Special Guest Lecture

Part 1: Big data, deep neural networks, and translational applications

Dr. James Hong

Big data

- No singular definition

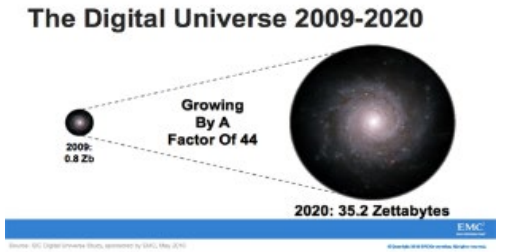
- Big data:
  - Scale
  - Diversity
  - Complexity
  - Requires new architecture and methods to manage and extract value from it

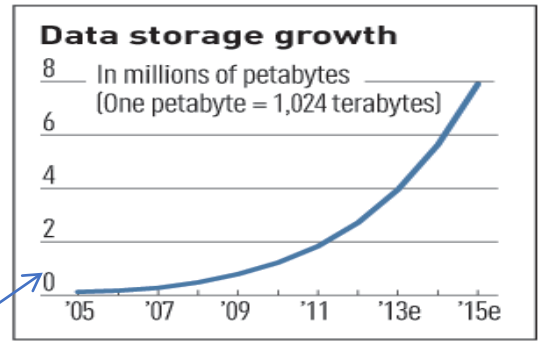
Big data has several characteristics

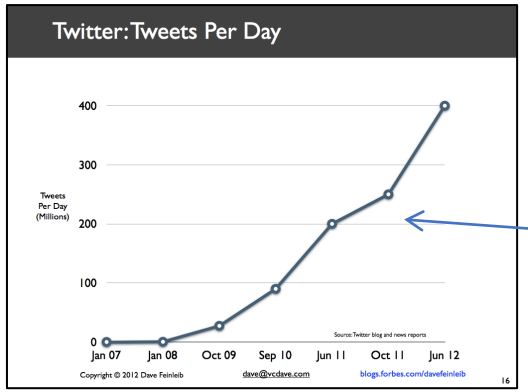
- Some refer to these as the “V’s”
- There are generally 10 V’s but we will focus on:
  - Volume
  - Variety
  - Velocity
  - Veracity

Volume

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially



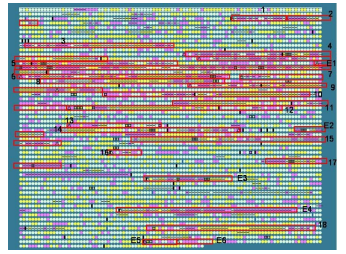


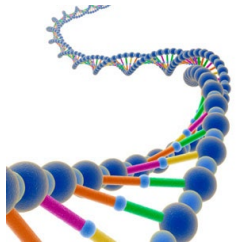


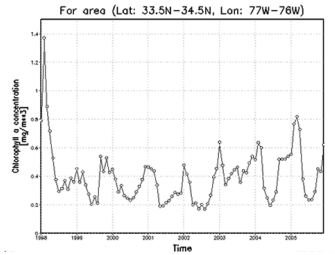
Exponential increase in collected/generated data

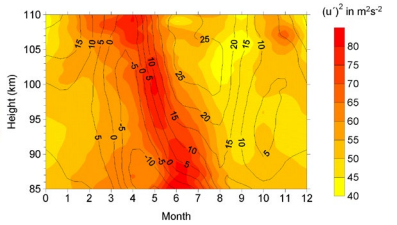
Variety

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data









To extract knowledge➔ all these types of data need to linked together

Velocity

- Data is being generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- Examples
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Veracity

- Confidence drops as data volume goes up
- Uncertainty due to data inconsistency
- Incompleteness
- Ambiguities
- Latency
- Model approximations

Model has changed for data generation/usage

Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data

Generation of big data in science and medicine

- Static imaging data (CT/MRI/Histology)
- Dynamic video data (Behaviour, Cell Migration Data)
- Static text data (OMICs data – RNASeq, Proteomics, Epigenomics, genomics, patient demographics)
- Dynamic text data (Sensors, electrophysiology, respiratory readouts)
- You have to often associate these to one another to gain more insight
    - E.g. Genetic predisposition of an Asian male to respiratory complications after COVID

Two main challenges in big data

- Storage (not our focus today)
  - Cloud storage in large server farms
- **Analysis**
  - Connecting and integrating multi-modal data
  - Real-time processing
  - Integration into the cloud storage framework
  - We need an “intelligent” system that can function in a similar manner to a human

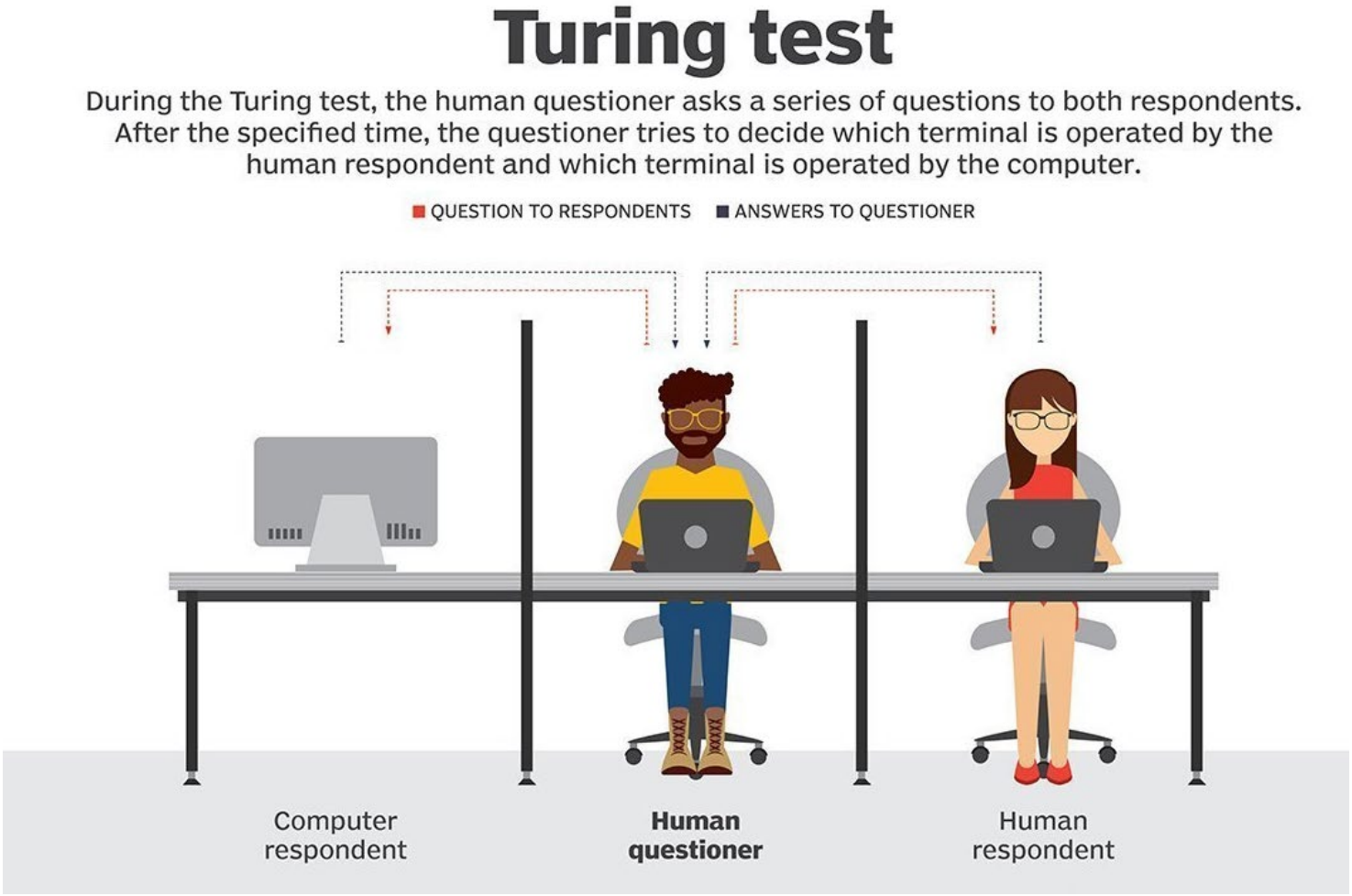
Human intellect

- What makes us smart?
  - Billions of years of evolution
  - The “dominant” species
  - We *often* learn from our mistakes
  - Adapt ourselves in the environment to perform better

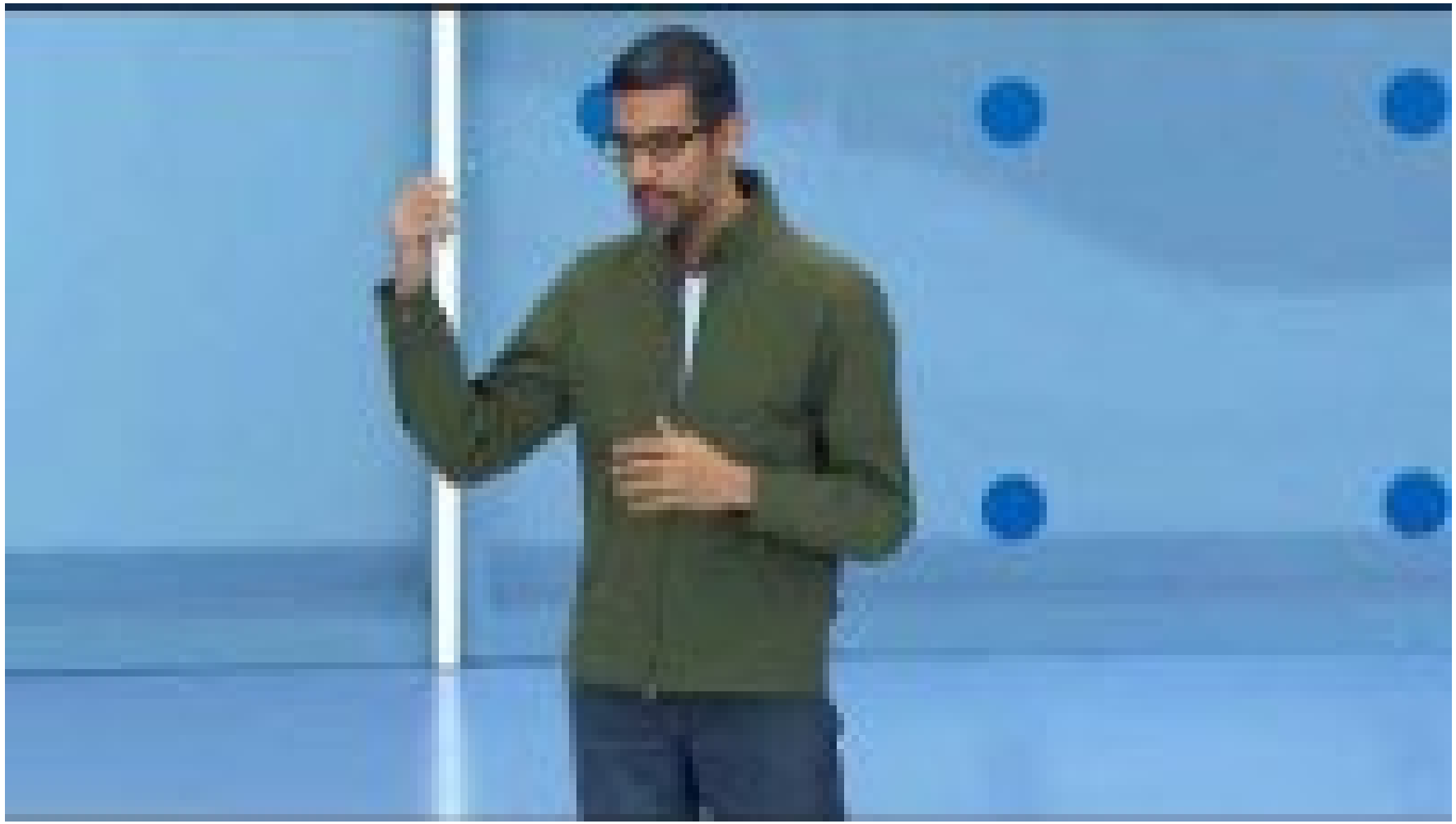
Definition of artificial intelligence

- Definition of intelligence:
  - “The ability to acquire and apply knowledge and skills”
- AI is defined as the capability of a machine to imitate intelligent human behavior
- AI uses computers to model intelligent behaviour with minimal human intervention
- AI should be able to store, compute, and learn

Turing test to differentiate humans from AI



Google Duplex makes a haircut appointment



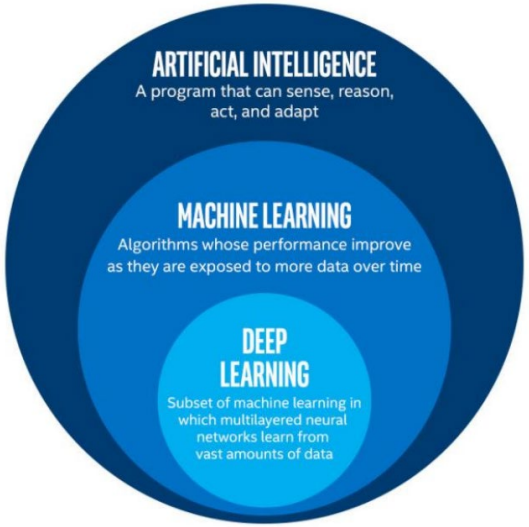
How do we model intelligent behaviour?

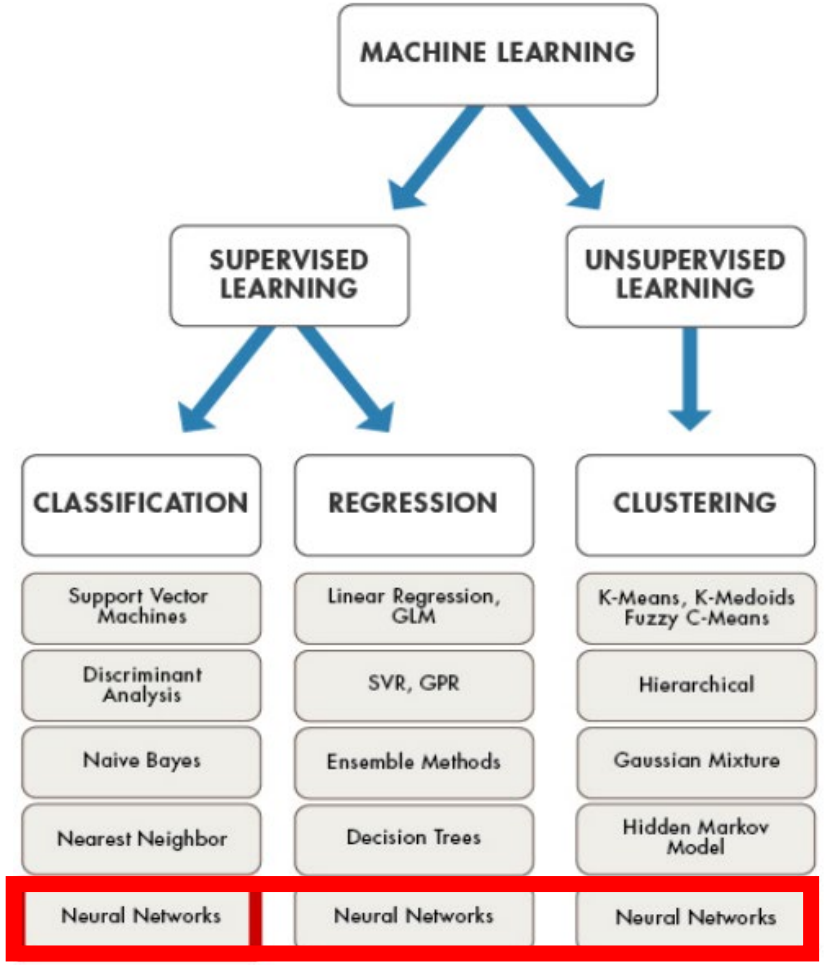
- Start with the brain
- Brain function as a result of an immense number of neurons that fire signals
- Neurons connect through synapses that propagate electrical impulses by releasing neurotransmitters
- Synaptic plasticity is the activity-dependent changes in the effectiveness of synapses, or how synapse can alter the “strength” of their connections – allows for **learning**

Neuroscience has been critical for AI

- An artificial neural network (ANN) mimics the biological brain by acquiring knowledge through learning
- It then stores this knowledge by adjusting the weights within the network (the strengths of the connection)
- In 1949, a neuroscientist Hebb describes that if one neuron repeatedly stimulates a second neuron, the connection between them would strengthen (the concept of “potentiation”)

Nomenclature

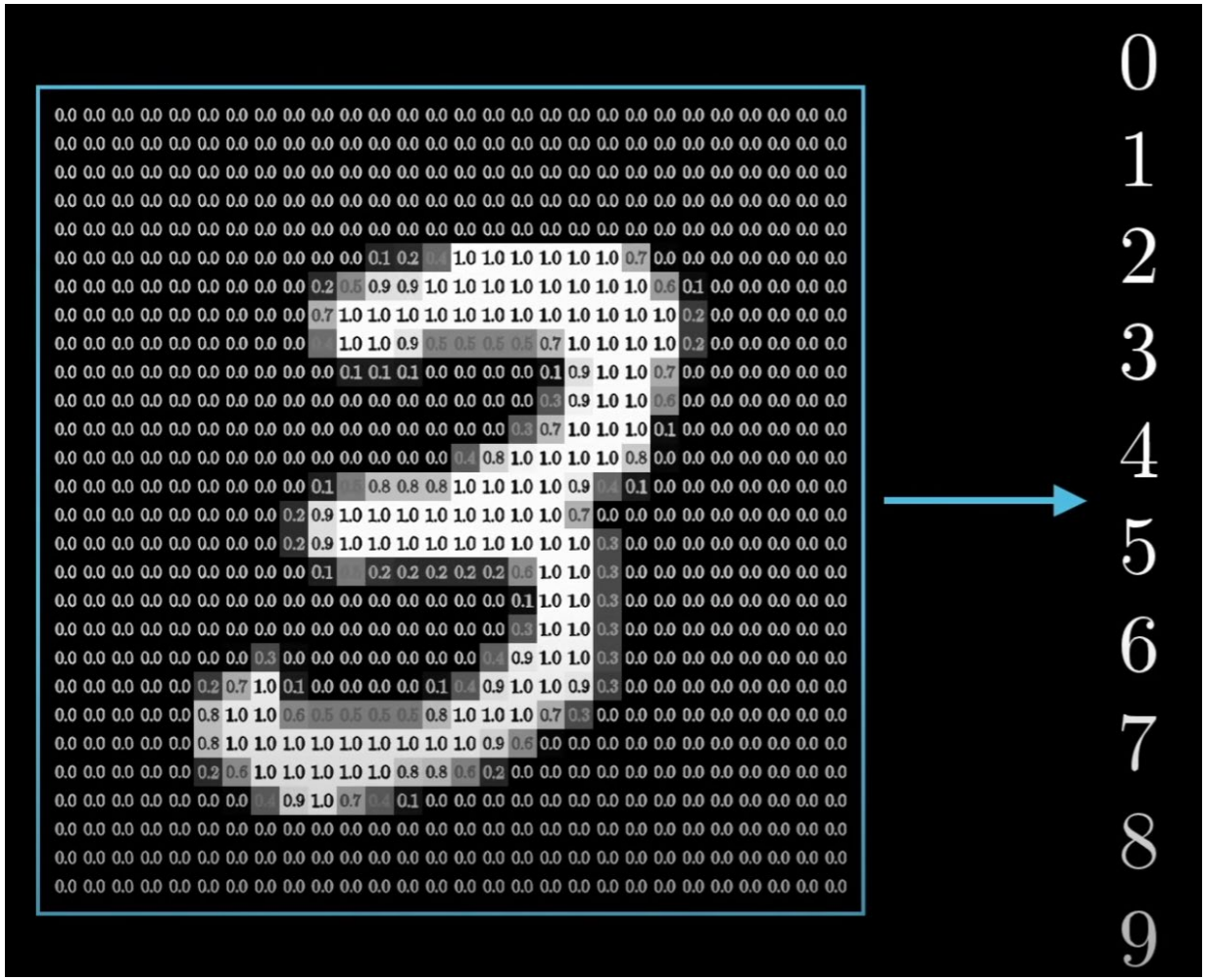




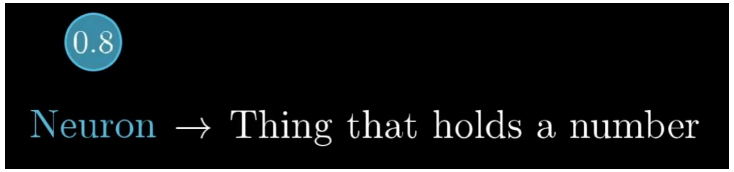
Crash course on neural networks

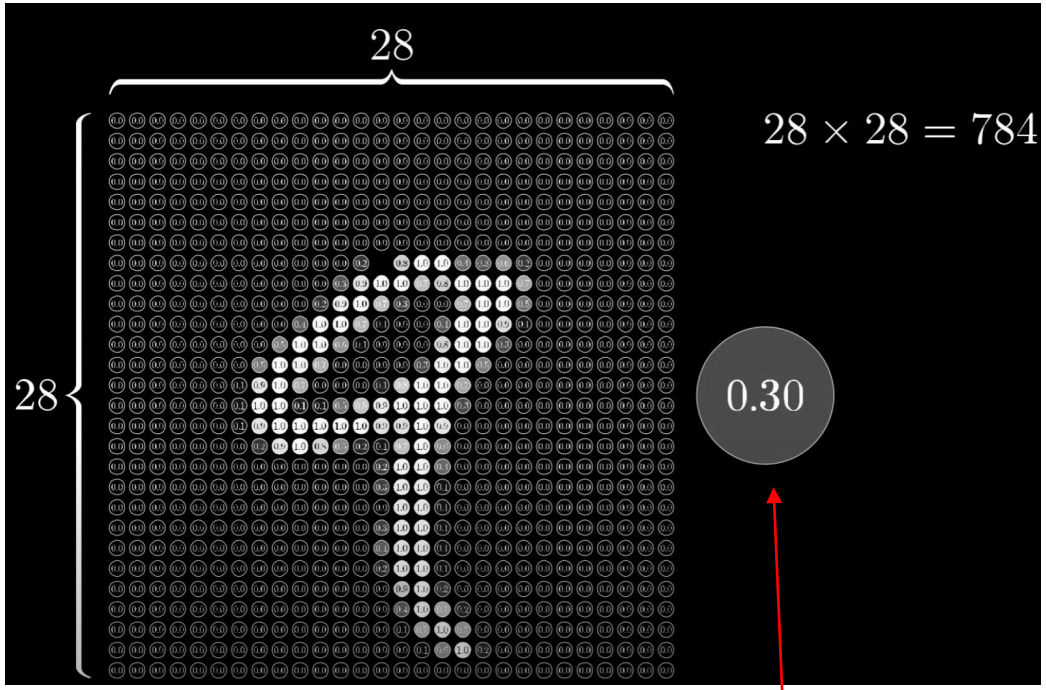
How does our brain know how to tell numbers?

How can we tell a machine how to tell numbers?



Neurons = Nodes

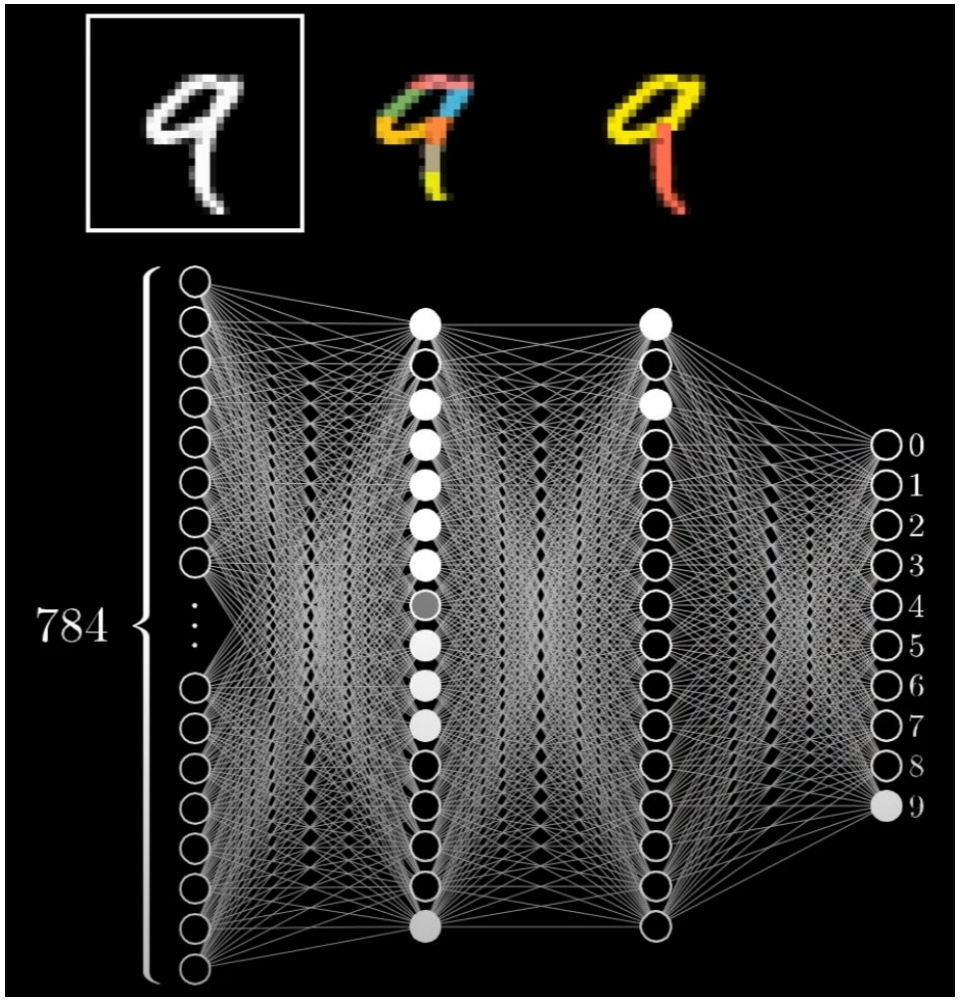




“Activation”
The higher this number, the more likely it will light up

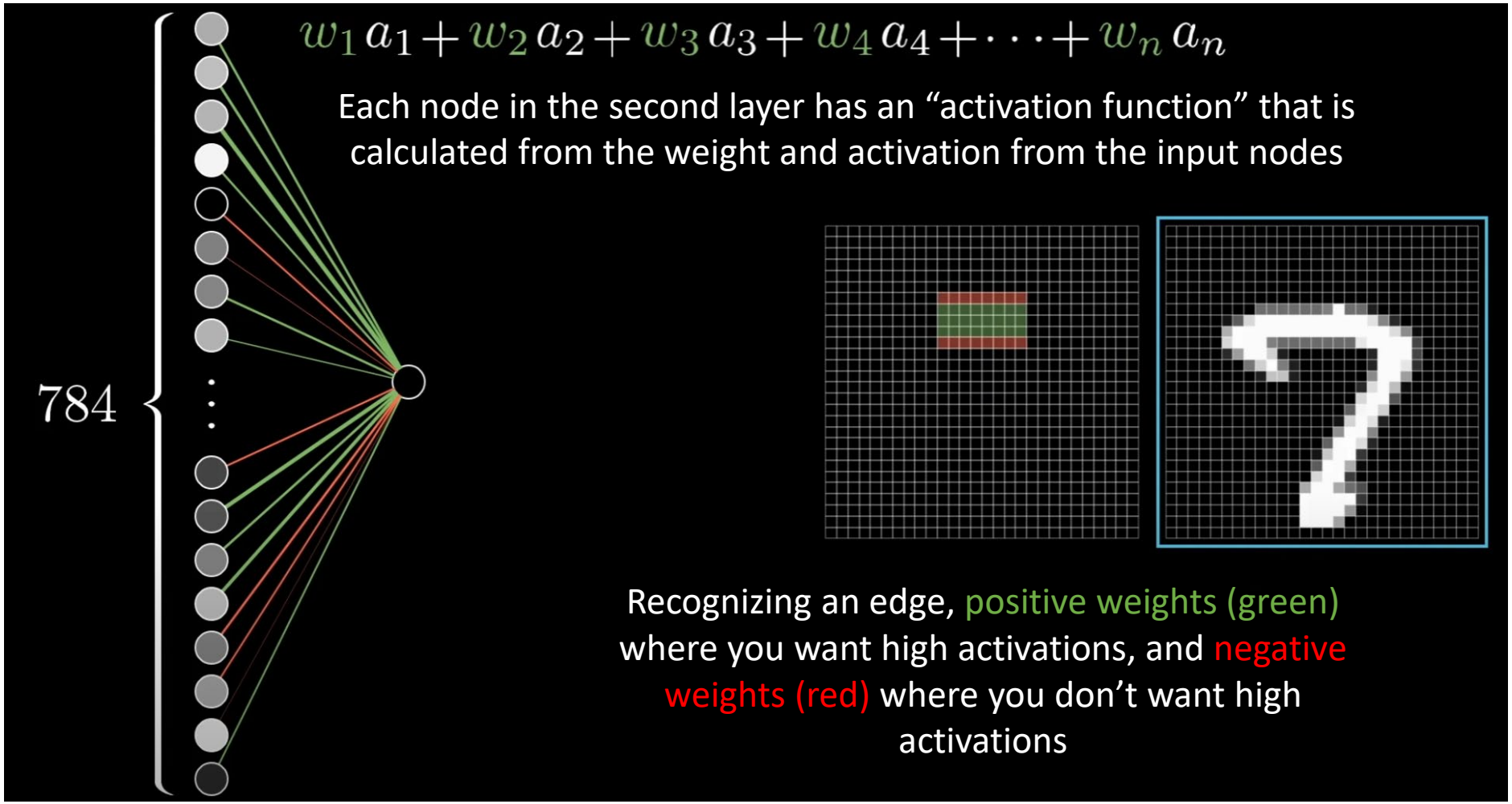
Layers of abstraction

28 x 28 pixel
= 784 units

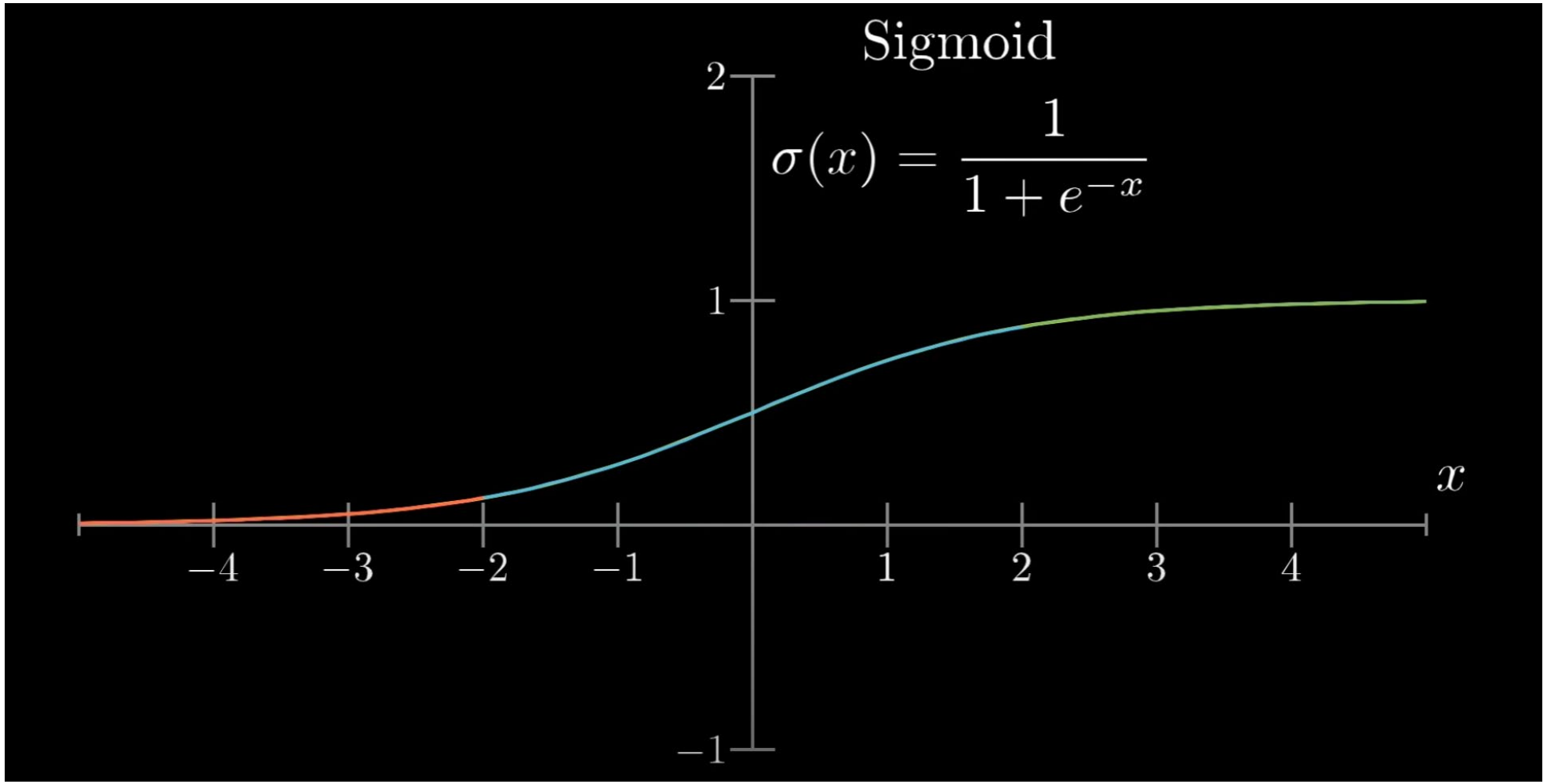


“hidden layers”

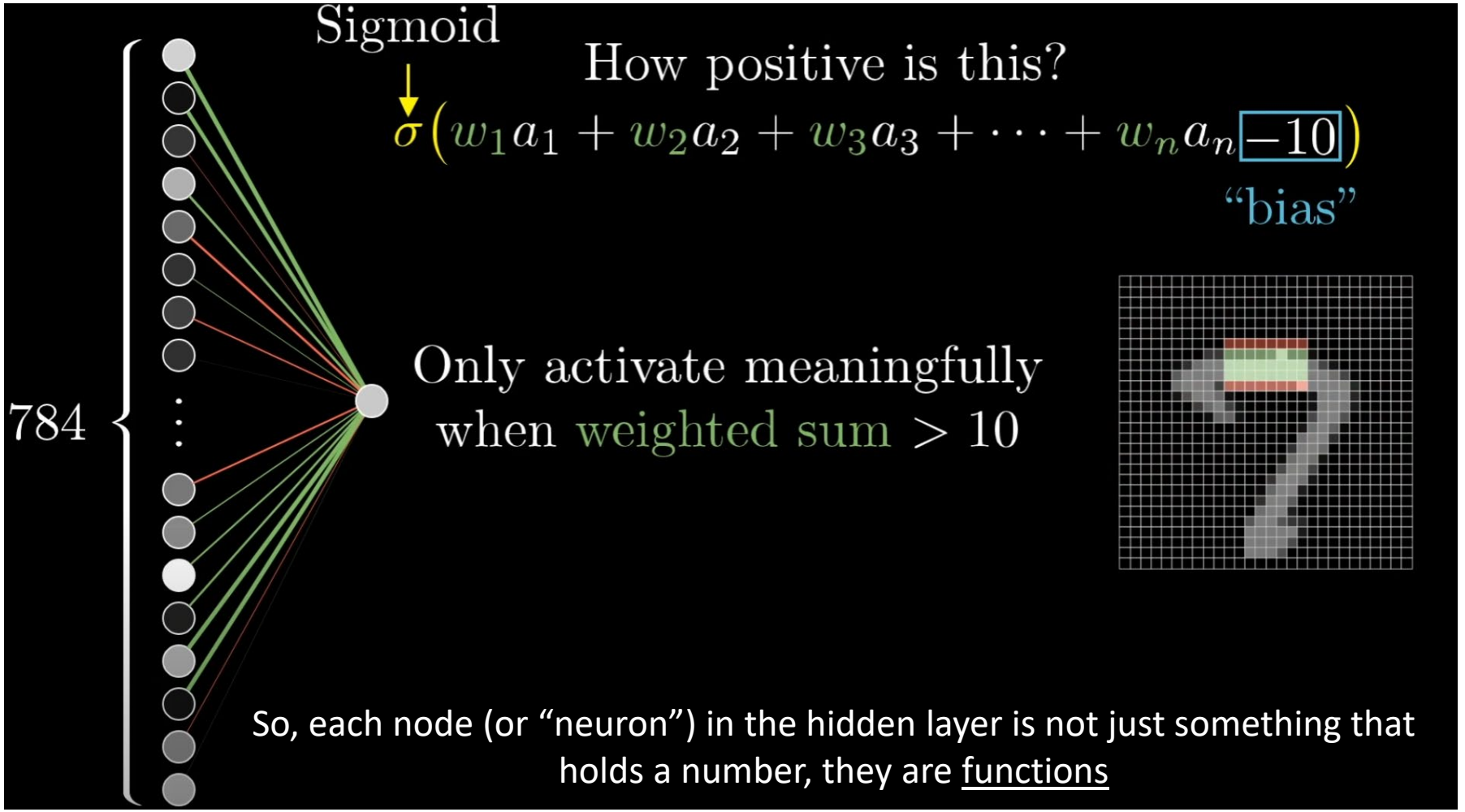
How do you program this?



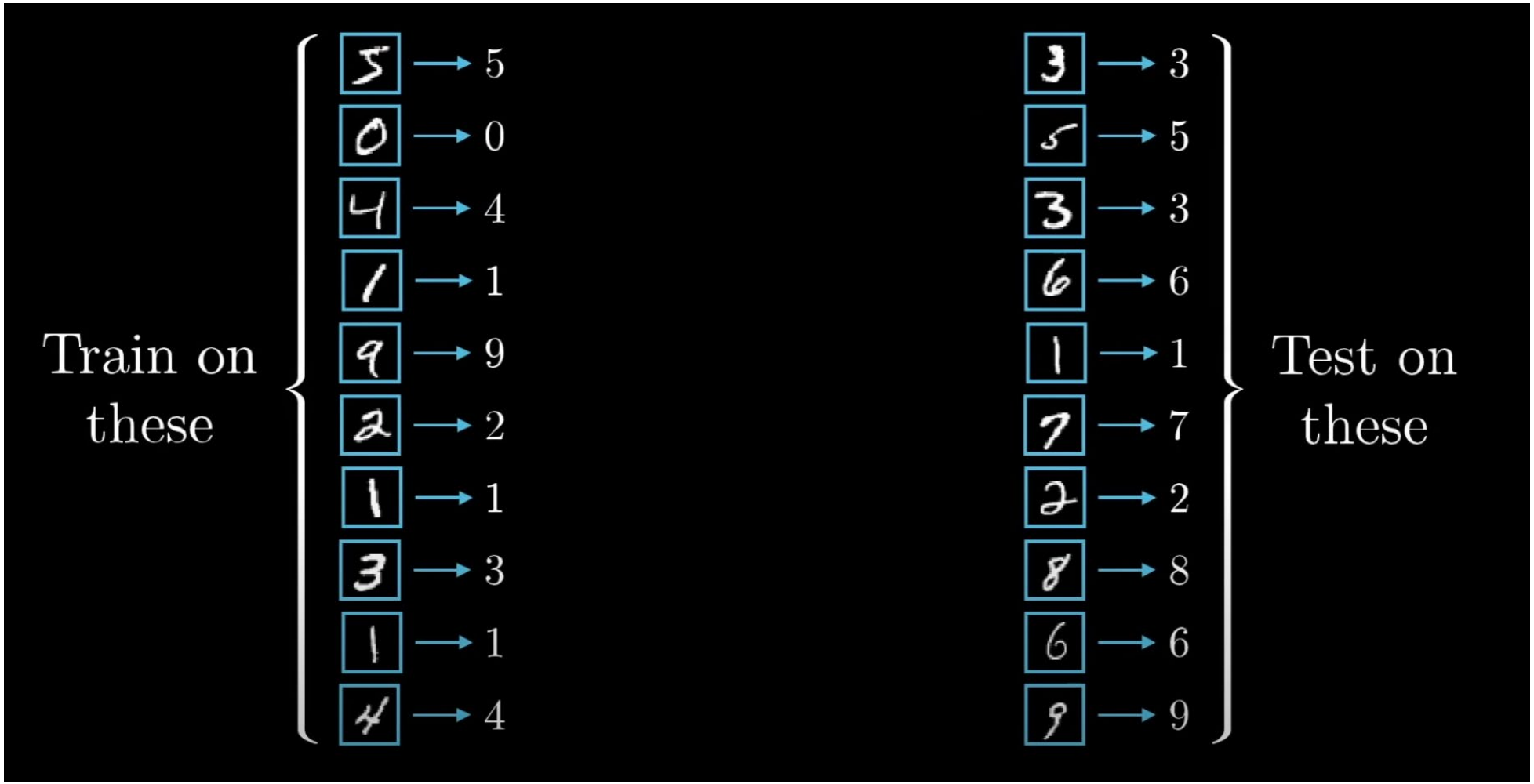
Use a sigmoid function to yield 0-1 values



Add a bias so that if the total sum is too low, the node doesn't activate

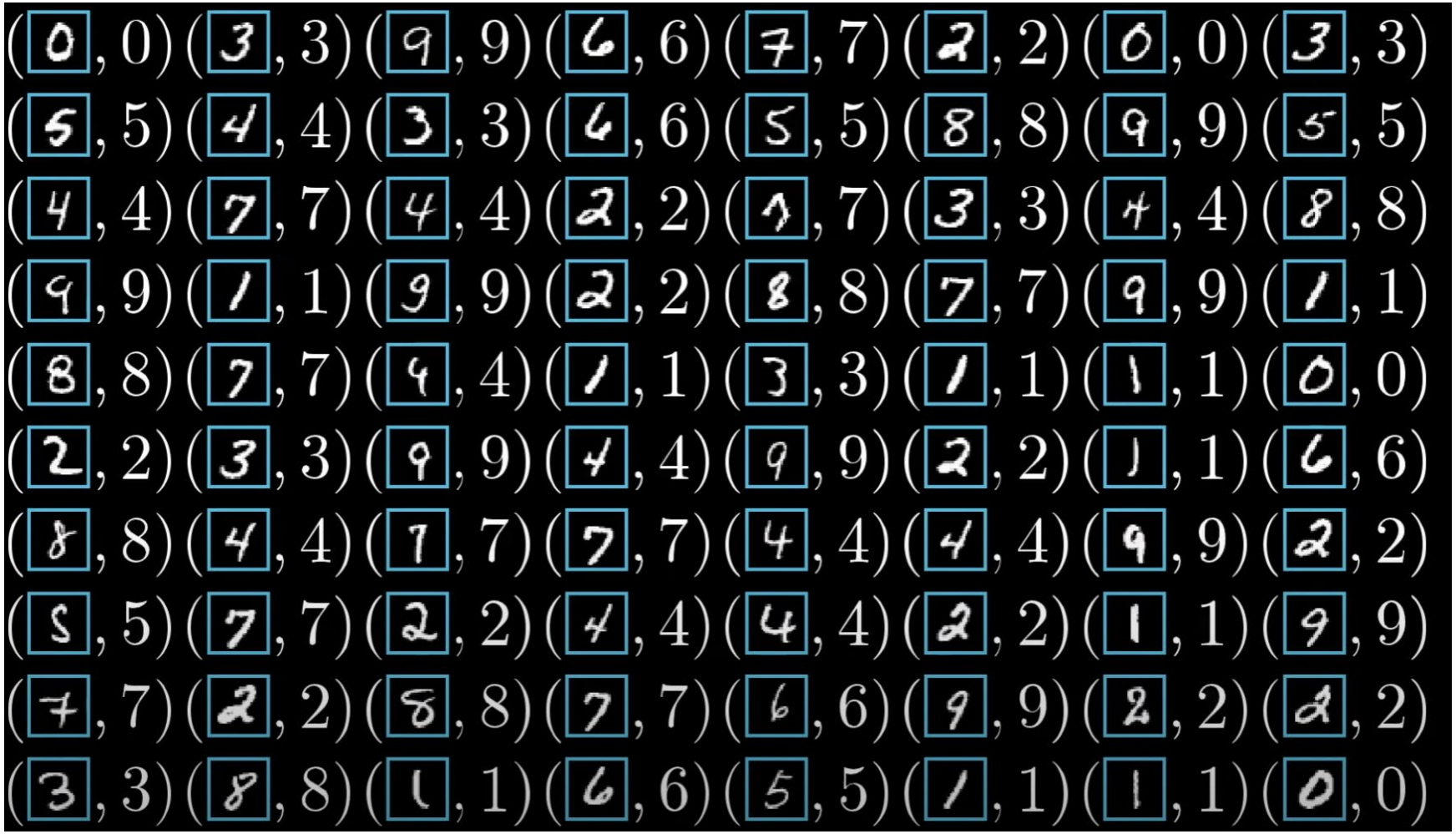


Neural network learning

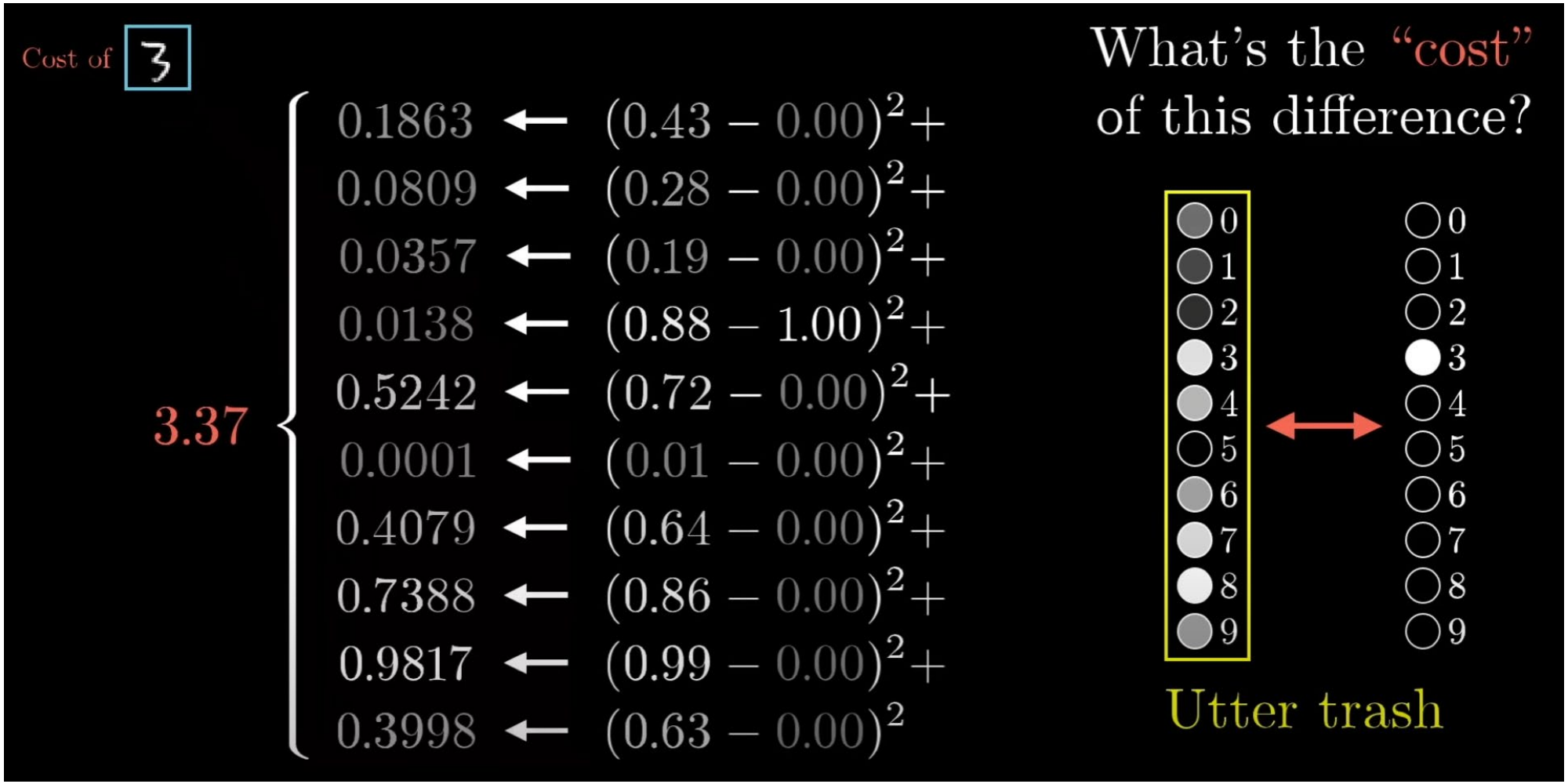


Training data, establishing “ground truths”

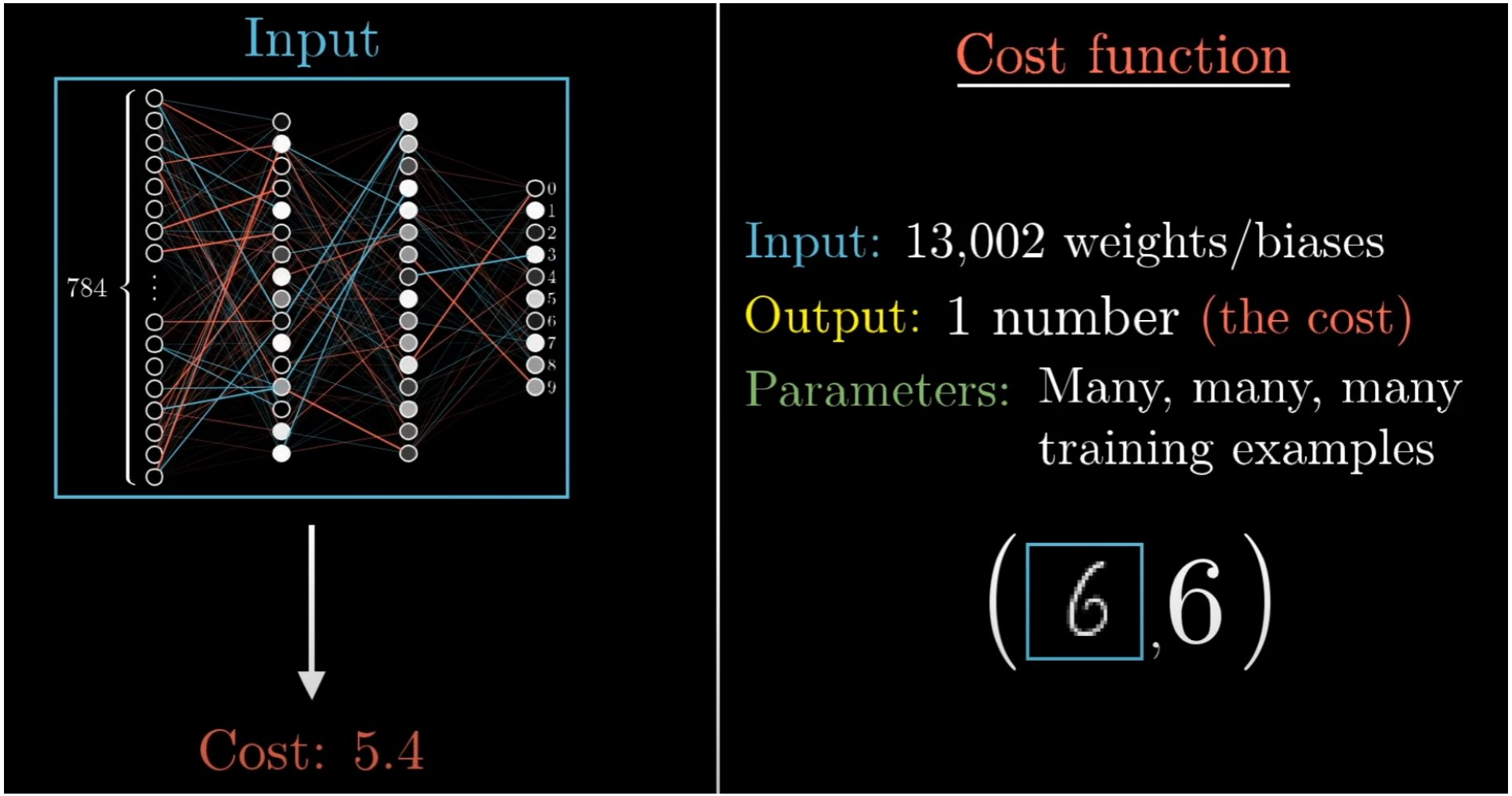
MNIST Dataset



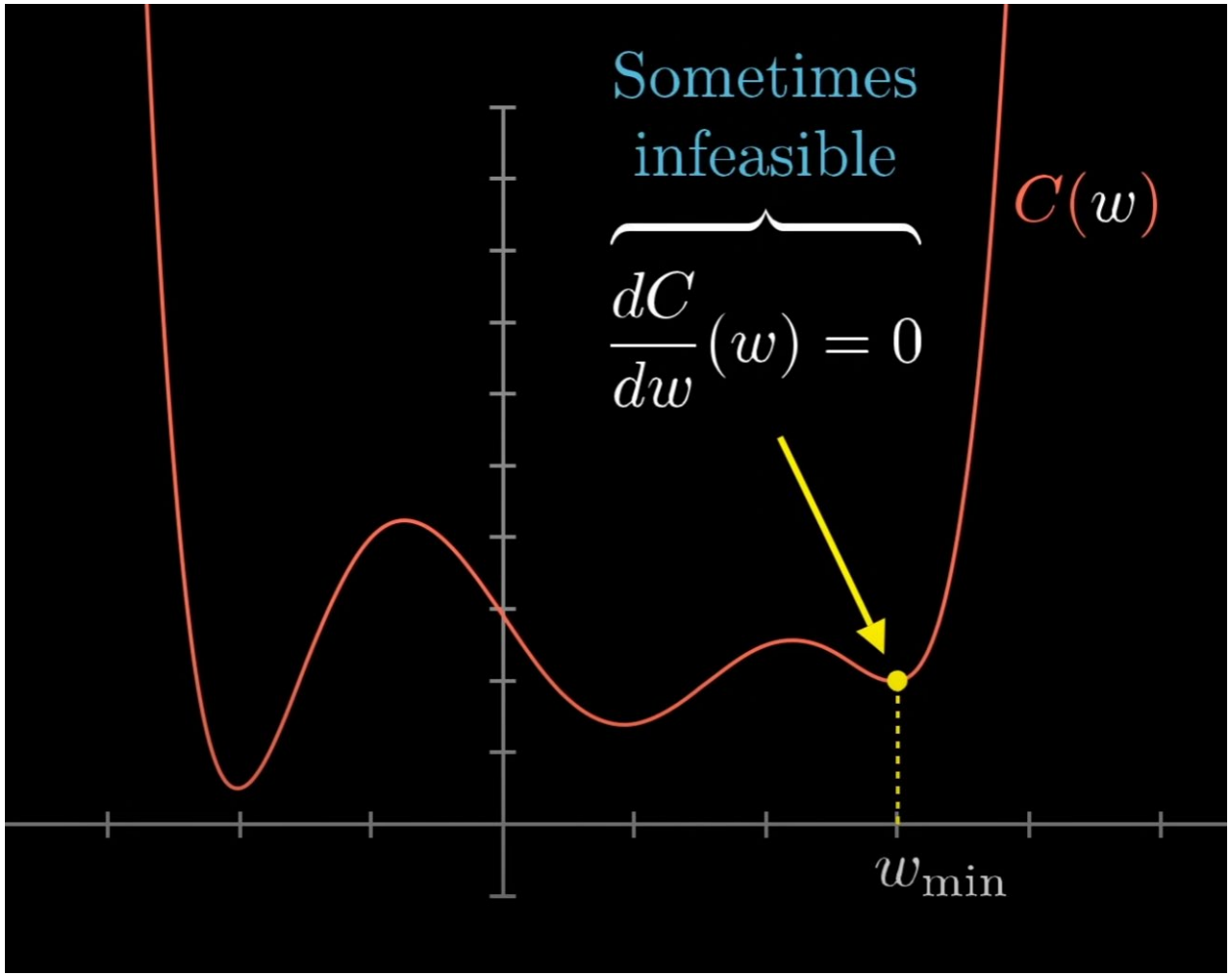
Cost functions



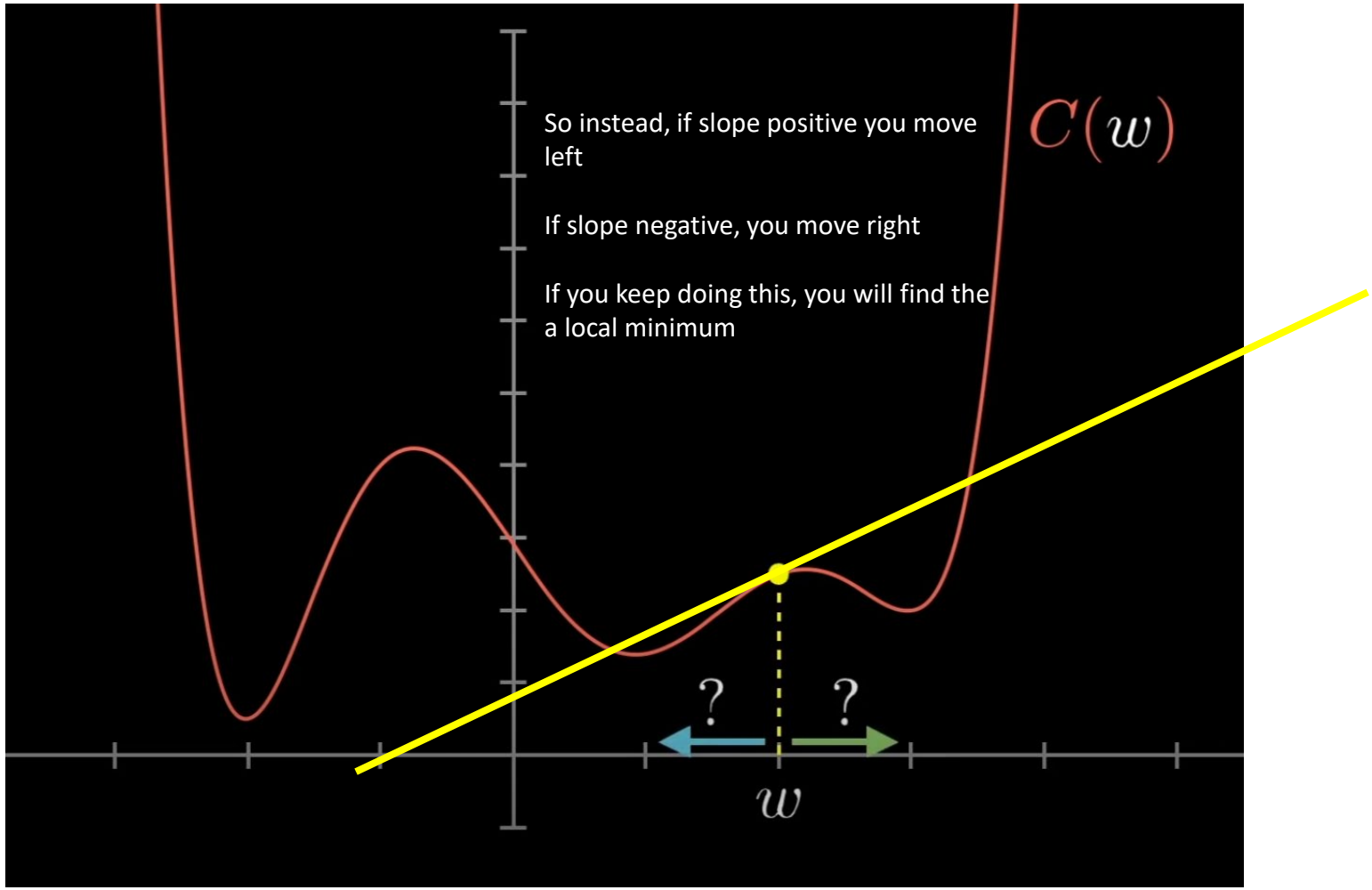
Average cost of a neural network = how good it is



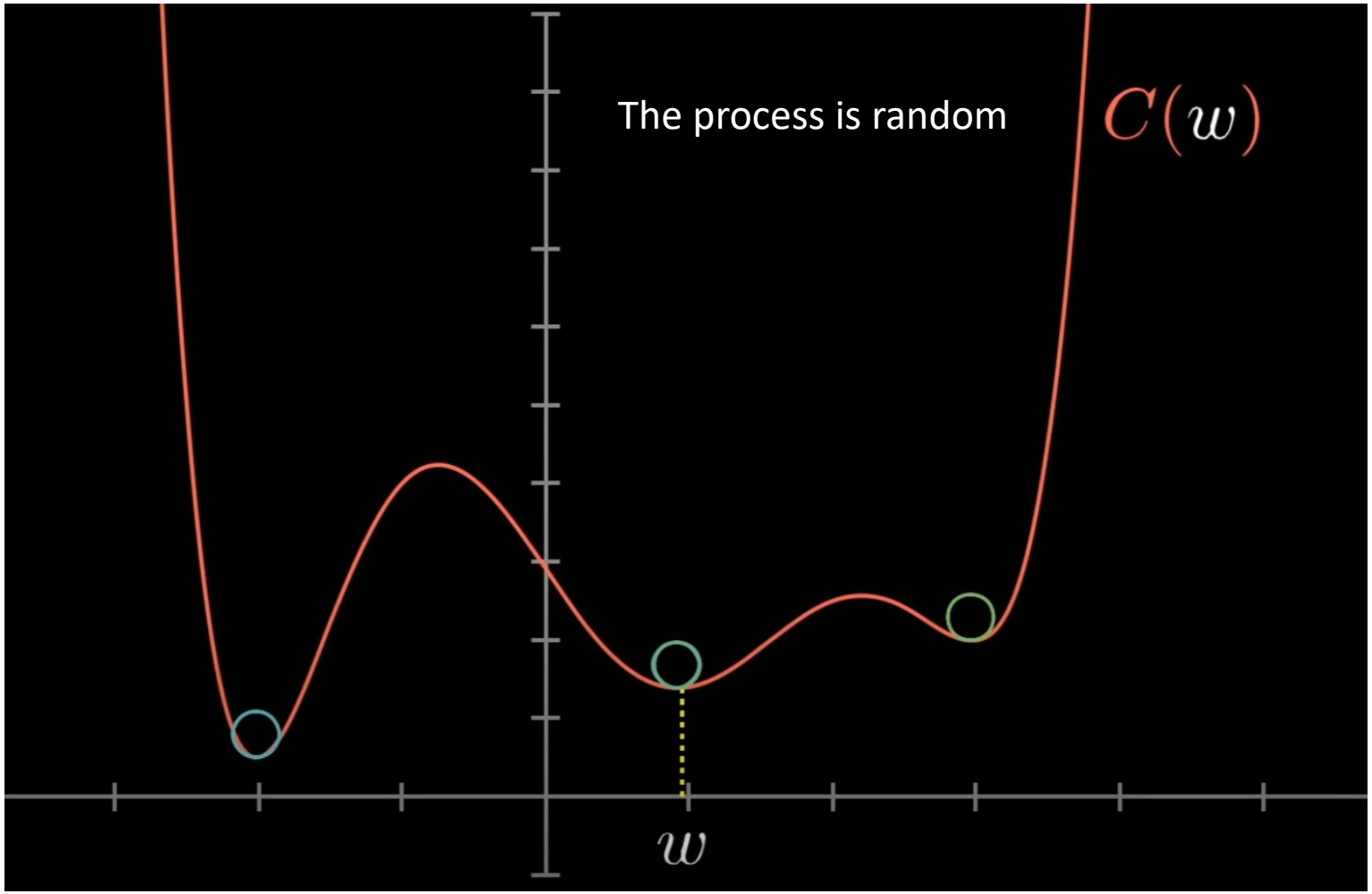
How do you tell the NN how to adjust weights to minimize cost?



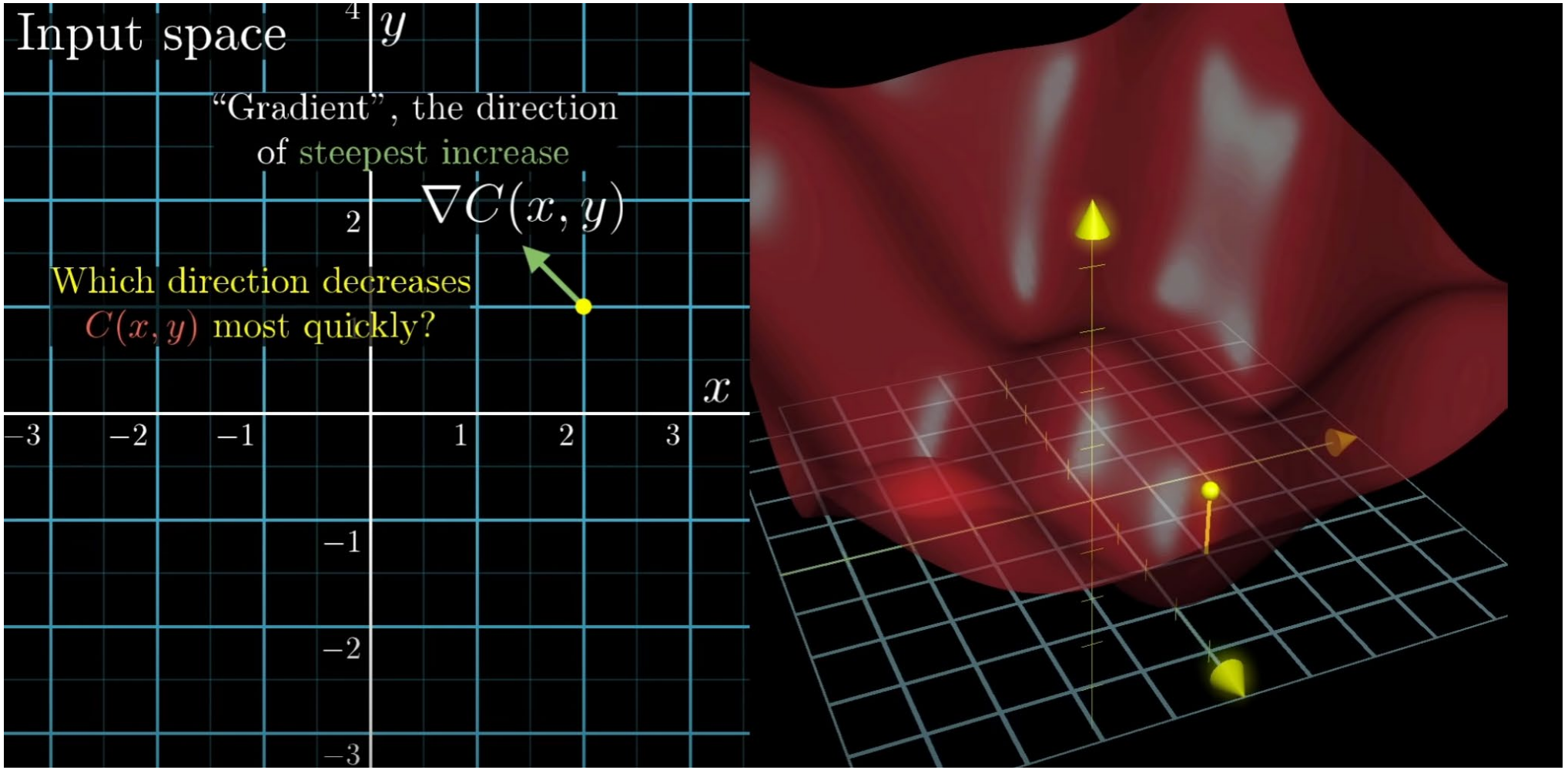
Cost functions are difficult to minimize using differential calculus



Very difficult to find the global minimum



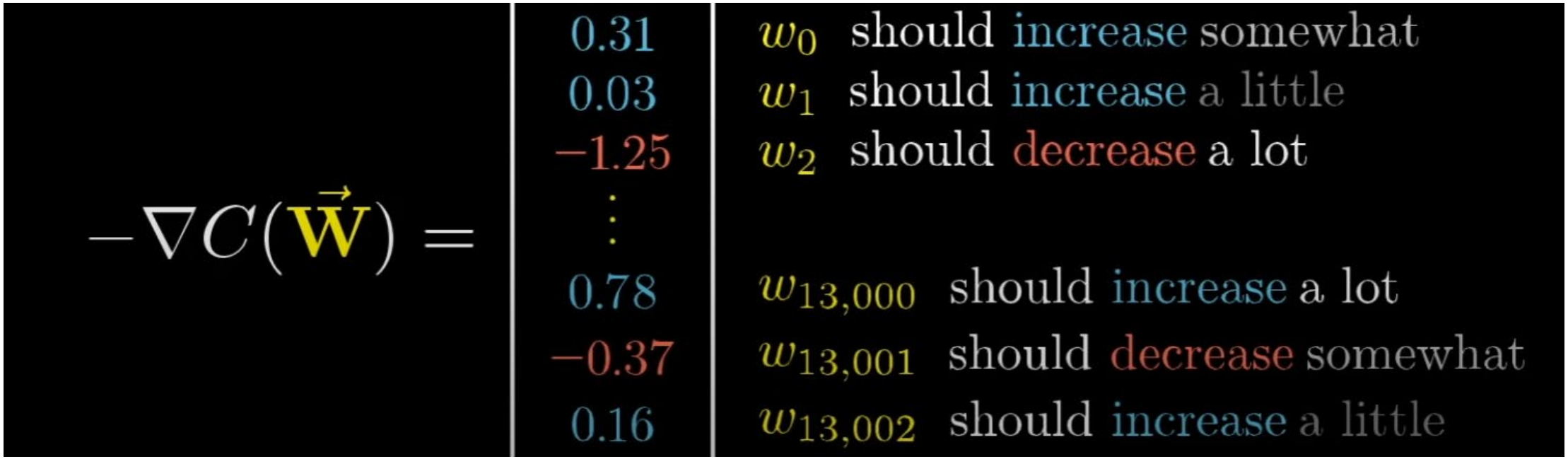
With multiple inputs the complexity increases



The concept of gradient descent

These gradient vectors allows the NN to know which weights are important to shift to minimize cost

Gradient vectors



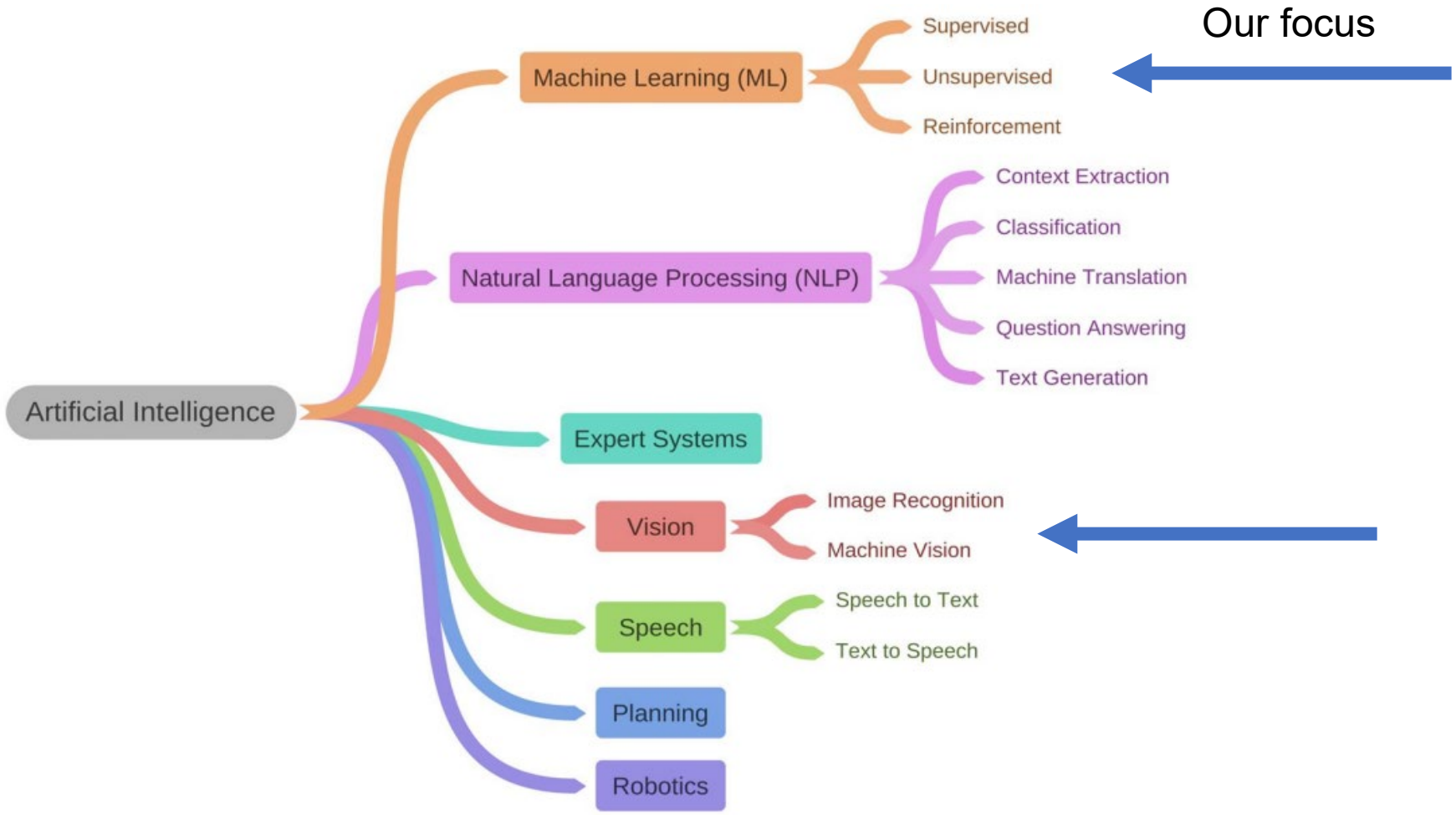
Backpropagation

- Refers to the changes in the weight and biases of nodes after a specific round of training
- The gradient vectors would tell you what weights and biases to shift give you the best minimization of cost
- By doing this after every round of training, this improves your chances at getting to a neural network with a lower average cost

Summary

- Neural networks are made up of layers of nodes
- The layers in the middle (not the input/output layer) all carry an activation function that is composed of the sum of the activation, weights and biases from the nodes that feed into it, this collectively determines whether or not that node activates
- Learning in neural network requires training, in training errors have “costs” that are computed with a cost function
- The goal of the training algorithm is to minimize the cost function
- In each iteration of the training, backpropagation adjusts the weights and biases based on the gradient vectors of the nodes such that it has a better chance at reducing error the next time around

Scope of artificial intelligence



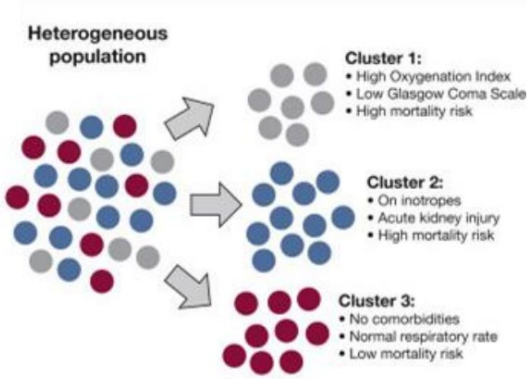
Types of machine learning

- Unsupervised learning
- Supervised learning
- Reinforcement learning

Unsupervised learning

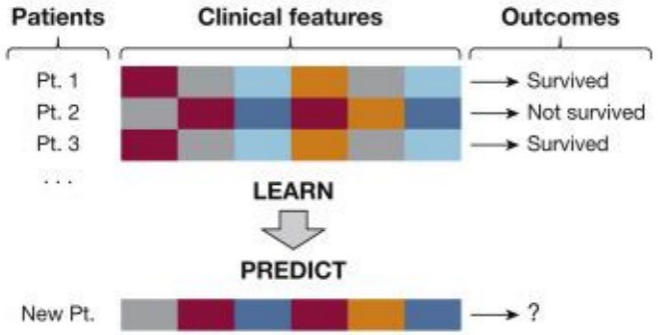
- Unsupervised learning algorithms are used to uncover naturally occurring patterns or groupings in the data, without targeting a specific outcome

- The most compelling use case of unsupervised learning in health care is in precision medicine, in which the goal is to uncover subsets of patients who share similar clinical or molecular characteristics and are, in theory, more likely to respond to targeted therapies directed at their shared underlying pathobiology



Supervised learning

- Supervised: Used to uncover the relationship between variables of interest and one or more target outcomes
- For supervised problems, the target outcome(s) must be known
- For example, if researchers want to know whether a set of clinical features (eg, vital signs, laboratory tests) can predict ICU mortality, they could apply a supervised learning algorithm to a dataset in which each patient record contains the set of clinical features of interest and a label specifying their outcome

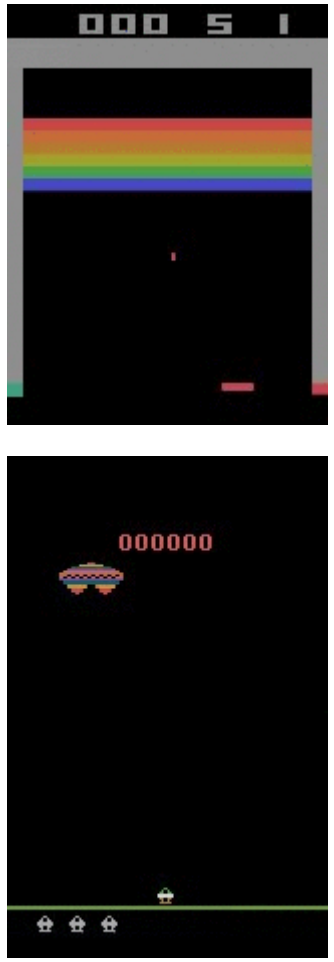


Reinforcement learning

- Reinforcement Learning(RL) is a type of machine learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences
- Unlike supervised learning where feedback provided to the agent is correct set of actions for performing a task, reinforcement learning uses rewards and punishment as signals for positive and negative behavior
- While the goal in unsupervised learning is to find similarities and differences between data points, in reinforcement learning the goal is to find a suitable action model that would maximize the total cumulative reward of the agent

Reinforcement learning achieves superhuman performance

Initial Performance

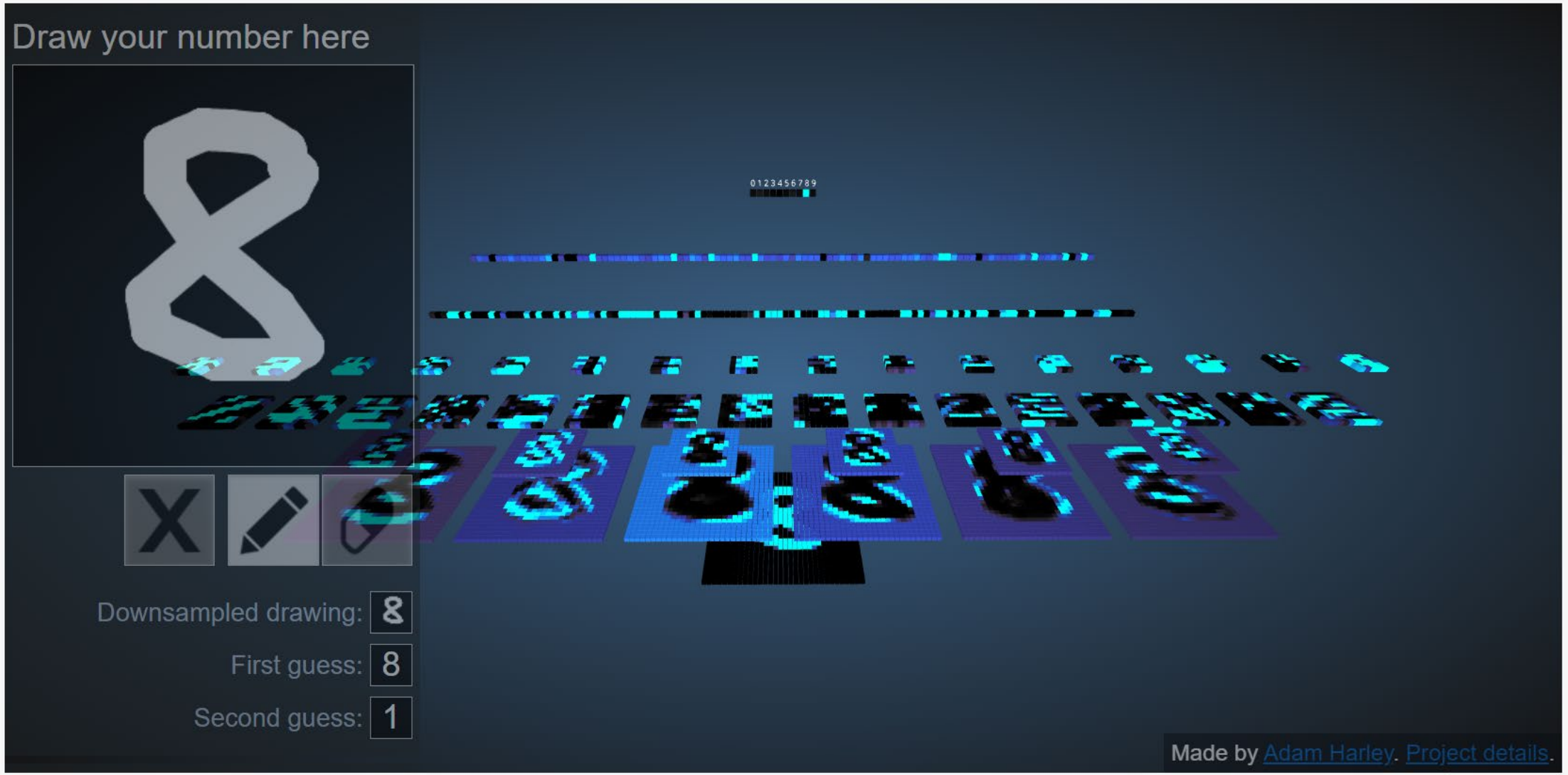


After 30 Minutes

Computer vision

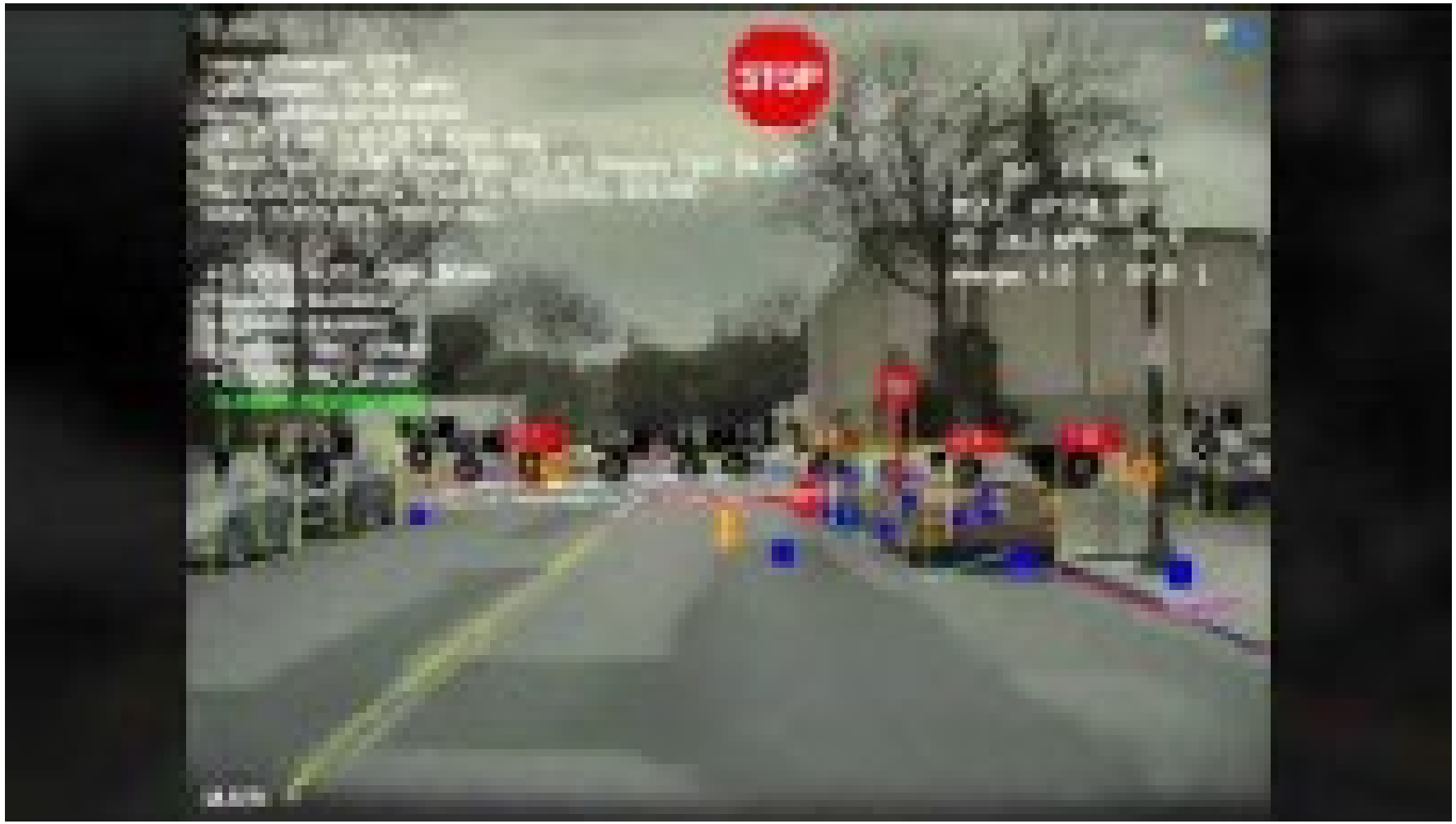
- The processing of an image to enable identification of image input and to provide an appropriate output

Demo of neural network classification of numbers



Source: https://www.cs.ryerson.ca/~aharley/vis/conv/

Application of a neural network to real-time classification / segmentation



Translational research applications of AI

- Cancer
  - Classification of tumours based on histology
  - Classification of tumours based on transcriptome
- Single cell RNA seq
  - Classifying cell clusters and labeling them
- Drug design
  - Prediction of protein structures from sequences

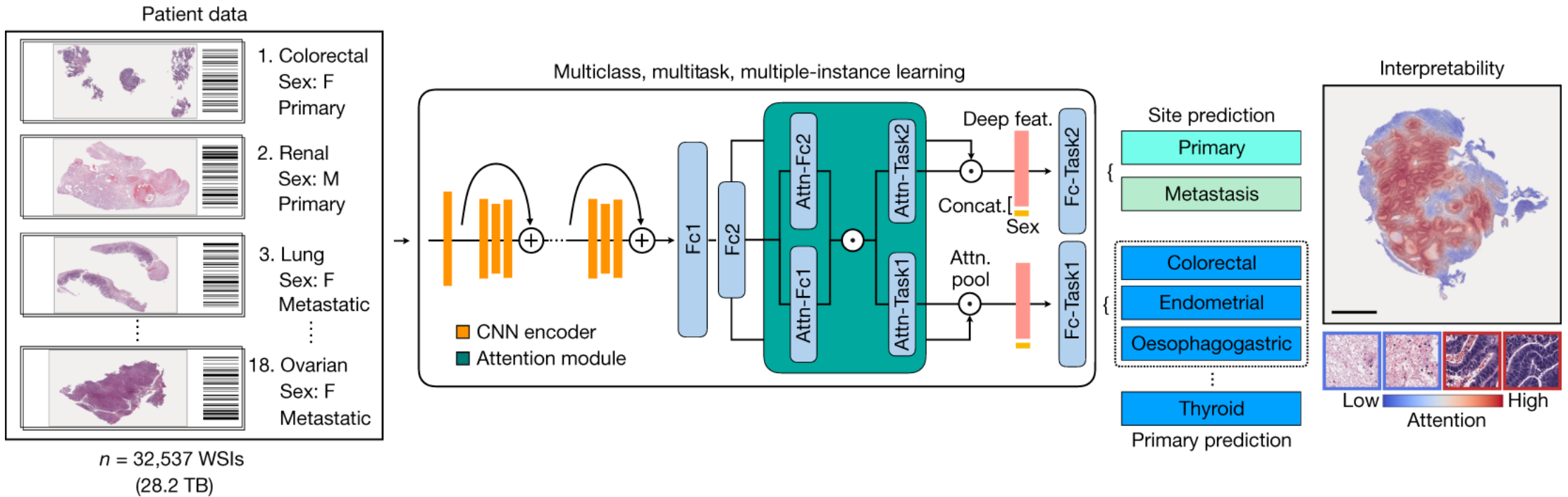
Case study: AI predicts origins of CUPs

Tumour Origin Assessment via Deep Learning

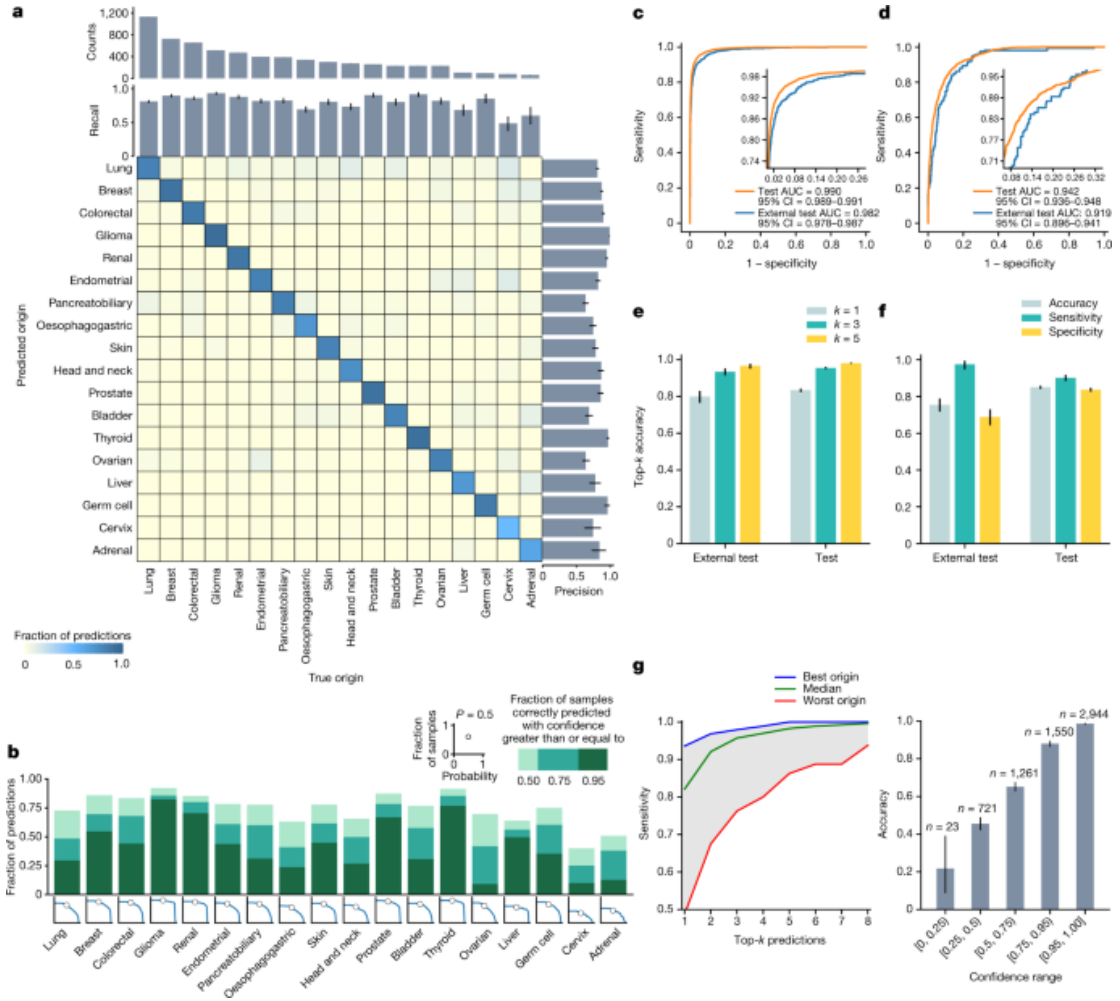
Article | Published: 05 May 2021

AI-based pathology predicts origins for cancers of unknown primary

Ming Y. Lu, Tiffany Y. Chen, Drew F. K. Williamson, Melissa Zhao, Maha Shady, Jana Lipkova & Faisal Mahmood



Performance of TOAD



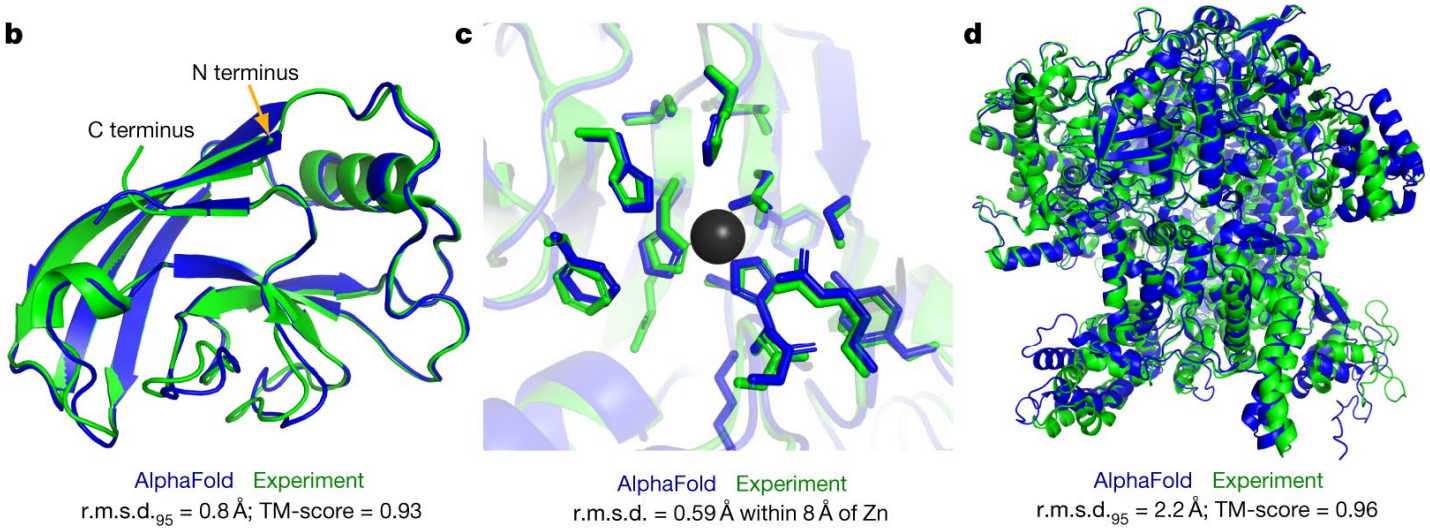
Case study: AI predicts protein 3D structure

Article | Open Access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

John Jumper, Richard Evans, ... Demis Hassabis + Show authors

Nature **596**, 583–589 (2021) | Cite this article



Case study: AI predicts the cell identities of scRNASeq

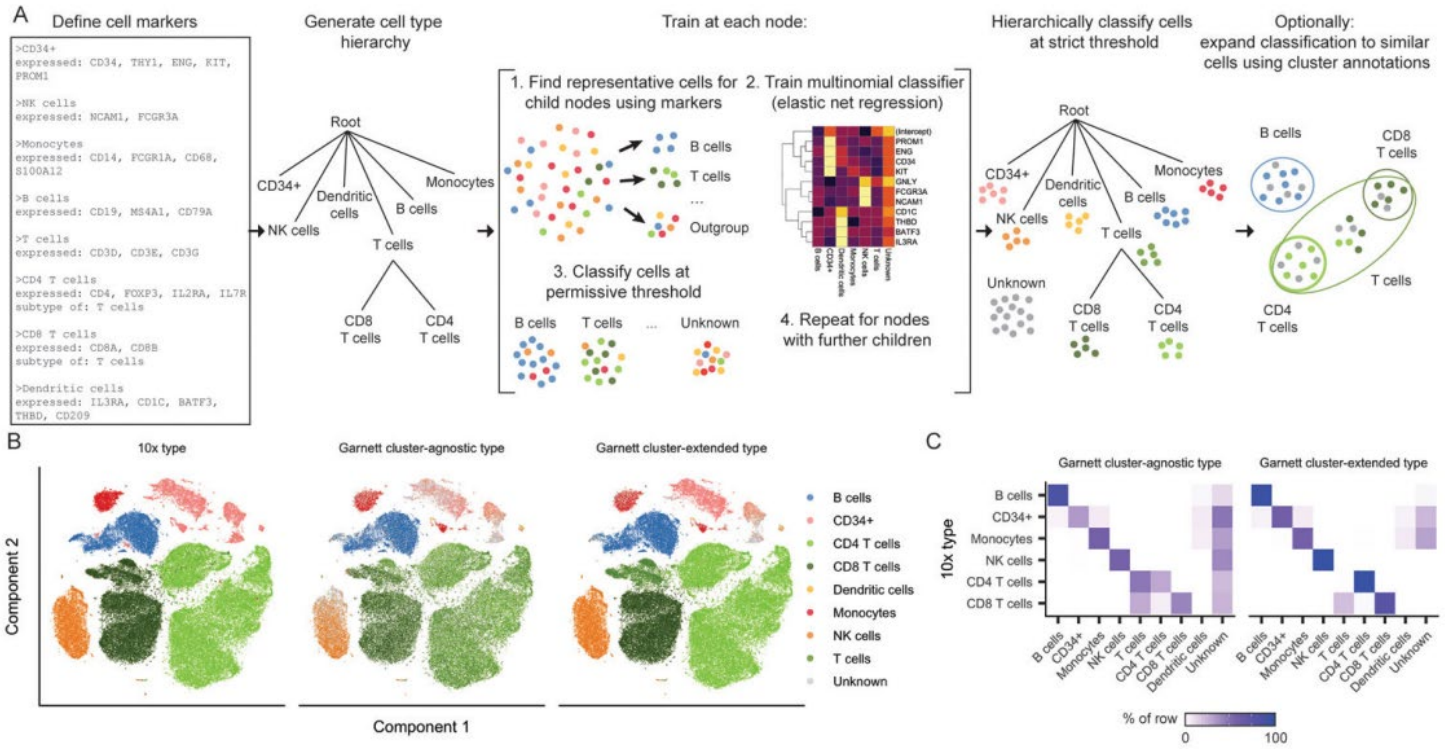
Brief Communication | Published: 09 September 2019

Supervised classification enables rapid annotation of cell atlases

Hannah A. Pliner, Jay Shendure & Cole Trapnell

Nature Methods **16**, 983–986 (2019) | Cite this article

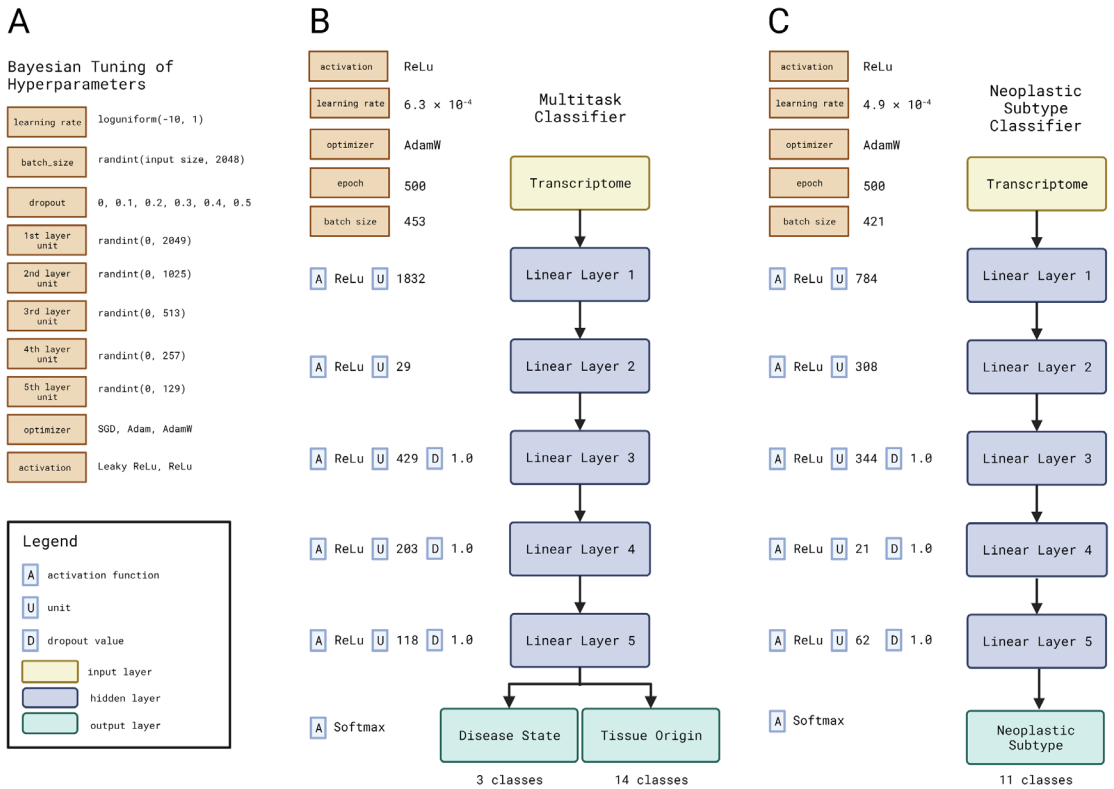
14k Accesses | **130** Citations | **84** Altmetric | Metrics



Case study: AI predicts tumour origin

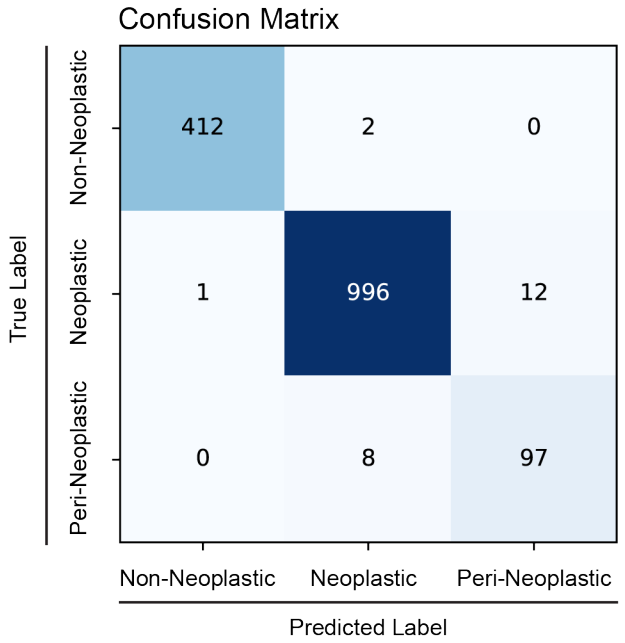
A deep learning model to classify neoplastic state and tissue origin from transcriptomic data

James Hong, PhD [1*], Laureen D. Hachem, MD [1,2*], Michael G. Fehlings, MD, PhD [1,2]

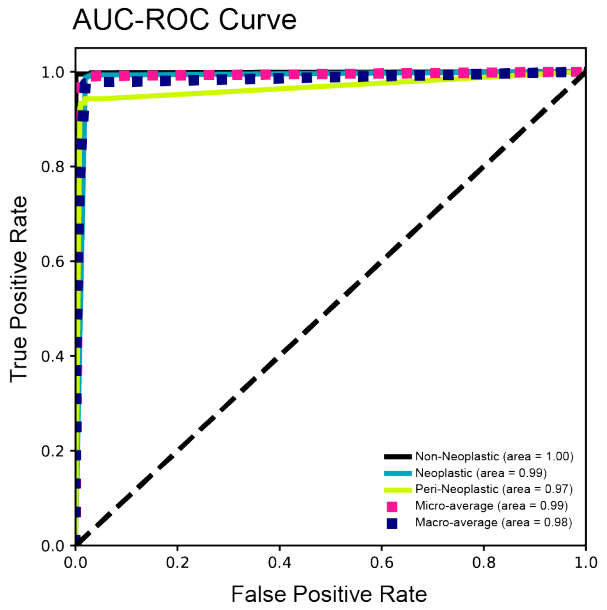


Performance of tissue model

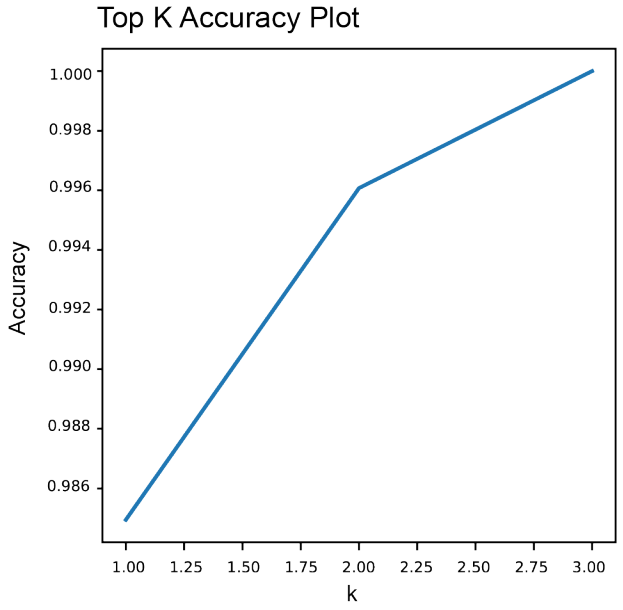
A



B

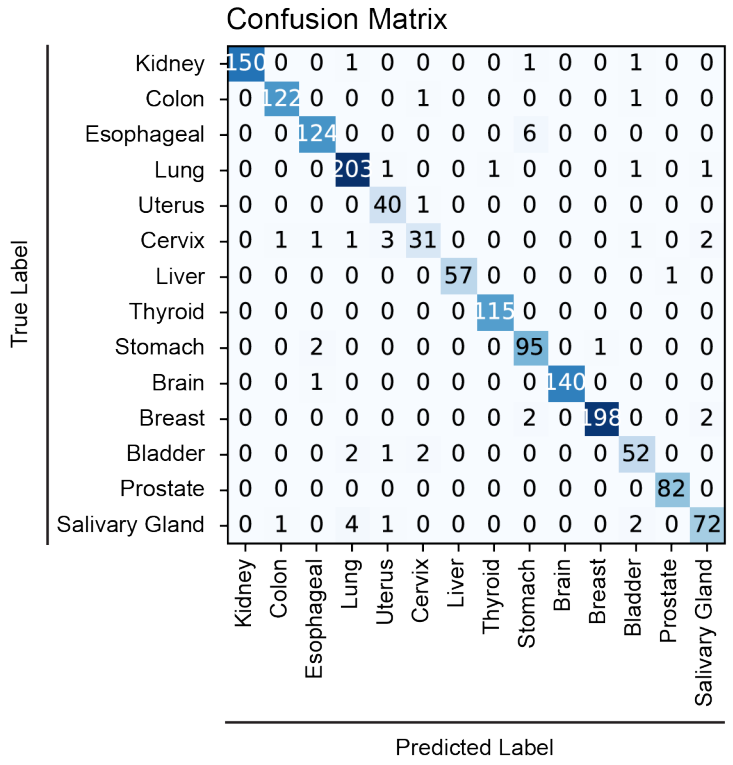


C

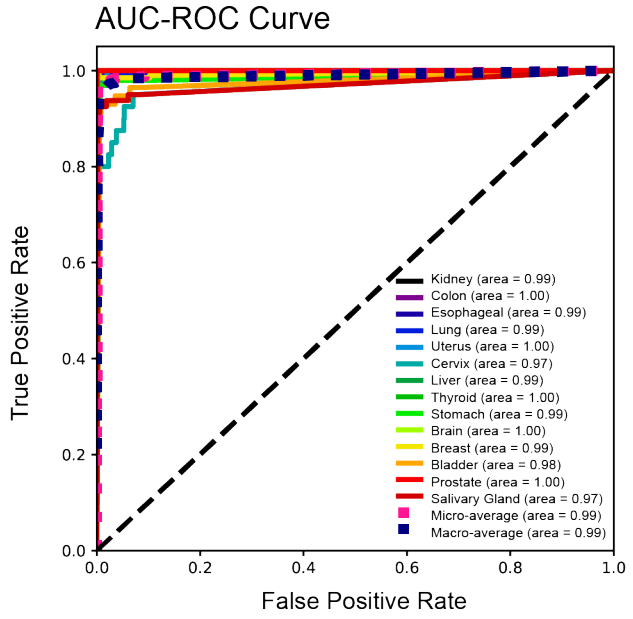


Performance of organ model

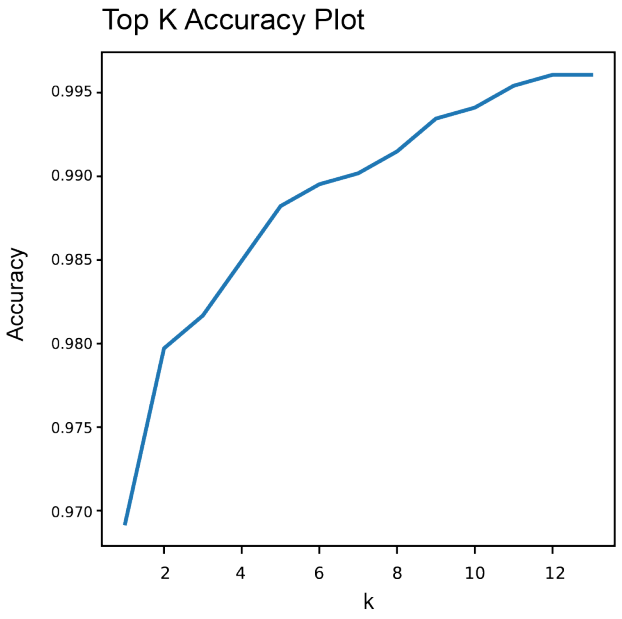
A



B



C



Performance relative to classic ML algorithms