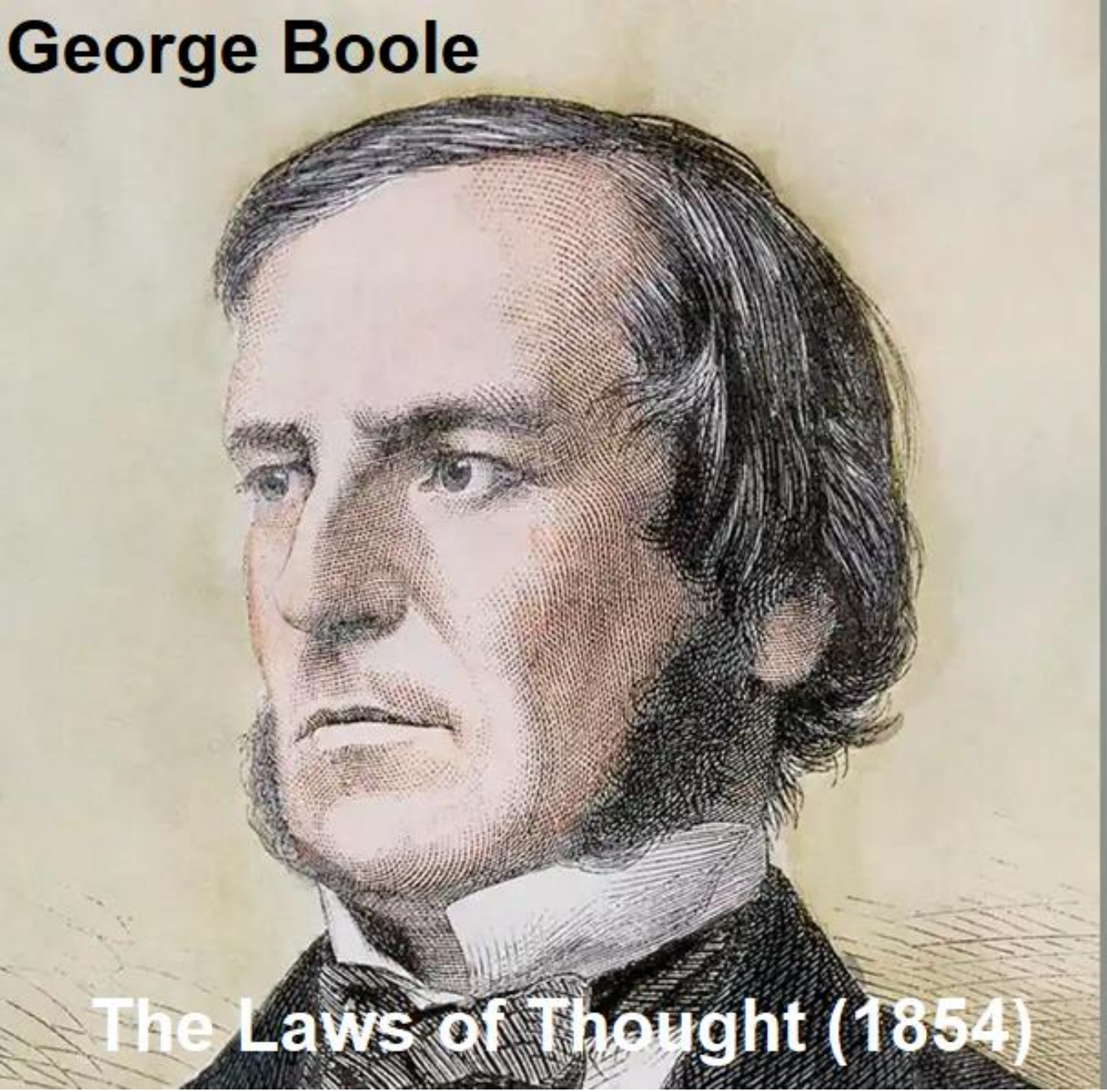
Artificial Intelligence

Neural Networks & Machine Learning

Jerry Sweeney

js@cix.ie

Founder and Managing Director, CloudCIX Limited



"I think we're moving into a period when, for the first time ever, we may have things more intelligent than us.“

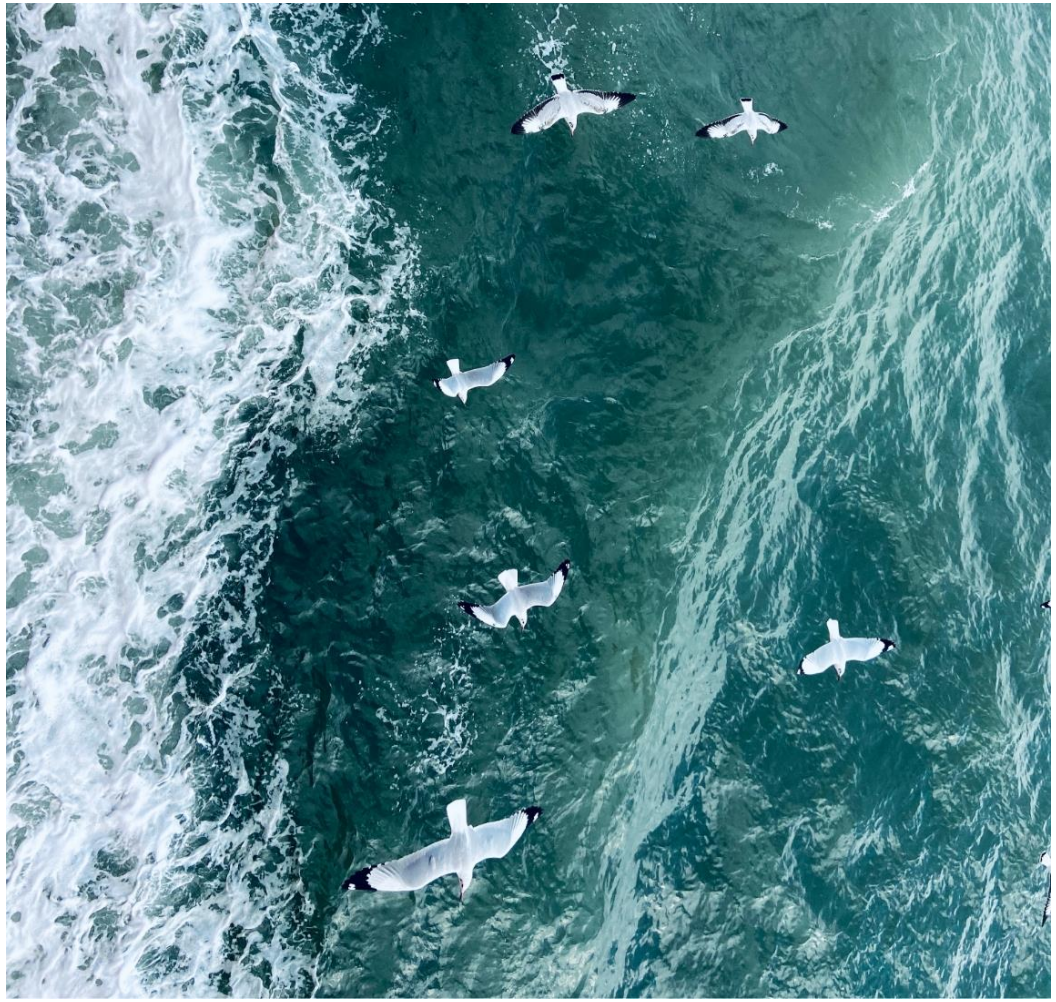
Geoffrey Hinton (2024)

CloudCIX Chatbot Research Projects

- Rare Disease Chatbot for GPs
  - 7,000 identified rare diseases
  - > 20,000 published papers
  - 10% of people have a rare disease

- Bilingual Irish-English Chatbot
  - Constitutional Requirement
  - Government Privacy Requirement

Potential Uses…



Nvidia A100 / H100 GPUs

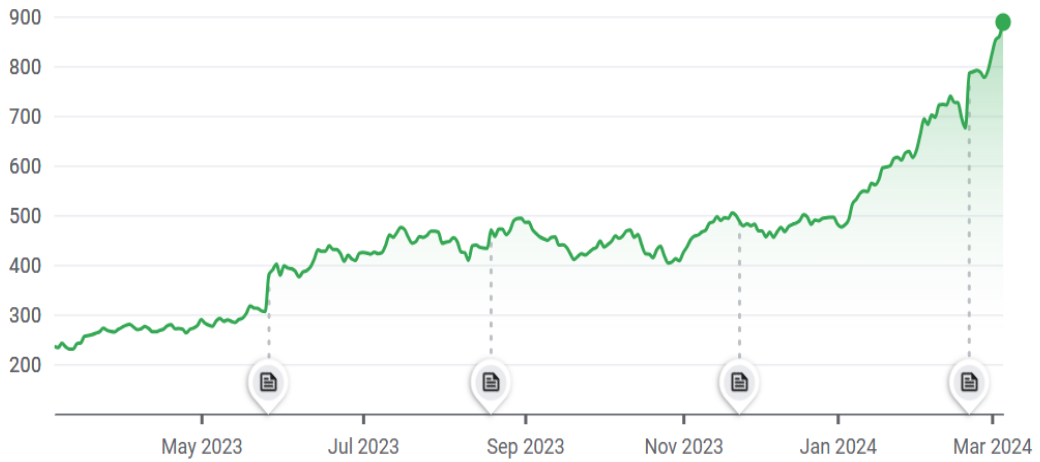
NVIDIA Corp

$887.89 ↑276.96% +652.35 1Y

Mar 6, 3:09:51 PM UTC-5 · USD · NASDAQ · Disclaimer

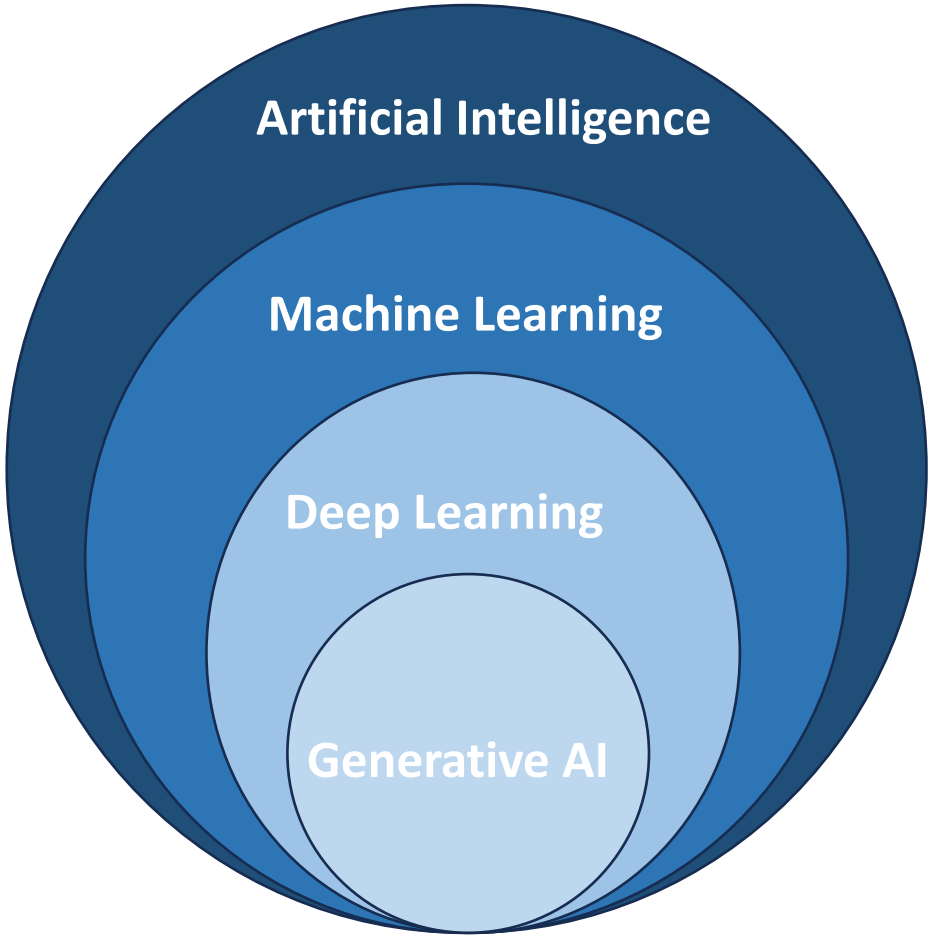
1D 5D 1M 6M YTD 1Y 5Y MAX

Key events



Agenda

- **1** Twin Cognitive Revolutions
- **2** The Brain and Brain Emulation
- **3** Neural Networks & Deep Learning
- **4** Neural Network Types
  - Semantic Vector Search ('Google' search)
  - Convoluted (image classification / recognition)
  - Transformer (large language model (LLM) e.g. ChatGPT)
- **5** LLMs in Detail
  - What LLMs do
  - Tokens and Attention
  - Prompt Engineering
- **6** Let's build a Chatbot trained on a specific corpus of data.



Objectives of the Presentation

1) Achieve a basic technical understanding of neural networks

2) Understand how a customised chatbot can be developed at low cost

Buckle up, you’re in for a ride!

(1) Intelligence

The Human Cognitive Revolution

The Machine Cognitive Revolution

Computer meets caveman!

Historical Perspective (Homo Genus)

- 6,000,000 Last common matriarch of chimpanzee and humans
- 2,000,000 Human species using tools spread out of Africa
- 300,000 Fire in common use by many species in the homo genus
- 200,000 Homo sapiens evolves in East Africa
- 70,000 - 40,000 Human Cognitive Revolution
- 45,000 Homo sapiens settle Australia
- 16,000 Homo sapiens settle Americas
- 13,000 All other species of Homo Genus become extinct

What happened during the ‘Cognitive Revolution’?

- No other species underwent this transition.

- “The most likely answer is the very thing that makes the debate possible: Homo sapiens conquered the world thanks to its unique language.”
  - Sapiens (A Brief History of Humankind)
  - Yuval Noah Harrari, 2011

Historical Perspective (Homo Sapiens)

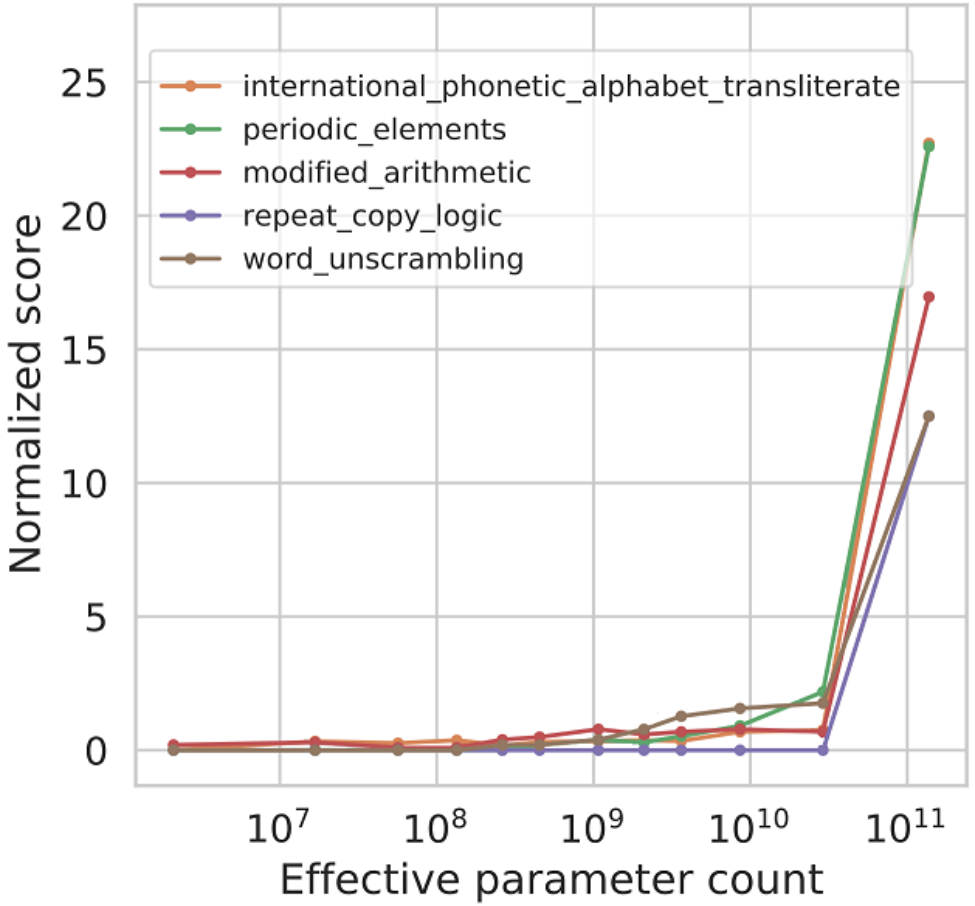
- 12,500 Agricultural Revolution
- 5,000 Kingdoms, Money
- 500 Scientific Revolution
- 200 Industrial Revolution
- 80 First computer
- 75 “Computer Machinery and Intelligence” (Alan Turing, 1950)
- 65 Integrated Circuit (Moore’s Law)

The Machine Cognitive Revolution

- 2010 AI resurgence (Deep Mind, 2010)
- 2012 Convolutional Neural Networks
- 2016 AlphaGo
- 2017 Attention is all you need. The Transformer LLM (Google)
- 2018 AlphaFold
- 2021 - ? Emergence of the Machine Cognitive Revolution
- 2022 ChatGPT / Midjourney / Dall-E
- 20?? Artificial General Intelligence

Machine ‘Cognitive Revolution’

As LLMs are scaled they hit a series of **critical scales** at which new abilities are suddenly “unlocked”. LLMs are not directly trained to have these abilities, and they appear in **rapid and unpredictable** ways as if *emerging* out of thin air. These *emergent* abilities include performing arithmetic, answering questions, summarizing passages, and more, which LLMs learn **simply by observing natural language**.



https://www.assemblyai.com/

Ryan O’Connor March 2023

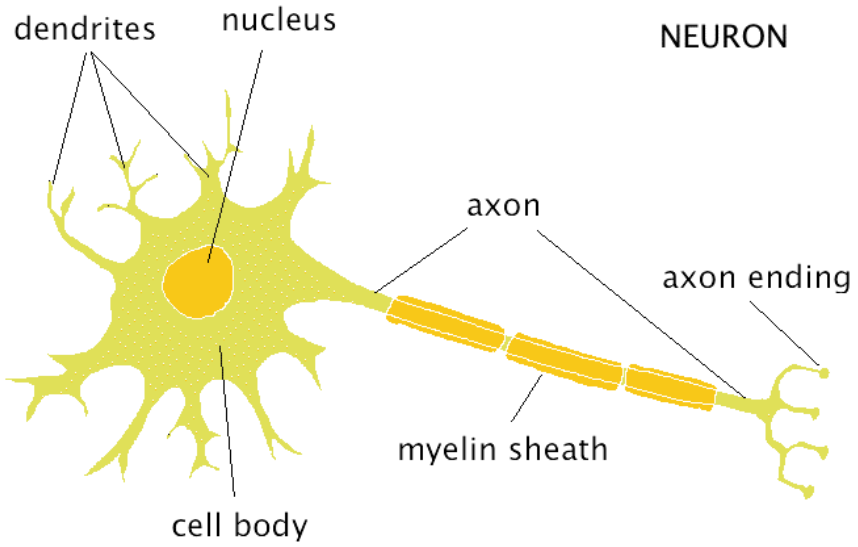
(2) Brains and Brain Emulation

Denis Hassabis (Founder, CEO of Deepmind)

- 1976 Born
- 1986 Chess Grandmaster (ELO 2300)
- 1992/3 Developed ‘Theme Park’
- 1993/7 Cambridge, Computer Science
- 1997/2004 Successful Games Development
- 2004/2009 PhD in Cognitive Neuroscience
- 2010 Founded Deepmind
- 2014 Sold Deepmind to Google
- 2014+ AlphaGo, AlphaFold

Demis Hassabis - Wikipedia

Neuroscience (How the brain works)

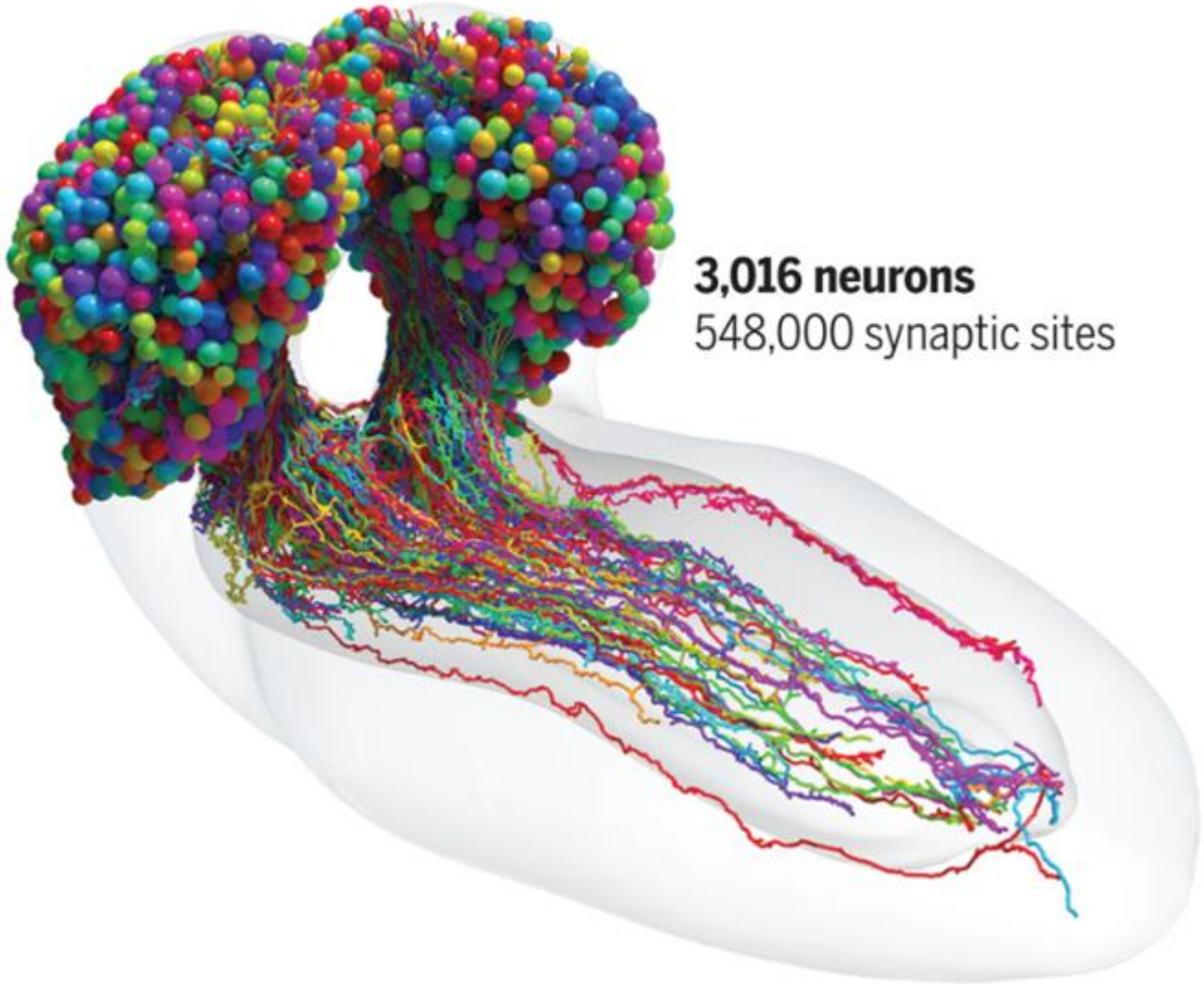


- Incoming signals on dendrites
- If the aggregate of inputs exceeds a threshold, then the output fires.
- Learning (memory) is the threshold of the firing.
- The axon endings are the outputs
- 3Types: Sensory, Interneuron, Motor.

Neuron

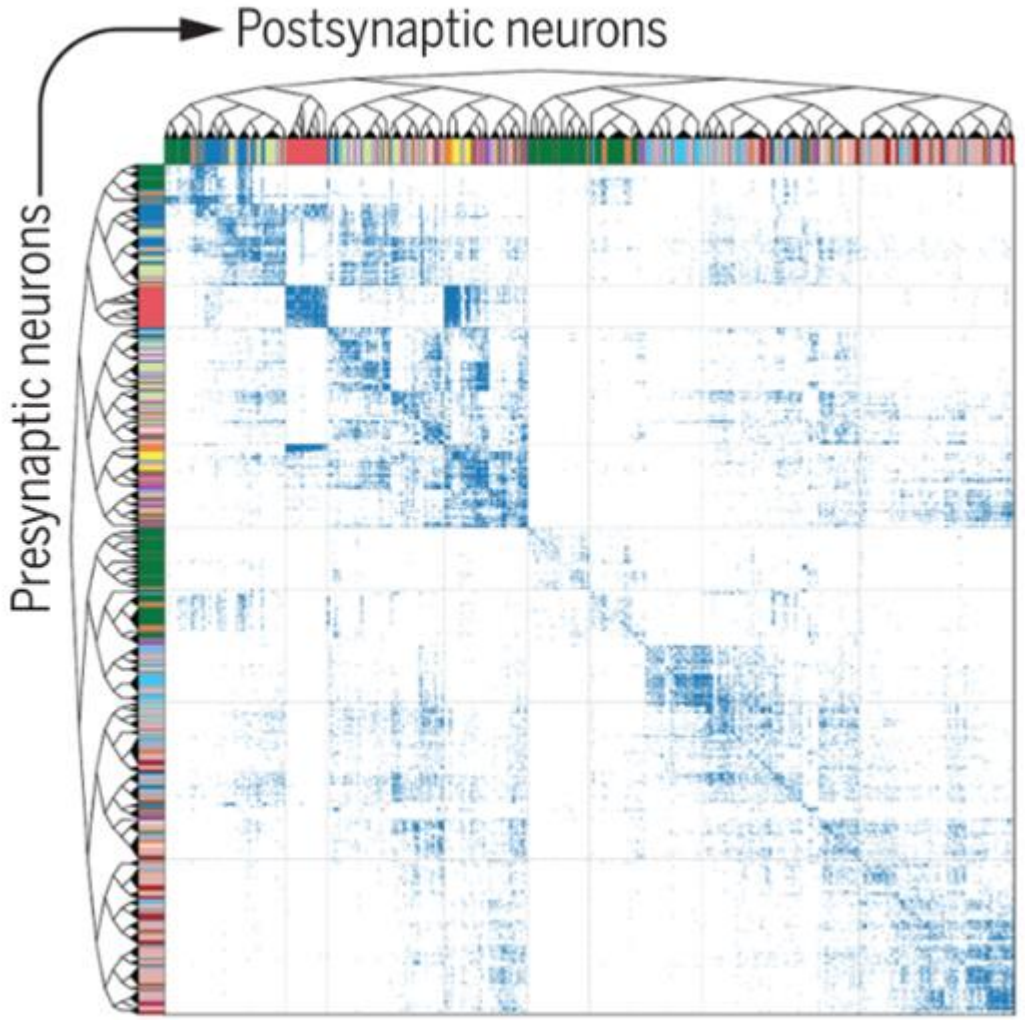
ChaLearn Connectomics Challenges - 5min Tutorial on Brain Science

Morphology



3,016 neurons
548,000 synaptic sites

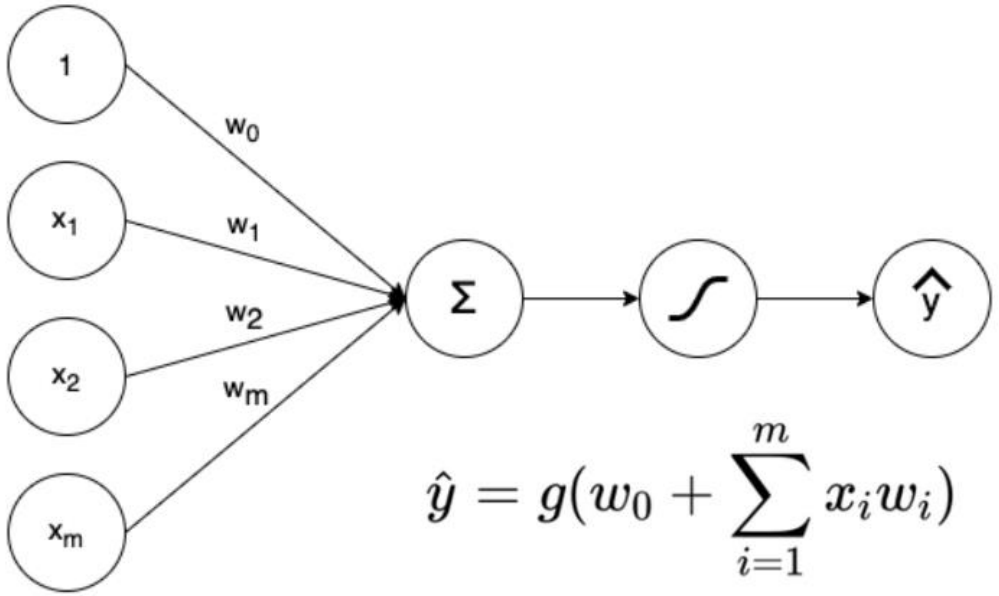
Connectivity



The connectome of the *Drosophila* larval brain.

Image Credits: Michael Winding

The Perceptron (Artificial Neuron Simulator)



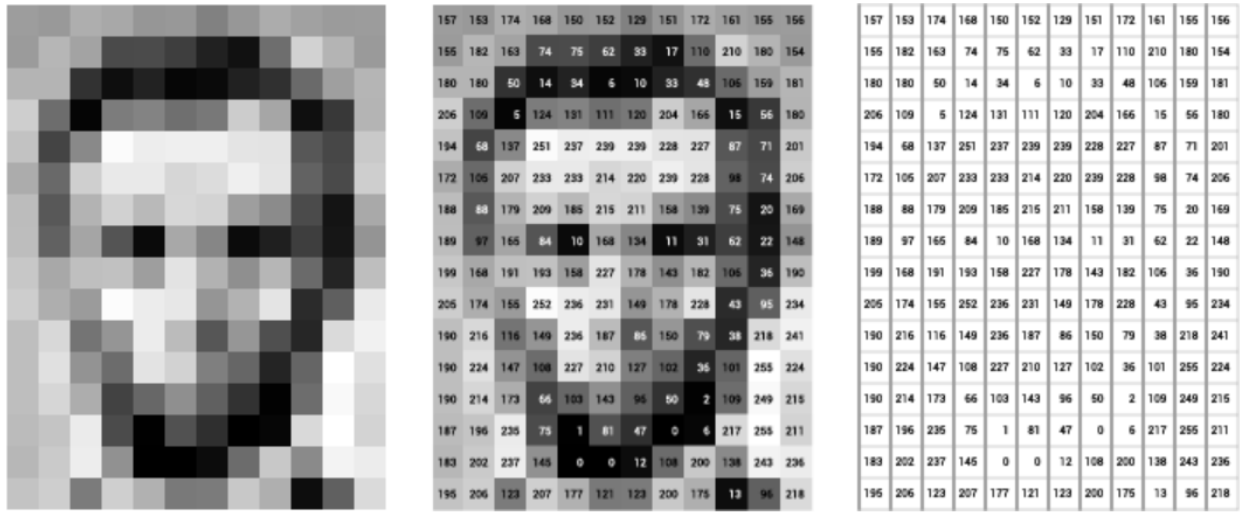
Perceptron

- Mathematical mimic of the Neuron
- Incoming signals x_1 to x_m.
- Training sets the weights of the offset w0 and values w1 to wm.
- The output goes through an activation function.
- Perceptrons are implemented in software.

Neural Network Basics: The Perceptron | Akshay Mahajan (makshay.com)

Neural Networks Use Numbers

- Text data is mapped to integers (tokens)
- Image and video data have RGB numbers for pixels
- Audio is a number sequence (e.g. MP4)



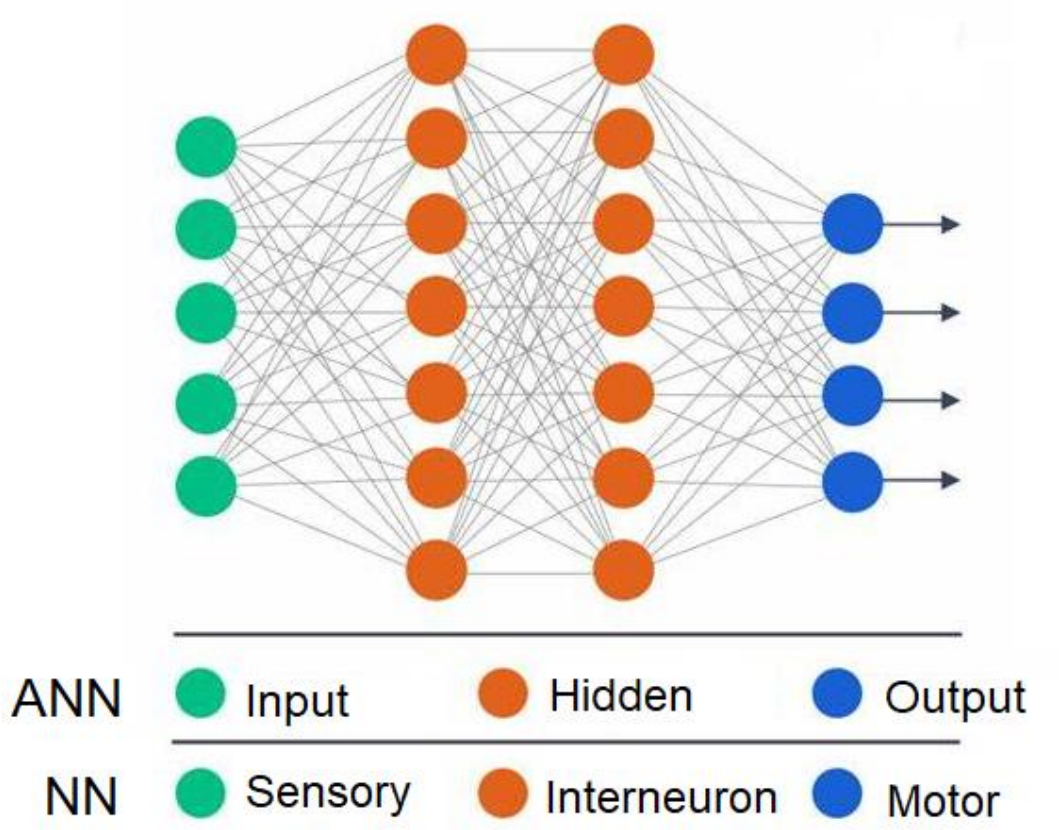
Source: Serena Young

Example of tokenizers (BPE):

original sentence :

- `Ní ghearrann formhór na n-iar-bhunscoileanna in Éirinn táillí.`
- original Llama 2 (English chatbot model) tokenizer:
- `_N í _g he ar ran n _form h ór _na _n - iar - b h uns co ile anna _in _É ir inn _t á ill í .`
- Irish-Llama 2 tokenizer (10k):
- `_Ní _g he ar ran n _formhór _na _n - iar - bhun scoileanna _in _Éirinn _tá il lí .`
- Irish-Llama 2 tokenizers (32k):
- `_Ní _g he ar rann _formhór _na _n - iar - bhunscoileanna _in _Éirinn _táillí .`

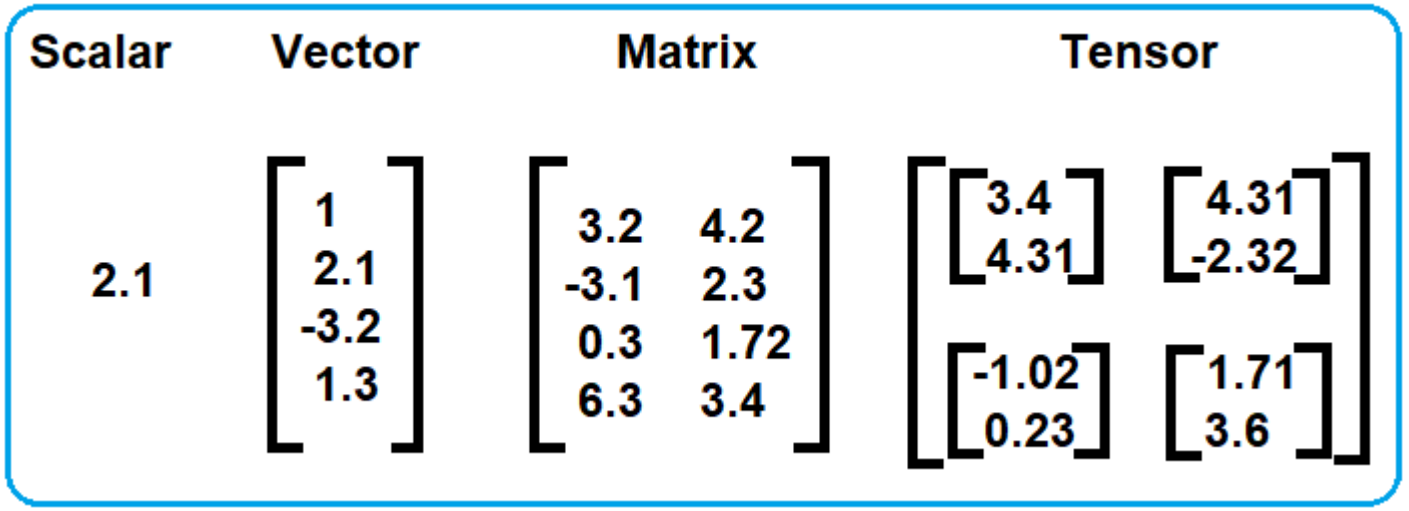
Artificial Neural Network (ANN) versus Neural Network (NN)



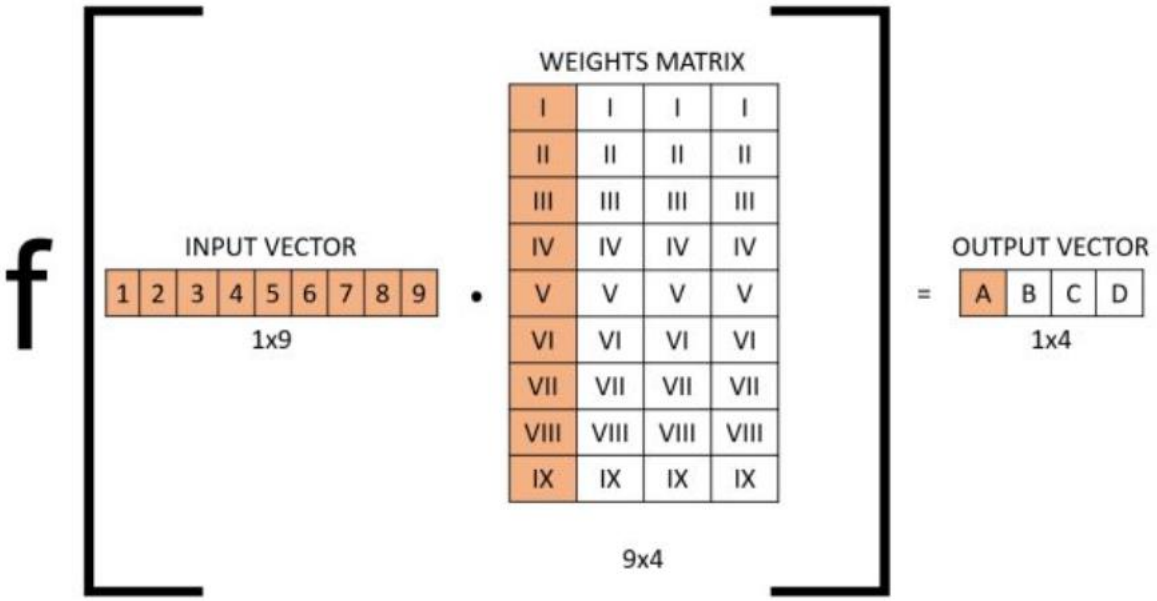
- ANN
  - Analog (70mV voltage pulses)
- NN
  - Digital (4, 8, 16 bit)
  - All data converted to numbers

Linear Algebra Terminology

- Scalar
  - A number
- Vector
  - A list of numbers
  - A ‘list’ in Python
- Matrix
  - A two-dimensional array of numbers
- Tensor
  - A three or higher order array of numbers



A fully connected layer
Dot Product aka Matrix Multiplication

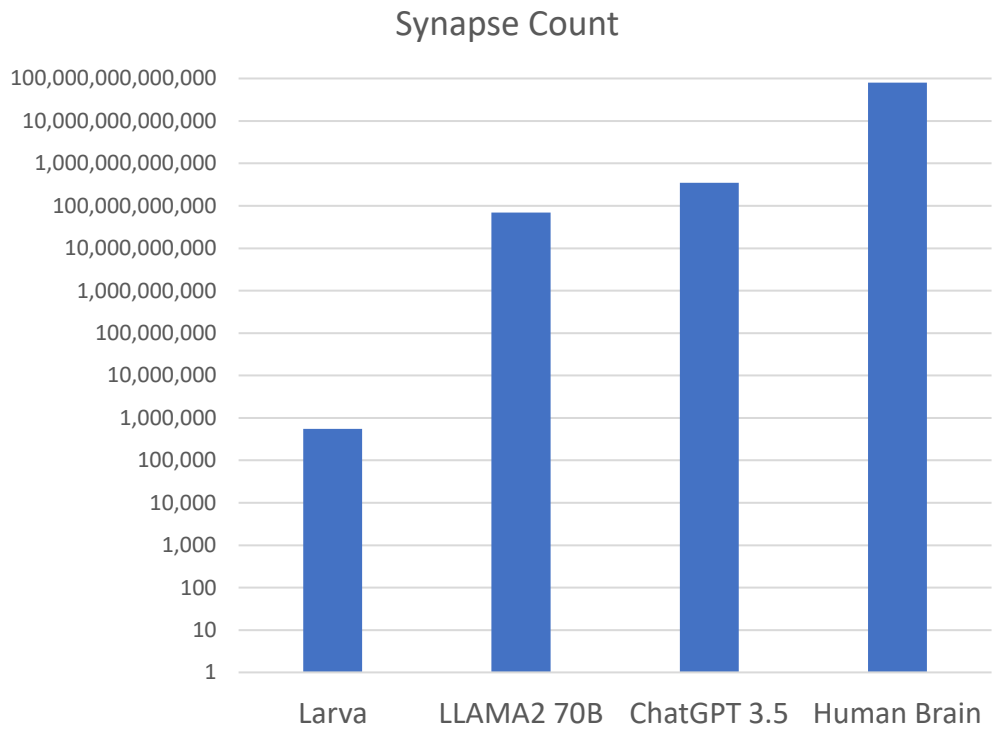


$$\hat{y} = g(w_0 + \sum_{i=1}^{m} x_i w_i)$$

- Tensorflow (Google)
  - Keras API
- PyTorch (Facebook)

Brain versus LLM size

- Fruitfly Larva
  - 548,000 Synapses
- Llama2 70B
  - 70B Parameters (Synapses)
- ChatGPT 3
  - ~ 350 Billion (Synapses)
- Human Brain
  - 80 Trillion Synapses

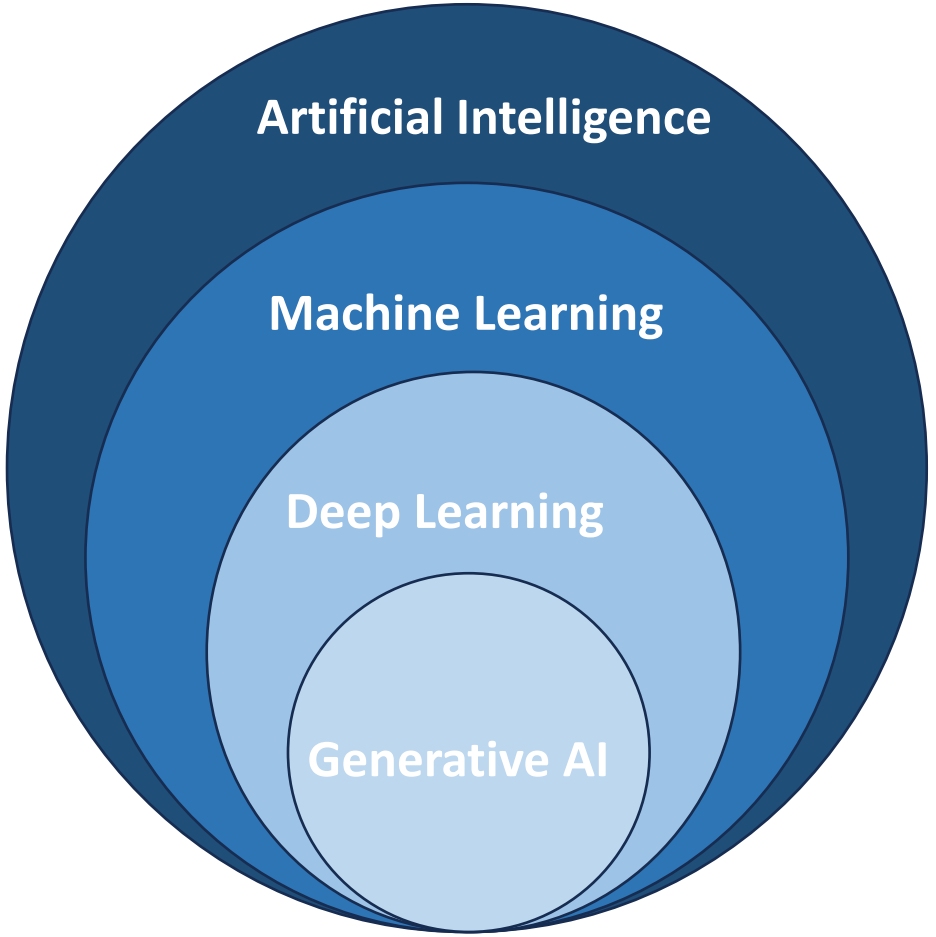


(3) Deep Learning and Generative AI

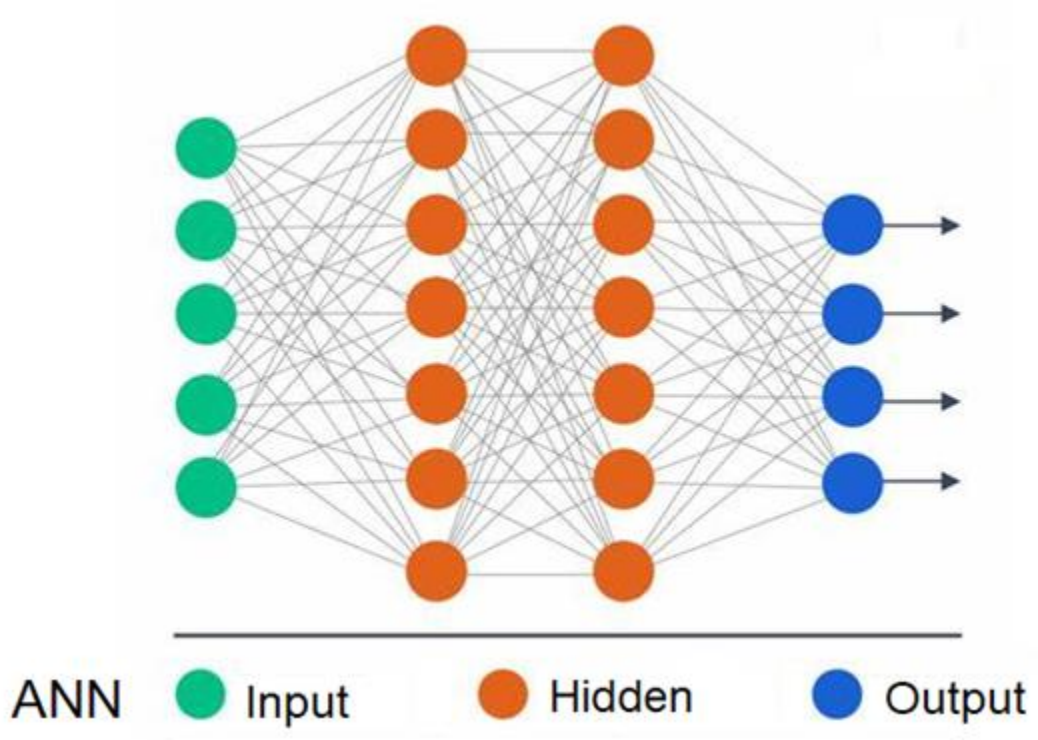
Generative AI and the Transformer Model

Artificial Intelligence Types

- Artificial Intelligence (Programmed)
  - Factory Automation
- Machine Learning (Program using statistics from growing data sets)
  - Recommendation Systems
  - SPAM filters
  - ‘Tay’ Microsoft’s Twitter Chatbot experiment
- Deep Learning (No Program, weights trained on vast datasets)
  - Discriminative AI
    - Classification, Vector Embedding
- Generative AI
    - Text, Audio, Image, Video

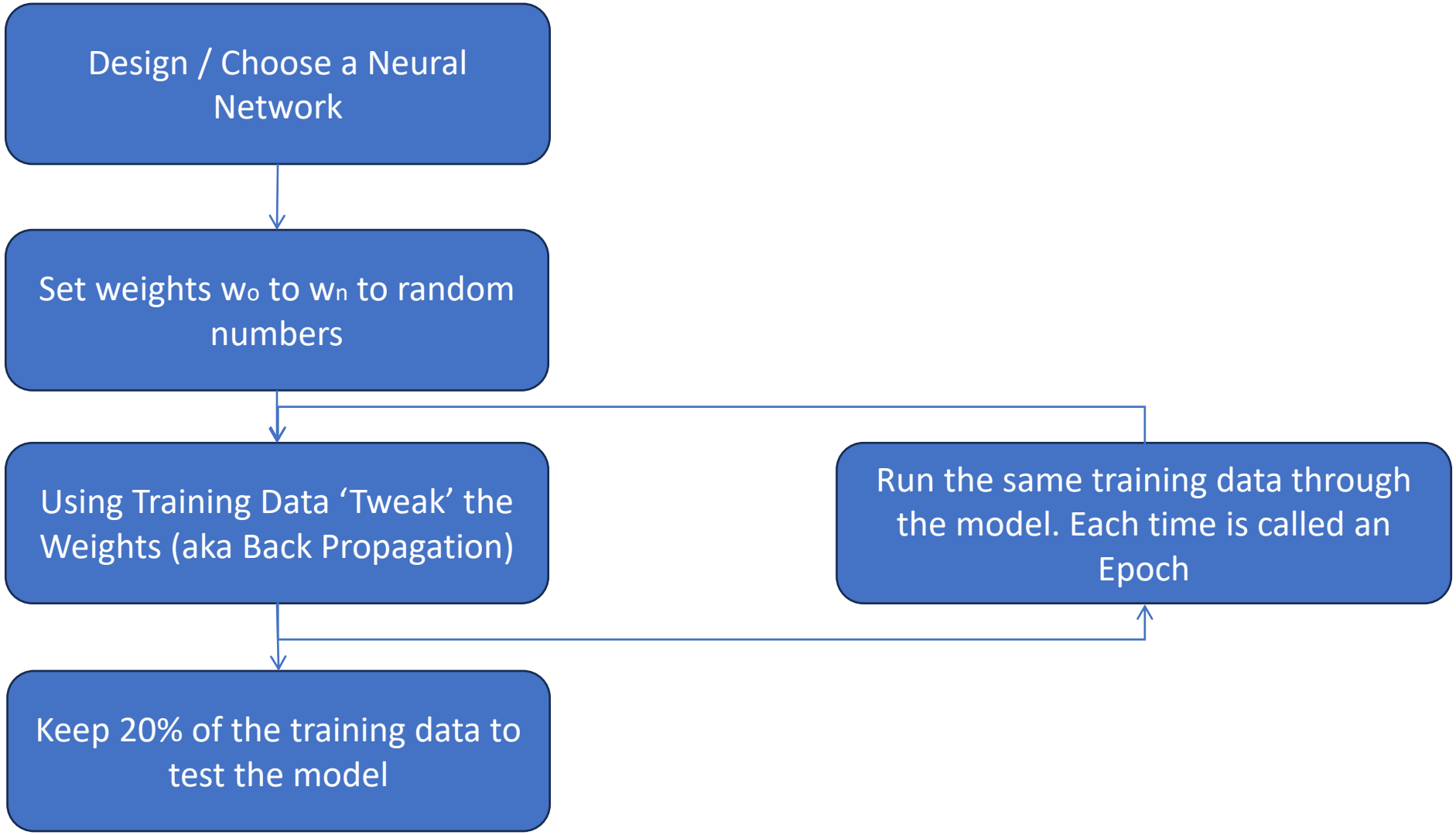


Training Deep Neural Networks



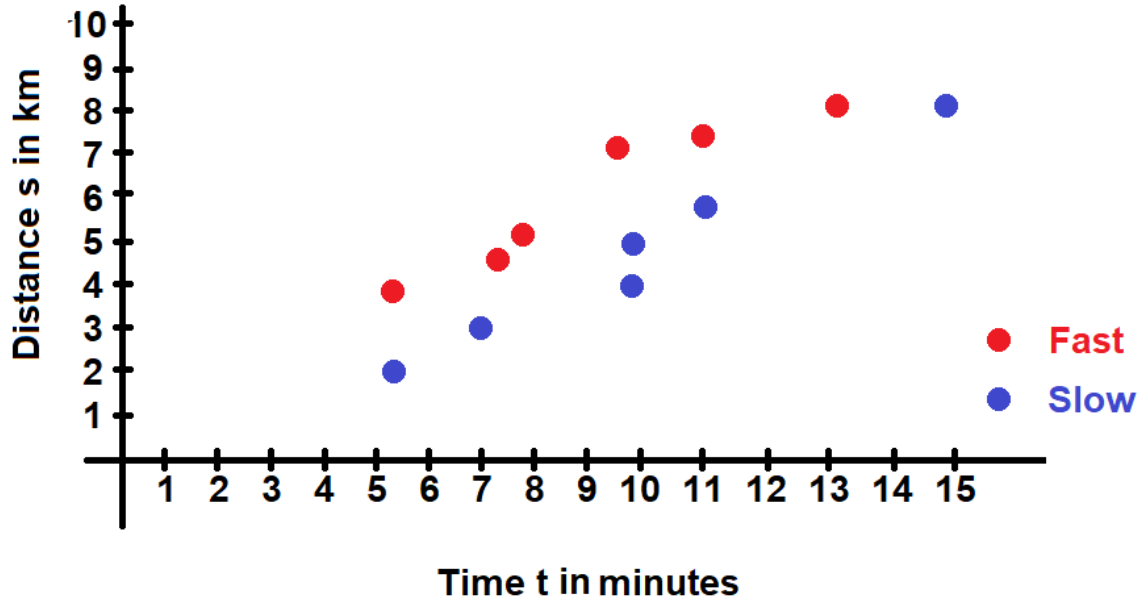
- Adjusting the weights (parameters)
  - A loss (reward) function
  - Back Propagation
- Functionality depends on training (i.e. deterministic)
- Training and Inference are separate processes
- Training takes a lot of time, energy and data

How do we train neural network…



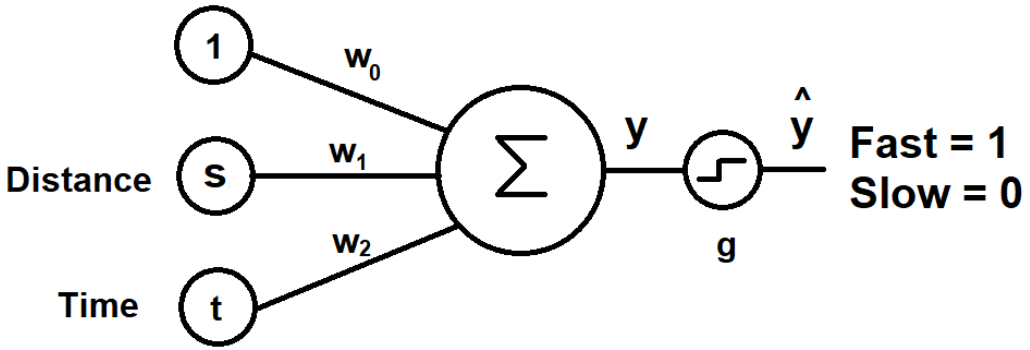
The simplest possible neural network…

- We want to create a neural network to decide (classify) if a racing cyclist is fast or slow given the distance travelled and time taken
- Here is the sample data that we have available to train our model. We have 12 data points and 6 are tagged slow and 6 are tagged fast

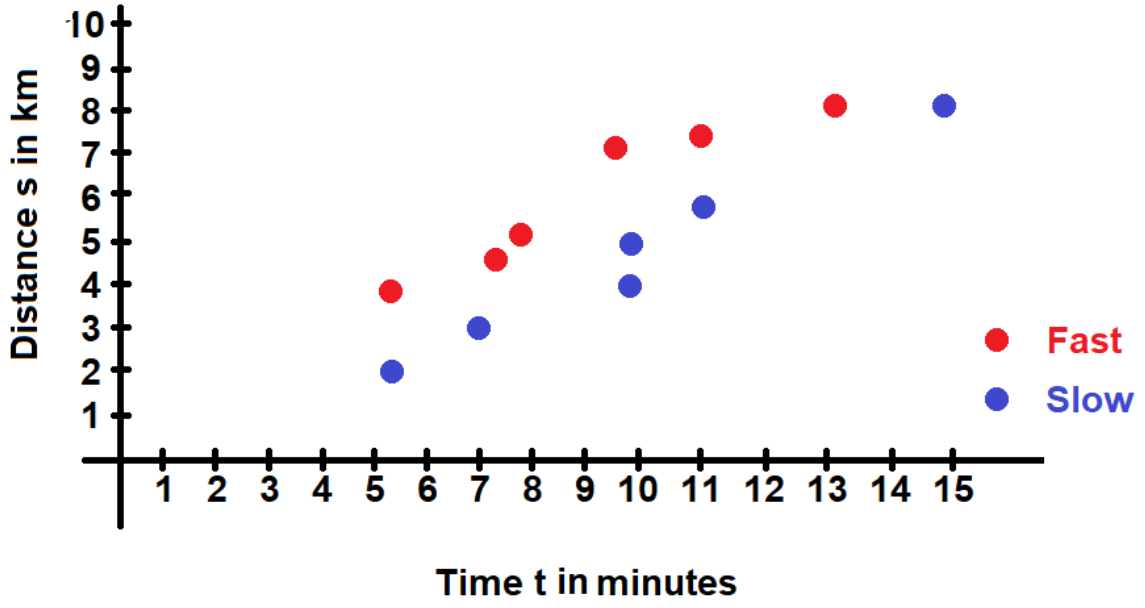


The simplest possible neural network…

- Let's try a single artificial neuron.

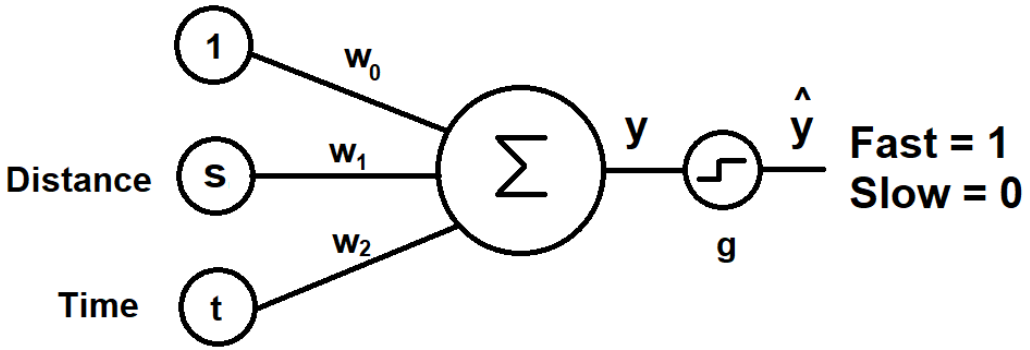


$$y = w_0 + sw_1 + tw_2$$

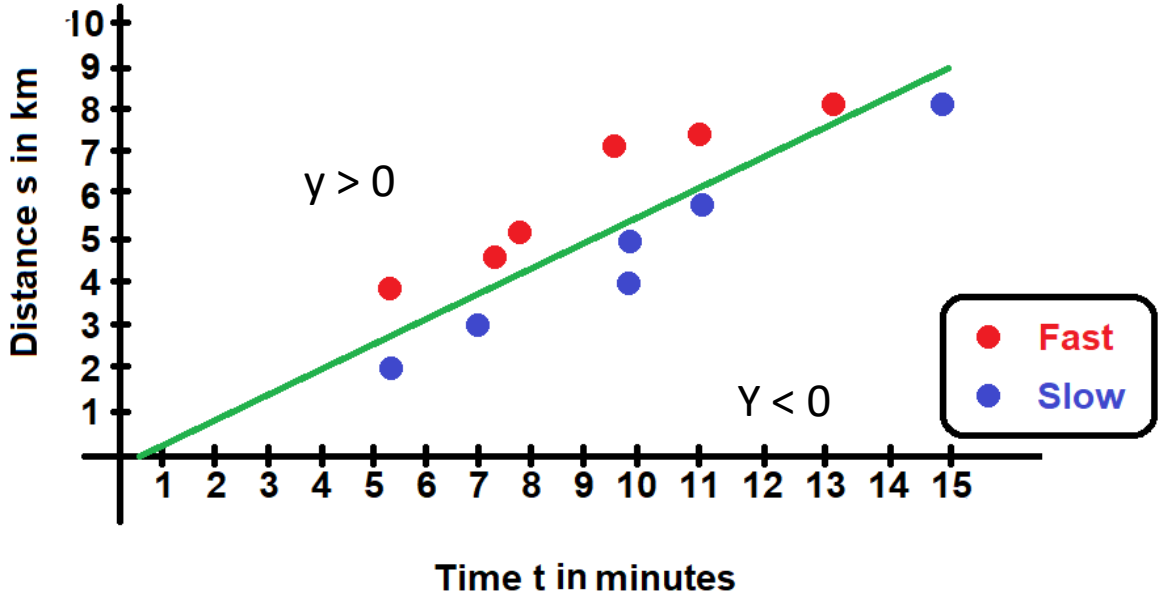


The simplest possible neural network…

- Let's try a single artificial neuron.



$$y = w_0 + sw_1 + tw_2$$



(4) Neural Network Types

Semantic Search (Vector Embedding)

Convolutional Neural Networks

Large Language Models

Morph a cute cat into a cute dog!

Semantic Search (Discriminative AI)

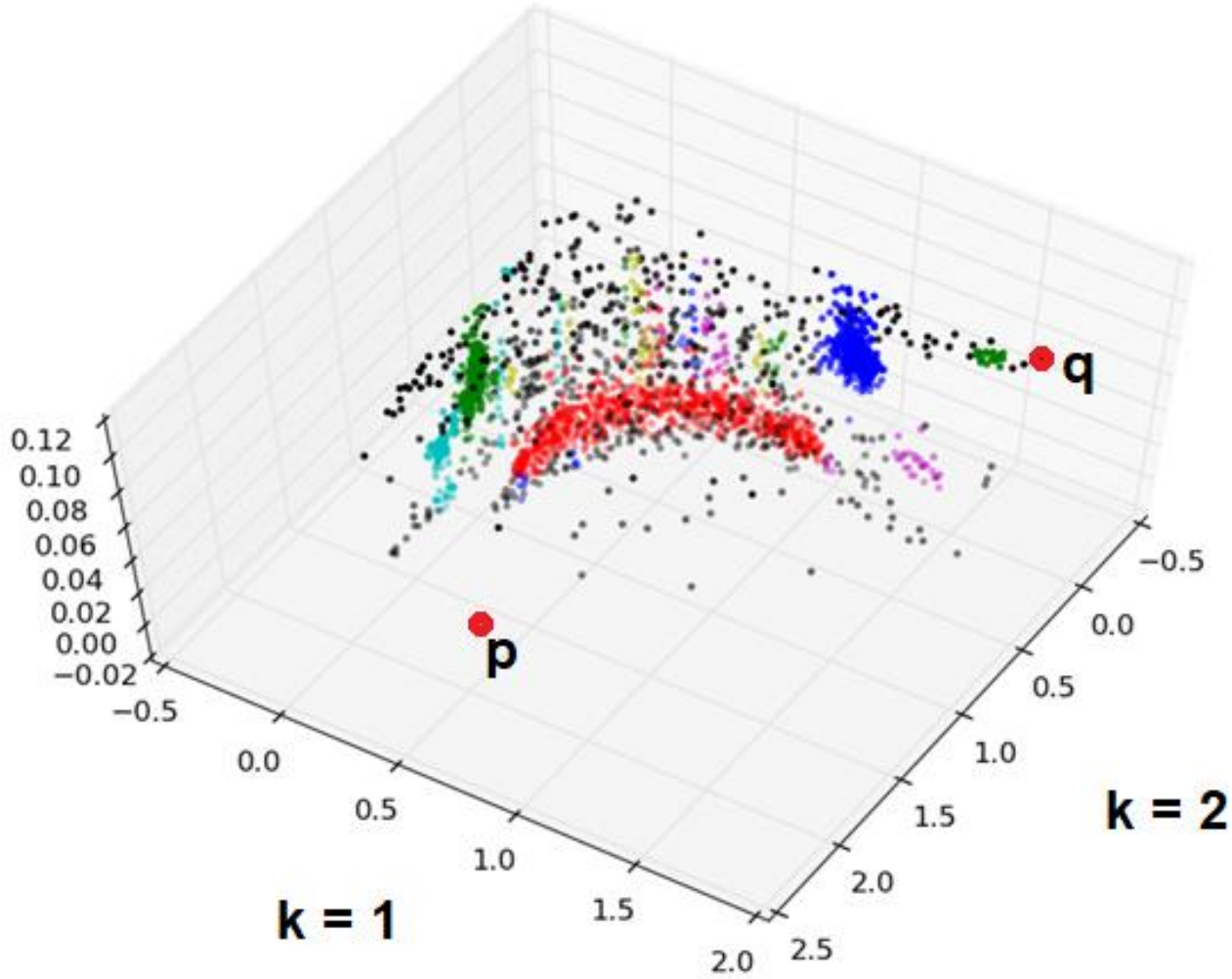
- Vector Embedding (Text, Audio, Image, Video)
- For text, turns a word, sentence, or chunk of text into a vector
  - Word2vec (Google, 2012)
  - Doc2vec (Google, 2014)
  - Universal Sentence Encoder USE (Google, 2018)
- Vector Database (Euclidian Distance)
  - Stores the object and its vector in a vector database
  - Search for the closest object to a new object
  - Hope that the closest question is close enough that its answer also answers this question

Embeddings in 3D visualisation

- p = [0.5, 2.0, 0.0]
- q = [1.7, -0.2, 0.07]

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) or data objects p and q.

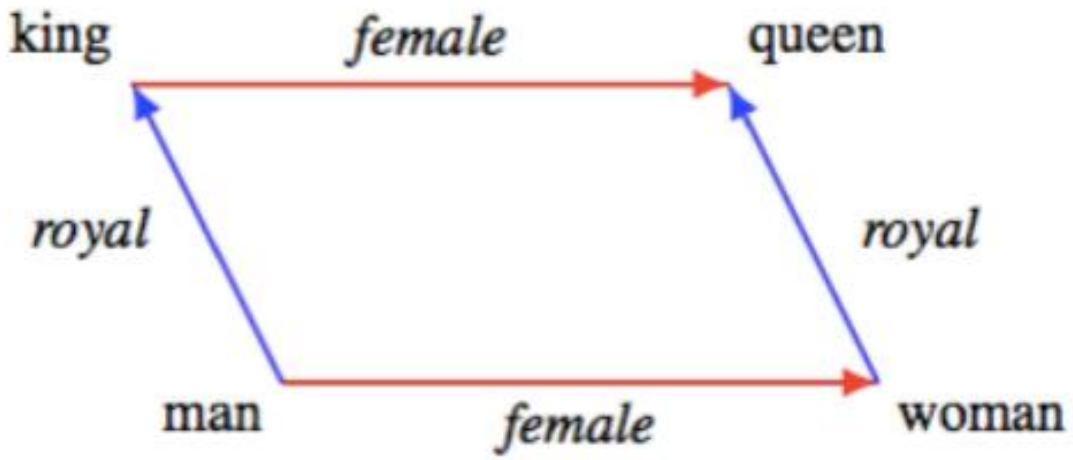


Embedding Neural Network

- Training Neural Network
  - Read lots of text and find similar usage
    - The King sat on the throne.
    - The Queen sat on the throne.
  - Read lots of text and find words that appear together.
    - The lion roared.
    - The fire roared.
    - The lion was on fire and it roared.
- Populating Reference Database
  - Create embeddings from reference data and place in vector database
- Perform Semantic Search
  - Create embedding from search term and find closest reference

What sorcery is this???

$$\overrightarrow{king} + (\overrightarrow{woman} - \overrightarrow{man}) \approx \overrightarrow{queen}$$



king = [2, 4 , 1, 3]
woman = [1, 3, 3, 1]
man = [2, 3, 3, 2]
queen = [1, 4, 1, 2]

Embedding Database Application Example
Frequently Asked Question (FAQ)

- Training
  - Collect all previous questions and the answers supplied
  - Create a USE vector for each question
  - Store the question, answer and the vector in a database
- Inference
  - Receive a new question from an end user.
  - Find the USE vector for the new question
  - Find the closest question to the new question, vector Euclidian distance
  - Offer the answer to the nearest question

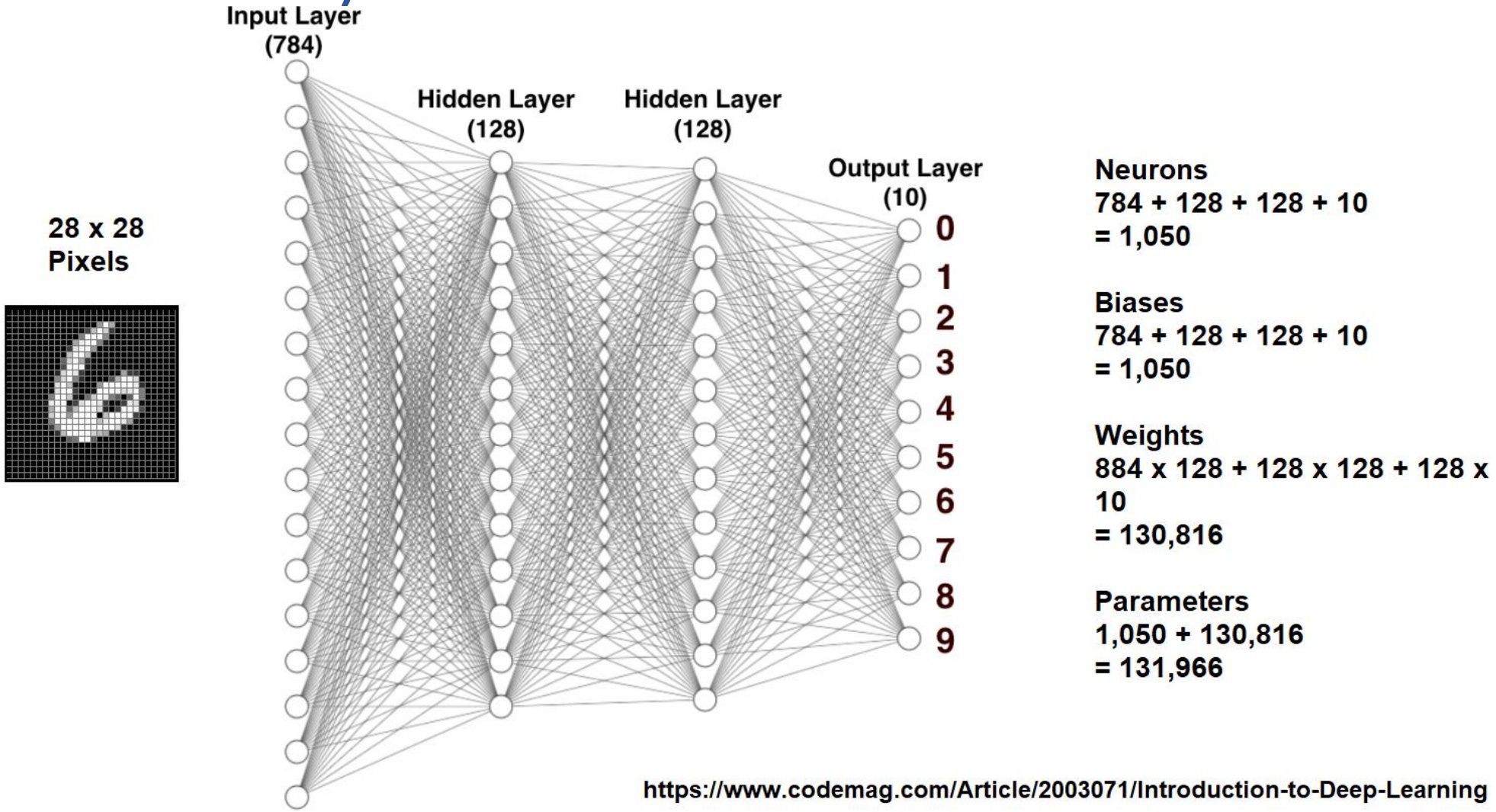
- Training Sample Data
  - Q: What is a 200 response?
  - A: Successful completion
  - Q: What is a 400 response?
  - A: Bad Request
  - Q: What is a 500 response?
  - A: Internal Server Error

- Inference Sample Data
  - Q: What does the HTML 500 error mean?
  - A: Internal Server Error

Convolutional Neural Networks

- NMIST data set
- Image Classification

The Deep Neural Network DNN
(as a Classifier)



Neurons
784 + 128 + 128 + 10
= 1,050

Biases
784 + 128 + 128 + 10
= 1,050

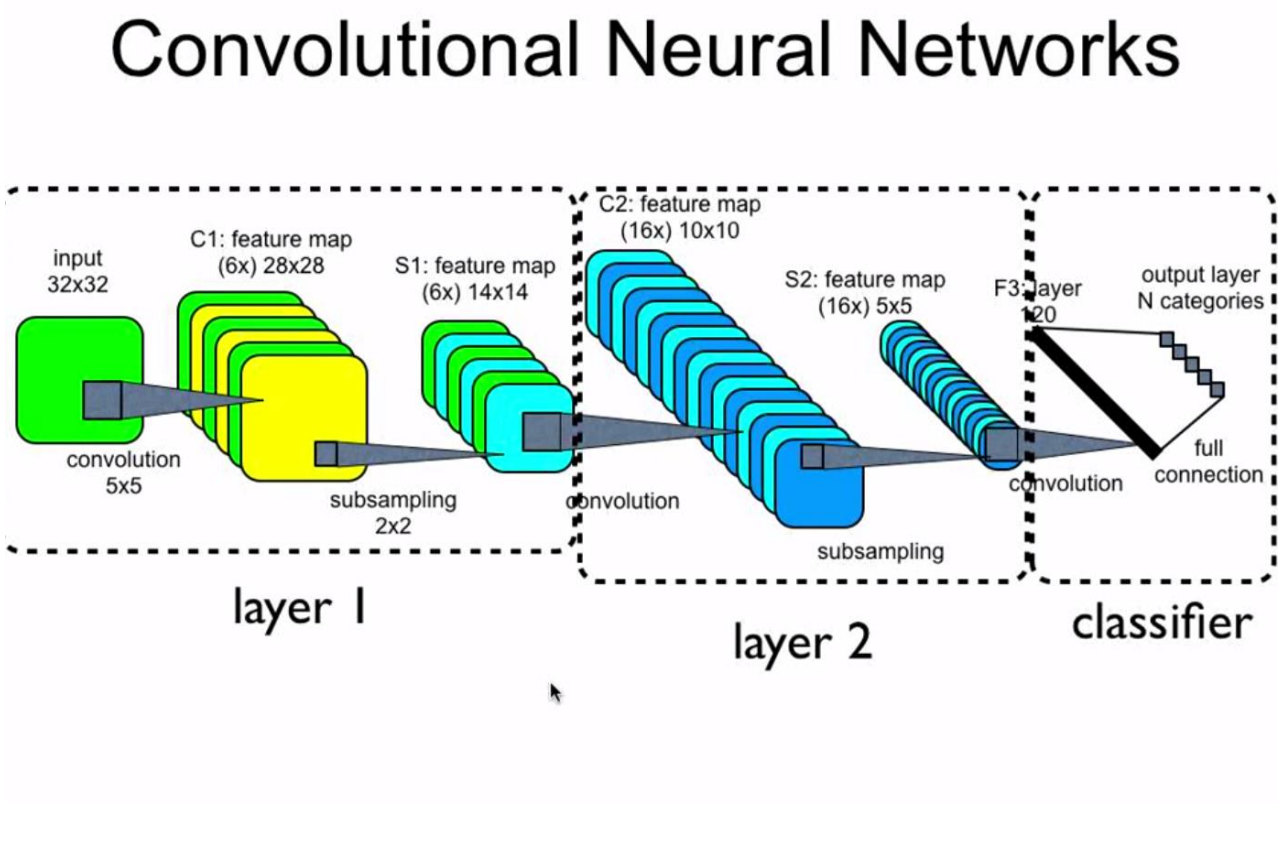
Weights
884 x 128 + 128 x 128 + 128 x 10
= 130,816

Parameters
1,050 + 130,816
= 131,966

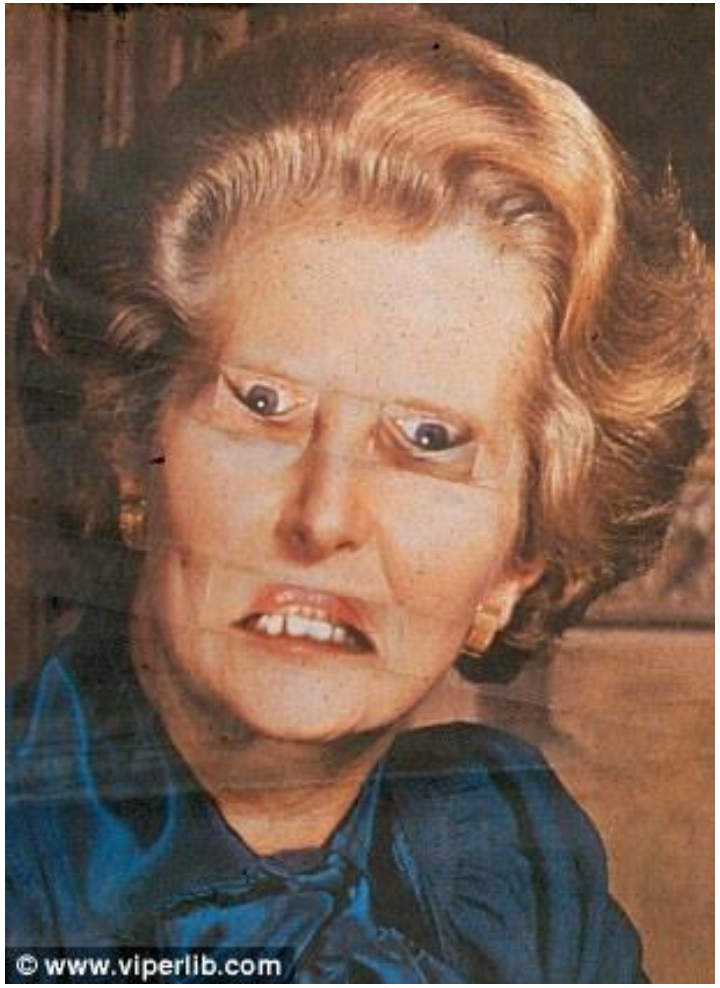
https://www.codemag.com/Article/2003071/Introduction-to-Deep-Learning

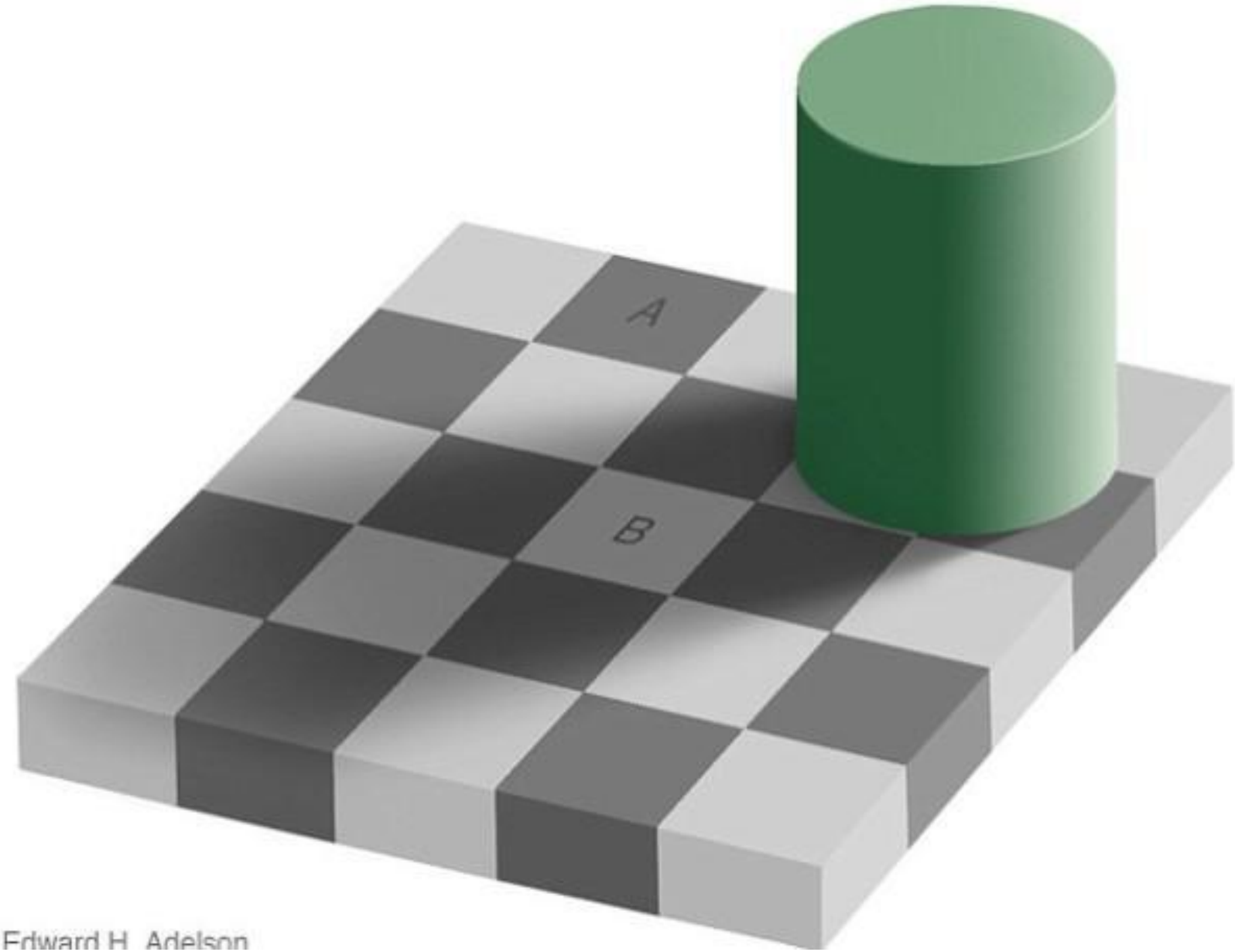
Convolutional Neural Networks

- Based on the Visual Cortex
- Curse of Dimensionality
- Convolutional Layers
  - Filters that scan the image.
- Pooling Layers
  - Reduce the size of the image
- Flat Layer fully connected.

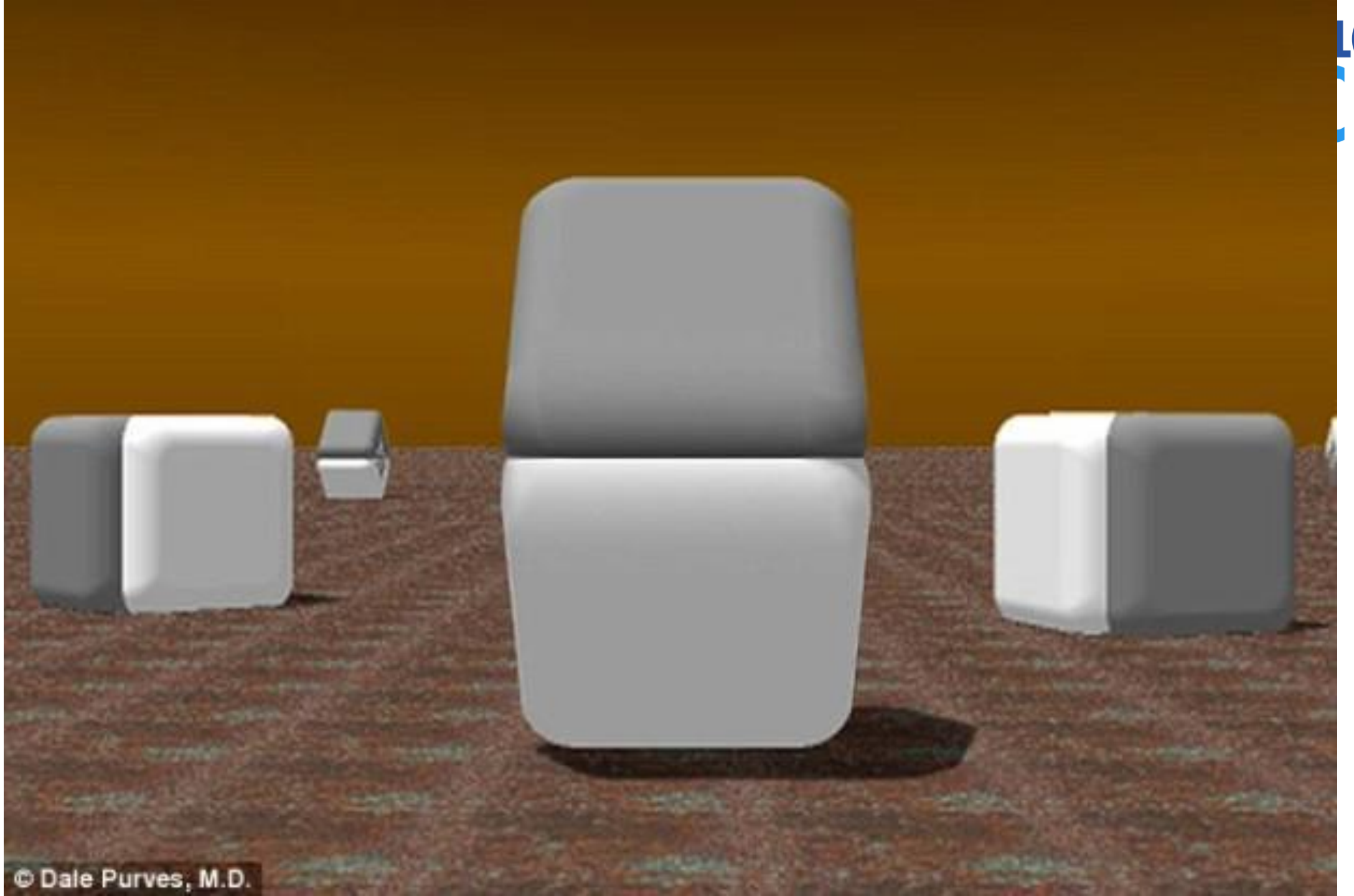


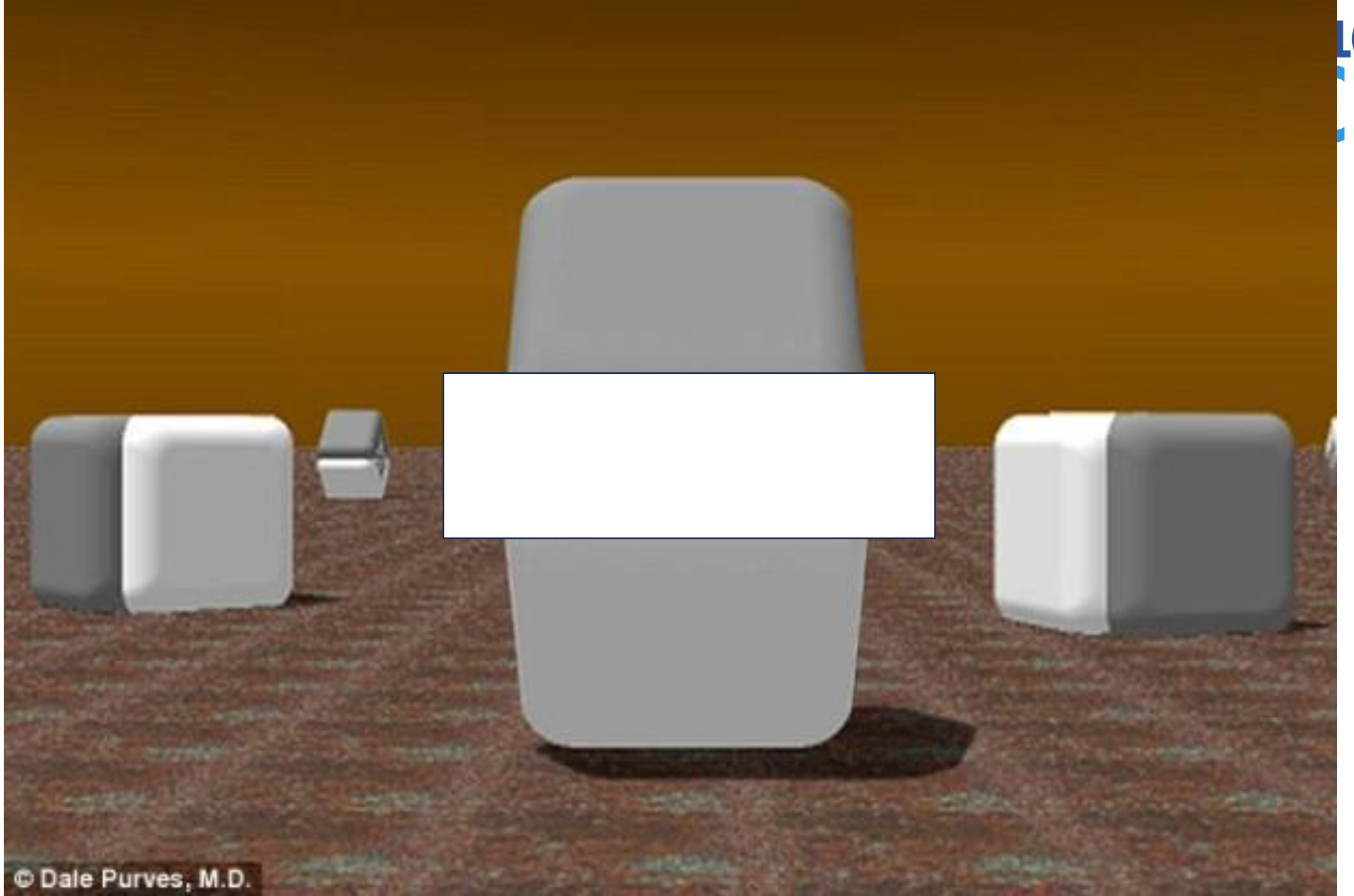
Detecting Patterns

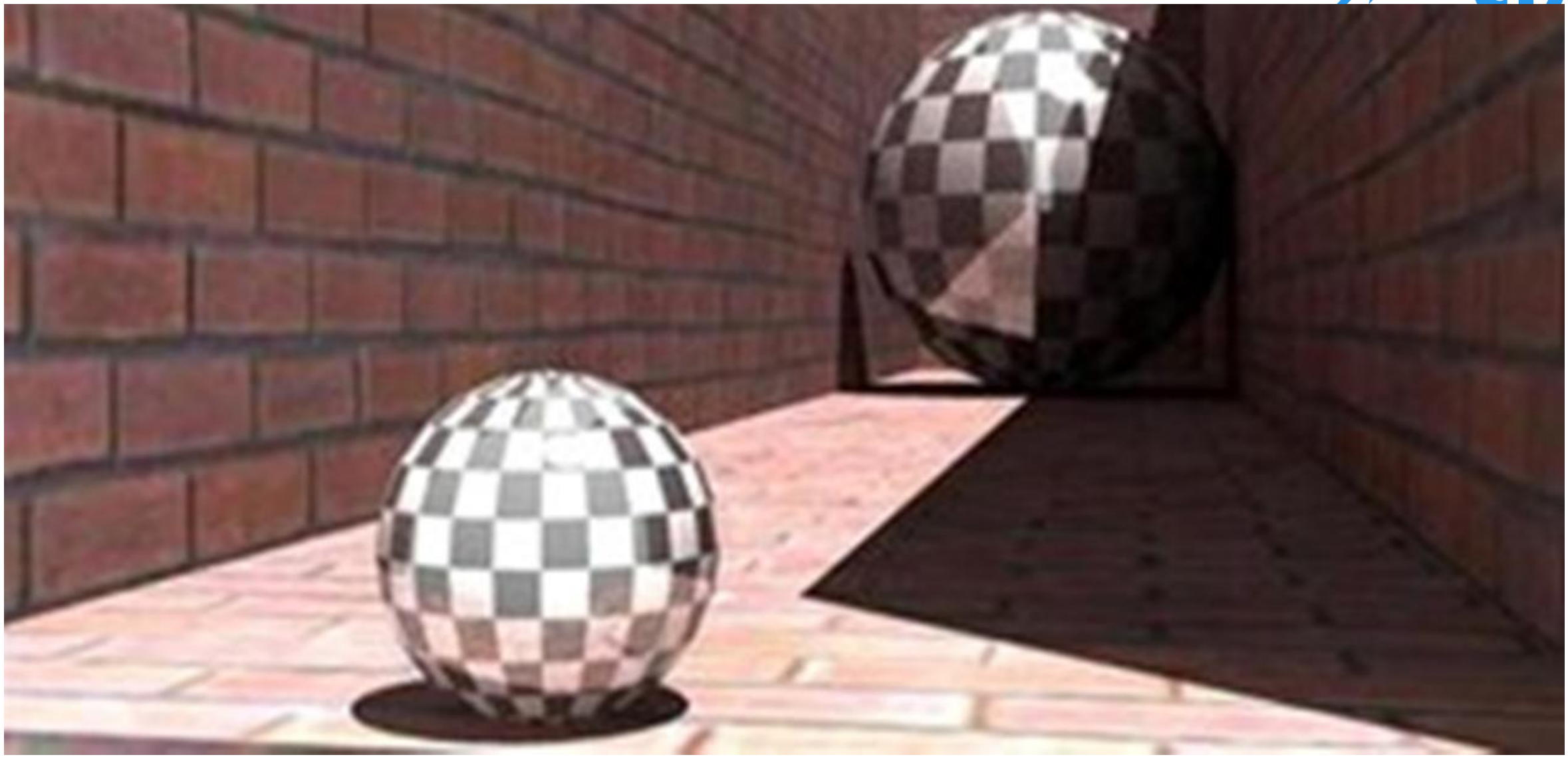




Edward H. Adelson

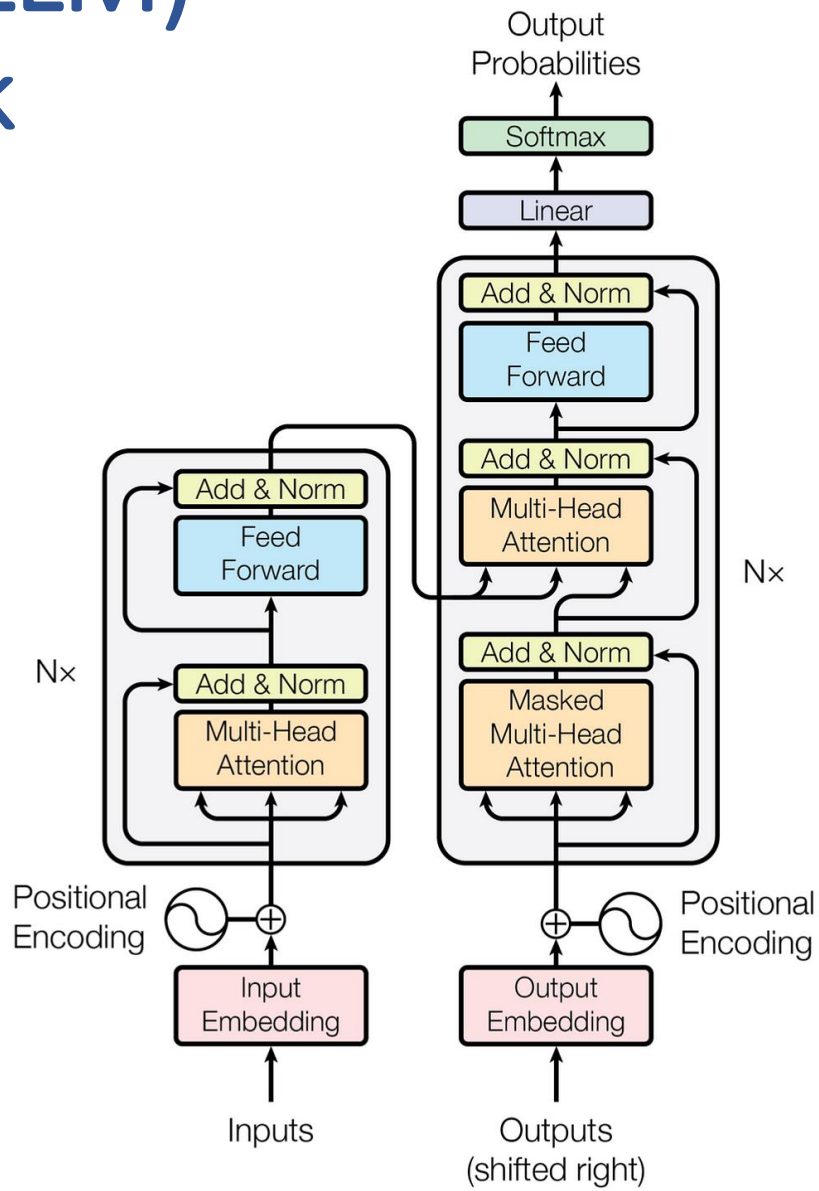

© Dale Purves, M.D.


© Dale Purves, M.D.

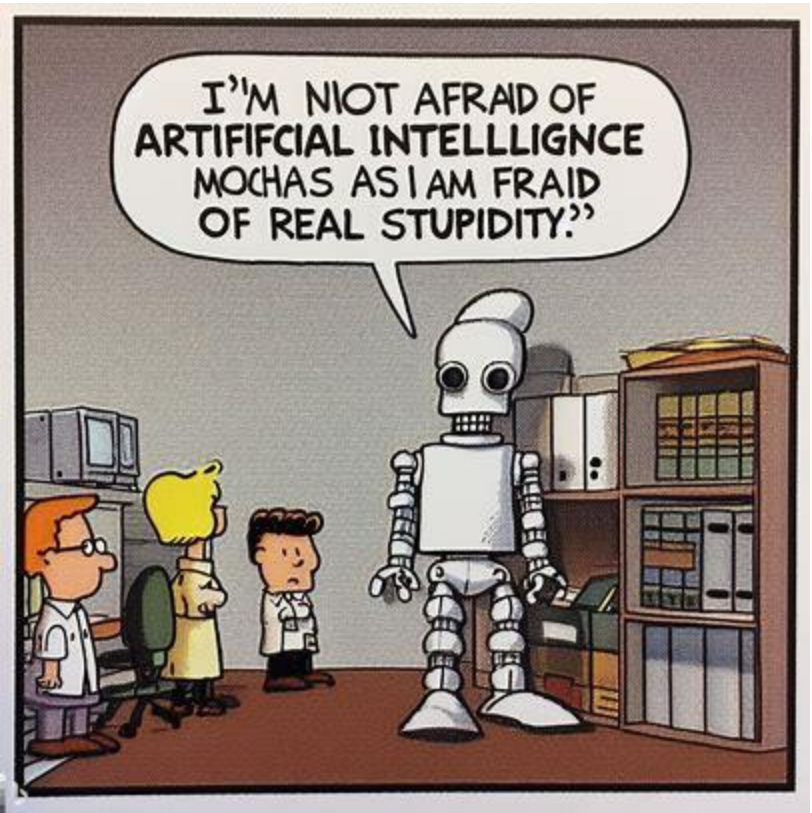


Large Language Models (LLM)
Transformer Neural Network

- Attention is all you Need, 2017, Google
- All LLMs today use the Transformer Neural Network
  - Generative Pretrained Transformer (GPT)
  - Google Bard



(5) LLMs in Detail



Predicting the next word (token)

Terminology

- Natural Language Processing (NLP)
  - Any software that processes language
- Large Language Model (LLM)
  - A text generative deep neural network
- Token
  - The vocabulary of the LLM
- Attention
  - The input buffer size of the LLM
- Prompt
  - The input to the LLM

Token

- The vocabulary of an LLM is its list of tokens. Each token is represented by an integer. LLMs process integers, not words or characters.
- Most LLMs have one token for simple common words. Complex words are constructed from subwords…
  - Unbelievable might have tokens ‘un’, ‘believ’ and ‘able’.
  - The token for ‘believe’ is not used.
- Capitalisation is ignored. ‘Apple’ and ‘apple’ receive the same token.

Auto Completion versus Question and Answer

- LLMs are initially trained by ‘reading’ a huge corpus of data. This gives them the ability to predict the next token.
  - ‘I like to eat ice’ is likely to be autocompleted with the token for ‘cream’.
- After autocompletion, the model is trained with examples of paired questions and answers.
- This requires prompt engineering to prepend the question with the text ‘question: ‘.

Training the LLM is a three step process

- Unsupervised Learning (Creates the ‘Base Model’ trillions of tokens)
  - Read the entire Internet, several times
  - The Model becomes an expert on predicting the next word (aka token)
- Supervised Fine Tuning (~ 10,000)
  - Train the model to answer questions by showing it examples of questions followed by correct answers
  - Starts answering everything as a question
- Reinforcement Learning with Human Feedback (~ 5,000)
  - A single question with non zero temperature = a bunch of answers
  - Ranked by humans 1 to 10
  - Build a ranking ‘reward model’
  - Train itself against the reward model

Prompt Engineering

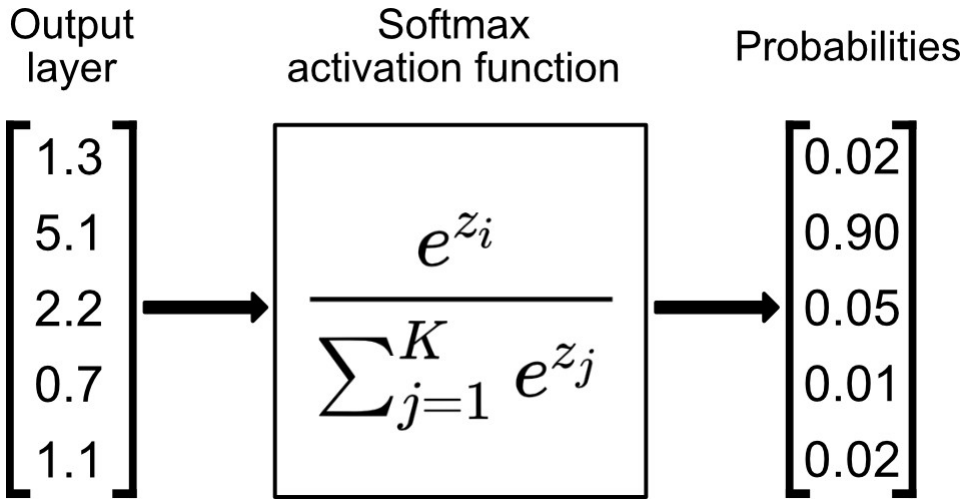
- Prompt engineering is the art and science of crafting prompts for LLMs.
- LLMs are deterministic. The same prompt always gives the same response (temperature will be discussed later).
- LLMs are stateless, in that they do no remember previous prompts.
- To create conversations, previous questions and answers must be included with the next prompt.

Attention

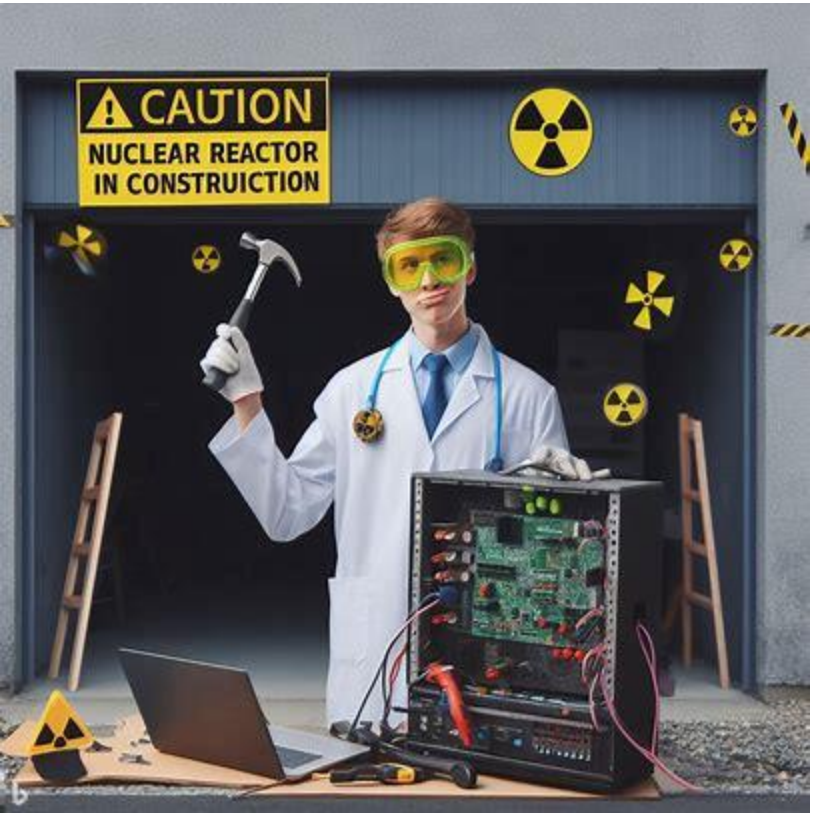
- Every LLM can process a maximum number of input tokens This number is called the ‘Attention’ of the neural network.
- y = f(x)
  - Attention is the dimension of the vector x.
  - An LLM produces a single token y on every loop.
- On the next cycle of the LLM, the previous prompt plus the freshly generated token are use as input to the LLM.
- If the prompt exceeds the attention, then the prompt is truncated.

Determinism & Hallucination

- LLMs are 100% Deterministic
  - Y = f(X)
  - Output = f(Prompt)
- Softmax & Temperature
  - Softmax turns outputs into probabilities
  - Temperature randomises the next token selection
- Hallucination
  - The function produces random rubbish
  - LLMs can hallucinate with great authority



(6) Let's Build a Custom Chatbot



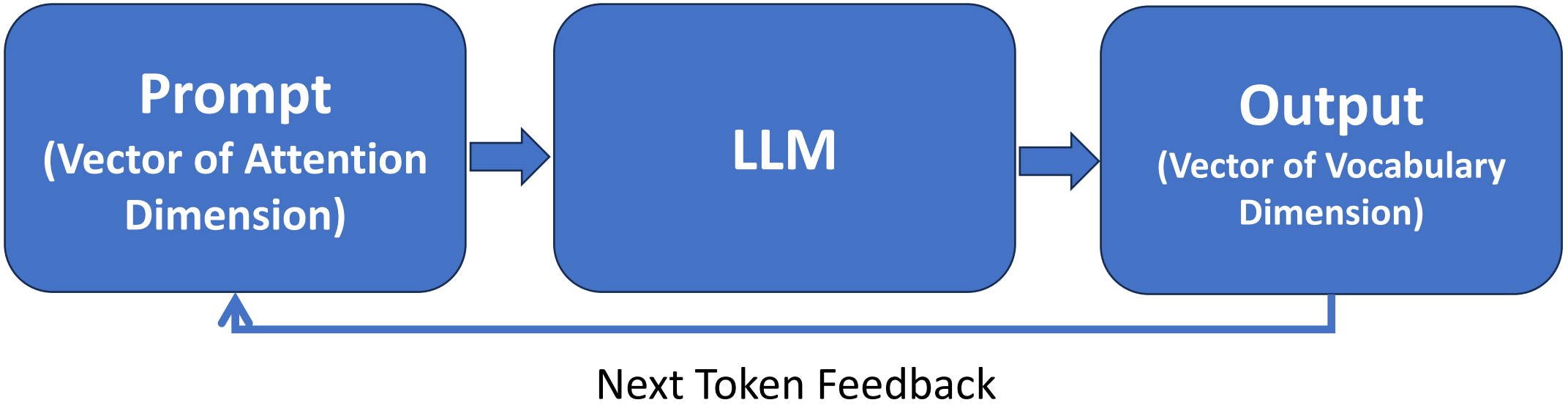
Trained on a corpus of data specific to our company.

Anatomy of a Custom Chatbot

- An LLM (after a Softmax layer) outputs a vector with a probability for each possible token.

Anatomy of a Custom Chatbot

- We take the most likely output token, prepend it to the prompt and feed the new prompt back into the LLM.



Anatomy of a Custom Chatbot

- Let’s put that feedback process inside the LLM box for simplicity.

Anatomy of a Custom Chatbot

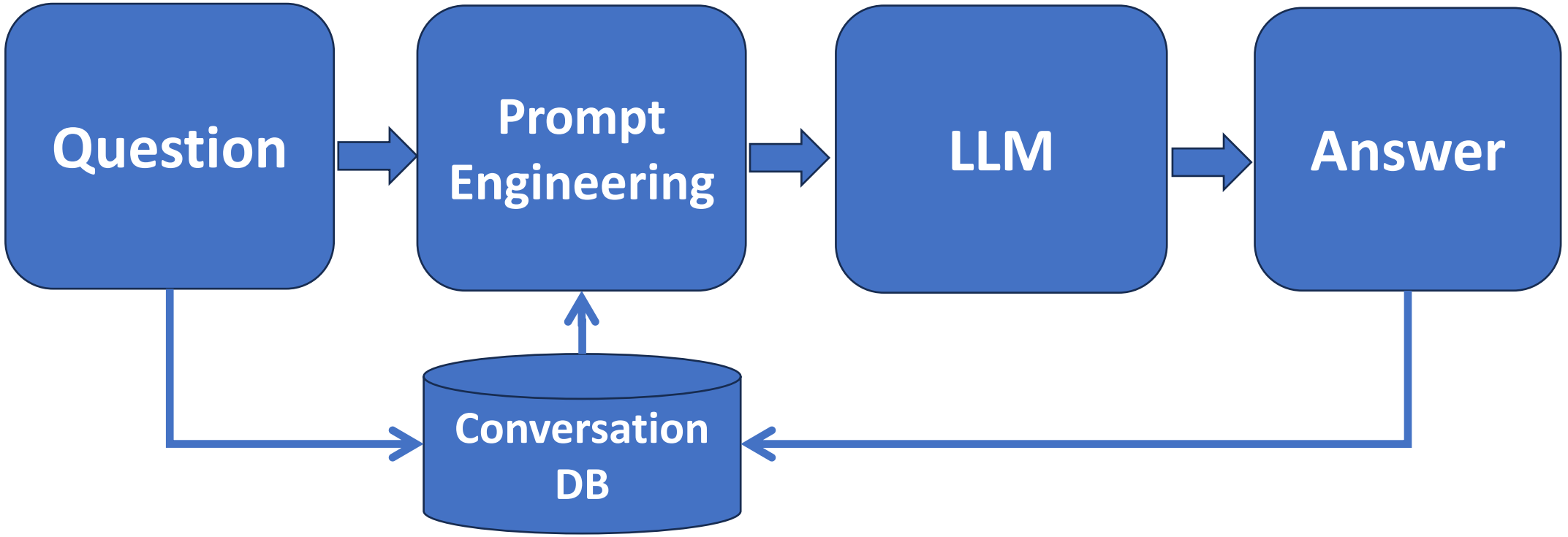
- The output is now the Answer, a complete list of predicted tokens.
- We now have our basic deterministic LLM model.

Anatomy of a Custom Chatbot

- With a basic LLM the next step it to add conversation capabilities.
- From here on we are in the realm of **Prompt Engineering**.

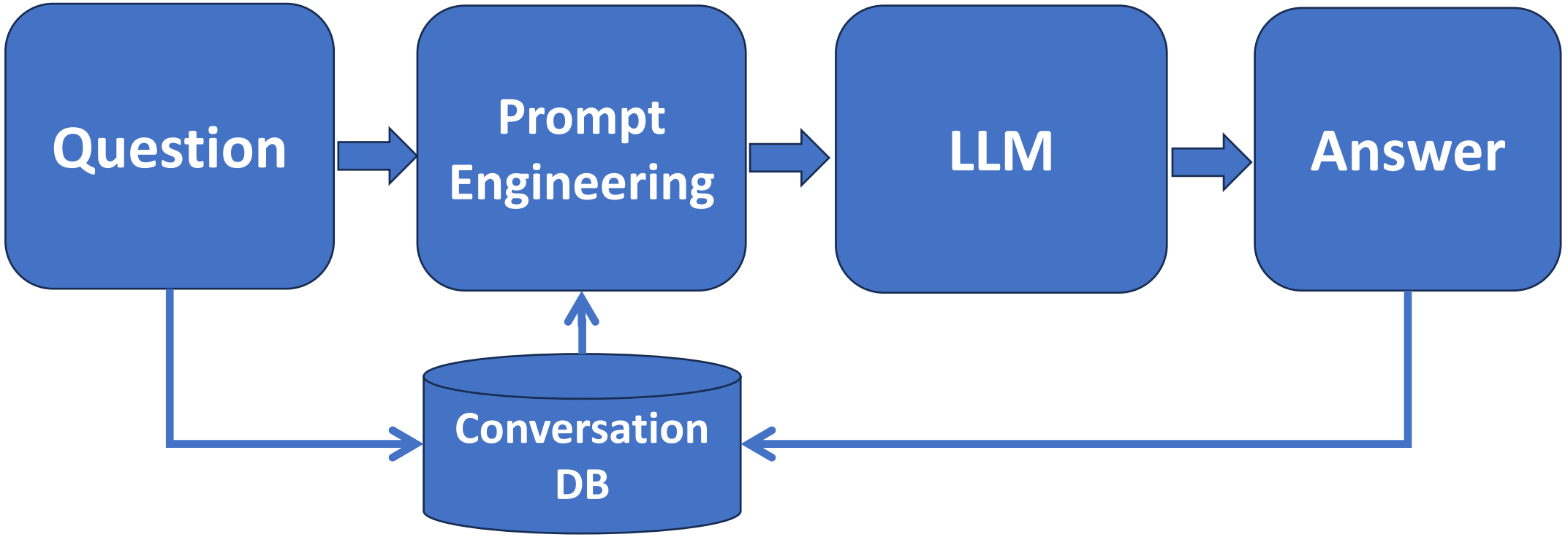
Anatomy of a Custom Chatbot

- To create a conversation, store previous Q&A and engineer a new prompt from old Q&A plus new question.



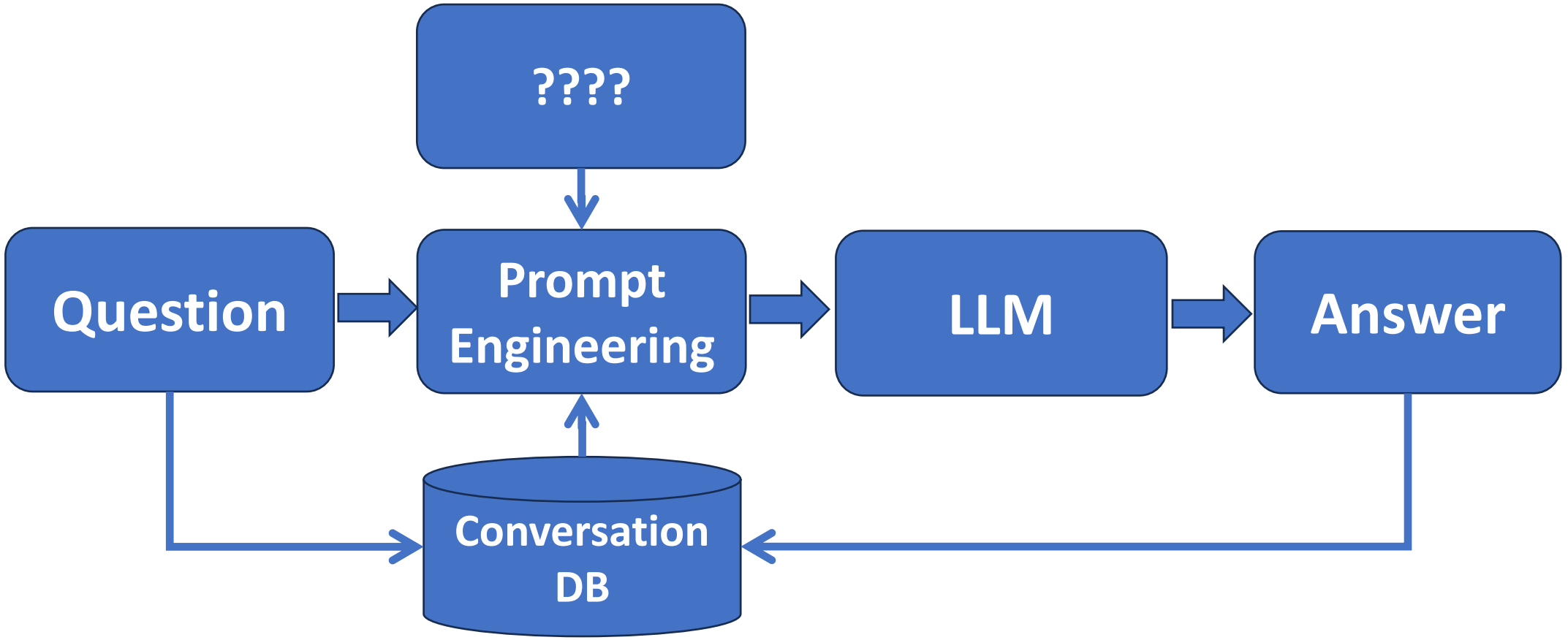
Anatomy of a Custom Chatbot

- What we have now is an application with the functionality of ChatGPT.
- How can we extend this model to ‘understand’ our custom use case?



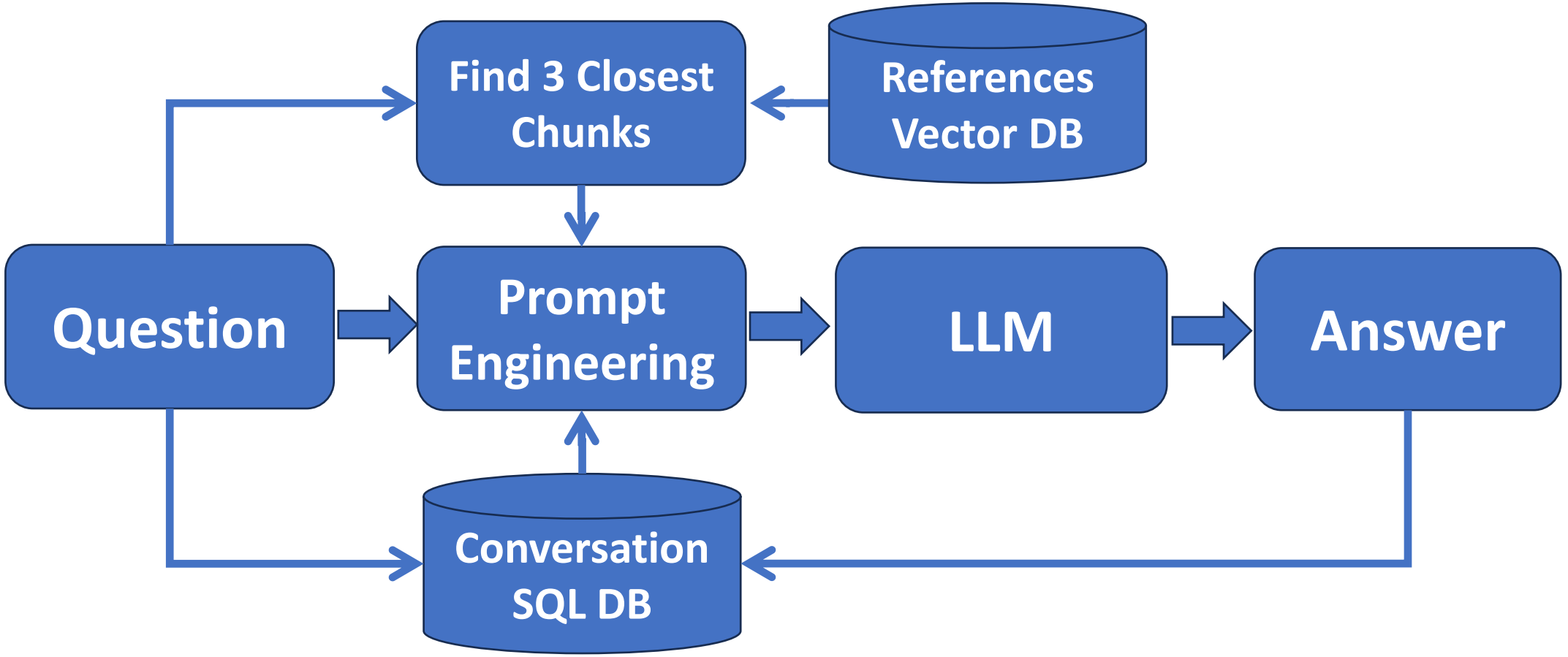
Anatomy of a Custom Chatbot

- We must incorporate our custom data into the prompt.

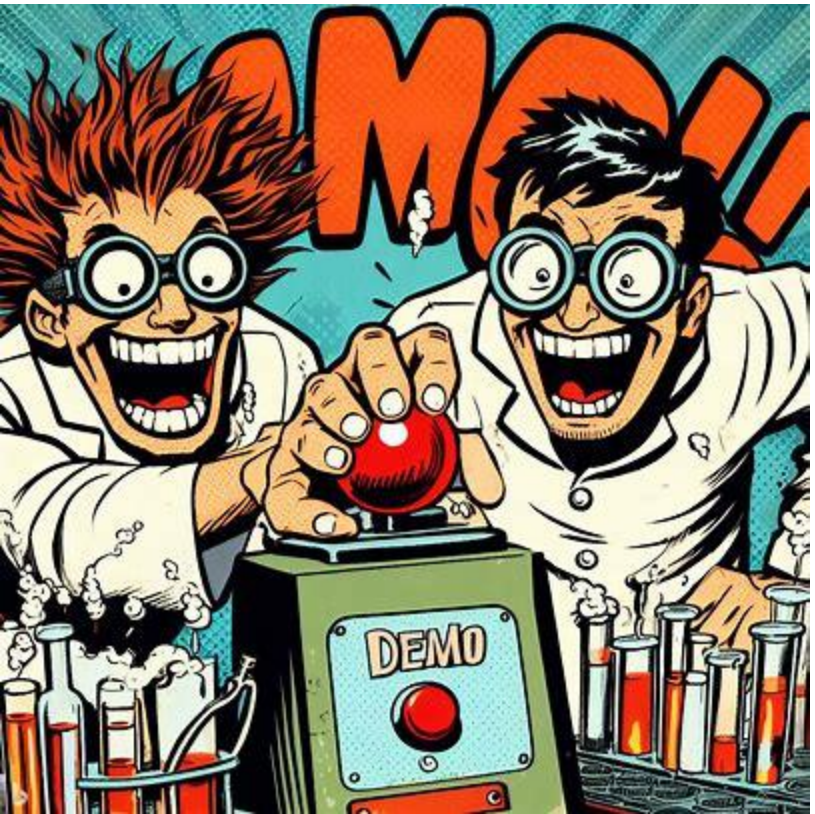


Anatomy of a Custom Chatbot

- We can use a vector database to select closest semantic ‘chunks’.



Demonstration



What could possibly go wrong?