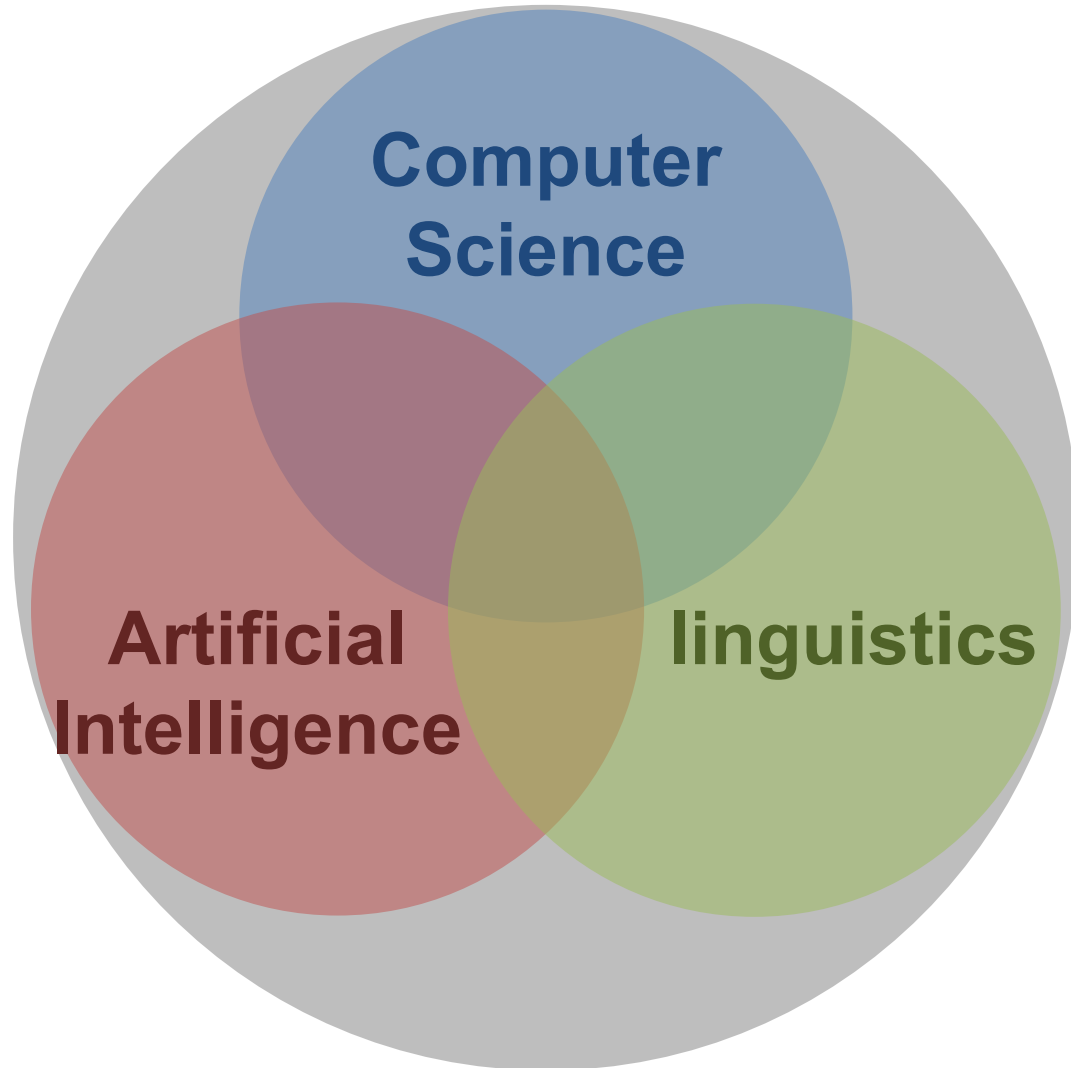# Introduction to NLP & Deep Learning

Kuan-Yu Chen (陳冠宇)
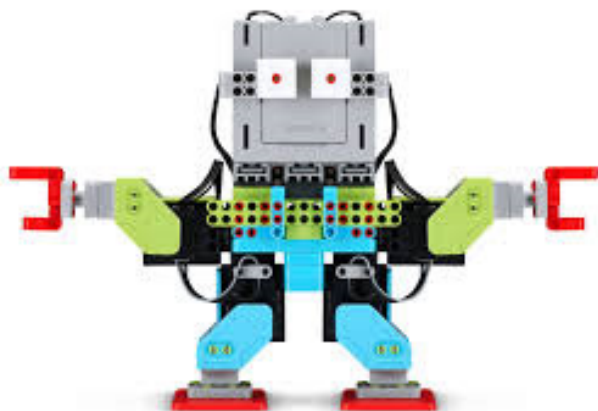
2018/03/08 @ TR-409, NTUST

Natural Language Processing

The Holy Grail of NLP

- Let machine understand human language and perform useful tasks
    - Challenges in natural-language processing frequently involve **speech recognition**, **natural-language understanding**, and **natural-language generation**

是個好天氣喔!

明天天氣怎樣?

Human Language

- A human language is a system specifically constructed to convey the speaker/writer's meaning

- A human language is mostly a **discrete**/**symbolic**/**categorical** signaling system
  - The categorical symbols of a language can be encoded as a signal for communication in several ways
    - Sound
    - Gesture
    - Writing/Images

- The large vocabulary, symbolic encoding of words creates a problem for machine learning
  - **sparsity!**

Conferences & Journals

- Conferences
  - Annual Meeting of the Association for Computational Linguistics (ACL)
  - ACM Conference on Information Knowledge Management (CIKM)
  - ACM Annual International Conference on Research and Development in Information Retrieval (SIGIR)
  - International Joint Conferences on Artificial Intelligence (IJCAI)
  - International Conference on Learning Representations (ICLR)

- Journals
  - ACM Transactions on Information Systems (TOIS)
  - IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)
  - Journal of the American Society for Information Science (JASIS)
  - Information Processing and Management (IP&M)
  - ACM Transactions on Asian Language Information Processing (TALIP)
 - Information Retrieval Journal (IRJ)

Major Topics for ACL 2018

ACL 2018 has the goal of a broad technical program. Relevant topics for the conference include, but are not limited to, the following areas (in alphabetical order):

- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Document Analysis
- Generation
- Information Extraction and Text Mining
- Linguistic Theories, Cognitive Modeling and Psycholinguistics
- Machine Learning
- Machine Translation
- Multidisciplinary
- Multilinguality
- Phonology, Morphology and Word Segmentation
- Question Answering
- Resources and Evaluation
- Sentence-level Semantics
- Sentiment Analysis and Argument Mining
- Social Media
- Summarization
- Tagging, Chunking, Syntax and Parsing
- Textual Inference and Other Areas of Semantics
- Vision, Robotics, Multimodal, Grounding and Speech
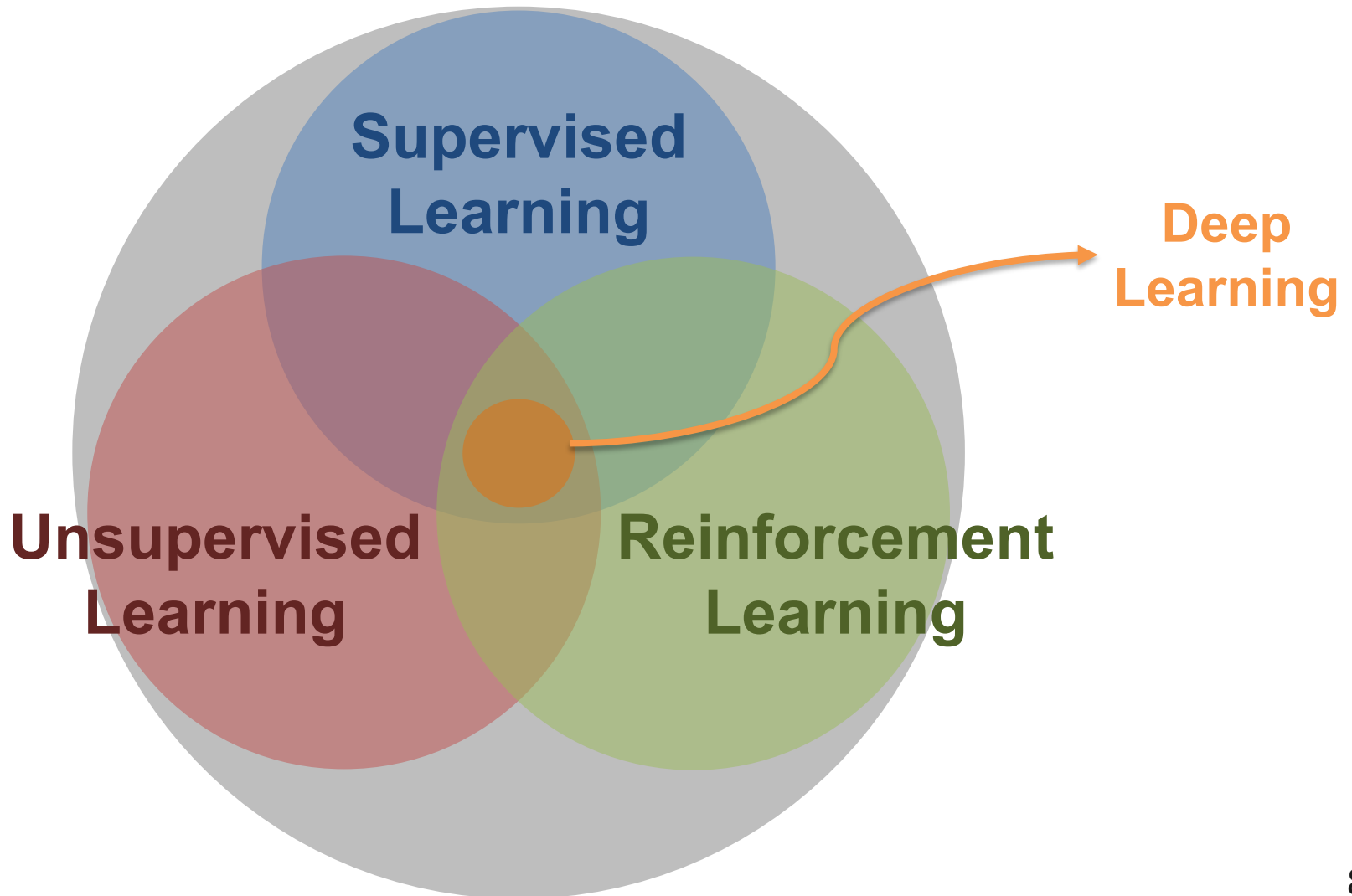- Word-level Semantics

是個好天氣喔!

明天天氣怎樣?
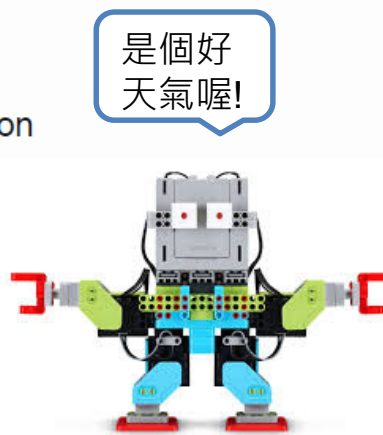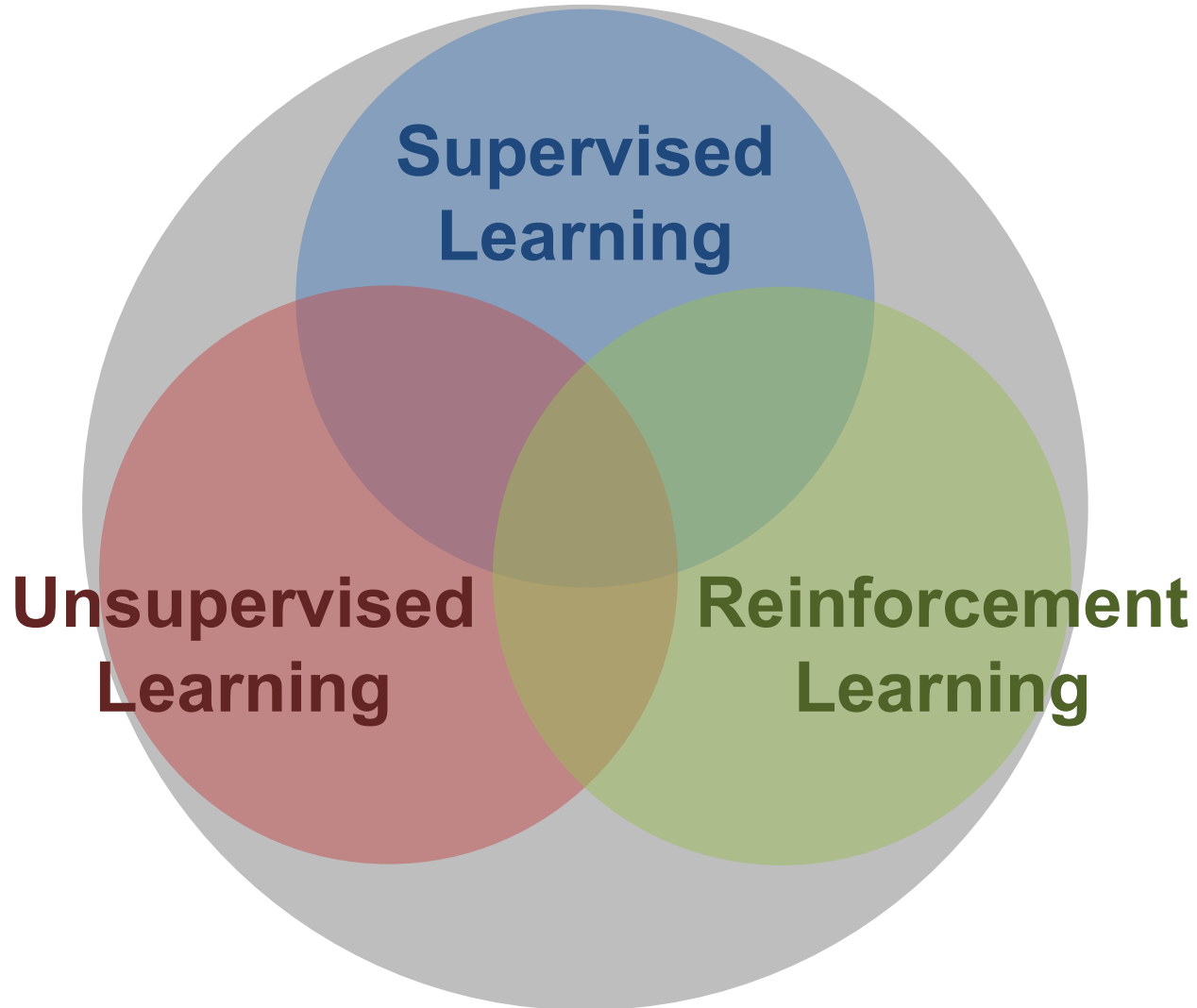
Machine Learning

Deep Learning
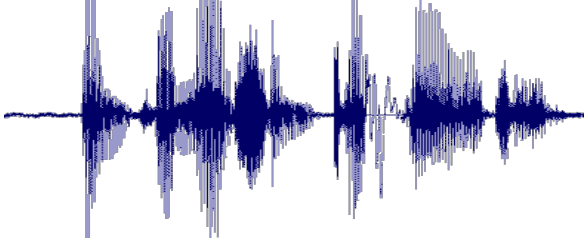
- Deep learning is a subfield of machine learning

What are we looking for?

- A **Wonderful** Function!
  - Speech Recognition

$$f\ (\quad \text{[speech waveform]}\quad) = \text{It is a nice day today}$$

  - Handwritten Recognition

$$f\ (\quad \text{It is a nice day today}\quad) = \text{It is a nice day today}$$
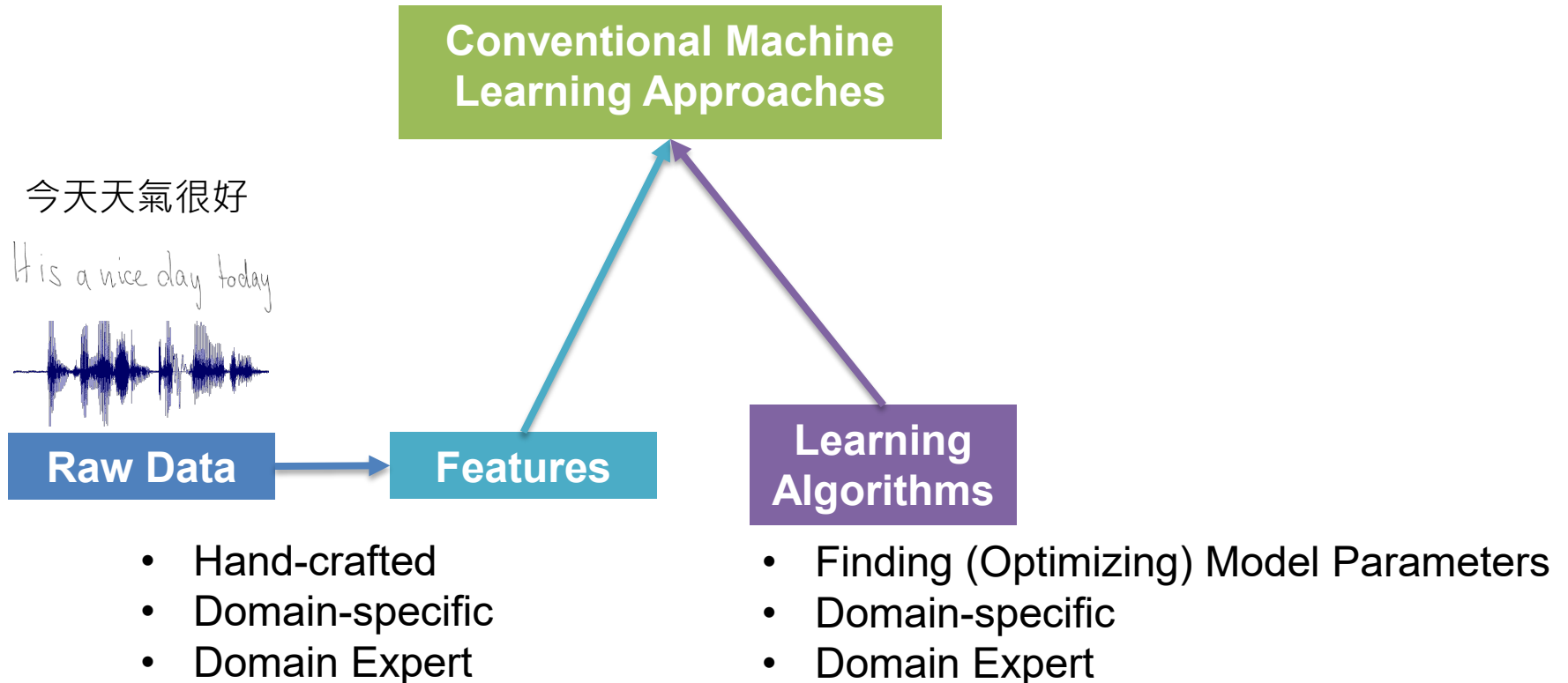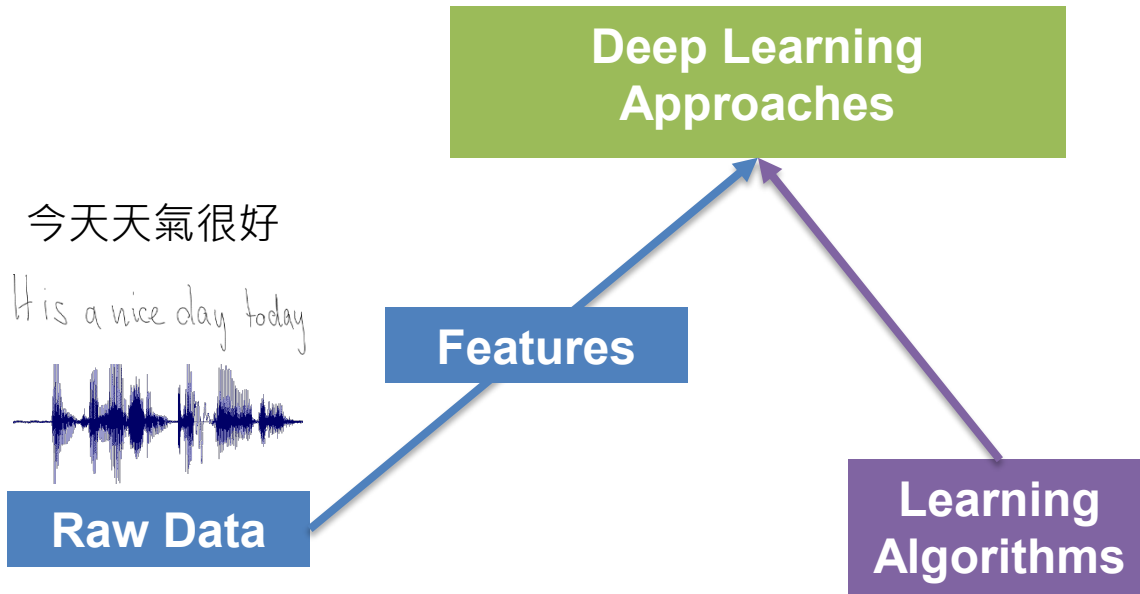
  - Machine Translation

$$f\ (\text{"今天天氣很好"}) = \text{It is a nice day today}$$

Shallow Learning & Deep Learning – 1

Conventional Machine Learning Approaches

今天天氣很好

It is a nice day today

Raw Data → **Features**

Learning Algorithms

- Hand-crafted
- Domain-specific
- Domain Expert

- Finding (Optimizing) Model Parameters
- Domain-specific
- Domain Expert

Shallow Learning & Deep Learning – 2

今天天氣很好

It is a nice day today

Raw Data

Features

**Deep Learning Approaches**

Learning Algorithms

- ~~Hand-crafted~~
- ~~Domain-specific~~
- ~~Domain Expert~~
- Features are learned by machine automatically

- Finding (Optimizing) Model Parameters
- ~~Domain-specific~~
- ~~Domain Expert~~

"Deep Learning" usually refers to neural network-based framework

The History

- Yoshua Bengio (UMONTRAL) @Deep Learning, 2015
  - Deep learning has a long and rich history. It only appears to be new, because it was relatively unpopular for several years preceding its current popularity, and because it has gone through many different names.
    - 1940s~1960s: cybernetics (McCulloch-Pitts Neuron, 1943; Perceptron, 1958)
    - 1980s~1990s: connectionism (Neocognitron, 1980)
    - 2006~: deep learning
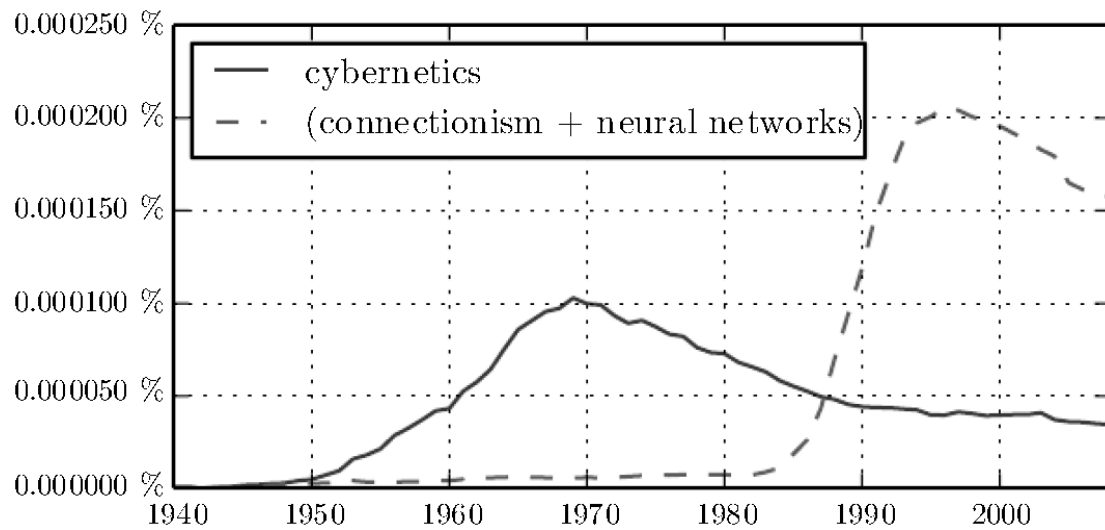
NEURAL NETWORK BASED LANGUAGE MODELS FOR HIGHLY INFLECTIVE LANGUAGES

Tomáš Mikolov, Jiří Kopecký, Lukáš Burget, Ondřej Glembek and Jan "Honza" Černocký

Speech@FIT, Faculty of Information Technology, Brno University of Technology, Czech Republic

{imikolov|kopecky|burget|glembek|cernocky}@fit.vutbr.cz

4.2. Real-time factor

Many promising language modeling techniques are not widely used because of their high computational requirements. Our implementation of training phase is not optimized in any way, except vocabulary size reduction by merging rare words. Training times are thus quite high: still, we are able to train full NN LM model in less than 2 weeks, the biggest 60 150 4 network was trained in approximately 4 weeks. More important factor for practical use is time needed to re-score N-best lists - see Table 4.

To obtain reasonable re-scoring times, two level cache was implemented. Since N-best variants of utterances share large amount of n-grams that need to be estimated, it is useful to remember already computed n-gram probabilities. The second level of cache is storing computed probability distributions - this cache is more memory hungry and so the number of stored entries is much smaller. With our cache model, we obtain cache hit 68% when rescoring 10-best lists and 92% when re-scoring 3000 best list.

Another interesting fact is that neural networks may take less disk space than conventional backoff LM (Table 5).
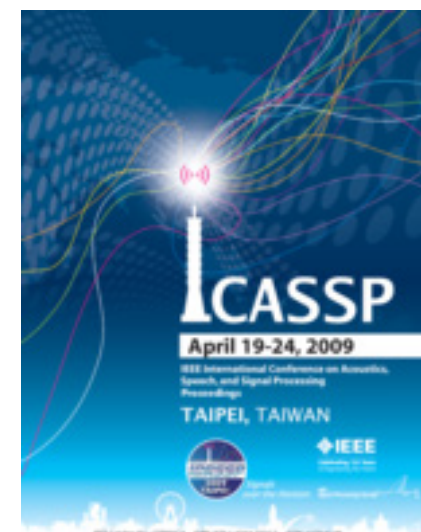
Reasons for Exploring Deep Learning

- In ~2010 deep learning techniques started outperforming other machine learning techniques
  - First in speech (2010) and vision (2012), then NLP (?)
    - Large amounts of training data favor deep learning
    - Faster machines and multicore CPU/GPUs favor Deep Learning
    - New models, algorithms, ideas
      - ✓ Better, more flexible learning of intermediate representations
      - ✓ Effective end-to-end joint system learning
      - ✓ Effective learning methods for using contexts and transferring between tasks
      - ✓ Better regularization and optimization methods

Breakthrough in Speech Recognition

132   7   Training and Decoding Speedup

Table 7.5 Model size, computation time, and word error rate (WER) with and without sparseness constraints on the SWB dataset

Acoustic model	# nonzero params	% nonzero params	Hub5'00 FSH	RT03S SWB (%)
GMM, BMMI	29.4M	–	23.6 %	27.4
DNN, CE	45.1M	Fully connected	16.4 %	18.6
	31.1M	69 %	16.2 v	18.5
	23.6M	52 %	16.1 %	18.5
	15.2M	34 %	16.1 %	18.4
	11.0M	24 %	16.2 %	18.5
	8.6M	19 %	16.4 %	18.7
	6.6M	5 %	16.5 %	18.7

The fully connected DNN contains 7 hidden layers each with 2,048 neurons (Summarized from Yu et al. [25])
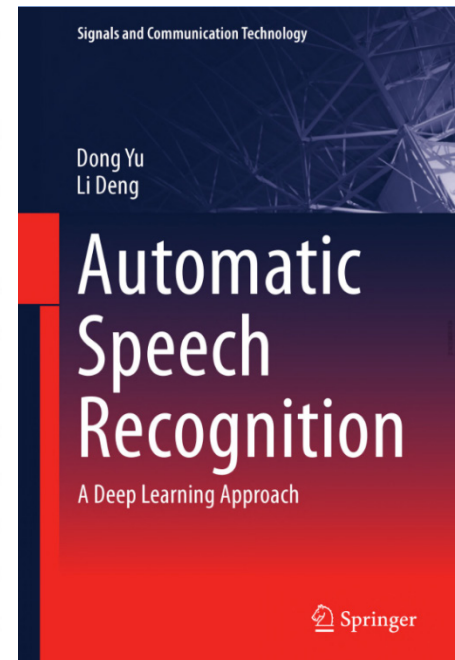
Context-Dependent Pre-Trained Deep Neural Networks for Large ...
ieeexplore.ieee.org/document/5740583/ ▼ 翻譯這個網頁
由 GE Dahl 著作 - 2012 - 被引用 1750 次 - 相關文章
2011年4月5日 - Abstract: We propose a novel context-dependent (CD) model for large-vocabulary speech recognition (LVSR) that leverages recent advances in using deep belief networks for phone recognition. We describe a pre-trained deep neural network hidden Markov model (DNN-HMM) hybrid architecture that ...

Manuscript received September 08, 2010; revised January 04, 2011, March 13, 2011; accepted March 14, 2011. Date of publication April 05, 2011; date of current version December 16, 2011. G. E. Dahl contributed to this work as an intern with Microsoft Research, Redmond, WA. This manuscript greatly extends the work presented at ICASSP 2011 [1]. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nelson Morgan.
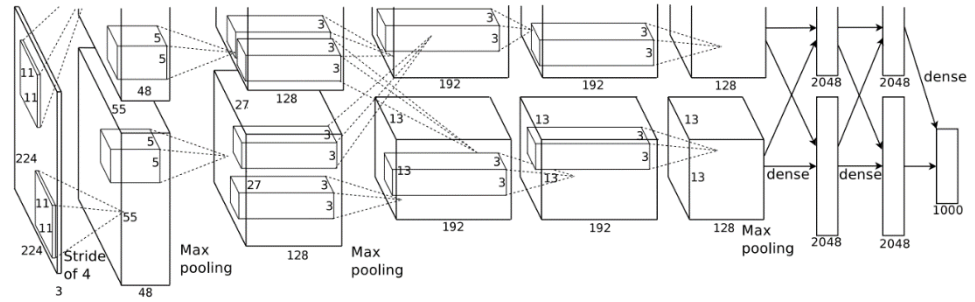
Signals and Communication Technology
Dong Yu
Li Deng
Automatic Speech Recognition
A Deep Learning Approach
Springer

Geoffrey Hinton   Li Deng   Dong Yu

G. E. Dahl, D. Yu, L. Deng and A. Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30-42, Jan. 2012.
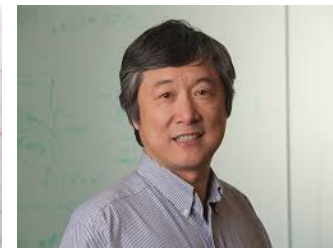
Breakthrough in Computer Vision

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

ImageNet classification with deep convolutional neural networks
https://dl.acm.org/citation.cfm?id=2999257 ▾
由 A Krizhevsky 著作 - 2012 - 被引用 20340 次 - 相關文章
2012年12月3日 - 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 ...
Abstract · Authors · References · Cited By

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12), F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.), Vol. 1. Curran Associates Inc., USA, 1097-1105.
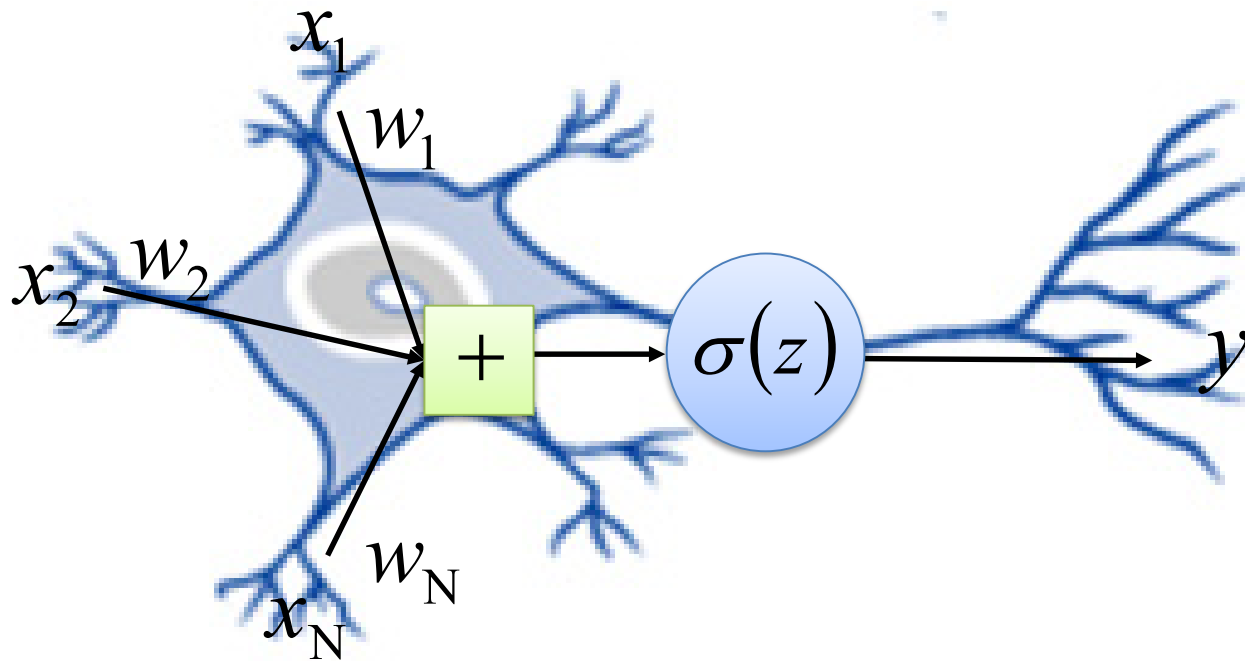
Evolution

- 1960s: Perceptron (single layer neural network)
- 1969: Perceptron has limitation
- 1980s: Multi-layer perceptron
- 1986: Backpropagation
- 1989: 1 hidden layer is "good enough", why deep?
- 2006: RBM initialization (breakthrough)
- 2009: GPU
- 2010: Breakthrough in Speech Recognition (Dahl et al., 2010)
- 2012: Breakthrough in ImageNet (Krizhevsky et al. 2012)
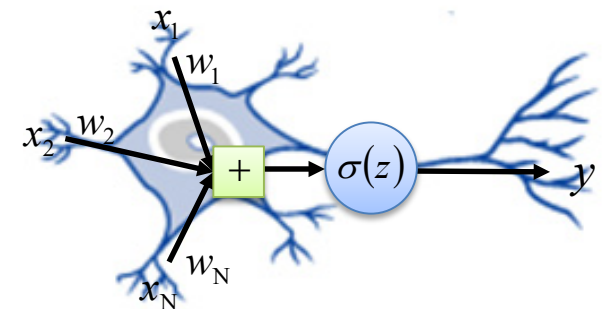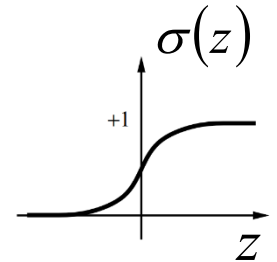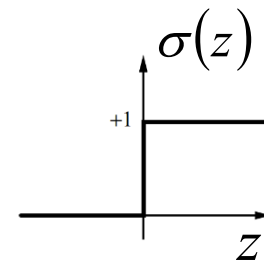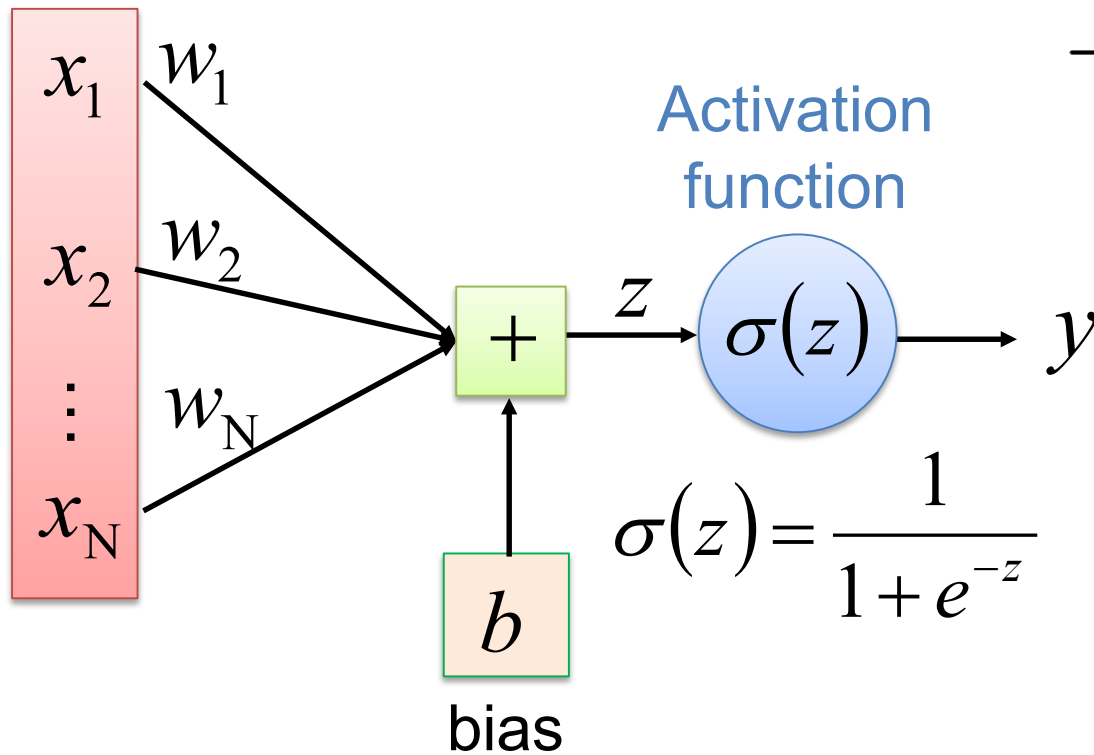- 2015: "superhuman" results in Image and Speech Recognition

Neuron

Bipolar   Unipolar   Multipolar

Dendrites (樹突)

Receptive (接收區)

Trigger (觸發區)

Axon (軸突)　**Conducting (傳導區)**

Output (輸出區)

Synapse (突觸)
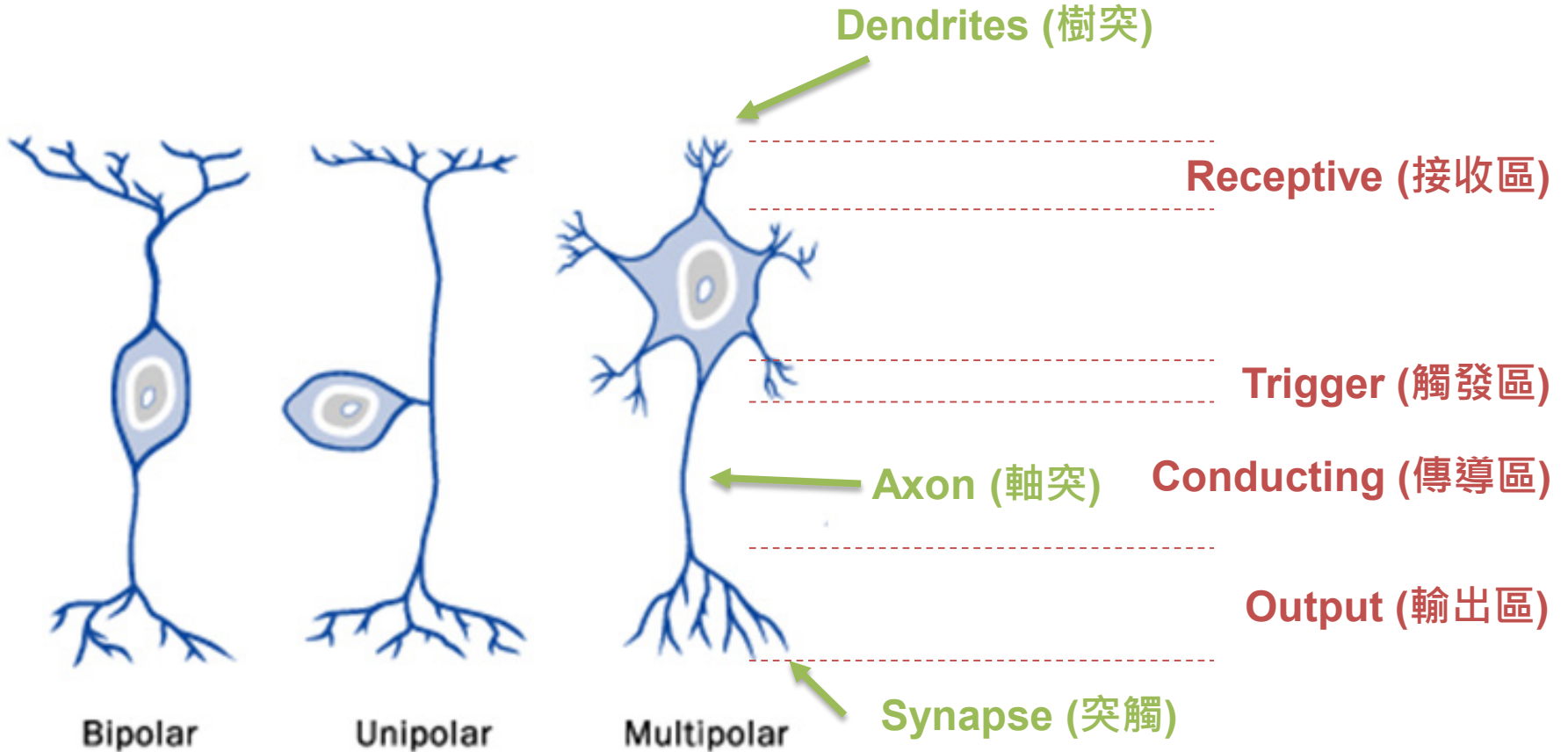
Handcrafted Neuron

Single Neuron (McCulloch-Pitts Neuron)

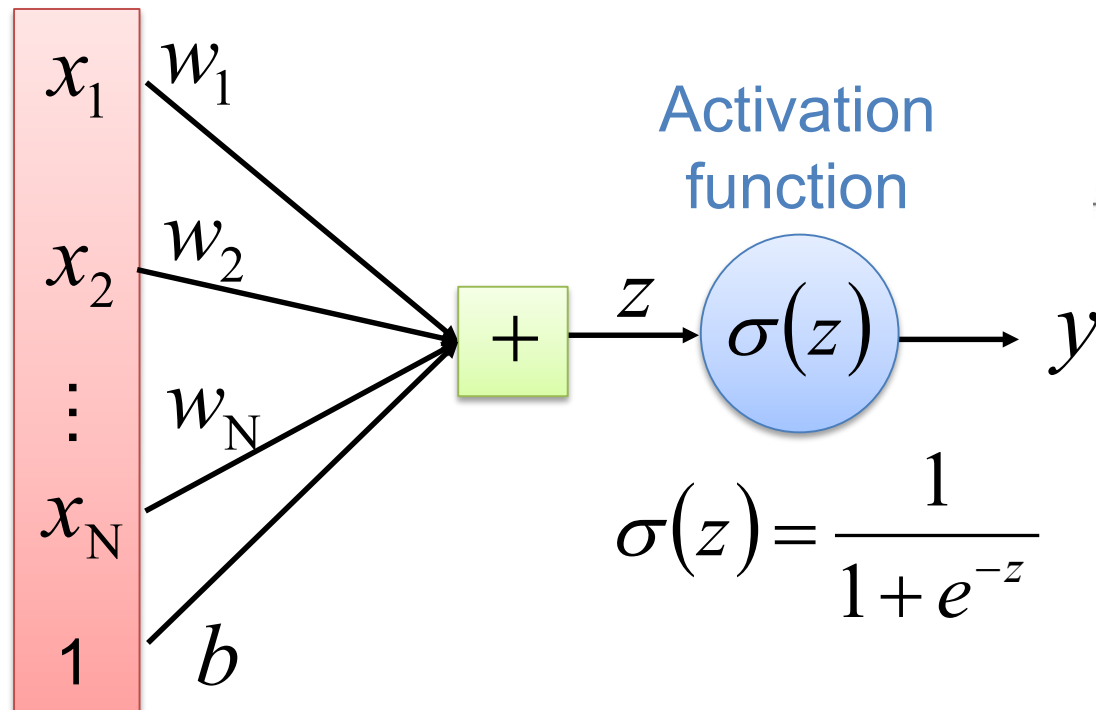- This vastly simplified model of real neurons is also known as a **Threshold Logic Unit**

$$y = \sigma(w^T x + b)$$

Input

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Activation function

x_1 w_1

x_2 w_2

$\vdots$ w_N

x_N

$+$ $\xrightarrow{z}$ $\sigma(z)$ $\longrightarrow$ y

b

bias

The Bias – 1

- The "bias" can be views as an extra feature
  - It is a shift!

$$y = \sigma(w^{\mathrm{T}}x + b) = \sigma(\tilde{w}^{\mathrm{T}}\tilde{x})$$

Input

x_1 w_1

x_2 w_2

$\vdots$

x_N w_N

1 b

Activation function

$+$ z $\sigma(z)$ y

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

$\sigma(z)$

1

0.5

0 z
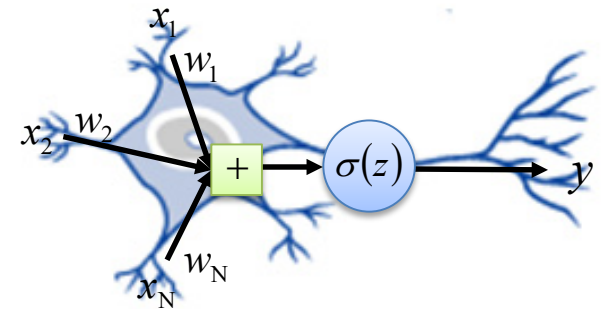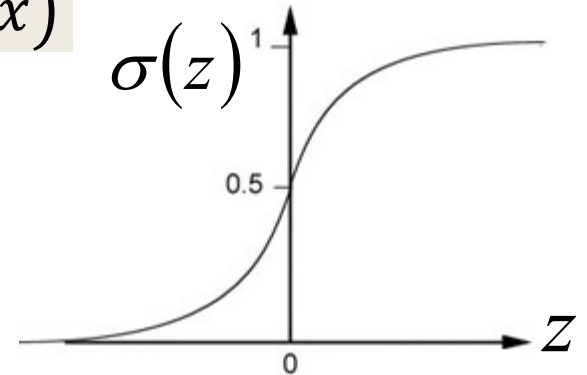
The Bias – 2

- A naïve explain
  - Without bias
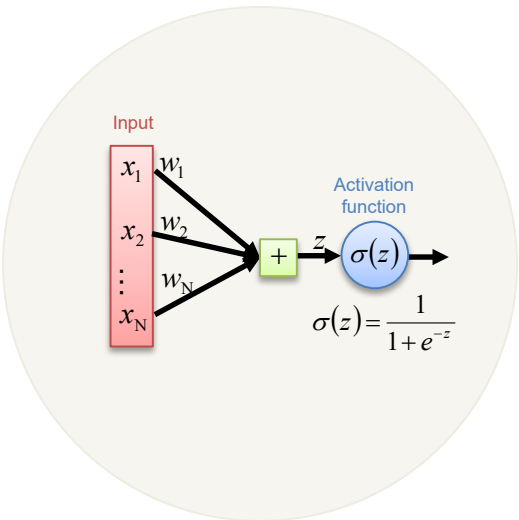
0.9

1.9

2.9

  - With bias=0.1
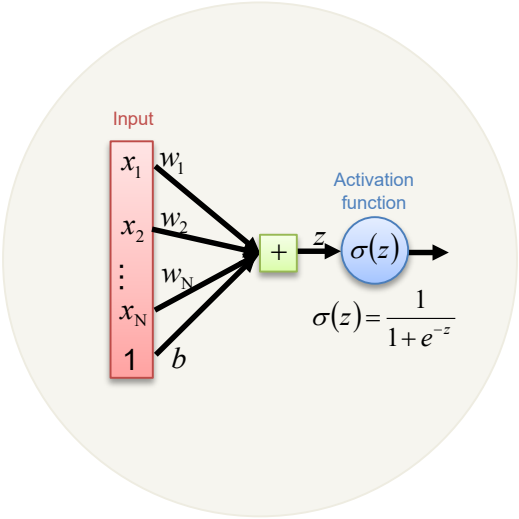
1

2

3

Single Neuron – Binary Classification
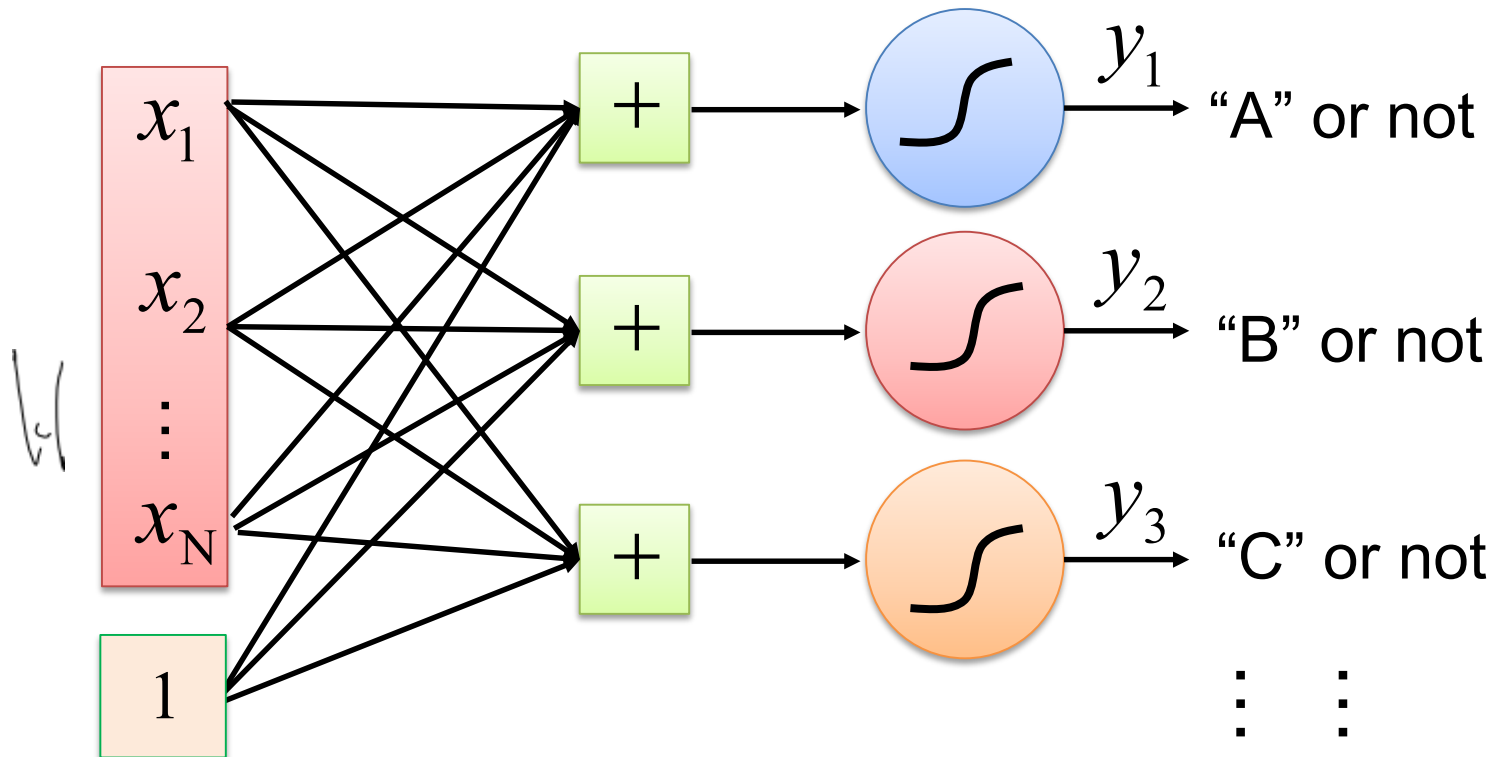
- A single neuron can only handle binary classification
  - Single output

$$\begin{cases} "\mathrm{H}", & if \ y \geq 0.5 \\ not \ "\mathrm{H}", & if \ y < 0.5 \end{cases}$$

Input

x_1 w_1

x_2 w_2

$\vdots$

x_N w_N

$+$ z

Activation function

$\sigma(z)$ y
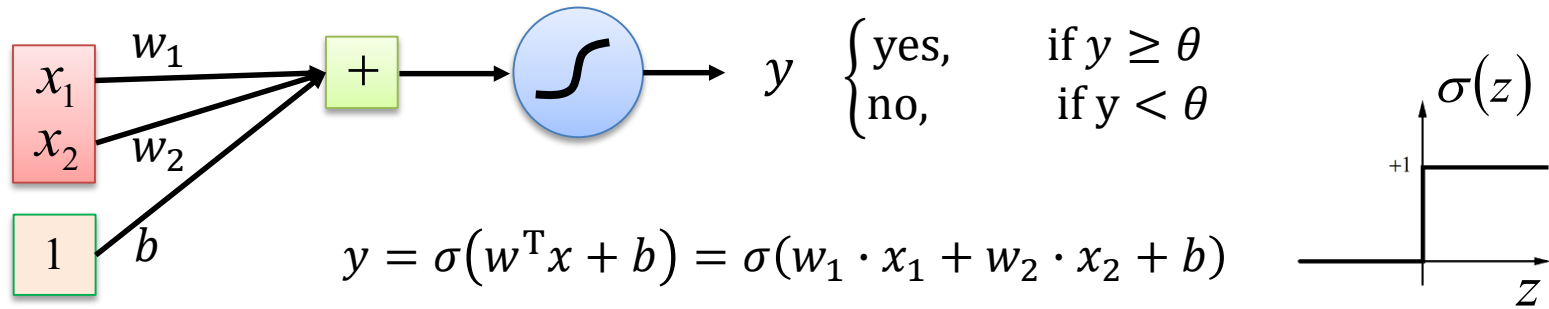
b

bias

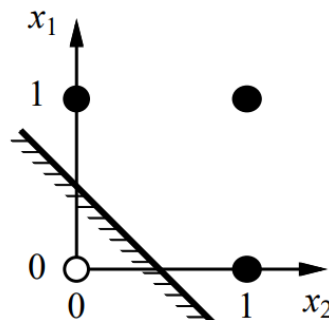$\sigma(z)$
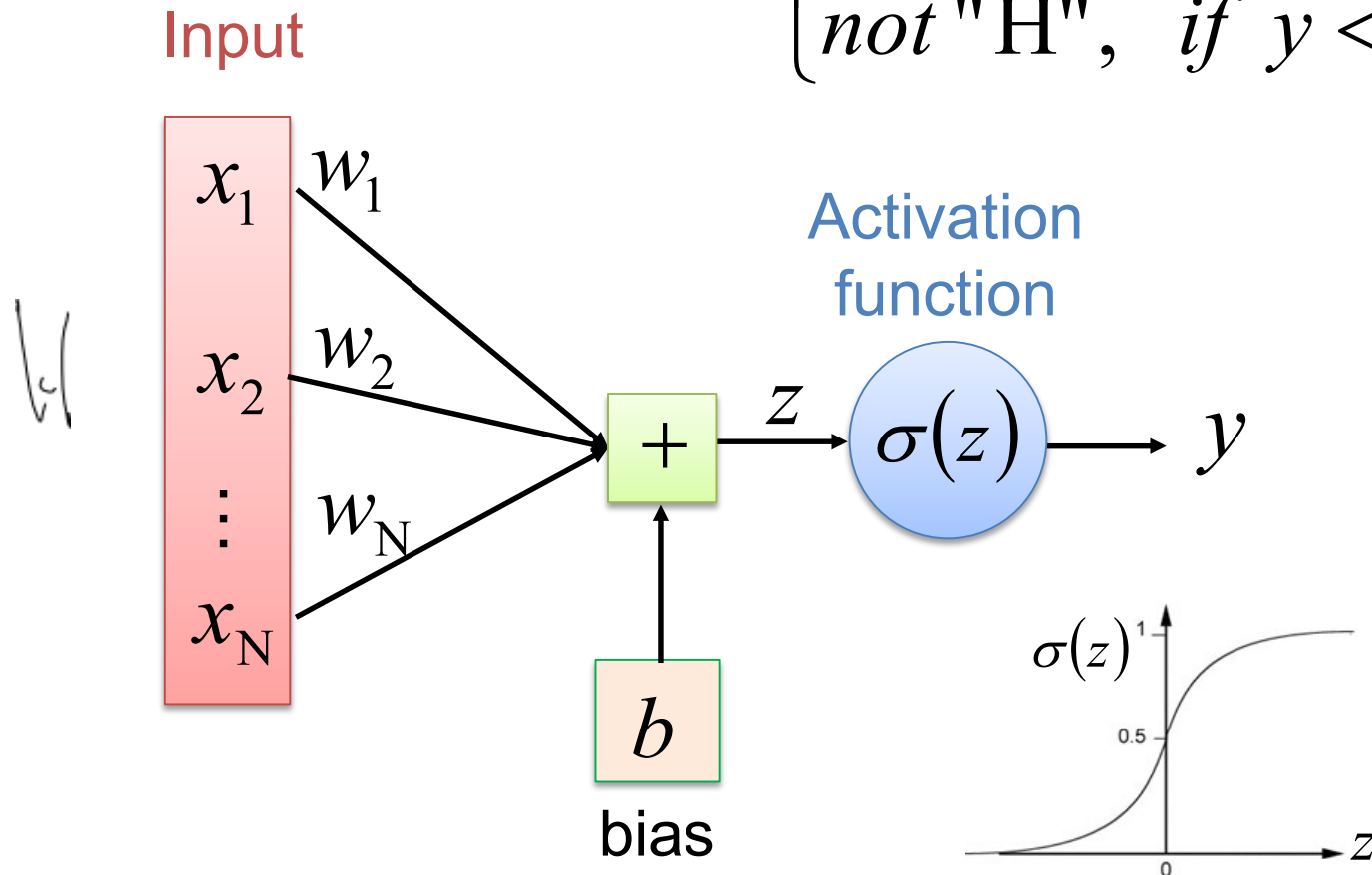
Multiple Neurons – Multi-class Classification

- A layer of neurons can handle multiple possible output, and the result depends on the max one
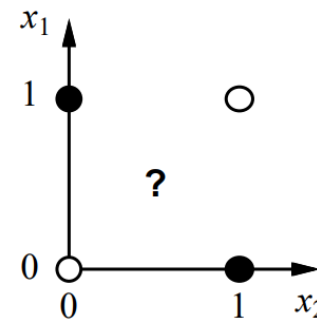 - Perceptron!

The Limitation – Linear Separator!

$$y \quad \begin{cases} \text{yes,} & \text{if } y \geq \theta \\ \text{no,} & \text{if } y < \theta \end{cases}$$

$$y = \sigma(w^{\mathrm{T}}x + b) = \sigma(w_1 \cdot x_1 + w_2 \cdot x_2 + b)$$

$\sigma(z)$

$+1$
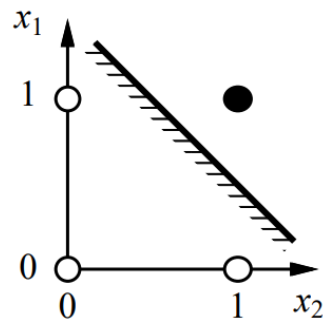
(a) x_1 **and** x_2

(b) x_1 **or** x_2

(c) x_1 **xor** x_2

$(0,0): \sigma(0.3 \times 0 + 0.3 \times 0 - 0.5) < 0$

$(0,1): \sigma(0.3 \times 0 + 0.3 \times 1 - 0.5) < 0$

$(1,0): \sigma(0.3 \times 1 + 0.3 \times 0 - 0.5) < 0$

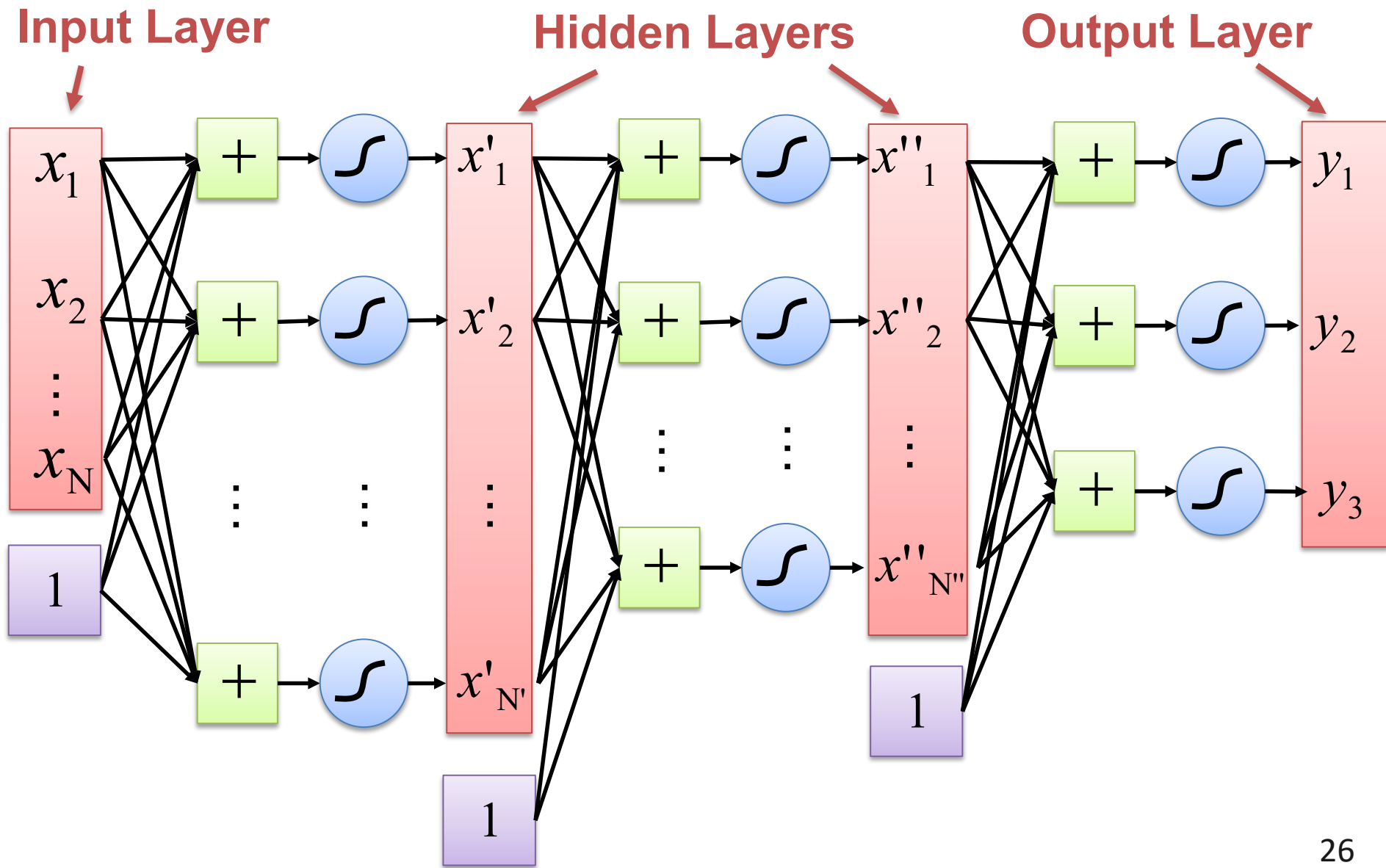$(1,1): \sigma(0.3 \times 1 + 0.3 \times 1 - 0.5) \geq 0$

$(0,0): \sigma(0.3 \times 0 + 0.3 \times 0 - 0.2) < 0$
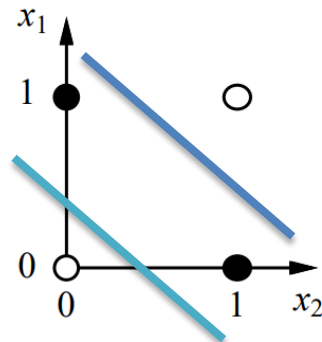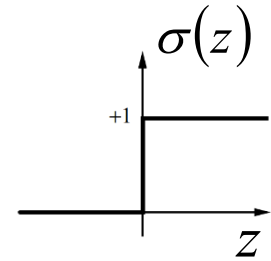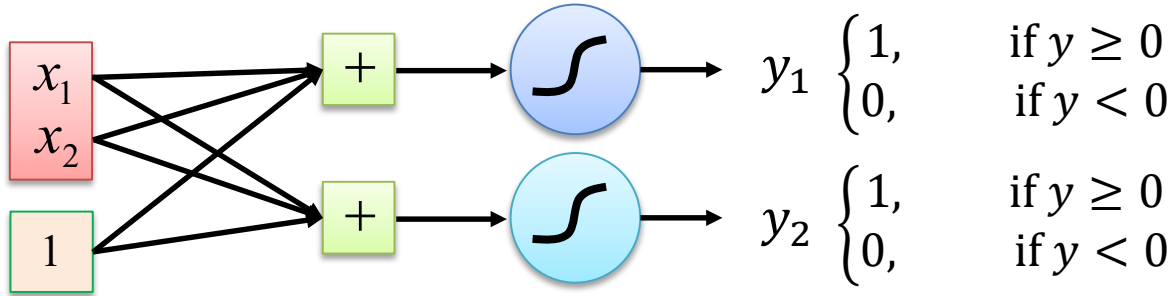
$(0,1): \sigma(0.3 \times 0 + 0.3 \times 1 - 0.2) \geq 0$

$(1,0): \sigma(0.3 \times 1 + 0.3 \times 0 - 0.2) \geq 0$

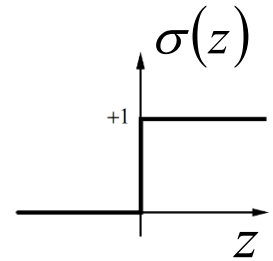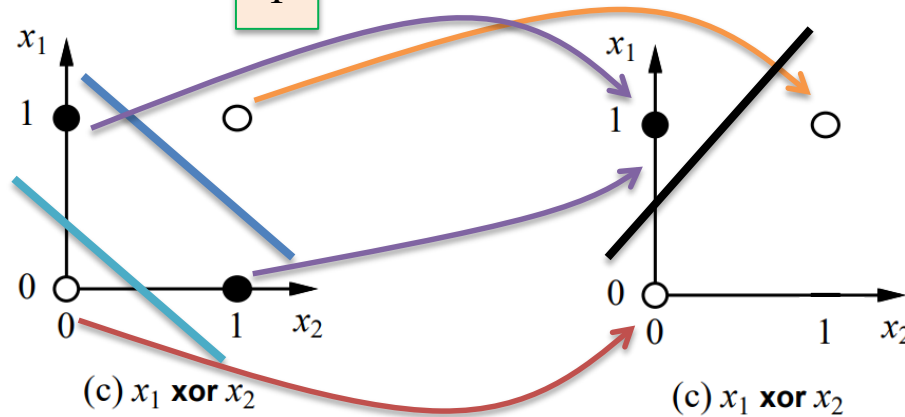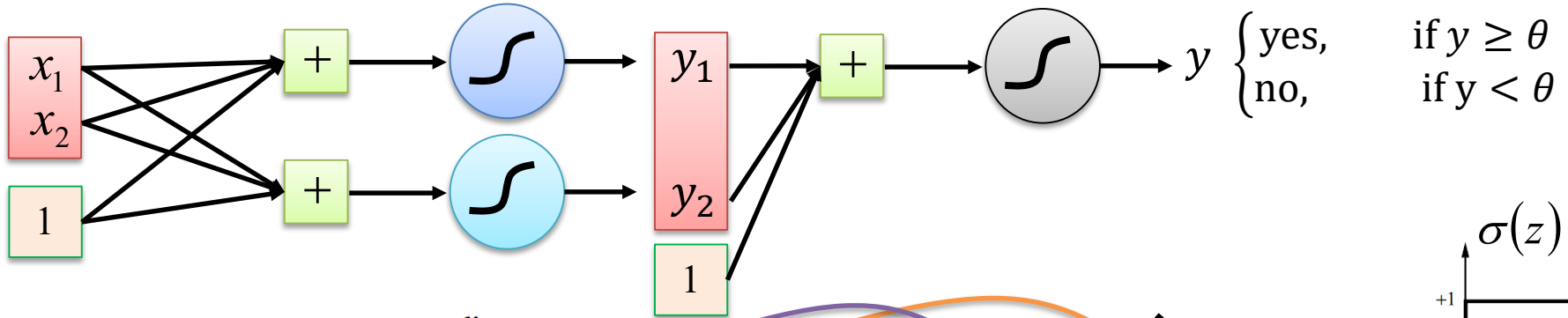$(1,1): \sigma(0.3 \times 1 + 0.3 \times 1 - 0.2) \geq 0$

Multi-Layer Perceptron

Input Layer

Hidden Layers

Output Layer

$$y_1 \begin{cases} 1, & \text{if } y \geq 0 \\ 0, & \text{if } y < 0 \end{cases}$$

$$y_2 \begin{cases} 1, & \text{if } y \geq 0 \\ 0, & \text{if } y < 0 \end{cases}$$

$$y = \sigma\left(w^{\mathrm{T}} x + b\right) = \sigma(w_1 \cdot x_1 + w_2 \cdot x_2 + b)$$

(c) x_1 **xor** x_2

$(0,0): \sigma(0.3 \times 0 + 0.3 \times 0 - 0.5) < 0$ $(0,0): \sigma(0.3 \times 0 + 0.3 \times 0 - 0.2) < 0$ ⟹ (0,0)

$(0,1): \sigma(0.3 \times 0 + 0.3 \times 1 - 0.5) < 0$ $(0,1): \sigma(0.3 \times 0 + 0.3 \times 1 - 0.2) \geq 0$ ⟹ (0,1)

$(1,0): \sigma(0.3 \times 1 + 0.3 \times 0 - 0.5) < 0$ $(1,0): \sigma(0.3 \times 1 + 0.3 \times 0 - 0.2) \geq 0$ ⟹ (0,1)

$(1,1): \sigma(0.3 \times 1 + 0.3 \times 1 - 0.5) \geq 0$ $(1,1): \sigma(0.3 \times 1 + 0.3 \times 1 - 0.2) \geq 0$ ⟹ (1,1)

MLP − 2

$$y \quad \begin{cases} \text{yes,} & \text{if } y \geq \theta \\ \text{no,} & \text{if } y < \theta \end{cases}$$

(c) x_1 **xor** x_2

(c) x_1 **xor** x_2

$(0,0): \sigma(0.3 \times 0 + 0.3 \times 0 - 0.5) < 0$ $(0,0): \sigma(0.3 \times 0 + 0.3 \times 0 - 0.2) < 0$ ⟹ $(0,0)$

$(0,1): \sigma(0.3 \times 0 + 0.3 \times 1 - 0.5) < 0$ $(0,1): \sigma(0.3 \times 0 + 0.3 \times 1 - 0.2) \geq 0$ ⟹ $(0,1)$

$(1,0): \sigma(0.3 \times 1 + 0.3 \times 0 - 0.5) < 0$ $(1,0): \sigma(0.3 \times 1 + 0.3 \times 0 - 0.2) \geq 0$ ⟹ $(0,1)$

$(1,1): \sigma(0.3 \times 1 + 0.3 \times 1 - 0.5) \geq 0$ $(1,1): \sigma(0.3 \times 1 + 0.3 \times 1 - 0.2) \geq 0$ ⟹ $(1,1)$

28

MLP – 3

$$y \begin{cases} \text{yes,} & \text{if } y \geq \theta \\ \text{no,} & \text{if } y < \theta \end{cases}$$
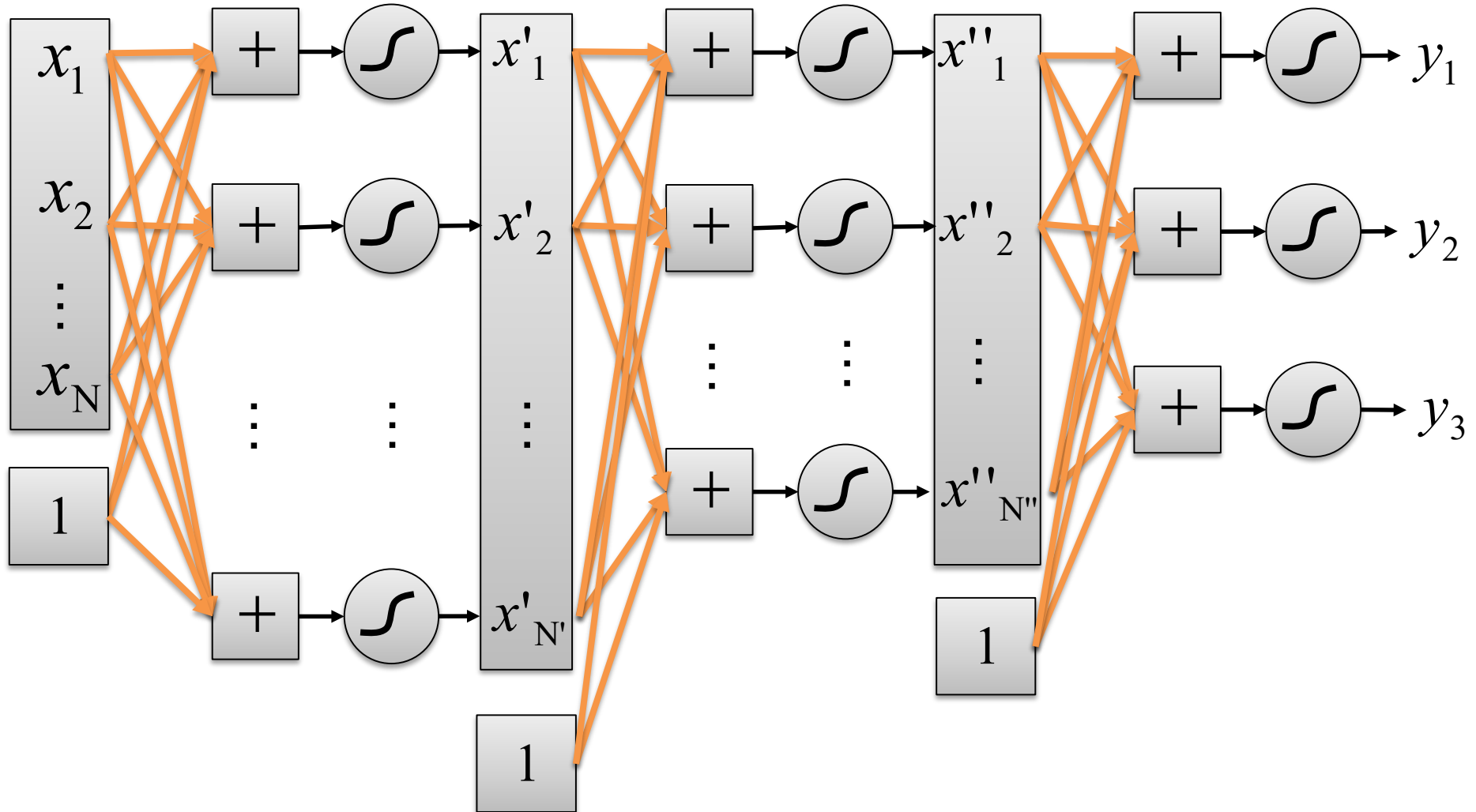
$\sigma(z)$

(c) x_1 **xor** x_2

(c) x_1 **xor** x_2

$(0,0): \sigma(0.3 \times 0 + 0.3 \times 0 - 0.5) < 0$ $(0,0): \sigma(0.3 \times 0 + 0.3 \times 0 - 0.2) < 0$ ⟹ (0,0)

$(0,1): \sigma(0.3 \times 0 + 0.3 \times 1 - 0.5) < 0$ $(0,1): \sigma(0.3 \times 0 + 0.3 \times 1 - 0.2) \geq 0$ ⟹ (0,1)

$(1,0): \sigma(0.3 \times 1 + 0.3 \times 0 - 0.5) < 0$ $(1,0): \sigma(0.3 \times 1 + 0.3 \times 0 - 0.2) \geq 0$ ⟹ (0,1)

$(1,1): \sigma(0.3 \times 1 + 0.3 \times 1 - 0.5) \geq 0$ $(1,1): \sigma(0.3 \times 1 + 0.3 \times 1 - 0.2) \geq 0$ ⟹ (1,1)
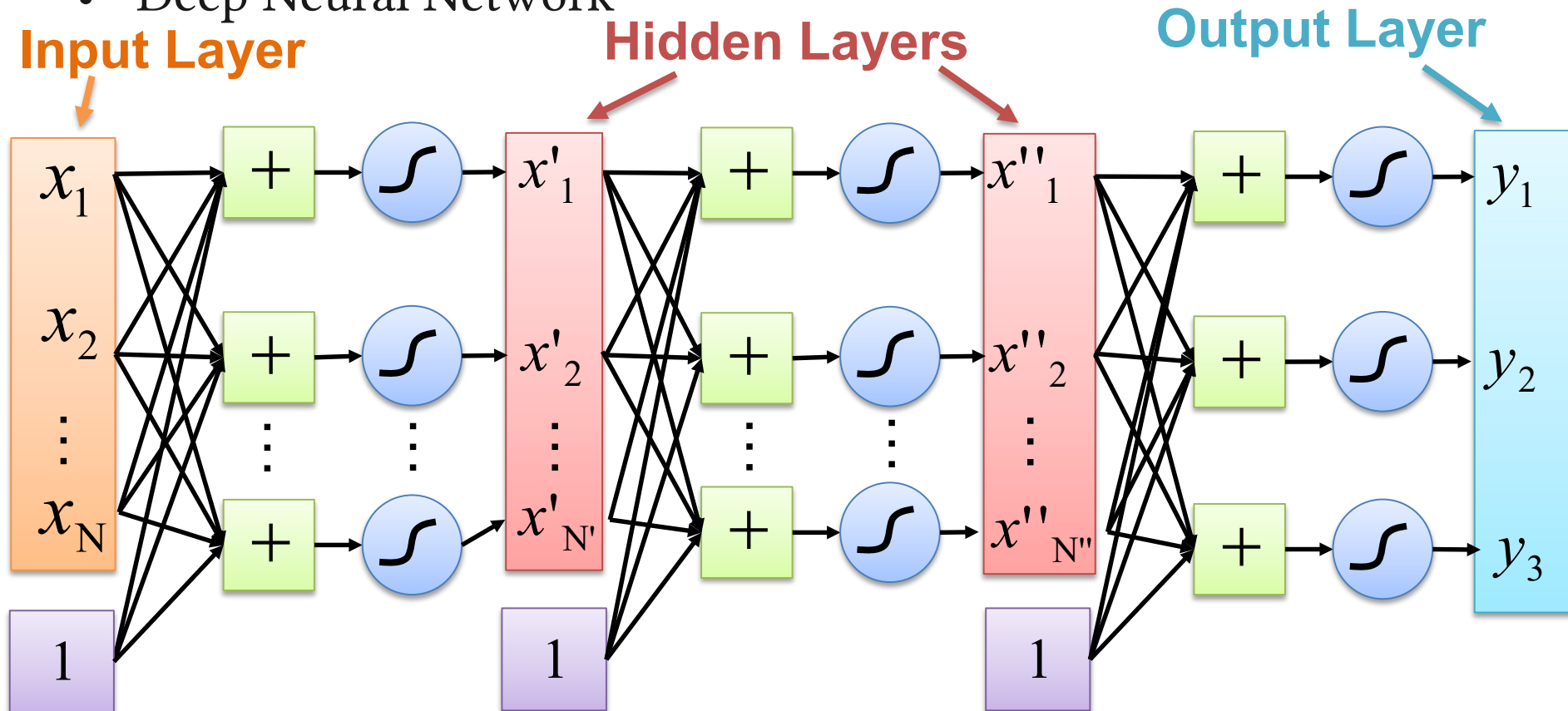
Feature Learning!

The Model Parameters

Deep Neural Network – 1

- Multilayer Perceptron
- Fully-connected Feed-forward Neural Network
- Vanilla Neural Network
- Deep Neural Network

Deep Neural Network – 2

- Deep Neural Networks
  - G. Hinton (UTORONTO & Google) @IEEE Signal Processing Magazine 2012
    - **more than one layer of hidden units**

  - D. Yu (Microsoft Research) @Automatic Speech Recognition 2015
    - The term deep neural network was originally introduced to mean **multilayer perceptron with many hidden layers**, but was later extended to mean any neural network with a deep structure

  - Rich Caruana (Microsoft Research) @ASRU2015
    - **three hidden layers**

Define Your Own Deep!

Questions?

kychen@mail.ntust.edu.tw