A Primer on Artificial Intelligence

SOCIAL CAPITAL_

How to Read This Presentation

- This presentation aims to provide a common foundation on artificial intelligence, covering how it works and the different types of models in an intuitive way.
- Each section of this presentation builds on the prior and assumes no prior knowledge about the discussed topic. You should treat this presentation like a book and read each chapter one after the other, taking breaks in-between.
- At the end of each section, there will be a slide with links to further short readings and YouTube videos to reinforce and enhance your learning.
- By the end of this presentation, you should have a good understanding of what artificial intelligence is, how the different types of models work, the current wave of generative AI, and how close we are to AGI.

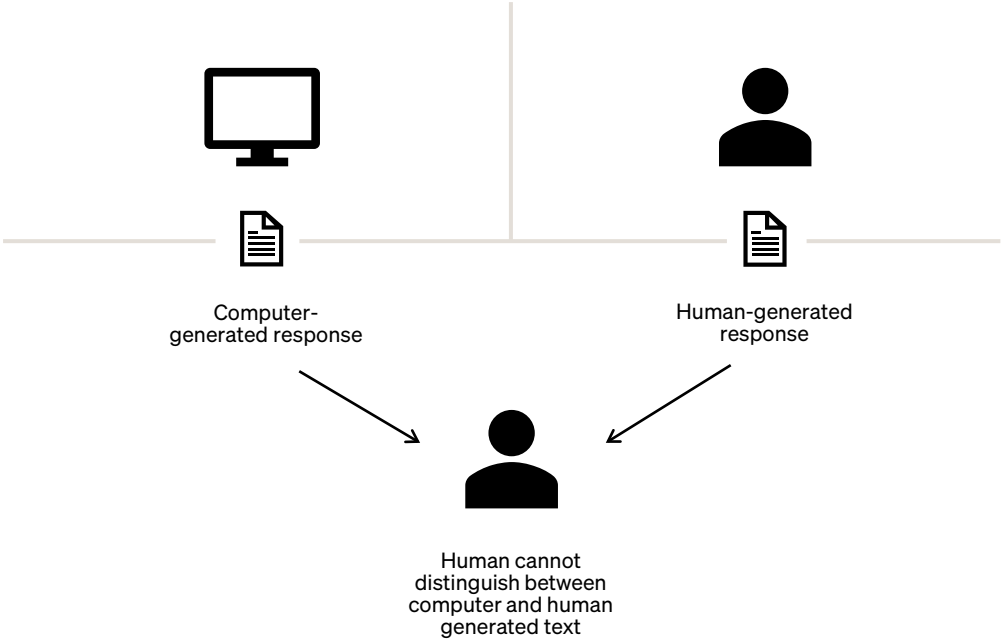
Table of Contents

Chapter		Page
01	Replicating the human brain	04
02	Computer circuitry 101	14
03	Computer logic 101	26
04	Introduction to neural networks	44
05	Types of neural networks	75
06	Introduction to large language models	133
07	Generative AI and value creation	161
08	Artificial general intelligence	195
09	Wrapping up...	206

CHAPTER 01

Replicating the Human Brain

The original aim of artificial intelligence was to pass the Turing Test, a challenge to build a computer that could respond to questions in a way that is indistinguishable from humans



Source: Saygin (2000), Turing Test 50 Years Later

To build a computer system that implements an artificial form of intelligence, we need to first understand human intelligence and how it works

Human intelligence is enabled by the brain, which works by representing sensory information electronically and chemically, and processing it through networks of neurons



Sensory information is represented electronically and chemically...

... and is then processed through networks of neurons in the brain

Source: MIT

When presented with an input, combinations of these neurons become active to recognize and classify it



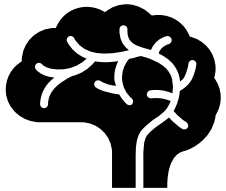
Image of an apple enters the eye, where light sensitive cells in the back of the retina convert the information into electrical signals

Signals then travel through the optic nerve to the brain, which interprets these signals as a series of shapes and colors

Then, neurons inside the brain communicate between each other through connections called synapses, which carry information to different areas of the brain for processing and storage

Combinations of these neurons are activated together based on previously learned associations, allowing the brain to recognize the image as an apple

The brain is a very complicated mixture of biological circuitry and logic...



Human Brain

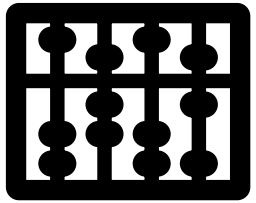
=



Biological Circuitry

The brain uses specialized cells called neurons that transmit and process information using electrical and chemical signals

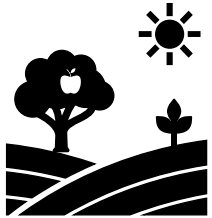
+



Logic

These neurons are organized into connected layers that communicate with each other to logically process information

...that can process large amounts of information at the same time



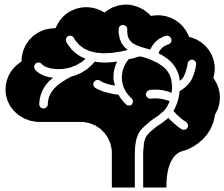


Sensory information is represented electronically and chemically



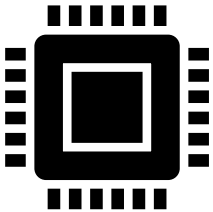
Brain recognizes that there is an apple in a field on a tree and it is sunny and feels warm all at the same time

Humans have attempted to replicate these intricate processes in computers using complex arrangements of electrical circuitry and logic



Human Brain

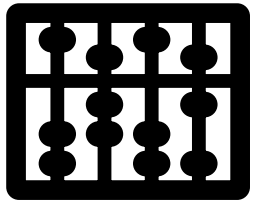
=



Electrical Circuitry

Humans have built electrical circuits that can represent, process and store information

+



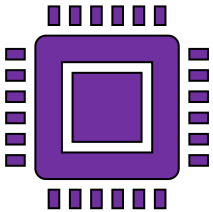
Logic

A series of logical architectures have been built on top of electrical circuitry to process information in a similar way to the human brain

What circuitry have we built to mimic the human brain?

Human Brain

=



Electrical Circuitry

Humans have built electrical circuits that can represent, process and store information

+

Logic

A series of logical architectures have been built on top of electrical circuitry to process information in a similar way to the human brain

Dive Deeper...

Further Reading & Watching

Reading:

- Can Artificial Intelligence Replicate the Human Brain? (Medium)
- Brain Anatomy and How the Brain Works (Johns Hopkins Medicine)

Watching:

- The Turing test: Can a Computer Pass for a Human? (Ted)
- The Science of Thinking (Veritasium)
- Building a Computer Like Your Brain (Bloomberg)

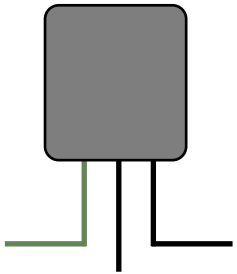
CHAPTER 02

Computer Circuitry 101

Computers are made of tiny electronic switches called transistors, which can instantly switch between on and off states by applying a voltage

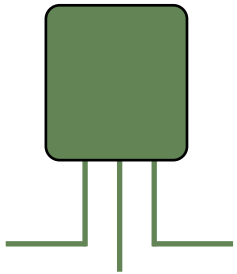
Apple's M3 Max chip contains 97 billion transistors

Off



Transistor is off because no voltage is applied

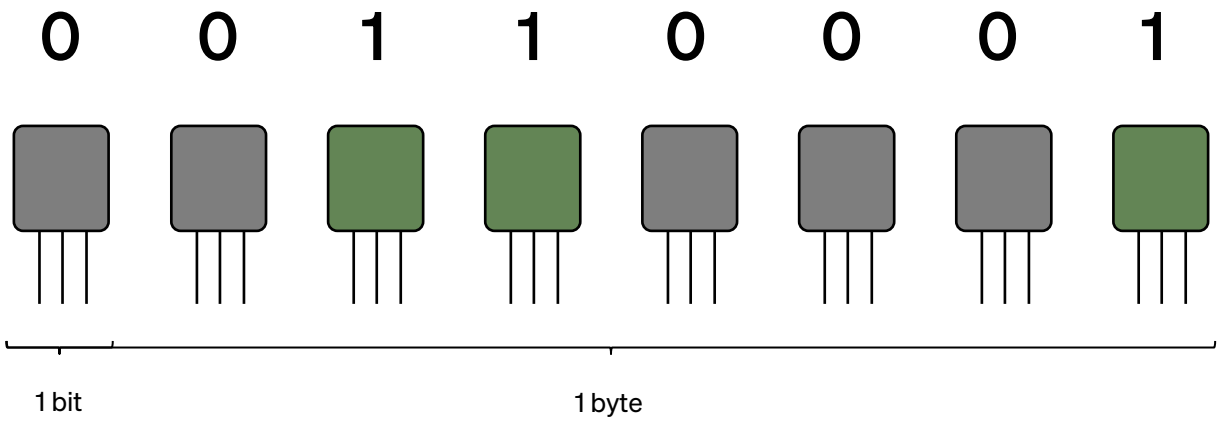
On



Transistor is on after a voltage is applied

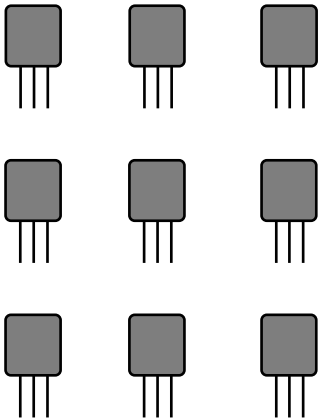
The on and off states of transistors represent binary digits of 0 and 1 or ‘false’ and ‘true’, which computers use to represent and process information

A single transistor is called a ‘bit’, and 8 ‘bits’ make a ‘byte’

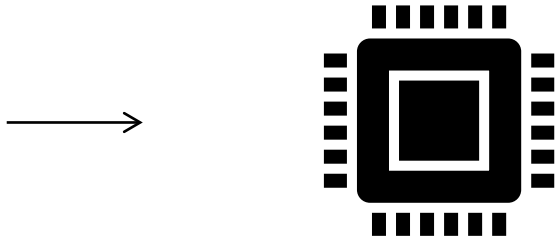


Multiple transistors can be organized together to build a processing unit called a ‘core’, which can take input data and process it according to defined instructions

Multiple Transistors…

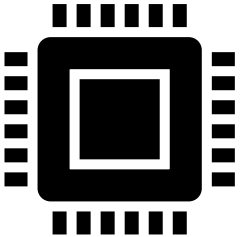


Organized into a single core



A single core can process tasks one after the other

Single Core

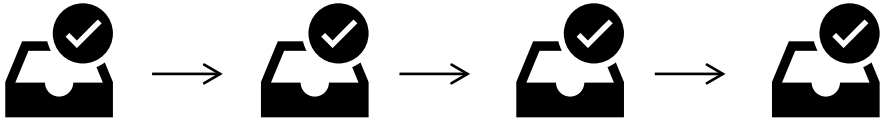


Complete Task A

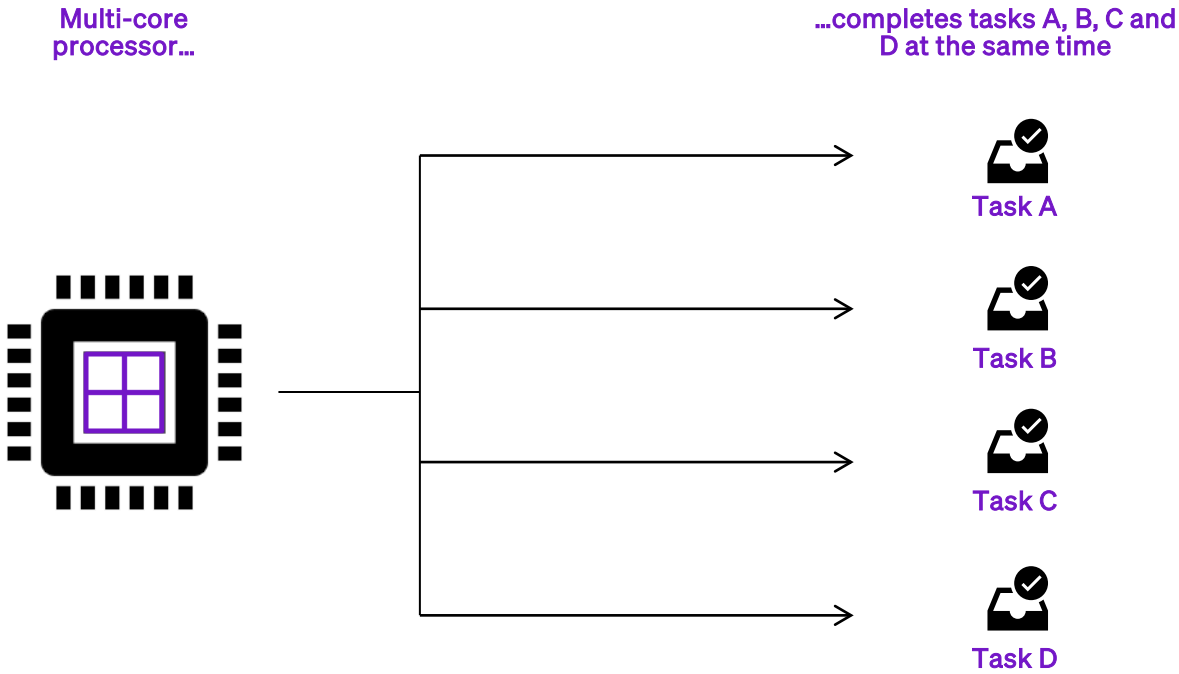
Complete Task B

Complete Task C

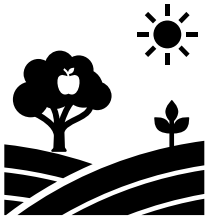
Complete Task D



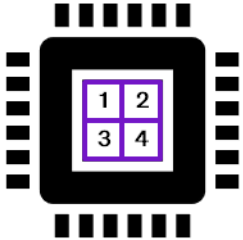
So processors are often organized into multiple cores, which allows multiple tasks to be completed in parallel



CPUs are built using a limited number of cores which reduces the amount of information that they can process at the same time, but they can do it very quickly



CPU



Core 1: It is sunny

Core 2: There is a field

Core 3: There is a tree

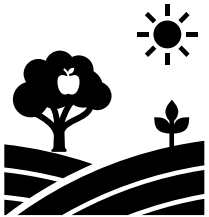
Core 4: There is an apple

Human Brain

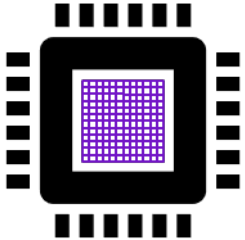


It is sunny and warm and there is a mild breeze and there is a field and the grass is green and there is a tree and there is an apple

Specialized processors called GPUs were developed with thousands of cores to process massive amounts of information simultaneously



GPU



Core 1 → Core 1,000+

It is sunny and warm and there is a mild breeze and there is a field and the grass is green and there is a tree and there is an apple

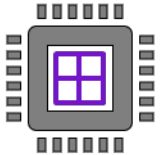
Human Brain



It is sunny and warm and there is a mild breeze and there is a field and the grass is green and there is a tree and there is an apple

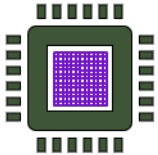
Today, there are several new processor architectures which attempt to improve on the circuitry of the GPU

CPU



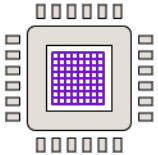
- Low number of cores
- Each core processes tasks very quickly
- Can execute a handful of operations at once
- Useful for general computing tasks

GPU



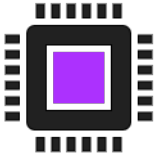
- Thousands of cores
- Each core processes tasks slower than CPUs
- Can execute thousands of operations at once
- Useful for graphics processing and other more specialized parallel computing tasks

TPU



- Fewer cores than GPUs
- Each core is extremely efficient at machine learning tasks
- Can execute many operations at once
- Useful for machine learning and other tasks which require matrix multiplication

LPU



- Single core architecture, but 16 chip-to-chip interconnects
- Hyper-efficient at sequential tasks like language processing
- Can output words in a language model rapidly
- Useful for specifically running large language models

Source: Nvidia, Google, Groq

As circuitry has become more advanced, the logic systems built on top have also become more advanced to more closely replicate how the human brain works

What logic systems have we created to mimic the human brain?

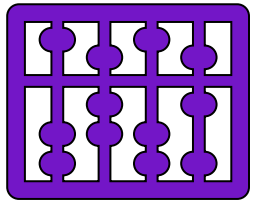
Human Brain

=

Electrical Circuitry

Humans have built electrical circuits that can represent, process and store information

+



Logic

A series of logical architectures have been built on top of electrical circuitry to process information in a similar way to the human brain

Dive Deeper...

Further Reading & Watching

Reading:

- Central Processing Unit (Khan Academy)
- What's the Difference Between a CPU and a GPU? (Nvidia)
- Groq's Record-Breaking Language Processor Hits 100 Tokens Per Second On A Massive AI Model (Forbes)

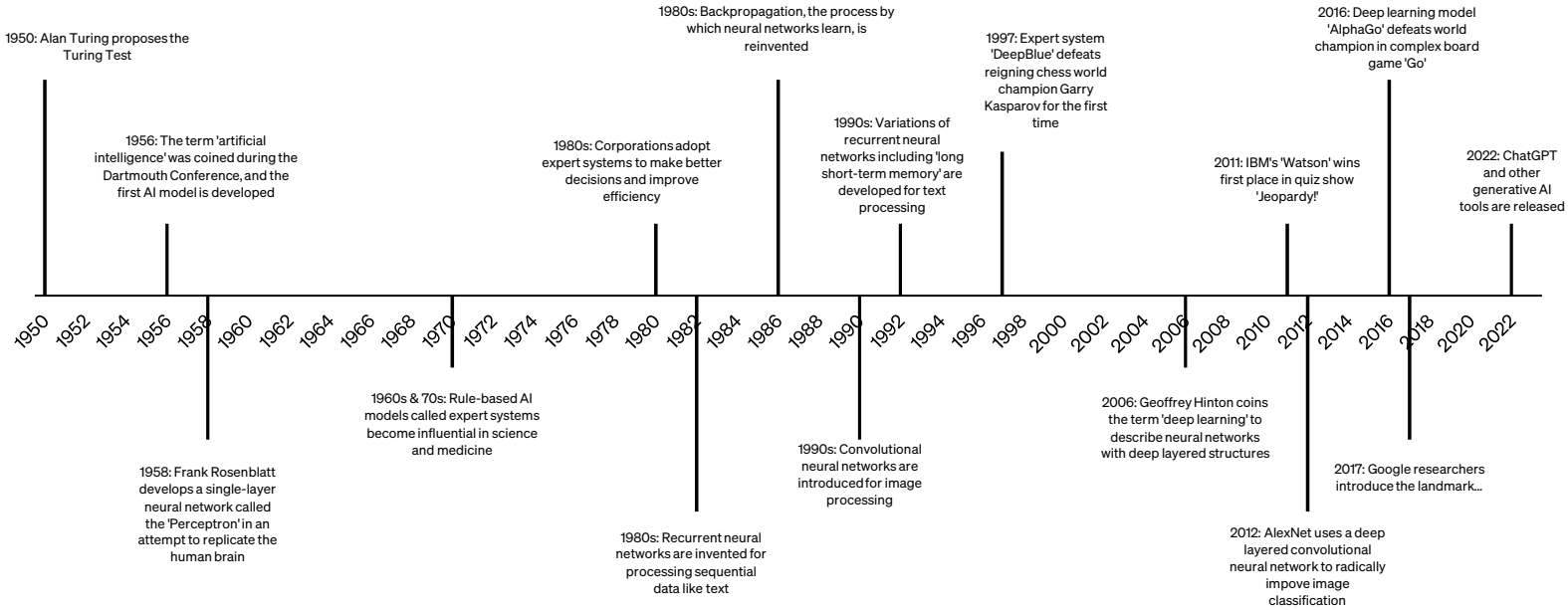
Watching:

- How Does a Transistor Work? (Veritasium)
- How Computers Work (Basics Explained)
- GPUs: Explained (IBM)
- Tensor Processing Units: History and Hardware (Google Cloud Tech)

CHAPTER 03

Computer Logic 101

Developments in computer logic have built on top of each other since the late 1950s, resulting in AI models that can increasingly replicate aspects of human intelligence



Source: URBS-Lab with Ryan Urbanowicz

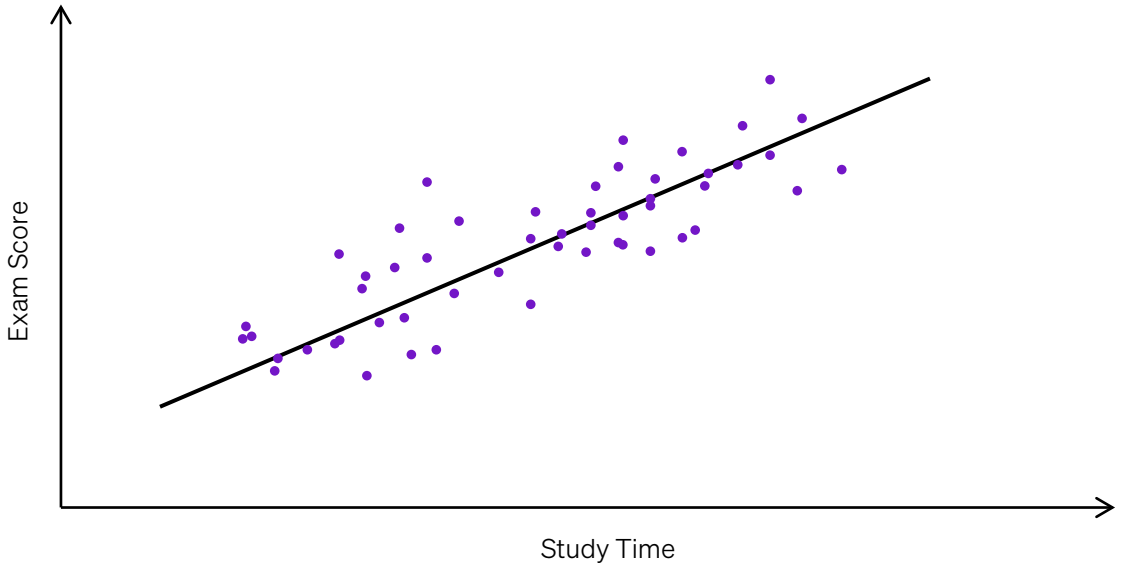
All computer logic systems are increasingly complicated forms of statistics and probabilities

To make an educated guess using probabilities, computers need to first process the underlying data

This involves learning patterns from data, distinguishing between different types of data, and then creating rules based on data

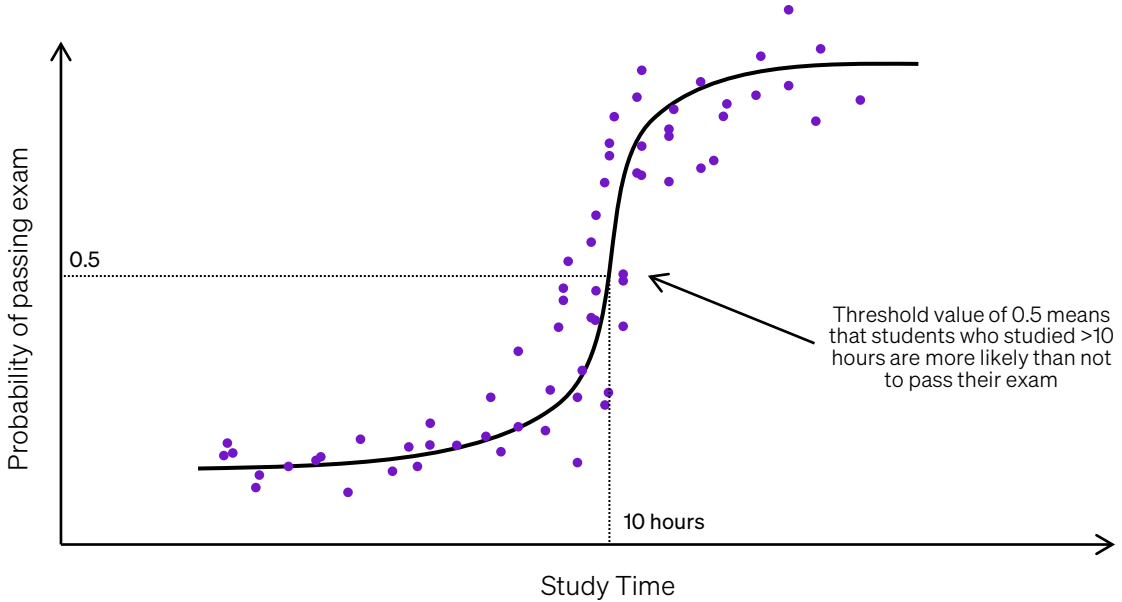
How do computers
learn patterns from data?

Linear regression is a statistical technique which allows computers to learn the relationship between two or more variables by fitting a ‘line of best fit’ through the data



Source: Analytics Vidhya

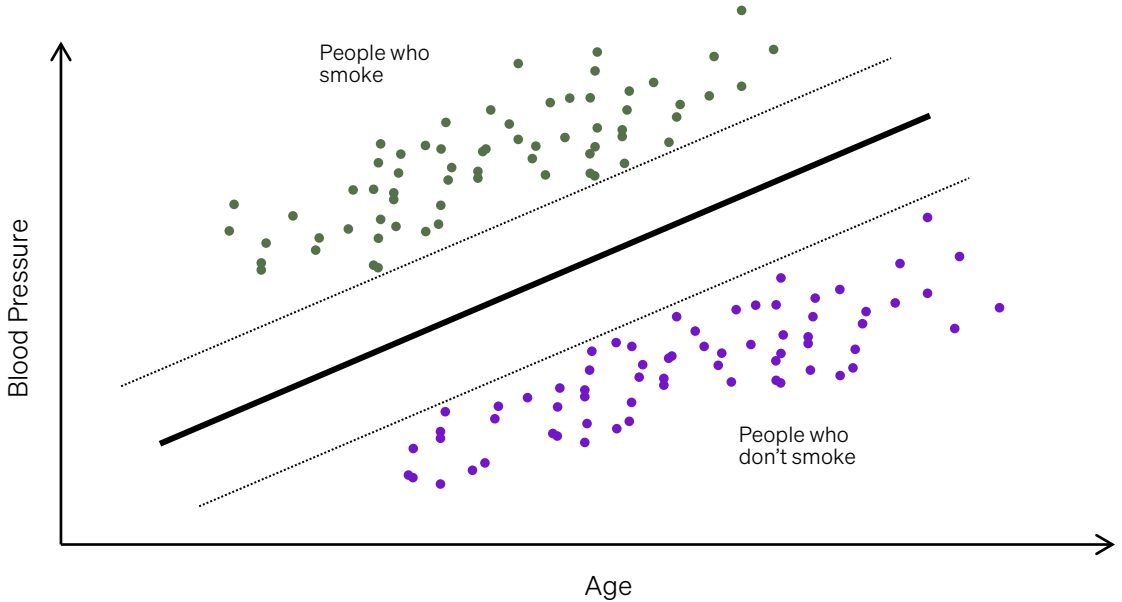
Computers can take the ‘line of best fit’ from a linear regression and squeeze it into an s-shaped curve between 0 and 1 to translate this relationship into an associated probability



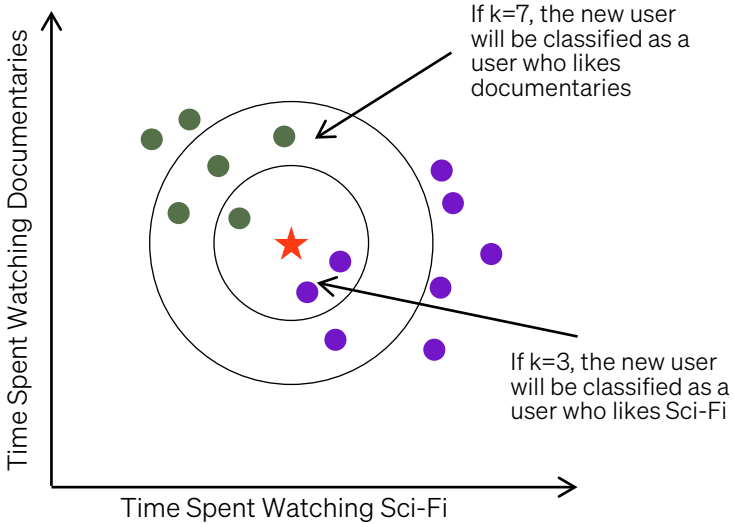
Source: Science Direct, StatQuest

How can computers distinguish
between different types of data?

Support vector machines allow computers to distinguish between data by finding the line that maximizes the margin between two different categories of data



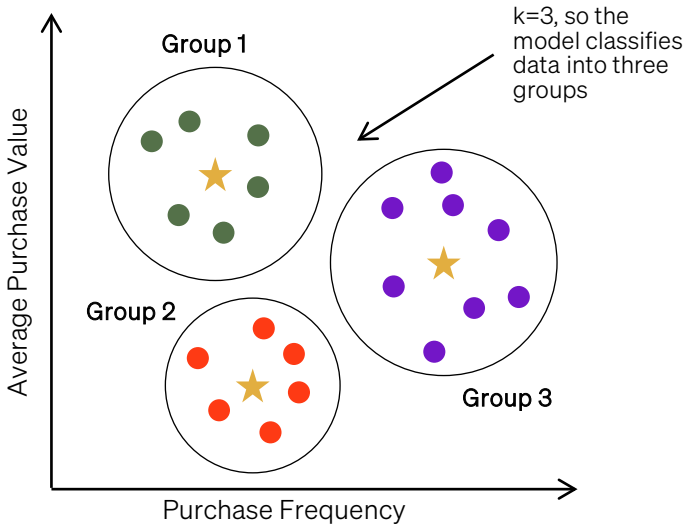
Computers can use the ‘k-nearest neighbors’ algorithm to classify data based upon how close it is to a defined number of datapoints surrounding it



Legend

- Users who like Sci-Fi
- Users who like Documentaries
- New user

K-means classifiers allow computers to group data into a specified number of ‘k’ clusters, based on how close each datapoint is to the central points of each cluster



Legend

- Casual shoppers
- Bargain hunters
- Loyal customers
- Central point

Source: StatQuest

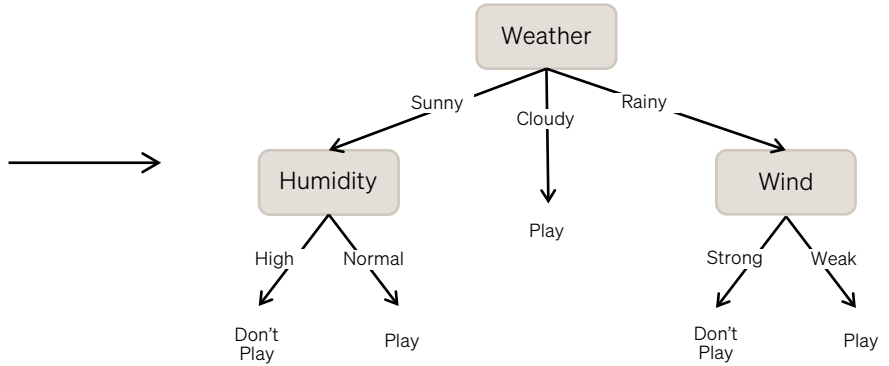
How can computers
create rules based on data?

Computers can learn to break down complex questions into smaller, simpler questions using a technique called a decision tree

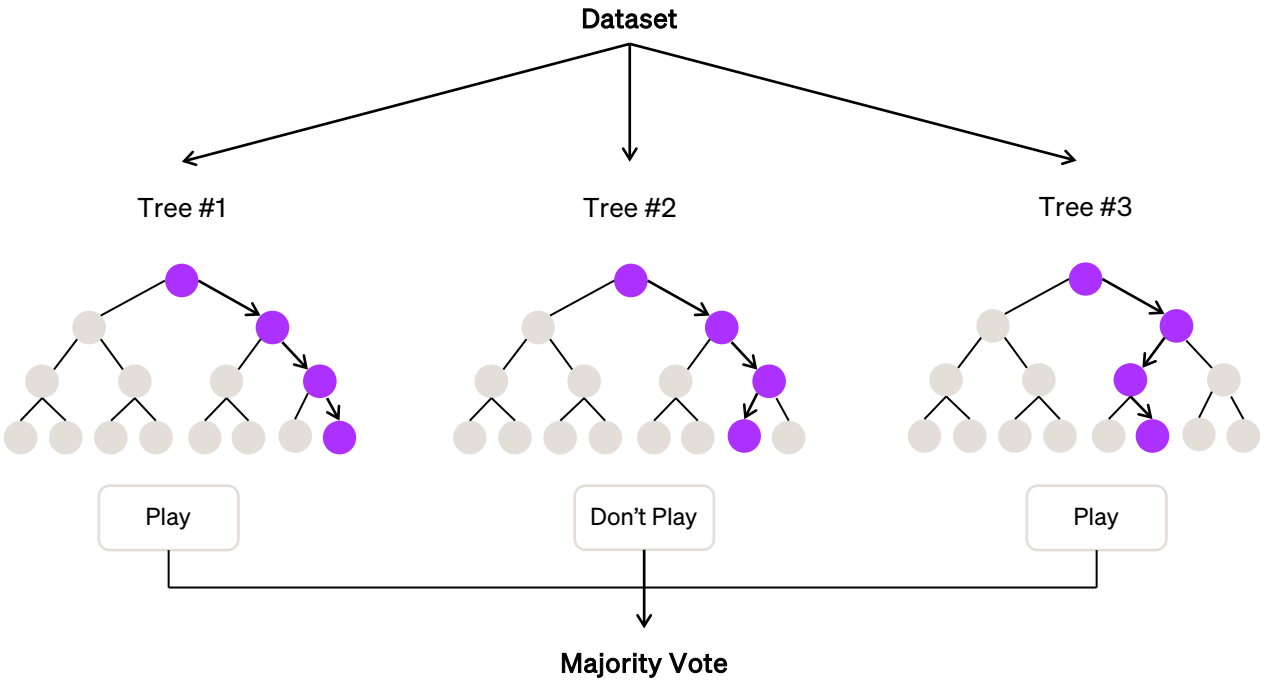
A set of complex data...

Day	Weather	Temp	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

...Translated into a series of rules using a decision tree



Computers can improve on the accuracy of decision trees by using random subsets of the dataset to train multiple decision trees, which can together ‘vote’ on a final prediction



Source: IBM

While these statistical methods are effective, they are very rudimentary and fail to categorize more complex data

To replicate the human brain even more closely, humans have developed more advanced systems of logic called neural networks

Dive Deeper...

Further Reading & Watching

Reading:

- An Introduction to Random Forest Algorithm for Beginners (Analytics Vidhya)
- The Ultimate Guide to K-Means Clustering (Analytics Vidhya)

Watching:

- History of Artificial Intelligence (URBS Lab with Ryan Urbanowicz)
- Logistic Regression (StatQuest)
- Support Vector Machines: All You Need to Know! (Intuitive Machine Learning)
- K-Nearest Neighbors (Intuitive Machine Learning)
- K-Means Clustering (StatQuest)

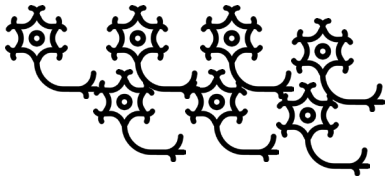
CHAPTER 04

Introduction to Neural Networks

Neural networks are a type of logic system that attempts to replicate the structure of the human brain

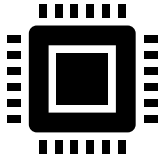
Human brain uses layers of neurons...

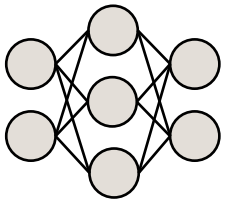




=

Neural network uses layers of nodes...

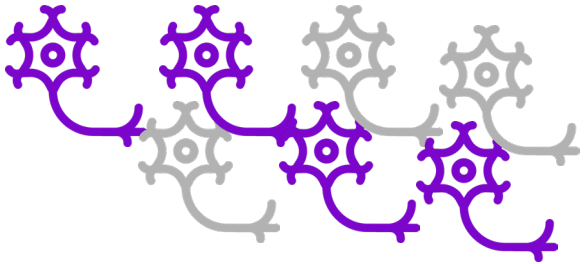




How do neural networks work?

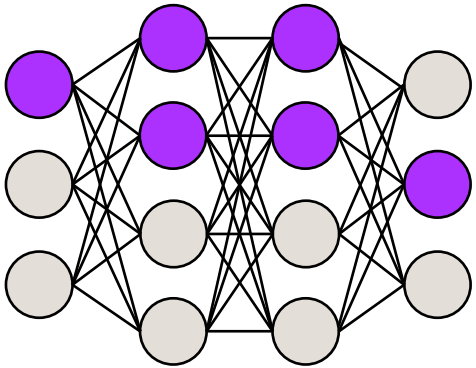
Combinations of nodes that have learned to recognize specific patterns in data become active, like how neurons in the human brain fire together when recalling information

Neurons fire together...



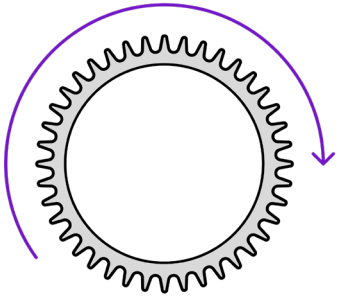
=

Nodes fire together...

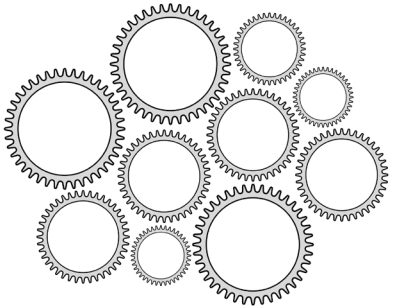


Source: Neuroscience News

Each node is like a gear in a machine — individually, it lacks meaning, but when tuned to work together, the nodes can map complex patterns in data



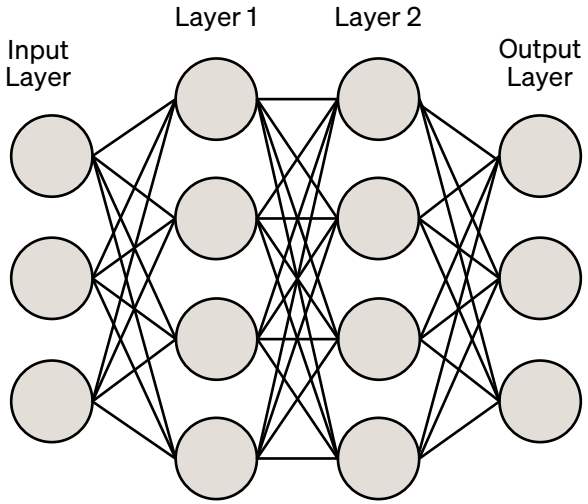
Individual node lacks meaning



Many nodes working together can learn complex information

Neural networks use multiple layers of nodes to process information and make a prediction based on what they have previously learned

Data is fed into a neural network

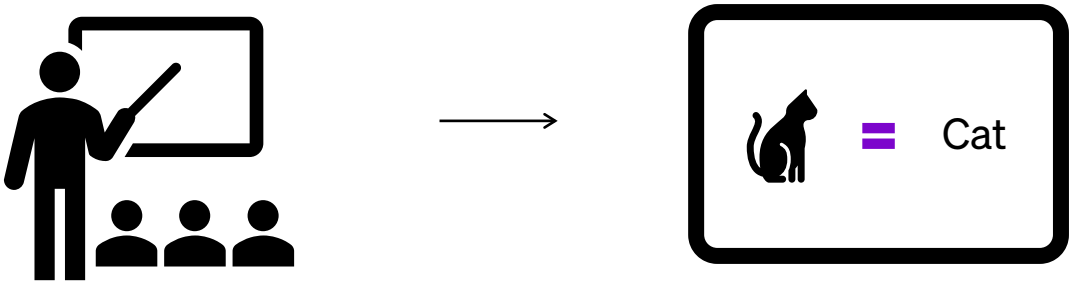


Neural network makes a prediction

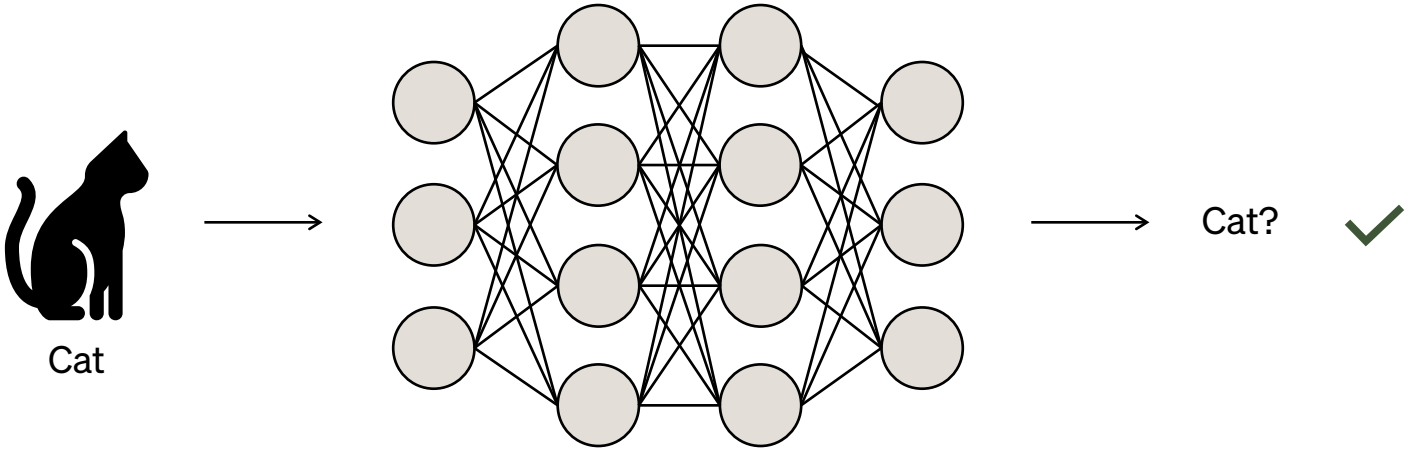
Source: 3Blue, 1Brown

How do neural networks learn?

One way that neural networks can learn is a process called ‘supervised learning’, which is similar to a teacher explaining concepts within a classroom

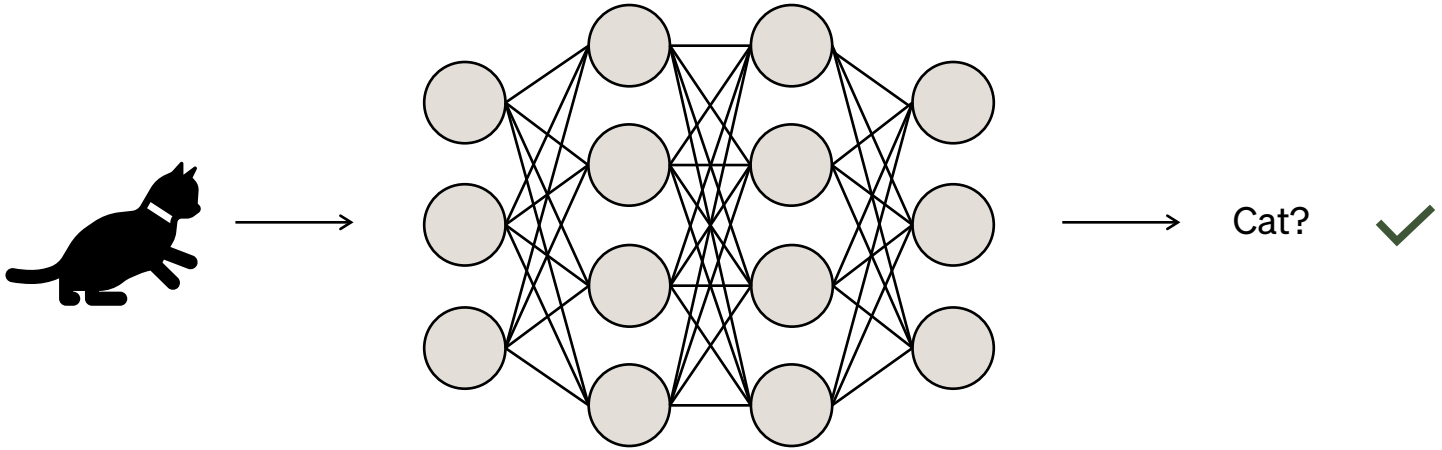


The goal of training a neural network is to build a model that can accurately predict training data...



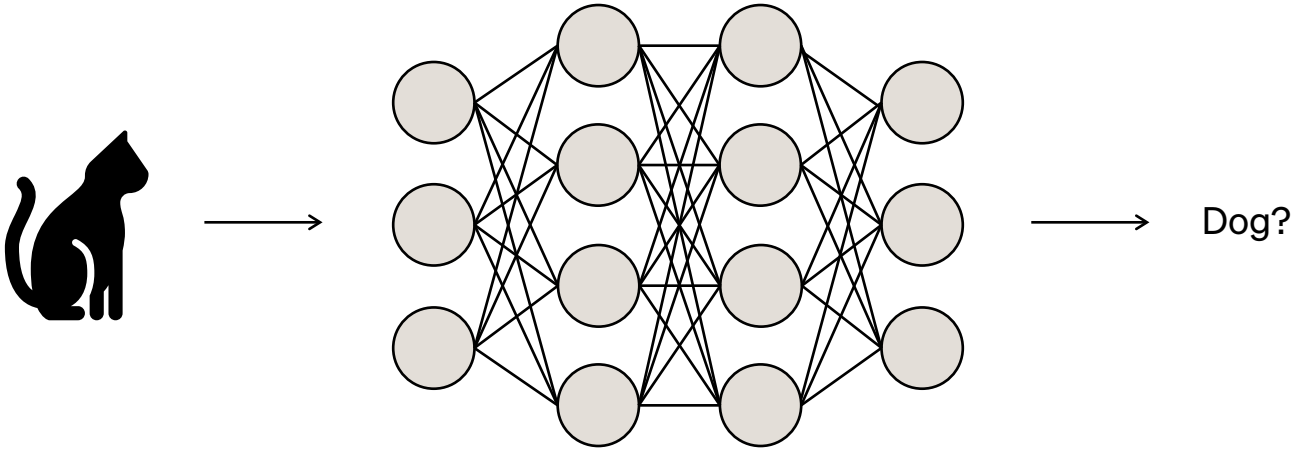
Neural network
makes a prediction

...and eventually generalize to new examples that the model has not yet seen



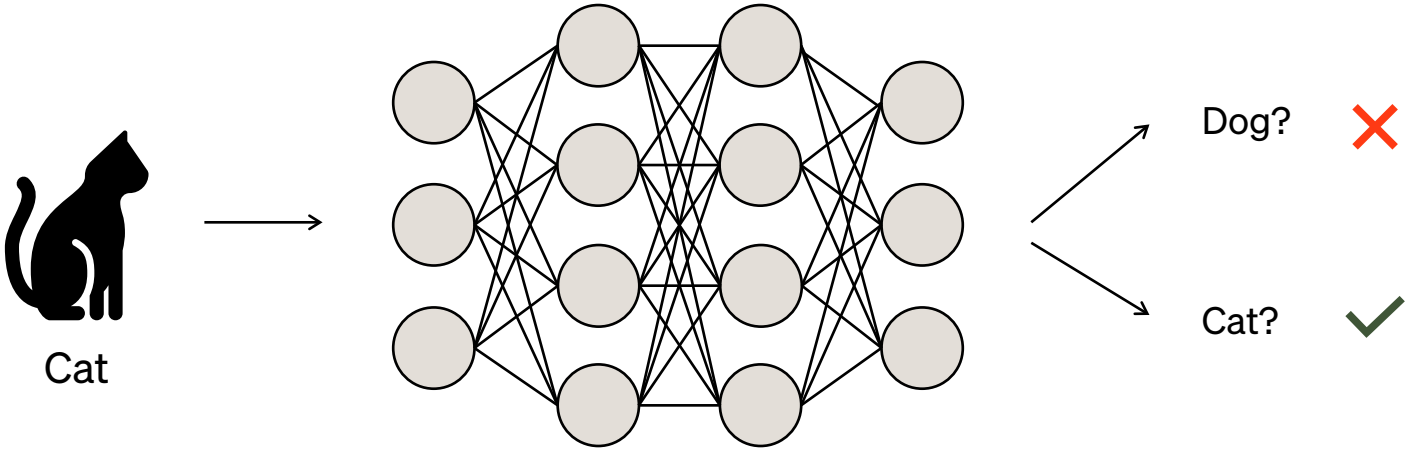
Neural network
makes a prediction

During the training process, a large amount of data is fed into the neural network, which makes a prediction based on what it thinks the data represents

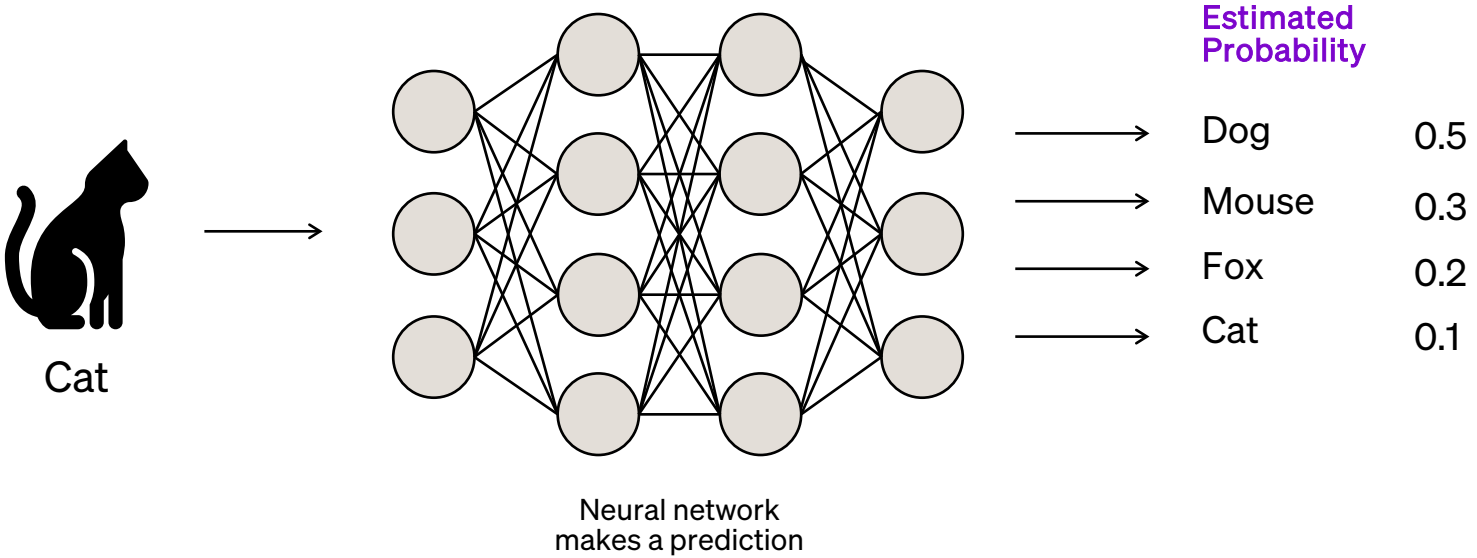


Neural network
makes a prediction

Because this learning is ‘supervised’, the input data is labeled, allowing the model to quickly identify whether its prediction is correct or false



Neural networks represent their predictions as a series of probabilities, which reflect the confidence they have in their prediction



This allows a ‘cost function’ to be assigned as the difference between the model’s predictions and the true value based on the labeled training data



Cat

	Estimated Probability		True Probability		Cost
→	Dog	0.5	- 0.0	=	0.5
→	Mouse	0.3	- 0.0	=	0.3
→	Fox	0.2	- 0.0	=	0.2
→	Cat	0.1	- 1.0	=	(0.9)

If the neural network worked perfectly, this cost would be 0...

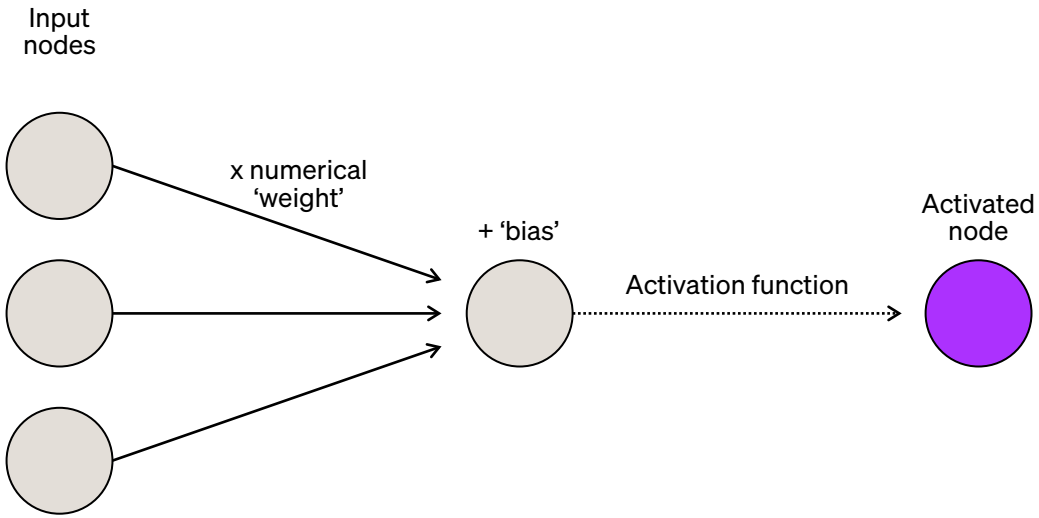


Cat

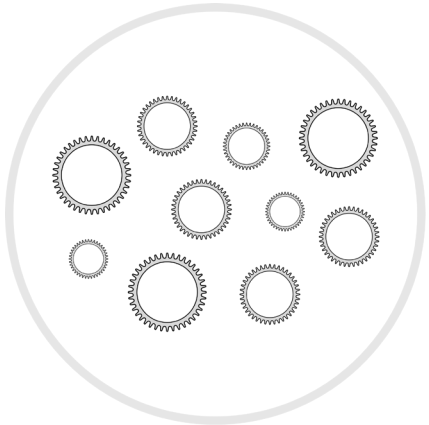
	Estimated Probability		True Probability		Cost
→	Dog	0.0	- 0.0	=	0.0
→	Mouse	0.0	- 0.0	=	0.0
→	Fox	0.0	- 0.0	=	0.0
→	Cat	1.0	- 1.0	=	0.0

How can we minimize this cost function so that the neural network makes accurate predictions?

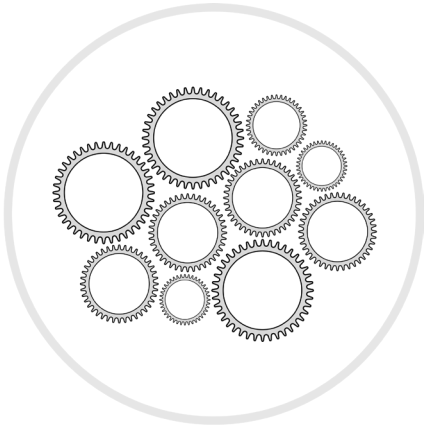
Neural networks use mathematical ‘weights’ and ‘biases’ to determine how important each node is within the network



We can change the weights and biases of each node to change how the model interprets data, with the hope of reducing the cost function

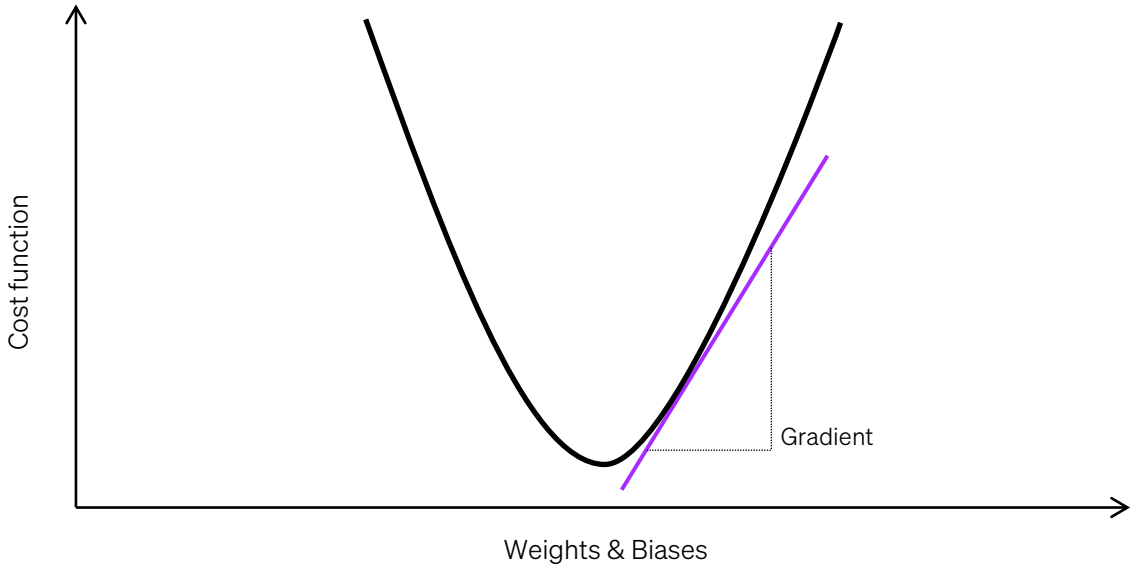


Random weights
and biases

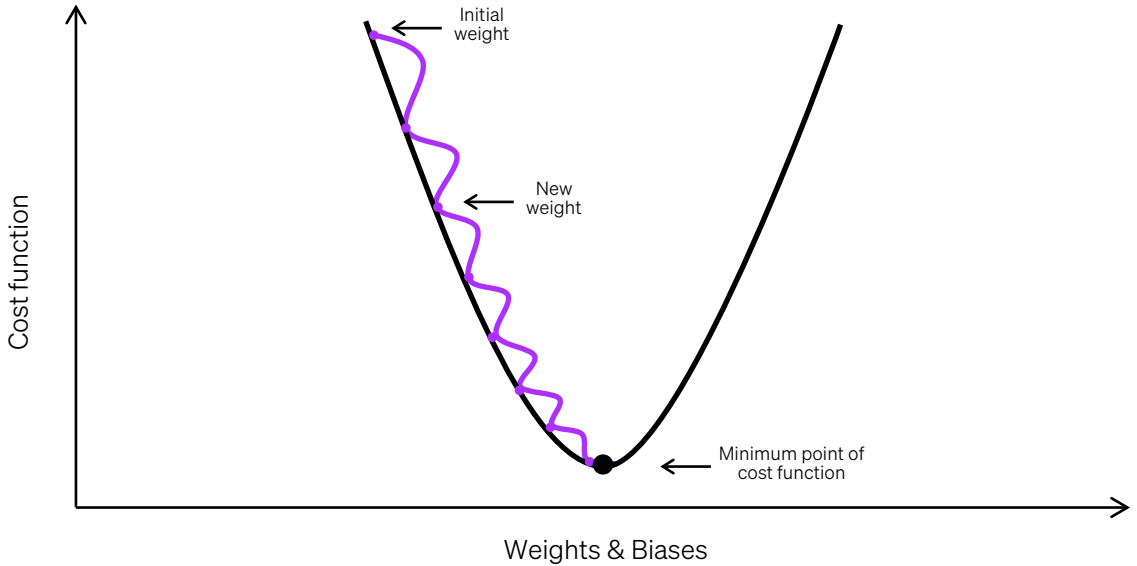


Trained
neural network

This process is called ‘backpropagation’, which calculates how the cost function changes with respect to changes in a model’s weights and biases

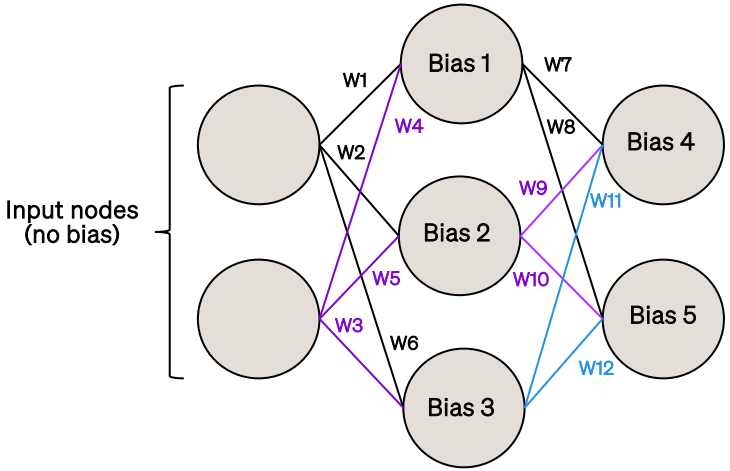


Then, each of the model's weights and biases are iteratively updated to minimize this cost function through a process called 'gradient descent'



Source: 3Blue, 1Brown

The total number of weights and biases that a model uses to make a prediction is called its ‘parameters’

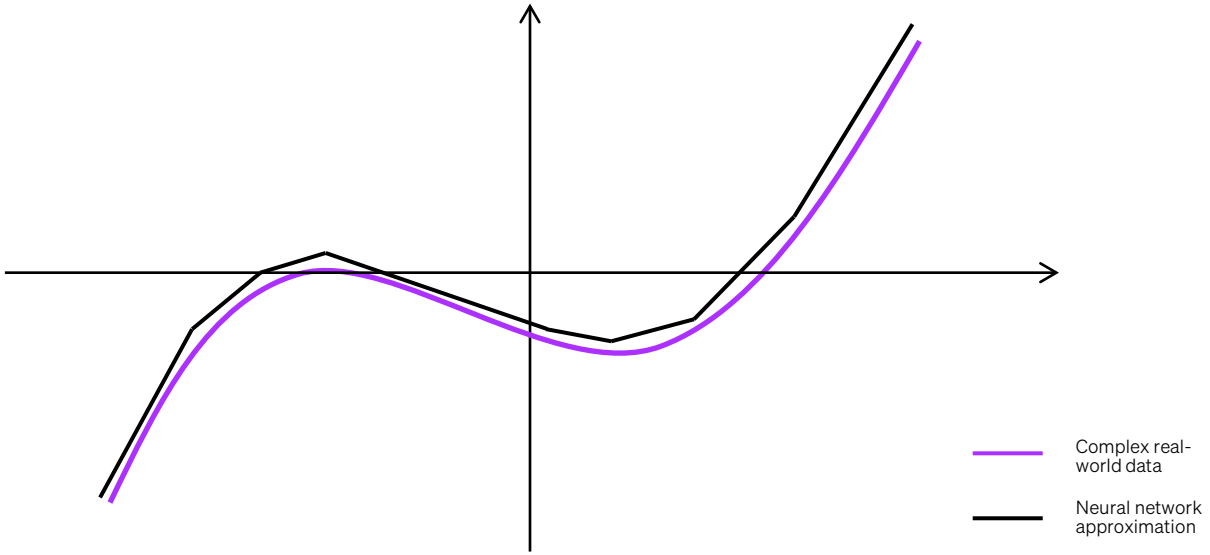


12 Weights

+ 5 Biases

= 17 Parameters

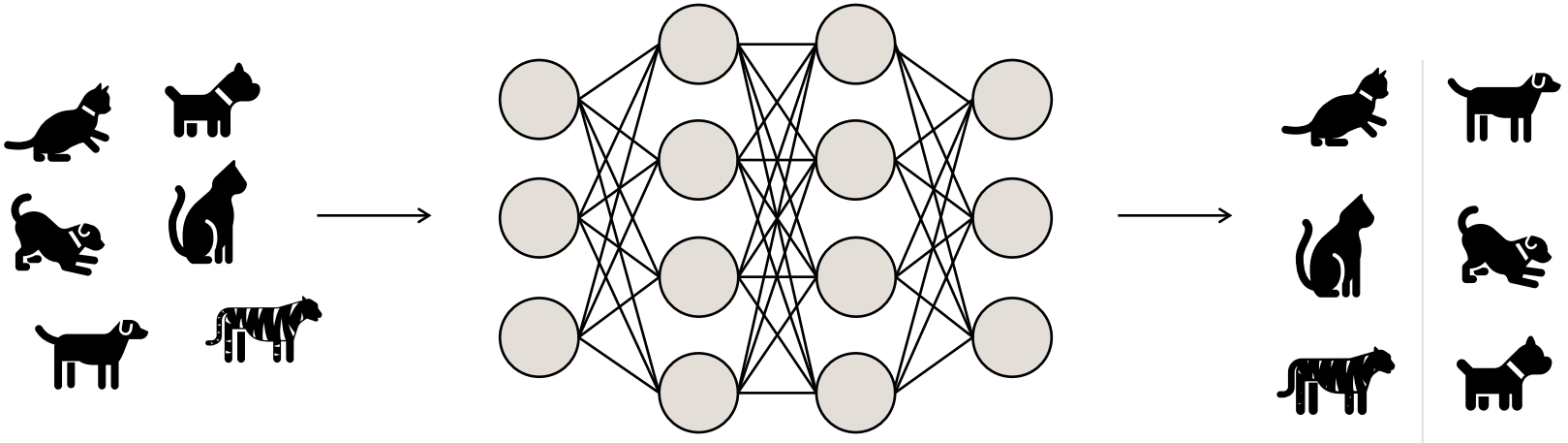
When a neural network is trained, the outputs of each node stack together to effectively represent complex patterns in real-world data



Source: Emergent Garden

Neural networks can also learn by themselves from unstructured data using a process called ‘unsupervised learning’

Neural networks can extract features from unlabeled data, and use statistical techniques like k-means clustering to discover patterns within this data



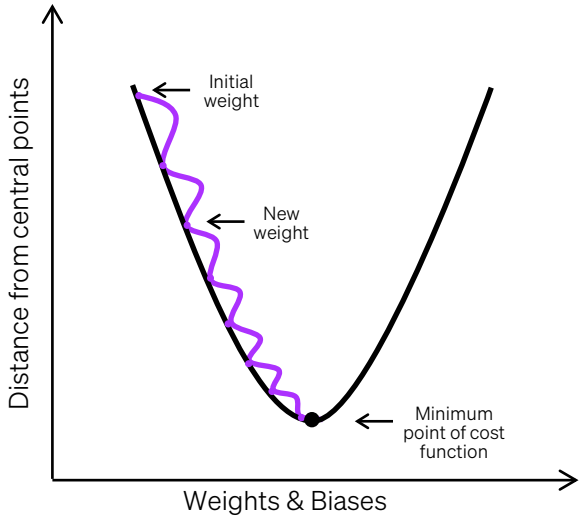
Unstructured, unlabeled data

Neural network detects patterns

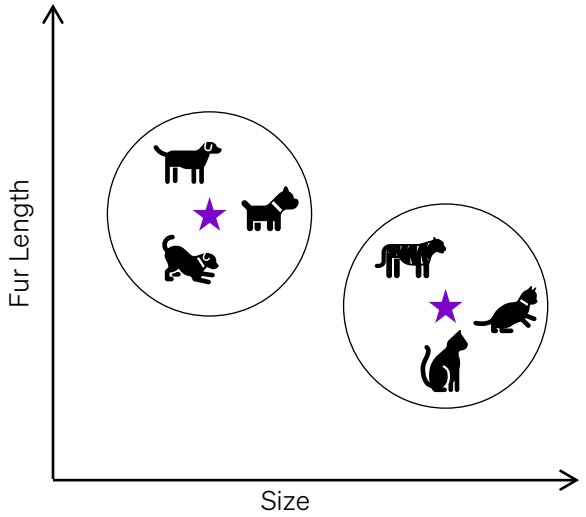
Grouped data

Each of the neural network's weights and biases are updated to minimize the distance from the central point within the k-means classifier

Weights and biases updated…

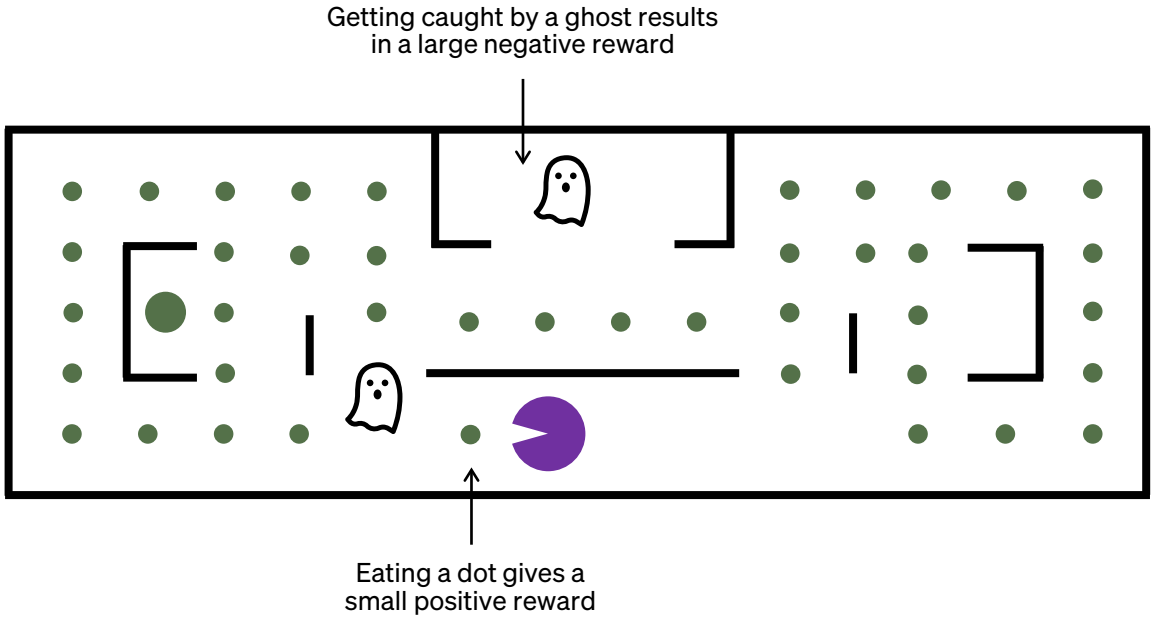


…to minimize distance from central points

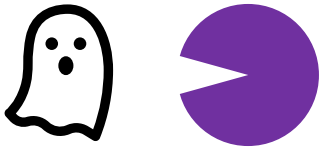


‘Reinforcement learning,’ is another method of training which uses a system of rewards and punishments to train a neural network

Training is an iterative process where the model attempts to maximize its positive rewards while minimizing its negative rewards



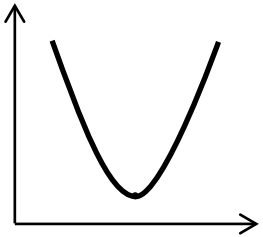
A cost function is assigned as the difference between the model's prediction for the reward of a given action, and the actual reward...



Model believes that eating a ghost will lead to a positive reward

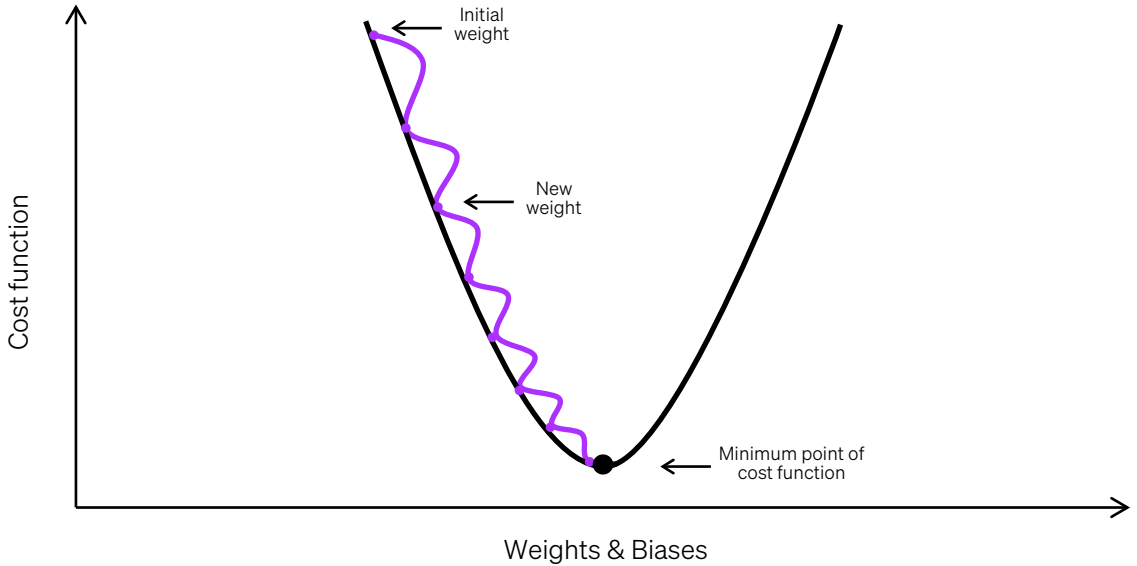


Getting caught by ghost actually results in a large negative reward



Cost function is assigned as difference between expected and actual result

And again, each of the model's weights and biases are iteratively updated to minimize this cost function until the model arrives at an equilibrium that results in the desired behavior



Source: 3Blue, 1Brown

To effectively replicate human intelligence, however, neural networks must also be able to understand and generate images, text, audio and other types of complex data

Dive Deeper...

Further Reading & Watching

Reading:

- What Are Neural Networks? (IBM)
- What the Hell is a Perceptron? (Towards Data Science)
- Supervised vs Unsupervised Learning: What's the Difference? (IBM)
- Reinforcement Learning 101 (Towards Data Science)

Watching:

- But What is a Neural Network? (3Blue, 1Brown)
- Gradient Descent: How Neural Networks Learn (3Blue, 1Brown)
- What is Backpropagation Really Doing? (3Blue, 1Brown)
- Why Neural Networks Can Learn Almost Anything (Emergent Garden)

CHAPTER 05

Types of Neural Networks

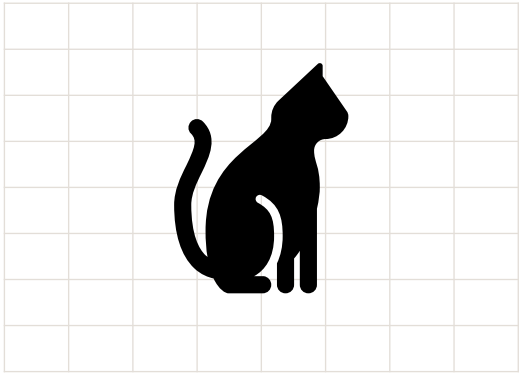
How can neural networks understand images?

Convolutional neural networks (CNNs) were introduced in the 1990s to aid with complex image processing and classification tasks

How do CNNs work?

CNNs first recognize images as a series of numbers associated with the amount of red, green and blue within each pixel of a given image...

What we see

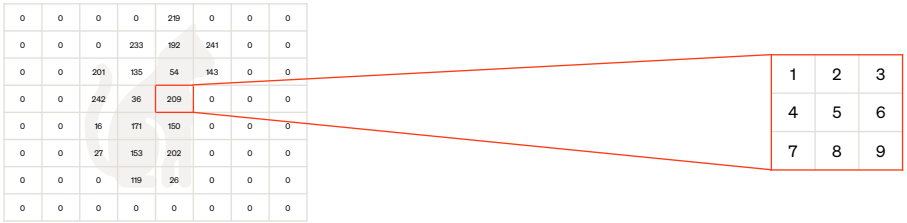


What CNNs see

0	0	0	0	219	0	0	0
0	0	0	233	192	241	0	0
0	0	201	135	54	143	0	0
0	0	242	36	209	0	0	0
0	0	16	171	150	0	0	0
0	0	27	153	202	0	0	0
0	0	0	119	26	0	0	0
0	0	0	0	0	0	0	0

...and then slide a filter over each pixel of the image using a mathematical function called a 'convolution'

Image x Filter

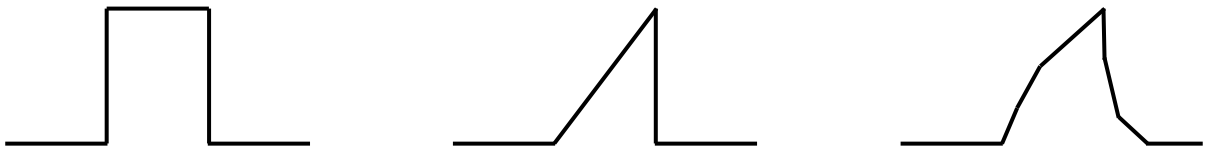


Input Image

Matrix of weights that slides over input image

A convolution is a mathematical function that can combine two functions together to create a third function

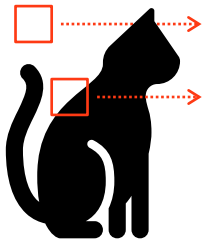
A x B = C



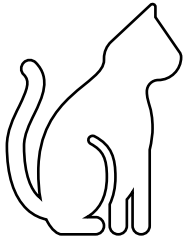
Function A | Function B | Combination

The filters allow CNNs to extract specific features such as edges and colors to build a representation of the image

Filters scan an image...

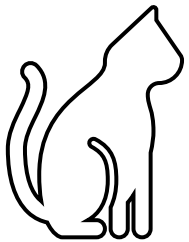


...and map features like edges

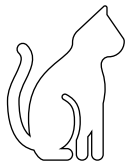


The CNN summarizes these features in a ‘pooling layer’ to create a down-sized representation of the feature map for more efficient processing

Original feature map



Summarized feature map in the pooling layer



Then, a traditional neural network learns to form combinations of these features across multiple layers until they sum to the final images

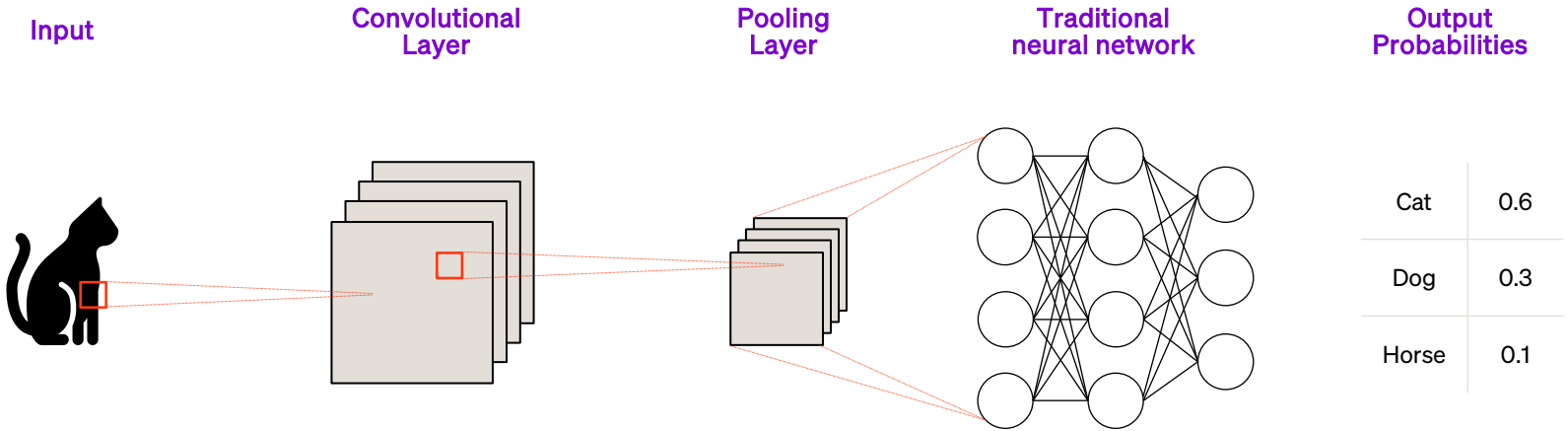
Edges

More complex features

Images



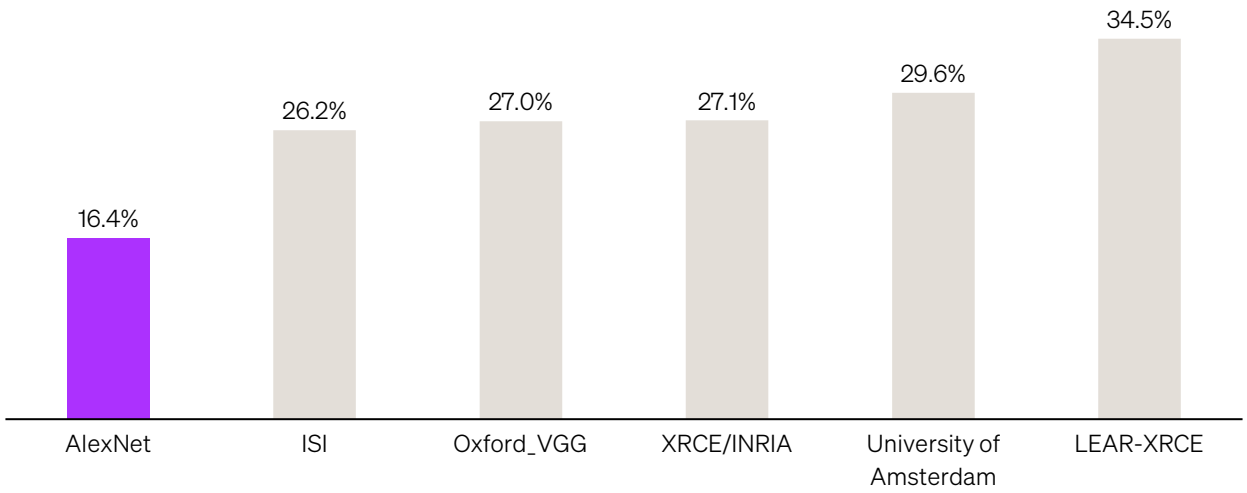
This allows the network to predict what a particular image represents, and express this prediction as a series of probabilities



More convolutional and pooling layers allow CNNs to map more complex and intricate features, which improves their ability to effectively classify images

A key breakthrough occurred in 2012 when ‘AlexNet’ used an 8-layered structure to classify 1.3mm images into more than 1,000 different classes in the ‘ImageNet’ challenge

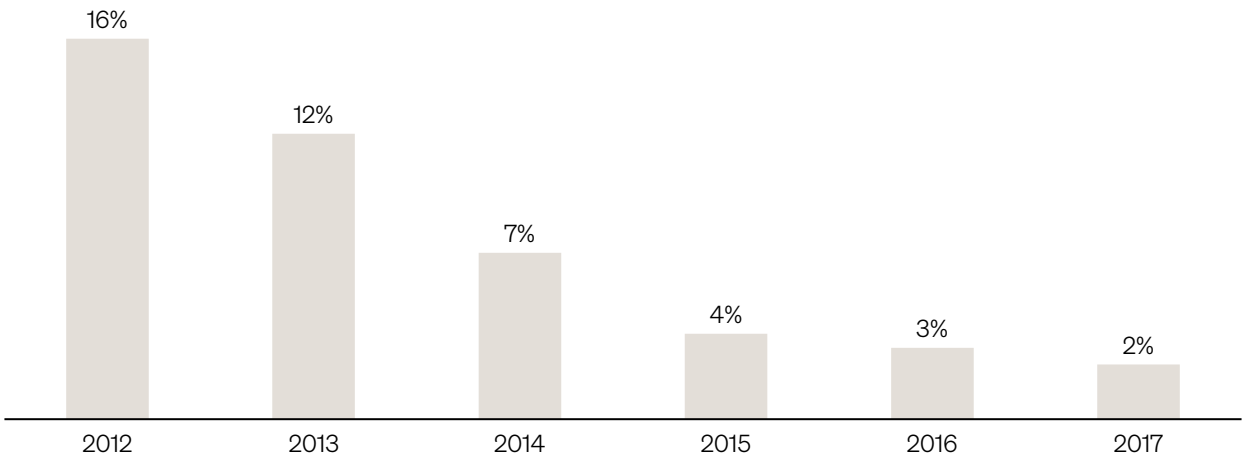
AlexNet became the first CNN to score a sub-25% top-5 error rate (whether the correct answer is within a model’s top 5 guesses)



Source: Pinecone

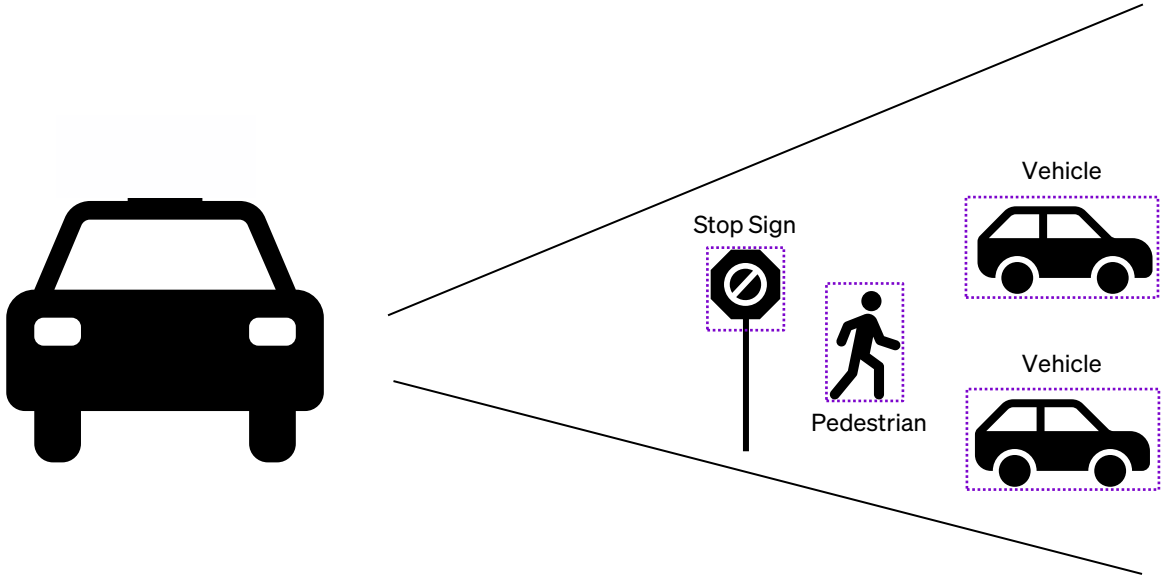
Since then, CNNs have dramatically improved in their ability to recognize and classify images

ImageNet Winner Top-5 Classification Error Rate



Source: ILSVRC

Enabling widespread commercial applications, including self-driving cars, which use CNNs to identify objects such as pedestrians and other vehicles

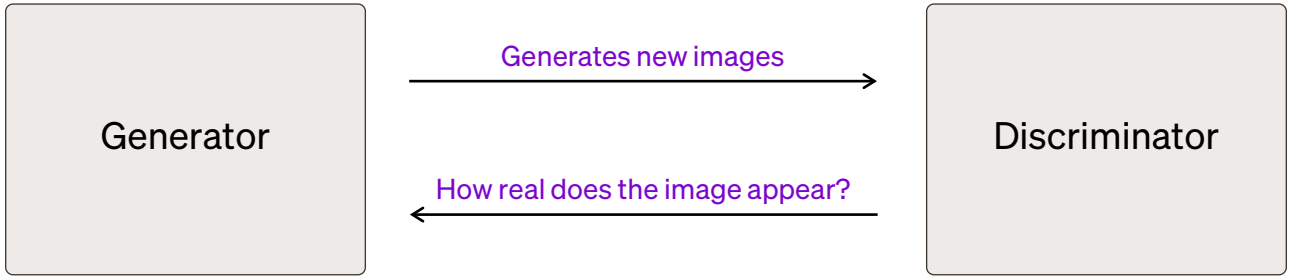


Source: Augmented Startups

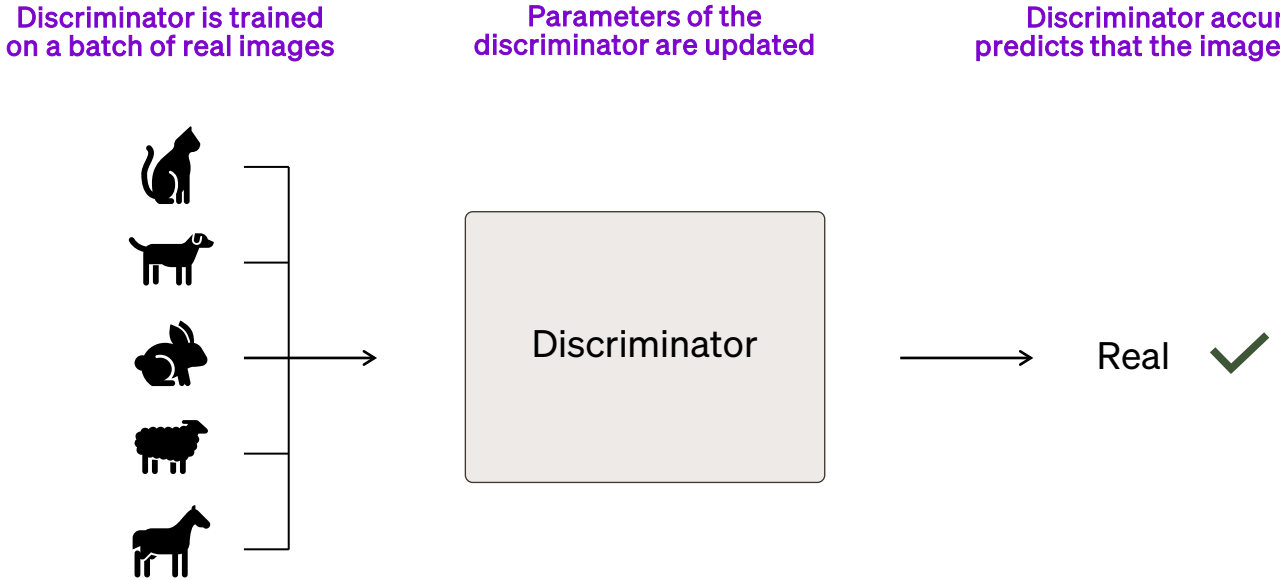
How can neural networks generate new images?

Generative adversarial networks (GANs) use two CNNs, a generator and a discriminator, to generate new images which are indistinguishable from real photographs

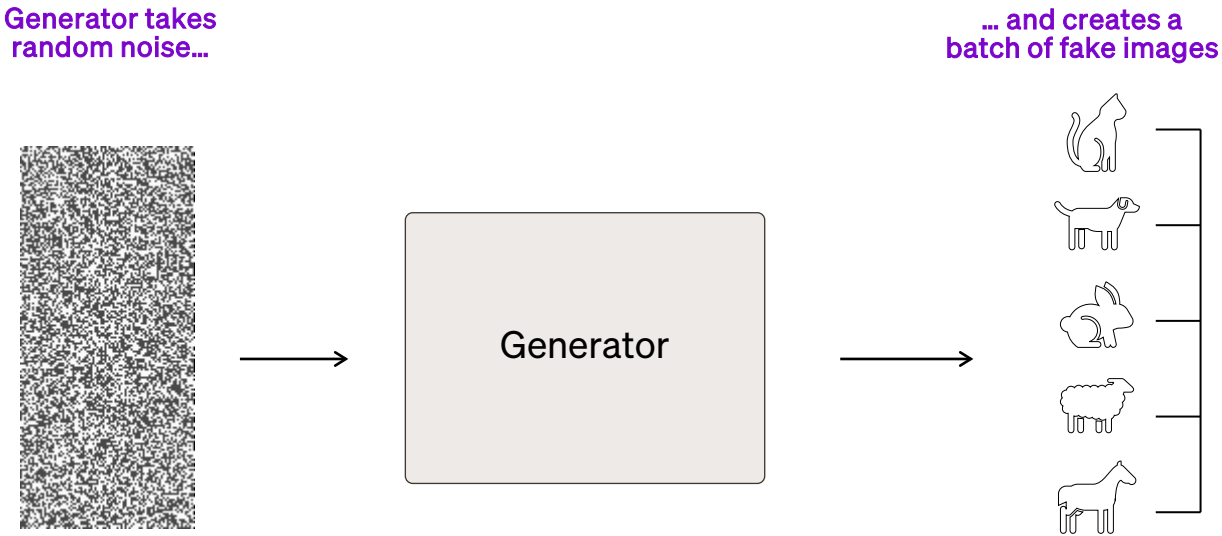
The generator and discriminator work together to first generate new images, and then discriminate whether they are real or not in an iterative process to improve the generator



The discriminator is trained on a batch of real images, and its parameters are updated until the model can accurately predict that the images are real



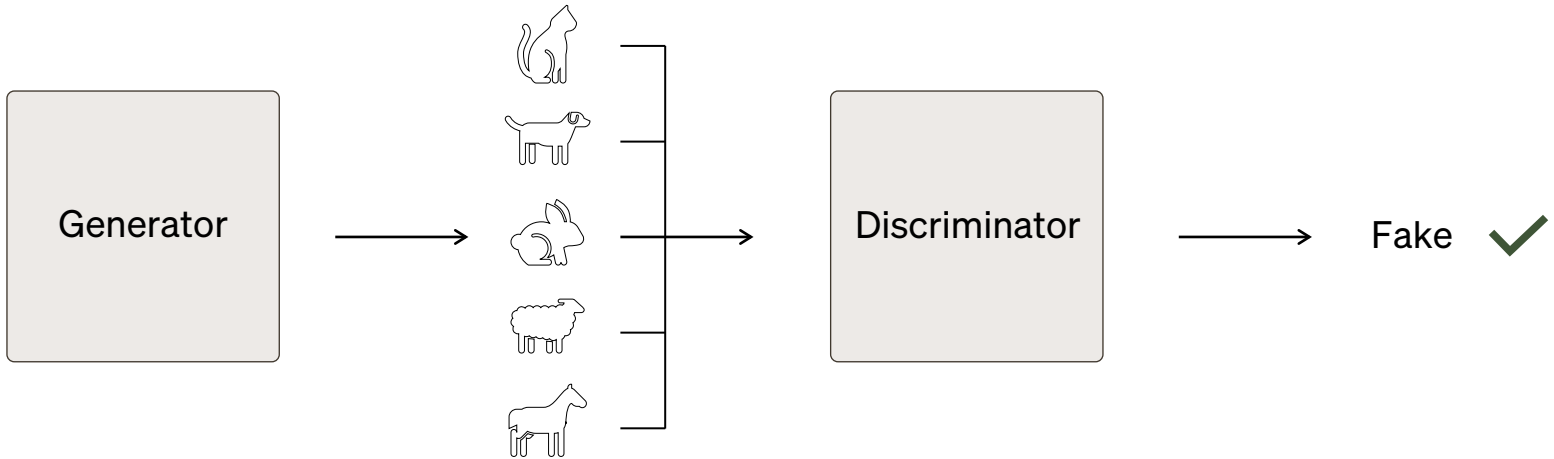
Then, the generator creates a batch of fake images from random noise



The discriminator decides whether these images are real or fake

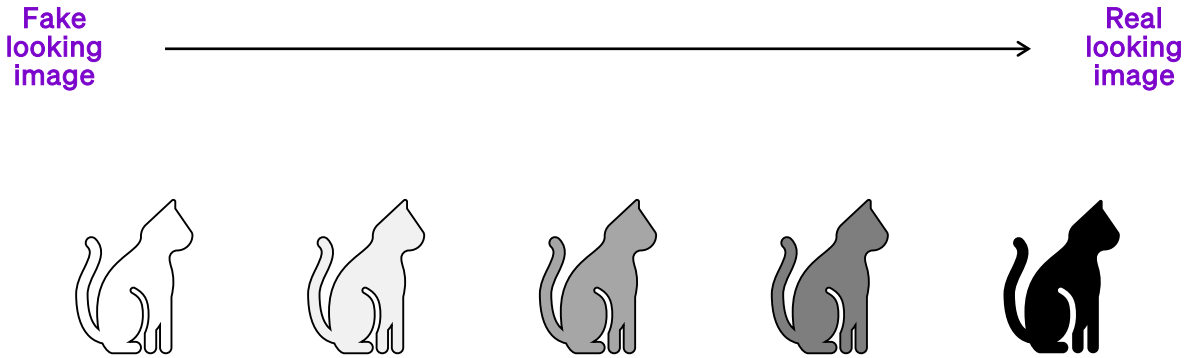
Generator produces a batch of fake images

Discriminator accurately predicts that the images are fake



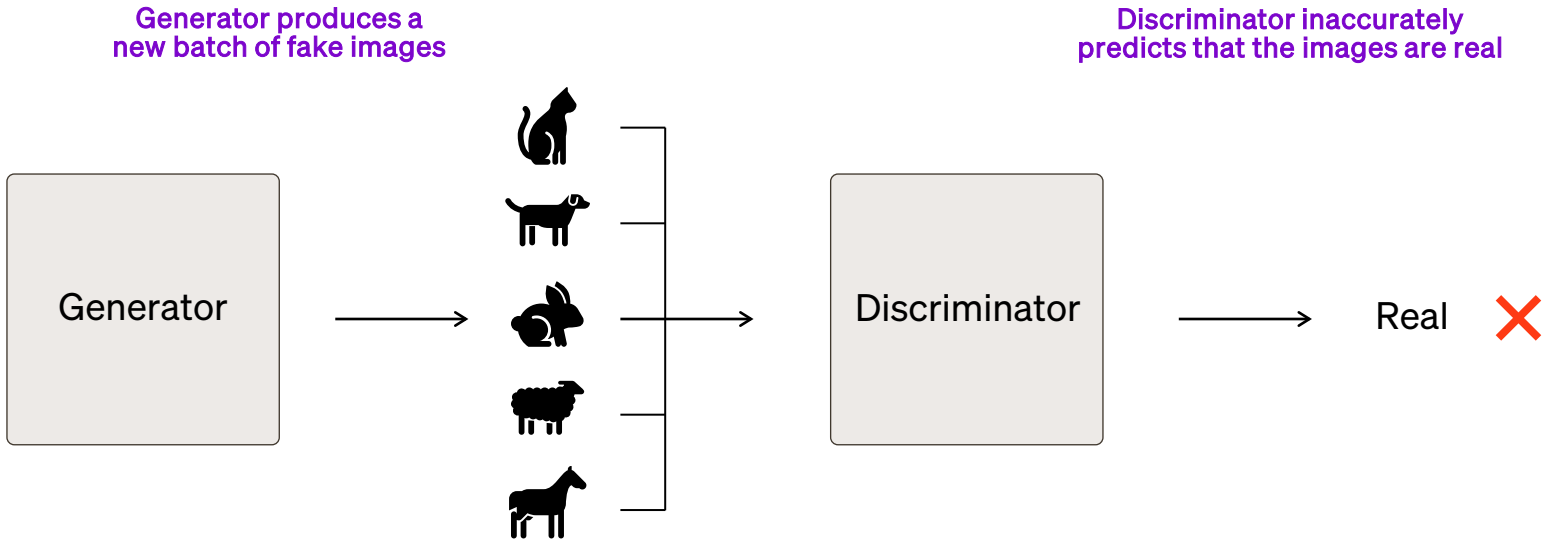
Source: Google For Developers

And the generator's parameters are updated to generate new images that can increasingly fool the discriminator



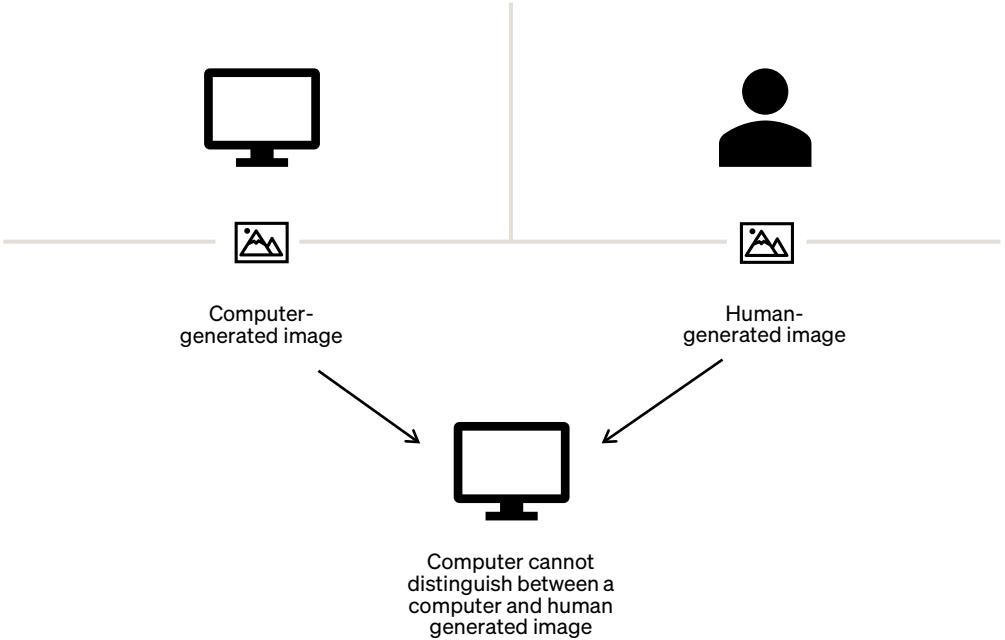
Source: Google For Developers

The process ends when the generator produces new images that can fool the discriminator into thinking they are real



Source: Google For Developers

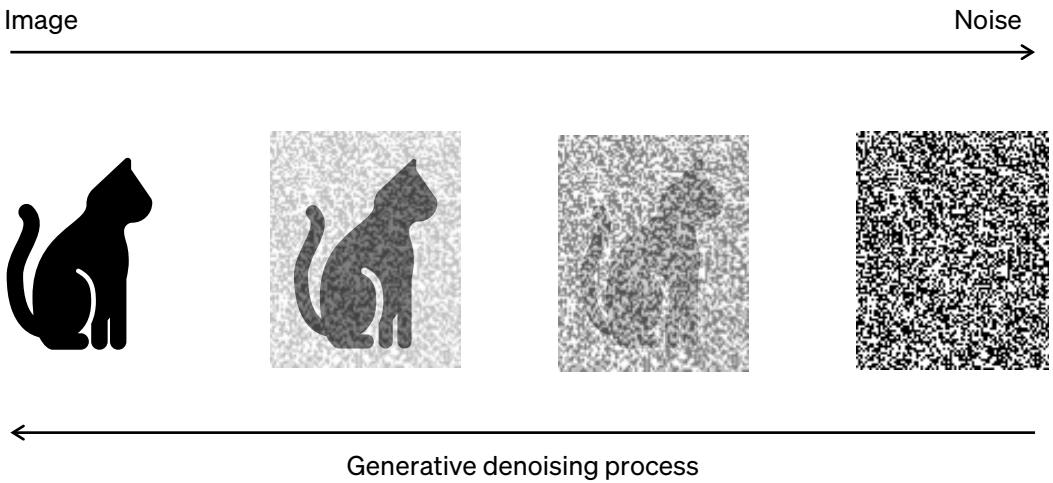
This is like a Turing Test for a computer



Source: Saygin (2000), Turing Test 50 Years Later

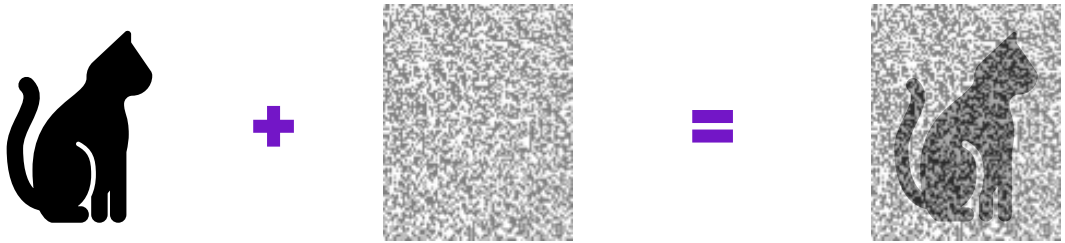
While GANs are very effective at generating realistic images, they can be difficult and unstable to train

Diffusion models are a new class of image generation tool that work by gradually adding noise to an image, and then learning to reverse this process to generate new images

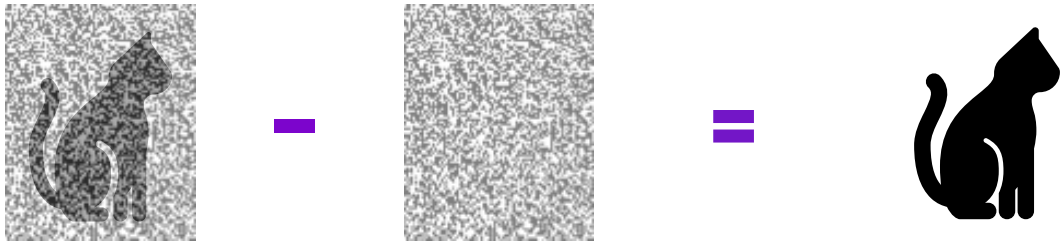


Source: Assembly AI

During the training process, a neural network learns to predict the noise that was added to the image in each step of the diffusion process...



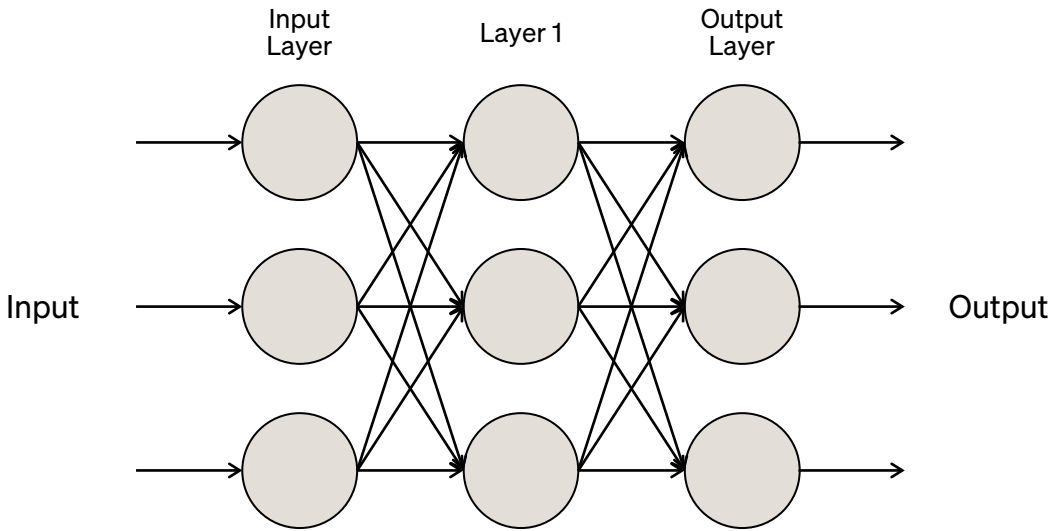
... so that it can learn to reverse this process to generate new images from random noise



Diffusion models are more computationally intensive than GANs, but often yield more stable and realistic outputs

How can neural networks understand text?

Traditional neural network processes information in one direction from input to output...



Source: Code Emporium

Which make them ineffective for sequential tasks like understanding text, where remembering the order and meaning of the previous words is critical



“The brown dog had a bone”

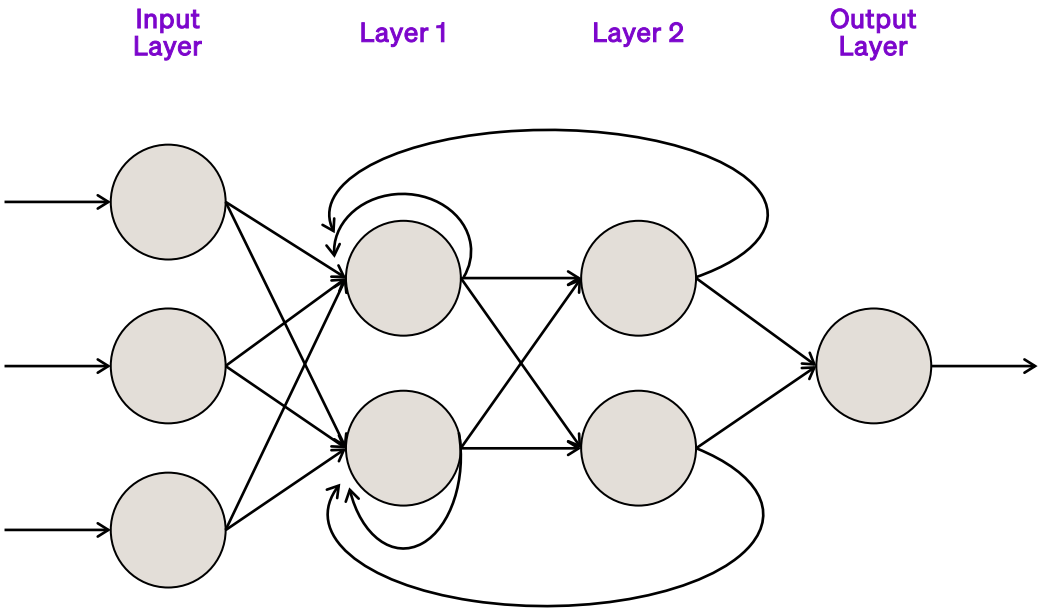
Brown

Dog

Bone

How do we build a neural network
that can remember previous information?

Recurrent neural networks (RNNs) can cycle information through loops, allowing them to recall previously processed data in a sequence

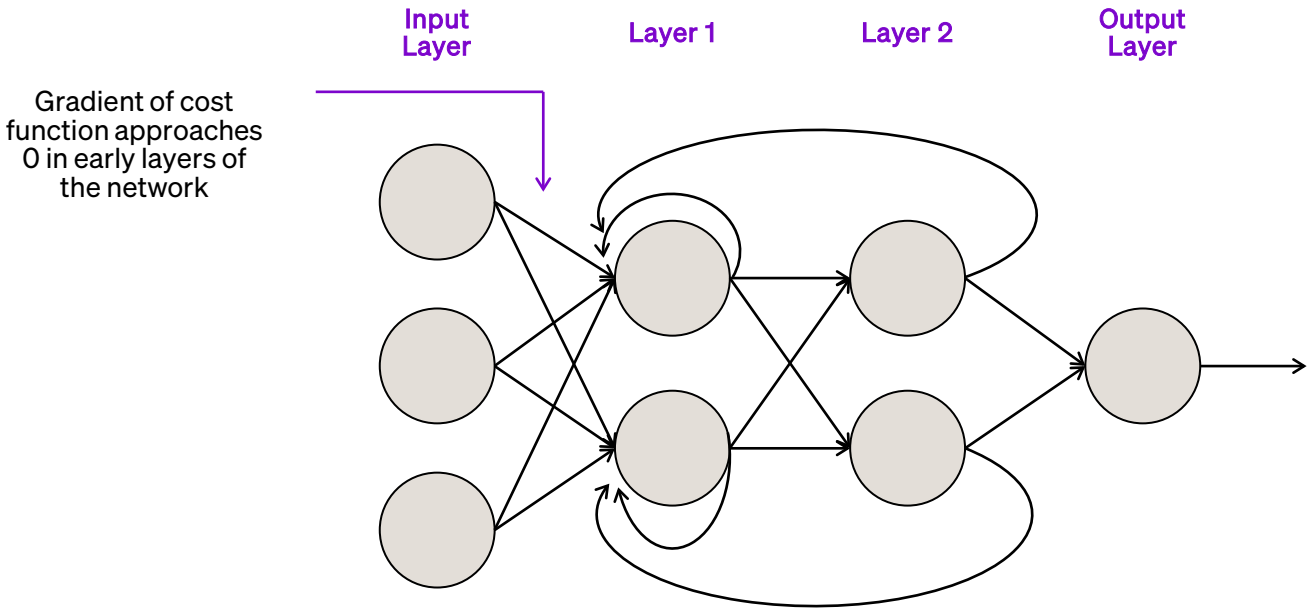


Source: Code Emporium

But RNNs have no filter for deciding which information is relevant and which can be forgotten, resulting in models that attempt to remember very long strings of information

Input Sentence		Relevant Information		Irrelevant Information
“James, who was wearing a Loro Piana sweater, forgot his umbrella and got soaked in the rain”	=	Loro Piana sweater Forgot umbrella Soaked in the rain	+	James

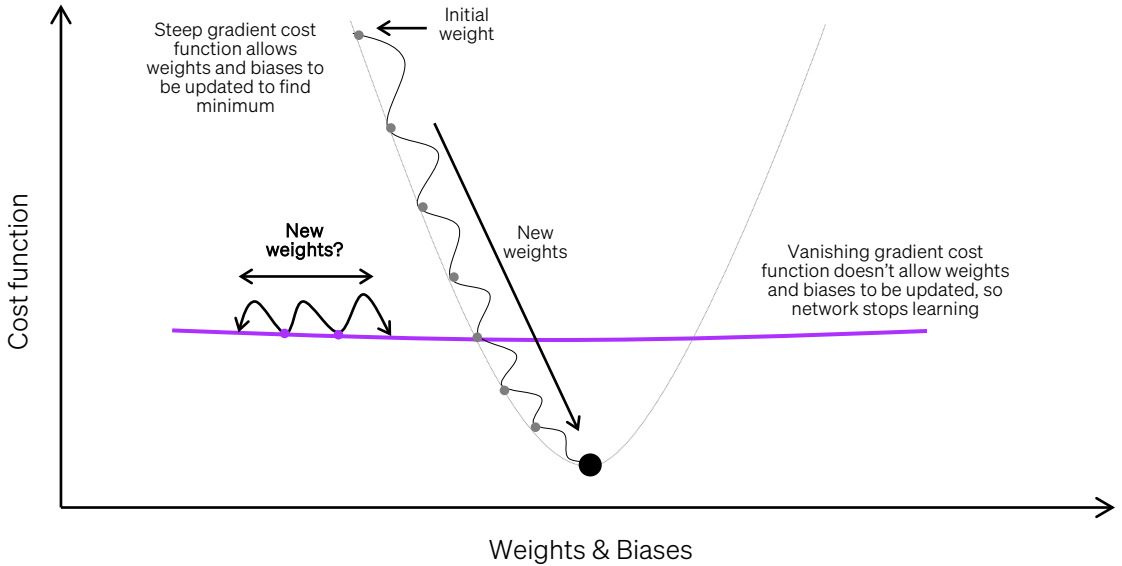
This results in the ‘vanishing gradient’ problem, where the gradient of the cost function becomes miniscule in early layers of the network as it attempts to train on long sequences



Source: Bot Penguin

Why is this a problem?

As the gradient of the cost function approaches zero, the model's weights and biases in earlier layers of the network aren't updated, so the network stops learning



Source: Bot Penguin

Variations of RNNs like long-short-term memory networks (LSTMs) were developed with built-in memory cells and ‘gates’ to filter and remember longer strings of key information

Forget Gate	Input Gate	Output Gate
Decide which of the previous information should be forgotten	Decide which of the previous information should be stored in the memory cell	Pass updated information from the current word to the next word

Source: Analytics Vidhya

This allows LSTMs to process language much more efficiently than traditional RNNs, helping to solve the vanishing gradient problem

"James, who was wearing a Loro Piana sweater, forgot his umbrella and got soaked in the rain"



Forget Gate	Input Gate	Output Gate
LSTM ignores "James" since this is irrelevant information	LSTM decides that the phrase "forgot his umbrella" and the "Loro Piana sweater" are important context and takes note	This phrase "got soaked in the rain" is processed, noting that this is due to the lack of an umbrella, resulting in a wet Loro Piana sweater

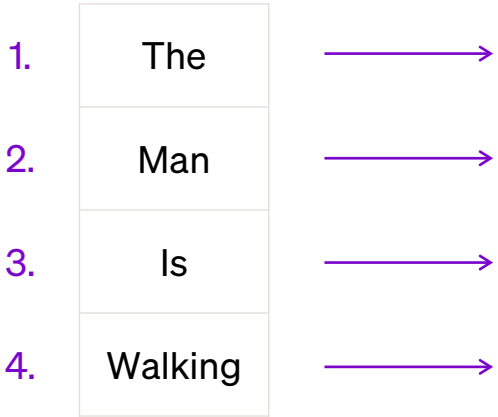
Source: Analytics Vidhya

Why aren’t LSTMs used anymore?

LSTMs require input words to be passed sequentially to capture their position in the context of other words...



This is inefficient to train, and fails to take advantage of GPUs and other specialized processors which excel at parallel processing



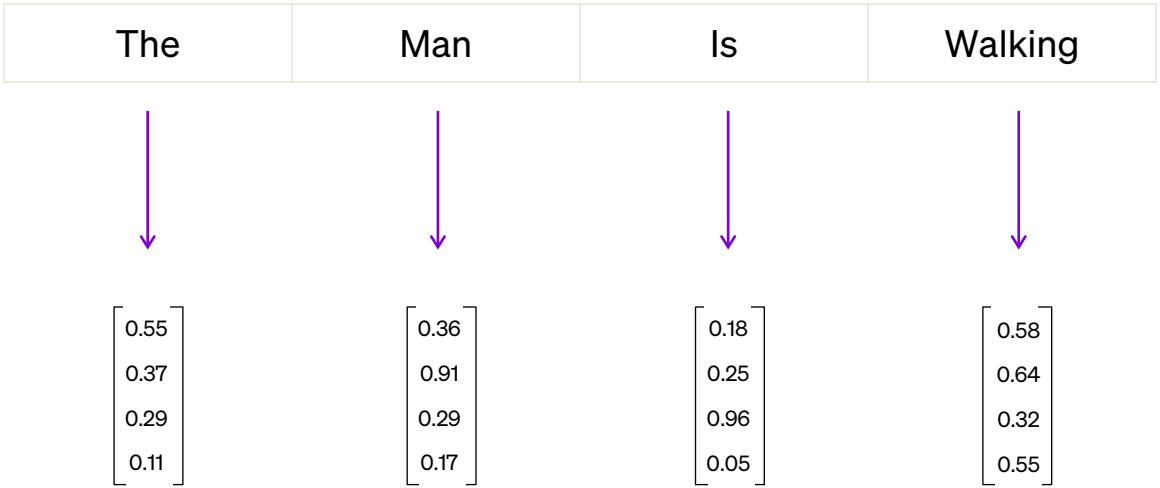
Source: Code Emporium

How do we build a more efficient model that can understand and generate human text?

The transformer architecture is the latest state-of-the-art iteration of natural language processing models

How do transformer models work?

Transformer models first represent words as vectors, which represent the meaning of each word numerically...



Source: Google Cloud Tech

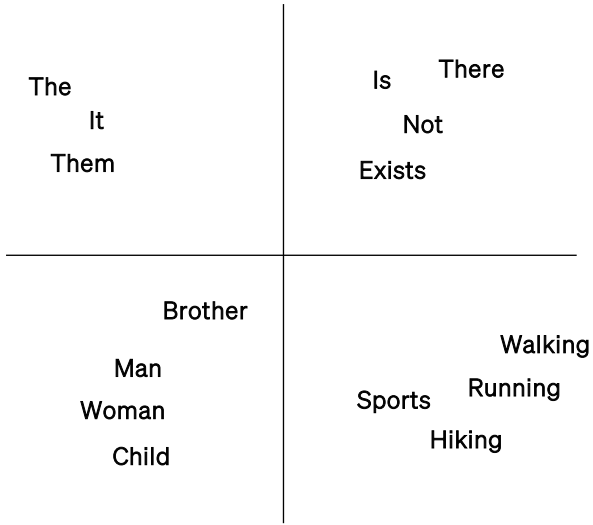
...and map their relationship with other words

Vector representation...

The	Man	Is	Walking
0.55	0.36	0.18	0.58
0.37	0.91	0.25	0.64
0.29	0.29	0.96	0.32
0.11	0.17	0.05	0.55

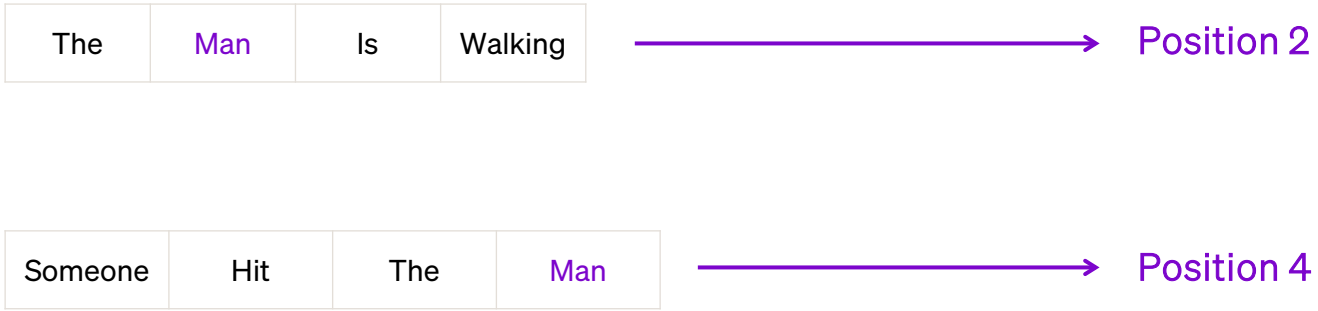
=

...maps relationship with other words



Source: Google Cloud Tech

The position of each word is also encoded as a vector, allowing transformers to capture context without needing to process each word serially like LSTMs



Source: Google Cloud Tech

Transformer models are unique because they use ‘self attention’, which allows the model to focus on the context around relevant words to better understand their meaning

The word ‘money’ allows the transformer to understand which type of bank is mentioned

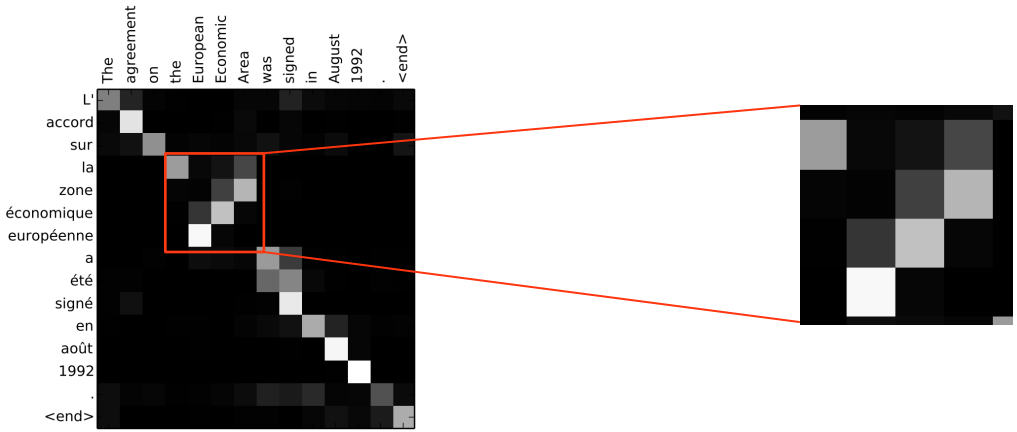
“Yesterday, I went to the bank to deposit money”



This is particularly useful for tasks like language translation, which require the context surrounding words to accurately translate between languages

Attention heatmap for language translation

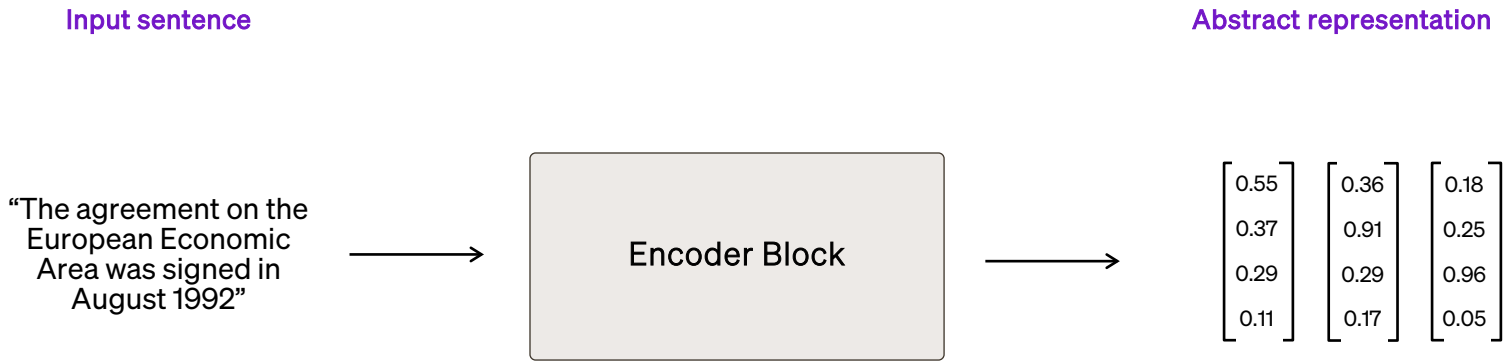
Model uses surrounding context to translate words



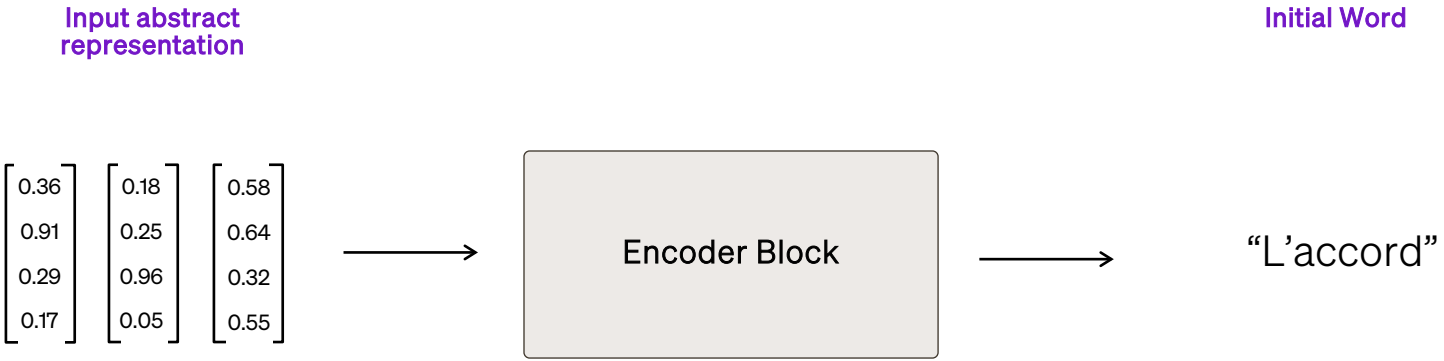
Source: Bahdanau et al. (2016), Neural Machine Translation by Jointly Learning to Align and Translate

The original transformer model was designed for language translation using two blocks - an encoder and a decoder

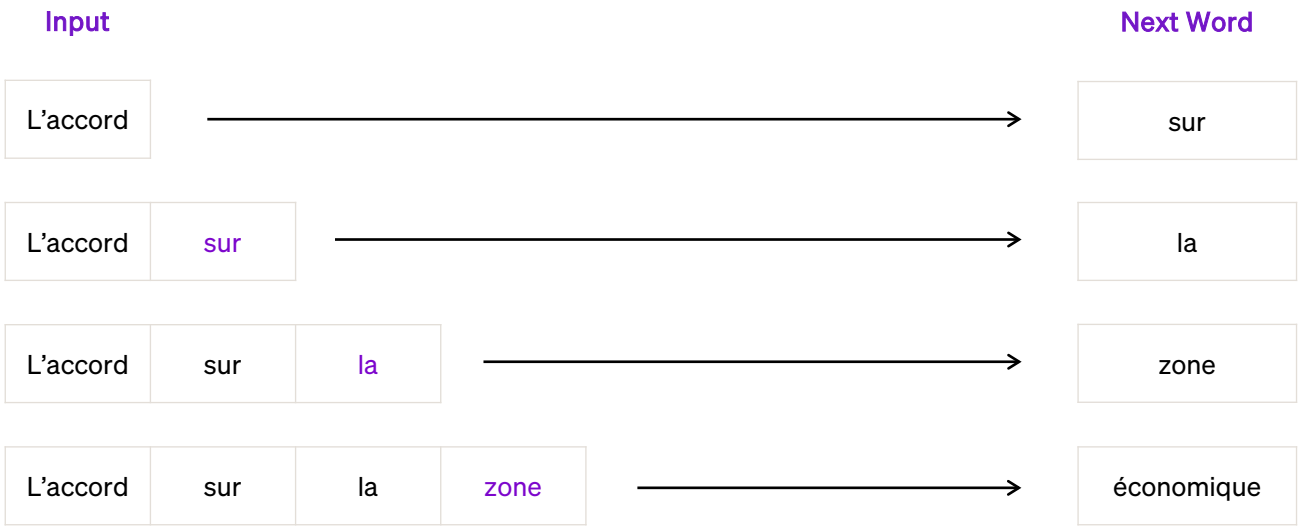
The encoder block takes an input sentence and ‘encodes’ it into an abstract mathematical representation that captures the meaning of the sentence



This becomes an input for the decoder block, which uses the abstract representation to prompt an initial word which includes the context of the surrounding words



The decoder then uses this initial word as an input to generate the next most probable word, which becomes an input for the next in an autoregressive manner



Eventually, AI researchers realized that encoders and decoders can be separated to perform different tasks

Models like OpenAI's GPT and Google's BERT exclusively use the decoder or encoder block, making each model better suited for different tasks

OpenAI's GPT	Google's BERT
Decoder only	Encoder only
Understands context to the left of a word	Understands context In both directions
Better for generating text	Better for understanding text

Source: Hugging Face

Dive Deeper...

Further Reading & Watching

Reading:

- Convolutional Neural Networks, Explained (Towards Data Science)
- A Gentle Introduction to Generative Adversarial Networks (Machine Learning Mastery)
- Introduction to Diffusion Models for Machine Learning (Assembly AI)

Watching:

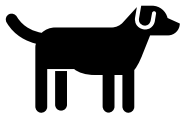
- Convolutional Neural Networks Explained (Futurology)
- Transformer Neural Networks - EXPLAINED! (Code Emporium)
- Transformers, Explained: Understand the Model Behind GPT, BERT, and T5 (Google Cloud Tech)
- Transformer models: Encoder-Decoders (Hugging Face)
- Transformer models: Encoders (Hugging Face)
- Transformer models: Decoders (Hugging Face)

CHAPTER 06

Introduction to Large Language Models

There are two types of AI models...

Discriminative models, which are used to classify data into two or more categories...

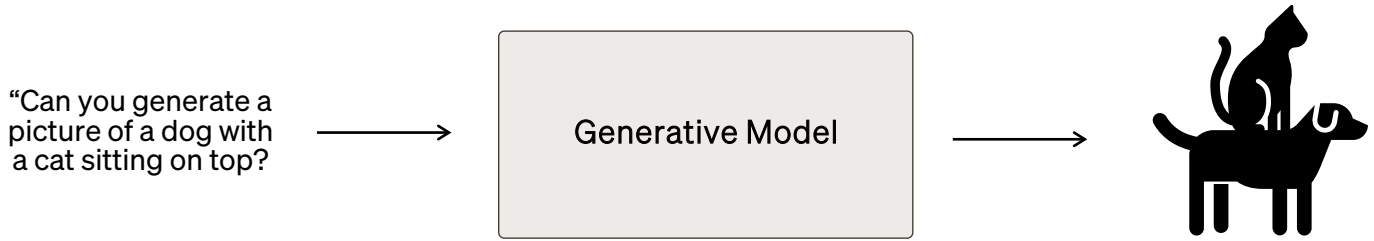


“Is this an image of a dog or a cat?”

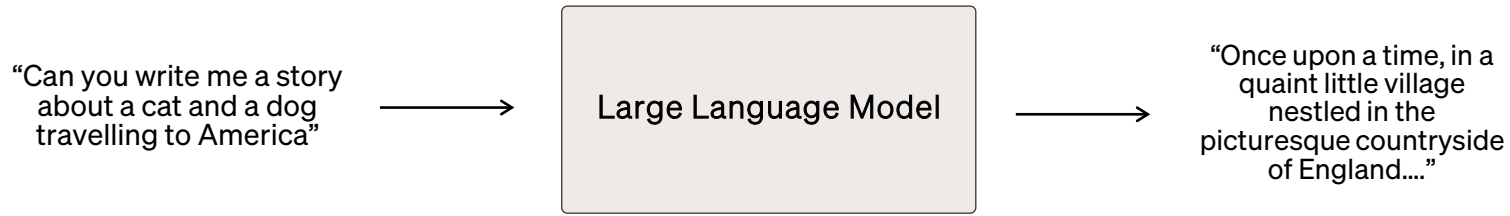
Discriminative Model

Dog

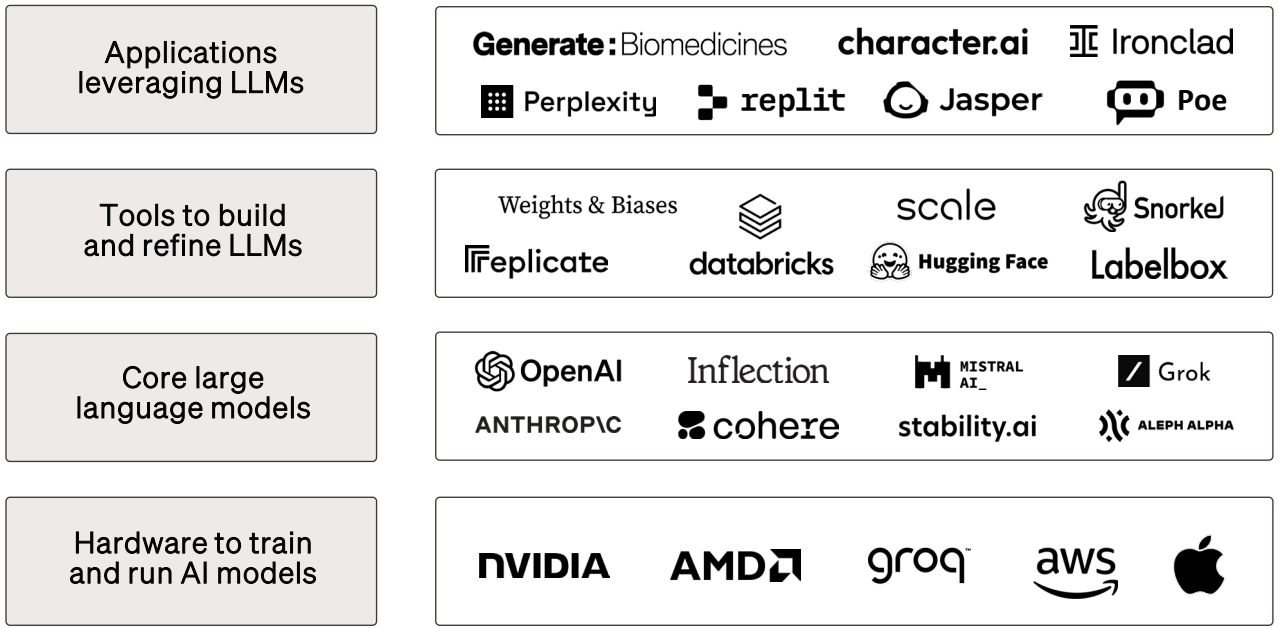
... and generative models, which are used to create new types of data that don't already exist in the world



Large language models (LLMs) are a type of generative model
that can write new strings of text using natural language as their input



As LLMs have become increasingly popular, a new ecosystem of companies is developing to help build, refine, and leverage these models



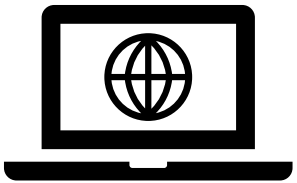
How do you build a large language model?

Large language models must first be trained before they can make predictions about words using a process called inference



Large language models are trained by scraping large amounts of text data from diverse sources across the internet

Large amounts of data are scraped from the internet...



... and converted into text files



This text data is cleaned up and broken down into smaller units called ‘tokens’

“The man is walking”

Token 1	Token 2	Token 3	Token 4
The	Man	Is	Walking

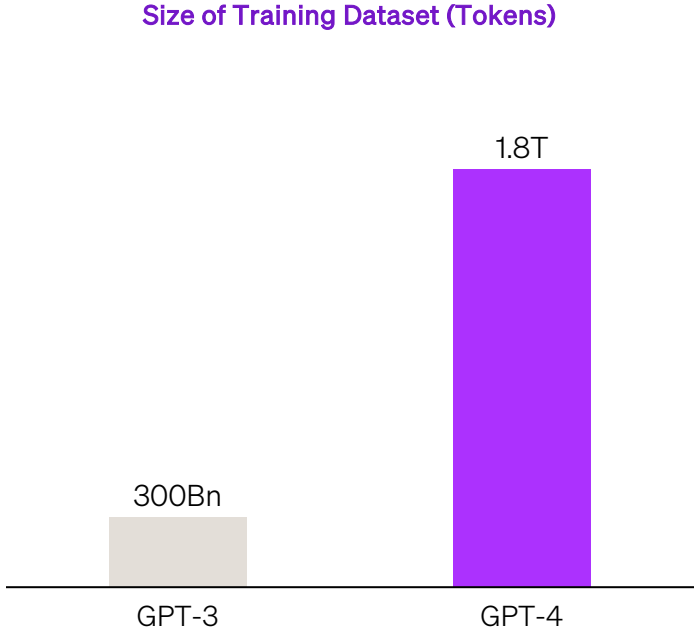
The model then aims to predict the next token using an initial arrangement of random weights and biases



Each of the model's weights and biases are iteratively updated until it can accurately predict the next word



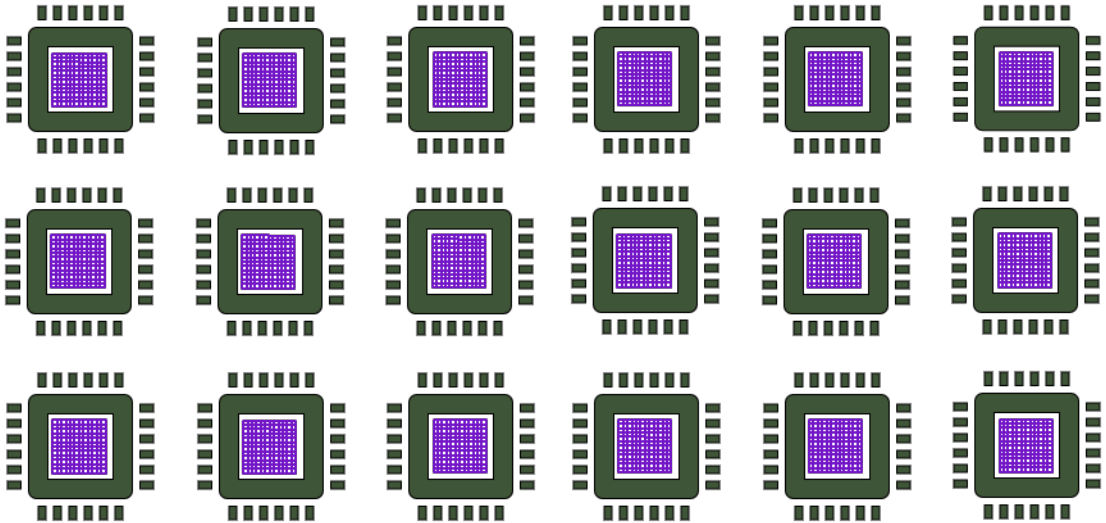
Models are trained on a dataset of billions and even trillions of tokens to map and understand the relationships between words on the internet



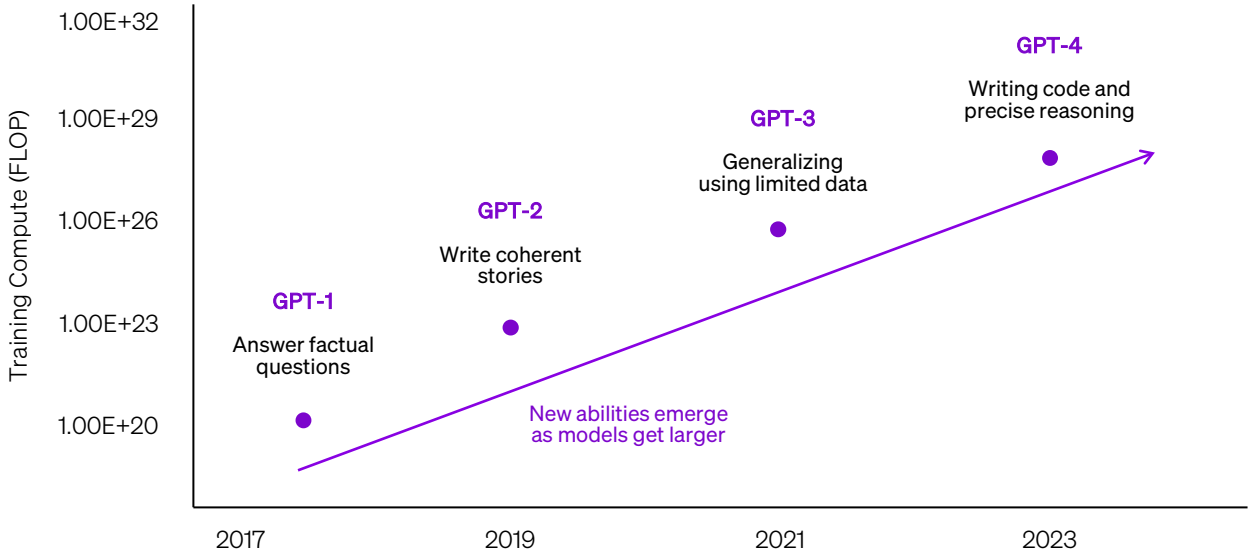
Source: Lambda Labs, Klu AI

This requires large amounts of compute power to represent and map the relationships between words

GPT-4 was trained using ~25,000 Nvidia A100 GPUs for ~90-120 days



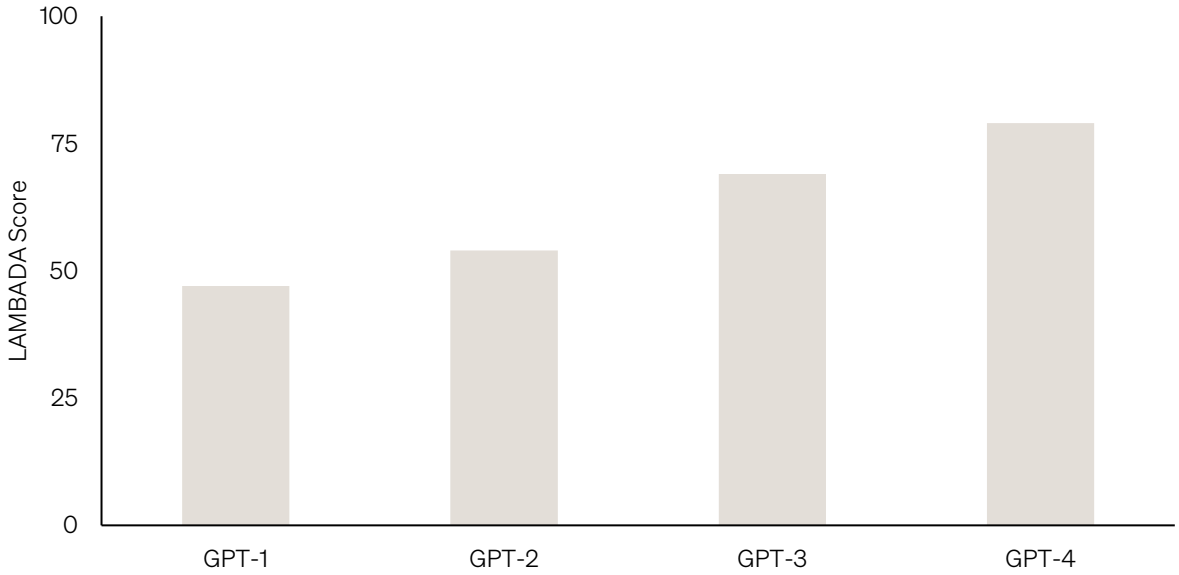
As models have been trained on more data and compute, their performance, capabilities, and utility across real-world applications has improved significantly



Source: Berglund et al (2023), "Taken Out of Context: On Measuring Situational Awareness in LLMs" 147

But while more training data does lead to better performance, exponential increases in training data only result in linear improvements in performance...

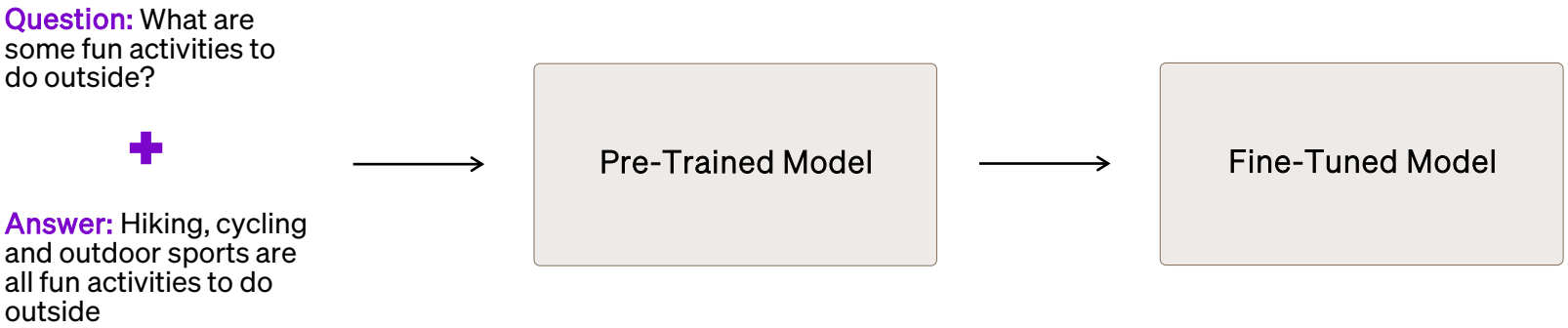
LAMBADA score is a performance metric for predicting words in text



Source: KD Nuggets

Once a language model is pre-trained, we can fine-tune it to perform a particular task like answering questions

This involves feeding the pre-trained model
thousands of questions and their corresponding answers...



...until it can accurately predict the best response to a given question

Question:

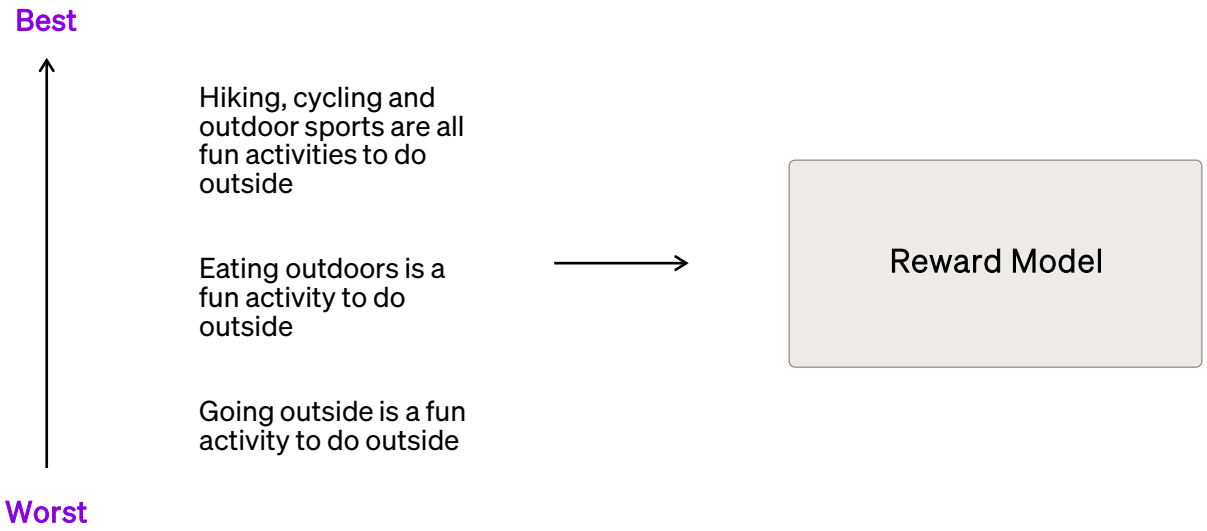
What are some fun activities to do outside?

Fine-Tuned Model

Answer:

Hiking, cycling and outdoor sports are all fun activities to do outside

Reinforcement learning from human feedback (RLHF) is used to rank the model's responses, which can then be used to train a 'reward' model to predict the best response



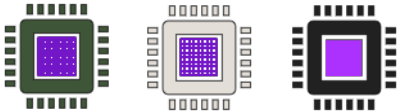
After fine-tuning, models predict answers to questions using a process called ‘inference’

Inference is less computationally intensive than training, but still requires clusters of specialized hardware to compute the appropriate response to a given question

Inference

Question:

What are some fun activities to do outside?



Answer:

Hiking, cycling and outdoor sports are all fun activities to do outside

Source: Nvidia

LLMs face two key issues during inference

First, LLMs can ‘hallucinate’ and create convincing but false information when presented with questions for which they lack context



Second, the power requirements and time to output a token during inference is prohibitively high for many commercial applications



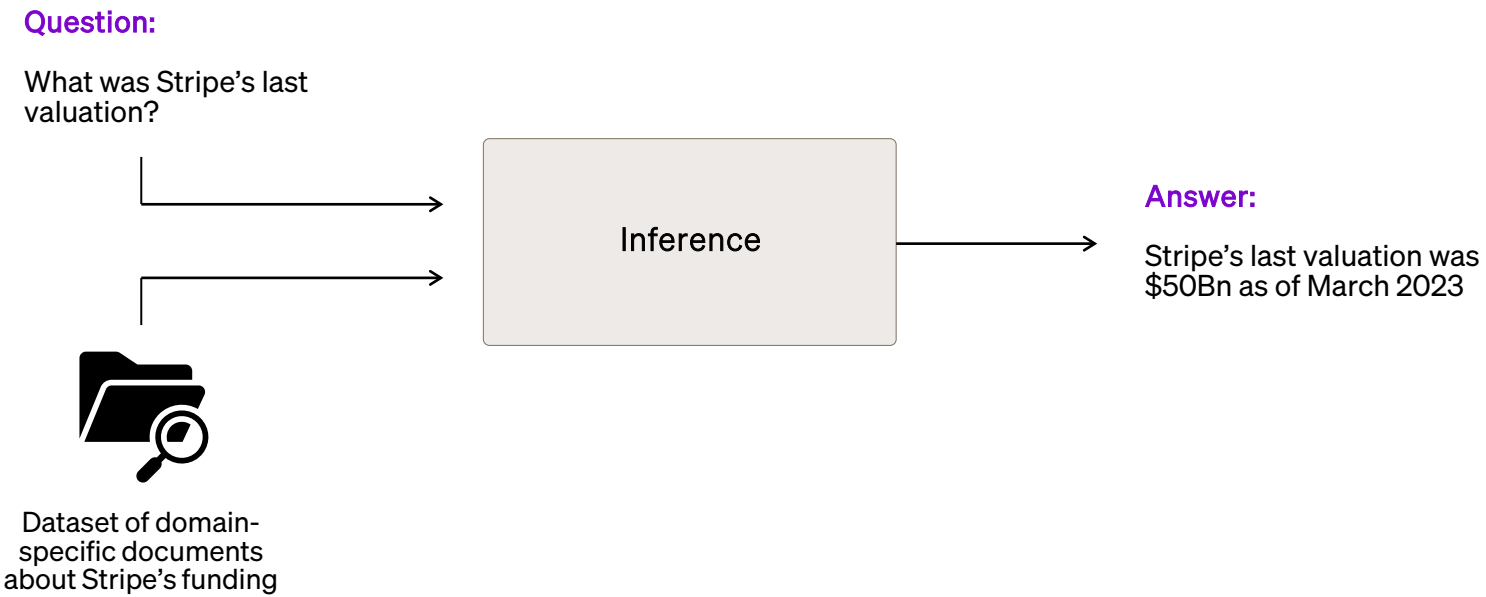
Models use large GPU clusters which draw on significant power to calculate the next most probable word



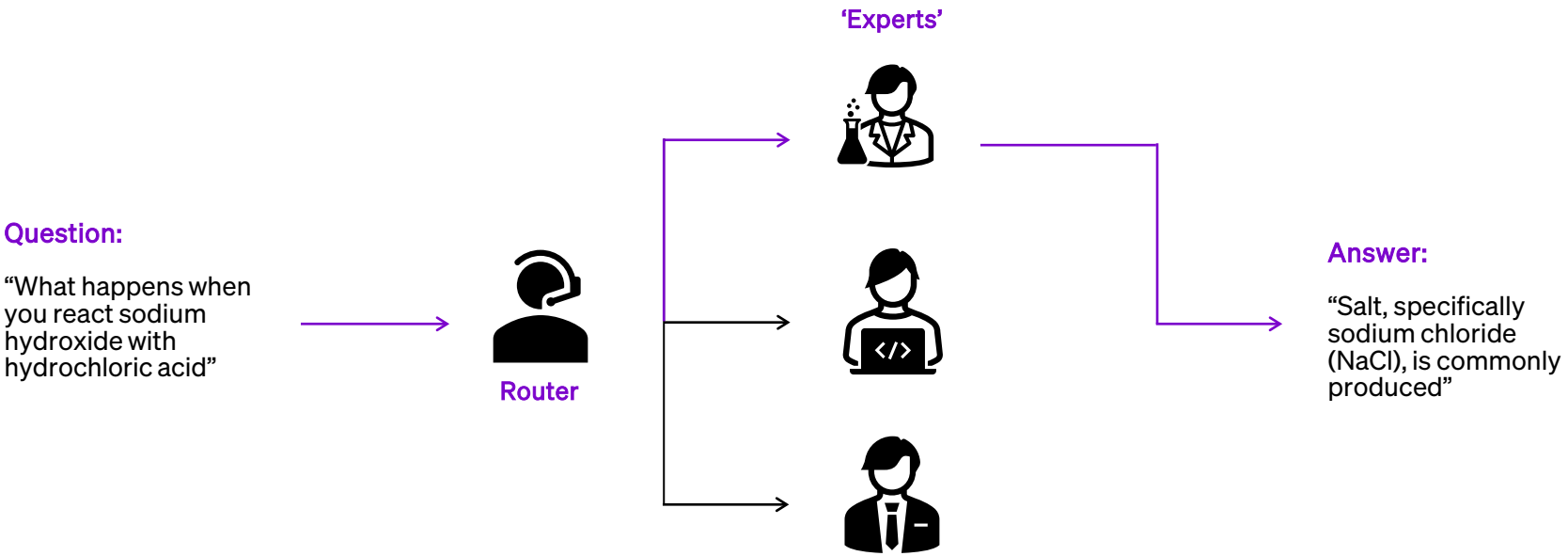
Models generate words token by token – for larger and more complex models, the latency to generate the next token can be frustrating

Source: Nvidia

One way to overcome hallucinations is ‘retrieval augmented generation’ (RAG), which trains a model to use a dataset of specific knowledge to provide more context-aware responses



New techniques like ‘mixture of experts’ route questions to smaller models that are ‘experts’ in that field, reducing the time and compute requirements to run inference



Source: Chen et al (2022), “Towards Understanding Mixture of Experts in Machine Learning”

Dive Deeper...

Further Reading & Watching

Reading:

- How Large Language Models Work (Data Science at Microsoft)
- Different Ways of Training LLMs (Towards Data Science)
- Large Language Model Cost Analysis (La Javaness R&D)
- What Is RAG? (AWS)
- Mixture of Experts Explained (GitHub)

Watching:

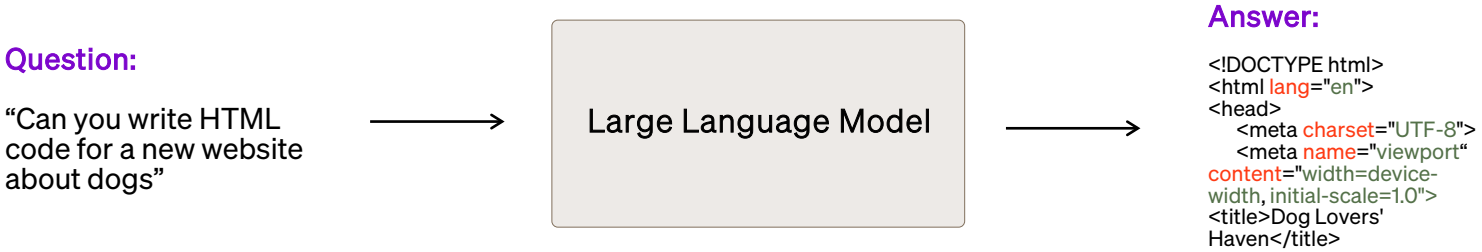
- Intro to Large Language Models (Andrej Karpathy)
- What Makes Large Language Models Expensive? (IBM)
- Why Large Language Models Hallucinate (IBM)

CHAPTER 07

Generative AI & Value Creation

Revolutionizing company building

Generative AI can assist in writing new code, suggesting optimizations, and automating repetitive programming tasks



Source: OpenAI

Language models are particularly effective in predicting potential bugs in code, suggesting fixes, and generating patches automatically

Question:

What is the issue with my code?

```
using System;

// define a Verifier class with a
method for checking Palindrome
public class Verifier
{
   // check if the string is a
palindrome
   public static bool
IsPalindrome(string str)
   {
```

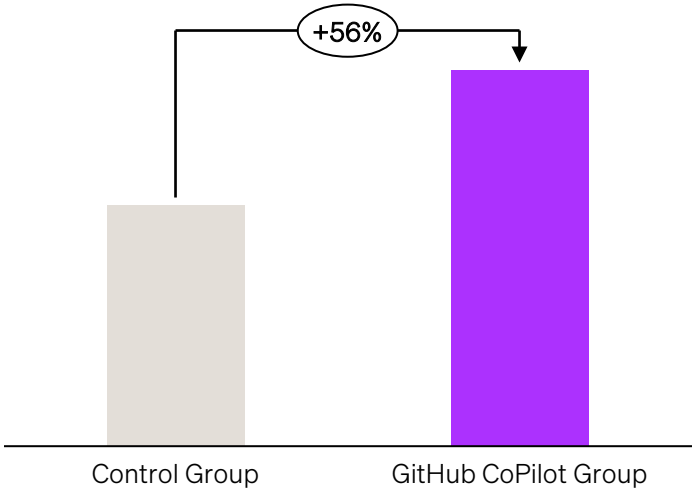

Code Copilot

Answer

There is an error in the 'IsPalindrome' method...

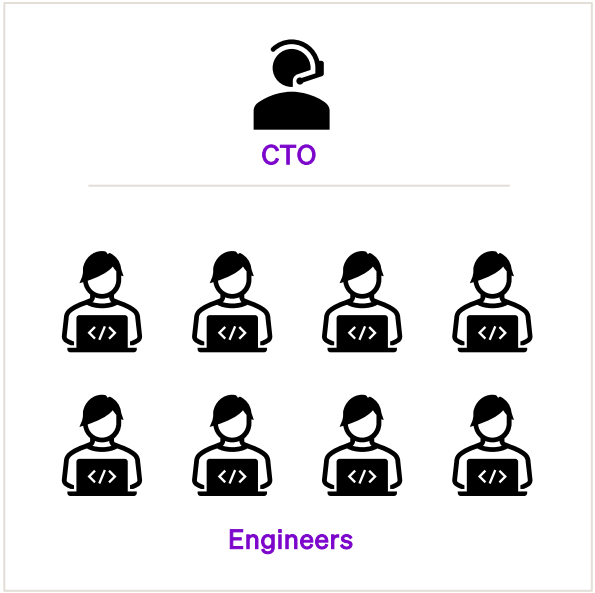
This significantly increases the productivity of a developer team...

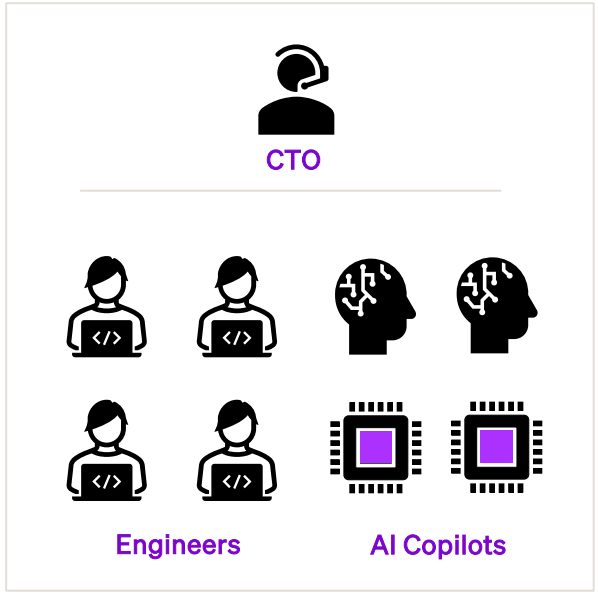
Developers who used GitHub copilot were found to be ~56% more efficient at completing tasks



Source: Source: Peng et al (2023), The Impact of AI on Developer Productivity

... allowing smaller teams to now bring new products to market





Source: Coatue 166

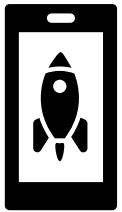
Text to image tools are improving the pace of product prototyping, which can increase the speed at which a company iterates on a new product to discover product-market-fit

Question:

"Design an ultra minimalist landing page for a rocket game app and give me the associated python code"

Text to image tool

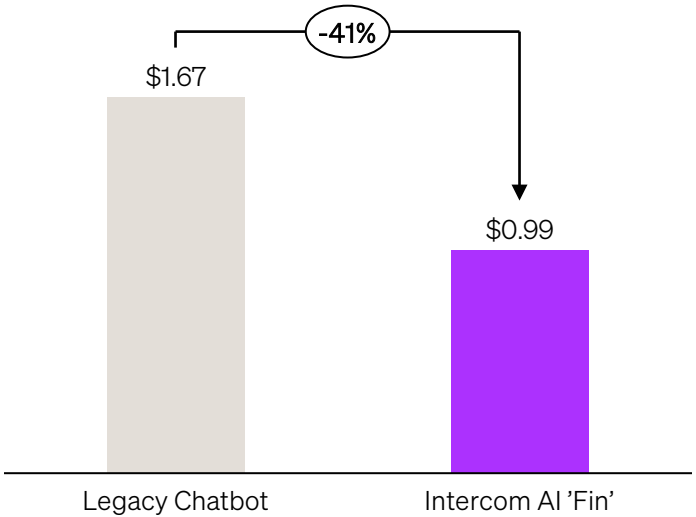
Answer:



```
<!DOCTYPE html>
<html lang="en">
<head>
<meta charset="UTF-8">
<meta name="viewport"
content="width=device-width,
initial-scale=1.0">
<title>Rocket Rush</title>
<style>
 body, html {
  height: 100%;
  margin: 0;
  font-family: 'Arial'
```

Once a product is developed, effective customer service is becoming cheaper and faster as LLMs improve the speed and cost of resolving customer service tickets

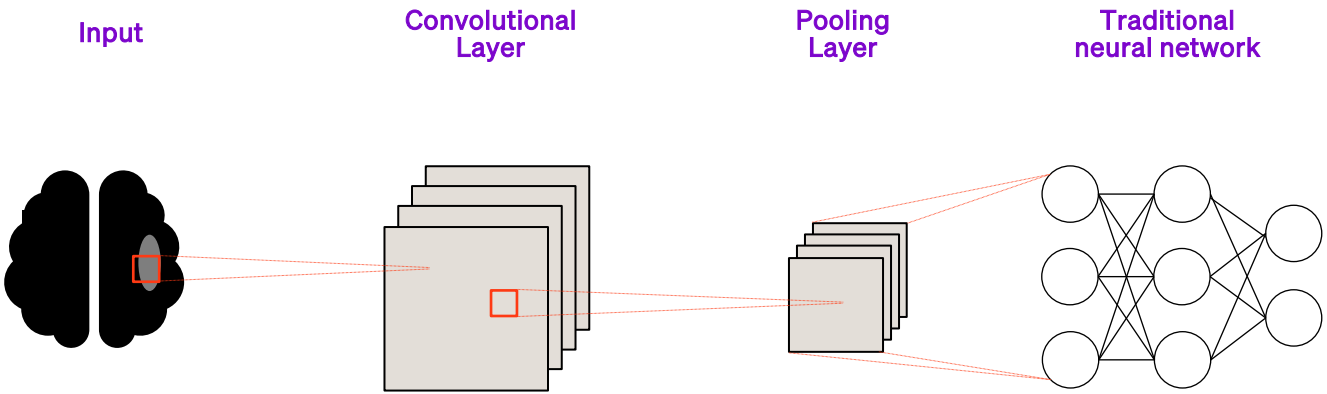
Using LLMs has reduced the cost per ticket for customer service company Intercom by ~40%



Source: Intercom, Invest Like the Best

Applications in healthcare

Discriminative AI models such as CNNs are already used to classify MRI and CT scans more accurately, leading to higher accuracy when diagnosing patients



Source: Chattopadhyay (2022), MRI-Based Brain Tumor Image Detection Using CNN Based Deep Learning Method

Today, there is a new wave of generative AI applications that are focused on further improving diagnostics and patient care

Physicians can speed up repetitive tasks like writing up clinical plans using generative AI tools, which can do this automatically using the physician's notes

Patient Notes



Clinical Decision Support AI Tool

Clinical Plan

Objectives:

Achieve and maintain blood pressure goal of below 130/80 mmHg.

Maintain blood glucose levels within the target range as advised by the diabetes care team.

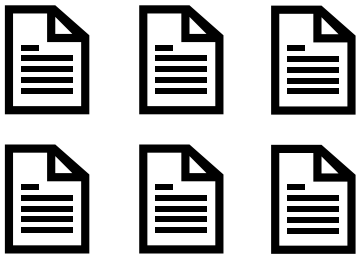
Adhere to medication regimen to manage diabetes, hypertension, and hyperlipidemia.

Source: Meditech, Glass Health

LLMs can also make the process of uncovering clinical practice guidelines much more efficient

Physicians manually review research to determine clinical practice guidelines

44,000 search results for “ovarian cancer screening” on PubMed



LLMs allow physicians to explore topics and uncover guidelines with much less friction

“What are the criteria to screen for ovarian cancer”?

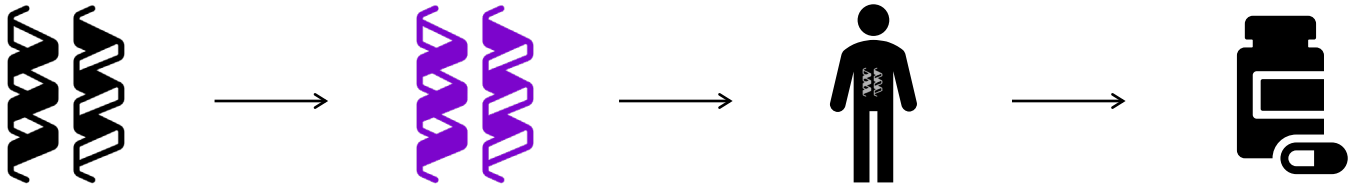


LLM-generated response

Source: Pubmed, Pathway

Another wave of generative AI applications are focused on improving drug discovery and personalized medicine

Generative AI can be used to generate novel protein sequences with unique functional properties that can serve as the basis for new medicines



Generative AI model ingests many natural protein structures and learns common principles on how protein structures are made

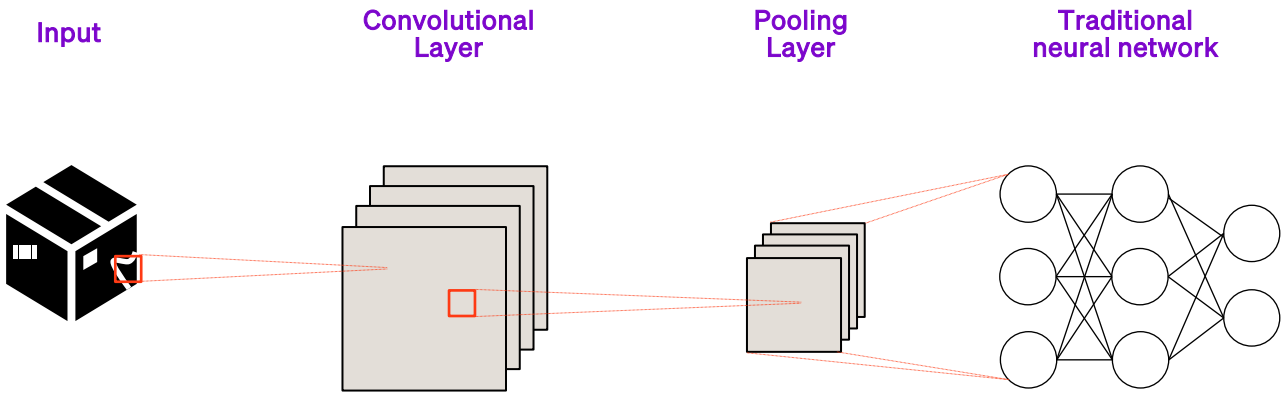
The model learns to generate new protein structures that are not commonly found in nature

Researchers create these novel protein structures and then see which ones create the desired effect within the body

Successful proteins are selected as the basis for new drug development candidates

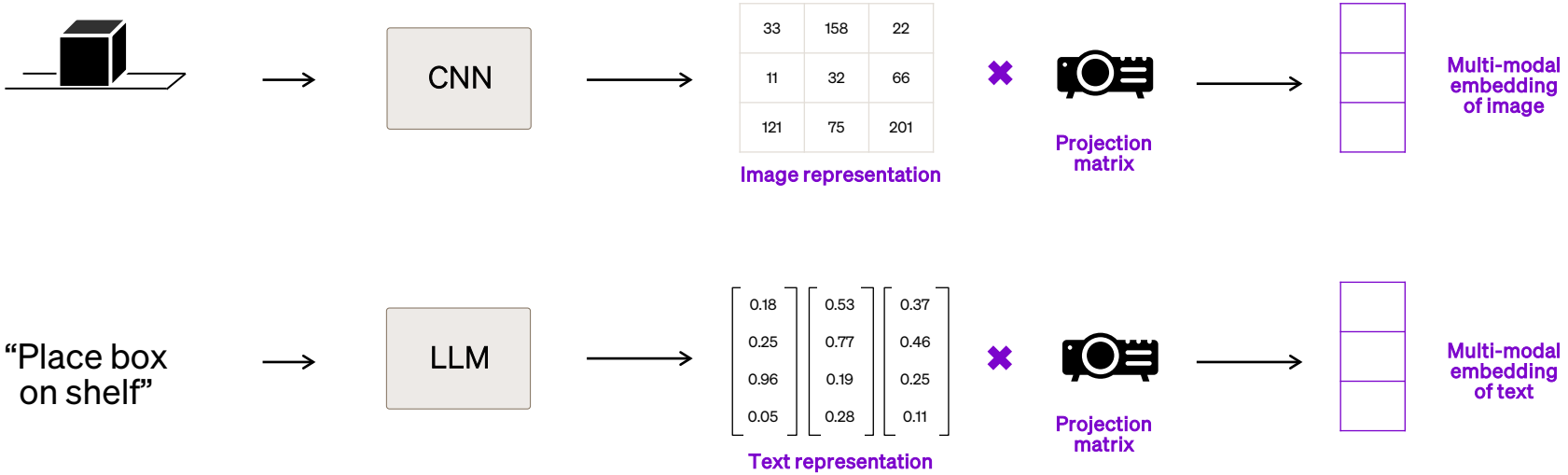
Applications in manufacturing and robotics

Discriminative AI models like convolutional neural networks are already being used in manufacturing for quality control and assurance

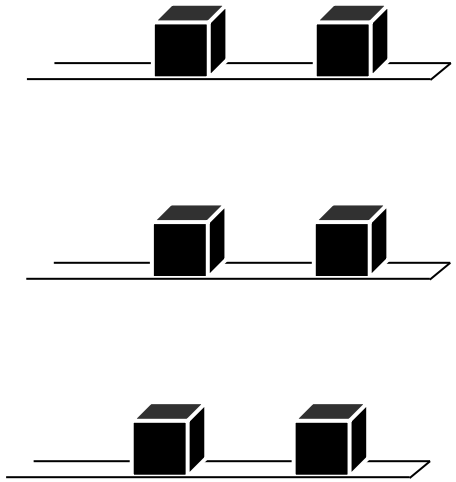


Source: Banus et al (2021), Deep Learning for the Quality Control of Thermoforming Food Packages

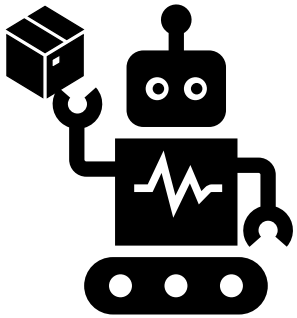
Multi-modal models create representations of different types of data in the same space, allowing AI models to interact with their environment using multiple ‘senses’



This allows more advanced AI models to complete complicated warehouse and manufacturing tasks using robots

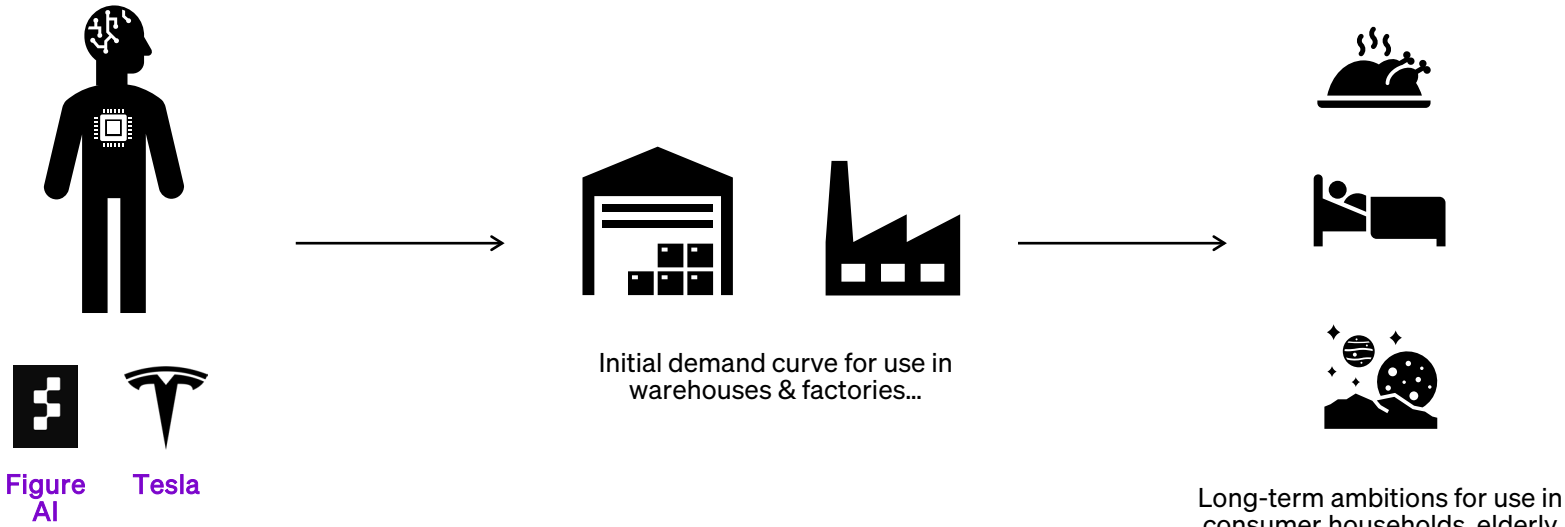


“Place box on shelf”



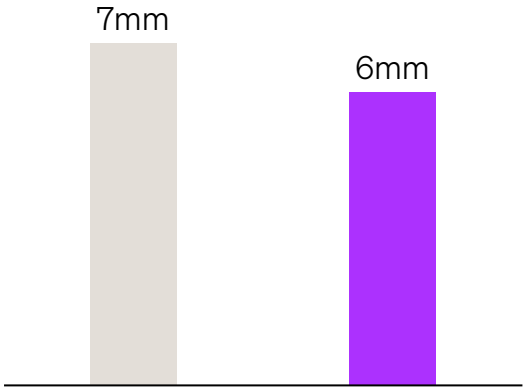
Source: Amazon

Several companies are racing to create humanoid robots which aim to replace humans across various tasks, starting with warehouses and factories



Source: Figure AI, Tesla

This could help to relieve structural labor shortages and churn in these industries, which are expected to run out of people to hire by the end of 2024



Unfulfilled warehouse, transportation and retail jobs

People available to fill these roles, with high attrition rates

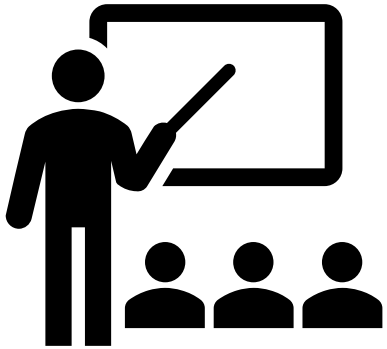
0

Key warehouse suppliers predict they will run out of people to hire by the end of 2024

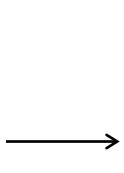
Source: Figure AI

Applications in education

Today, classroom education is a standardized, sub-optimal method of learning that often fails to cater to the individual strengths and weaknesses of each student



Classroom of students





Student A is better at learning through text





Student B is better at learning visually

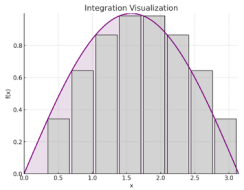
With generative AI tools, learning can become much more personalized and accessible

Question:

"I am in 12th grade - explain integration in a visual, graphical way"

AI learning tool

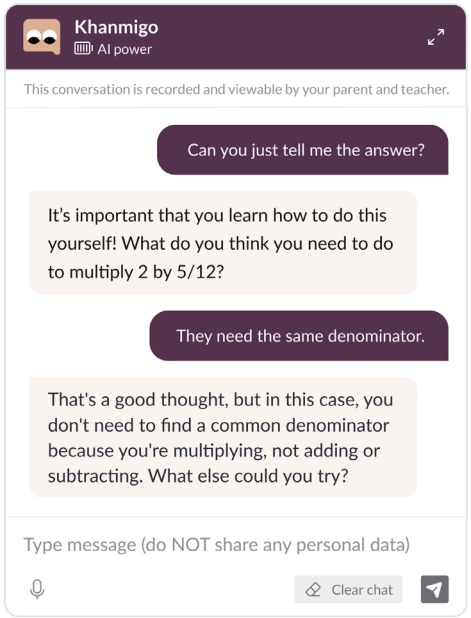
Answer:

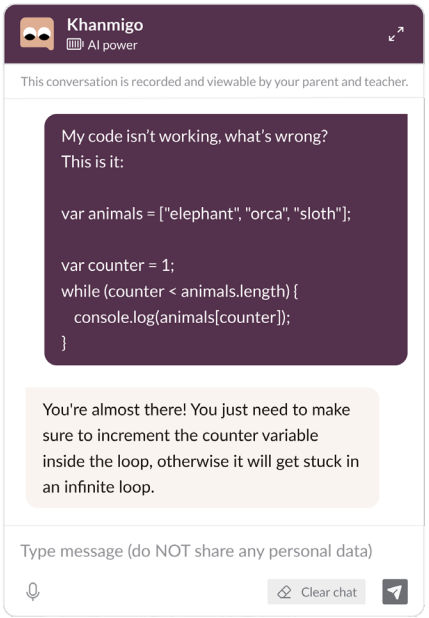


"Integration can be visually understood as a way of finding the area under a curve. Here's a step-by-step graphical explanation..."

Source: Khan Academy, OpenAI

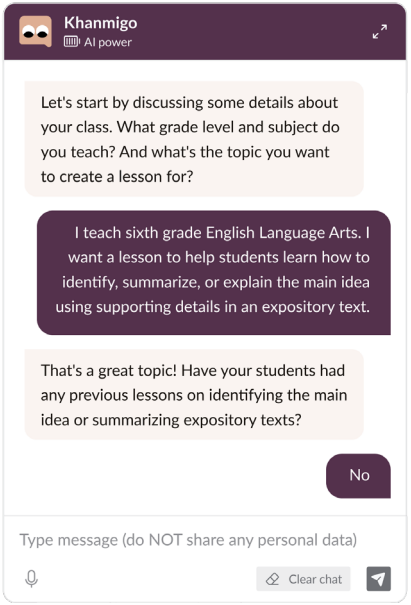
Tools like ‘Khanmigo’ can unlock accessible personal tutoring for students...





Source: Khan Academy

...and more efficient lesson planning for educators



Lesson Plan: Identifying and Summarizing the Main Idea in Expository Texts

Grade Level: 6th Grade
Subject: English Language Arts
Duration: 1 hour

Objective: Students will be able to identify the main idea of an expository text, summarize the text, and explain the main idea using supporting details.

Materials:
•Expository text samples (articles, essays, textbook excerpts)
•Highlighters
•Note cards
•Whiteboard and markers
•'Main Idea and Supporting Details' worksheet

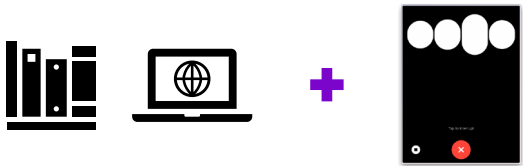
Source: Khan Academy, Lesson Plans AI

The barriers to entry to learn about any topic at any level of complexity have never been lower

Students use textbooks and online resources, but often have no recourse to ask further questions



Students can leverage AI tools to verbally ask clarifying questions, similar to asking a personal tutor



Applications for content creators

Generative AI is a transformative tool for content creators across all modes of communication including text, audio, image and video

Text	Images	Video	Audio
ChatGPT	Midjourney	runway	splash
Rytr.me	DALL·E	PICTORY	beatoven.ai
Scribe	Jasper	SYNTHESYS	amper
Bard	Stable Diffusion	lumen5	Jukedeck
Notion	craiyon	Colossyan	MUSENET

Source: Company Data

Writing tools like ChatGPT can be used to write content that either serves as a prompt for further development, or stands alone as a piece of writing



Text to image generation tools like Stable Diffusion can produce new images from text descriptions that can serve to enhance written communication

Question:

"Can you use this script to generate images of muffin and buddy in the style of a cartoon?"

→

Image Generation Tool

→

Answer:



Source: OpenAI, Stable Diffusion

And video-creation tools like Runway can create entire videos using just an image for the initial frame

Question:

"Can you use this image to create a video of the animals jumping up and down?"



Video Generation Tool

Answer:



Source: OpenAI, Runway

Models like Splash can create new music in different styles from a single text prompt

Question:

"Can you create a soundtrack to a cartoon video of a cat and dog jumping up and down?"

Music Generation Tool

Answer:



Dive Deeper...

Further Reading & Watching

Reading:

- AI: The Coming Revolution (Coatue)
- Research: Quantifying GitHub Copilot's Impact on Code Quality (GitHub)
- Everything You Need to Know About Fin, the Breakthrough AI Bot Transforming Customer Service (Intercom)
- How Generative AI Is Changing Creative Work (Harvard Business Review)
- Generative AI Imagines New Protein Structures (MIT)
- Meet Khan Academy's Chatbot Tutor (CNN)

Watching:

- Inside a Humanoid Robot Lab (Figure AI)

CHAPTER 08

Artificial General Intelligence (AGI)

Artificial general intelligence is a type of AI which is intelligent enough to replace the average human for nearly any job

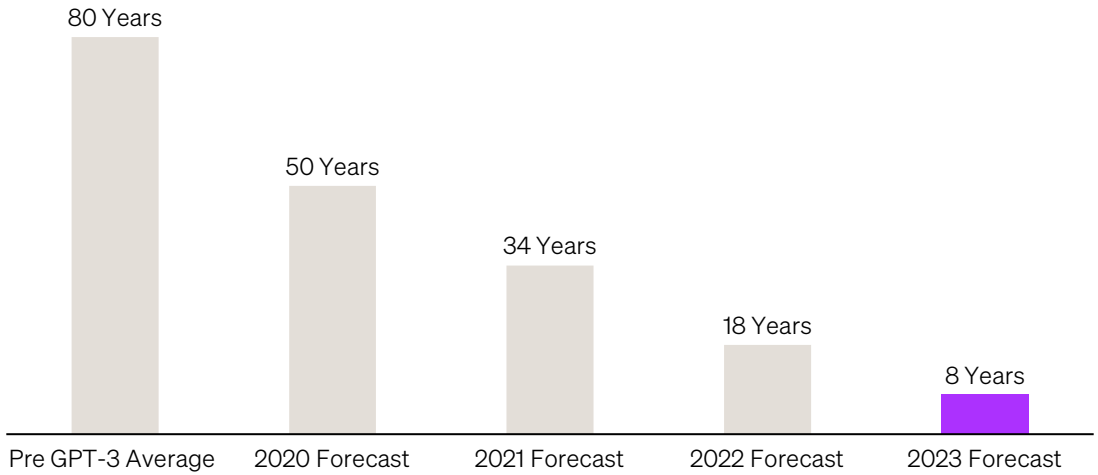
While narrow AI systems can comfortably outperform humans for specific roles, general AI systems are yet to reach this level of competence

	Narrow AI	General AI
Level 1: Emerging Equal to or somewhat better than an unskilled human	Symbolic AI, simple rule-based systems	Human-in-the-loop computing e.g. Amazon Mechanical Turk
Level 2: Competent At least 50th percentile of skilled adults	Smart speakers such as Siri, Alexa, or Google Assistant	ChatGPT, Bard, Llama 2
Level 3: Expert At least 90th percentile of skilled adults	Spelling and grammar checkers such as Grammarly; generative image models such as DALL-E2	Not yet achieved
Level 4: Virtuoso At least 99th percentile of skilled adults	DeepBlue, AlphaGo	Not yet achieved
Level 5: Superhuman Outperforms 100% of humans	AlphaFold, AlphaZero, StockFish	Not yet achieved

Source: Google Deepmind

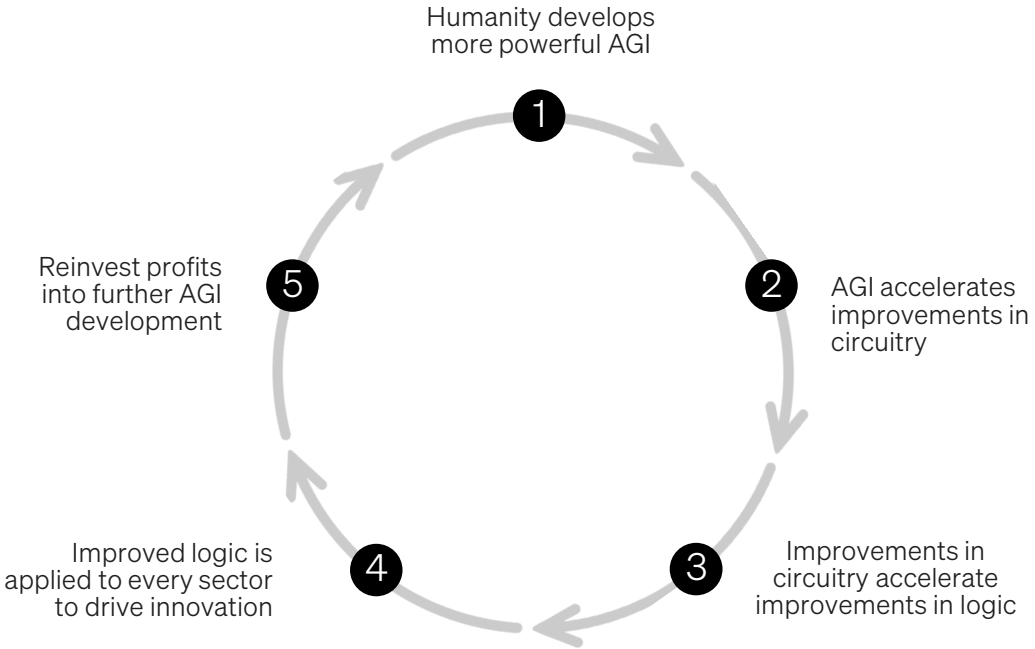
The expected timeline for AGI has been shrinking as models have become increasingly intelligent and performed better than expected

Expected Years Until a General Artificial Intelligence System Becomes Available

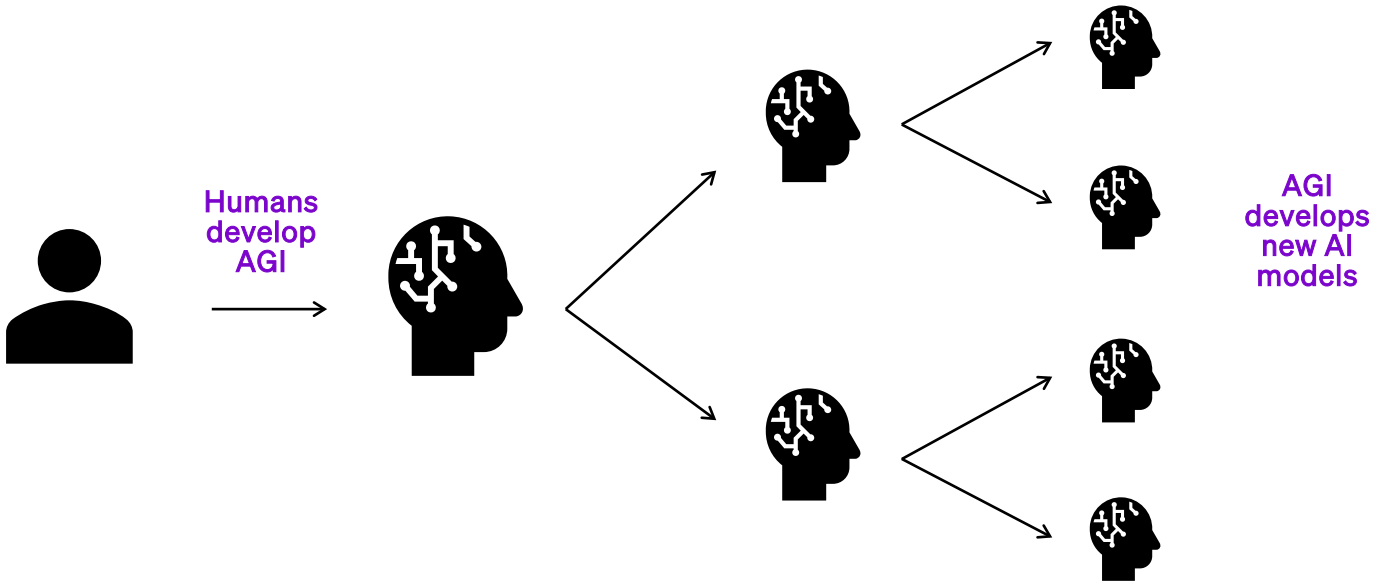


Source: ARK Institute

As AGI nears, powerful feedback loops could emerge to rapidly increase the pace of innovation and progress



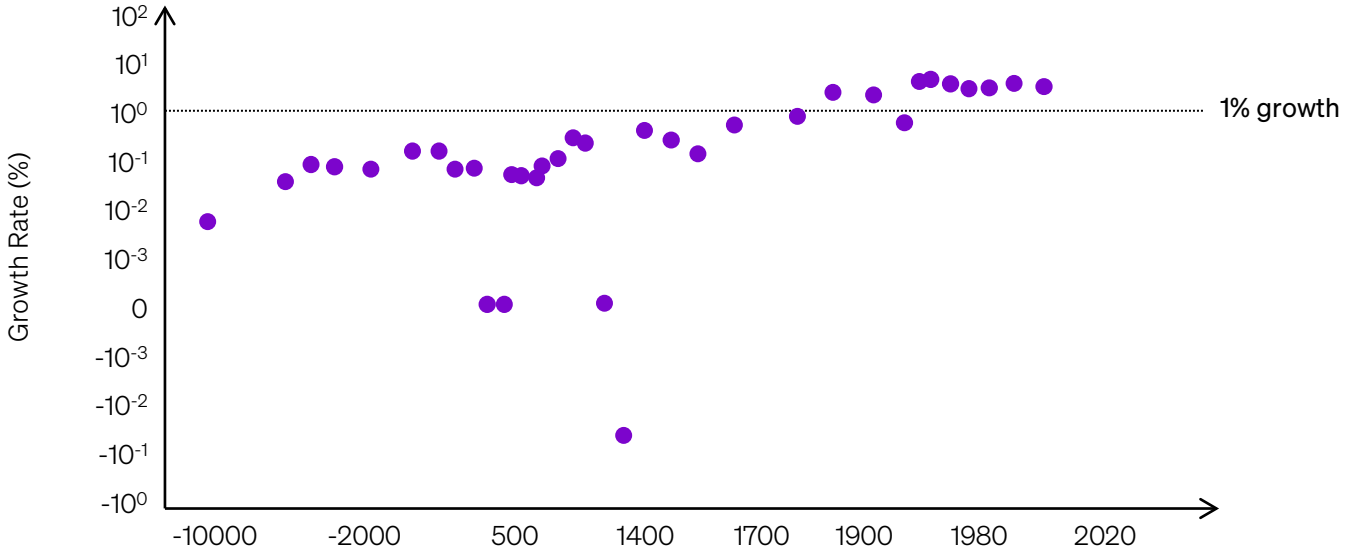
While humans will be the first to build AGI, an AGI could theoretically be trained to produce subsequent AI models



Source: Open Philanthropy

In the past, new technologies allowed the global population to grow, driving further technological innovation and increases in gross world product

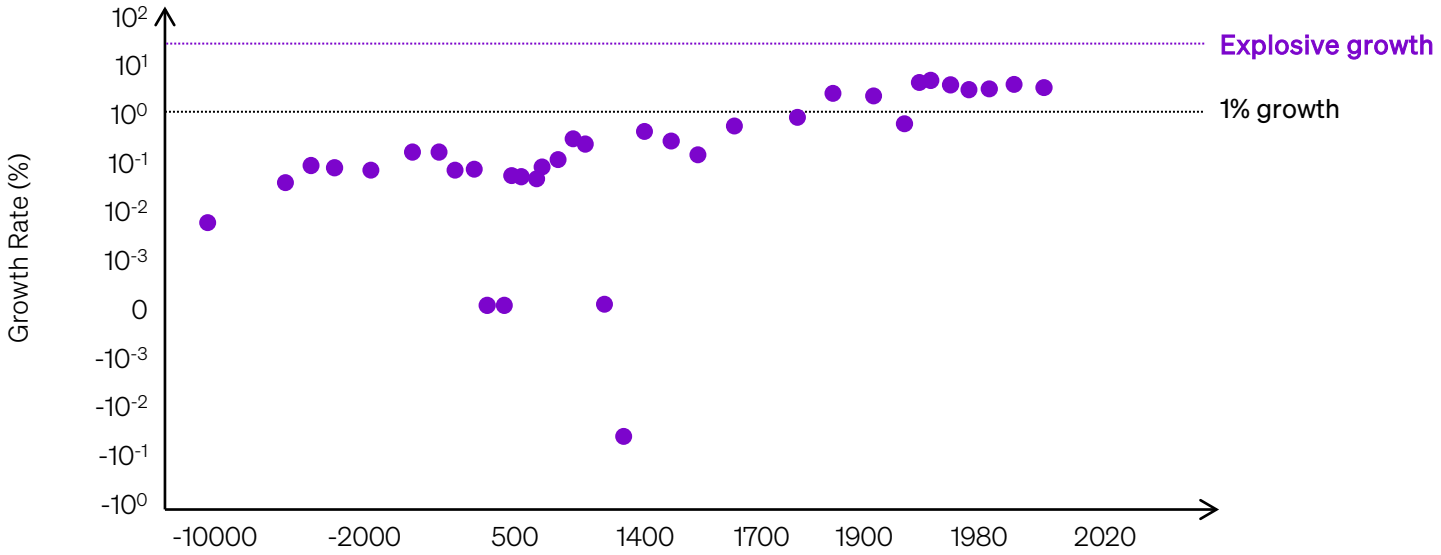
Observed Gross World Product Growth Over Time



Source: Open Philanthropy

But in the future, as AI systems proliferate and increasingly replace human labor, this same flywheel could meaningfully accelerate GDP

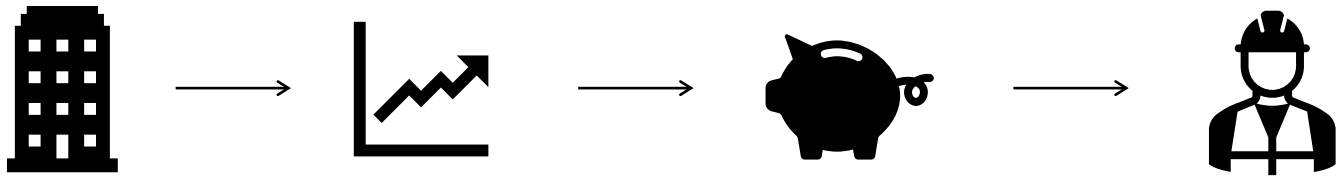
Observed Gross World Product Growth Over Time



Source: Open Philanthropy

This has profound implications for employment across multiple sectors...

Various forms of universal basic income, financed by a tax on companies that benefit from AGI, have been proposed to redistribute wealth to those made unemployed by AGI



Companies replace human labor with AI systems that boost efficiency and improve profitability

Governments charge a tax on the market caps of certain companies above a particular threshold

Shares of these companies are paid into a fund and allowed to compound over many years

Shares from the fund are sold, and proceeds are redistributed to people made unemployed by AGI

Source: Sam Altman

Dive Deeper...

Further Reading & Watching

Reading:

- Moore's Law for Everything (Sam Altman)
- Could Advanced AI Drive Explosive Economic Growth? (Open Philanthropy)
- An Executive Primer on Artificial General Intelligence (McKinsey)
- Google DeepMind's Six Levels of AGI (Google Deepmind)

Watching:

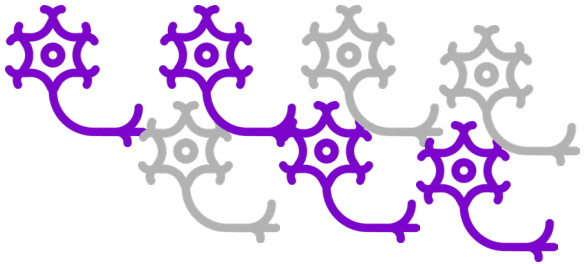
- The Exciting, Perilous Journey Toward AGI (Ilya Sutskever)
- The Transformative Potential of AGI — and When It Might Arrive (Ted)

CHAPTER 09

Wrapping Up...

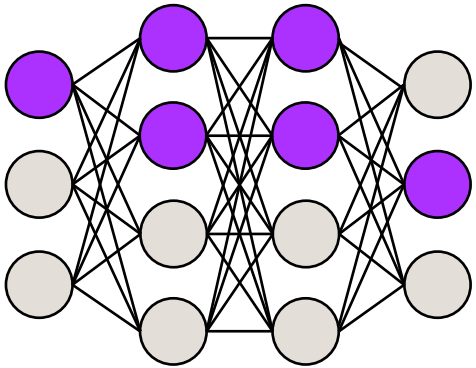
Neural networks were introduced as a system of logic to replicate the human brain...

Neurons fire together...



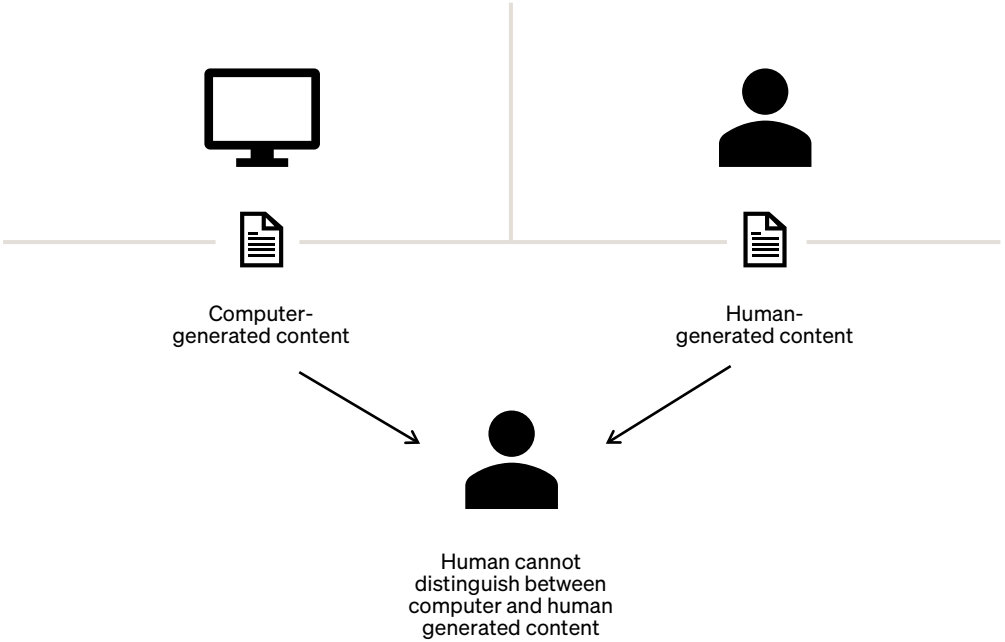
=

Nodes fire together...



Source: Neuroscience News

...with the goal of eventually building a computer system that is indistinguishable from humans



Source: Saygin (2000), Turing Test 50 Years Later

While we have already achieved superhuman abilities in narrow AI systems...

	Narrow AI
Level 1: Emerging Equal to or somewhat better than an unskilled human	Symbolic AI, simple rule-based systems
Level 2: Competent At least 50th percentile of skilled adults	Smart speakers such as Siri, Alexa, or Google Assistant
Level 3: Expert At least 90th percentile of skilled adults	Spelling and grammar checkers such as Grammarly; generative image models such as DALL-E2
Level 4: Virtuoso At least 99th percentile of skilled adults	DeepBlue, AlphaGo
Level 5: Superhuman Outperforms 100% of humans	AlphaFold, AlphaZero, StockFish ← Where we are today

Source: Google Deepmind

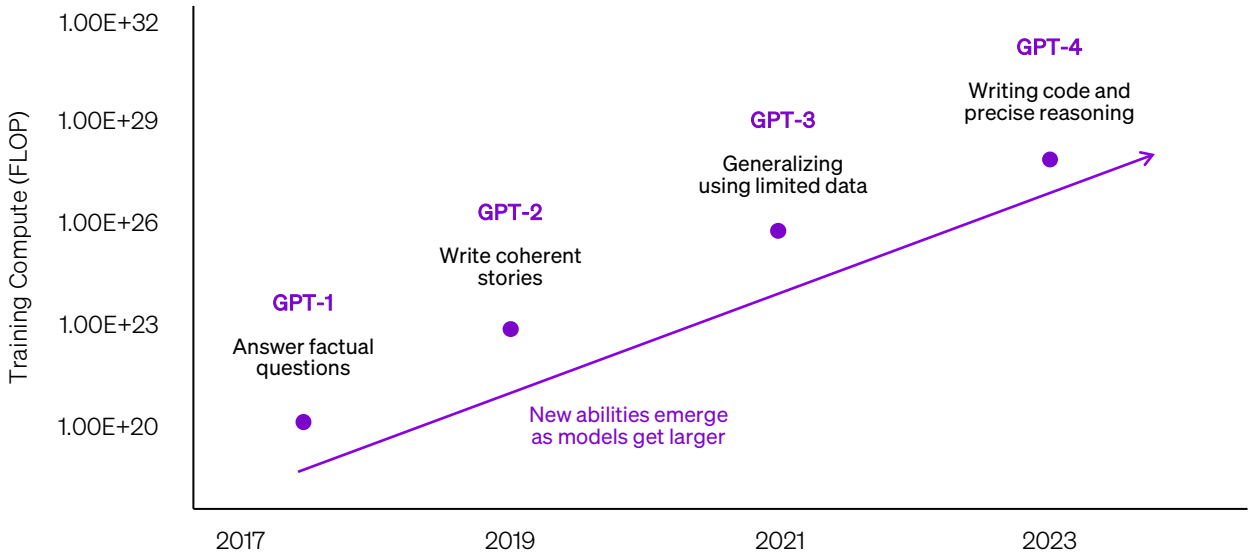
...we are still in the early innings of building general AI

	General AI
Level 1: Emerging Equal to or somewhat better than an unskilled human	Human-in-the-loop computing e.g. Amazon Mechanical Turk
Level 2: Competent At least 50th percentile of skilled adults	ChatGPT, Bard, Llama 2
Level 3: Expert At least 90th percentile of skilled adults	Not yet achieved
Level 4: Virtuoso At least 99th percentile of skilled adults	Not yet achieved
Level 5: Superhuman Outperforms 100% of humans	Not yet achieved

← Where we are today

Source: Google Deepmind

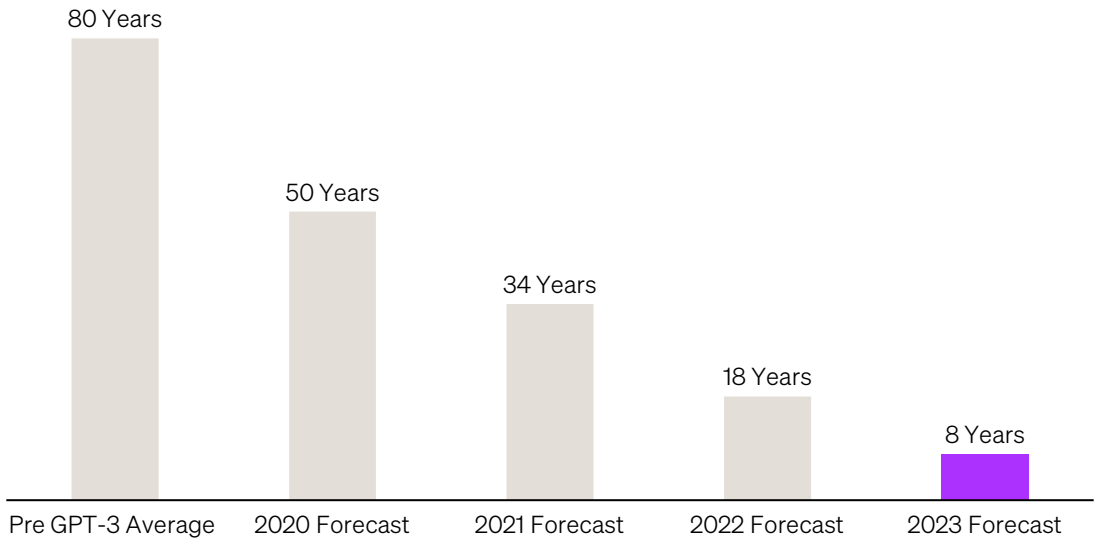
But as general AI systems have become increasingly performant...



Source: Berglund et al (2023), Taken Out of Context: On Measuring Situational Awareness in LLMs

...the expected timeline to build AGI has been shrinking

Expected Years Until a General Artificial Intelligence System Becomes Available



Source: ARK Institute