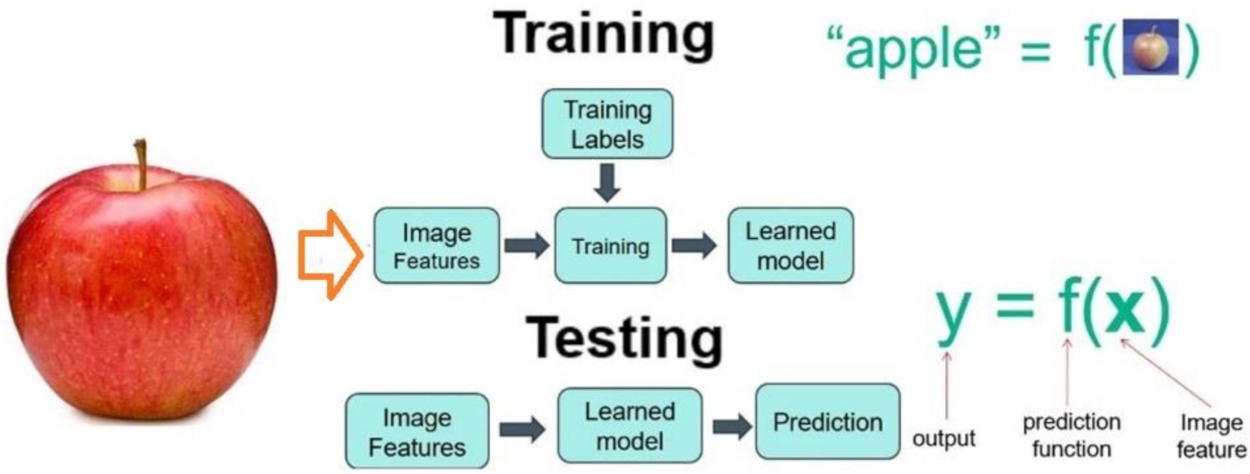
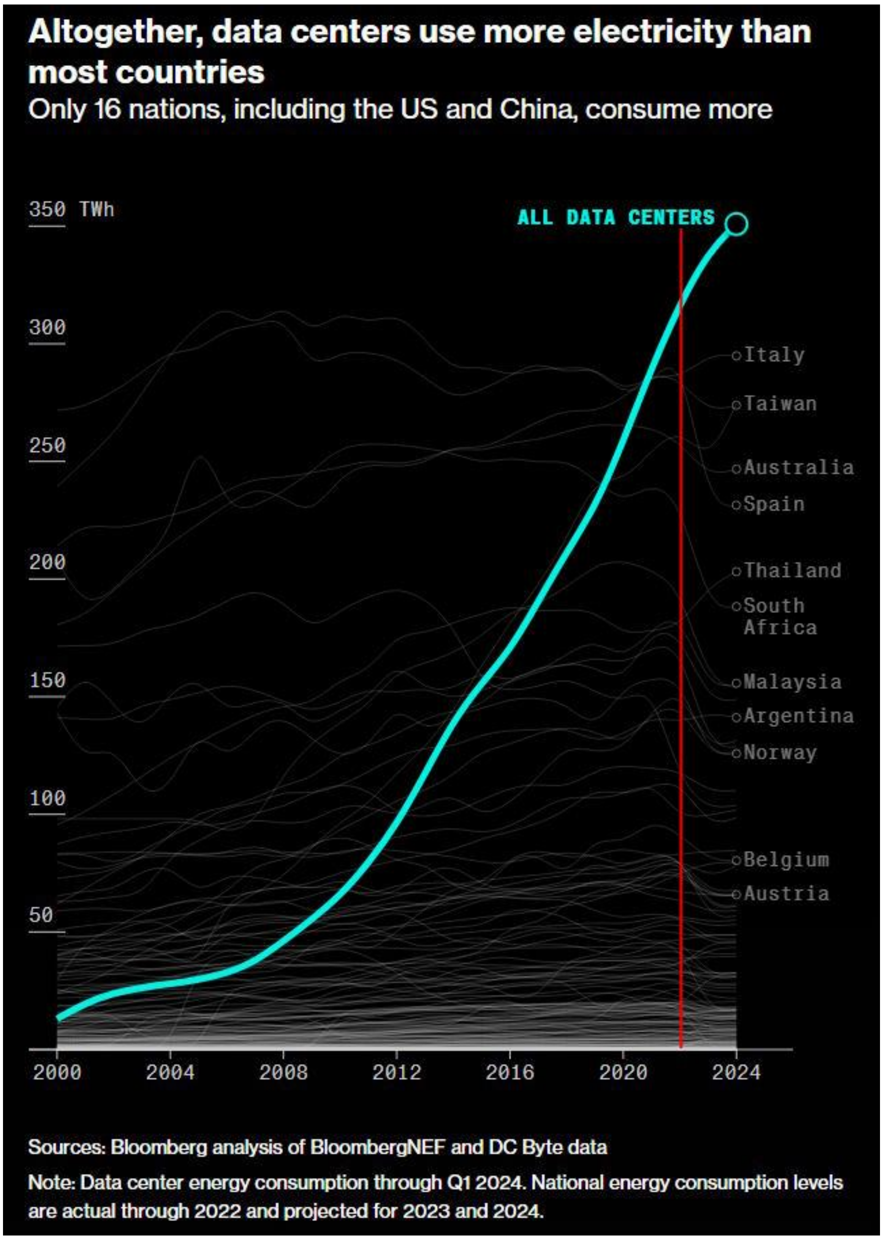
Sustainable AI

Contributed by Arvind Bhagwath – Founder of microML.IN (https://microml.in)



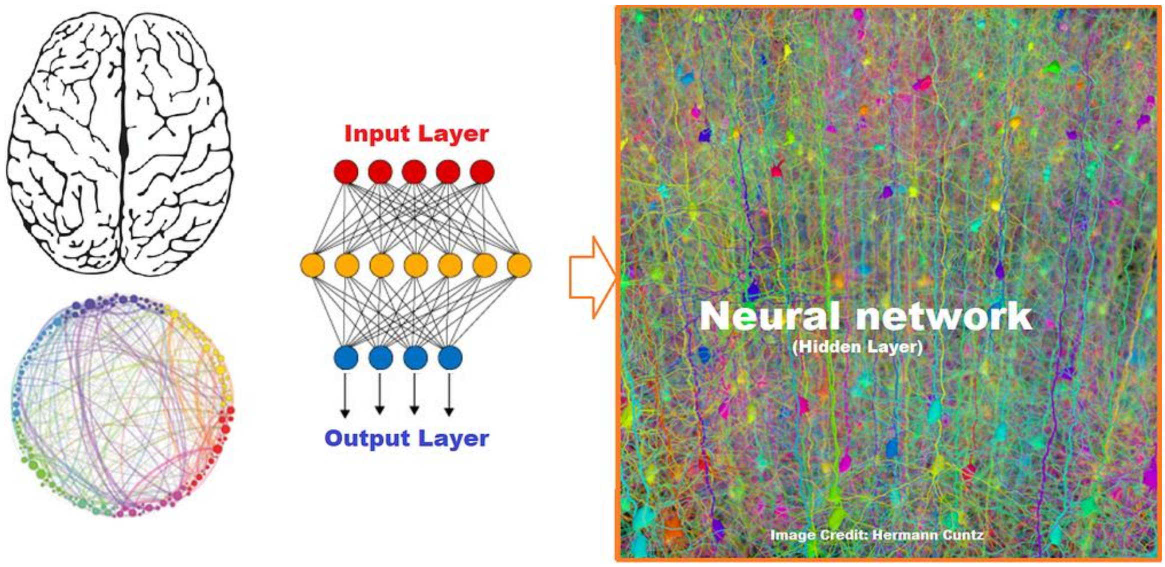
The rush to implement **Artificial Intelligence (AI)** in industries has disrupted long-established sustainability plans. Few research firms estimated that AI accelerators will **consume 2318 TWh of global power** in next four years, reaching a **1.5% share of global electricity consumption** and raises big concern on **Artificial Intelligence (AI) data centers** contributing to **global warming.**

For the fact, our Human brain uses **just 20 watts** of energy, however as per current data available in internet by research firms, Human created AI systems currently consume a lot of power and altogether, data centers use more electricity than most of countries.



To reduce this high energy use and carbon emissions, using more efficient smaller Machine learning processes and change in current AI System architecture is key for developing sustainable AI Systems. We can reduce the size of many AI models by an order of magnitude if we design the AI architecture just the way like how our Human brain works.

For the fact, memories in Human Brain aren't stored in just one part of the brain. Different types of sensory data(memory) are stored across different, interconnected brain regions. These data stored in networks may be used for predictive modelling, decision control that normally come out as emotional output from Brain. As a part of evolution, self-learning resulting from experience occurs and each time data would be stored within networks, which later is used to derive conclusions from these complex set of information stored in **neural networks**.



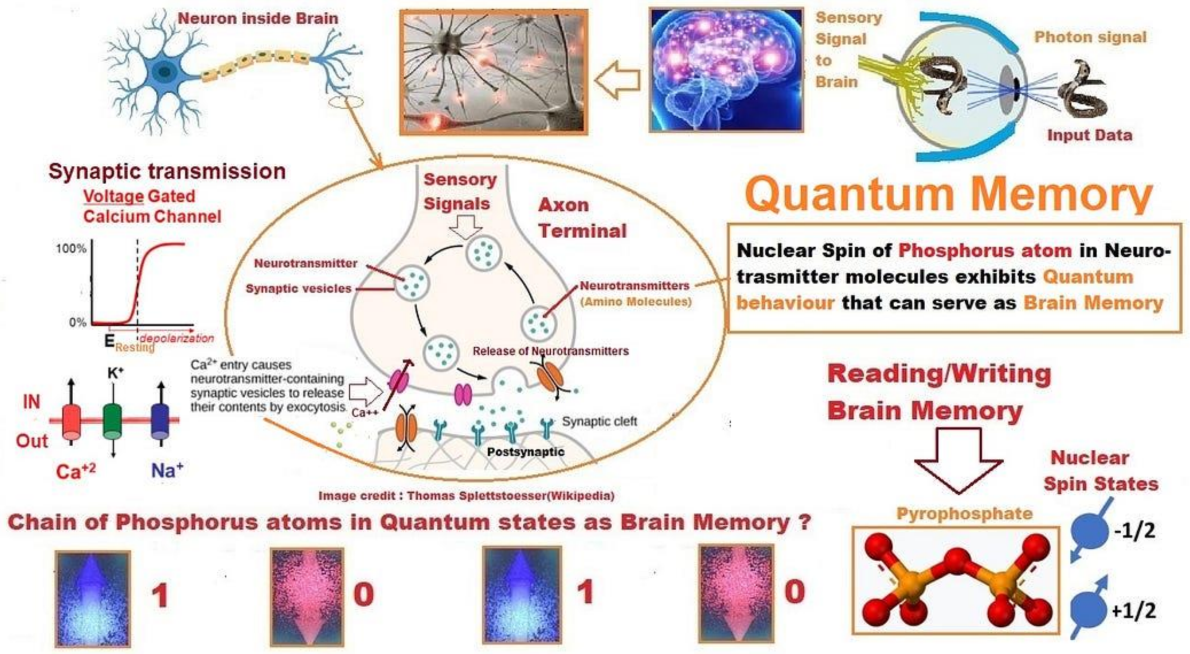
Quantum Memory for Sustainable AI Systems?

Memory is the faculty of the brain by which data or information is encoded, stored, and retrieved when needed. Brain is made of neurons and a single neuron is connected to many other neurons and the total number of neurons and connections is extensive. **Human brain** contains almost **100 billion neurons.** Each neuron has on average **7,000 synaptic**

connections to other neurons. It has been estimated that the brain of a three-year-old child has about 1015 synapses.

Back in 1998, a physicist Bruce Kane from Australia had suggested that **phosphorus atoms** embedded in silicon would be the ideal way to store and manipulate **quantum information** that later put forward a design for a **Kane quantum computer**. The Kane computer is based on an array of individual phosphorus donor atoms embedded in a pure silicon lattice. Both the nuclear spins of the donors and the spins of the donor electrons participate in the computation.

With current deeper understanding on behaviour of **Phosphorus atoms** and **Quantum mechanism**, possibility of quantum processing with nuclear spins could be one of main reason **on how brain stores and reads memory** needs to be explored. As we are aware, when a memory is created, information flows from the cortex, the part of the brain rich in nerve cells, to **the hippocampus**, the central switching point for memories in the brain. The information flows in the opposite direction when we retrieve a memory. Brain memories underlie our ability to learn, to tell stories, even to recognize each other's. Researchers have been able to trace memory down to the structural and even the molecular level in recent years and this should be possible only if they show Quantum behaviour and hence role of Phosphorus atoms play an important role. Phosphorus has only one naturally occurring stable isotope, **^{31}P, which like ^{1}H has spin = ½** and thus two discrete energy states which becomes ideal candidate to store memory at molecular level something represented in image below:



In the **nervous system**, a synapse is a structure that permits a **neuron** to pass an electrical or chemical signal to another neuron or to the target effector cell. Synapses are essential to the transmission of nervous impulses from one neuron to another. As shown in image above, **communication** at chemical synapses requires **release of neurotransmitters**. When the presynaptic membrane is depolarized, **voltage-gated Ca2+** channels open and allow Ca2+ to enter the cell. The calcium entry causes **synaptic vesicles** to **fuse (contract)** and thus helps in releasing neurotransmitter molecules into the **synaptic cleft**. The neurotransmitter diffuses across the synaptic cleft and binds to ligand-gated ion channels in the postsynaptic membrane, resulting in a localized depolarization or hyperpolarization of the postsynaptic neuron.

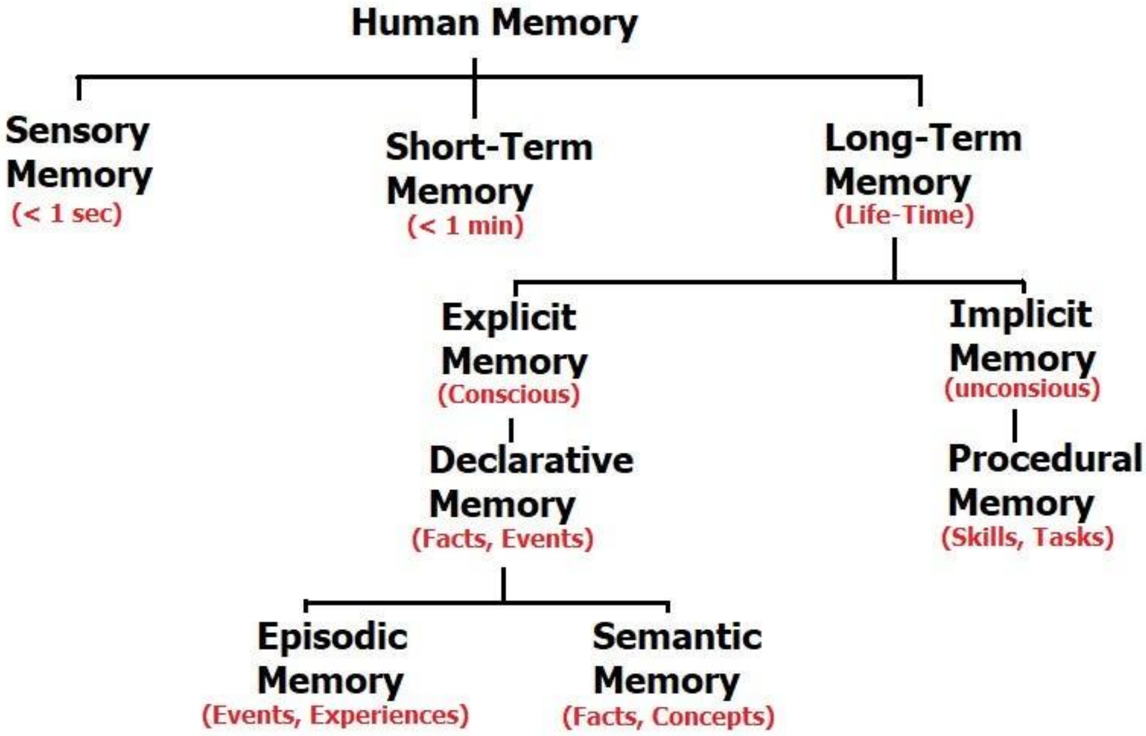
Neurons are specialized to pass signals to individual target cells, and synapses are the means by which they do so. Both the presynaptic and postsynaptic sites contain extensive arrays of molecular machinery that link the two membranes together and carry out the signalling process/transfer of data with **release of neurotransmitters**. With availability of **Phosphorus atoms** in **Neurotransmitter molecules**, Phosphate bonds might exhibit **Quantum behaviour** which Brain might be using to store memory and this mechanism could be almost like how qubit operations are possible by manipulating the electric fields in Quantum computers using **Phosphorus atoms.**

Using lesser variants in current Machine Learning models?

Current centralized Machine learning LLM models generates lot of variants which might actually not be required at all for normal use and training one such single machine learning (ML) model can equal the combined carbon emissions of multiple cars. We need to focus on sustainable AI systems on reducing the negative environmental impacts associated with current design of artificial intelligence technologies such as creating independent Machine Learning models on standalone microchips instead of depending on centralized models for every request. microML.IN focuses in developing several such standalone modules on low cost and low powered microcontrollers for actual AI requirement of any such industries.

Next step, we can further reduce the size of unnecessary dataset used to train a model to minimize energy use and carbon emissions involved leveraging the idea on how Human brain works.

Memory in Human Brain is often understood as an informational processing system with explicit and implicit functioning that is made up of a **sensory memory, short-term memory** and **long-term memory.**



Sensory memory holds information, derived from the senses, less than one second after an item is perceived. Short-term memory is also known as working memory. Short-term memory allows recall for a period of several seconds to a minute without rehearsal. For example, recalling a ten-digit - telephone number is a kind of Short-term memory. In Long-term memory, declarative, or explicit memory is the conscious storage and non-declarative or implicit memory is the unconscious storage and recollection of information. Implicit memory uses past experiences to remember things without thinking about them and it is used to perform certain procedural tasks, such as driving, riding a bike, playing a musical instrument etc.

Long-term memory can store much larger quantities of information for potentially unlimited duration (sometimes a whole life span). Its capacity is immeasurable. It is maintained by more stable and permanent changes in neural connections widely spread throughout the brain.

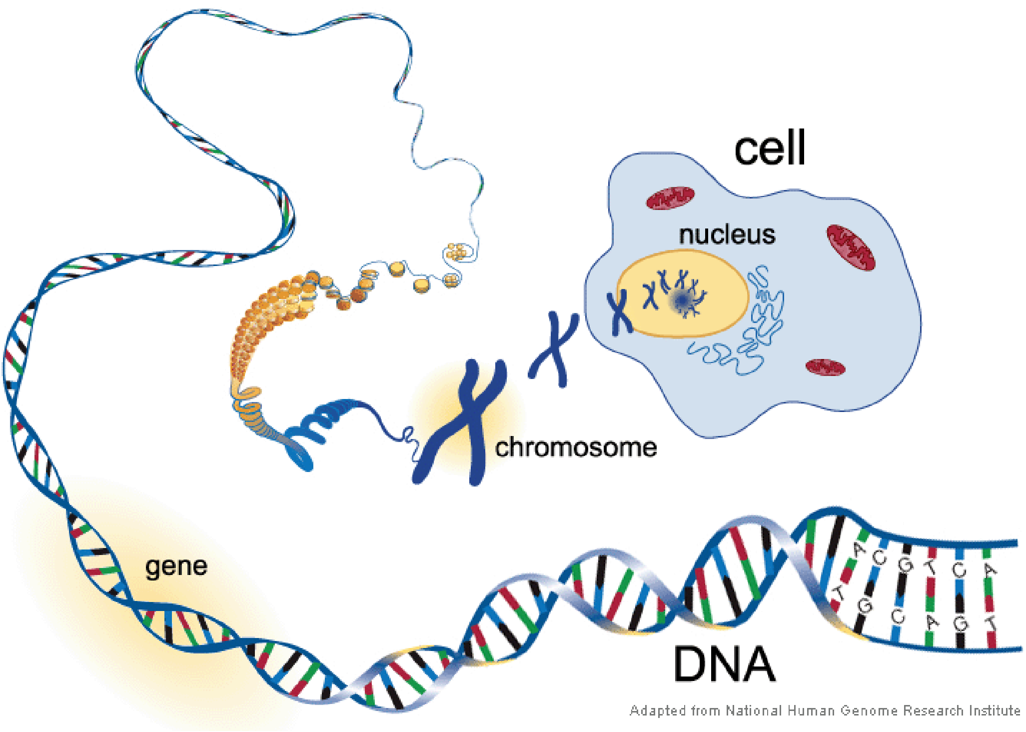
We can leverage this kind of storage mechanism instead of pushing all the data to one centralized data center that currently consumes huge energy. The rise of generative AI and surging GPU shipments is causing data centers to scale to thousands of accelerators, shifting the emphasis to power as a mission-critical problem to solve.

Role of Phosphorus in Genetic or DNA Memory?

Several genes, proteins and enzymes have been extensively researched for their association with Long-term memory. Long-term memory, unlike short-term memory, is dependent upon the synthesis of new proteins.

For instance, Rats when exposed to an intense learning event may retain a life-long memory of the event. When such an exposure was experimentally applied, more than 5,000 differently methylated DNA regions appeared in the hippocampus neuronal genome of the rats at one and at 24 hours after training. These alterations in methylation pattern occurred at many genes that were down-regulated, often due to the formation of new 5-methylcytosine sites in CpG rich regions of the genome.

Genome size is the total number of the DNA base pairs in one copy of a haploid genome. Genome size varies widely across species. Invertebrates have small genomes; this is also correlated to a small number of transposable elements. In humans, the nuclear genome comprises approximately 3.2 billion nucleotides of DNA, divided into 24 linear molecules, the shortest 50 000 000 nucleotides in length and the longest 260 000 000 nucleotides, each contained in a different chromosome.

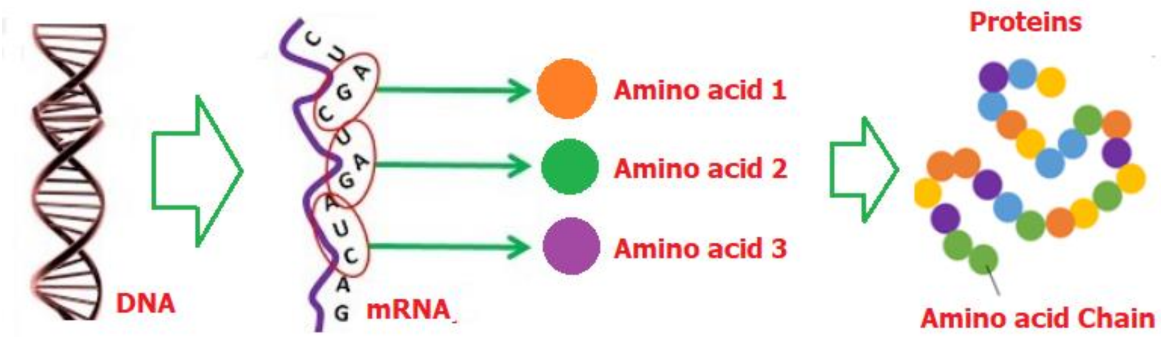


Unlike Neural memory, **DNA** or **Genetic memory** is straight forward. Genetic memory is a memory present at birth that exists in the absence of sensory experience and is incorporated into the genome over long spans of time.

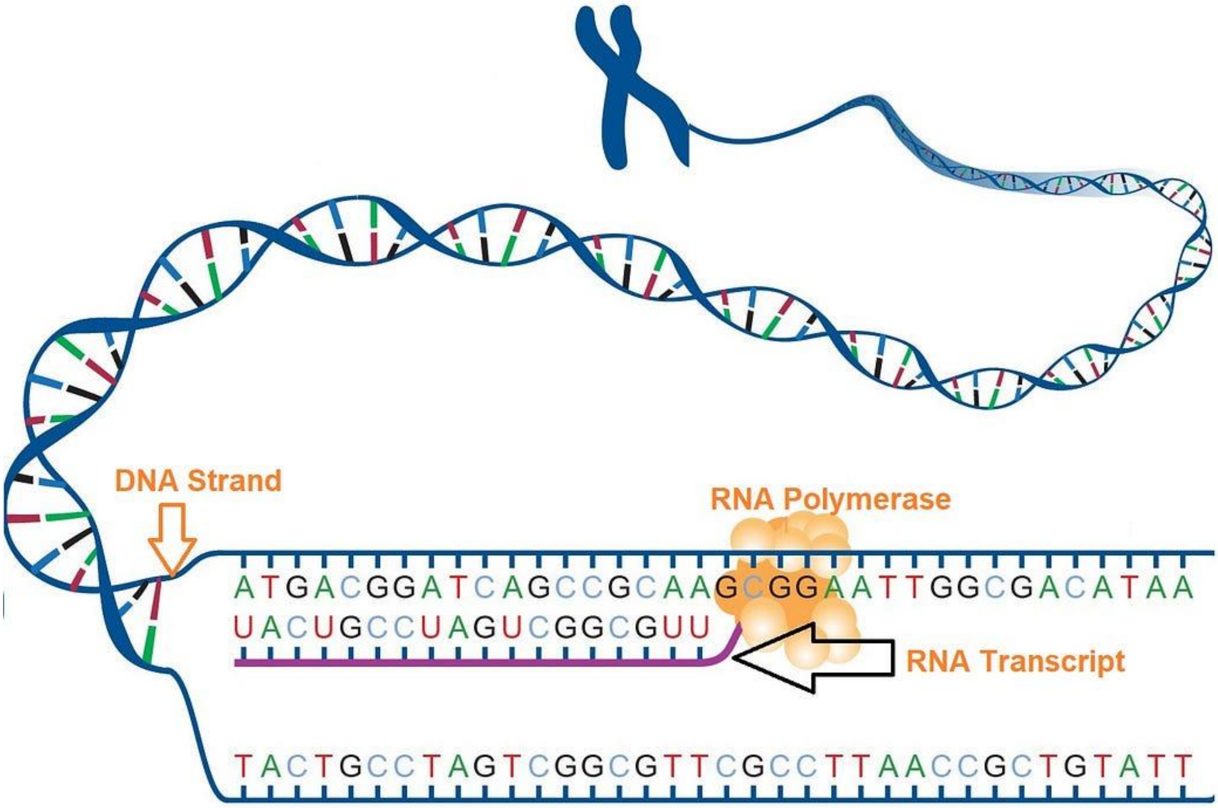
Phosphorus here too plays a very important role in building up this memory. As we are aware, **DNA** and **RNA** structures here are connected by **phosphorous bonds** and this DNA memory is like the flash drive that is used to store biological data almost like computers, which use 0 and 1. DNA uses A, C, G and U/T (the 'nucleotides', 'nucleosides' or 'bases').

This memory got developed with several million years of evolution based on events recorded in DNA which is just stored in code of four chemical bases **adenine (A), guanine (G), cytosine ©,** and **thymine (T)**.

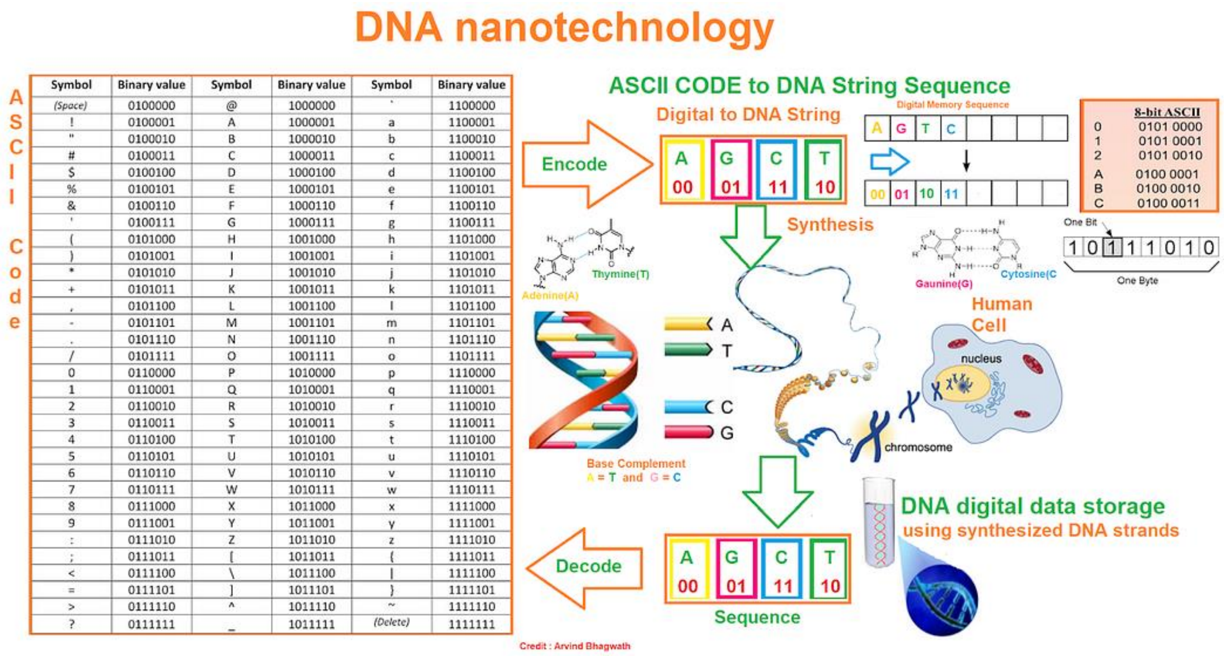
A, C, G and **U/T** (the 'nucleotides', 'nucleosides' or 'bases') helps to build different Amino Acids. **Proteins** are molecules made of amino acids and a gene is a segment of a DNA molecule that contains the instructions needed to make a **unique protein.**



Example: If you cut the leg off a salamander, it will grow back. Similarly in roses, they have thorns to protect them from being eaten by animals. Extra protection functions like Thorns for Roses or Regeneration is triggered by DNA from earlier events of damage. This is how memory from these phosphate bonds helps for survival and growth.



The idea of DNA data storage is not merely theoretical. Scientists have mastered to decode DNA memory that is helping for several reverse engineering projects like development of life saving vaccines. In vaccine production process, DNA code has to be uploaded to a DNA printer which then converts the bytes to actual DNA molecules. CRISPR DNA-editing technology is available now to record images of a human hand into the genome. It was showcased by Church's group at Harvard for E. coli, which were read out with higher than 90 percent accuracy. Researchers at the University of Washington and Microsoft Research have developed a fully automated system for writing, storing and reading data encoded in DNA. The genetic material DNA has garnered considerable interest as a medium for digital information storage because its density and durability.



Conclusion:

To meet the current demand of AI systems, businesses are investing billions of dollars to grow GPU capacity which is definitely not sustainable and not green. The rush to implement AI in business has disrupted long-established sustainability plans with growing cost of electricity and challenges facing energy grids. In my view, understanding why human brain uses just 20 watt of power and performs better than AI systems should give new definition on how memory is processed, defined and stored. Quantum way to store the data and processing needs to be explored. It posits that quantum-mechanical phenomena, such as entanglement and superposition of **Phosphorus atoms** may play an important role in storing and processing the data in much sustainable way.