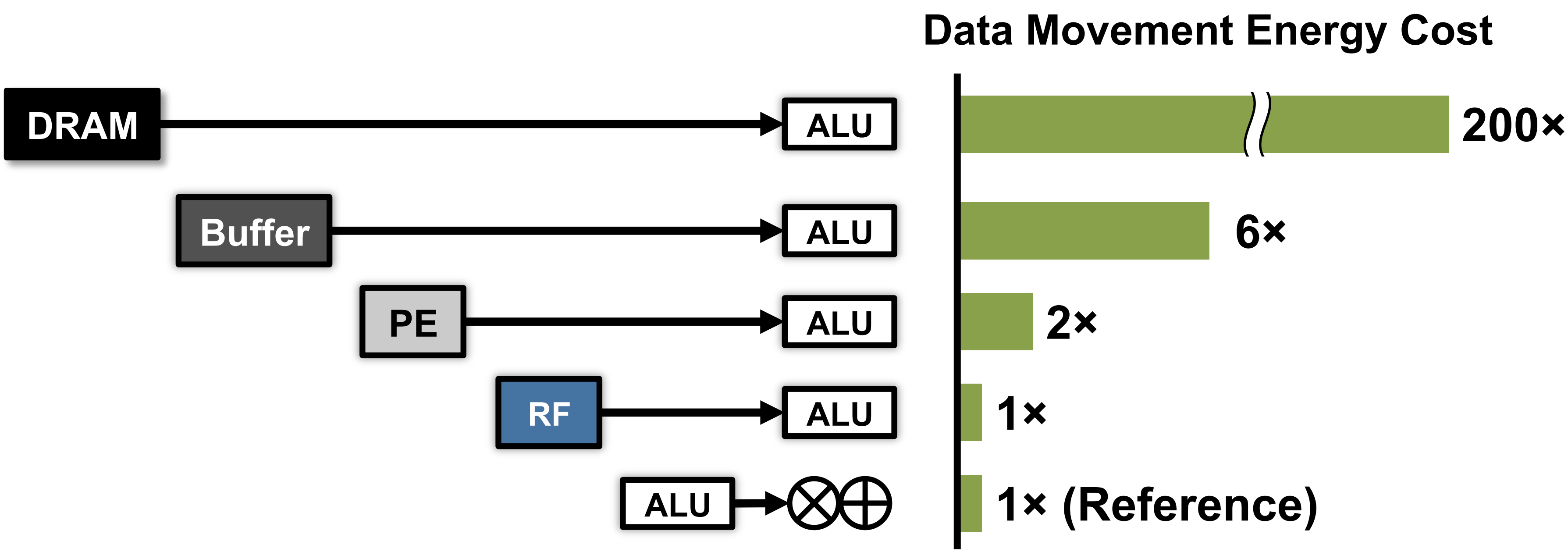
Energy-Efficient Dataflow

Yu-Hsin Chen, Joel Emer, Vivienne Sze, ISCA 2016

Maximize data reuse and accumulation at RF

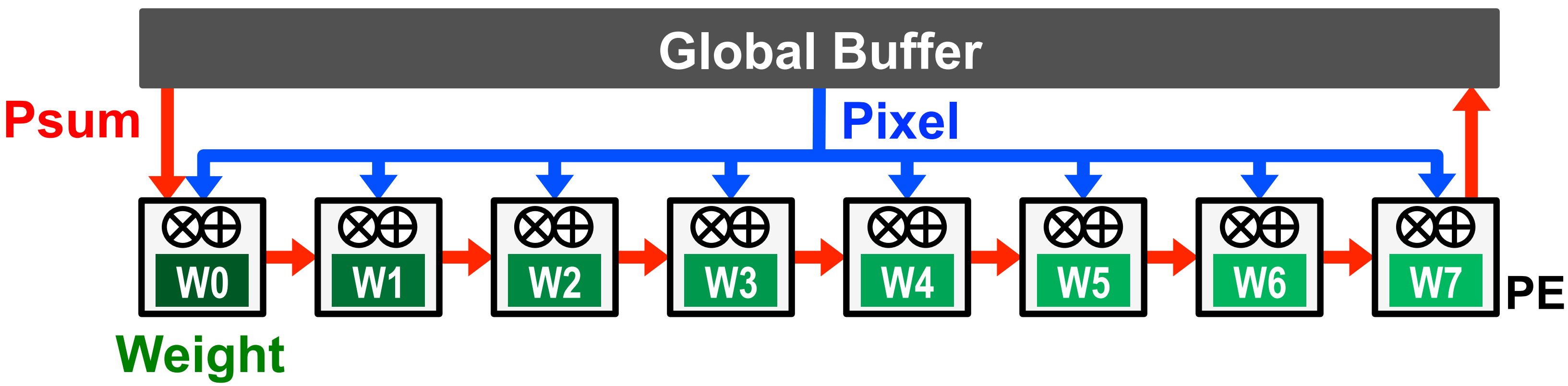
Data Movement is Expensive





Maximize data reuse at lower levels of hierarchy

Weight Stationary (WS)

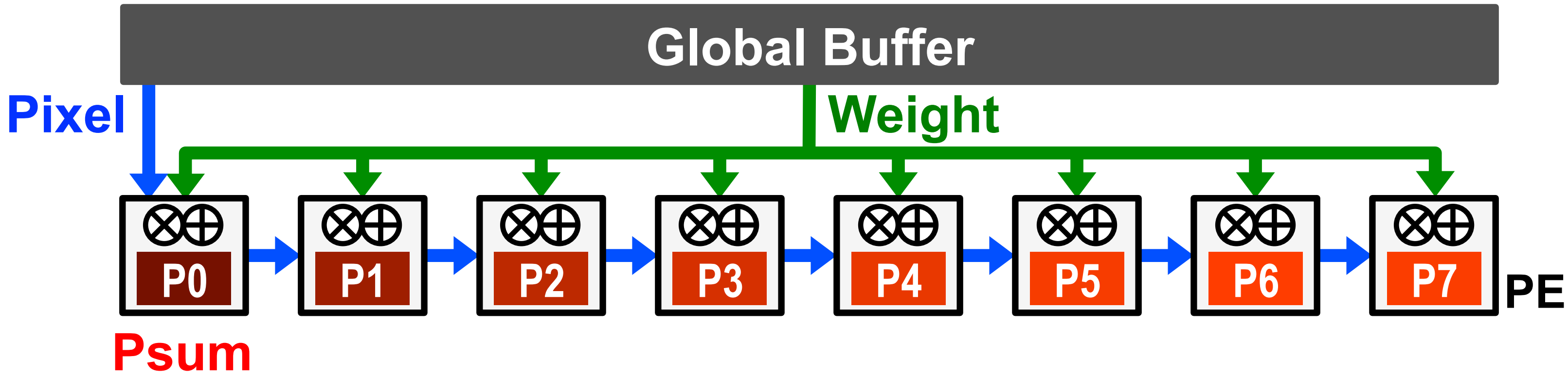


- **Minimize weight** read energy consumption
  - maximize convolutional and filter reuse of weights
- **Examples:**

  [**Chakradhar**, *ISCA* 2010] [**nn-X (NeuFlow)**, *CVPRW* 2014]

  [**Park**, *ISSCC* 2015] [**Origami**, *GLSVLSI* 2015]

Output Stationary (OS)



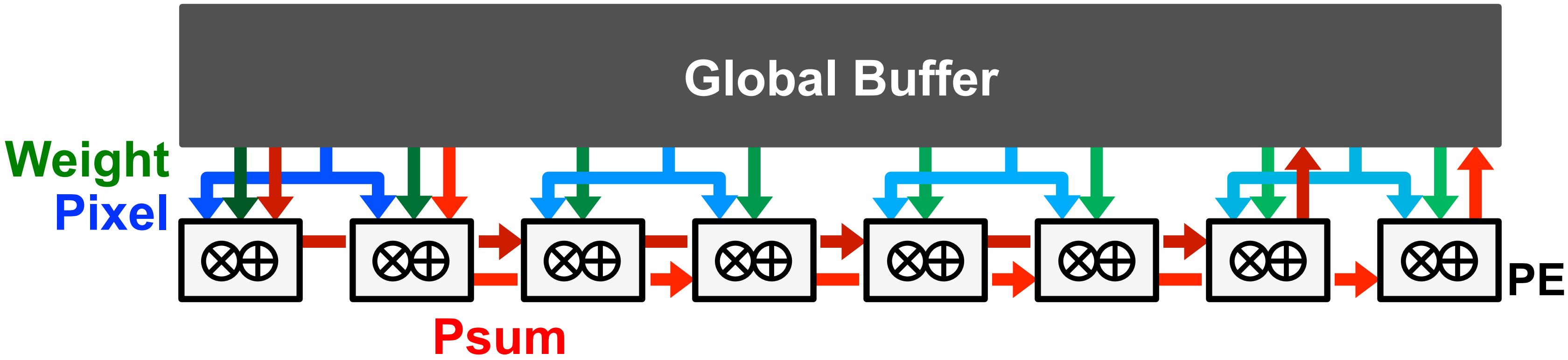
- **Minimize** **partial sum** R/W energy consumption
  - maximize local accumulation

- **Examples:**

  [**Gupta**, *ICML* 2015] [**ShiDianNao**, *ISCA* 2015]

  [**Peemen**, *ICCD* 2013]

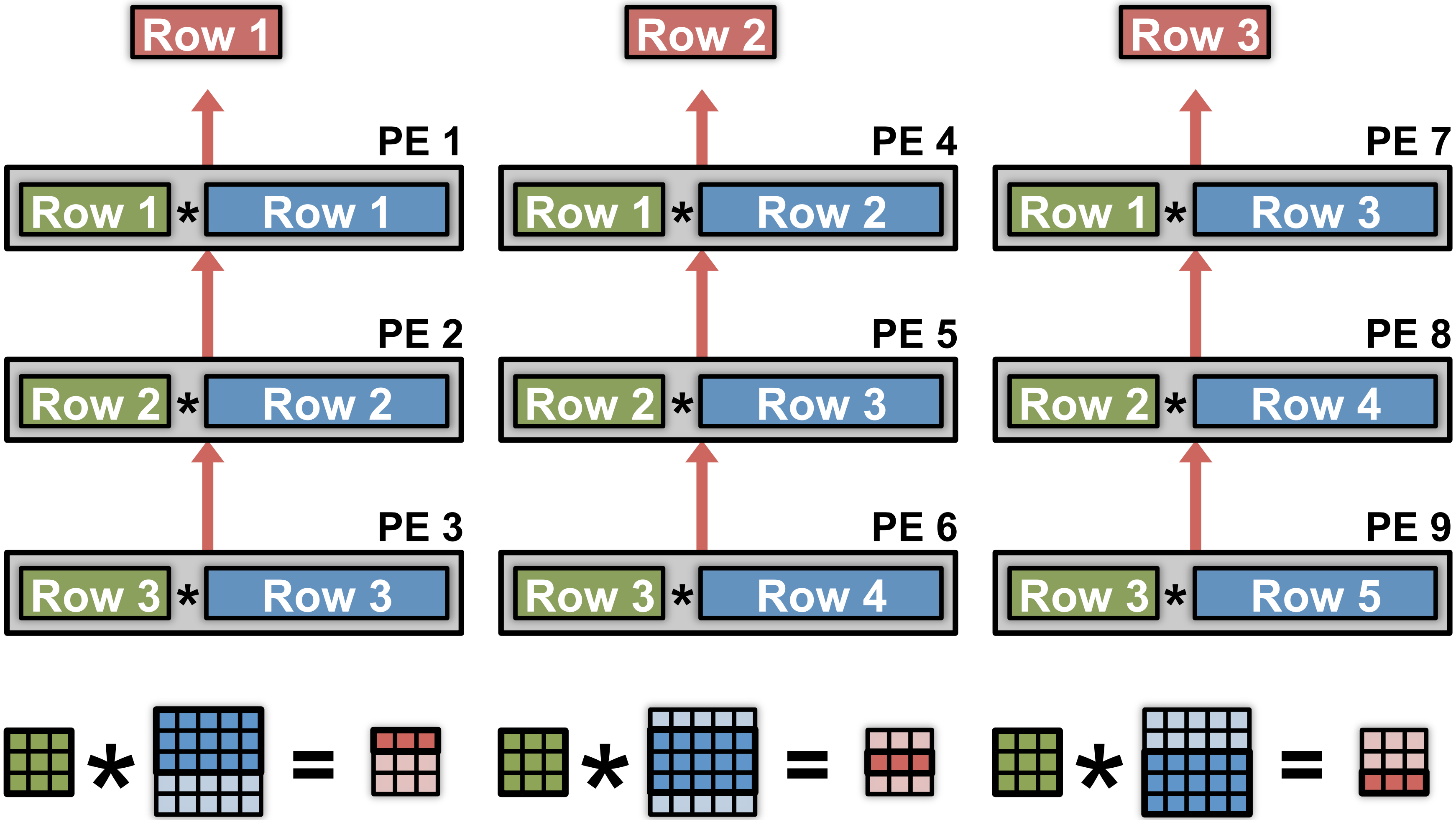
No Local Reuse (NLR)



- Use a **large global buffer** as shared storage
  - Reduce **DRAM** access energy consumption
- **Examples:**

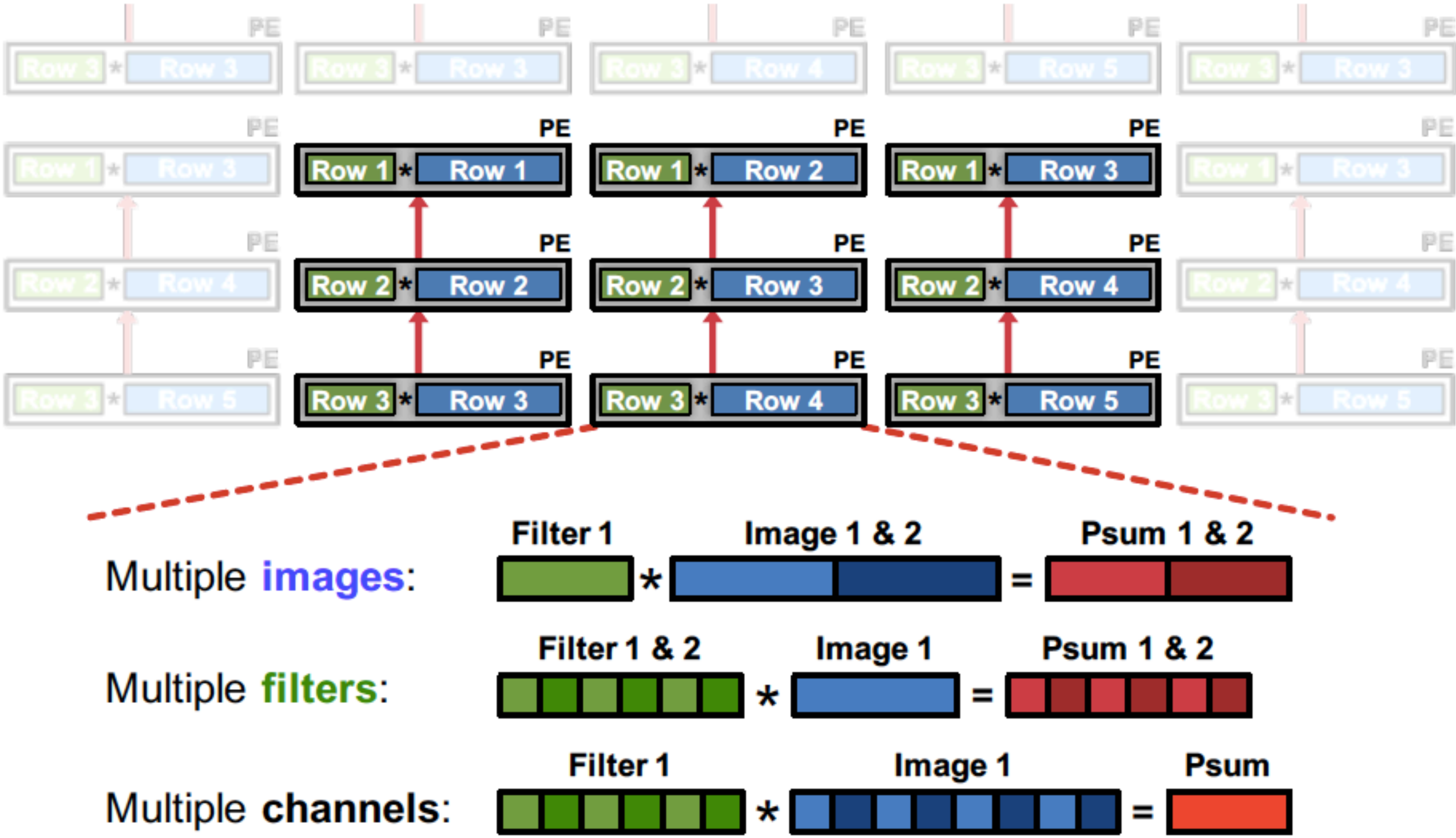
  [**DianNao**, *ASPLOS* 2014] [**DaDianNao**, *MICRO* 2014]
  [**Zhang**, *FPGA* 2015]

Row Stationary Dataflow



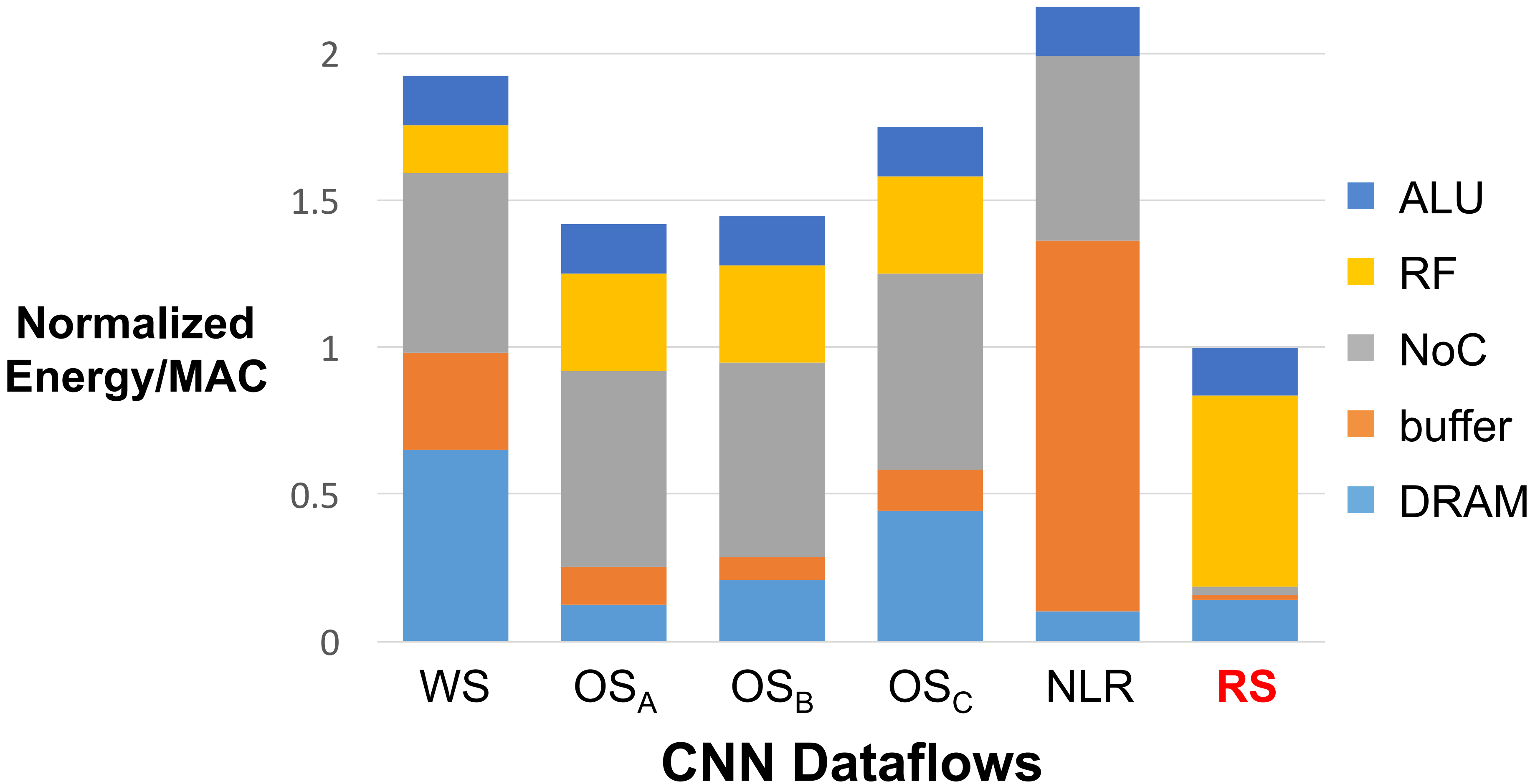
Optimize for **overall energy efficiency** instead for only a certain data type

CNN Convolution – The Full Picture



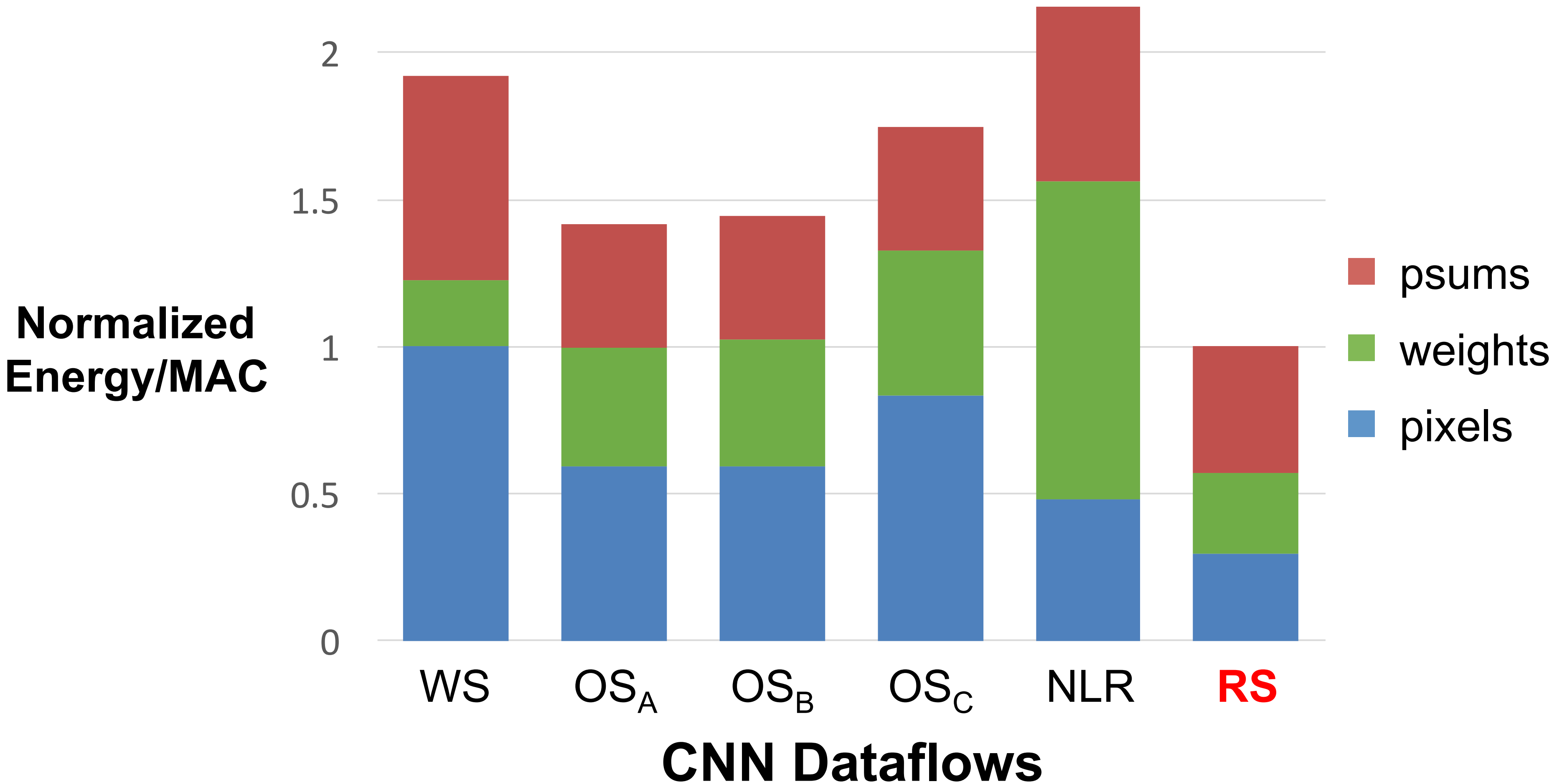
Map rows from **multiple images**, **filters** and **channels** to same PE to exploit other forms of reuse and local accumulation

Dataflow Comparison: CONV Layers



RS uses **1.4× – 2.5× lower** energy than other dataflows

Dataflow Comparison: CONV Layers



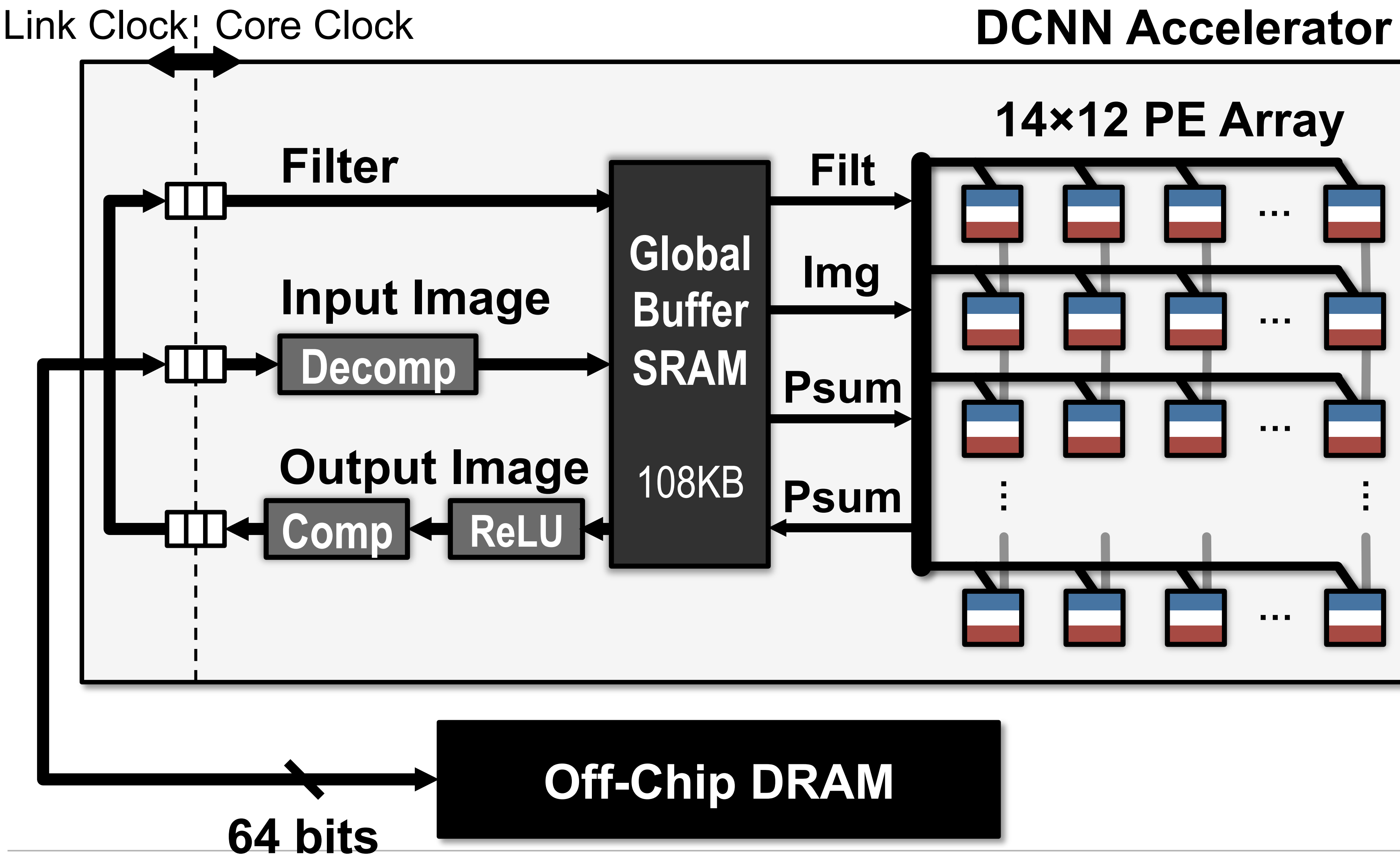
RS optimizes for the best **overall** energy efficiency

Energy-Efficient Accelerator

Yu-Hsin Chen, Tushar Krishna, Joel Emer, Vivienne Sze, ISSCC 2016

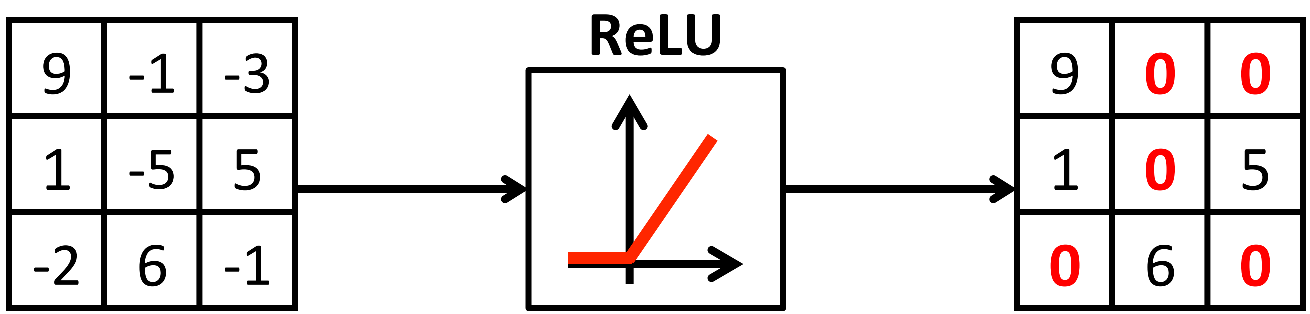
Exploit data statistics

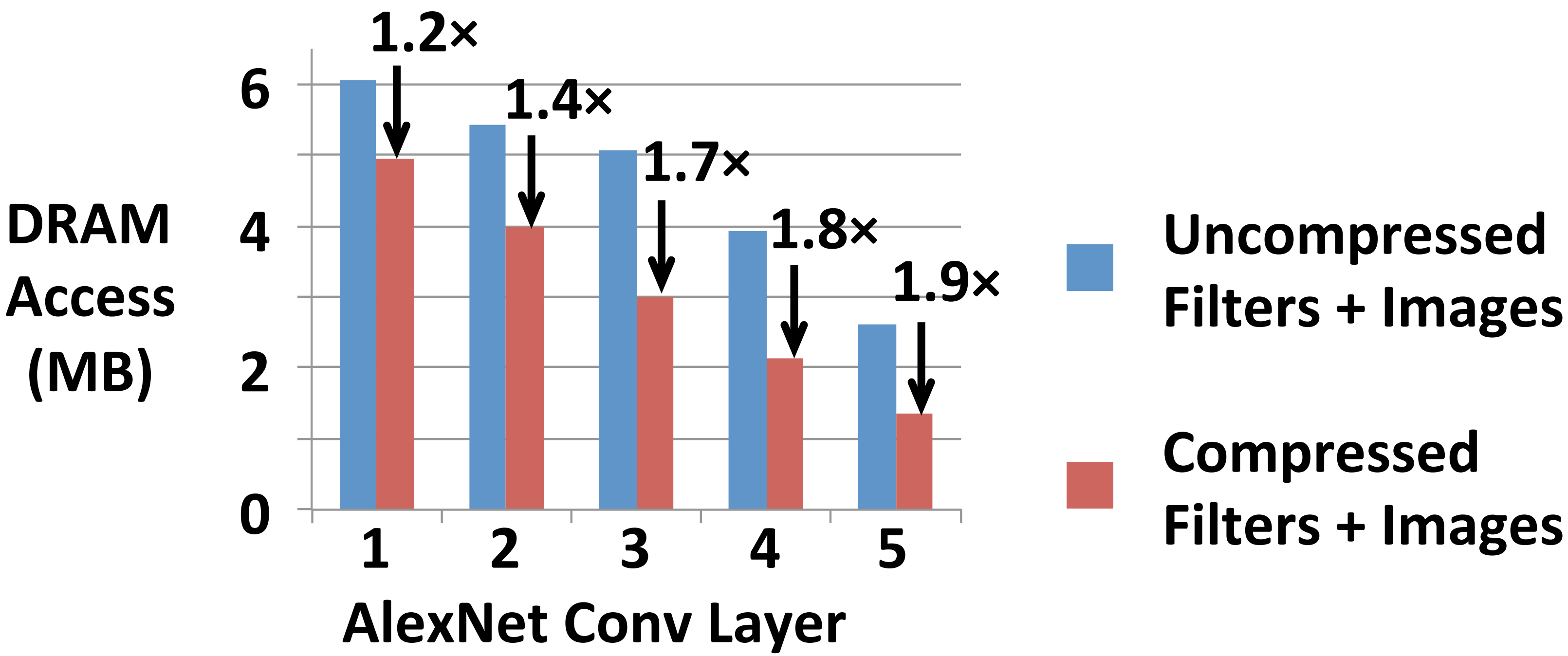
Eyeriss Deep CNN Accelerator



Data Compression Saves DRAM BW

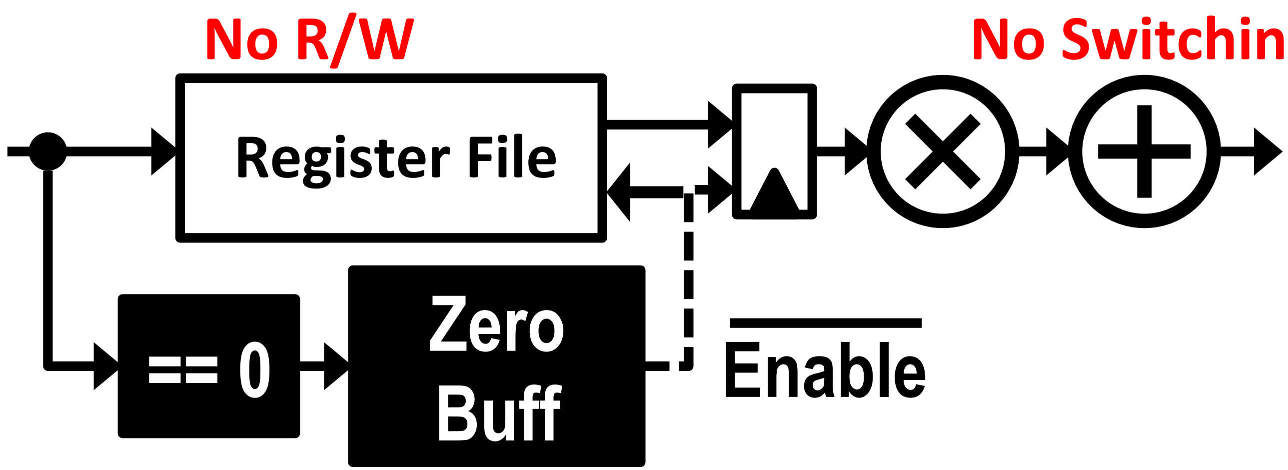
Apply Non-Linearity (**ReLU**) on Filtered Image Data





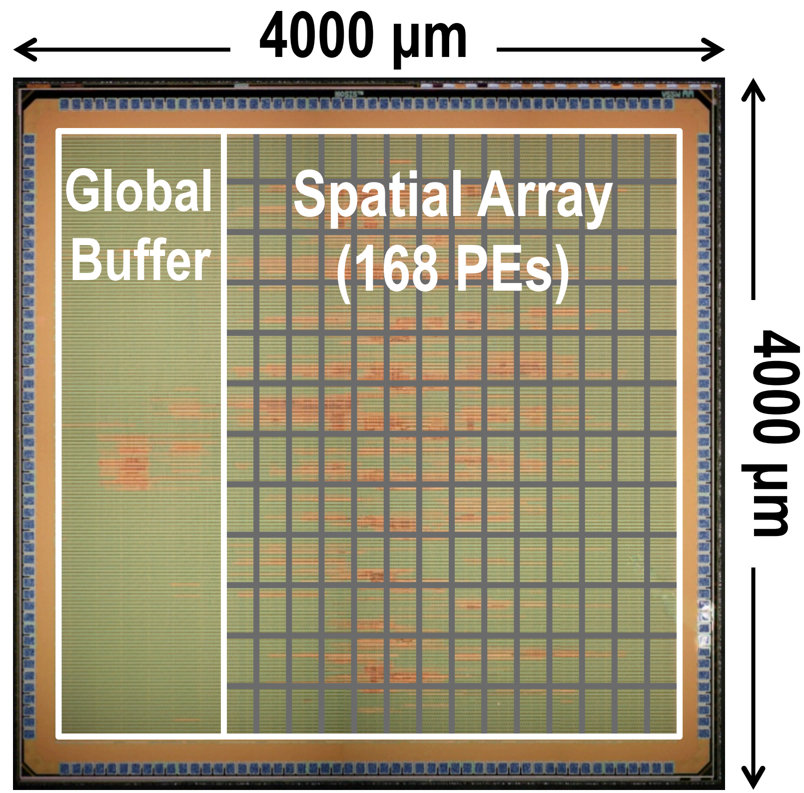
Zero Data Processing Gating

- Skip PE local **memory access**
- Skip MAC **computation**
- Save PE processing power by 45%



Eyeriss Chip Spec & Measurement Results

Technology	TSMC 65nm LP 1P9M
On-Chip Buffer	108 KB
# of PEs	168
Scratch Pad / PE	0.5 KB
Core Frequency	100 – 250 MHz
Peak Performance	33.6 – 84.0 GOPS
Word Bit-width	16-bit Fixed-Point
Natively Supported CNN Shapes	Filter Width: 1 – 32 Filter Height: 1 – 12 Num. Filters: 1 – 1024 Num. Channels: 1 – 1024 Horz. Stride: 1–12 Vert. Stride: 1, 2, 4



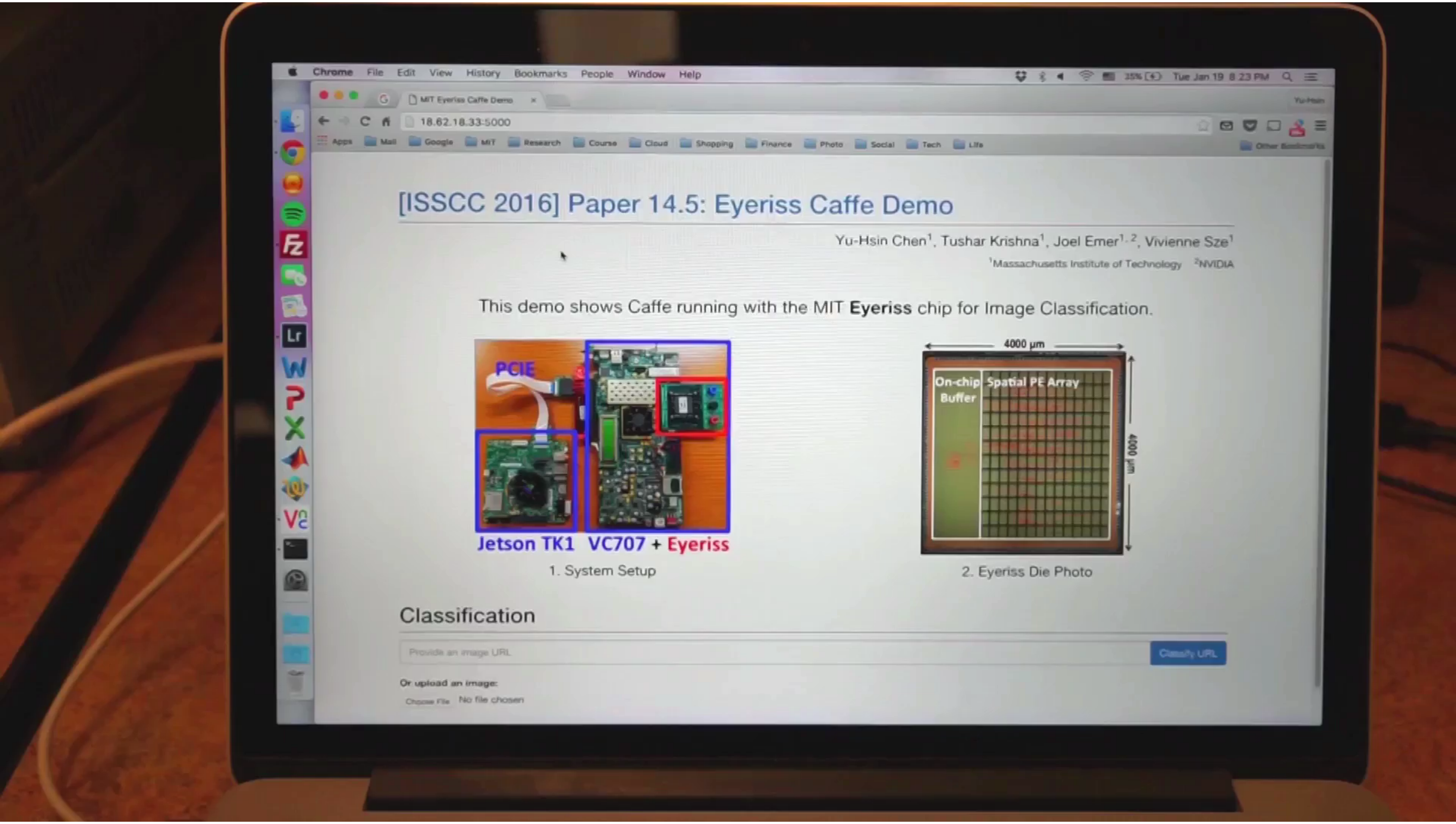
AlexNet: For 2.66 GMACs [8 billion 16-bit inputs (**16GB**) and 2.7 billion outputs (**5.4GB**)], only requires **208.5MB** (buffer) and **15.4MB** (DRAM)

Comparison with GPU

	This Work	NVIDIA TK1 (Jetson Kit)
Technology	65nm	28nm
Clock Rate	200MHz	852MHz
# Multipliers	168	192
On-Chip Storage	Buffer: 108KB Spad: 75.3KB	Shared Mem: 64KB Reg File: 256KB
Word Bit-Width	16b Fixed	32b Float
Throughput[1]	34.7 fps	68 fps
Measured Power	278 mW	Idle/Active[2]: 3.7W/10.2W
DRAM Bandwidth	127 MB/s	1120 MB/s [3]

1. AlexNet Convolutional Layers Only
2. Board Power
3. Modeled from [Tan, SC11]

Demo of Image Classification on Eyeriss



https://vimeo.com/154012013

Integrated with BVLC Caffe DL Framework

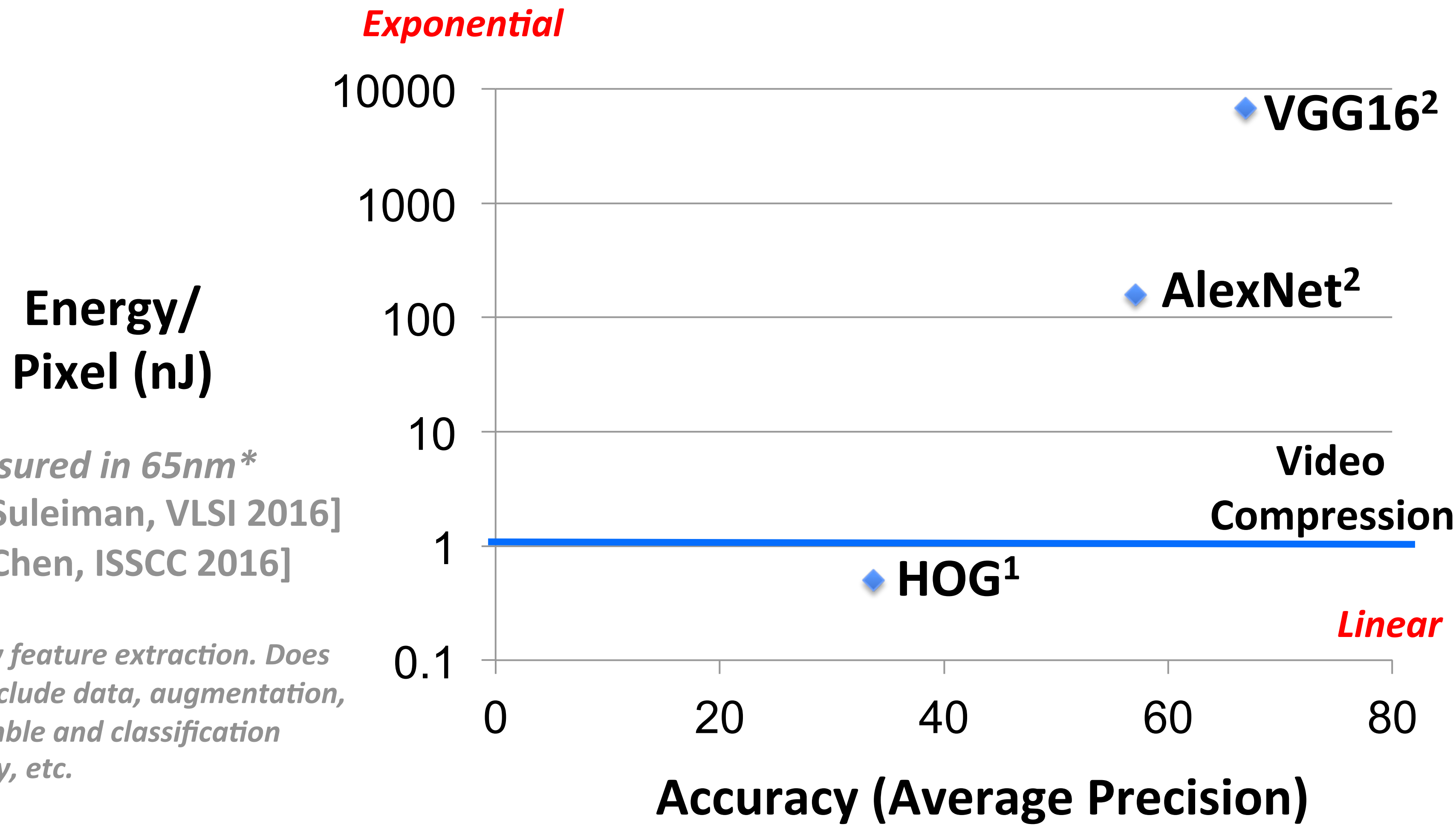
Summary of Eyeriss Deep CNN

- **Eyeriss:** a **reconfigurable** accelerator for state-of-the-art deep CNNs **at below 300mW**
- Energy-efficient **dataflow to reduce data movement**
- **Exploit data statistics** for high energy efficiency
- **Integrated** with the **Caffe DL framework** and demonstrated an image classification system

More info about **Eyeriss** and **Tutorial on DNN Architectures** at

http://eyeriss.mit.edu

Features: Energy vs. Accuracy



*Measured in 65nm**

1. [Suleiman, VLSI 2016]
2. [Chen, ISSCC 2016]

** Only feature extraction. Does not include data, augmentation, ensemble and classification energy, etc.*

Measured in on VOC 2007 Dataset

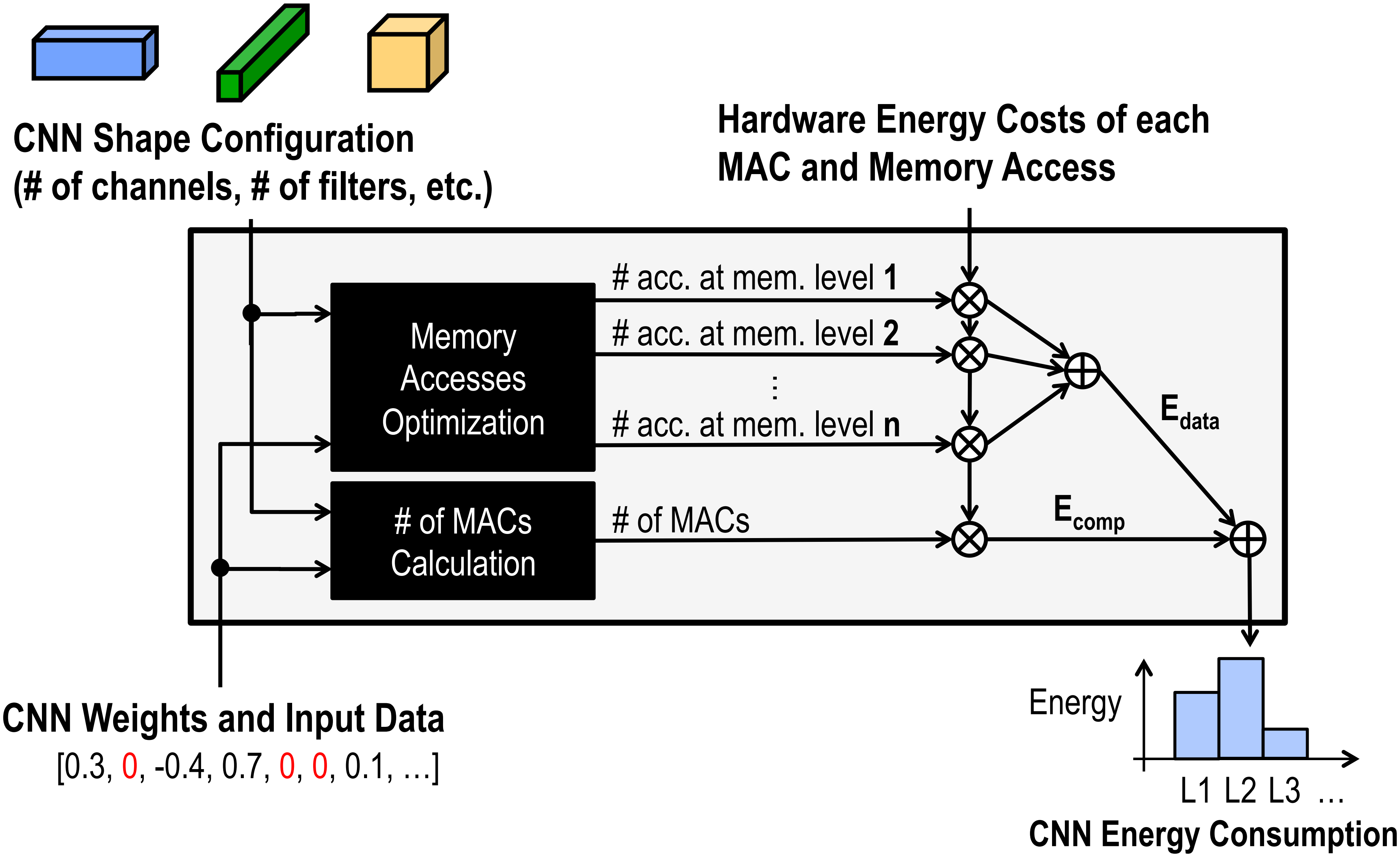
1. DPM v5 [Girshick, 2012]
2. Fast R-CNN [Girshick, CVPR 2015]

Designing Energy-Efficient CNNs using Energy-Aware Pruning

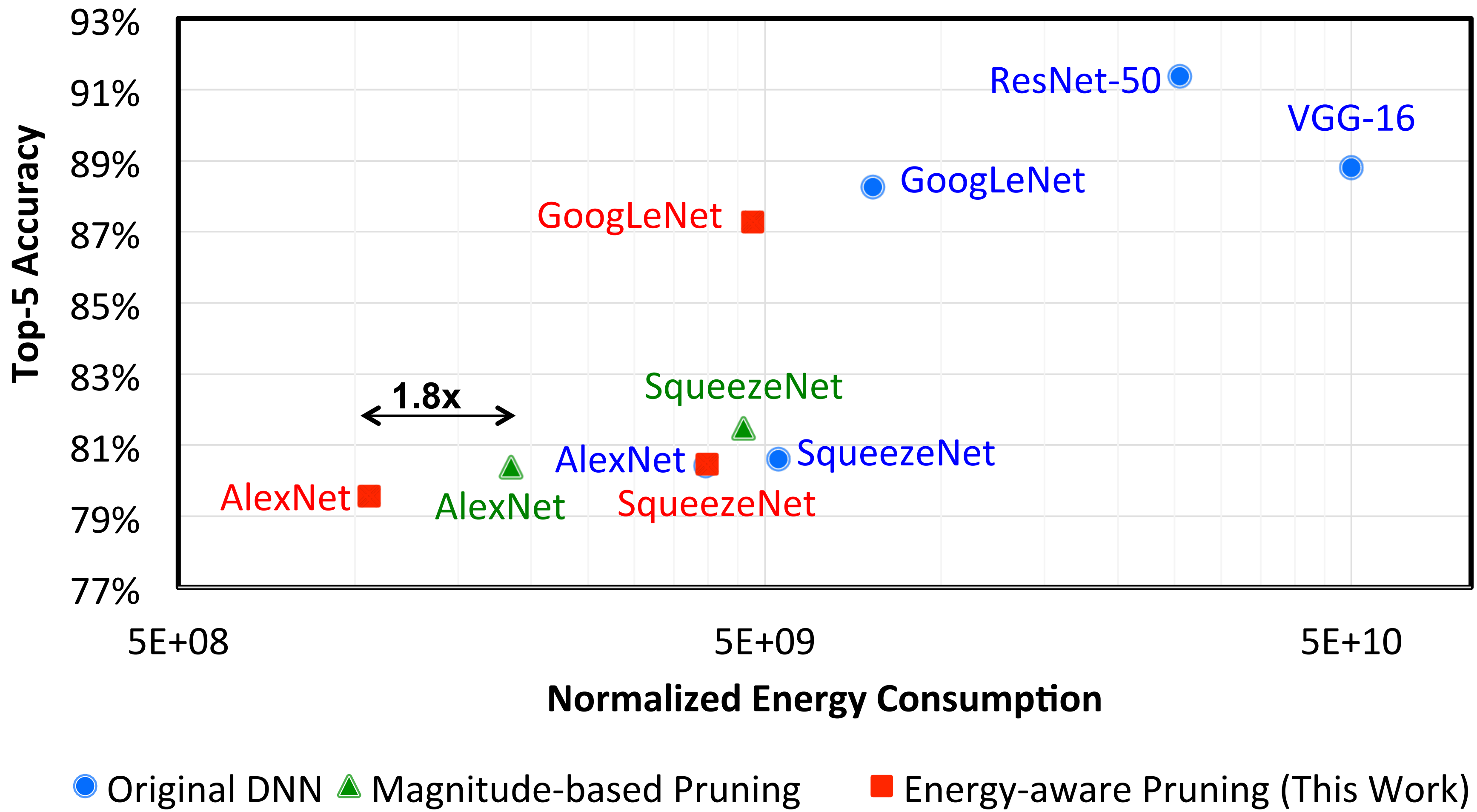
Tien-Ju Yang, Yu-Hsin Chen, Vivienne Sze, CVPR 2017



Energy-Evaluation Methodology



Energy-Aware Pruning



Remove weights from layers **in order of highest to lowest energy**
3.7x reduction in AlexNet / 1.6x reduction in GoogLeNet

Enable real-time navigation on nanoDrone

Image source: Cheerson

Big battery

Mobile GPU

Enable energy-efficient navigation for **Search and Rescue**

Acknowledgements

Research conducted in the **MIT Energy-Efficient Multimedia Systems Group** would not be possible without the support of the following organizations:

References

More info about **Eyeriss** and
Tutorial on DNN Architectures at
http://eyeriss.mit.edu

More info about research in the **Energy-Efficient Multimedia Systems Group @ MIT**
http://www.rle.mit.edu/eems

For updates Follow @eems_mit

http://mailman.mit.edu/mailman/listinfo/eems-news