MACHINE LEARNING: ALGORITHMS & HARDWARE

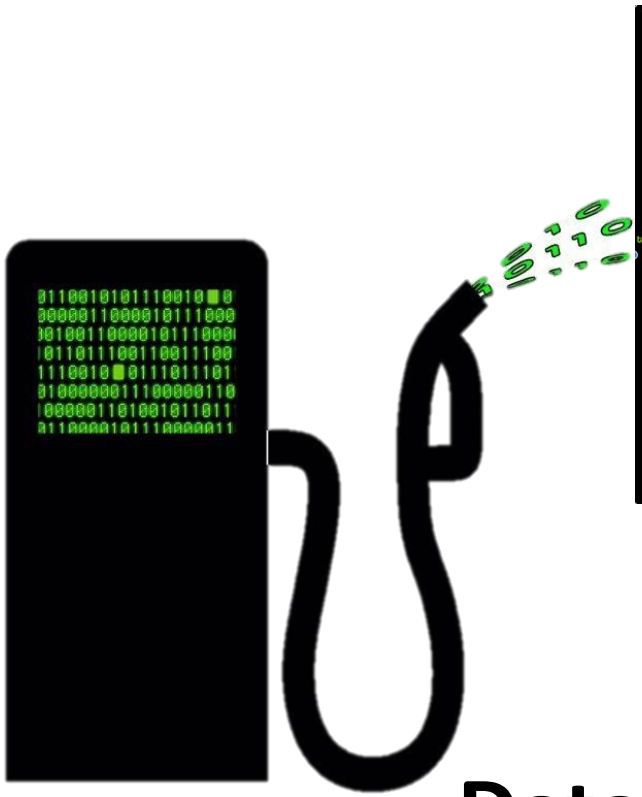
KAUSHIK ROY
CENTER FOR BRAIN-INSPIRED COMPUTING (C-BRIC)
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA

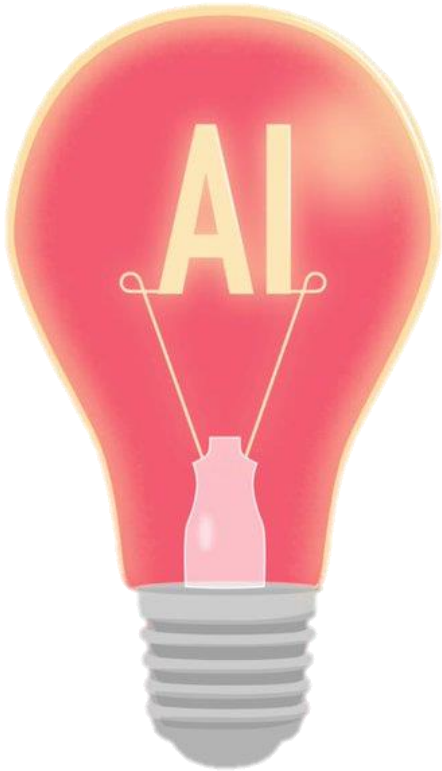
Adopted from lectures by Andrew Ng

We live in a data-driven world!!!

C-BRIC

Artificial Intelligence (AI): Why now?



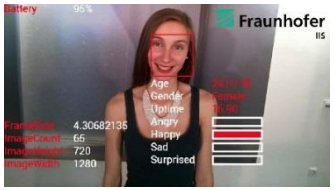


Data is the new oil and AI is the new electricity!!!!

C-BRIC

Artificial Intelligence (AI): Why now?

Recognition
Analytics
Inference



Digital data explosion has created a great demand for embedded platforms: mobile, wearables, Internet of Things (IoT) devices etc. to perform cognition

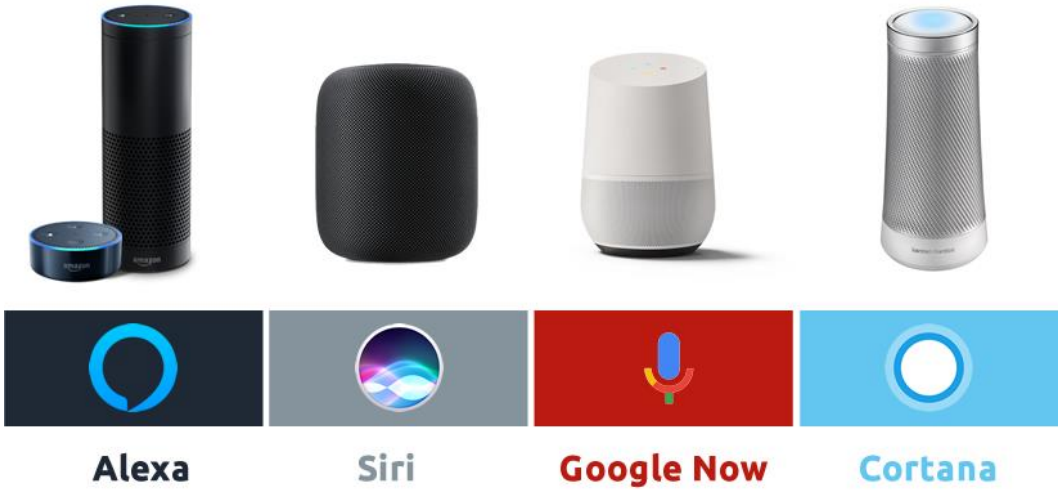
Cognition –

The device/computing system has to understand the data, make sense of it, and then, reason, decide or make a prediction based on what it sees.

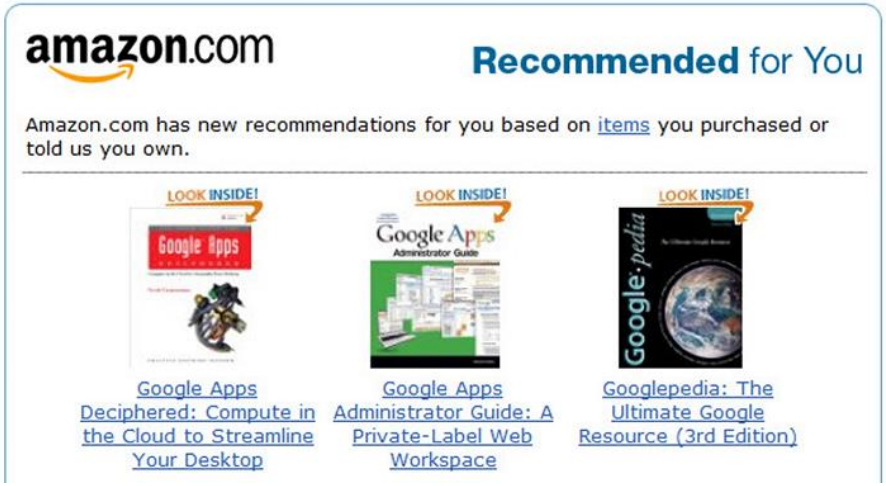
C-BRIC

AI enables Cognition

Facebook photo-tagging



Self-driving Car



C-BRIC

AI has even defeated humans!!

1997

IBM Deep Blue vs. Kasparov

2011

IBM Watson vs. Brad Ritter & Ken Jennings

2016

Google AlphaGo vs. Lee Sedol

https://www.youtube.com/watch?v=jGyCsVhtW0M

A great documentary on Alphago and AI in general!!

Key driver of AI: Neural Networks/Deep learning

C-BRIC

More recent success stories …

Identifying snow leopards with AI

Google AlphaGo vs. Lee Sedol

(1920 CPUs, 280 GPUs)

IBM's AI loses to Harish Natarajan, but still is persuasive…

Artificial intelligence used to recognize primate faces in the wild

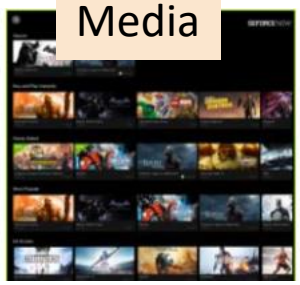
AI is scaling…

- Across applications
- Investments
- and people…

Image analytics

Medicine

Media



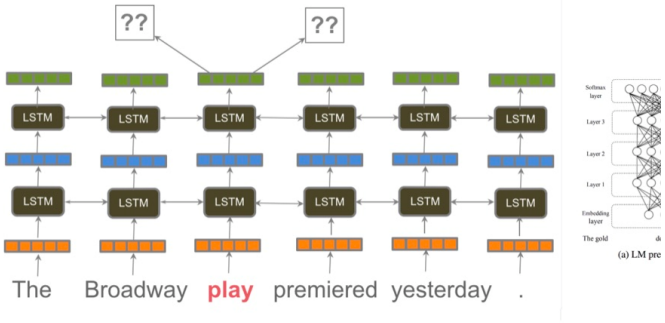
Security/Defense

Transportation

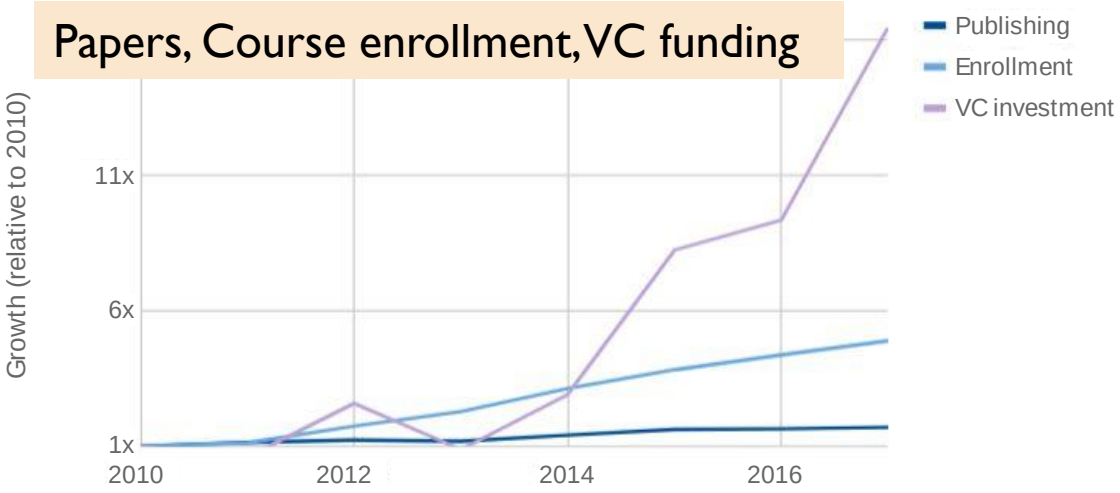
Customer Care

NLP's ImageNet moment has arrived

08.JUL.2018



Papers, Course enrollment, VC funding



However… Many challenges yet to be addressed

AI Stats News: Chatbots Lead To 80% …nd

Forbes CommunityVoice Connecting expert communities to the Forbes audience. What is This?

2,400 views | Feb 22, 2019, 08:30am

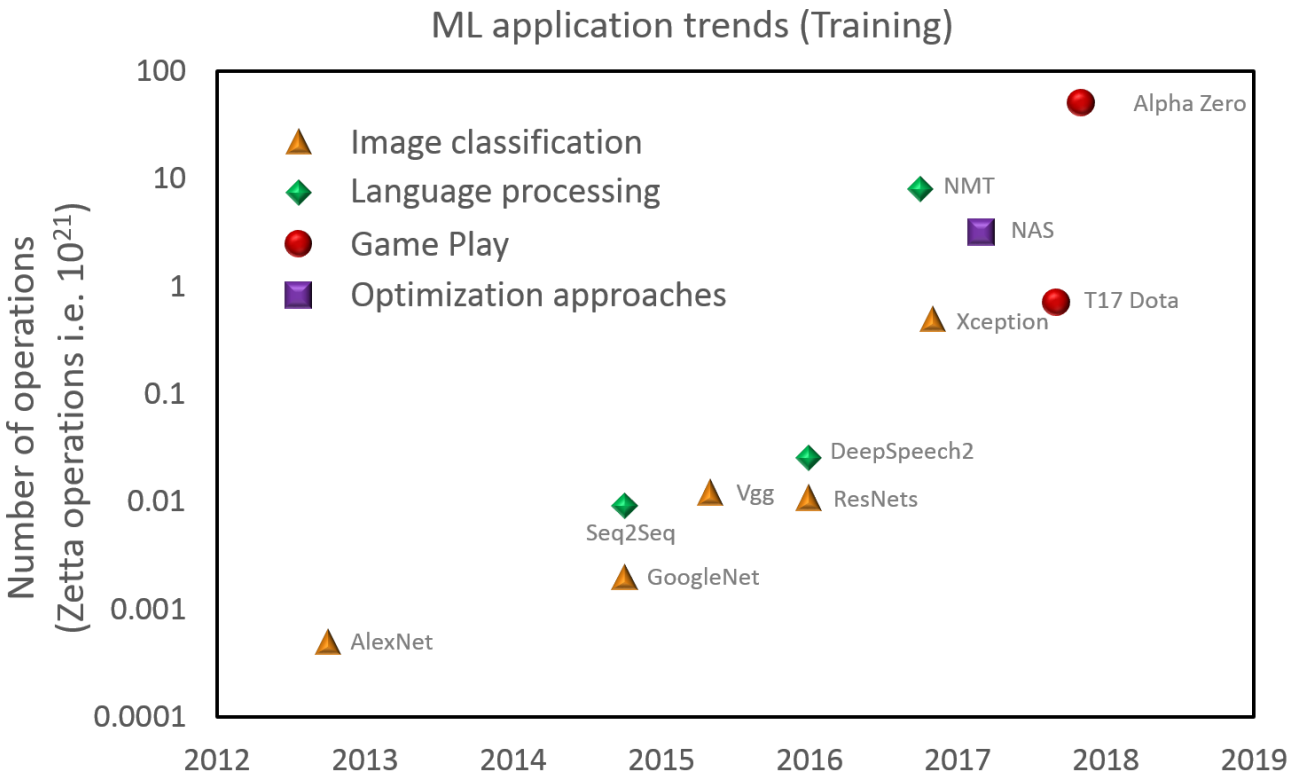
Exp…
Ope…

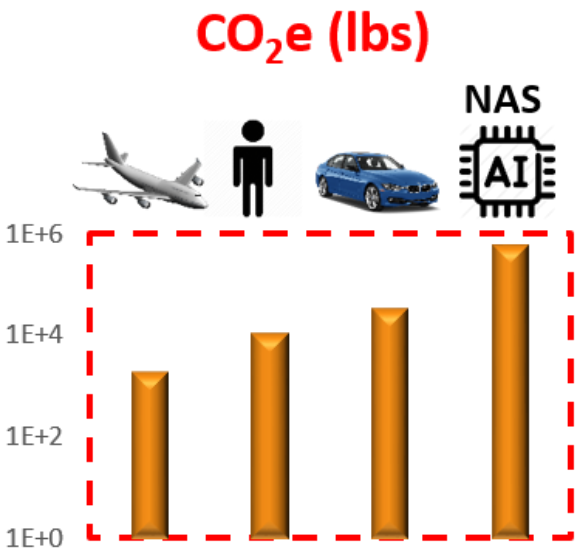
Forbes Technology Council

Consumer Concerns About Self-Driving Cars

% of respondents naming the following reasons for their reluctance to use self-driving cars

AI Compute Demands (Training)



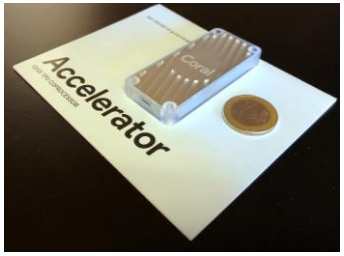


- Estimated CO_2 emissions from **NAS on Transformer (big)** is:
  - **315x** higher than **Air Travel from NY to SF/passenger**
  - **17x** higher than **average American (1 year)**
  - **5x** higher than **Car (1 lifetime)**

Efficiency Gap in AI

- Case study: Object recognition in a smart glass with a state-of-the-art accelerator

+



Google Edge TPU

Retinanet DNN* on a smart glass

Performance	
Frames/sec	13.3
Battery Life	
Energy/op	0.5 pJ/op
Energy/frame	0.15 J/frame
Time-to-die (2.1WH)	**64 mins**

*300 GOPs/inference

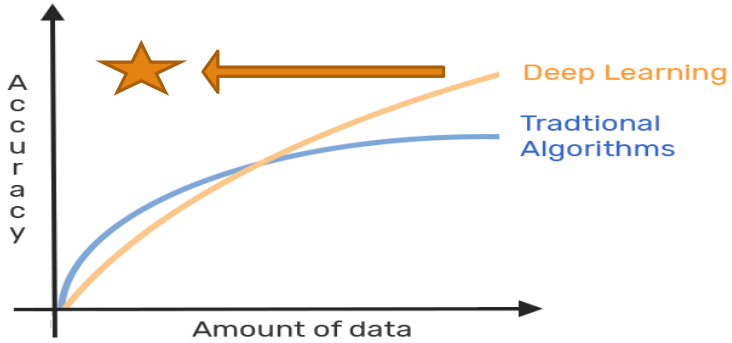
Where do the in-efficiencies come from?

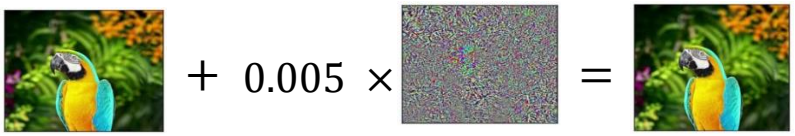
Algorithms Hardware Architecture Circuits and Devices

Ref: Venkataramani, S., Roy, K. and Raghunathan, A. "Efficient embedded learning for IoT devices." In *2016 21st Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 308-311. IEEE.

Beyond compute efficiency….

- Learning with less data
- Generalization & Robustness
- Lifelong learning

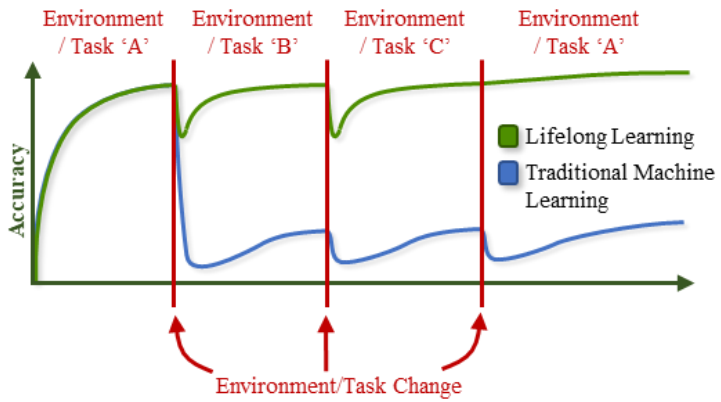


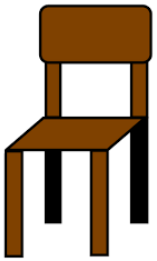


X_{clean}
97.3% confidence
Macaw

Δ
Adversarial Perturbation

$X_{adversary}$
88.9% confidence
Bookcase





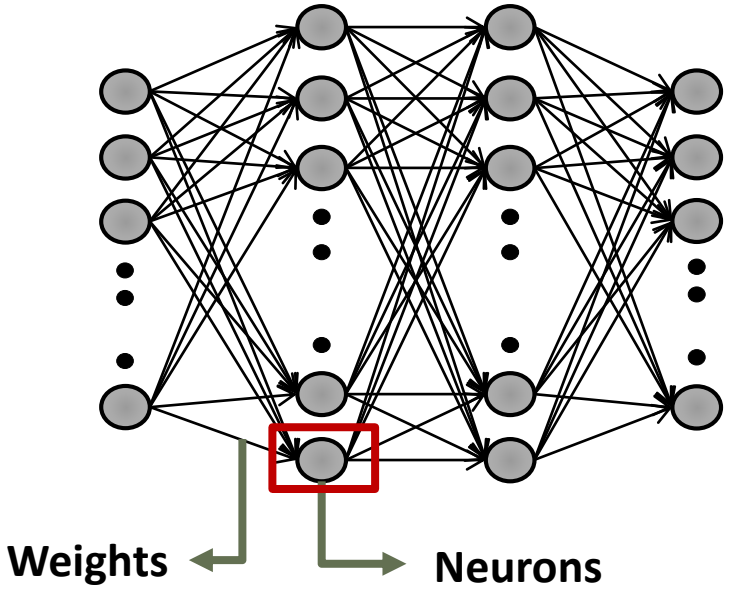
Neural Networks- Loosely brain-inspired*

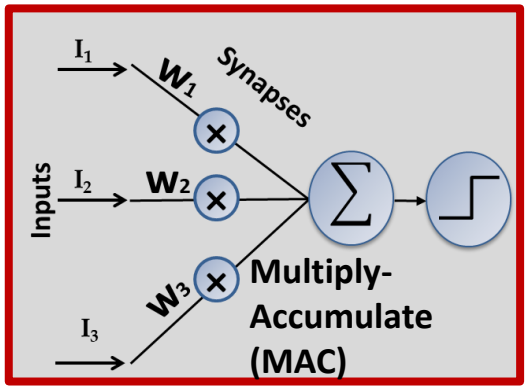
Biological Neural Network

Interconnected web of neurons/synapses

- **Neurons** are the computing elements
- **Synapses/Weights** store memory and take part in learning/training → Intelligence

Artificial Neural Network



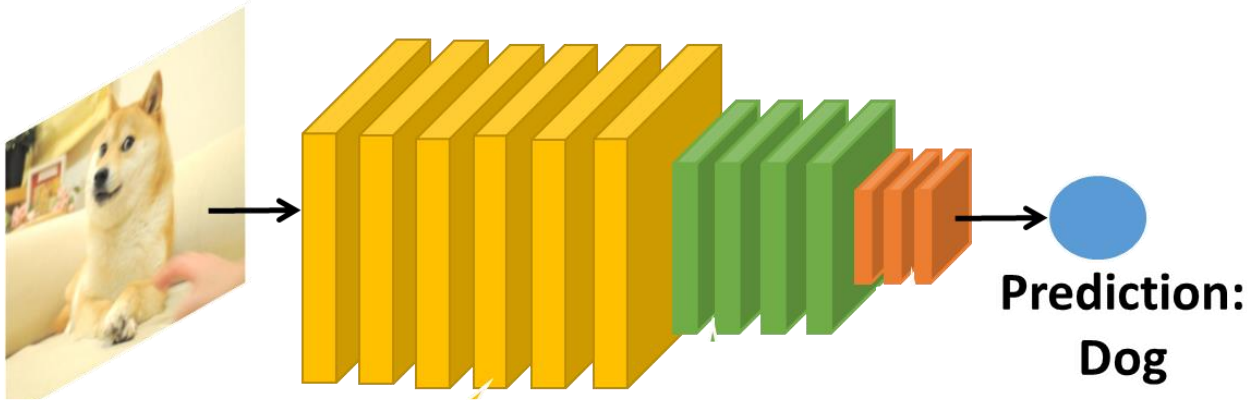


Key operation:

- **MAC/ Dot-product** $(\text{or } \Sigma_i w_i * I_i)$
- **Non-Linearity** (e.g. threshold, sigmoid, etc.)

C-BRIC

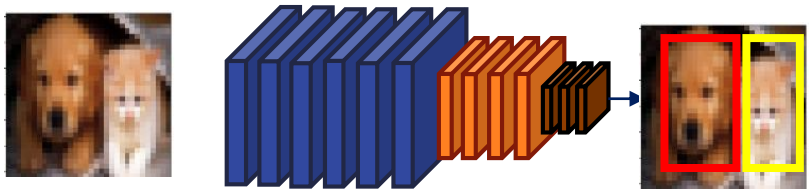
Training a Neural Network → Intelligence



Network discovers features from these pixel values,
Pretty incredible!!!

Neural Networks: Different Levels of Bio-fidelity

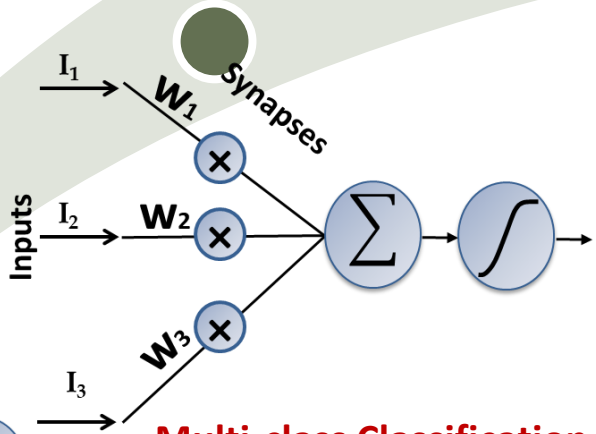
Deep Learning Revolution



Convolutional Neural Networks for Complex Recognition

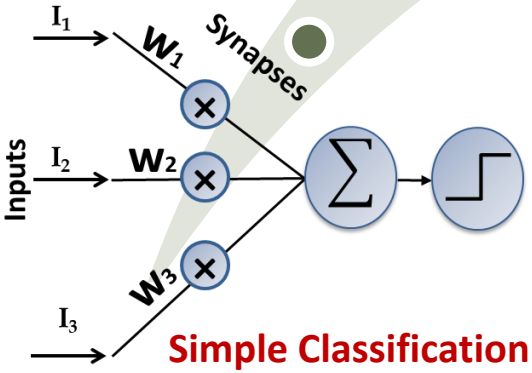
Spiking Neural Networks

Artificial Neural Networks

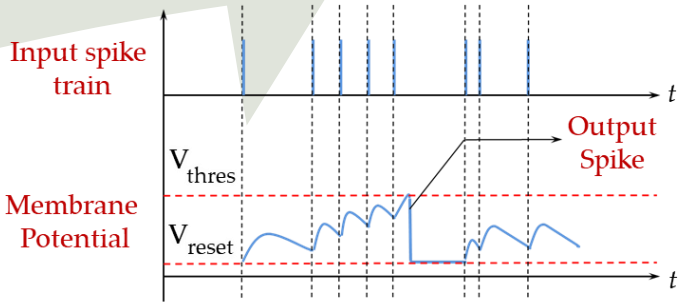


Multi-class Classification- Sigmoid, Tanh, ReLU

Perceptrons



Simple Classification – Threshold Non-linearity

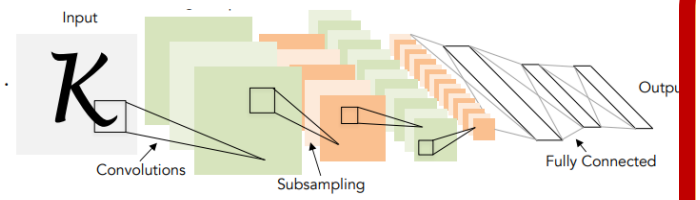


- **Neural computation through spikes**
- **Energy savings due to computation with sparse spiking events**

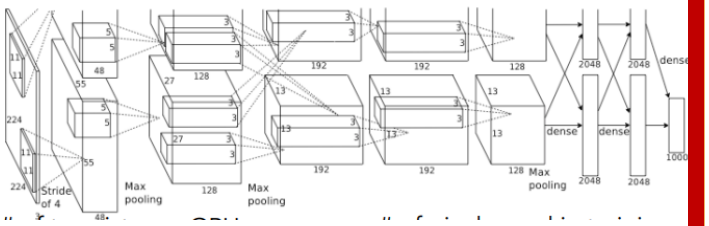
C-BRIC

Key Enablers of Deep Learning

LeCun et al. 1990



Krizhevsky et al. 2012



ALGORITHMS- Improved with better training, regularization, optimization strategies

# of transistors

10^6

~ 1Million Dot Product Operations

# of transistors GPUs

10^9

~ 1Billion Dot Product Operations

HARDWARE- Improved with architectural innovations, transistor scaling etc. for more compute power

# of pixels used in training

10^7 NIST

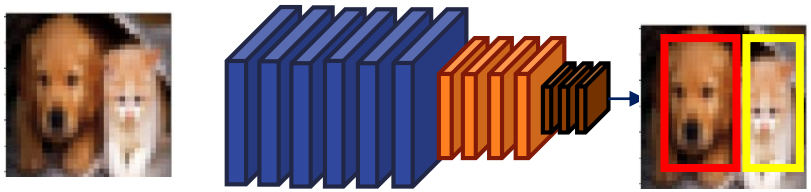
# of pixels used in training

10^{14} IMAGENET

DATA- Improved training/learning with more data

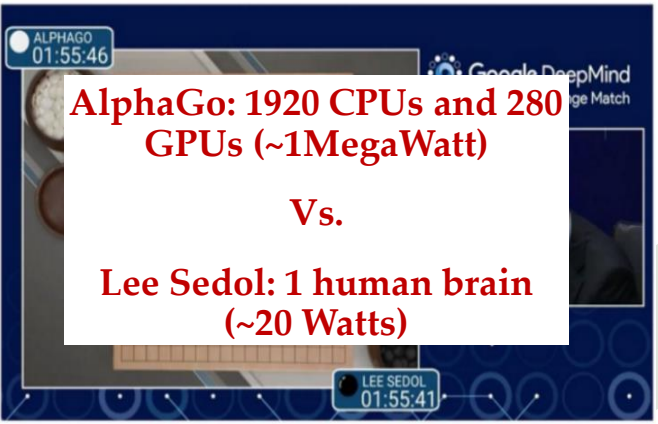
C-BRIC

Neural Networks: Different Levels of Bio-fidelity



Convolutional Neural Networks for Complex Recognition

Spiking Neural Networks

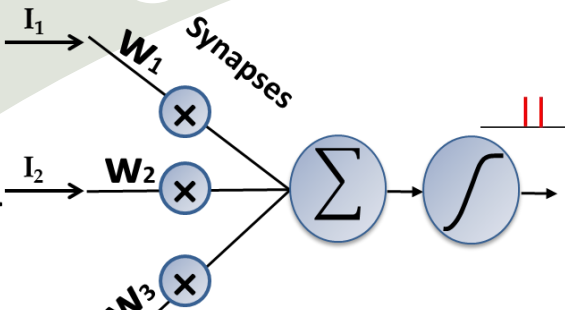


AlphaGo: 1920 CPUs and 280 GPUs (~1MegaWatt)

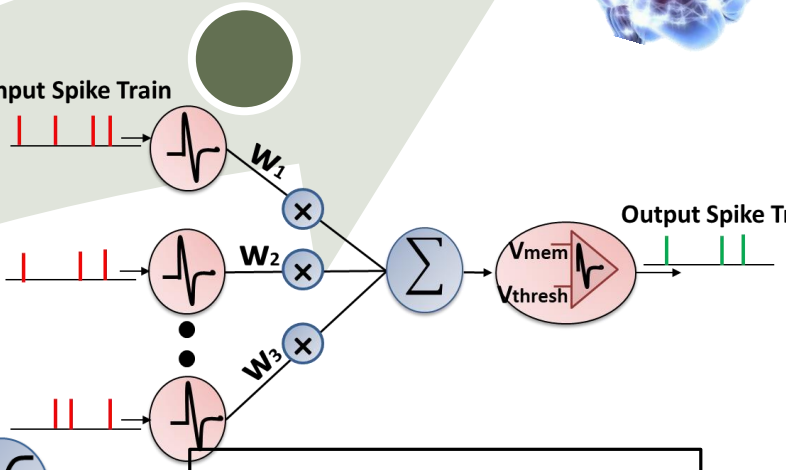
Vs.

Lee Sedol: 1 human brain (~20 Watts)

Artificial Neural Networks

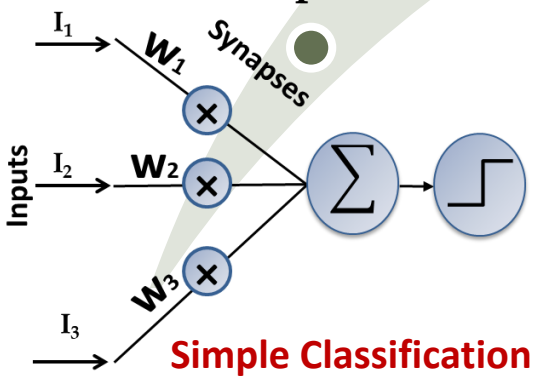


Multi-class Classification- Sigmoid, Tanh, ReLU



Neuromorphic Computing- Integrate-and-Fire Non-linearity

Perceptrons



Simple Classification – Threshold Non-linearity

C-BRIC

Why is Energy-Efficiency a Concern?

Case study: Object recognition in a smart glass

Overfeat DNN on a smart glass

Performance*	
OMAP 4430	1.5 fps (ideal)
Battery Life	
Energy/op (mobile GPU)	5 x 10^{-2} nJ/op
Energy/frame	0.16 J/frame
Time-to-die (2.1WH)	**25 min (ideal)**

*3.2 GigaOPS/inference
1 OPS ≈ 1 dot product operation

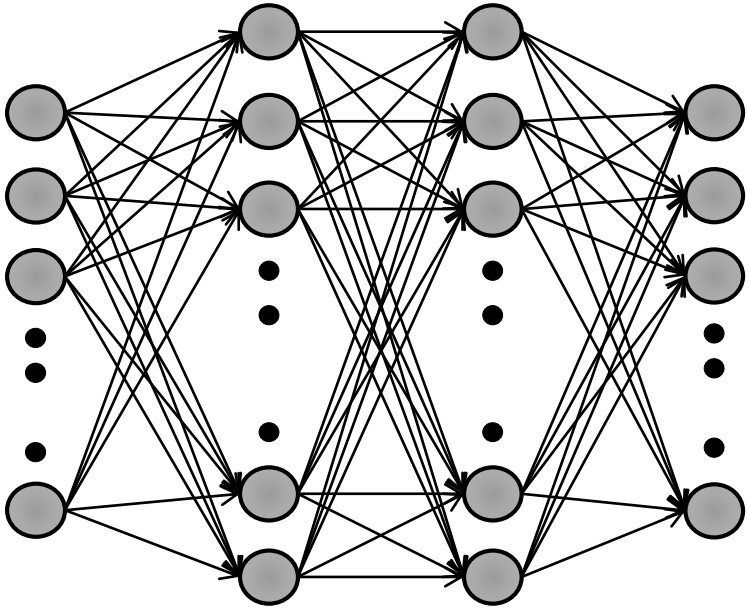
- **Battery powered devices (smart-phones, smart-watches, drones etc.) have resource or energy constraints.**
- **Enabling intelligence on these platforms necessitates energy-efficient deep learning or neural network implementations**

C-BRIC

Where do Inefficiencies Come From?

Biological Neural Network

Artificial Neural Network



Algorithms, Computing Architecture, Neurons, Synapses……..

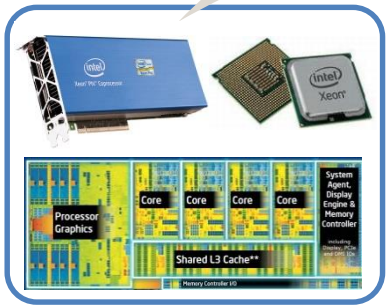
- Vastly interconnected web of neurons/synapses (1 billion neurons, 10000 synapses per neuron) → Compute and Memory are all intertwined and co-located
- Approximate and stochastic computation
- Sparse, irregular, event-driven, massively parallel networks

C-BRIC

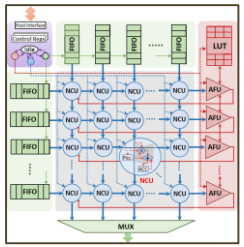
Hardware for Addressing Inefficiencies

- CMOS and Post-CMOS neuro-mimetic devices and interconnects
- Compute-near-memory / Compute-in-memory
- Approximate and stochastic neuronal and synaptic hardware
- Architectures that embody computing principles from the brain (sparse, irregular, event-driven, massively parallel)
- Programming and evaluation frameworks

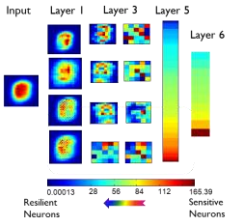
Multicores/GPUs



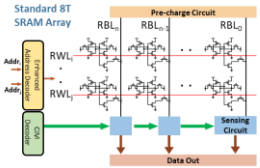
Accelerators



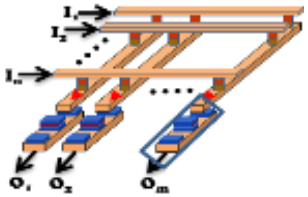
Approximate & Stochastic Hardware



In-memory computing

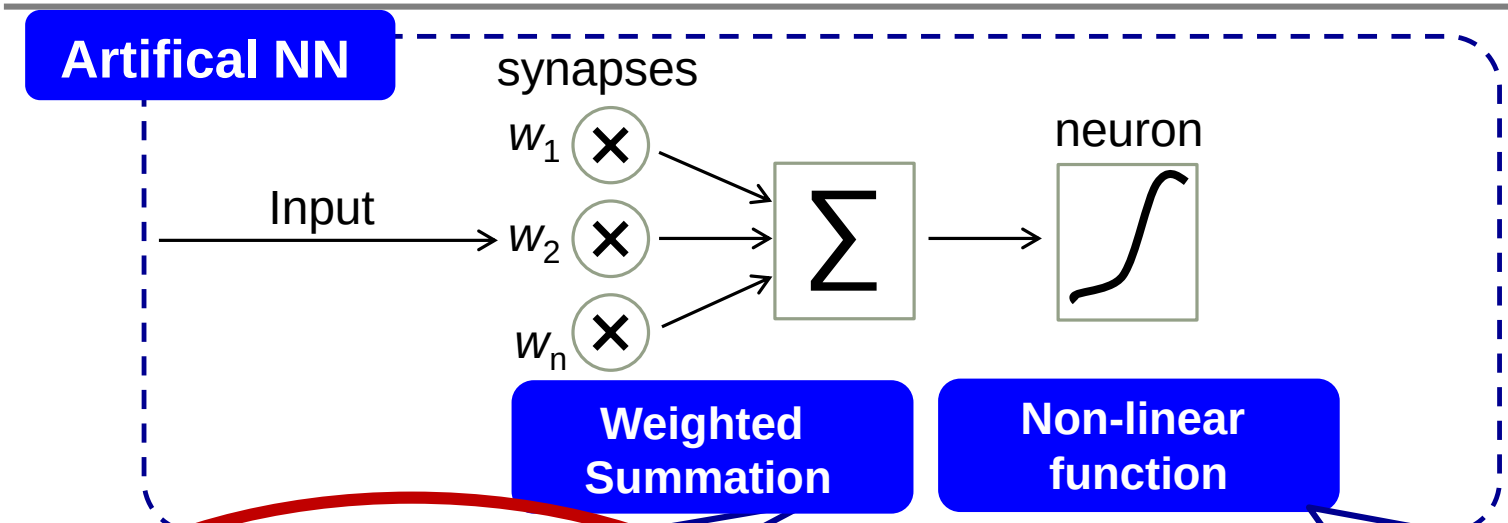


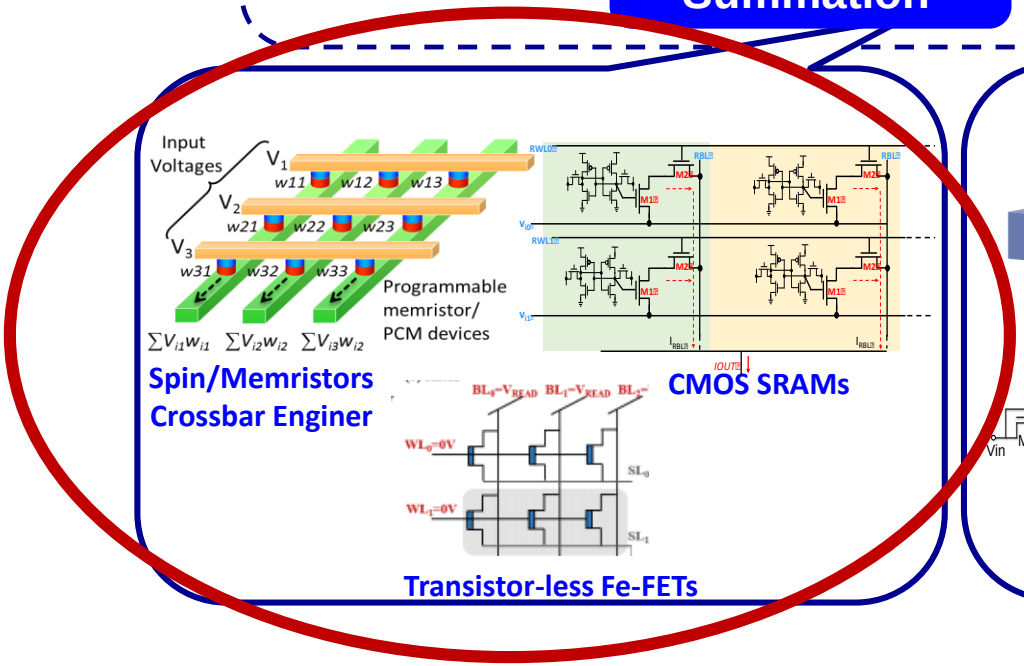
Post-CMOS Devices

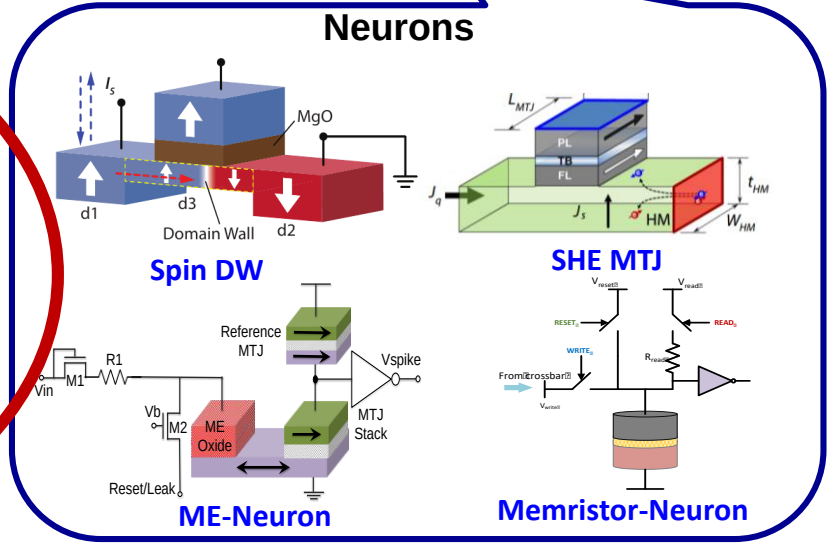


~10^4 Energy Gap

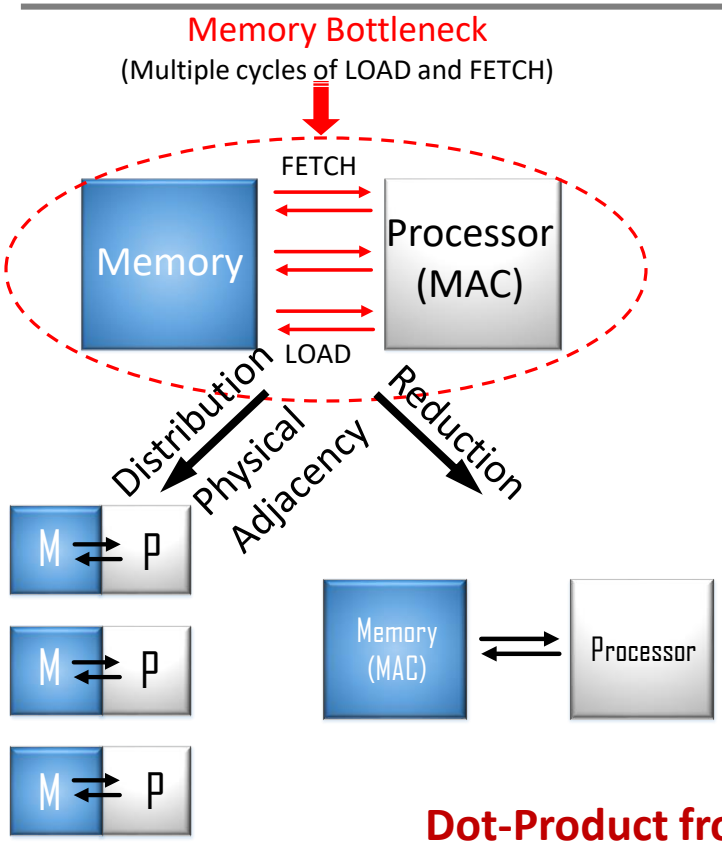
Neural Networks: Simple Hardware Model

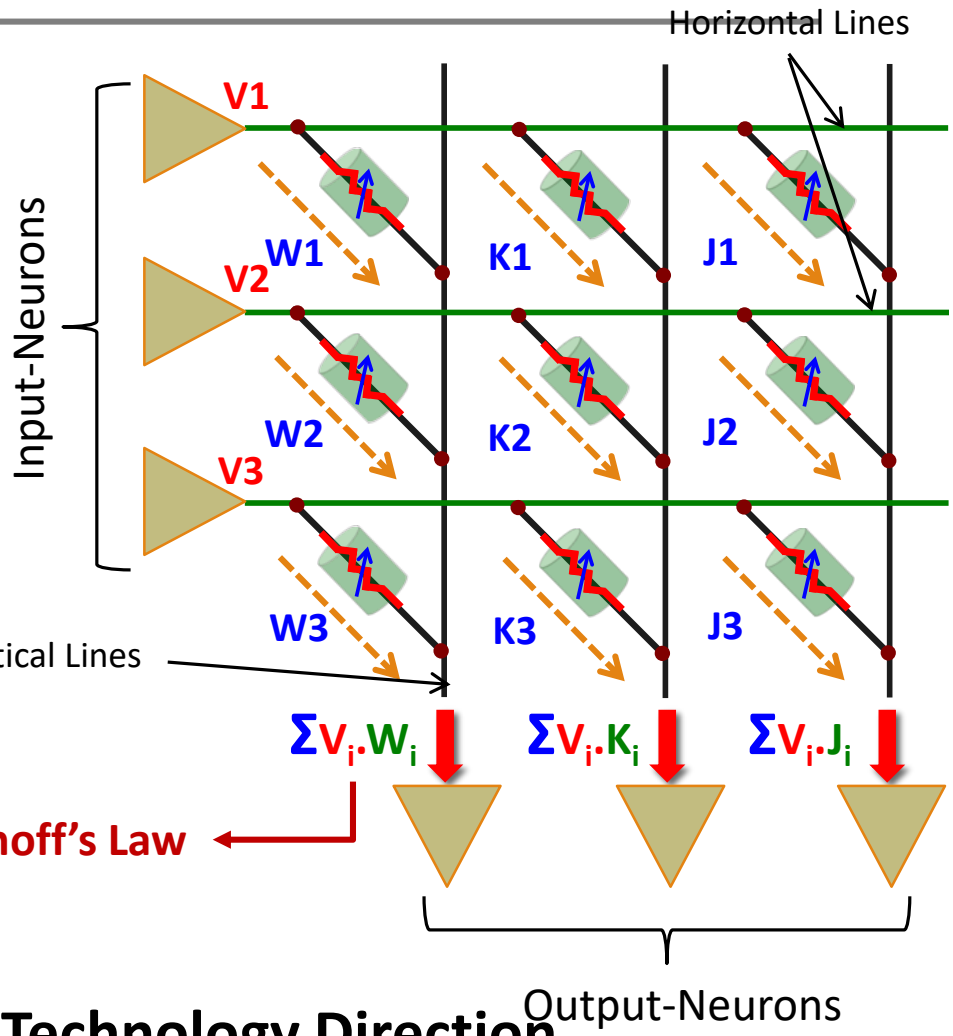






Memory Architecture/Circuits



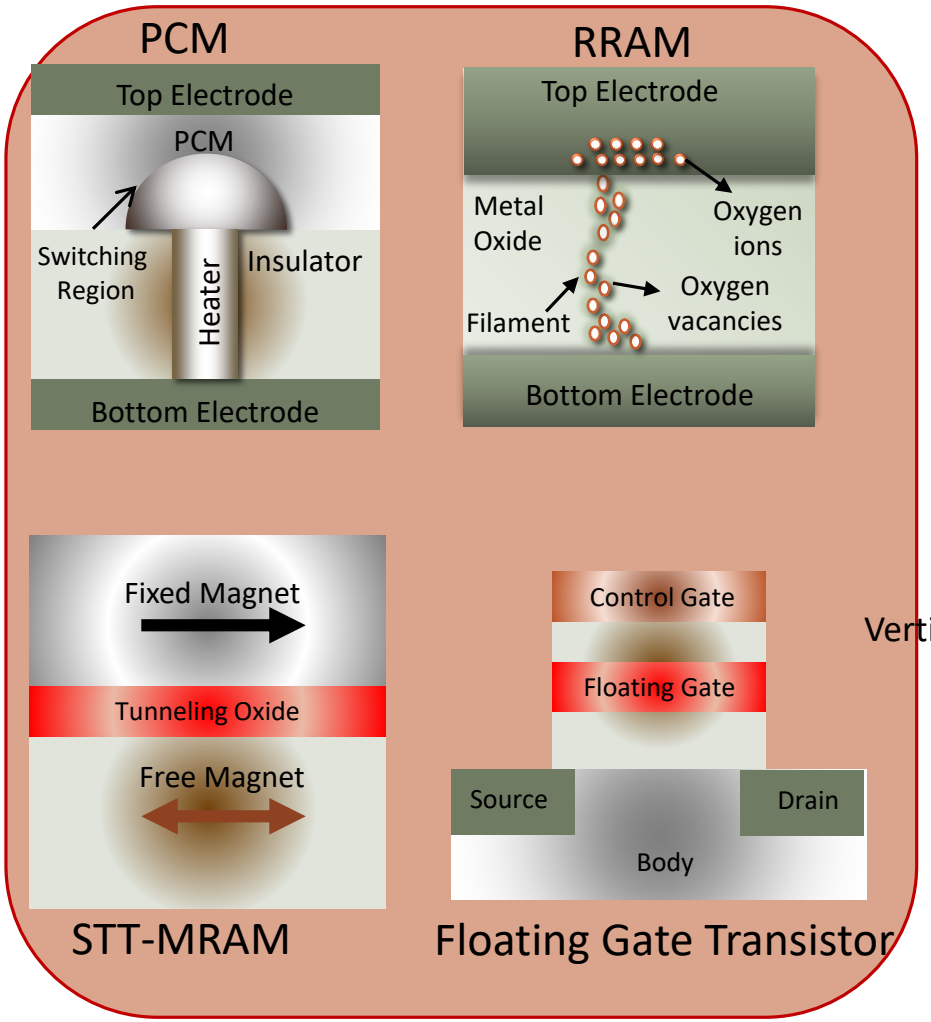


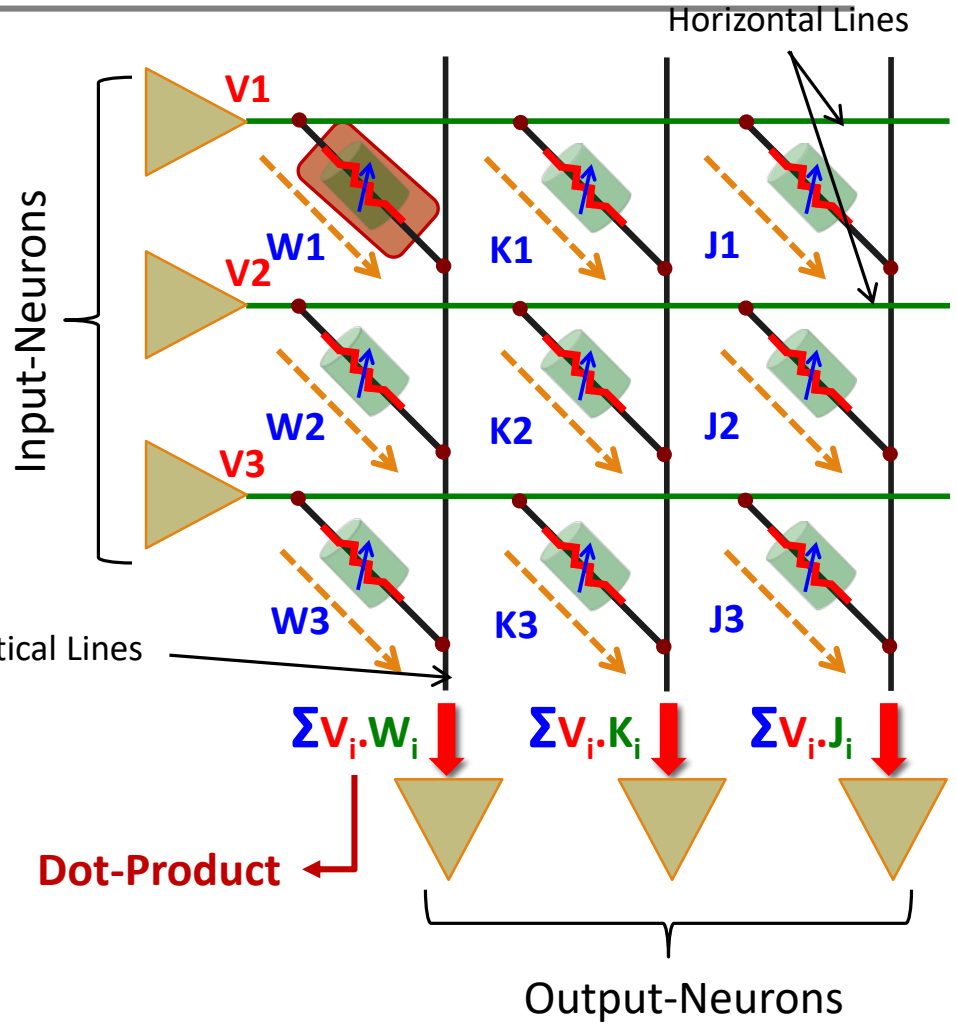
Dot-Product from Kirchoff's Law

von Neumann CMOS-to-
Non-von Neumann Emerging Technology Direction

C-BRIC

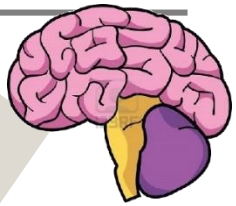
Post-CMOS Devices as Synaptic Memory Elements





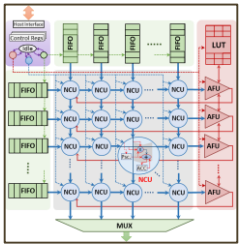
C-BRIC

Hardware for Addressing Inefficiencies

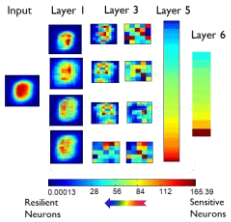


Multicores/GPUs

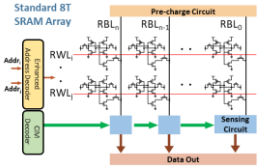
Accelerators



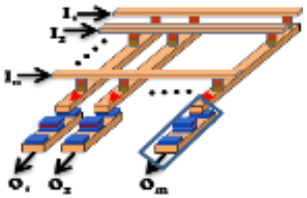
Approximate & Stochastic Hardware



In-memory computing



Post-CMOS Devices



~10^4 Energy Gap

- Can we have algorithms that can yield energy-efficiency?
- Can the algorithms be hardware compatible?

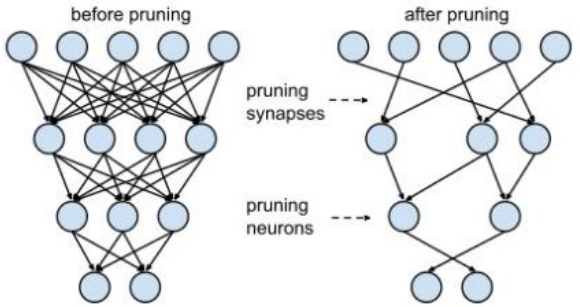
C-BRIC

Algorithms for Addressing Inefficiencies

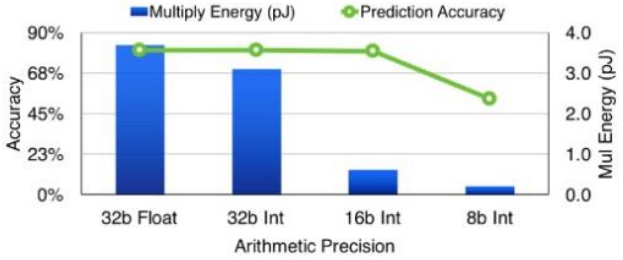
- Pruning/ Compression
- Quantization
- Binarized Networks – Dot Product to simplified XNOR
- Early Exit
- Spiking Networks

Compatible with post-CMOS implementations

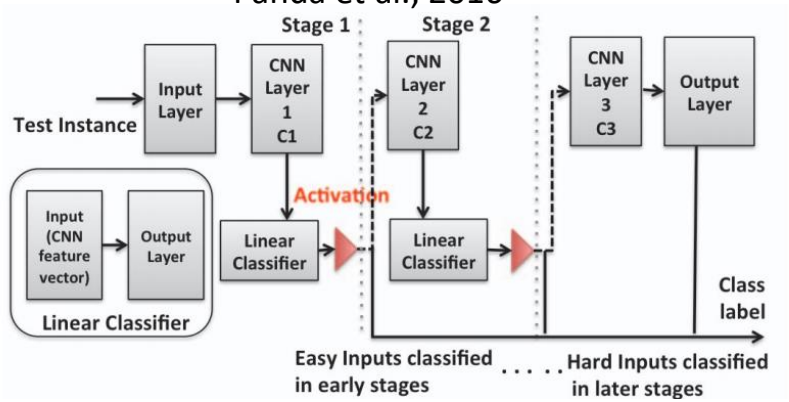
Han et al., 2016

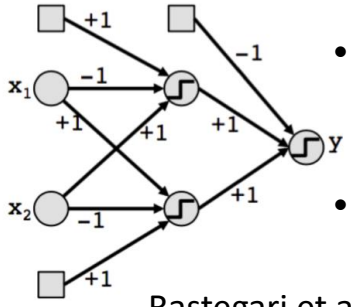


Han et al., 2016, 2017



Panda et al., 2016





- 32x less memory with XNOR than MAC
- 23x faster than MAC

Rastegari et al., 2016

C-BRIC

Algorithms, Systems, Circuits, & Devices

Top-Down: Device-driven Algorithms and Models

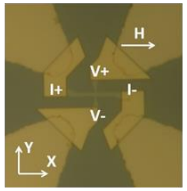
Spiking Networks Ising Networks/Boltzmann Machine Bayesian Networks

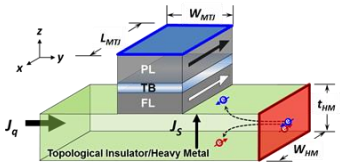
Algorithm-Hardware Co-Design is necessary to reap full benefits!!

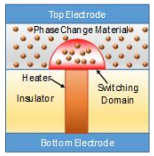
Bottom-Up: Neuro-mimetic Device Design & Fabrication

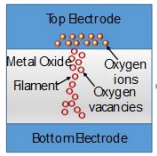
Device physics to mimic probabilistic functionalities

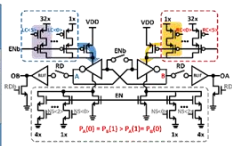
Device Fabrication TI/HM Driven MTJ Switching PCM RRAM CMOS





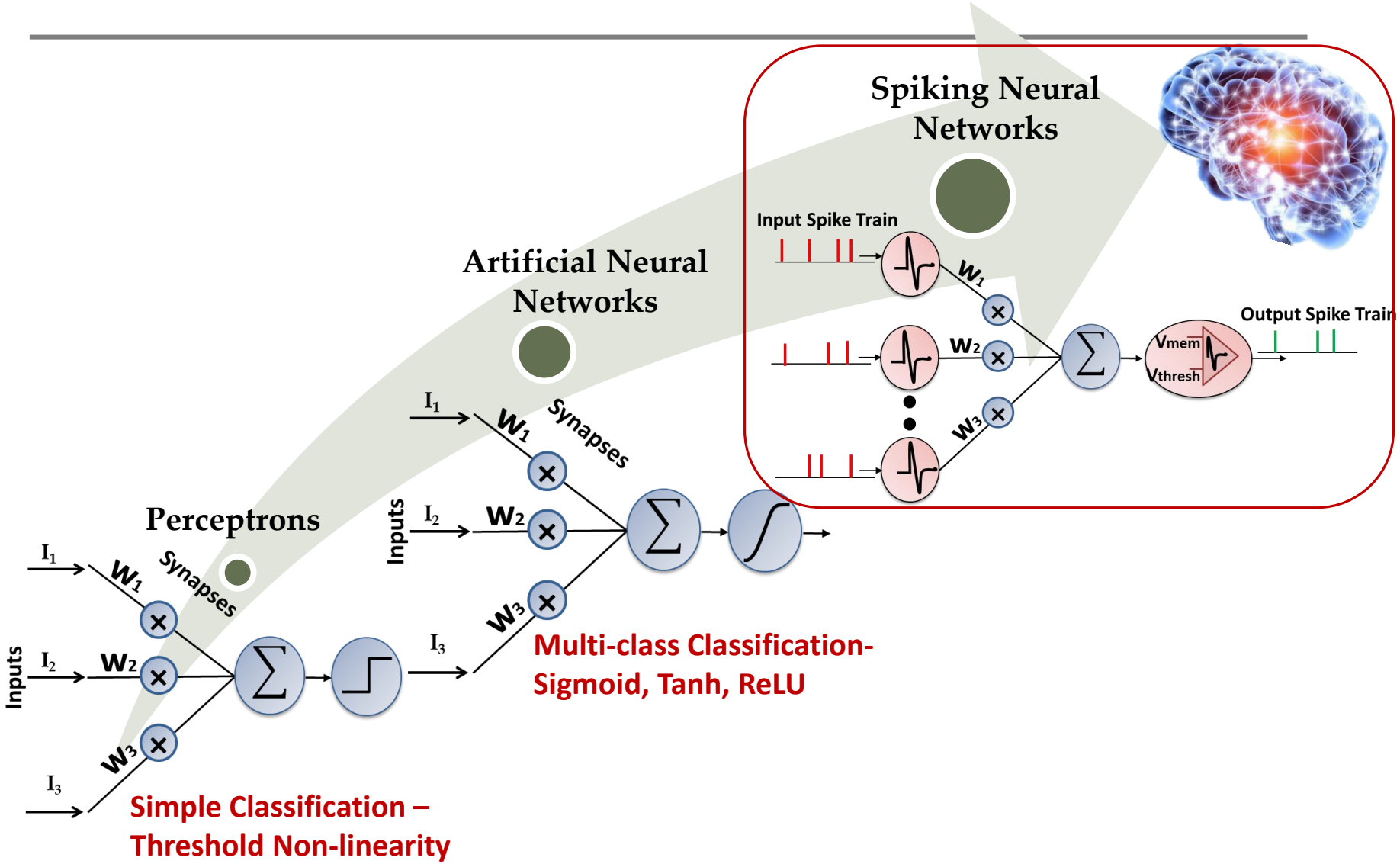






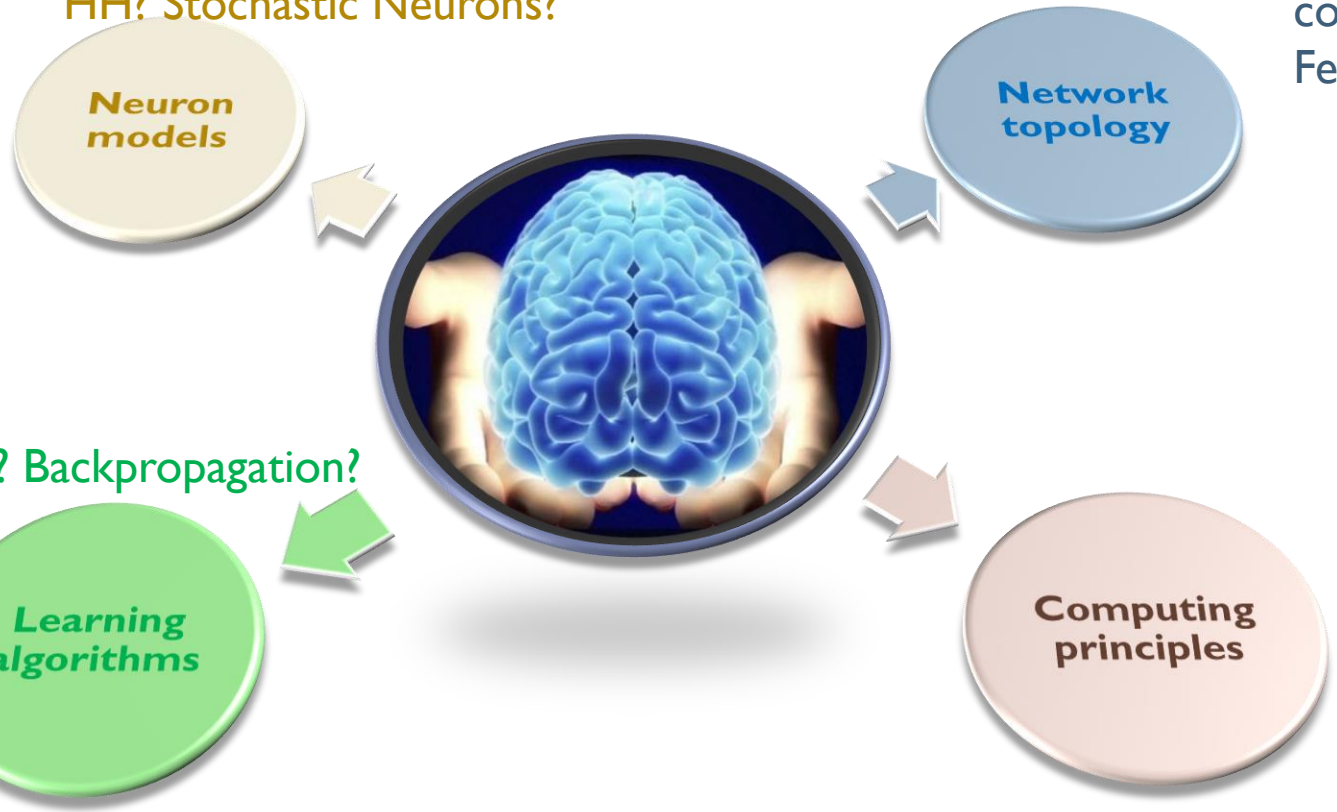
Device-Circuit

Taking the Bio-plausible Route



Which Cues from the Brain can yield Energy-Efficiency?

Discrete Spiking (LIF)?
HH? Stochastic Neurons?



Long-range connections?
Feedback?

STDP? Backpropagation?

Comm. & comp. with spikes? Tightly integrated compute & memory?
Asynchronous, event-driven?
Time-based coding?

C-BRIC

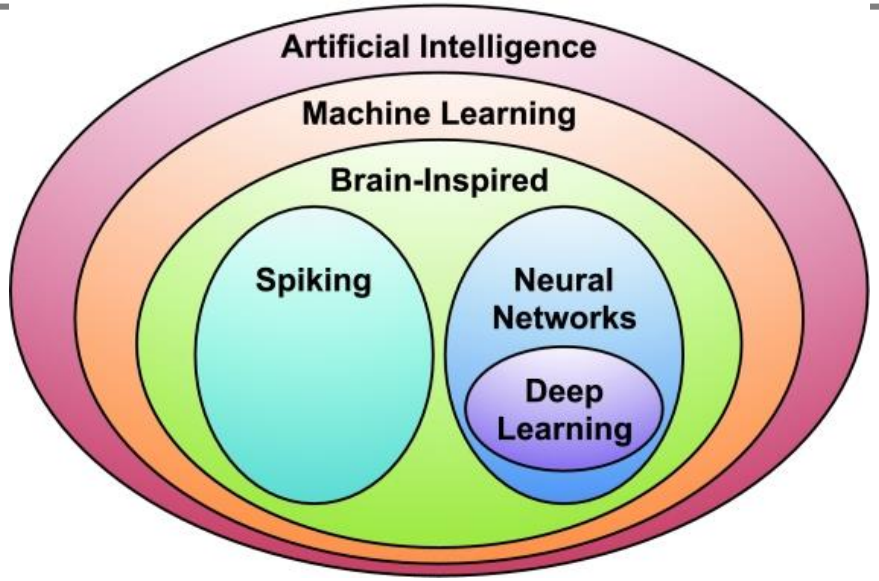
What is Machine Learning

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to progressively improve their performance on a specific task. Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions “without being explicitly programmed” to perform the task.

"A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*."

C-BRIC

AI, ML, Deep Learning, Neural Networks, Spiking



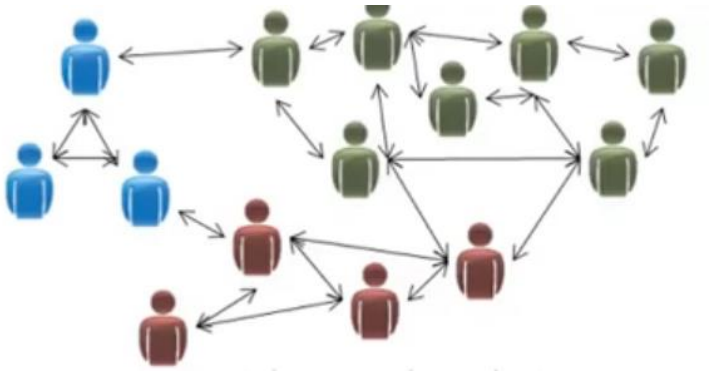
Solving some aspect of 'intelligence'

- Machine Learning - Field of study that gives computers ability to learn without being explicitly programmed
- Brain-inspired computation is a program or algorithm that takes some aspects of its basic form or functionality from the way the brain works ('Brain' is the best source of inspiration for intelligent applications)
- Neural Networks, Deep Learning have dot product or weighted summation of inputs, notion inspired from brain's synaptic/neuronal configuration
- Spiking – More deeply inspired from brain-like computations 'spikes' or 'events'

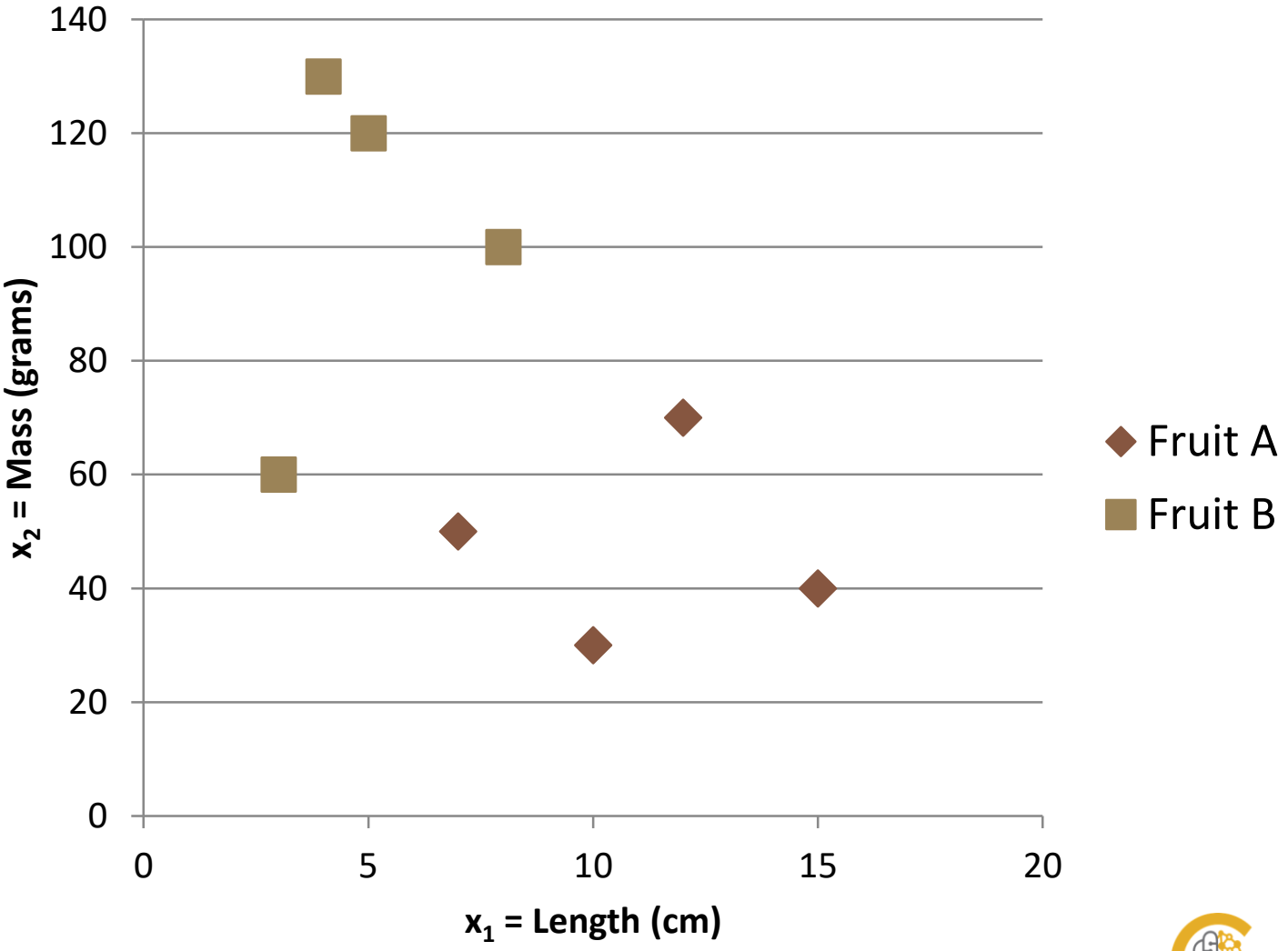
C-BRIC

ML algorithms

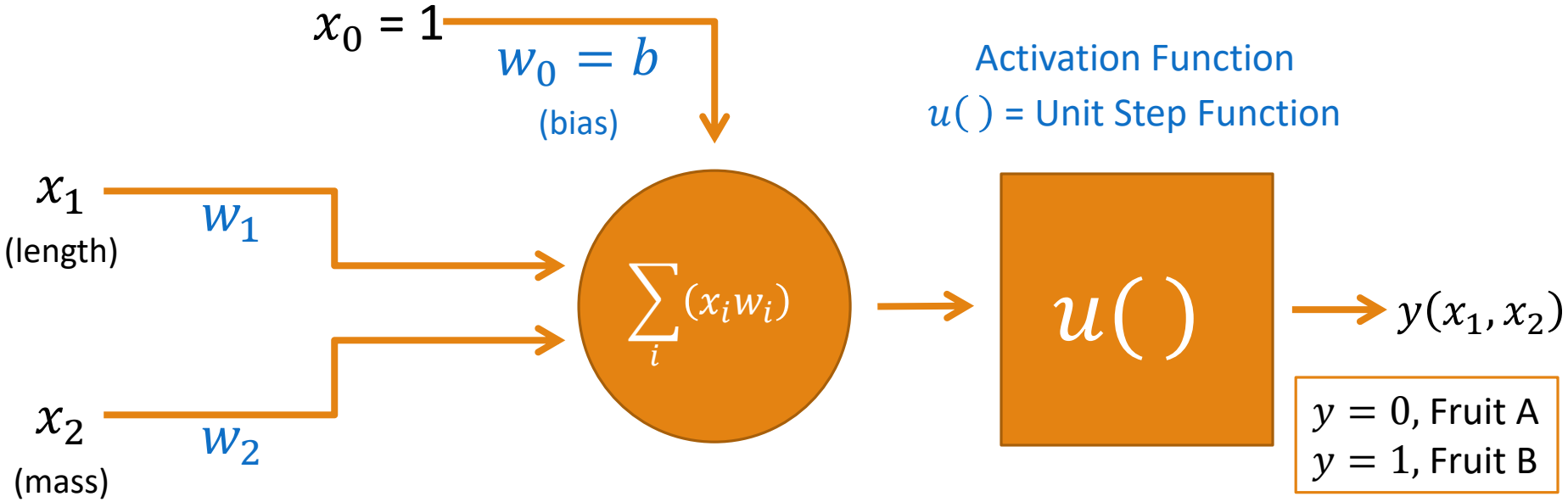
- Supervised Learning – Regression, Classification (Labels given)
- Unsupervised Learning – Clustering (No labels)
  - Application – Google News, Social Network Analysis, Market Segmentation etc.
- Others: Reinforcement Learning, Semi-supervised Learning



VERY SIMPLE NEURAL NETWORK: FRUIT CLASSIFICATION



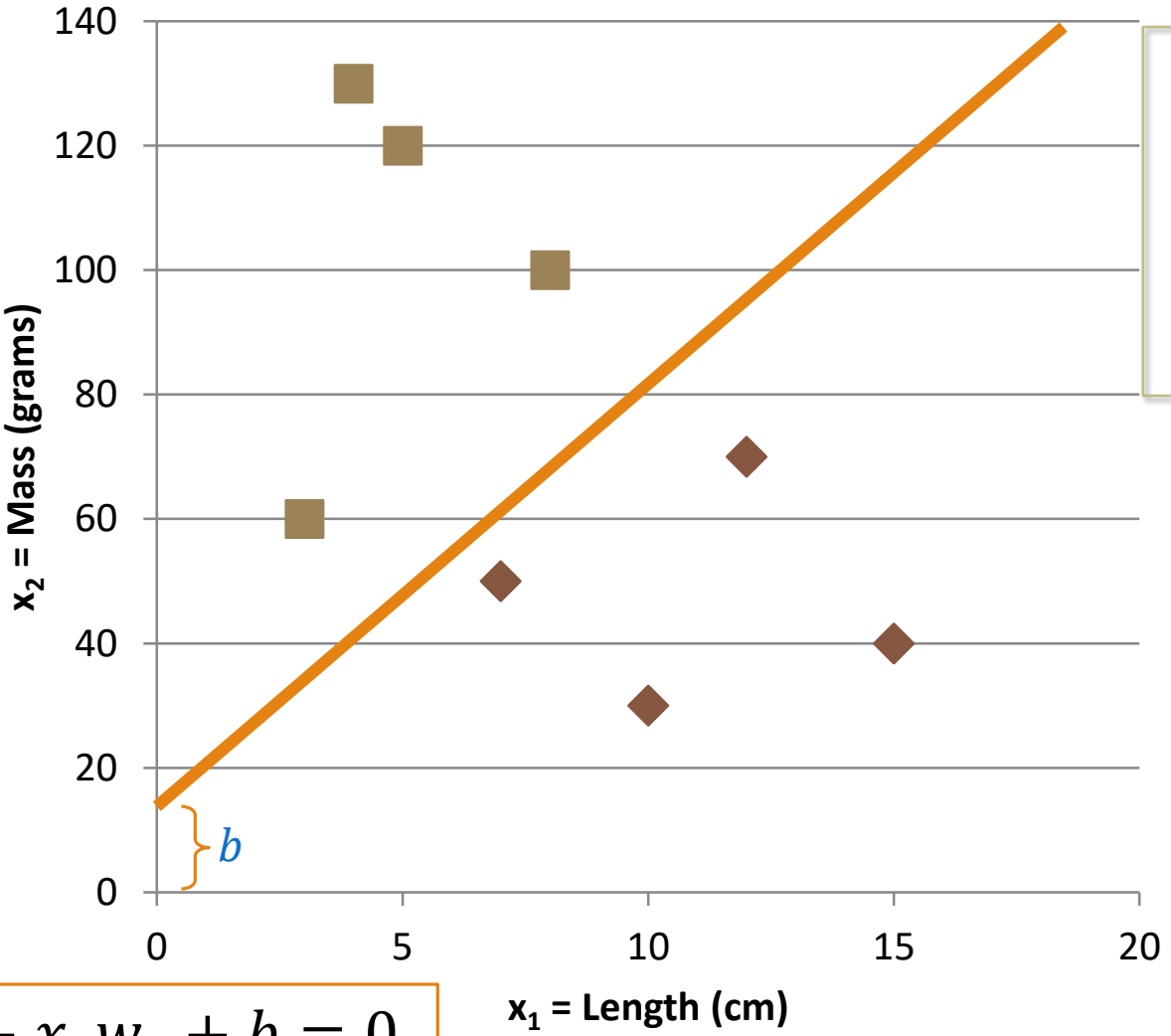
EXAMPLE: FRUIT CLASSIFICATION



Note: a single neuron ANN is called a "perceptron" and can model only linear functions.

$$x_1 w_1 + x_2 w_2 + b = 0$$

EXAMPLE: FRUIT CLASSIFICATION



Note: adding additional layers to the NN allows for non-linear function modeling.

$$b = 15$$
$$w_1 = 6.5$$
$$w_2 = -1$$

$$x_1 w_1 + x_2 w_2 + b = 0$$

C-BRIC