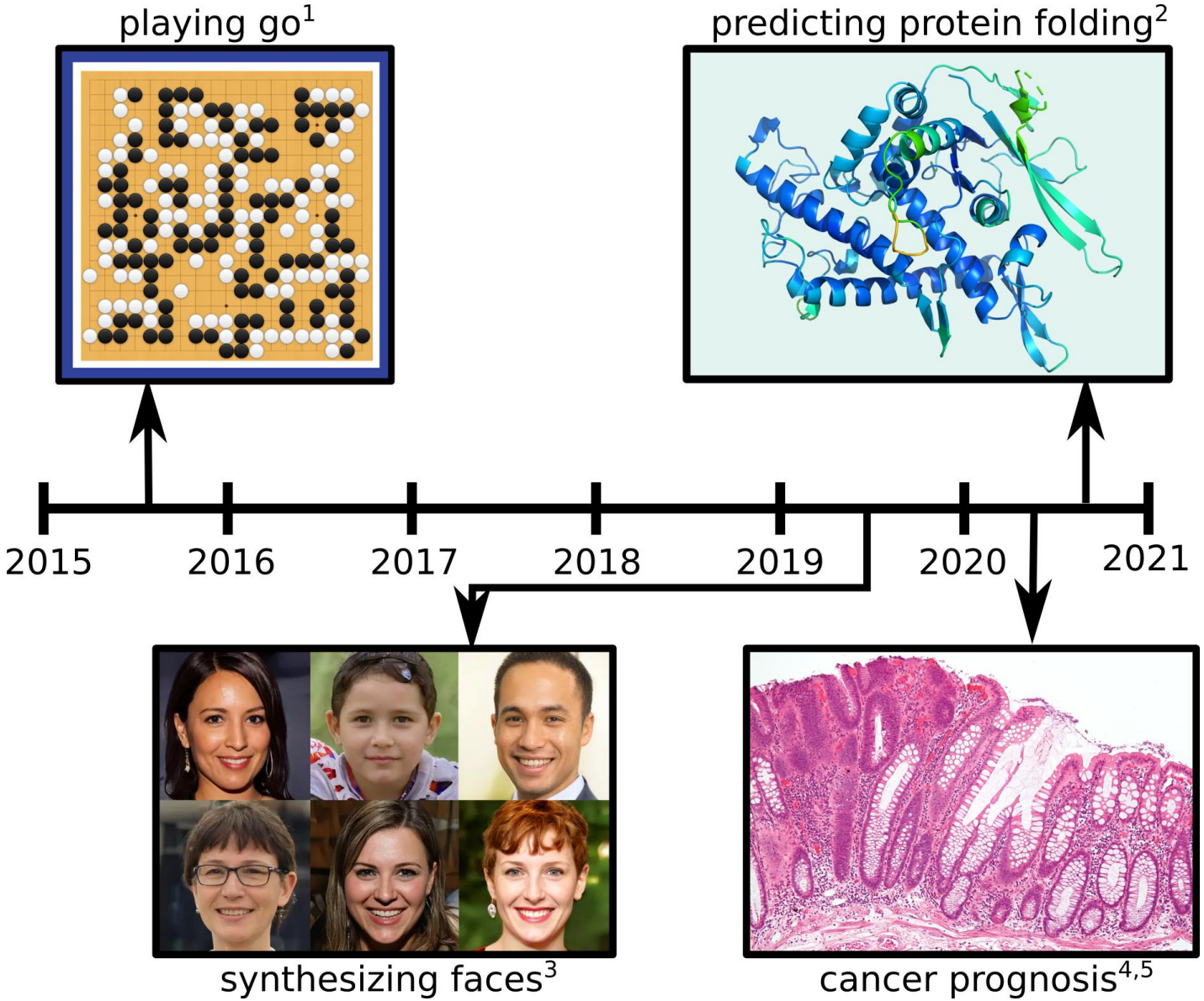
Machine Learning Fundamentals

Computational Physiology Summer School 2021

Gabriel Balaban

Machine Learning techniques have powered many recent technological breakthroughs



1) https://deepmind.com/research/case-studies/alphago-the-story-so-far
2) https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology
3) Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks.
4) Skrede OJ et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study.
5) Image source: Nephron, https://commons.wikimedia.org/w/index.php?curid=8273894

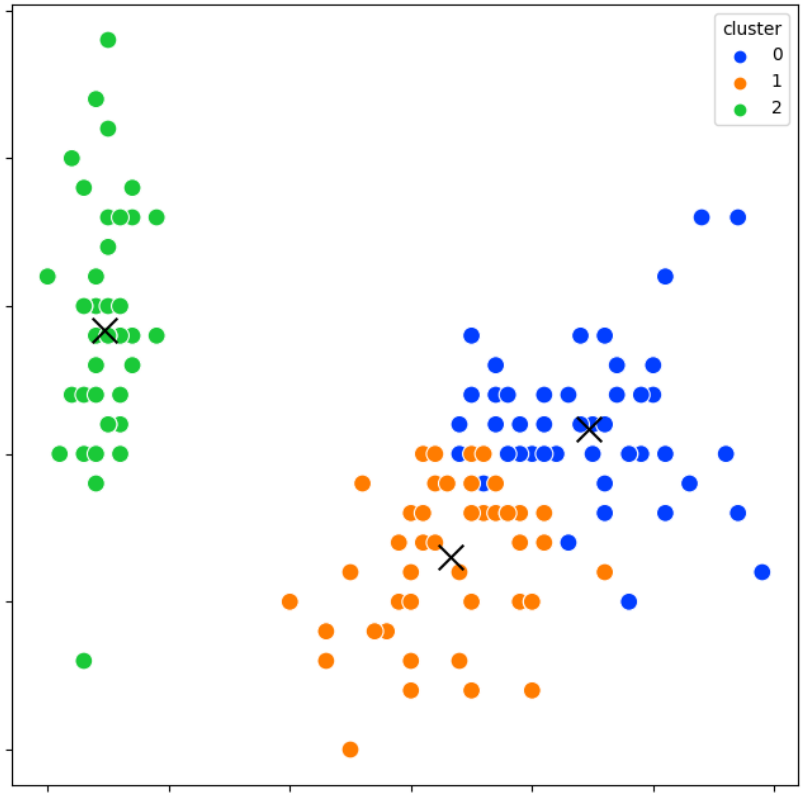
Definition:

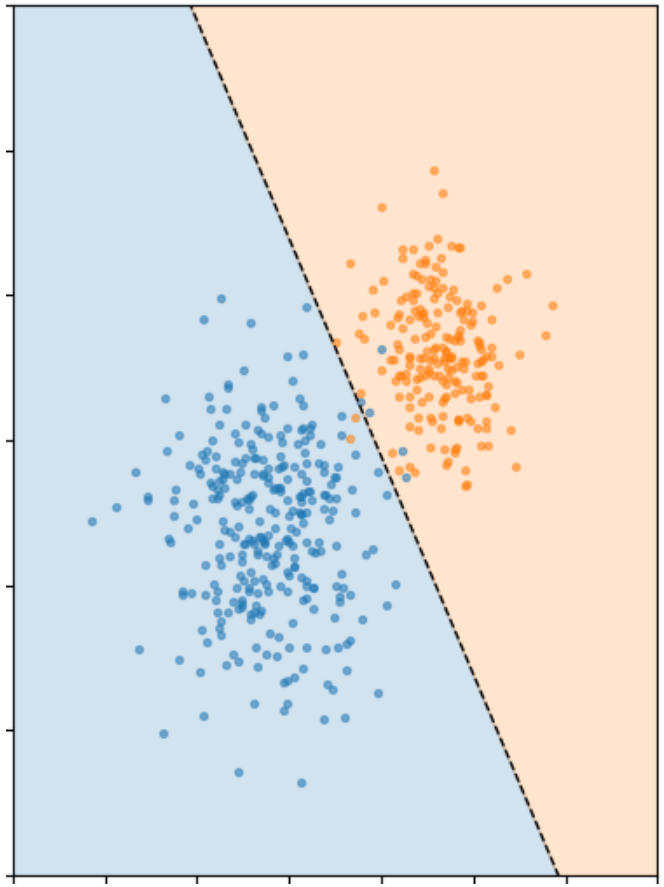
Machine learning is a branch of artificial intelligence focused on building applications that learn from data and improve their accuracy over time without being programmed to do so.

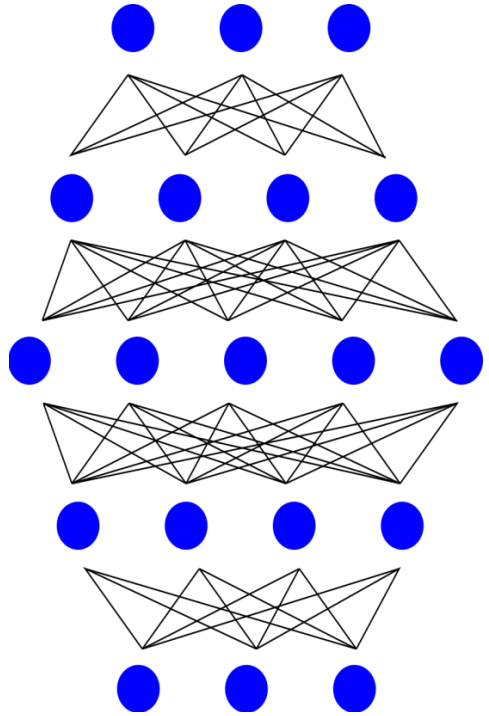
Spam filter: Hard to program but straight-forward to train using machine learning

In this module you will learn about 3 fundamental machine learning algorithms

K-means clustering	Logistic regression	Feed-forward neural networks







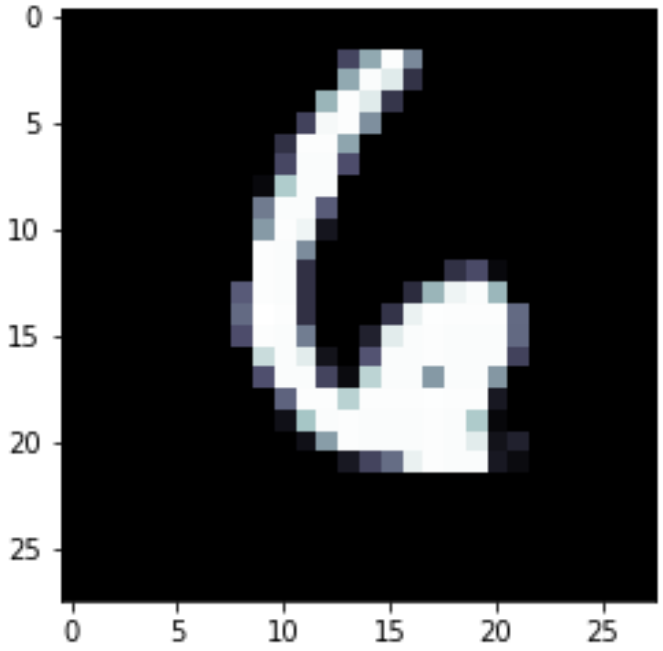
To complete the exercises you will need basic data manipulation skills for tabular and image data

Tabular Data

	Name	Sex	Fare
0	Braund, Mr. Owen Harris	male	7.2500
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	71.2833
2	Heikkinen, Miss. Laina	female	7.9250
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	53.1000
4	Allen, Mr. William Henry	male	8.0500
...	...	...	...
886	Montvila, Rev. Juozas	male	13.0000
887	Graham, Miss. Margaret Edith	female	30.0000
888	Johnston, Miss. Catherine Helen "Carrie"	female	23.4500
889	Behr, Mr. Karl Howell	male	30.0000
890	Dooley, Mr. Patrick	male	7.7500

pandas https://pandas.pydata.org/

Image Data

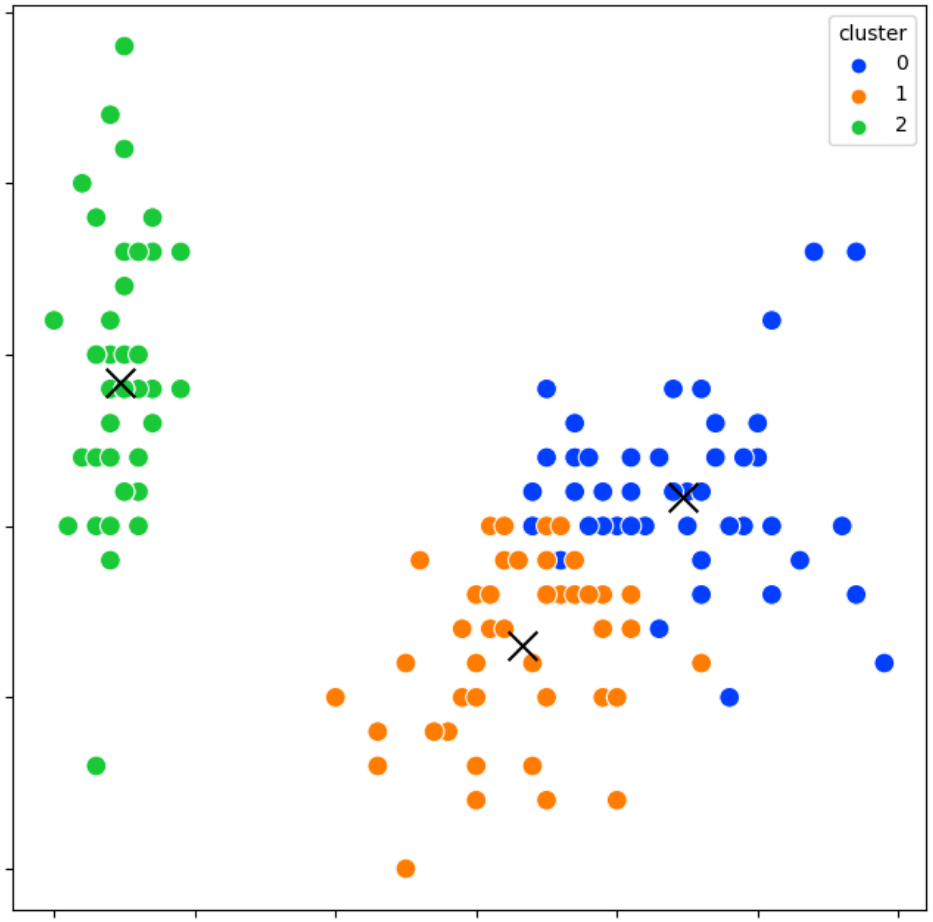


matplotlib https://matplotlib.org/

There are a few other python packages that are used in the exercises. It can help you to get familiar with them before the 3 hour practice session.

Package Name	Website	Function
Scikit-learn	https://scikit-learn.org/stable/	Basic machine learning tools and algorithms
pytorch	https://pytorch.org/	Deep learning / neural networks
seaborn	https://seaborn.pydata.org/	Statistical data visualization
numpy	www.numpy.org	Array computations
scipy	www.scipy.org	Scientific computing algorithms (optimization)

K-means clustering is a fundamental algorithm to find clusters in data



Given X = data matrix (num datapoints x num features)

Want class membership k = vector of (num datapoints)

K-means algorithm

Initialize cluster centres $X_c^l \quad l = \{1 \ldots k\}$

Iterate until convergence

1) $k_i = \underset{l}{\operatorname{argmin}} \|X_i - X_c^l\|$

2) $X_c^l = \frac{1}{\#\{k_i = l\}} \sum_{i, k_i = l} X_i$

Logistic regression is a fundamental supervised machine learning algorithm for making binary predictions

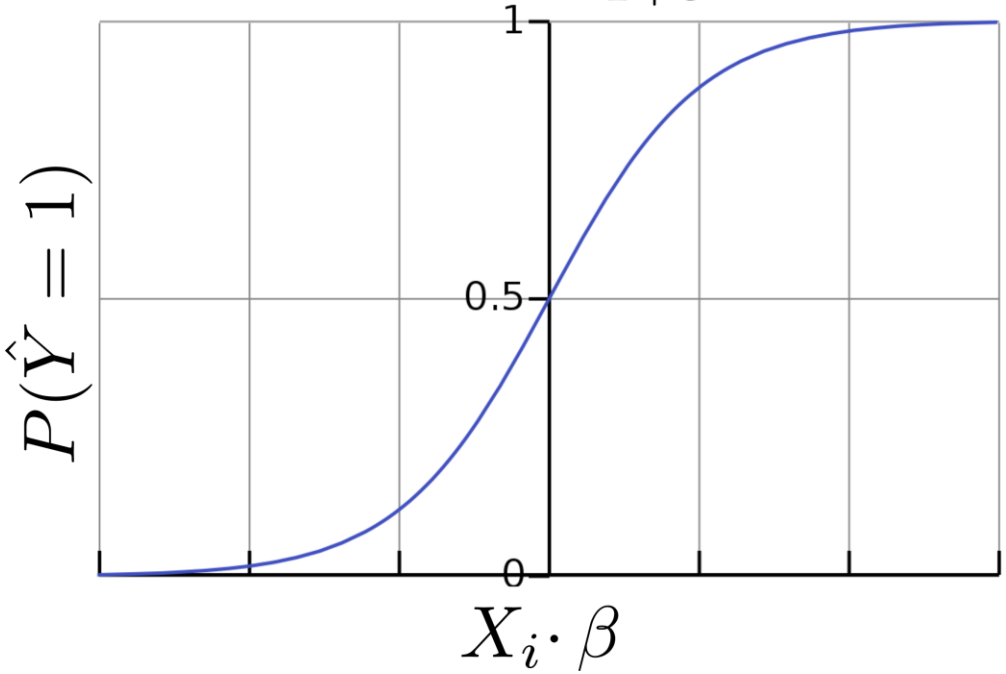
Given X = data matrix (num datapoints x num features)

Y = binary data labels (vector of num features)

Want β = regression coefficients (vector of num features)

$\hat{Y}$ = model predictions

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

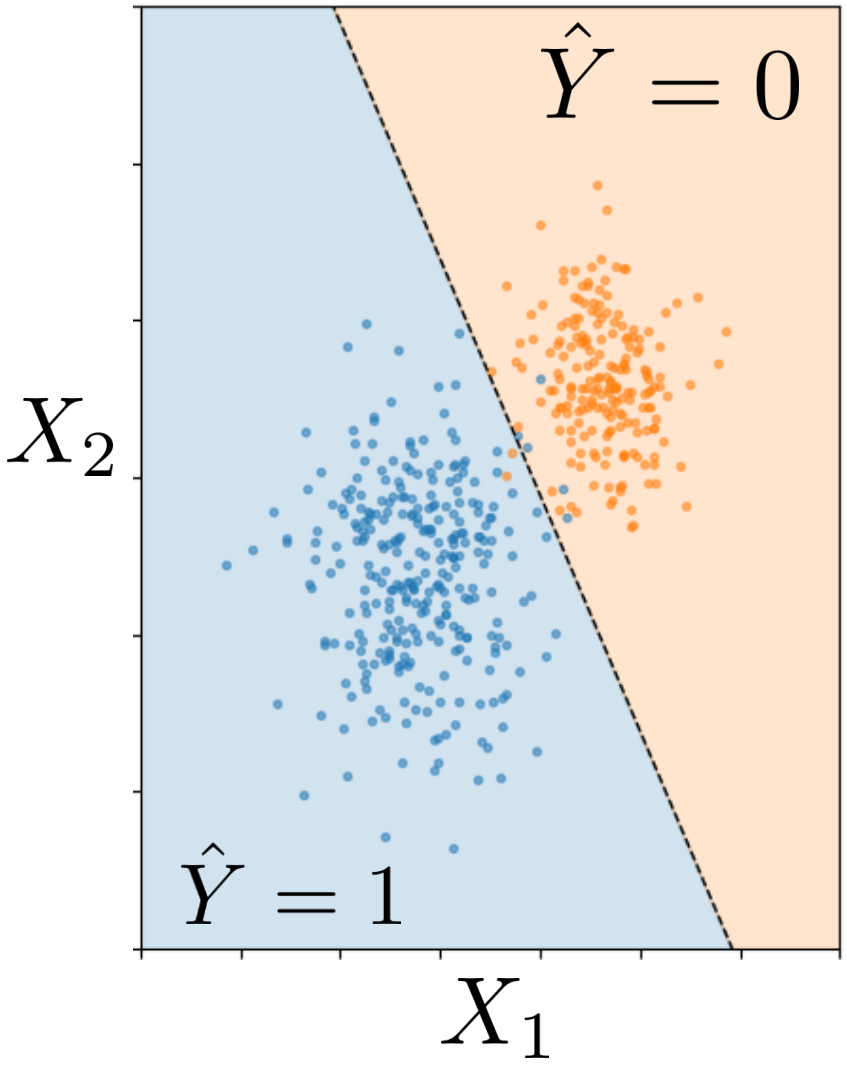


Training algorithm (minimum cross-entropy)

$\min_{\beta} l(\beta, Y, X)$ Optimization problem

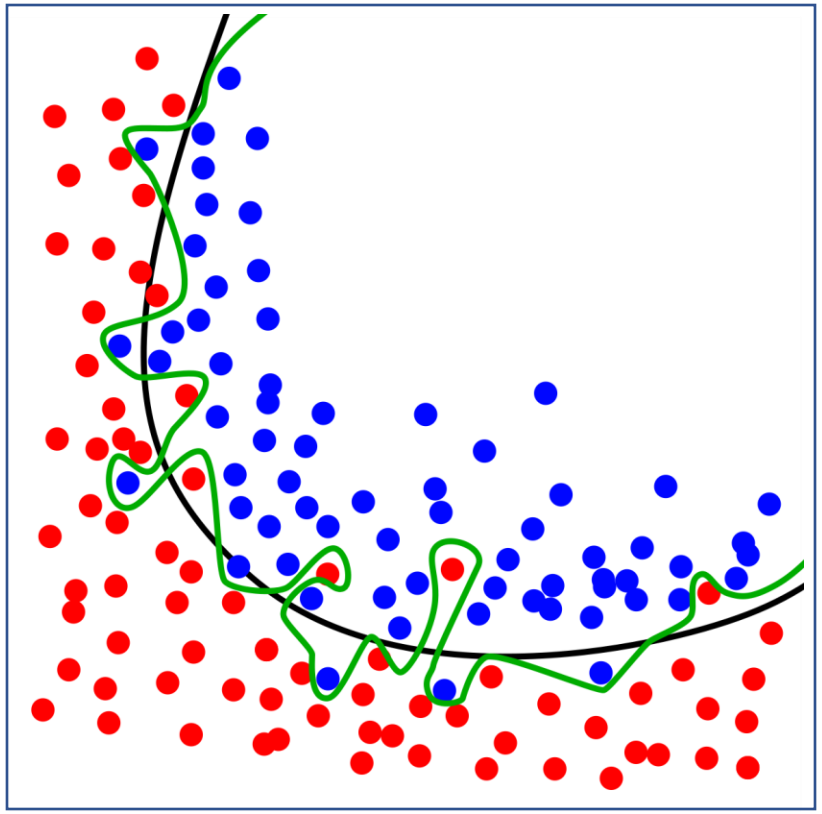
$l(\beta, Y, X) = -1 * \sum_{i=1}^{N_{data}} y_i \log(\hat{Y}_i)) + (1 - y_i) \log(1 - \hat{Y}_i)$ cross entropy loss function

Logistic regression creates linear decision boundaries

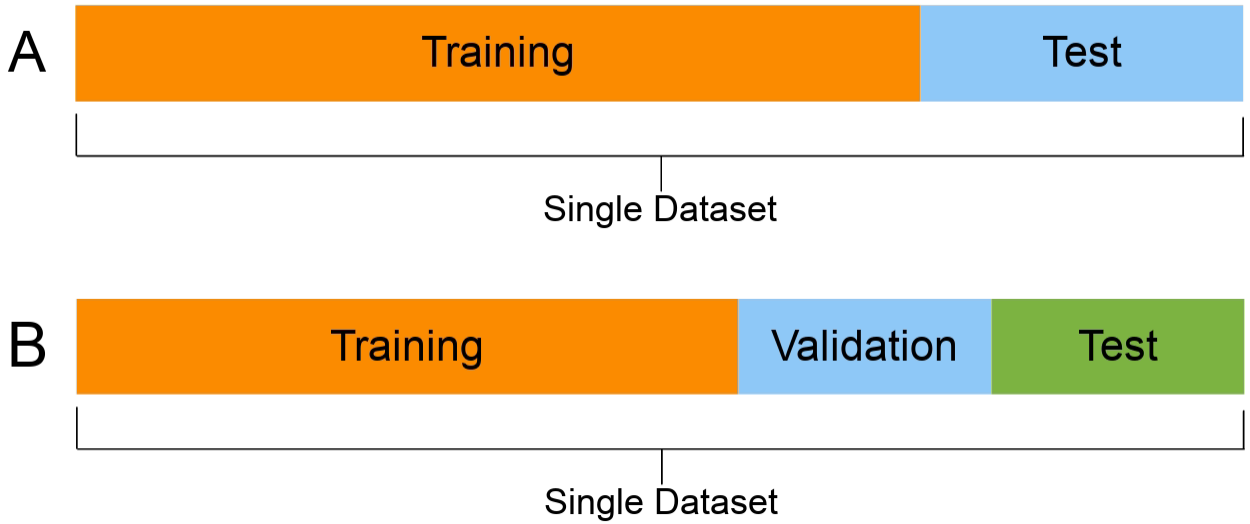


Over-fitting can occur when the number of features is too high in relation to the number of data-points

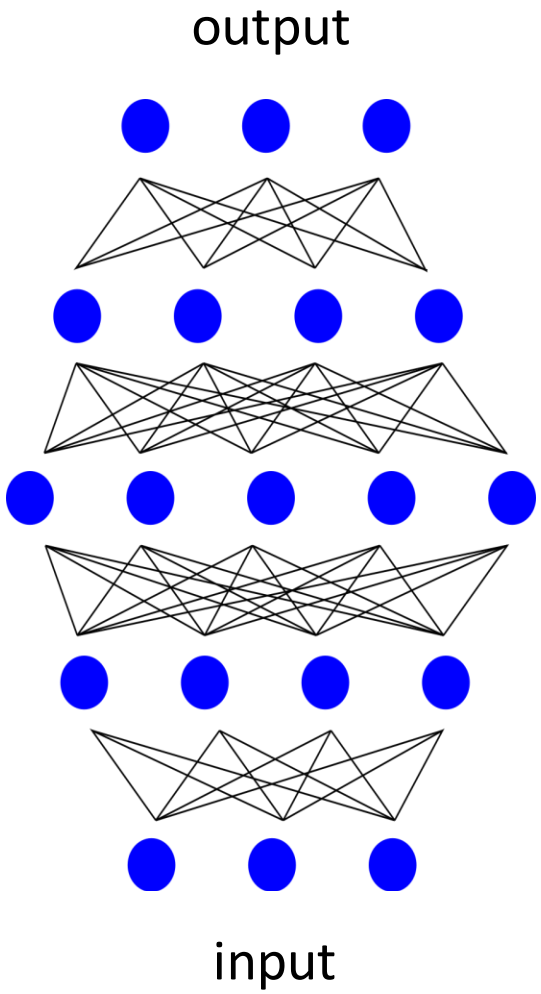
The problem



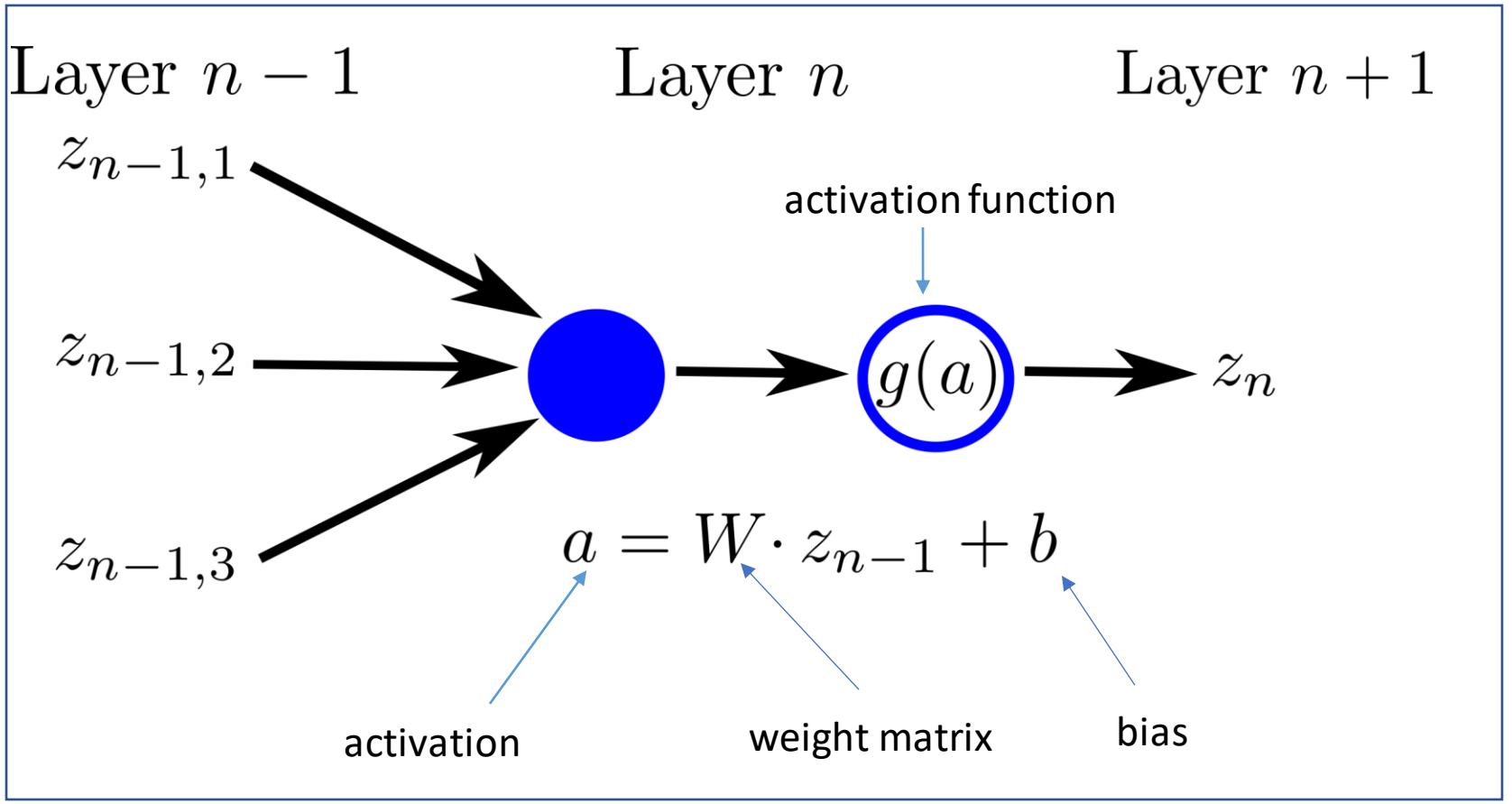
The solution: use a validation set to compare performance



Neural networks are flexible function approximators consisting of a directed graph with artificial neurons



How information is transformed at a neuron



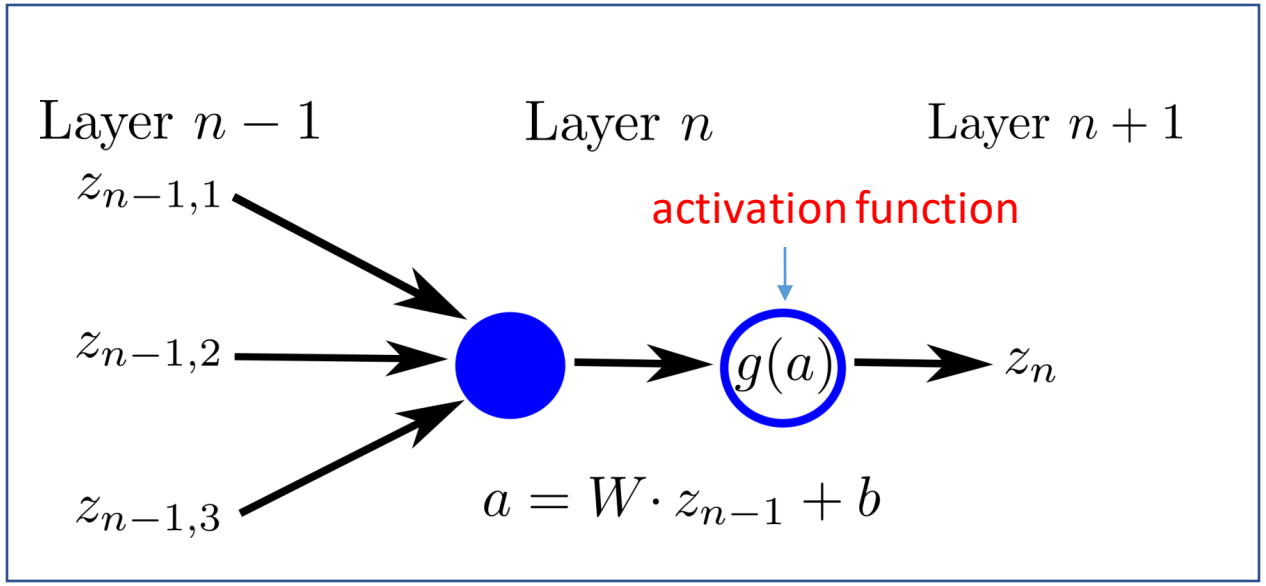
The universal approximation theorem helps explain why neural networks are so powerful

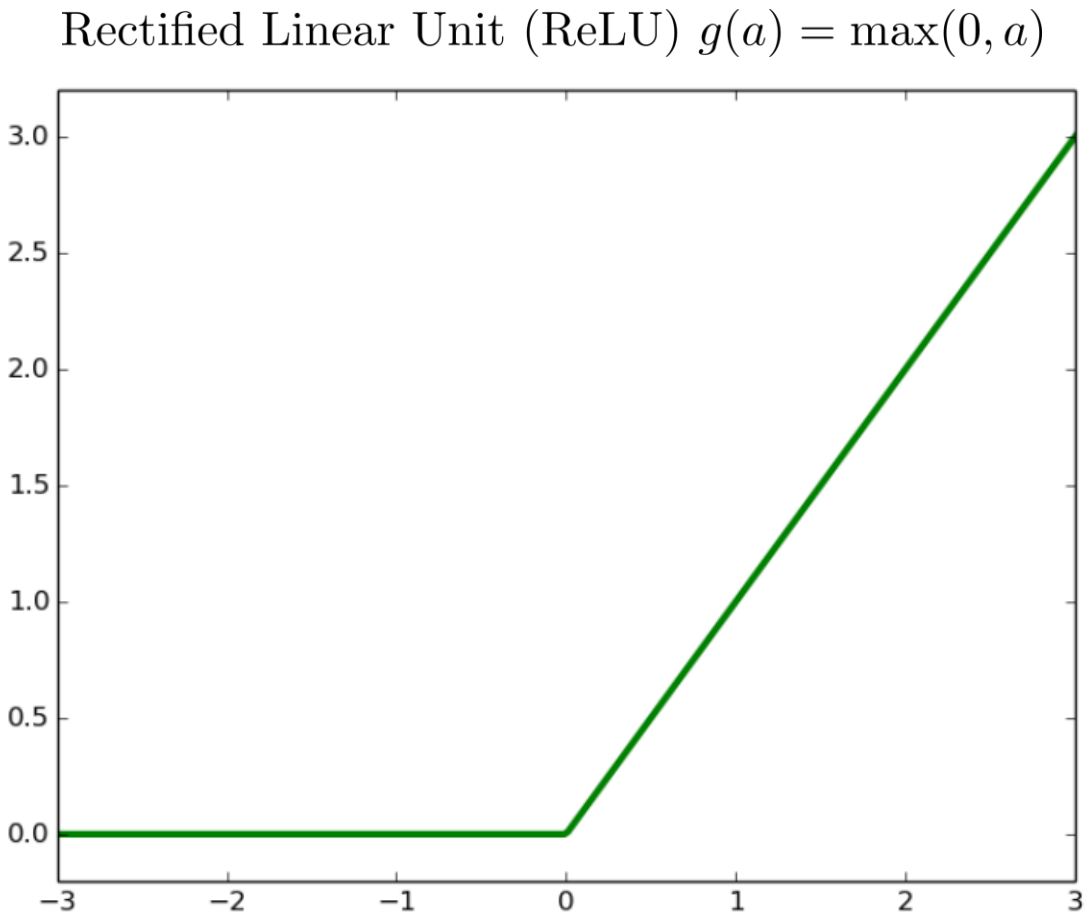
Universal approximation theorem (a version)

Let $g(\cdot)$ be a continuous function on a m-dimensional hypercube $[0,1]^m$. Let $a(\cdot)$ be a non-constant, bounded, continuous (activation) function. Then $g(\cdot)$ can be approximated arbitrarily well, that is, for every maximal deviation $\epsilon > 0$, there exists a set of weights u_i, w_i and biases b_i such that:

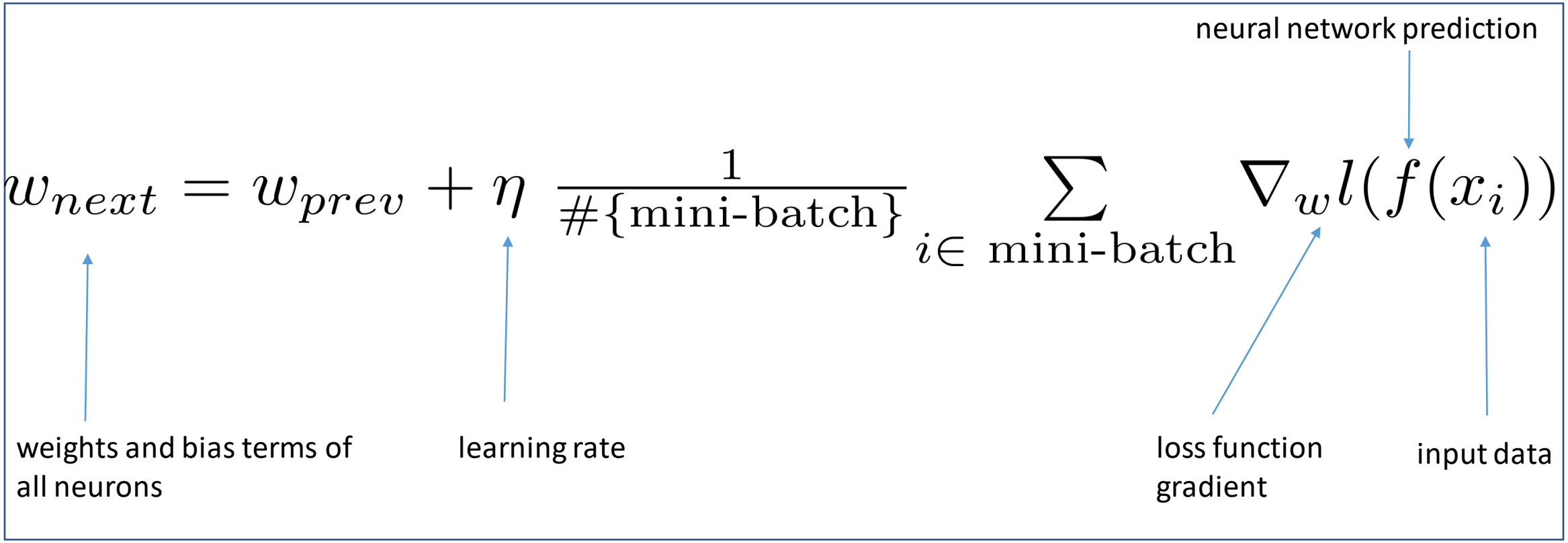
$$\forall x \in [0,1]^m : \ |g(x) - (\sum_{i=1} u_i a(w_i x + b_i) + b)| < \epsilon$$

Rectified linear units are a good choice of activation function





Neural networks can be trained with stochatic gradient descent using small subsets of the data called "mini-batches"



Terminology: Going through the data once is called an "epoch" of training

In the excercises I have provided a lot of code for you to get you started. I have also given you a few specific coding tasks to help you learn.

Please google bits of code that you are curious about and feel free to experiment/explore.

Good luck and have fun experimenting with machine learning.
I'll see you at the zoom session!

K-means clustering	Logistic regression	Feed-forward neural networks