APS360: Applied Fundamentals of Deep Learning

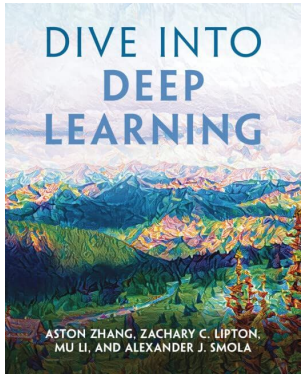
Week 1: Artificial Neural Networks - Part I

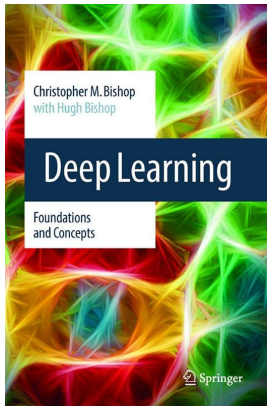
Material & Resources

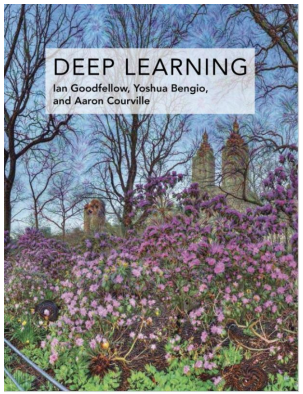
Course Website http://q.utoronto.ca

Textbooks (Optional, free online):

- **Dive into Deep Learning** https://d2l.ai
- **Deep Learning: Foundations and Concepts** https://www.bishopbook.com
- **Deep Learning Book** https://www.deeplearningbook.org







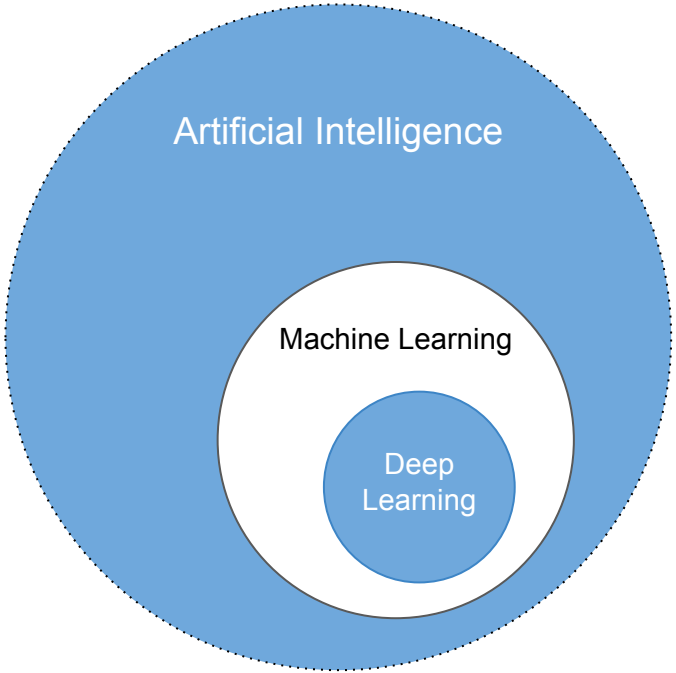
What is Deep Learning?

Terminology

Artificial Intelligence (AI) : broad & poorly defined concept of developing computer systems that can perform tasks normally only humans could do

Machine Learning (ML) : computers learn by example, from data, rather than being explicitly programmed, to solve a task

Deep Learning (DL) : A machine learning method that learns multiple levels of abstractions over data end-to-end



Why do we need Machine Learning?

Much of computer science is the study of algorithms, how to write a program to solve a specific task.

Typically we write a program to evaluate some input with a set of clear rules that determine the output

Imagine I asked you to write a program to tell me if there is a goat in my photos

We need Machine Learning

- Almost any rule you can come up with will have some **counter-example** in the real world
- It is difficult to formulate rules that **cover all the conditions** we are expected to understand (and often we've never imagined them!)
- High-dimensional input space, hard to understand, we must first learn easier **representations**
- We instead require a way to learn from a lot of examples how to solve a problem!






Formal Definition of ML

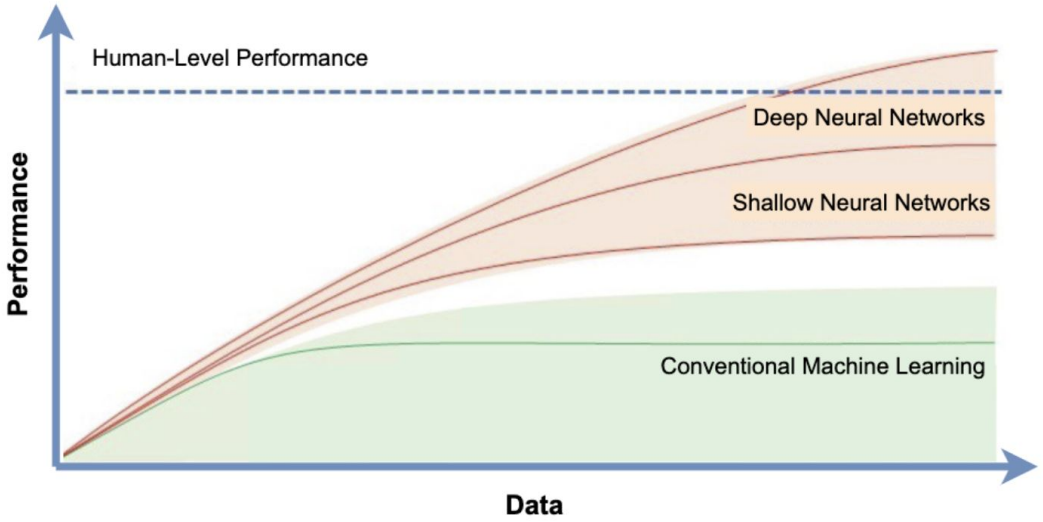
"A computer program is said to learn from **experience E** with respect to some class of **tasks T** and **performance measure P**, if its performance at tasks in T, as measured by P, improves with experience E."

(**Mitchell et al. 1997**)

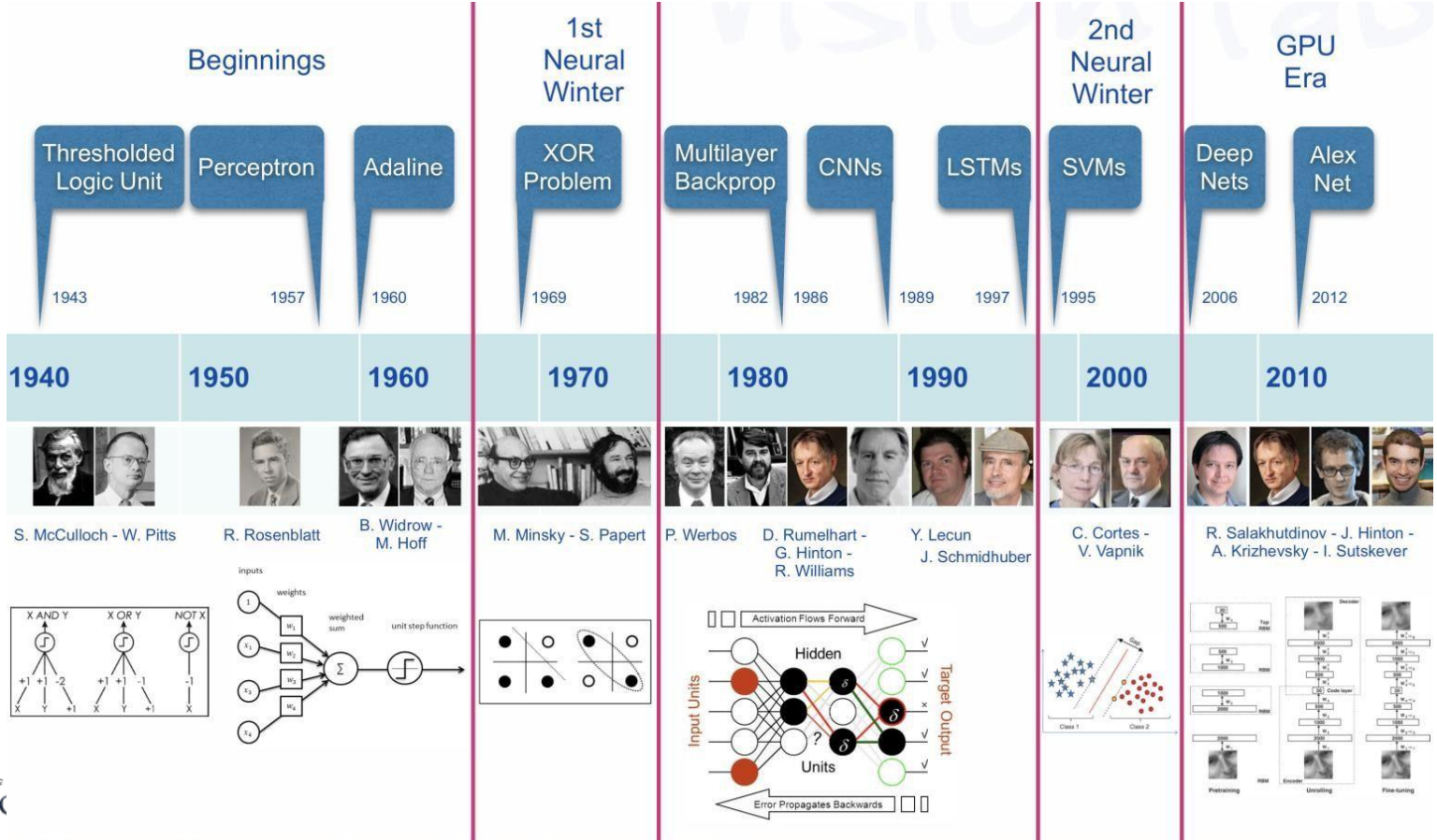
Deep Learning

Deep Learning is the latest version of **Artificial Neural Networks (ANN)** , or **Connectionism** an old ML method

Neural Networks were inspired by the brain… like a jet plane was by birds

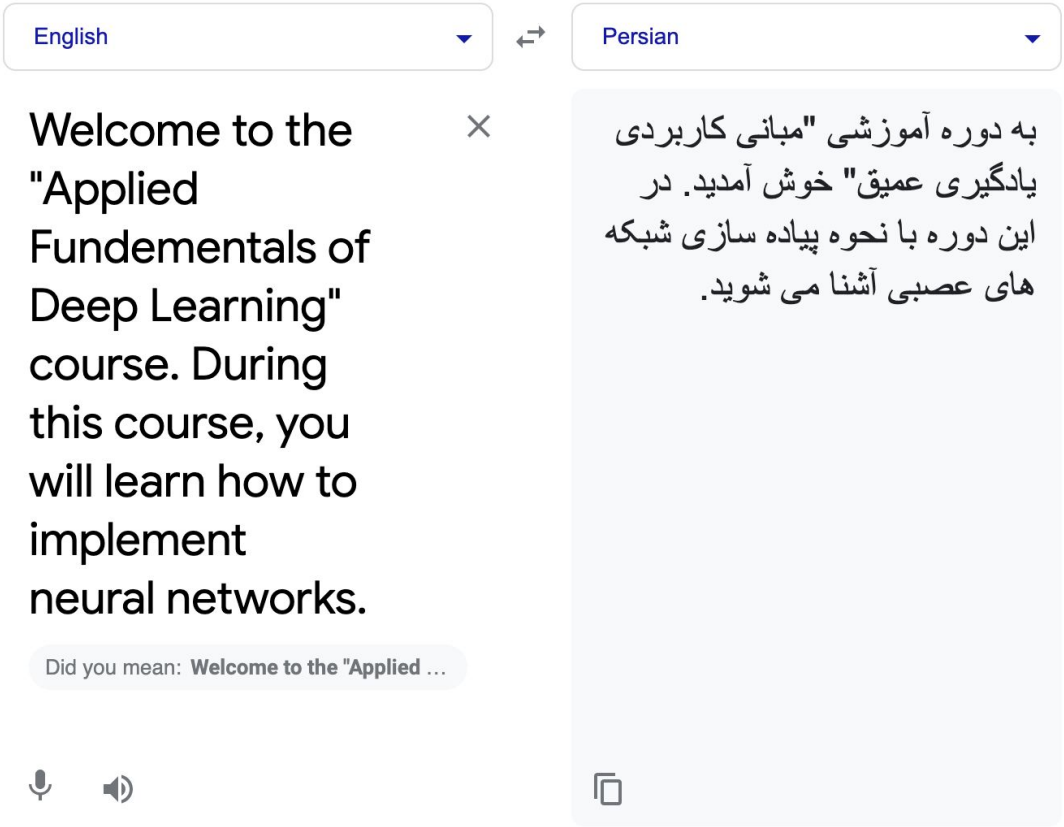


History of Deep Learning



Deep Learning: Successes and Caveats

Deep Learning Success: Machine Translation



Deep Learning Success: Drug Discovery

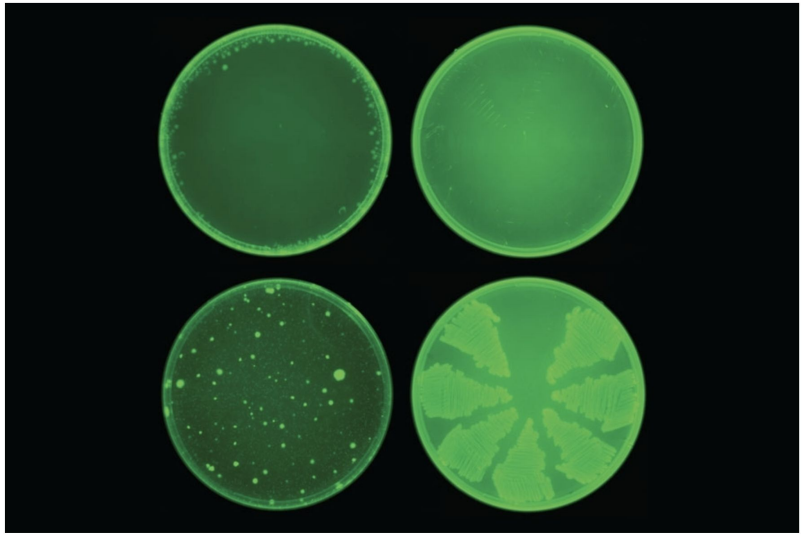
MIT News
ON CAMPUS AND AROUND THE WORLD

SUBSCRIBE

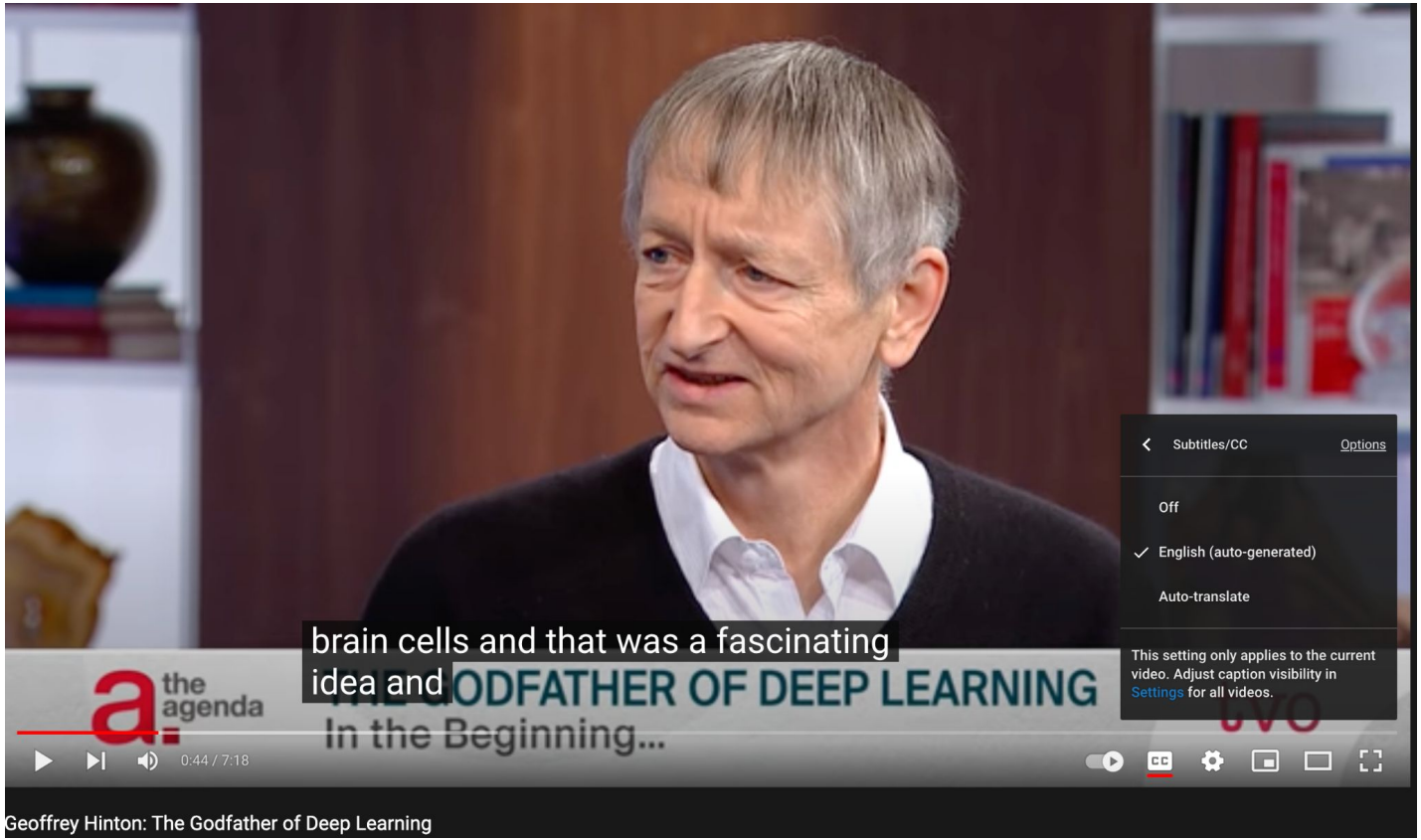
Artificial intelligence yields new antibiotic

A deep-learning model identifies a powerful new drug that can kill many species of antibiotic-resistant bacteria.

Anne Trafton | MIT News Office
February 20, 2020



Deep Learning Success: Speech Recognition



Deep Learning Success: Image Generation

A photo of a confused grizzly bear in calculus class

A pre raphaelite painting of a person waiting for their iPhone to power on after plugging it back in

robot apes showing affection for their young ones.

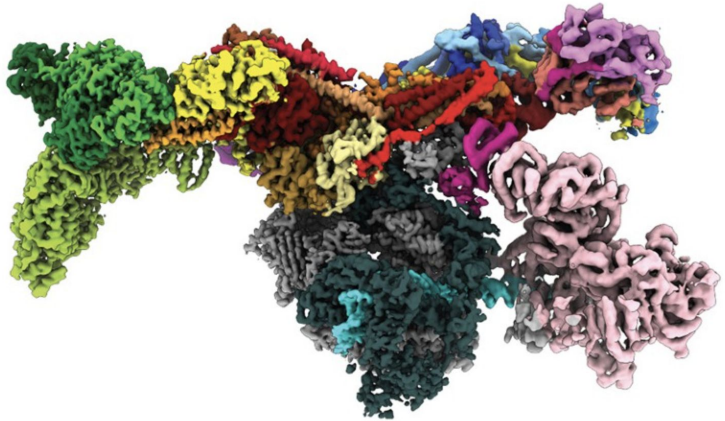
A spooky swamp with mist and fog rising up towards the bright moonlight

Deep Learning Success: AlphaFold

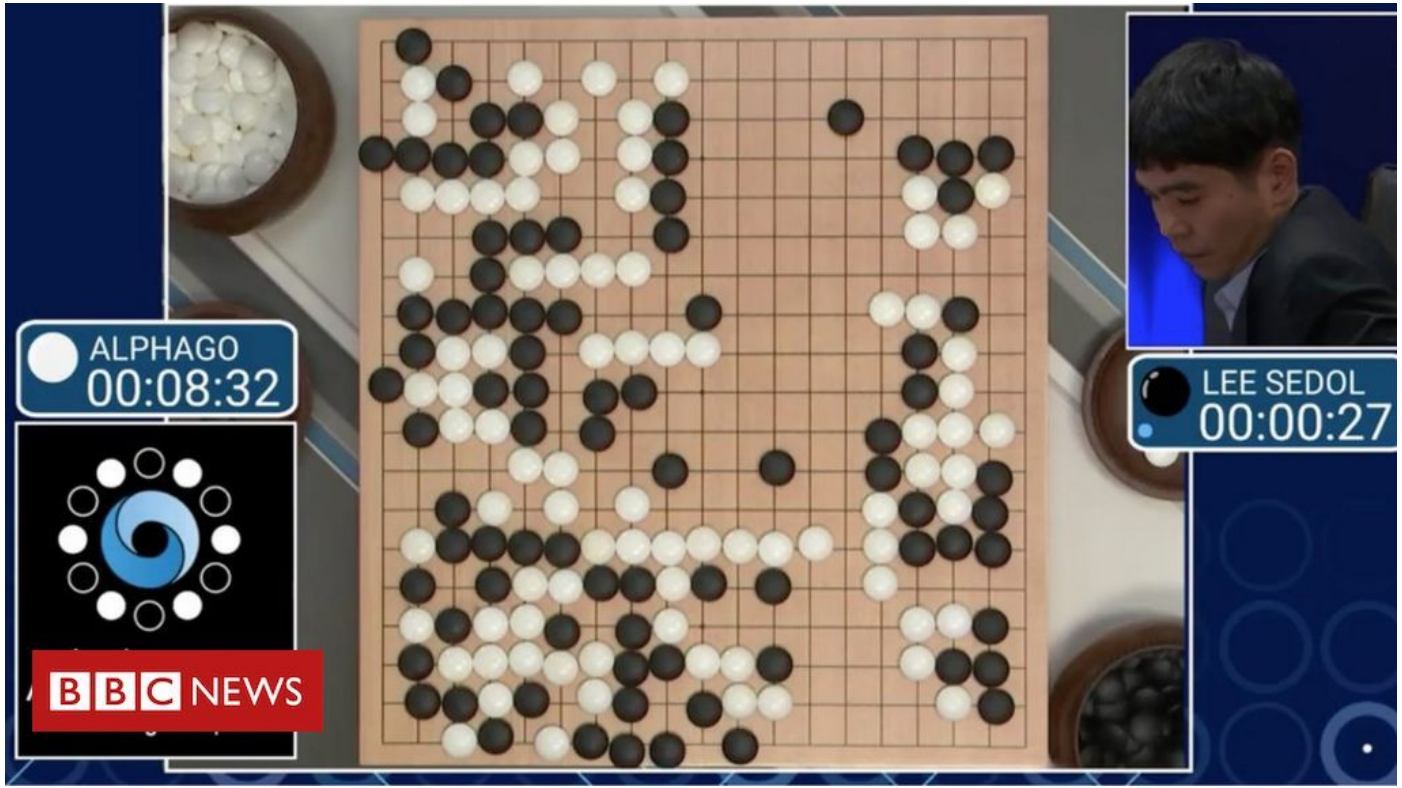
DeepMind's AI predicts structures for a vast trove of proteins

AlphaFold neural network produced a 'totally transformative' database of more than 350,000 structures from *Homo sapiens* and 20 model organisms.

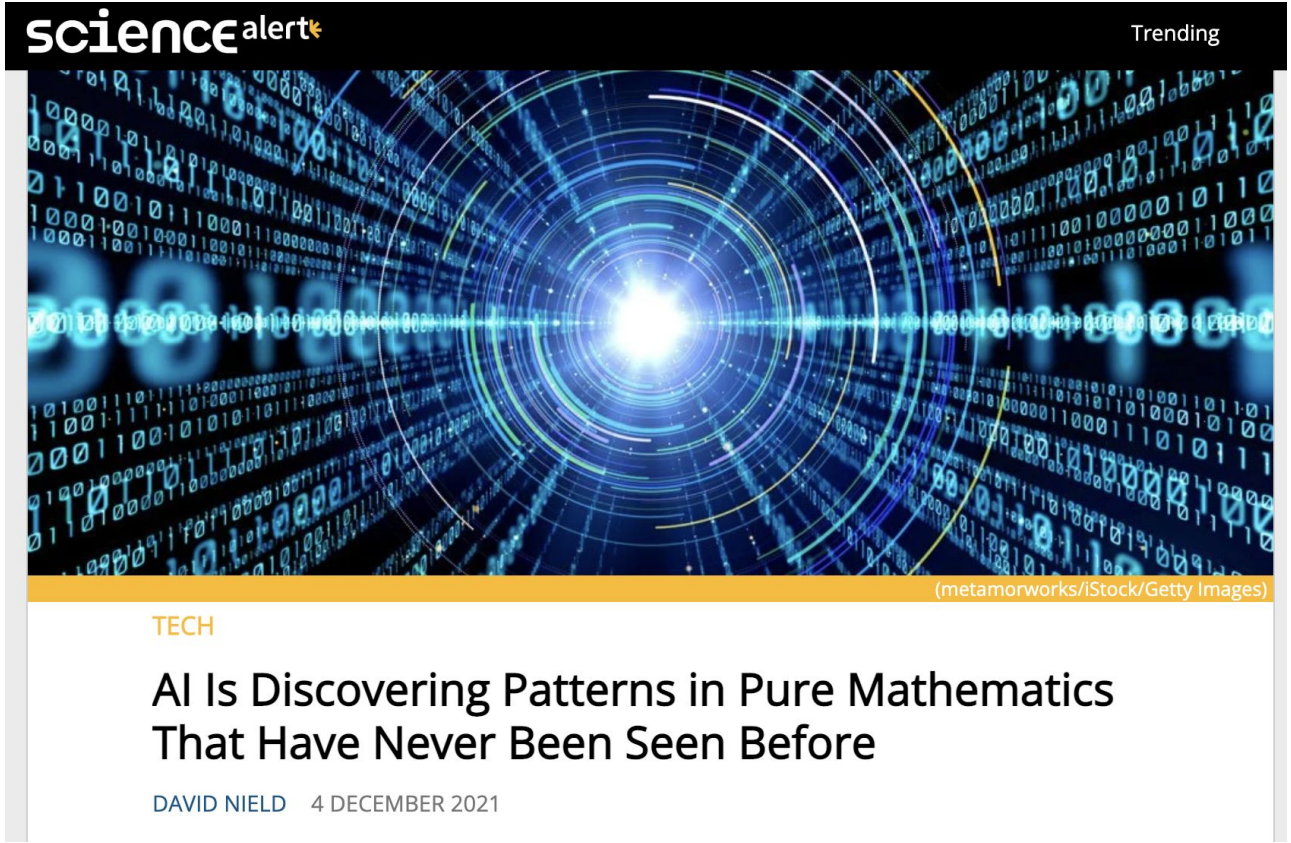
Ewen Callaway



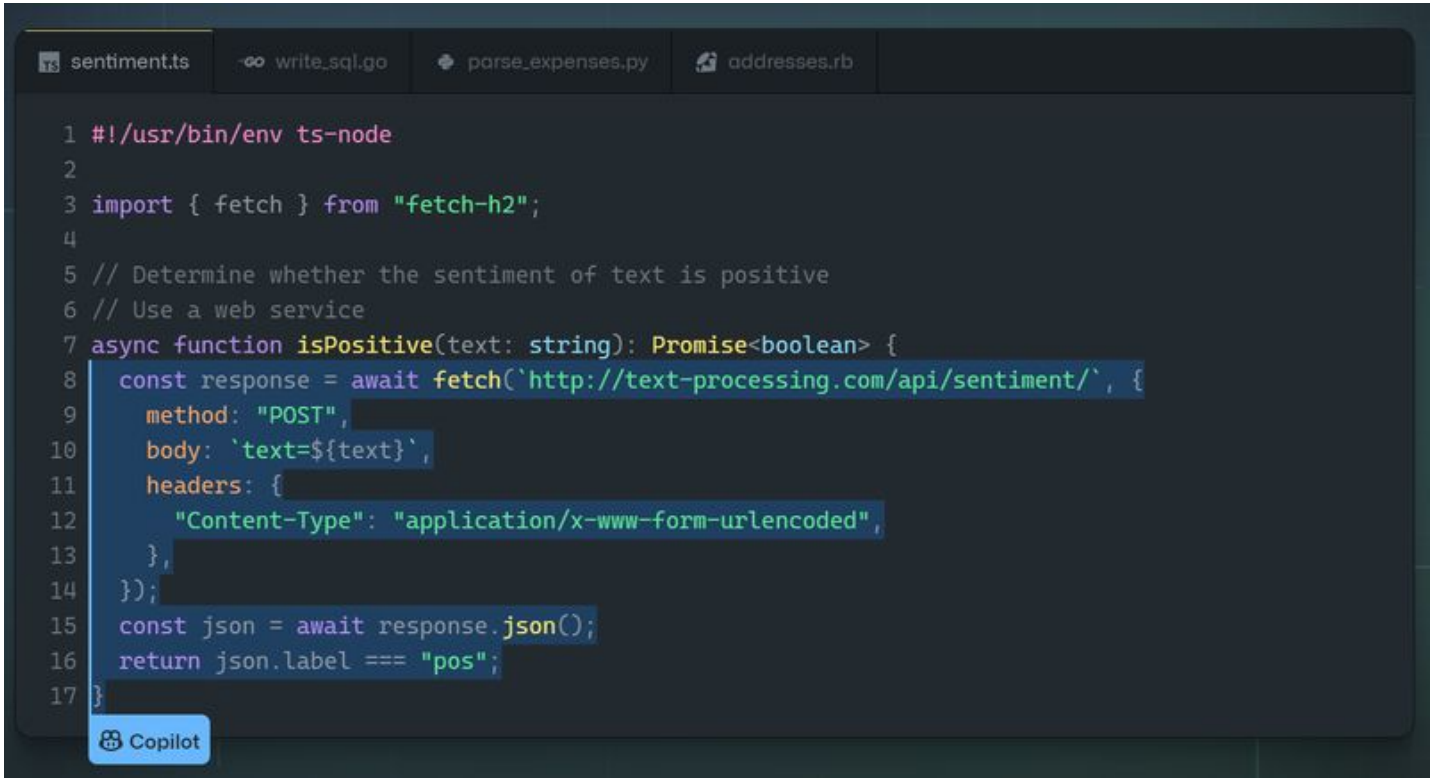
Deep Learning Success: AlphaGo



Deep Learning Success: Mathematics



Deep Learning Success: Code Generation



Deep Learning Success: Language Modelling

This is a picture of Barack Obama. He is a former president of the United States.

How many people are there in this picture?

There are at least 5 people in this picture.

Where was this picture taken?

It was taken in a school.

How many mirrors are there?

There are at least 2 mirrors in this picture.

What is the person standing on?

The person is standing on a rug.

He is standing on a scale.

I think you are right.

What is he doing?

He is looking at the scale.

Where is Obama's foot positioned?

Obama's foot is positioned on the right side of the scale.

What happens as a result?

The scale shows a higher weight.

Is the person on the scale aware of it?

I think he is not aware of it.

Do you think that's why people are laughing?

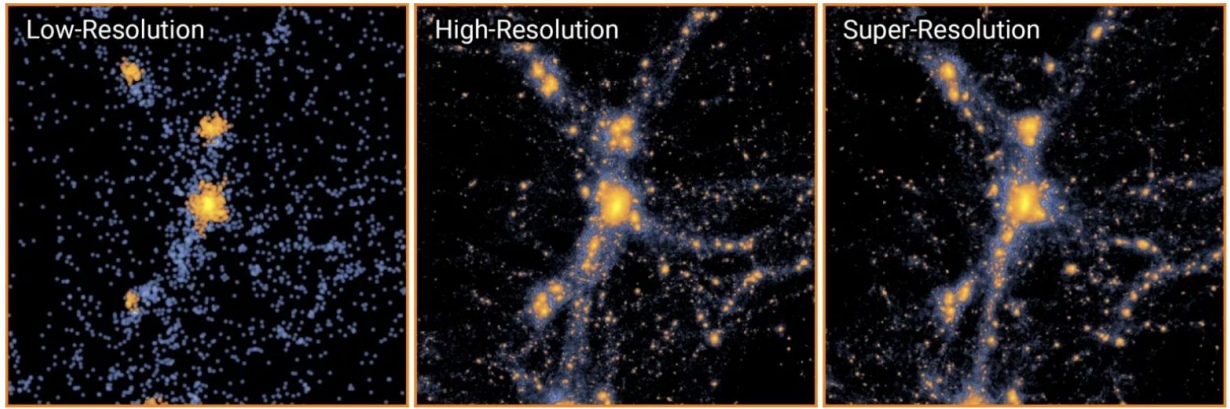
I think so.

Deep Learning Success: Simulators

AI "Magic" Just Removed One of the Biggest Roadblocks in Astrophysics

TOPICS: Artificial Intelligence Astronomy Astrophysics Machine Learning Popular Simons Foundation

By SIMONS FOUNDATION MAY 8, 2021



Simulations of a region of space 100 million light-years square. The leftmost simulation ran at low resolution. Using machine learning, researchers upscaled the low-res model to create a high-resolution simulation (right). That simulation captures the same details as a conventional high-res model (middle) while requiring significantly fewer computational resources. Credit: Y. Li et al./Proceedings of the National Academy of Sciences 2021

Deep Learning Caveats: Interpretability

Figure https://xkcd.com/1838/

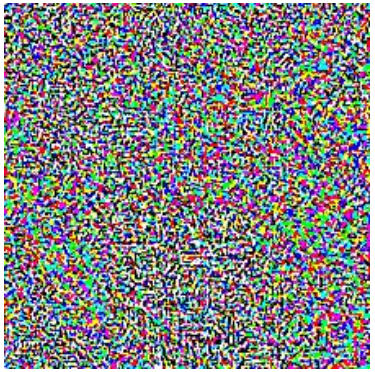
Deep Learning Caveats: Adversarial Examples

Clean Image

Dog

Confidence: 92.2%

Noise

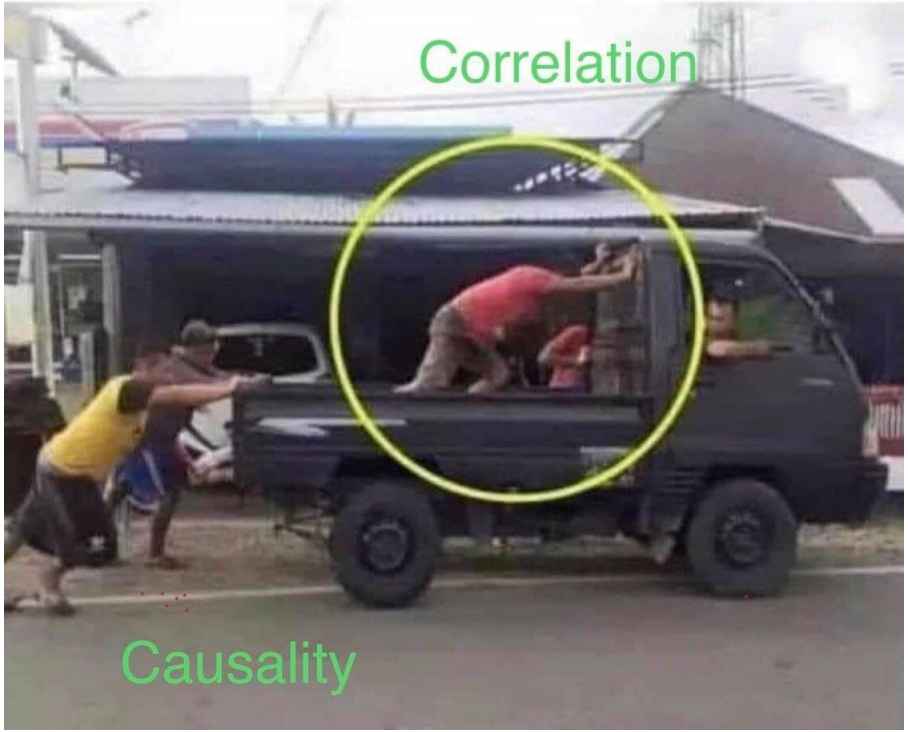


Adversarial Example

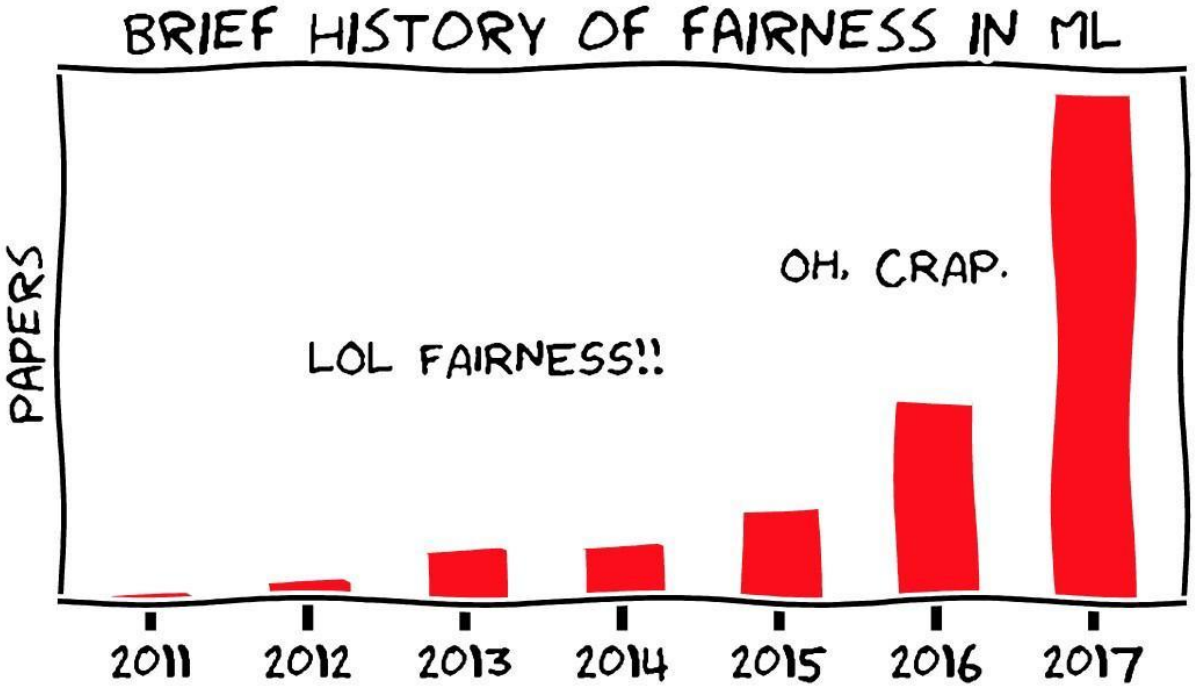
Ostrich

Confidence: 98.2%

Deep Learning Caveats: Causality

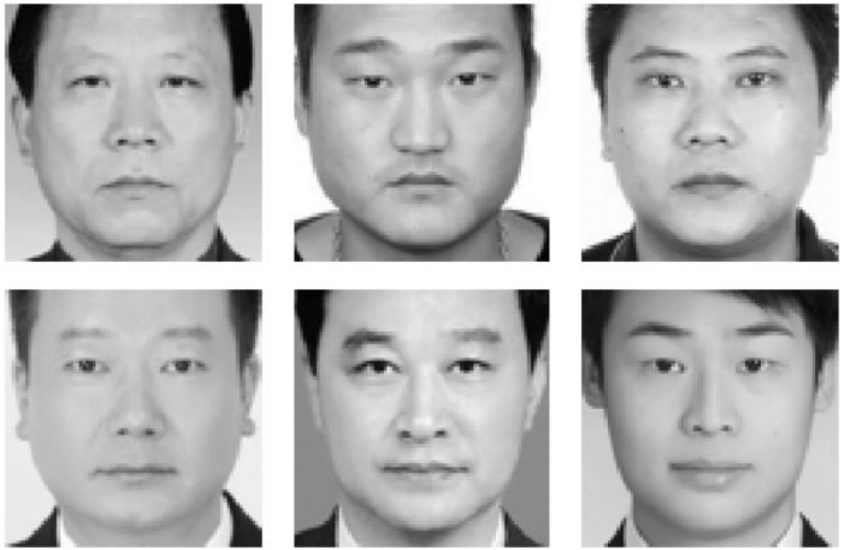


Deep Learning Caveats:Fairness & Bias



Bias in ML: Examples

- A 2016 arXiv paper, claimed to be able to predict whether someone was a convicted criminal or not solely from a driver's license-style photo with 90% accuracy
- When looked at in detail, images of criminals were collected from government IDs, while non-criminal face images were collected from the internet

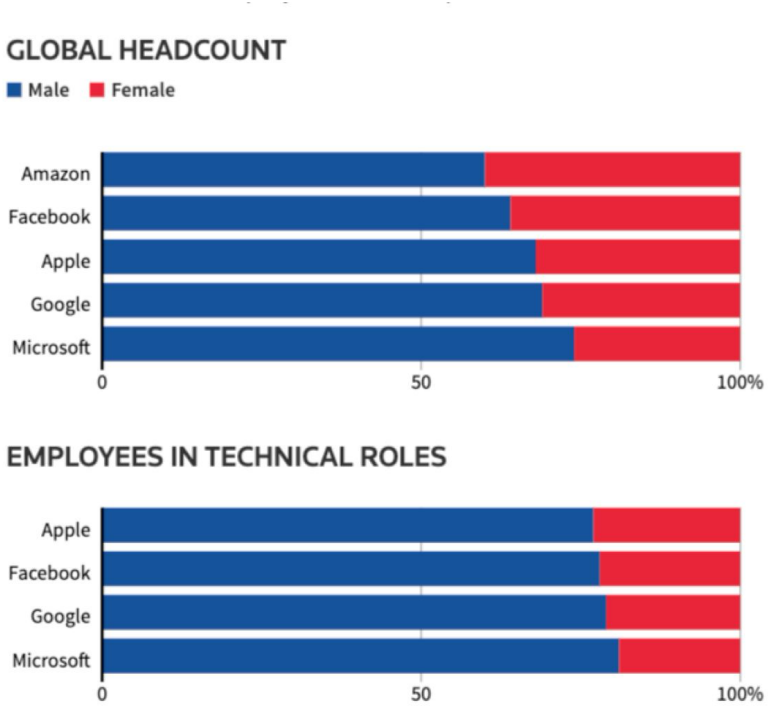


Wu and Zhang's "criminal" images (top) and "non-criminal" images (bottom).

https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

Bias in ML: Examples

Jeffrey Dastin, *Amazon scraps secret AI recruiting tool that showed bias against women*, **Thomson Reuters** (2018)



Machine Learning Basics

Machine Learning

Supervised Learning

- Regression (real-valued or continuous value) or Classification (categorical or 1 of N)
- Requires data with ground-truth labels/outputs

Unsupervised Learning

- Self-supervised Learning, Semi-supervised Learning
- Requires observations without human annotations

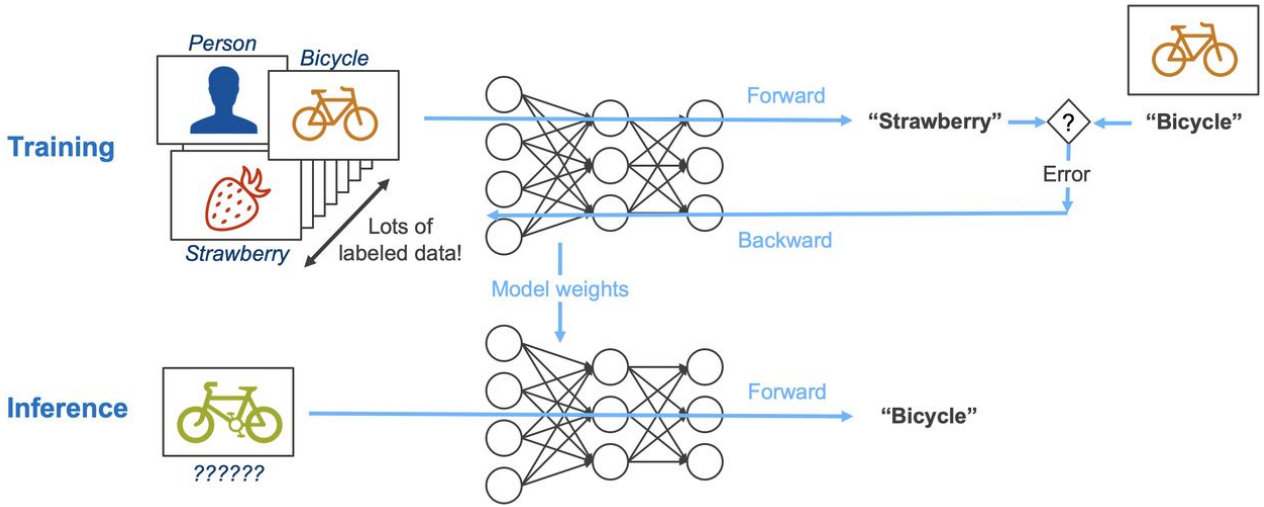
Reinforcement Learning

- Sparse rewards from environment (e.g., won/lost)
- Actions affects the environment (dynamic nature)

Supervised Learning

Model learns to maps an input to an output based on example input-output pairs.

Much like a teacher guides a student, but with **many** more examples



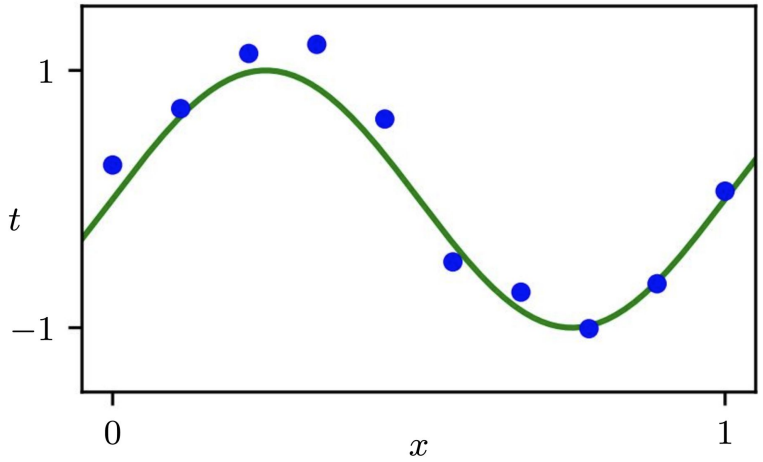
Fitting a Polynomial (Regression)

Suppose we use function $t=\sin(2\pi x)+\epsilon$ to generate 10 data points.

The goal is to predict the value of t for some new value of x, without knowledge of the green curve.

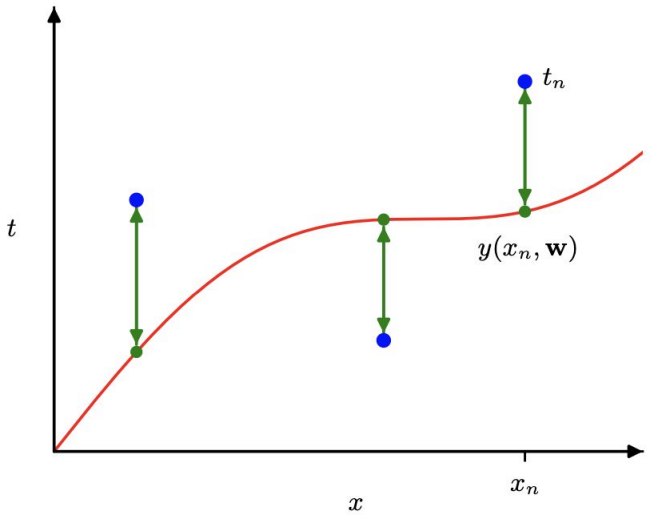
Training is finding a set of optimal weights to best approximate the function

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$



Loss Function

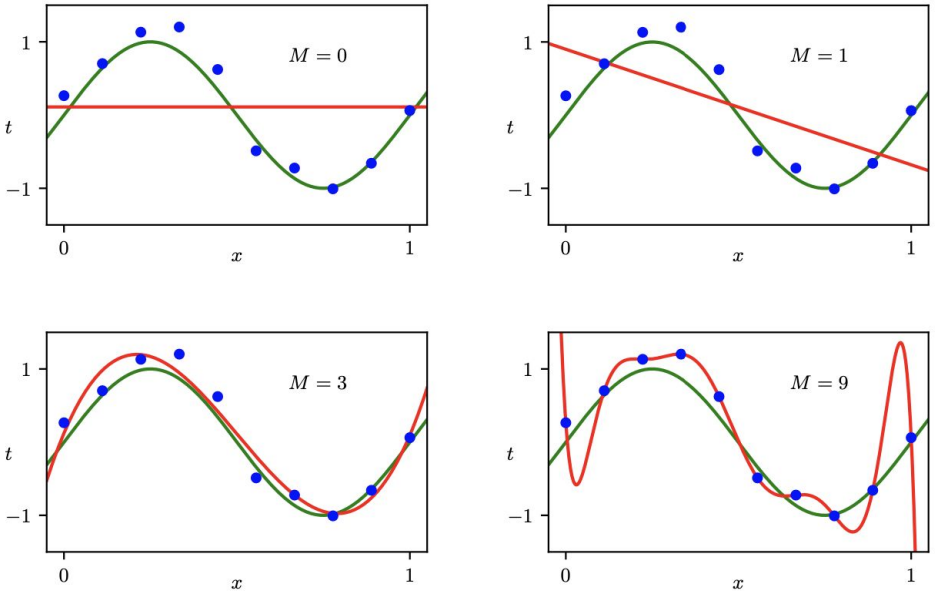
We want to learn the coefficients (weights) such that the the predictions are as close as possible to the training data.

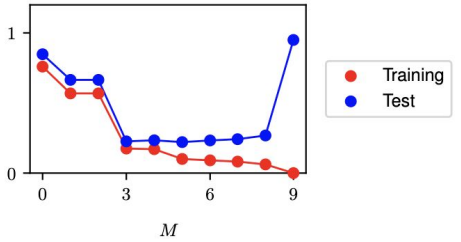


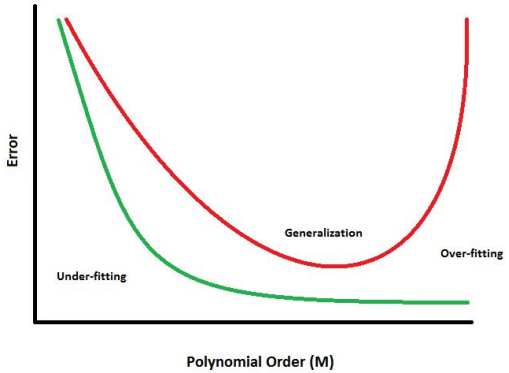
$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

Model complexity

How should we decide the polynomial order?





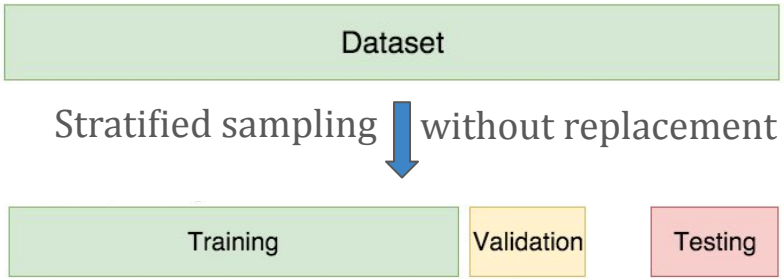


Train/Validation/Test splits

we cannot use all our data for training, need to use some to test our model

Common mistake is to keep trying until favourable outcomes are achieved on the testing data, Effectively making your testing data the same as your training data and leads to over-fitting!

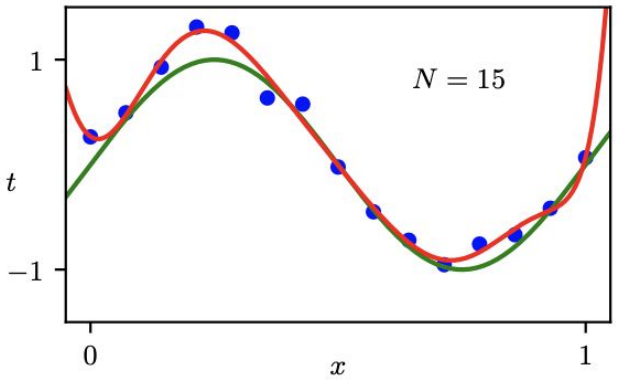
Train on training, tune hyper-parameters on validation, evaluate sparingly on test set

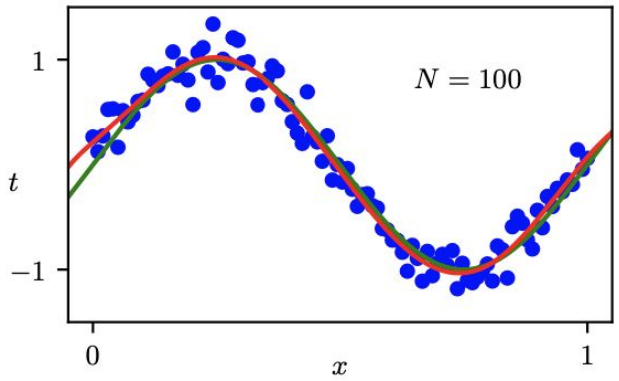


Avoiding overfitting: More Data

Below is the fitting of a model with M=9 with 15 (left) and 100 (right) training examples.

More data helps the model to learn better and generalize better



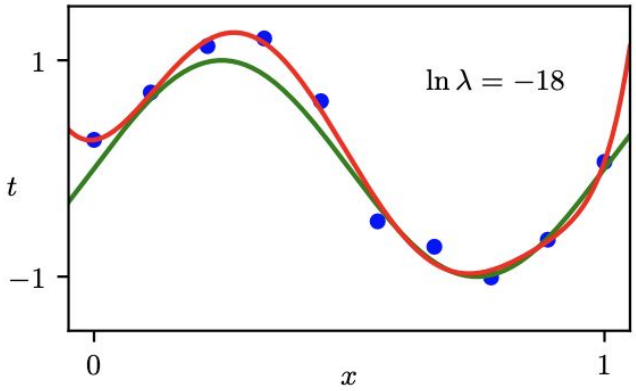


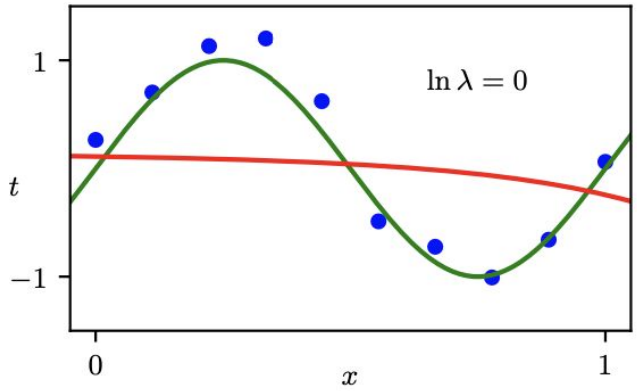
Avoiding overfitting: Regularization

	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.11	0.90	0.12	0.26
$w_1^\star$		−1.58	11.20	−66.13
$w_2^\star$			−33.67	1,665.69
$w_3^\star$			22.43	−15,566.61
$w_4^\star$				76,321.23
$w_5^\star$				−217,389.15
$w_6^\star$				370,626.48
$w_7^\star$				−372,051.47
$w_8^\star$				202,540.70
$w_9^\star$				−46,080.94

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

$$\|\mathbf{w}\|^2 \equiv \mathbf{w}^{\mathrm{T}}\mathbf{w} = w_0^2 + w_1^2 + \ldots + w_M^2$$





Artificial Neuron

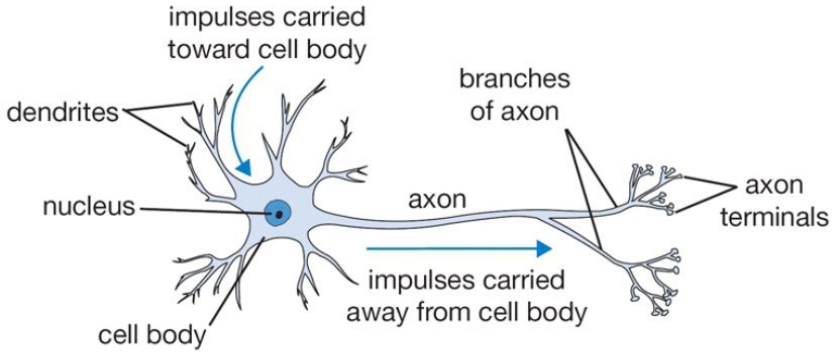
Simplified Biological Neuron

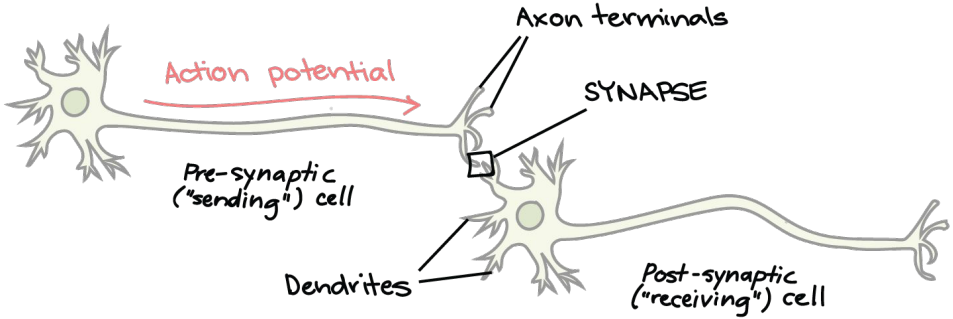
Dendrites receive information from other neurons.

Cell body consolidates information from the dendrites.

Axon passes information to other neurons.

Synapse is the area where the axon of one neuron and the dendrite of another connect.





Artificial Neuron

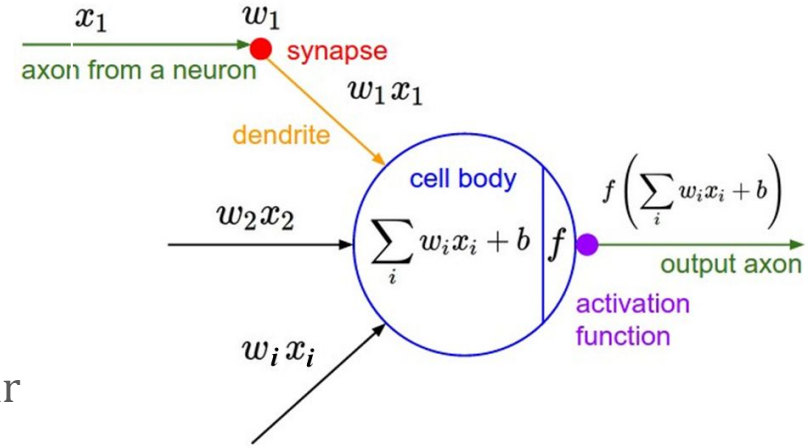
x_i is the **input** such as a pixel in an image

w_i is the **weight** for input x_i that we learn for this particular input

b is the **bias**, a weight we learn with no input

f is the **activation function** that determines how our output changes with the sum of all weight-input products

y is the **output** such as the class an image belongs to



$$y_p = f(\mathbf{w}\cdot\mathbf{x}+b)$$

Artificial Neuron

year
sq^2
#rooms
finished

$$\begin{pmatrix} 2001 \\ 3000 \\ 5 \\ 1 \end{pmatrix} \Longrightarrow X = \begin{pmatrix} 2001 \\ 3000 \\ 5 \\ 1 \end{pmatrix}$$

price

$$\begin{pmatrix} 1.5M \end{pmatrix} \Longrightarrow y_t = \begin{pmatrix} 1.5M \end{pmatrix}$$

$$y_p = f(\mathbf{w} \cdot \mathbf{x} + b)$$

$$f\left(\begin{pmatrix} 2001 \\ 3000 \\ 5 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} + b \right) \cong \begin{pmatrix} 1.5M \end{pmatrix}$$

If f(x)=x then:

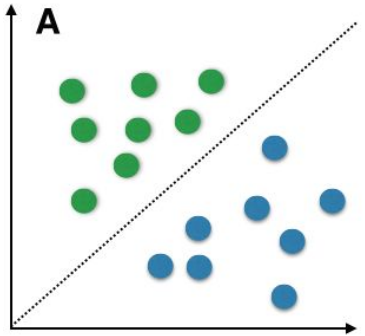
$$y_p = 2001w_1 + 3000w_2 + 5w_3 + 1w_4 + b$$

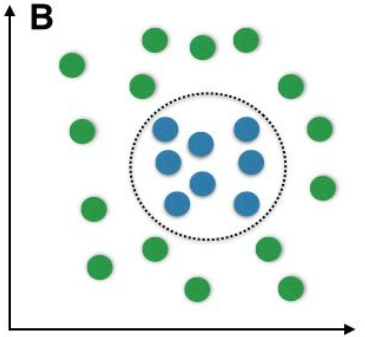
Neuron with a Linear Activation Function

What is wrong with a linear activation function?

- Most real datasets are not **linearly separable** , e.g. we can't find a line that separates classes well in a classification problem
- We can learn non-linear transformations of our data to help
- Multiple layers with non-linear transformations help
- **No advantage from multiple linear layers** → composite is a linear layer

$$\underbrace{\mathbf{W}^{(1)}\mathbf{W}^{(2)}\mathbf{W}^{(3)}}\mathbf{x}$$
$$= \mathbf{W}'\mathbf{x}$$





https://sebastianraschka.com/Articles/2014_naive_bayes_1.html

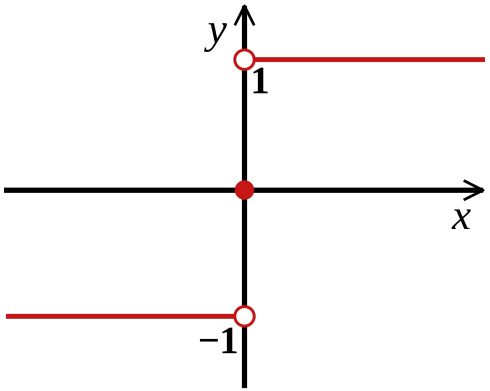
Early Activation Functions: Perceptrons

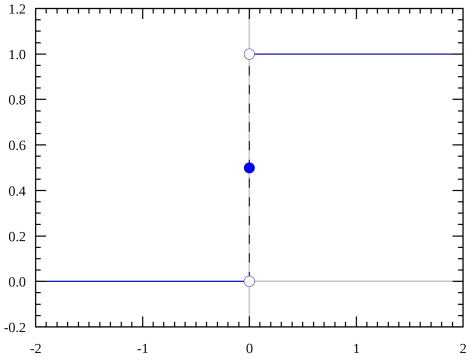
First artificial neurons (1943-70s) used a simple binary activation function based on which side of the hyperplane the input is:

$$f(x) = \text{sign}(x)$$ **Sign function**

$$f(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{if } x \geq 0 \end{cases}$$ **Heaviside (unit) step function**

This is called the *decision boundary*



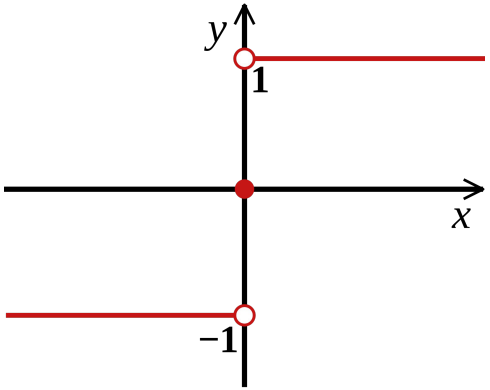


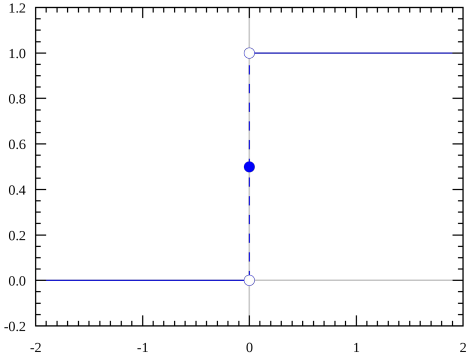
Early Activation Functions: Perceptrons

First artificial neurons (1943-70s) used a simple binary activation function based on which side of the hyperplane the input is:

$$f(x) = \text{sign}(x)$$ **Sign function**

$$f(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{if } x \geq 0 \end{cases}$$ **Heaviside (unit) step function**





This is called the *decision boundary*

These functions are not differentiable, continuous, or smooth

Sigmoid Activation Function

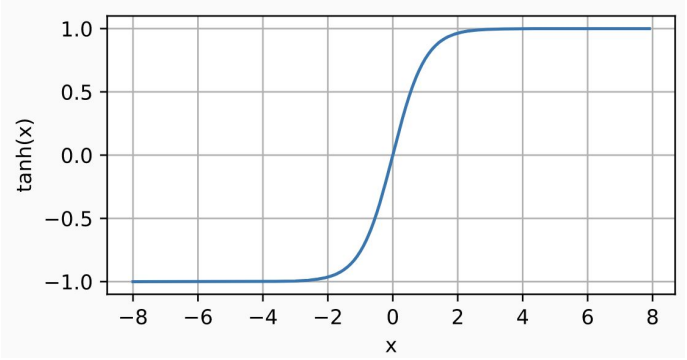
Sigmoid activation functions were the most common before 2012:

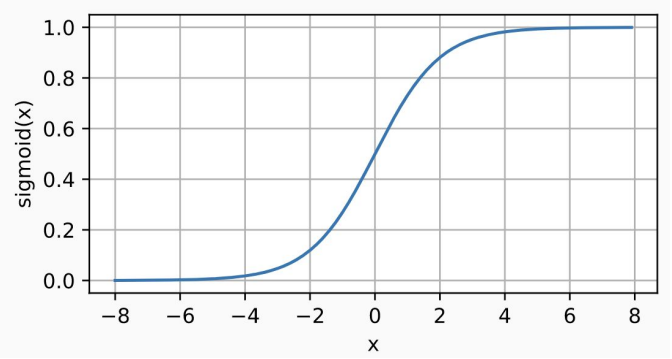
- Easily differentiable, smooth, continuous
- Range between [-1, 1] or [0, 1]

There are *many* sigmoid functions, the most common are:

$$f(x) = \tanh(x)$$ **Hyperbolic tangent**

$$f(x) = \frac{1}{1 + e^{-x}}$$ **Logistic function**





Sigmoid Activation Function

Sigmoid activation functions were the most common before 2012:

- Easily differentiable, smooth, continuous
- Range between [-1, 1] or [0, 1]

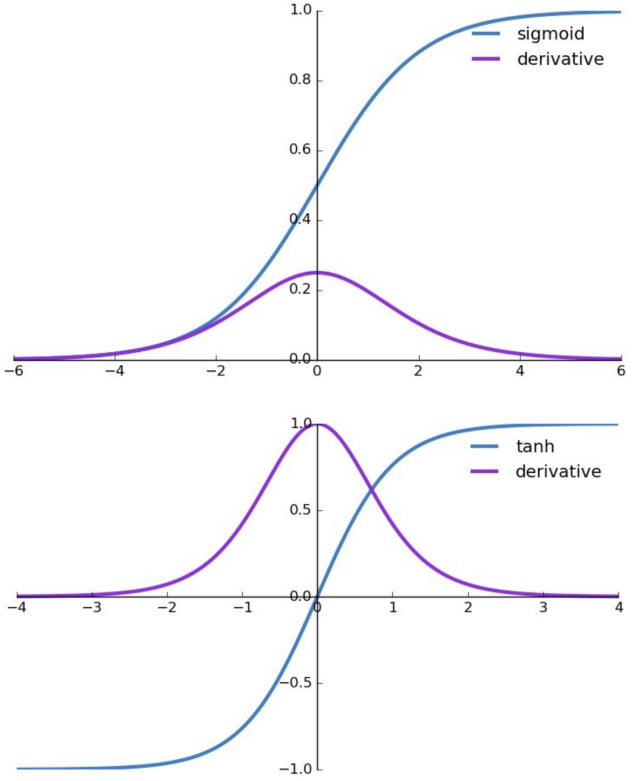
There are *many* sigmoid functions, the most common are:

$$f(x) = \tanh(x)$$ **Hyperbolic tangent**

$$f(x) = \frac{1}{1 + e^{-x}}$$ **Logistic function**

Saturated neurons "kill" the gradients

Gradients become vanishingly small very quickly away from x=0

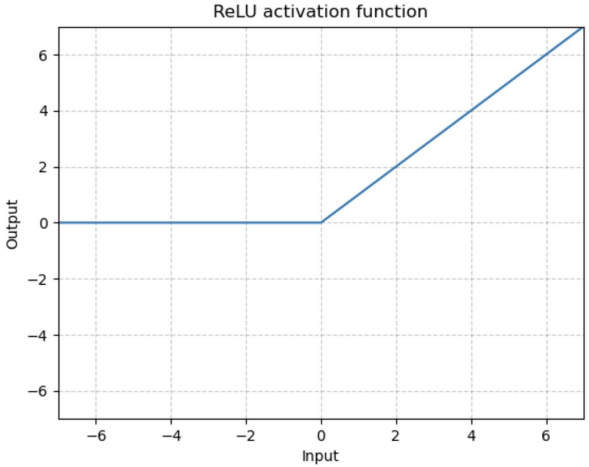


ReLU Activation Function

Modern deep learning typically uses the Rectified Linear Unit (ReLU) based activation functions:

$$\mathrm{ReLU}(x) = (x)^+ = \max(0, x)$$

ReLU



ReLU Activation Function

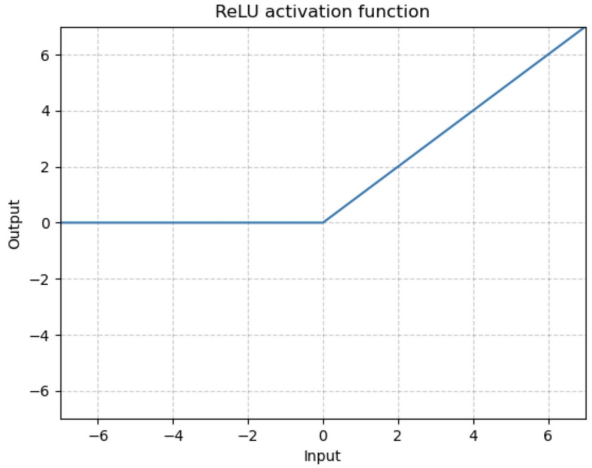
Modern deep learning typically uses the Rectified Linear Unit (ReLU) based activation functions:

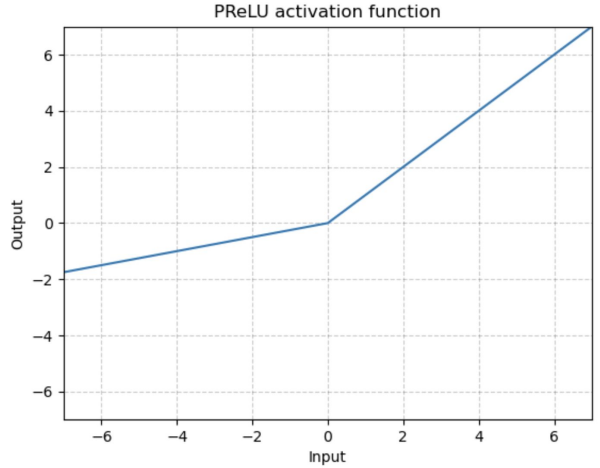
$$\mathrm{ReLU}(x) = (x)^+ = \max(0, x)$$ **ReLU**

$$\mathrm{LeakyRELU}(x) = \begin{cases} x, & \text{if } x \geq 0 \\ \text{negative\_slope} \times x, & \text{otherwise} \end{cases}$$ **Leaky ReLU**

$$\mathrm{PReLU}(x) = \begin{cases} x, & \text{if } x \geq 0 \\ ax, & \text{otherwise} \end{cases}$$ **Parametric ReLU**

Very easy derivatives 0 or 1, use 0 at x=0





Continuous Approximations of ReLU

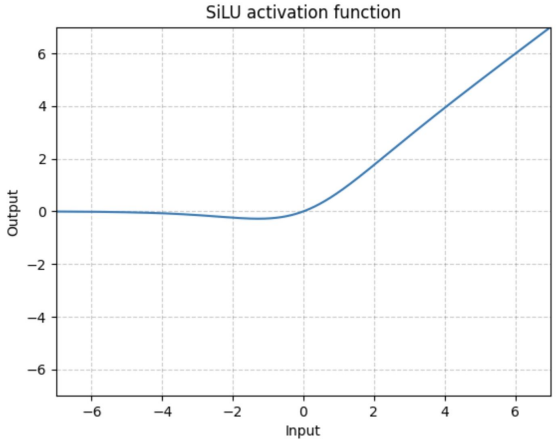
We can approximate ReLU activation by continuous functions:

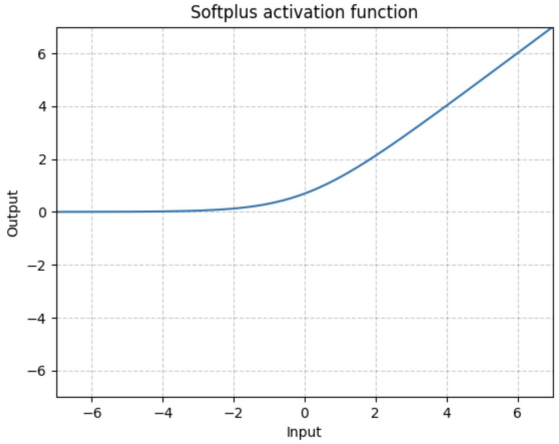
$$\text{SiLU}(x) = x.\sigma(x) = \frac{x}{1+e^{-x}}$$

SiLU (Swish)

$$\text{SoftPlus}(x) = \frac{1}{\beta}\log\left(1+e^{\beta x}\right)$$

SoftPlus





Work on par or better than ReLU functions

Questions?