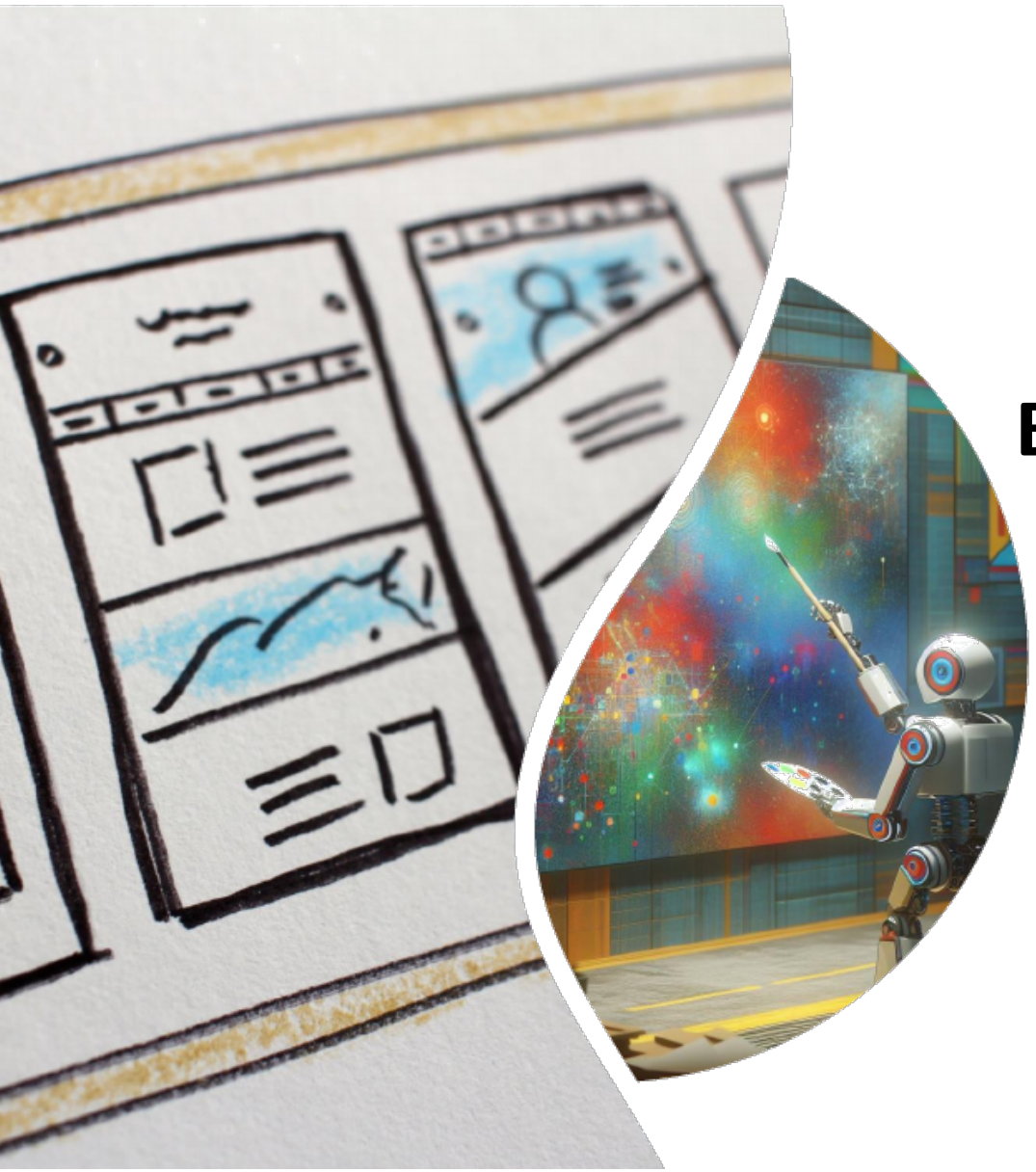


DGIN 5201
Digital Transformation
Lecture 15

Emerging Technology 1: AI and Deep Learning

Time and date:
13:05–14:25, 4-Mar-2025
Location: LSC C236

Image: DALL-E. Bing Image Creator. Generated by AI

Previous Lecture

Tuesday lecture last week:

- Guest Speaker: Tapajyoti Das (Tukan)
  - Project ideas in the startup area
- Project discussion
- Technical requirements of project discussion

Thursday lecture and Friday labs last week:

- Team meetings 1
- Discussion about team ideas, project specification

Emerging Technologies

- First topic: AI and Deep Learning

Emerging Technologies: AI and Deep Learning

- AI—Artificial Intelligence is intelligence demonstrated by machines
- Coined by John McCarthy in 1956, workshop at Darthmouth College
- AI Goal: Building an *intelligent agent* (intelligent = human or rational)
- Definitions can be divided into four categories (Russel and Norvig 2010 3ed.):
  - Thinking Humanly
  - Thinking Rationally
  - Acting Humanly
  - Acting Rationaly

AI Research Field

- Three functionalities of an intelligent agent:

1. Sense, perception
   - Computer Vision, Audio and Speech Processing
   - NLP Analysis
   - Sensor and other data analysis and mining
2. Understanding, reasoning, inference
   - Planning, problem solving, search
   - Machine Learning, NLP
3. Acting
   - Planning, NLP generation, speech generation

General AI Methodology

- Symbolic and Knowledge-based AI
  - based on logic rules for reasoning
  - monotonic and certain
  - requires exploration and elimination of many possiblities
  - works well on small problems but hard to scale
- Stochastic and Probabilistic AI
  - based on probabilities or other scoring schemes
  - non-monotonic and uncertain
  - requires computational model for evaluating or generating possibilities
  - scales well with a lot of data and computational power, generally not explainable

Emerging AI Applications in Digital Transformation

- Automated data analysis and automated reporting
- Automated communications which provide data that can be analyzed for better AI-based decision support
- Eliminating repetitive tasks

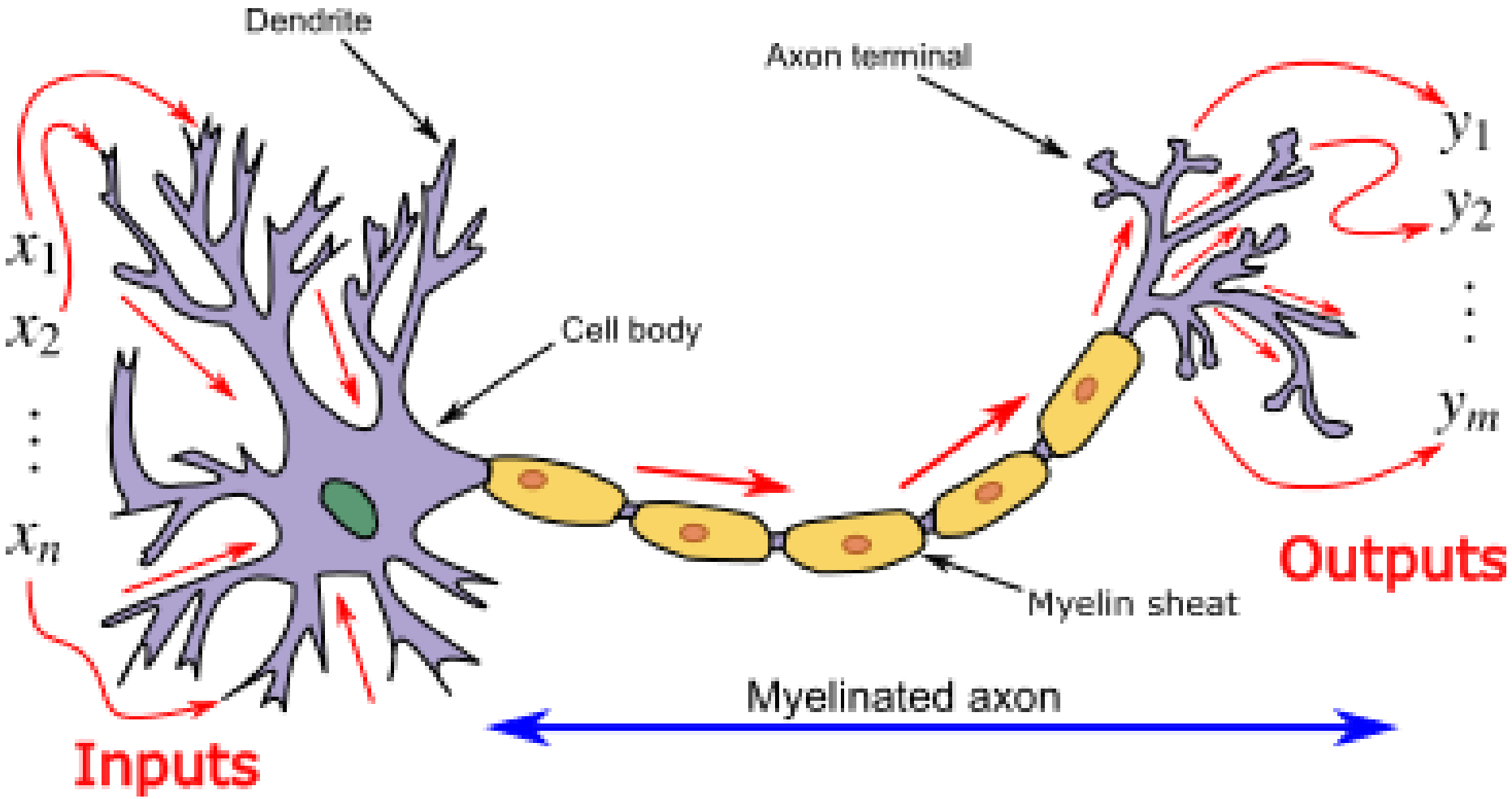
AI as Emerging Technology

- In large due to recent Machine Learning advances
- Machine Learning: learning patterns based on large amount of data, called training data
- Advances in areas:
  - Computer vision: recognizing images, object in images, video analysis, self-driving cars
  - NLP: text analysis and generation using models trained on Internet datasets, machine translation
  - Other data: speech recognition, genome mining, behavioural analysis
- Machine Learning APIs provided as a service

Deep Learning

- based on Artificial Neural Networks
- known since 1957 (Rosenblatt)
- backpropagation as training known since 1975
- slower progress for a couple decades
- new models after 2000
- 2012: ImageNet competition and Krizhevsky et al. AlexNet
- deep learning for NLP: word2vec (2013), BERT (2018), GPT-2, GPT-3 (2020), ...

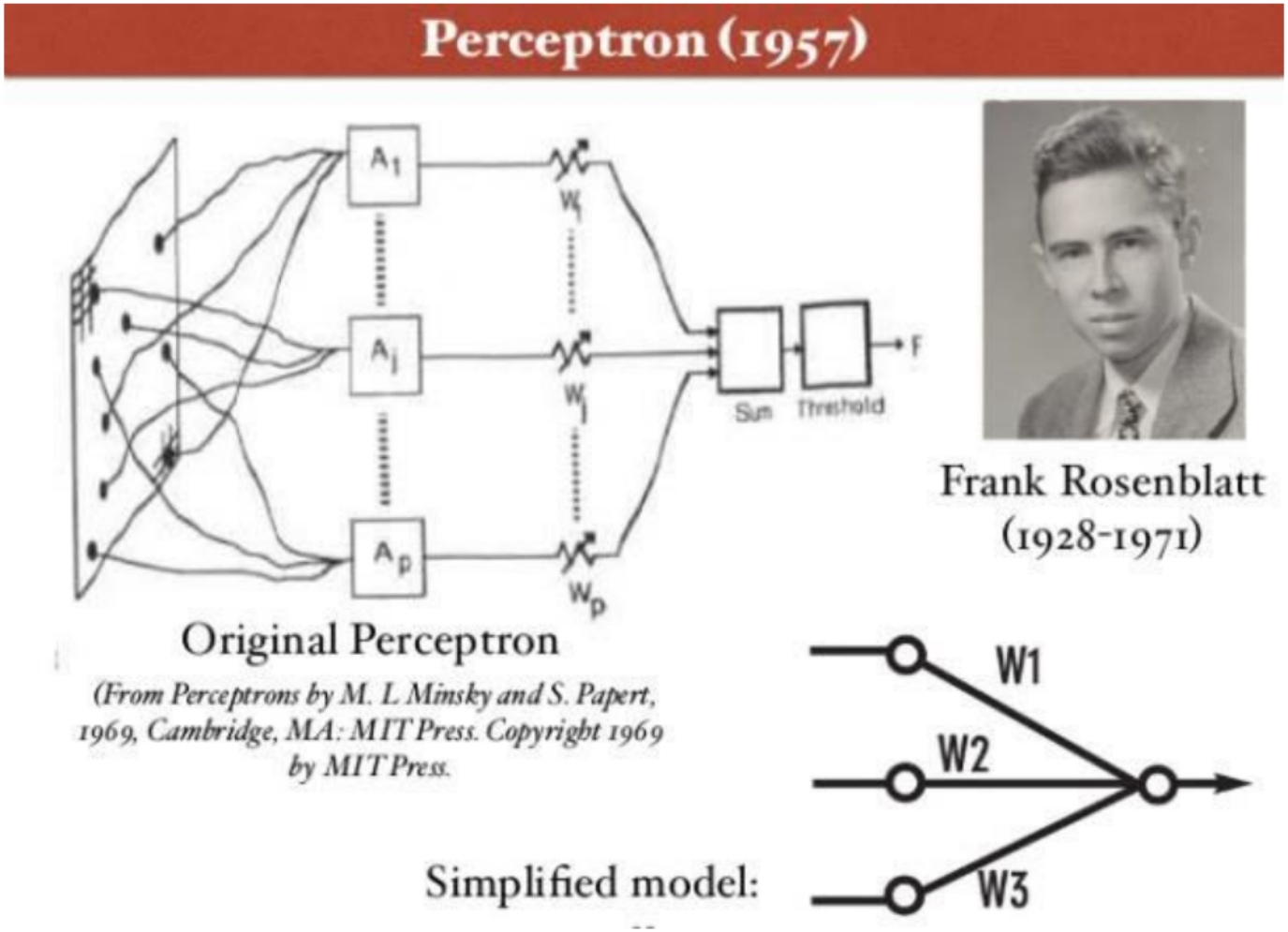
Biological Neuron



By Egm4313.s12 (Prof. Loc Vu-Quoc) - Own work, CC BY-SA 4.0,

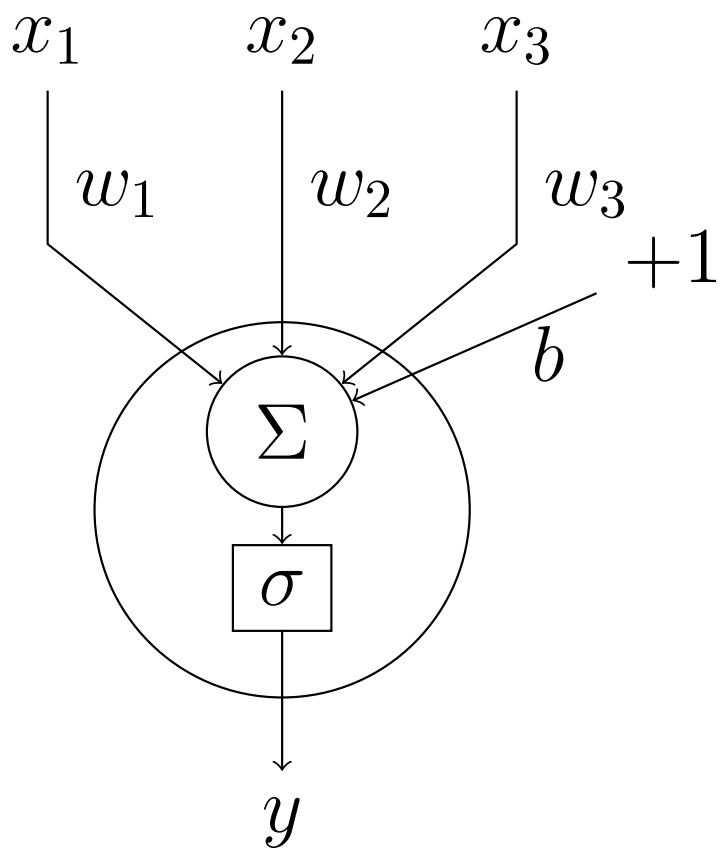
`https://commons.wikimedia.org/w/index.php?curid=72816083`

Traditional Perceptron (Artificial Neuron)



https://www.simplilearn.com/what-is-perceptron-tutorial

Computation in Artificial Neuron (Perceptron)



— input layer

— weights

— (b) bias

— weighted sum

— activation function

— output value

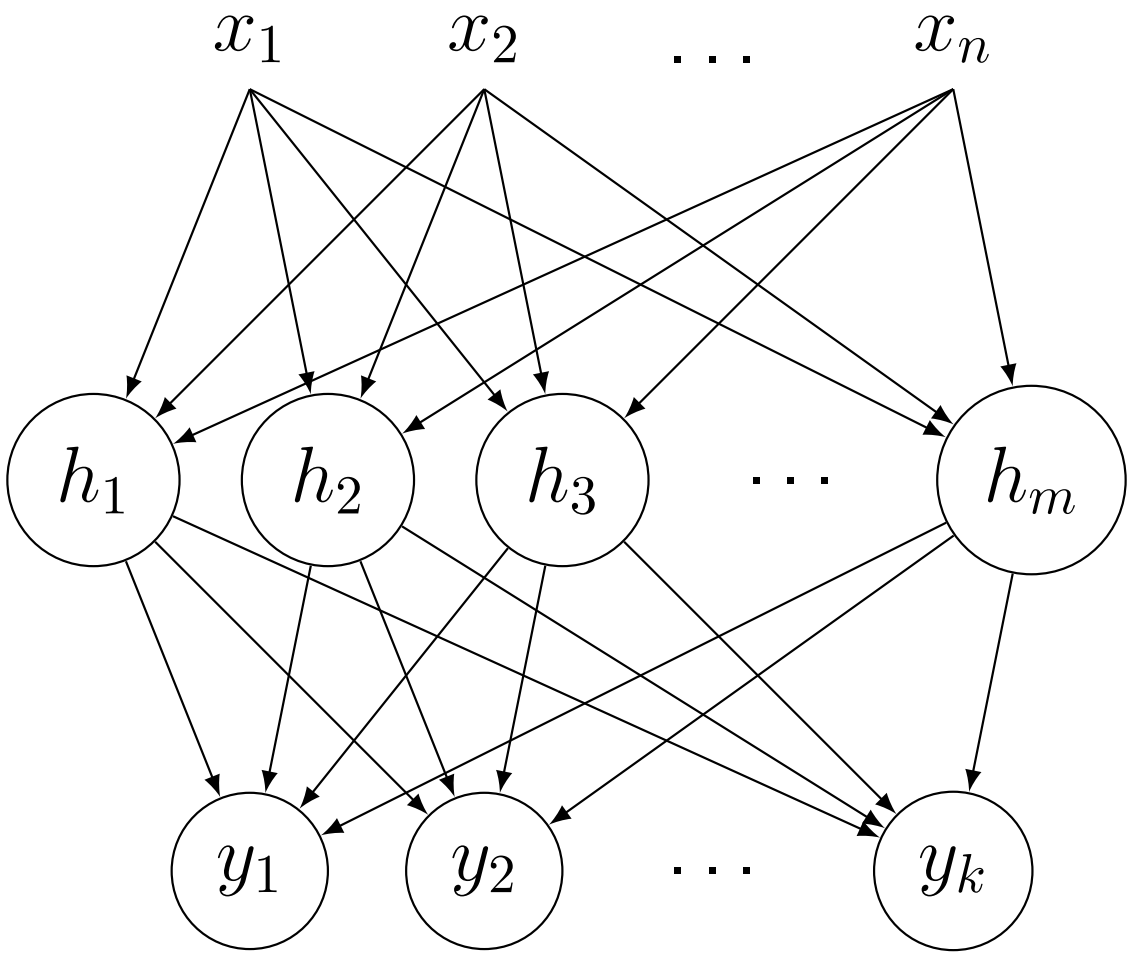
$$y = \sigma(b + \sum_i x_i w_i) = \sigma(b + x_1 w_1 + x_2 w_2 + x_3 w_3)$$

Perceptron Properties

- Biological neurons would imply activation function (non-linear transform) to be step function, or at least monotonically non-decreasing
- Could use identity function or linear function, but not a good idea
- If used as classifier ($y \geq 0$ or $y < 0$), similar to Naïve Bayes, SVM (Support Vector Machines), and logistic regression
  - linear separability
- Connected to make Neural Networks (brain analogy)

Feedforward Neural Network

also called *multi-layer perceptron*

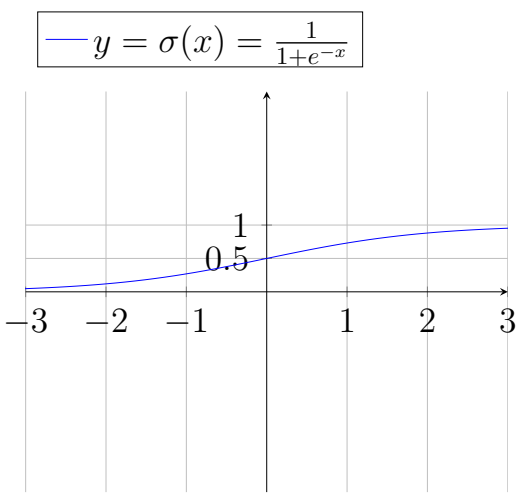


Activation Function

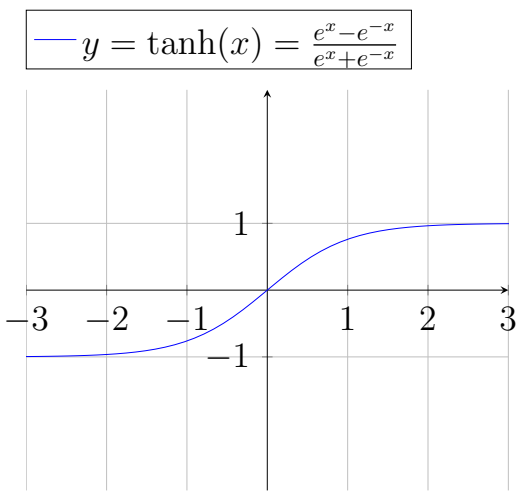
- must be non-linear
  - otherwise, the whole neural network would collapse into one neuron
- should be monotonically non-decreasing
- useful to be differentiable and relatively simple for speed of training
- Best known activation functions: sigmoid, tanh, ReLU (Rectified Linear Unit)

Common Activation Functions

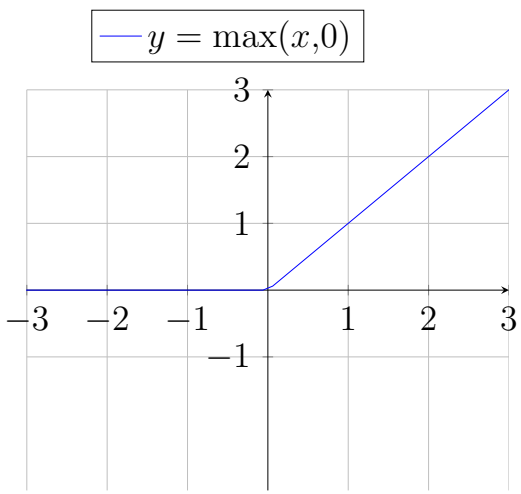
Sigmoid



tanh

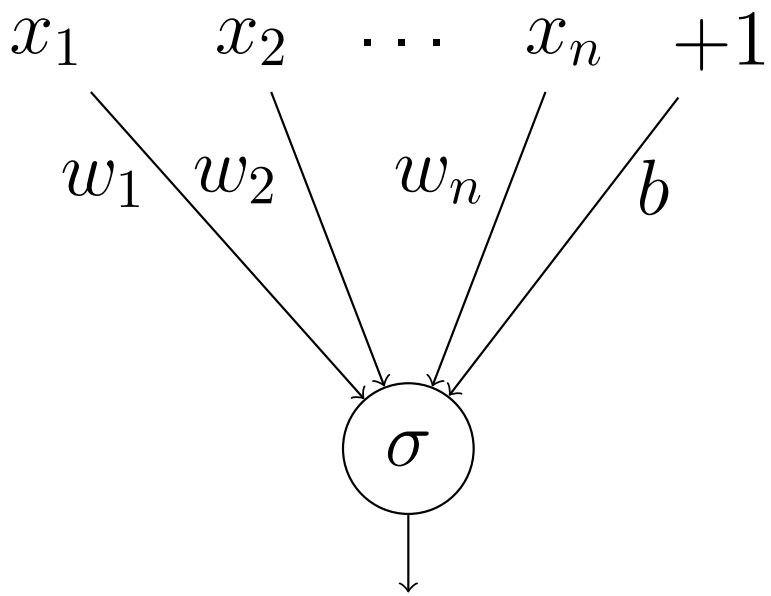


ReLU



Binary Classification with One Layer

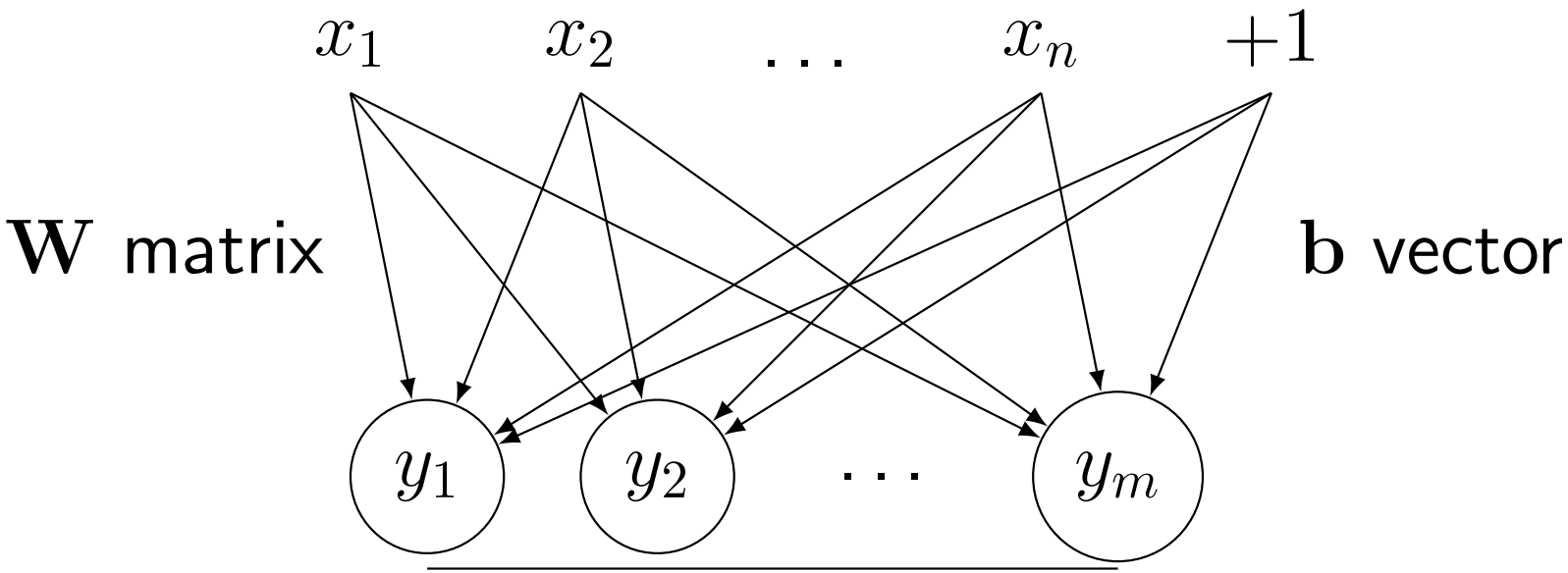
- same as binary logistic regression



$$y = \sigma(\boldsymbol{w} \cdot \boldsymbol{x} + \boldsymbol{b})$$

Multinomial Logistic Regression

- achieved with one-layer classification



$$\mathbf{y} = \text{softmax}(\mathbf{W}\mathbf{x} + \mathbf{b})$$

Softmax Function

- Softmax transforms numbers into positive domain using e^x; i.e., $\exp(x)$, function, and normalizing numbers into a probability distribution

$$\text{softmax}(\mathbf{x}) = [\frac{\exp(x_1)}{\sum_{i=1}^{n}\exp(x_i)}, \frac{\exp(x_2)}{\sum_{i=1}^{n}\exp(x_i)}, \cdots \frac{\exp(x_n)}{\sum_{i=1}^{n}\exp(x_i)}]$$

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^{n}\exp(x_j)}$$

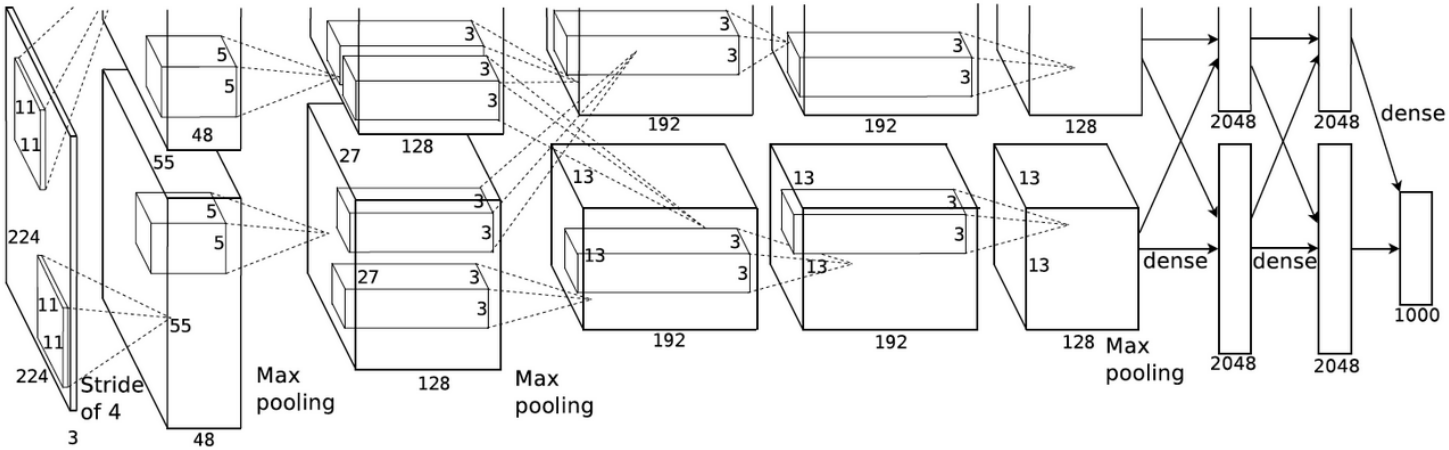
- Example (Jurafsky and Martin):

$$\mathbf{x} = [0.6, 1.1, -1.5, 1.2, 3.2, -1.1]$$

$$\text{softmax}(x) = [0.055, 0.09, 0.006, 0.099, 0.74, 0.01]$$

Deep Learning

- Achieved with many network layers
- Example, AlexNet schema:



- Driven by previous ML (Machine Learning) advances and hardware advances (GPU)

Another View to Popularity of Deep Learning Models

- Artificial Neural Networks research, 1958 perceptron
- Backpropagation training 1986
- Neural Networks used since then but no significant success in NLP
- Important milestone: AlexNet winning ImageNet competition on Sep 30, 2012
- word2vec 2013, Mikolov et al. at Google
- Development of larger models since then

New Network Architectures

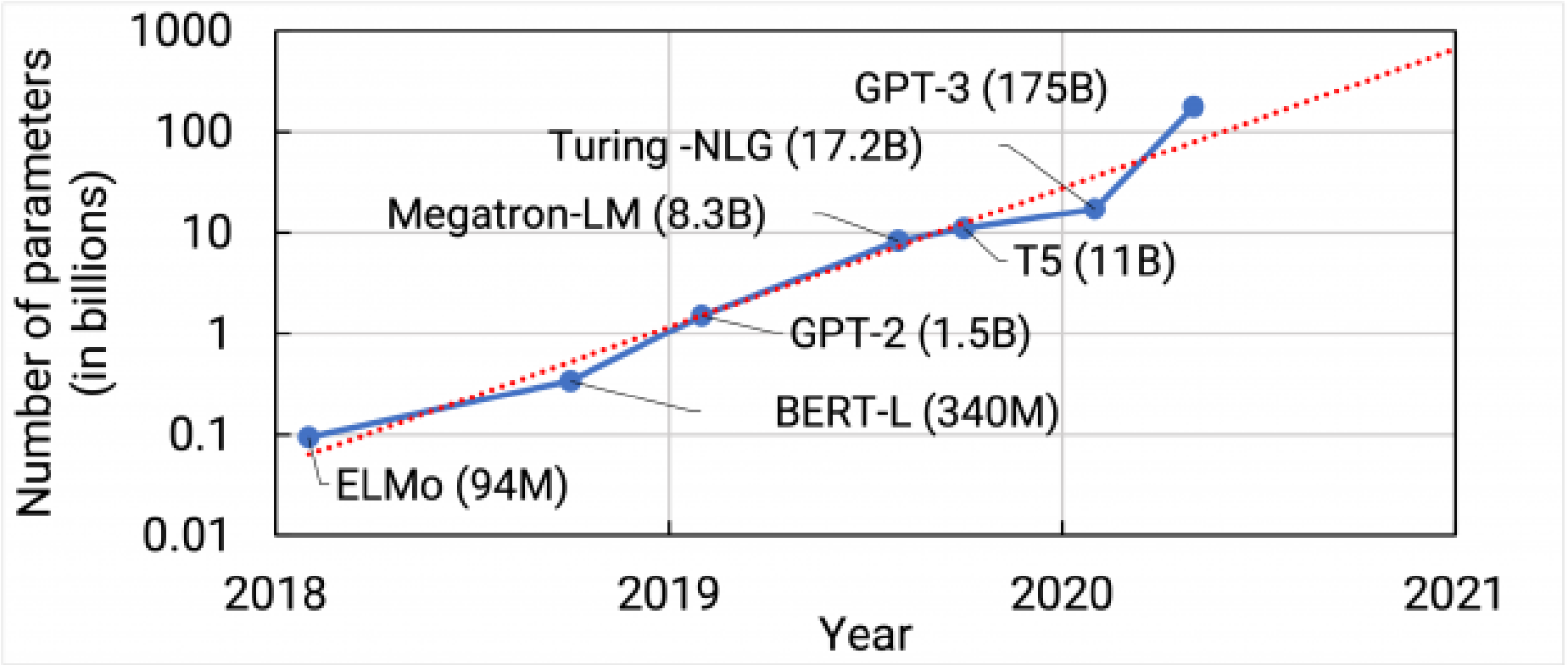
- Word embeddings (based on NN)
- RNN (Recurrent Neural Networks)
- LSTM (Long Short-Term Memory Networks)
- BERT (Bidirectional Transformers, Google)
- GPT-2, GPT-3 (OpenAI)
- etc. . .

Large Deep Learning Models in NLP

- ELMo (Embedding from Language Model) 2018 by Allen Institute for Artificial Intelligence and University of Washington, 94mil parameters
- BERT (Bidirectional Encoder Representations from Transformers) 2018 by Google, 340mil par.
- GPT-2 by OpenAI in 2019, 1.5bil. param.
- Megatron-LM bu NVIDIA, 8.3bil. param.
- Turing-NLG by Microsoft, 17.2bil. param.
- GPT-3 in 2020 by OpenAI, 175bil. param.
- Exponential growth in number of parameters
- GPT-3 is not open, with exclusive licence to Microsoft

Deep Learning Language Model Sizes

- Exponential growth:



Report 1: Seminar Report Reminder

- Seminar Report 1 due on Monday, 10-Mar-2025 by midnight
- Submit on Brightspace
- Read general report specifications and use given Word template
- Based on both: Tukan's presentation and this presentation
- Answer two questions:
  1. Summarize some directions on further development described in Tukan's presentation.
  2. Desribe how can this technology presented in Tukan's and this presentation be applied in a domain of your interest, related to the certificate you are in.
- Write in your own words.