AI FOR EVERYONE

Presented to: American Indian Science and Engineering Society

Presenter: Meghana Rao, Intel Corporation

25th March, 2020

NOTICES AND DISCLAIMERS

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

No product or component can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks .

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/benchmarks .

Intel® Advanced Vector Extensions (Intel® AVX) provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

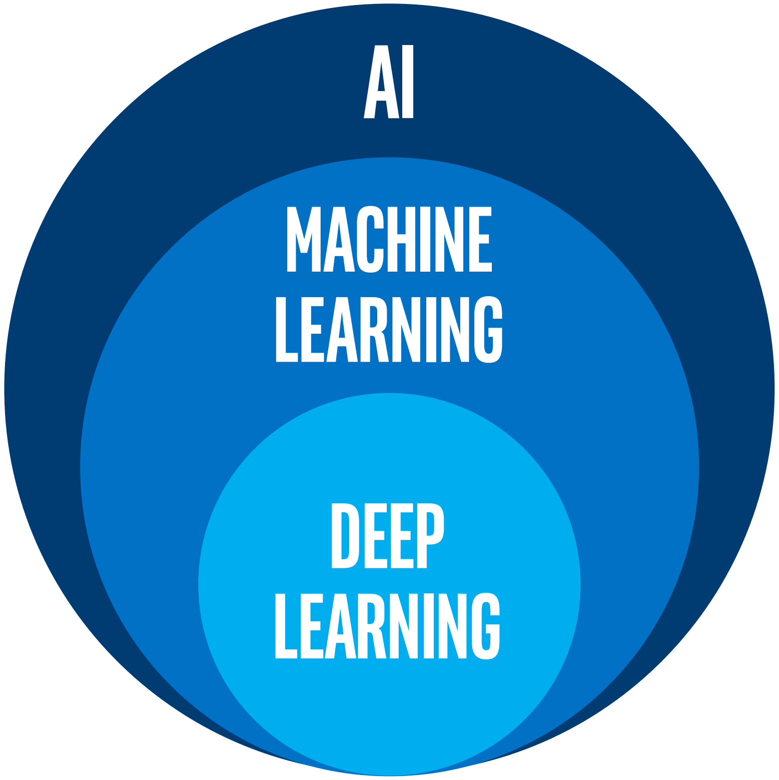
Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

AGENDA

- Introduction to Artificial Intelligence
- AI in the past and present day
- Intel and AI
- AI Journey
- Introduction to Machine Learning
- Introduction to Deep Learning
- Challenges in solving problems through AI
- Community Support
- QnA

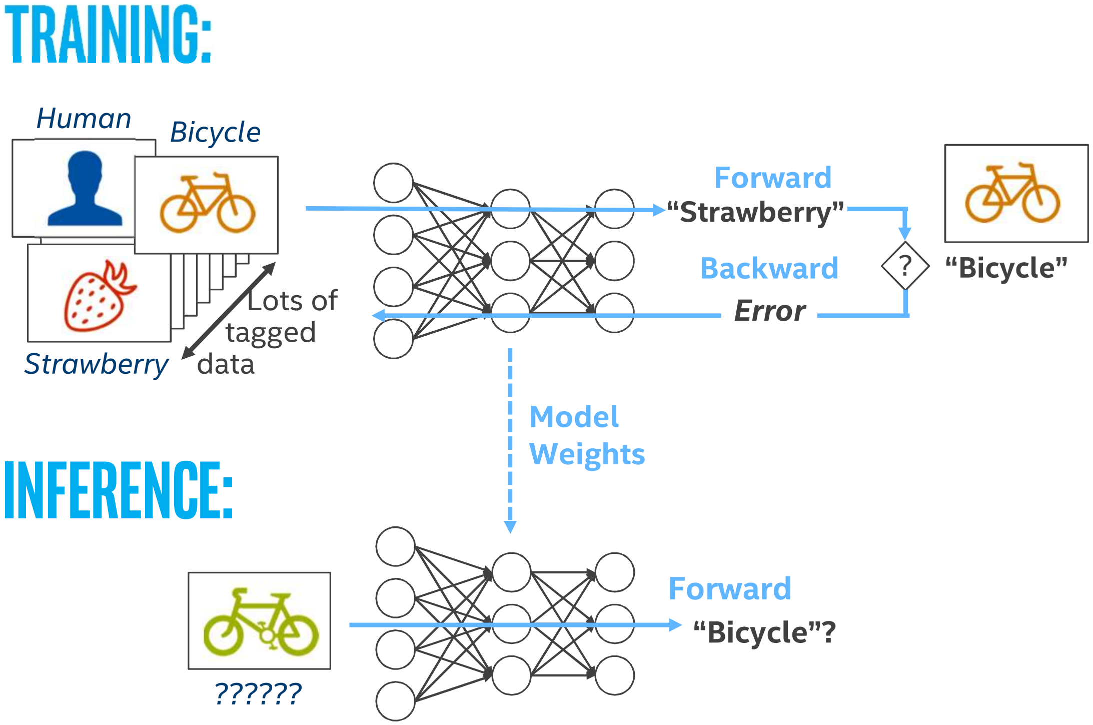
INTRODUCTION TO AI

WHAT IS AI?


AI

MACHINE LEARNING

DEEP LEARNING


TRAINING:

Human

Bicycle

Strawberry

Lots of tagged data

Forward

"Strawberry"

?

"Bicycle"

Backward

Error

Model Weights

INFERENCE:

??????

Forward

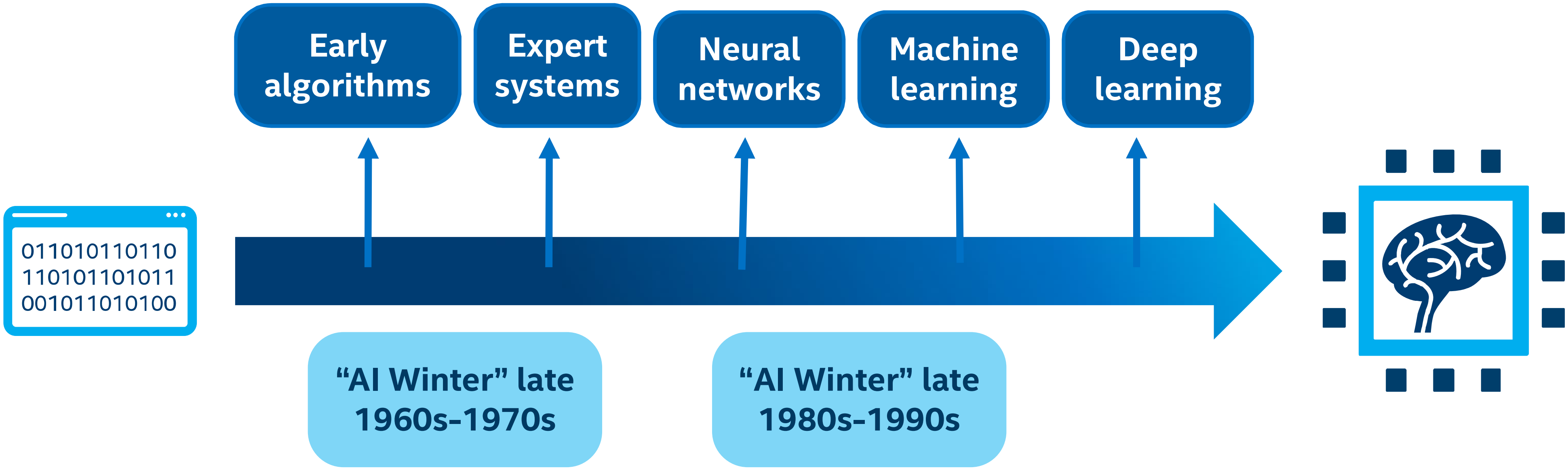
"Bicycle"?

Many different approaches to AI

HISTORY AND REASONS FOR CURRENT MOMENTUM

HISTORY OF AI

AI has experienced several hype cycles, where it has oscillated between periods of excitement and disappointment.



Early algorithms

Expert systems

Neural networks

Machine learning

Deep learning

“AI Winter” late 1960s–1970s

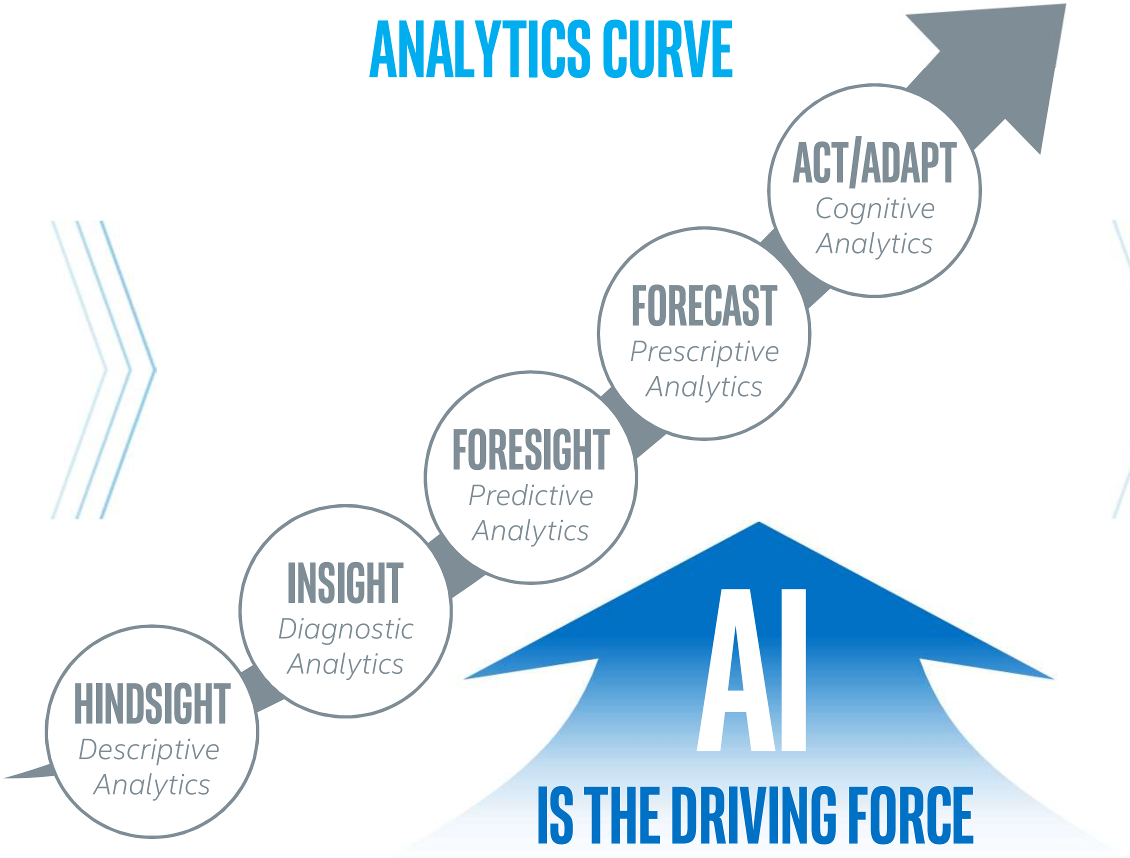
“AI Winter” late 1980s–1990s

WHY AI NOW? ACCESS TO DATA

DATA DELUGE (2019)

- 25 GB[1] per month Internet User
- 50 GB[2] per day Smart Car
- 3 TB[2] per day Smart Hospital
- 40 TB[2] per day Airplane Data
- 1 PB[2] per day Smart Factory
- 50 PB[2] per day City Safety

ANALYTICS CURVE



HINDSIGHT Descriptive Analytics

INSIGHT Diagnostic Analytics

FORESIGHT Predictive Analytics

FORECAST Prescriptive Analytics

ACT/ADAPT Cognitive Analytics

AI IS THE DRIVING FORCE

INSIGHTS

BUSINESS

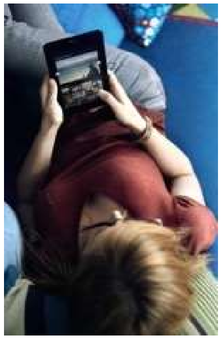
OPERATIONAL

SECURITY

[1] Source: http://www.cisco.com/c/en/us/solutions/service-provider/vni-network-traffic-forecast/infographic.html

[2] Source: https://www.cisco.com/c/dam/m/en_us/service-provider/ciscoknowledgenetwork/files/547_11_10-15-DocumentsCisco_GCI_Deck_2014-2019_for_CKN__10NOV2015_.pdf

AI TRANSFORMATION ACROSS INDUSTRIES



CONSUMER

- Smart Assistants
- Chatbots
- Search
- Personalization
- Augmented Reality
- Robots

HEALTH

- Enhanced Diagnostics
- Drug Discovery
- Patient Care
- Research
- Sensory Aids

FINANCE

- Algorithmic Trading
- Fraud Detection
- Research
- Personal Finance
- Risk Mitigation

RETAIL

- Support
- Experience
- Marketing
- Merchandising
- Loyalty
- Supply Chain
- Security

GOVERNMENT

- Defense
- Data Insights
- Safety & Security
- Resident Engagement
- Smarter Cities

ENERGY

- Oil & Gas Exploration
- Smart Grid
- Operational Improvement
- Conservation

TRANSPORT

- In-Vehicle Experience
- Automated Driving
- Aerospace
- Shipping
- Search & Rescue

INDUSTRIAL

- Factory Automation
- Predictive Maintenance
- Precision Agriculture
- Field Automation

OTHER

- Advertising
- Education
- Gaming
- Professional & IT Services
- Telco/Media
- Sports

Source: Intel forecast

ACCESS TO HARDWARE

END POINT

User-touch end point devices with lower power requirements such as laptops, tablets, smart home devices, drones

EDGE

Small scale data centers, small business IT infrastructure, to few on-premise server racks and workstations

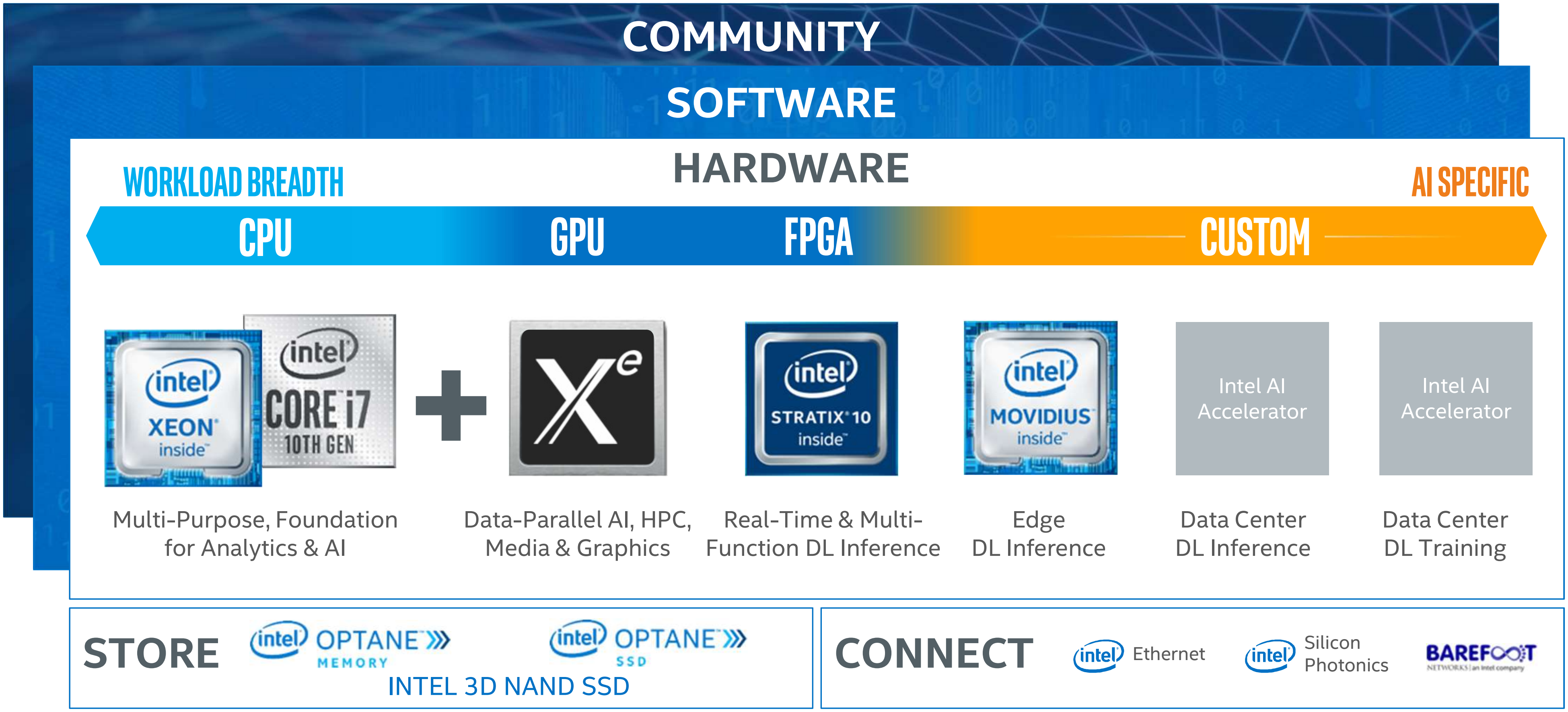
DATA CENTER

Large scale data centers such as public cloud or comms service providers, gov't and academia, large enterprise IT

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

INTEL AI PORTFOLIO

ONE INTEL ANALYTICS & AI PRODUCTS



COMMUNITY

SOFTWARE

HARDWARE

WORKLOAD BREADTH

AI SPECIFIC

CPU + GPU, FPGA, CUSTOM

CPU	GPU	FPGA	CUSTOM	CUSTOM	CUSTOM
Xeon + Core i7 10th Gen	Xe	Stratix 10	Movidius	Intel AI Accelerator	Intel AI Accelerator
Multi-Purpose, Foundation for Analytics & AI	Data-Parallel AI, HPC, Media & Graphics	Real-Time & Multi-Function DL Inference	Edge DL Inference	Data Center DL Inference	Data Center DL Training

STORE: Optane Memory, Optane SSD, INTEL 3D NAND SSD

CONNECT: Ethernet, Silicon Photonics, Barefoot Networks

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

INTEL AI SOFTWARE

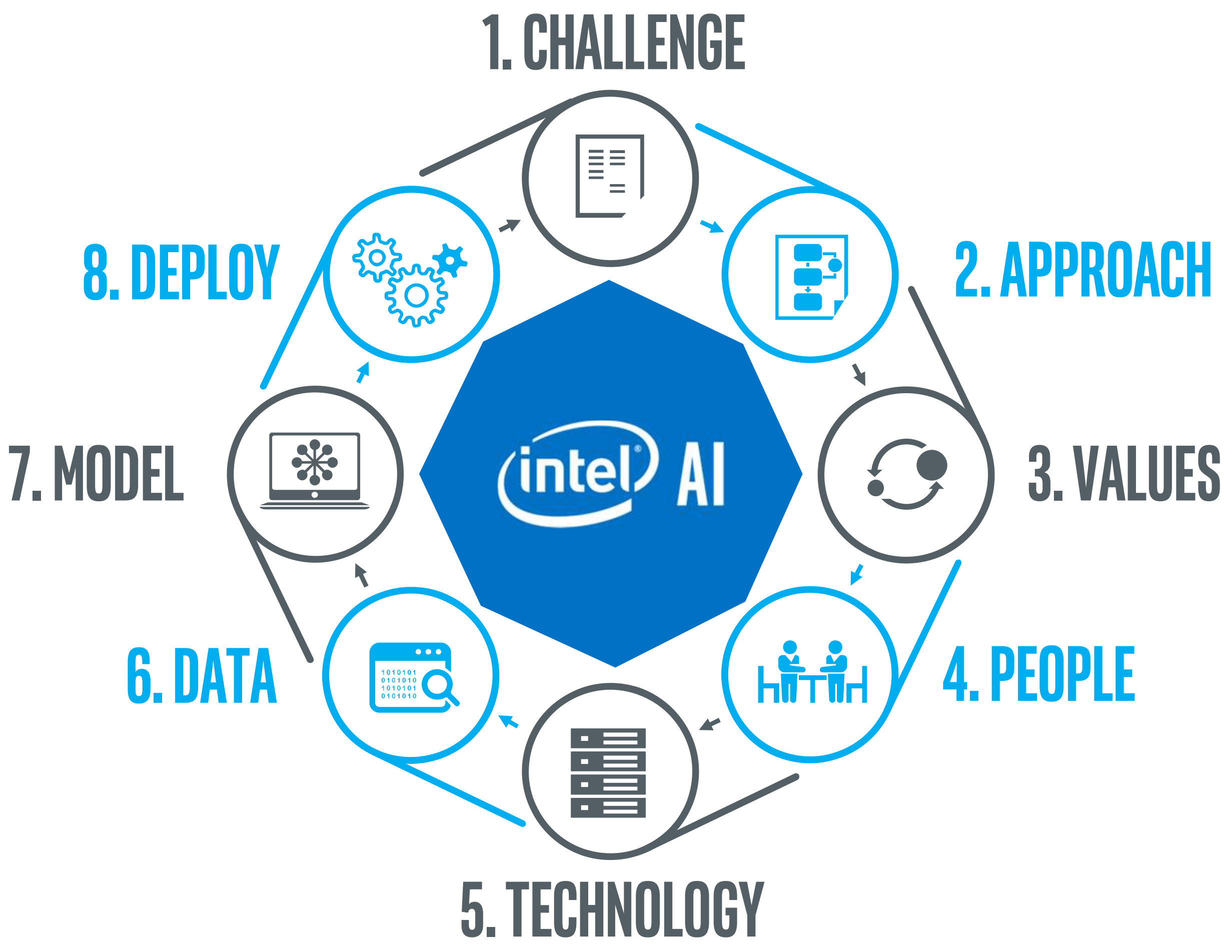
AI = ML + DL + MT

	MACHINE LEARNING	DEEP LEARNING		MANAGEMENT TOOLS	
DEVELOPER TOOLS App Developers SW Platform Developer	ANALYTICS ZOO	ModelZoo OpenVINO		Containers	
TOPOLOGIES & MODELS Data Scientist **FRAMEWORKS** Data Scientist	- **Intel Distribution for Python** (SKlearn, Pandas) python - MILib and Mahout on Spark Spark	TensorFlow PyTorch PaddlePaddle mxnet Caffe BigDL ONNX	**DL Tools** NN Distiller, NLP Architect, RL Coach, CVAT	NAUTA kubernetes Kubeflow Deep Learning Reference Stack	Architect & DevOps
GRAPH ML Performance Engineer **KERNEL** ML Performance Engineer	- Intel Data Analytics Acceleration Library (Intel DAAL) - Intel Math Kernel Library (Intel MKL)	- nGraph - PlaidML - Intel Machine Learning Scaling Library (Intel MLSL) - Deep Learning Boost - Intel® Deep Neural Network Library (DNNL)		Data Analytics Reference Stack	
	CPU	CPU ▪ GPU ▪ FPGA ▪ CUSTOM			

Red font products are the most broadly applicable SW products for AI users

AI JOURNEY

THE AI JOURNEY



MACHINE LEARNING

MACHINES LEARN IN TWO WAYS

Supervised Learning & Unsupervised Learning

SUPERVISED LEARNING

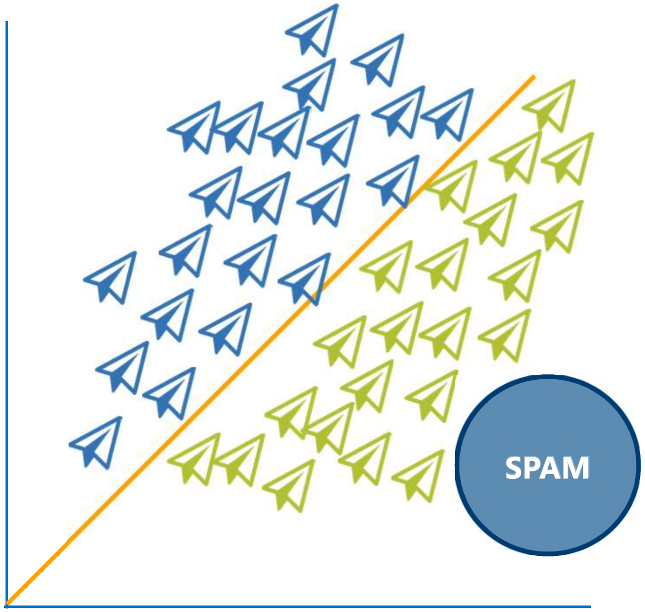
We train the model. We feed the model with correct answers. Model Learns and finally predicts.

We feed the model with “ground truth”.

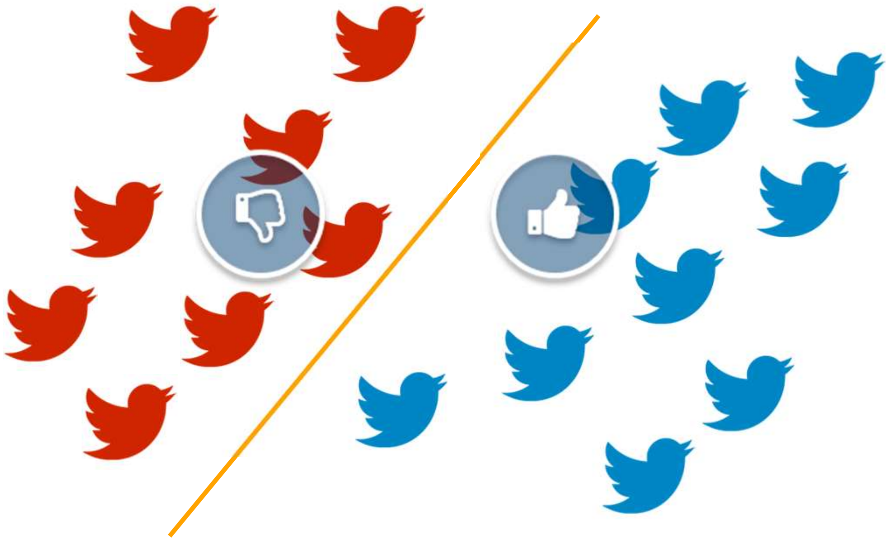
EXAMPLES OF SUPERVISED LEARNING - CLASSIFICATION

Predict a **label** for an entity with a given set of features.

PREDICTION



SENTIMENT ANALYSIS



EVALUATION METRIC

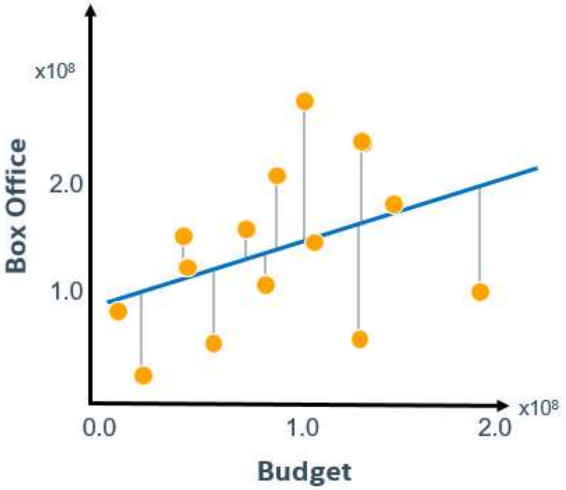
There are many metrics available* to measure performance, such as:

- **Accuracy**: how well predictions match true values.

Accuracy target

- **Mean Squared Error**: average square distance between prediction and true value.

$$\min_{\beta_0,\beta_1} \frac{1}{m} \sum_{i=1}^{m} \left(\left(\beta_0 + \beta_1 x_{obs}^{(i)} \right) - y_{obs}^{(i)} \right)^2$$



*The wrong metric can be misleading or not capture the real problem.

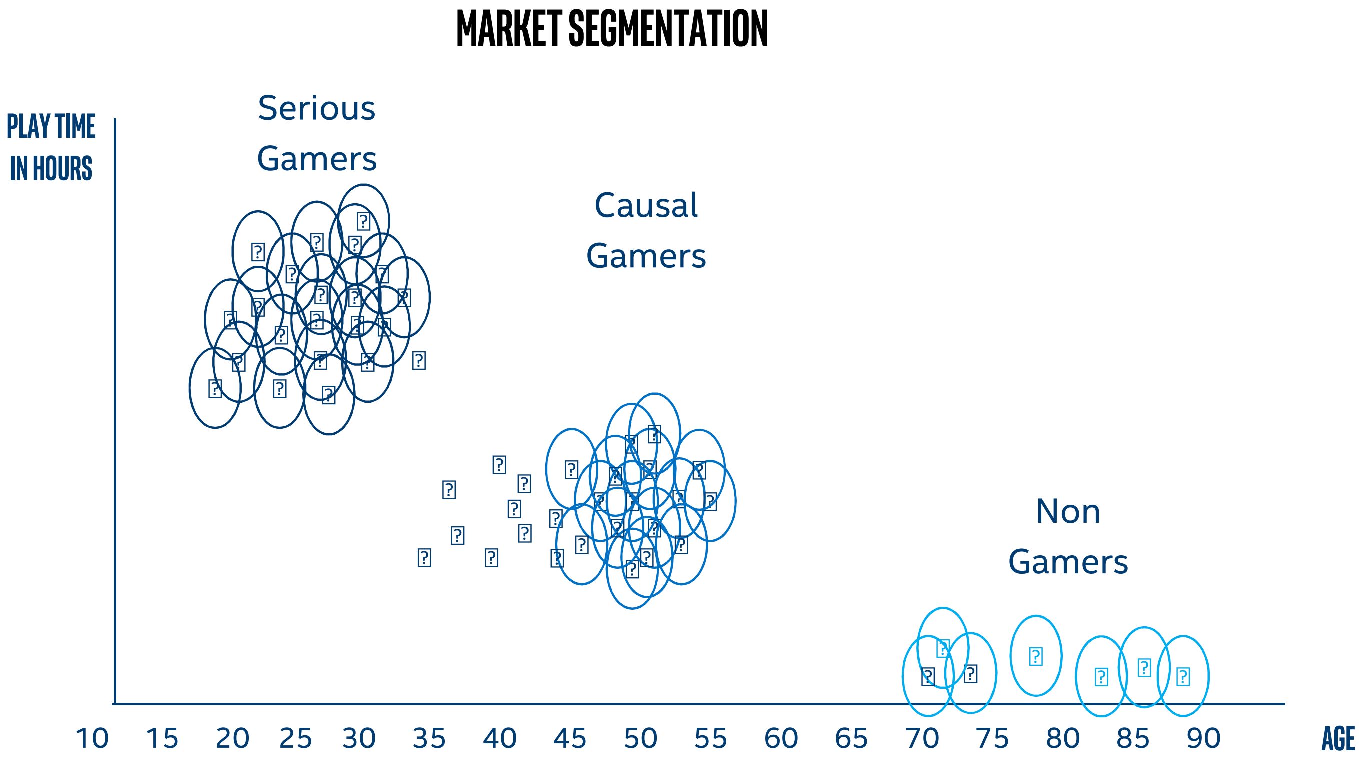
UNSUPERVISED LEARNING

Data is given to the model. Right answers are not provided to the model. The model makes sense of the data given to it.

Can teach you something you were probably not aware of in the given dataset.

EXAMPLE OF UNSUPERVISED LEARNING - CLUSTERING

Group entities with similar features



ADDITIONAL MACHINE LEARNING EXAMPLES

Fraud Detection

Movie Recommendation

Recommending

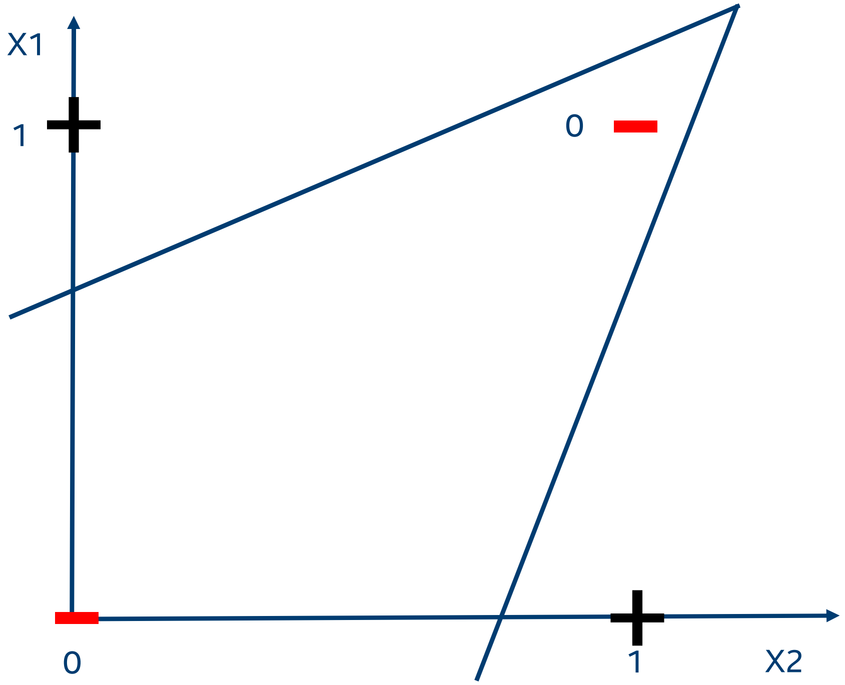
Similar news articles

CHC
3 0
CLE
World Series Live: Cubs Grab Early Lead
A&M Sneaks in at No. 4 in First CFP Rankings
B/R's Midseason NFL All-Rookie Team
Thunder Already Have an Eye on Trade Deadline
Week 9 Fantasy Football Preview

Other brand names can be claimed as the property of others

WHAT IS THE LIMITATION WITH LINEAR CLASSIFIERS?

X1	X2	y
0	0	0
0	1	1
1	0	1
1	1	0



XOR
The counter example to all models

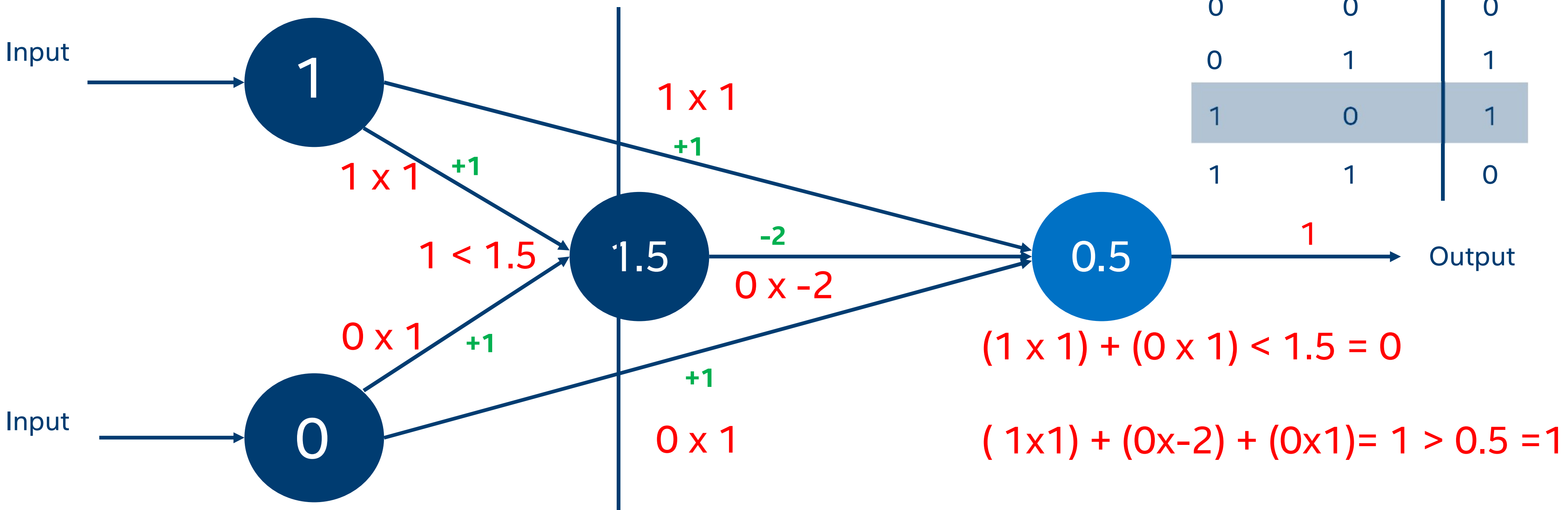
We need non-linear functions

Source: https://medium.com/towards-data-science/introducing-deep-learning-and-neural-networks-deep-learning-for-rookies-1-bd68f9cf5883

WE NEED LAYERS USUALLY LOTS WITH NON-LINEAR TRANSFORMATIONS

XOR = (X1 and not X2) OR (Not X1 and X2)

X1	X2	y
0	0	0
0	1	1
1	0	1
1	1	0



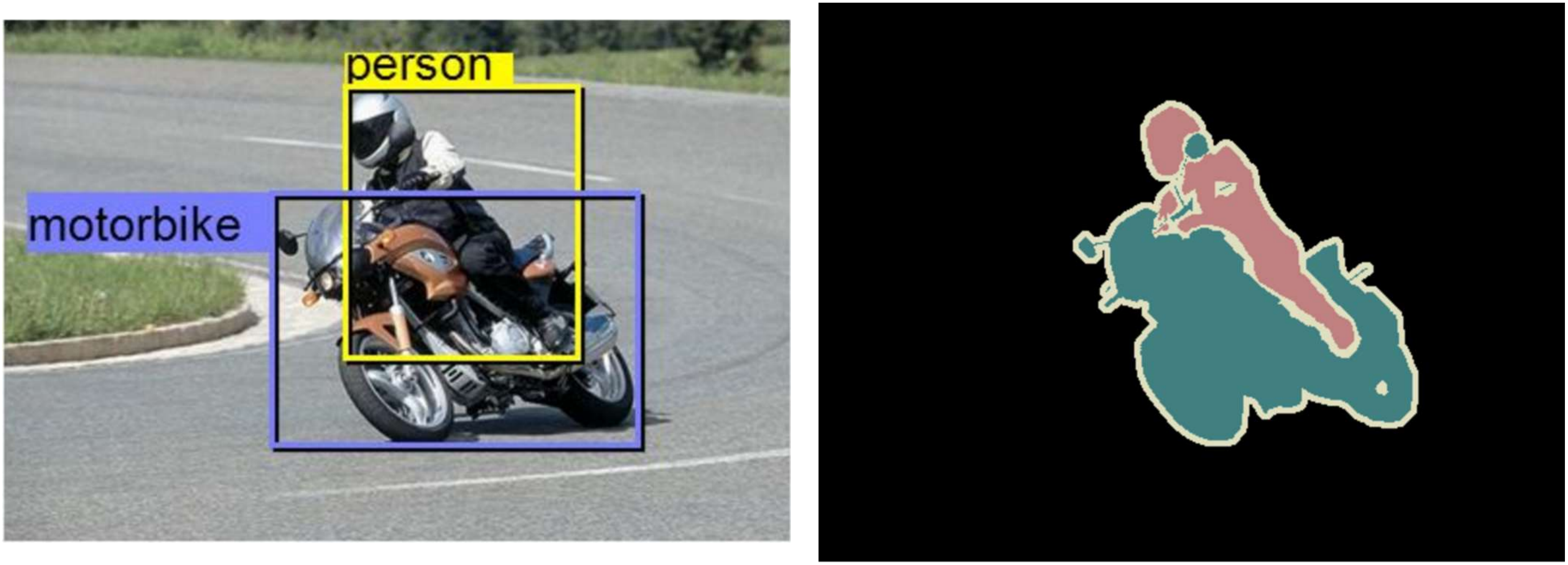
(1 x 1) + (0 x 1) < 1.5 = 0

(1x1) + (0x-2) + (0x1)= 1 > 0.5 =1

Threshold to 0 or 1

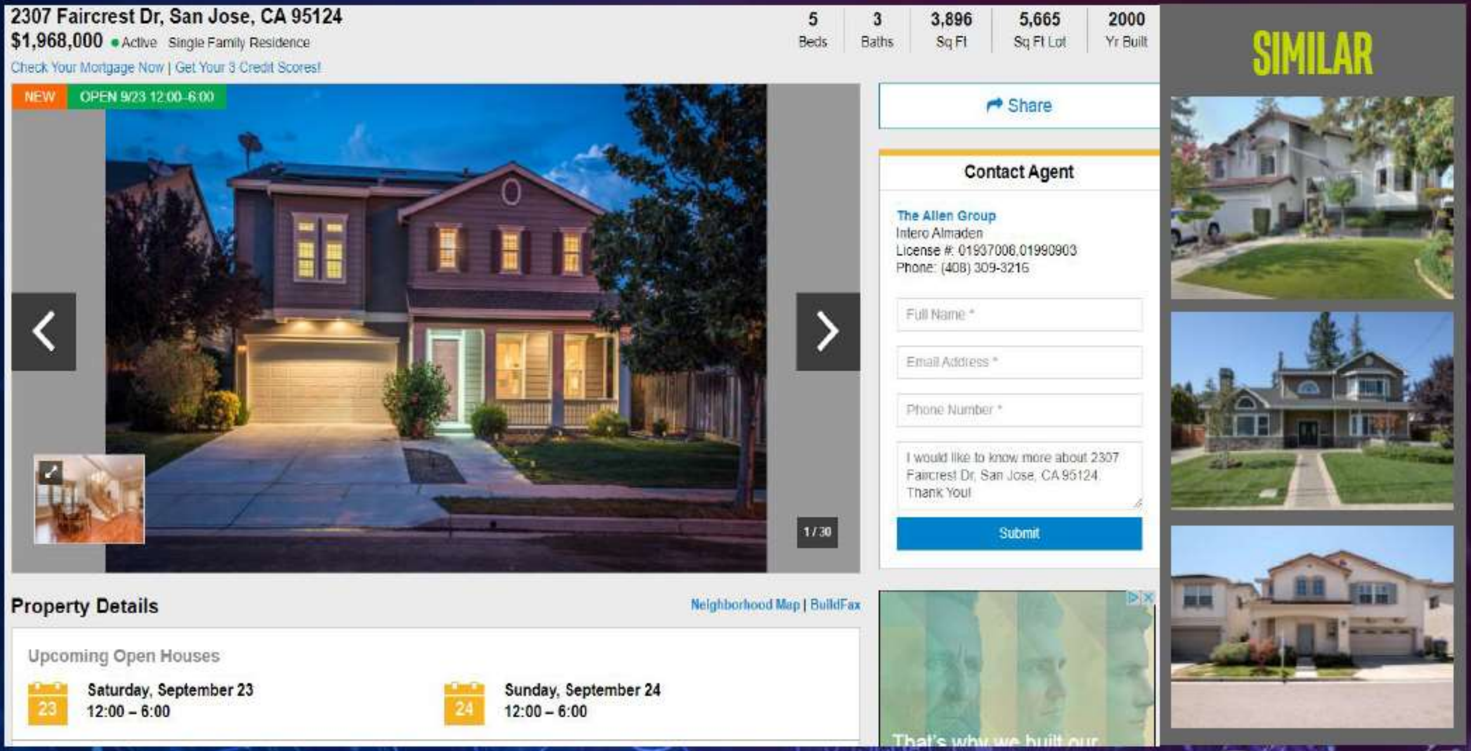
DEEP LEARNING

CLASSIFICATION / DETECTION / SEMANTIC SEGMENTATION



https://people.eecs.berkeley.edu/~jhoffman/talks/lsda-baylearn2014.pdf

HOME BUYING ASSISTANT: 10 CPU NODES



2307 Faircrest Dr, San Jose, CA 95124

$1,968,000 ● Active Single Family Residence

Check Your Mortgage Now | Get Your 3 Credit Scores!

5	3	3,896	5,665	2000
Beds	Baths	Sq Ft	Sq Ft Lot	Yr Built

Contact Agent

The Allen Group
Intero Almaden
License #: 01937006,01990903
Phone: (408) 309-3216

Full Name *

Email Address *

Phone Number *

I would like to know more about 2307 Faircrest Dr, San Jose, CA 95124. Thank You!

Submit

Property Details

Neighborhood Map | BuildFax

Upcoming Open Houses

23 Saturday, September 23
12:00 – 6:00

24 Sunday, September 24
12:00 – 6:00

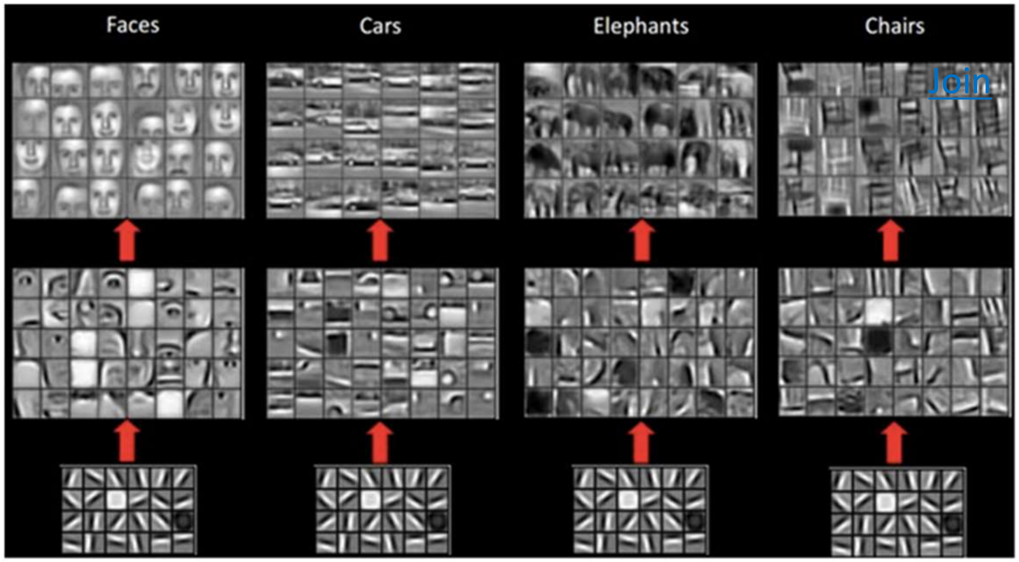
J. Dai, Y. Yuhao and J. Wang, "Using BigDL to build image similarity-based house recommendations." Nov. 2017.
https://software.intel.com/en-us/articles/using-bigdl-to-build-image-similarity-based-house-recommendations

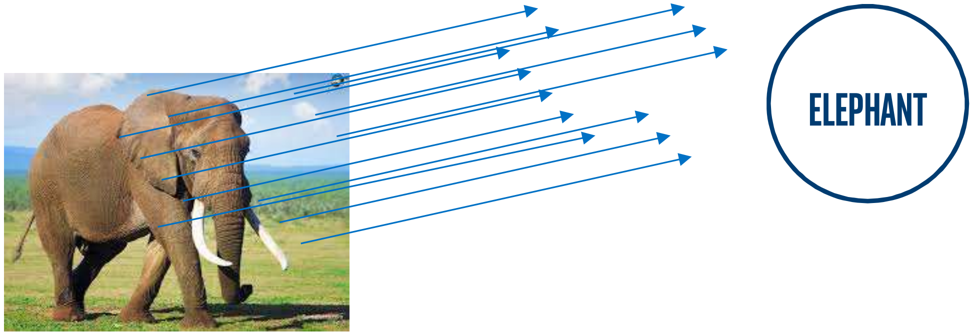
FRAMEWORK HARDWARE

* Other names and brands may be claimed as the property of others.

HOW DO DEEP LEARNING NETWORKS LEARN? EACH LAYER LEARNS SOMETHING

Layer 1 ▶ Layer 2 .... Layer N ▶ Prediction



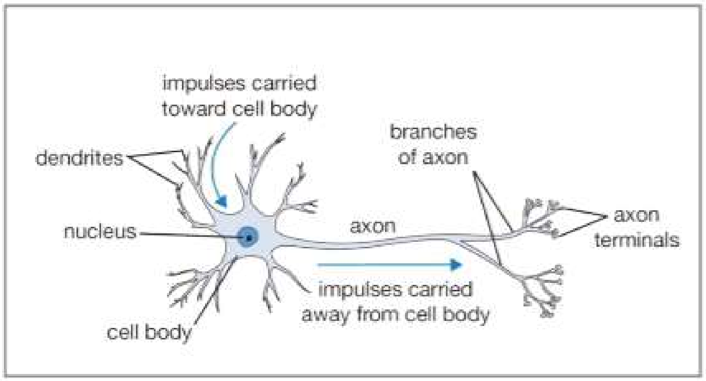


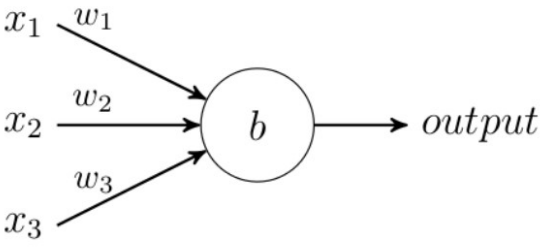
Zeiler & Fergus '13

HOW CAN I BUILD A NEURAL NETWORK?

MOTIVATION FOR NEURAL NETS

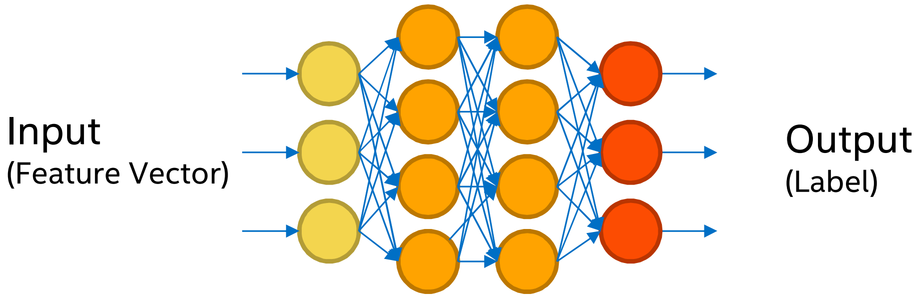
- Use biology as inspiration for mathematical model
- Get signals from previous neurons
- Generate signals (or not) according to inputs
- A neuron fires when it's output > threshold
- Pass signals on to next neurons
- By layering many neurons, can create complex model


impulses carried toward cell body
branches of axon
dendrites
nucleus
axon
axon terminals
impulses carried away from cell body
cell body



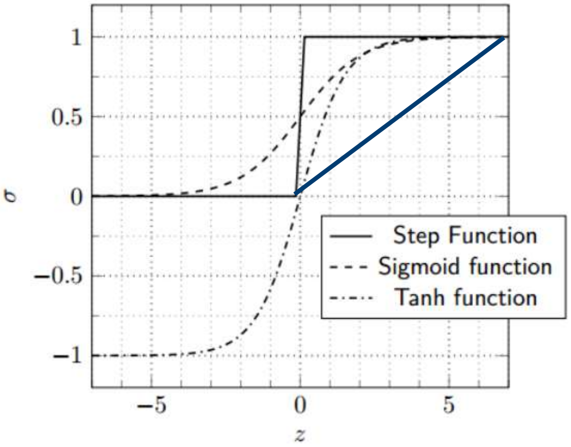
FULLY CONNECTED NEURAL NETWORK

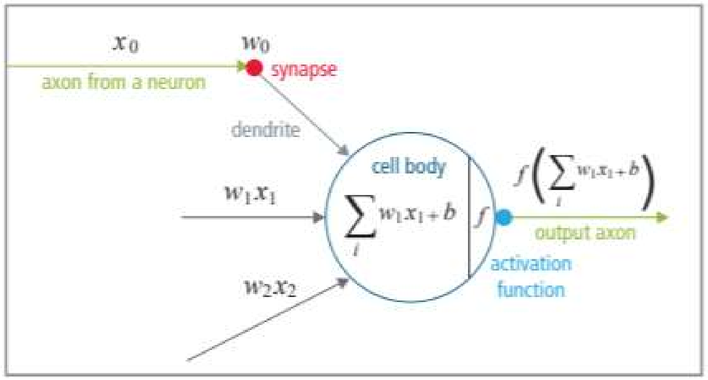
- Multiple layers of stacked neurons forming a network (topology)
- Each neuron is connected to every neuron in subsequent layers
- Network topologies are constantly evolving based on complexity of problems being solved by AI



WHAT IS AN ACTIVATION FUNCTION?

- The output of a neuron could range from –infinity to + infinity
- How does it know when to fire?
- An activation function establishes a boundary for the output
- Many types of activation functions exist





MATRIX REPRESENTATION OF COMPUTATION

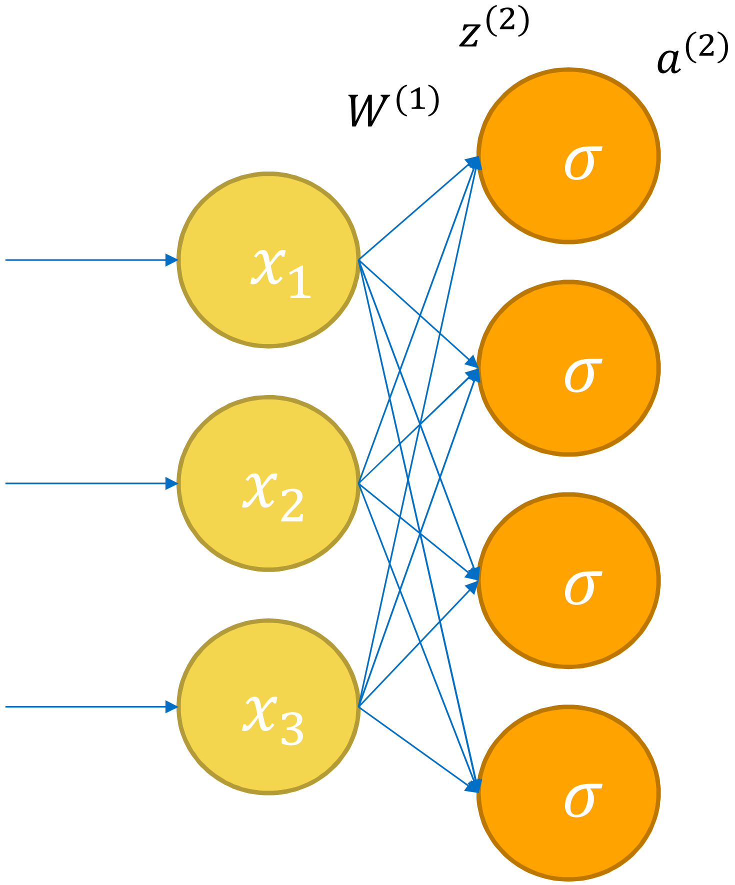
$$z^{(2)} = xW^{(1)}$$

$$a^{(2)} = \sigma(z^{(2)})$$

$W^{(1)}$ is a 3x4 matrix

$z^{(2)}$ is a 4-vector

$a^{(2)}$ is a 4-vector



CONTINUING THE COMPUTATION

For a single training instance (data point)

Input: vector x (a row vector of length 3)

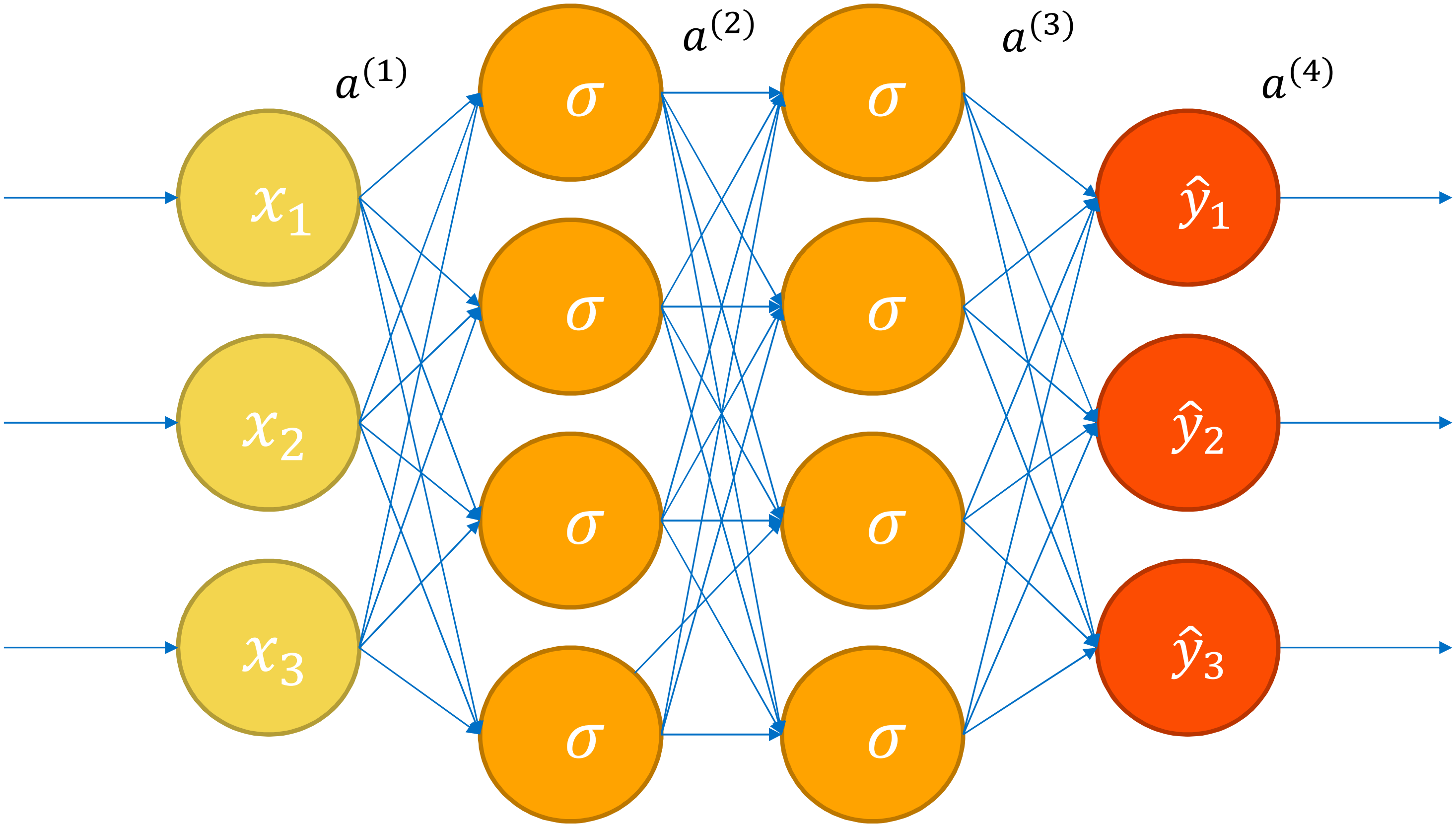
Output: vector $\hat{y}$ (a row vector of length 3)

$$z^{(2)} = xW^{(1)} \qquad a^{(2)} = \sigma(z^{(2)})$$

$$z^{(3)} = a^{(2)}W^{(2)} \qquad a^{(3)} = \sigma(z^{(3)})$$

$$z^{(4)} = a^{(3)}W^{(3)} \qquad \hat{y} = softmax(z^{(4)})$$

A FULLY CONNECTED NEURAL NETWORK WITH ACTIVATIONS

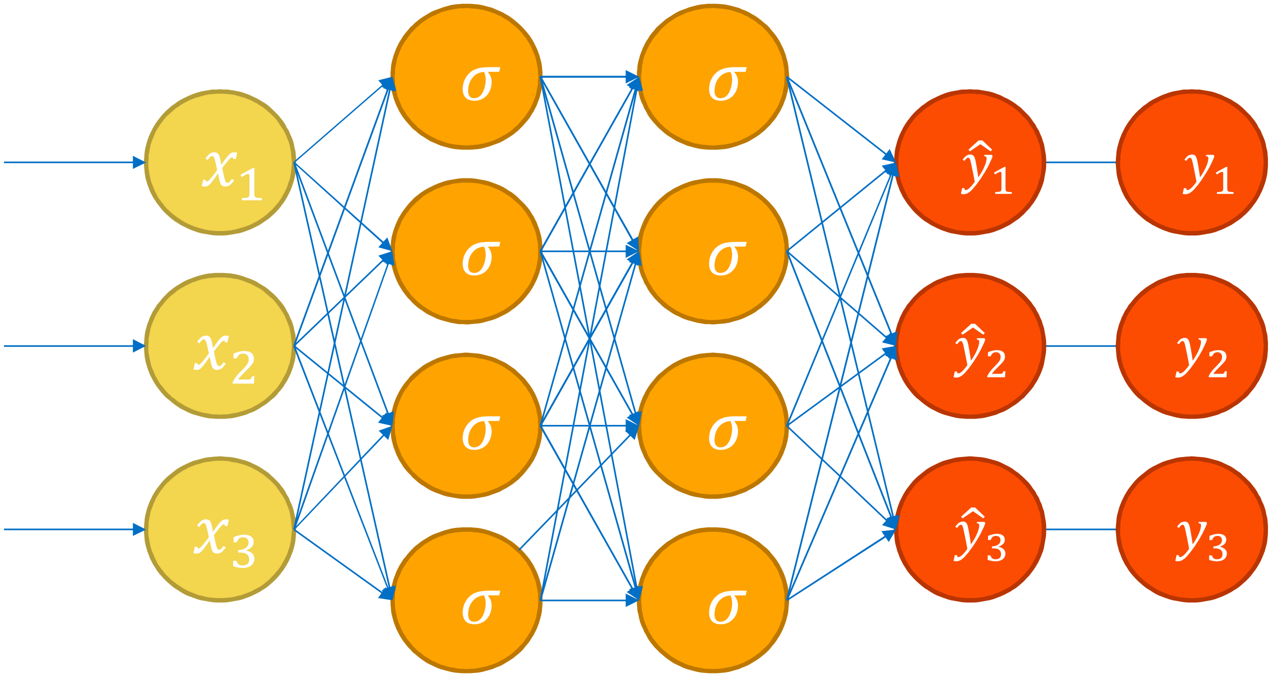


$\hat{y}, the\ actual\ ouput$ may not be the expected output.
The Network needs to be trained to get better accuracy

HOW CAN I TRAIN A NEURAL NETWORK?

FORWARD PROPAGATION

- Calculate the Loss Function – compare the predictions to the ground truth

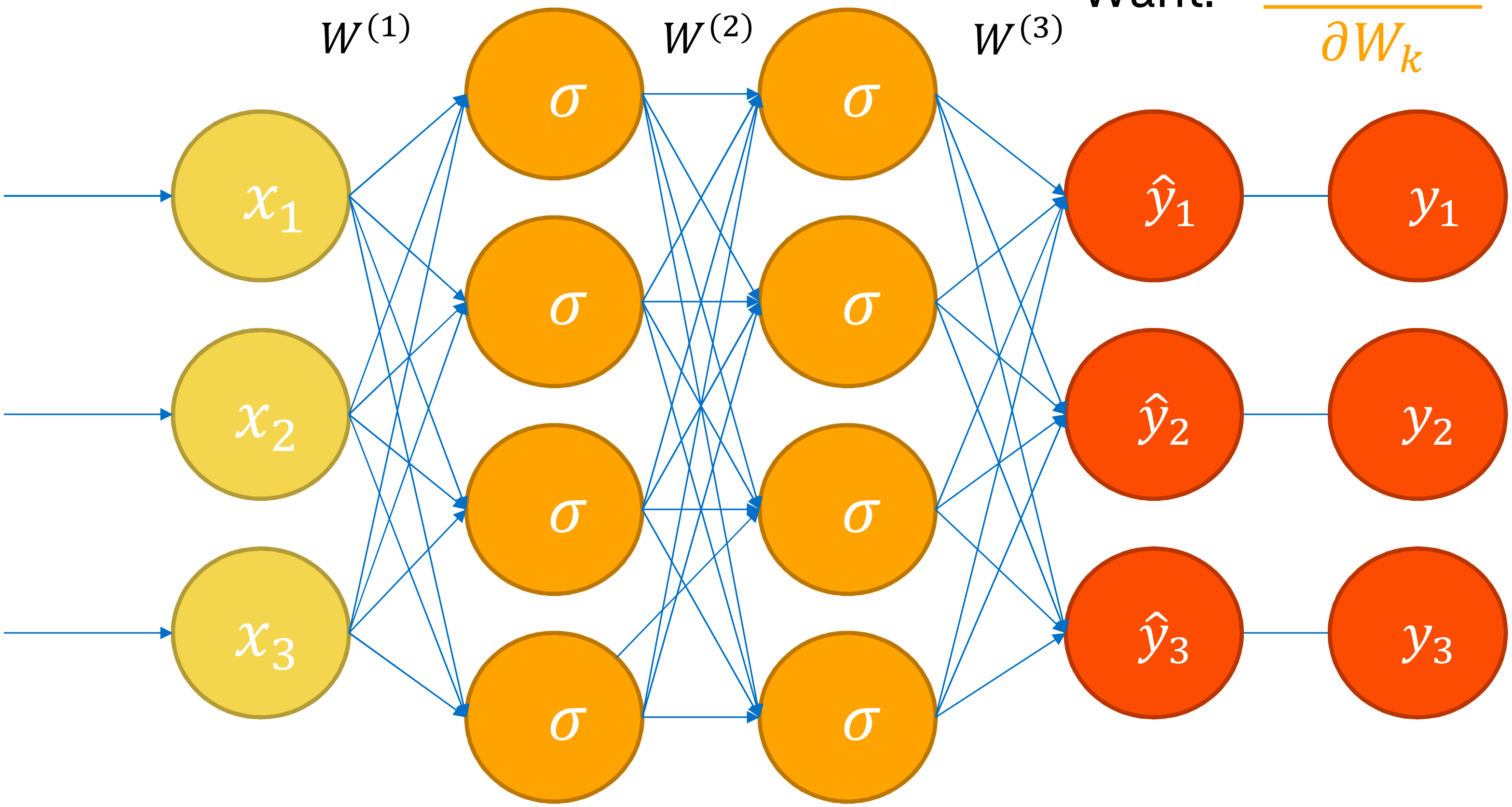


Evaluate: $J(y_i, \hat{y}_i)$

- How far the “actual output” is from “Ground Truth” determines how much more the network needs to learn to adjust it’s output to minimize loss

BACKPROPAGATION

Want: $\dfrac{\partial J(y_i, \hat{y}_i)}{\partial W_k}$



APPLY GRADIENTS TO EVERY WEIGHT IN THE NETWORK

$$\frac{\partial J}{\partial W^{(3)}} = (\hat{y} - y) \cdot a^{(3)}$$

$$\frac{\partial J}{\partial W^{(2)}} = (\hat{y} - y) \cdot W^{(3)} \cdot \sigma'\left(z^{(3)}\right) \cdot a^{(2)}$$

$$\frac{\partial J}{\partial W^{(1)}} = (\hat{y} - y) \cdot W^{(3)} \cdot \sigma'\left(z^{(3)}\right) \cdot W^{(2)} \cdot \sigma'\left(z^{(2)}\right) \cdot X$$

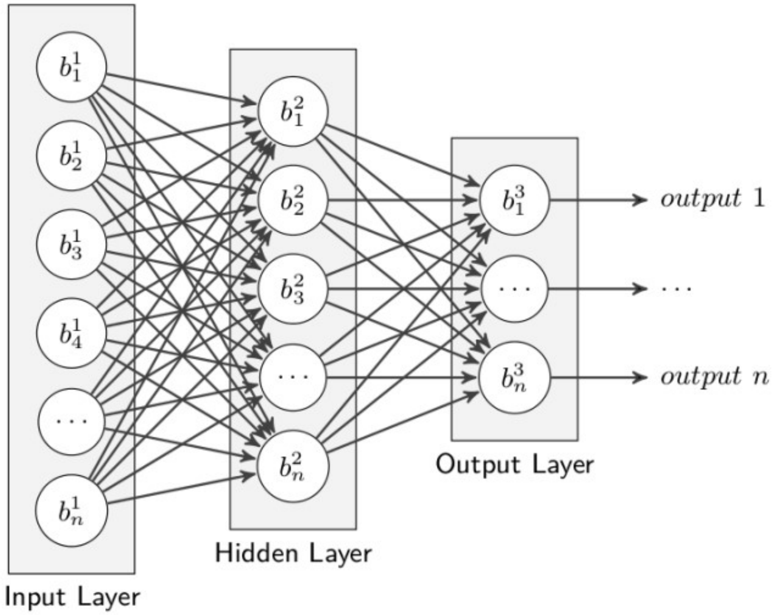
- Recall that: $\sigma'(z) = \sigma(z)(1-\sigma(z))$ (Sigmoid activation function)
- Though they appear complex, above are easy to compute!

FULLY CONNECTED NEURAL NETWORK VS. CONVOLUTED NEURAL NETWORKS

FULLY CONNECTED NETWORK

More complicated problems can be solved by connecting multiple neurons together and using more complicated activation functions.

- Organized into layers of neurons.
- Each neuron is connected to every neuron in the previous layer.
- Each layer transforms the output of the previous layer and then passes it on to the next.
- Every connection has a separate weight



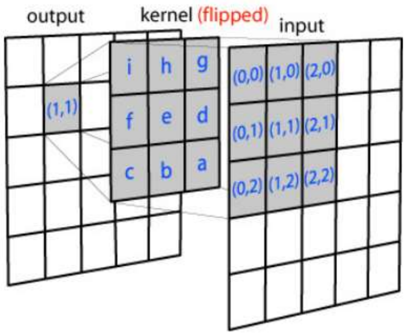
http://svail.github.io/mandarin/

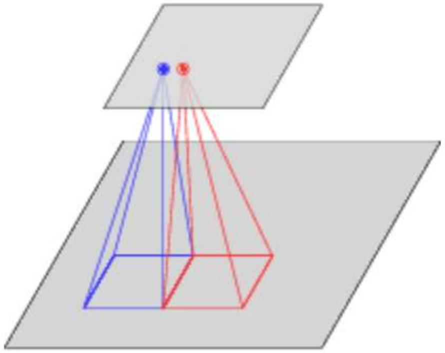
CONVOLUTIONAL NEURAL NETWORK

CONVOLUTIONAL NEURAL NETWORK (CNN)

Convolutional neural networks reduce the required computation and are good for detecting features.

- Each neuron is connected to a small set of nearby neurons in the previous layer
- The same set of weights are used for each neuron
- Ideal for spatial feature recognition, Ex: Image recognition
- Cheaper on resources due to fewer connections





http://svail.github.io/mandarin/

CNN FOR RECOGNIZING DIGITS

CNN FOR DIGIT RECOGNITION

PROC. OF THE IEEE, NOVEMBER 1998 7

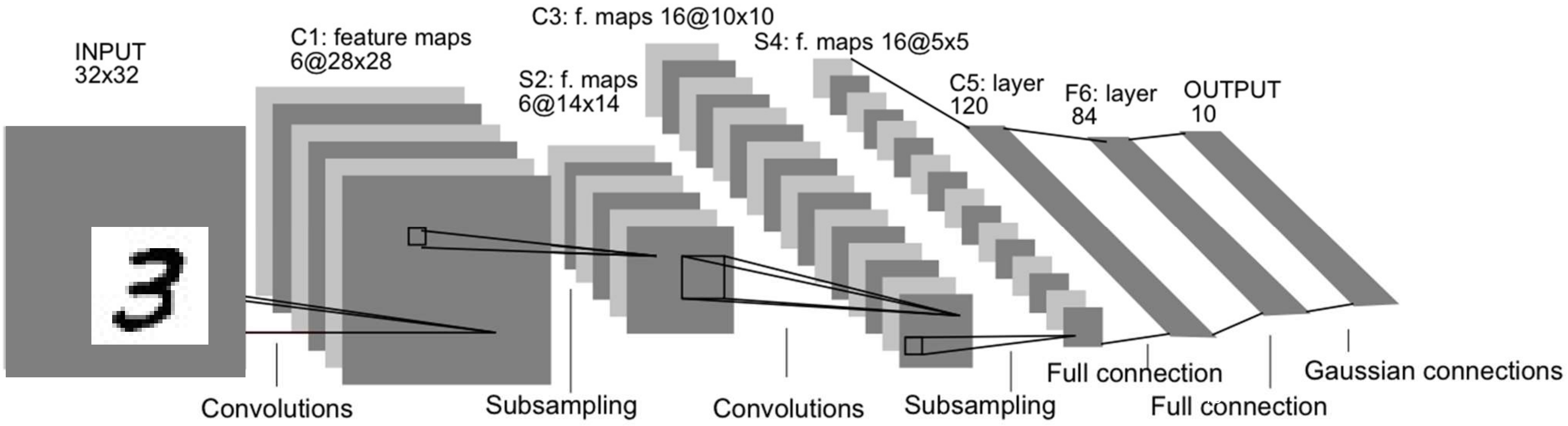
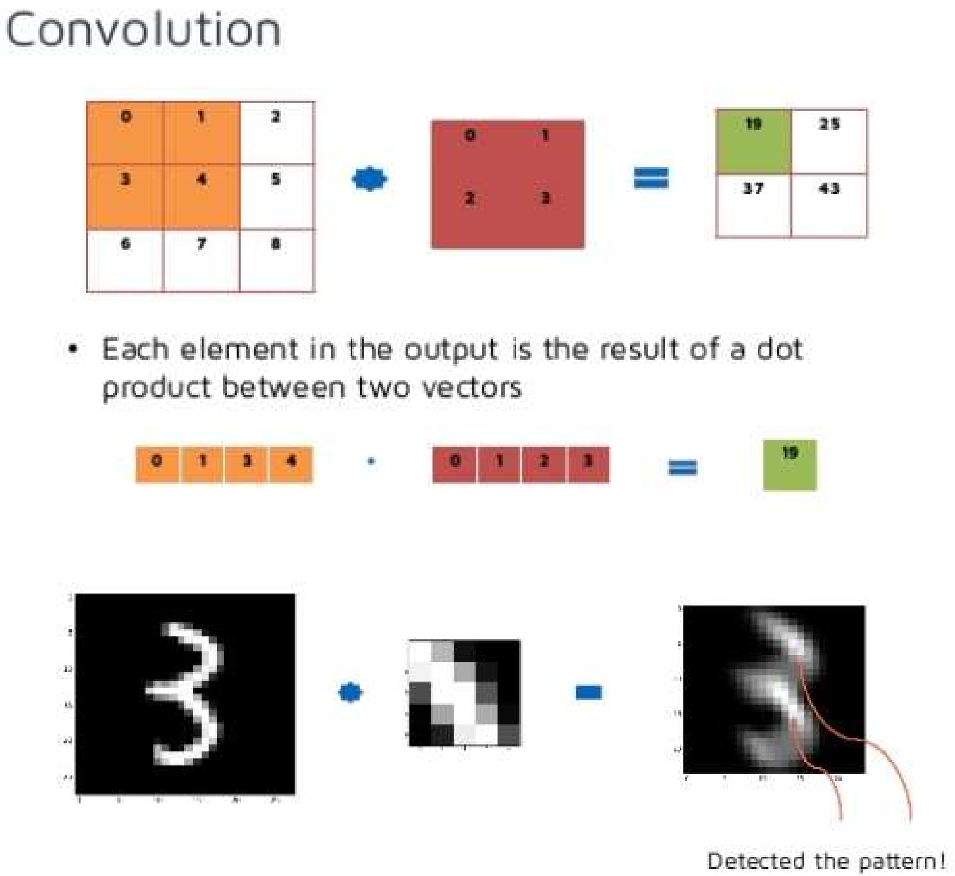


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Source: http://cs231n.github.io/

IDENTIFYING FEATURES



Convolution

$$\begin{bmatrix}0&1&2\\3&4&5\\6&7&8\end{bmatrix} * \begin{bmatrix}0&1\\2&3\end{bmatrix} = \begin{bmatrix}19&25\\37&43\end{bmatrix}$$

- Each element in the output is the result of a dot product between two vectors

$$\begin{bmatrix}0&1&3&4\end{bmatrix} \cdot \begin{bmatrix}0&1&2&3\end{bmatrix} = \begin{bmatrix}19\end{bmatrix}$$

Detected the pattern!

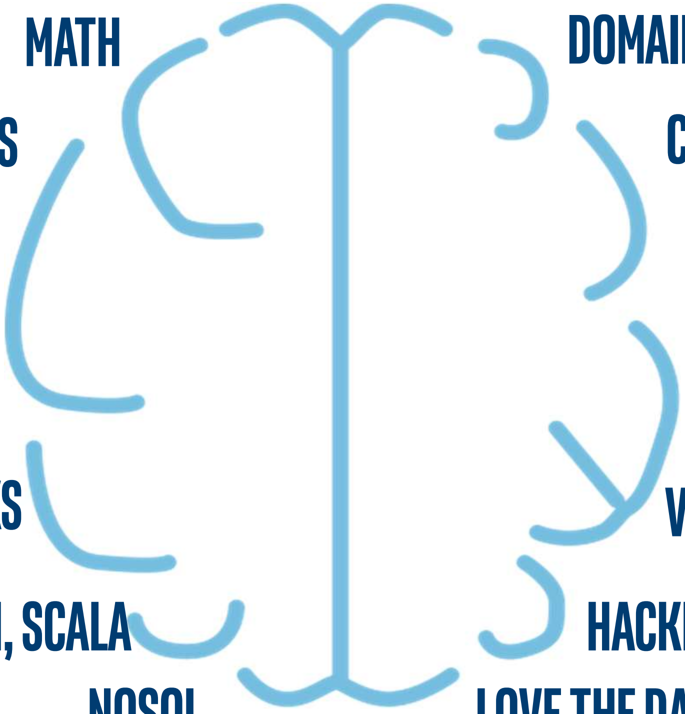
CHALLENGES

CHALLENGES

- Availability of data
  - Data Sources
  - Data Shapes
  - Amount of Data
  - Data Preprocessing
  - Labelling the data
- Reducing possibilities for overfitting and under-fitting
- Human error in data labelling
- Human Bias

HOW CAN YOU LEARN MORE?

DATA SCIENTIST SKILL SET



MATH

STATISTICS

MACHINE LEARNING

DEEP LEARNING

NEURAL NETWORKS

R, PYTHON, SCALA

NOSQL

PASSION

LOVE THE DATA

HACKER MINDSET

VISUALIZATION

STORY TELLER

ENGAGE WITH “C” LEVEL

COMMUNICATION

DOMAIN KNOWLEDGE

LEARN MORE AT THE INTEL® AI ACADEMY

For developers, students, instructors and startups

Get smarter using online tutorials, webinars, student kits and support forums

Educate others using available course materials, hands-on labs, and more

Get 4-weeks FREE access to the Intel® AI DevCloud, use your existing Intel® Xeon® Processor-based cluster, or use a public cloud service

Showcase your innovation at industry & academic events and online via the Intel AI community forum

software.intel.com/ai

RESOURCES

- Intel Developer Zone
  - https://software.intel.com
- Intel® AI Academy

  https://software.intel.com/ai-academy
- Intel® AI Student Kits

  https://software.intel.com/ai-academy/students/kits/

Q&A