The AI renaissance:
Why it has taken off and where it is going

CONTENTS

Foreword

The technological enablers for widespread adoption of artificial intelligence (AI) have fallen into place. Early movers in technology industries have driven initial applications, but the scene is now set for widespread and accelerated adoption.

For ambitious businesses, this is good news.

There is now a democratization of access to AI techniques, thanks to faster and cheaper compute, better data and significant advances in machine learning techniques.

AI has firmly taken its place as a transformative technology with limitless potential. I believe it has the potency to transform diverse industries – especially as we respond to the challenges brought about by the COVID-19 pandemic.

We're seeing an AI renaissance, but not everything is solved yet. There is a complex landscape of hardware, data and algorithms to navigate (some off-the-shelf and others custom built for true competitive differentiation). And of course, making the new AI systems robust, responsible and explainable is not necessarily straightforward.

In this paper, we look at what is happening in today's market to spark this AI innovation. Perhaps more importantly, we explore where the opportunities lie for ambitious businesses in the coming years.

Much of this innovation will be driven by the interplay of AI with other technologies. For example, there are many synergies between AI and 5G, and a growing opportunity to take advantage of advances in semiconductors to deploy AI that is smaller, faster and at the edge.

This technological convergence leaves me highly optimistic that, despite these challenging times, it's increasingly realistic for many businesses to enhance their strategy and operations with AI. This could include increasing operational efficiencies through more robust processes or developing completely new digital services to reignite a business and transform a market.

We've been honored to work with many businesses across the globe to harness the potential of AI. I hope this paper proves to be a useful resource for you and perhaps inspires the next innovation in your business. If you have any feedback or would like to discuss these ideas in more detail, please don't hesitate to send me an email.

Tim Ensor
Director of Artificial Intelligence
tim.ensor@cambridgeconsultants.com

The AI renaissance: why it has taken off and where it is going

the great technological leaps in history have all been the culmination and alignment of multiple enablers. The first industrial revolution, from about 1760 to 1840, led to a huge increase in economic output and with it came better wages and a higher standard of living for much of the population.

It did not happen in the mid-eighteenth century by chance. Many different technologies had been advancing gradually over the preceding decades, and only when the necessary contributing factors were in place could the revolution gain pace.

- Improvements in **transportation**, such as the steam locomotive and steamship, enabled raw materials to be brought from their sources – in many cases from global locations – to a central point of manufacture and the finished products to be sent to their destination markets
- The use of **power sources** advanced from water wheels and windmills to steam-powered engines which could provide a much more reliable and higher power output for a more centralized factory
- Advances in **materials and mechanics**, including the mass production of steel and manufacturing techniques, such as the use of milling machines and screw-cutting lathes, enabled the development of more efficient specialized tools

1.1 Technological enablers have fallen into place

More recently, the rapid uptake of smartphones since the launch of the iPhone in 2007 has demonstrated how quickly a new technology can become pervasive.

Before the arrival of the smartphone there had been early precursors which indicated the applications and market demand, but which failed to take off. It was only when all the necessary technological (and market) building blocks were in place that the scene was set for smartphones as we know them today:

- Advances **in battery technology** were needed so that a device could last for at least a day between charges
- **Touchscreen displays** were needed to provide the simple and intuitive interface for a mass consumer product
- **Low-power processing** was needed to handle the many data-rich applications which we regularly use on our smartphones
- All this was needed at **a low enough cost** that early adopters could drive initial volumes high enough to allow later lower cost devices to reach more consumers

Similarly, the technological enablers for widespread adoption of AI have been falling into place in recent years. Early adopters in technology industries have driven initial applications but the scene is now set for widespread adoption to accelerate.

In this paper we explore the technological enablers which are contributing to the revolution in AI, as summarized in Figure 1.

1.2 Advances in AI techniques

The term AI is used loosely to describe a wide range of computational techniques of varying complexity. It is defined by its systems' behaviors, performance and results rather than the ways they are achieved. A selection of these is illustrated in Figure 2.

AI algorithms have improved significantly in recent years. Faster compute and better data don't just enable algorithms to be trained for better accuracy and inference, but also facilitate easier and more practical experimentation. As a result, architectures that were originally designed in the 1980s, 1990s and 2000s but were too impractical to implement, can now be fully explored.

Machine learning (ML) is a general framework and family of techniques that enable computer systems to computationally improve with increasing experience of performing a task. The experience is typically built from examples of task data, but it can also be simulated or real-world experience from an agent-environment interaction. ML is by far the largest field within AI and is the focus of the most active research. Many of the recent advances in ML have arisen from the development of artificial neural networks (ANN) which are inspired by the workings of animal brains. Table 1 outlines some of the most prominent ANN techniques currently available.

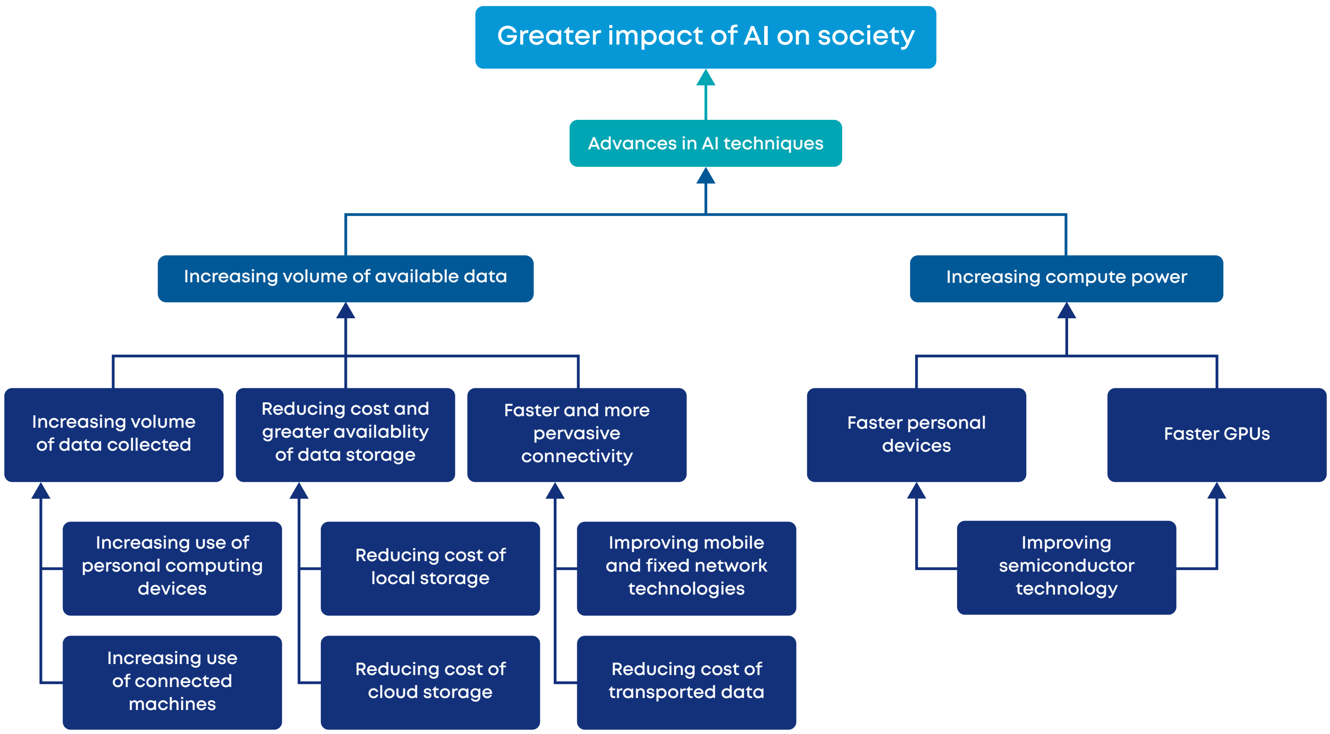
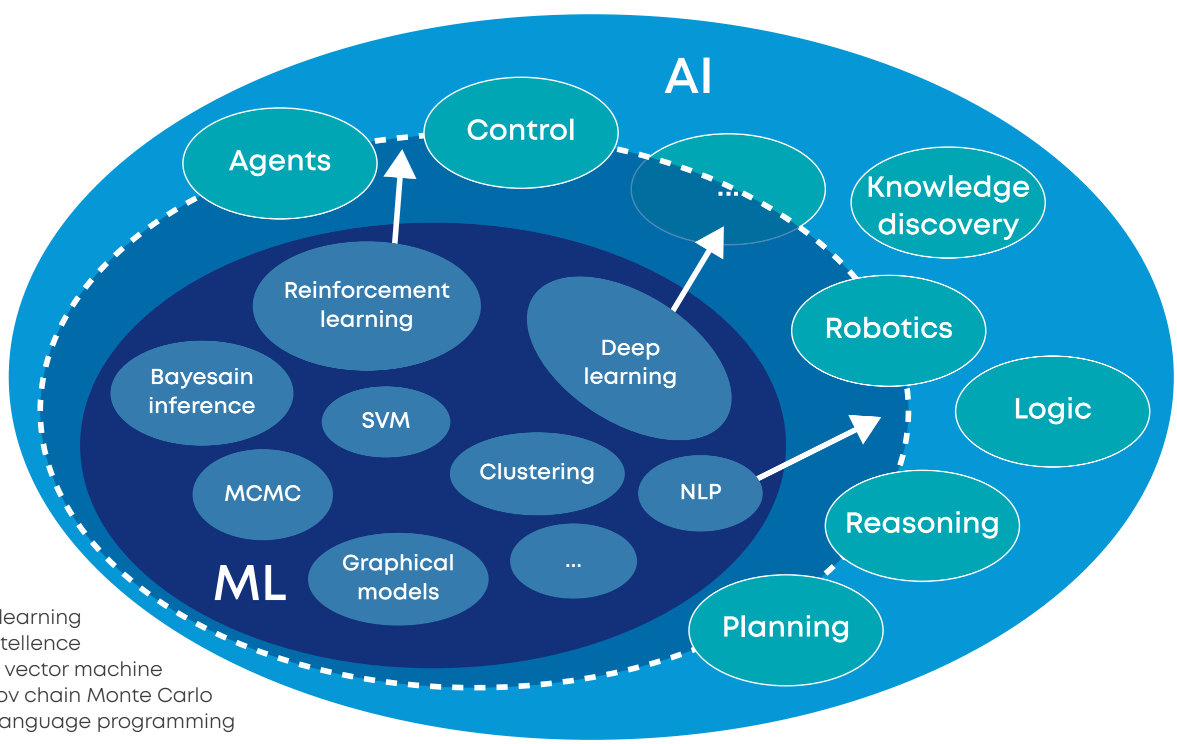


Figure 1: The technological enablers for AI are driving a greater impact on society
Source: Cambridge Consultants



ML – Machine learning
AI – Artificial intellence
SVM – Support vector machine
MCMC – Markov chain Monte Carlo
NLP – Natural language programming

Figure 2: AI covers a range of different techniques and areas of application
Source: Cambridge Consultants

Increasingly, other subfields of AI are being tackled with ML methods. Prominent applications of ML are object recognition, speech recognition, clustering and recommender systems. ML can be further subdivided according to the kind of data that is available to it:

- **Supervised learning** – each example consists of input data and label pairs. The system has access to the correct label and must learn the mapping from input to label
- **Unsupervised learning** – each example consists of input data with no label. The system must learn the underlying structure of the data, probability distributions over parameters to generate the data
- **Reinforcement learning** – input data is gathered by the system through interaction with the dynamic environment. Some interactions generate positive or negative rewards. The system must select actions to maximize the rewards

In addition, the explosion of data has allowed machines to be exposed to ever larger and more comprehensive training sets. The greater compute power has allowed machines to analyze the data in a reasonable time frame, draw conclusions and then use these conclusions to make decisions.

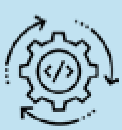
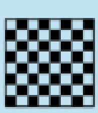
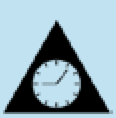
Selected Techniques	Description	Applications
Convolutional Neural Network (CNN) 	▪ Like all neural networks, CNNs comprise many layers of interconnected neurons. In a CNN each layer acts as a filter for the following layer ▪ Each successive layer of a CNN recognizes more complexity and abstract detail than the last	▪ Classifying music and video to improve recommendations for OTT content subscribers ▪ Image and video recognition in autonomous vehicles or security systems ▪ Analyzing the drivers for consumer churn, predicting churners based on behavior in order to intervene ▪ Natural language processing tasks such as search query retrieval
Reinforcement Learning (RL) 	▪ Reinforcement learning involves machines learning from feedback – either in the form of positive and negative reinforcement ▪ The purpose is for the machine to learn how to reach a particular goal or outcome, using trial and error	▪ Machines that are used in manufacturing to move objects can continually learn about the object to move it with greater precision and speed ▪ Delivery management e.g. optimal route to serve customers with just one vehicle ▪ Dimension and optimize networks by learning the drivers for each class of device ▪ Optimize network policy in response to a changing environment
Generative Adversarial Networks (GAN) 	▪ One system generates, another recognizes. This feedback loop allows both systems to improve ▪ Both systems improve by competing and the machine improves its ability to generate scenarios	▪ Creating images from descriptive language ▪ Improving the quality of images – applications in a number of fields including astronomy ▪ Generate parameters to create network slices on demand for different users on the network ▪ Training anomaly detection systems – e.g. medical imaging, network security
Long Short Term Memory (LSTM) 	▪ A type of neural network which is well suited to classify, process and predict time series with time lags of unknown size ▪ Good at learning from sequences and predicting rare events	▪ Speech recognition and sentence classification – understanding the semantic meaning of a sentence from keywords and context ▪ Predictive maintenance of infrastructure ▪ Predicting rare events, such as flooding, based on sensory data ▪ Anomaly detection based on expected configuration of a network

Table 1: Selected neural network techniques
Source: Cambridge Consultants

1.2.1 Aficionado: LSTM and RNN to classify musical genre

Recurrent Neural Networks (RNN) and Long-Short Term Memory (LSTM) are both classes of neural network that use directionality between nodes to handle data that includes a temporal element.

Cambridge Consultants has developed a demonstration of these techniques that is able to classify music by genre. This could be equally applicable to detecting faults in an industrial system or rapidly assessing patient health from sensor waveforms.

Our ***Aficionado*** demonstration was tested by having a pianist play a variety of music – covering baroque, classical, ragtime and jazz genres – in a live demonstration. Applying complex algorithms, the system then searched for different influences and assessed the likely genre in real time. It overwhelmingly outperformed conventional hand-coded software, painstakingly written by humans.

1.2.2 GANs to go beyond human vision

Technology has never enabled machines to interpret real-world scenes the way that humans can.

Generative Adversarial Networks (GAN) heralds a new era of sensing, realizing the potential of deep learning to advance practical AI and opening the way to countless applications for real-world businesses.

GANs combine two AI systems. One generates output, the other assesses or classifies the output of the first. This feedback loop allows both systems to improve.

Cambridge Consultants has demonstrated the potential of this technique through its SharpWave™ technology. The system has been trained to know what the real world looks like, with both optically perfect and distorted images. Then, when it is presented with something distorted, the technology can form a real-time judgement of the 'true' scene.

Beyond imaging, this technique can be applied to virtually any source of signal or data. It can be applied to audio signals, radar, LiDAR, or any system that needs to maintain the quality of sensor data, regardless of external factors.

The real-world perception delivered by SharpWave will make machines much more valuable thanks to enhanced decision-making and services. Empowering healthcare professionals with more accurate medical imaging, equipping manufacturing facilities with higher-grade sensing and predictive maintenance systems, helping farmers detect crop health, providing autonomous vehicles with more accurate and complete data… the potential is incalculable.

We are now applying a strain of this deep learning research to enhance the images surgeons view during medical procedures. Early results show real promise.

1.3 Increasing volume of data available

A critical enabler for AI has been the huge volume of data that is now routinely generated, harvested and stored by digital systems. These huge data sets can be used to train machines and allow them to continuously learn and improve. The increase in accessible data can be attributed to three factors:

- **Digitalization of services** – increasing volume of data collected
- **Cloud compute** – reducing cost and greater availability of data storage
- **Internet of Things** – more pervasive connectivity

The digitization of retail, media and telecommunications means that these areas are benefitting from AI first. The Internet of Things – wearables, autonomous vehicles and smart cities – will only expand the domains that AI can be applied to.

1.3.1 More data is being collected

The amount of data we are generating has been increasing rapidly since the dawn of the internet. Figure 3 shows the exponential increase in internet traffic over the last 25 years and the predicted increase up to 2021.

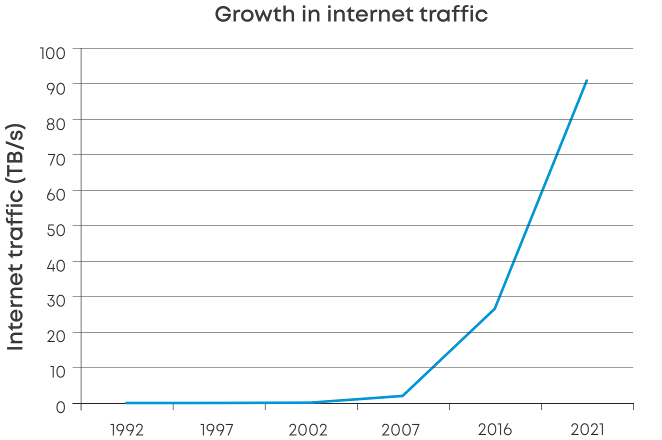


Figure 3: The amount of Internet traffic has grown at a phenomenal rate
Source: Cisco VNI[1], Cambridge Consultants

This increase in data has come from a combination of humans and machines generating and accessing data:

- The increased ownership of personal computing devices such as smartphones and tablets, shown in Figure 4, has meant that more and more data is generated by each user each day. Smartphone users provide data such as browsing history, images and videos which can be used by many online service providers such as Google and Facebook to train AI algorithms[2]
- Data is also generated through the ever increasing use of machine to machine connections. These are connections that do not require any human intervention in order to collect or use data. This data can be analyzed by AI algorithms, enabling them to monitor and learn from the physical world

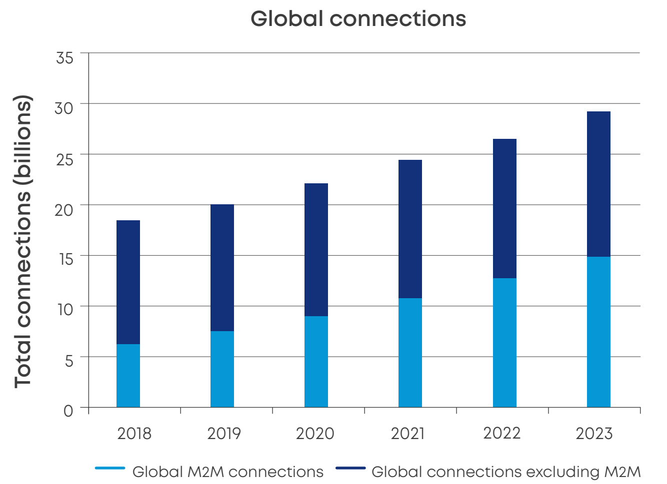


Figure 4: The total number of connected personal computing devices and machine to machine connections is expected to continue to grow rapidly
Source: Cisco[3], Global connections excluding M2M includes smartphones, PCs, tablets TVs and other personal devices

1.3.2 It is cheaper to get data to the cloud and store it there

Lots of data is being generated by users and machines, and at the same time the cost of storing all that data has reduced significantly.

Hard disk drive storage has greatly reduced in cost, as shown in Figure 5. This is largely down to improvements in technology allowing greater capacity hard drives to be manufactured more easily. There is a similar trend in the cost of flash storage which is used in almost all portable devices. This has meant that devices can have large amounts of local storage for the data they are generating and processing.

Increases in the availability and pervasiveness of fast connections to the internet also mean that large amounts of data can be transmitted to cloud servers and stored by cloud service providers. Figure 6 shows how the cost of storing data in the cloud has fallen, which in turn is reducing the barriers to organizations storing the large amounts of data they are gathering as a digital asset.

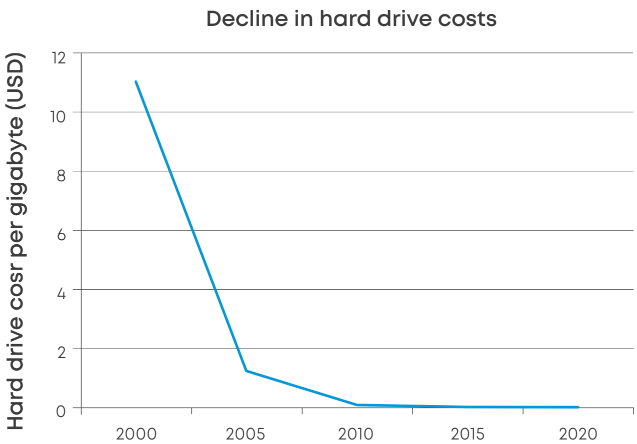


Figure 5: The cost of hard drive storage has fallen dramatically
Source: Statistic Brain[4], Cambridge Consultants

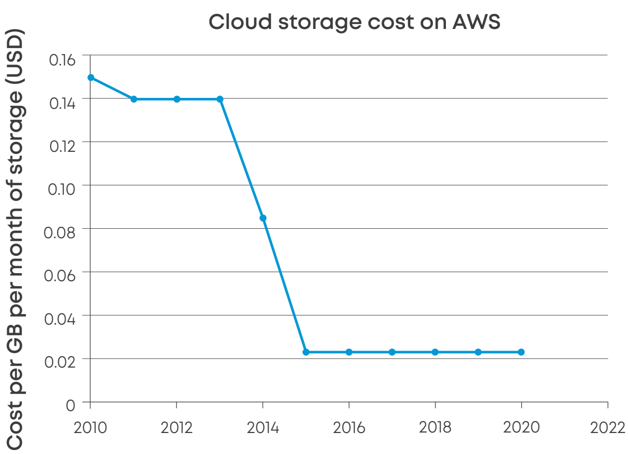


Figure 6: The cost of cloud storage has bottomed out at a low price
Source: Amazon Web Services data[5] accessed via Wayback machine

1.3.3 Connectivity is faster and more pervasive

The increasing ability to connect devices to the internet is driving the ability not only to collect more data but also to provide control of those devices by AI algorithms. Connectivity is becoming faster and more widely available. The evolution of mobile technologies from the first 2G digital systems of the early 1990s through to today's 4G and 5G has provided a rapid increase in the data rate available to mobile devices. Figure 7 shows how this trend is expected to continue across the globe. Mobile connectivity enables devices to send and receive data while on-the-move or in locations which are difficult to access with wires, but still most of the world's data is transmitted through fixed connections, thanks in part to WiFi offload.[6] Residential and commercial premises have enjoyed higher data rates through their fixed broadband connections, and similarly to mobile connections, this is expected to continue as seen in Figure 8.

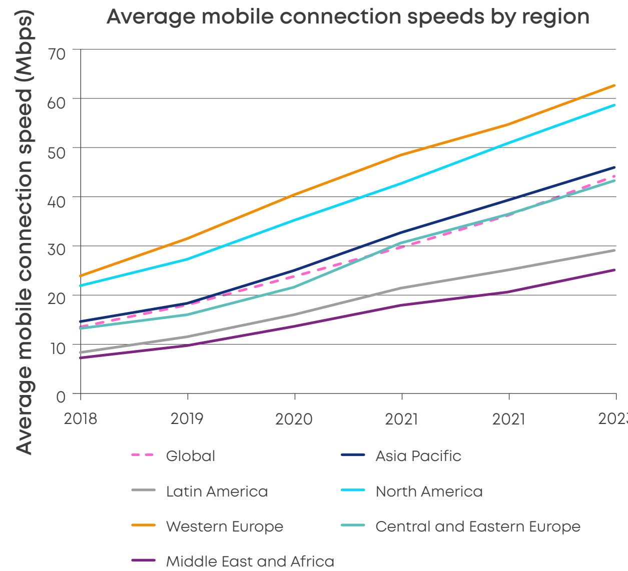


Figure 7: Average mobile connection speeds by region
Source: Cisco[7]

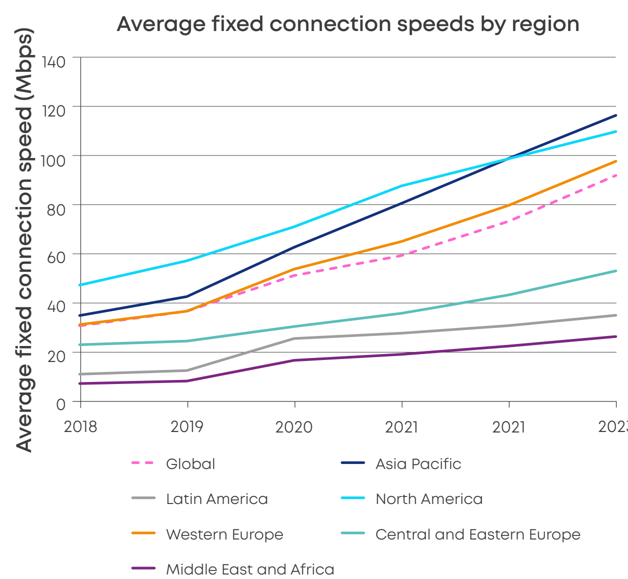


Figure 8: The average speed of fixed broadband is expected to continue to increase globally
Source: Cisco[8]

1.4 Increasing compute power

The evolution of AI techniques and the increased gathering of data require an increasing ability to perform computations. Without a corresponding increase in the availability of compute power, AI would not be able to take advantage of these.

There are two areas where this increase in computing power has enabled AI to progress. The first is an increase in power of Graphical Processing Units (GPUs) which are used by servers to undertake machine learning. The second is the increase in specification of personal computing devices which provide more local processing of information and implementation of AI models without the reliance on a connection to a cloud server.

This increase in processing power is underpinned by the continuing advances in semiconductor technology and the shrinking size of transistors within processors.

As shown in Figure 9, the size of transistors has been reducing year-on-year. TSMC announced in early 2019 that its 5nm technology was close to being ready, with volume production expected in 2020[9]. This will allow more transistors to be placed on a single die improving the processing speed of chips. A note of caution is required here: node size has traditionally been used to denote the size of transistors inside a CPU, with smaller node sizes being more desirable as they correspond to more powerful CPUs. However, as node sizes have shrunk to the width of tens of atoms, feature geometries have changed to address the manufacturing and performance limitations that arise, so there is no longer a direct correlation between node size and chip performance. As a result, it is possible that a 14 nm chip from one manufacturer will be comparable to a so-called 7 nm chip from another vendor.

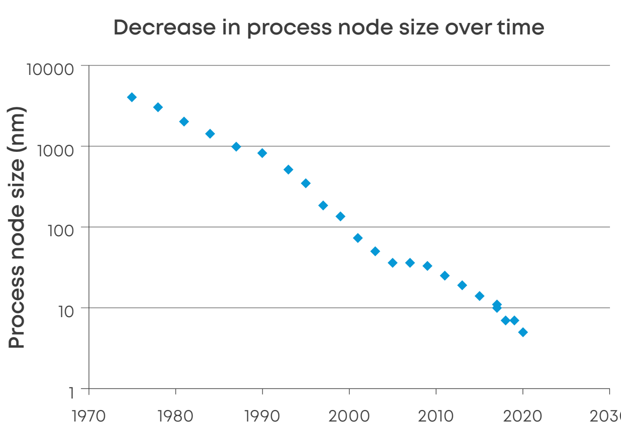
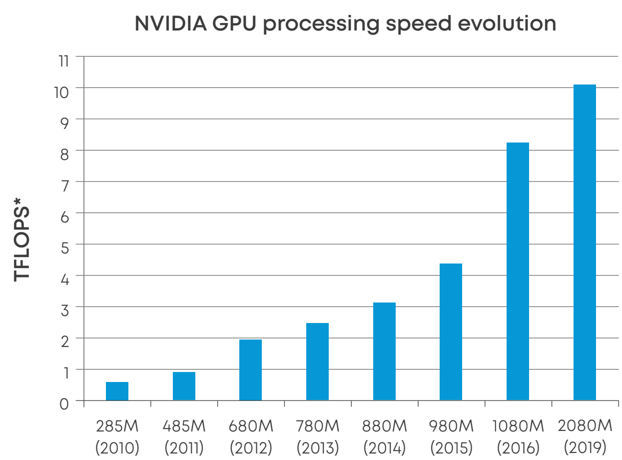


Figure 9: The size of transistors within integrated circuits continues to fall
Source: ExtremeTech[10], Huawei[11] AnandTech[12]

1.4.1 Faster GPUs

Increasing computing power has become available through the use of GPUs for machine learning. These were initially designed to carry out millions of calculations in parallel to meet the demand for fast-moving graphics in video games. Their massively parallel architectures are ideally suited to performing the calculations required by machine learning algorithms, and so GPUs have been harnessed to enable machine learning algorithms to run quickly and efficiently. The company best known for its GPUs, Nvidia, has released increasingly powerful GPUs, as shown in Figure 10, increasing the power available for machine learning applications. Nvidia has also released its Volta technology which it claims can achieve 125 TFLOPS for training and inference applications[13].



*trillions of floating point operations per second, a measure of computing speed

Figure 10: The processing speed of GPUs has grown rapidly
Source: Nvidia, GTX range[14], [15]

1.4.2 Higher specification personal devices

The increasing processing power available to personal computing devices enables more local processing of information using AI algorithms. For example, voice recognition is becoming more effective on personal devices. Increasingly, personal computing device chipsets will contain AI hardware accelerators to specifically target improved execution of AI algorithms, rather than just an increased amount of general-purpose compute. This will allow increasingly complex AI algorithms to be run on small, mobile devices incentivizing app developers to push the limits of AI at the edge by utilizing these resources. Figure 11 shows the focus that is being placed on AI execution by mobile SoC developers. The SoCs include the latest offerings from Hisilicon (Huawei), MediaTek, Qualcomm and Samsung and have been scored according to an AI benchmark.[16] The selected benchmark consists of 21 Computer Vision and AI Tests performed by Neural Networks including tasks such as object recognition and classification, image deblurring and playing Atari games.[17]

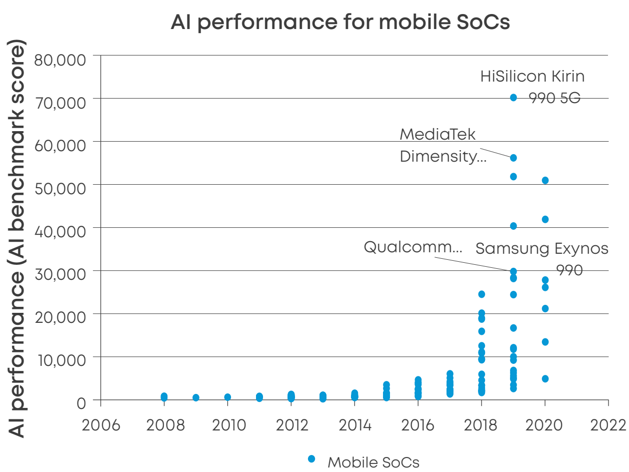


Figure 11: The AI performance of mobile SoCs has increased dramatically over the past 5 years
Source: AI-Benchmark[18]

1.5 Where it is going?

A large part of the innovation that we will see in the coming years will arise from the interplay of AI with other technological and commercial trends. For example, there are many synergies between AI, 5G, and the Internet of Things. The adoption of network function virtualization (NFV) and distributed computing (edge, cloud and fog computing) will support the transition to network architectures with suitably low latencies and low-cost processing power. We discuss some further trends below.

Neural processors

Neural or AI-specific processors emerged in 2017 with announcements from Google[19], Huawei[20], IBM[21], Intel[22], Nvidia[23], Samsung[24] and Tesla[25] about development work or products on the way. All high end smartphones now contain SoCs that are optimized for AI.

Until now GPUs, originally designed for rapid processing of images and video, have delivered massive gains in performance in AI tasks. Bespoke neural chips will continue to drive gains in processing performance and power efficiency through designs that are better optimized for neural networks and other machine learning tasks.

These developments will support advances both in training and developing AI systems and also deploying AI processes in user devices. For example, the type of neural processor embedded in Apple's latest iPhone enables more voice and image recognition tasks to be carried out locally. This type of system architecture reduces latency and power consumption, reduces reliance on connectivity and offers benefits for privacy and security since data does not need to leave the device.

Blockchain

Blockchain technology will have a role to play in expanding the application of AI. Blockchains can be used to provide verification and audit, to data and to models. Some proponents suggest that this will enable the sharing of training data, testing data and models as auditability, verification and protection can be provided. In turn this is expected to lead to better AI systems and applications.[26]

The combination of AI and blockchain also opens up the possibility of 'smart contracts' or AI Decentralized Autonomous Organizations (DAOs). These could be entirely autonomous software based legal entities encoded in blockchain and executed by AIs. Applications could range from self-executing stock trading to self-owned, autonomous taxis.

Quantum computing

The possibilities of AI together with quantum computing are only just beginning to be explored. Right now, vital progress is being stimulated by academic research, early-stage investment and the trailblazing work of the world's tech giants.

Quantum technology already enables atomic clocks, lasers, sophisticated encryption tools and more. But today's quantum tools tend to rely on bulk effects, controlling billions of fundamental particles, rather than the precision of single particles. Tomorrow's quantum technologies, with precision control of single particles, promise nothing less than a new technology revolution.

AI will continue to deliver commercial value

The technological breakthroughs within AI over the last decade have driven value through the wide range of commercial use cases they have been applied to, across a range of different sectors. AI is a key technology in a large number of consumer applications, such as streaming service recommendation engines and smart speakers, and it has also become pervasive in company activities through applications such as virtual customer service assistants and automated operations.

This has meant that AI, alongside other technologies, has begun to disrupt a range of markets that, historically, have not been technology focused. It is likely to reach a point in the not too distance future where AI technology will become an essential part of how we conduct business, and AI will be required for companies to operate on a level playing field, across a broad range of markets. Companies that do not consider the effects of AI on their business model, regardless of the current state of their industry, risk falling behind. Those that can identify novel applications of AI in their sector first, and execute quickly, have the potential to capture the substantial rewards that will come with the new capabilities AI technologies are enabling.

To discuss advances in AI and ML contact:

Tim Ensor Director of Artificial Intelligence
tim.ensor@cambridgeconsultants.com

Michal Gabrielczyk Head of Edge AI
michal.gabrielczyk @cambridgeconsultants.com

References

1 https://cyrekdigital.com/pl/blog/content-marketing-trendy-na-rok-2019/white-paper-c11-741490.pdf

2 https://techcrunch.com/2016/04/30/why-image-recognition-is-about-to-transform-business/

3 https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

4 http://www.statisticbrain.com/average-cost-of-hard-drive-storage/

5 https://aws.amazon.com/s3/pricing/

6 https://blogs.cisco.com/sp/wi-fi-wanderlust-whats-happening-with-the-other-mobile-technology

7 https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

8 https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

9 https://www.tsmc.com/english/dedicatedFoundry/technology/5nm.htm

10 https://www.extremetech.com/wp-content/uploads/2013/07/transistor-gate-length-scaling.jpg

11 https://consumer.huawei.com/en/campaign/kirin980/

12 https://www.anandtech.com/show/15219/early-tsmc-5nm-test-chip-yields-80-hvm-coming-in-h1-2020

13 https://www.nvidia.com/en-us/data-center/tensorcore/, reviewed 29 April 2020

14 https://www.nvidia.com

15 https://www.techpowerup.com/gpu-specs/geforce-gtx-1080-ti.c2877, reviewed 29 April 2020

16 http://ai-benchmark.com/ranking_processors.html

17 http://ai-benchmark.com/tests.html

18 http://ai-benchmark.com/index.html

19 https://cloud.google.com/blog/big-data/2017/05/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

20 http://consumer.huawei.com/en/press/news/2017/ifa2017-kirin970/

21 http://fortune.com/2017/12/05/ibm-ai-chip-nvidia/

22 https://www.theverge.com/circuitbreaker/2017/10/17/16488414/intel-ai-chips-nervana-neural-network-processor-nnp

23 http://fortune.com/2017/12/05/ibm-ai-chip-nvidia/

24 https://www.pocket-lint.com/phones/news/samsung/142673-samsung-galaxy-s9-to-get-dedicated-neural-engine-for-ai-tasks

25 https://www.theverge.com/2017/12/8/16750560/tesla-custom-ai-chips-hardware

26 http://dataconomy.com/2016/12/blockchains-for-artificial-intelligence/

About Cambridge Consultants

Cambridge Consultants has an exceptional combination of people, processes, facilities and track record. Brought together, this enables innovative product and services development and insightful technology consulting. We work with companies globally to help them manage the business impact of the changing technology landscape.

We're not content to deliver business strategy based on target specifications, published reports or hype. We pride ourselves on creating real value for clients by combining commercial insight with engineering rigor. We work with some of the world's largest blue-chip companies as well as with innovative start-ups that want to change the status quo fast.

With a team of around 800 staff in Cambridge (UK), Boston, San Francisco and Seattle (USA), Singapore and Tokyo, we have all the in-house skills needed to help you – from creating innovative concepts right the way through to taking your product into manufacturing. Most of our projects deliver prototype hardware or software and trial production batches. Equally, our technology strategy consultants can help you to optimize your product portfolio and technology roadmap, investigate new opportunities or refine your operations.

For further information or to discuss your requirements, please contact:

Tim Ensor, Director of Artificial Intelligence
tim.ensor@cambridgeconsultants.com

cambridge consultants
Part of Capgemini Invent

UK — USA — SINGAPORE — JAPAN

www.cambridgeconsultants.com

Cambridge Consultants is part of Capgemini Invent, the innovation, consulting and transformation brand of the Capgemini Group. www.capgemini.com

Lit-WHI-044 v0.3 '2022_The AI renaissance_Whitepaper'